Incorporating Heterogenous Data Sources in Phylogenetic Modeling

April Wright

LBRN Annual Meeting

January 18

SOUTHEASTERN

LOUISIANA UNIVERSITY

LBRN

Big Data

Mazberry.com

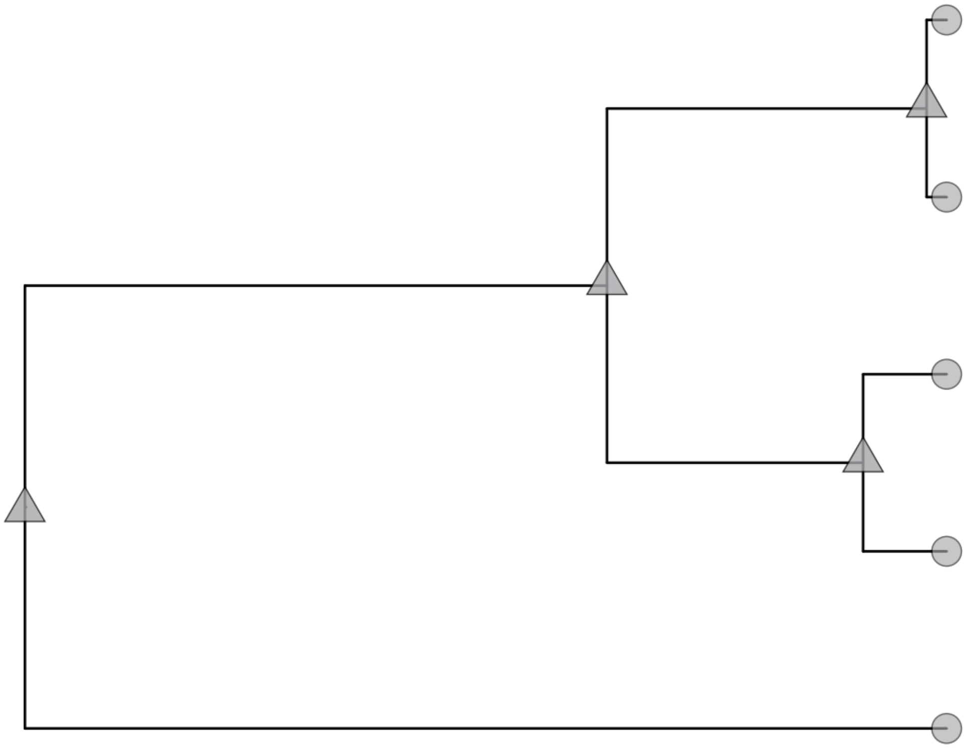
Datanami.com

Big Data

Mzberry.com

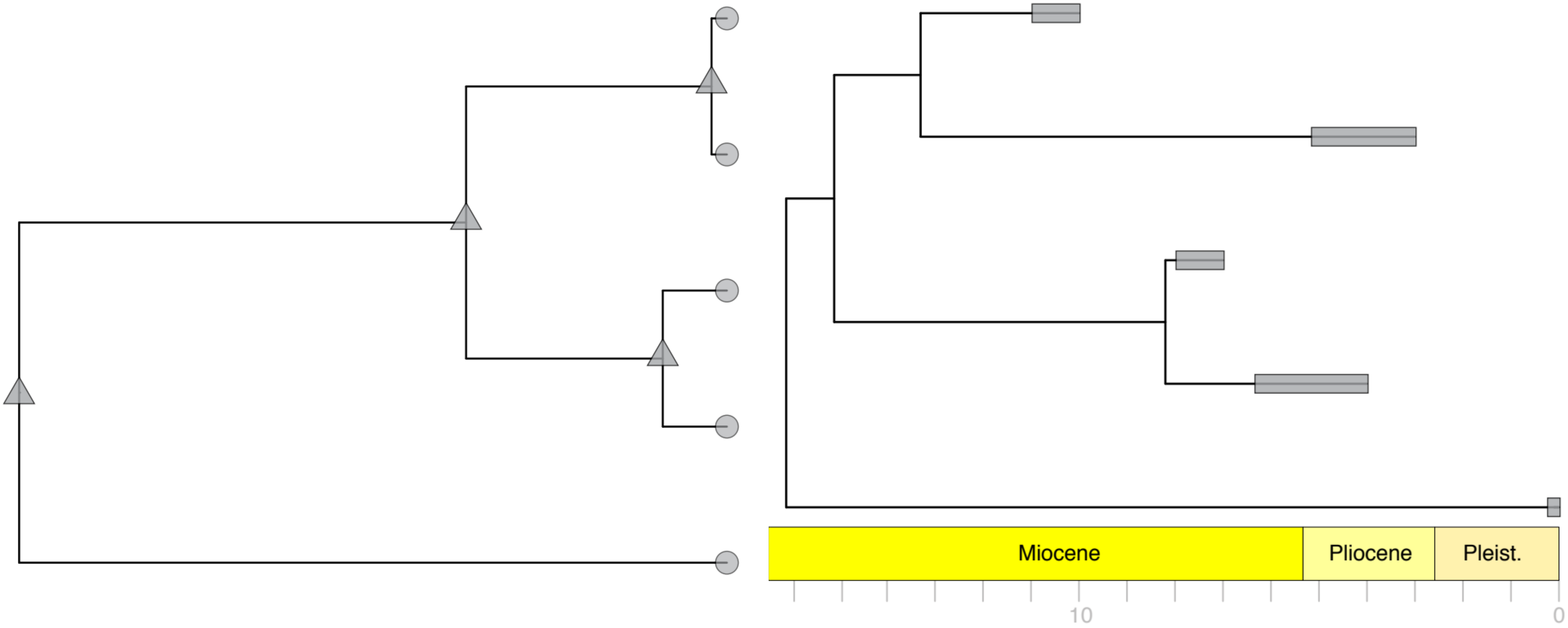
Dtnmi.com

Heterogeneous data = data coming from multiple sources, each with their own generating and collection process



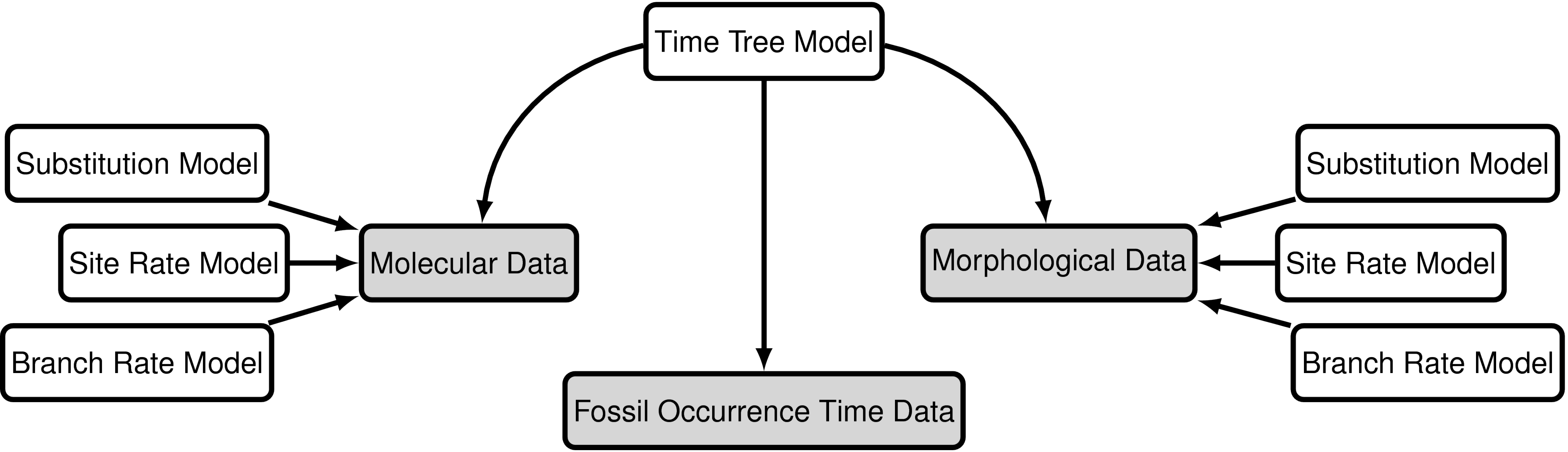
Estimating phylogenetic trees usually involves modeling the evolution of nucleotide sequences over time

Wright and Warnock, in review

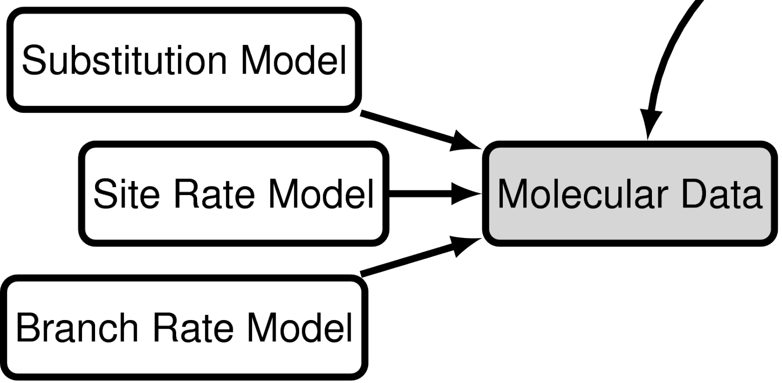


Scaling phylogeny to time involves adding in age information, and often phenotypic information

Wright and Warnock, in review

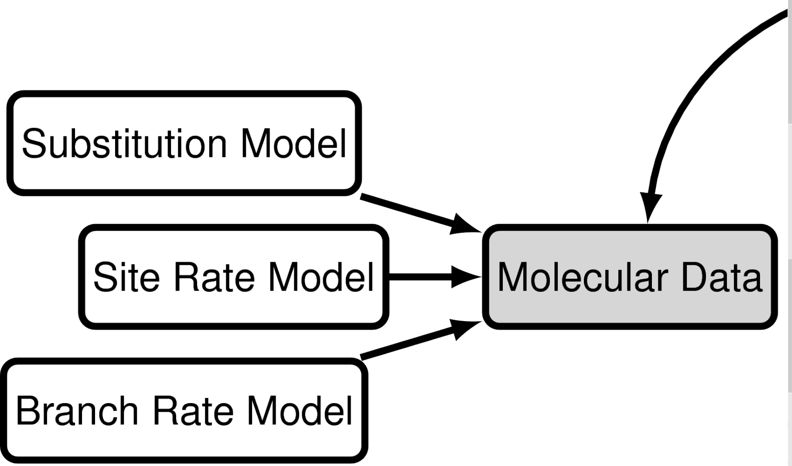


Wright, Pett
and Heath

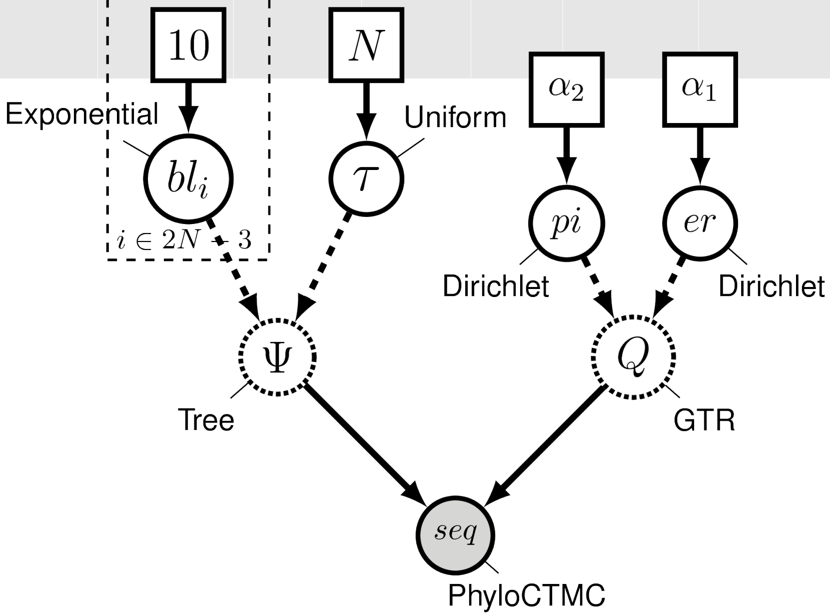


Taxon 1	A	C	T	A	C	T	C	G	
Taxon 2	A	C	T	A	A	T	G	T	C
Taxon 3	A	T	T	A	C	T	G	T	G
Taxon 4	G	G	A	A	C	T	G	G	T
Taxon 5	G	G	C	T	C	T	G	A	A

Wright, Pett
and Heath

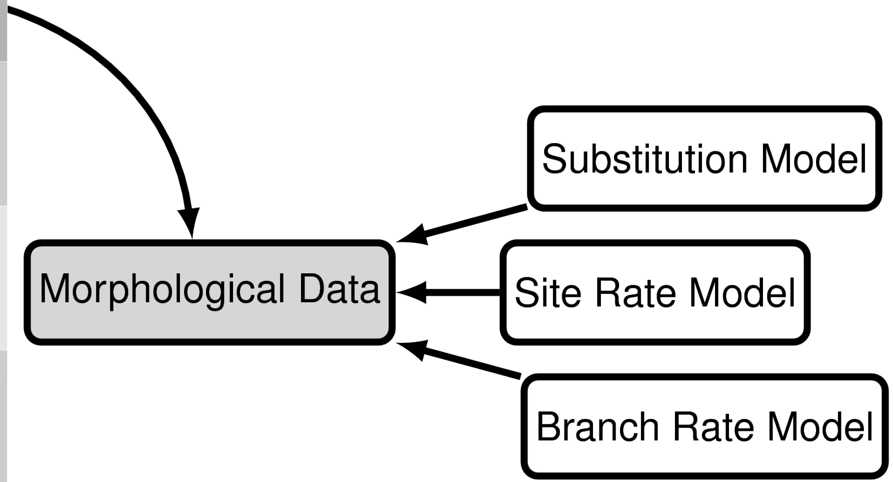


Taxon 1	A	C	T	A	C	T	C	G	
Taxon 2	A	C	T	A	A	T	G	T	C
Taxon 3	A	T	T	A	C	T	G	T	G
Taxon 4	G	G	A	A	C	T	G	G	T
Taxon 5	G	G	C	T	C	T	G	A	A



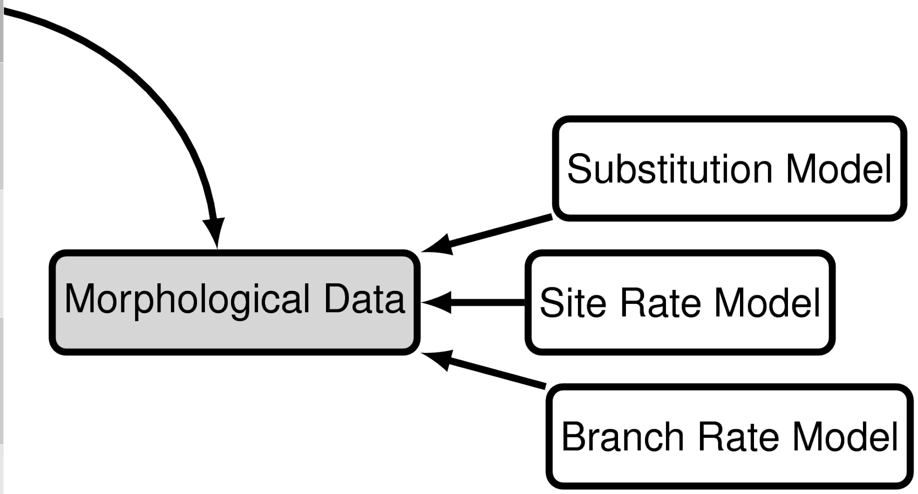
Wright, Pett and Heath

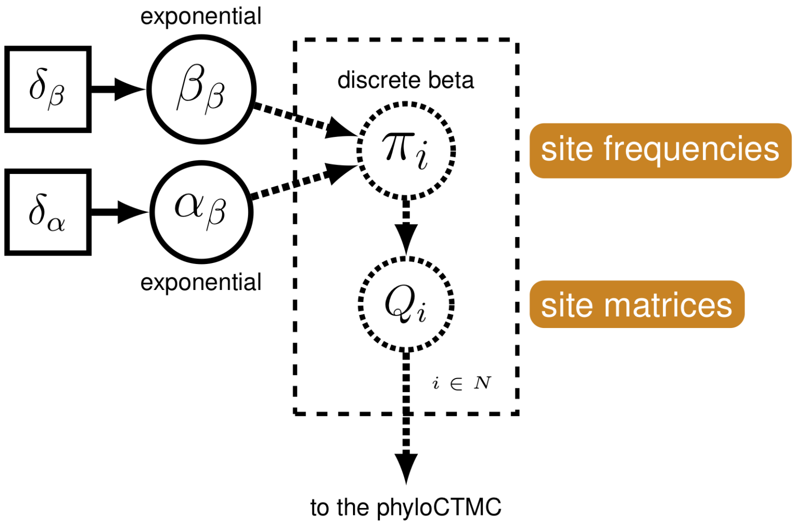
Taxon 1	0	0	0	0	0	0	0	0	0
Taxon 2	0	2	0	2	1	1	0	0	0
Taxon 3	1	2	0	2	3	0	0	1	1
Taxon 4	1	1	1	1	3	1	0	1	1
Taxon 5	1	1	0	1	2	1	1	1	0



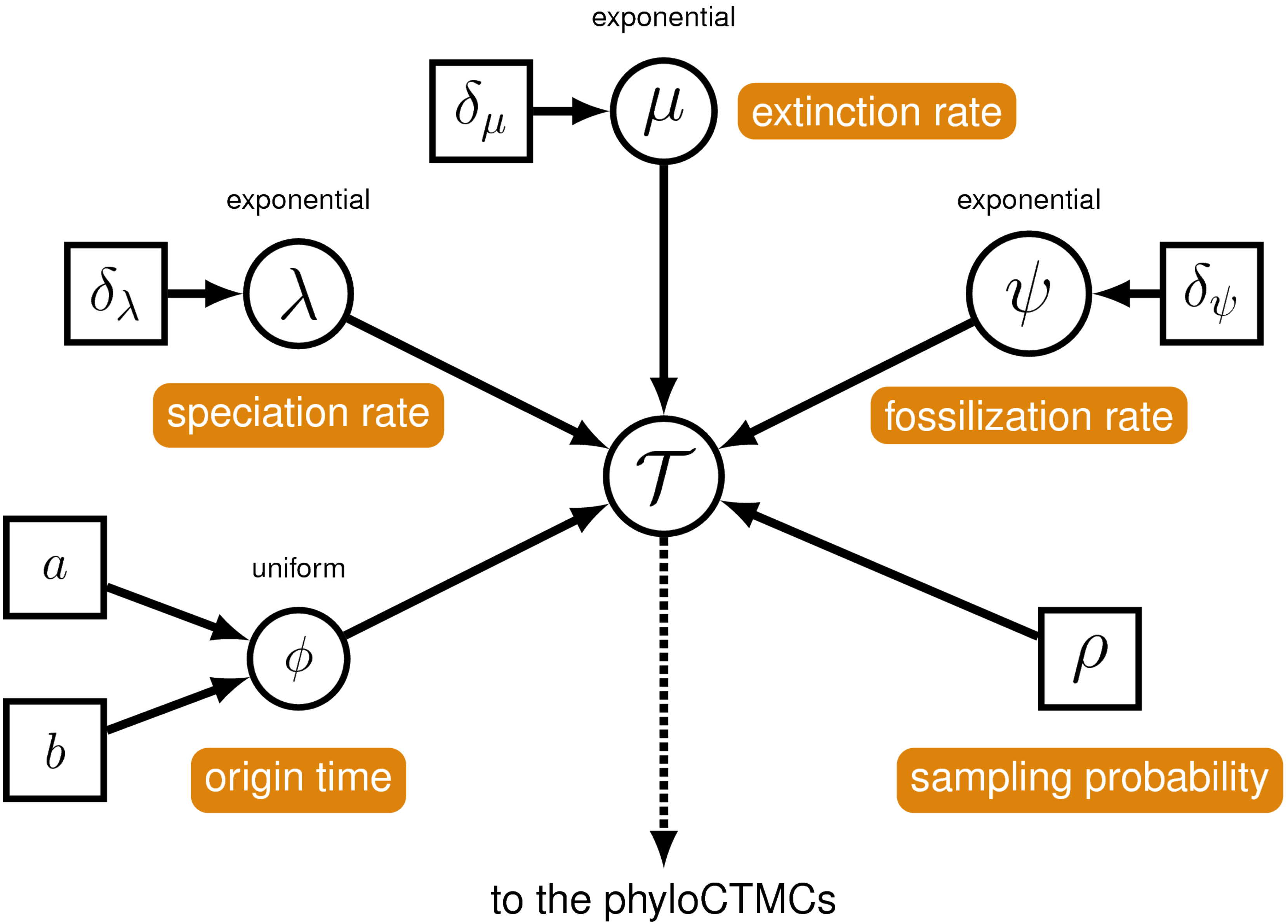
Wright, Pett and Heath

Taxon 1	0	0	0	0	0	0	0	0	0
Taxon 2	0	2	0	2	1	1	0	0	0
Taxon 3	1	2	0	2	3	0	0	1	1
Taxon 4	1	1	1	1	3	1	0	1	1
Taxon 5	1	1	0	1	2	1	1	1	0

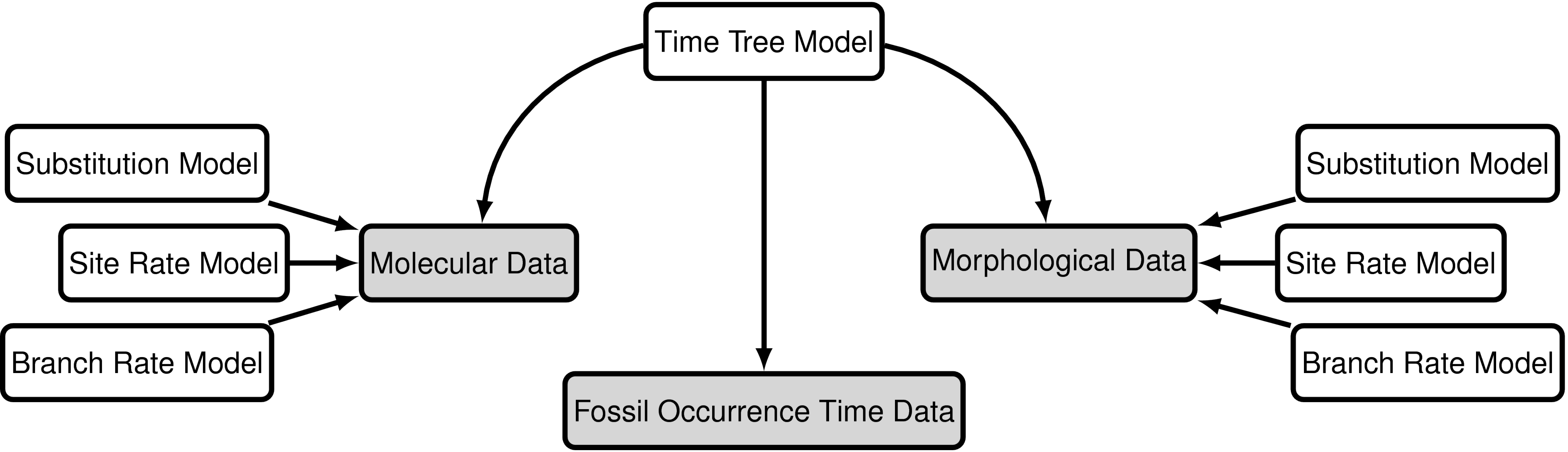




Wright, Pett and Heath

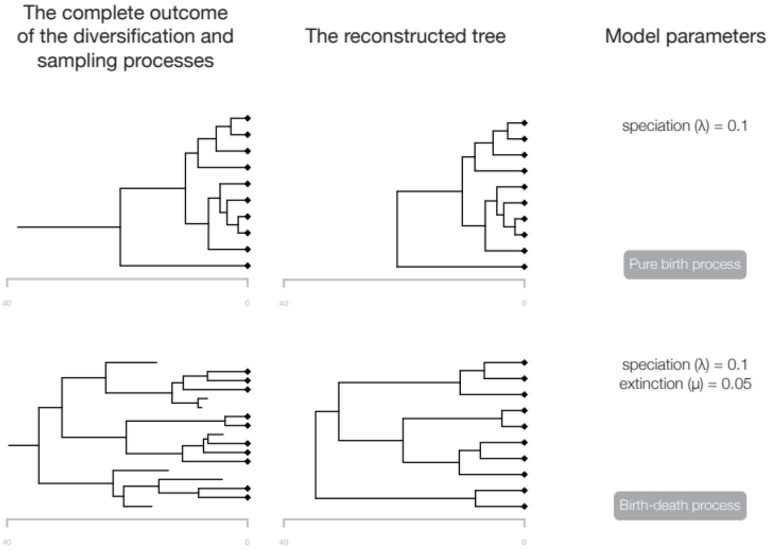


Wright, Pett
and Heath

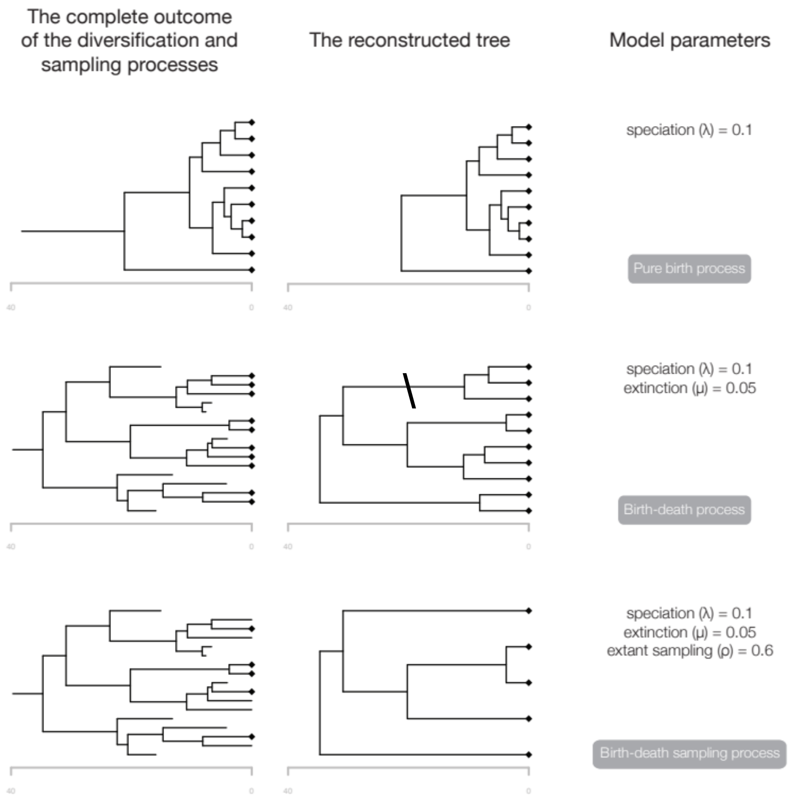


Wright, Pett
and Heath

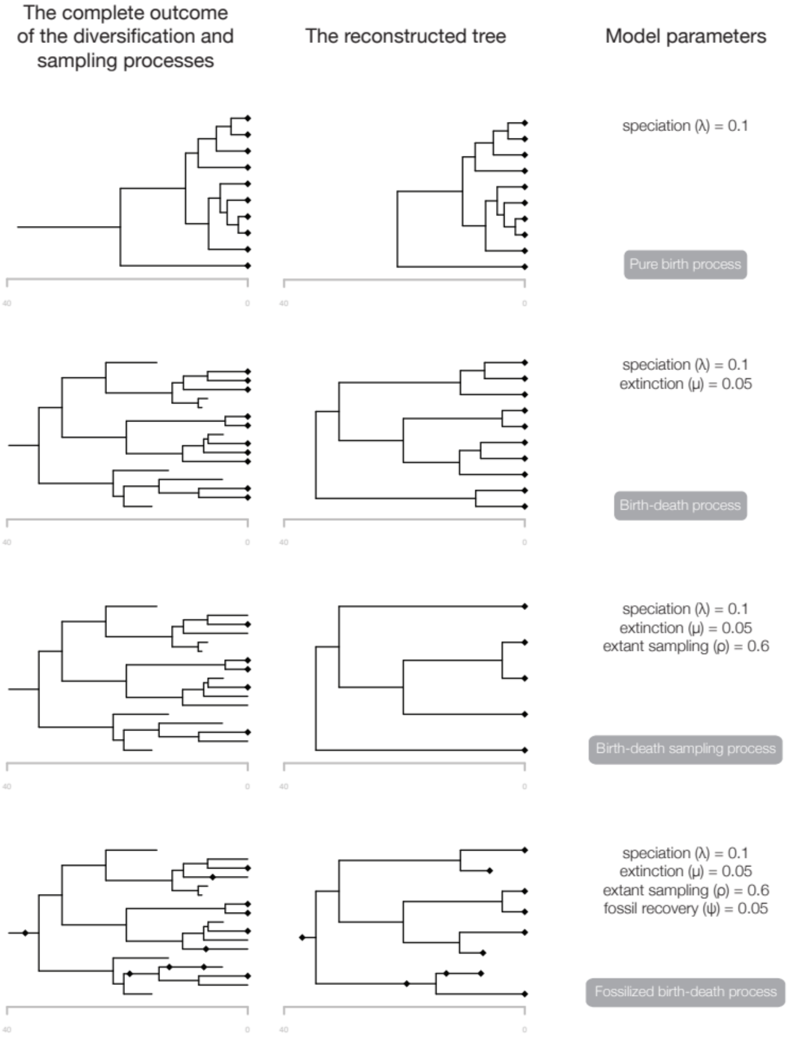
Wright and Warnock, in review



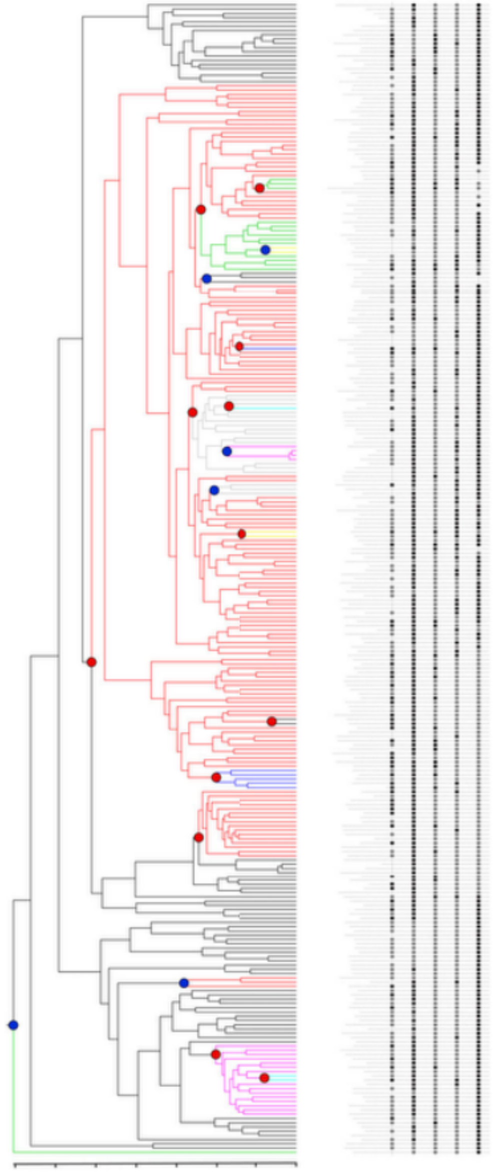
Wright and Warnock, in review



Wright and Warnock, in review



Wright and Warnock, in review

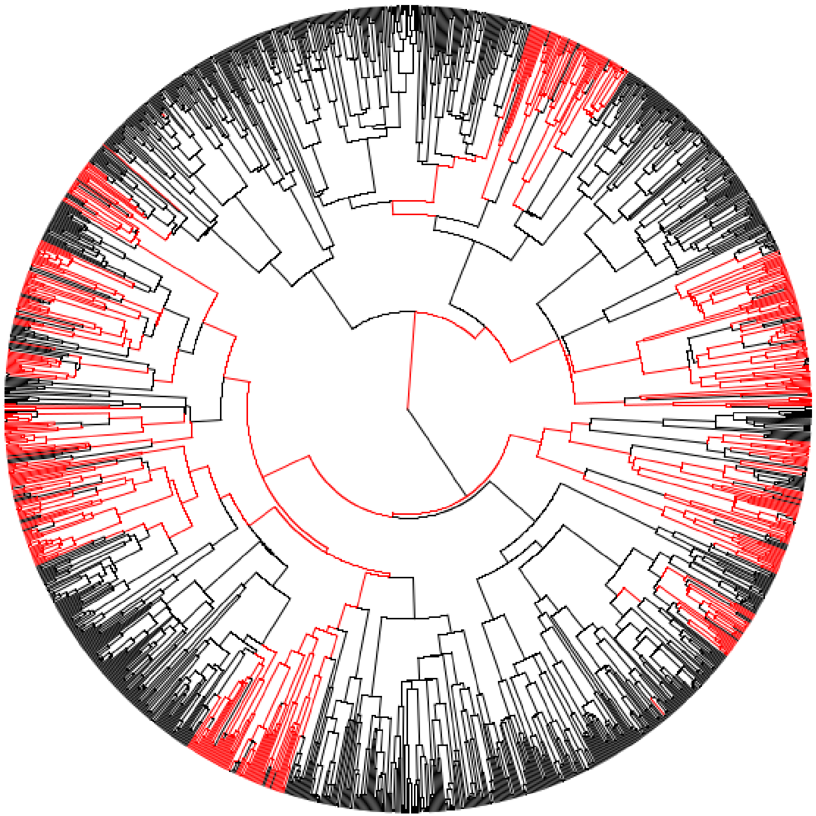


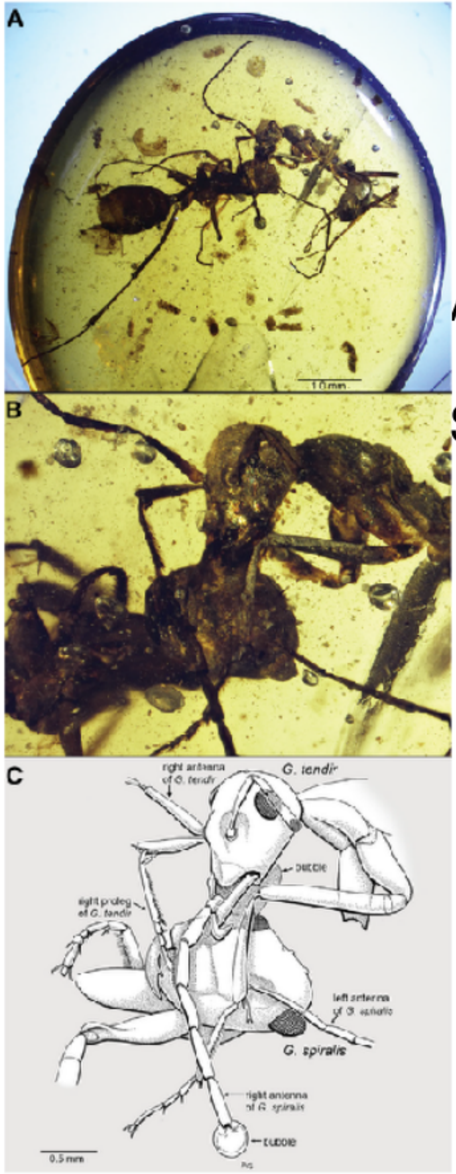
Why Ants?

Abundant molecular resources

666-tip multi-gene phylogeny from Blanchard and Moreau (2017)

Why Ants?



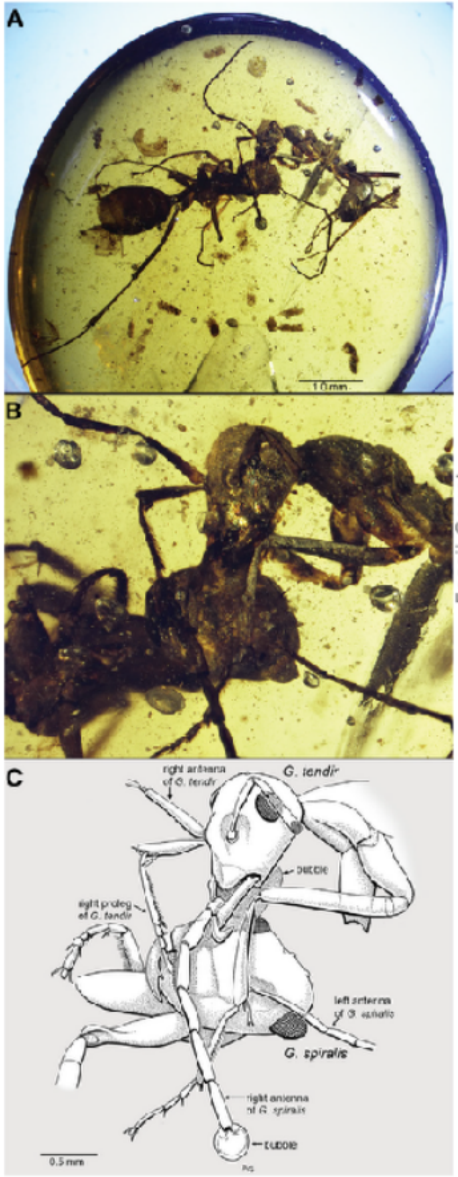


Why Ants?

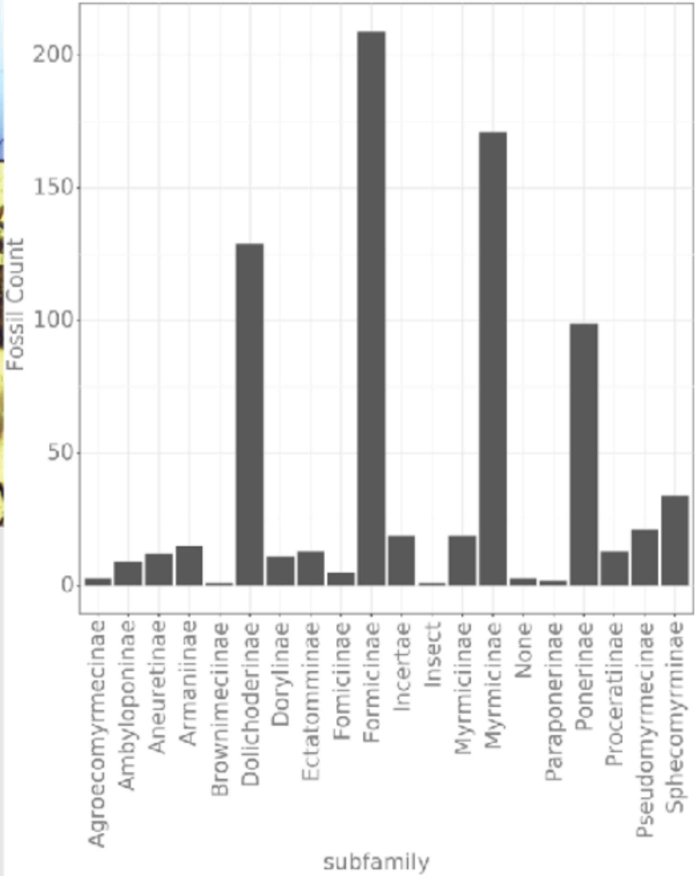
Abundant morphological resources

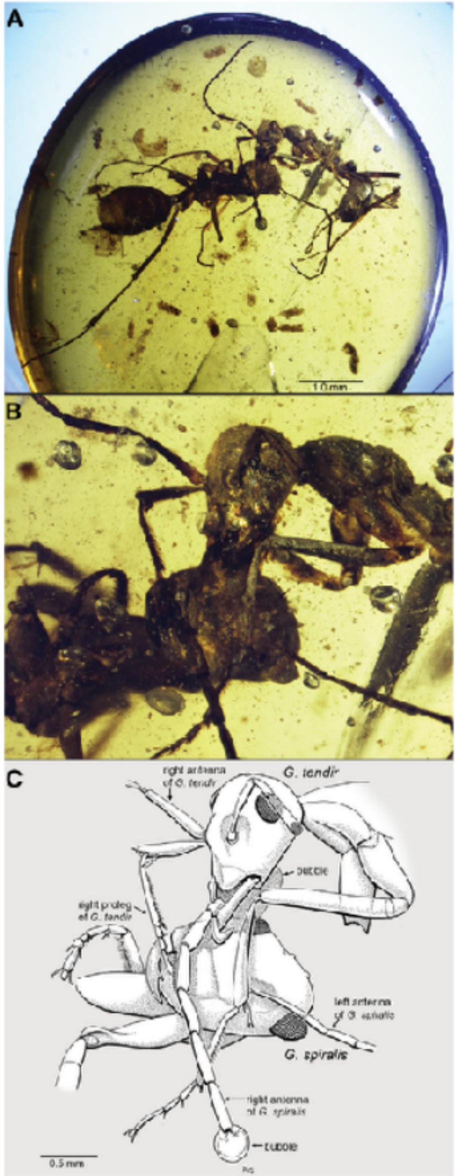
Samples both with character data
and without character data

Barden & Grimaldi, 2017


A
10 mm
B
C
right antenna of G. tendir
G. tendir
bubble
right proleg of G. tendir
left antenna of G. spiralis
G. spiralis
right antenna of G. spiralis
bubble
0.5 mm

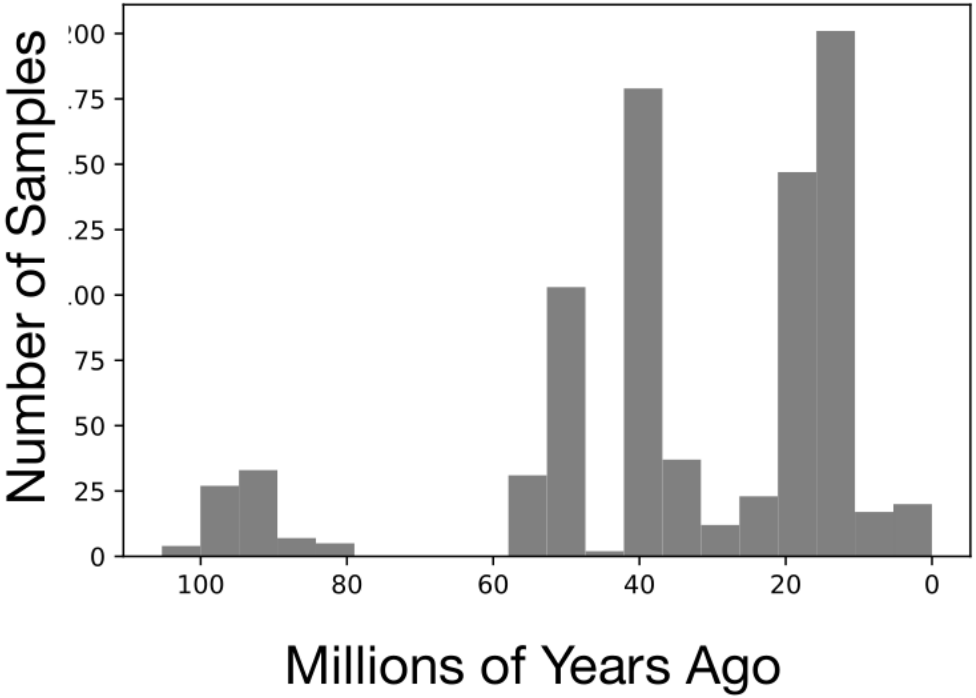
Why Ants?


Fossil Count
200
150
100
50
0
Agroecomyrmecinae
Ambyloponinae
Aneuretinae
Armaniinae
Brownimeciinae
Dolichoderinae
Dorylinae
Ectatomminae
Fomiciinae
Formicinae
Incertae
Insect
Myrmiciinae
Myrmicinae
None
Paraponerinae
Ponerinae
Proceratiinae
Pseudomyrmecinae
Sphecomyrminae
subfamily



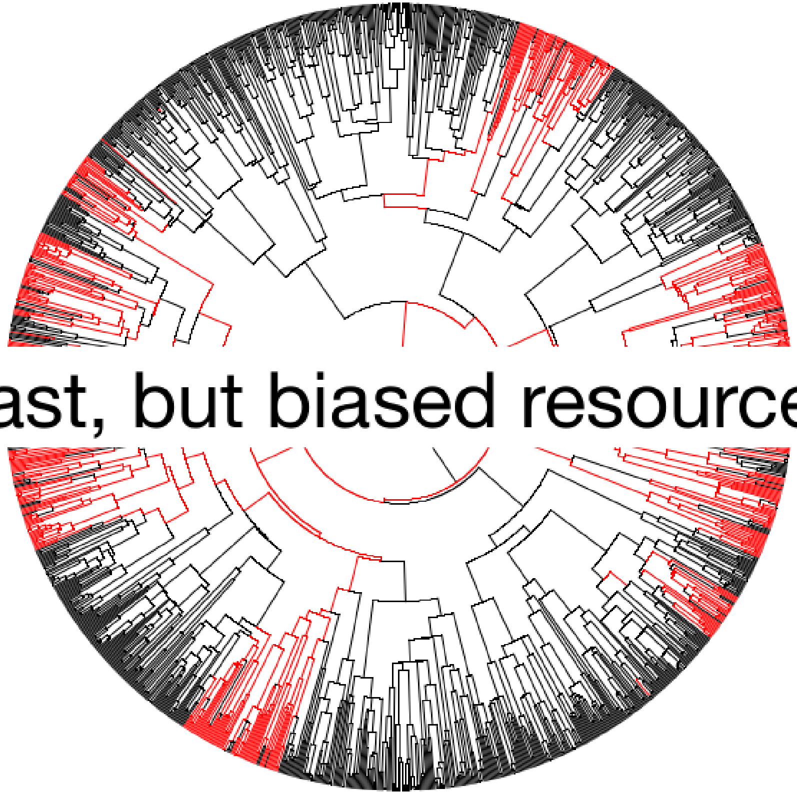
A
B
C
G. tendir
G. spiralis

Why Ants?

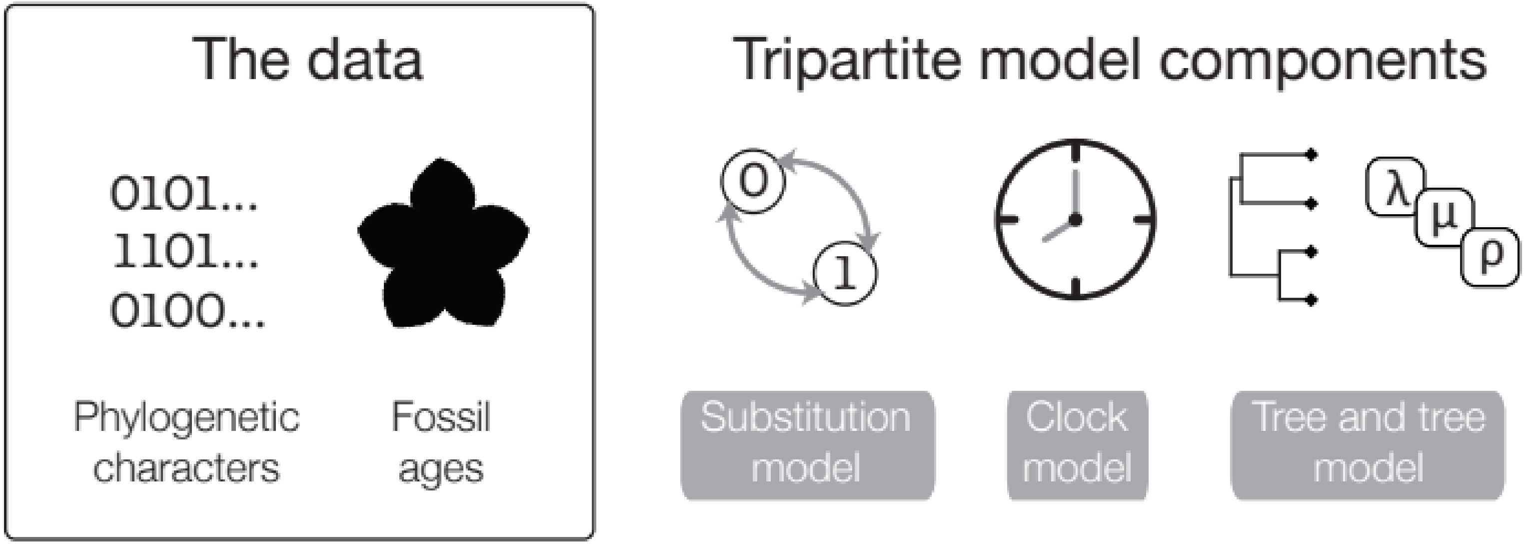


Number of Samples
Millions of Years Ago
100
80
60
40
20
0
25
50
75
.00
.25
.50
.75

Why Ants?



Vast, but biased resources



Wright and Warnock, in review

Systematic Conflict

Molecular evidence supports each of these

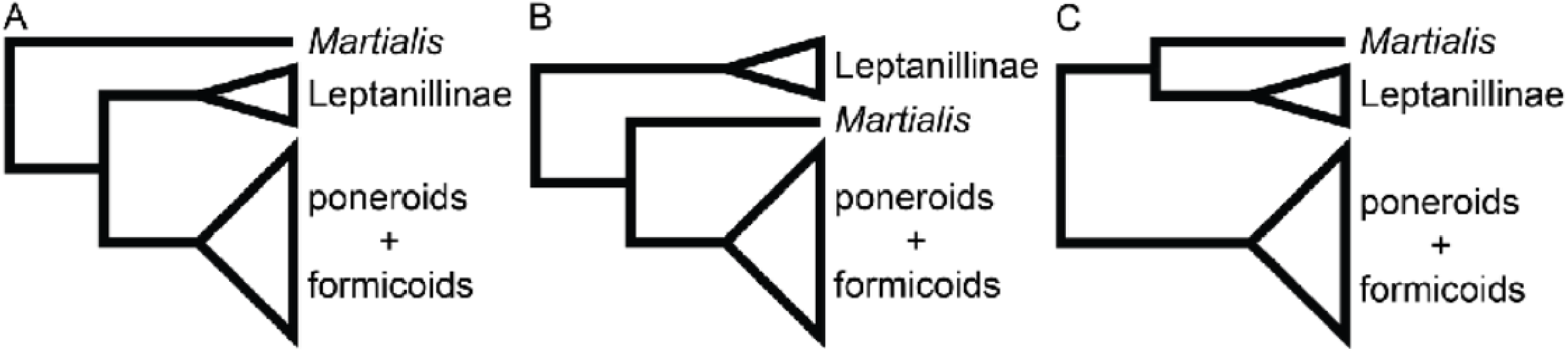
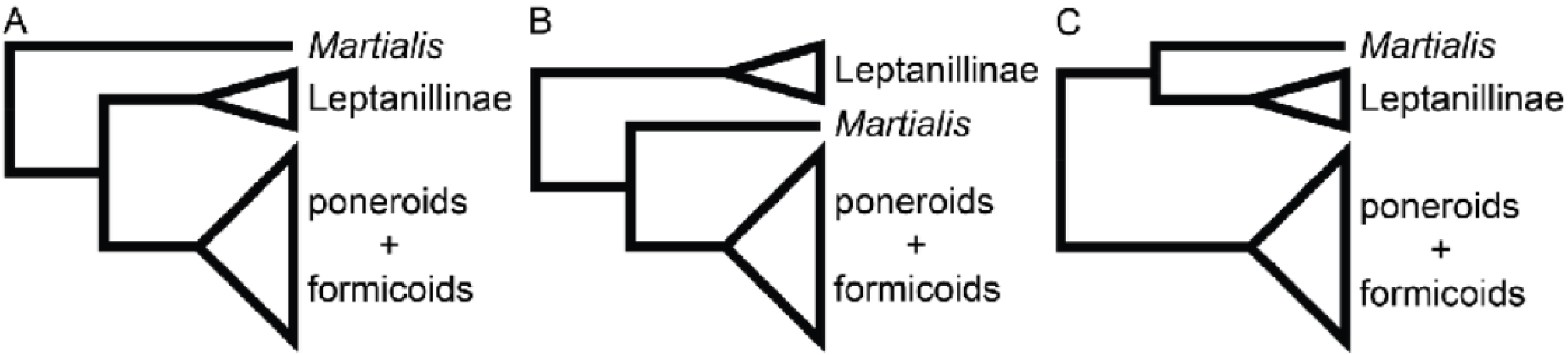


Figure adapted from Borowiec et al 2019

Systematic Conflict

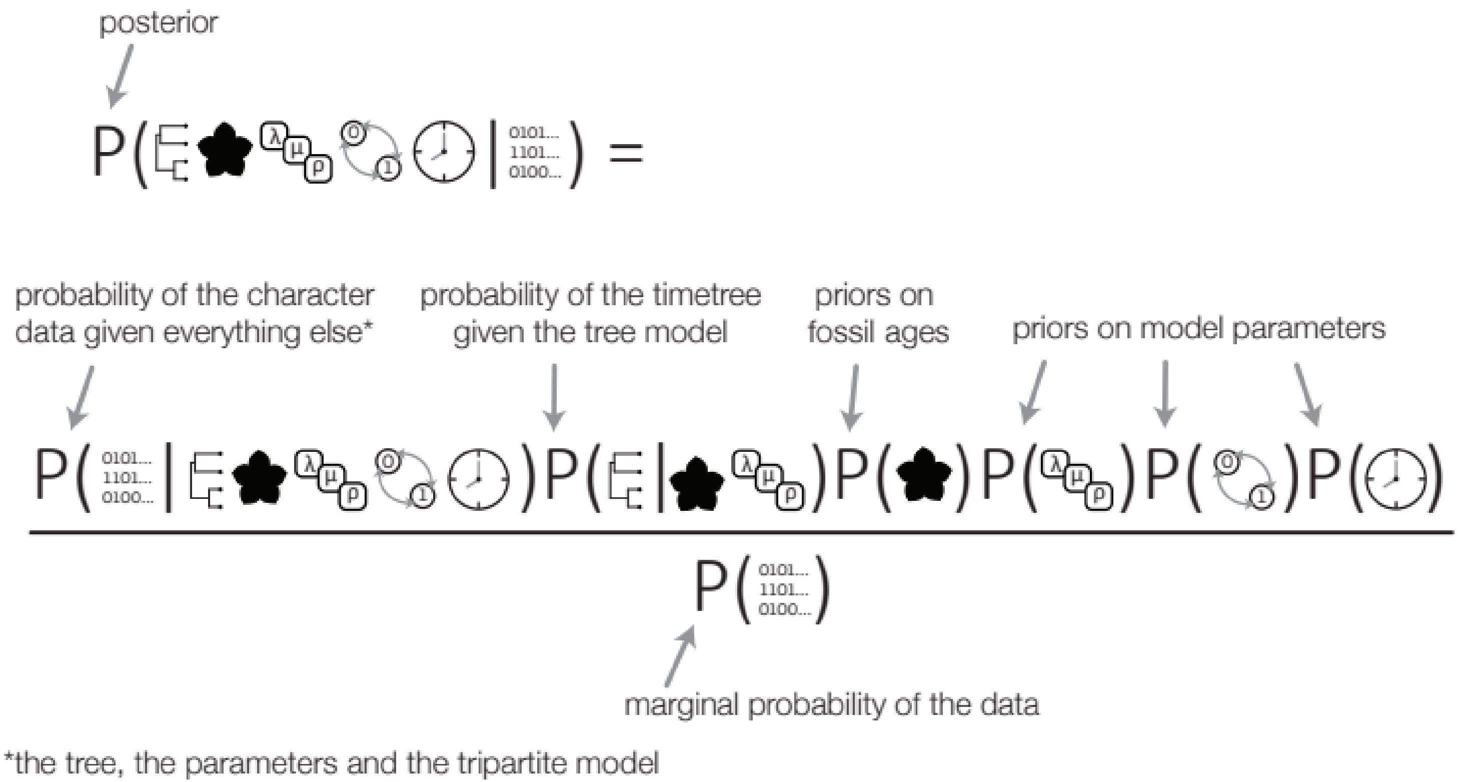
Molecular evidence supports each of these



Morphology supports none of them

Figure adapted from Borowiec et al 2019

Putting everything together



Wright and Warnock, in review

Depending on the assumptions made about evolution, we recover support for every one of these topologies

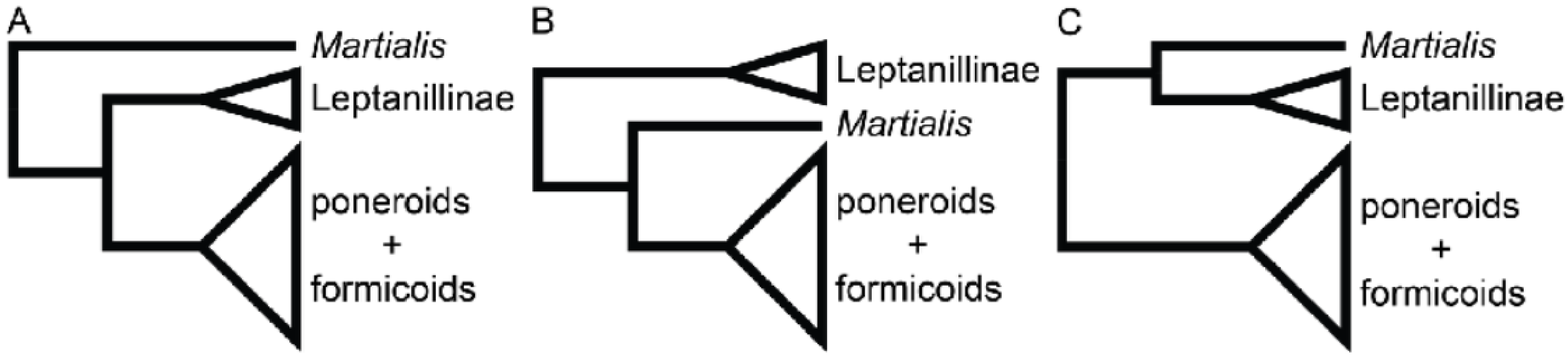
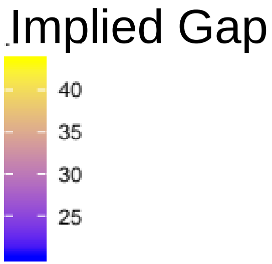


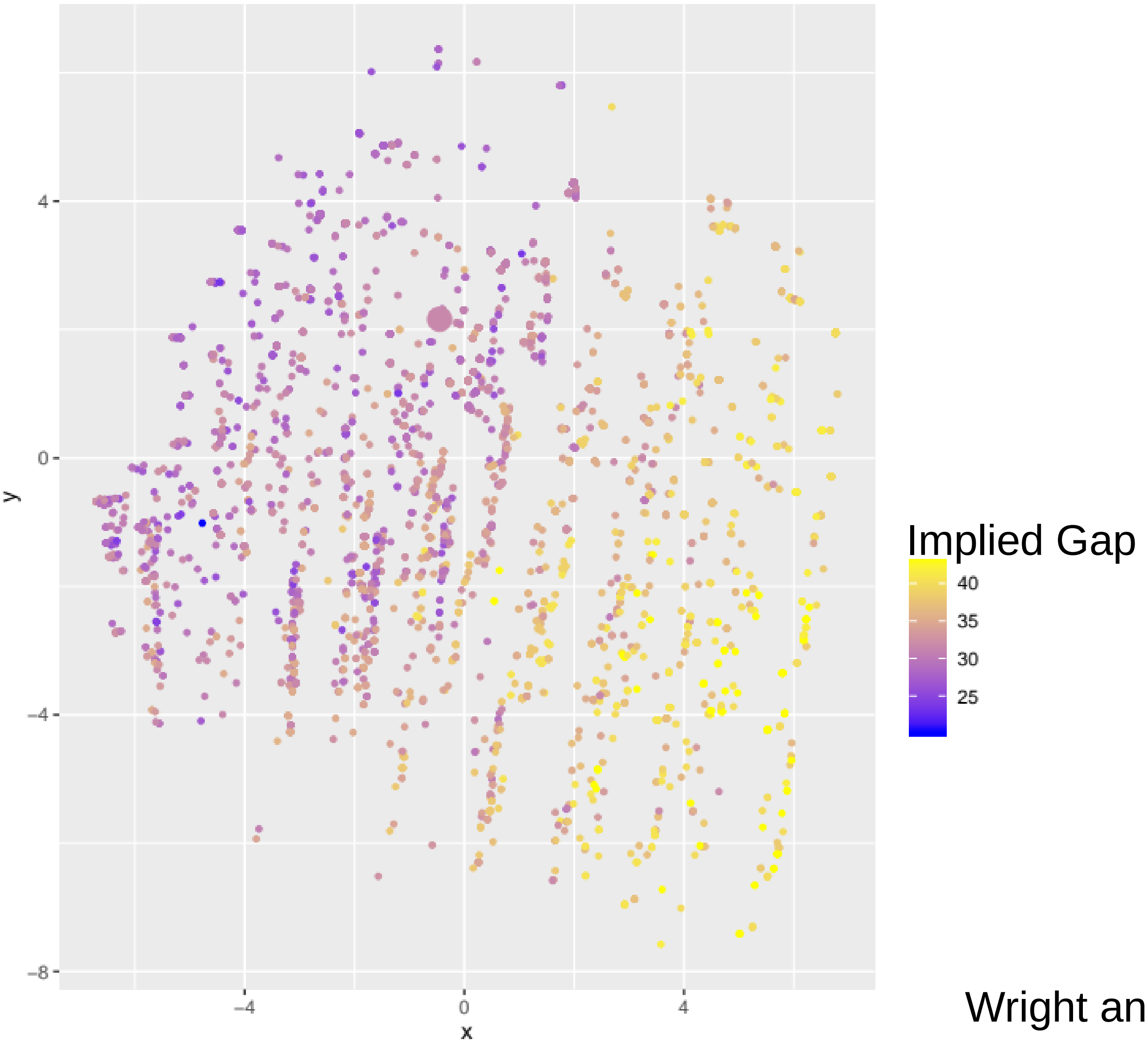
Figure adapted from Borowiec et al 2019


SOUTH BAY
SB
© "Mike" Michael L. Baird flickr.bairdphotos.com

If every hypothesis has some support, how do we know which is true?

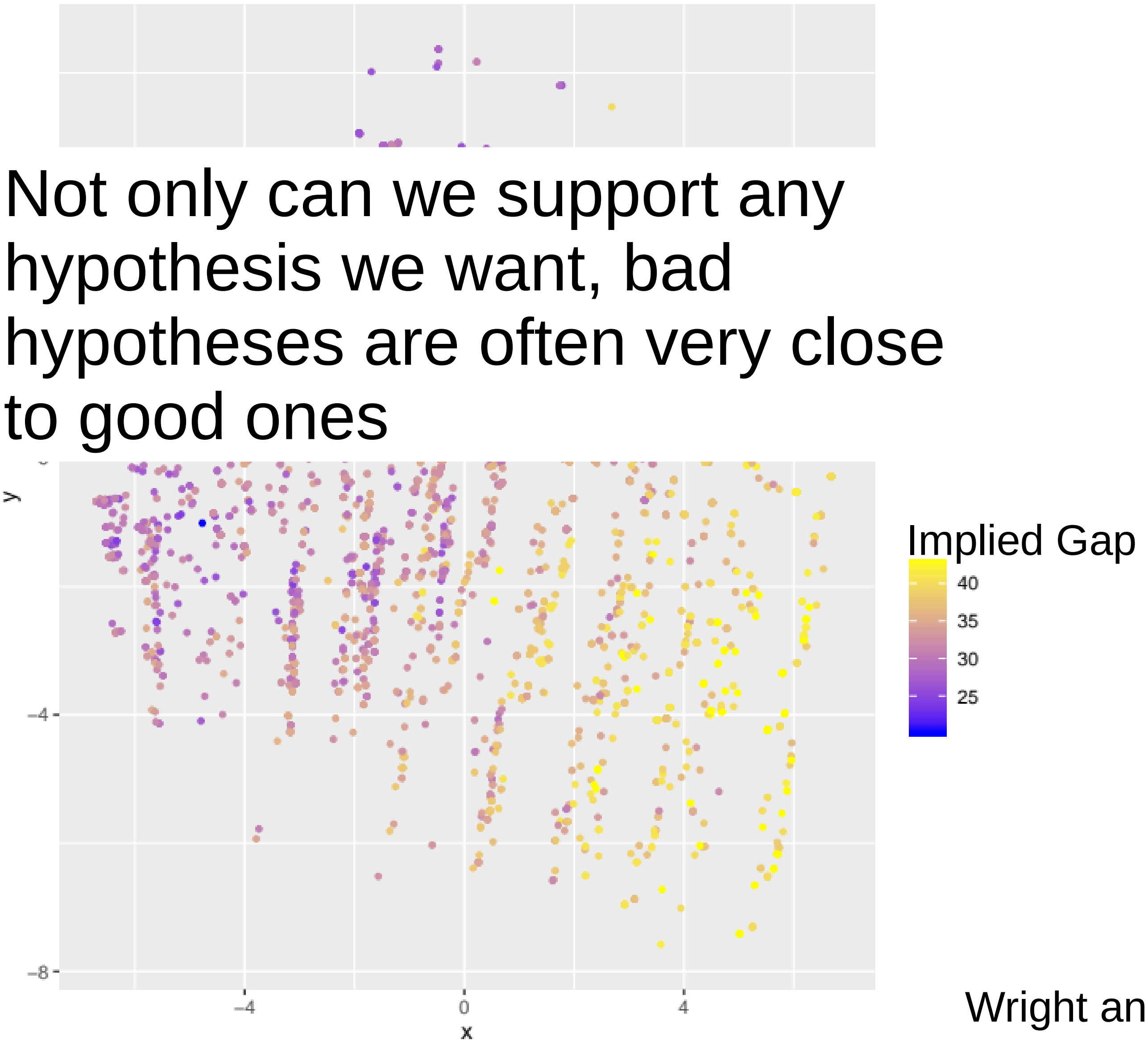


Wright and Lloyd, in review



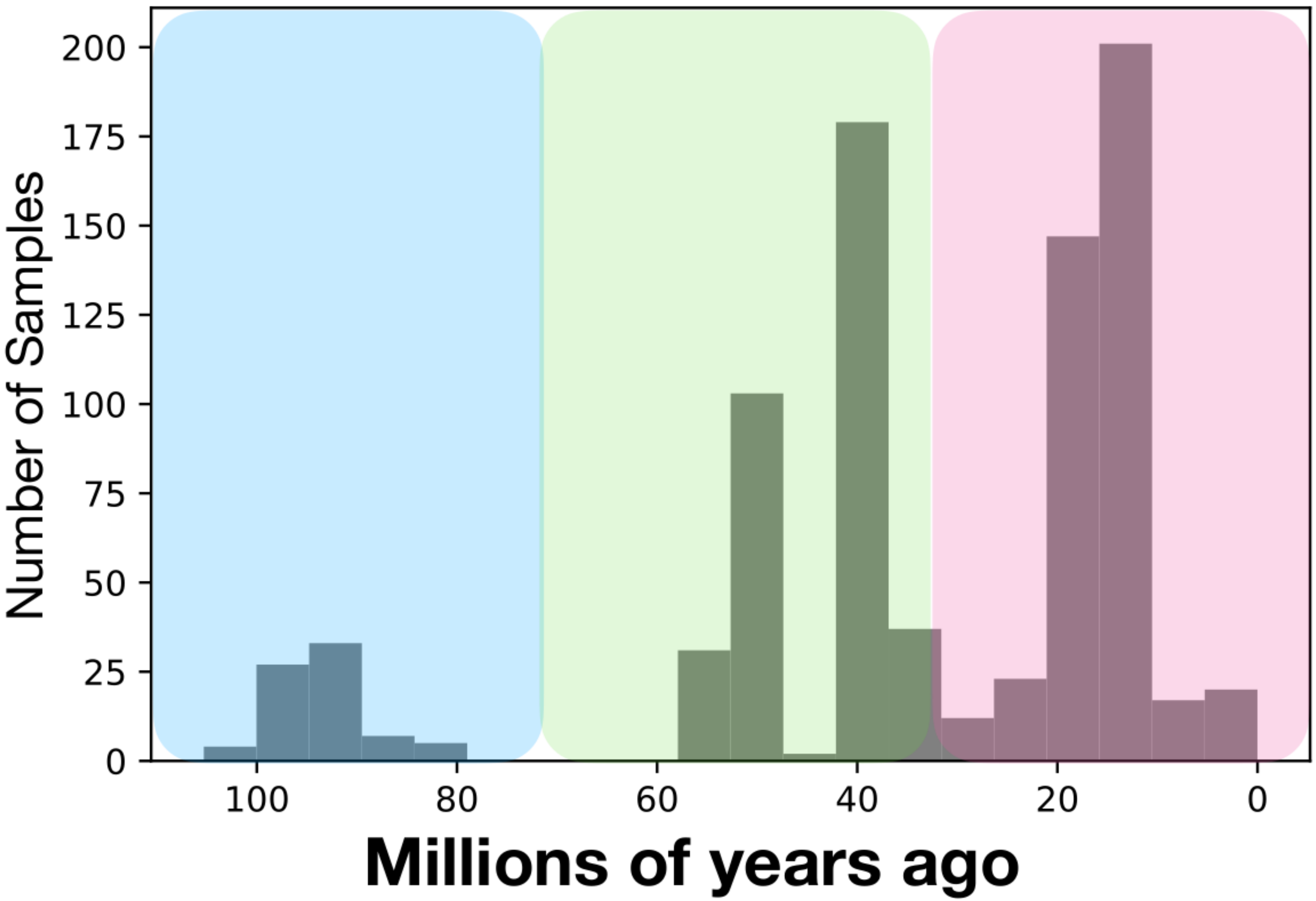
Wright and Lloyd, in review

Not only can we support any hypothesis we want, bad hypotheses are often very close to good ones

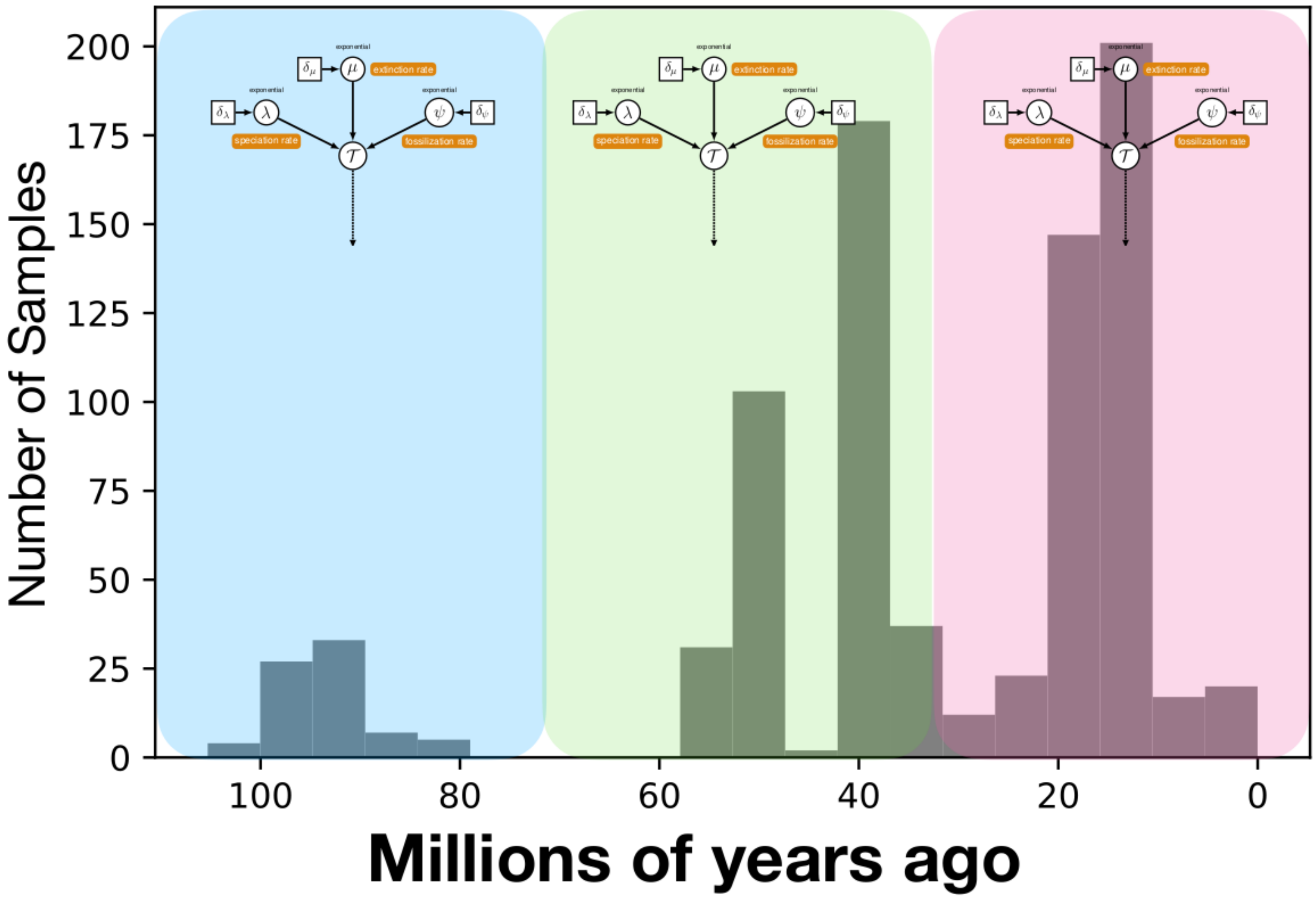


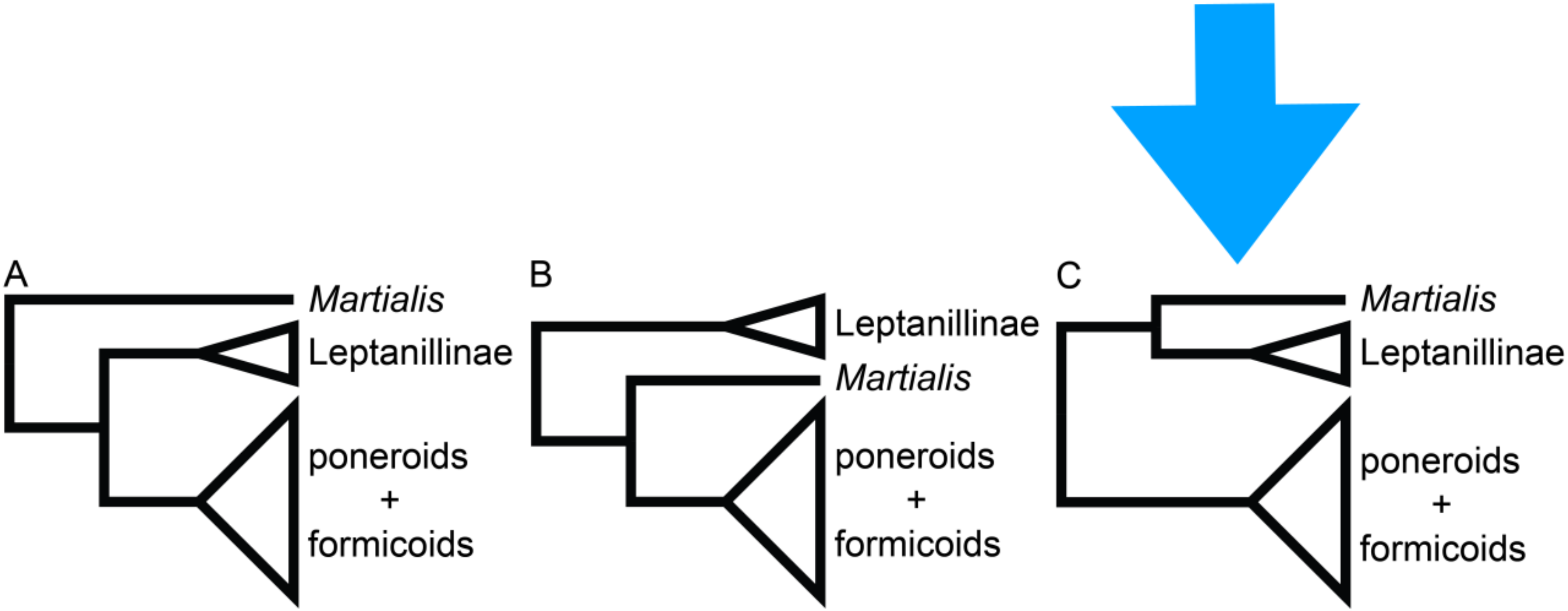
Wright and Lloyd, in review

Accounting for Uneven Fossil Sampling



Accounting for Uneven Fossil Sampling




A
Martialis
Leptanillinae
poneroids
+
formicoids
B
Leptanillinae
Martialis
poneroids
+
formicoids
C
Martialis
Leptanillinae
poneroids
+
formicoids

Introduction to Posterior Prediction

Assessing the fit of Normal distributions to trait data

Jeremy M. Brown and Christina L. Kolbmann

Last modified on October 10, 2019

Training students to be data detectives

Computational Biology

The why, when, and how of computing in biology classrooms [version 1; peer review: 1 approved]

April M. Wright [1], Rachel S. Schwartz [2], Jamie R. Oaks[3], Catherine E. Newman[4], Sarah P. Flanagan [5]

Author details

This article is included in the Bioinformatics Education and Training Collection collection.

Training students to be data detectives

Computational Biology

The why, when, and how of computing in biology classrooms [version 1; peer review: 1 approved]

April M. Wright[1], Rachel S. Schwartz[2], Jamie R. Oaks[3], Catherine E. Newman[4], Sarah P. Flanagan[5]

Author details

This article is included in the Bioinformatics Education and Training Collection collection.

Lessons learned in this class are being applied to our intro biology sequence

Training students to be data detectives

Training students to be data detectives

Hands-On Learning with RevBayes

Primary organizer: Dr. April Wright, Southeastern Louisiana University

Content: This workshop will focus on using the phylogenetic estimation software RevBayes in an instructional setting. We will first introduce the graphical model framework used by the software. Graphical models can be used to introduce the fundamentals of probability, while also enabling transparent and flexible assembly of new phylogenetic models. Then, we will discuss using RevBayes in hands-on exercises for systematics and phylogenetics courses. Topics will include making use of the robust RevBayes tutorial library, tailoring pre-existing tutorials to your course, contributing your tutorials to the tutorial bank, using interactive computing notebooks, and integrating the software with R and Python.

Training students to be data detectives

Discrete morphology - Models and Tree Inference

April M. Wright

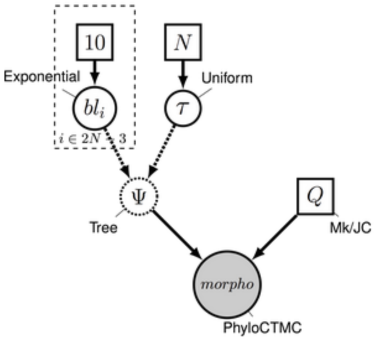
Source: `vignettes/module_05_TripartiteModel1_morph_change_models/05_RB_MCMC_Discrete_Morph.Rmd`

Introduction to phylogenetic models of morphological evolution

Morphological data is commonly used for estimating phylogenetic trees from fossils. This tutorial will focus on estimating phylogentic trees from *discrete* characters, those characters which can be broken into non-overlapping character states. This type of data has been used for estimation of phylogenetic trees for many years. In the past twenty years, Bayesian methods for estimating phylogeny from this type of data have become increasingly common.

This tutorial will give an overview of common models and assumptions when estimating a tree from discrete morphological data. We will use a dataset from Zamora, Rahman, and Smith (2013). This dataset contains 23 extinct echinoderm taxa and 60 binary and multistate characters.

Overview of Discrete Morphology Models




```
for (i in 1:n_branches) {
        bl[i] ~ dnExponential(10.0)
}
topology ~ dnUniformTopology(taxa)
psi := treeAssembly(topology, bl)

Q_morpho <- fnJC(2)

phyMorpho ~ dnPhyloCTMC( tree=psi,
Q=Q, type="Standard",
coding="variable" )

phyMorpho.clamp( morpho )
```

Graphical model showing the Mk model (left panel). Rev code specifying the Mk model is on the right-hand panel.

Contents

Thank You!

- Jeremy Brown
- Christina Kolbmann
- Basanta Khakurel
- Courtney Grigsby
- Tyler Tran
- Rachel Warnock
- Tracy Heath
- Graeme Lloyd
- Corrie Moreau
- David Bapst

Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM12345.

SOUTHEASTERN LOUISIANA UNIVERSITY

LBRN

MS:

- Barido-Sottani J, Saupe E, Smiley TM, Soul, LC, Wright AM, Warnock RCM. In review. Seven rules for simulations in paleobiology.
- Wright AM, Lloyd, GT. In review. Bayesian analyses in phylogenetic paleontology: Interpreting the posterior sample. Preprint: https://github.com/graemetlloyd/ProjectWhalehead/blob/master/vignettes/MS.pdf
- Warnock RCM, Wright AM. In review. Understanding the tripartite approach to Bayesian divergence time estimation. Preprint: https://www.overleaf.com/read/cbdxvgvxdkdq
- Wright AM, Schwartz RS, Oaks JM, Newman CM, and Flanagan SP. Accepted. The Why, When, and How of Computing in Biology Classrooms. Preprint: https://www.overleaf.com/read/cnnpbfzgzvfd
- Wright AM 2019. A systematist's guide to estimating Bayesian phylogenies from morphological data. Insect Systematics and Diversity 3: https://doi.org/10.1093/isd/ixz006.
- Wright AM 2019. treesiftr: An R package and server for viewing phylogenetic trees and data Journal of Open Source Education, 2(11), 35, https://doi.org/10.21105/jose.00035
- Devitt TJ, Wright AM, Cannatella DC, Hillis, DM. 2019. Species delimitation in endangered groundwater salamanders: Implications for aquifer management and biodiversity conservation. Proceedings of the National Academy of Sciences 116: 2624-2633

Courses:

- Biological Data Analysis: https://biologicaldataanalysis2019.github.io/2019/
- Introduction to Biodiversity Data Science: https://paleantology.github.io/GBIO153H/index.html
- Systematics: https://wrightaprilm.github.io/Systematics2020/index.html

Service:
Assistant Editor: Systematic Biology
Organizer, iEvoBio