Discrete & Categorical data & models

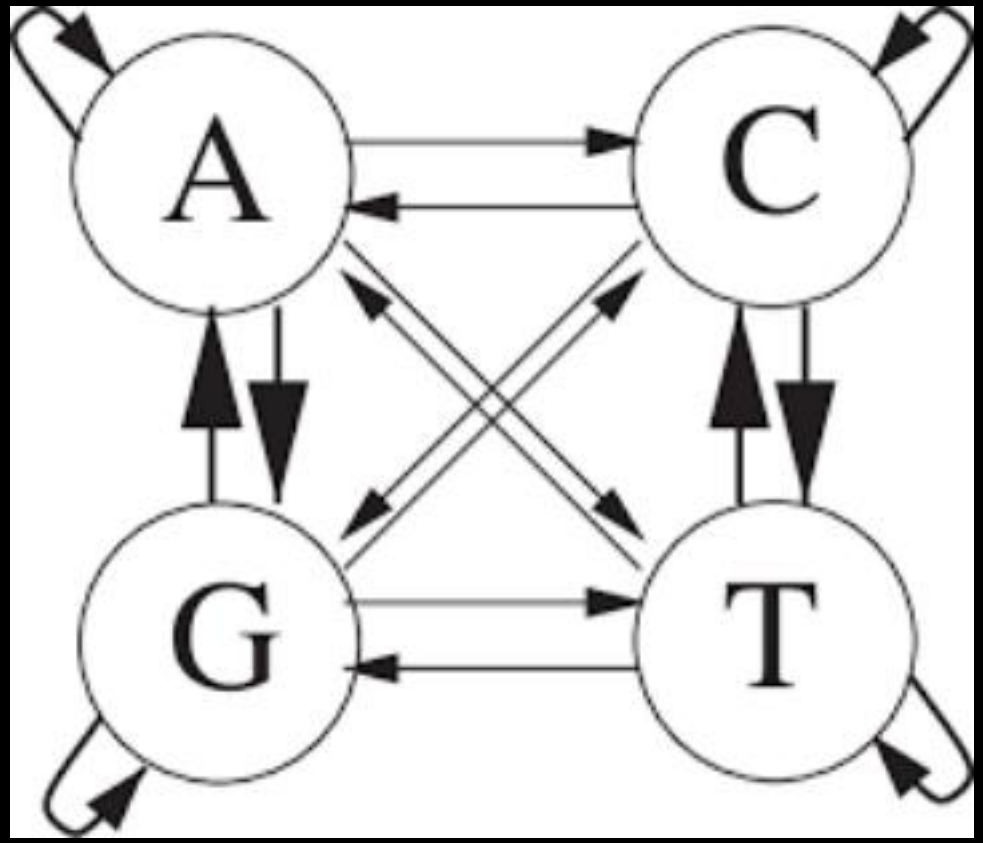
....and how they connect to continuous data and models

Evolutionary Inferences from Phylogenies: A Review of Methods

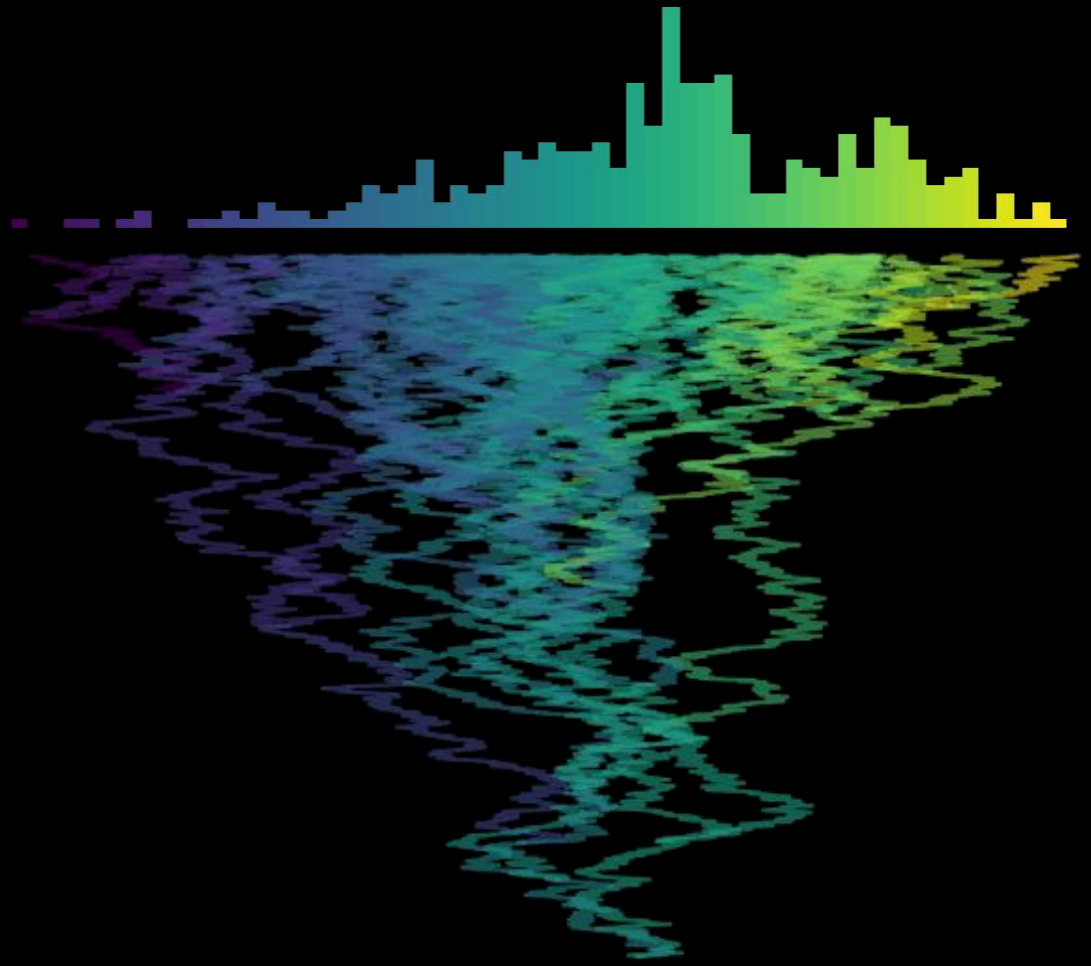
Brian C. O'Meara

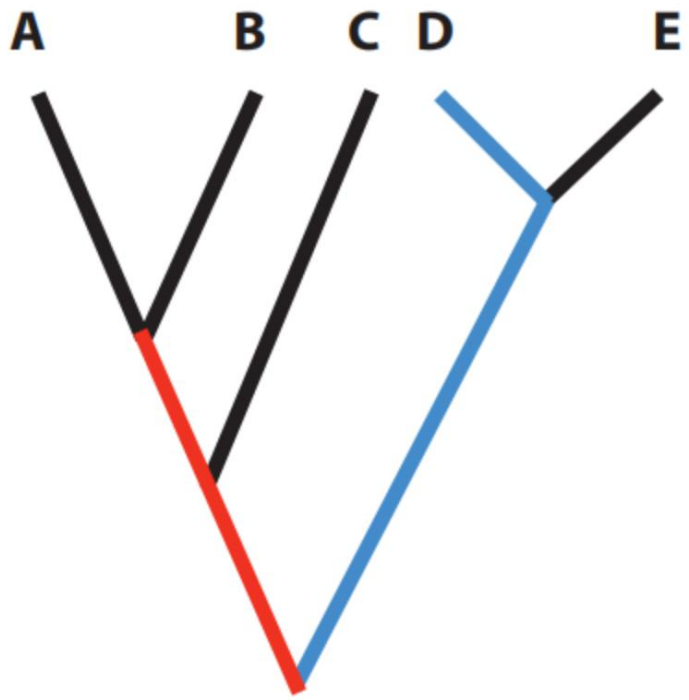
Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee 37996; email: bomeara@utk.edu, **http://www.brianomeara.info**

Markov Models



Gaussian Models





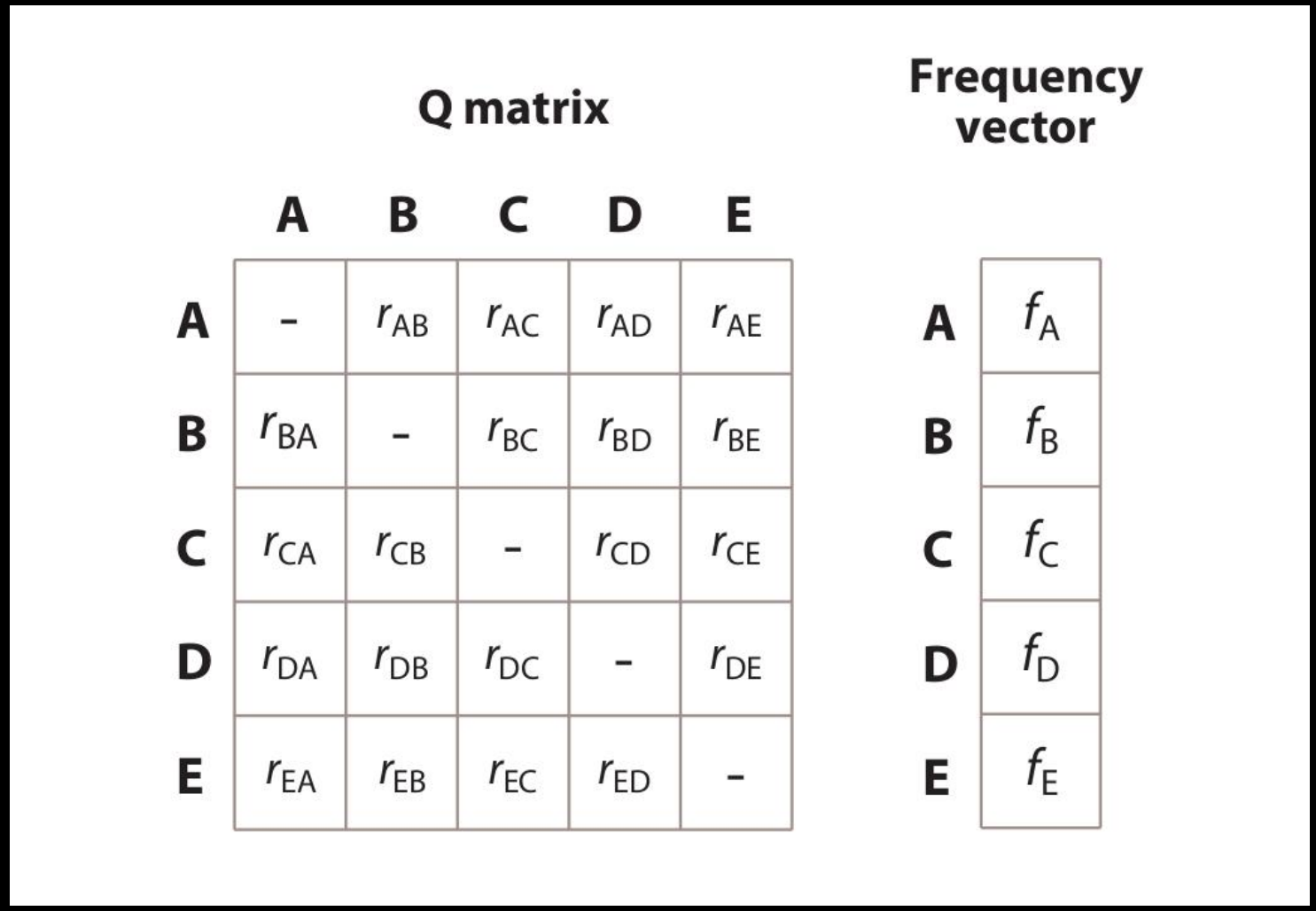
Variance-covariance matrix

	A	B	C	D	E
A	var(A)	cov(A,B)	cov(A,C)	cov(A,D)	cov(A,E)
B	cov(B,A)	var(B)	cov(B,C)	cov(B,D)	cov(B,E)
C	cov(C,A)	cov(C,B)	var(C)	cov(C,D)	cov(C,E)
D	cov(D,A)	cov(D,B)	cov(D,C)	var(D)	cov(D,E)
E	cov(E,A)	cov(E,B)	cov(E,C)	cov(E,D)	var(E)

Mean vector

A	μ_A
B	μ_B
C	μ_C
D	μ_D
E	μ_E

Our general CTMC model:



This model can be use to estimate rates of transition, ancestral states, and perform a variety of hypothesis testing questions (and is the model we use to estimate phylogenies)

Frequency vector: For sequence evolution, frequency vectors will show up as base frequencies.

For time-irreversible models, they are primarily used as a “prior” on the root state.

	A	G	C	T
A	-	r_{AG}	r_{AC}	r_{AT}
G	r_{AG}	-	r_{GC}	r_{GT}
C	r_{AC}	r_{GC}	-	r_{CT}
T	r_{AT}	r_{GT}	r_{CT}	-

	00	01	11	10
00	-	r_A	0	r_B
01	r_C	-	r_D	0
11	0	r_E	-	r_F
10	r_G	0	r_H	-

	A+	T+	A–	T–
A+	-	r_{AT}	δ	0
T+	r_{TA}	-	0	δ
A–	$k\delta$	0	-	0
T–	0	$k\delta$	0	-

	0	1	2	3
0	-	r_{01}	0	0
1	r_{10}	-	r_{12}	0
2	0	r_{21}	-	r_{23}
3	0	0	r_{32}	-

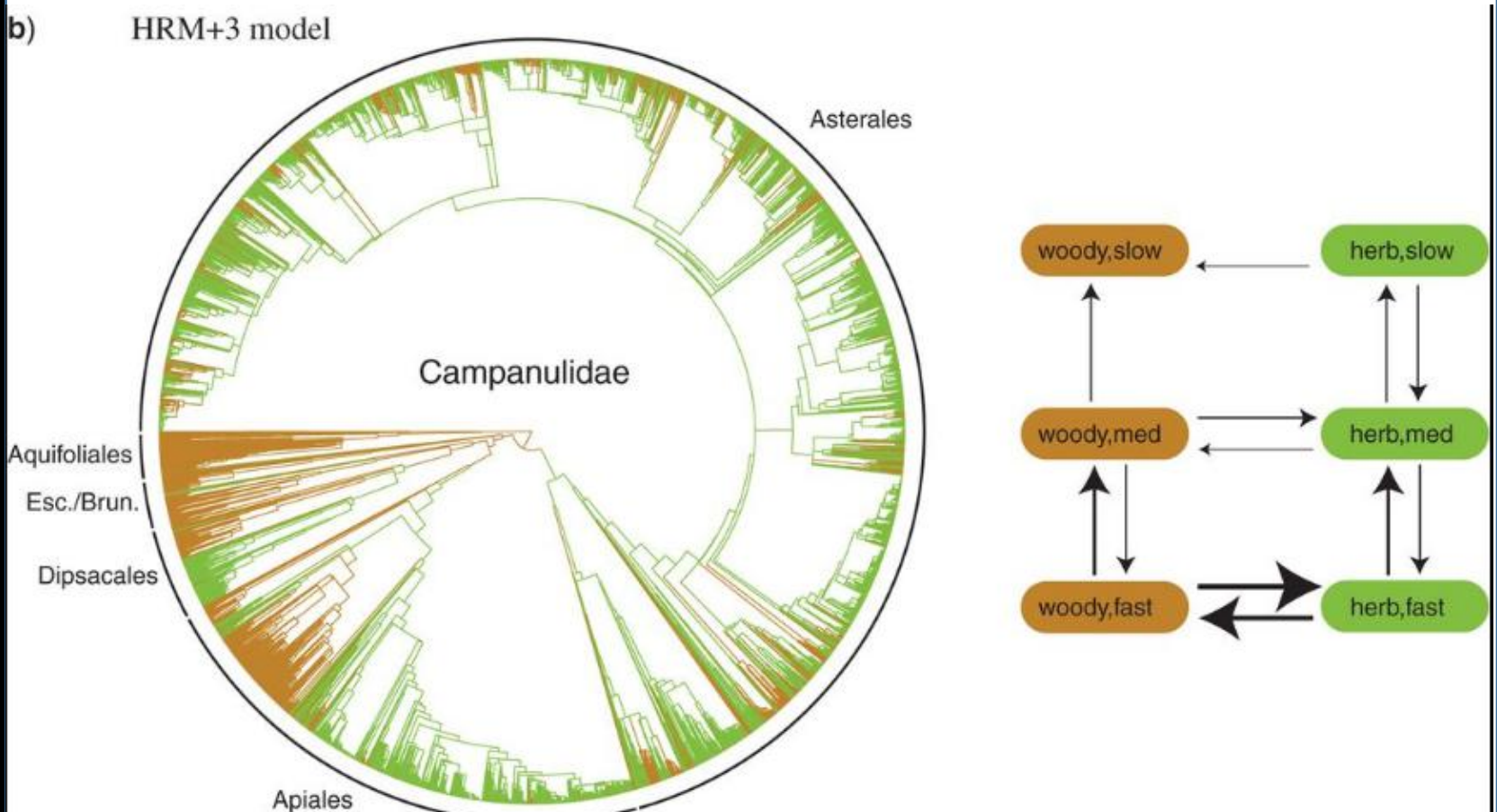
What do these models mean?

Hidden state models

Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms

Jeremy M. Beaulieu ✉, Brian C. O'Meara, Michael J. Donoghue
Author Notes

Systematic Biology, Volume 62, Issue 5, September 2013, Pages 725–737, https://doi.org/10.1093/sysbio/syt034
Published: 14 May 2013 **Article history** ▾



But what are the "Hidden states"?

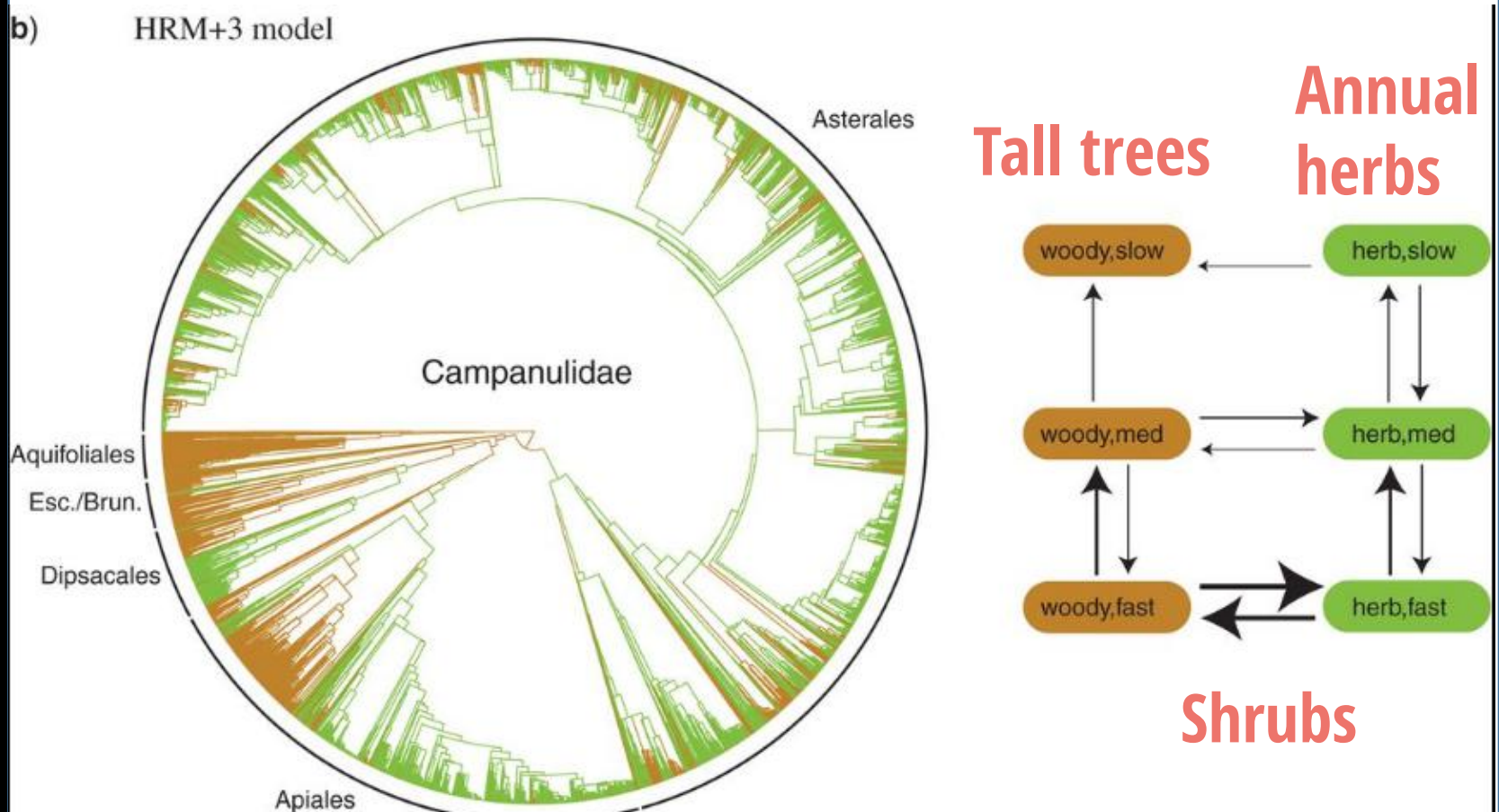
States with unobserved variation - Not all woody plants are the same

Different transition rates

Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms FREE

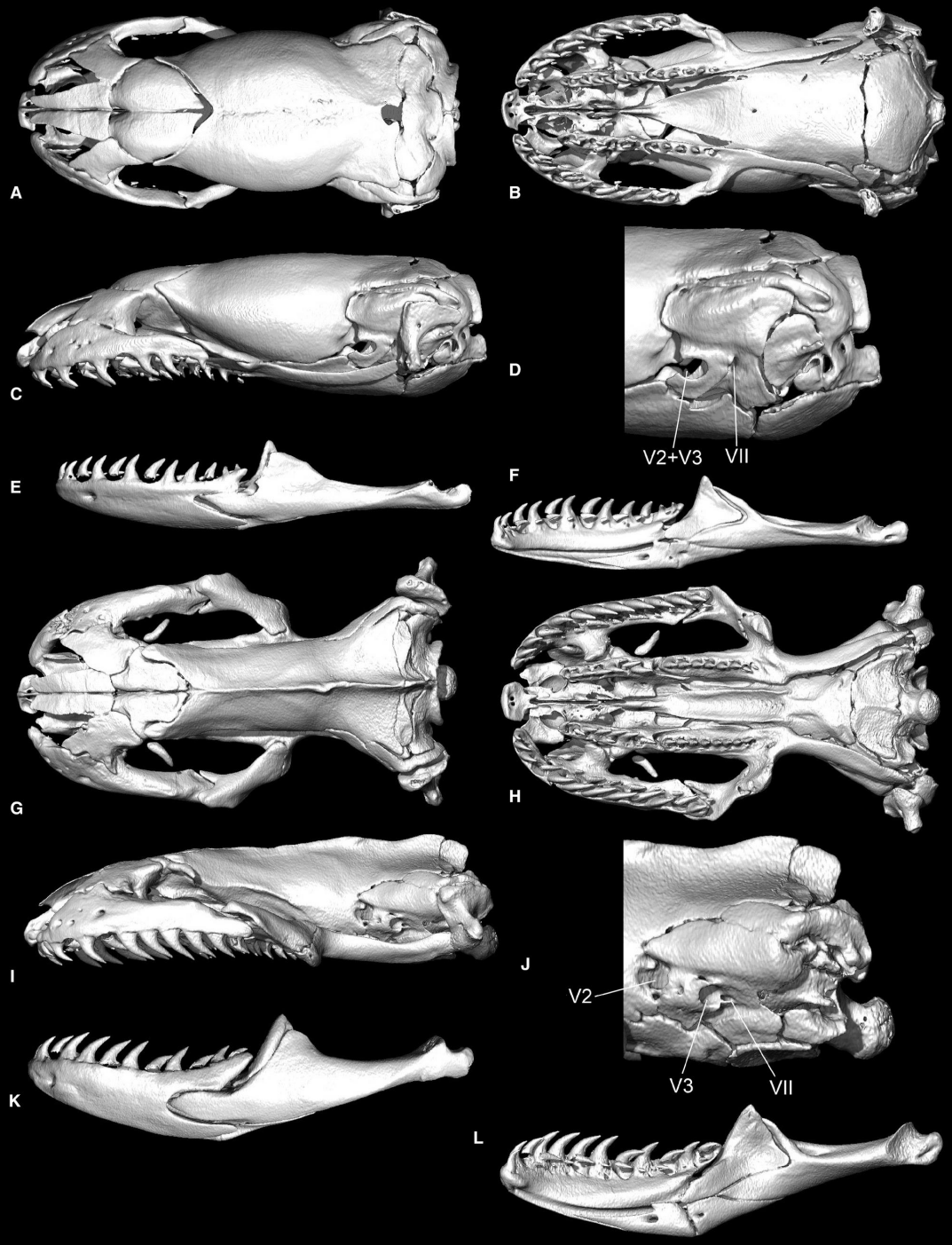
Jeremy M. Beaulieu ✉, Brian C. O'Meara, Michael J. Donoghue
Author Notes

Systematic Biology, Volume 62, Issue 5, September 2013, Pages 725–737, https://doi.org/10.1093/sysbio/syt034
Published: 14 May 2013 **Article history** ▾



Systematics & Biodiversity Science

Character construction:
The “hidden work” of phylogenetics & comparative methods


A
B
C
D
V2+V3
VII
E
F
G
H
I
J
V2
V3
VII
K
L


JCM Natural History Photography © John C. Murphy

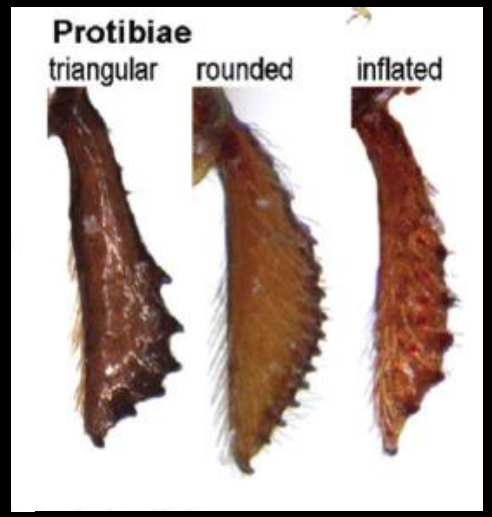

© KENNETH P. WRAY

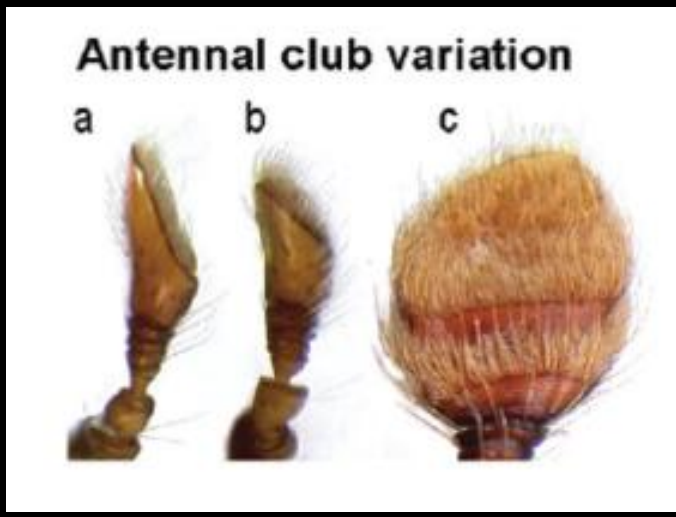

Copyright 2010 Chad Lane


© Gary Nafis

Genomes	Phenomes
Well-defined nucleotide state space (ACGT) amenable to automation	Many possible ways to represent through measurements

Biodiversity science is uniquely reliant on expert knowledge to quantify its primary data sources: species & trait measurements

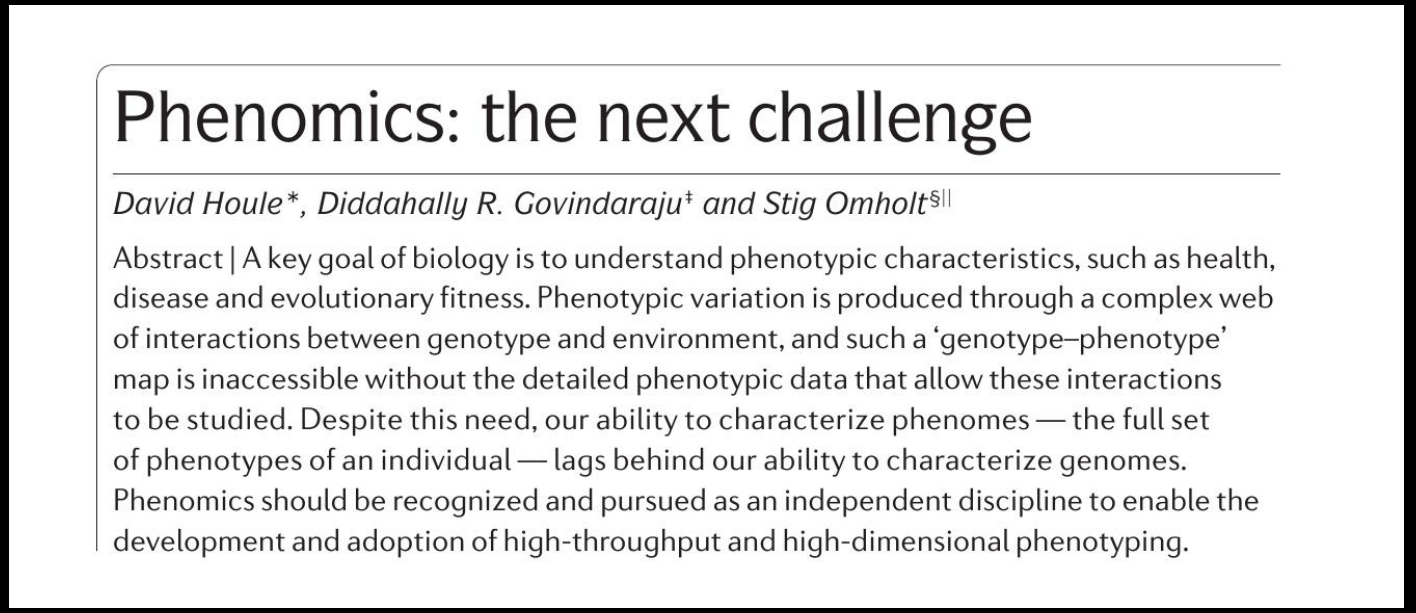




	C1	C2	C3	C4	C5	C6	C7	C8
Sp. 1								
Sp. 2								
Sp. 3								
Sp. 4								

Rapid progress on any biological problem rests on the hope that there is at least one viewpoint to each problem that makes causation relatively simple.

- **Houle et al 2010, Phenomics: The Next Challenge**



Phenomics: the next challenge

David Houle, Diddahally R. Govindaraju‡ and Stig Omholt§||*

Abstract | A key goal of biology is to understand phenotypic characteristics, such as health, disease and evolutionary fitness. Phenotypic variation is produced through a complex web of interactions between genotype and environment, and such a 'genotype–phenotype' map is inaccessible without the detailed phenotypic data that allow these interactions to be studied. Despite this need, our ability to characterize phenomes — the full set of phenotypes of an individual — lags behind our ability to characterize genomes. Phenomics should be recognized and pursued as an independent discipline to enable the development and adoption of high-throughput and high-dimensional phenotyping.

For every complex problem there is a solution which is clear, simple, and wrong

- Often attributed to H.L. Mencken, reporter and supporter of teaching evolution in the “Scopes Monkey Trial”

What happens when we get the traits “wrong”?

(And what can we do about it)

Syst. Biol. 0(0):1–19, 2019

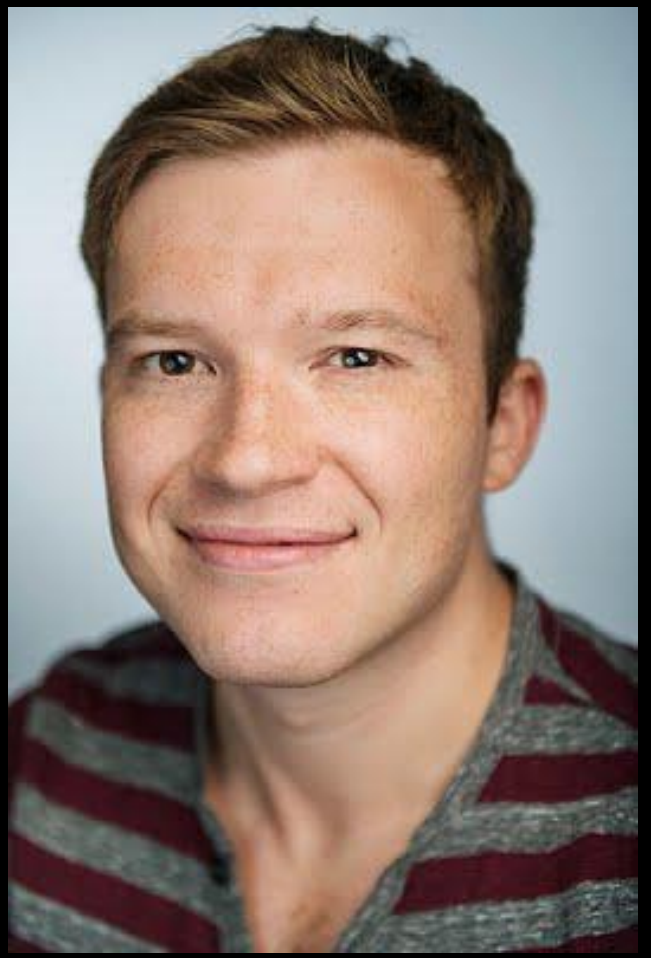
DOI:10.1093/sysbio/syz005

Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov Models Suggests a New Framework for Modeling Discrete Phenotypic Traits

SERGEI TARASOV[1,2]

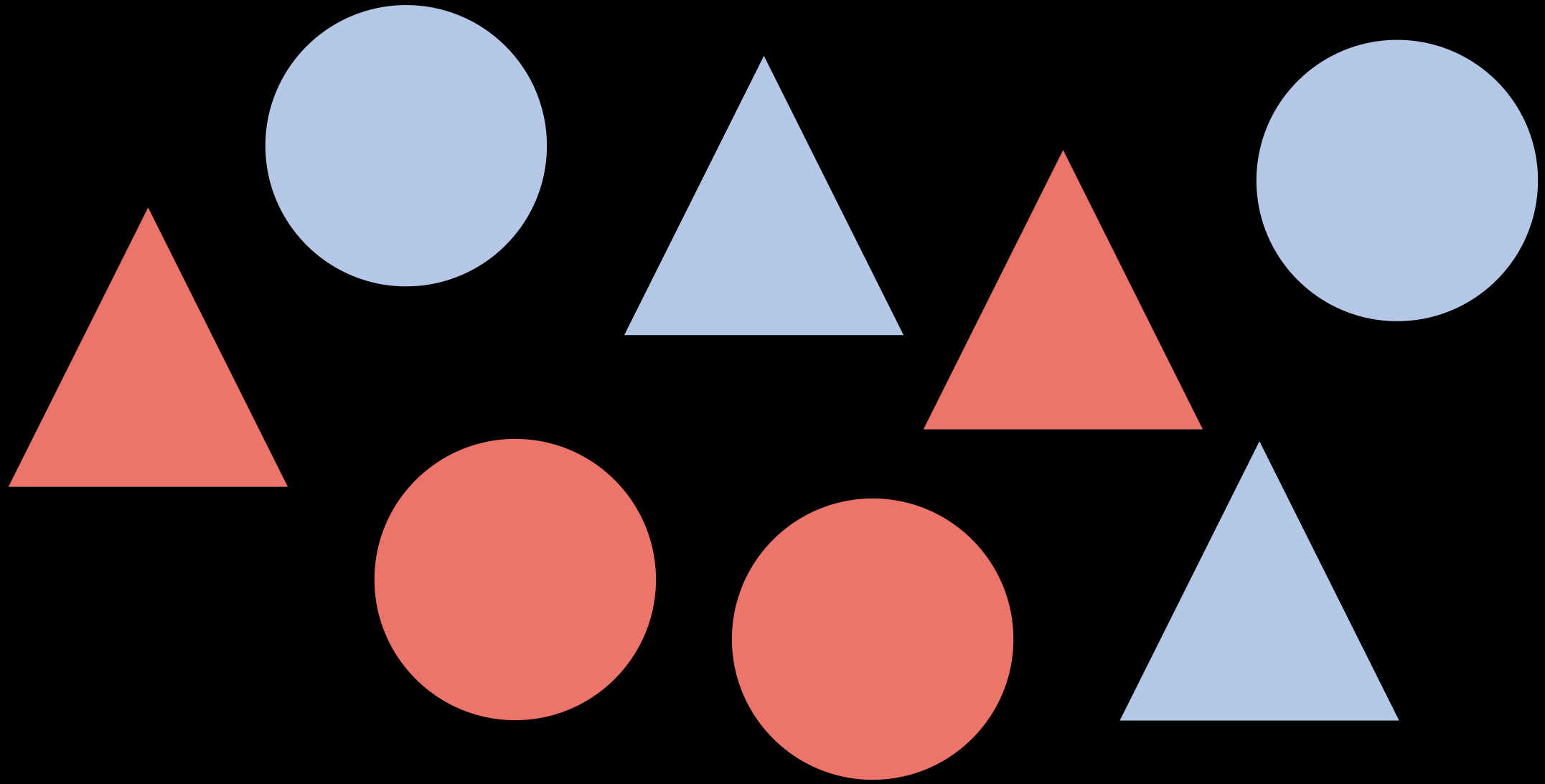
[1]*National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN 37996, USA; and* [2]*Department of Biological Sciences, Virginia Tech, 4076 Derring Hall, 926 West Campus Drive, Blacksburg, VA 24061, USA*
E-mail: sergxf@yandex.ru.

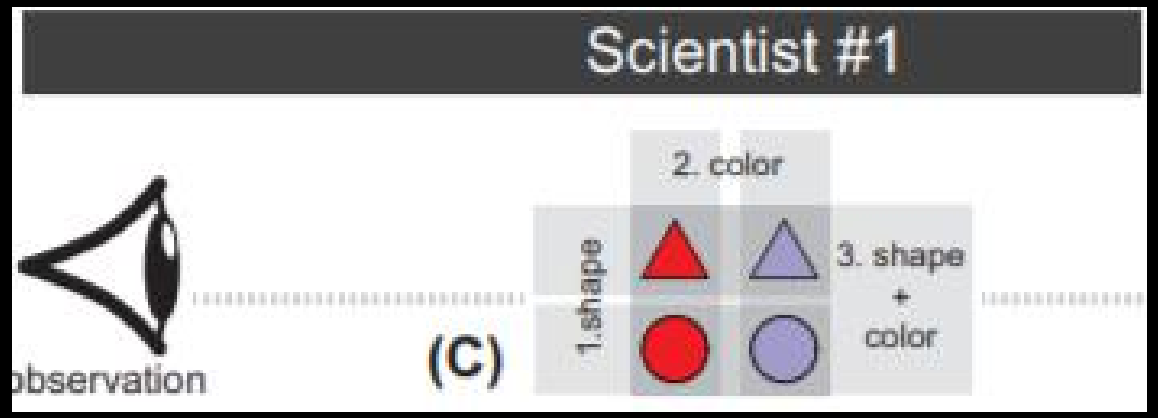
Received 19 September 2017; reviews returned 6 January 2019; accepted 15 January 2019
Associate Editor: Emma Goldberg

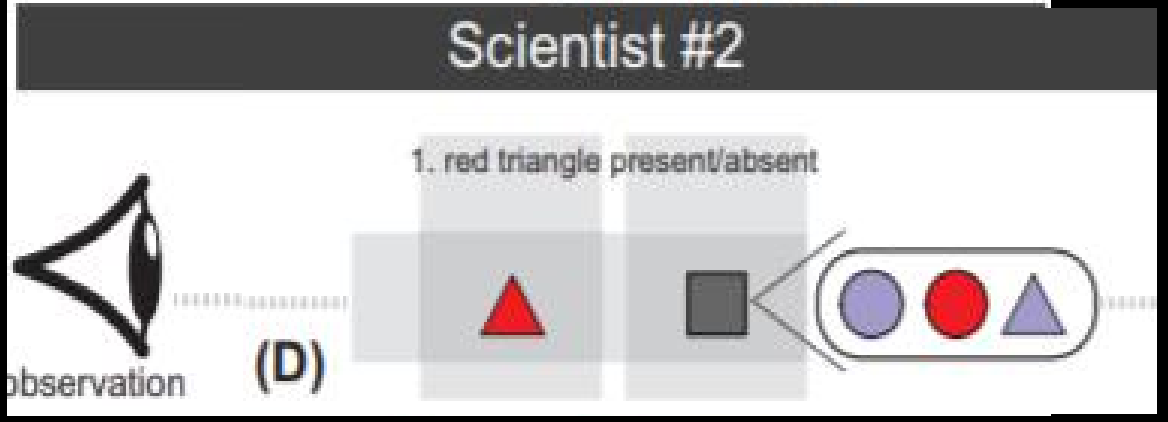


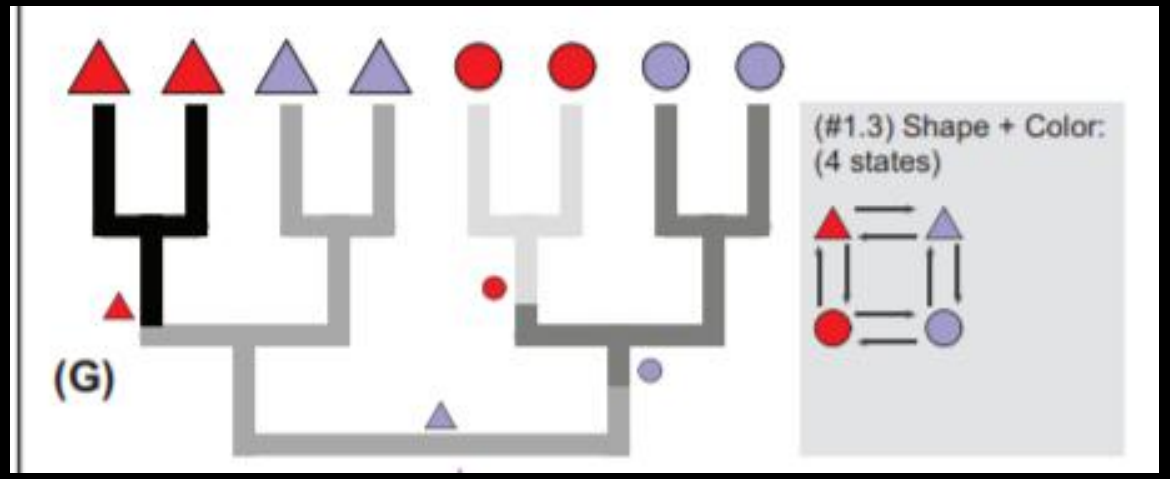
Sergei Tarasov
Finnish Museum of Natural History

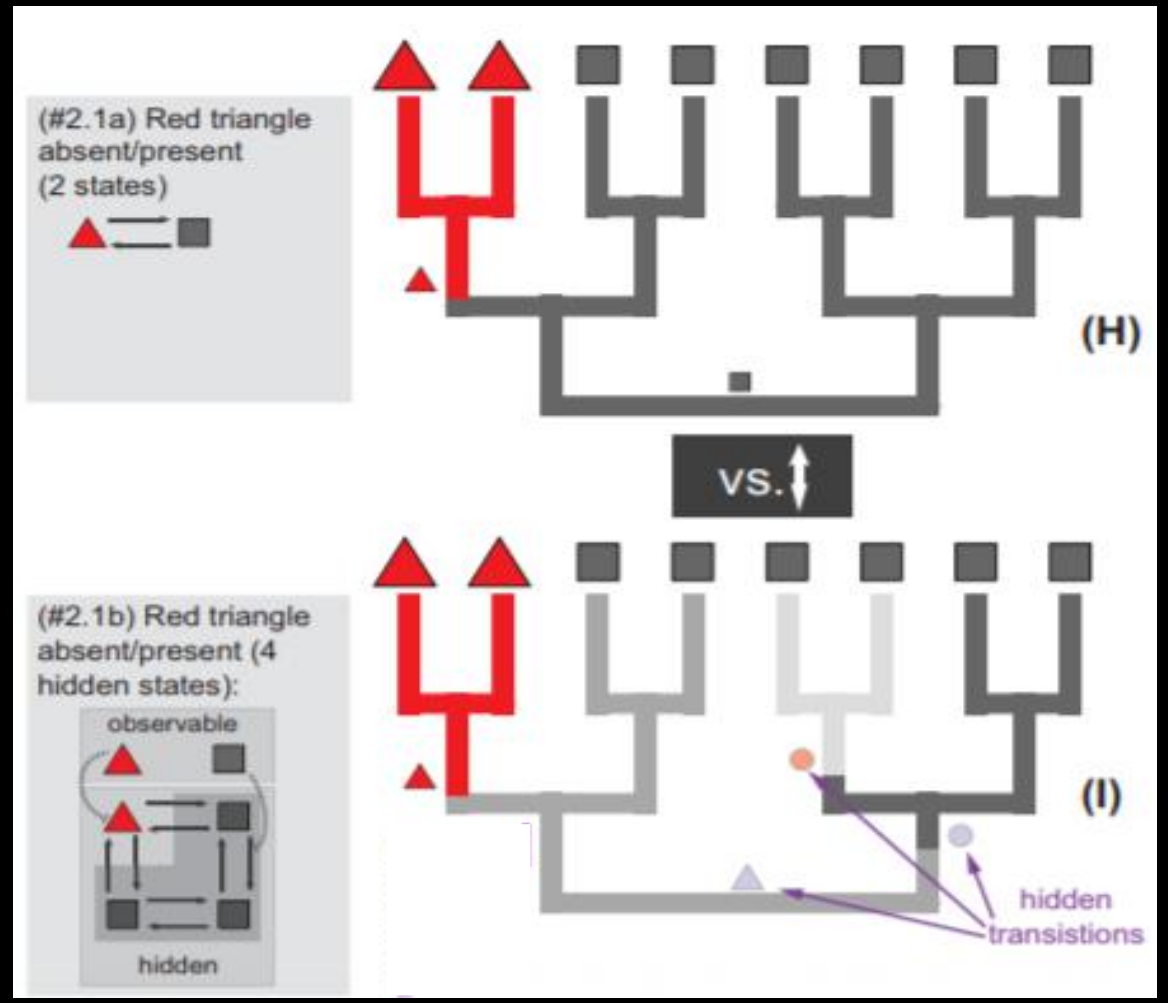
Two-Scientist Paradox (Tarasov, 2018)



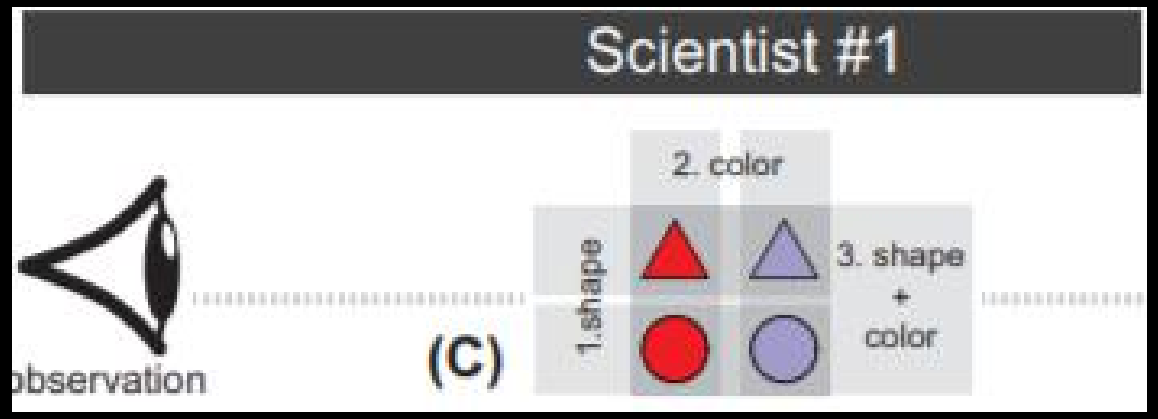


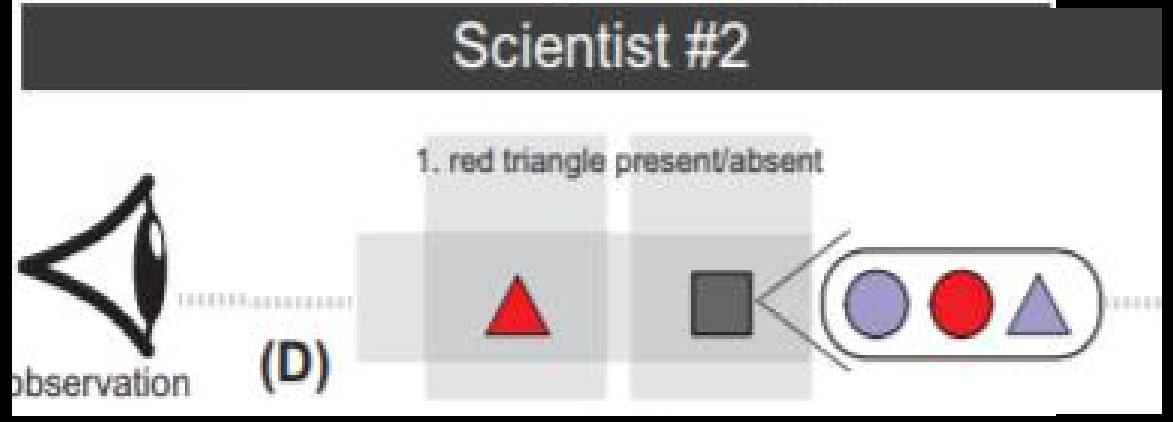


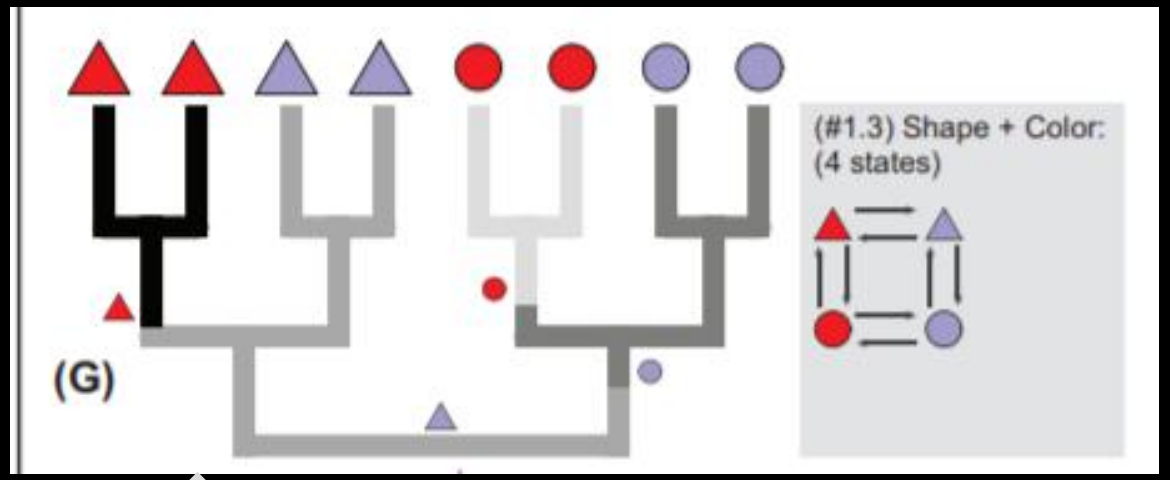


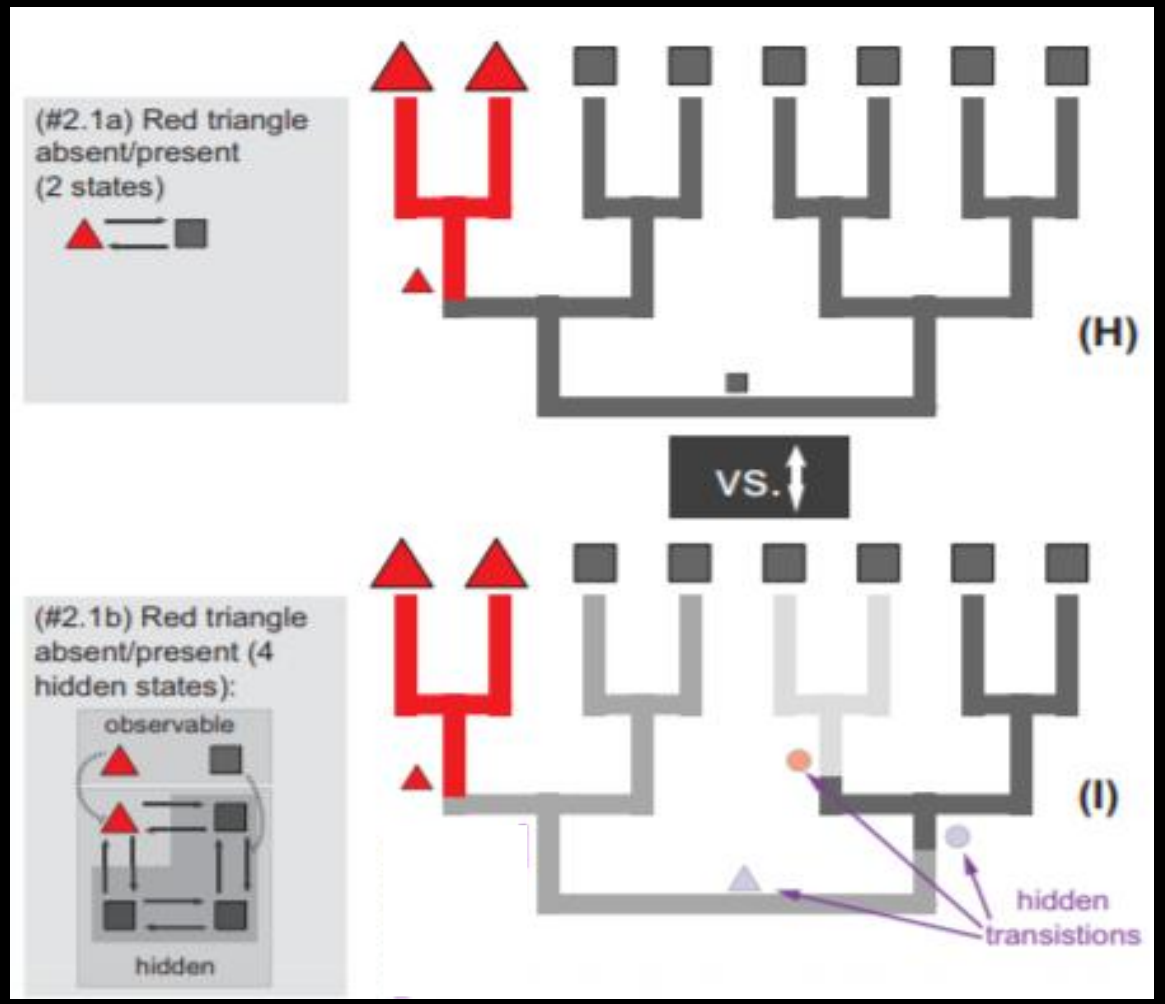


Scientist #2 supports HIdden State Model (Rate heterogeneity)





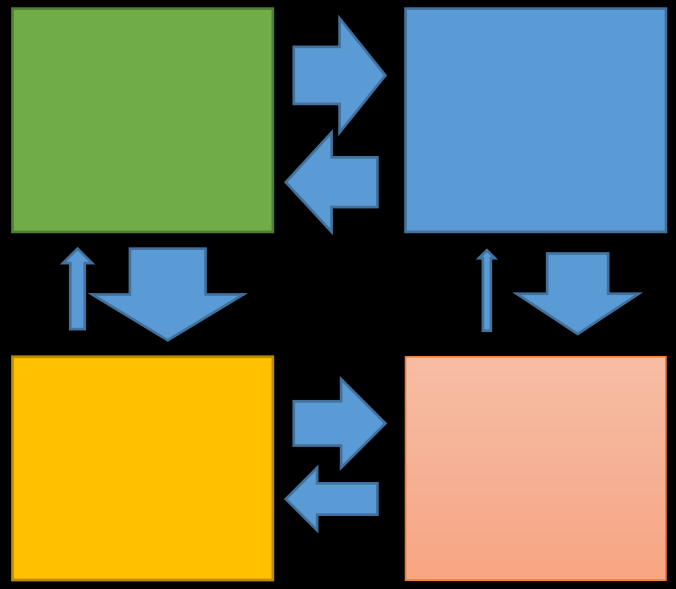




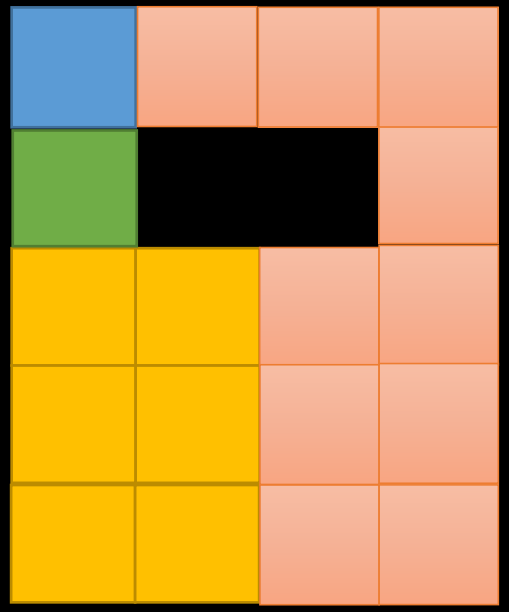
Truth is no rate heterogeneity, the true state space was misrepresented

Many comparative questions can be recast as questions about the *state space of evolution*

Framing questions as rates vs. states



Unequal rates
equal state spaces

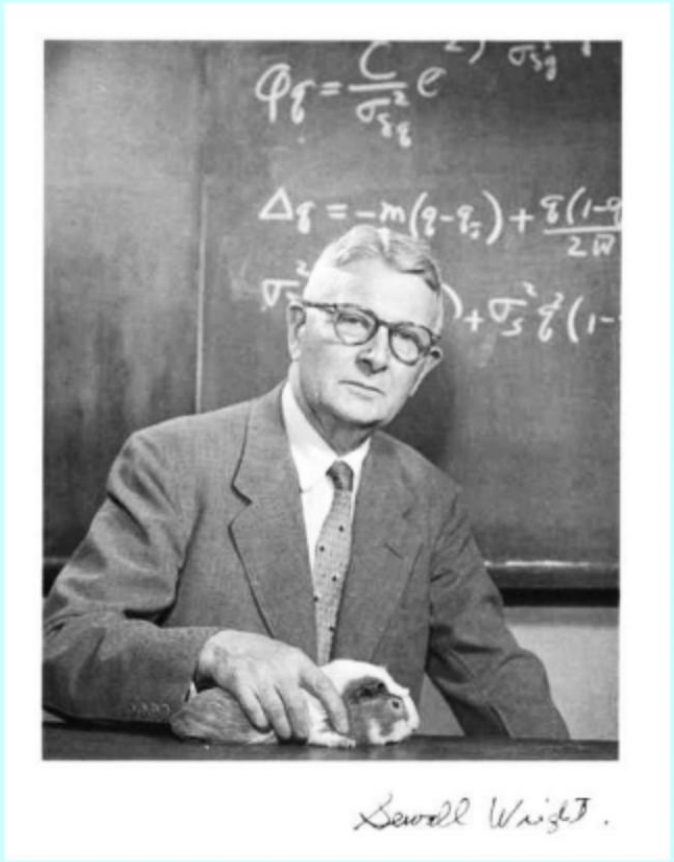


Equal rates
Unequal state spaces

“Macroevolutionary Architecture of Traits”

The threshold model

A relevant model was invented in 1934 by



Sewall Wright (1889-1988)
shown here in the 1950's

Dear Joe,

I thought you might like the ability to teach like Sewall Wright, complete with guinea pig eraser, though this one works on whiteboards and is IACUC-approved. Though parts separately are washable, there's an internal support plate of cardboard that probably shouldn't get wet.

Best,
Brian [O'Meara]

Dear Joe,

I thought you might like the ability to teach like Sewall Wright, complete with guinea pig eraser, though this one works on whiteboards and is IACUC-approved. Though parts separately are washable, there's an internal support plate of cardboard that probably shouldn't get wet.

Best,
Brian [O'Meara]

► Philos Trans R Soc Lond B Biol Sci. 2005 Jul 7;360(1459):1427–1434. doi: 10.1098/rstb.2005.1669

Using the quantitative genetic threshold model for inferences between and within species

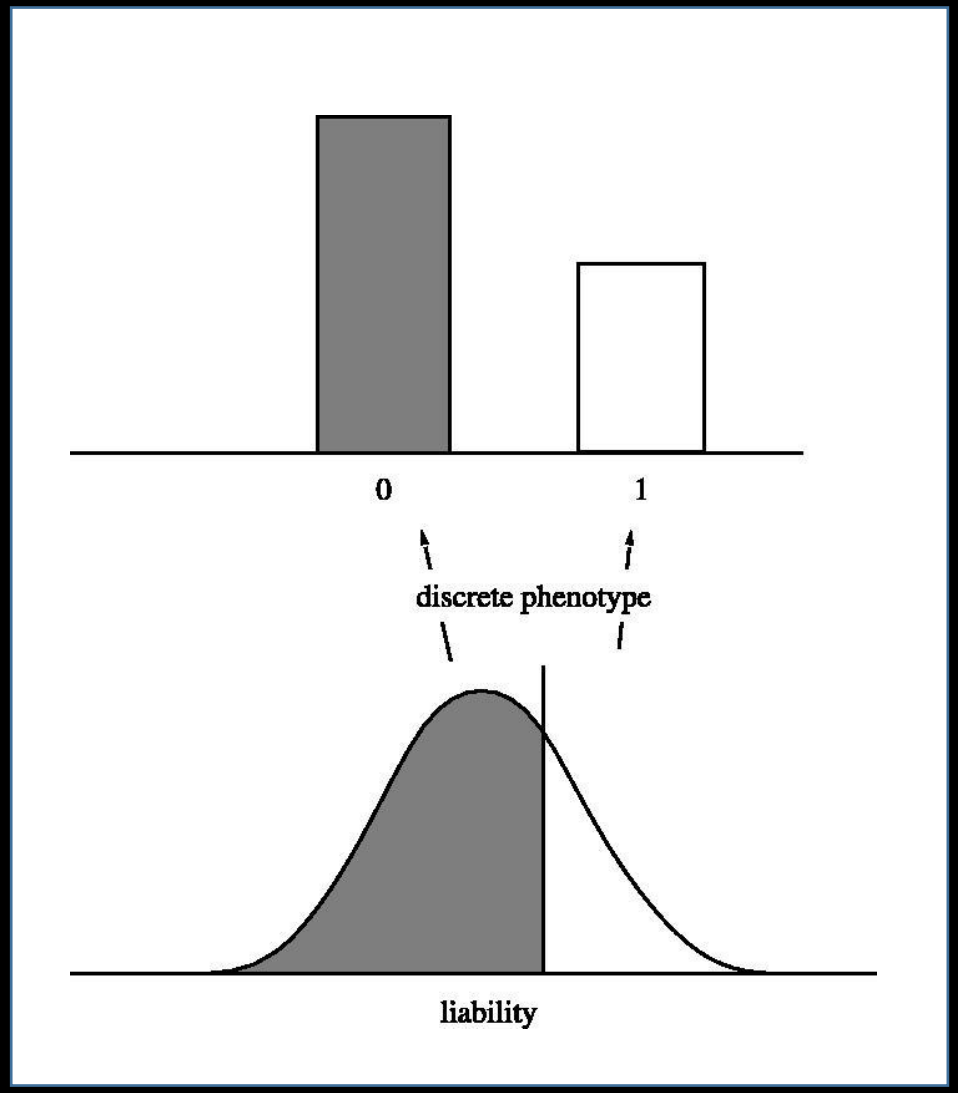
Joseph Felsenstein [1,*]

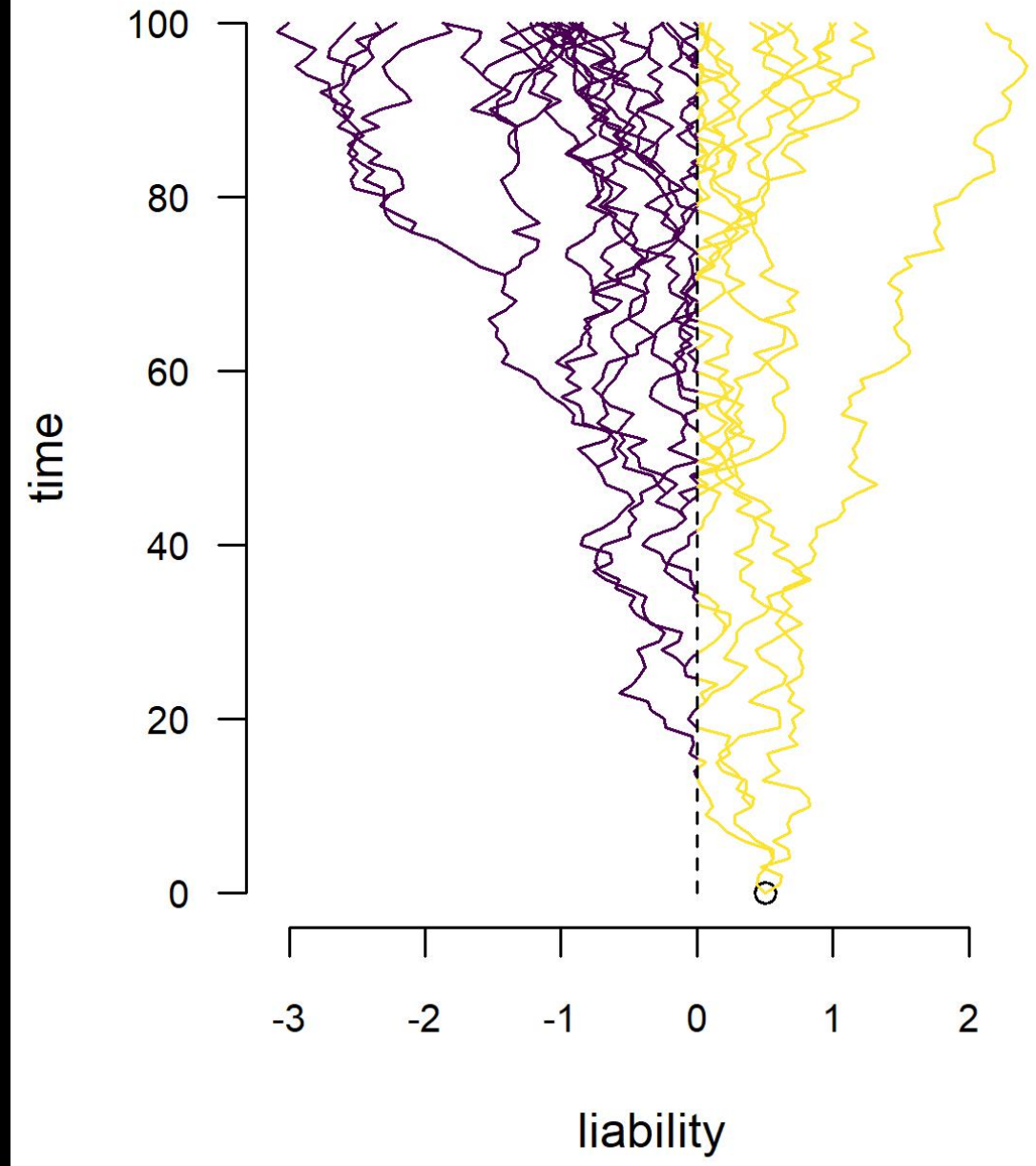
► Author information ► Article notes ► Copyright and License information

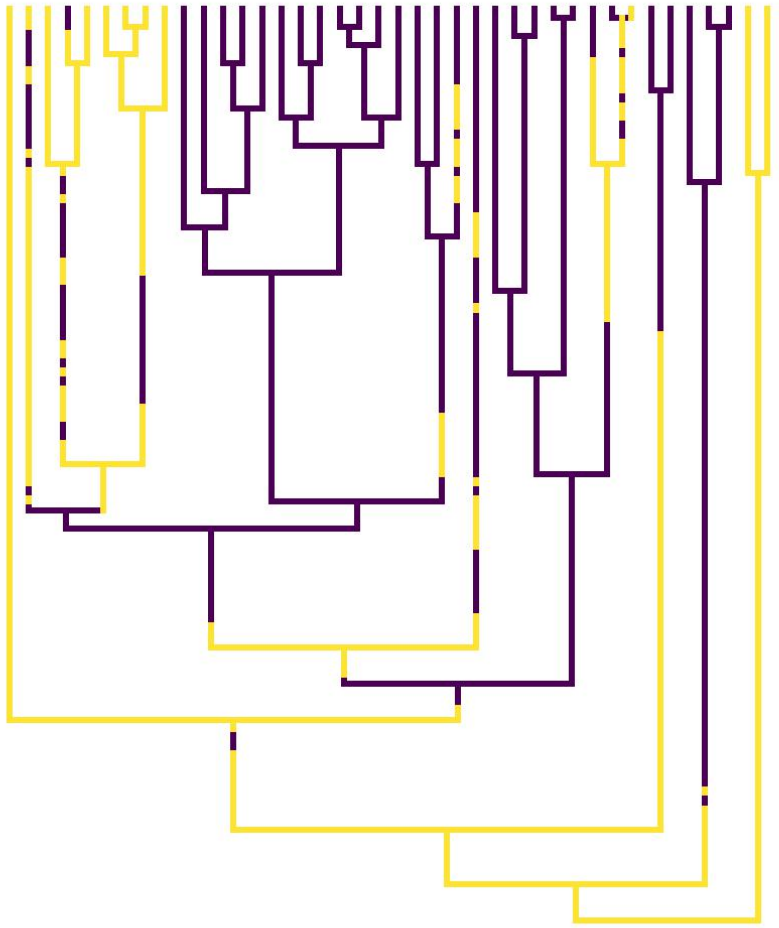
PMCID: PMC1569509 PMID: 16048785

What is a liability?

How is the behavior of a threshold trait on a phylogeny different than a standard Markov trait?

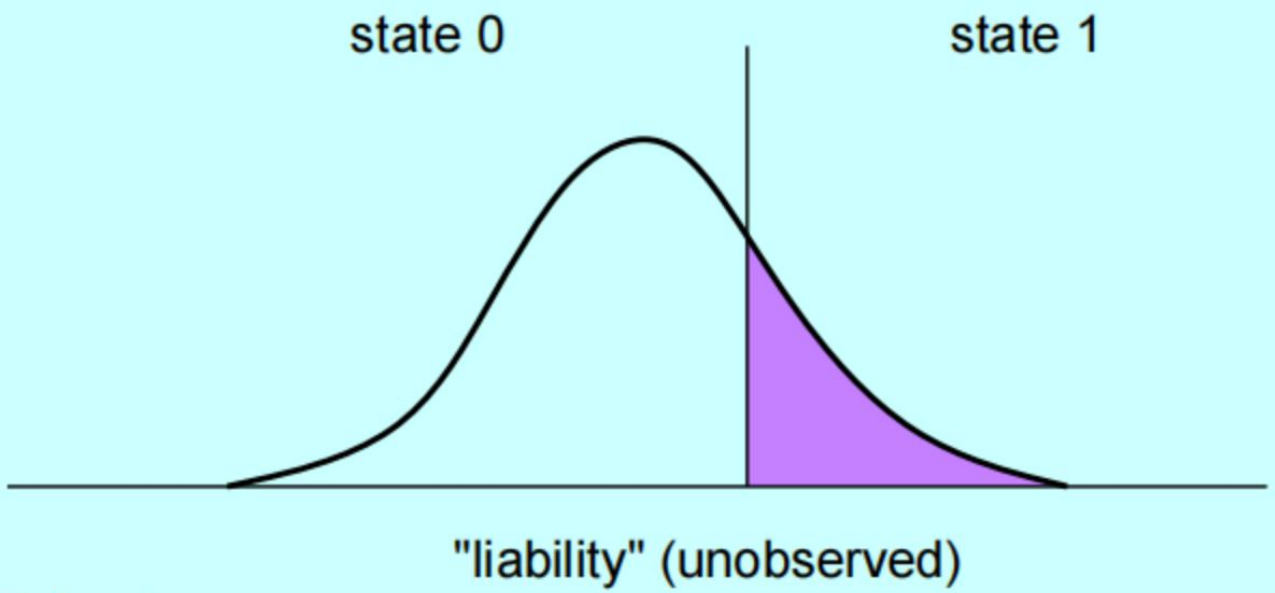



100
80
60
40
20
0
time
-3
-2
-1
0
1
2
liability



The threshold model

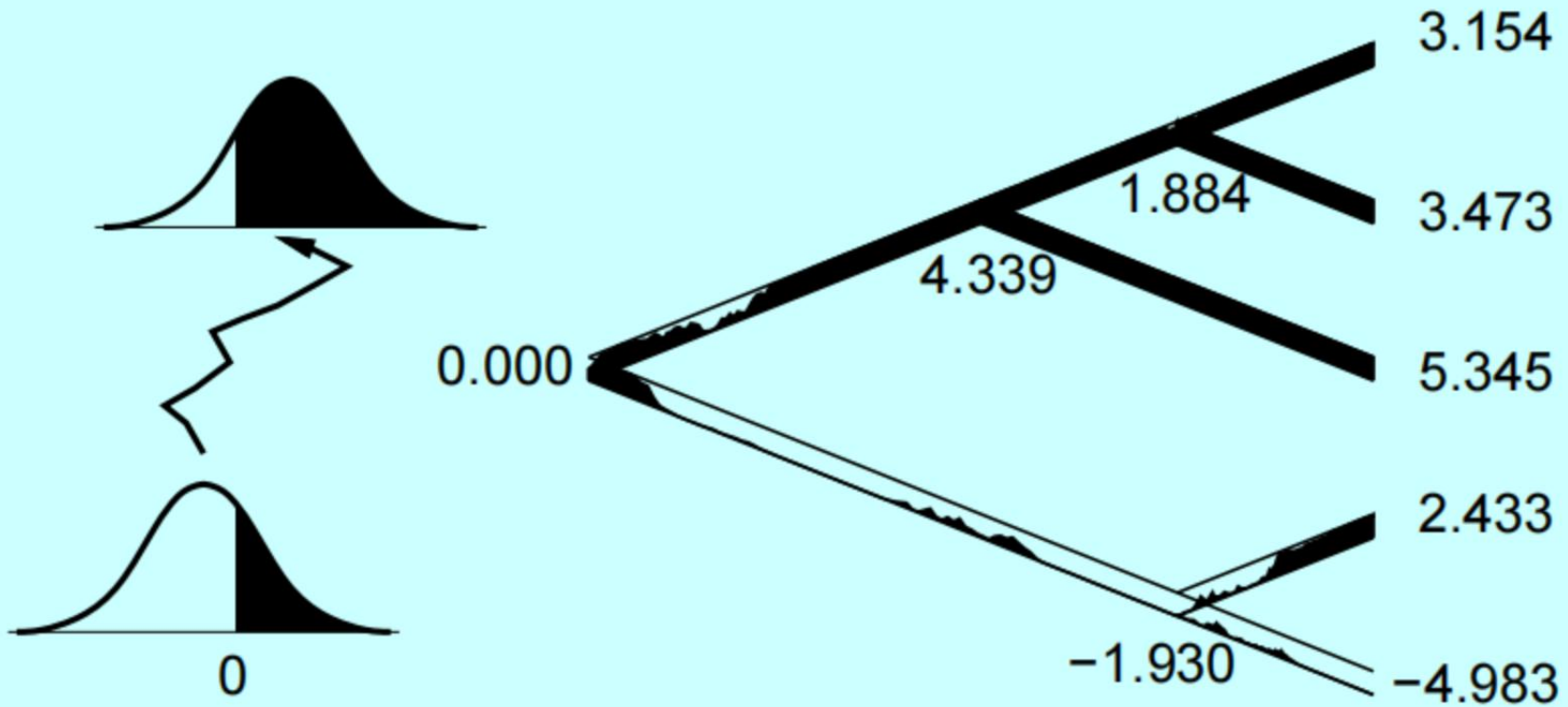
The threshold model (Wright, 1934; Falconer, 1965), plus Brownian motion



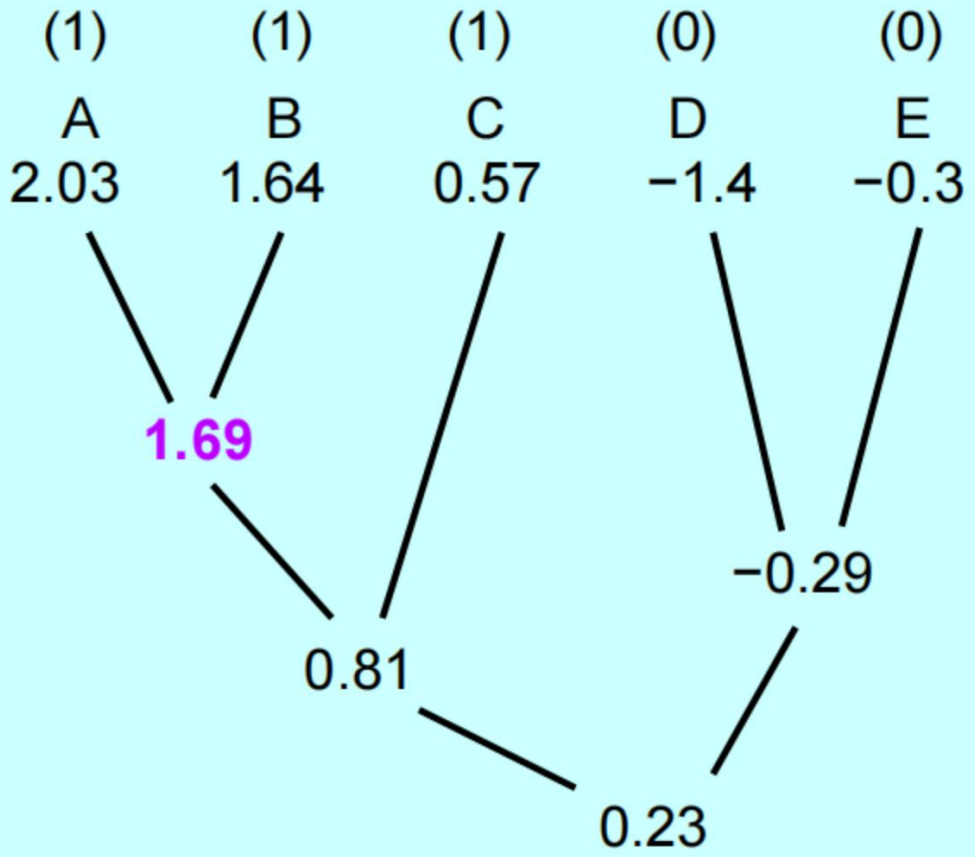
Advantages:

1. Predicts polymorphism as a lineage crosses the threshold
2. Soon after the threshold is crossed, one is more likely to revert. Less later.
3. Can allow covariation of characters

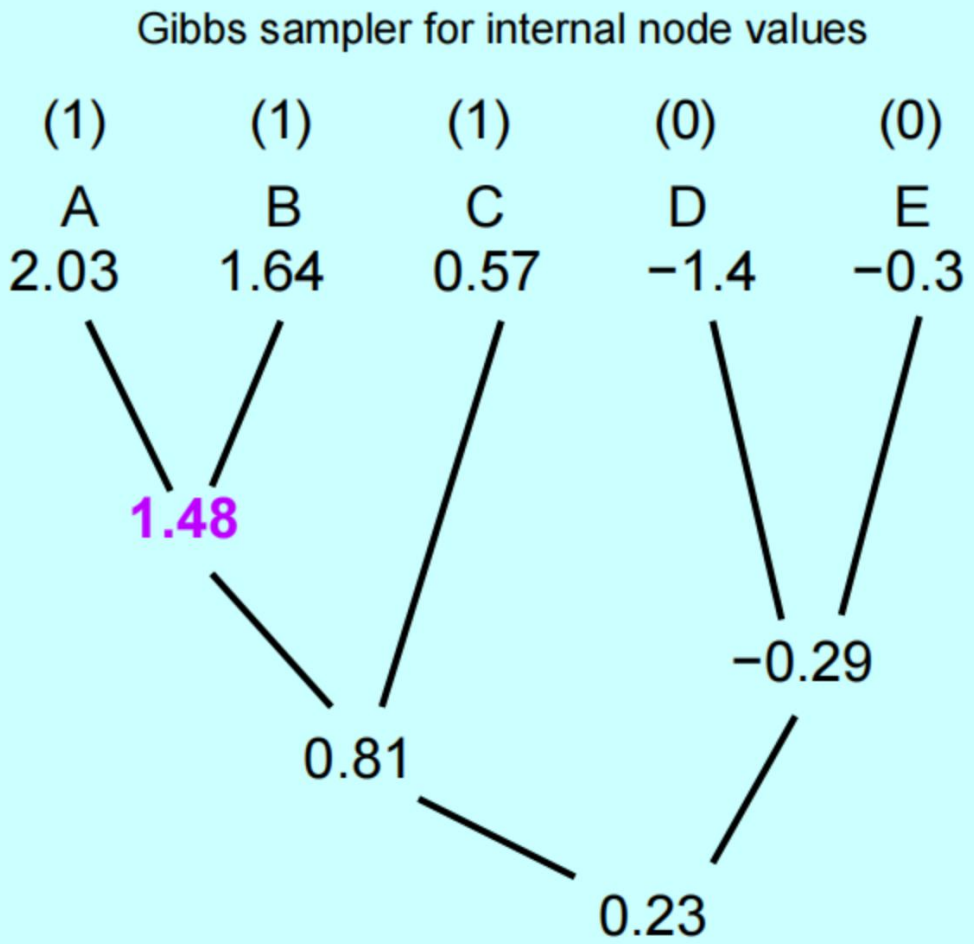
The threshold model on a tree



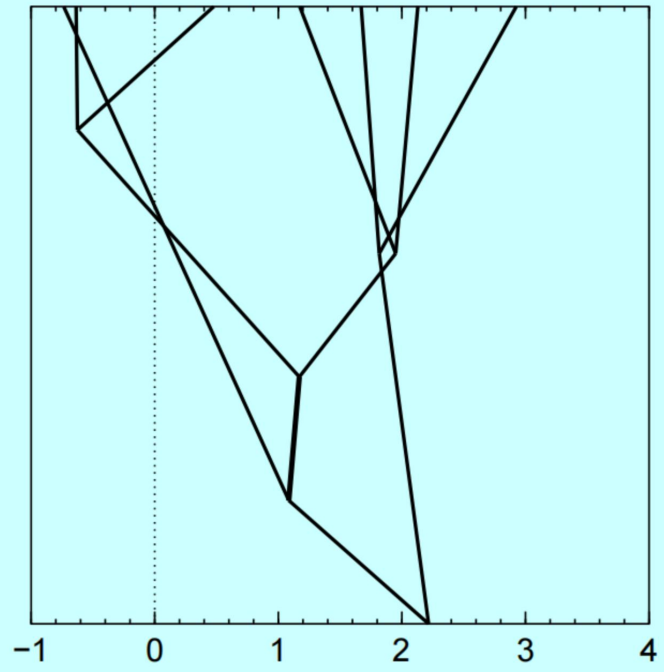
MCMC on liabilities



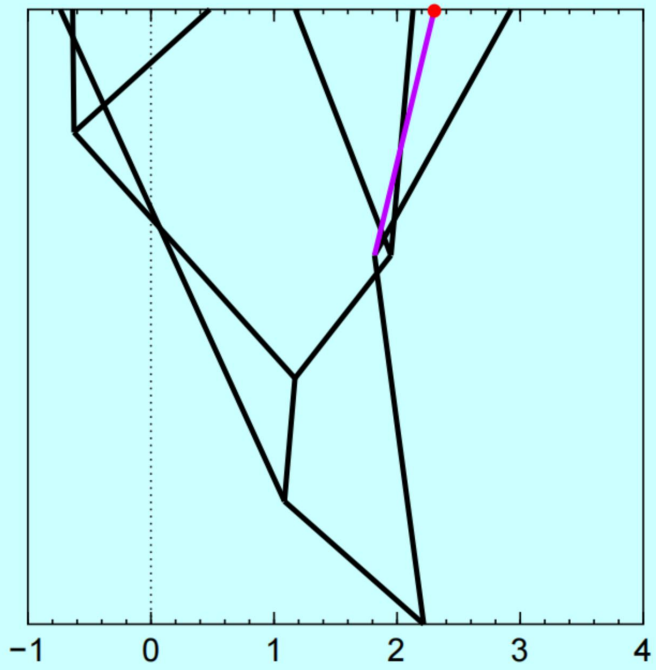
MCMC on liabilities: result of Gibbs sampling



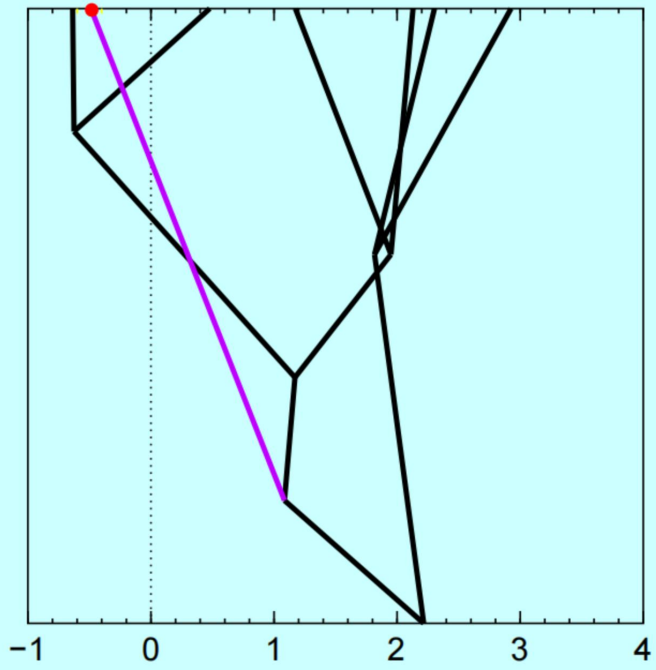
An example



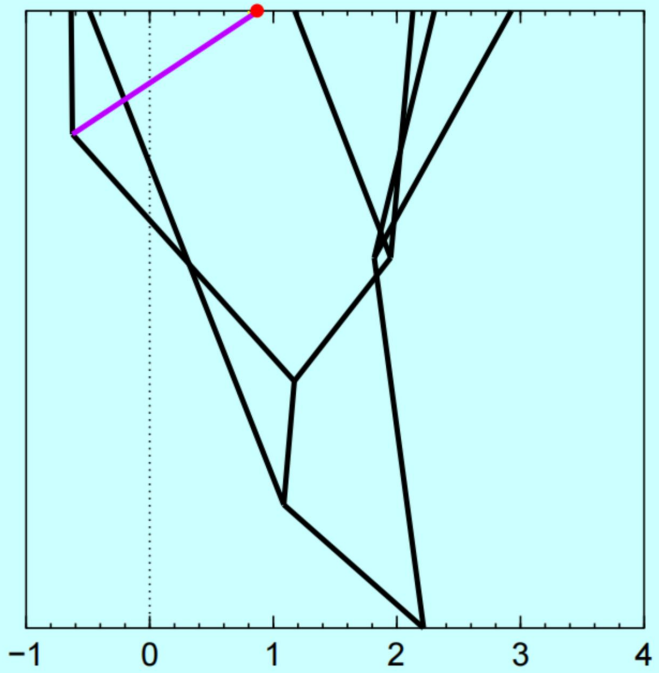
An example



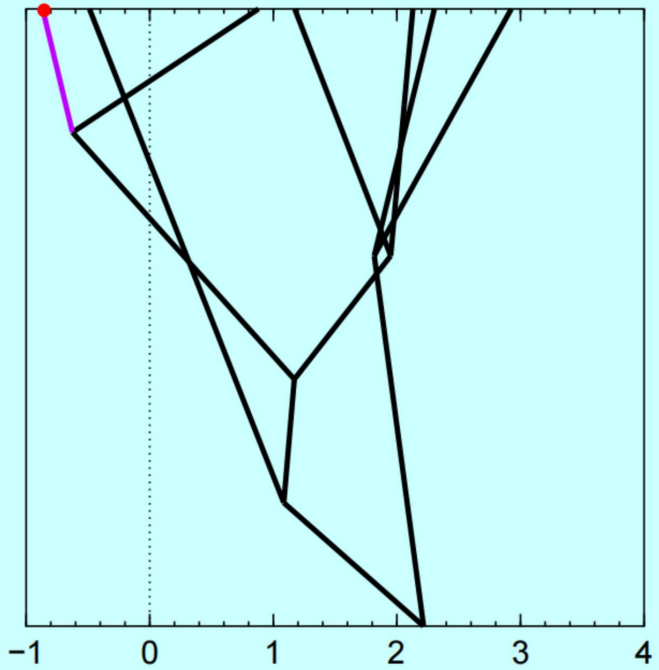
An example



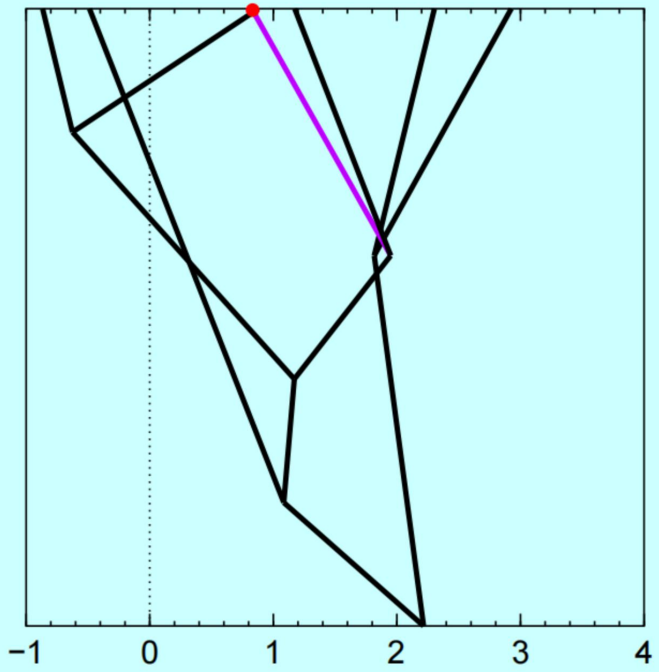
An example



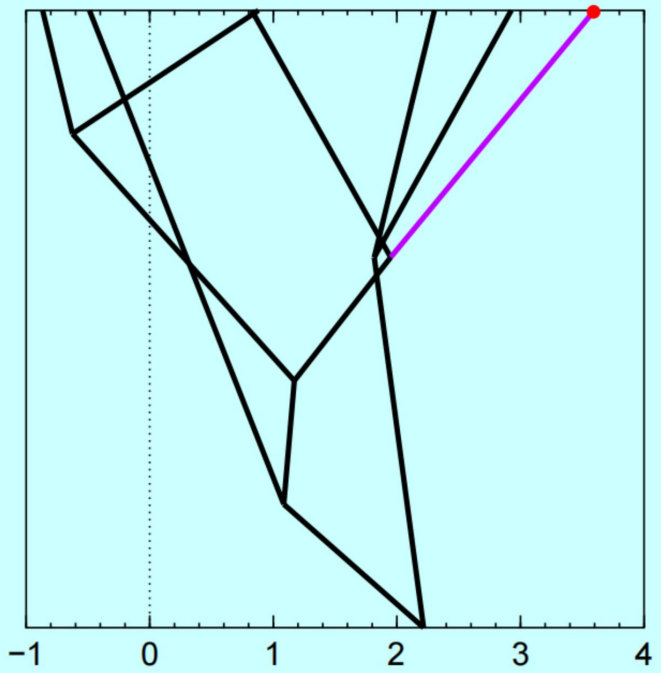
An example



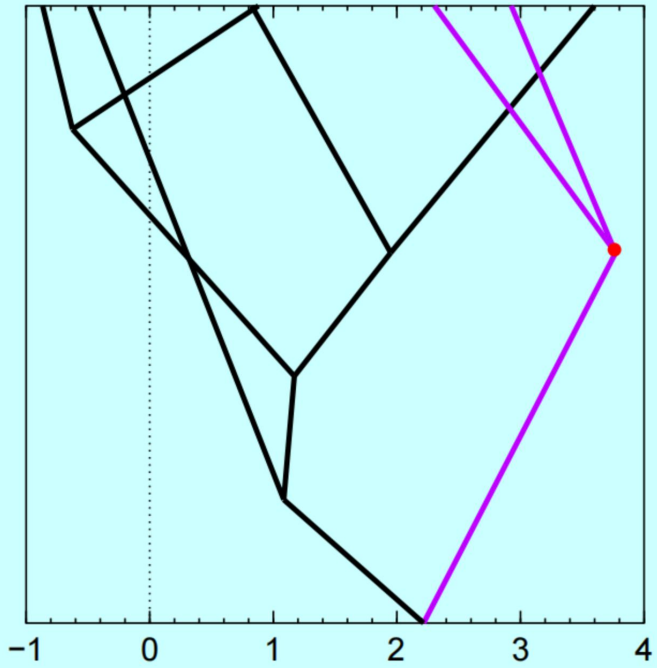
An example



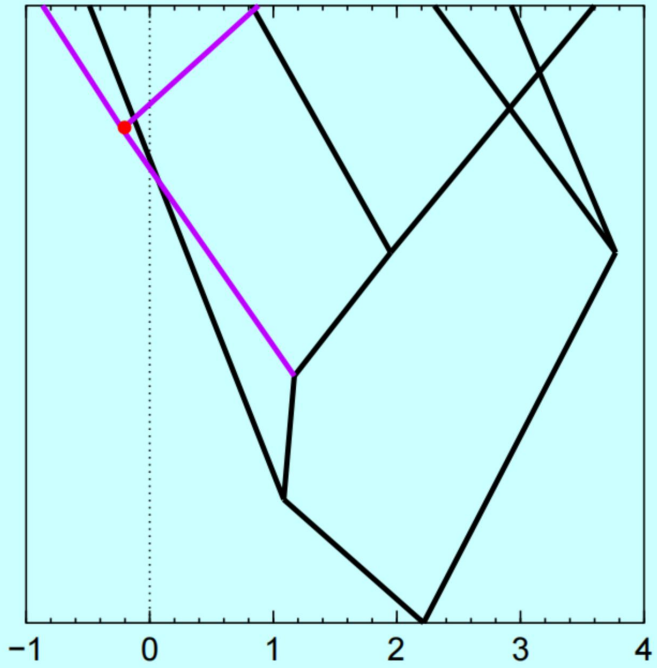
An example



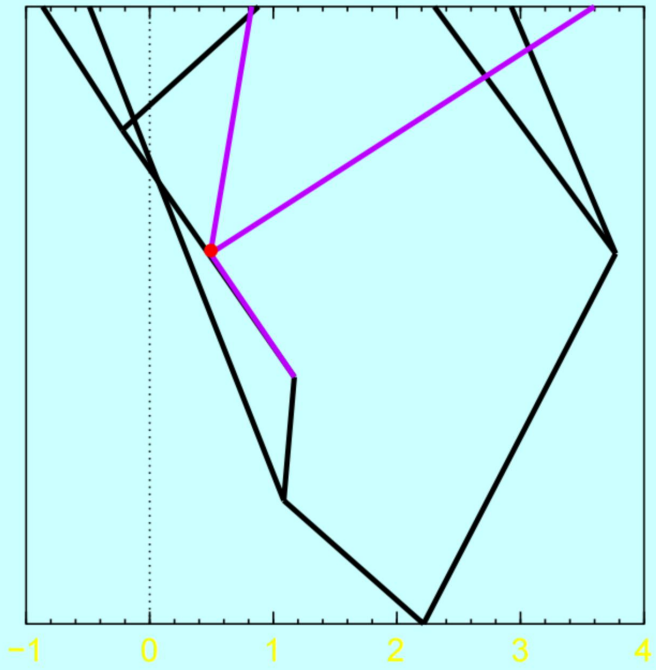
An example



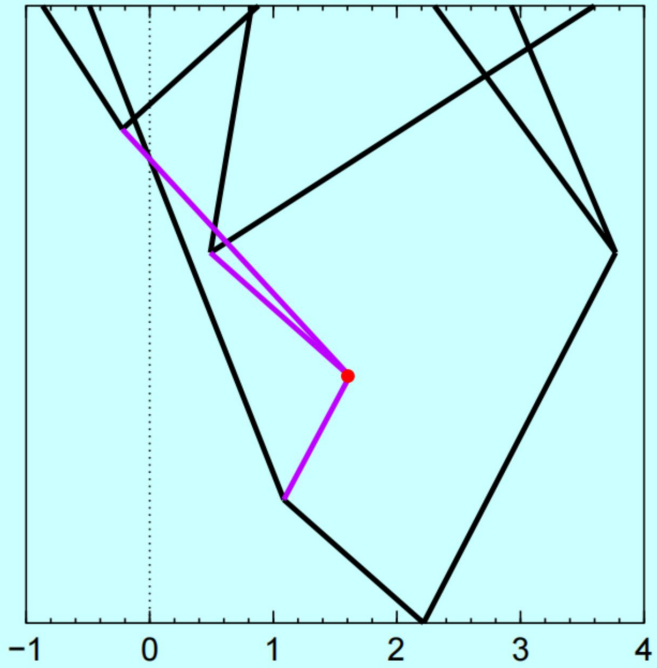
An example



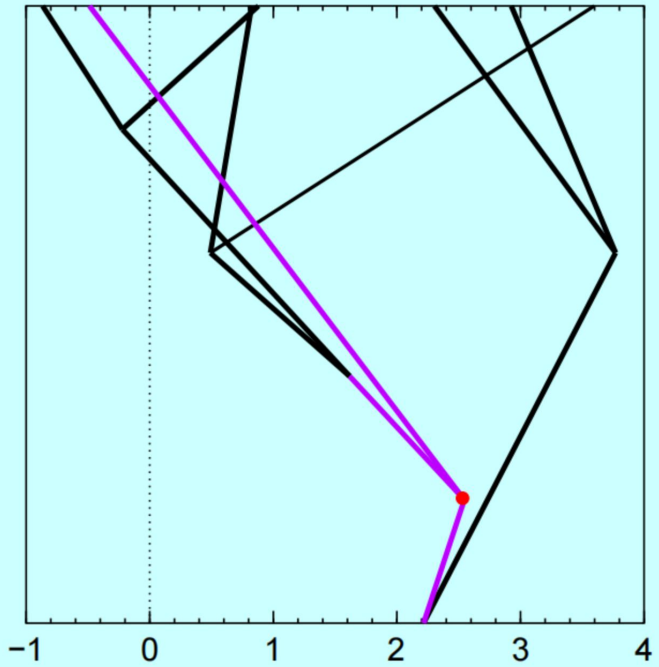
An example



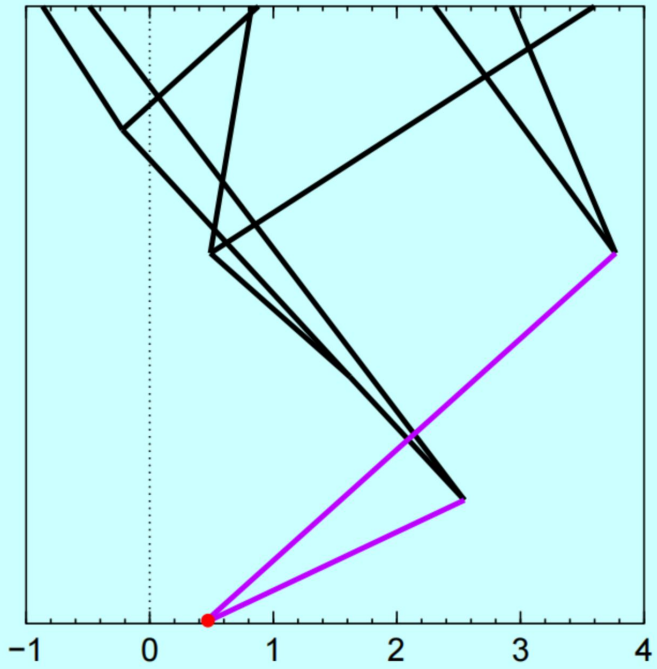
An example



An example



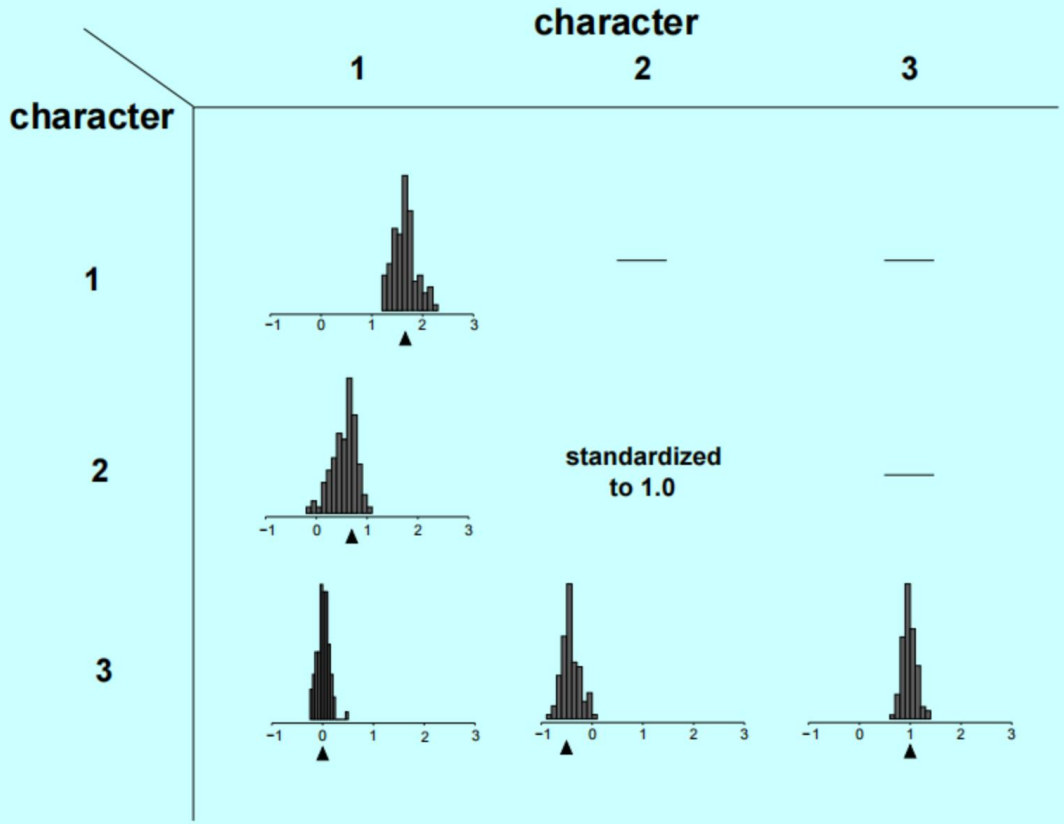
An example



Easy to combine with continuous characters!

Easily and naturally handles missing data!!

Simulations with both discrete and continuous characters



Characters 1 and 3 are continuous, character 2 is discrete. The inferred covariances are shown for the 100 simulated data sets.

We can model the threshold models with hidden state models as well!

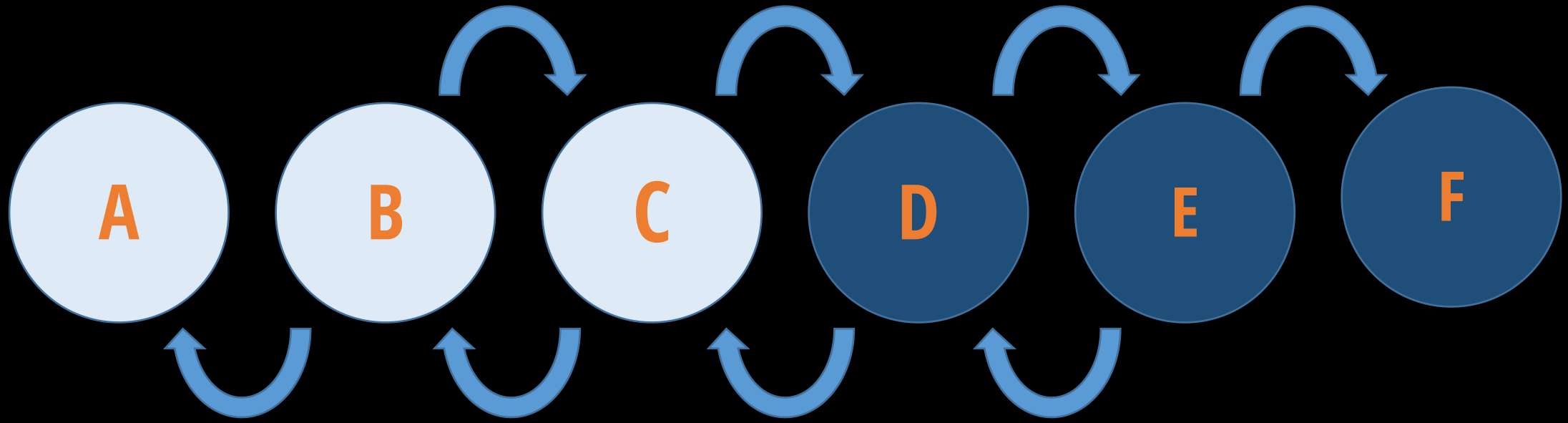
JOURNAL ARTICLE

Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms

Get access >

Jeremy M. Beaulieu ✉ , Brian C. O'Meara , Michael J. Donoghue

Systematic Biology, Volume 62, Issue 5, September 2013, Pages 725–737,
https://doi.org/10.1093/sysbio/syt034
Published: 14 May 2013 **Article history** ▾



ORIGINAL ARTICLE

doi:10.1111/evo.13947

Phylogenetic signal and evolutionary correlates of urban tolerance in a widespread neotropical lizard clade*

Kristin M. Winchell,[1,2] Klaus P. Schliep,[3,4] D. Luke Mahler,[5] and Liam J. Revell[3,6]

[1]Department of Biology, Washington University, St. Louis, Missouri 63130

[2]E-mail: kmwinchell@wustl.edu

[3]Department of Biology, University of Massachusetts Boston, Boston, Massachusetts 02125

[4]Institute of Computational BiotechnologyGraz University of Technology, Graz, Austria

[5]Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario M5S 3B2, Canada

[6]Facultad de Ciencias, Universidad Católica de la Santísima Concepción, Concepción, Chile

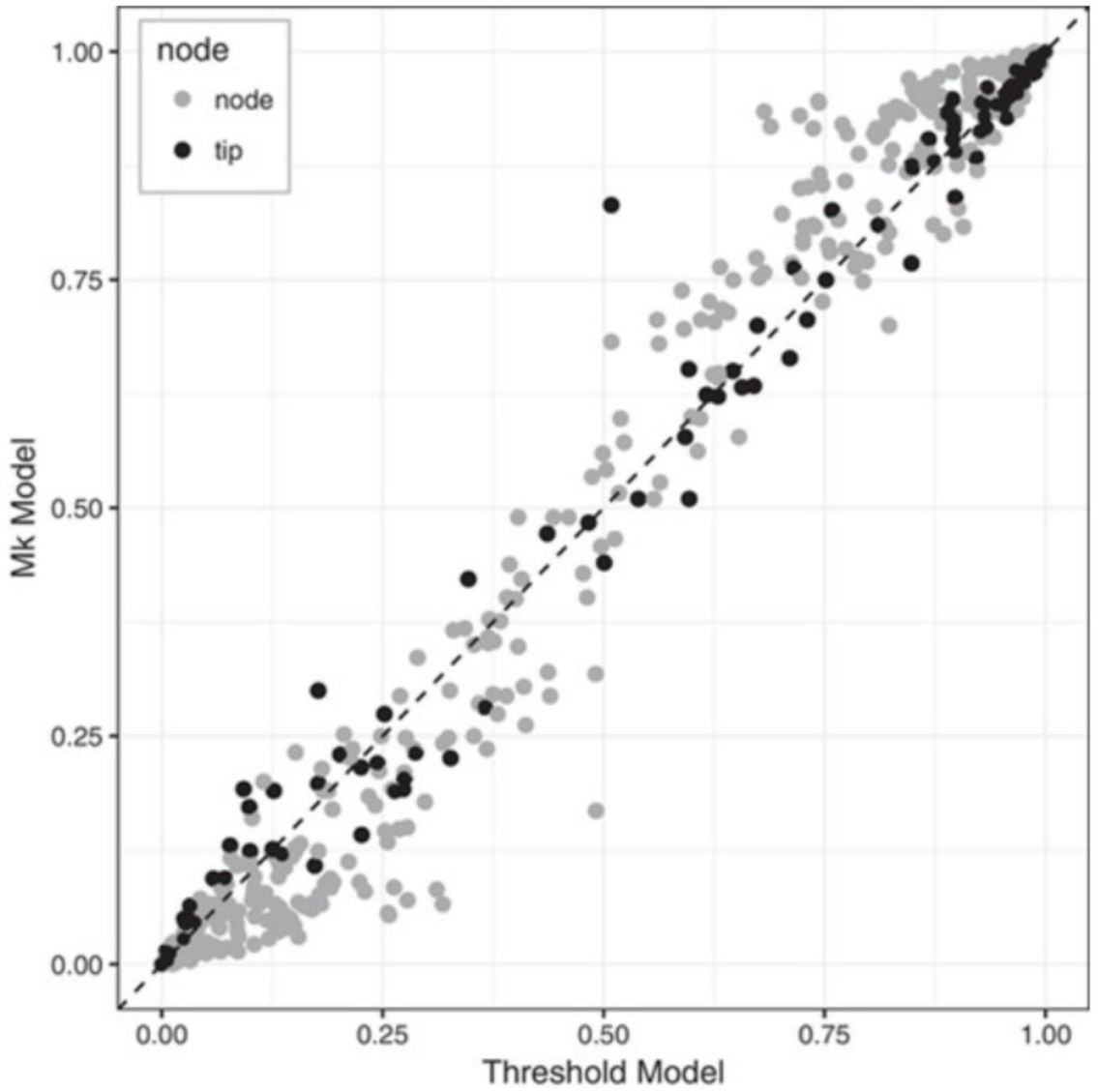


Figure 1 The threshold and Mk models yield similar results for posterior probabilities of tip (black) and node states (gray).

Syst. Biol. 65(4):651–661, 2016

DOI:10.1093/sysbio/syw015
Advance Access publication February 10, 2016

Inferring Bounded Evolution in Phenotypic Characters from Phylogenetic Comparative Data

FLORIAN C. BOUCHER[1,*] AND VINCENT DÉMERY[2]

[1]*Institute of Systematic Botany, University of Zurich, Zurich, Switzerland;* [2]*Laboratoire de Physico-Chimie Théorique, UMR Gulliver 7083, CNRS and ESPCI-ParisTech, Paris, France*

**Correspondence to be sent to: Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, 8010 Zurich, Switzerland; E-mail: florian.boucher@systbot.uzh.ch*

Received 15 June 2015; reviews returned 25 January 2016; accepted 25 January 2016
Associate Editor: Tanja Stadler

Abstract.—Our understanding of phenotypic evolution over macroevolutionary timescales largely relies on the use of stochastic models for the evolution of continuous traits over phylogenies. The two most widely used models, Brownian motion and the Ornstein–Uhlenbeck (OU) process, differ in that the latter includes constraints on the variance that a trait can attain in a clade. The OU model explicitly models adaptive evolution toward a trait optimum and has thus been widely used to demonstrate the existence of stabilizing selection on a trait. Here we introduce a new model for the evolution of continuous characters on phylogenies: Brownian motion between two reflective bounds, or Bounded Brownian Motion (BBM). This process also models evolutionary constraints, but of a very different kind. We provide analytical expressions for the likelihood of BBM and present a method to calculate the likelihood numerically, as well as the associated R code. Numerical simulations show that BBM achieves good performance: parameter estimation is generally accurate but more importantly BBM can be very easily discriminated from both BM and OU. We then analyze climatic niche evolution in diprotodonts and find that BBM best fits this empirical data set, suggesting that the climatic niches of diprotodonts are bounded by the climate available in Australia and the neighboring islands but probably evolved with little additional constraints. We conclude that BBM is a valuable addition to the macroevolutionary toolbox, which should enable researchers to elucidate whether the phenotypic traits they study are evolving under hard constraints between bounds. [BBM; bounds; evolutionary constraints; macroevolution; maximum likelihood estimation; phylogenetic comparative data.]

Bounded Brownian Motion

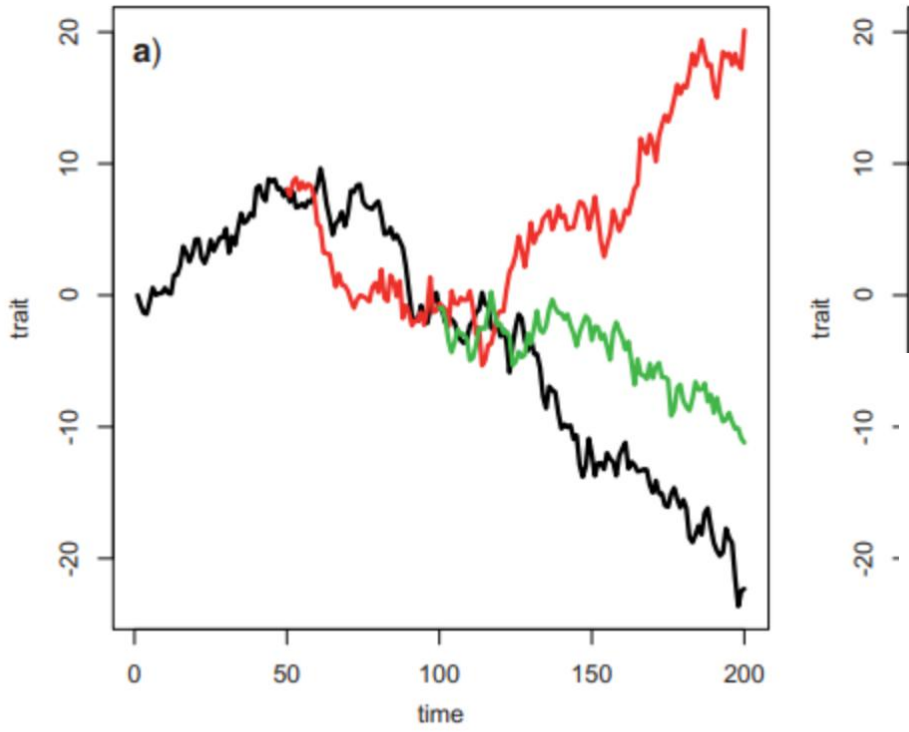
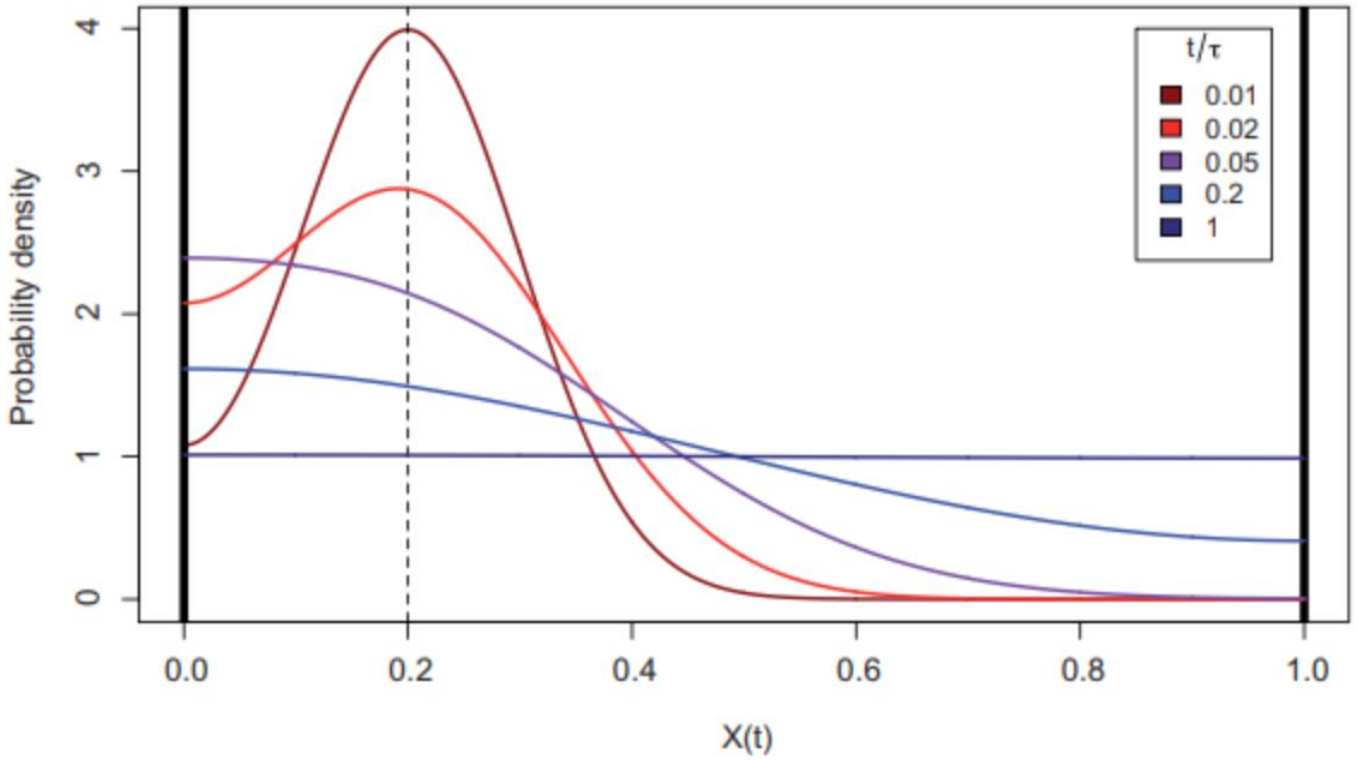


FIGURE 1. Behaviour of the BBM model. The two panels show the evolution of
Motion without bounds (BM). b) Brownian Motion with two bounds (horizontal
multiple hits on the bounds occur. Both plots were generated using the same val
higher in BM.



Syst. Biol. 67(2):304–319, 2018

DOI:10.1093/sysbio/syx075
Advance Access publication September 22, 2017

A General Model for Estimating Macroevolutionary Landscapes

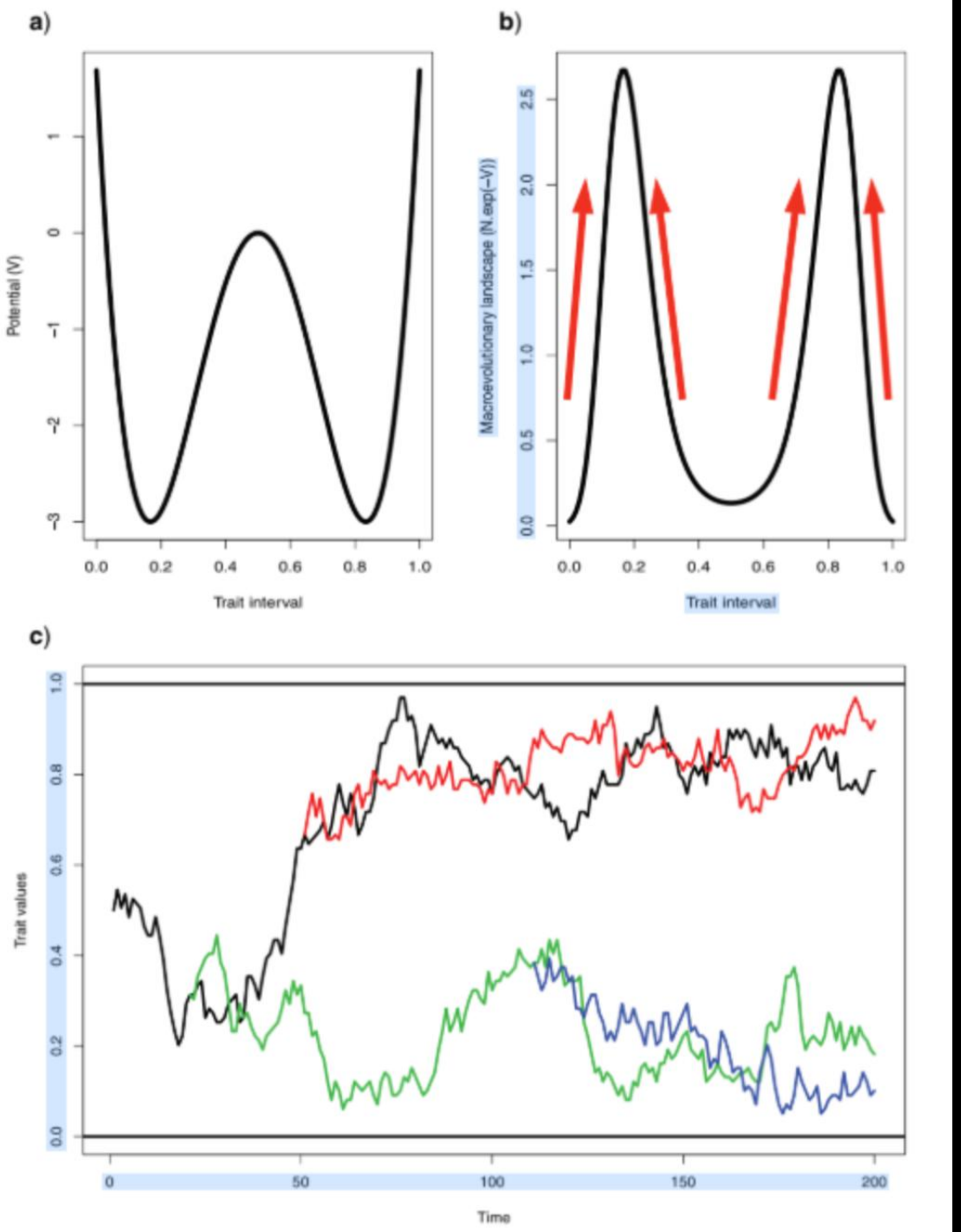
FLORIAN C. BOUCHER[1,2,*], VINCENT DÉMERY[3], ELENA CONTI[1], LUKE J. HARMON[4,5], AND JOSEF UYEDA[4]

[1]*Department of Systematic and Evolutionary Botany (ISEB), University of Zurich, Zurich, Switzerland;* [2]*Department of Botany and Zoology, University of Stellenbosch, Stellenbosch, South Africa;* [3]*Gulliver, CNRS, ESPCI Paris, PSL Research University, 10 rue Vauquelin, Paris, France;*
[4]*Department of Biological Sciences and Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, ID, USA; and*
[5]*Department of Fish Ecology and Evolution, Swiss Federal Institute of Aquatic Science and Technology (Eawag), Center for Ecology, Evolution, and Biogeochemistry, 6047 Kastanienbaum, Switzerland*
**Correspondence to be sent to: Department of Botany and Zoology, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa; E-mail: floflo boucher@gmail.com.*

Received 10 January 2017; reviews returned 4 September 2017; accepted 8 September 2017
Associate Editor: Richard Ree

Abstract.—The evolution of quantitative characters over long timescales is often studied using stochastic diffusion models. The current toolbox available to students of macroevolution is however limited to two main models: Brownian motion and the Ornstein–Uhlenbeck process, plus some of their extensions. Here, we present a very general model for inferring the dynamics of quantitative characters evolving under both random diffusion and deterministic forces of any possible shape and strength, which can accommodate interesting evolutionary scenarios like directional trends, disruptive selection, or macroevolutionary landscapes with multiple peaks. This model is based on a general partial differential equation widely used in statistical mechanics: the Fokker–Planck equation, also known in population genetics as the Kolmogorov forward equation. We
be used to
describe macr
how it can
be fitted to em
rimination
from alternati
data using
either maximu
les of body
mass evolutio
ied since it
opens the way
usion; FPK
model; macroevolution; maximum-likelihood estimation; MCMC methods; phylogenetic comparative data; selection.]

"To compute the likelihood of FPK, we instead discretize the trait interval by considering only a set of n points equally spaced between two extreme values, Bmin and Bmax, a procedure already used for the BBM model (Boucher and Démery, 2016)."


a)
b)
c)
Potential (V)
Trait interval
Macroevolutionary landscape (N.exp(−V))
Trait interval
Trait values
Time

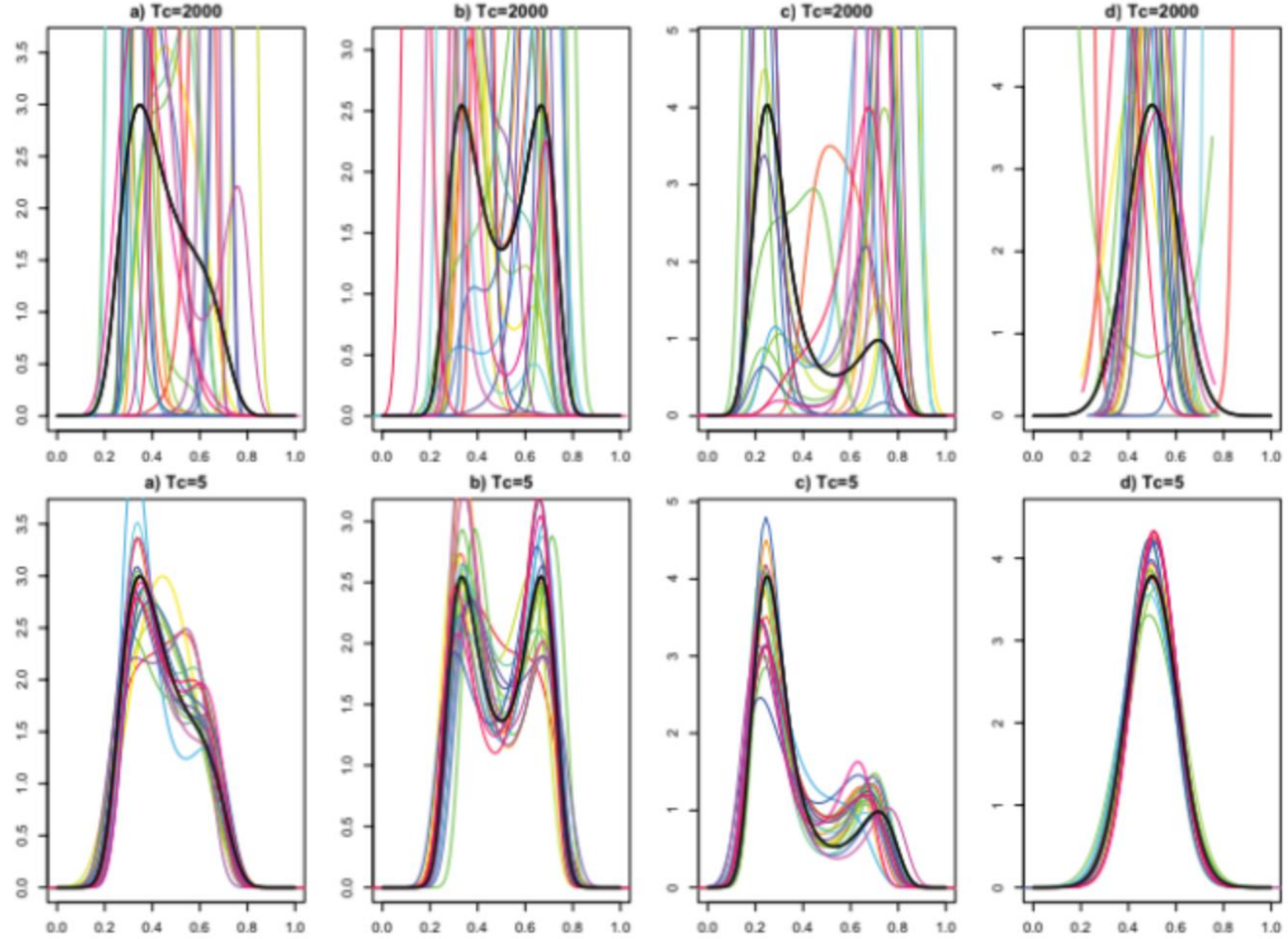


FIGURE 3. Estimation of the macroevolutionary landscape in different versions of the FPK model. Thin lines in each plot show the macroevolutionary landscapes estimated in 20 simulations, each one in a different color, while the simulated macroevolutionary landscape is shown by the thick black line. Only results for trees with 100 tips are shown. Top row: simulations with $Tc = 2,000$, in which stationarity was not reached. Bottom row: simulations with $Tc = 5$, in which stationarity was reached. From left to right, columns show simulation for scenarios a to d.

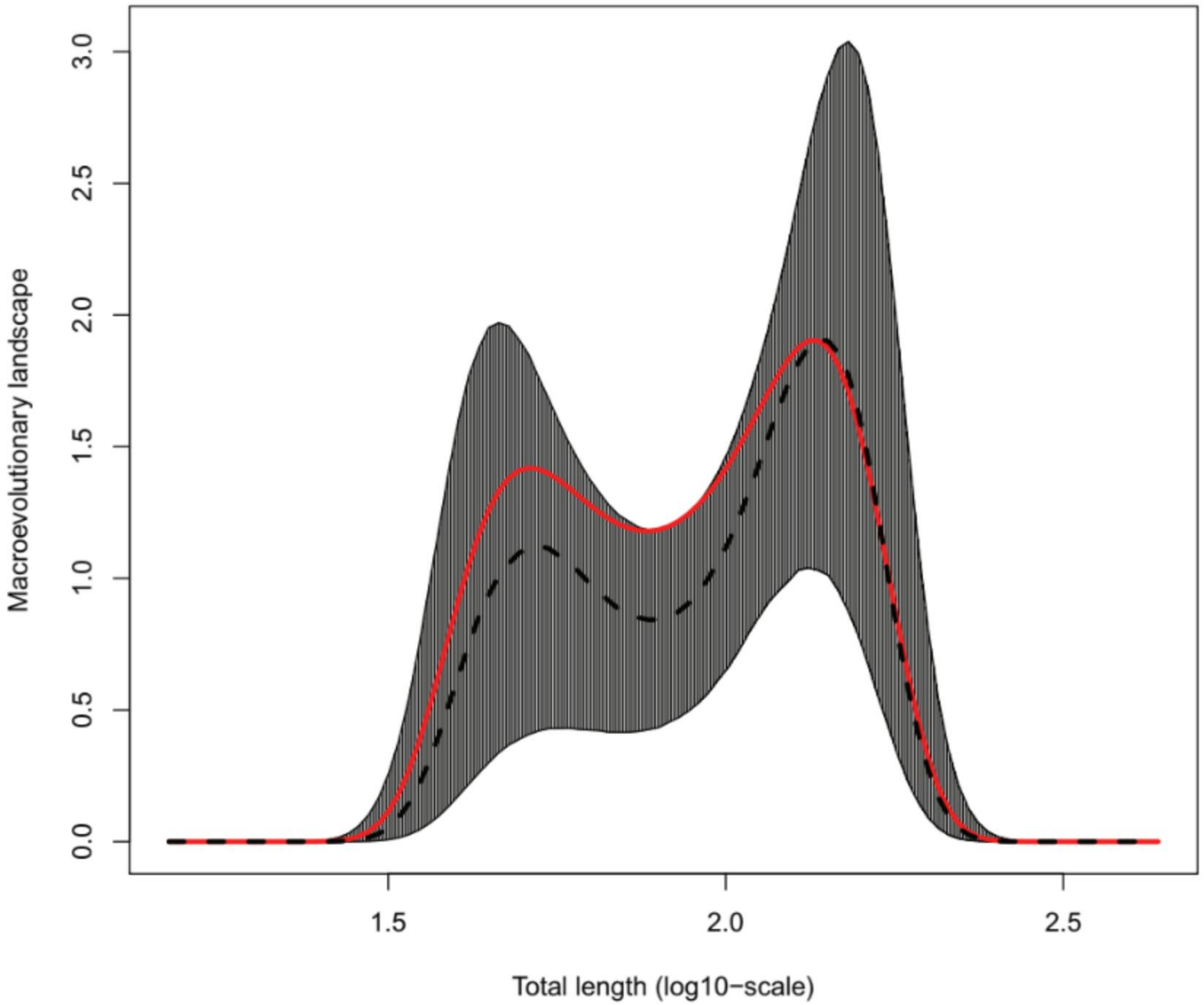


FIGURE 5. Posterior distribution of the macroevolutionary landscape estimated for body length evolution in watersnakes (tribe Thamnophiini). This posterior distribution was obtained by concatenating the two MCMC chains after the first 10% of samples were discarded as burnin (800,000 MCMC steps in total). The figure shows the value of the macroevolutionary landscape (N.exp($-V(x)$)) on the y-axis as a function of log10(total length) measured in centimeters. The dashed black line shows the median value of the macroevolutionary landscape over the posterior, while the grey area ranges from the 25% to the 75% quantiles. The solid red line shows the maximum-likelihood estimate of the macroevolutionary landscape.

BBMV is a flexible method for estimating macroevolutionary landscapes

Discretization could blow up the number of parameters, but by maintaining 1) 1-dimensional change and 2) using a FPK equation so that transition rates depend only on the estimated parameters of a simpler function

We can do this for lots of PCMs, and make things simpler

Problem:

PGLS assumes:

The *expected value* of trait Y

is predicted by the *current value* of X

with phylogenetically correlated residuals

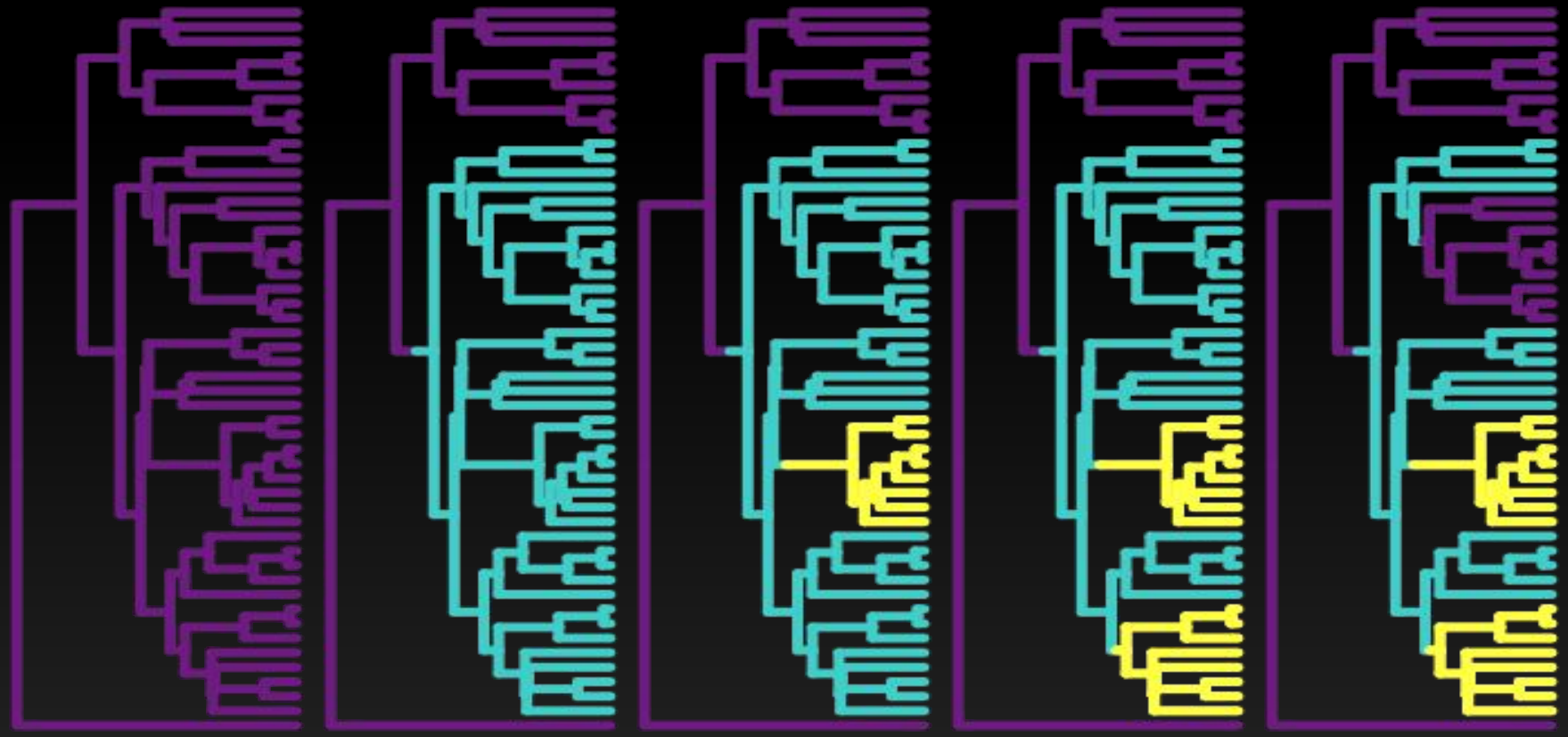
History matters

...

...must model
evolution of predictor

Ornstein-Uhlenbeck models

“Painting regimes”



$Ln(conduit\ diameter) \sim Temp_{min}$

$Temp_{min}$ Freeze / Not Freeze

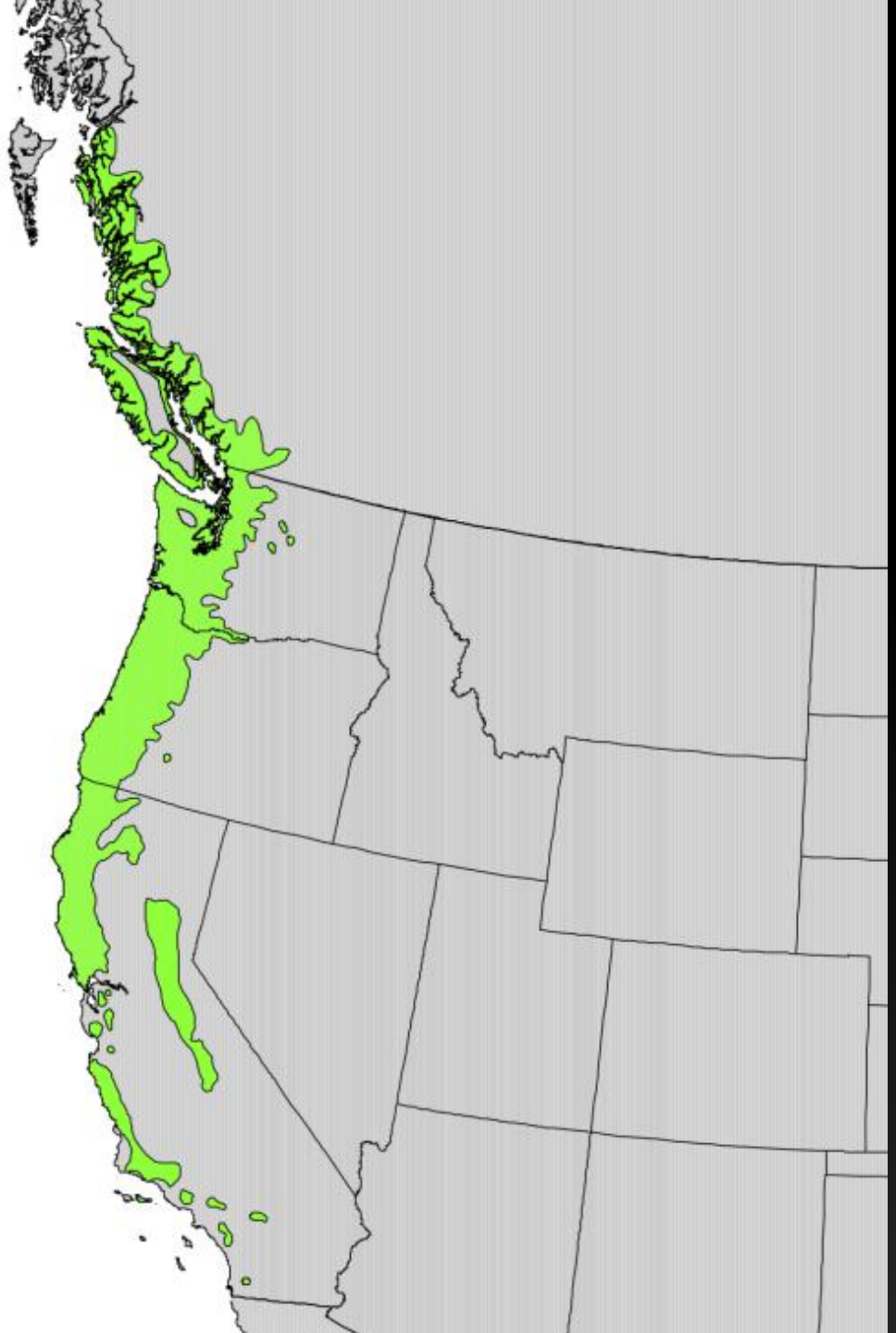
Phenology Deciduous / Evergreen

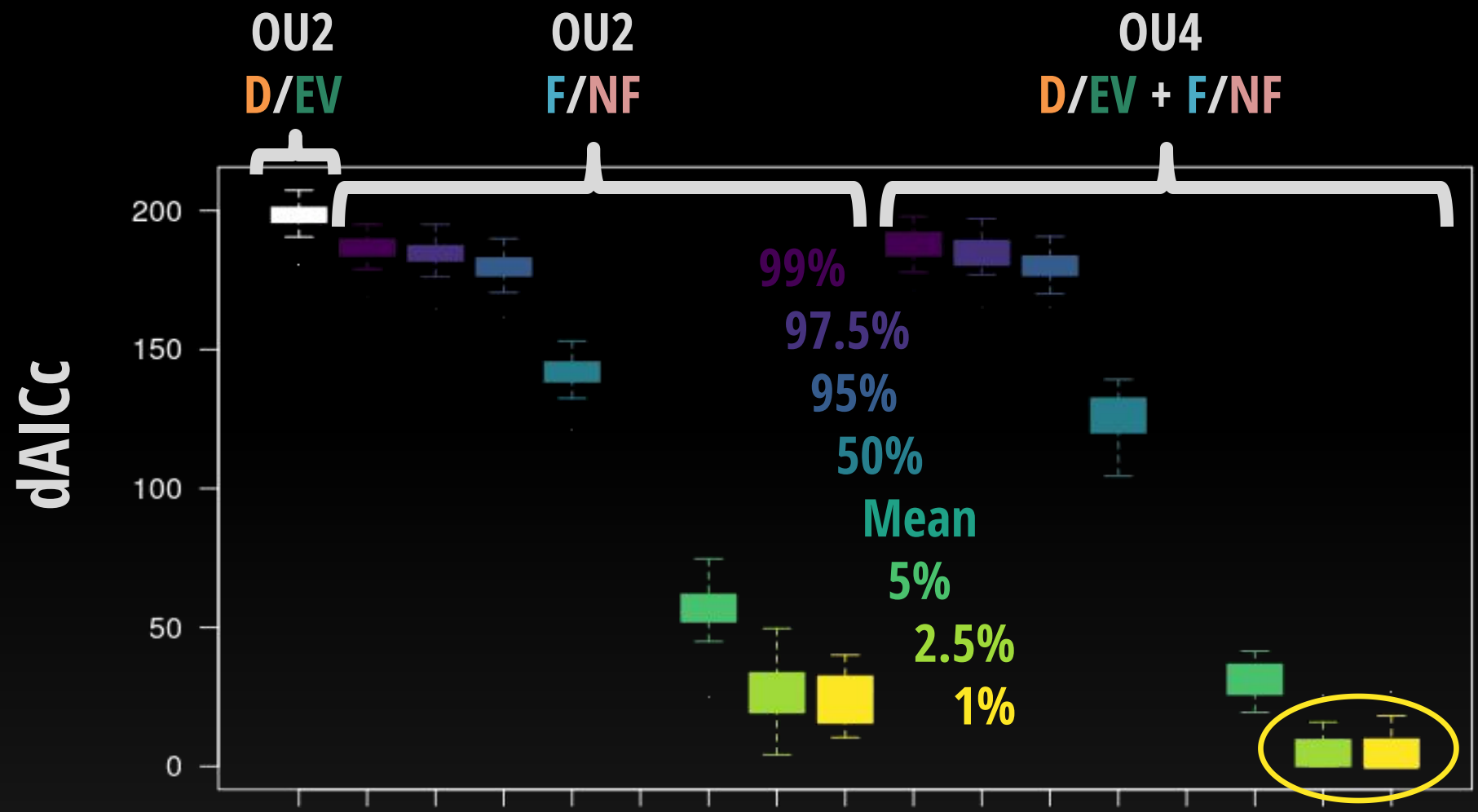
398 species

Analyzed in OUwie (Beaulieu et al. 2012)

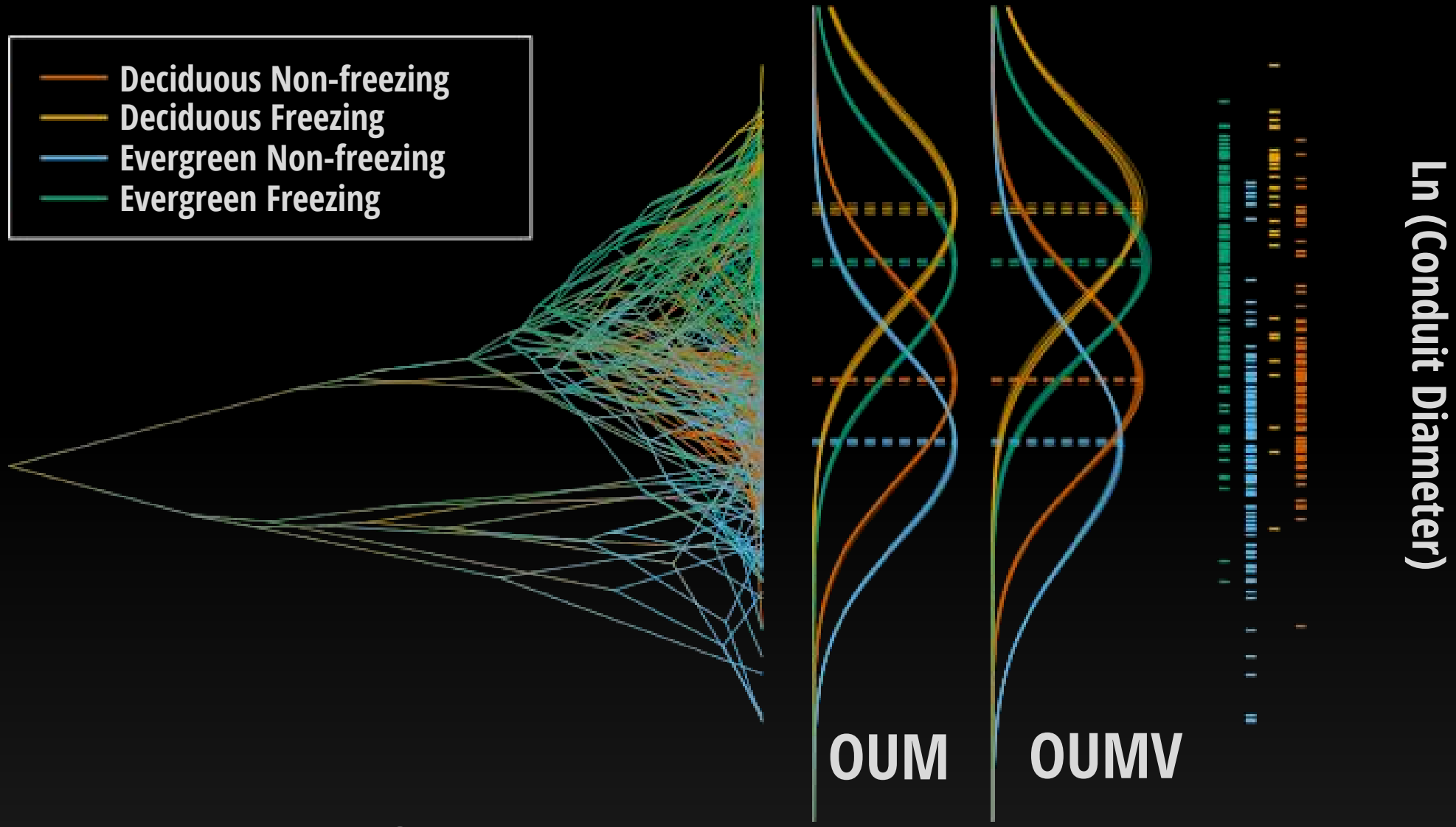
Will Pearse

GBIF
Global Biodiversity Information Facility





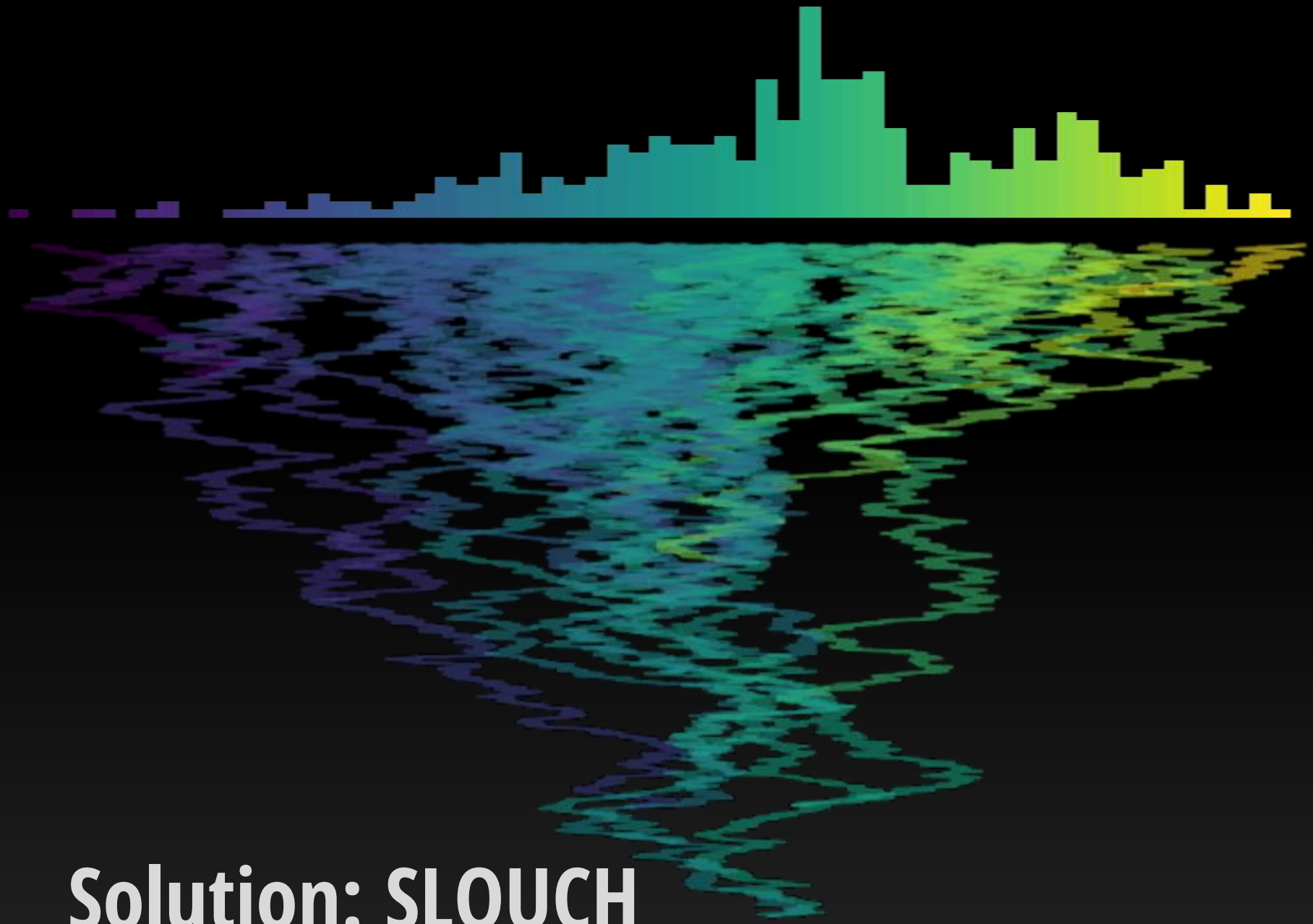
Climate at range margins + leaf phenology best predicts conduit diameter



No evidence of change in variance, only mean (OUM best model)

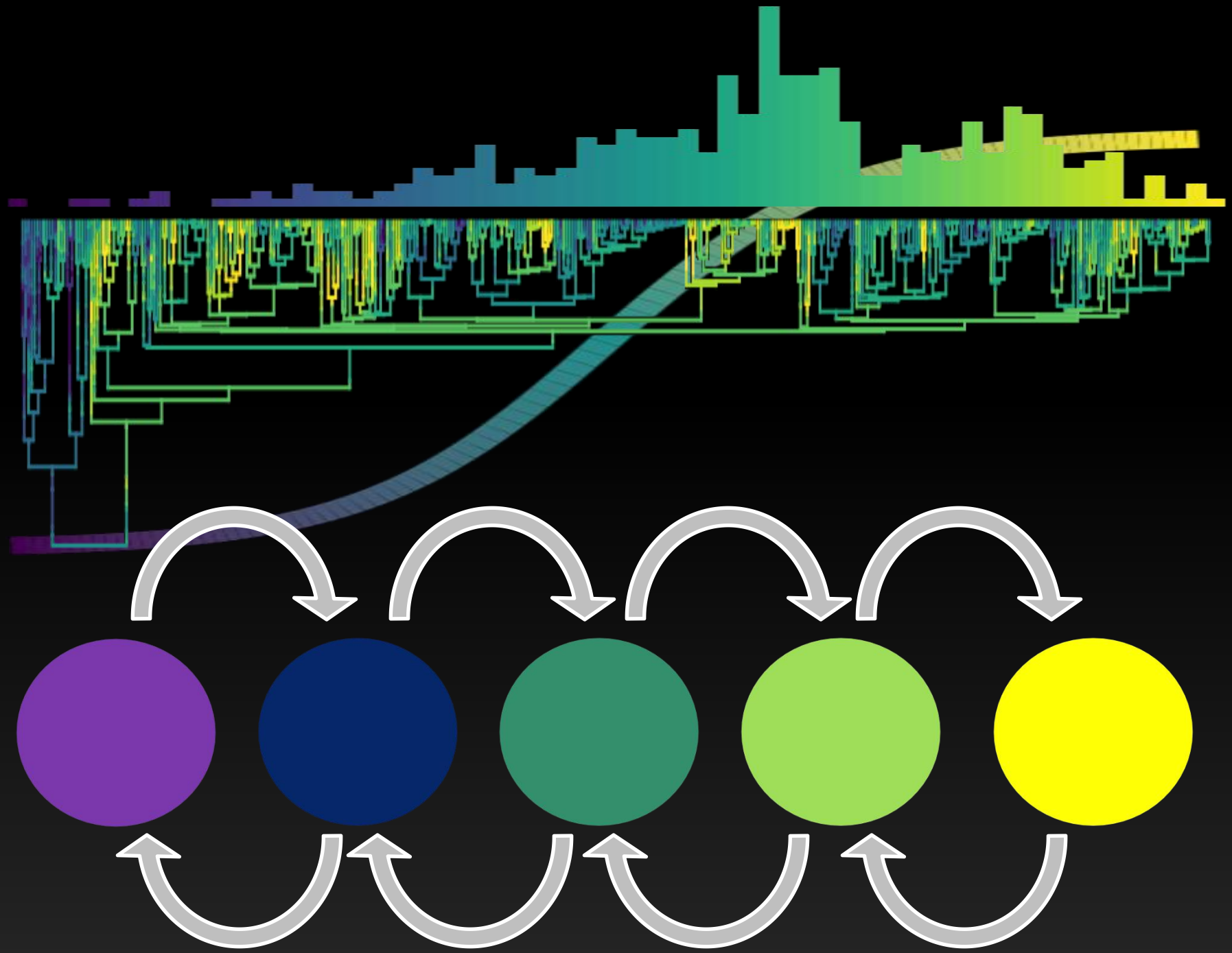
Discrete predictors are
...unsatisfying...

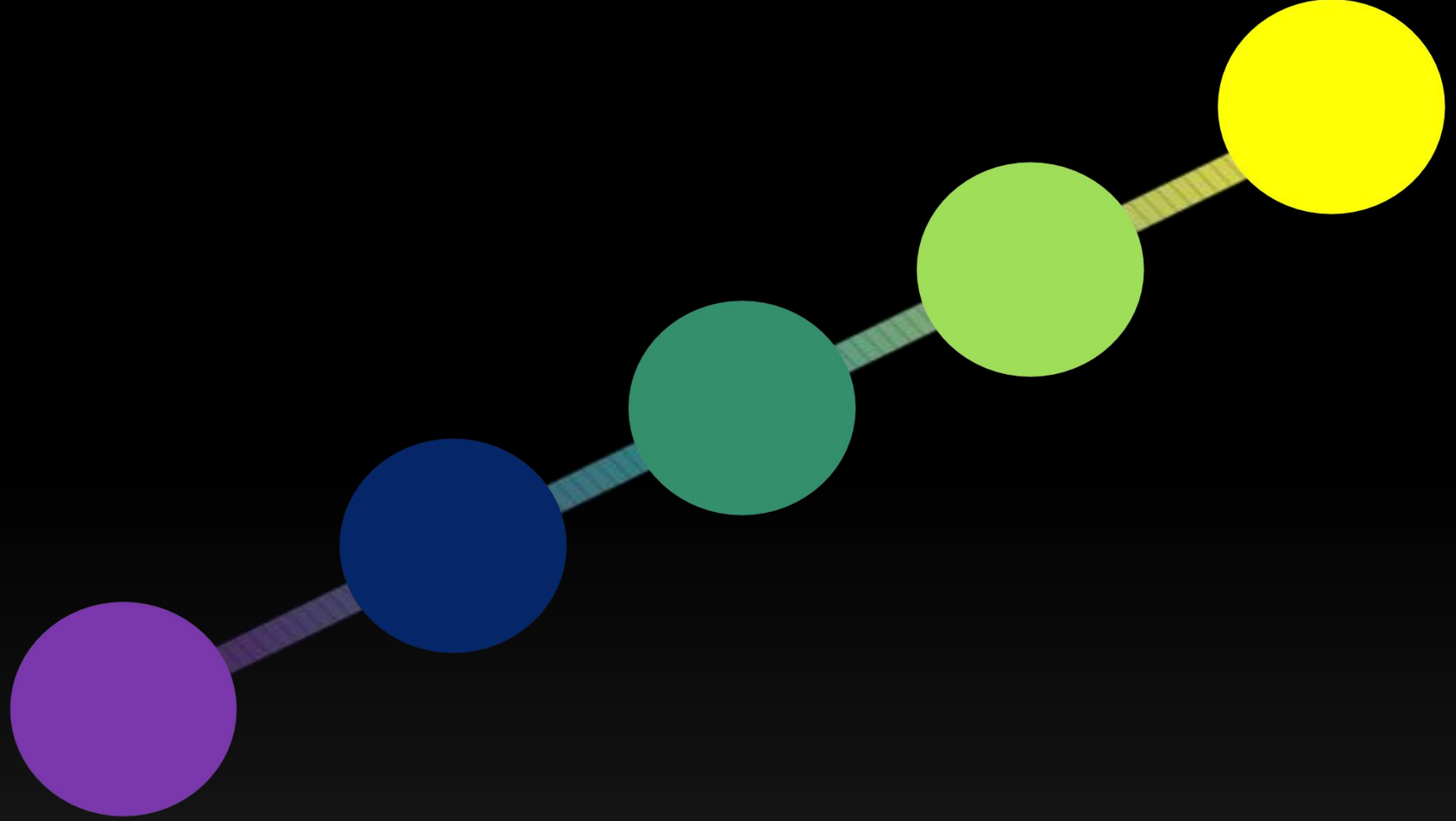
Does temperature act on adaptive optima beyond the effect of freezing?

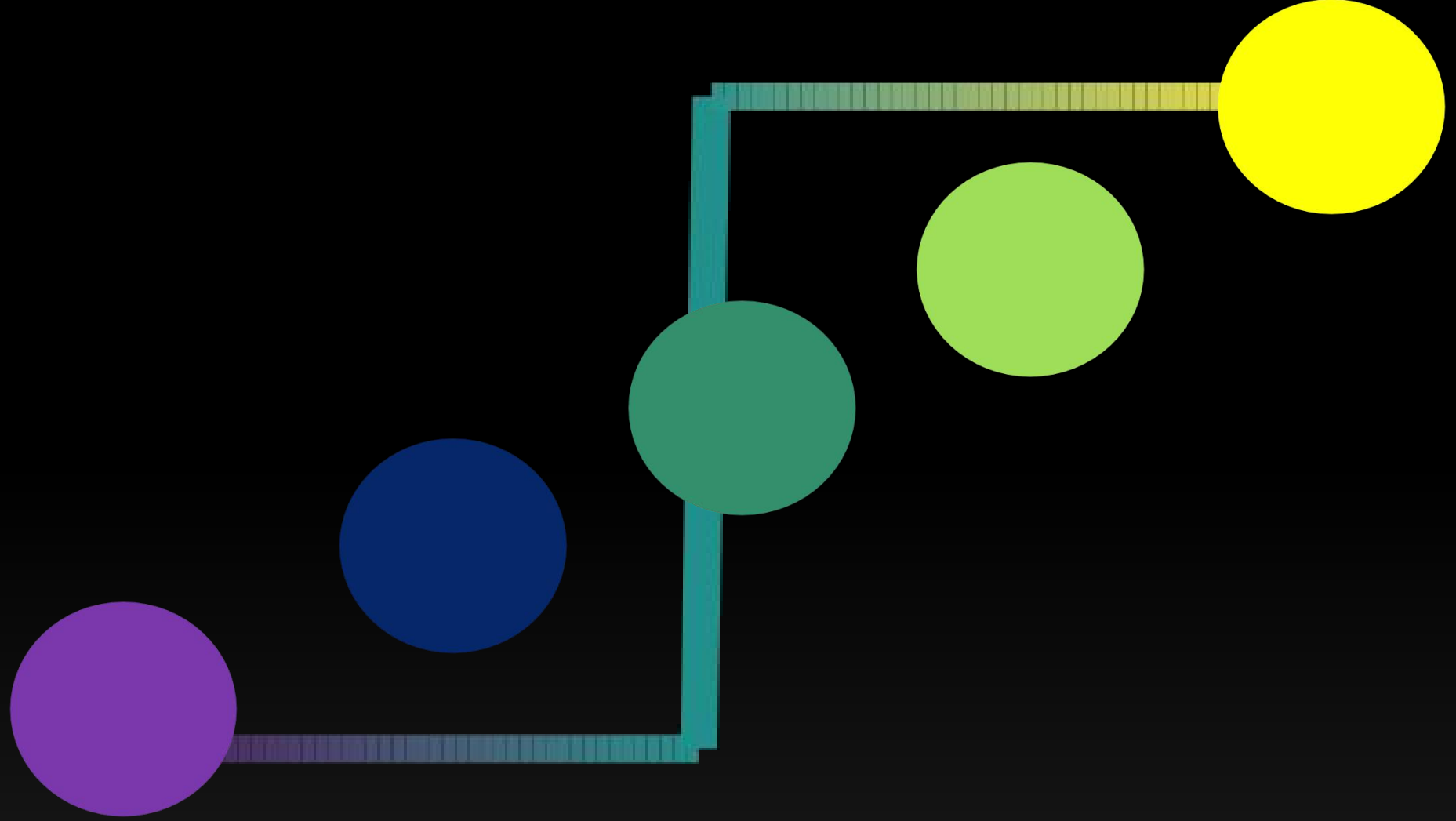


Solution: SLOUCH

Hansen, Pienaar & Orzack, 2008







Functions

Step
Step0 (Center = 0°C)
Sigmoid
Sigmoid0 (Center = 0°C)
Linear
Linear + step

} X 100

Simmap Trees w/ 10 bins

Functions

Step

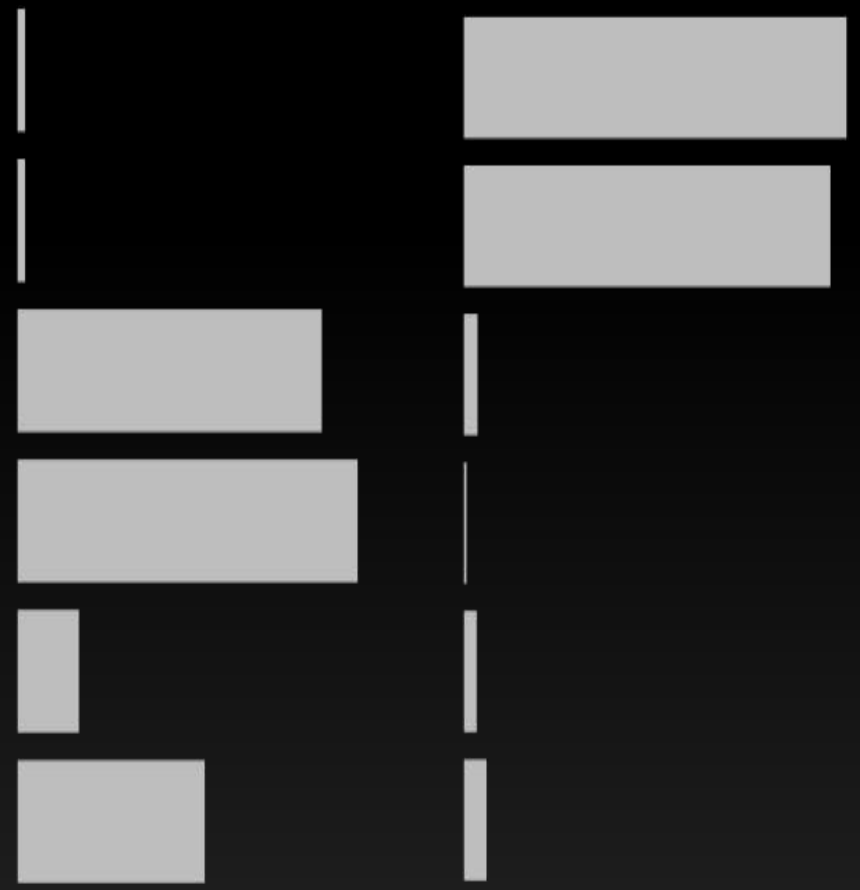
Step0 (Center = 0°C)

Sigmoid

Sigmoid0 (Center = 0°C)

Linear

Linear + step



But what about leaf phenology?

Split the dataset into

Deciduous and Evergreen species

Deciduous

Functions

AICw

dAICc

Step

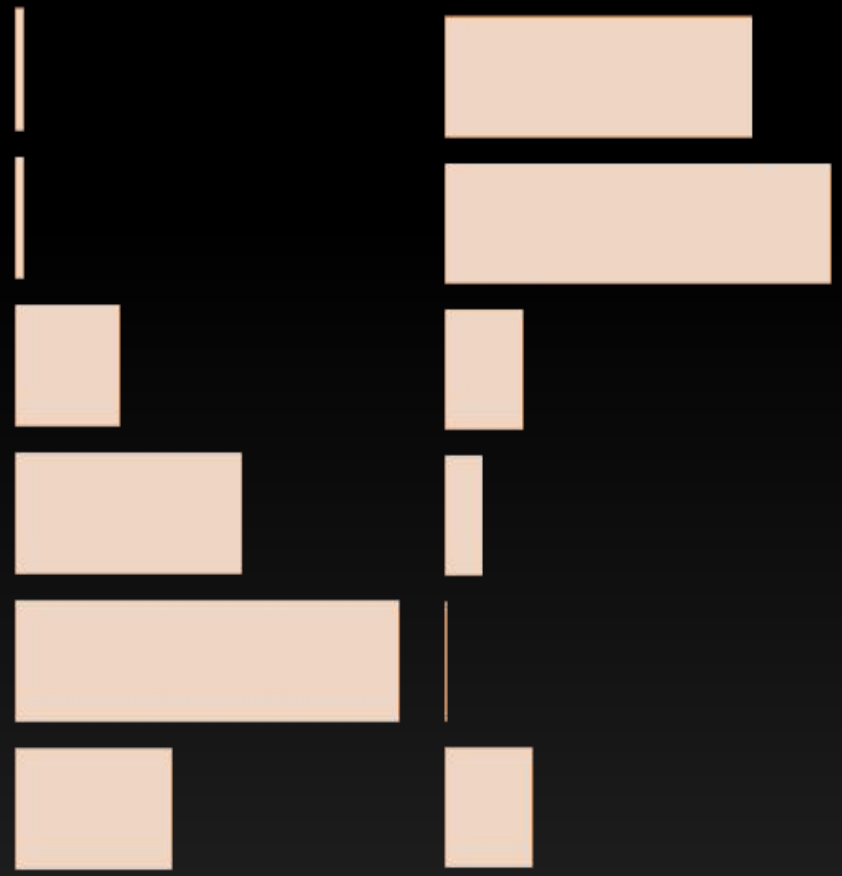
Step0 (Center = 0°C)

Sigmoid

Sigmoid0 (Center = 0°C)

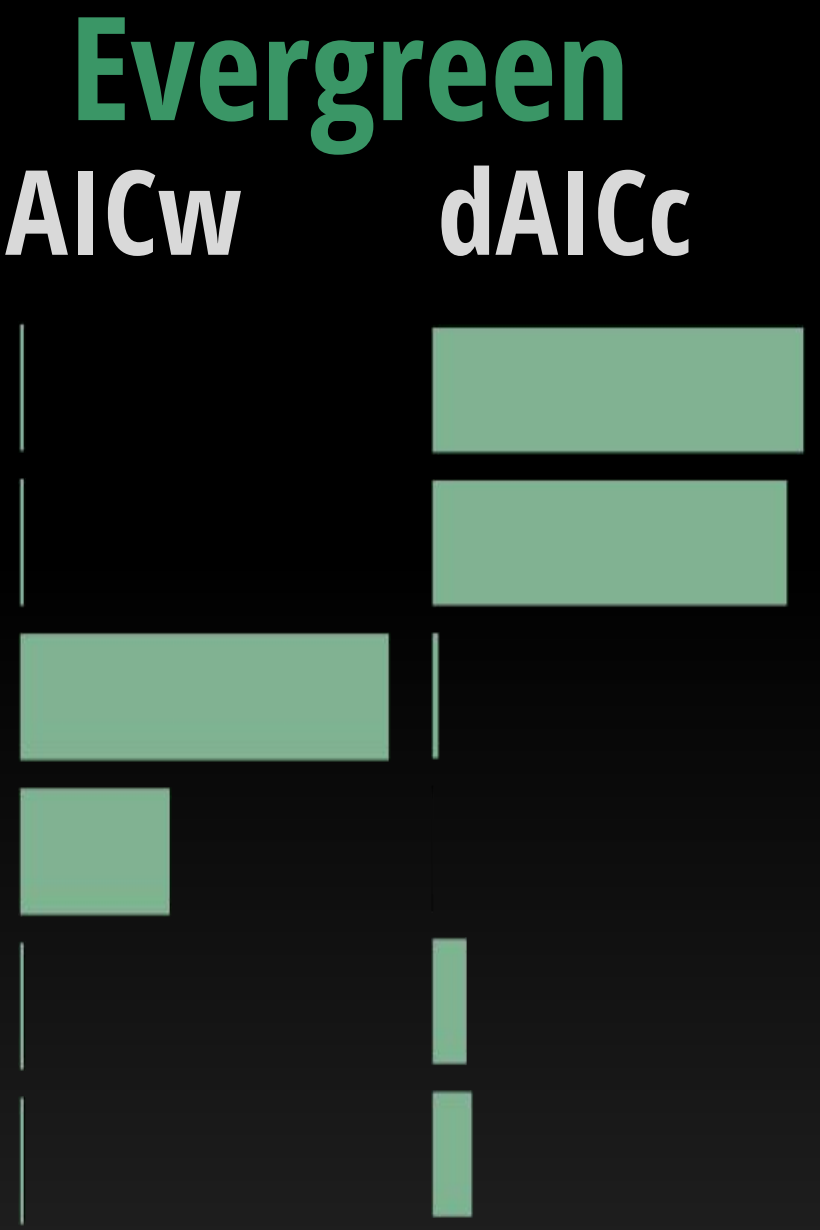
Linear

Linear + step



Functions

Evergreen



Step

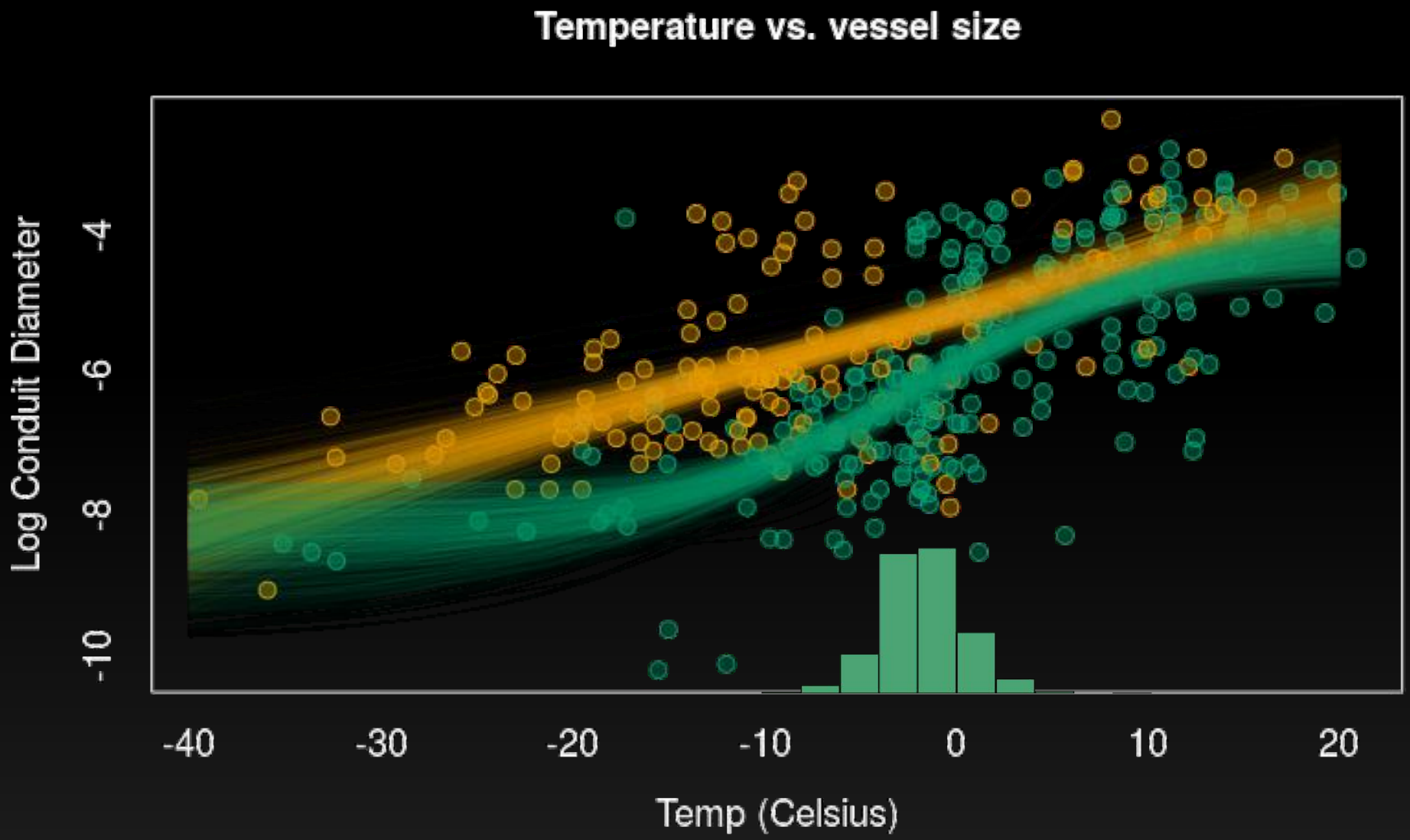
Step0 (Center = 0°C)

Sigmoid

Sigmoid0 (Center = 0°C)

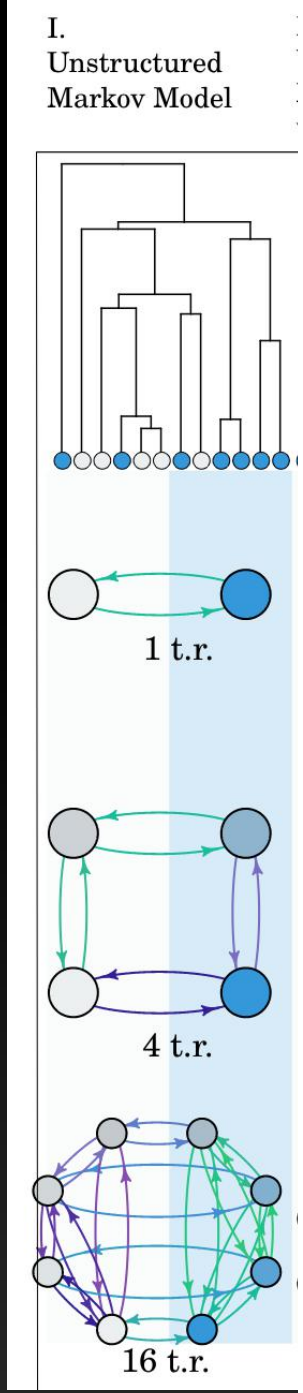
Linear

Linear + step


Temperature vs. vessel size
Log Conduit Diameter
-4
-6
-8
-10
-40
-30
-20
-10
0
10
20
Temp (Celsius)

Suppose you hypothesize an ecological factor (cold or warm) is associated with increased rates of evolution of a discretely measured trait (light and dark).

In reality, the trait is threshold trait. Therefore rate variation comes from two sources:
1. Being close to the threshold
and
2. The ecological factor



A. No Hidden States

B. Two Hidden States

C. Four Hidden States

nature ecology & evolution

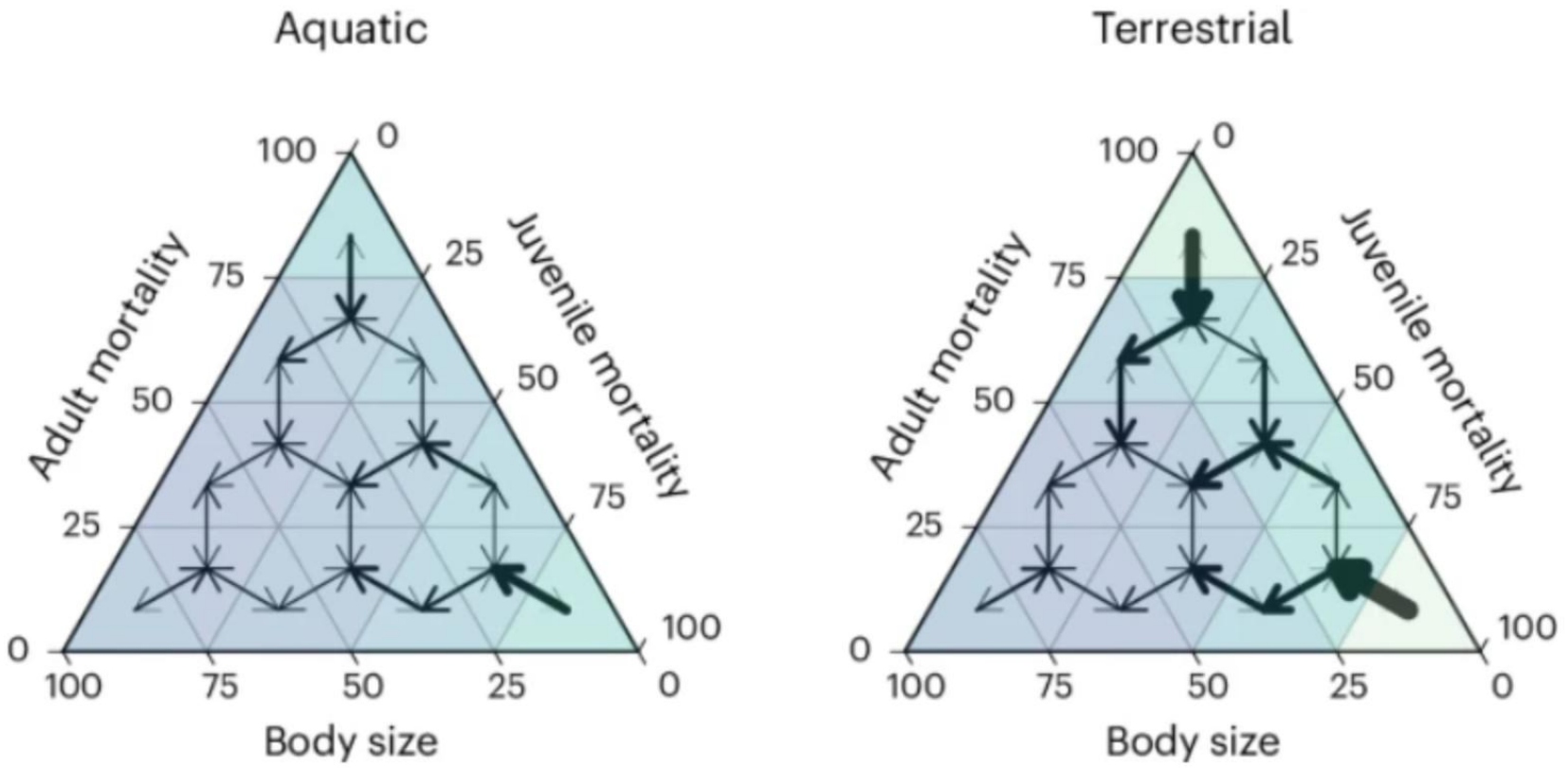
Article | Published: 12 March 2025

Fundamental constraints on vertebrate life history are shaped by aquatic–terrestria reproductive mode

George C. Brooks ✉, Josef C. Uyeda, Nicholas J. Bone, H Kindsvater

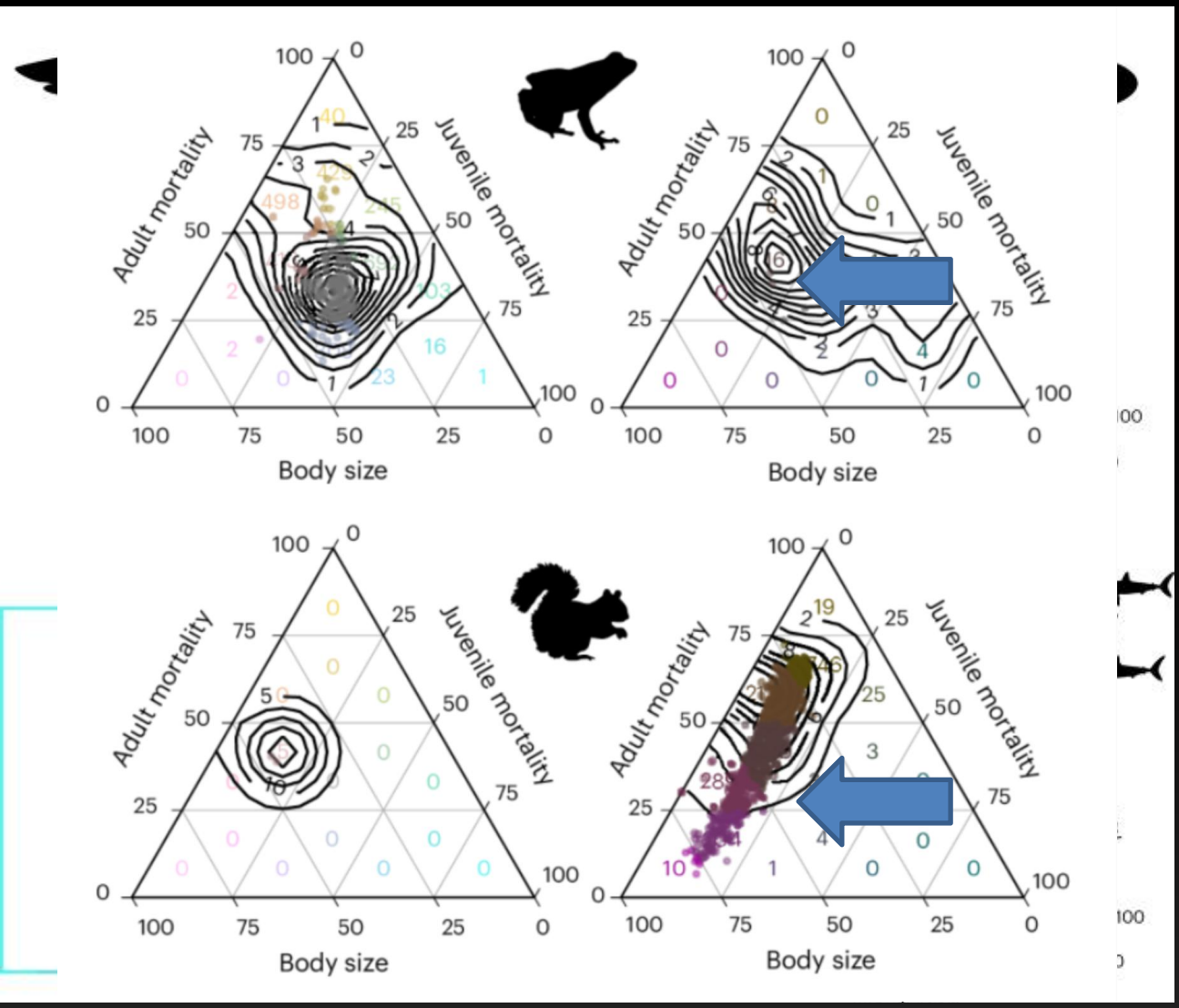
Nature Ecology & Evolution **9**, 857–866 (2025) | Cite this

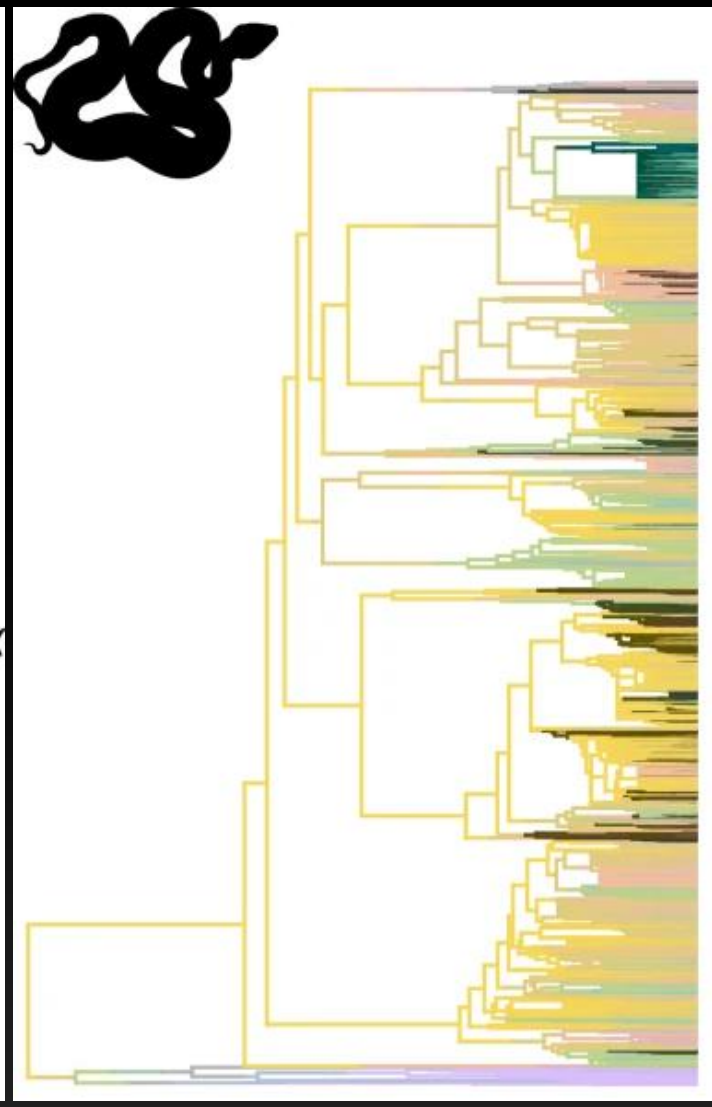
Fig. 2: Movement through life-history space for aquatic and terrestrial vertebrates.

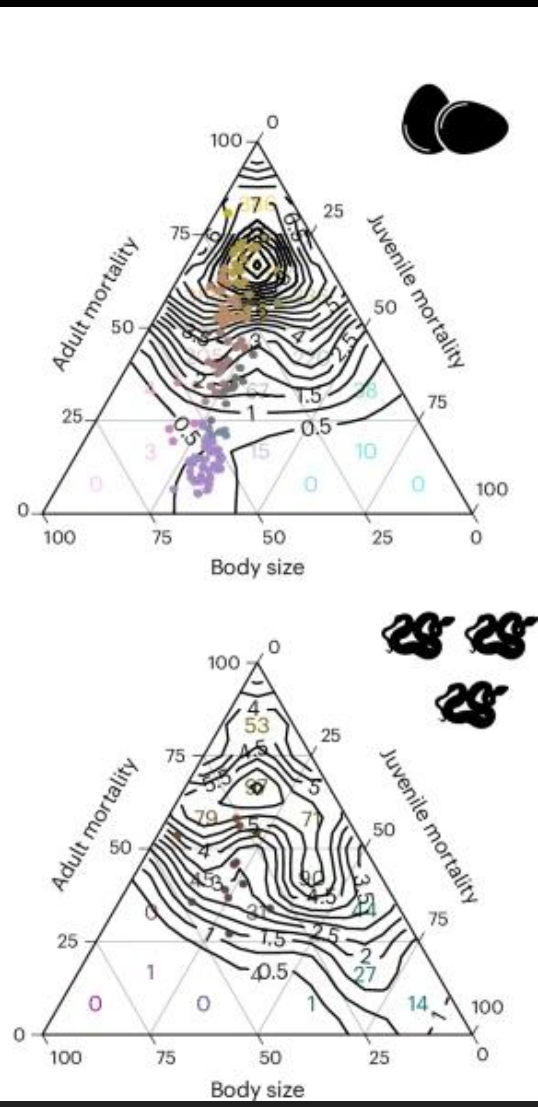


Reconstructing ancestral states (handles ambiguous states!)

?







Summary

Threshold models bridge continuous and discrete models, reflecting potentially realistic G-P maps

Hidden state models can also represent structured hypotheses about the underlying state space

Discretization of continuous data can enable powerful tools for studying macroevolutionary landscapes

While structured models are special cases of generalized Markov models, they avoid blowing up the number of parameters, and can be informed by *other things we know about biology*.

You can change any parameter of the model (or the model itself) at the shift

Brownian Motion - σ^2

Brownian Motion w/trend - σ^2, μ

OU - θ, σ^2, α

EB - σ^2, b

Example questions:

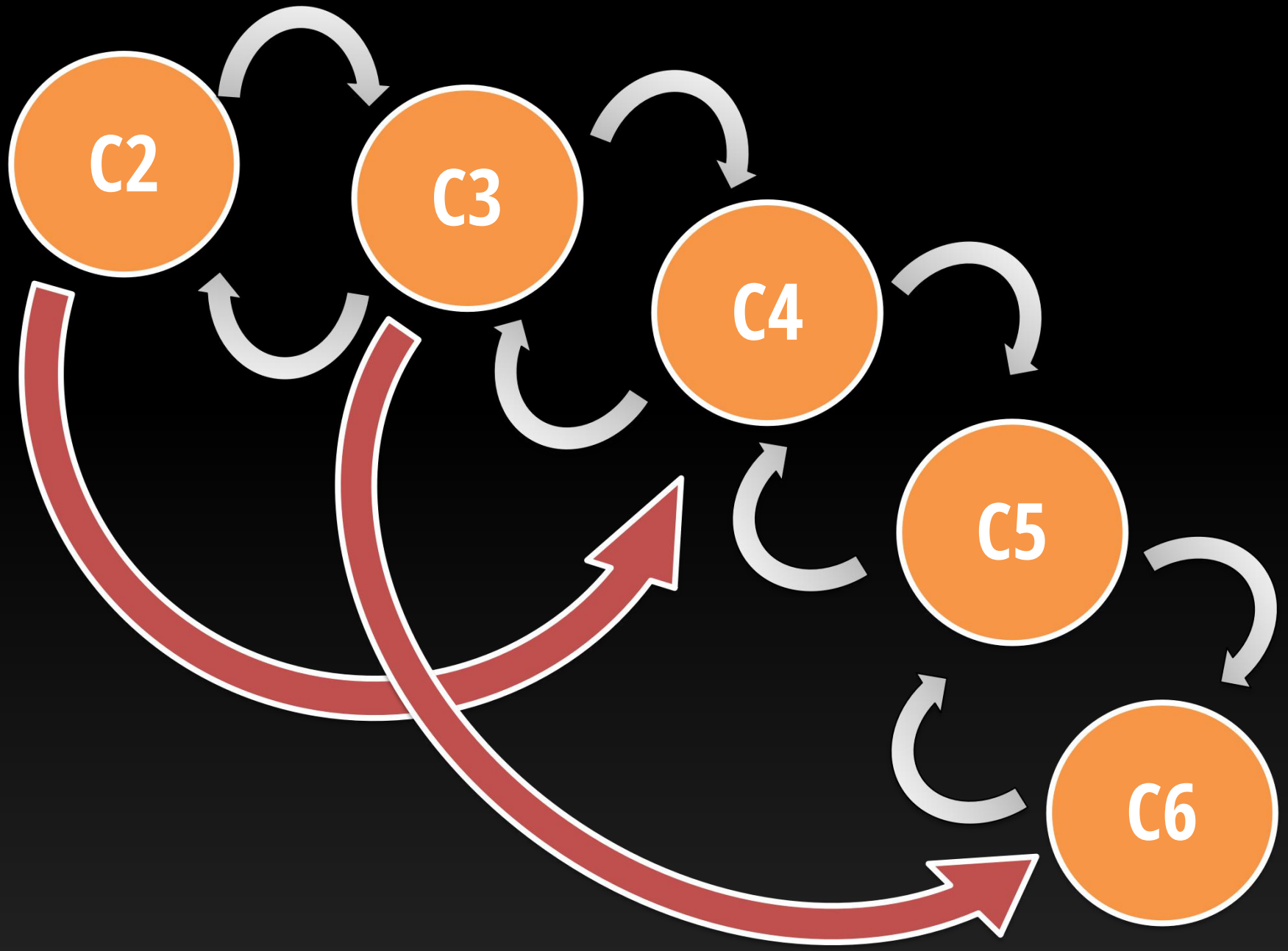
Does the rate of evolution change with habitat?

Does niche space expand on islands?

Do ectotherms have more constrained trait evolution than endotherms?

Is gene expression more constant early or late in cancer progression?

Are Anolis lizard ecomorphs convergent?


C2
C3
C4
C5
C6