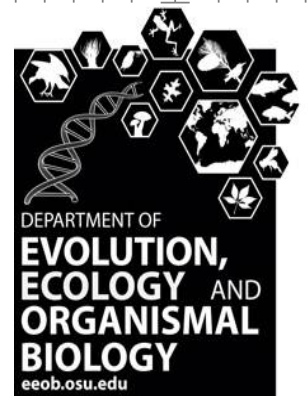




Model-selection as a tool for evolutionary inference.



Bryan Carstens



Model selection as a tool for evolutionary inference.



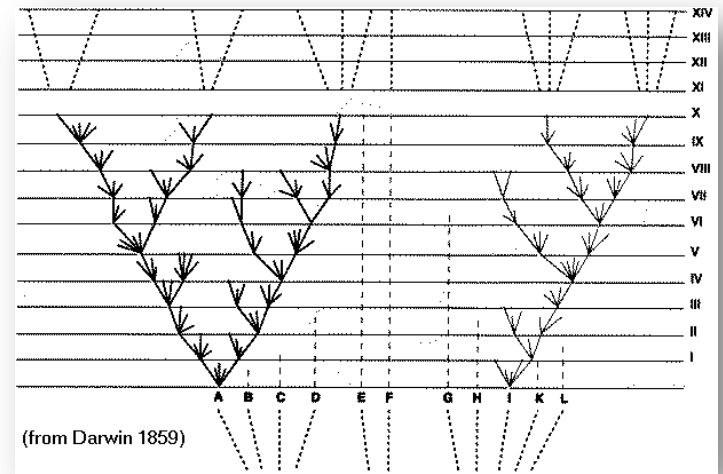
- NSF (DEB-1257784; DEB-0918212; DEB-0956069)



Research in the Carstens lab

understand **diversification** . . .

. . . descent with modification from
a common ancestor



Our investigations occur at the interface between population genetics and systematics, where population-level processes produce phylogenetic patterns.

Evolution, 43(6), 1989, pp. 1192–1208

GENE TREES AND ORGANISMAL HISTORIES: A PHYLOGENETIC APPROACH TO POPULATION BIOLOGY¹

JOHN C. AVISE
Department of Genetics, University of Georgia, Athens, GA 30602

Ann. Rev. Ecol. Syst. 1987, 18:489–522
Copyright © 1987 by Annual Reviews Inc. All rights reserved

INTRASPECIFIC PHYLOGEOGRAPHY: The Mitochondrial DNA Bridge Between Population Genetics and Systematics

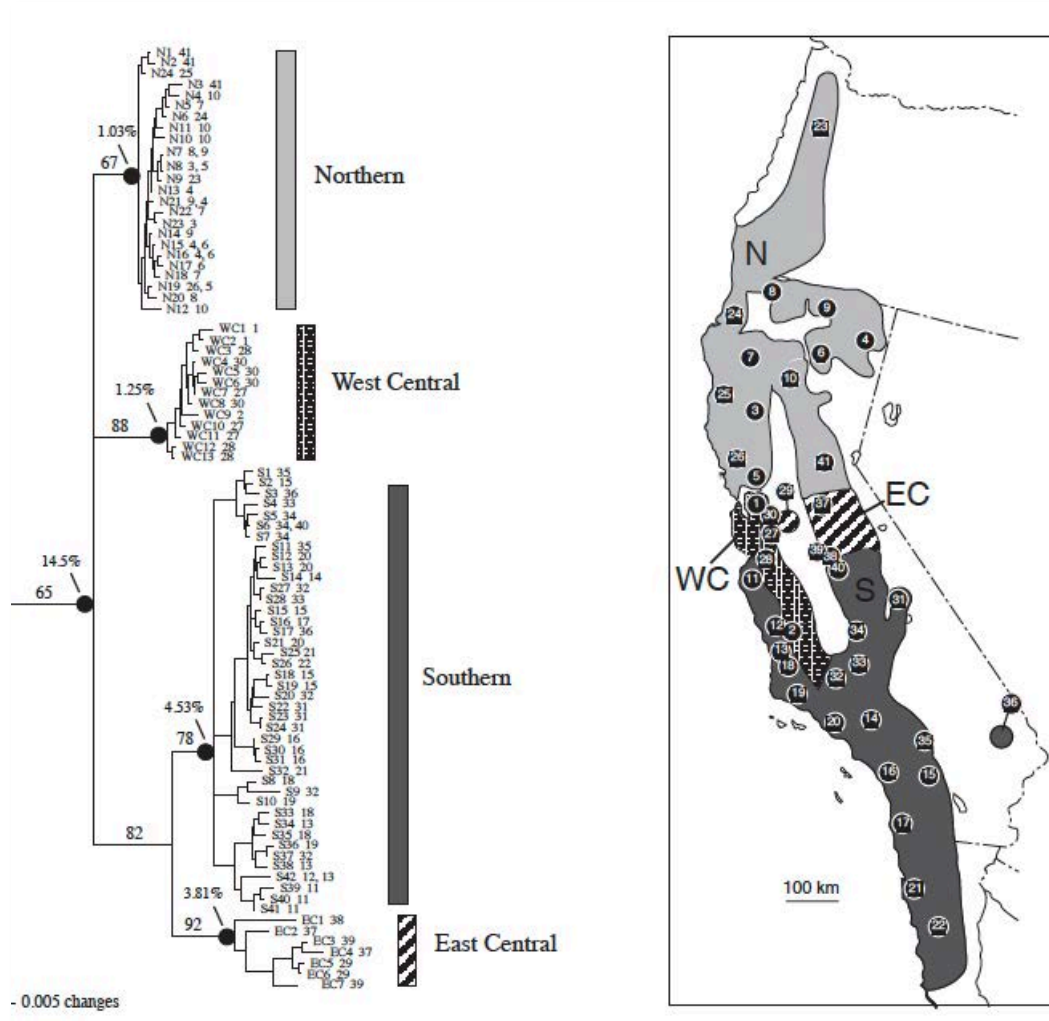
John C. Avise¹, Jonathan Arnold¹, R. Martin Ball¹, Eldredge Bermingham^{1,2}, Trip Lamb^{1,3}, Joseph E. Neigel^{1,4}, Carol A. Reeb¹, and Nancy C. Saunders^{1,5}

¹Department of Genetics, University of Georgia, Athens, Georgia 30602; ²NMFS/CZES, Genetics, 2725 Montlake Boulevard East, Seattle, Washington 98112; ³Savannah River Ecology Laboratory, Drawer E, Aiken, South Carolina 29801; ⁴Department of Microbiology and Immunology, School of Medicine, University of California, Los Angeles, California 90024; ⁵School of Veterinary Medicine, Virginia Tech University, Blacksburg, Virginia 24046

- developed from systematics (trees)
- reliant on mtDNA
- early investigations qualitative
- **inferences are intuitive**

Summarize genetic variation in some way

- F_{ST} , *Tajima's D*
- estimate gene trees from the data



trees + maps => inference



“The geographical distribution of distinct clades suggests that a combination of topographic barriers and the expansion and contraction of suitable habitat during the past 2 million years, especially along particular mountain ranges, have played a major role in the diversification of *N. fuscipes*.” (Matocq, 2002)

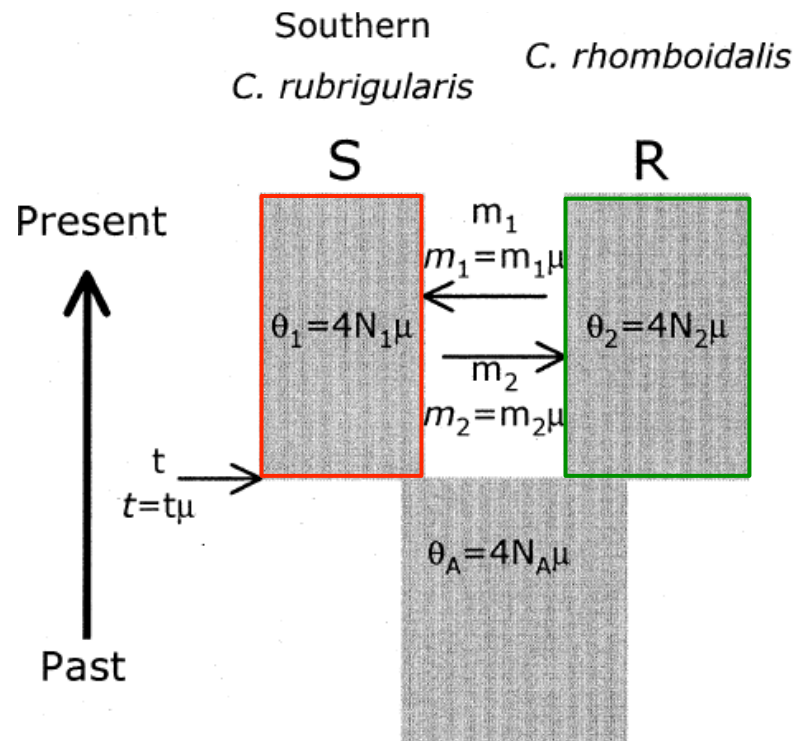
Summarize genetic variation in some way

- F_{ST} , *Tajima's D*
- estimate gene trees from the data

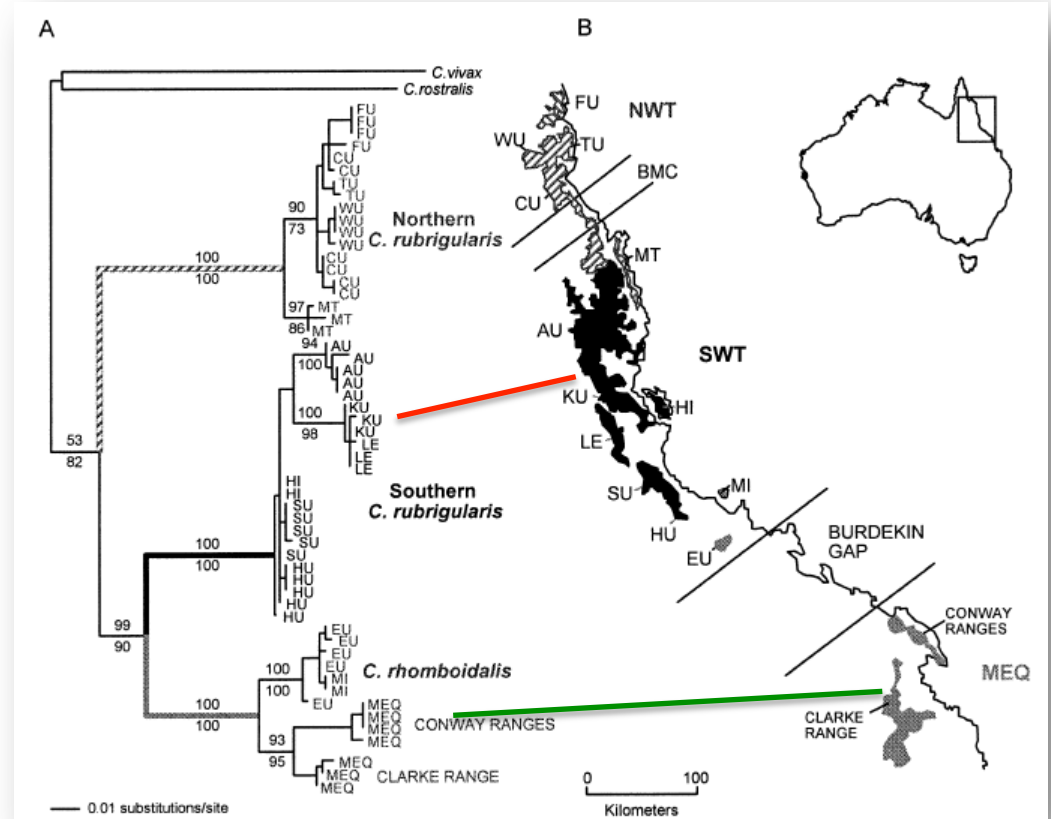
Estimate parameters using some available model
(*assumed to fit data*)

- Nm with Wright's Island model
- migration rates with a coalescent-model

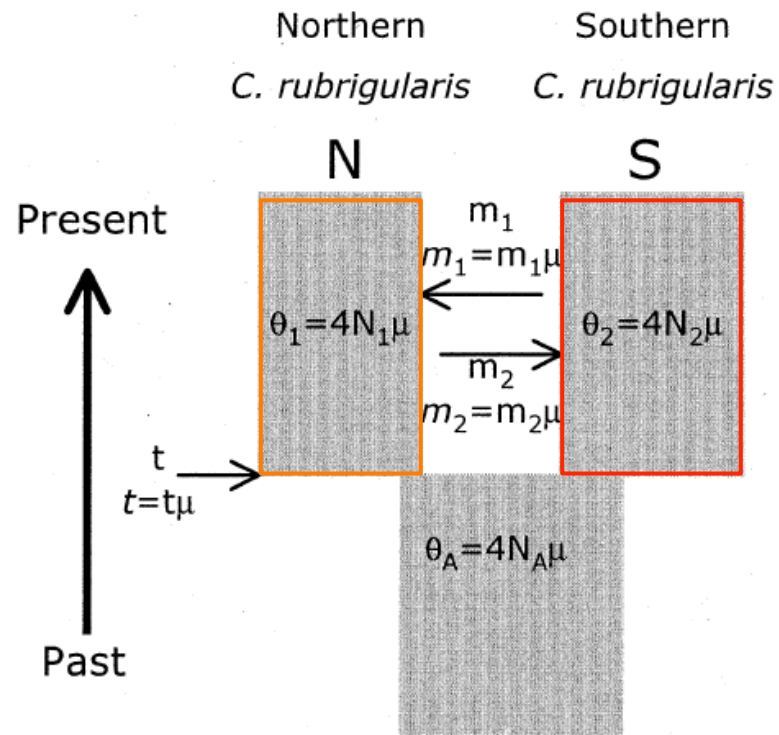
Dolman & Moritz 2006



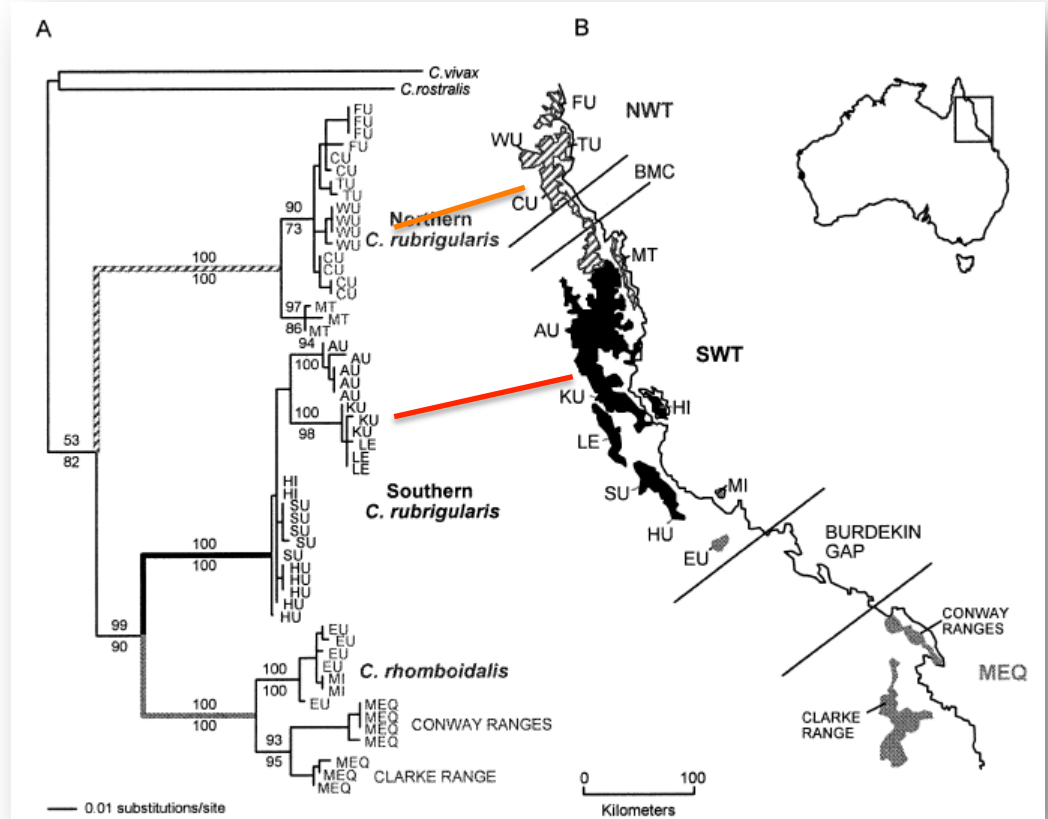
S-R



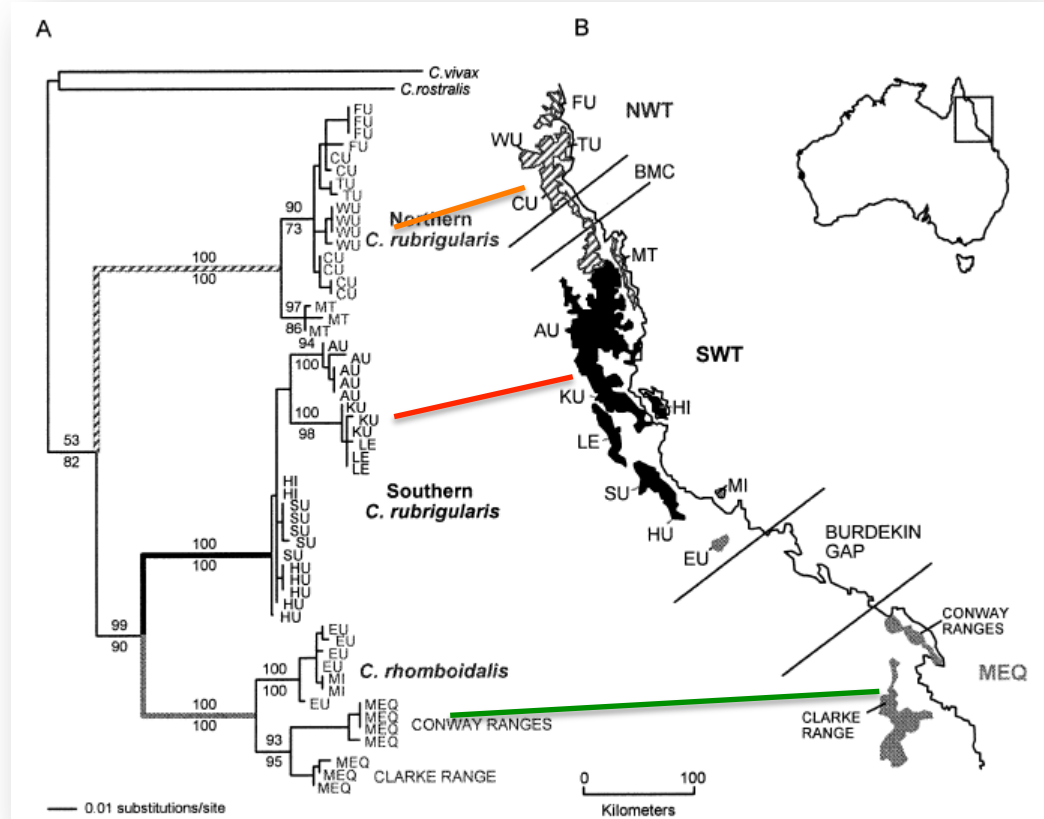
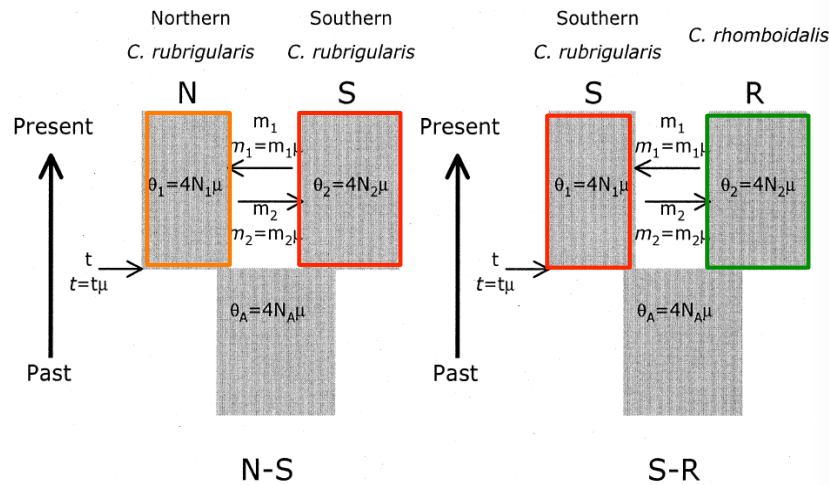
Dolman & Moritz 2006



N-S

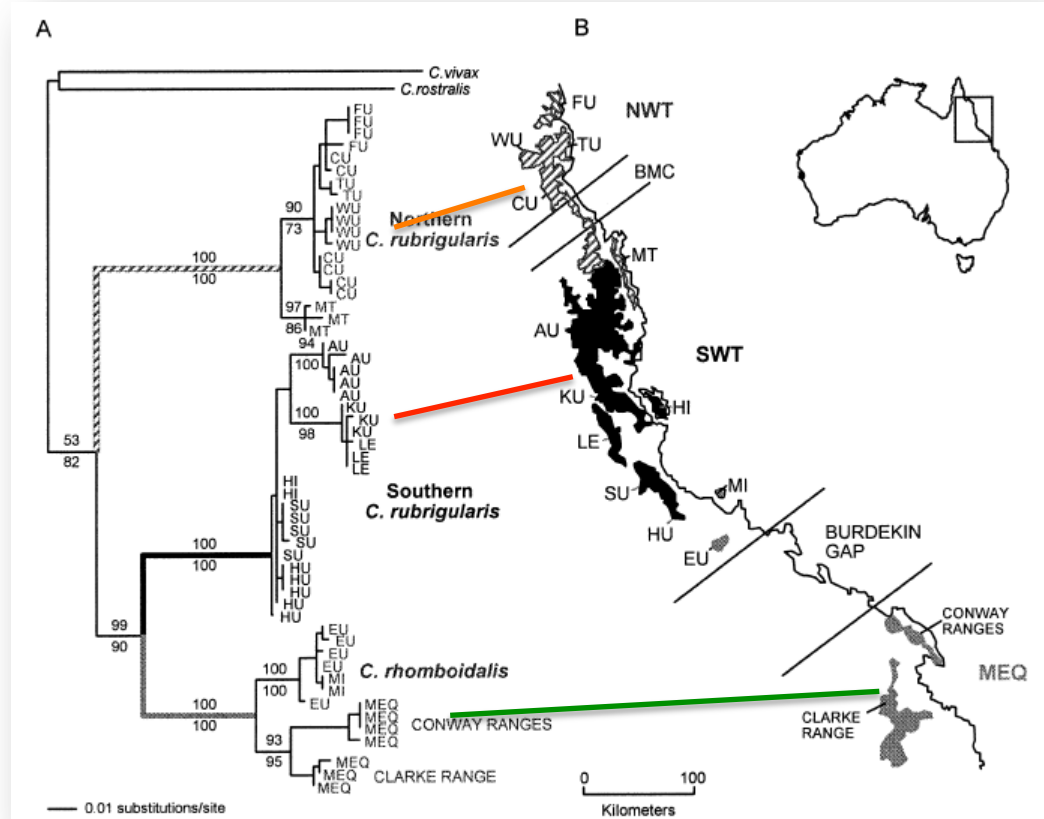
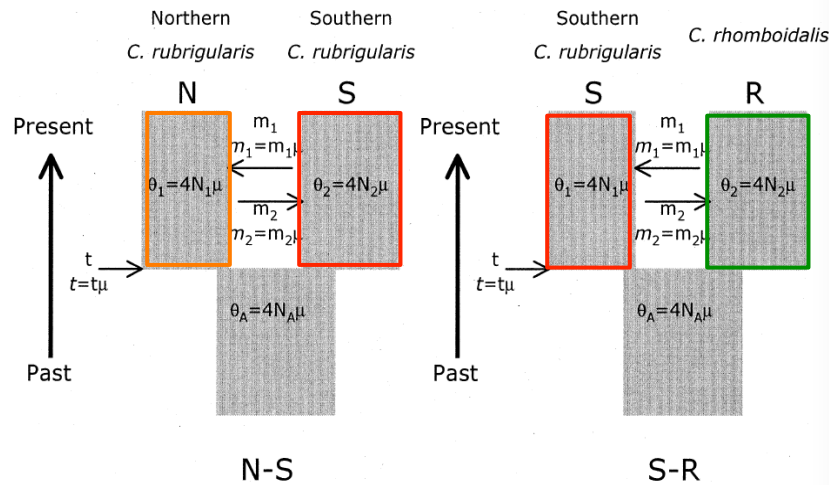


Dolman & Moritz 2006



	Population size			Time	Migration $m = m/\mu$		Migration ($2Nm$)	
	N-S	N	S		S to N	N to S	S to N	N to S
90% HPD		0.674	0.727	0.676	0.758	0.118	0.040	0
S-R		0.410-1.066	0.449-1.106	?	?	0-1.073	0-0.361	0-0.510
		S	R	Ancestral	S-R	R to S	R to S	S to R
90% HPD		0.871	0.602	0.306	0.323	0	0	0
		0.471-1.344	0.317-1.064	?	?	0-1.478	0-0.644	0-0.220

Dolman & Moritz 2006



	Population size			Time		Migration $m = m/\mu$		Migration ($2Nm$)	
	N-S	N	S	Ancestral	N-S	S to N	N to S	S to N	N to S
90% HPD		0.674	0.727	0.676	0.758	0.118	0	0.040	0
S-R		0.410-1.066	0.449-1.106	?	?	0-1.073	0-1.403	0-0.361	0-0.510
		S	R	Ancestral	S-R	R to S	S to R	R to S	S to R
90% HPD		0.871	0.602	0.306	0.323	0	0	0	0
		0.471-1.344	0.317-1.064	?	?	0-1.478	0-0.733	0-0.644	0-0.220

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

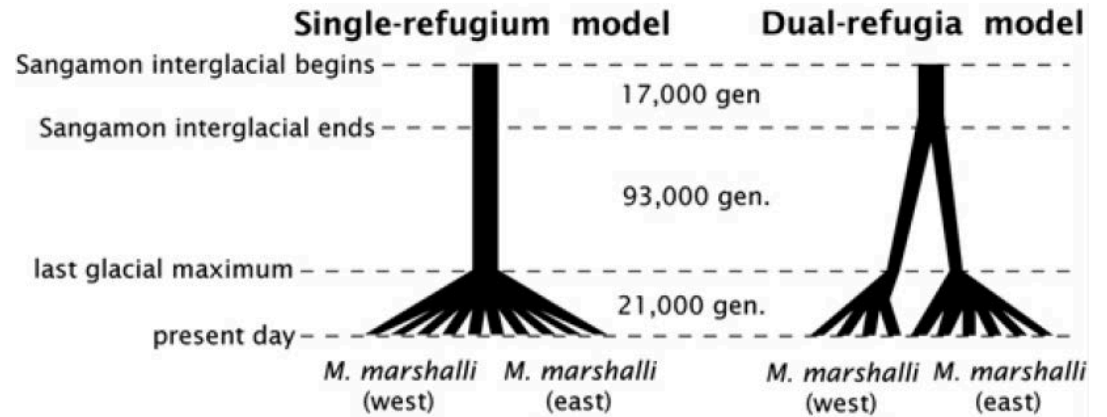
- *over-interpretation* – more detailed historical scenarios are proposed than the data support (Knowles & Maddison 2002)

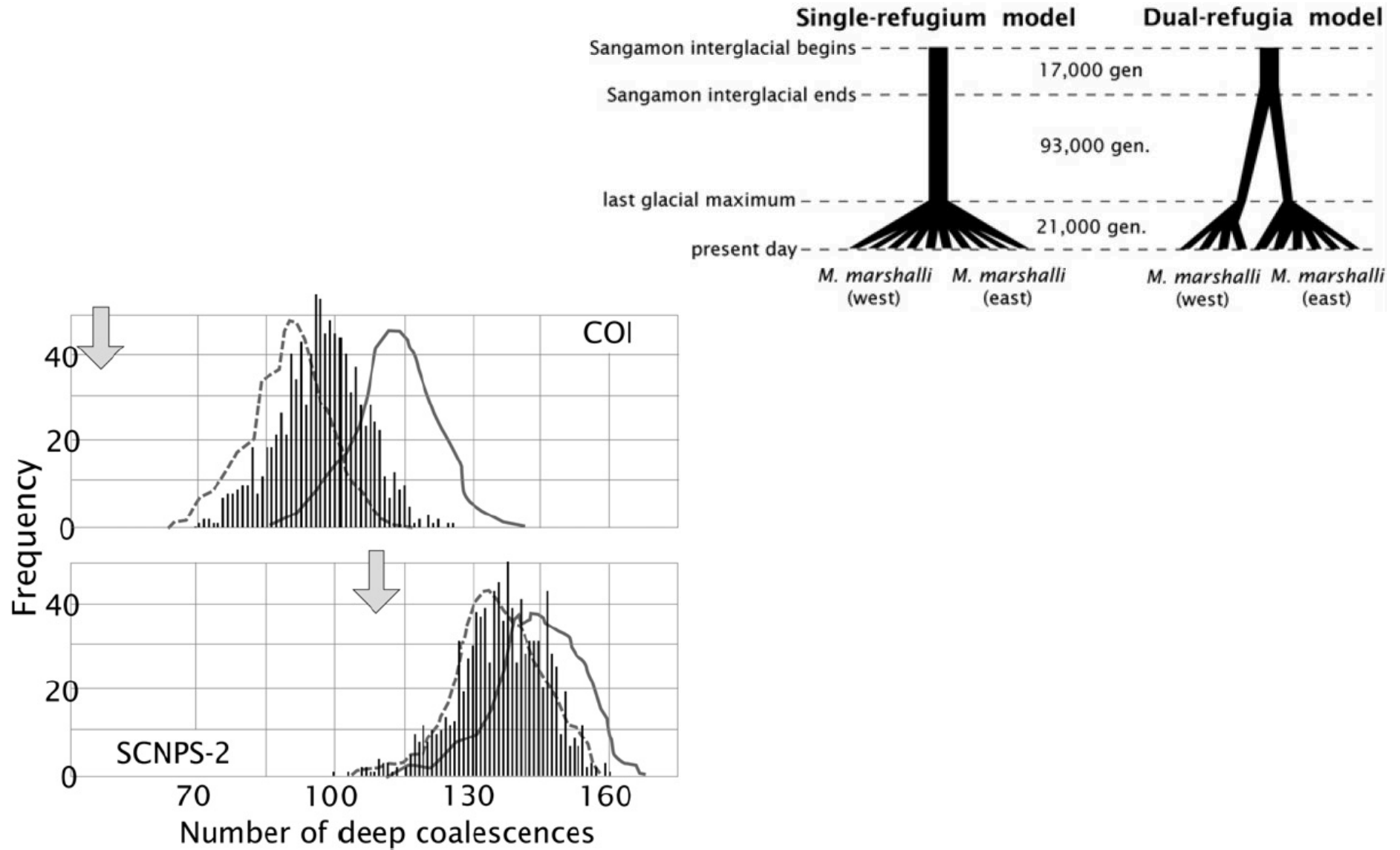
Summaries and estimates are formally generated, but interpreted by researchers in a **qualitative** manner.

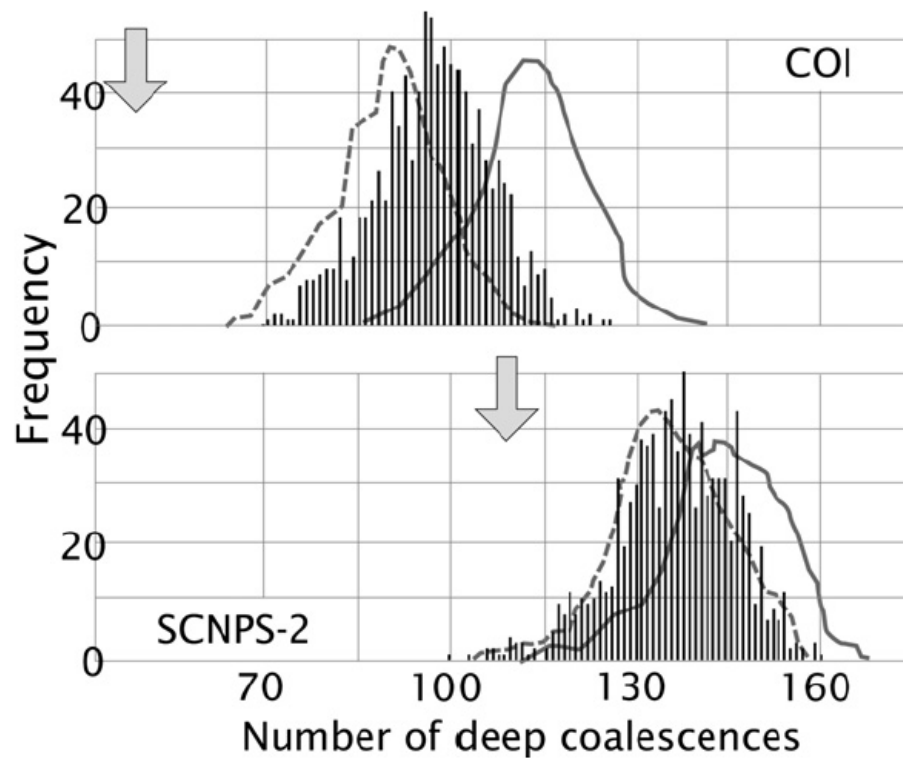
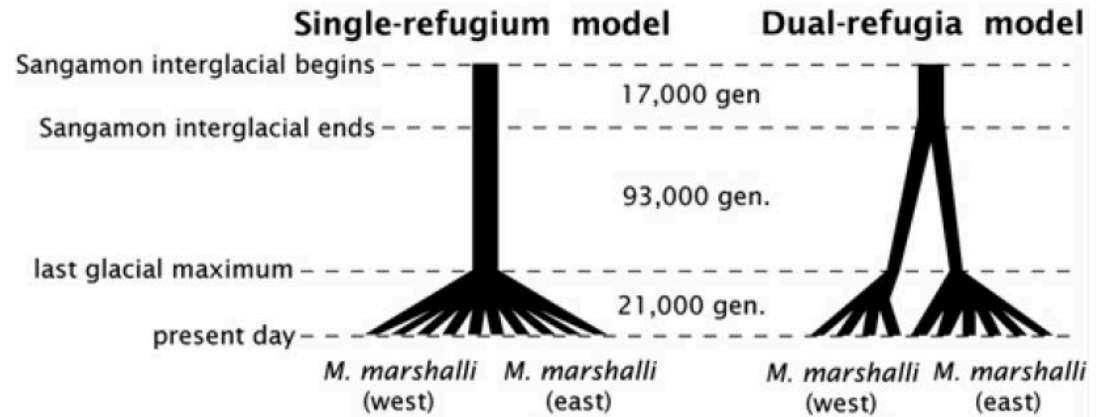
- *over-interpretation* – more detailed historical scenarios are proposed than the data support (Knowles & Maddison 2002)
- *confirmation bias* – novel information is interpreted in a manner consistent with preconceived ideas (Nickerson 1998)

How should we analyze our data? Goal is to understand how genetic diversity is partitioned across the landscape structure and identify the forces that led to this pattern.

- population structure
- population size ($\theta = 4N_e\mu$)
- divergence time (τ)
- magnitude of population size change (γ)
- gene flow (m)







Assumptions

- accuracy of θ_i , other values
- adequacy of sampling strategy
- timing of population model
- topology of population model
- adequacy of summary statistics

Hypothesis-testing is not the best way to move beyond qualitative data analysis.

- Rejecting an unrealistic hypothesis tells us nothing about an empirical system, and may promote false confidence regarding our understanding of the system.
- It is also impossible to differentiate among hypotheses that can not be rejected.

Phylogeography is a *historical* discipline . . .

. . . that uses statistical tools developed for *experimental* research.

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

Phylogeography is a *historical* discipline . . .

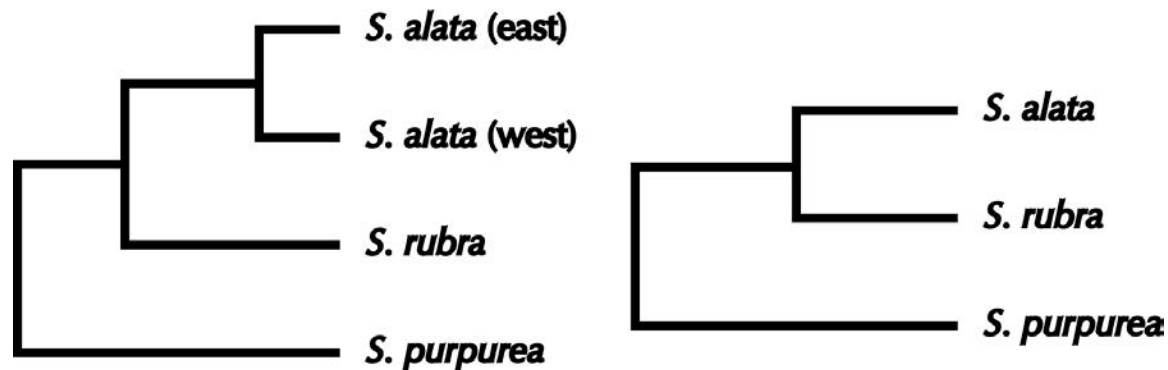
. . . that uses statistical tools developed for *experimental* research.

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

Information theoretic approach. Calculate ***Prob*** (**model**; | **data**) for multiple models, rank using AIC or other metrics.

Species delimitation using species trees

- Compare the probability of models where putative lineages are separate to the probability of models where they are the same.



Syst. Biol. 56(6):887–895, 2007
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701701091

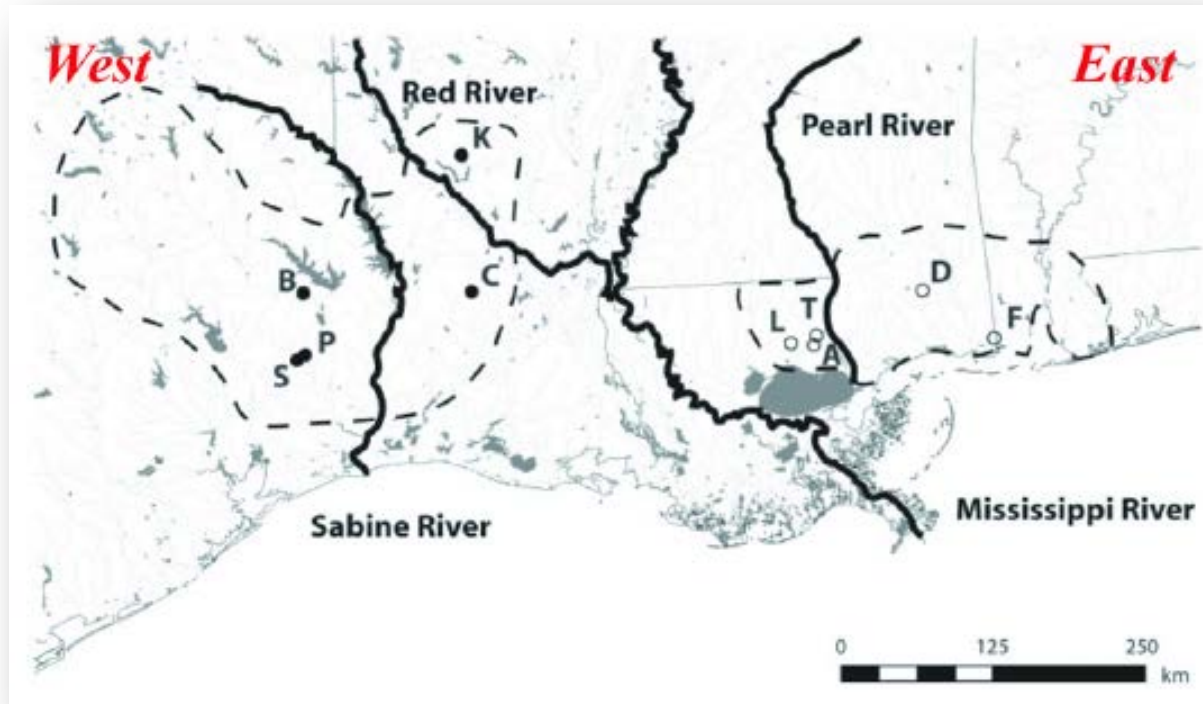
Delimiting Species without Monophyletic Gene Trees

L. LACEY KNOWLES AND BRYAN C. CARSTENS

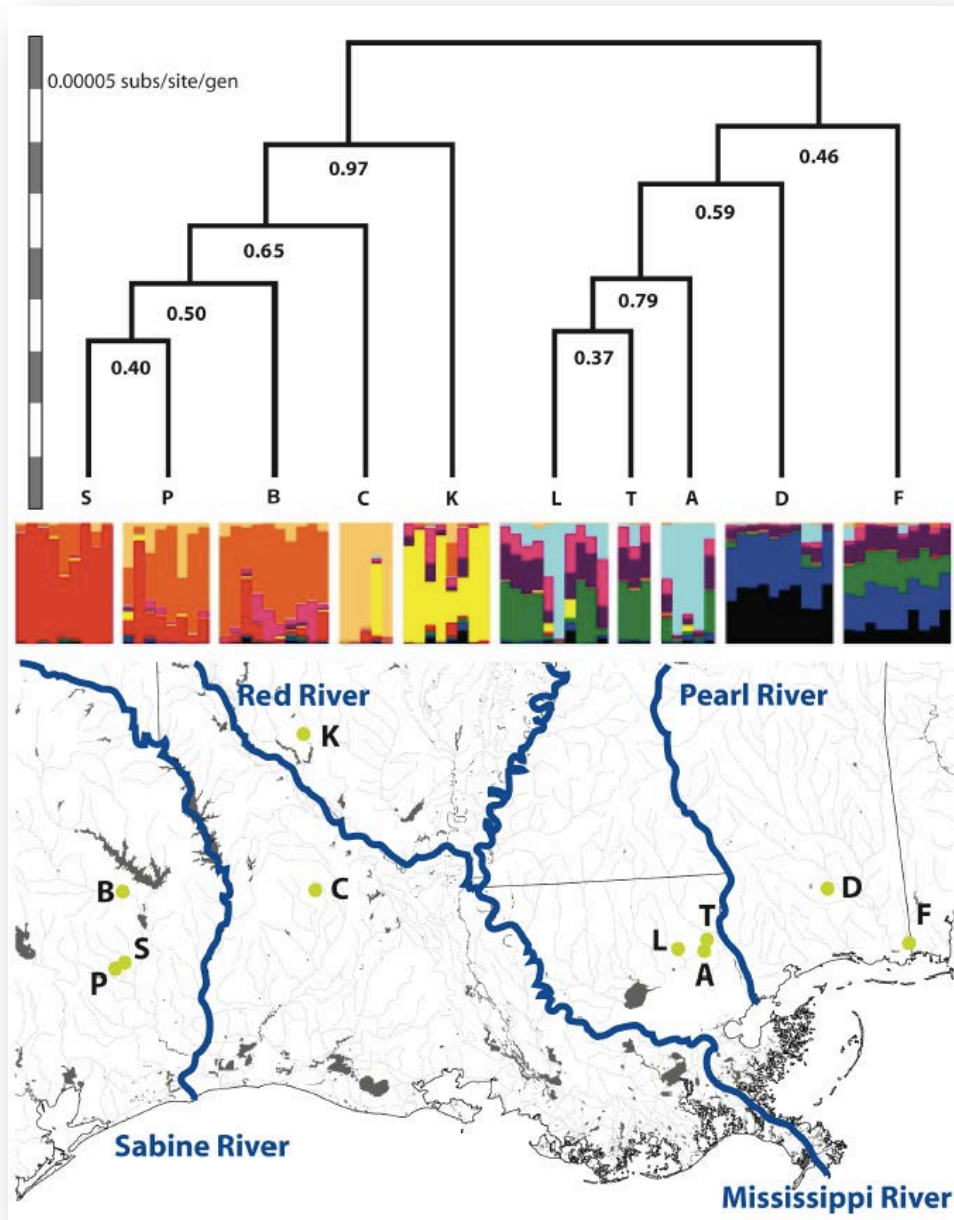
Department of Ecology and Evolutionary Biology, Museum of Zoology, 1109 Geddes Avenue, University of Michigan,
Ann Arbor, MI 48109-1079, USA; E-mail: knowlesl@umich.edu (L.L.K.)

Species delimitation.

Sarracenia alata



21,147 permutations of 10 populations!



Jordan Satler

Structurama (Huelsenbeck et al. 2011)

BPP (Yang and Rannala 2010)

spedeSTEM (Ence & Carstens 2011)

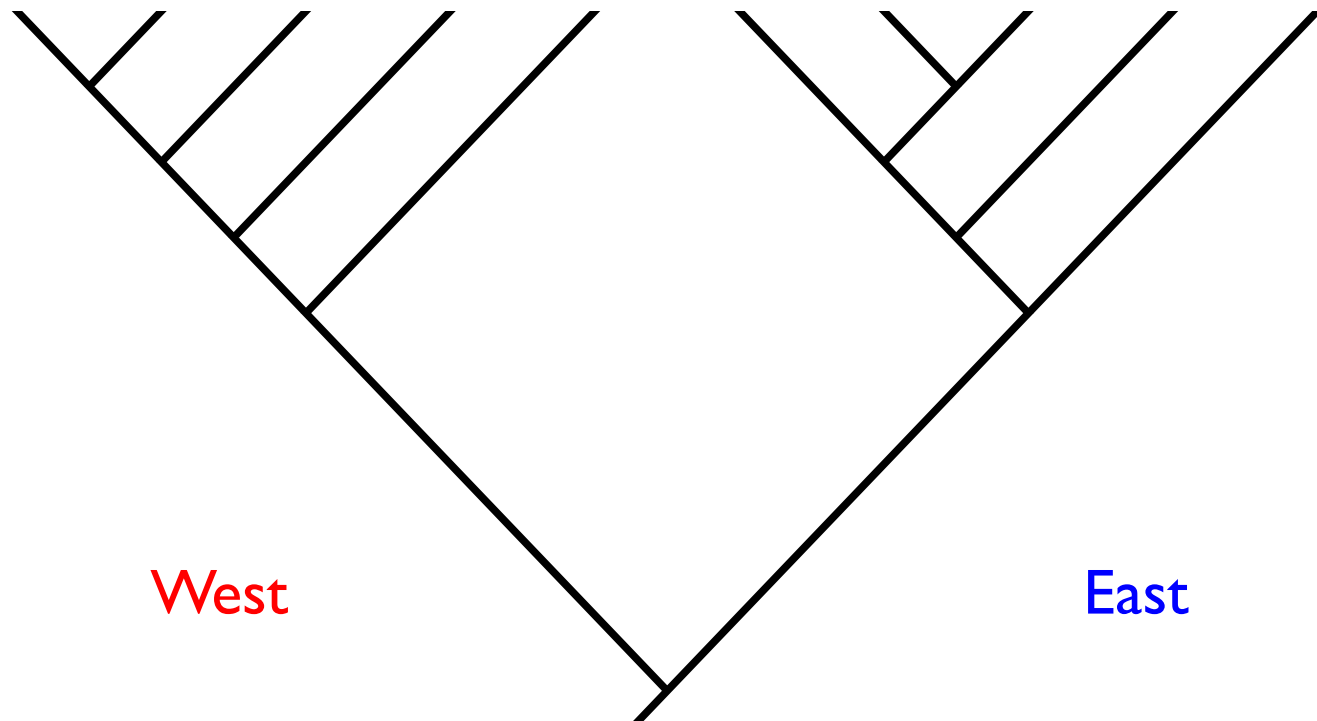
BP&P



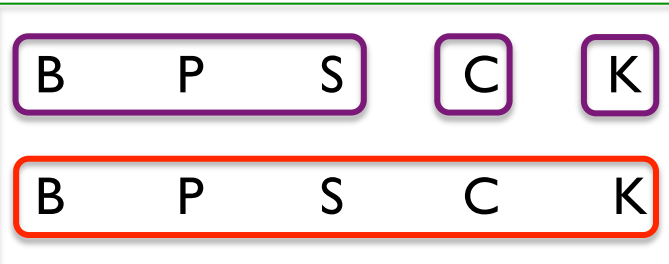
SpedeSTEM



Structurama



BP&P



Structurama



Biological Journal of the Linnean Society, 2013, **109**, 737–746. With 2 figures

The carnivorous plant described as *Sarracenia alata* contains two cryptic species

BRYAN C. CARSTENS* and JORDAN D. SATLER

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

Received 17 January 2013; revised 19 February 2013; accepted for publication 19 February 2013

Limitations of existing methods

- phylogentic models that do not allow gene flow / population expansion
- genetic clustering methods do not model temporal divergence



MOLECULAR ECOLOGY

Molecular Ecology (2013) 22, 4369–4383

doi: 10.1111/mec.12413

INVITED REVIEWS AND META-ANALYSES

How to fail at species delimitation

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

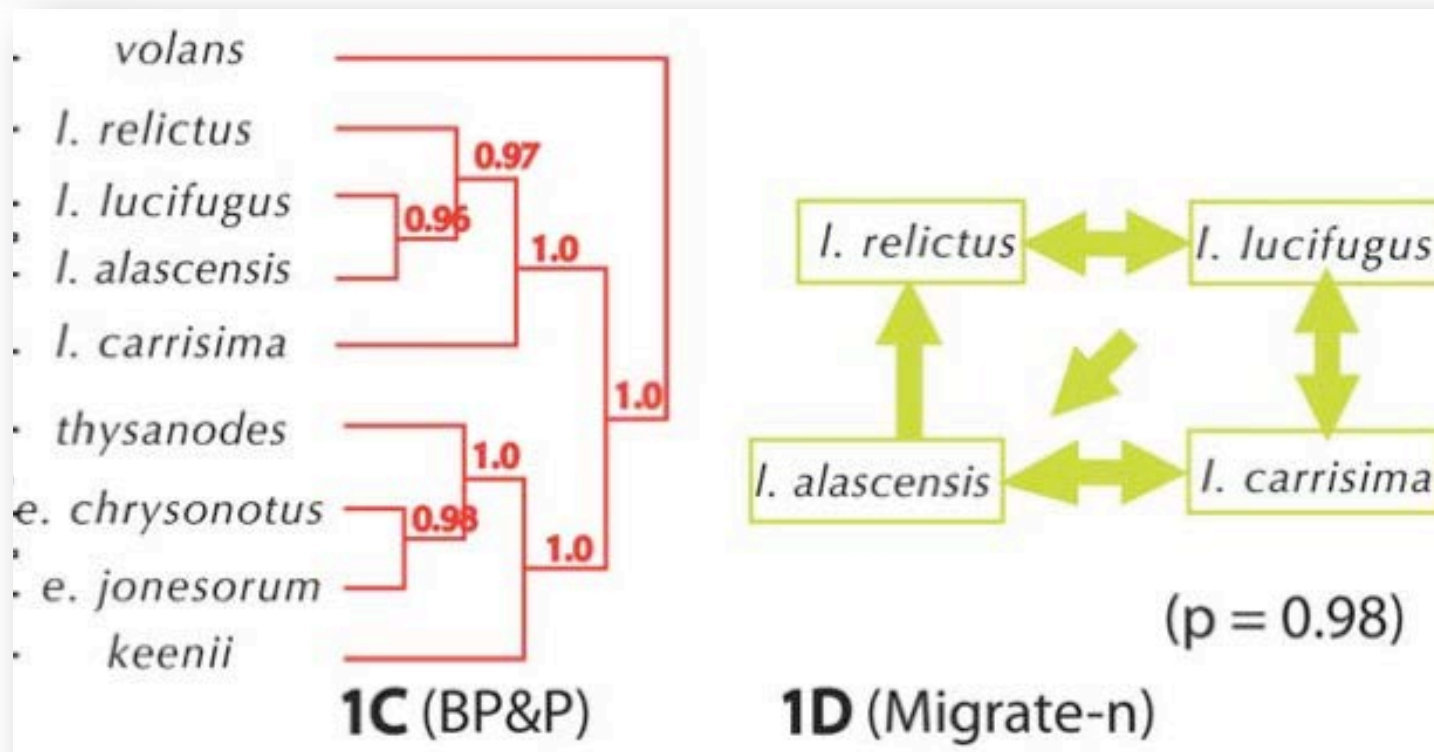
*Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Avenue, Columbus, OH43210-1293, USA, †Department of Biological Sciences, Louisiana State University, Life Sciences Building, Baton Rouge, LA 70803, USA

Species delimitation.

Myotis lucifugus



Ariadna Morales-Garcia



lightning talk – bryan.c.ca

SpedeSTEM Web Interface

← → ↻ ⌂

spedestem.asc.ohio-state.edu

🔍 ☆ ☰

SpedeSTEM Online

Run SpedeSTEM

Resources

Contact Us


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](#).
[See this file for more help](#)


[Sign up](#) [Login](#)

About us

Research in the Carstens lab seeks to understand how biological diversity is generated using computational approaches. We investigate empirical systems by identifying the limits of evolutionary lineages, in order to evaluate the relative contributions of evolutionary processes and infer the ecological and environmental forces that have contributed to the formation of population genetic structure.



Department of Ecology Evolution and Organismal Biology

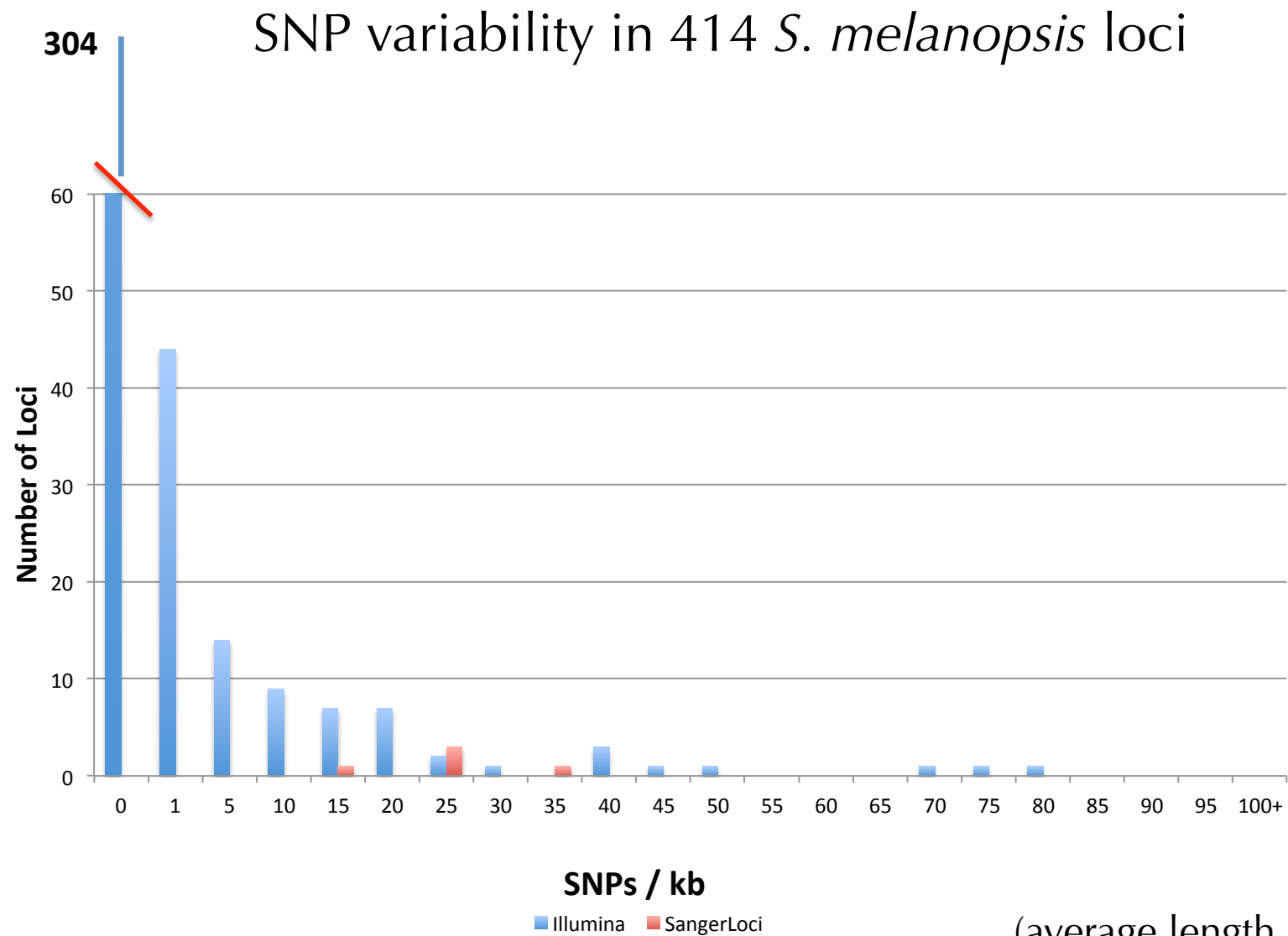


National Science Foundation
WHERE DISCOVERIES BEGIN

Funded by NSF DEB 0918212



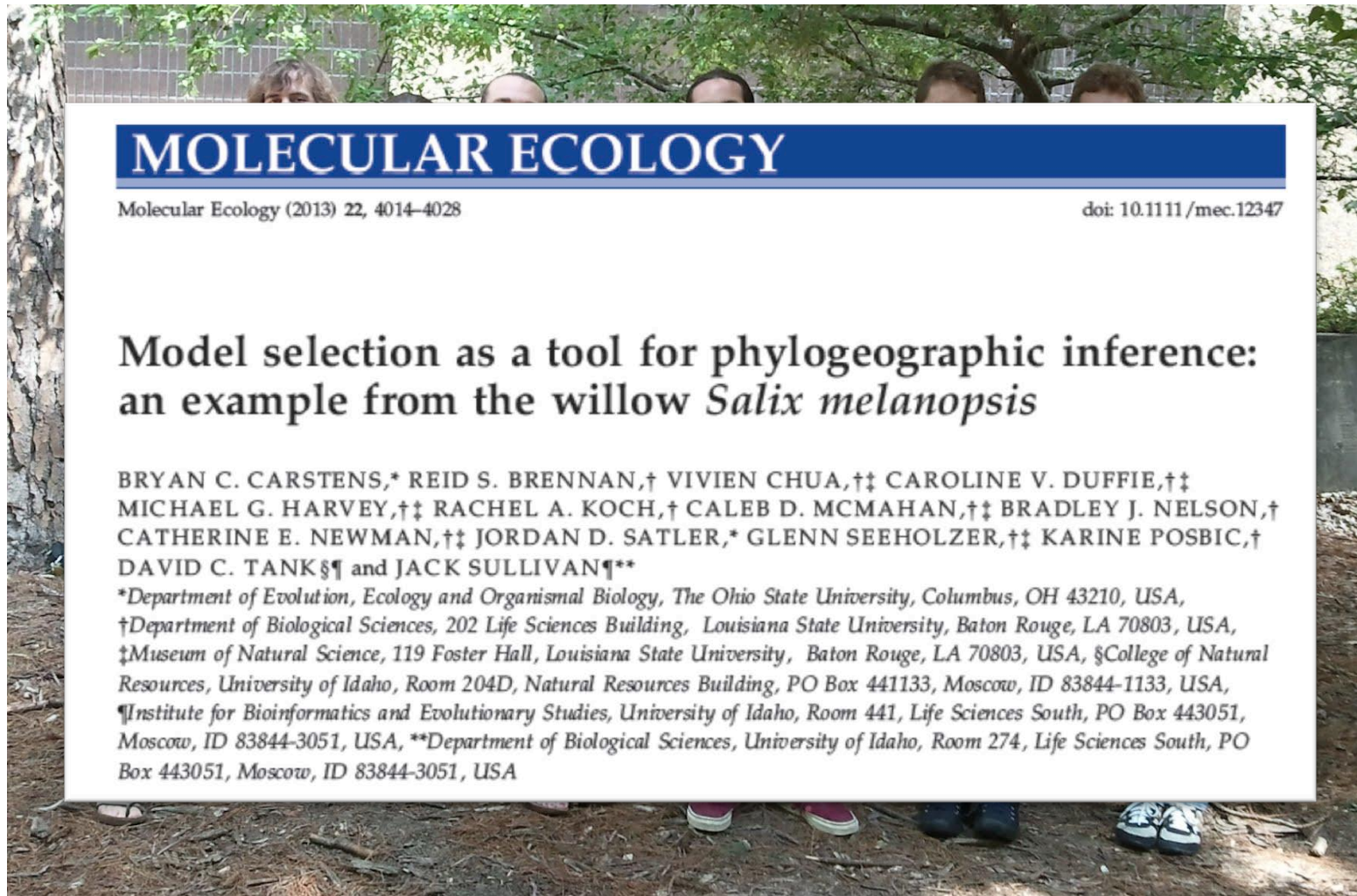
- Genomic DNA extracted from leaf tissue, sheared with **Bioruptor** (Diagenode)
- Sequenced ~400 bp fragments using Roche 454 sequencing
- **RepeatMasker** (Smit 2008) to screen resulting sequencing for low-complexity DNA and repetitive elements
- **MYbaits** custom enrichment; probes designed by Jean-Marie Rouilliard (Mycroarray, Ann Arbor).
- sequenced 39 *S. melanopsis* and 1 *S. alba* on Illumina GAiiX (8 samples / lane); $\sim 3.76 \times 10^6$ high quality (108 bp) reads per sample
- reference genome using *Populus trichocarpa* genome and *de novo* assembly (**Velvet**; Zerbino and Birney 2008)
- Assembly and SNP-calling (100x coverage) using **SAMtools** (Li et al. 2009)



(average length 620 bp)



Brad Nelson, Jordan Satler, Caleb McMahon, Glen Seeholzer, Mike Harvey (back row)
Rachel Koch, Caroline Duffy, Cathy Newman, Reid Brennan, Vivian Chua, Karine Probsic (front row)



MOLECULAR ECOLOGY

Molecular Ecology (2013) 22, 4014–4028

doi: 10.1111/mec.12347

Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*

BRYAN C. CARSTENS,* REID S. BRENNAN,† VIVIEN CHUA,†‡ CAROLINE V. DUFFIE,†‡
MICHAEL G. HARVEY,†‡ RACHEL A. KOCH,† CALEB D. MCMAHAN,†‡ BRADLEY J. NELSON,†
CATHERINE E. NEWMAN,†‡ JORDAN D. SATLER,* GLENN SEEHOLZER,†‡ KARINE POSBIC,†
DAVID C. TANK§¶ and JACK SULLIVAN¶**

*Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA,

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

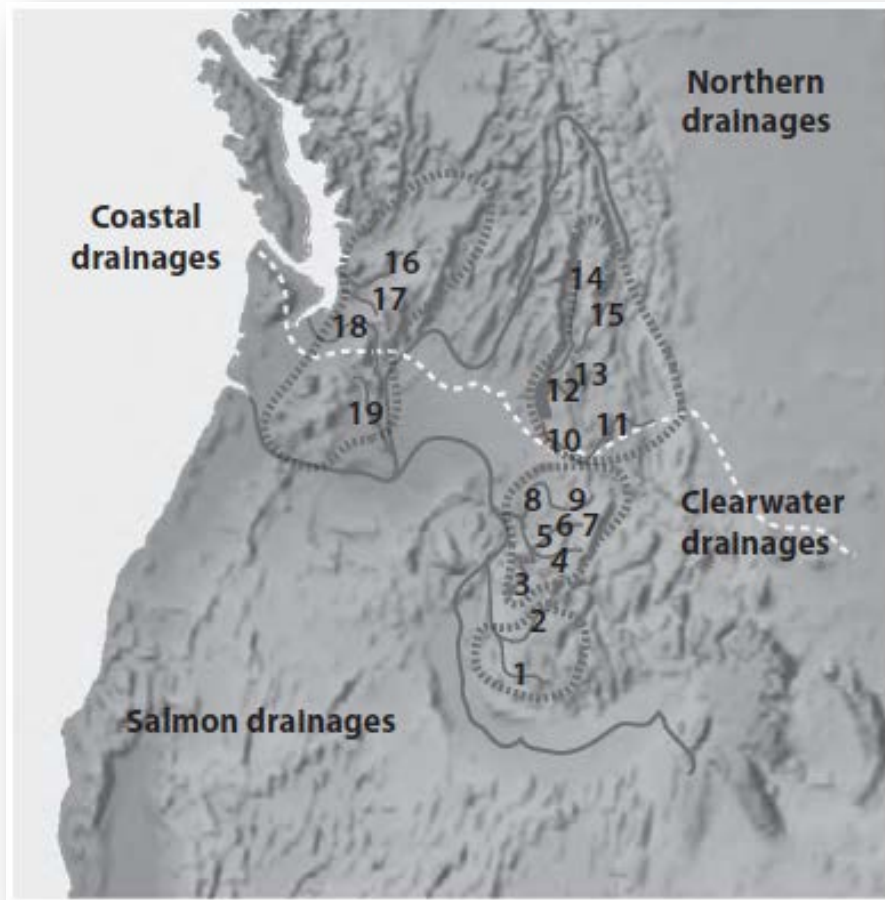
‡Museum of Natural Science, 119 Foster Hall, Louisiana State University, Baton Rouge, LA 70803, USA, §College of Natural

Resources, University of Idaho, Room 204D, Natural Resources Building, PO Box 441133, Moscow, ID 83844-1133, USA,

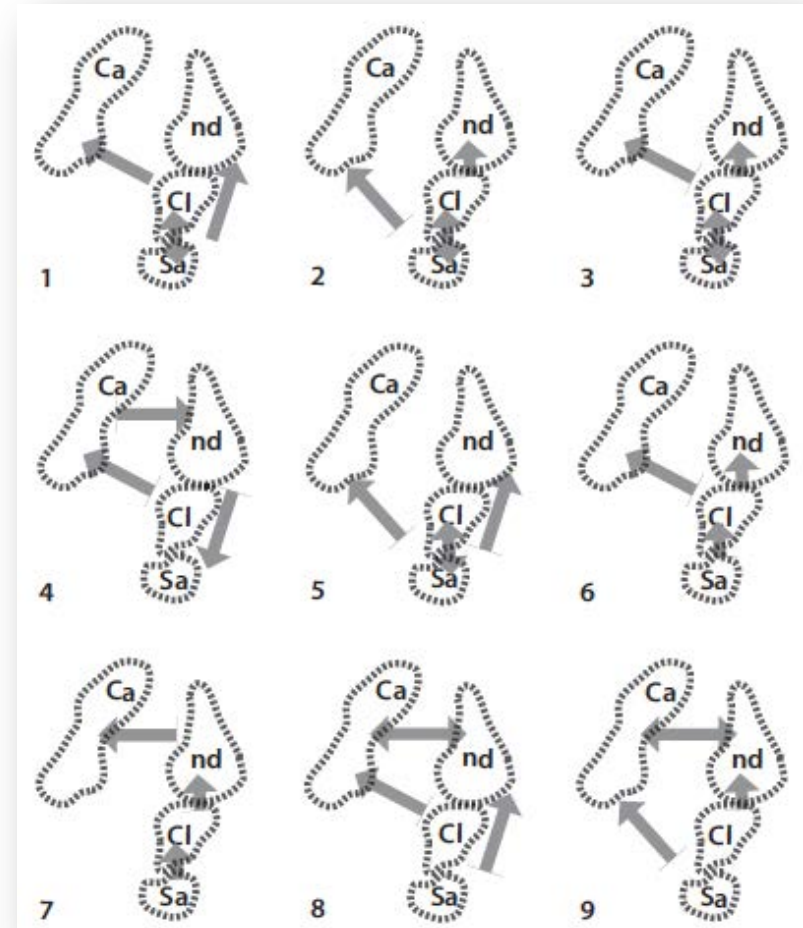
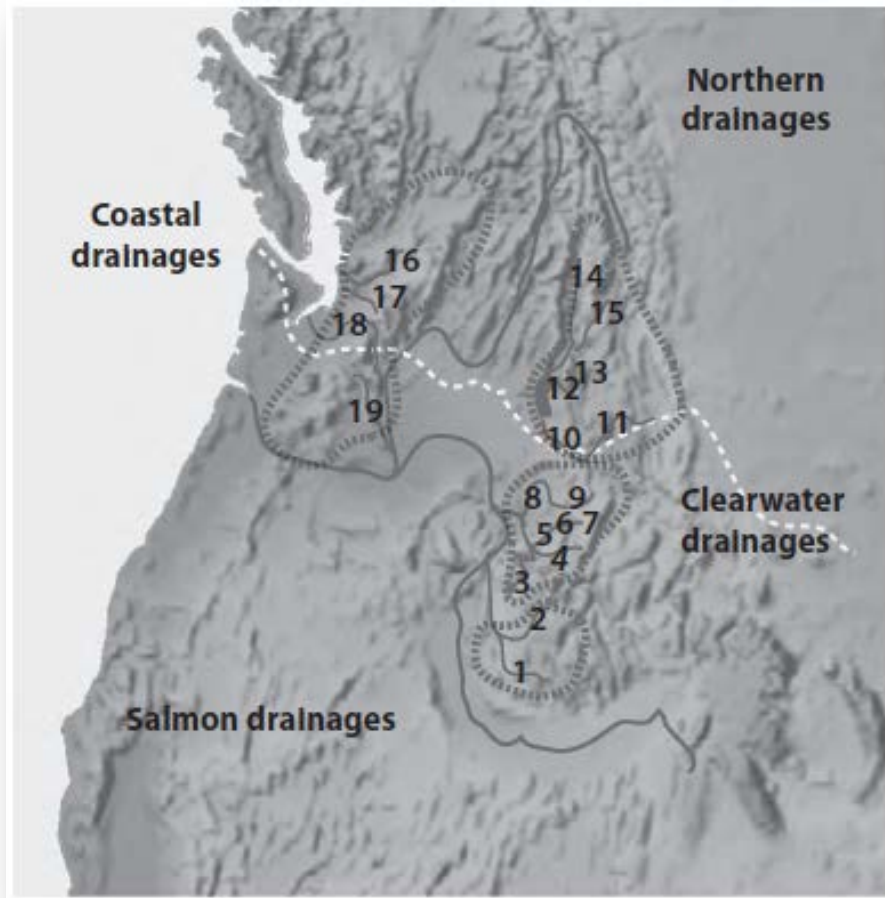
¶Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Room 441, Life Sciences South, PO Box 443051,

Moscow, ID 83844-3051, USA, **Department of Biological Sciences, University of Idaho, Room 274, Life Sciences South, PO

Box 443051, Moscow, ID 83844-3051, USA

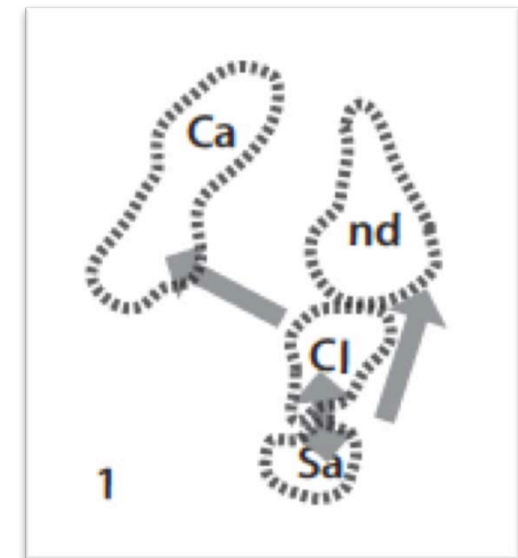
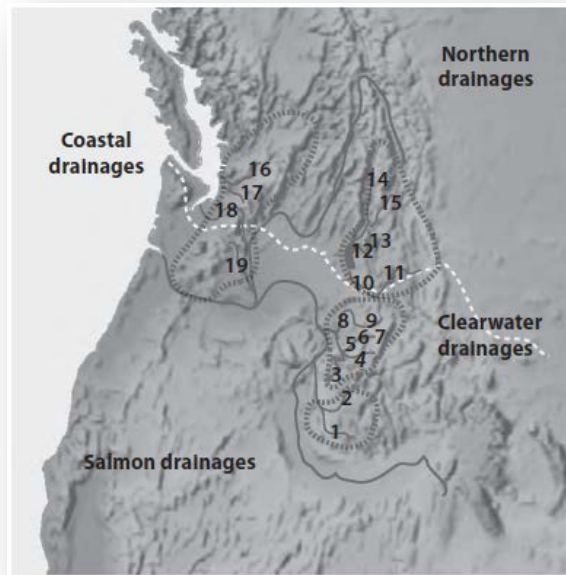


$$\mathcal{P} = \begin{pmatrix} \Theta_1 & \mathcal{M}_{21} & \mathcal{M}_{31} & \dots & \mathcal{M}_{n1} \\ \mathcal{M}_{12} & \Theta_2 & \mathcal{M}_{32} & \dots & \mathcal{M}_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ \mathcal{M}_{1n} & \mathcal{M}_{2n} & \dots & \mathcal{M}_{n-1,n} & \Theta_n \end{pmatrix}, \quad \text{Migrate-n (Peter Beerli et al.)}$$



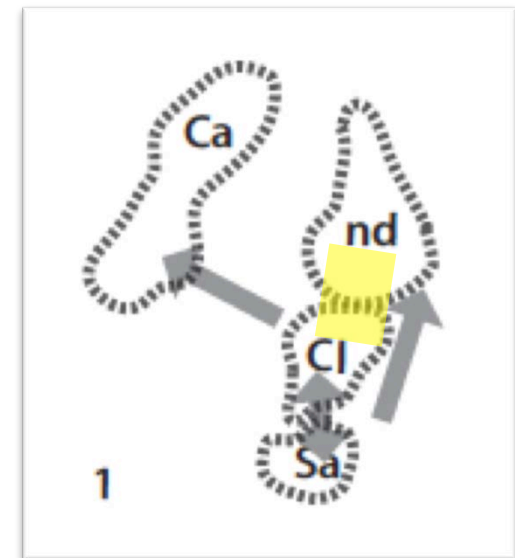
$$\mathcal{P} = \begin{pmatrix} \Theta_1 & \mathcal{M}_{21} & \mathcal{M}_{31} & \dots & \mathcal{M}_{n1} \\ \mathcal{M}_{12} & \Theta_2 & \mathcal{M}_{32} & \dots & \mathcal{M}_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ \mathcal{M}_{1n} & \mathcal{M}_{2n} & \dots & \mathcal{M}_{n-1,n} & \Theta_n \end{pmatrix},$$

Migrate-n (Peter Beerli *et al.*)

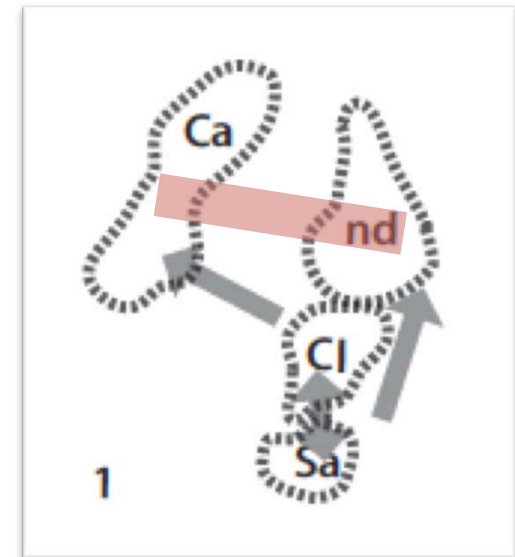


No.	Description of model	Migration pattern	Harmonic ImL	P
1	Two refugia, with Clearwater as source of Cascades and Salmon as source for northern	$Ca \leftrightarrow Cl \leftrightarrow Sa \rightarrow No$	-43872.35	0.99976
2	Two refugia, with Salmon as source of Cascades and Clearwater as source for northern	$Ca \leftrightarrow Sa \leftrightarrow Cl \rightarrow No$	-43880.67	0.00024
3	Single refuge (Clearwater)	$Sa \leftrightarrow Cl \rightarrow N; Cl \rightarrow Ca$	-43950.06	0.00000
4	Stepping stone (Clearwater)	$Cl \rightarrow Ca \rightarrow N \rightarrow Sa$	-43966.94	0.00000
5	Single refuge (Salmon)	$Cl \leftrightarrow Sa \rightarrow N; Sa \rightarrow Ca$	-43997.26	0.00000
6	Stepping stone (Salmon), with Clearwater source for Cascades	$Sa \rightarrow Cl \rightarrow Ca; Cl \rightarrow N$	-44039.82	0.00000
7	Stepping stone (Salmon)	$Sa \rightarrow Cl \rightarrow N \rightarrow Ca$	-44122.03	0.00000
8	Clearwater and Salmon are sources, sending migrants to Cascades and northern, respectively, and those exchange migrants	$Cl \rightarrow Ca \leftrightarrow N \leftarrow Sa$	-44150.8	0.00000
9	Clearwater and Salmon are sources, sending migrants to northern and Cascades, respectively, and those exchange migrants	$Cl \rightarrow N \leftrightarrow Ca \leftarrow Sa$	-44259.9	0.00000

Parameter	2.50%	Mode	97.50%
θ_1	0.002	0.00397	0.00574
θ_2	0.00247	0.00417	0.0058
θ_3	0.0022	0.00404	0.0058
θ_4	0.00067	0.00237	0.00407
$M_{2 \rightarrow 1}$	1.18	2.92	4.95
$M_{3 \rightarrow 1}$	1.03	2.56	4.48
$M_{4 \rightarrow 1}$	1.22	2.96	5.19
$M_{1 \rightarrow 2}$	1.27	2.64	4.56
$M_{3 \rightarrow 2}$	1.49	3.06	5.21
$M_{4 \rightarrow 2}$	1.37	2.89	4.92
$M_{1 \rightarrow 3}$	1.33	2.85	5.22
$M_{2 \rightarrow 3}$	1.53	3.45	5.80
$M_{4 \rightarrow 3}$	1.49	3.47	5.77
$M_{1 \rightarrow 4}$	0.34	1.45	3.18
$M_{2 \rightarrow 4}$	0.37	1.58	3.41
$M_{3 \rightarrow 4}$	0.31	1.41	3.03



Parameter	2.50%	Mode	97.50%
θ_1	0.002	0.00397	0.00574
θ_2	0.00247	0.00417	0.0058
θ_3	0.0022	0.00404	0.0058
θ_4	0.00067	0.00237	0.00407
$M_{2 \rightarrow 1}$	1.18	2.92	4.95
$M_{3 \rightarrow 1}$	1.03	2.56	4.48
$M_{4 \rightarrow 1}$	1.22	2.96	5.19
$M_{1 \rightarrow 2}$	1.27	2.64	4.56
$M_{3 \rightarrow 2}$	1.49	3.06	5.21
$M_{4 \rightarrow 2}$	1.37	2.89	4.92
$M_{1 \rightarrow 3}$	1.33	2.85	5.22
$M_{2 \rightarrow 3}$	1.53	3.45	5.80
$M_{4 \rightarrow 3}$	1.49	3.47	5.77
$M_{1 \rightarrow 4}$	0.34	1.45	3.18
$M_{2 \rightarrow 4}$	0.37	1.58	3.41
$M_{3 \rightarrow 4}$	0.31	1.41	3.03



Diffusion Approximate Demographic Inference (*dadi*)

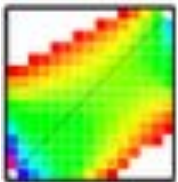
OPEN ACCESS Freely available online

PLOS GENETICS

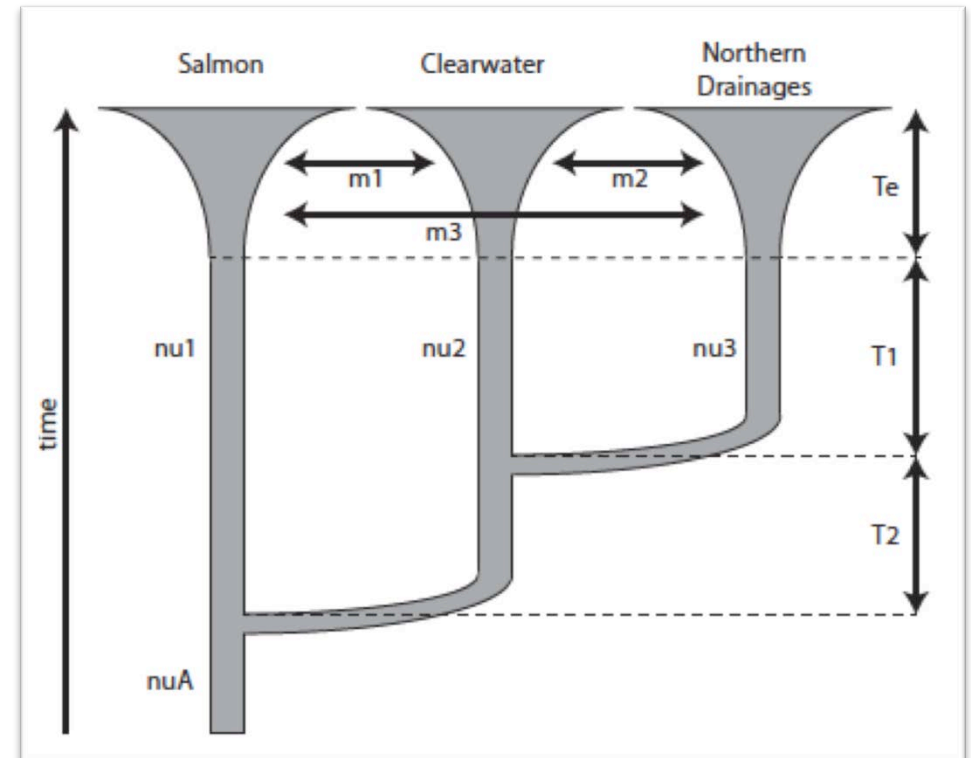
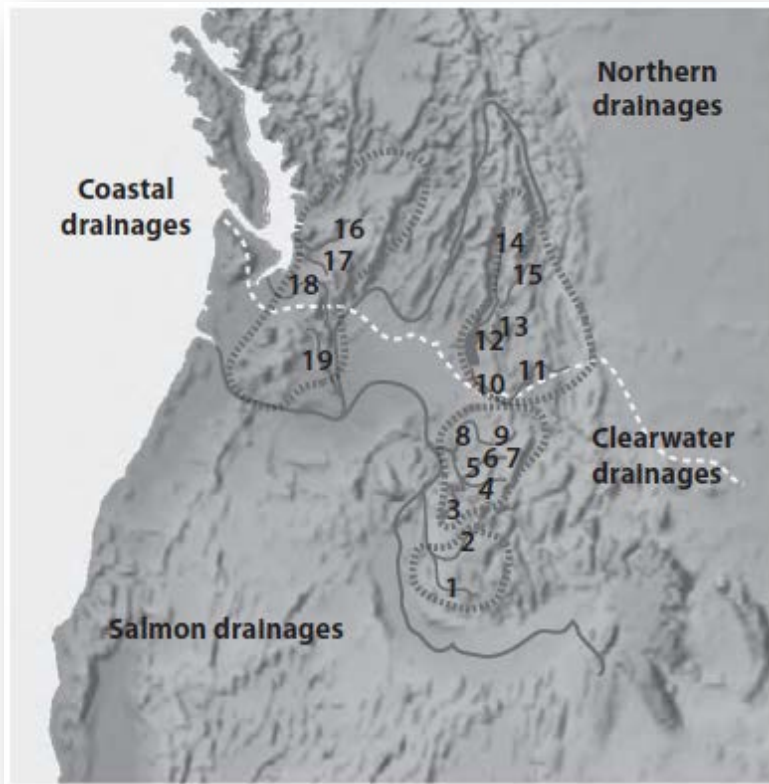
Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data

Ryan N. Gutenkunst^{1*}, Ryan D. Hernandez², Scott H. Williamson³, Carlos D. Bustamante³

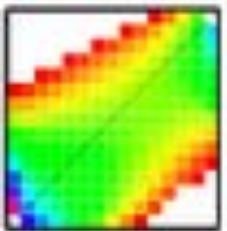
¹Theoretical Biology and Biophysics and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, ²Human Genetics, University of Chicago, Chicago, Illinois, United States of America, ³Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America



- empirical data are summarized by allele frequency spectra
- expected AFS is calculated given demographic model, parameters using diffusion approximation
- probability of data | model is calculated by composite likelihood



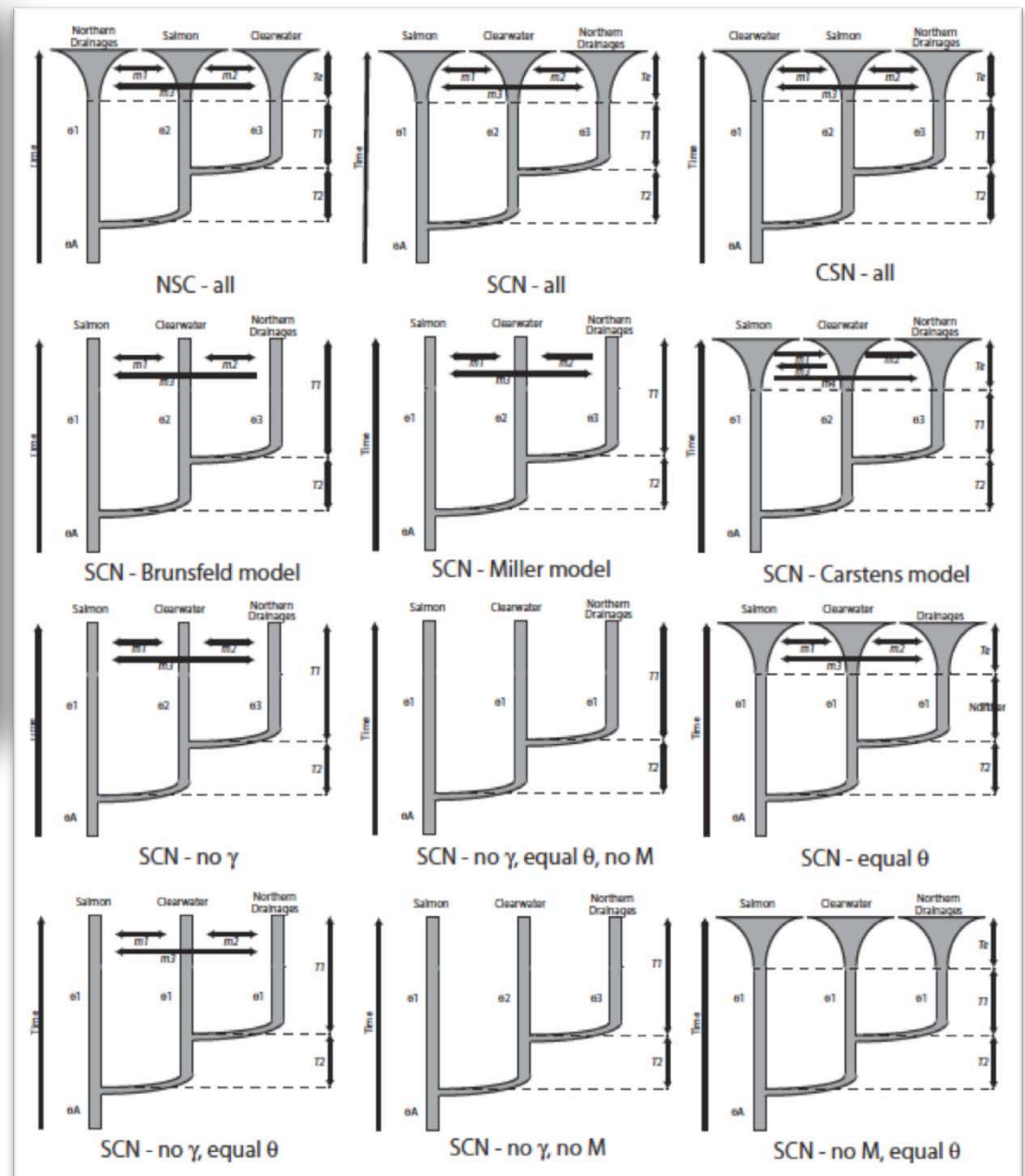
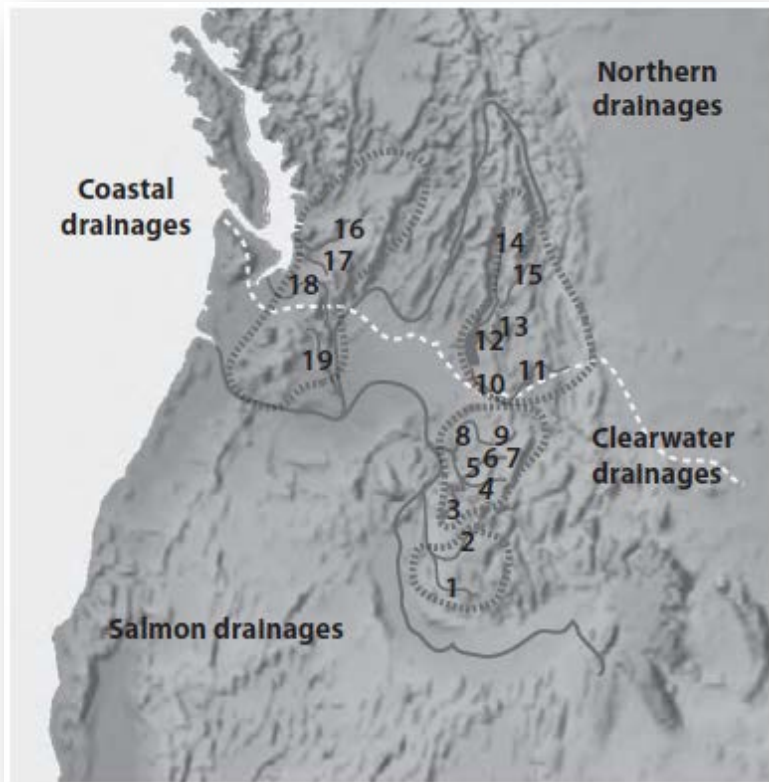
$$\ln L = - 336.939$$



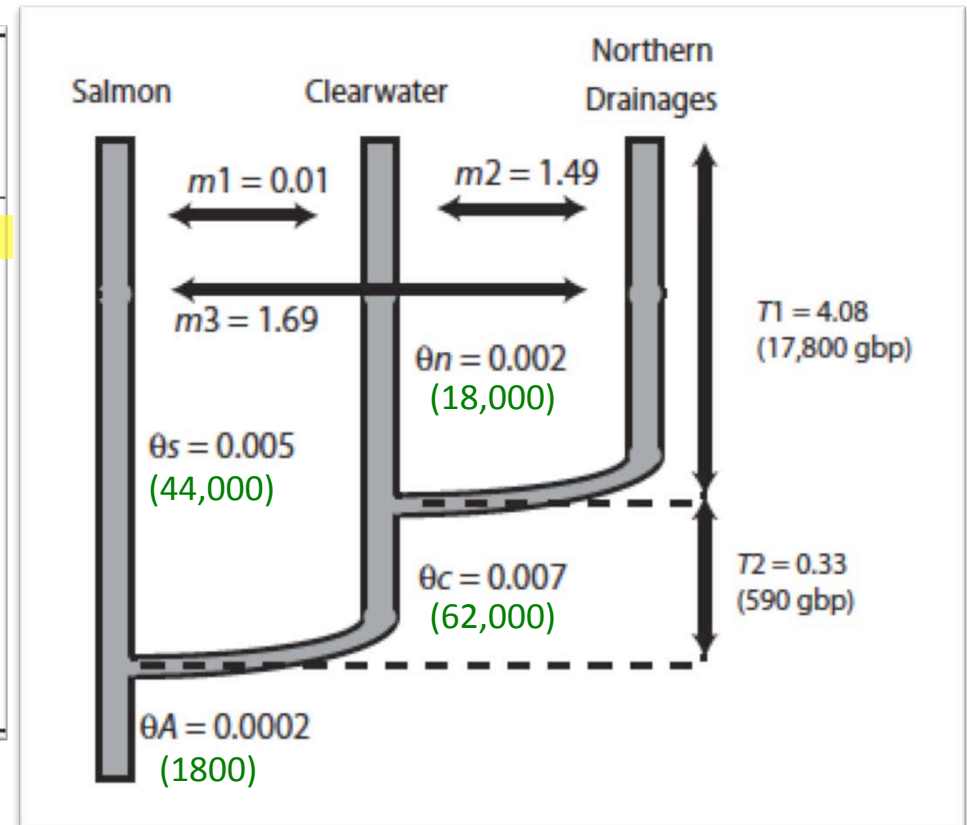
- derived from Hudson's ms, so extremely flexible in terms of the demographic models that can be specified.

Model-based inference.

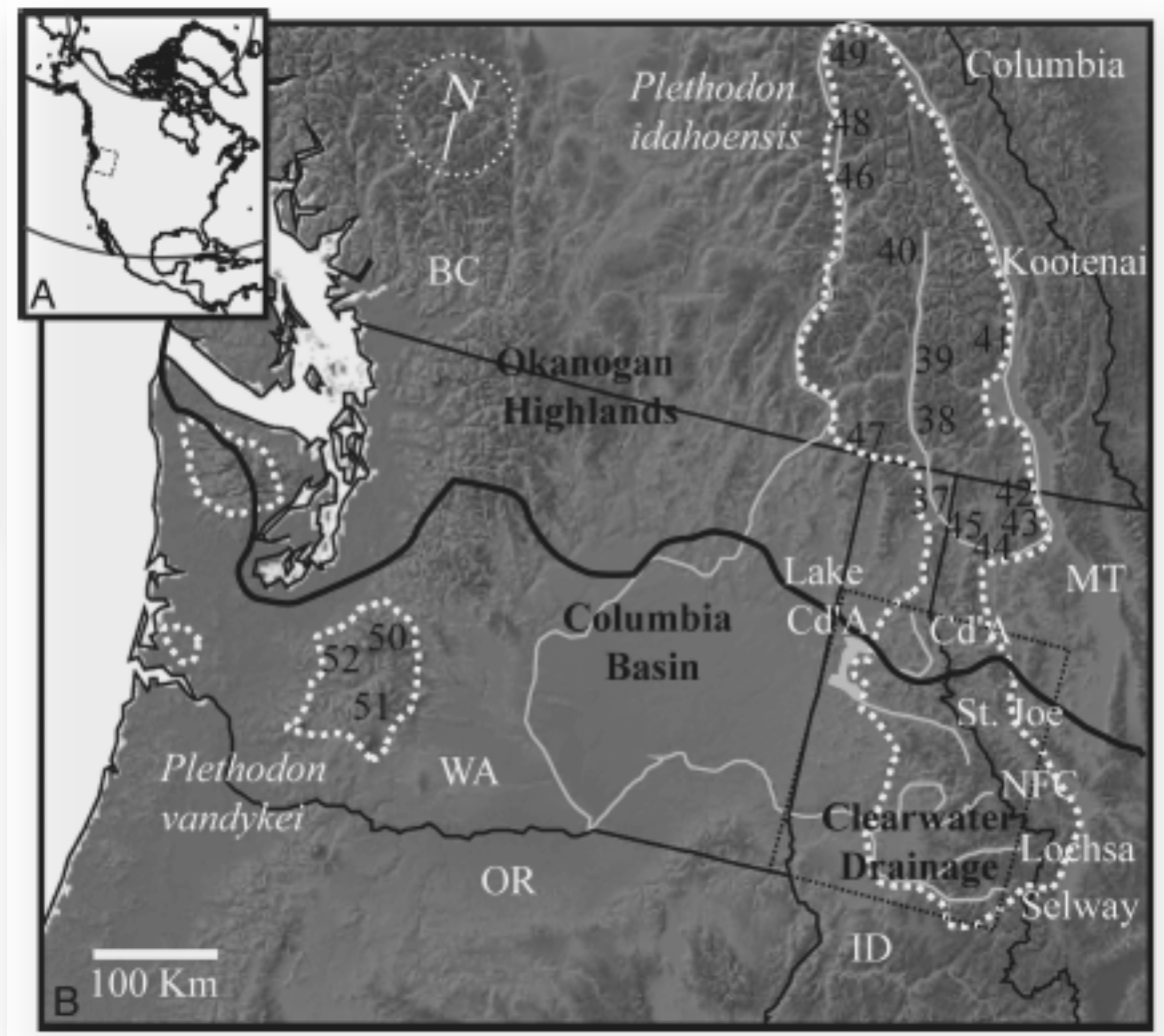
Salix melanopsis



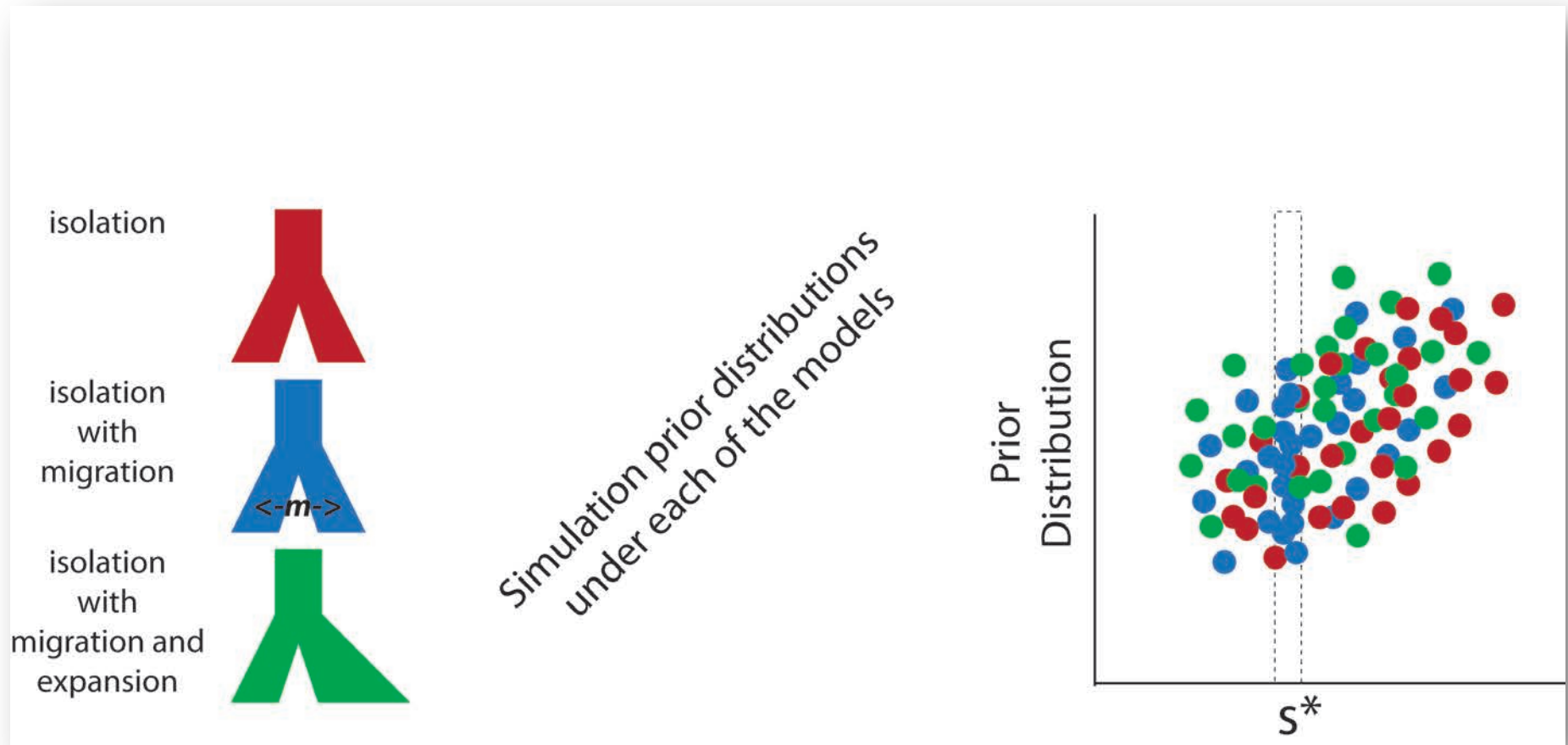
Inland temperate rainforest—three-dimensional AFS					
Model name	$-\ln L$	k	AIC	Δ_i	w_i
SCN—no γ	331.664	10	683.33	0.00	0.950
SCN—equal θ	337.202	8	690.40	7.08	0.028
SCN—no γ , no M	338.687	7	691.37	8.05	0.017
SCN—all	336.939	10	693.88	10.55	0.005
CSN—all	342.744	10	705.49	22.16	0.000
SCN—Brunsfield model	345.453	9	708.91	25.58	0.000
SCN—no γ , equal θ	349.778	7	713.56	30.23	0.000
SCN—no γ , no M, equal θ	354.255	5	718.51	35.18	0.000
SCN—Carstens model	371.995	11	765.99	82.66	0.000
NSC—all	409.803	10	839.61	156.28	0.000
SCN—Miller model	492.43	9	1002.86	319.53	0.000



- Identifying a model with a good fit to the data allows us to make inferences from estimates of the parameters that are relevant to our data.
- A very small ancestral population $\sim 18,400$ gbp gave rise to extant *S. melanopsis* in a S-N direction.



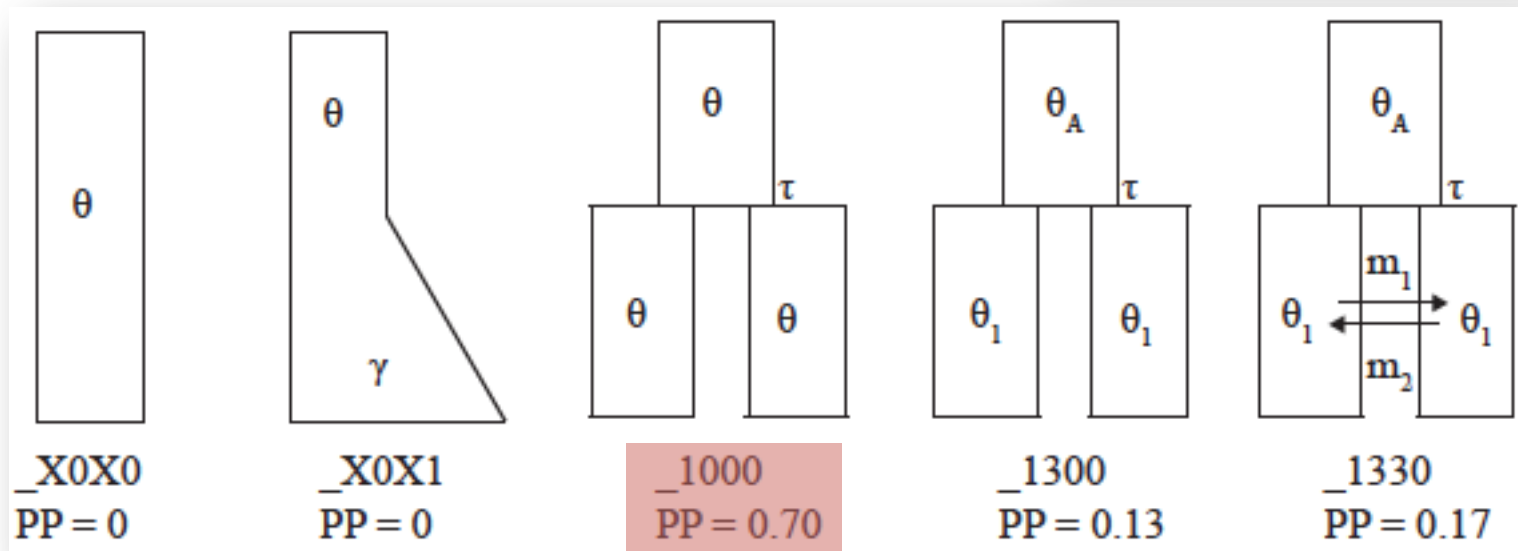
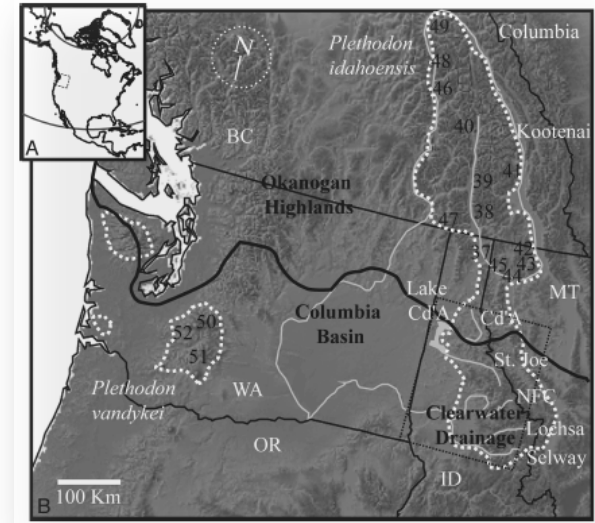
(Pelletier & Carstens, in review)



- simulate a prior distribution under a set of models using MS (Hudson 2002)
- MSBAYES (Hickerson et al. 2007) to perform rejection step

Model choice with ABC.

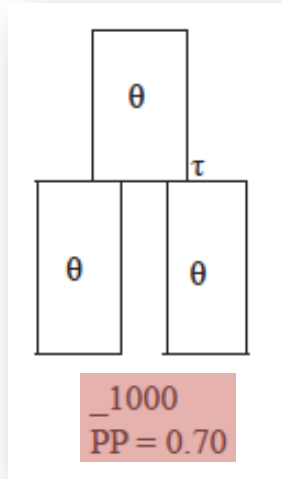
Plethodon idahoensis



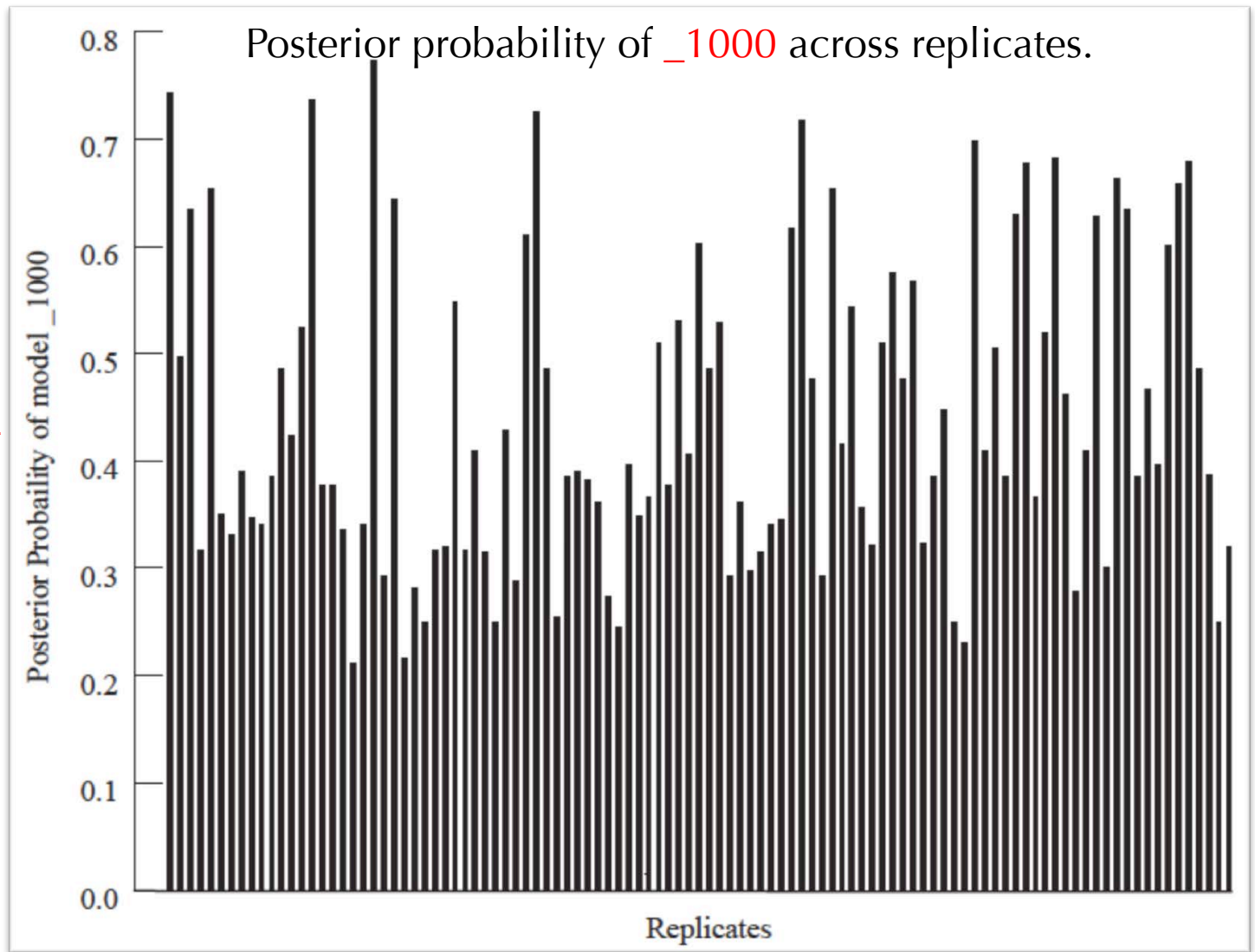
For each model: $\tau\theta m\gamma$

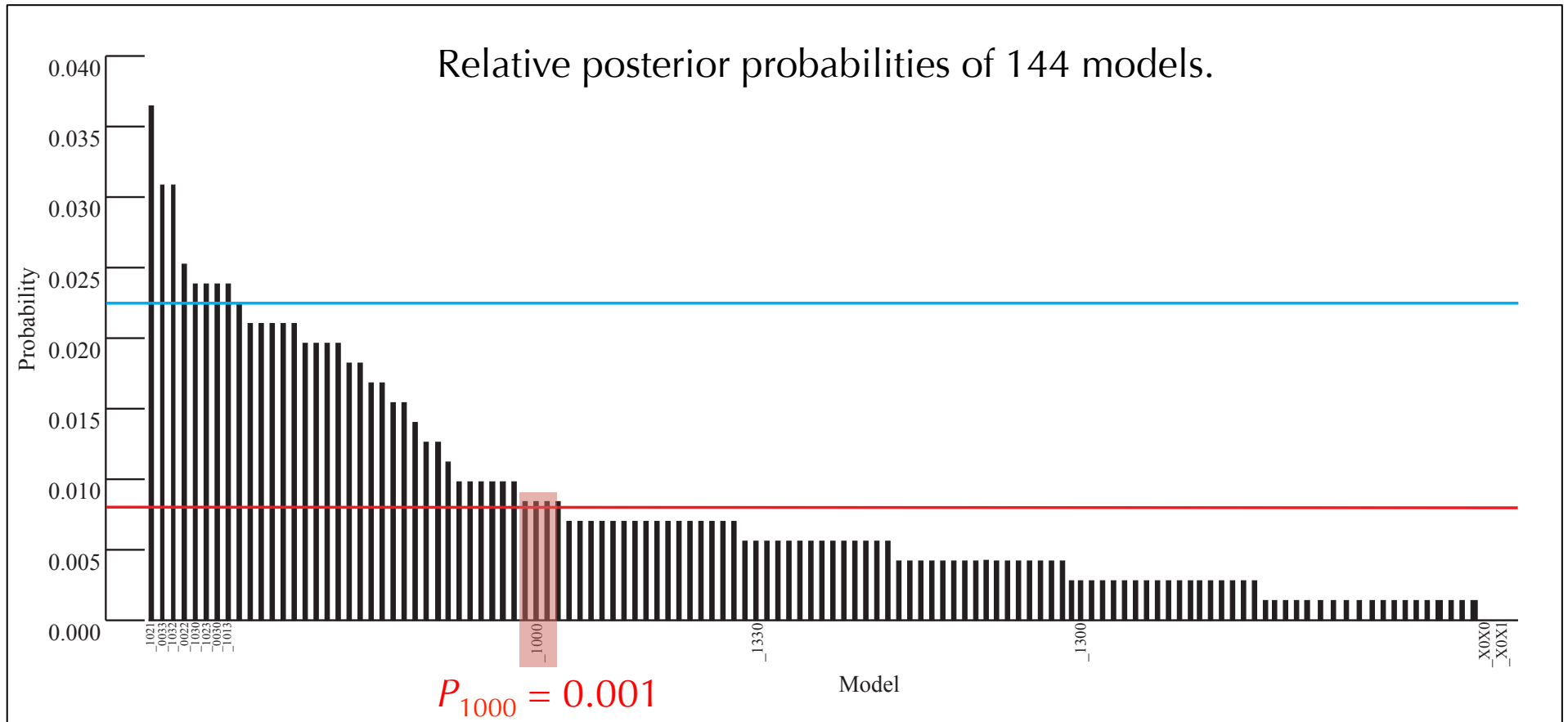
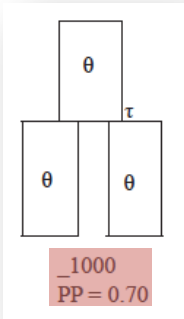
Divergence time (τ)	Theta (θ)	Migration (m)	Population expansion (γ)
0: island model 1: divergence at time (τ) X: panmixia	0: $\theta_A = \theta_1 = \theta_2$ 1: $\theta_A = \theta_1, \theta_2$ 2: $\theta_A = \theta_2, \theta_1$ 3: $\theta_A, \theta_1 = \theta_2$ 4: $\theta_A, \theta_1, \theta_2$	0: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

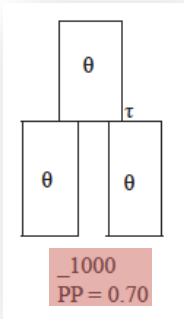
How does the composition of the model comparison set influence the *relative* posterior probability?



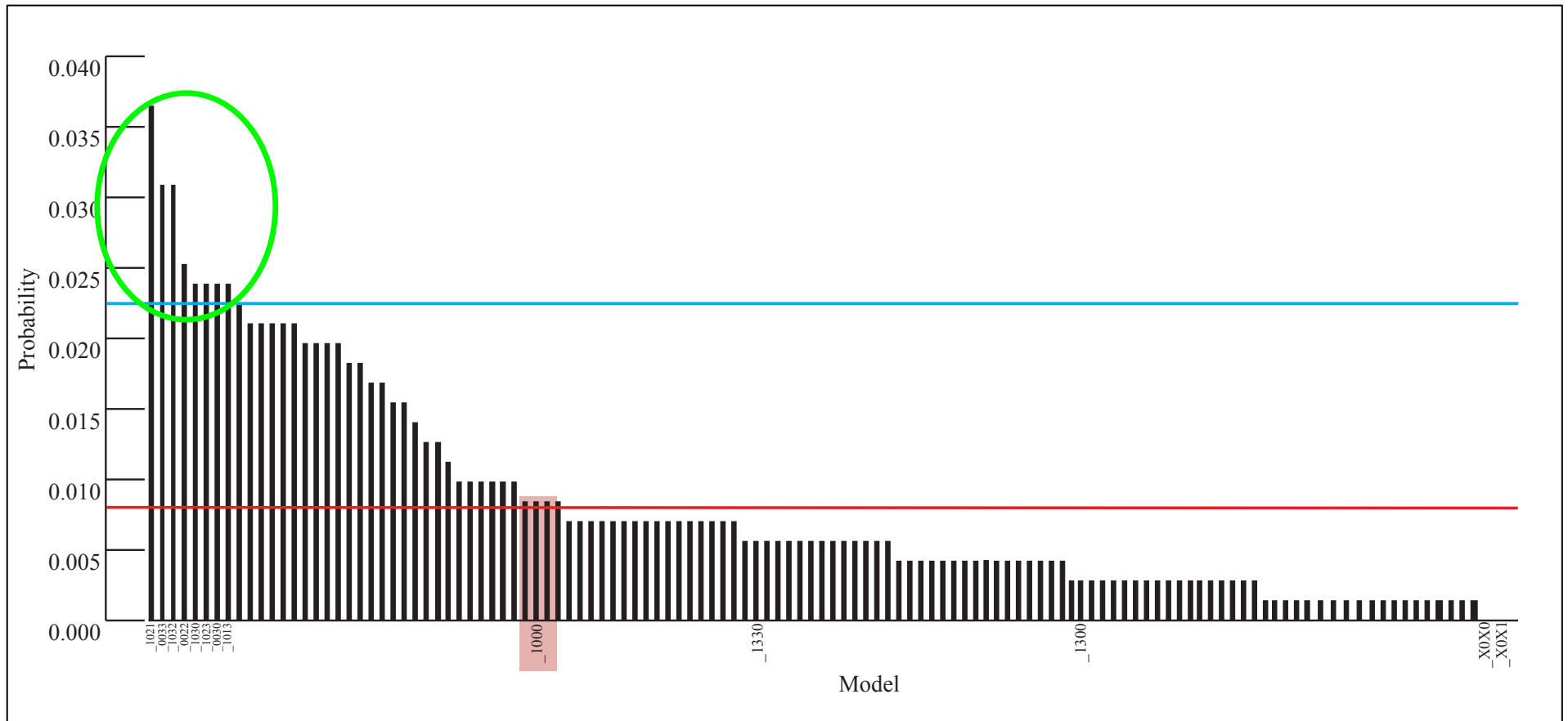
average $P_{1000} = 0.44$

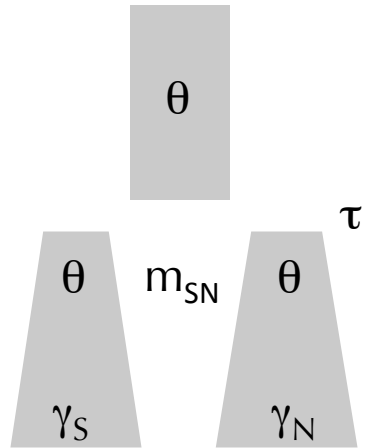






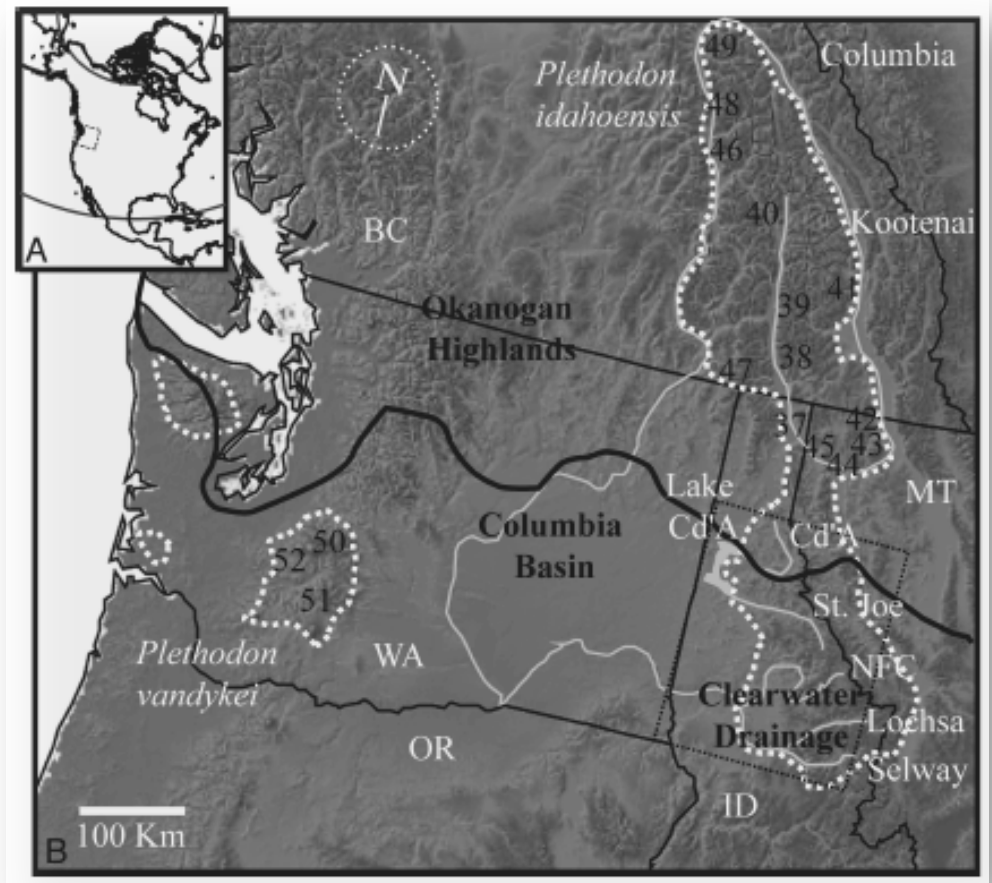
- measured absolute model fit using posterior predictive simulation and mean Euclidean distance
- selected models for comparison that were **measurably better** than the best score ever observed by chance
- compared 8 models in nested (island, isolation) comparison

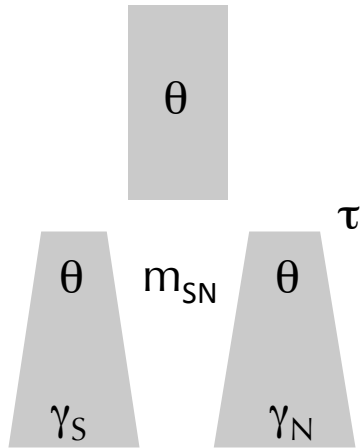




$$P_{1023} = 1.0$$

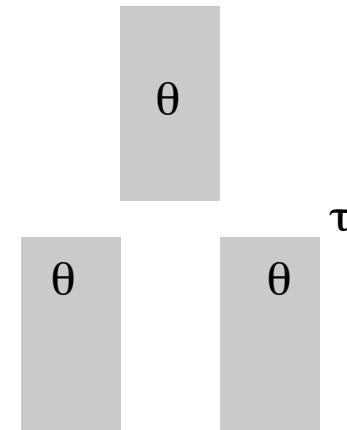
- gene flow from south to north
- population expansion in north and south, but $\gamma_N \gg \gamma_S$
- vastly different than m_{1000}





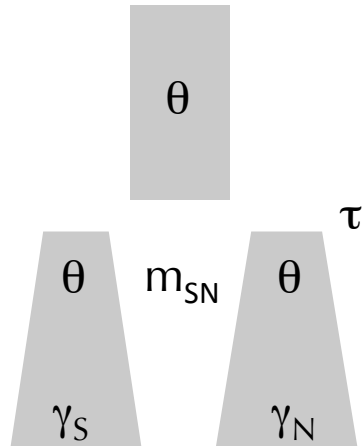
$$P_{1023} = 1.0$$

- gene flow from south to north
- population expansion in north and south, but $\gamma_N \gg \gamma_S$
- vastly different than m_{1000}

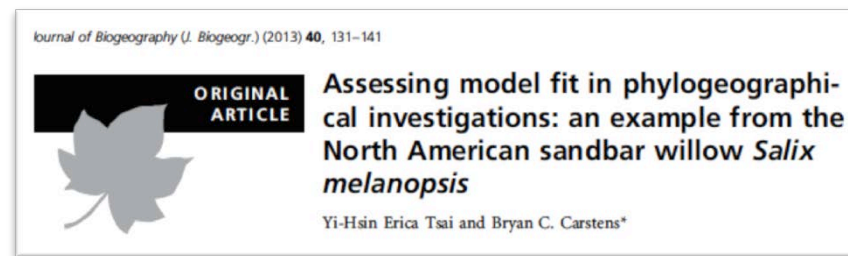


$$P_{1000} = 1.0$$

(naïve analysis)



- model choice is complicated by biases inherent to these vectors, to the point that some doubt its consistency (Robert *et al.* 2010)



- vector of summary statistics used to summarize the data in the prior distribution and empirical data – chosen following Tsai & Carstens (2013)
- complicates questions with inherent differences in dimensionality
- O'Meara (2010) showed that the $L(D|M)$ can be approximated by calculating the proportion of times that genealogies observed in the empirical data are found in a distribution simulated under some model



Phylogeographic Inference using Approximated Likelihoods

1

R - based package < data description > set of all possible models based on the number of free parameters



2

modifiers generated (population size, divergence times, migration regimes, expansion*)

3

ms (Hudson) < ms call generated by *phrapl* > simulates gene trees under each model

((3,(1,(8,10))),((5,(6,9)),(7,(2,4))));
 (((7,10),(2,5)),(4,(9,((1,6),(3,8)))));
 (((3,(4,6)),((10,(1,5)),(7,8))),((2,9)));
 ((2,3),((4,7),(5,(9,10))),((6,(1,8))));
 (1,((5,(6,(4,(2,8))),(3,7),(9,10))));
 ((3,6),((2,4),((9,(5,8)),(7,(1,10)))));

((3,(1,7)),((2,8),((4,9),(6,(5,10)))));
 ((7,(5,6)),((2,(10,(9,(4,8))),(1,3)));
 ((7,(9,10)),(1,((2,8),(6,(5,(3,4)))));
 ((6,(10,(9,(3,7))),(2,4),(1,(5,8))));
 ((1,4),(3,(7,((6,10),(8,(5,(2,9))))));
 ((9,10),((7,8),((4,5),(2,(6,(1,3))))));

((2,(4,(7,9))),((3,5)),(8,(1,(6,10))));
 ((3,(7,(6,9))),((10,((5,8),(2,(1,4)))));
 (8,(5,((6,(1,4)),(3,(2,(10,(7,9))))));
 ((7,(9,(6,10))),((8,(1,5)),(2,(3,4))));
 ((4,(5,(9,(2,6))),(1,(8,(10,(3,7)))));
 ((7,(6,(10,(3,8))),(2,(9,(4,(1,5)))));

4

for each model, perl script calculates the proportion of simulated trees that match the empirical gene trees; this value approximates the $-lnL$ (data | model_i)

model	locus 1	locus 2	locus i	-lnL
M1	0.00031	0.00013 ...	0.00028	-10.94753692
M2	0.00023	0.00008 ...	0.00068	-10.90267326
M3	0.00342	0.00542 ...	0.00254	-7.327140891

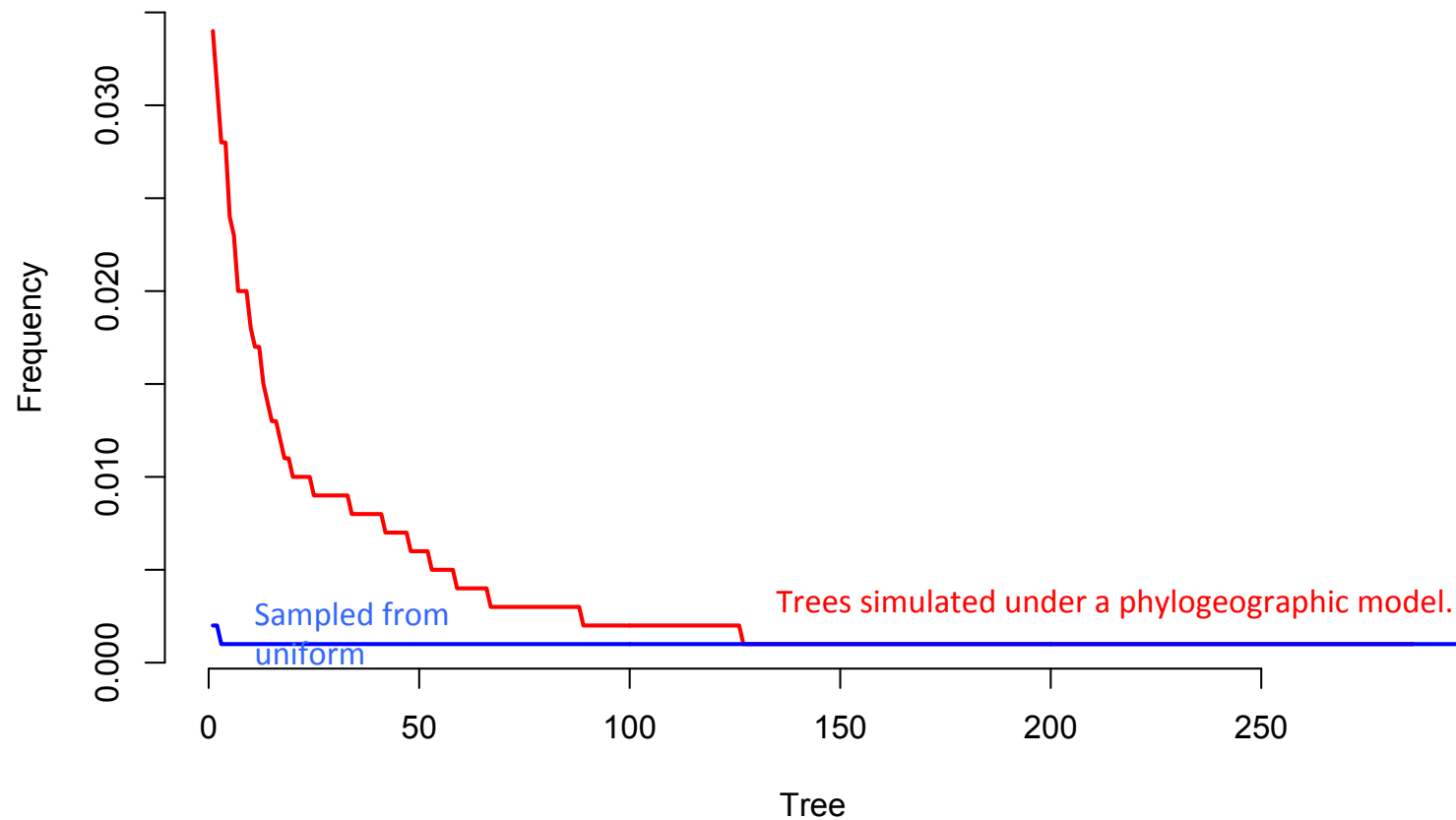
5

set of models evaluated using AIC or other approaches

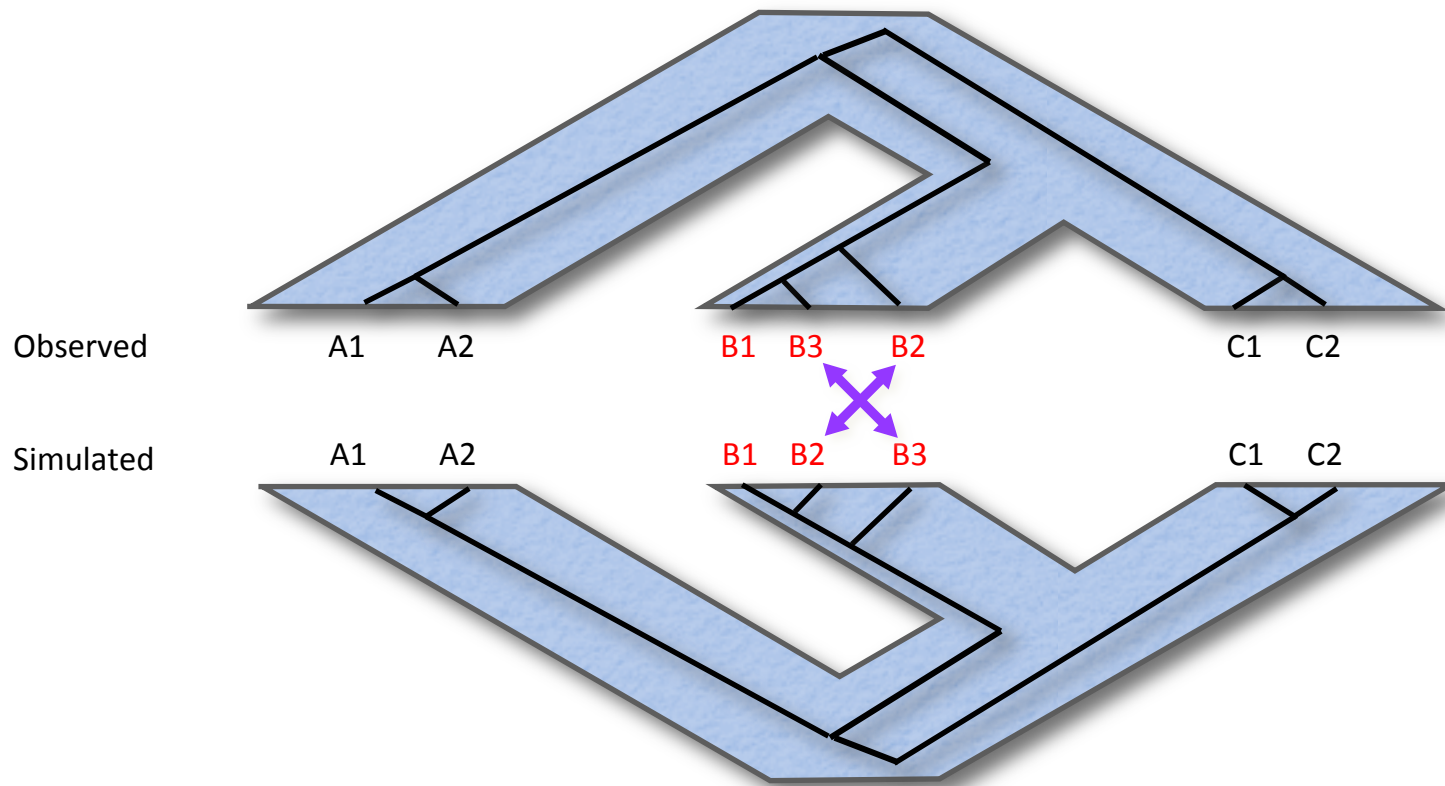
*to be added



Tree probabilities not uniform
(some trees *much* more likely than others)

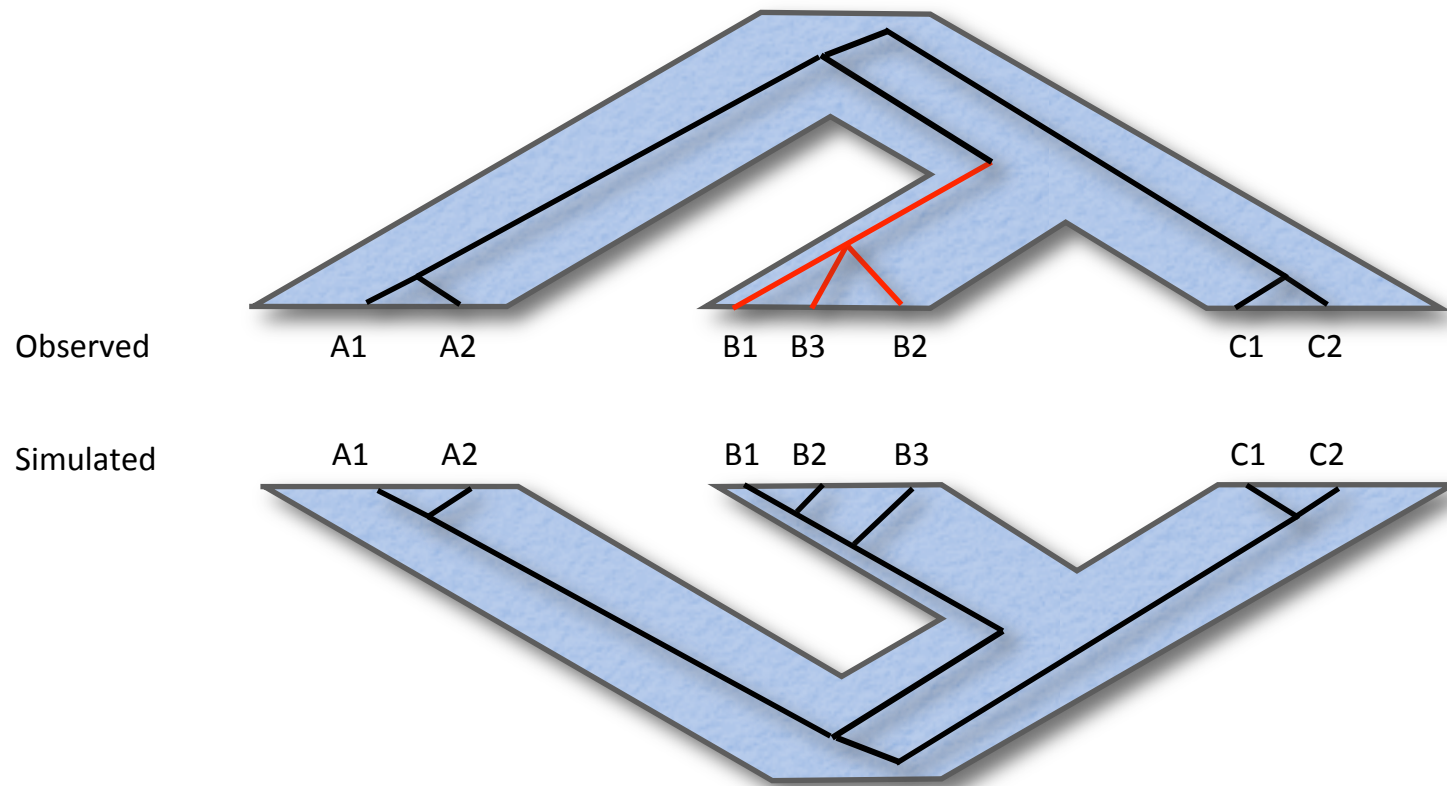


Simplification 1: Sample labels within populations arbitrary

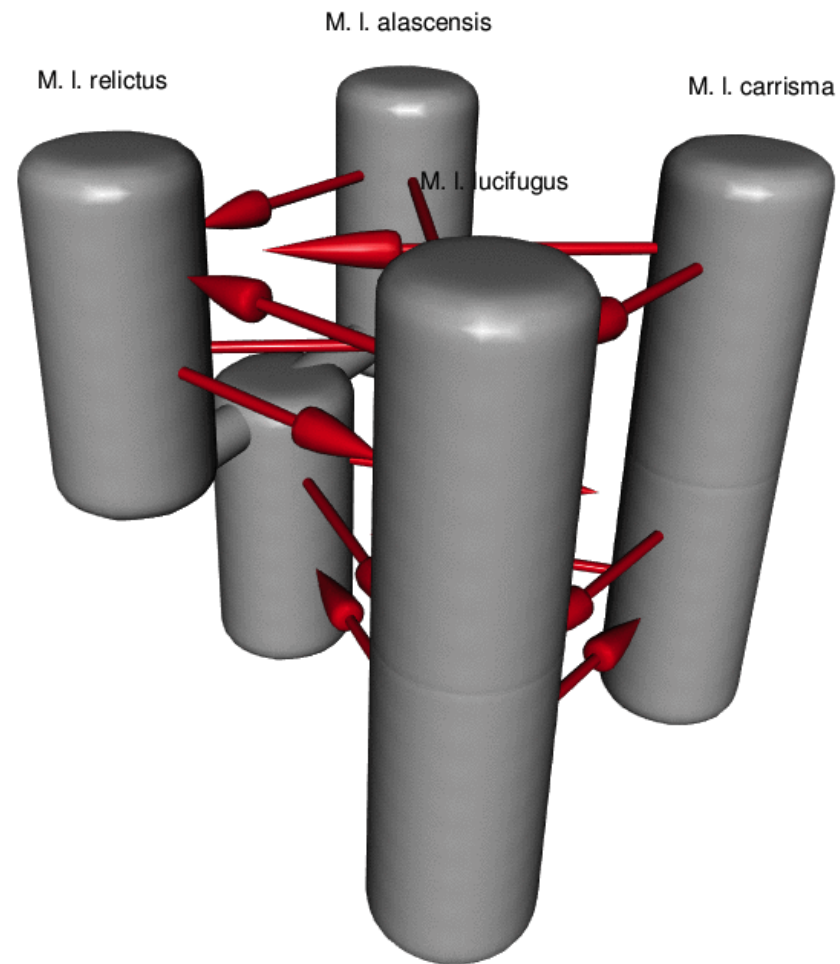


Match based on all possible labeling, then correct for this
i.e., three possible permutations, so if there is a match divide by 3 to get probability

Simplification 2: Polytomies are soft in gene trees (optional)



Match based on all possible resolutions, then correct



- Model selection allows us to identify models that are a reasonable fit to the data.
- Parameter estimates should only be used as the basis for inference when they are made from appropriate models.
- Model selection also can provide clues that appropriate models are not in the model comparison set.

- Model selection allows us to identify models that are a reasonable fit to the data.
- Parameter estimates should only be used as the basis for inference when they are made from appropriate models.
- Model selection also can provide clues that appropriate models are not in the model comparison set.

Model	k	AIC	Δ_i	w_i
AAC00	2	8769.0782	0	0.471458846
AACDD	3	8769.5574	0.4792	0.291964382
ABC00	3	8770.9324	1.8542	0.073820156
ABC0D	4	8771.3364	2.2582	0.049285593
AACDE	4	8771.3892	2.311	0.046750821
ABCDD	4	8771.5317	2.4535	0.04054173
ABCD0	4	8772.3866	3.3084	0.017243441
FULL	5	8773.332	4.2538	0.006699493
ABB00	1	8775.8409	6.7627	0.000545055
ABA00	2	8776.0193	6.9411	0.000455997
ABBDD	2	8776.0647	6.9865	0.000435758
AAA00	3	8776.4832	7.405	0.000286743
ABADD	3	8776.75	7.6718	0.000219595
AAADD	2	8777.3167	8.2385	0.000124597
ABBDE	3	8777.589	8.5108	9.49E-05
ABADE	3	8778.0647	8.9865	5.90E-05
AAADE	4	8779.5084	10.4302	1.39E-05

- **ABC** (Fagundes *et al.* 2007; Peter *et al.* 2009)
- **BP&P** (Yang & Rannala 2010)
- ***dadi*** (Gutenkunst *et al.* 2009)
- **IMa** (Hey & Nielsen 2007; Carstens *et al.* 2009)
- **Migrate-n** (Beerli & Palczewski 2010)
- **MSBAYES** (Hickerson *et al.* 2007)
- **STRUCTURE** – Evanno's k (Evanno *et al.* 2005)
- **spedeSTEM** (Ence & Carstens 2010)
- **BEAST** (spatial diffusion models) Lemey *et al.* 2010
- **Phrapl** (O'Meara, Carstens *et al.* in prep)
- **heuristic search of very complex model space** (Carstens lab, *under dev.*)

heuristic search of very complex model space

- MEDs used as model optimality criteria in a heuristic search of complex model space
- Matt Demarest is writing the code to heuristically search the models that Pelletier & Carstens (*in rev.*) calculated exhaustively
- once performance using small model space is satisfactory, expand model space to 3 and 4 diverging lineages

Populations	Number of Models
1	3
2	240
3	70,200
4	24,701,040
5	9,396,476,180

heuristic search of very complex model space

- MEDs will be used as model optimality criteria in a heuristic search of complex model space
- Matt Demarest is writing the code to heuristically search the models that Pelletier & Carstens (*in rev.*) calculated exhaustively
- once performance using small model space is satisfactory, expand model space to 3 and 4 diverging lineages

new methods for comparative phylogeography

- novel methods seek to cluster codistributed species into some number of groups, each defined as the product of a particular evolution history
- goal is to identify evolutionary communities (groups of organisms that interact throughout evolutionary time)

Acknowledgements.

Margaret Koopman
Yi-Hsin Erica Tsai
Amanda Zellmer

Sarah Hird
Noah Reid
John McVay

Tara Pelletier
Jordan Satler
Ariadna Morales-Garcia

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

- NSF (DEB-1257784; DEB-0918212;
DEB-0956069)

