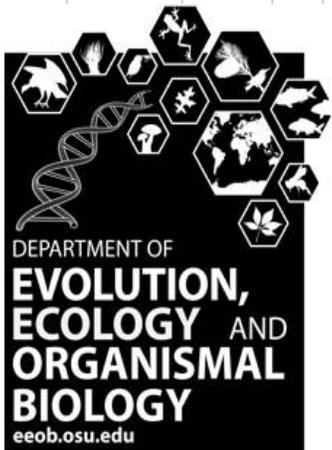
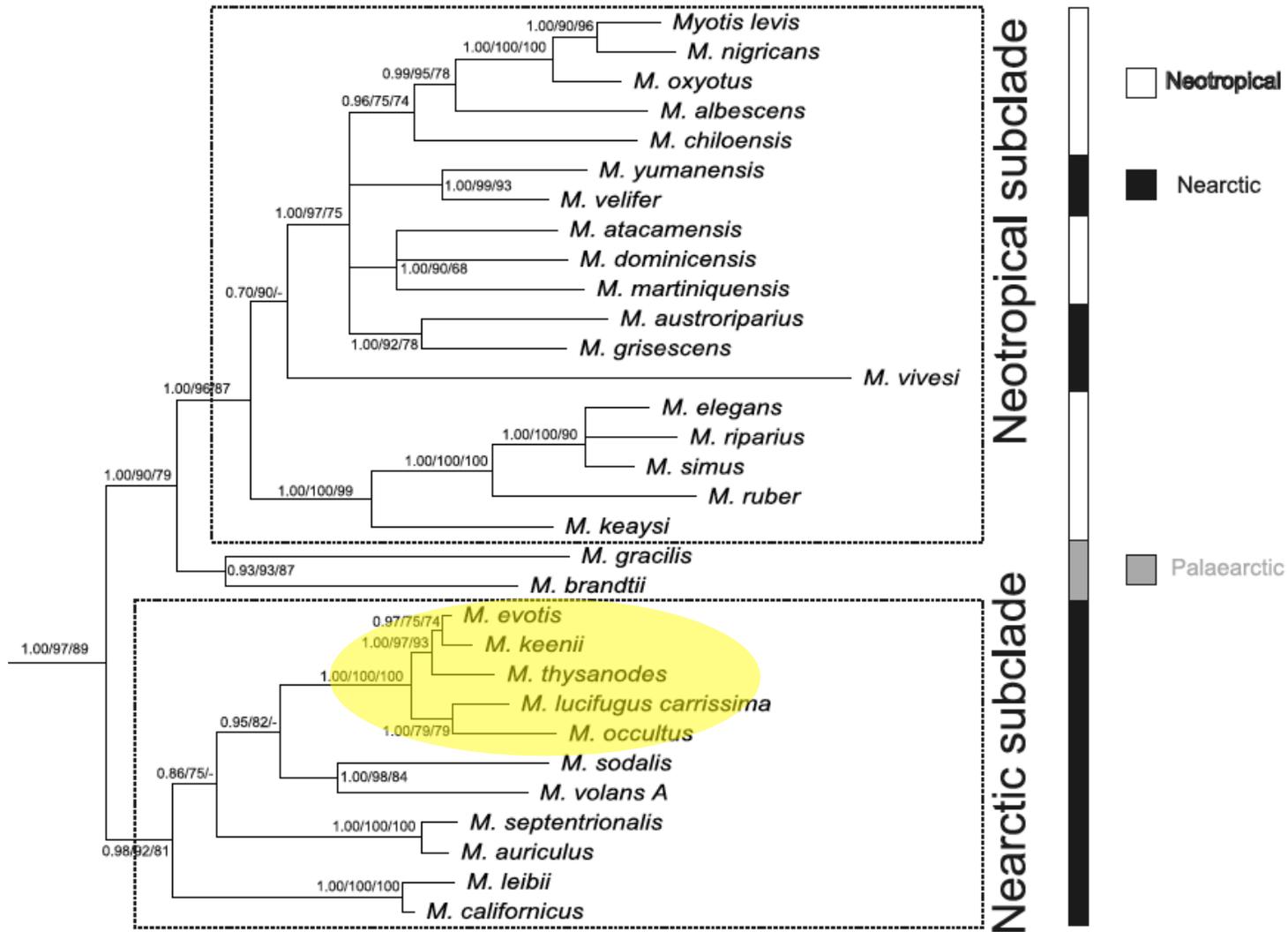


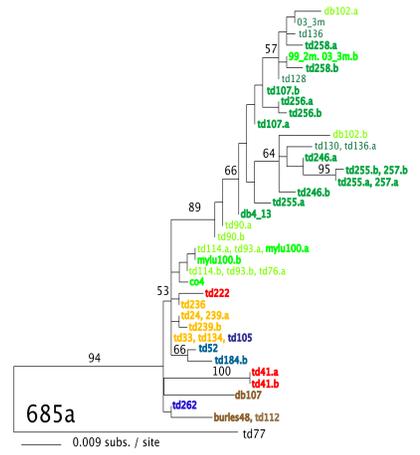
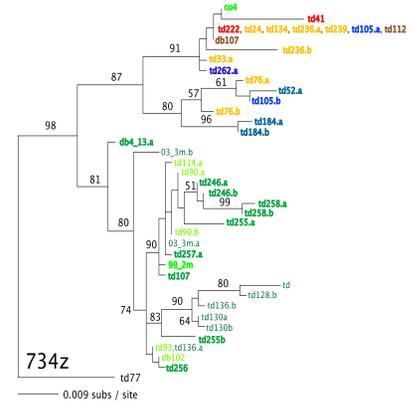
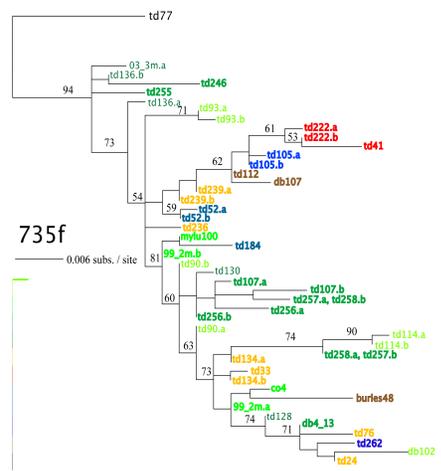
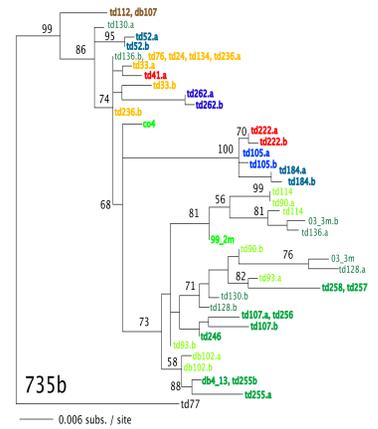
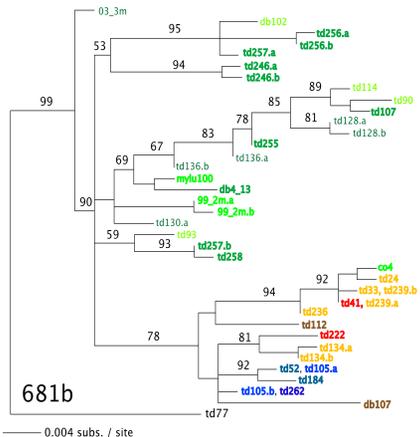
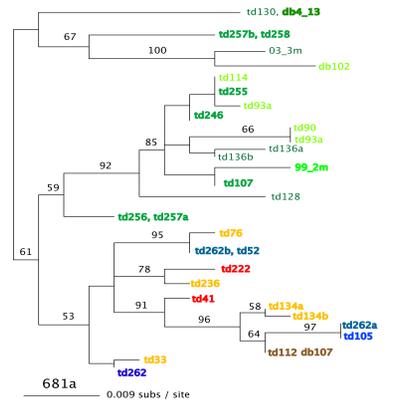
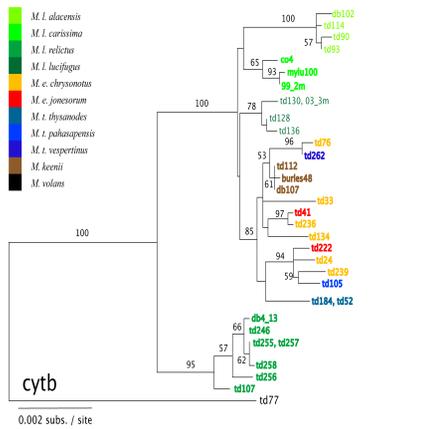
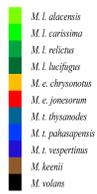
# Model selection as a tool for inference in phylogeographic research.



@bryanccarstens  
[carstens.12@osu.edu](mailto:carstens.12@osu.edu)  
<http://carstenslab.org.ohio-state.edu>







*Myotis lucifugus* (*alascensis*, *carissima*, *lucifugus*, *relictus*)

*Myotis evotis* (*evotis*, *pacificus*, *jonesorum*, *chryonotus*)

*Myotis thysanodes* (*aztecus*, *thysanodes*, *pahasapensis*, *vespertinus*)

*Myotis keenii*

*Syst. Biol.* 59(4):400–414, 2010

© The Author(s) 2010. Published by Oxford University Press on behalf of Society of Systematic Biologists.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.5>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI:10.1093/sysbio/syq024

Advance Access publication on May 24, 2010

## Species Delimitation Using a Combined Coalescent and Information-Theoretic Approach: An Example from North American *Myotis* Bats

BRYAN C. CARSTENS<sup>1,\*</sup> AND TANYA A. DEWEY<sup>2</sup>

<sup>1</sup>Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70808, USA; and <sup>2</sup>Department of Ecology and Evolutionary Biology, Museum of Zoology, University of Michigan, 1109 Geddes Avenue, Ann Arbor, MI 48109-1079, USA;

\*Correspondence to be sent to: Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70808, USA;  
E-mail: [carstens@lsu.edu](mailto:carstens@lsu.edu).

Received 20 May 2009; reviews returned 18 August 2009; accepted 11 December 2009

Associate Editor: Marshal Hedin

P. Myers



S. Altenbach

Phylogenetics

**STEM: species tree estimation using maximum likelihood for gene trees under coalescence**

Laura S. Kubatko<sup>1,\*</sup>, Bryan C. Carstens<sup>2</sup> and L. Lacey Knowles<sup>3</sup>

<sup>1</sup>Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, <sup>2</sup>Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803 and

<sup>3</sup>Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

Received on November 28, 2008; revised and accepted February 04, 2009

Associate Editor: Martin Bishop

$$L(S, \tau) = \prod_{j=1}^N f(g_j | S, \tau)$$

**STEM:** Analytical calculation of phylogeny under a coalescent model that accounts for the loss of ancestral polymorphism due to genetic drift.

**MOLECULAR ECOLOGY  
RESOURCES**

Molecular Ecology Resources (2011) 11, 473–480

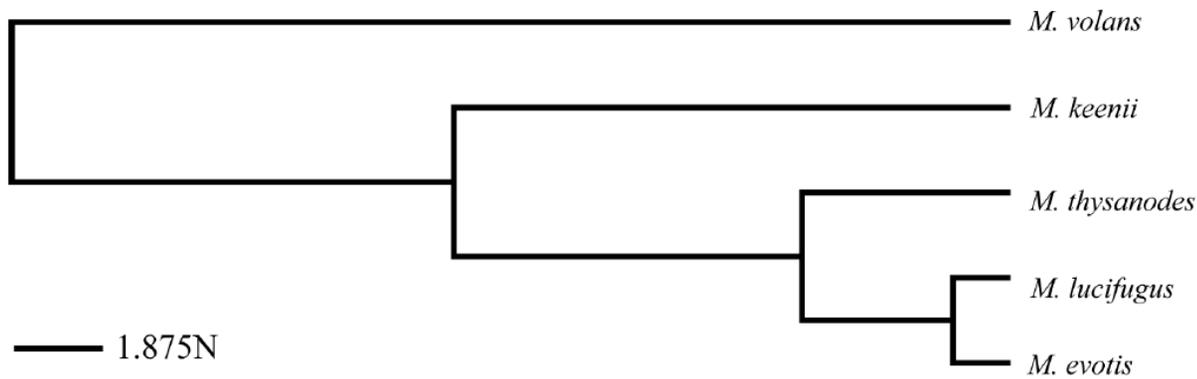
doi: 10.1111/j.1755-0998.2010.02947.x

**TECHNICAL ADVANCES**

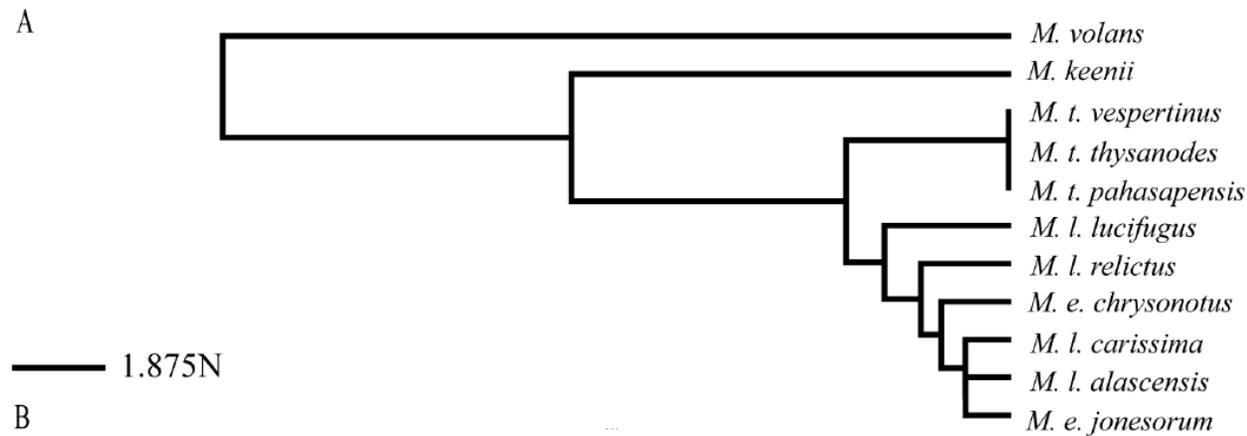
**SpedeSTEM: a rapid and accurate method for species delimitation**

DANIEL D. ENCE and BRYAN C. CARSTENS

Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA



$$\ln L = -664.695$$



$$\ln L = -570.533$$

- two extremes (species as OTUs, subspecies as OTUs)
- 148 other hierarchical permutations for these data . . .

Permutation	lnL	k	AIC	$\Delta_i$	L(Mjd)	wi
Mp_Mp_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar	-570.52259	8	1157.06318	0	1	0.68478536
Mp_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar	-570.52528	9	1159.06556	1.999976	0.13532366	0.092687151
Mp_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar	-570.52568	9	1159.065136	1.999956	0.13541238	0.092679736
Mp_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar	-570.52575	9	1159.06515	1.99997	0.135339343	0.092678438
Mp_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar	-573.375457	7	1160.750814	3.683774	0.02519786	0.01717375
Mp_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar_Ar	-570.52561	10	1161.065122	3.999942	0.018316701	0.012540014
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-573.375395	8	1162.75079	5.68561	0.00339462	0.002334479
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-573.375435	8	1162.75087	5.68569	0.00339419	0.002334293
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-573.375442	8	1162.750884	5.685704	0.003394143	0.00233426
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-573.375427	9	1164.750854	7.685674	0.00049361	0.000314564
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-576.69759	6	1165.397598	8.332418	0.00054859	0.000164752
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-579.50394	7	1173.362188	16.09708	1.021315e-07	6.90795e-08
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-579.503932	8	1175.362864	18.09684	1.382376e-08	9.466235e-09
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-579.503971	8	1175.362142	18.09692	1.382262e-08	9.465515e-09
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-579.503979	8	1175.362158	18.096978	1.382245e-08	9.465362e-09
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.400817	6	1176.861634	19.766454	2.526446e-09	1.730075e-09
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-579.503664	9	1177.362128	20.096948	1.878712e-09	1.283946e-09
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.415915	7	1178.83183	21.7665	3.523615e-10	2.412315e-10
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.40754	7	1178.861508	21.796728	3.41965e-10	2.346395e-10
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.40794	7	1178.861588	21.796488	3.419325e-10	2.345157e-10
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.40802	7	1178.861604	21.796424	3.419273e-10	2.345475e-10
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-583.284497	7	1180.508964	23.508114	6.200455e-11	4.245885e-11
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.415853	8	1180.831796	23.766526	4.767925e-11	3.262511e-11
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.415893	8	1180.831786	23.766606	4.767945e-11	3.262474e-11
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.415901	8	1180.831802	23.766622	4.767945e-11	3.262469e-11
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-582.40797	8	1180.861574	23.796394	4.627625e-11	3.168935e-11
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-583.284434	8	1182.508868	25.506388	8.392462e-12	5.749315e-12
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-583.284474	8	1182.508848	25.507058	8.391785e-12	5.746575e-12
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-583.284482	8	1182.508864	25.507384	8.391655e-12	5.746481e-12
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-583.415866	0	1183.831770	35.766590	6.495525e-13	4.418411e-13

# Information theoretic metrics for 150 models of lineage composition

Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-588.731557	8	1193.463114	36.397934	1.558045e-16	1.066928e-16
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-594.149378	6	1200.282156	43.216976	1.702576e-19	1.16595e-19
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-595.018018	6	1202.030876	44.978896	2.947655e-20	2.018111e-20
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-594.149315	7	1203.282833	45.21685	2.364476e-20	1.578075e-20
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-594.149355	7	1203.282811	45.21693	2.364295e-20	1.577945e-20
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-594.149363	7	1203.282126	45.216946	2.364253e-20	1.577925e-20
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-595.017976	7	1204.035852	46.978772	3.98888e-21	2.731545e-21
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-595.018016	7	1204.036032	46.978852	3.988855e-21	2.731325e-21
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-595.018023	7	1204.036046	46.978866	3.988525e-21	2.731285e-21
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-594.149348	8	1204.282806	47.216916	3.118545e-21	2.132545e-21
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-595.018008	8	1206.036016	48.978736	5.396045e-22	3.69625e-22
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-600.479384	5	1210.943668	53.878788	3.987662e-24	2.730835e-24
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-600.479322	6	1212.943844	55.878664	5.397655e-25	3.696235e-25
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-600.479361	6	1212.943822	55.878742	5.397235e-25	3.695945e-25
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-600.479369	6	1212.943838	55.878758	5.397145e-25	3.695885e-25
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-600.479354	7	1214.943808	57.878728	7.364455e-26	5.082995e-26
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-602.181369	7	1218.362798	61.297538	2.992565e-27	1.638195e-27
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-602.181306	8	1220.362812	63.297432	3.237985e-28	2.217325e-28
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-602.181346	8	1220.362692	63.297512	3.237725e-28	2.217155e-28
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-602.181354	8	1220.362708	63.297528	3.237675e-28	2.217115e-28
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-605.058643	6	1222.117286	65.052106	5.600535e-29	3.831625e-29
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-602.181339	9	1222.362678	65.297498	4.381845e-29	3.006625e-29
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-605.058581	7	1224.117342	67.051982	7.588445e-30	5.190975e-30
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-605.058621	7	1224.117342	67.052062	7.578835e-30	5.190525e-30
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-605.058628	7	1224.117256	67.052076	7.579725e-30	5.190465e-30
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-605.058614	8	1226.117228	69.052048	1.025835e-30	7.024765e-31
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.229872	6	1234.459764	73.394564	2.443255e-34	1.673115e-34
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.229881	7	1236.45962	75.394444	3.366995e-35	2.264585e-35
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.22985	7	1236.4597	75.39452	3.366725e-35	2.264465e-35
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.229857	7	1236.459714	75.394534	3.366685e-35	2.264375e-35
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.457633	7	1236.915266	79.850886	2.084762e-35	1.435835e-35
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-614.114033	5	1238.228006	81.102826	5.641995e-36	3.805025e-36
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.229842	8	1238.459684	81.394594	4.475245e-36	3.064585e-36
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.457571	8	1238.915142	81.849692	2.838015e-36	1.943415e-36
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.457561	8	1238.915122	81.85004	2.837785e-36	1.943275e-36
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-611.457618	8	1238.915236	81.850556	2.837745e-36	1.943245e-36
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-614.113394	6	1240.227281	83.1627	7.636573e-37	5.228415e-37
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-614.113398	6	1240.227296	83.16278	7.635962e-37	5.228965e-37
Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp_Mp	-614.113388	6	1240.227296	83.162796	7.635845e-37	5.228915e-37

Lineage composition	$-\ln L$	k	AIC	$\Delta i$	L (Model data)	$w_i$
<i>Clock-like loci</i>						
Mtp_Mtt_Mtv, Mla, Mlc, Mll, Mlr, Mej, Mec	-570.533	8	1157.065	0.000	1.000	0.685
Mtp_Mtv, Mtt, Mla, Mlc, Mll, Mlr, Mej, Mec	-570.533	9	1159.065	2.000	0.135	0.093
Mtt_Mtp, Mtv, Mla, Mlc, Mll, Mlr, Mej, Mec	-570.533	9	1159.065	2.000	0.135	0.093
Mtt_Mtv, Mtp, Mla, Mlc, Mll, Mlr, Mej, Mec	-570.533	9	1159.065	2.000	0.135	0.093
Mtp_Mtt_Mtv, Mla_Mlc, Mll, Mlr, Mej, Mec	-573.375	7	1160.751	3.686	0.025	0.017
Mtp, Mtt, Mtv, Mla, Mlc, Mll, Mlr, Mej, Mec	-570.533	10	1161.065	4.000	0.018	0.013
Mtp_Mtv, Mtt, Mla_Mlc, Mll, Mlr, Mej, Mec	-573.375	8	1162.751	5.686	0.003	0.002
Mtt_Mtp, Mtv, Mla_Mlc, Mll, Mlr, Mej, Mec	-573.375	8	1162.751	5.686	0.003	0.002
Mtt_Mtv, Mtp, Mla_Mlc, Mll, Mlr, Mej, Mec	-573.375	8	1162.751	5.686	0.003	0.002
Mtp, Mtt, Mtv, Mla_Mlc, Mll, Mlr, Mej, Mec	-573.375	9	1164.751	7.686	0.000	0.000

## Information theory metrics for 10 models!

- four models account for **96.4%** of the total model probability
- all treat subspecies within *M. evotis* and *M. lucifugus* as evolutionary lineages
- difference among top models derived from pretending variable . . .



# SPECIES DELIMITATION WITH ABC AND OTHER COALESCENT-BASED METHODS: A TEST OF ACCURACY WITH SIMULATIONS AND AN EMPIRICAL EXAMPLE WITH LIZARDS OF THE *LIOLAEMUS DARWINII* COMPLEX (SQUAMATA: LIOLAEMIDAE)

Arley Camargo,<sup>1,2</sup> Mariana Morando,<sup>3</sup> Luciano J. Avila,<sup>3</sup> and Jack W. Sites, Jr.<sup>1</sup>

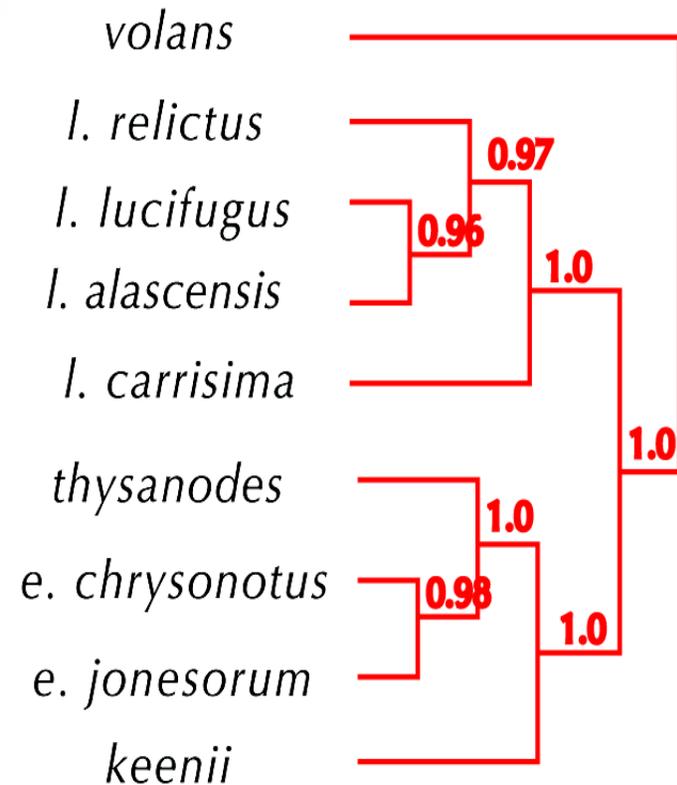
<sup>1</sup>Department of Biology & Monte L. Bean Museum, Brigham Young University, Provo, Utah 84602

<sup>2</sup>E-mail: arley.camargo@gmail.com

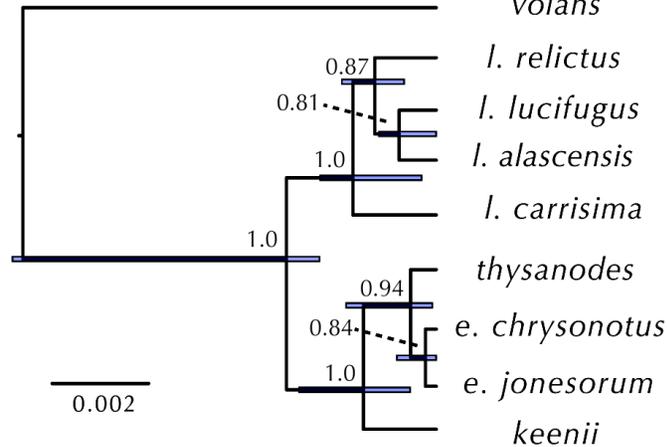
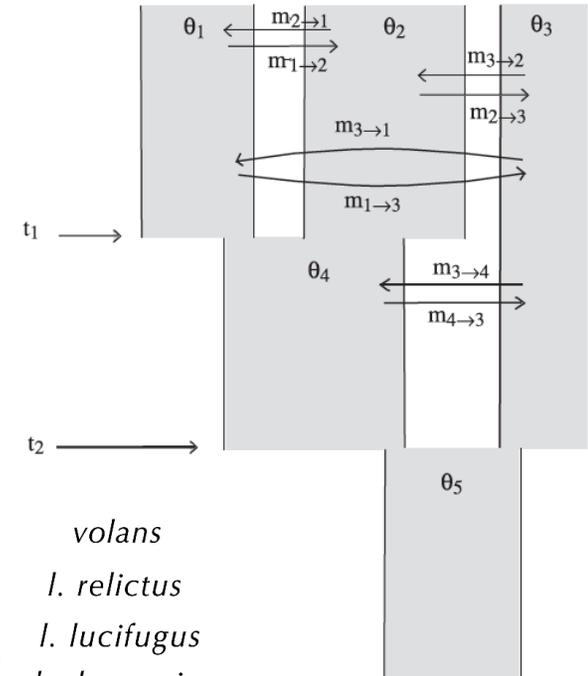
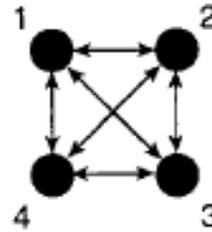
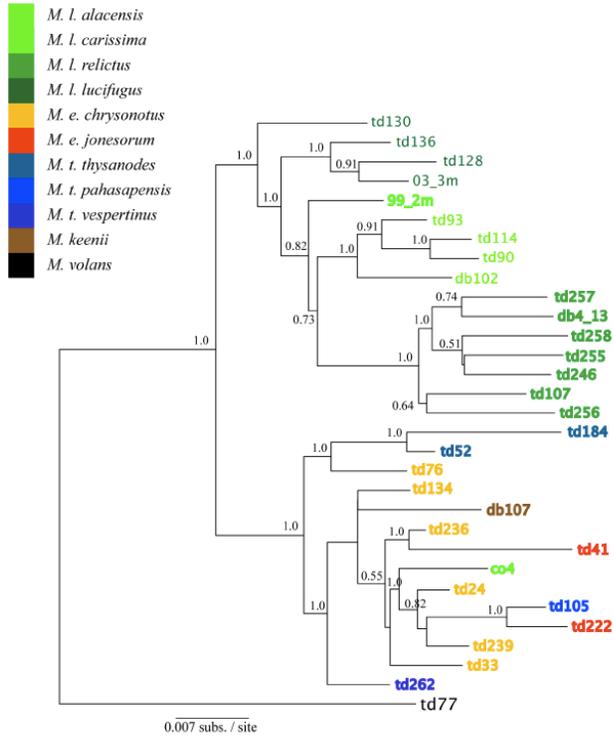
<sup>3</sup>CONICET-CENPAT, Boulevard Almirante Brown 2915, U9120ACD, Puerto Madryn, Chubut, Argentina

Species delimitation?  
ABC > BPP > spedeSTEM

BPP (Yang & Rannala 2010)



may you live in interesting times...



Why do we choose certain models to analyze our data?

Phylogenetics

**STEM: species tree estimation using maximum likelihood for gene trees under coalescence**

Laura S. Kubatko<sup>1,\*</sup>, Bryan C. Carstens<sup>2</sup> and L. Lacey Knowles<sup>3</sup>

<sup>1</sup>Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, <sup>2</sup>Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803 and

<sup>3</sup>Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

Received on November 28, 2008; revised and accepted February 04, 2009

Associate Editor: Martin Bishop

$$L(S, \tau) = \prod_{j=1}^N f(g_j | S, \tau)$$

**STEM:** Analytical calculation of phylogeny under a coalescent model that accounts for the loss of ancestral polymorphism due to genetic drift.

*Assumptions of model*

- $\theta$  is constant (equal on each branch of ST)
- data are evolving in a manner consistent with the molecular clock
- shared polymorphism results from incomplete lineage sorting (**no gene flow**)

- species tree methods **do not consider** population level processes such as gene flow or population expansion
- gene flow will **decrease** the accuracy of phylogeny estimation

Molecular Phylogenetics and Evolution 49 (2008) 832–842

Contents lists available at ScienceDirect

 **Molecular Phylogenetics and Evolution** 

journal homepage: [www.elsevier.com/locate/ympev](http://www.elsevier.com/locate/ympev)

---

Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow

Andrew J. Eckert<sup>a</sup>, Bryan C. Carstens<sup>b,\*</sup>

<sup>a</sup>Section of Evolution and Ecology, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA  
<sup>b</sup>Department of Biological Sciences, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70803, USA

*Syst. Biol.* 63(1):17–30, 2014  
 © The Author(s) 2013. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.  
 For Permissions, please email: [journals.permissions@oup.com](mailto:journals.permissions@oup.com)  
 DOI:10.1093/sysbio/syt049  
 Advance Access publication August 13, 2013

**The Influence of Gene Flow on Species Tree Estimation: A Simulation Study**

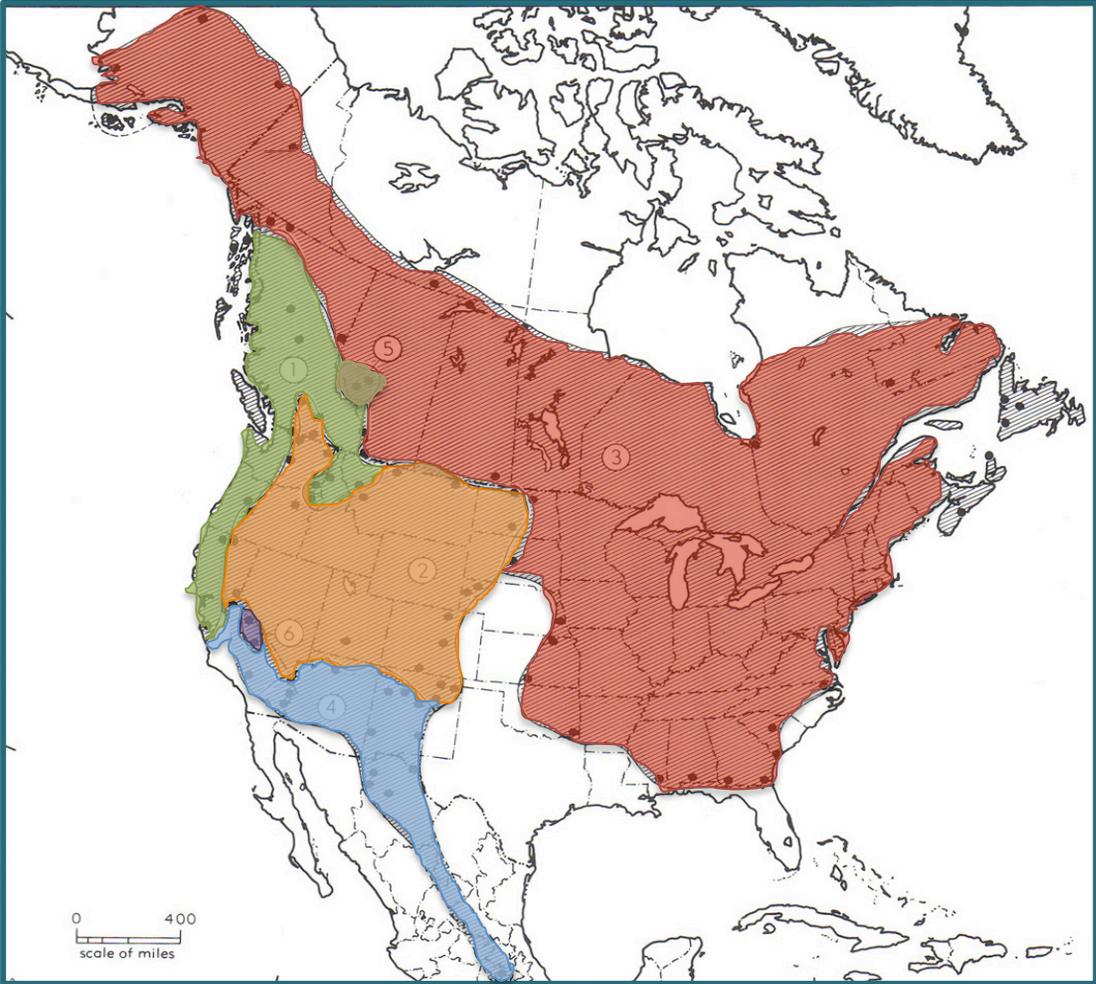
ADAM D. LEACHE<sup>1,\*</sup>, REBECCA B. HARRIS<sup>1</sup>, BRUCE RANNALA<sup>2,3</sup>, AND ZIHENG YANG<sup>3,4</sup>

<sup>1</sup>Department of Biology and Burke Museum of Natural History and Culture, University of Washington, Seattle, WA 98195 USA;  
<sup>2</sup>Genome Center and Department of Evolution & Ecology, University of California, Davis, CA 95616, USA;  
<sup>3</sup>Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China; and  
<sup>4</sup>Department of Biology, University College London, Gower Street, London WC1E 6BT, UK

\*Correspondence to be sent to: Department of Biology, University of Washington, Seattle, WA 98195, USA;  
 E-mail: [leache@uw.edu](mailto:leache@uw.edu).

Received 15 February 2013; reviews returned 10 May 2013; accepted 2 August 2013  
 Associate Editor: Laura Kubatko

Little brown bat subspecies (*Myotis lucifugus*)



*M.l.alacensis*

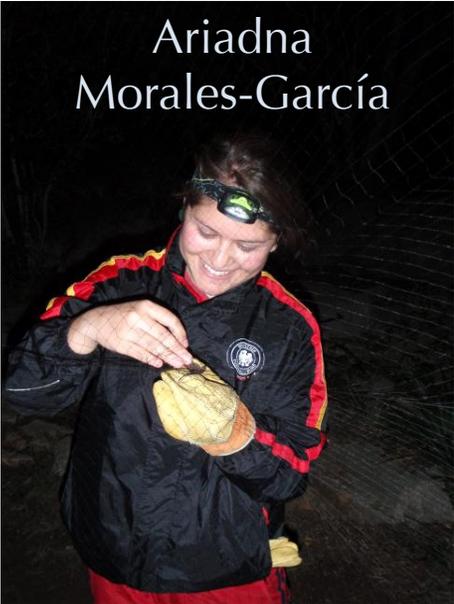
*M.l.carissima*

*M.l.relictus*

*M.l.lucifugus*

(Hall 1981)

*M.l.pernox*



How do we detect gene flow? Use a program such as Migrate-*n* to estimate it . . .

## Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach

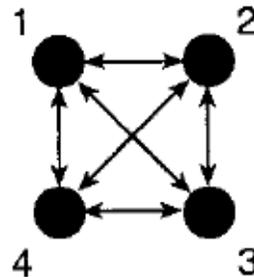
Peter Beerli\* and Joseph Felsenstein

Department of Genetics, University of Washington, Box 357360, Seattle, WA 98195-7360

Contributed by Joseph Felsenstein, February 9, 2001

PNAS | April 10, 2001 | vol. 98 | no. 8 | 4563-4568

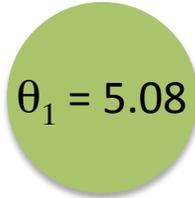
Estimates  $\theta = 4N_e\mu$  and  $M = m / \mu$  using an *n*-island model.



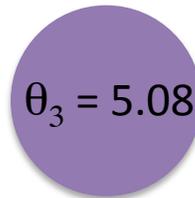
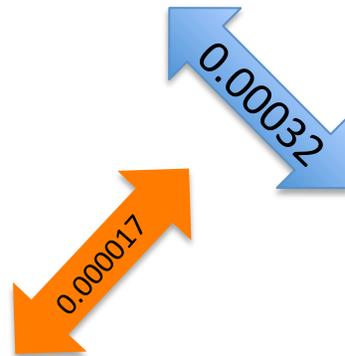
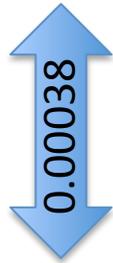
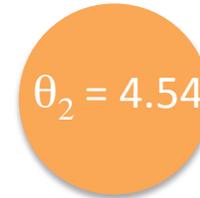
**Fig. 1.** *n*-island model with four populations of equal size, exchanging migrants with equal rates.

How do we detect gene flow? Use a program such as Migrate-*n* to estimate it . . .

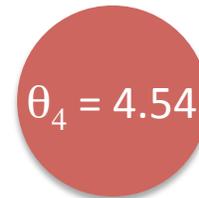
*M. l. alacensis*



*M. l. carissima*



*M. l. relictus*



*M. l. lucifugus*

$$\theta = 4N_e\mu$$
$$M = m / \mu$$

## Unified Framework to Evaluate Panmixia and Migration Direction Among Multiple Sampling Locations

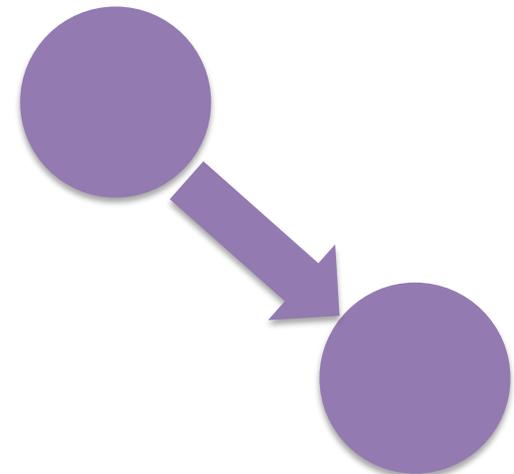
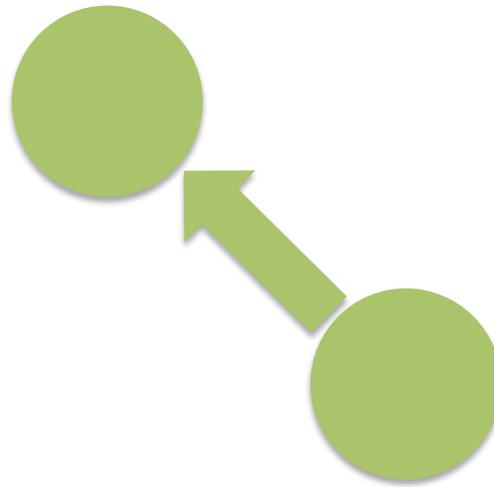
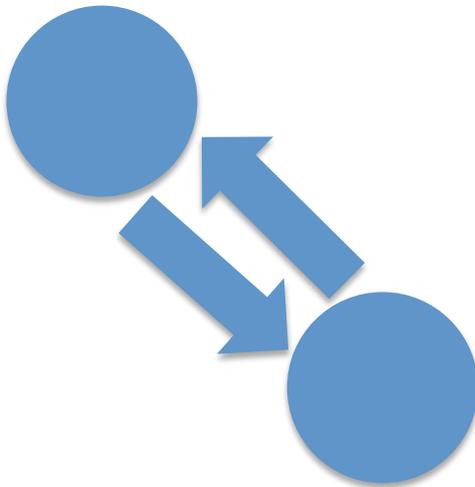
Peter Beerli<sup>1</sup> and Michal Palczewski

*Department of Scientific Computing, Florida State University, Tallahassee, Florida 32306*

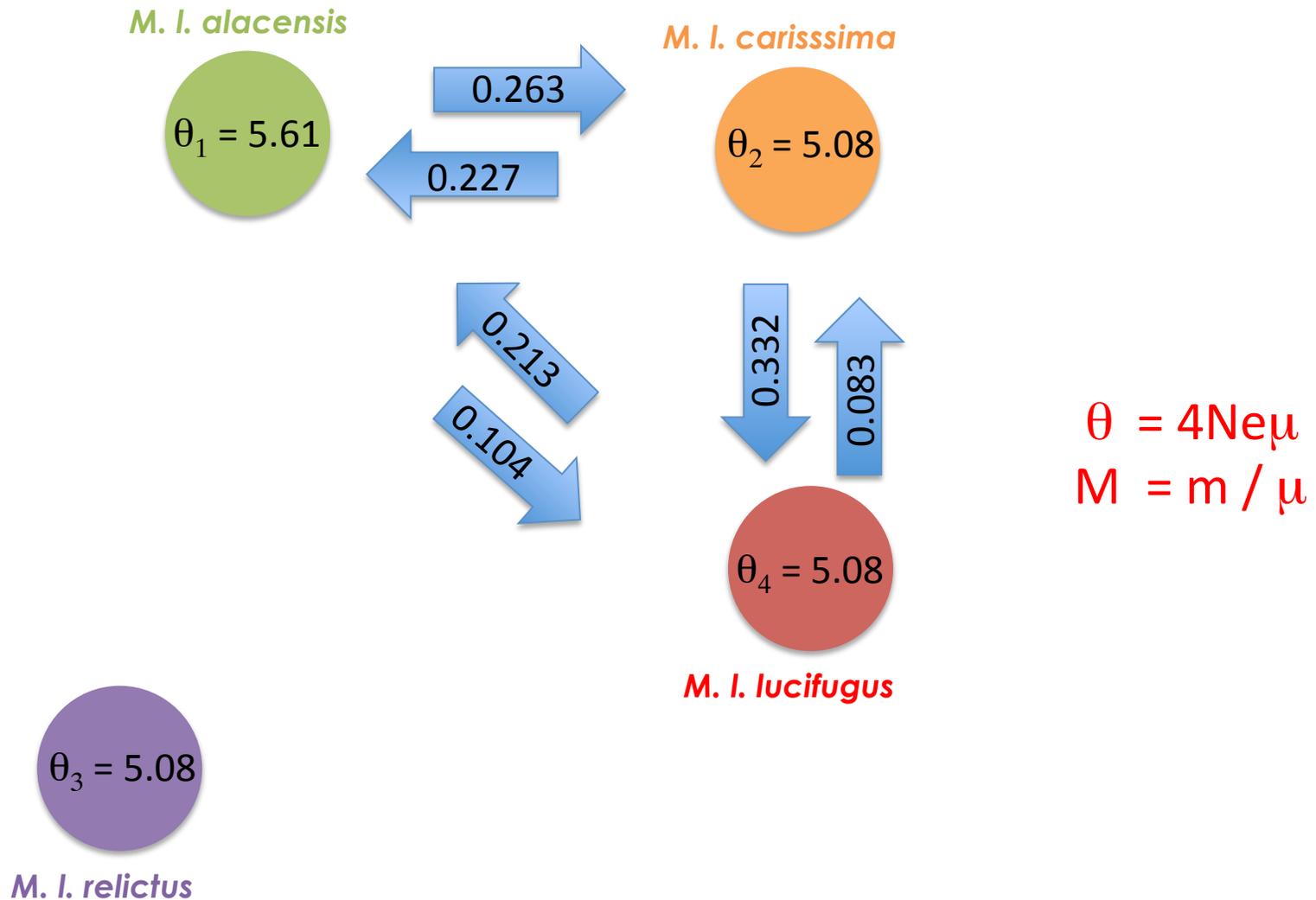
Manuscript received November 27, 2009

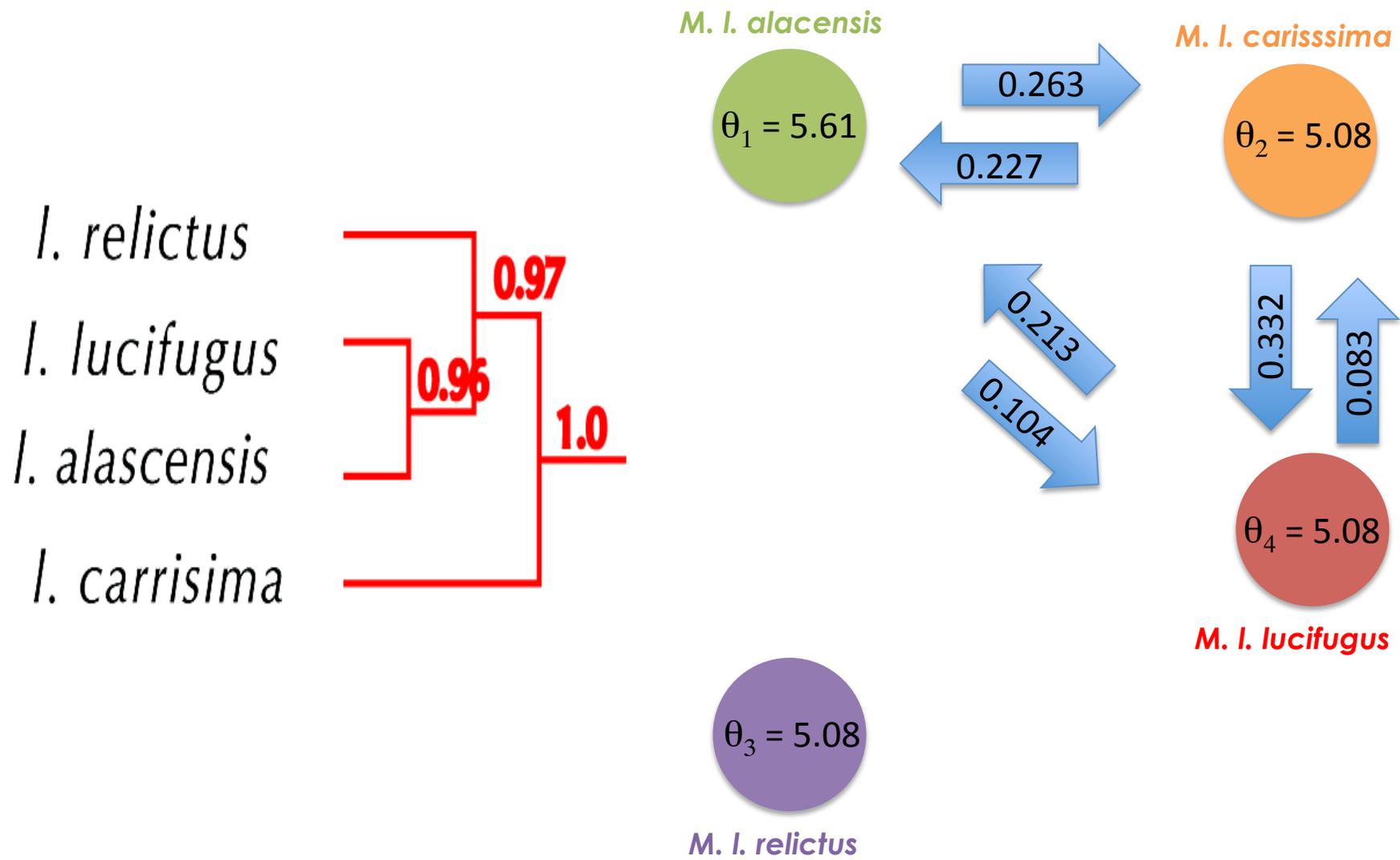
Accepted for publication February 17, 2010

**Migrate-n 3.6** has the ability to calculate marginal likelihoods, so migration models can be evaluated using information theory.



How do we detect gene flow? Use a program such as Migrate-*n* to estimate it . . .





Same data, different interpretations . . .

- If we choose a species tree / delimitation approach, we infer that **each** subspecies within *M. lucifugus* is an independent evolutionary lineage (and probably assume that these lineages do not exchange alleles).
- If we choose an  $n$ -island migration model, we infer that **three of the four** subspecies exchange alleles at a substantial rate (and thus that they are not independent).

# How do we analyze genetic data in phylogeography?

Summarize genetic variation with statistics

- $F_{ST}$ ,  $\theta_w$

Estimate parameters using some model

- $Nm$  with Wright's Island model
- rates of gene flow with a coalescent-model
- genealogies using a phylogenetic model

Use these statistics or estimates to *understand* or *infer* the evolutionary history that produced the genetic variation.



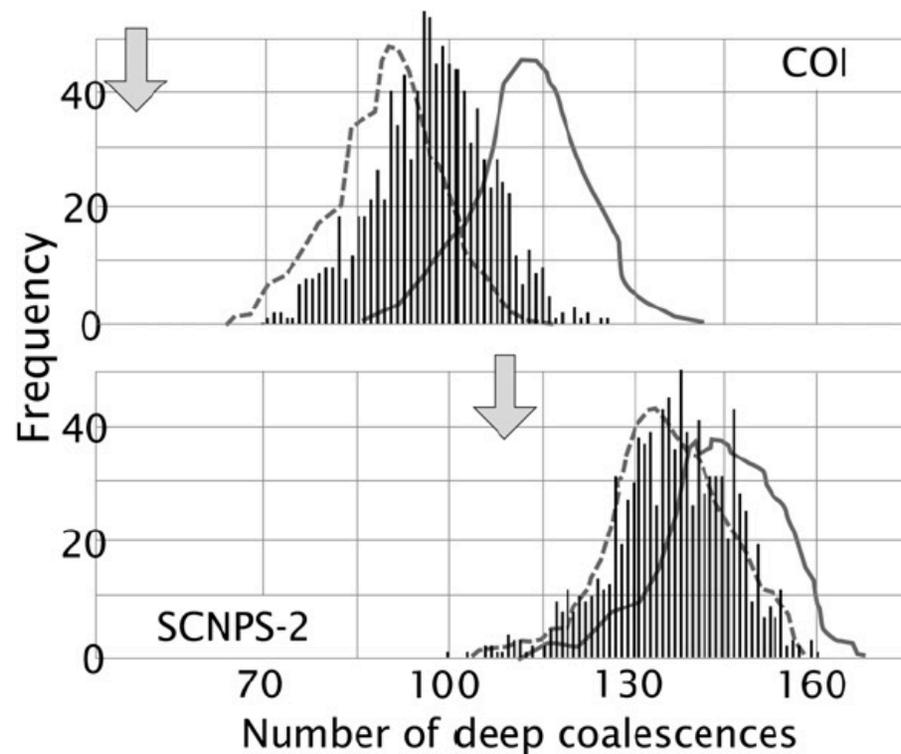
Summaries and estimates are formally generated, but interpreted by researchers in a **qualitative** manner.

- *over-interpretation* – more detailed historical scenarios are proposed than the data support (Knowles & Maddison 2002)
- *confirmation bias* – novel information is interpreted in a manner consistent with preconceived ideas (Nickerson 1998)

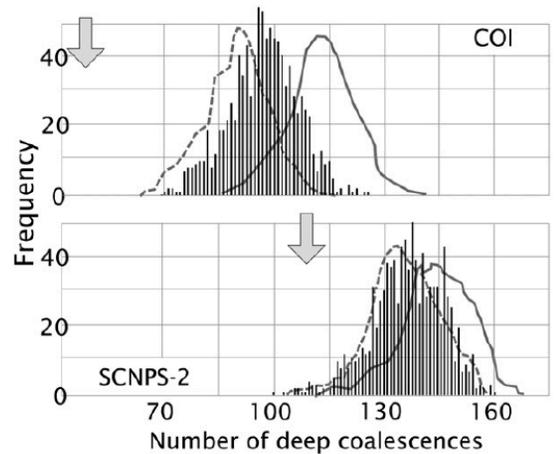
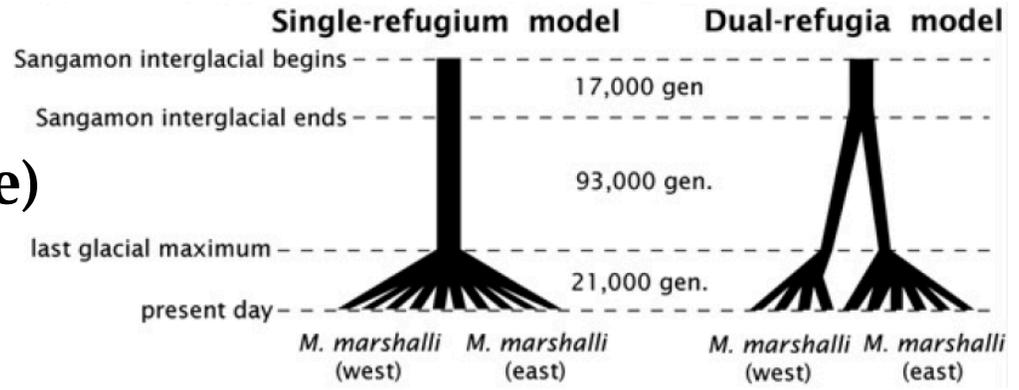
Knowles & Maddison (2002) suggest that phylogeographic hypothesis testing should be used to test demographic models.

## Phylogeographic Hypothesis-testing

**Prob (data | null model is true)** is calculated, but because genetic data are not independent and identically distributed, parametric simulations are used to construct the test distribution.



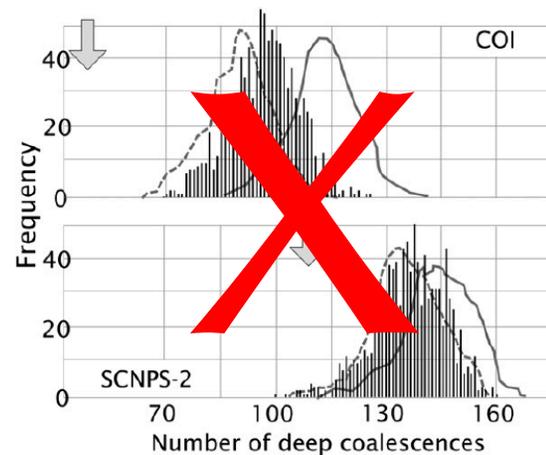
***Prob* (data | null model is true)**



**Assumptions**

- $\theta_i$ , other values
- adequate sampling strategy
- timing of population model
- topology of population model
- adequacy of summary statistics

- rejecting an unrealistic hypothesis tells us nothing useful about an empirical system
- $H_0$  testing may promote false confidence regarding our understanding of the system, and we can not differentiate among hypotheses that can not be rejected



Our goal is to identify the historical forces that generate biodiversity – this *requires* that we understand the historical demography of the species.

- *We can not replicate evolutionary history.*
- *We do not have experimental controls.*

Phylogeography is a *historical* discipline . . . that has relied on statistical tools developed for *experimental* research.

## How can we move past phylogeographic hypothesis testing?

- we can assess the fit of the models that we utilize  
(Bayesian model checking)
- we can choose the best model among a bunch of choices  
(phylogeographic model selection)



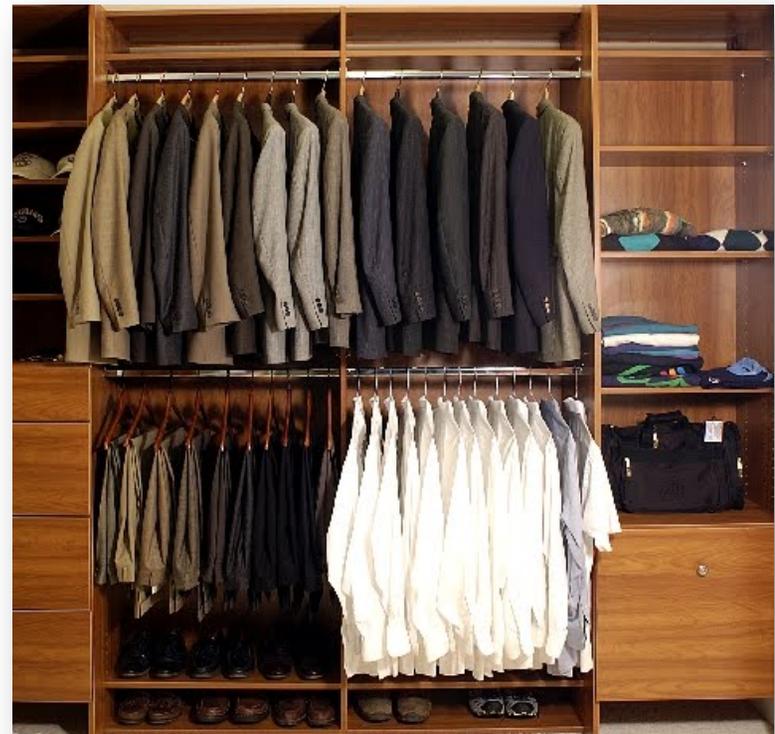
**Bayesian model-checking** allows an evaluation of the statistical fit of complex models to the data.

Integrate over the uncertainty in parameter estimates by simulating from the posterior distribution.

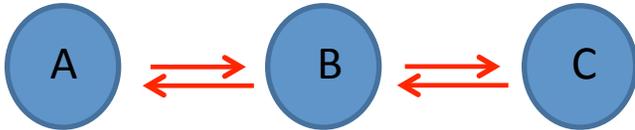
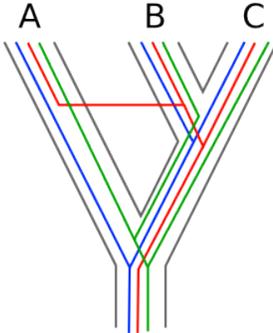
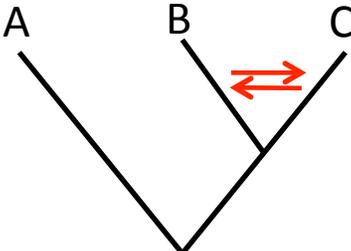


*Bayesian model-checking*

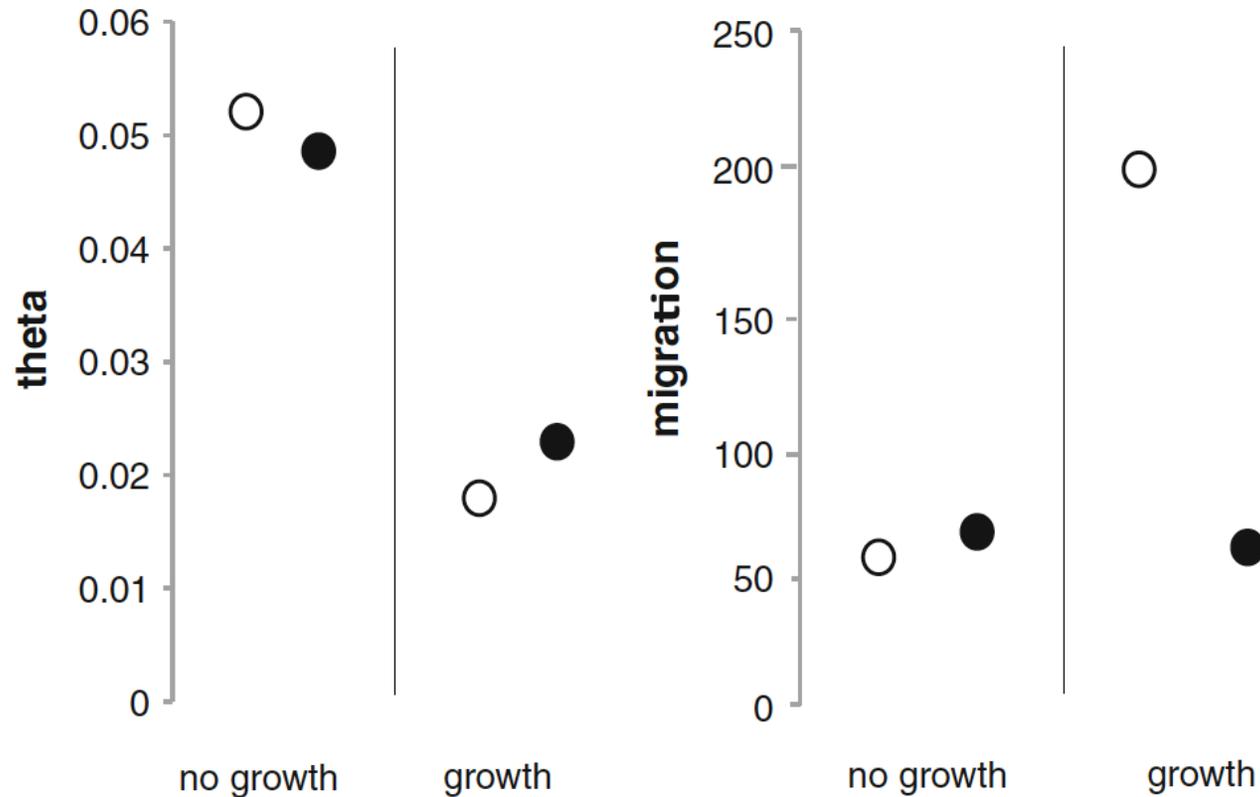
- **Information theory** is a statistical framework developed for quantifying the loss of information that occurs when a model is used to describe reality (*KL distance*; Kullback & Leibler 1951).
- Akaike (1973) linked K-L distance and maximum likelihood.
- We calculate  $Prob(H_j | \text{data})$  for  $j$  hypotheses and rank them using AIC.



*Information theoretic approaches*

Method	Parameters estimated	Parameters NOT estimated	Model
Migrate-n	$m, \theta$	$\tau, topology$	 <p>A diagram showing three blue circles labeled A, B, and C arranged horizontally. Red double-headed arrows connect A to B and B to C, representing migration between all three populations.</p>
*BEAST	topology, $\tau, \theta$	$m$	 <p>A phylogenetic tree with three tips labeled A, B, and C. The tree is rooted at the bottom. A red horizontal line connects the branches leading to B and C, representing migration between these two populations.</p>
IMa2	$\tau, m, \theta$	topology	 <p>A phylogenetic tree with three tips labeled A, B, and C. The tree is rooted at the bottom. A red double-headed arrow is placed on the branch leading to C, representing migration between B and C.</p>

Parameter estimation depends on the parameters included in the model used to estimate the parameters.



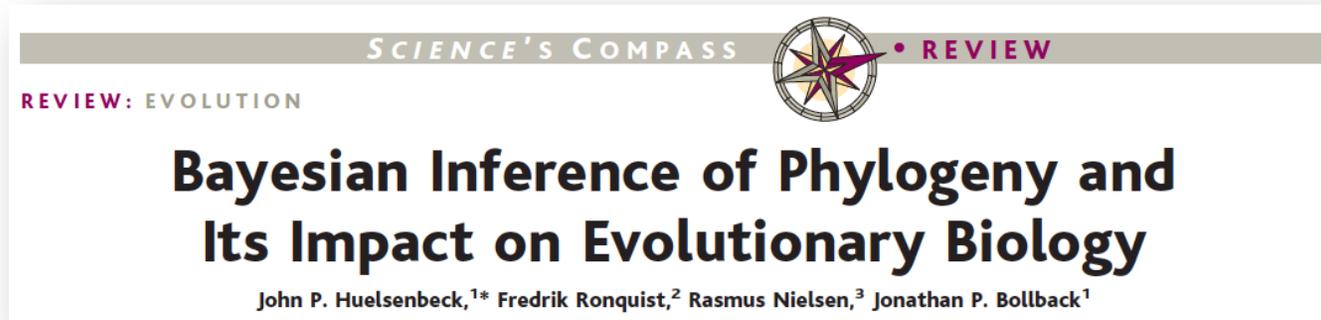
Conserv Genet  
DOI 10.1007/s10592-010-0095-7

RESEARCH ARTICLE

**Conservation genetic inferences in the carnivorous pitcher plant  
*Sarracenia alata* (Sarraceniaceae)**

Margaret M. Koopman · Bryan C. Carstens

# Bayesian Model-checking using Posterior Predictive Simulation



Too big...



...too small...



too Bowie ...

posterior  
probability

likelihood

prior

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$



```
#NEXUS
[ID: 0852508174]
begin trees;
  translate
    1 Anrm,
    2 Bnrm,
    3 Cnrm,
    4 Cnc,
    5 Dnc;
  [
tree rep.1 = ((2:0.100000,(4:0.100000,5:0.100000):0.100000):0.100000,3:0.100000,1:0.100000);
tree rep.1000 = ((3:0.006993,(4:0.006555,5:0.007229):0.00554):0.014269,2:0.001265,1:0.007926);
tree rep.2000 = ((5:0.012335,(3:0.000672,4:0.000022):0.004186):0.016851,2:0.002214,1:0.005338);
tree rep.3000 = (((4:0.013396,3:0.001861):0.010962,5:0.000552):0.001771,2:0.000542,1:0.001639);
tree rep.4000 = ((4:0.002758,(3:0.005498,5:0.003617):0.005864):0.006643,2:0.005980,1:0.025120);
tree rep.5000 = ((4:0.001777,(3:0.000789,5:0.001393):0.006475):0.013680,2:0.004508,1:0.006280);
tree rep.6000 = ((5:0.002306,(4:0.002026,3:0.000966):0.003021):0.016065,2:0.008722,1:0.011203);
tree rep.7000 = (2:0.005251,((5:0.004186,4:0.003543):0.002246,3:0.001565):0.007210,1:0.002549);
tree rep.8000 = (2:0.003825,((3:0.000630,5:0.003034):0.001699,4:0.006023):0.022671,1:0.025138);
tree rep.9000 = (2:0.000986,((5:0.013872,4:0.005184):0.001416,3:0.003382):0.005159,1:0.003640);
tree rep.10000 = (2:0.004103,((3:0.000307,5:0.003384):0.000849,4:0.010198):0.002563,1:0.019618);
tree rep.11000 = ((3:0.001570,(5:0.009439,4:0.003157):0.008600):0.008988,2:0.020539,1:0.001156);
tree rep.12000 = (2:0.005935,(5:0.005158,(3:0.001101,4:0.003551):0.000527):0.012832,1:0.001782);
tree rep.13000 = (((3:0.000084,4:0.001978):0.001340,5:0.005619):0.021711,2:0.003141,1:0.004153);
tree rep.14000 = ((3:0.002721,5:0.003063):0.000965,4:0.002150):0.017916,2:0.002912,1:0.001911);
tree rep.15000 = ((5:0.003662,4:0.008229):0.001214,3:0.004921):0.003048,2:0.003570,1:0.005086);
tree rep.16000 = (2:0.001223,(4:0.009145,(5:0.002650,3:0.005159):0.000760):0.023695,1:0.005769);
```

posterior  
probability

likelihood

prior

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$

```
#NEXUS
[ID: 052508174]
begin trees;
  kronstat;
  1 Anm;
  2 Bnm;
  3 Cnm;
  4 Cnc;
  5 Dnc;
  6 ;
tree rep.1 = ((2:0.100000,(4:0.100000,5:0.100000):0.100000):0.100000,3:0.100000,1:0.100000);
tree rep.1000 = ((3:0.000993,(4:0.000553,5:0.007229):0.000554):0.014269,2:0.001265,1:0.007920);
tree rep.2000 = ((5:0.022335,(3:0.000075,4:0.000022):0.001180):0.016851,2:0.002216,1:0.003382);
tree rep.3000 = (((4:0.013396,3:0.001861):0.010662,5:0.000552):0.001771,2:0.000542,1:0.001639);
tree rep.4000 = ((4:0.002758,(3:0.005498,5:0.003617):0.005664):0.006663,2:0.000980,1:0.002120);
tree rep.5000 = ((4:0.001777,(3:0.000785,5:0.001393):0.000675):0.012680,2:0.000508,1:0.000280);
tree rep.6000 = ((5:0.002306,(4:0.002026,3:0.000966):0.003021):0.016005,2:0.000722,1:0.011203);
tree rep.7000 = ((2:0.005251,(3:0.004186,4:0.003543):0.002246,3:0.001565):0.007210,1:0.002549);
tree rep.8000 = ((2:0.003825,(3:0.000305,5:0.000305):0.001899,4:0.000033):0.021071,1:0.023183);
tree rep.9000 = ((2:0.000986,(3:0.013872,4:0.005164):0.001416,3:0.003382):0.005159,1:0.003640);
tree rep.10000 = ((2:0.004103,(3:0.000307,5:0.003384):0.000849,4:0.010188):0.002563,1:0.010618);
tree rep.11000 = ((3:0.001570,(5:0.009439,4:0.001577):0.000600):0.000048,2:0.003939,1:0.001150);
tree rep.12000 = ((2:0.000935,(5:0.005158,(3:0.001181,4:0.003553):0.000572):0.012810,1:0.001782);
tree rep.13000 = (((3:0.000084,4:0.001978):0.001340,5:0.005619):0.021711,2:0.003141,1:0.004153);
tree rep.14000 = (((3:0.002721,5:0.003065):0.000905,4:0.002130):0.017916,2:0.002912,1:0.001311);
tree rep.15000 = (((5:0.003662,4:0.000229):0.001216,3:0.000912):0.000304,2:0.003570,1:0.000800);
tree rep.16000 = ((2:0.001223,(4:0.009145,(3:0.002650,3:0.005159):0.000700):0.023695,1:0.005769);
```

- simulate from posterior
- generate predictive distribution
- compare to empirical data

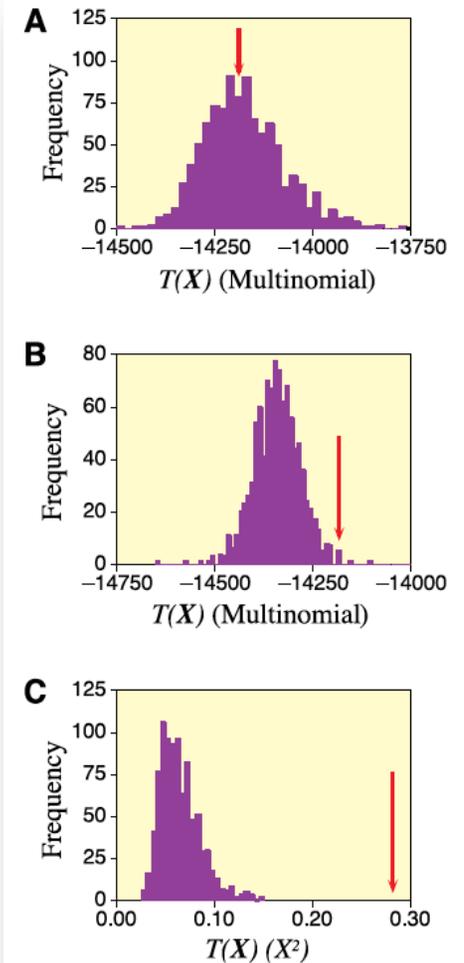


Fig. 3. The posterior predictive distributions for tests of (A) the adequacy of the GTR model, (B) of the adequacy of the Jukes-Cantor model, and (C) the hypothesis of constant nucleotide frequencies over time. The arrows above the distributions show the observed value of the test statistics.

SCIENCE'S COMPASS • REVIEW

REVIEW: EVOLUTION

# Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology

John P. Huelsenbeck,<sup>1\*</sup> Fredrik Ronquist,<sup>2</sup> Rasmus Nielsen,<sup>3</sup> Jonathan P. Bollback<sup>1</sup>



posterior probability =  $\frac{\text{likelihood} \times \text{prior}}{\text{Pr[Data]}}$

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$

Multispecies Coalescent model (implemented in \*Beast)

$L(g_i) = P(d_i | g_i)$ . Likelihood of the data<sub>i</sub> given genealogy<sub>i</sub>

$L(u_i) = P(g_i | u_i)$ . Likelihood of genealogy<sub>i</sub> given the molecular clock<sub>i</sub>

$L(S) = P(u_i | S_i)$ . Likelihood of the molecular clock<sub>i</sub> given the species tree\*

$$P(S|D) \propto \prod_{i=1}^n \int_{g_i} \int_{u_i} P(d_i | g_i) P(g_i | u_i) P(u_i | S) P(S) du_i dg_i.$$

P(S) is the joint prior probability distribution on the species tree\*

## Poor Fit to the Multispecies Coalescent is Widely Detectable in Empirical Data

NOAH M. REID<sup>1,\*</sup>, SARAH M. HIRD<sup>1</sup>, JEREMY M. BROWN<sup>1</sup>, TARA A. PELLETIER<sup>2</sup>, JOHN D. McVAY<sup>1</sup>, JORDAN D. SATLER<sup>2</sup>,  
 AND BRYAN C. CARSTENS<sup>2</sup>

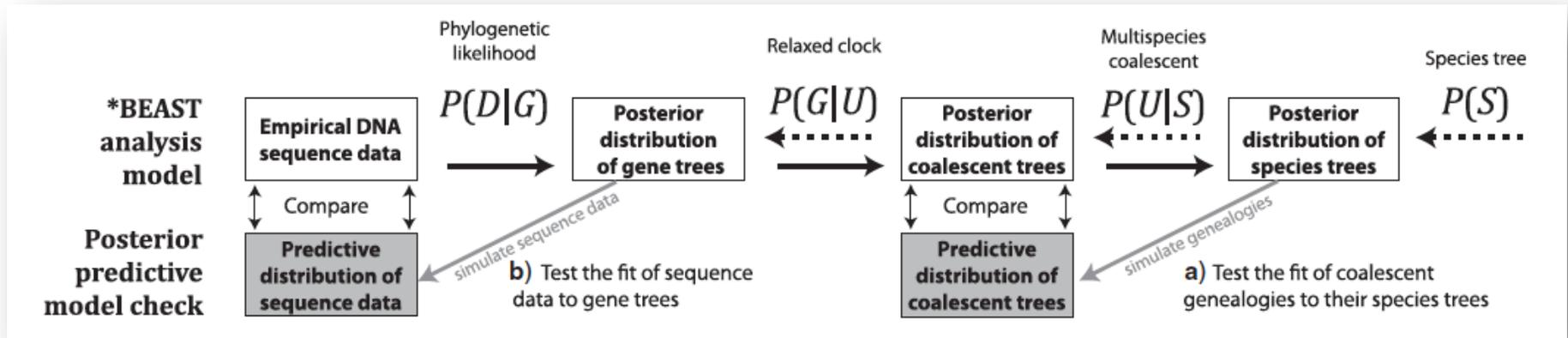
<sup>1</sup>Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; and

<sup>2</sup>Department of Evolution, Ecology & Organismal Biology, Ohio State University, Columbus, OH 43210, USA

\*Correspondence to be sent to: 202 Life Sciences Building, Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; E-mail: nreid1@tigers.lsu.edu.

Received 26 November 2012; reviews returned 2 March 2013; accepted 1 August 2013

Associate Editor: Laura Kubatko



4 / 25 data sets had poor fit on the species tree level  
 44 / 240 loci were outliers on the sequence data level

### Poor Fit to the Multispecies Coalescent is Widely Detectable in Empirical Data

NOAH M. REID<sup>1,\*</sup>, SARAH M. HIRD<sup>1</sup>, JEREMY M. BROWN<sup>1</sup>, TARA A. PELLETIER<sup>2</sup>, JOHN D. MCVAY<sup>1</sup>, JORDAN D. SATLER<sup>2</sup>,  
AND BRYAN C. CARSTENS<sup>2</sup>

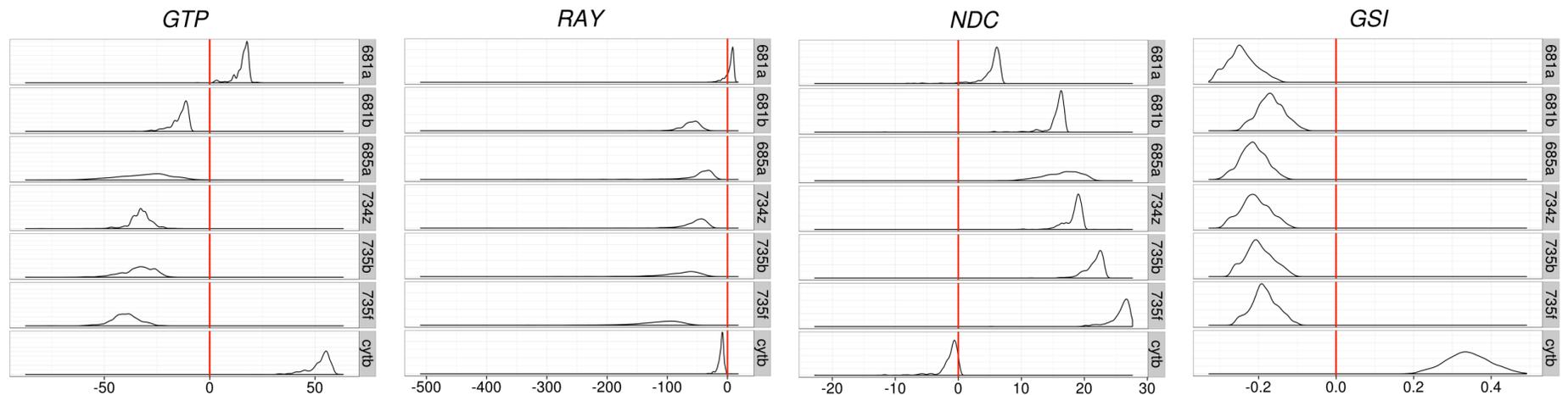
<sup>1</sup>Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; and

<sup>2</sup>Department of Evolution, Ecology & Organismal Biology, Ohio State University, Columbus, OH 43210, USA

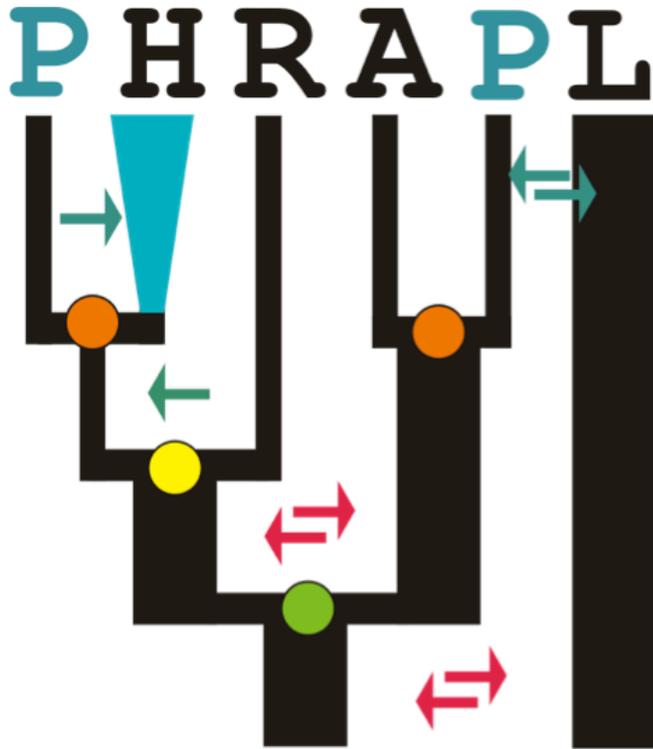
\*Correspondence to be sent to: 202 Life Sciences Building, Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; E-mail: nreid1@tigers.lsu.edu.

Received 26 November 2012; revisions returned 2 March 2013; accepted 1 August 2013  
Associate Editor: Laura Kubacko

- analyzed data using \*Beast (species tree model)
- 50 million generations represented in posterior distribution
- posterior predictive simulations using our R-package (Gruenstaeudl et al. in review)



The multi-species coalescent model IS NOT a good fit to the *Myotis* data.



≈

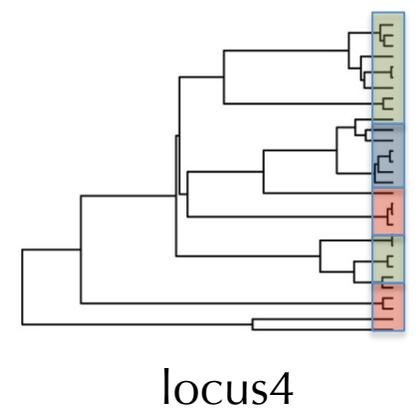
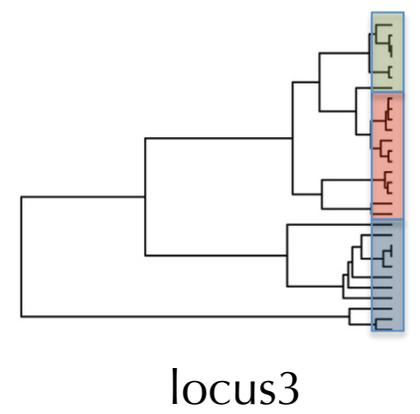
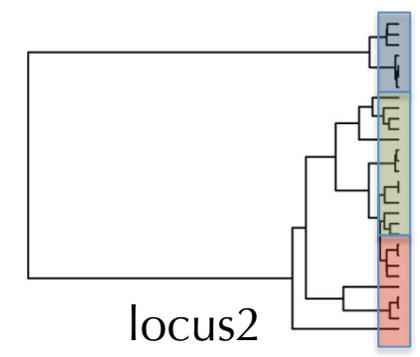
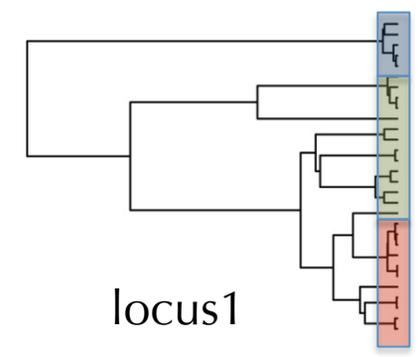
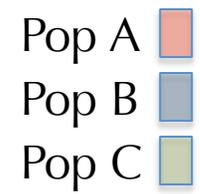


PHylogeographic InfeRence using APproximated Likelihoods

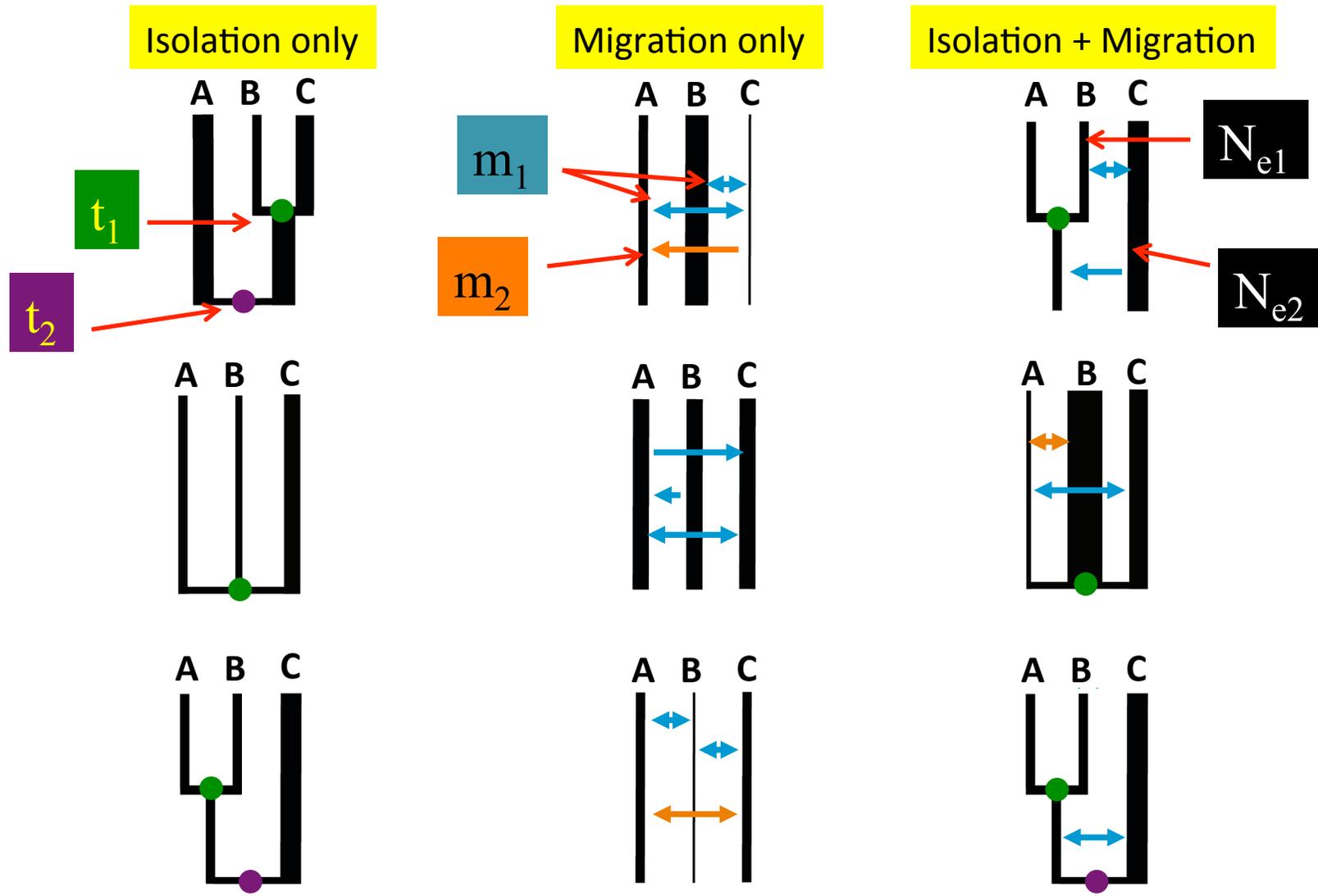
with Brian O'Meara, Nathan Jackson, Ariadna Morales García

# Input

1. Gene trees
2. Population assignments
3. Max K (max number of free parameters;  $t, m, N_e$ )

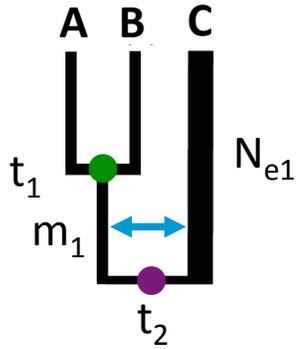


# Phrapl functions: Define all possible models

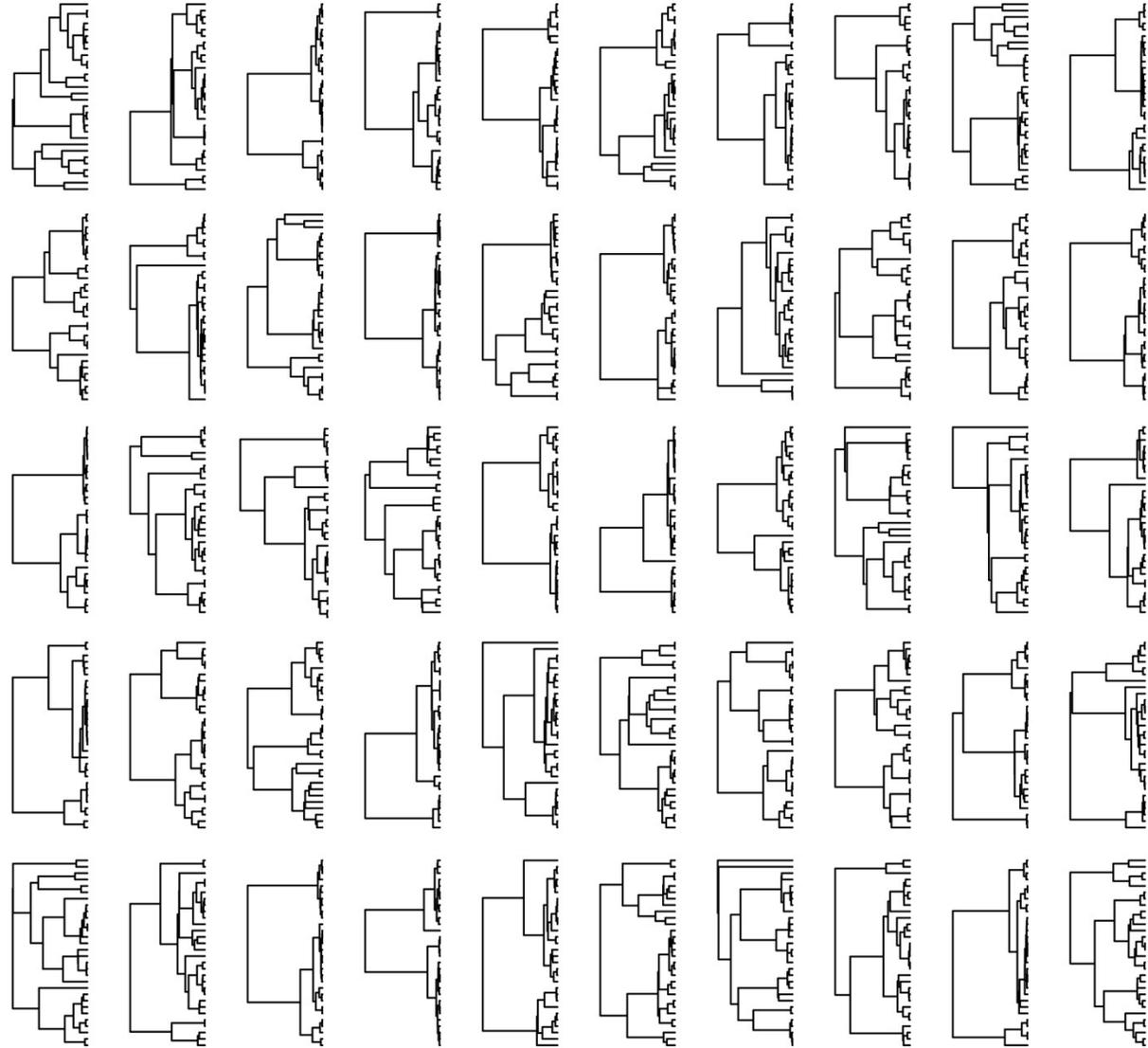


# Phrapl functions: approximate likelihood

For each model...

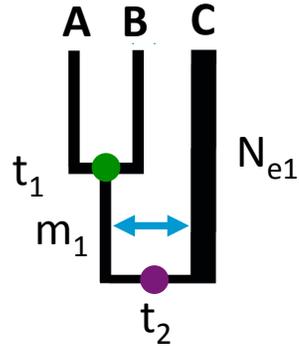


Simulate large number  
of trees

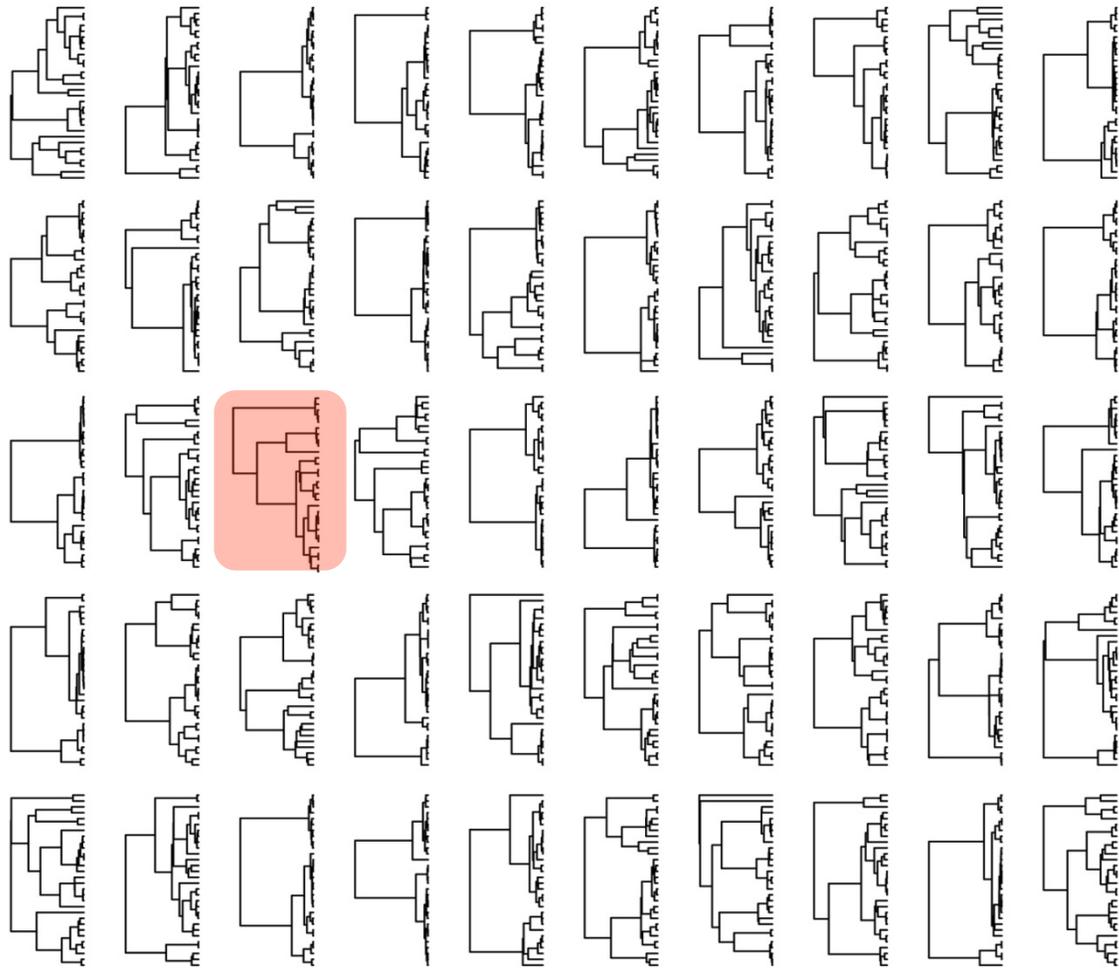
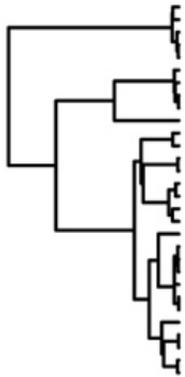


# Phrapl functions: approximate likelihood

For each model...



observed tree

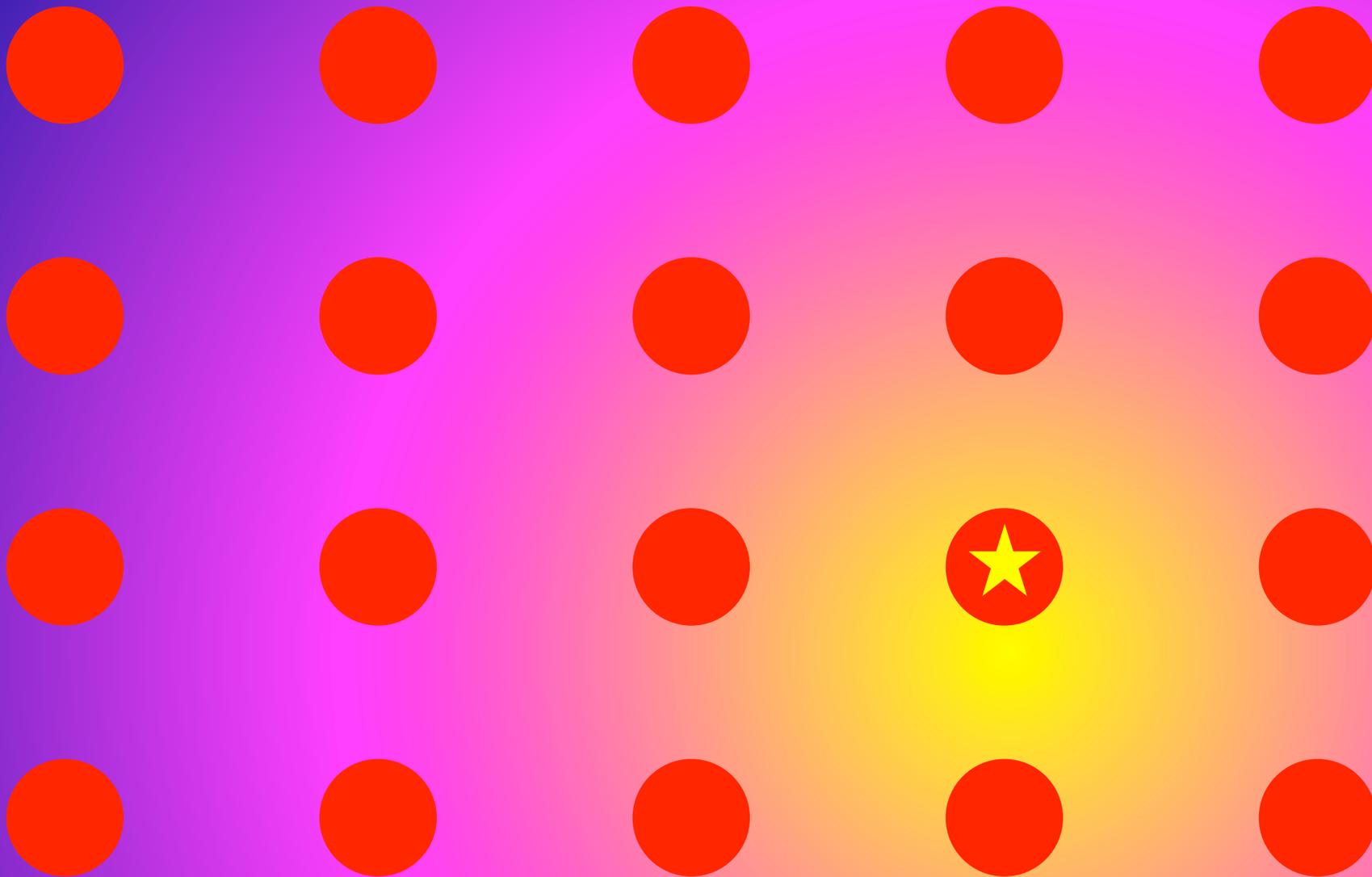


expected trees

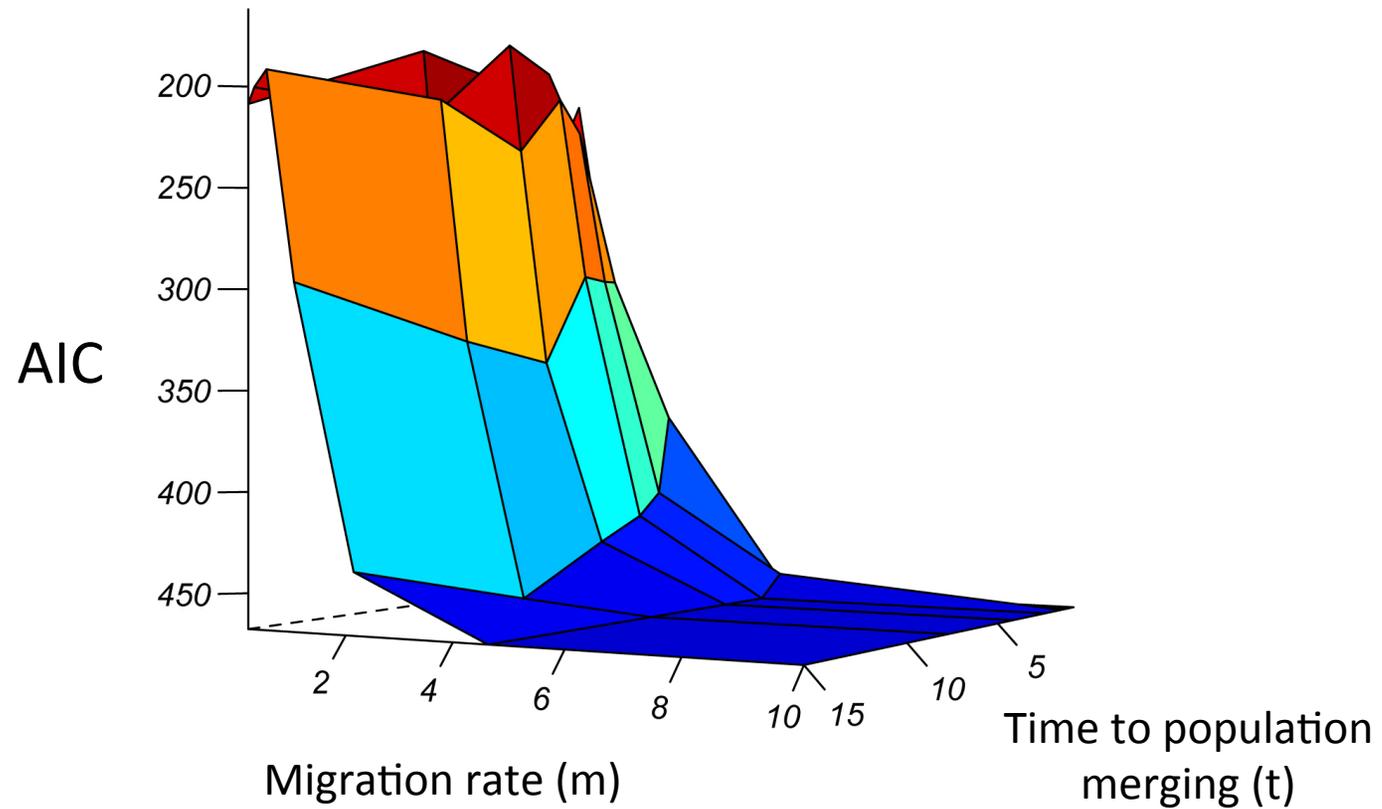
Calculate # of topological matches

1 match / 50 simulated trees  $\approx$   
 $\text{prob}(\text{topology}(\text{observed}) \mid \tau_1, m_1, N_{e1}) = \text{likelihood}$

Model Optimization required for approximation of the L (D|M)

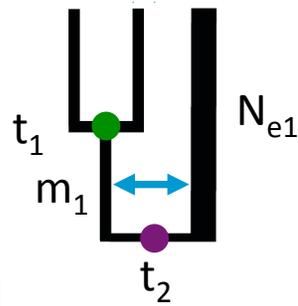


# Phrapl functions: model parameters are optimized using grid values



# Phrapl functions: approximate likelihood

model 1



expected trees



observed tree

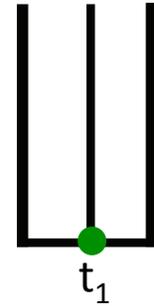


$$\ln L \approx \ln(5/50) = -2.303$$

$$\text{AIC} = 2.303 + 2 * 1 = 4.303$$

$N_{e1}$

model 2



expected trees



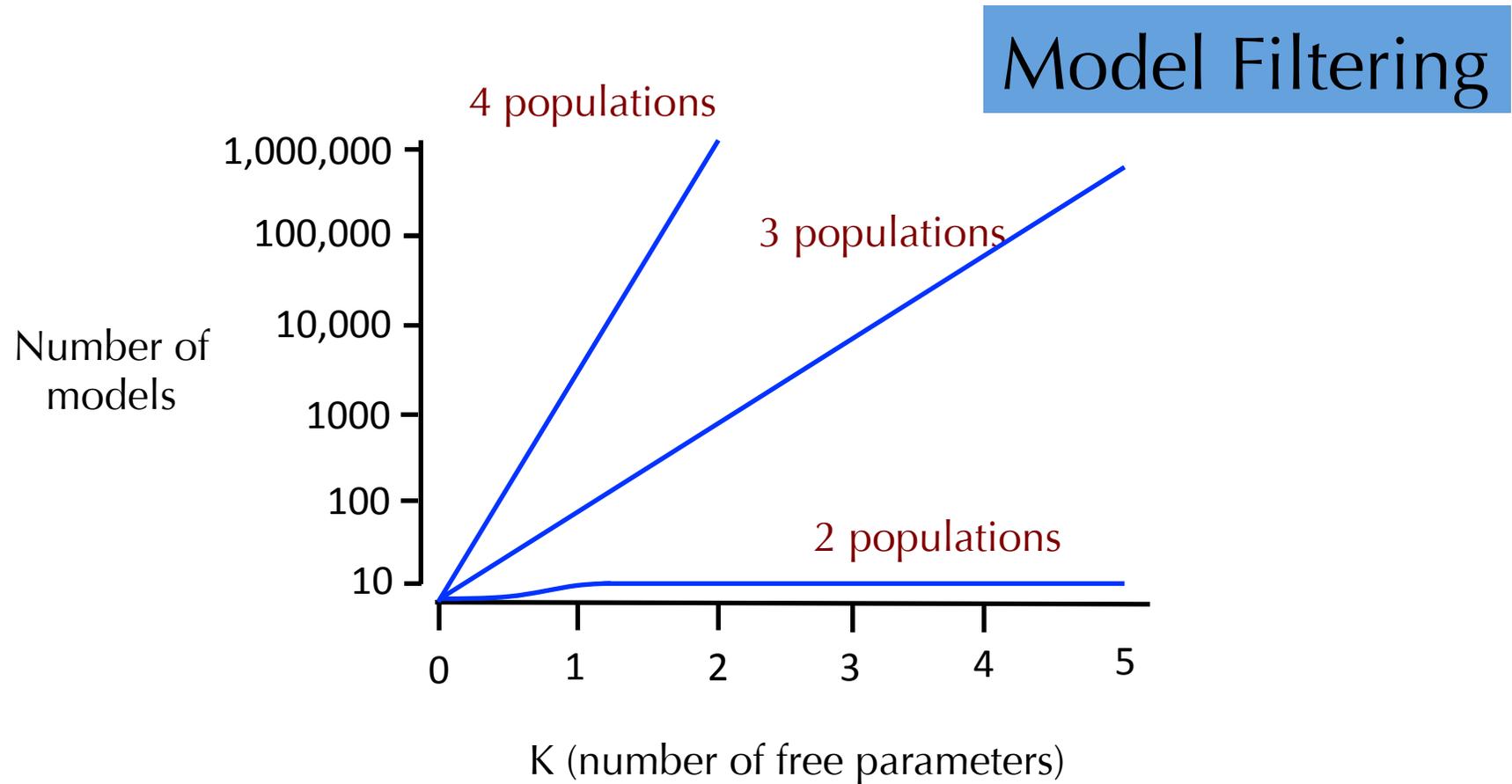
$$\ln L \approx \ln(1/50) = -3.912$$

$$\text{AIC} = 3.912 + 2 * 3 = 9.912$$

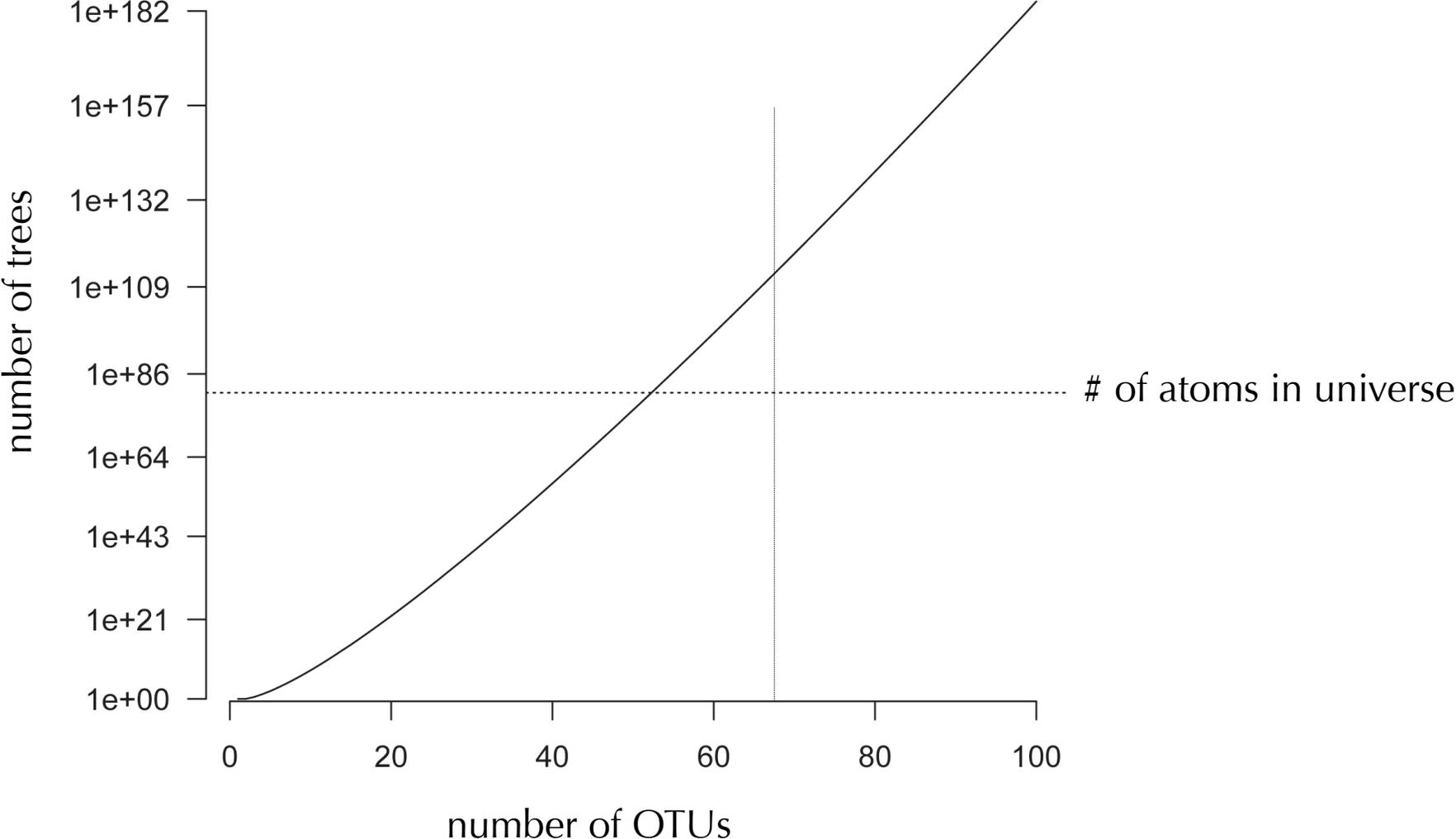
# Phrapl output: comparing model fit using AIC

model	K	lnL	AIC	dAIC	AIC_weights	parameters
546	3	-52.79157	111.58314	0.00000	0.55011	tau1, tau2, Ne1
376	4	-53.23296	114.46593	2.88300	0.13014	tau1, Ne1, Ne2, m1
394	4	-54.81089	117.62177	6.03900	0.02686	tau1, tau2, Ne1, m1
330	4	-54.86685	117.73369	6.15100	0.02540	tau1, Ne1, Ne2, m1
288	4	-55.09833	118.19665	6.61400	0.02015	tau1, Ne1, m1, m2
91	3	-56.14528	118.29055	6.70700	0.01923	Ne1, Ne2, m1
399	4	-55.19640	118.39280	6.81000	0.01827	tau1, tau2, Ne1, Ne2
200	4	-55.49627	118.99254	7.40900	0.01354	tau1, Ne1, m1, m2
30	4	-55.52415	119.04829	7.46500	0.01317	Ne1, m1, m2, m3
69	3	-56.91221	119.82441	8.24100	0.00893	Ne1, m1, m2
25	3	-56.91288	119.82575	8.24300	0.00892	Ne1, m1, m2
615	4	-55.92477	119.84954	8.26600	0.00882	tau1, Ne1, m1, m2
291	4	-55.93903	119.87806	8.29500	0.00869	tau1, Ne1, m1, m2
322	4	-56.12679	120.25357	8.67000	0.00721	tau1, Ne1, Ne2, m1
549	4	-56.14120	120.28239	8.69900	0.00710	tau1, tau2, Ne1, m1
72	4	-56.15484	120.30967	8.72700	0.00700	Ne1, m1, m2, m3
632	4	-56.19445	120.38889	8.80600	0.00673	tau1, Ne1, m1, m2
635	4	-56.43589	120.87178	9.28900	0.00529	tau1, Ne1, m1, m2
97	4	-56.44301	120.88601	9.30300	0.00525	Ne1, Ne2, m1, m2
272	4	-56.49789	120.99579	9.41300	0.00497	tau1, Ne1, m1, m2
240	3	-57.50589	121.01177	9.42900	0.00493	tau1, Ne1, m1
415	4	-56.62749	121.25498	9.67200	0.00437	tau1, Ne1, m1, m2
560	4	-56.66308	121.32615	9.74300	0.00421	tau1, tau2, Ne1, m1

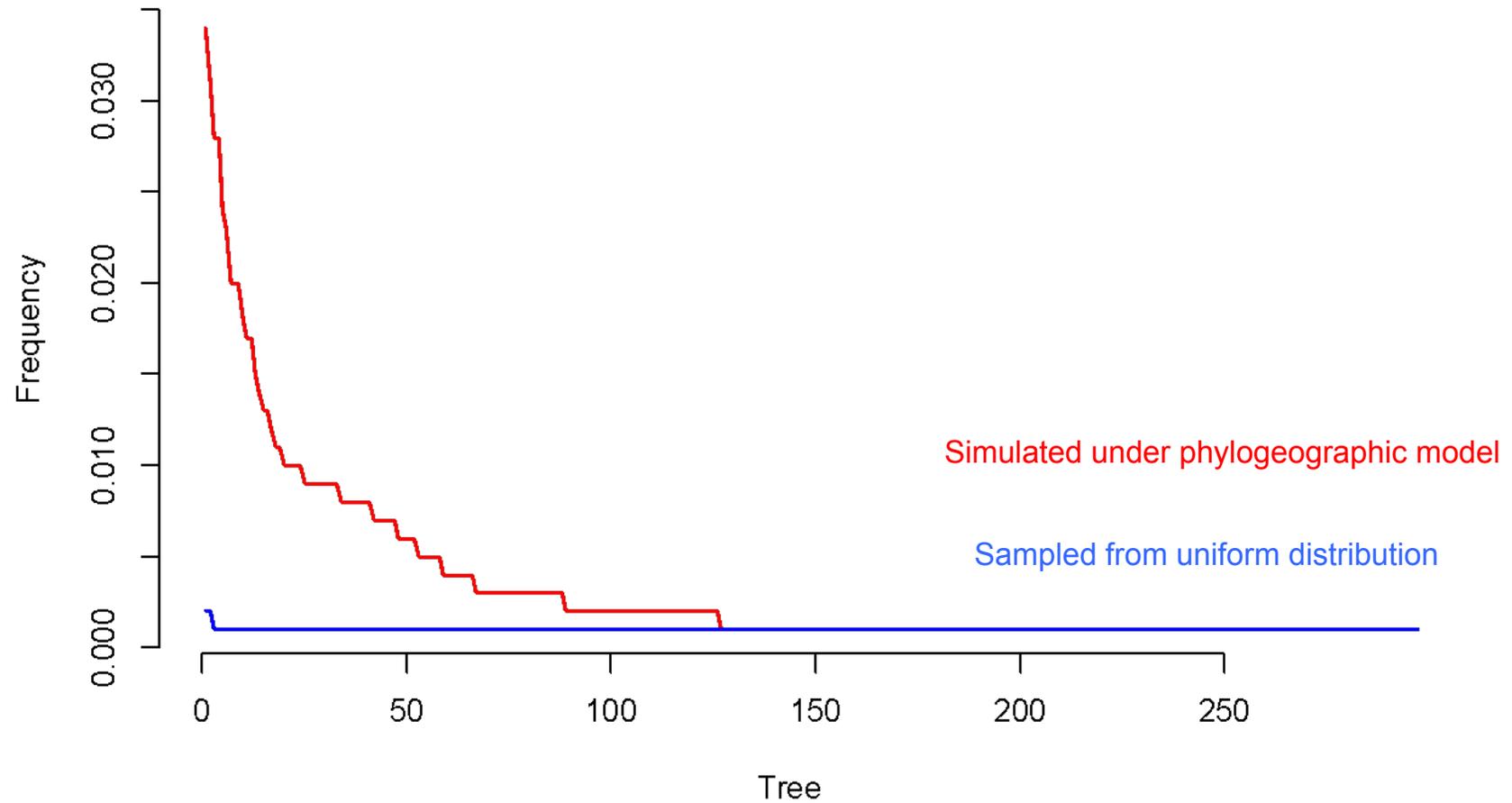
# Challenge #1: model space quickly gets enormous



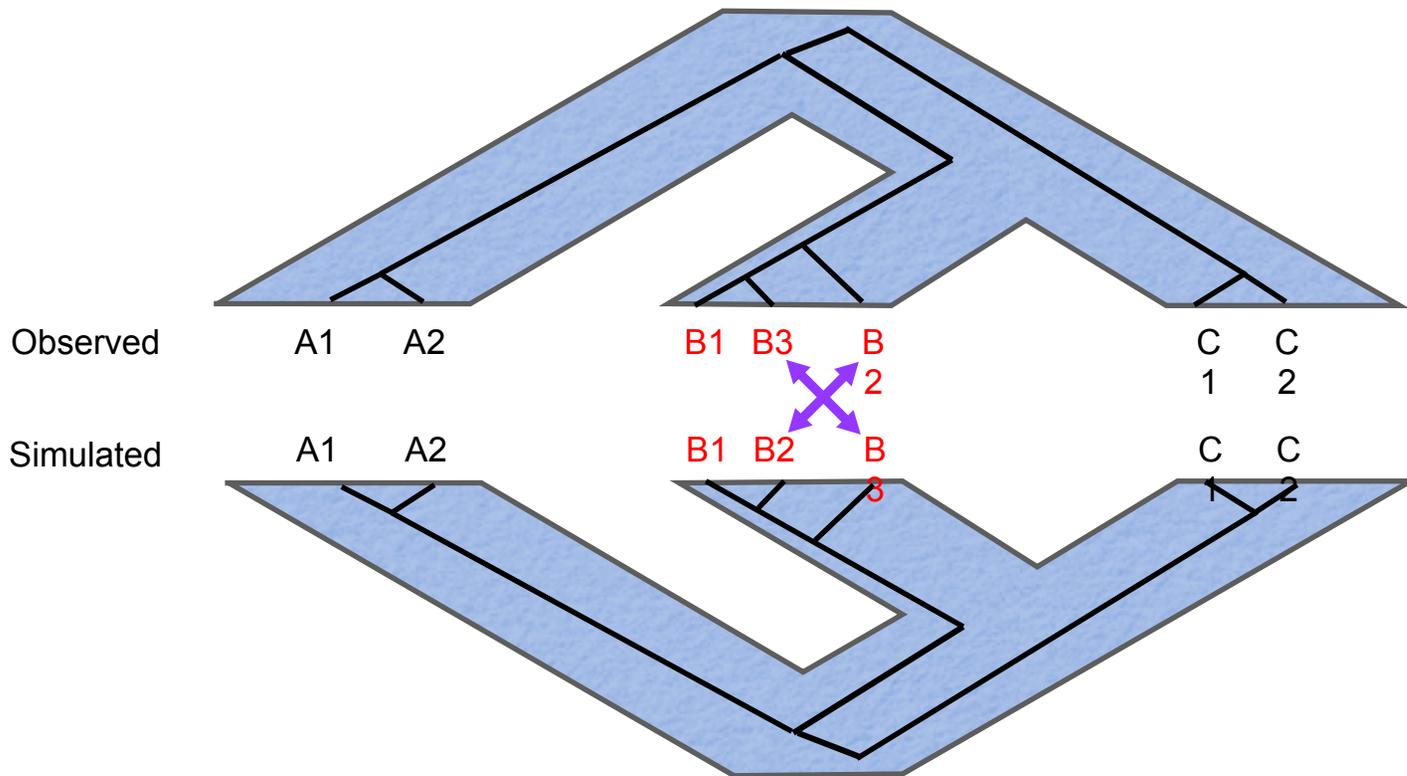
# Challenge #2: tree space quickly gets enormous



# Tree probabilities are not uniform

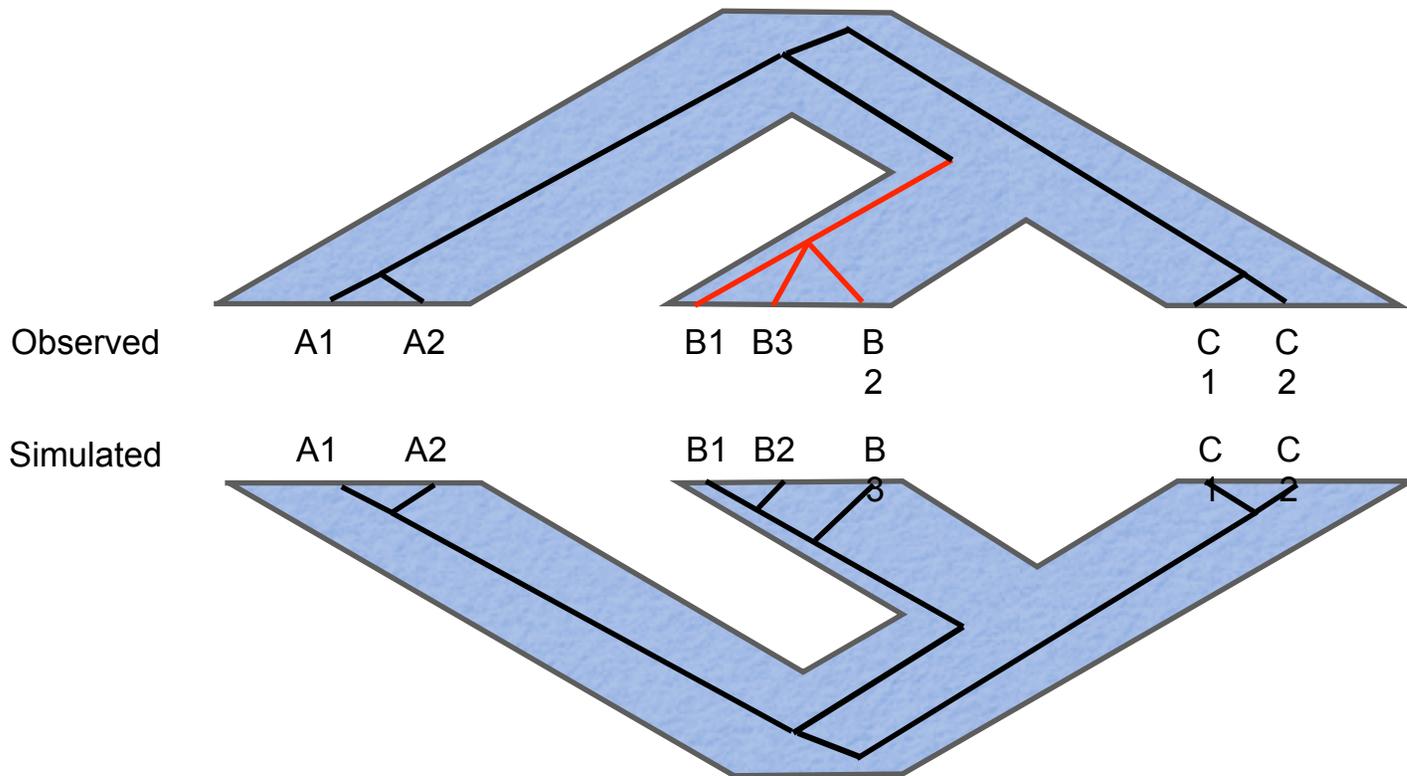


# Sample labels within populations arbitrary



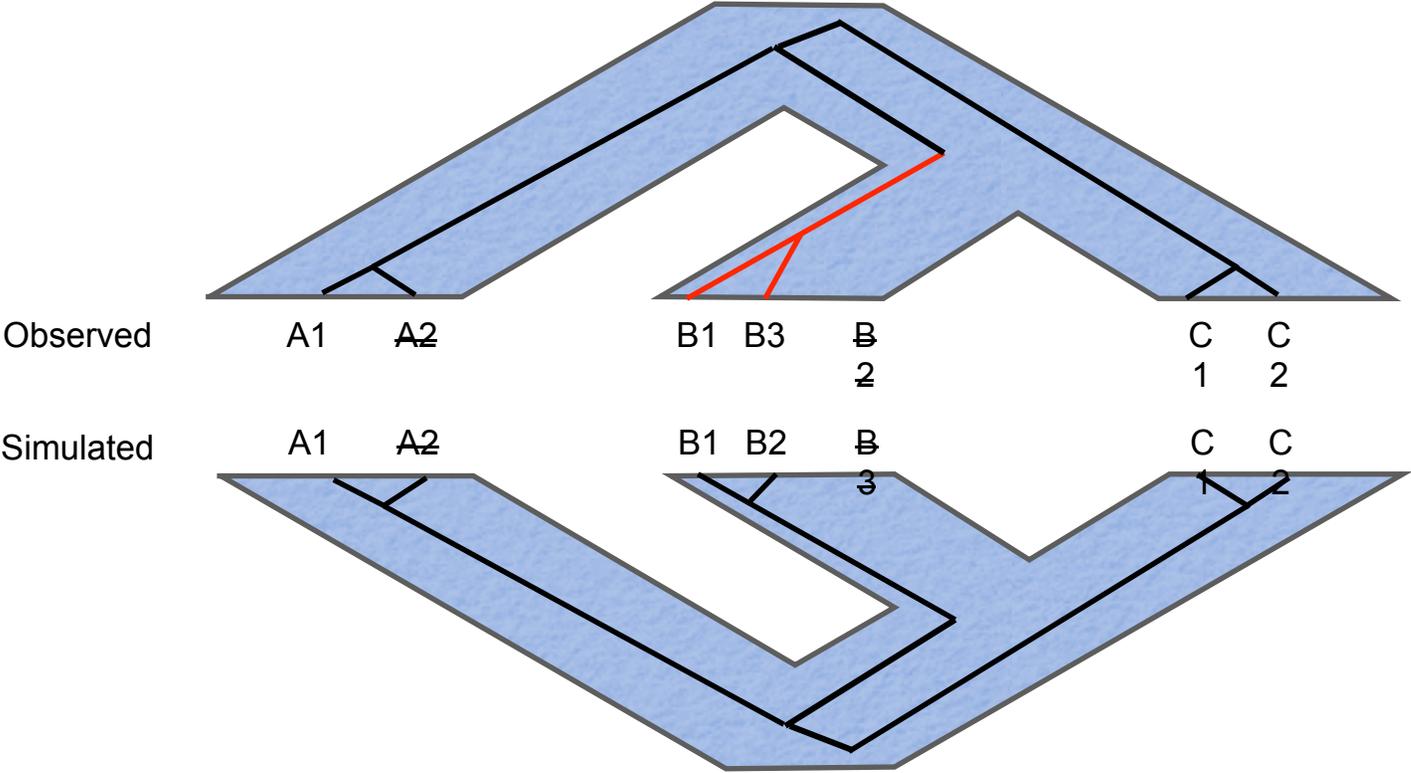
Match based on all possible labeling, then correct for this  
i.e., three possible permutations, so if there is a match divide by 3 to get probability

# Polytomies treated as soft in gene trees (optional)



Match based on all possible resolutions, then correct for this

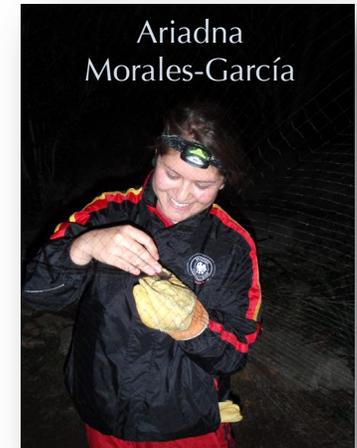
# Subsampling reduces gene tree space and sampling variance





# Custom sequence capture probes synthesized by MYcroarray.

	Sequenced Exons	Ultra Conserved Elements	Anonymous loci	Total / Sample
<b>Mluc34</b>	1055	15	0	1070
<b>Mluc37</b>	2067	76	23	2166
<b>Mevo3</b>	2814	180	31	3025
<b>Mevo4</b>	1184	42	6	1232
<b>Mevo5</b>	3824	317	40	4181
<b>Mvol6</b>	3964	309	37	4310
<b>Mvol7</b>	3220	229	31	3480
<b>Mvol8</b>	3645	180	25	3850
<b>Average</b>	<b>2721.6</b>	<b>168.5</b>	<b>24.1</b>	<b>2914.3</b>



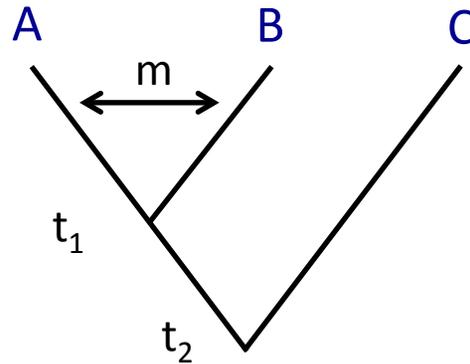
**MYcroarray**

Custom molecular baits,  
probes and building blocks



How does PHRAPL perform?

# Analyzing simulated data

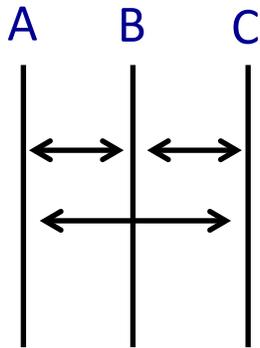


$m = 0$  to  $0.5$

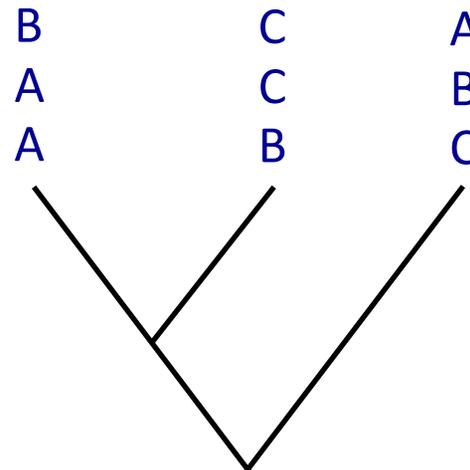
$t_1 = 0.5$  to  $2$

$t_2 = 1.25$  to  $4$

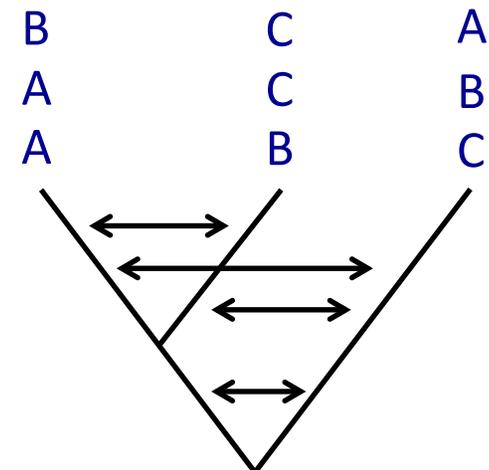
migration only



isolation only

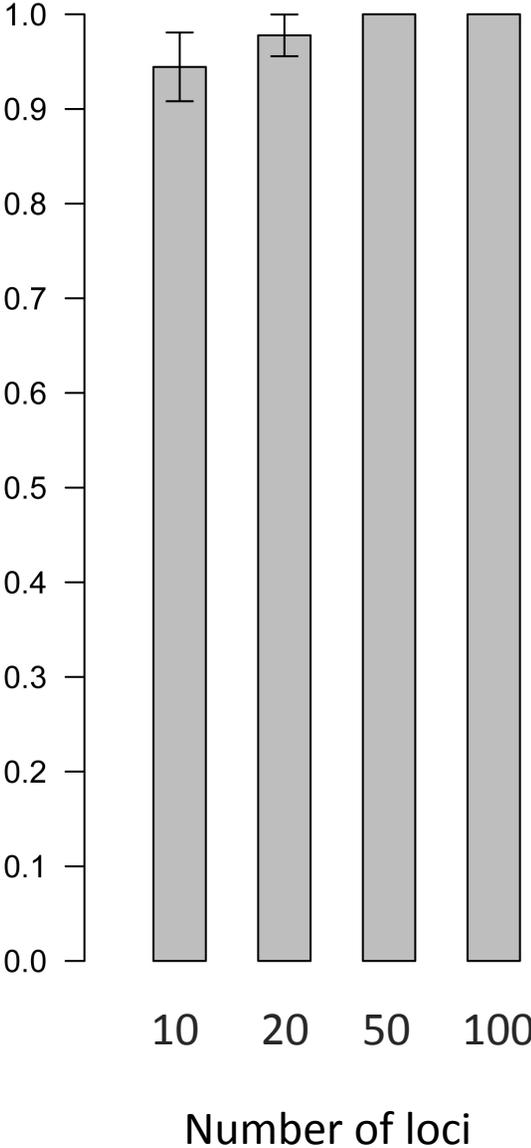


isolation + migration



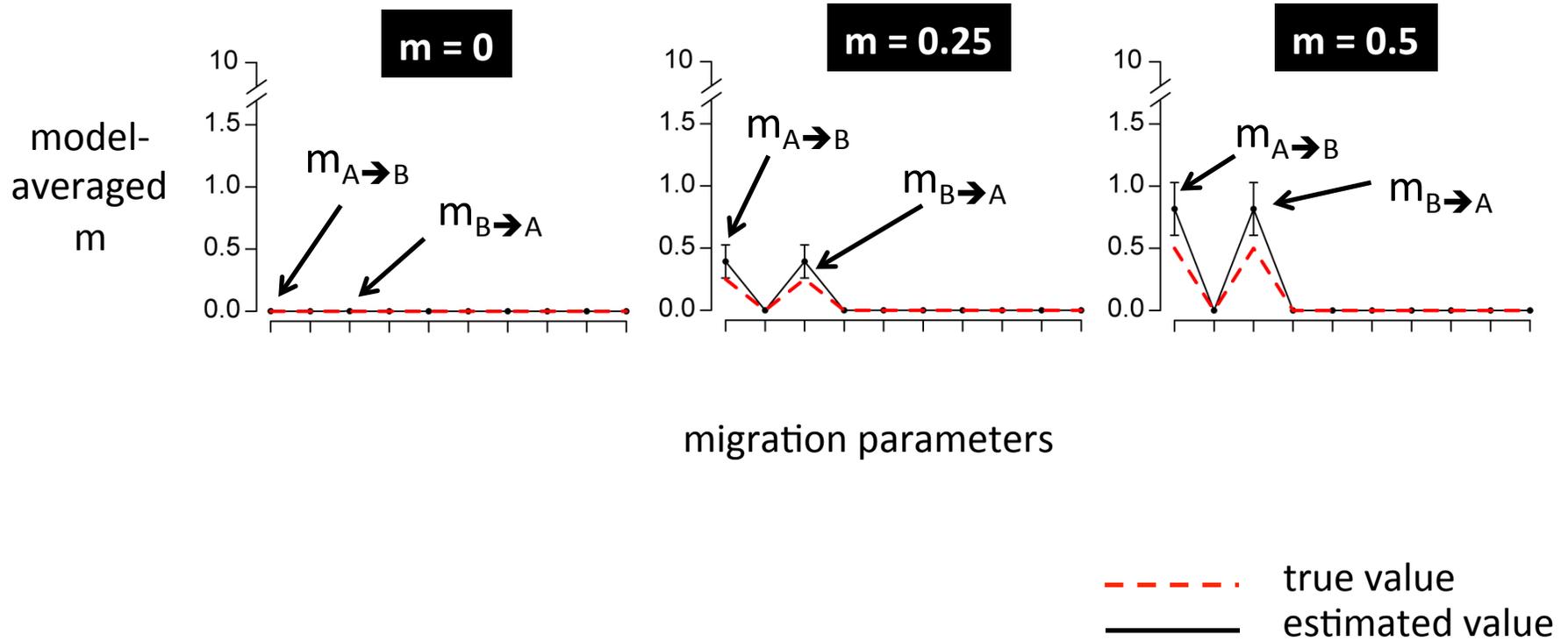
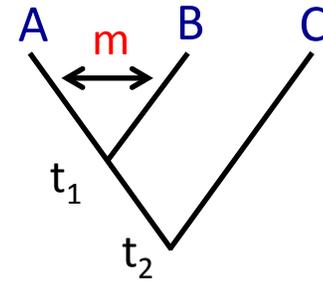
# Analyzing simulated data

Proportion of analyses where top AIC model = true model



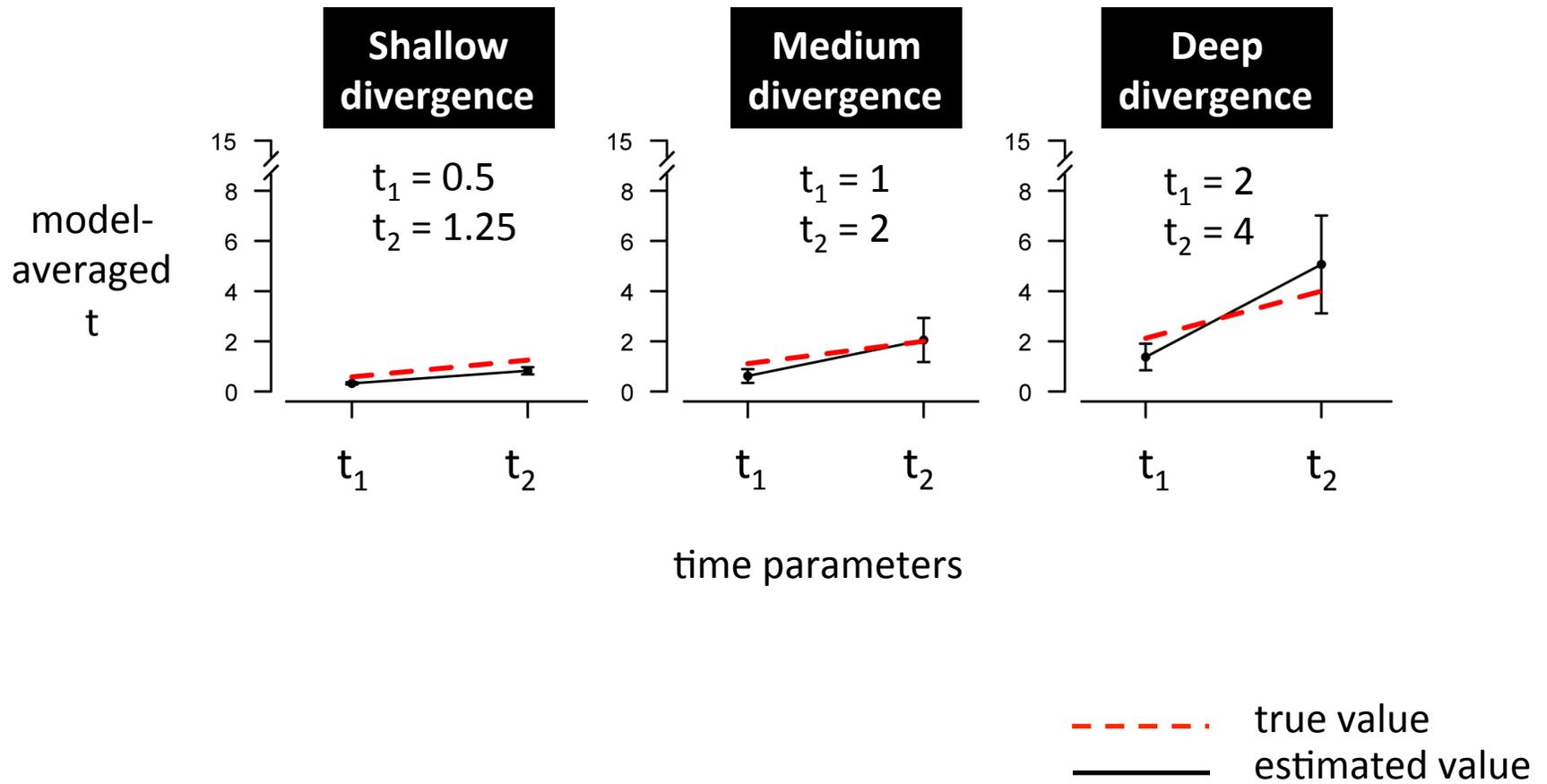
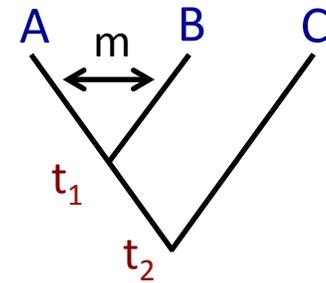
# Analyzing simulated data

## Migration parameter estimation

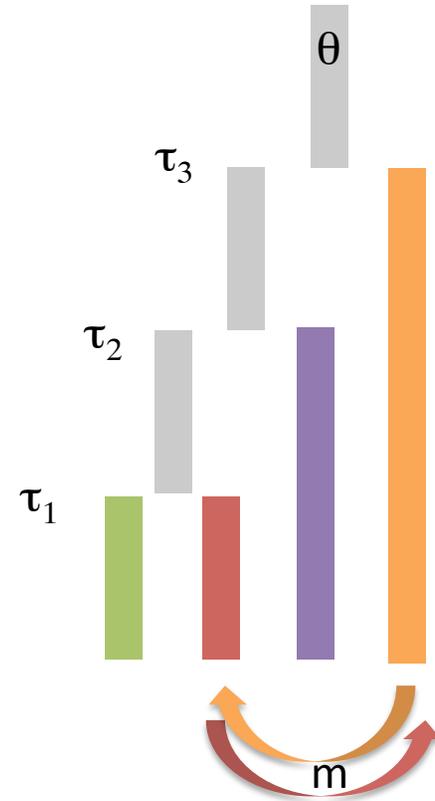
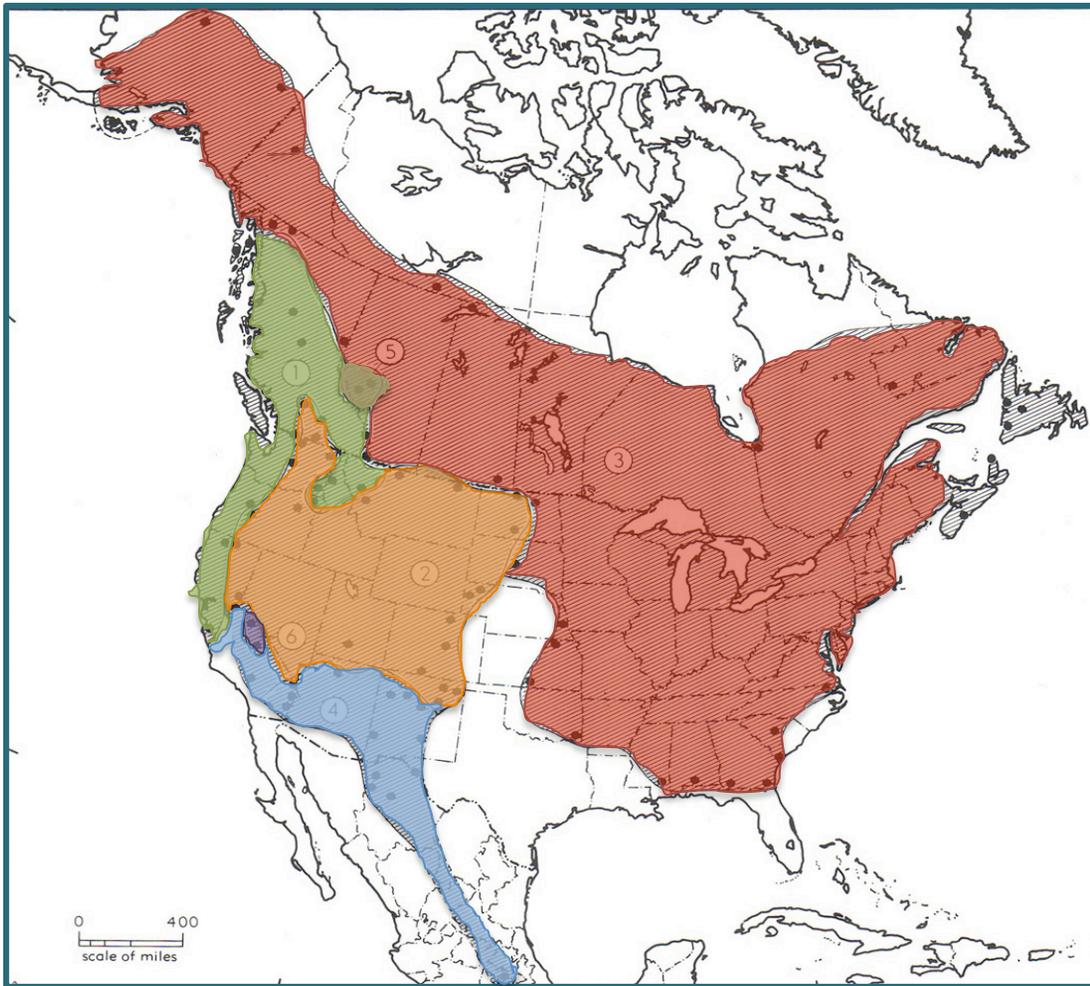


# Analyzing simulated data

## Time parameter estimation



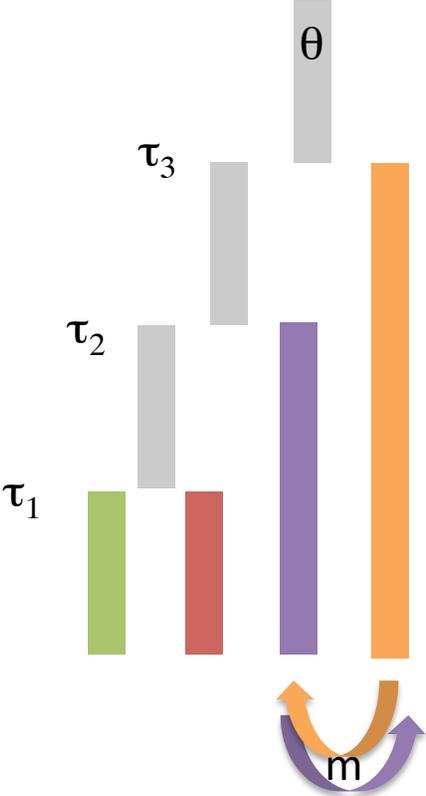
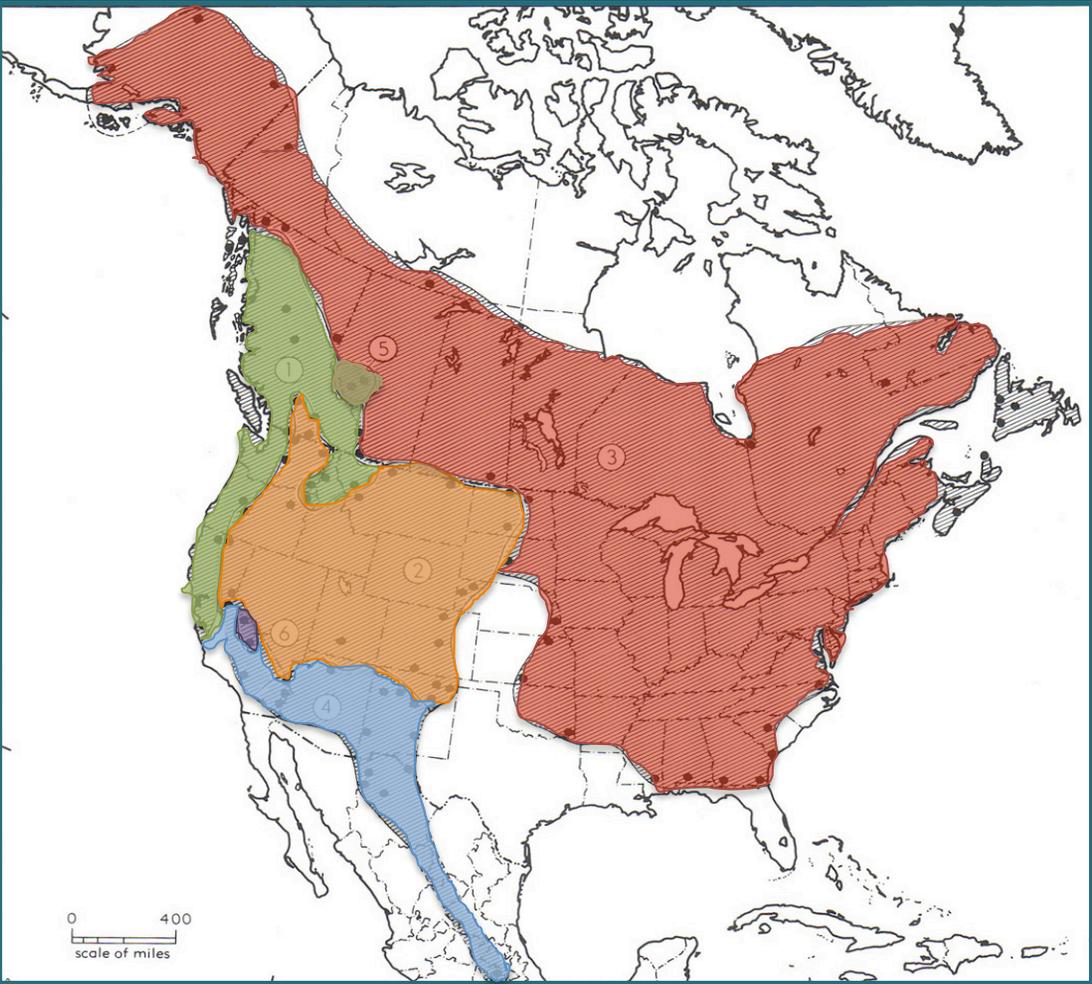
# Little brown bat subspecies (*Myotis lucifugus*)



$$w_{4101} = 0.405$$

- M.I. alacensis* (Hall 1981)
- M.I. carissima*
- M.I. relictus*
- M.I. lucifugus*
- M.I. pernox*

Little brown bat subspecies (*Myotis lucifugus*)



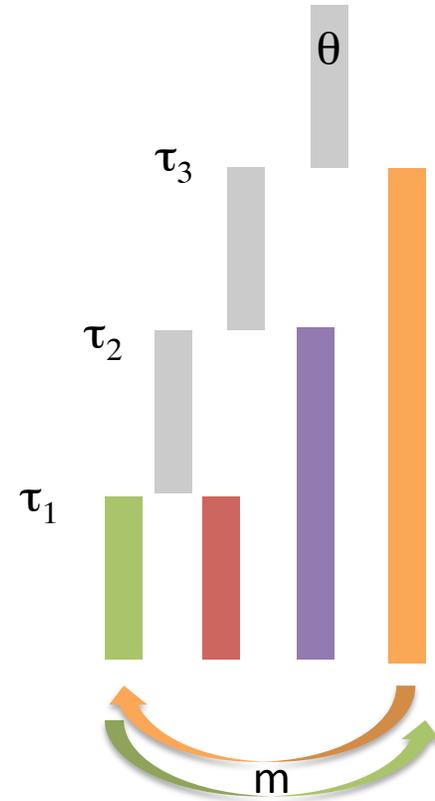
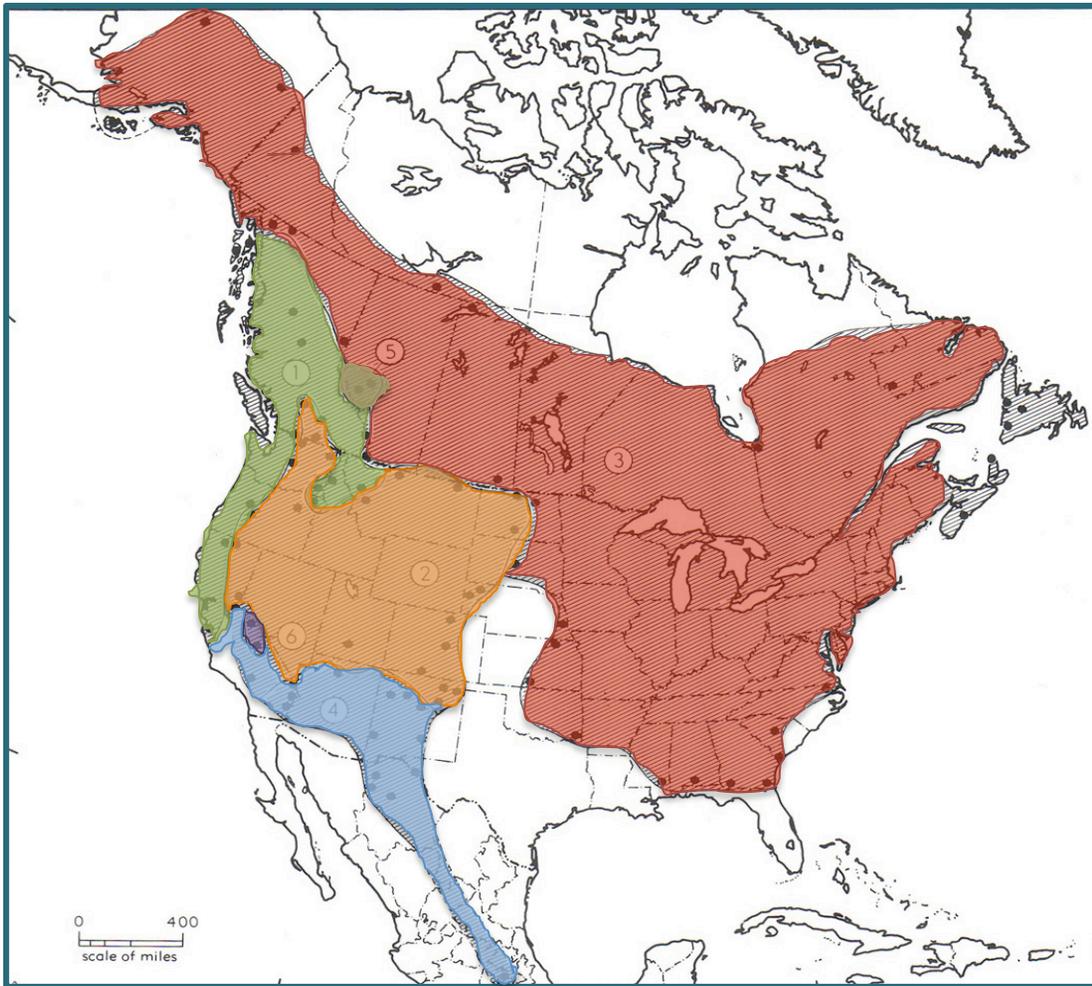
$w_{4113} = 0.352$

- M.I. alacensis*
- M.I. carissima*
- M.I. relictus*
- M.I. lucifugus*

(Hall 1981)

*M.I. pernox*

# Little brown bat subspecies (*Myotis lucifugus*)



$$w_{4162} = 0.095$$

*M.I.alacensis*

(Hall 1981)

*M.I.carissima*

*M.I.relictus*

*M.I.pernox*

*M.I.lucifugus*

Of the 216 models considered by the PHRAPL analysis, the top 5 (which account for  $> 0.98$  of the total model probability):

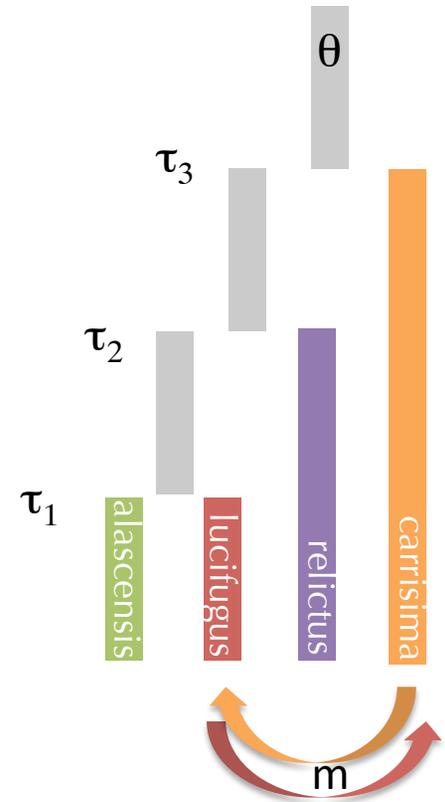
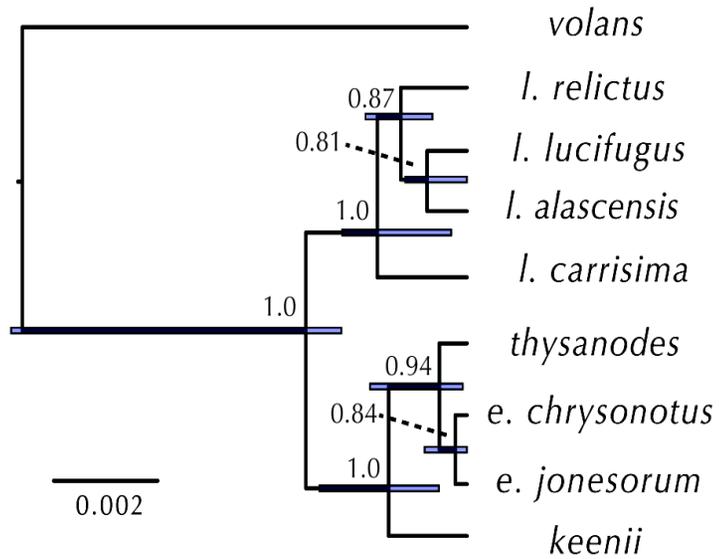
- all have the same topology
- all include migration
- none include a change in population size

migration_model	topology	models	AIC	lnL	$\Delta$ AIC	wi	Cumulative
Mlc-Mll migration	(((a,l)r)c)	4101	123.2317919	-57.61589594	0	0.404526853	0.404526853
Mlc-Mlr migration	(((a,l)r)c)	4113	123.5109318	-57.75546588	0.279	0.351854634	0.756381487
Mla-Mlc migration	(((a,l)r)c)	4162	126.1114743	-59.05573715	2.88	0.095843641	0.852225127
Mlr isolated	(((a,l)r)c)	4104	126.7170439	-59.35852193	3.485	0.07082543	0.923050557
Mla-Mll migration	(((a,l)r)c)	4099	127.0635014	-59.5317507	3.832	0.059544154	0.982594711



but . . .

The topology is the same as that estimated by \*Beast!



$$w_{4101} = 0.405$$



Margaret Koopman  
Yi-Hsin Erica Tsai  
Amanda Zellmer  
**Theresa Thomé**  
**Michael Gruenstaeudl**

Sarah Hird  
Noah Reid  
John McVay  
**Tara Pelletier**  
**Jordan Satler**  
**Ariadna Morales-García**  
**Greg Wheeler**

Danielle Fuselier  
Holly Stoute  
Dan Ence  
Jen Carstens  
Matt Demarest  
Maxim Kim  
**Edwin Rice**  
**Brandon Peterson**

**DEB-1257784**  
**DEB-0918212**  
**DEB-0956069**  
**DEB-1403034**  
**OISE-1118408**

