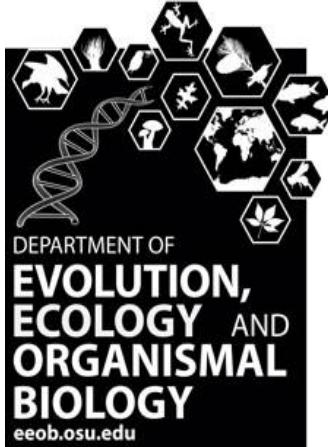
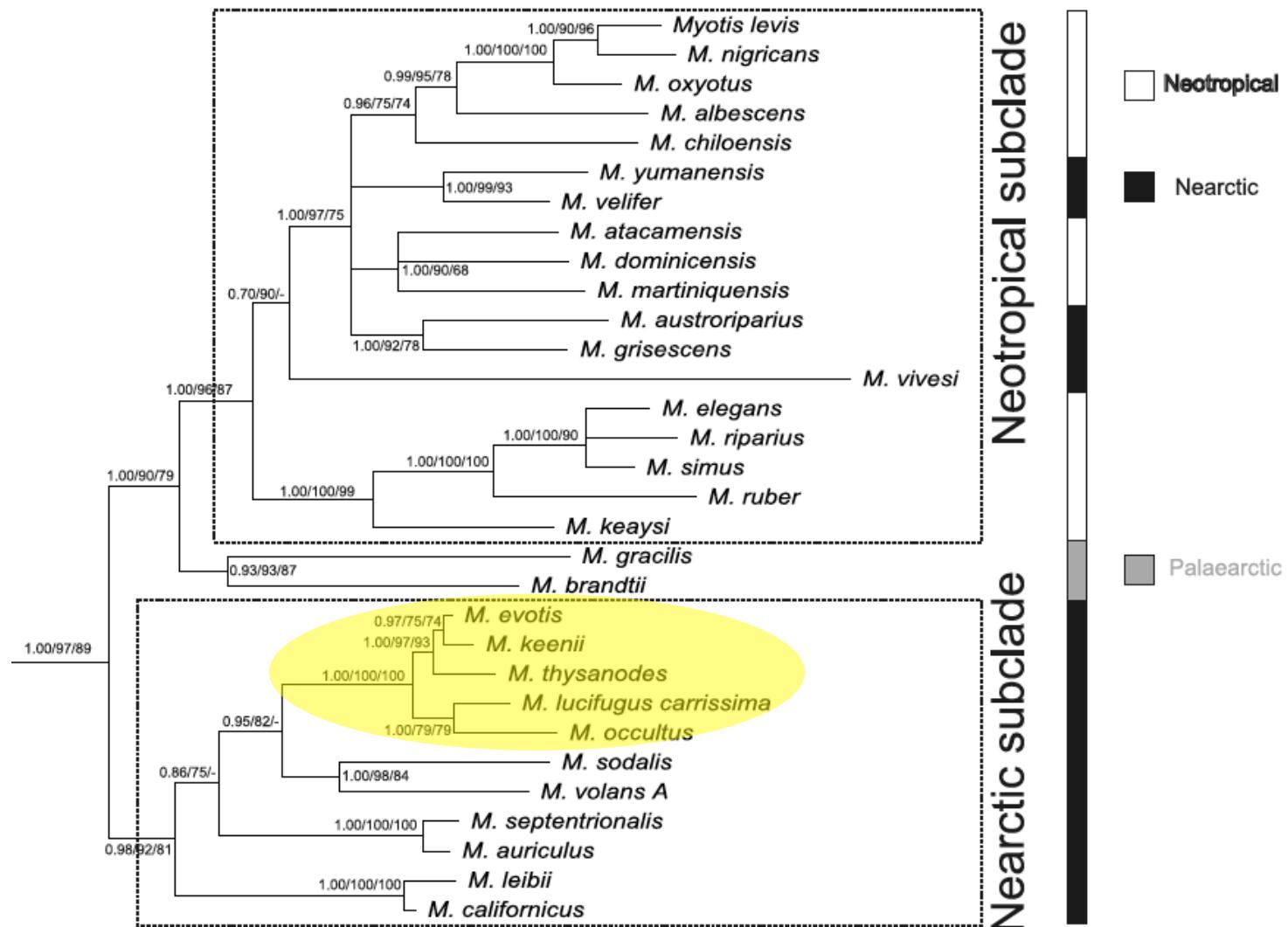


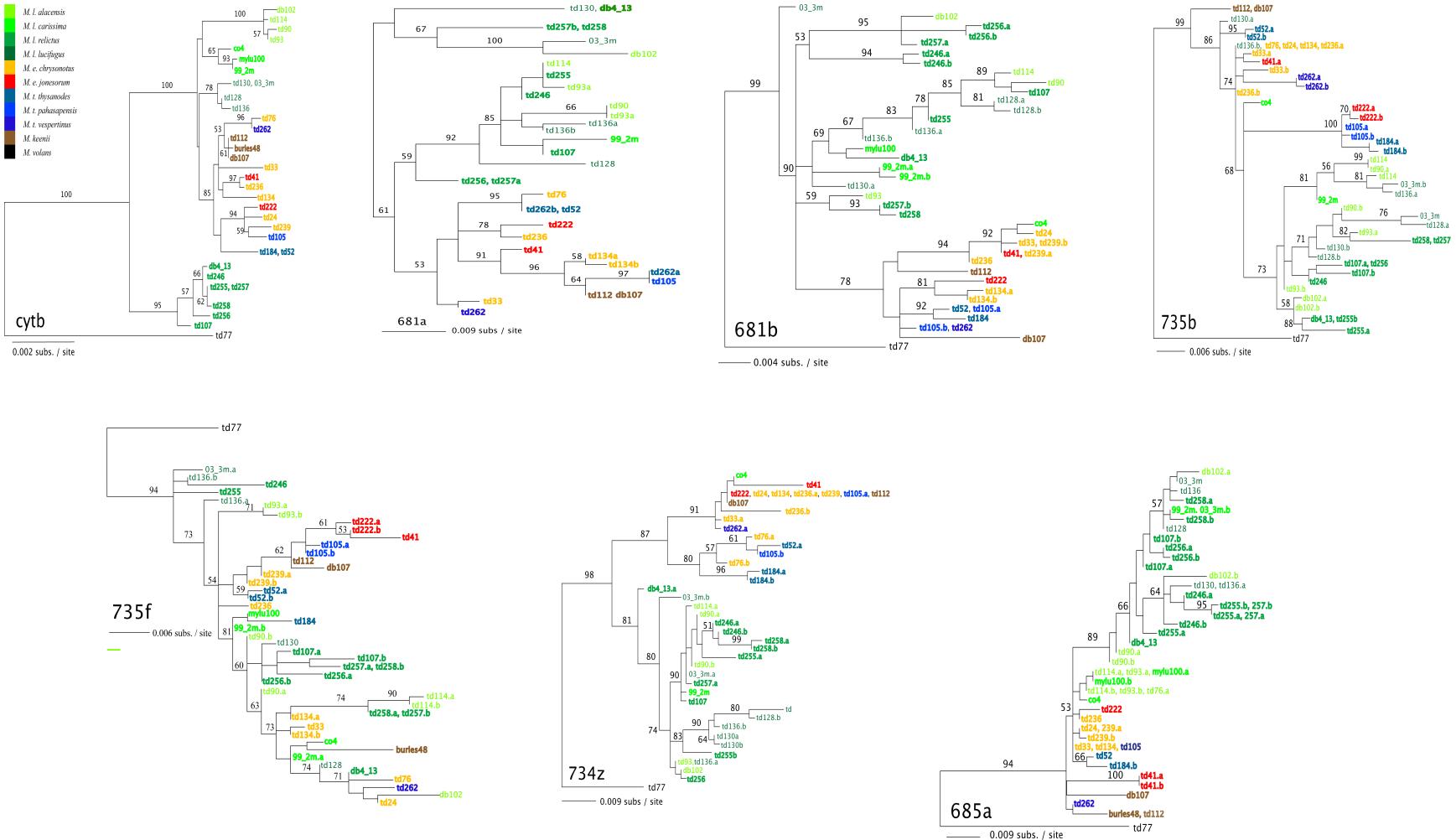
The stories that we tell ourselves about our phylogeographic data.



Bryan Carstens
carstens.12@osu.edu
[@bryancarstens](https://twitter.com/bryancarstens)
http://carstenslab.org.ohio-state.edu/OSU/Carstens_Lab.html







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^{1,*} AND TANYA A. DEWEY²

¹Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70808, USA; and ²Department 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.

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^{1,*}, Bryan C. Carstens² and L. Lacey Knowles³

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

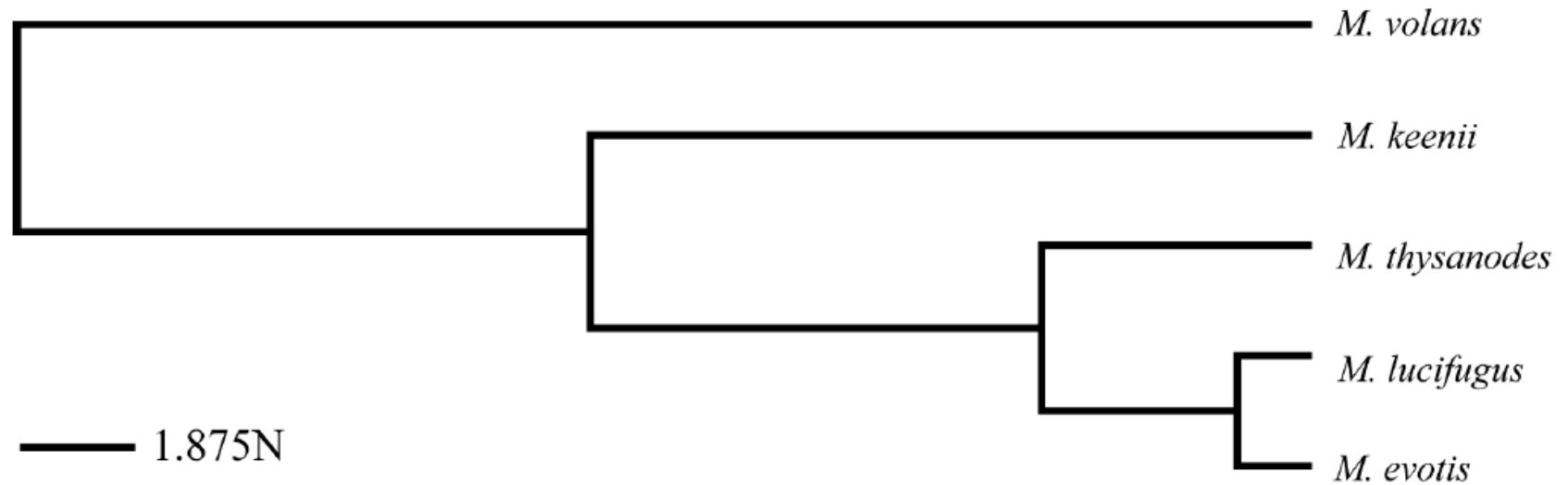
³Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

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

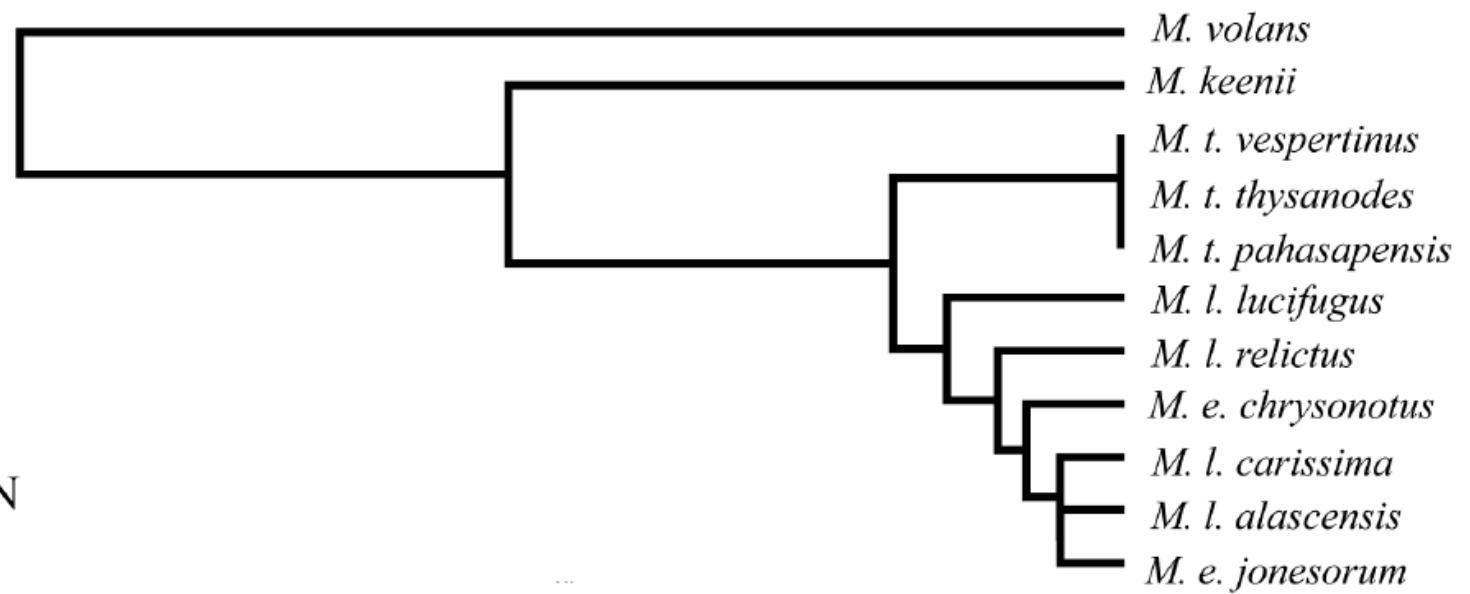
Associate Editor: Martin Bishop

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

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



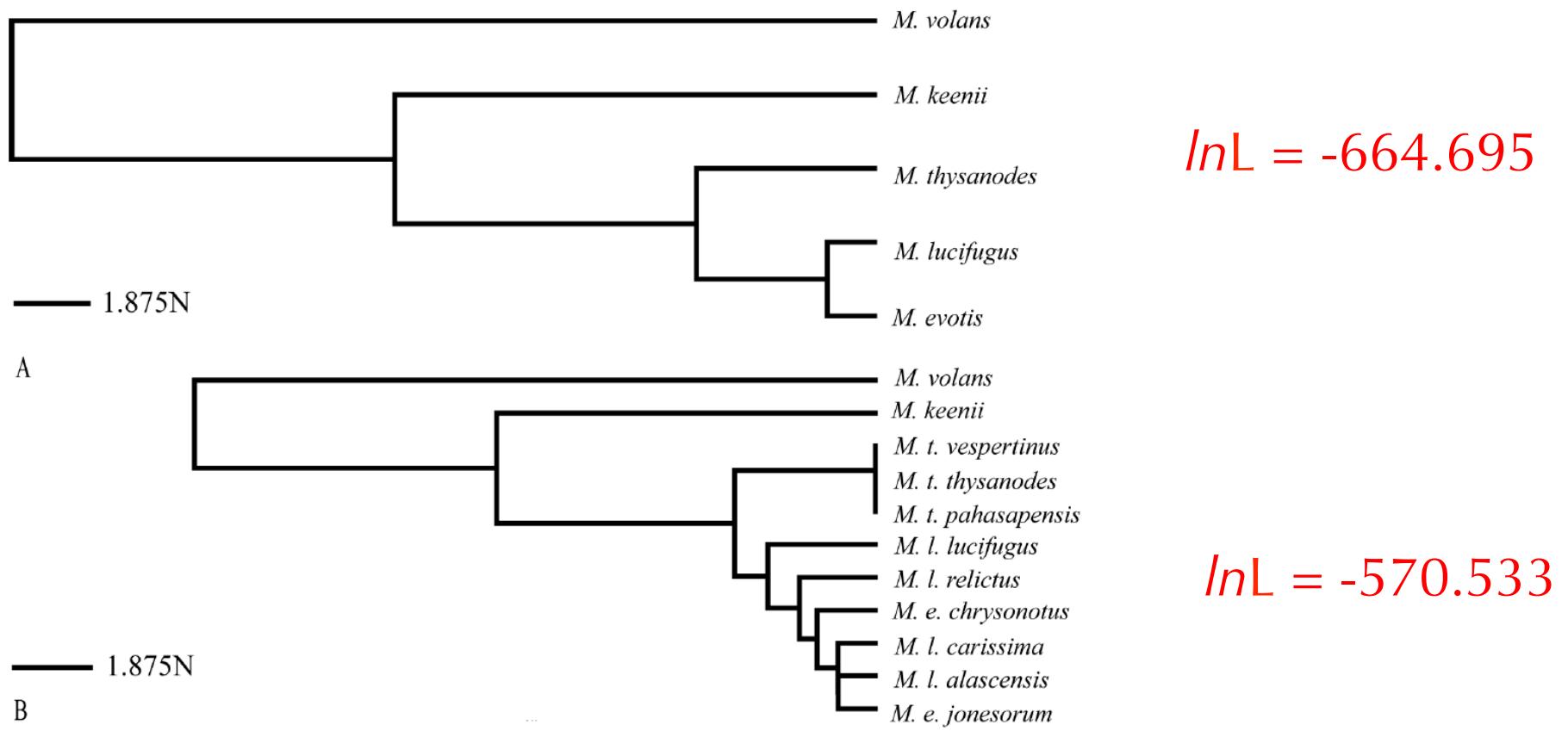
A



B

Carstens & Dewey, 2010

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



Carstens & Dewey, 2010

Permutation	InL	k	AIC	Δi	L(M d)	wi	Permutation	InL	k	AIC	Δi	L(M d)	wi
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-570.53259	8	1157.05118	0	1	0.06975036	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.19963	7	1240.39995	83.174746	7.54153E-37	5.1668E-37
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-570.53259	9	1159.06205	1.99976	0.135512066	0.009087151	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-611.47603	9	1240.915296	83.459266	3.8465E-37	2.6399E-37
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-570.53260	9	1159.06136	1.99995	0.135514228	0.009267936	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-611.41373	7	1242.27946	85.162766	1.0334E-37	7.67677E-38
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-570.53279	9	1159.05131	1.99997	0.135513943	0.009267948	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-611.1199	8	1241.2398	85.17462	1.0333E-37	6.9331E-38
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-570.37547	7	1166.750914	3.685754	0.025679709	0.007173575	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.11994	8	1242.2398	85.17474	1.62117E-37	6.9292E-38
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-570.53261	10	1161.06512	3.99942	0.138116780	0.023540014	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-611.1994	8	1242.39998	85.174716	1.62117E-37	6.9271E-38
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-573.37539	8	1162.75079	5.0551	0.023394462	0.002334479	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.11933	9	1244.239966	87.174686	1.3021E-37	9.4630E-39
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-573.37545	8	1162.75097	5.05569	0.02339419	0.002334599	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-612.94514	6	1253.89020	96.803948	9.9134E-43	6.2255E-43
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-573.37546	8	1162.75084	5.05574	0.023394143	0.002334546	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.93485	7	1255.86984	98.803734	1.2055E-38	6.4561E-44
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-573.37547	8	1162.75085	5.05574	0.023394156	0.002334546	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.94469	7	1255.86994	98.803804	1.2055E-43	6.4571E-44
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-576.09799	6	1165.39759	8.33248	0.02564529	0.009164752	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-612.93449	7	1255.86998	98.803818	1.2055E-43	6.4520E-44
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-576.58139	7	1171.86218	16.99798	1.021318E-10	6.99797E-08	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.54948	8	1257.86936	106.803788	1.6032E-44	1.1403E-44
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-576.58103	8	1171.86204	16.99884	1.020376E-10	6.99887E-09	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-612.57071	6	1263.54169	105.470214	1.5670E-46	1.6668E-46
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-576.58107	8	1171.862142	16.99902	1.020376E-10	6.99915E-09	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.570685	7	1264.54137	107.474319	2.1079E-47	1.4491E-47
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-579.59179	8	1175.162150	18.09678	1.3022E-08	9.4035E-08	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-625.27072	7	1264.541448	107.476268	2.1069E-47	1.4427E-47
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-579.59181	8	1175.162164	19.96458	2.5564E-09	1.7500E-09	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-625.24881	7	1264.541444	107.476284	2.1069E-47	1.4427E-47
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-579.59164	9	1177.162120	20.96948	1.7031E-09	1.2301E-09	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-626.69973	6	1265.869146	108.802966	5.6155E-48	3.7154E-48
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.41915	7	1178.01013	21.76665	3.5232E-10	2.4232E-09	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.70717	8	1265.841434	109.470254	2.8514E-48	1.9526E-48
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.43975	7	1178.010150	21.76628	3.1406E-10	2.3486E-09	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-612.94961	7	1267.890242	116.802942	7.3439E-49	5.0207E-49
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.43979	7	1178.010158	21.76608	3.4935E-10	2.3415E-09	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.66951	7	1267.890183	116.802922	7.3438E-49	5.0203E-49
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.43980	7	1178.010160	21.76624	3.4929E-10	2.3415E-09	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-613.66959	7	1267.890181	116.802918	7.3438E-49	5.0203E-49
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.43987	7	1180.502994	23.50314	6.2004E-11	4.2495E-10	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-614.49944	8	1268.890308	112.832268	9.9711E-50	6.8352E-50
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.43987	8	1180.502996	23.50465	6.2079E-11	4.2505E-10	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-610.81688	6	1272.033778	114.968558	1.1743E-50	6.8416E-51
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.41989	8	1180.503176	23.50666	6.20795E-11	3.5497E-10	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-610.81687	6	1273.62077	116.968094	2.3640E-51	1.7922E-51
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.41990	8	1180.503180	23.50662	6.20795E-11	3.54969E-10	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-610.81685	7	1274.033463	116.968083	2.0846E-51	1.6676E-51
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.43987	8	1180.503187	23.50674	6.20795E-11	3.54925E-10	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-610.81685	7	1274.033469	116.968075	2.0835E-51	1.6676E-51
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.43987	8	1180.503188	23.50678	6.20795E-11	3.54935E-10	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-610.81685	7	1274.033476	116.968073	2.0835E-51	1.6676E-51
Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-582.43987	8	1181.503474	23.50358	6.3947E-11	3.54935E-10	Mp_Mt_Mr_Mh_Ma_Ml_Mt_Mr_Mq_Ma	-610.830975	7	1275.66117	116.968777	2.1361E-52	

Information theoretic metrics for 150 models of lineage composition

Lineage composition	-lnL	k	AIC	Δi	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 . . .

NSF DEB-0918212

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

<https://spedestem.osu.edu>

The screenshot shows a web browser window titled "SpedeSTEM". The address bar displays the URL <https://spedestem.osu.edu>. The page header includes the "SpedeSTEM" logo, a "Run SpedeSTEM" button, a "Resources" link, a "Contact" link, and a "Sign in" button. Below the header, a main content area features a section titled "Species delimitation using Maximum Likelihood". This section contains a detailed description of the software's purpose and functionality. At the bottom of the page, there are logos for "OHIO STATE UNIVERSITY" and "National Science Foundation (NSF)", along with the text "Funded by NSF DEB 0918212".

Species delimitation using Maximum Likelihood

spedeSTEM is a program that delimits species using maximum likelihood and information theory. Specifically, the probabilities of multiple permutations of putative evolutionary lineages are calculated using STEM (Kubatko et al. 2009) and ranked by model probability (see Anderson 2004). spedestem takes as input ultrametric gene trees from multiple loci and an estimate of theta, and returns a table of models ranked by model probability. The web-based software here conducts both discovery and validation analyses, and also generates the set up files and allows the users to subsample alleles from large nexus files. spedestem does not estimate gene trees; for this, we suggest PAUP or Garli.

Department of Ecology Evolution and Organismal Biology

National Science Foundation (NSF)
WHERE DISCOVERIES BEGIN

Funded by NSF DEB 0918212



doi:10.1111/j.1558-5646.2012.01640.x

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)

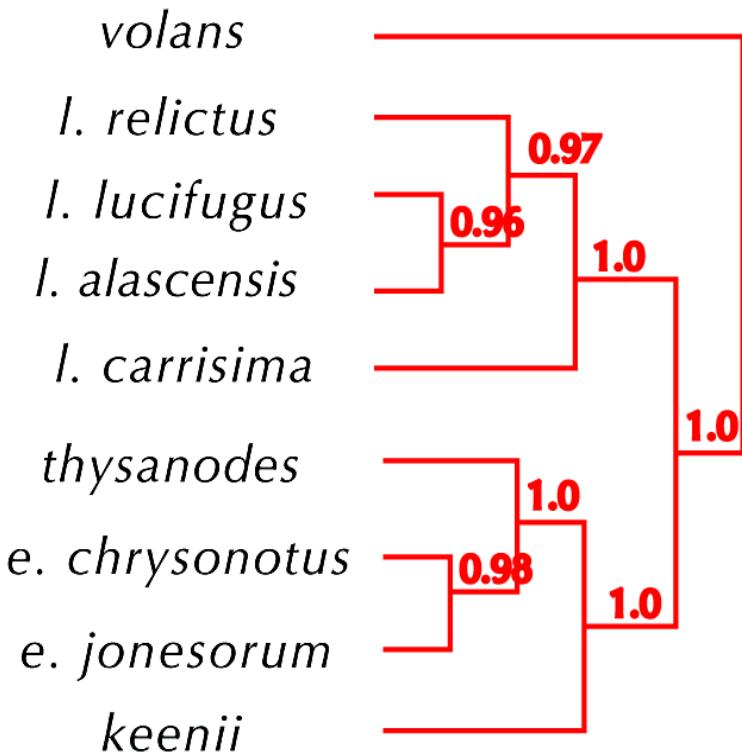
Arley Camargo,^{1,2} Mariana Morando,³ Luciano J. Avila,³ and Jack W. Sites, Jr.¹

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

²E-mail: arley.camargo@gmail.com

³CONICET-CENPAT, Boulevard Almirante Brown 2915, U9120ACD, Puerto Madryn, Chubut, Argentina

Species delimitation?
ABC > BPP > speDeSTEM



BPP (Yang & Rannala 2010)

MOLECULAR ECOLOGY

Molecular Ecology (2013) 22, 4369–4383

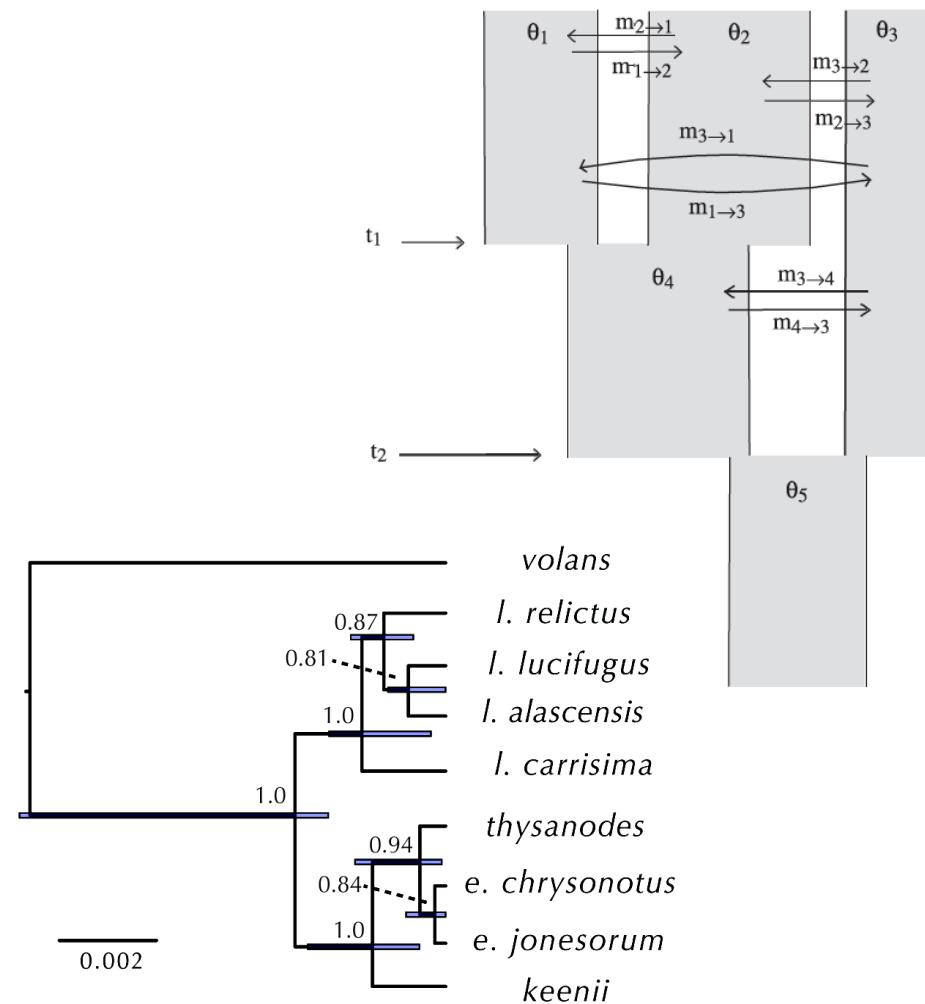
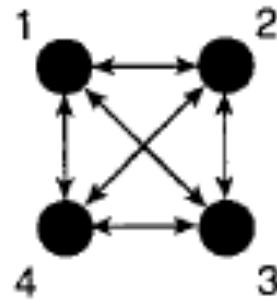
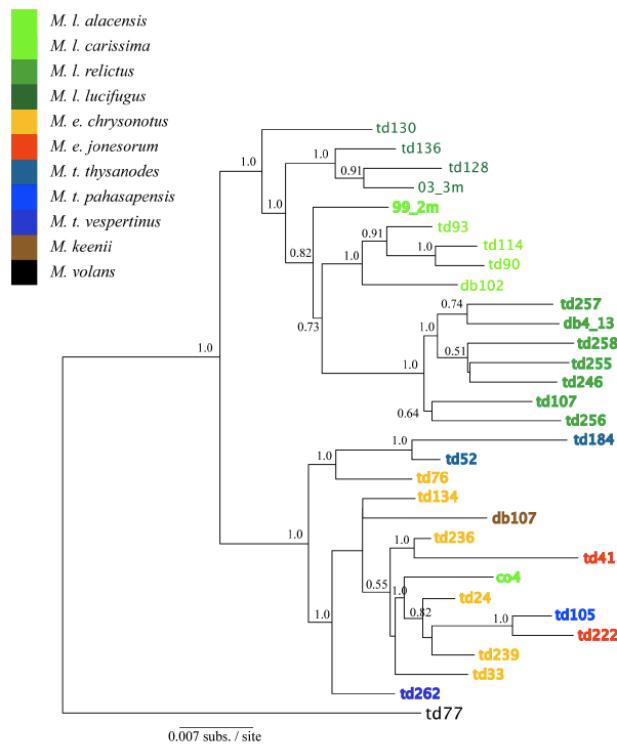
doi: 10.1111/mec.12413

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

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

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

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^{1,*}, Bryan C. Carstens² and L. Lacey Knowles³

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

³Ecology 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

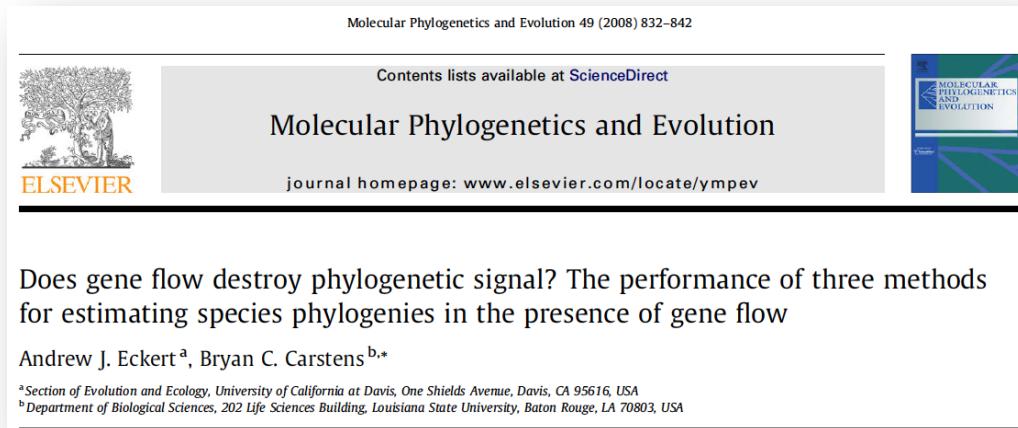
$$L(S, \tau) = \prod_{j=1}^N f(g_j | S, \tau)$$

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

Assumptions of model

- θ 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



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
 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^{1,*}, REBECCA B. HARRIS¹, BRUCE RANNALA^{2,3}, AND ZIHENG YANG^{3,4}

¹Department of Biology and Burke Museum of Natural History and Culture, University of Washington, Seattle, WA 98195 USA;

²Genome Center and Department of Evolution & Ecology, University of California, Davis, CA 95616, USA;

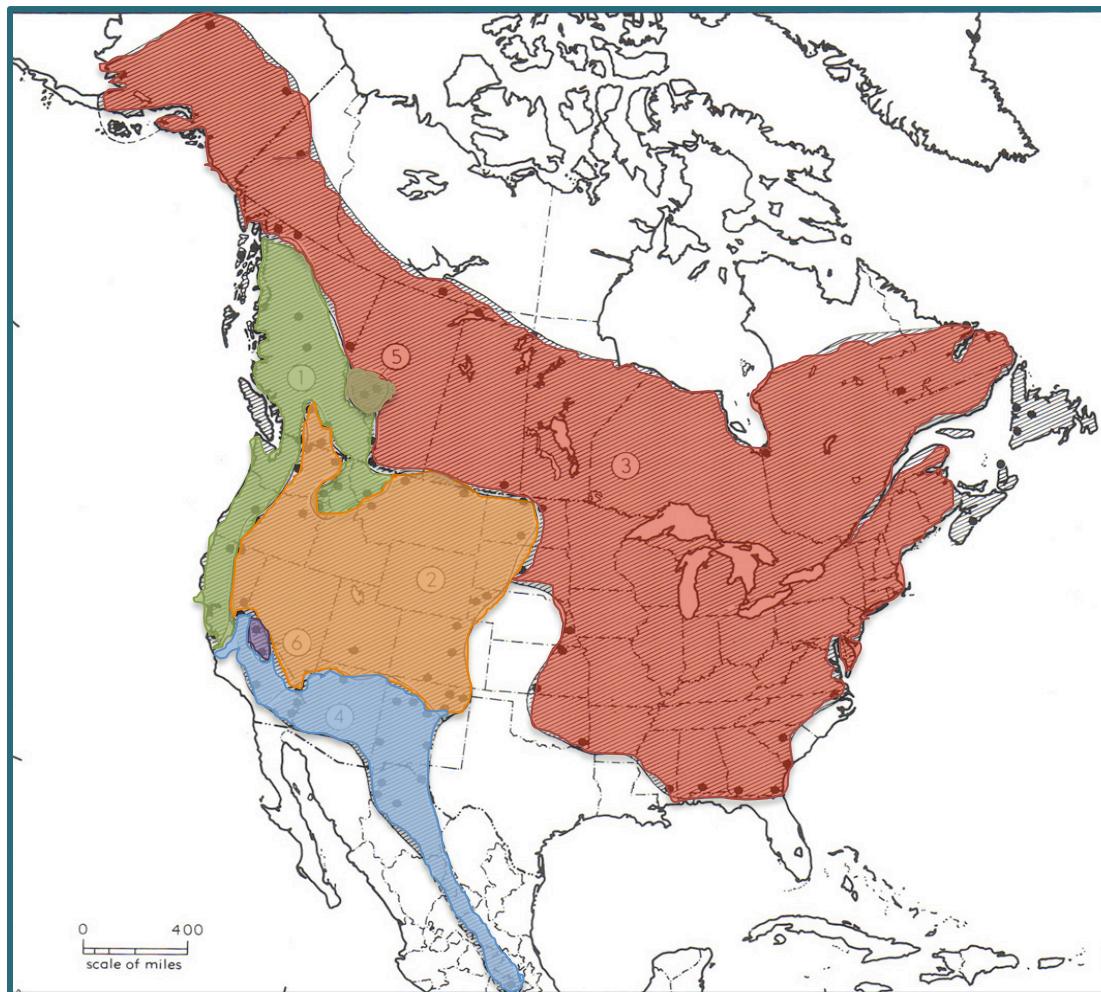
³Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China; and

⁴Department 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@u.washington.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

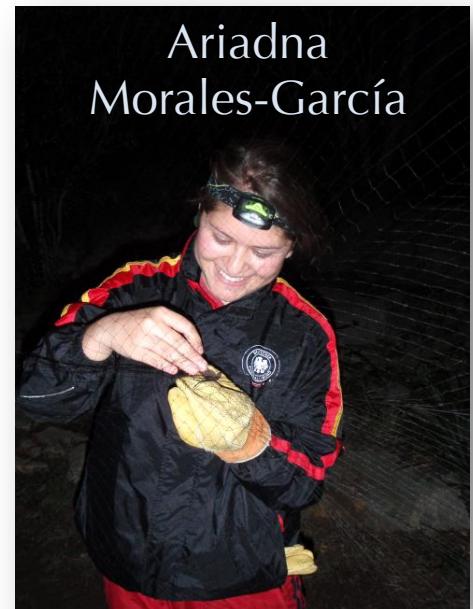
M.l.carissima

M.l.relictus

M.l.lucifugus

(Hall 1981)

Ariadna
Morales-García



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

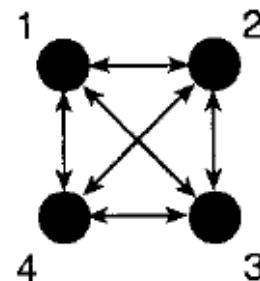
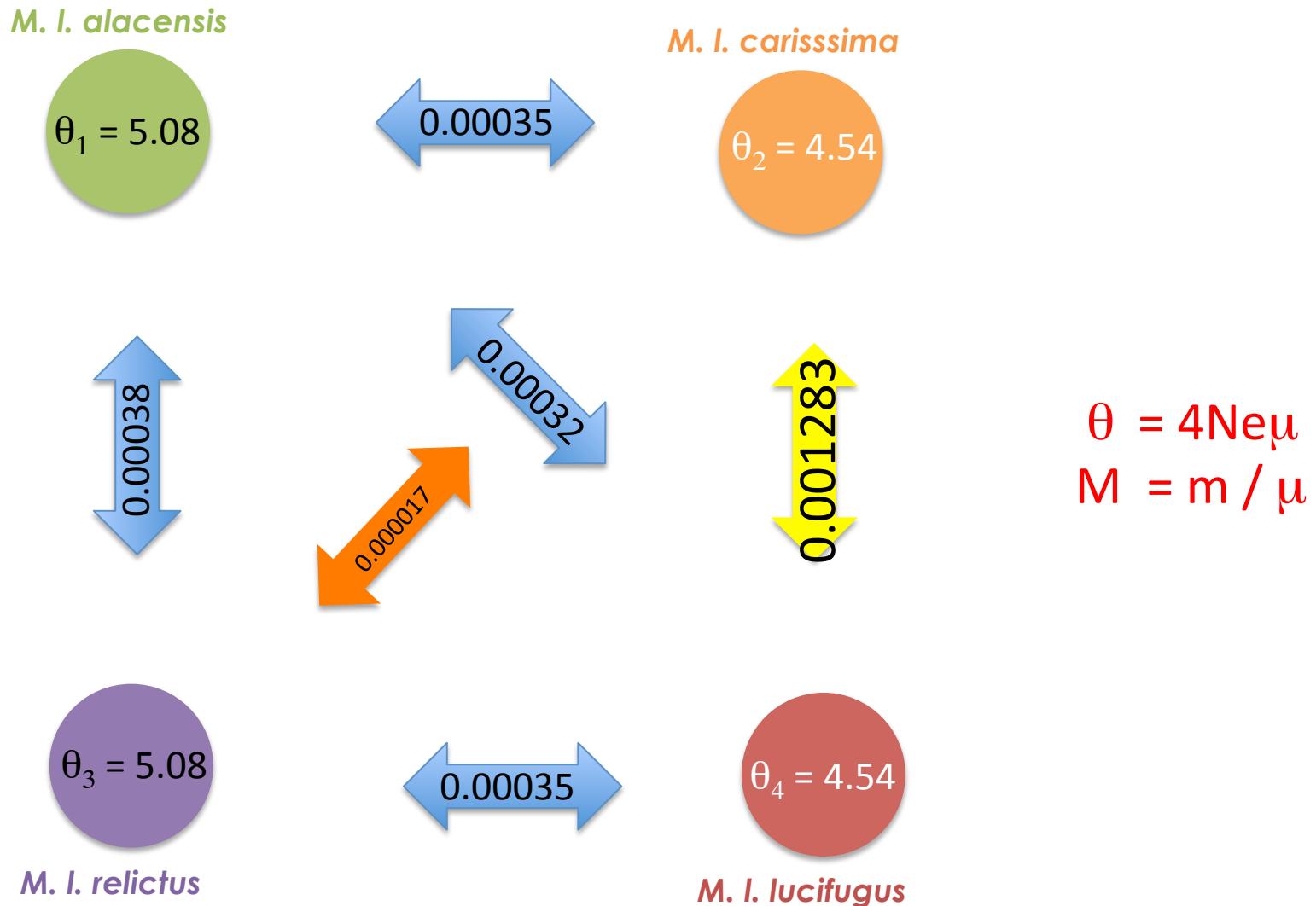


Fig. 1. n -island model with four populations of equal size, exchanging migrants with equal rates.

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



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

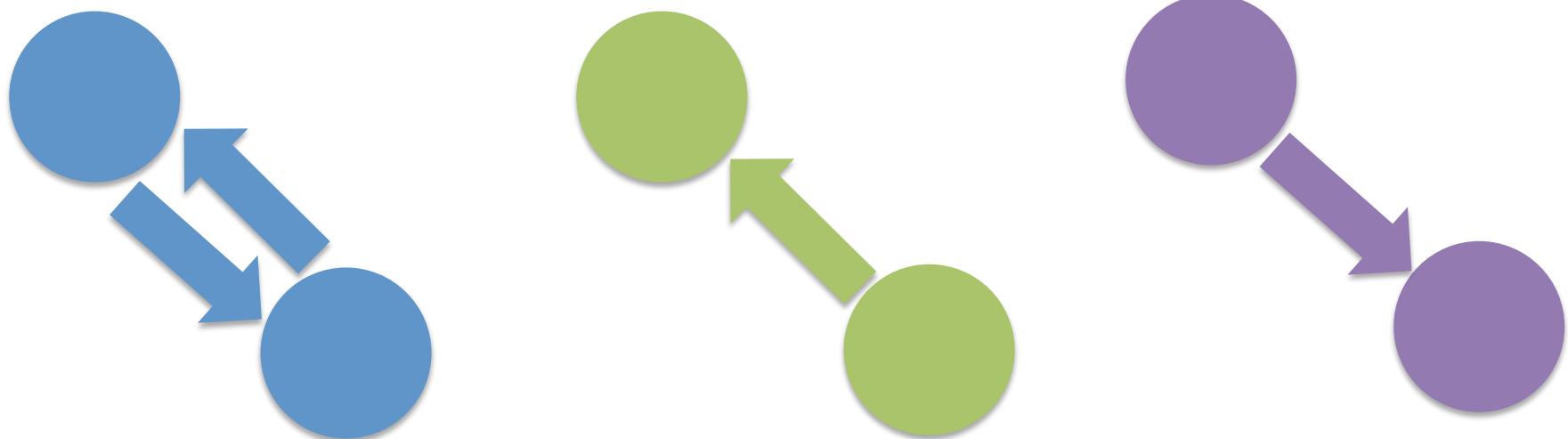
Peter Beerli¹ and Michal Palczewski

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

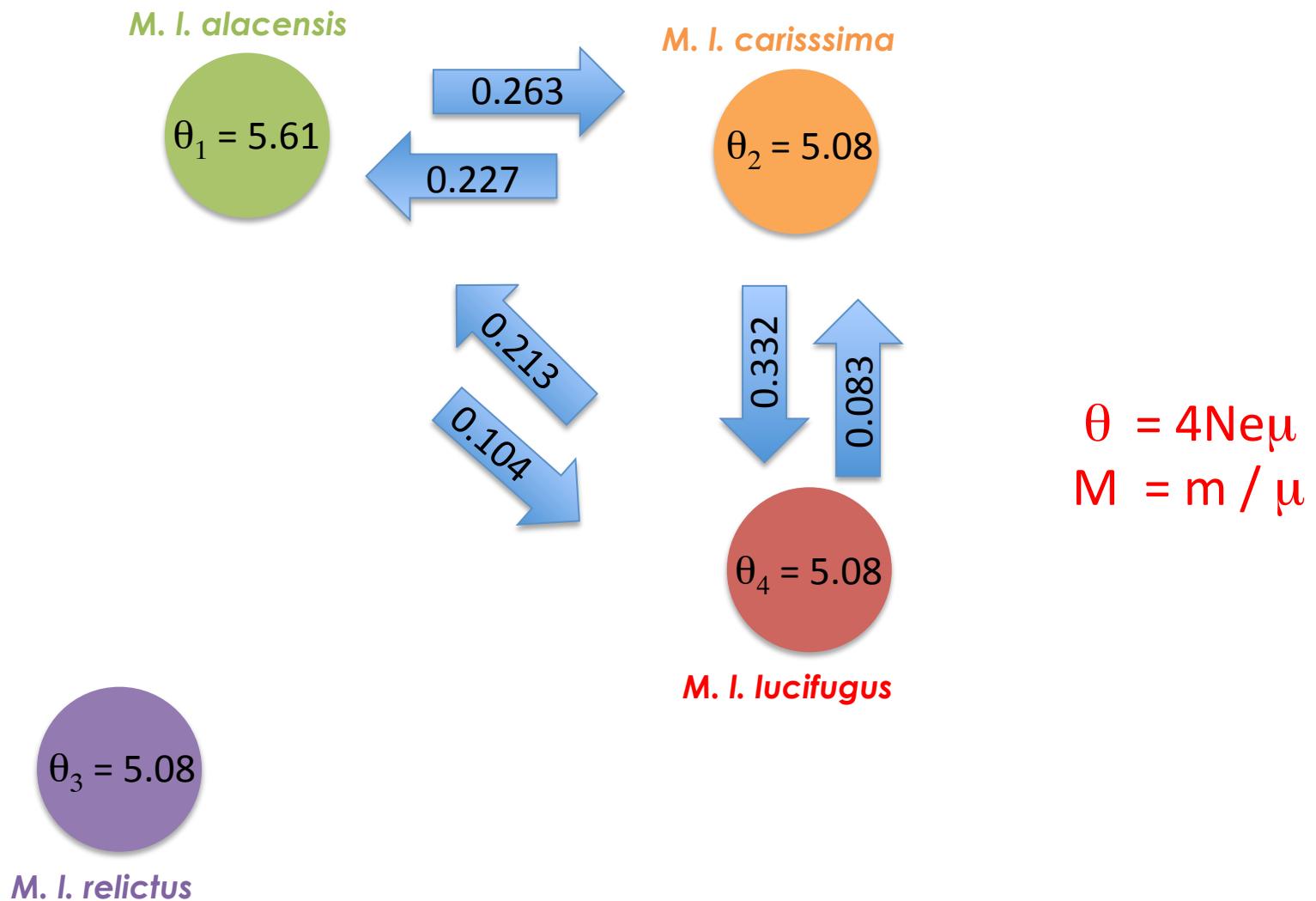
Manuscript received November 27, 2009

Accepted for publication February 17, 2010

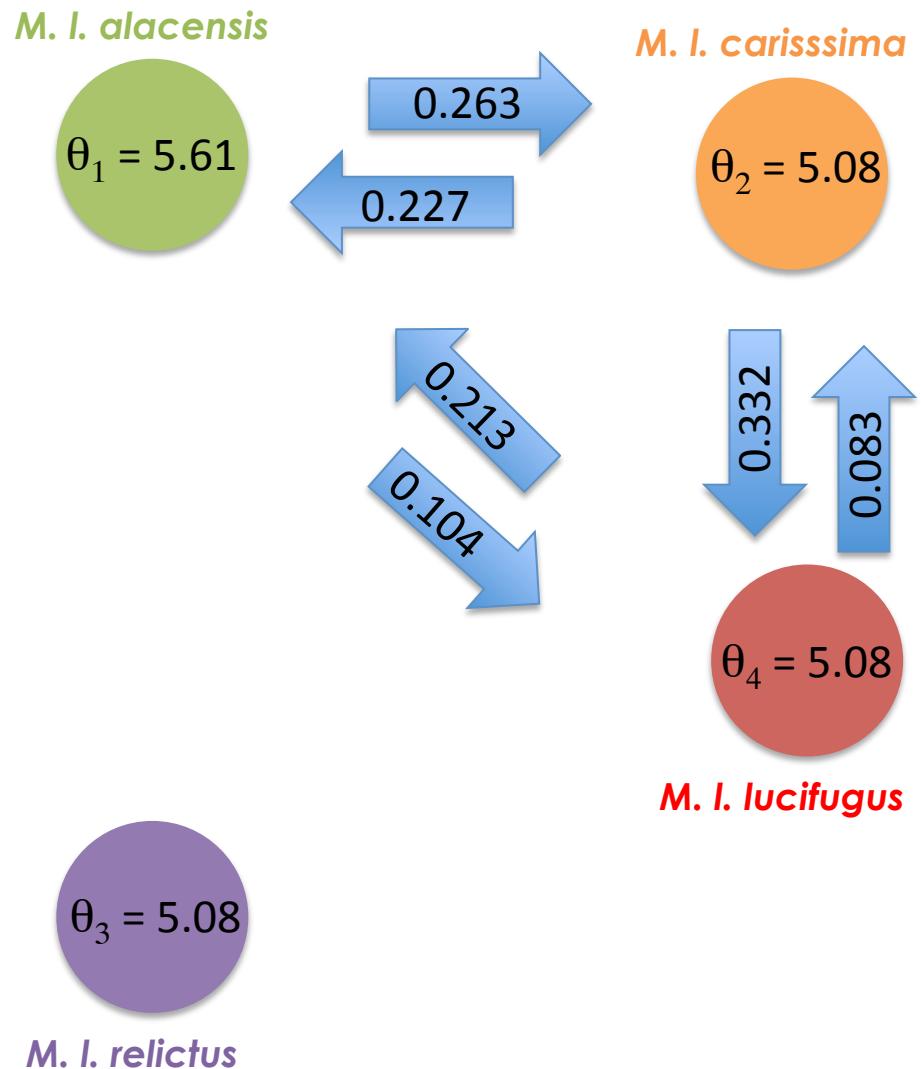
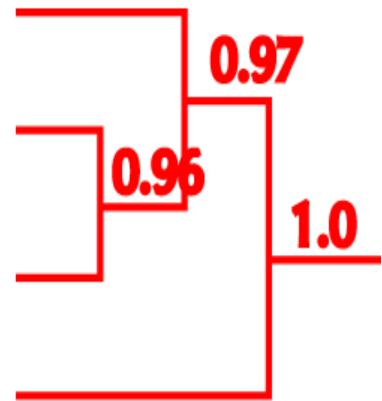
Migrate-n 3.6 has the ability to calculate marginal likelihoods, so migration models can be evaluated using information theory.



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



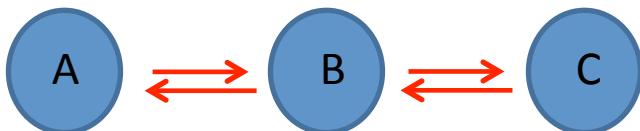
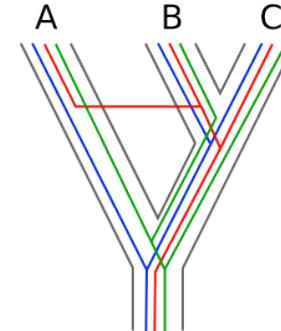
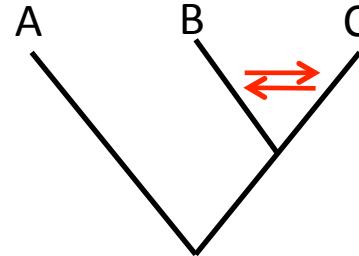
I. relictus
I. lucifugus
I. alascensis
I. carissima



Same data, different interpretations . . .

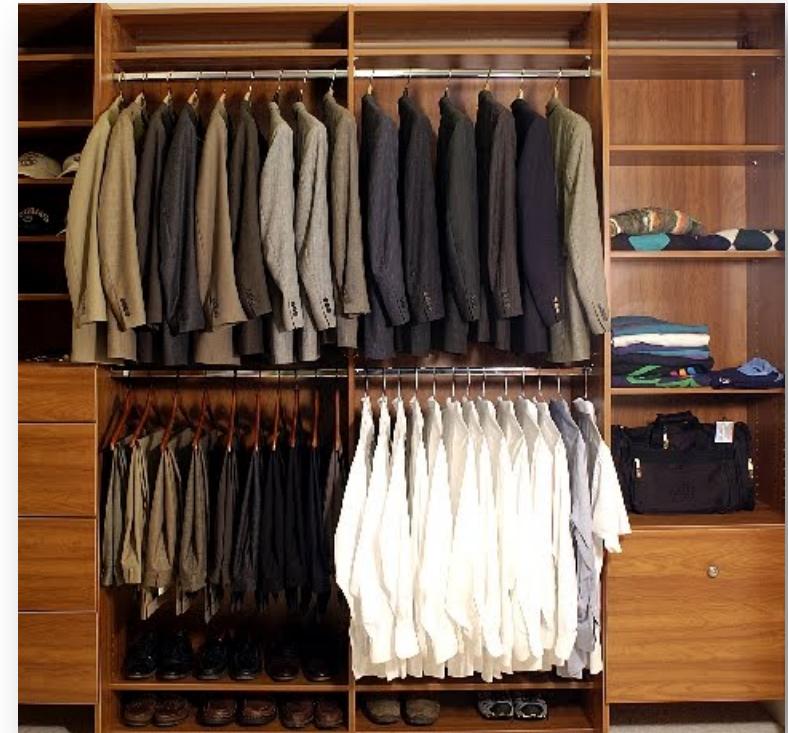
- If we choose a species tree / delimitation approach, we infer that **each** subspecies within *M. lucifugus* is an independent evolutionary lineage (and probably assume that these lineages do not exchange alleles).
- If we choose an *n*-island migration model, we infer that **three of the four** subspecies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

Full likelihood/Bayesian methods can not fully model all evolutionary processes.

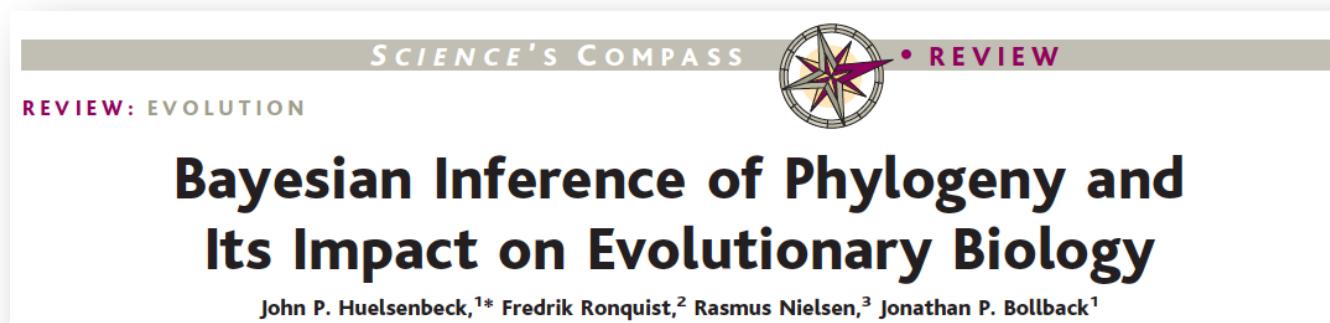
Method	Parameters estimated	Parameters NOT estimated	Model
Migrate-n	m, θ	τ	
*BEAST	topology, τ, θ	m	
IMa2	τ, m, θ	topology	

How do we justify our choices we analyze our data?

- we can assess the fit of the models that we choose
- we can choose the best model among a bunch of choices



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):0.000554):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: 852508374]
begin trees;
tree rep.1 = ((2:0.100000,(4:0.100000,5:0.100000):0.100000):0.100000,3:0.100000,1:0.100000);
tree rep.3000 = ((3:0.000100,(1:0.000100,5:0.000100):0.000100):0.014269,2:0.001265,1:0.000100);
tree rep.30000 = ((((1:0.000100,2:0.000100):0.000100):0.000100):0.000100,3:0.000100,4:0.000100);
tree rep.300000 = (((((1:0.000100,2:0.000100):0.000100):0.000100):0.000100):0.000100,5:0.000100);
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,(3:0.000789,5:0.001397):0.000475):0.013688,2:0.004580,1:0.026280);
tree rep.6000 = ((5:0.002360,(1:0.002026,3:0.000960):0.003821):0.016055,2:0.007722,1:0.011203);
tree rep.7000 = ((6:0.002026,(1:0.001630,3:0.000960):0.003821):0.016055,2:0.007722,1:0.011203);
tree rep.8000 = ((7:0.001825,(1:0.000630,5:0.003030):0.001169,4:0.006023):0.022671,1:0.025134);
tree rep.9000 = ((2:0.000986,(3:0.013872,4:0.005180):0.003140,3:0.003320):0.005159,1:0.003640);
tree rep.10000 = ((3:0.000307,(5:0.003334):0.000640):0.000640,4:0.010198):0.002653,1:0.019618;
tree rep.11000 = ((4:0.000100,(1:0.000100,5:0.000100):0.000100):0.000100,3:0.000100,2:0.000100);
tree rep.12000 = ((2:0.005955,(3:0.005154,(1:0.001154,4:0.005552):0.000527):0.013815,1:0.002150);
tree rep.13000 = (((((1:0.000884,4:0.001974):0.001340,5:0.005650):0.021711,2:0.003141,1:0.004153));
tree rep.14000 = (((((1:0.002721,5:0.003603):0.000965,4:0.002150):0.017912,2:0.002912,1:0.001911));
tree rep.15000 = (((((1:0.003662,4:0.008220):0.001214,3:0.004942):0.003044,2:0.003570,1:0.005086));
tree rep.16000 = ((2:0.001225,(3:0.009145,(5:0.002659,3:0.005159):0.000768):0.003695,1:0.005769));

```

SCIENCE'S COMPASS • REVIEW

REVIEW: EVOLUTION

Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology

John P. Huelsenbeck,^{1*} Fredrik Ronquist,² Rasmus Nielsen,³ Jonathan P. Bollback¹

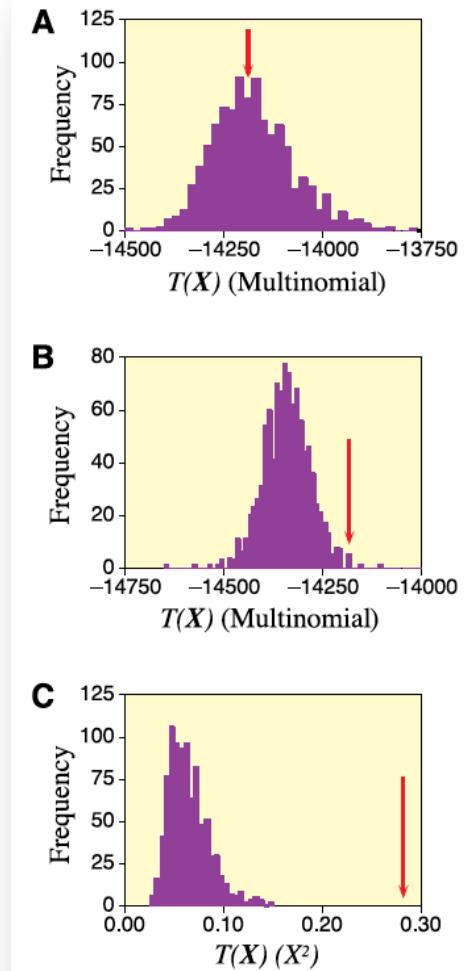


Fig. 3. The posterior predictive distributions for tests of (A) the adequacy of the GTR model, (B) of the adequacy of the Jukes-Cantor model, and (C) the hypothesis of constant nucleotide frequencies over time. The arrows above the distributions show the observed value of the test statistics.

posterior
probability

likelihood

prior

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



Multispecies Coalescent model

$L(g_i) = P(d_i|g_i)$. Likelihood of the data_i given genealogy_i

$L(u_i) = P(g_i|u_i)$. Likelihood of genealogy_i given the molecular clock_i

$L(S) = P(u_i|S_i)$. Likelihood of the molecular clock_i 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^{1,*}, SARAH M. HIRD¹, JEREMY M. BROWN¹, TARA A. PELLETIER², JOHN D. MCVAY¹, JORDAN D. SATLER², AND BRYAN C. CARSTENS²

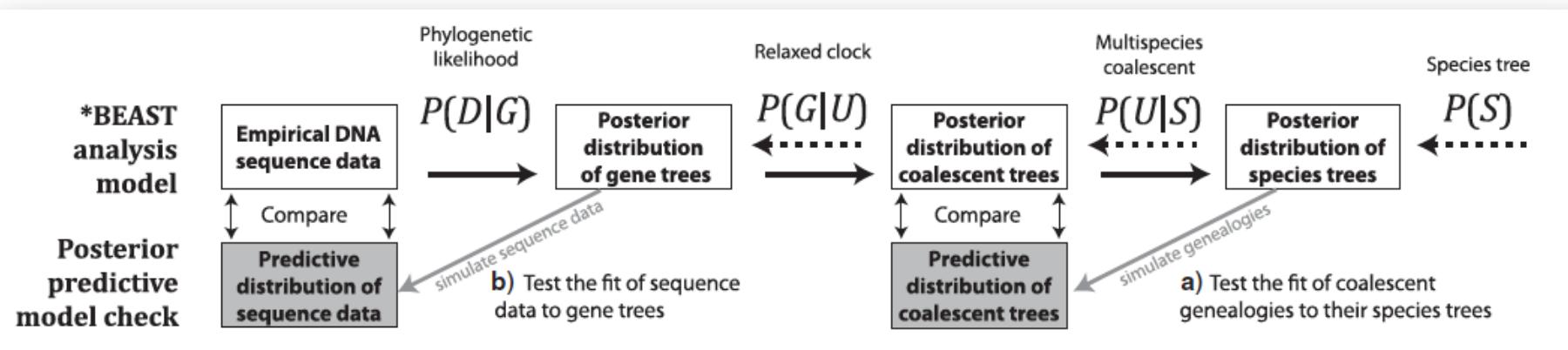
¹Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; and

²Department 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^{1,*}, SARAH M. HIRD¹, JEREMY M. BROWN¹, TARA A. PELLETIER², JOHN D. MCVAY¹, JORDAN D. SATLER²,
AND BRYAN C. CARSTENS²

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

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

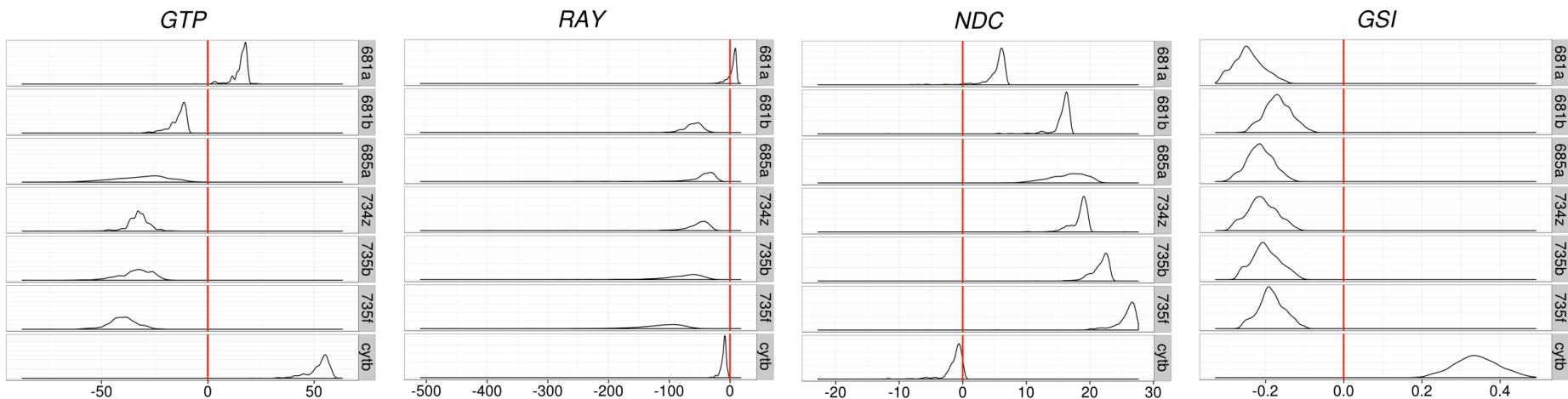
*Correspondence to be sent to: 202 Life Sciences Building, Department of Biological Sciences, Louisiana State University, Baton Rouge,

LA 70803, USA; E-mail: nreid1@ers.lsu.edu.

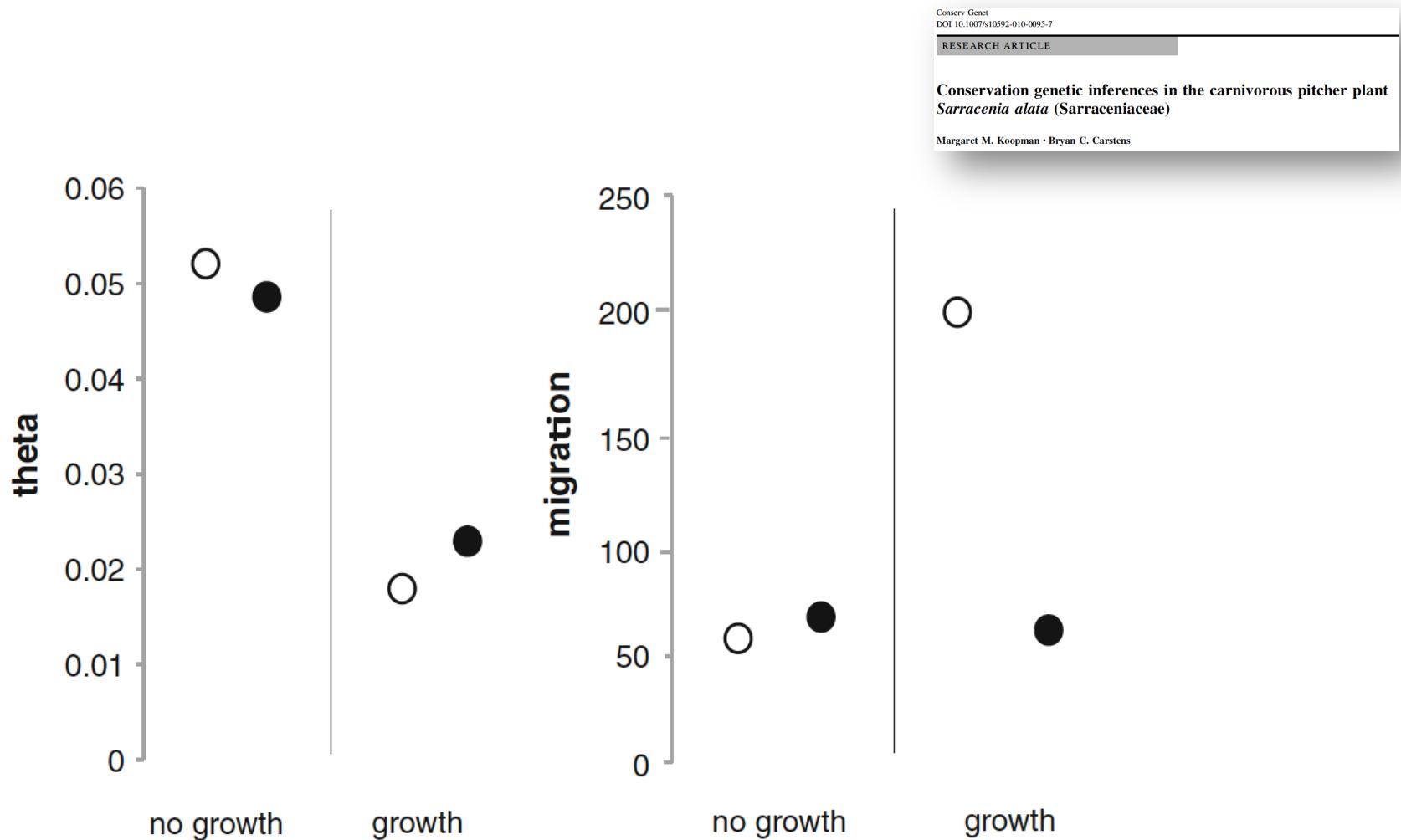
Received 26 November 2012; reviews returned 2 March 2013; accepted 1 August 2013

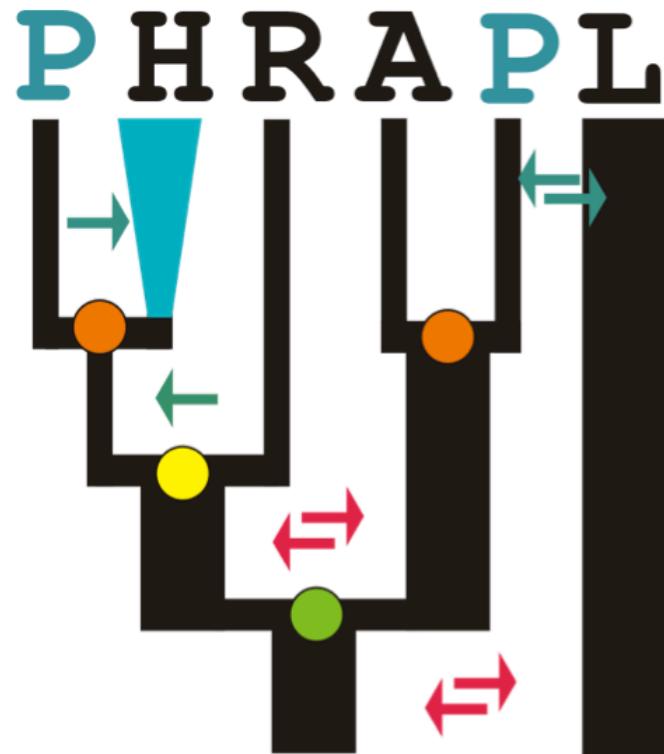
Associate Editor: Laura Kubatko

- 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 prep)

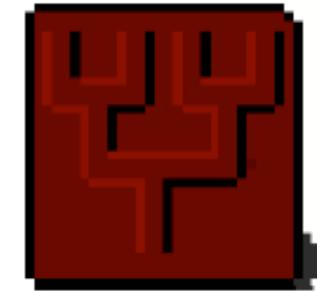


Parameter estimation depends on the parameters included in the model used to estimate the parameters.





≈



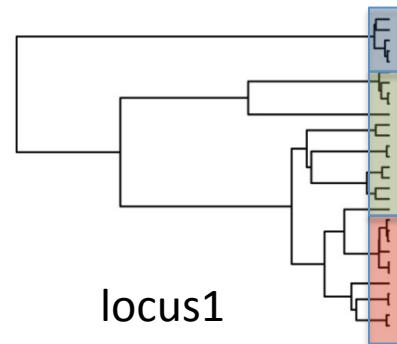
PHylogeographic InfeRence using APproximated Likelihoods

with Brian O'Meara and Nathan Jackson

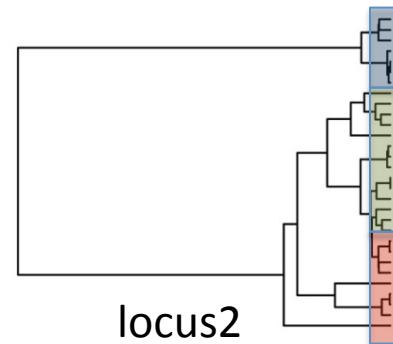
Input

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

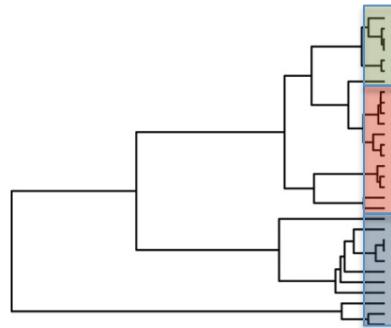
Pop A 
Pop B 
Pop C 



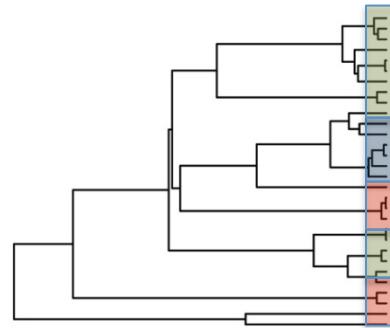
locus1



locus2

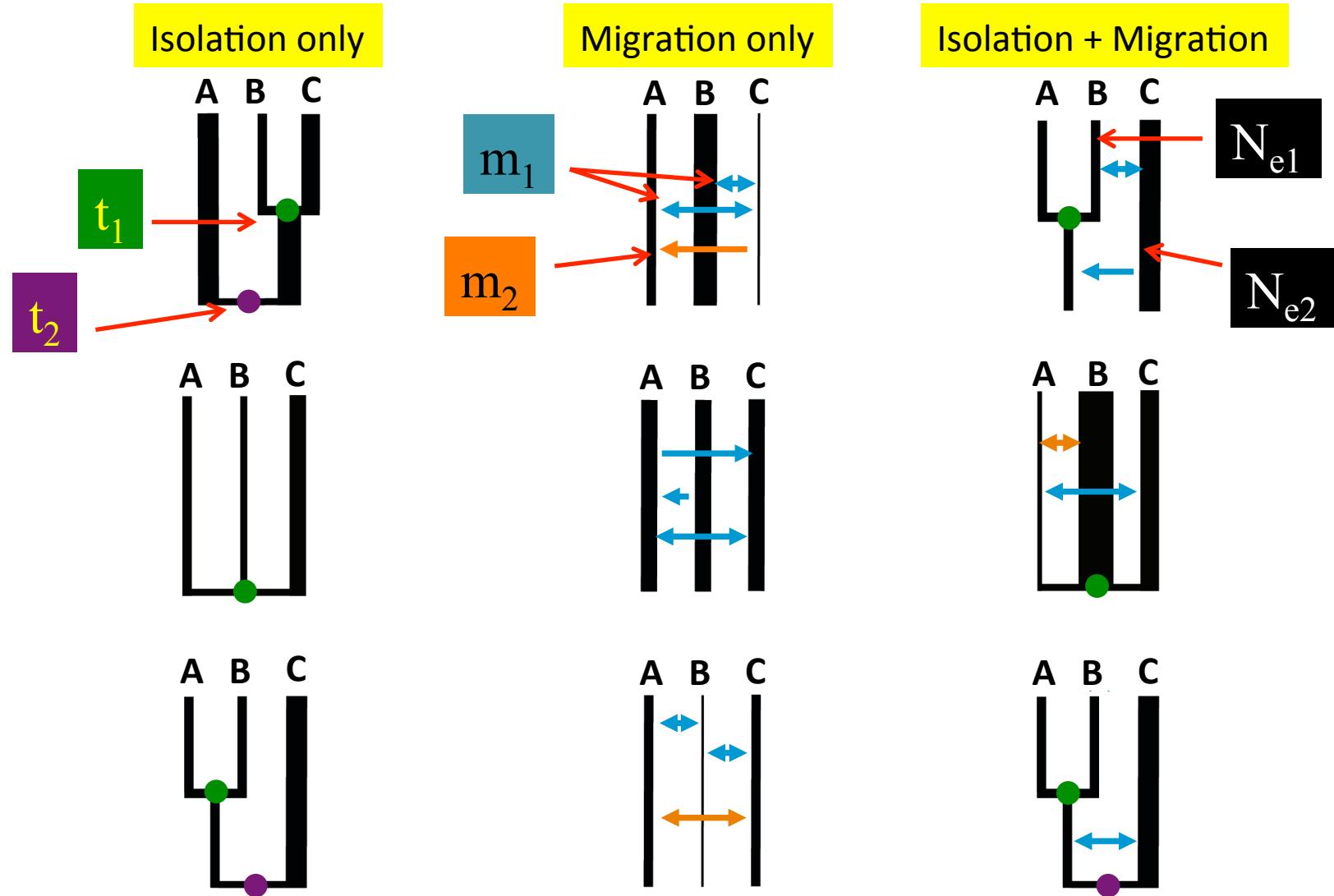


locus3



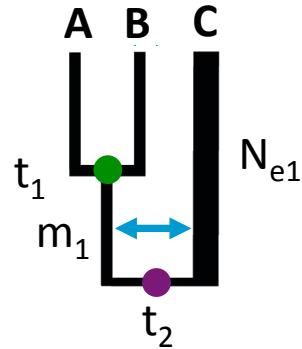
locus4

Phrapl functions: Define all possible models

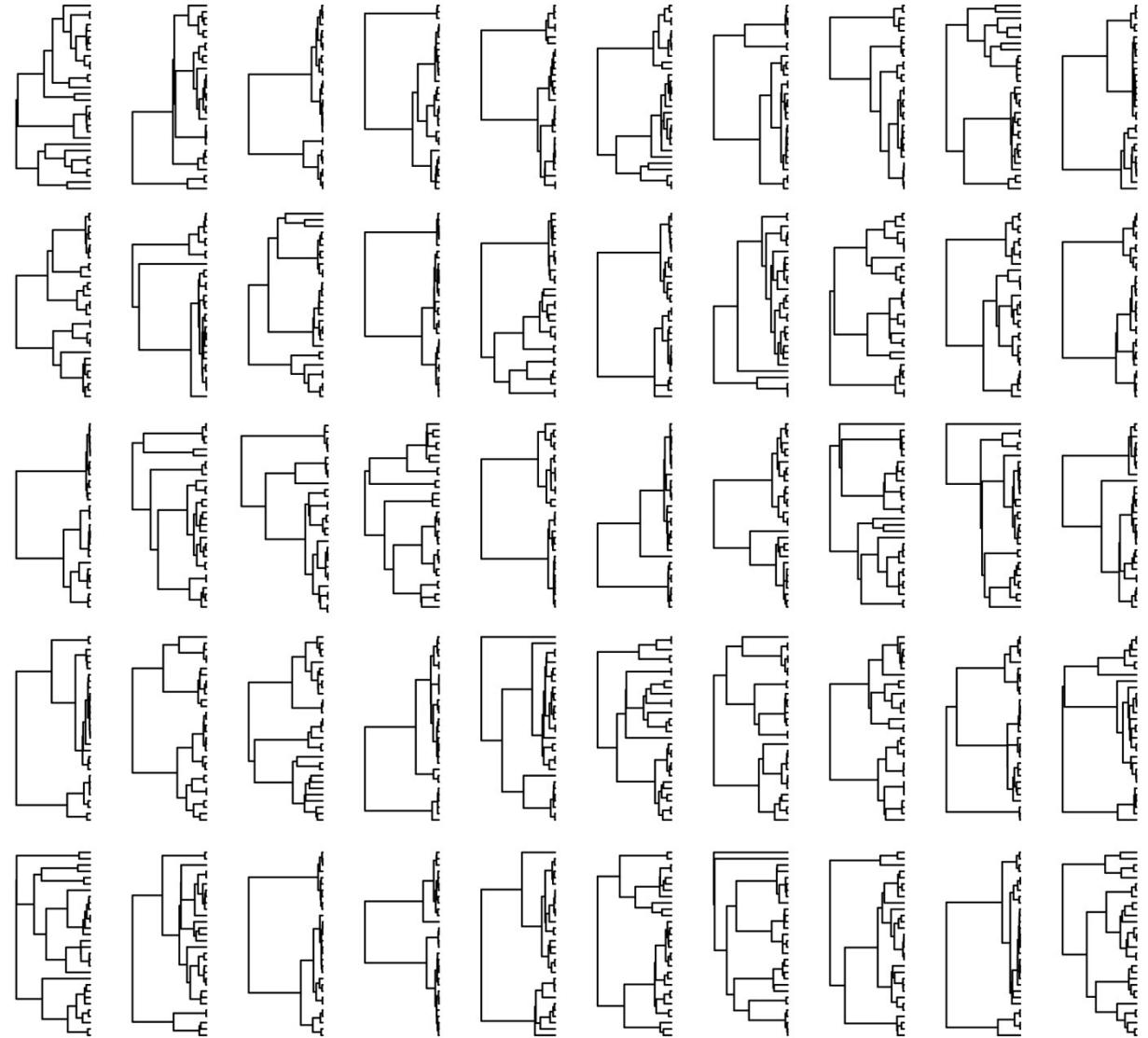


Phrapl functions: approximate likelihood

For each model...

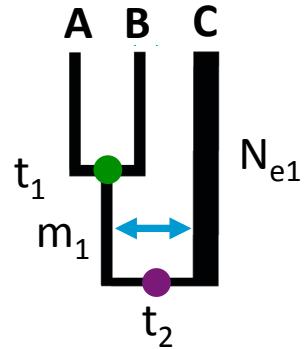


Simulate large number
of trees

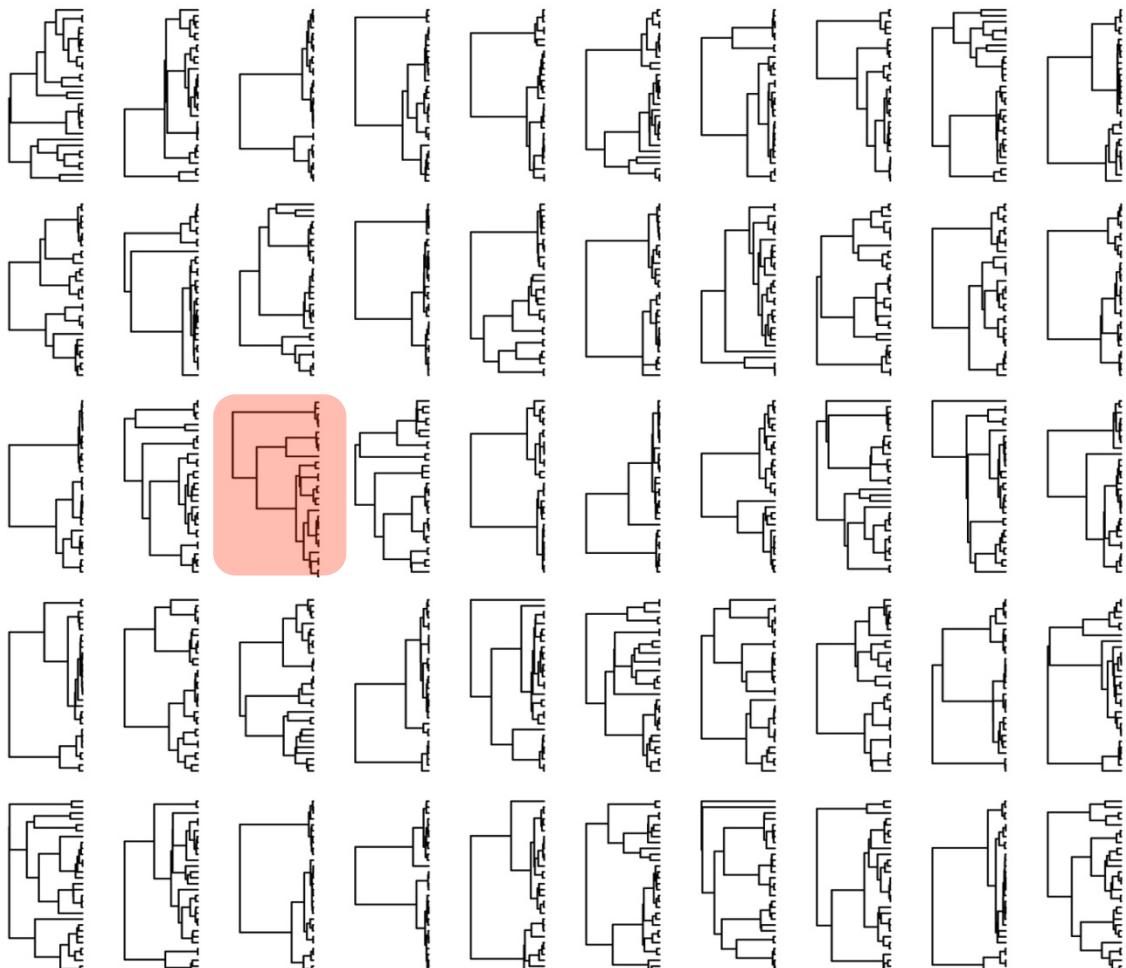


Phrapl functions: approximate likelihood

For each model...



observed tree

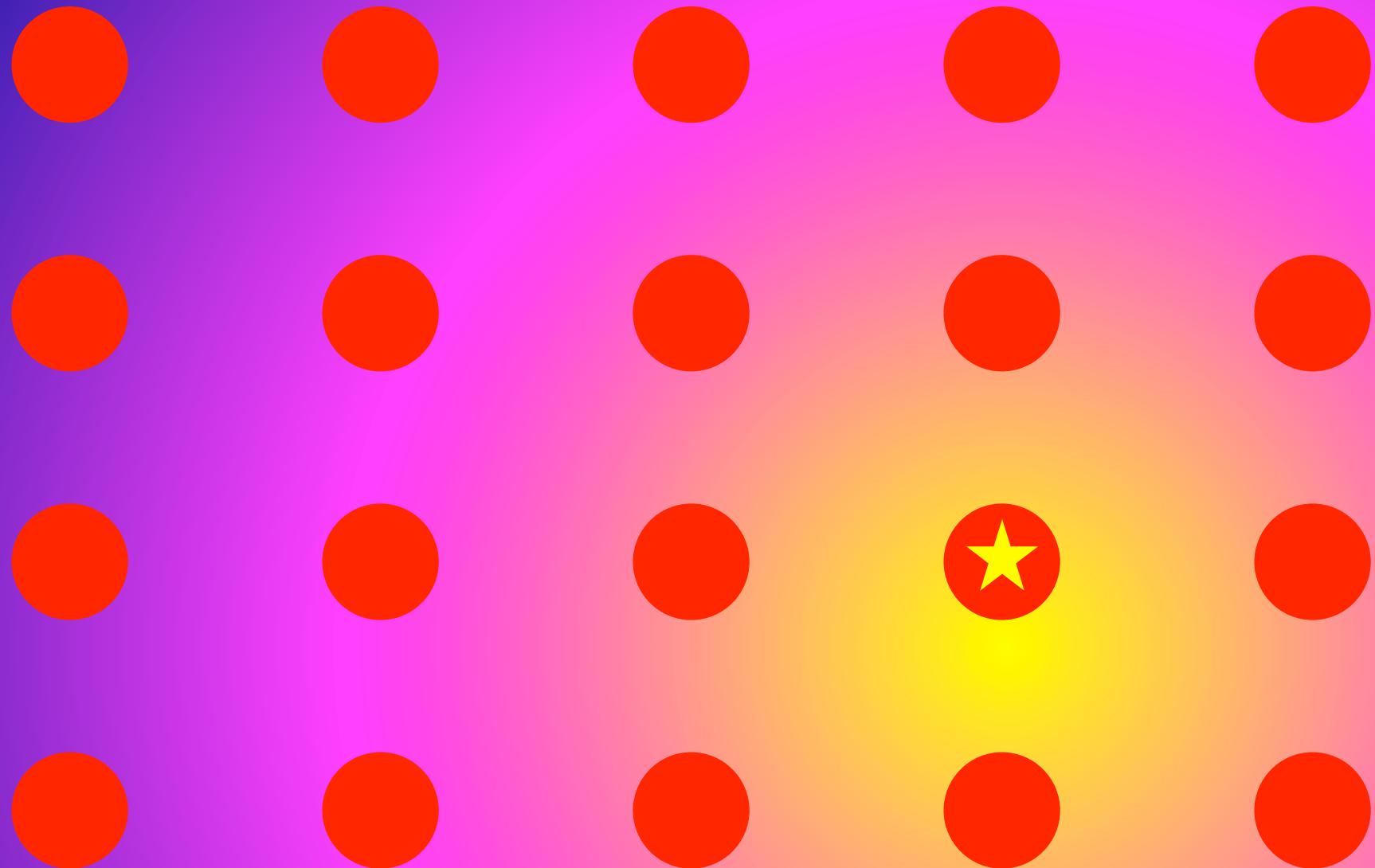


expected trees

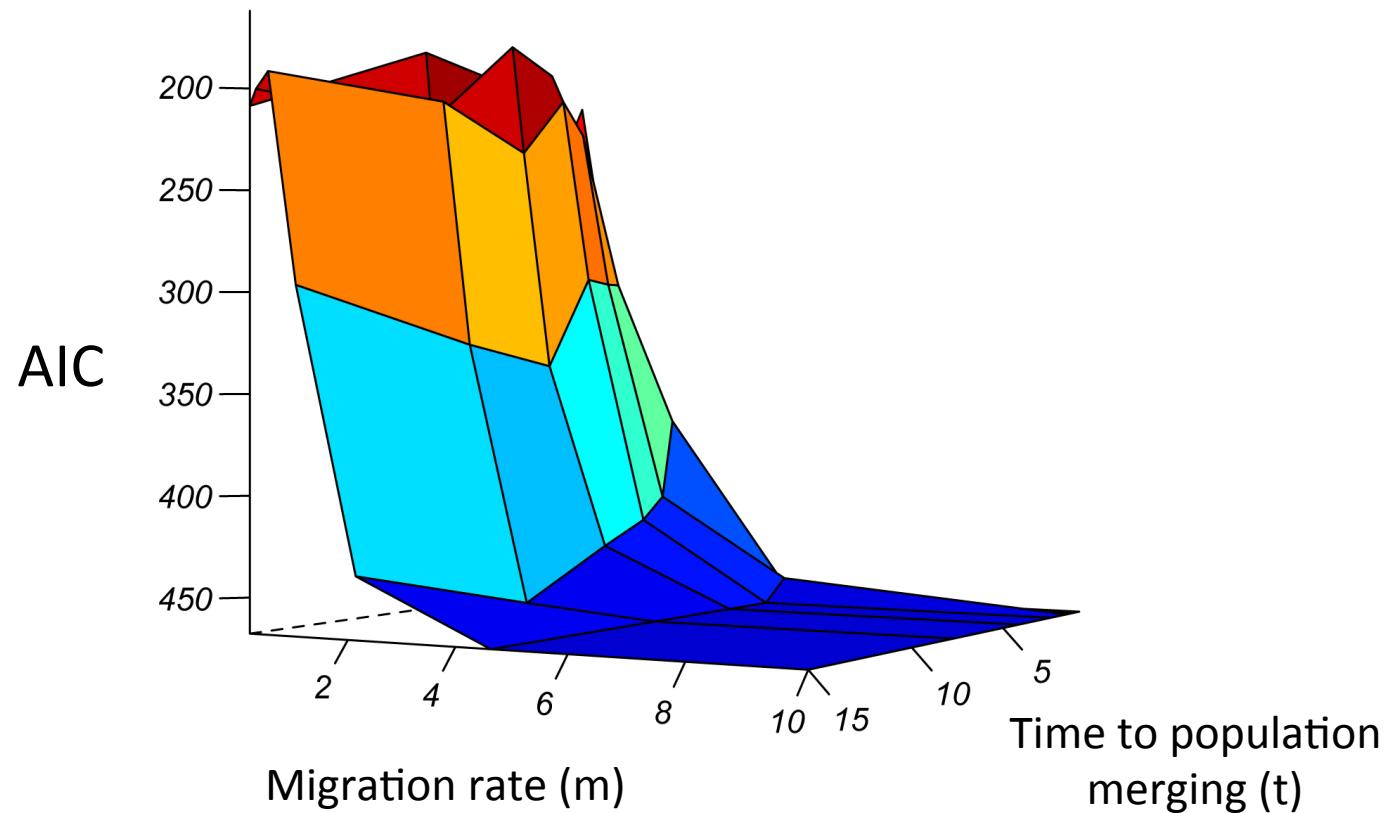
Calculate # of topological matches

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

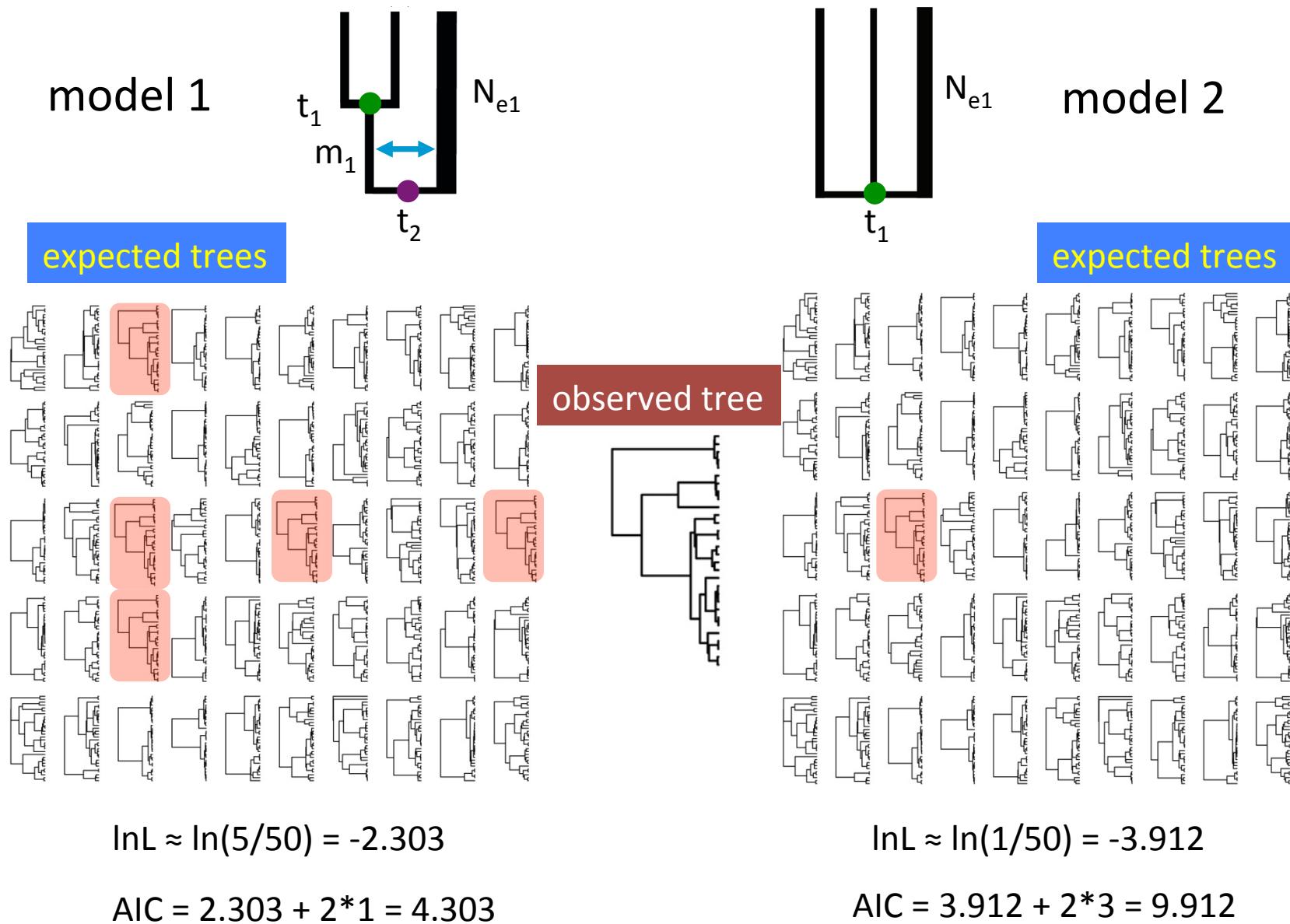
Model Optimization required for approximation of the L (D|M)



Phrapl functions: model parameters are optimized using a grid of values



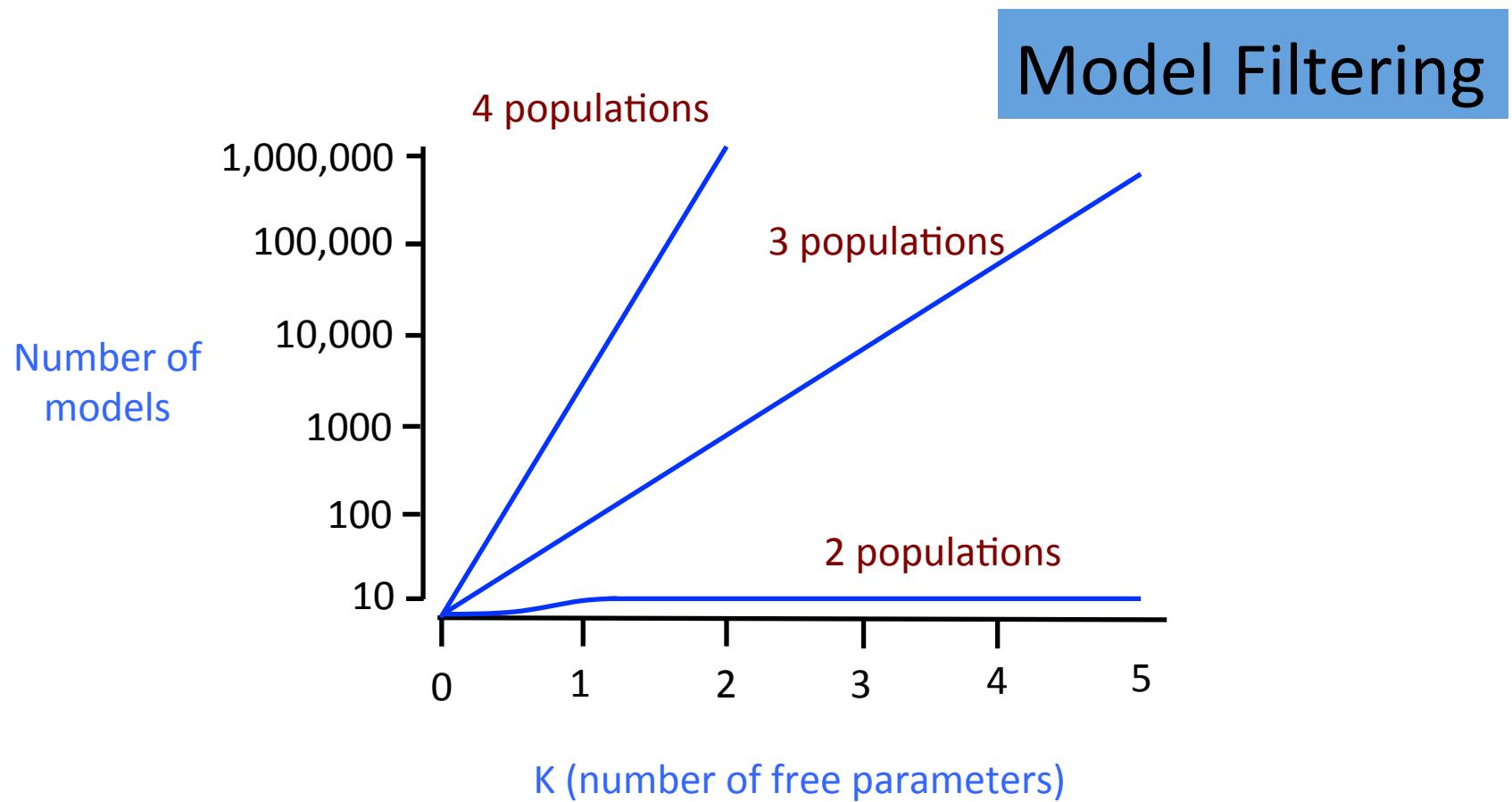
Phrapl functions: approximate likelihood



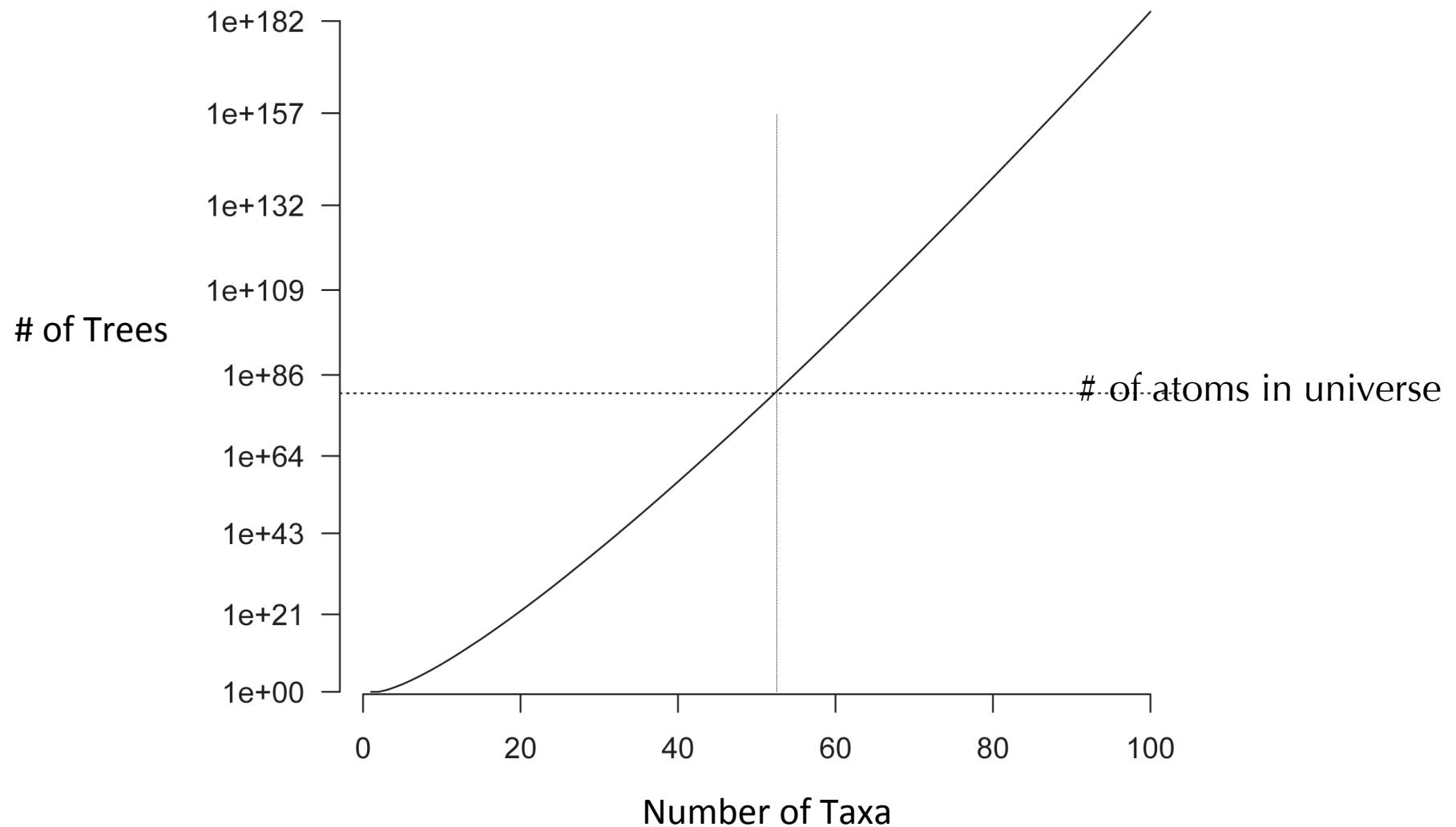
Phrapl output: comparing model fit using AIC

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

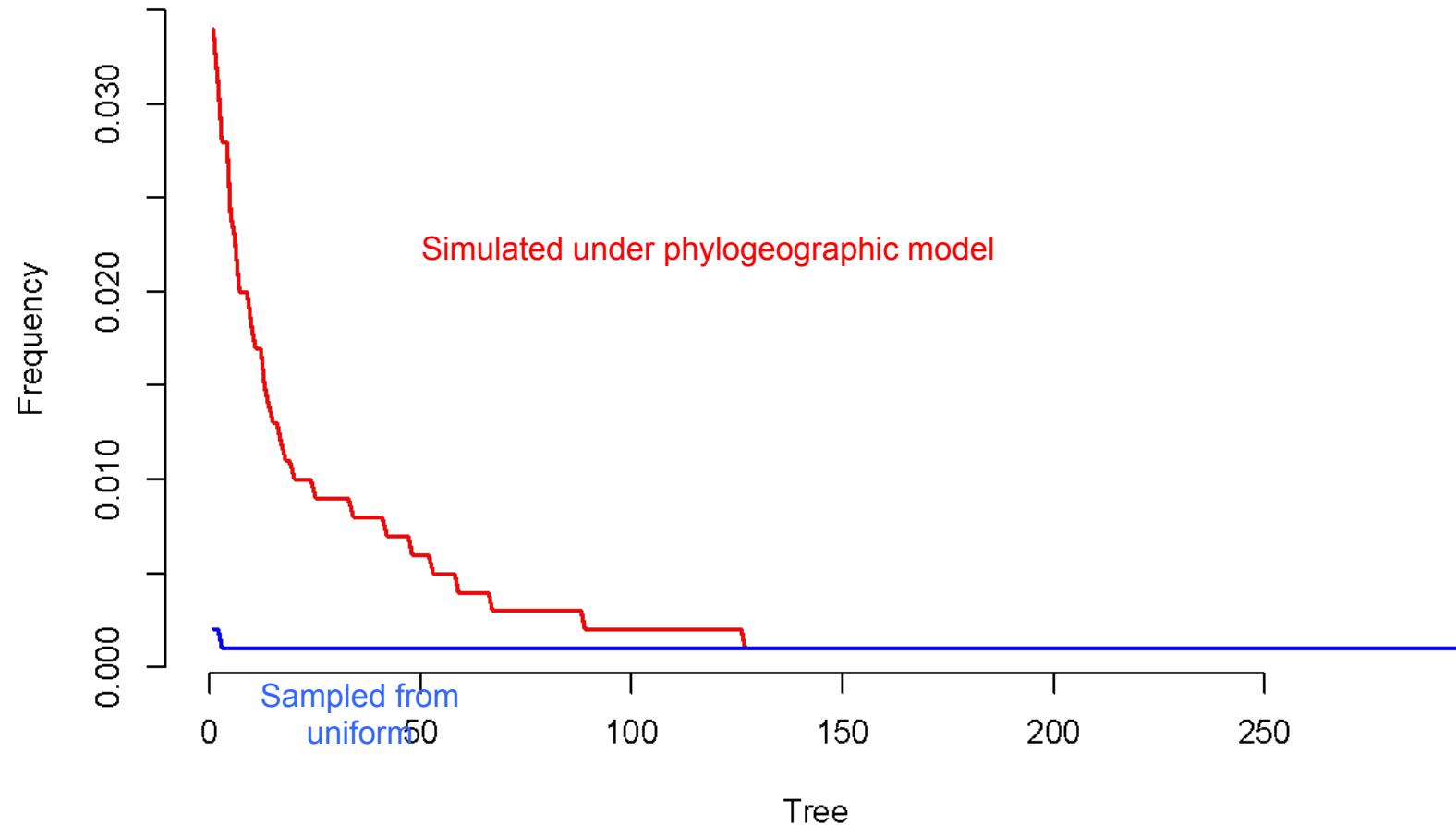
Challenge #1: model space quickly gets enormous



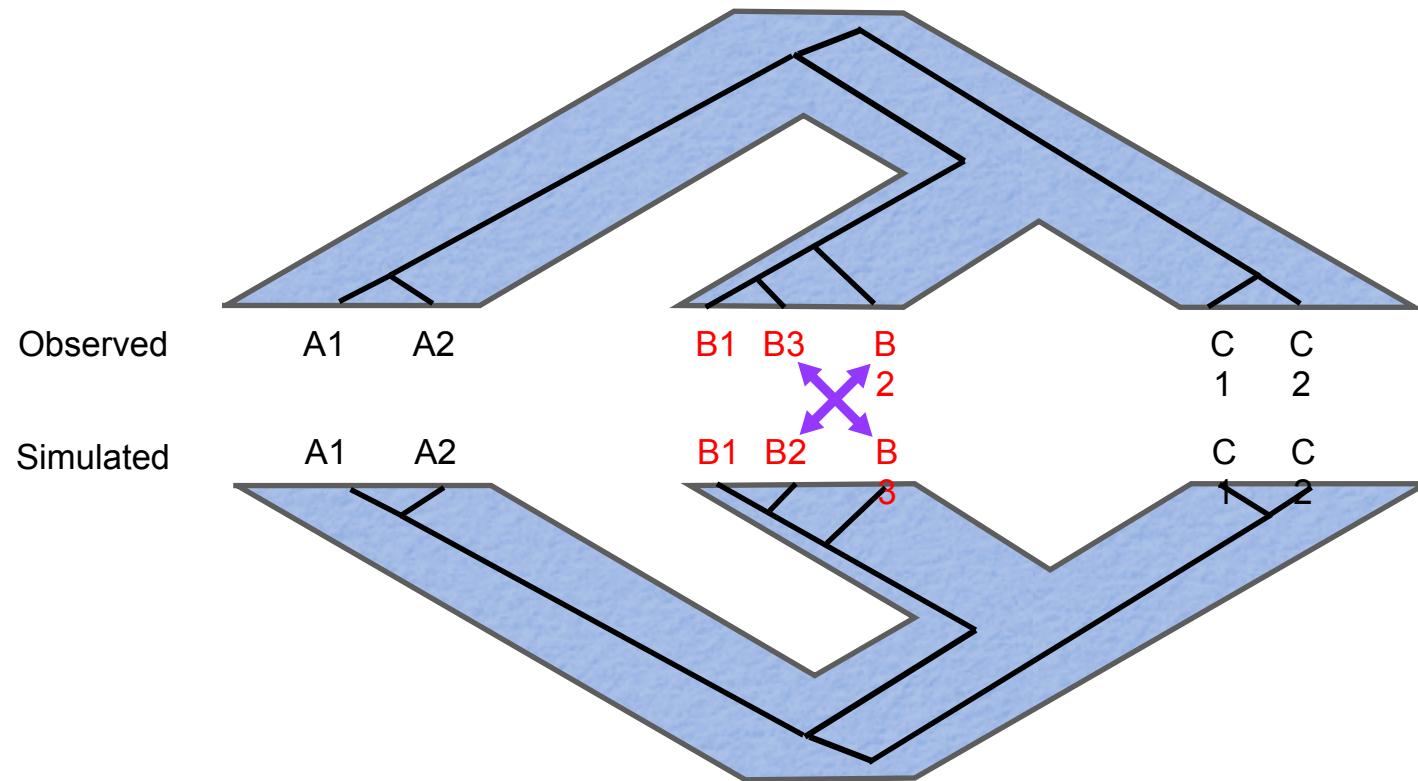
Challenge #2: tree space quickly gets enormous



Tree probabilities are not uniform

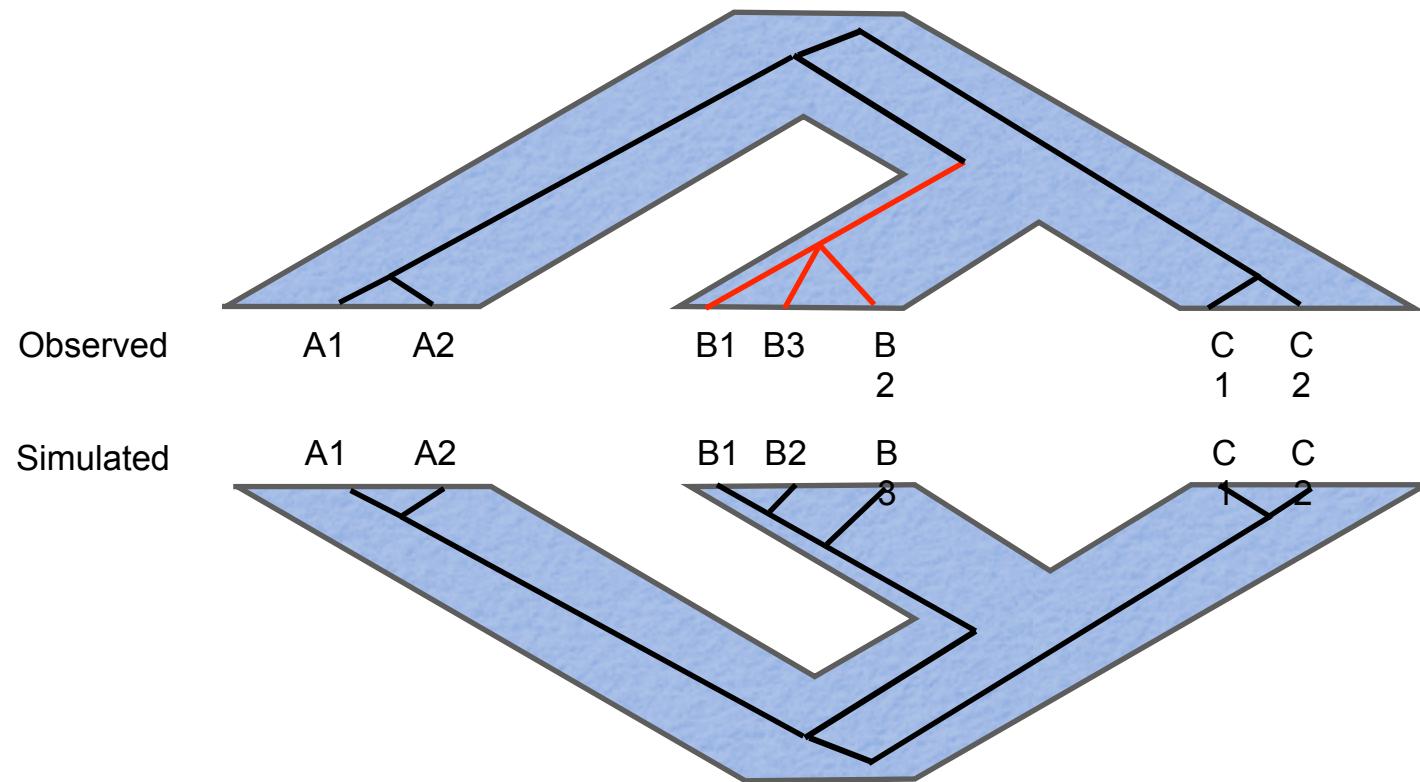


Clever idea 1: Sample labels within populations arbitrary



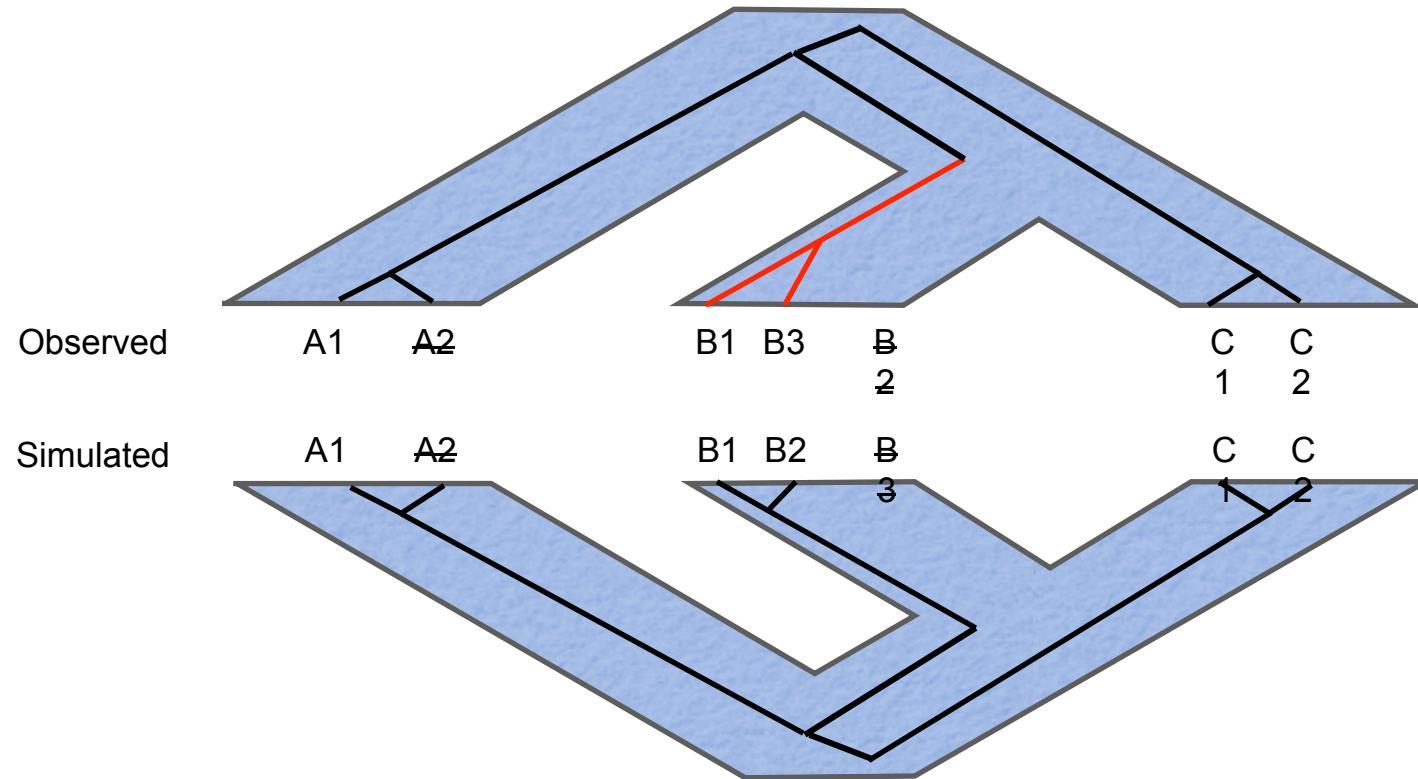
Match based on all possible labeling, then correct for this
i.e., three possible permutations, so if there is a match divide by 3 to get probability

Clever idea 2: Polytomies are soft in gene trees (optional)



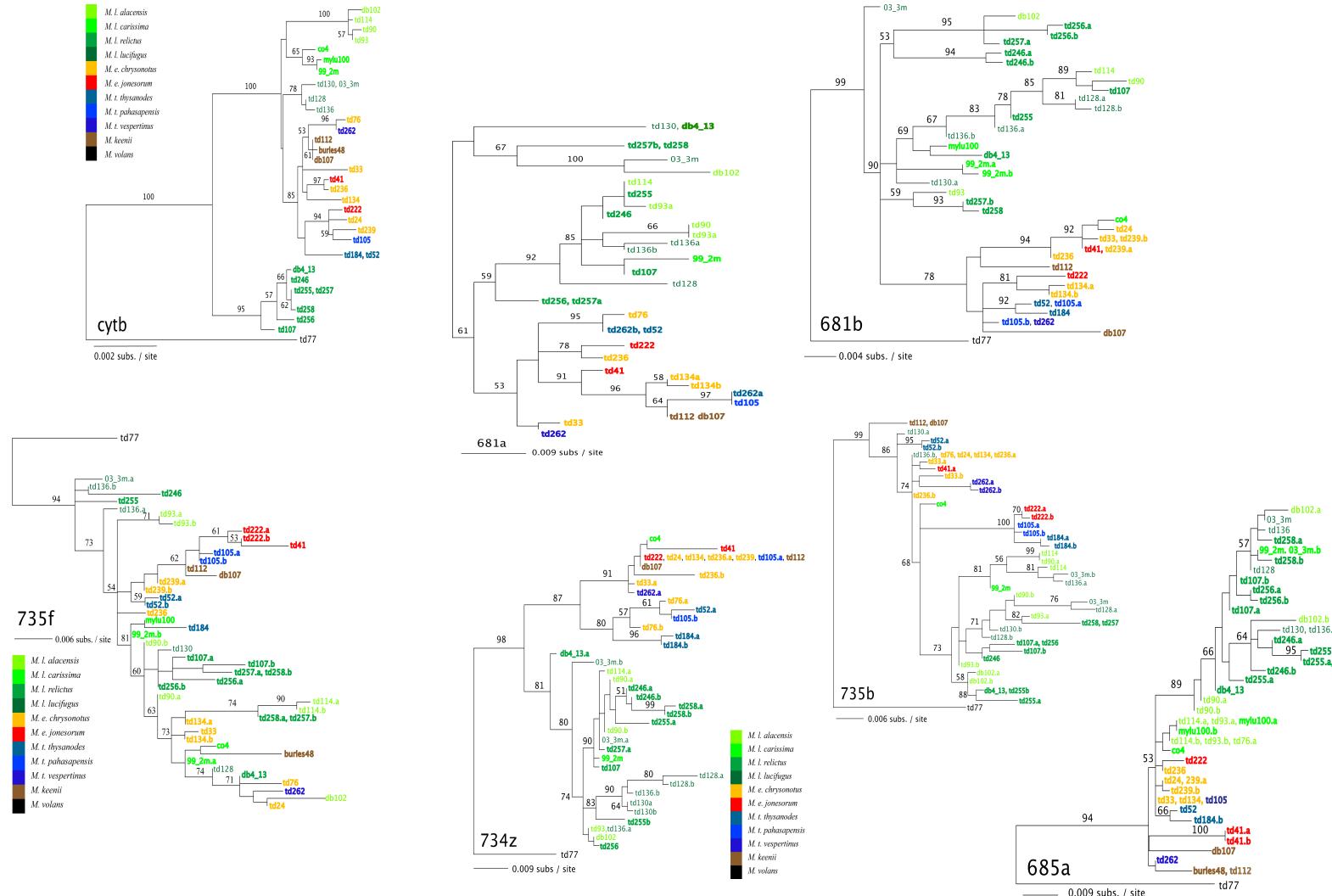
Match based on all possible resolutions, then correct for this

Clever idea 3: Subsample



Reduce the gene tree size (speed gain, precision loss)

Clever idea 4: Do many empirical loci, sample the same topologies multiple times

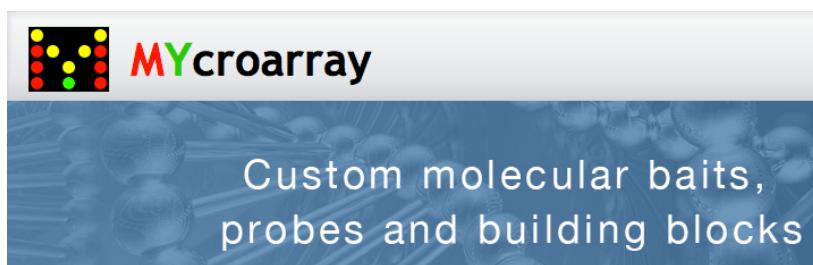


Run time increases linearly with the number of loci.

Custom sequence capture probes synthesized by MYcroarray.

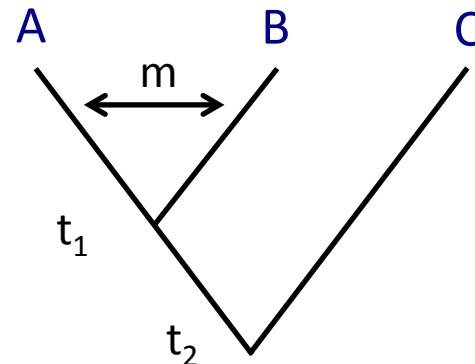
results of test experiment

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



How does PHRAPL perform?

Analyzing simulated data

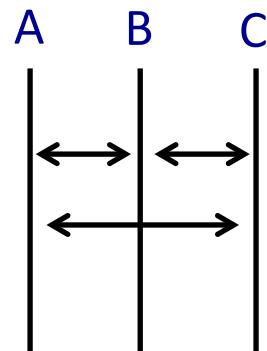


$m = 0$ to 0.5

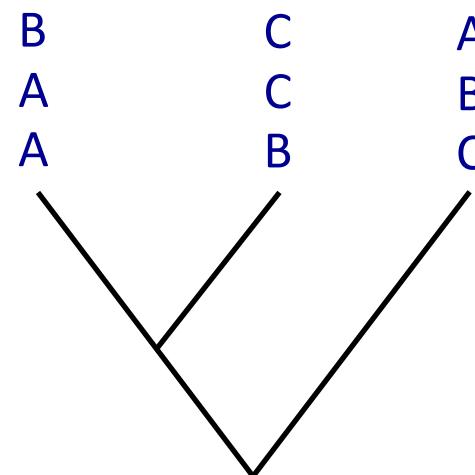
$t_1 = 0.5$ to 2

$t_2 = 1.25$ to 4

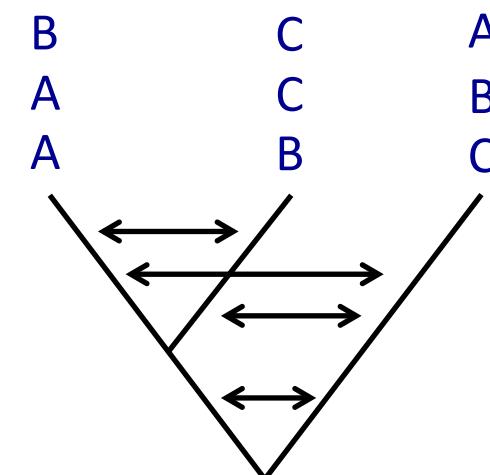
migration only



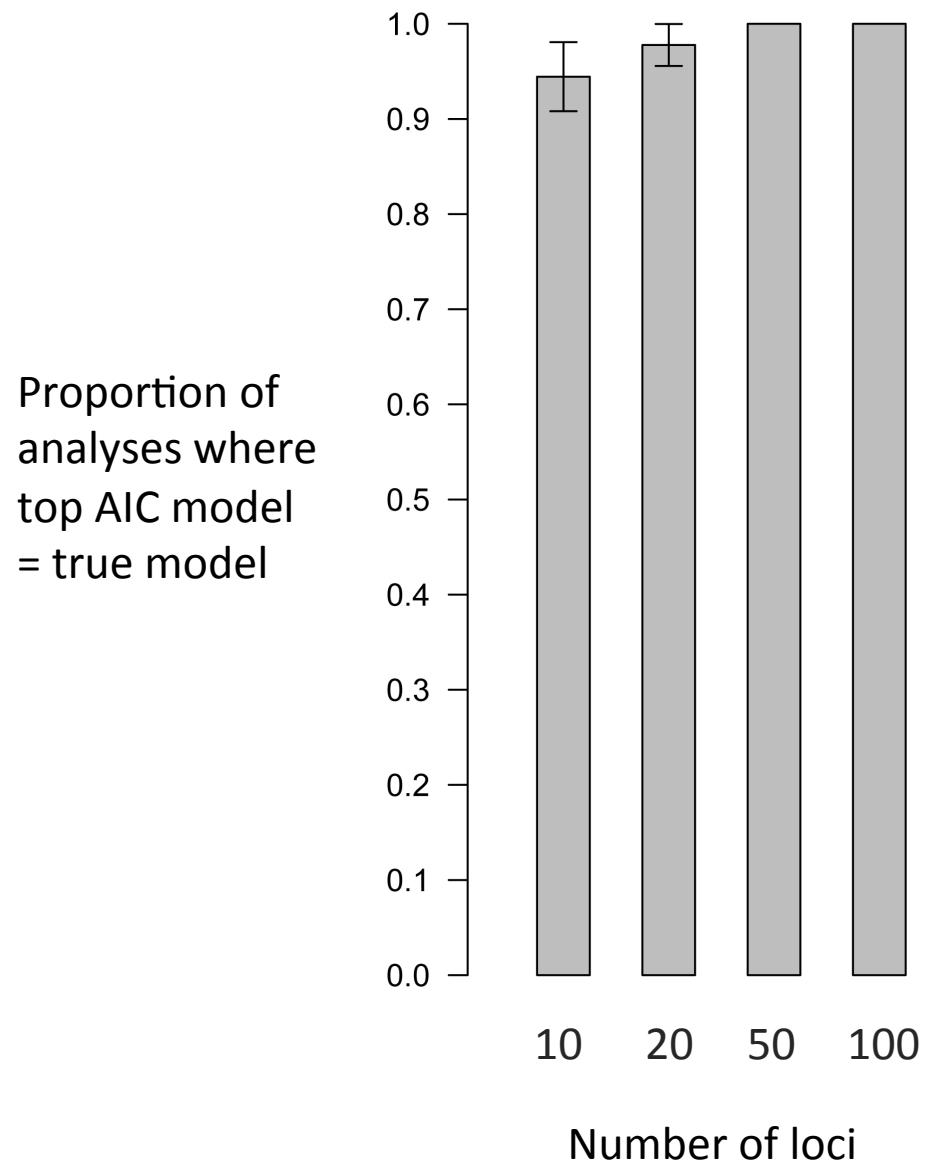
isolation only



isolation + migration

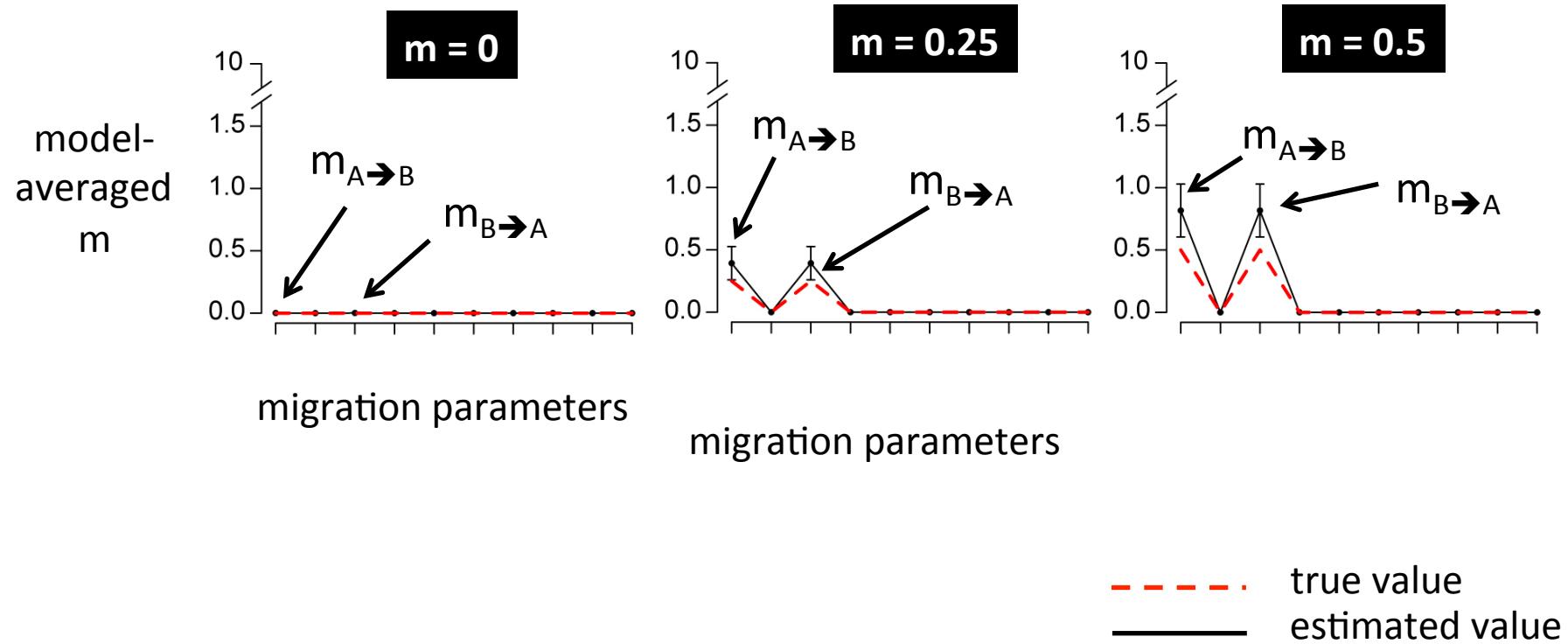
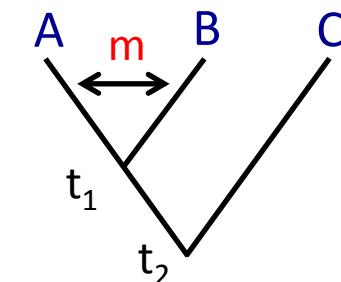


Analyzing simulated data



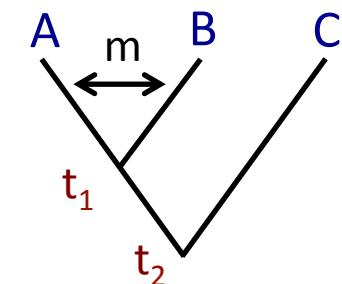
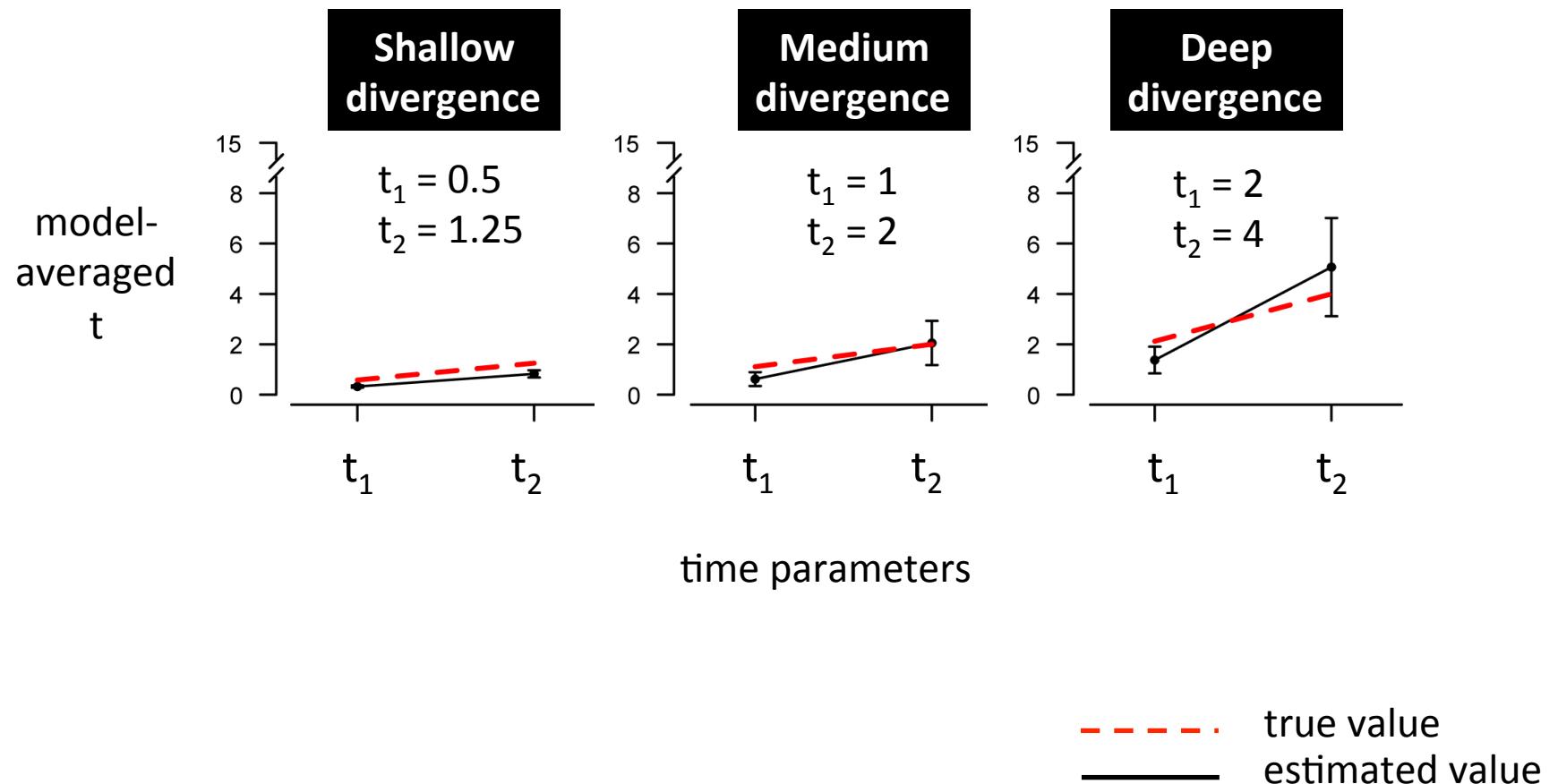
Analyzing simulated data

Migration parameter estimation

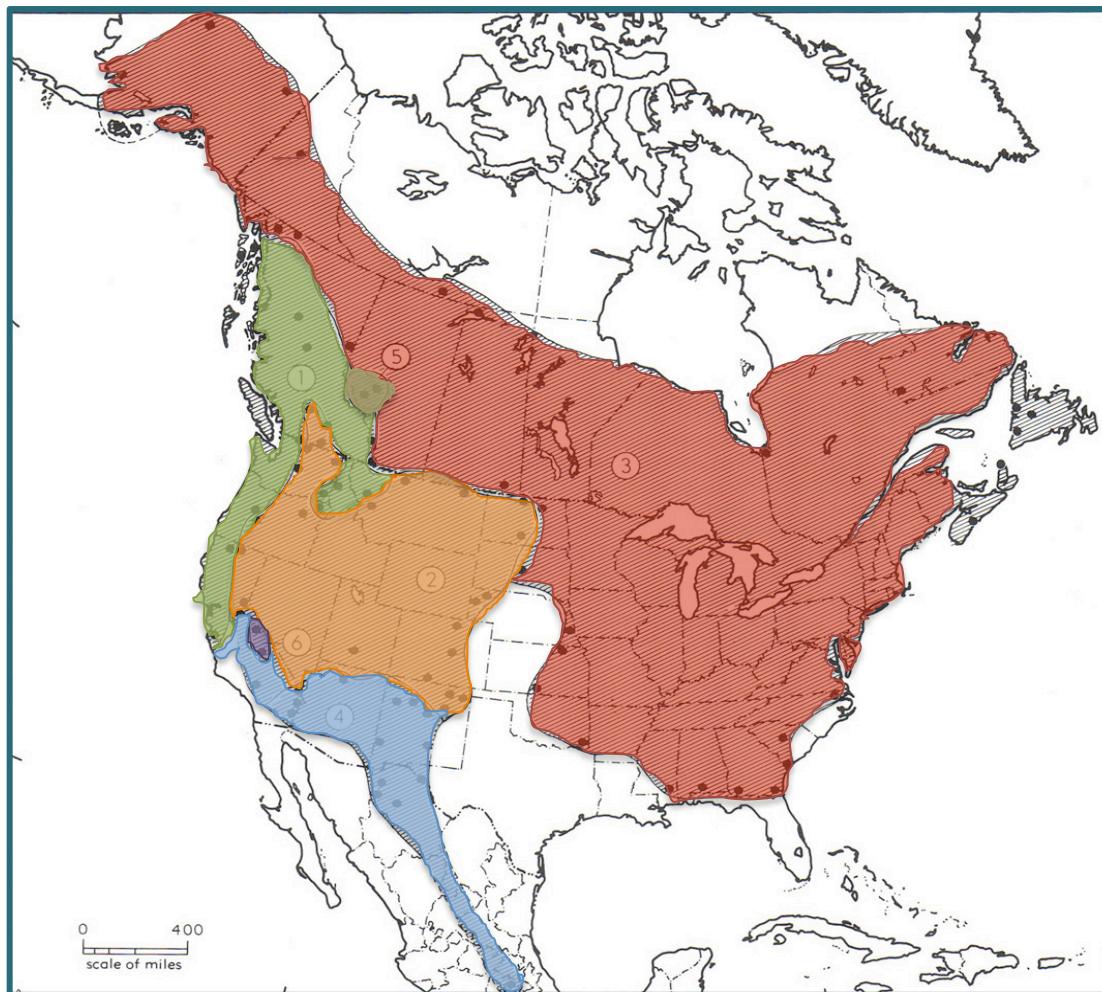


Analyzing simulated data

Time parameter estimation



Little brown bat subspecies (*Myotis lucifugus*)



M.l.alacensis

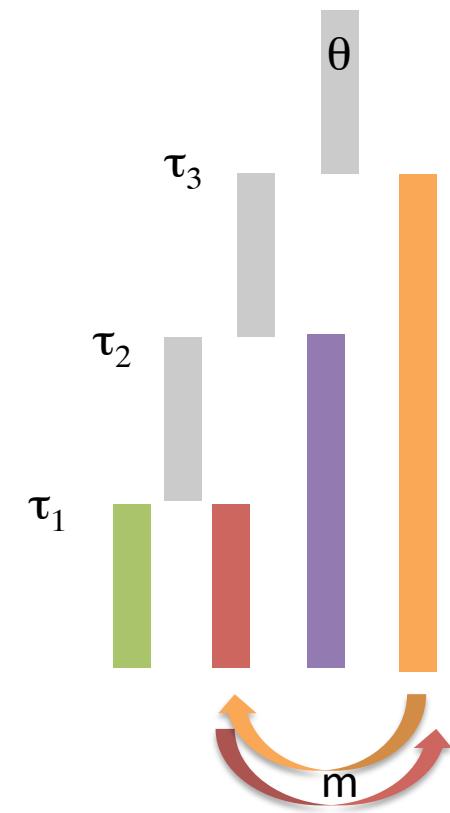
M.l.carissima

M.l.relictus

M.l.lucifugus

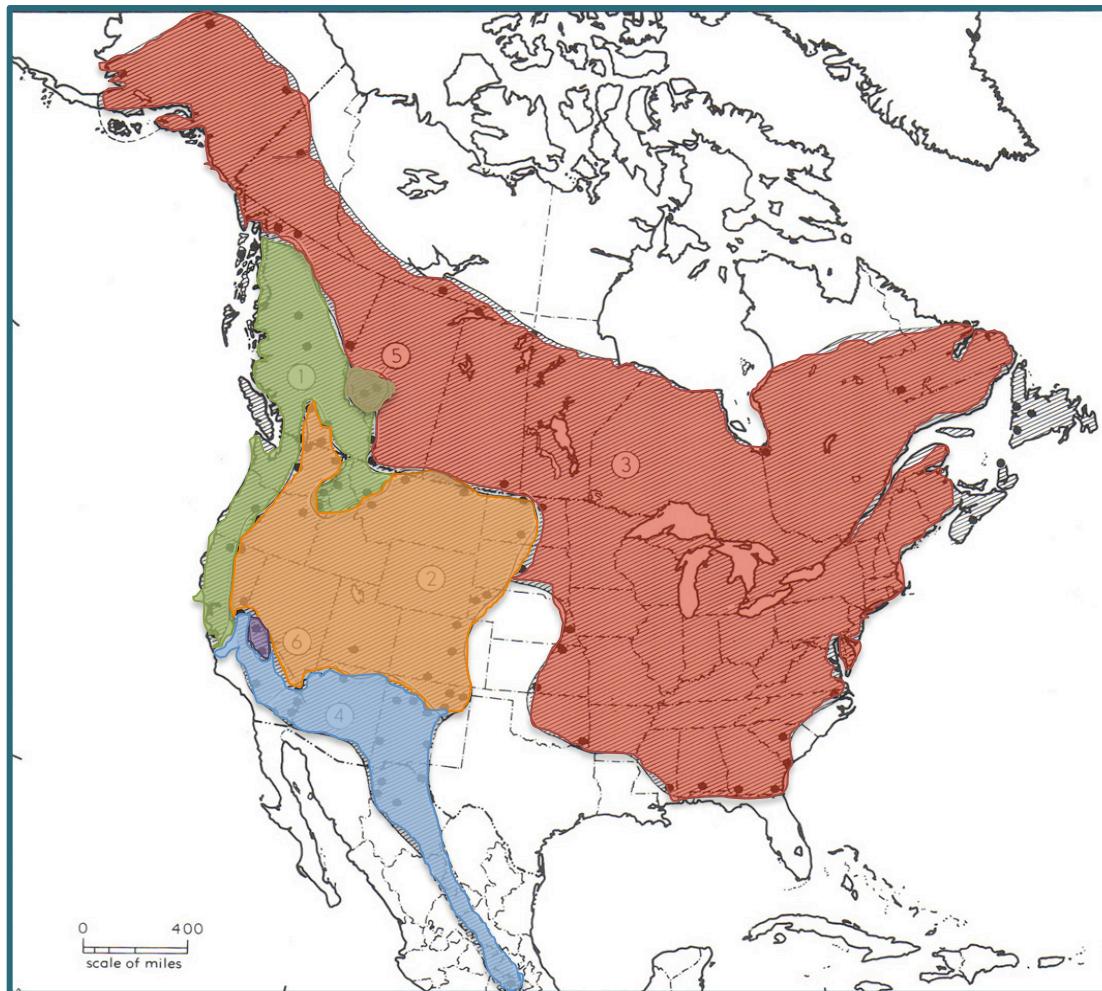
(Hall 1981)

M.l.pernox



$$w_{4101} = 0.405$$

Little brown bat subspecies (*Myotis lucifugus*)



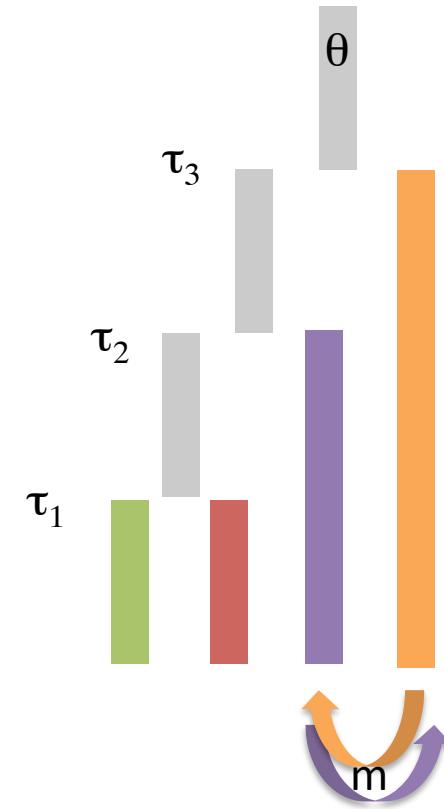
M.l.alacensis

M.l.carissima

M.l.relictus

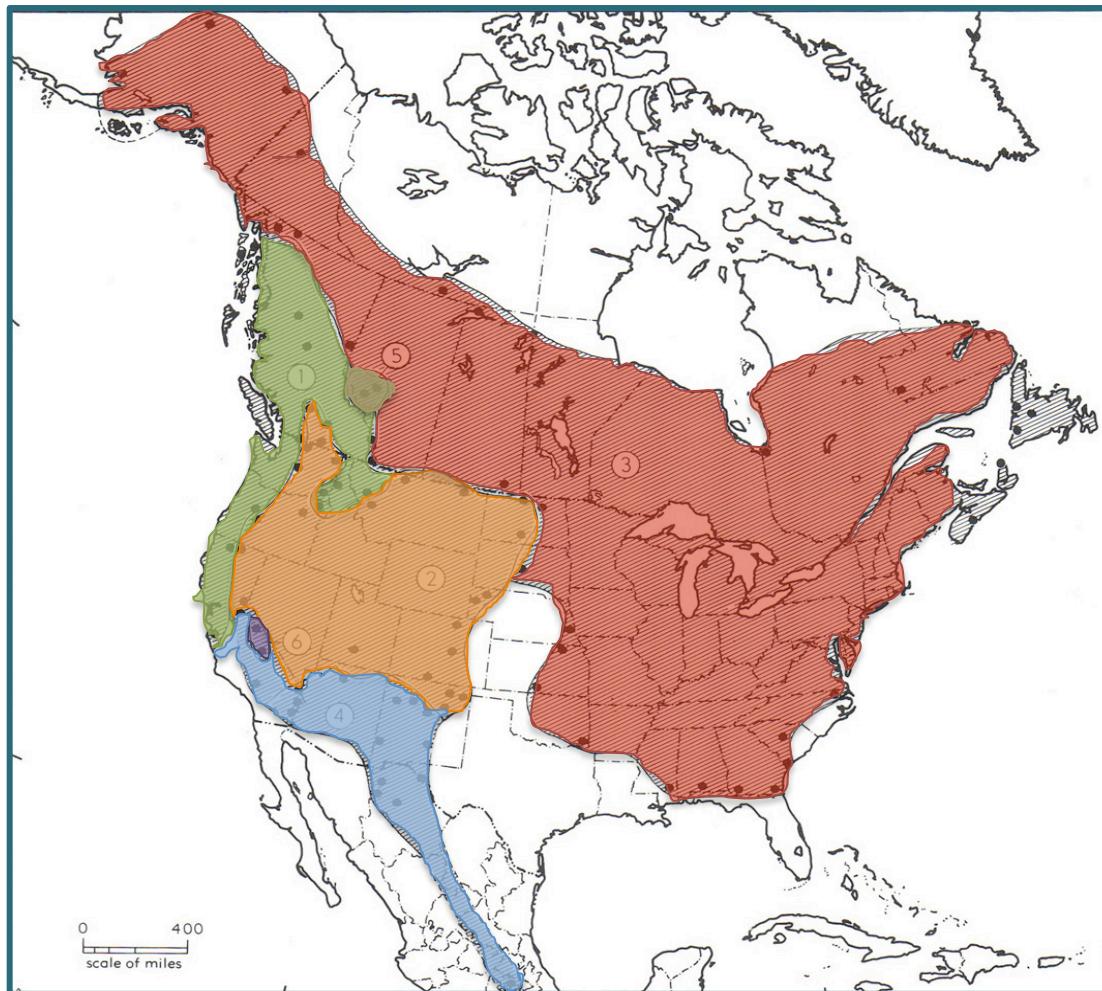
M.l.lucifugus

(Hall 1981)



$$w_{4113} = 0.352$$

Little brown bat subspecies (*Myotis lucifugus*)



M.l.alacensis

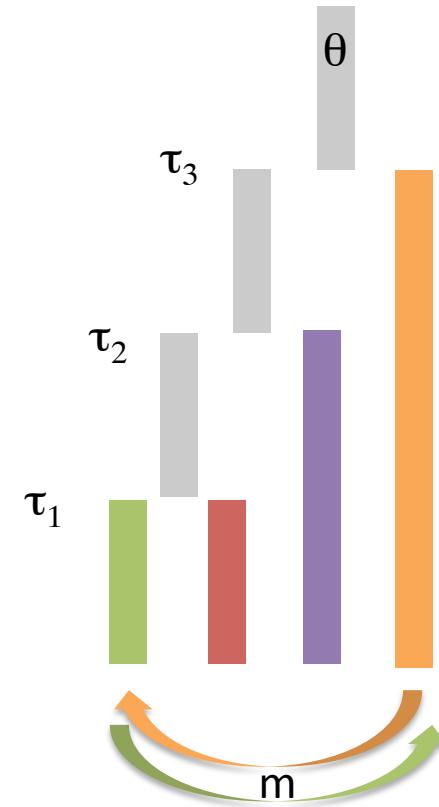
(Hall 1981)

M.l.carissima

M.l.relictus

M.l.pernox

M.l.lucifugus



$$w_{4162} = 0.095$$

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

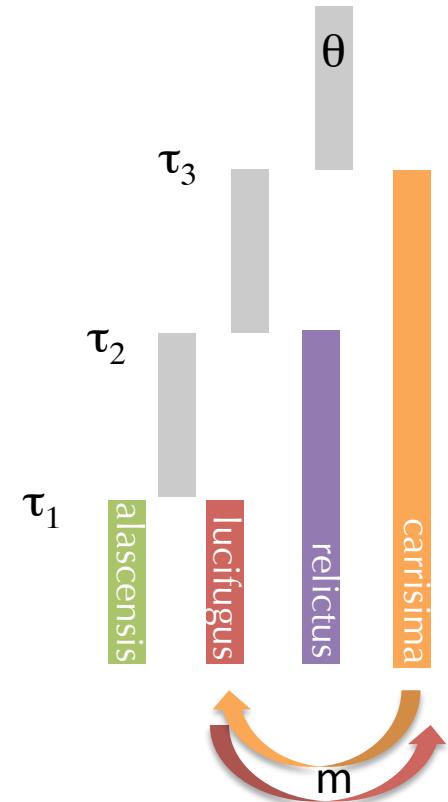
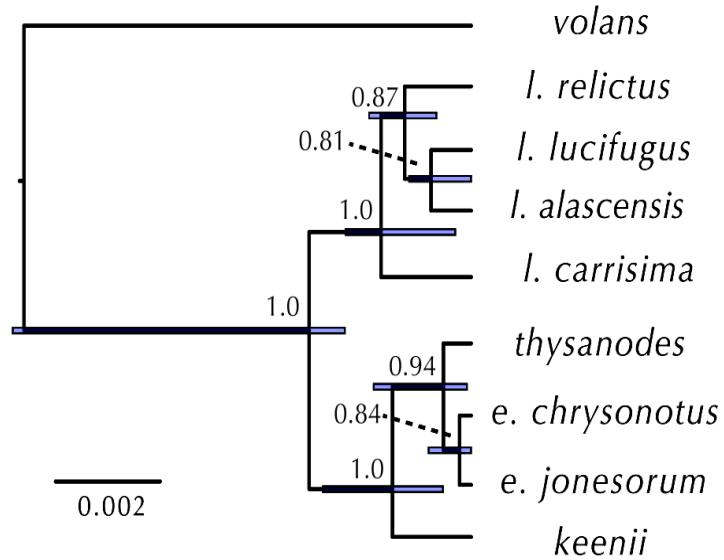
- all have the same topology
- all include migration
- none include a change in population size

migration_model	topology	models	AIC	InL	ΔAIC	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.



$$w_{4101} = 0.405$$

Future directions . . .

- R package P2C2M will be on CRAN as of October 2014
- Paper describing P2C2M will be submitted next week
- PHRAPL manuscript will be submitted soon (I hope . . .)
- PHRAPL extension for species delimitation is in the works
- The *Myotis* capture probe data are being sequenced as we sit here, we'll begin working on the analysis of these data once we get them back from Beckman Coulter Genomics

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

