June 4, 2018 **SSB Standalone Meeting, OSU**

Workshop Schedule:

Morning session	9:00-9:30am	Lacey Knowles, University of Michigan
(Talks)	9:30-9:50am	Paul Hime, University of Kansas
	9:50-10:10am	Stacey Smith, University of Colorado, Boulder
	10:10-10:30am	Break
	10:30-10:50am	Steven Smith, University of Michigan
	10:50-11:10am	Cecile Ane, University of Wisconsin-Madison
	11:10-11:30am	Melissa DeBiasse, University of Florida
	11:30-12:00pm	Laura Kubatko, The Ohio State University
Lunch	12:00-1:30pm	On your own
Afternoon session	1:30-2:00pm	ASTRAL -- Tandy Warnow, University of Illinois
(Software tutorials)	2:00-2:45pm	PhyloNet -- Luay Nakhleh, Rice University
	2:45-3:15pm	Break
	3:15-4:00pm	SVDQuartets -- Dave Swofford and Laura Kubatko
	4:00-5:00pm	Open lab

mbi

THE OHIO STATE UNIVERSITY DEPARTMENT OF EVOLUTION, ECOLOGY, AND ORGANISMAL BIOLOGY

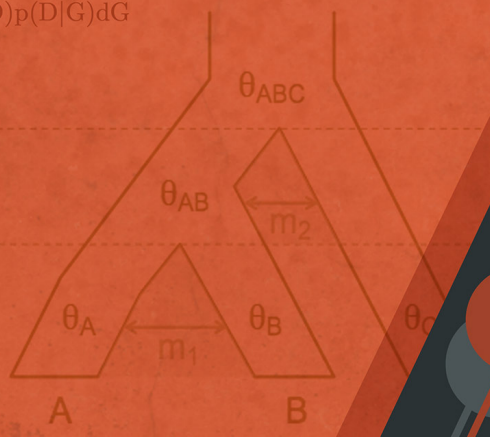
Species tree inference: empirical challenges of today and tomorrow

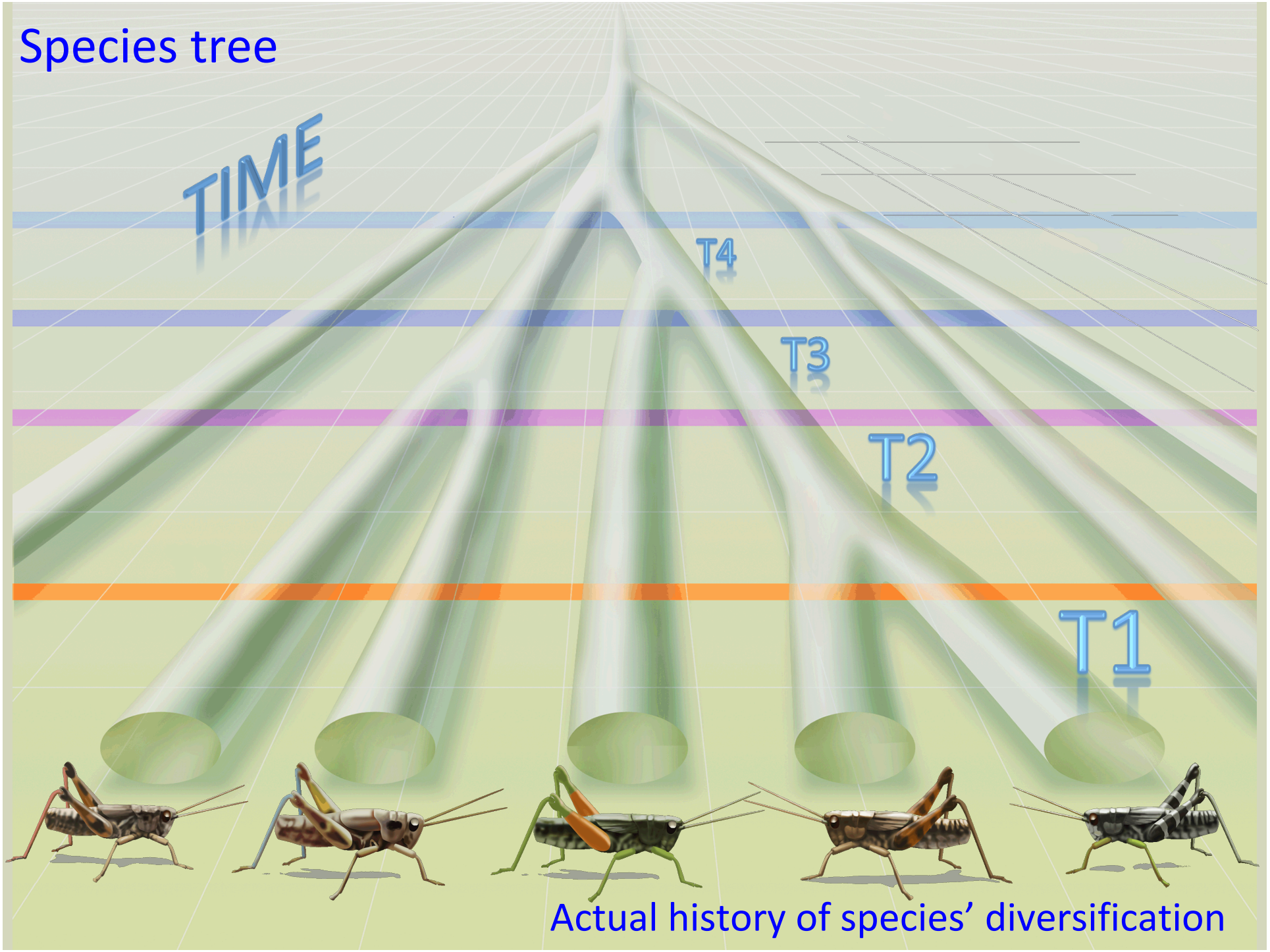
L. Lacey Knowles

Dept. of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor MI

Species tree inference: theoretical challenges of today and tomorrow

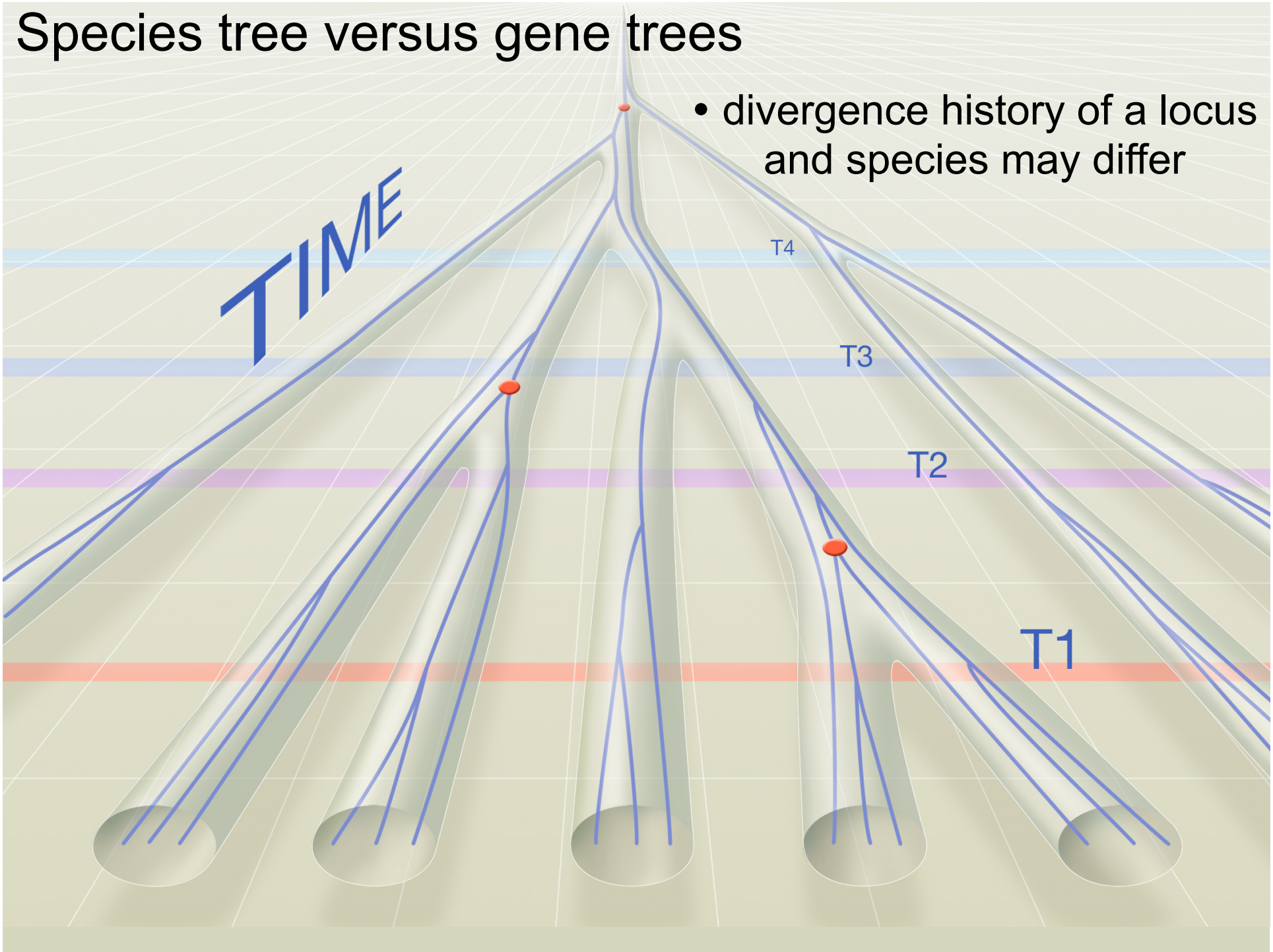
Laura Kubatko




Species tree
TIME
T4
T3
T2
T1
Actual history of species' diversification

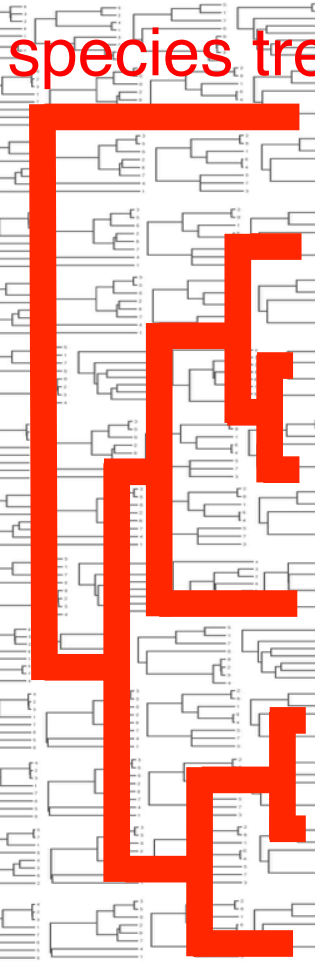
Species tree versus gene trees

- divergence history of a locus and species may differ



Species tree versus gene trees

- the divergence history of individual loci may differ:
  - incomplete lineage sorting
  - hybridization
  - duplication and loss of gene regions
  - selection
  - little information



Description

Recent computational and modeling advances have produced methods for estimating species trees directly, avoiding the problems and limitations of the traditional phylogenetic paradigm where an estimated gene tree is equated with the history of species divergence. The overarching goal of the volume is to increase the **visibility and use of these new methods by the entire phylogenetic community** by specifically addressing several challenges: *(i) firm understanding of the theoretical underpinnings of the methodology, (ii) empirical examples demonstrating the utility of the methodology as well as its limitations, and (iii) attention to technical aspects involved in the actual software implementation of the methodology*. As such, this volume will not only be poised to become the quintessential guide to training the next generation of researchers, but it will also be instrumental in ushering in a new phylogenetic paradigm for the 21st century.

Estimating Species Trees:

Practical and Theoretical Aspects

L. L. Knowles and L. S. Kubatko, eds.

2010

Methodological advances

SVDquartets

Phylonet

ASTRAL

Than, Ruths, Nakhleh (2008)

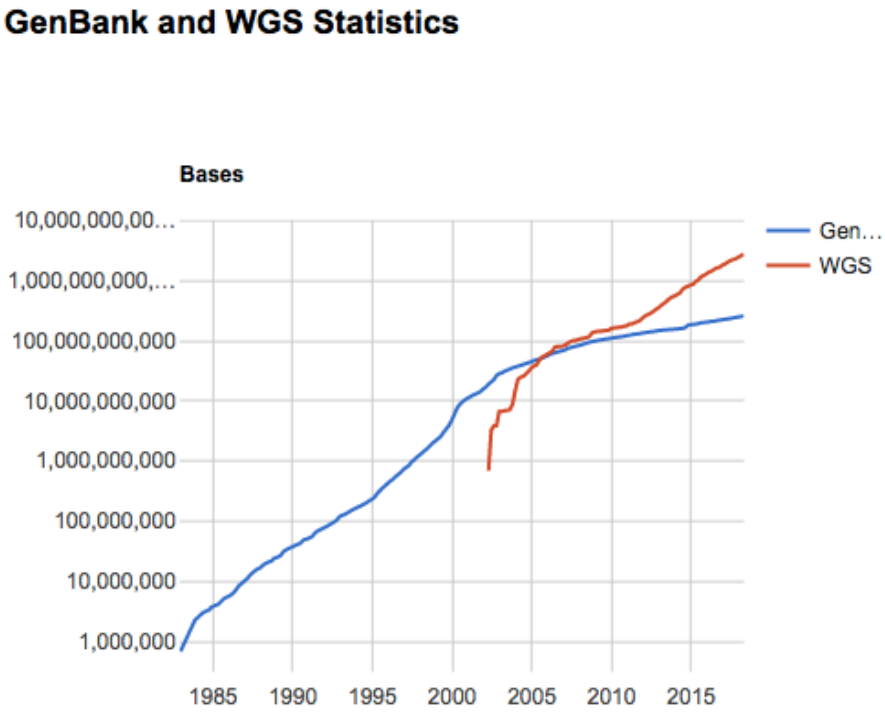
Mirarab et al. (2014)

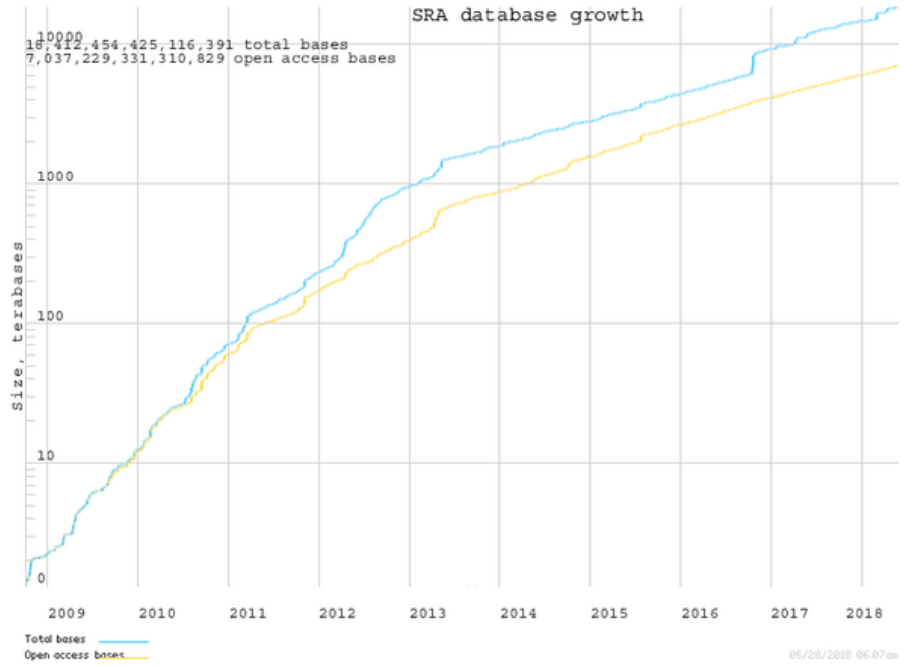
Chifman & Kubatko (2014)

Tons of data

- WGS (whole genome shotgun): 2,784,740,996,536 bases in GenBank

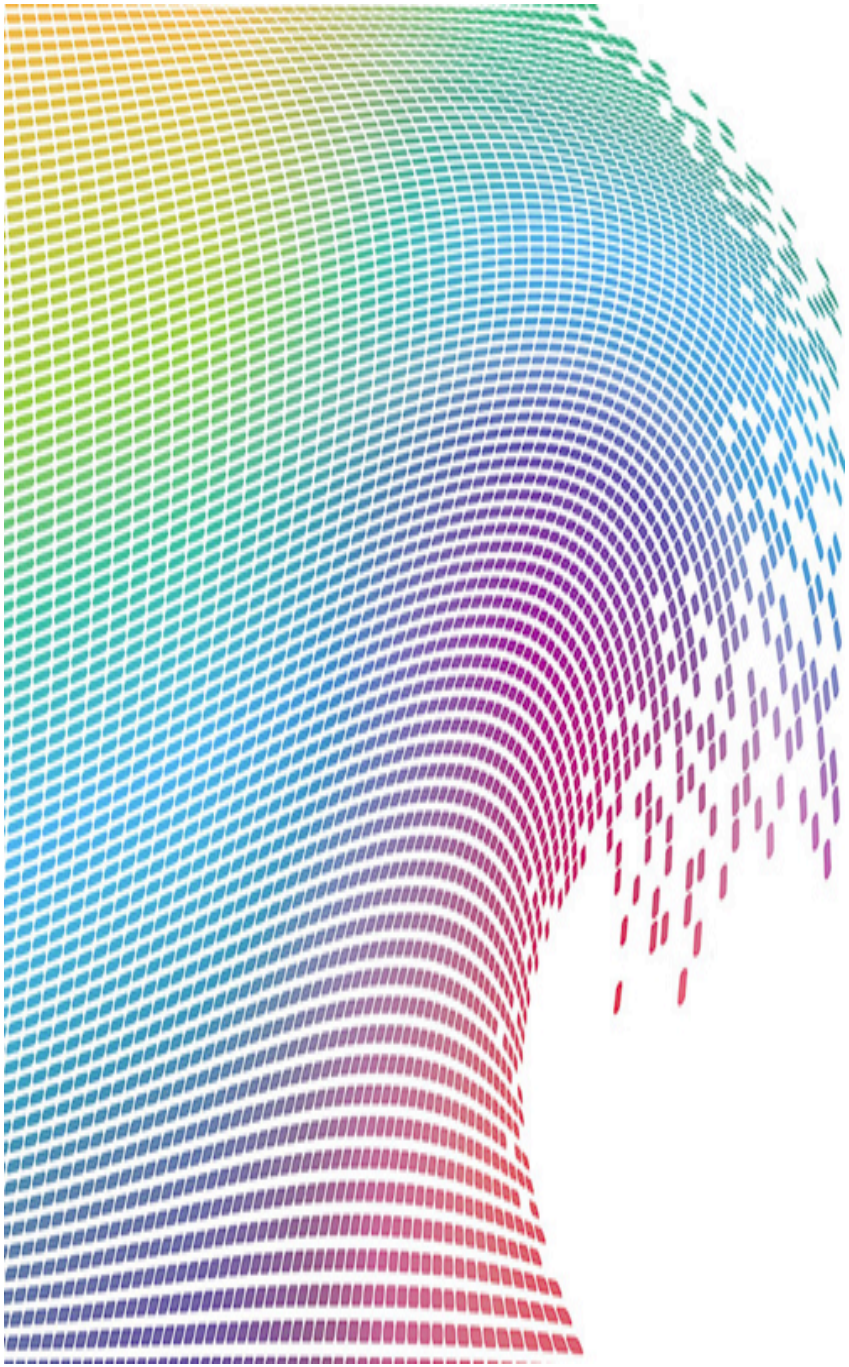
- SRA (short read archive) :16,267,243,120,778,112 bases (4-5 orders of magnitude more than WGS)





- thousands of transcriptomes (2-5K in plants alone)

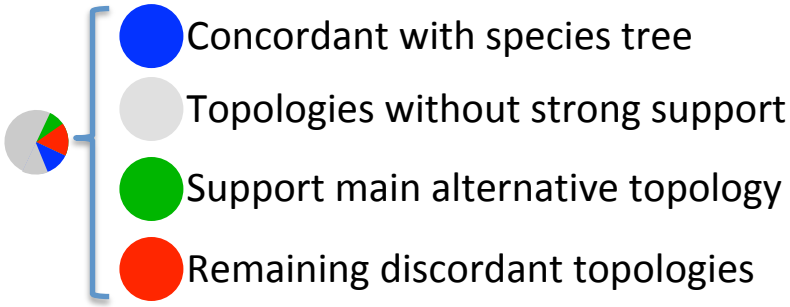
Genomic data



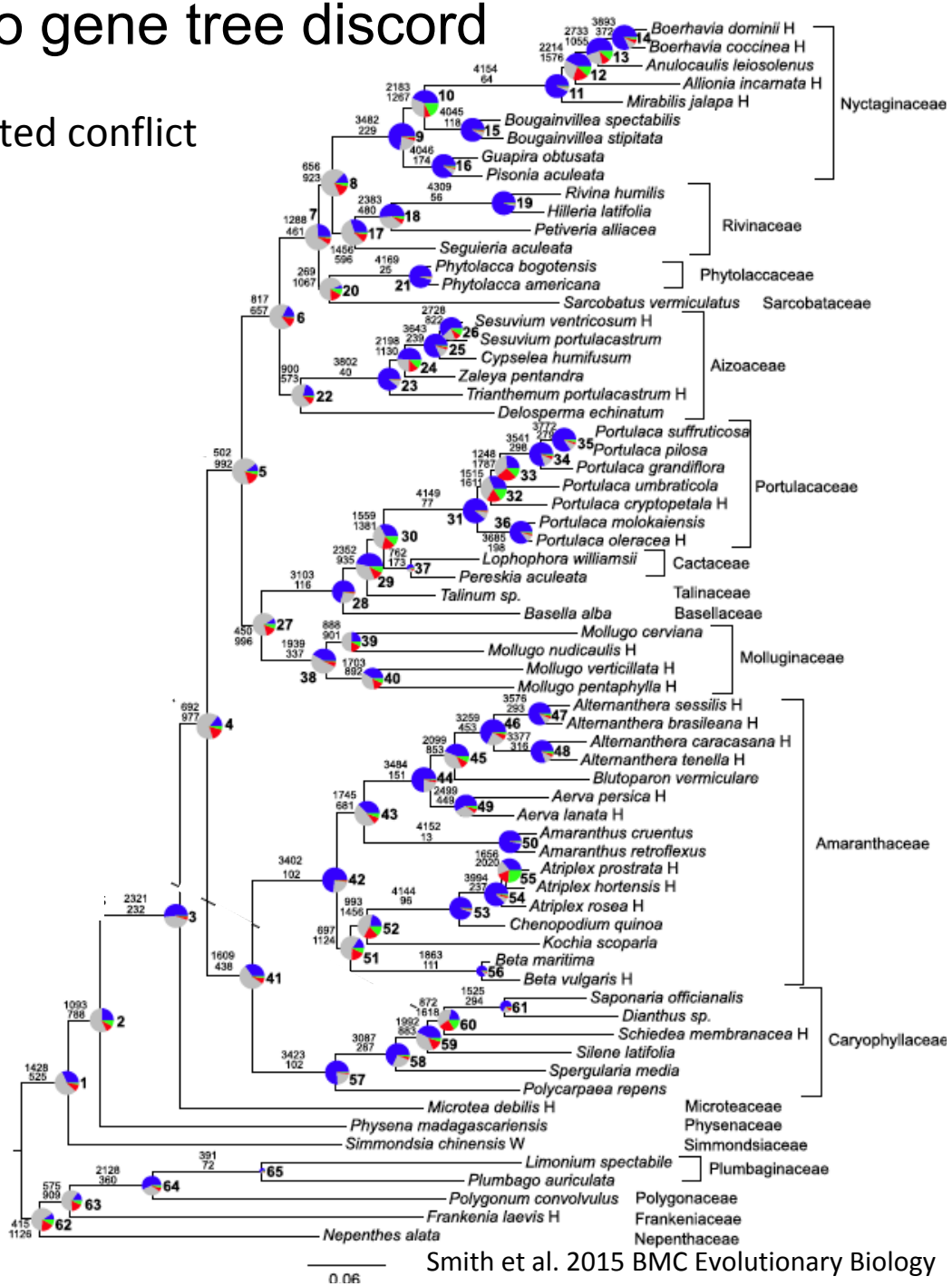
Recalcitrant nodes across the tree of life!

Multiple processes contribute to gene tree discord

- highly elevated levels of strongly supported conflict



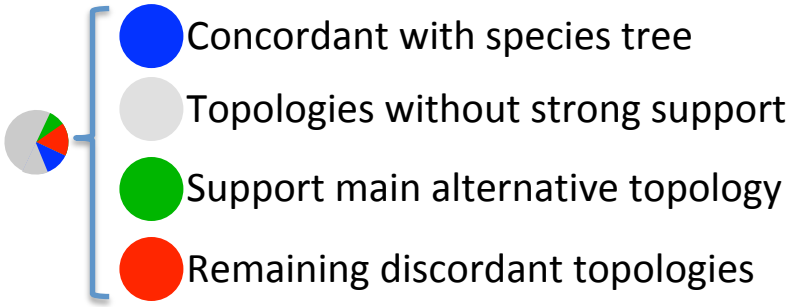
Caryophyllales



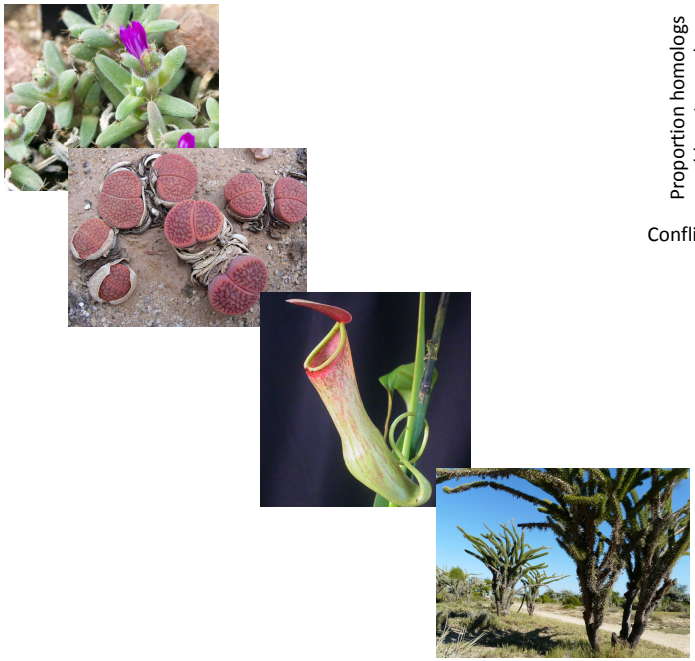
Smith et al. 2015 BMC Evolutionary Biology

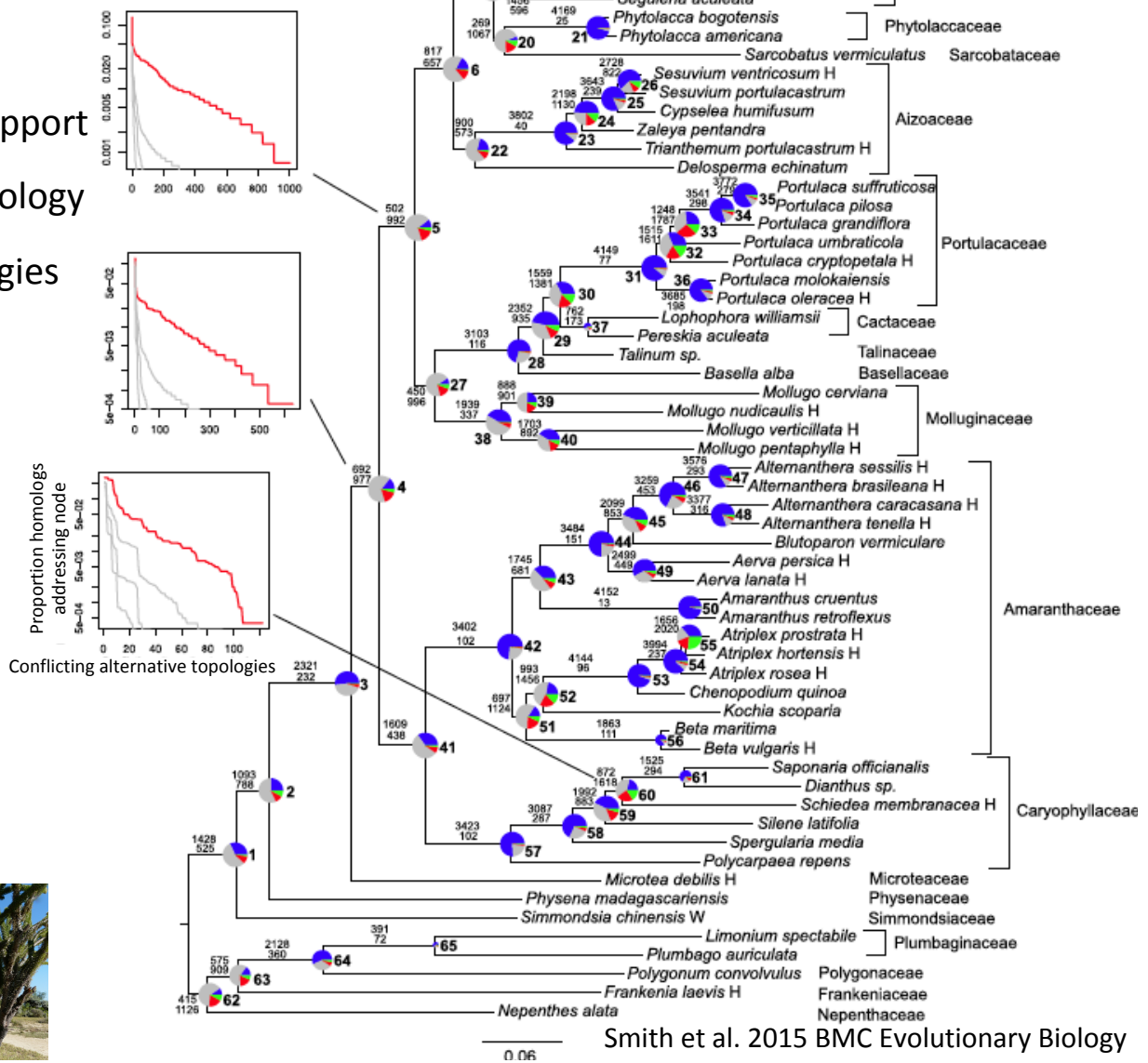
Multiple processes contribute to gene tree discord

- highly elevated levels of strongly supported conflict that cannot be explained by ILS alone at some nodes



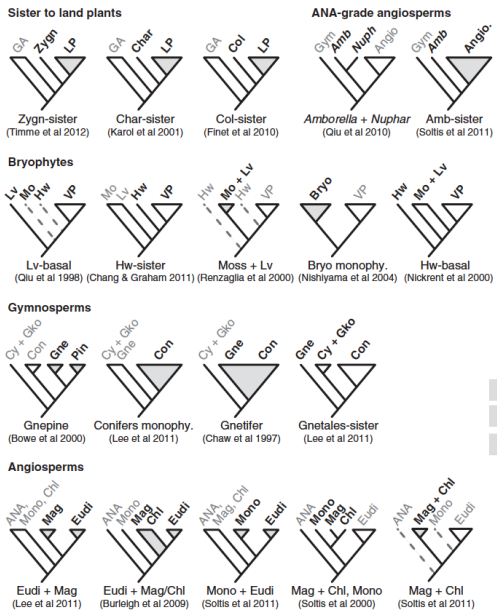
Caryophyllales





Smith et al. 2015 BMC Evolutionary Biology

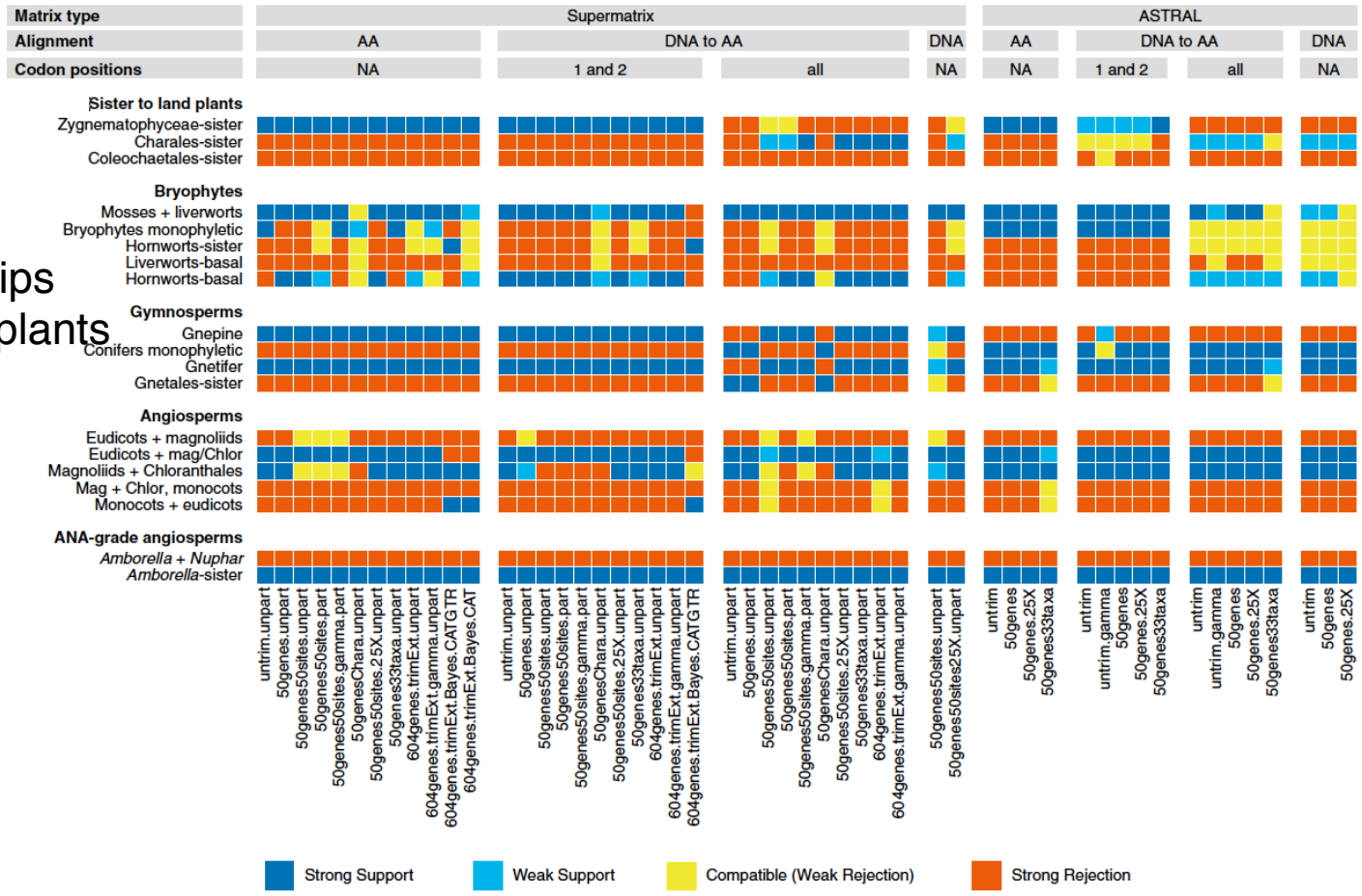
Systematic errors in phylogenetic inference caused by model misspecification



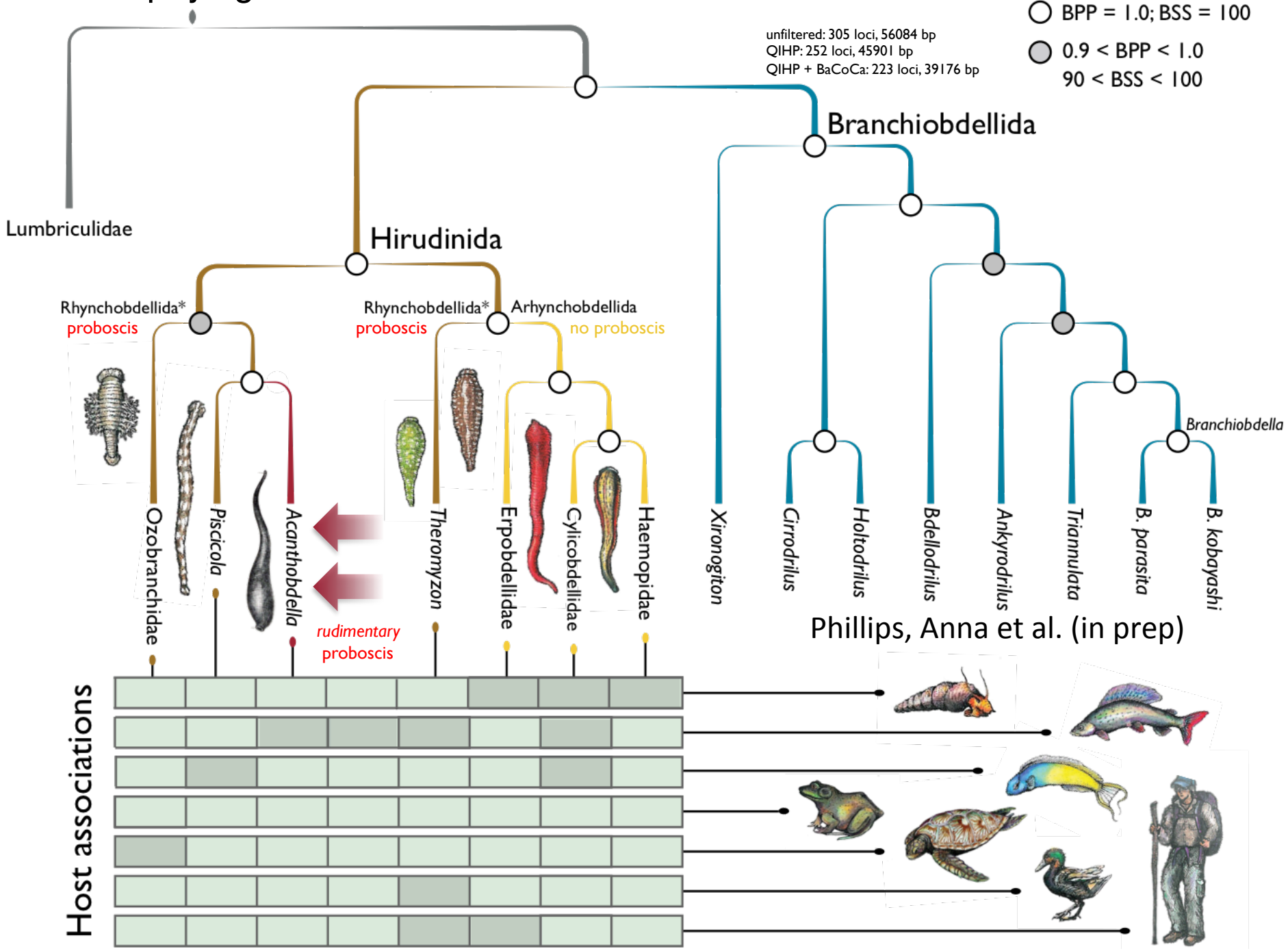
hypothesized relationships among major clades of plants

Wickett et al. 2014 PNAS

69 analyses of 92 taxa



Different phylogenetic estimates with inclusion/exclusion of loci



Phillips, Anna et al. (in prep)

Gene tree discord?

More data?

Data versus model problem?

Filter data?

Alignment?

Subsets of data?

Solutions to the data problem / model problem?

Basis for empirical data decisions

*characterized discord patterns from 90 published empirical studies that range is dataset size, taxa, and marker type

- Distribution of discord

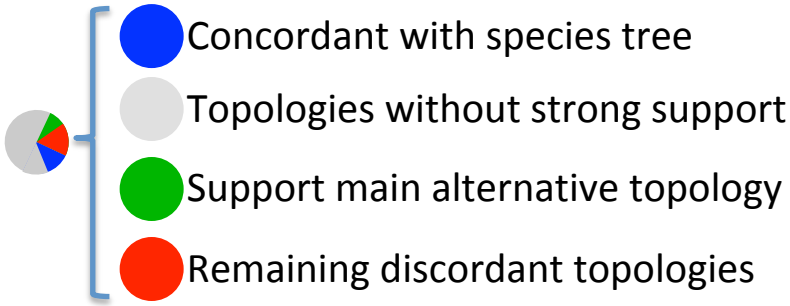
  - Correlated with properties of species trees

  - Restricted to “bad” nodes (concentrated in taxonomic groups)

  - Evidence of “bad” loci (differences among marker types)

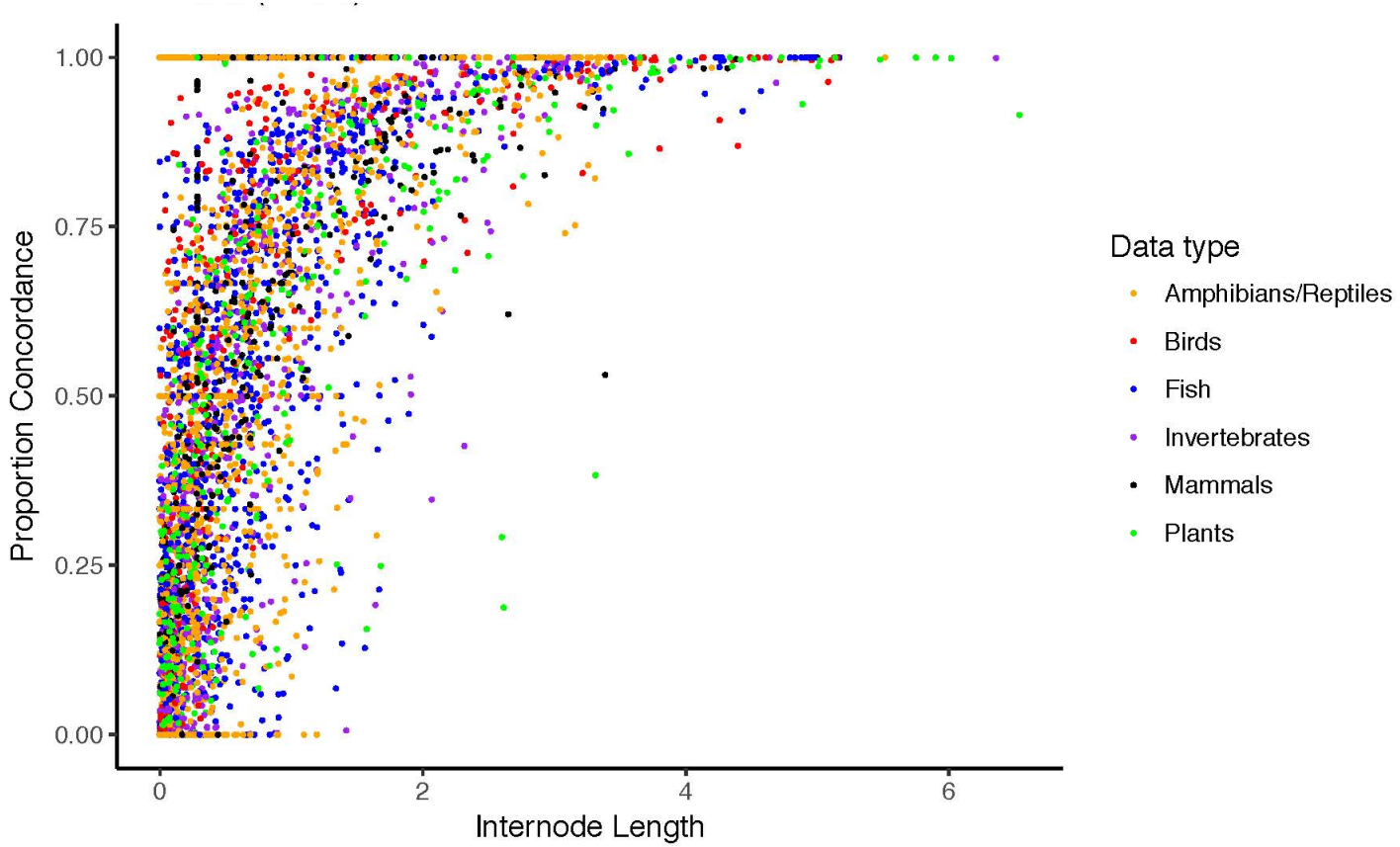
Phyparts (Smith et al. 2015) node-specific concordance/discord across loci

	node 1	node 2	node 3	node 4	node 5
gene 1					
gene 2					
gene 3					
gene 4					
gene 5					
gene 6					
gene 7					
gene 8					
gene 9					
gene 10					



“bad” nodes or loci: those associated with a disproportionate amount of discord

Concordance related to species tree shape

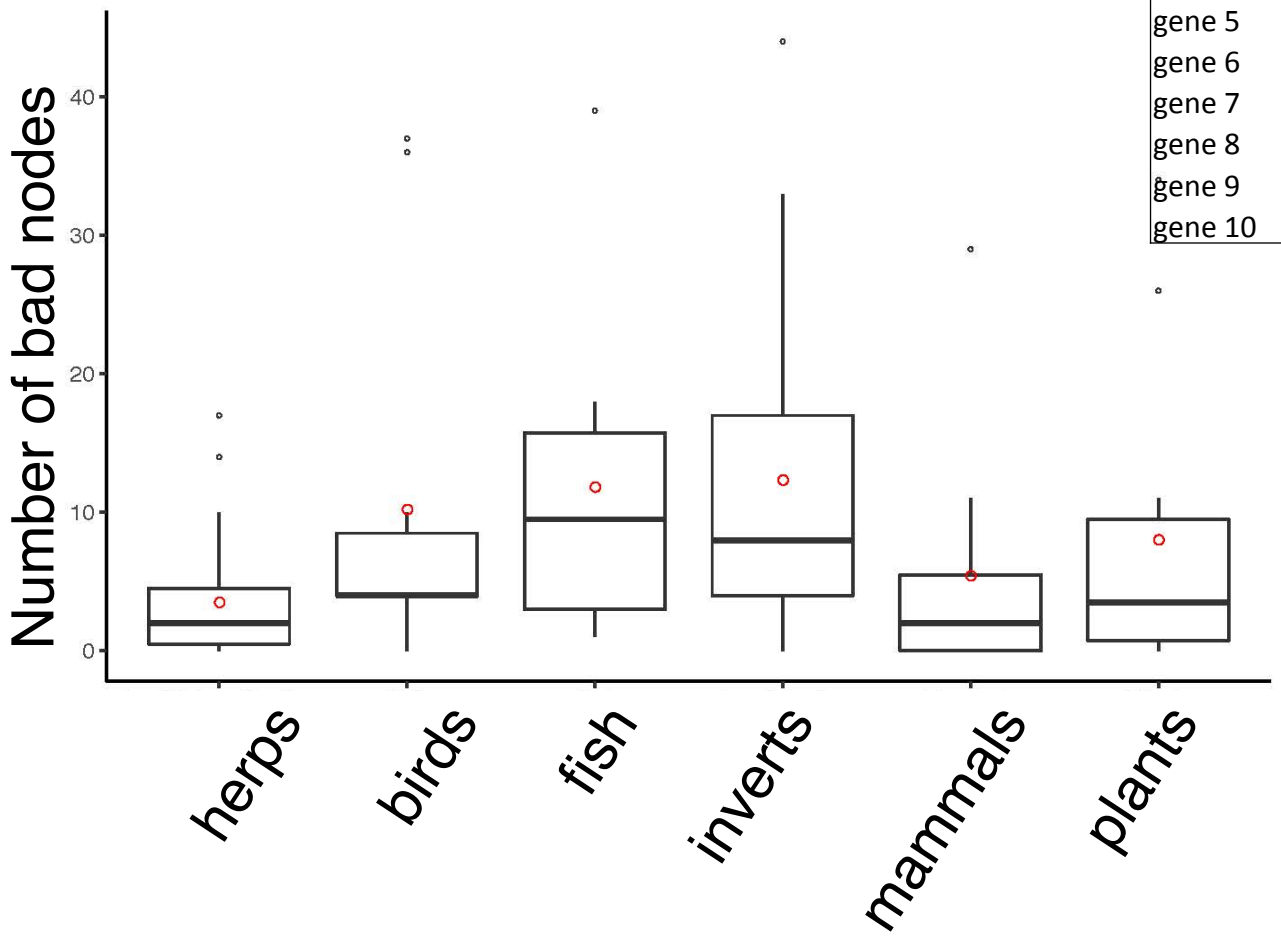


Practical relevance: account for ILS in phylogenetic models

“Bad” nodes

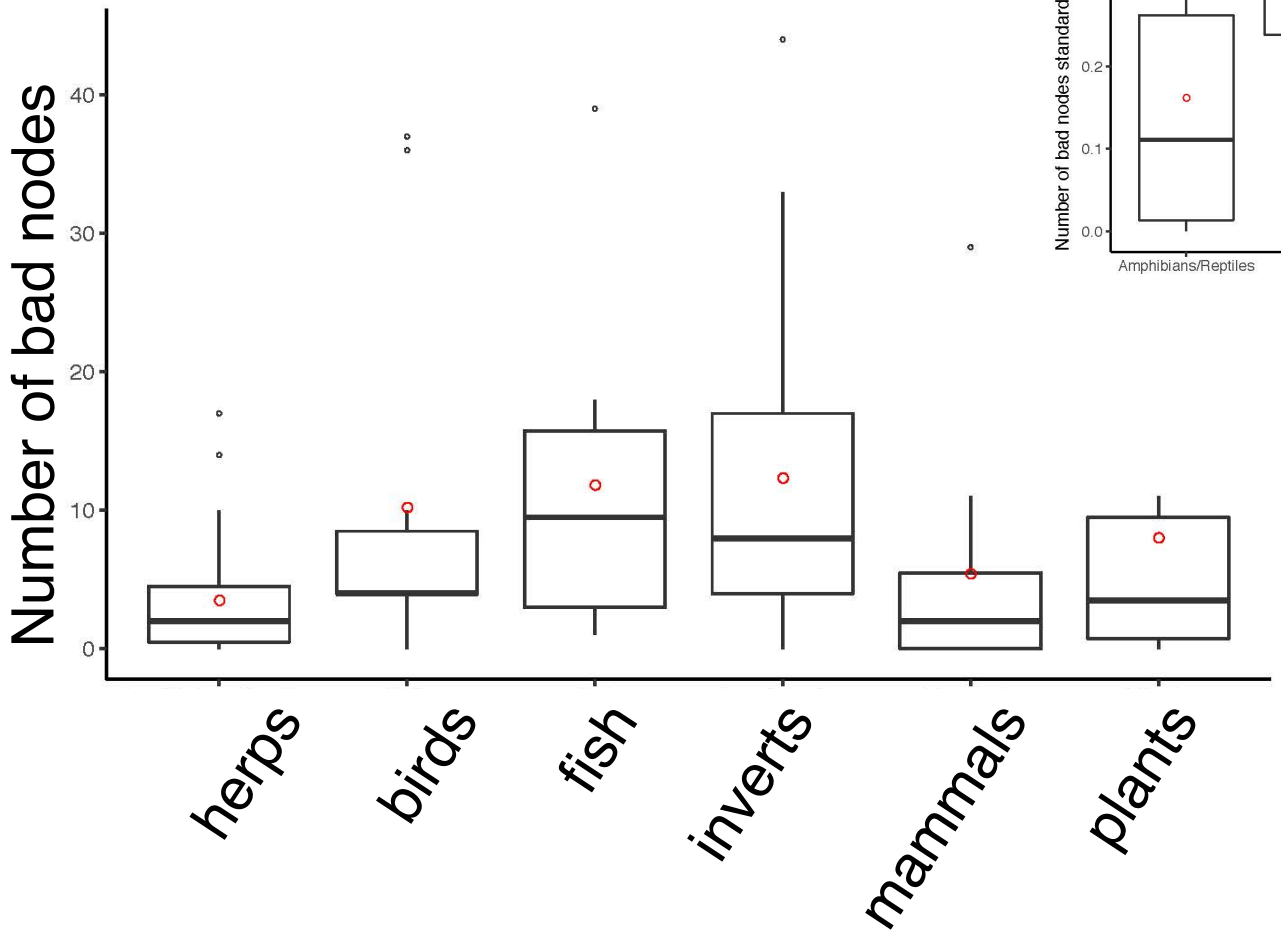
“bad” nodes: those associated with a disproportionate amount of discord

	node 1	node 2	node 3	node 4	node 5
gene 1					
gene 2					
gene 3					
gene 4					
gene 5					
gene 6					
gene 7					
gene 8					
gene 9					
gene 10					

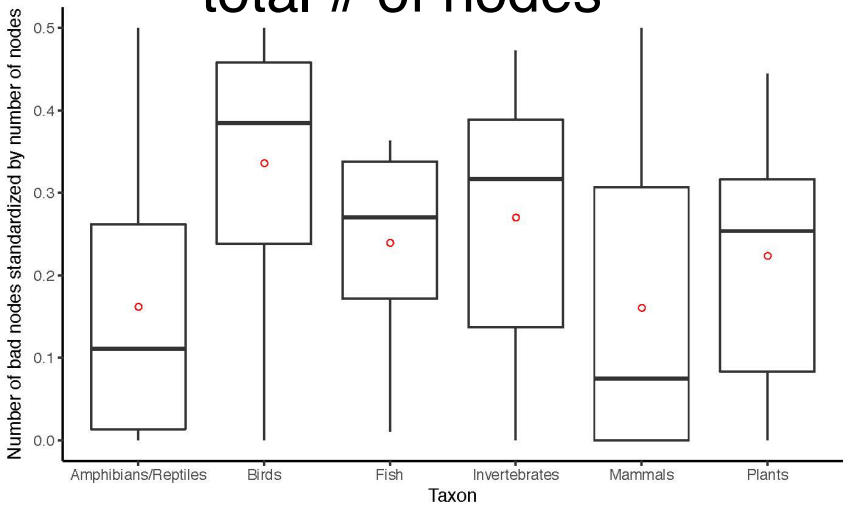


Practical relevance: no obvious evidence for concentration of discord in specific taxonomic groups

“Bad” nodes

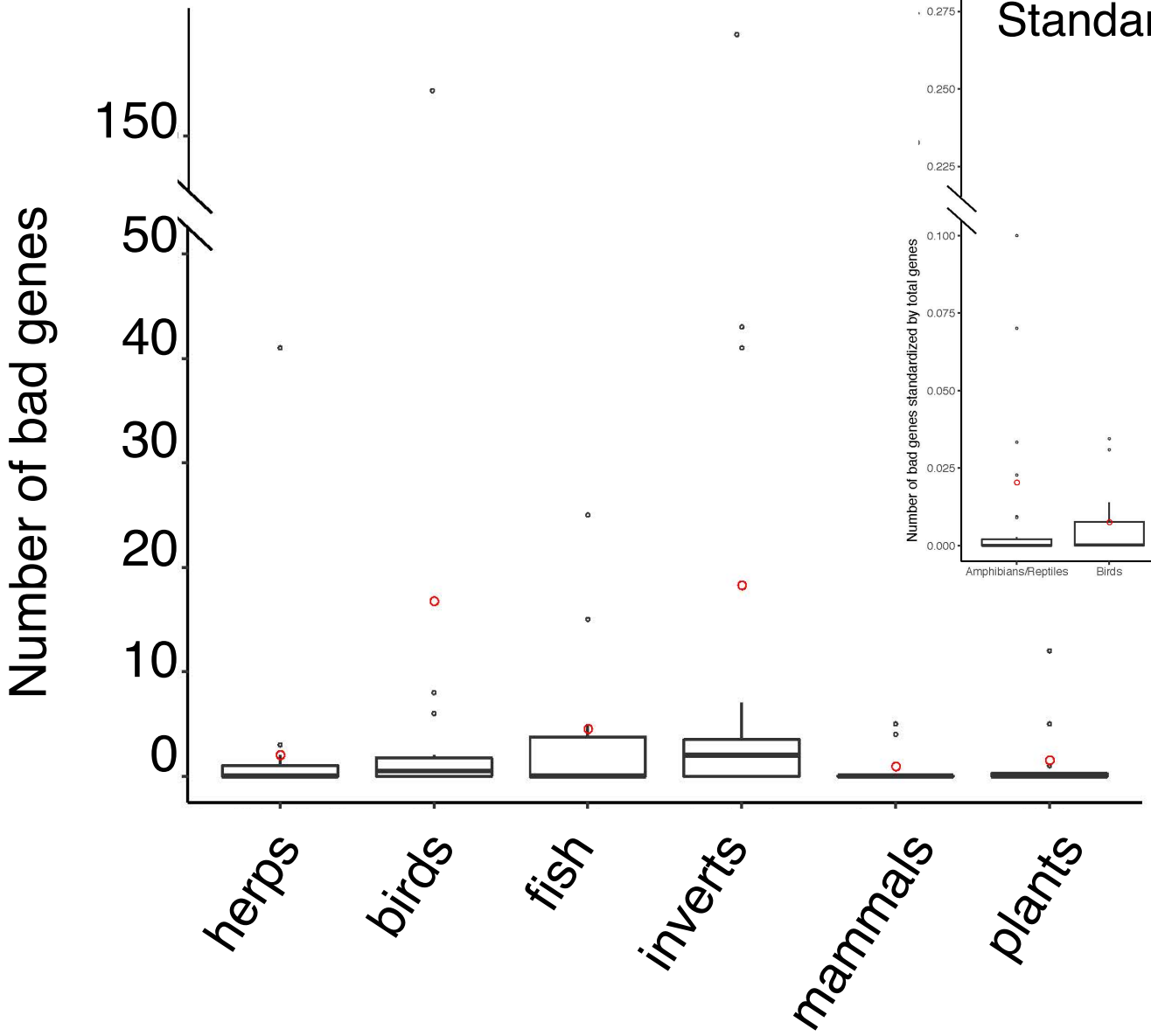


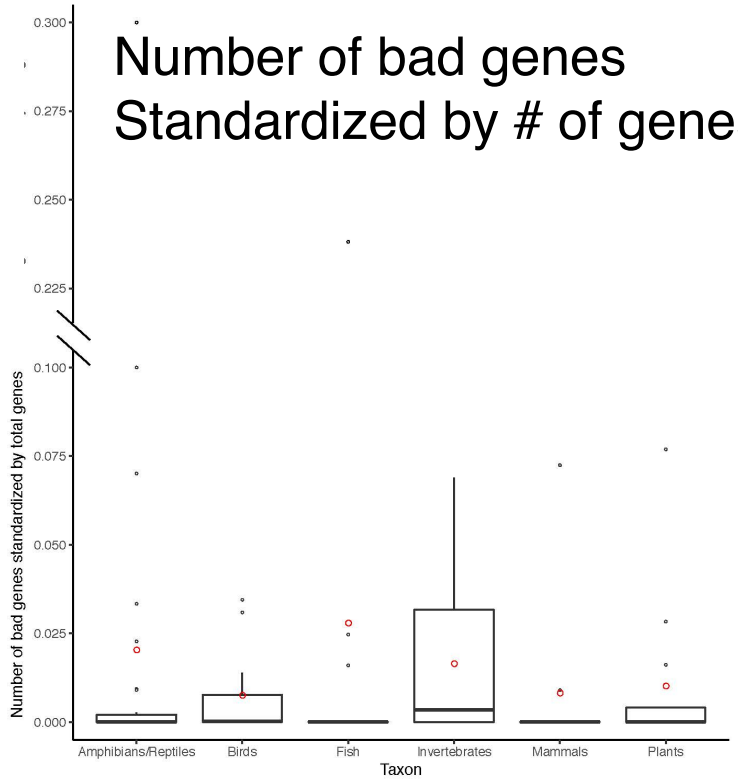
“Bad” nodes standardized by total # of nodes



Practical relevance: no obvious evidence for concentration of discord in specific taxonomic groups

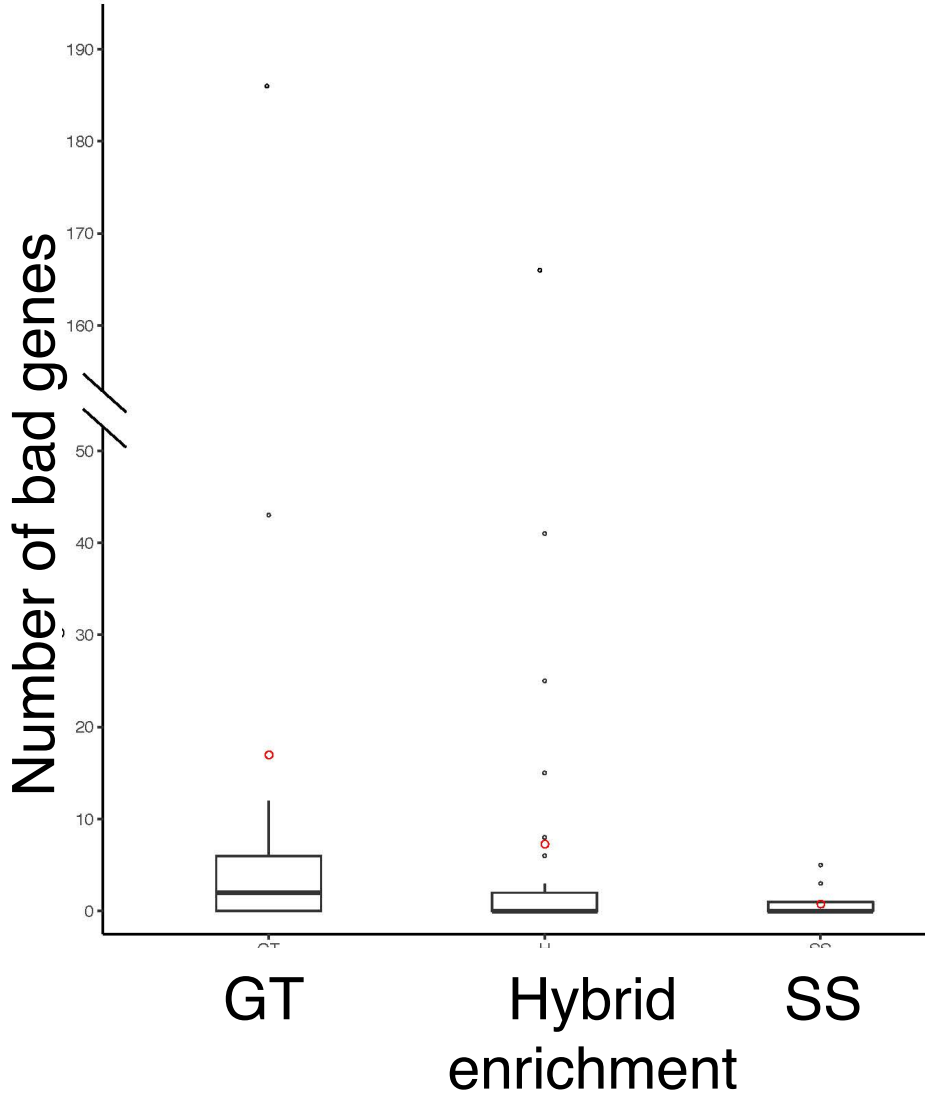
Number of bad genes

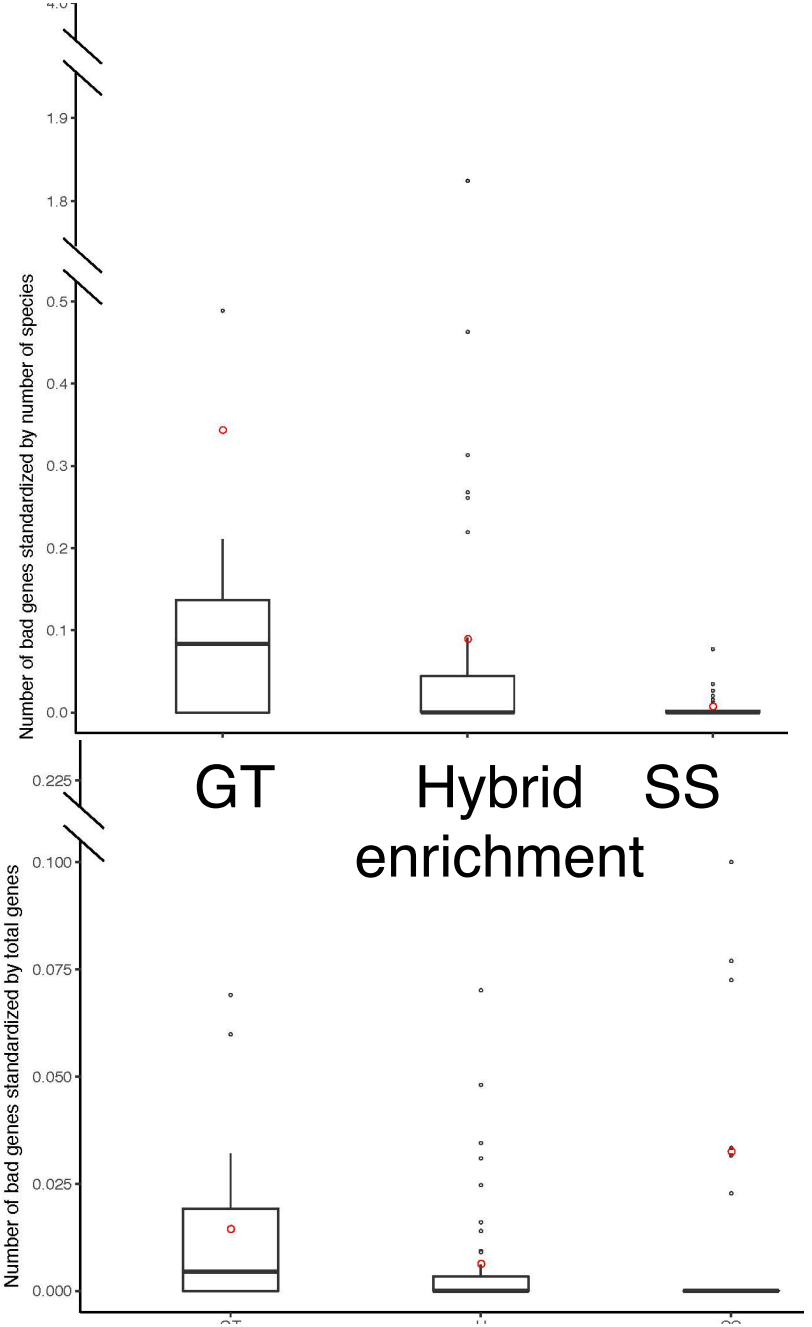




Practical relevance: include or exclude loci?

“Bad” genes by marker type





Practical relevance: quality checks of data could be improved

What is the empiricists to do?

Gene tree discord?

More data?

Data problem versus model problem?

Filter data?

Heterogeneity of processes?

Alignment?

Subsets of data?

Goal: practices to improve phylogenetic accuracy

Questionaire:

Dear Workshop Participants,

We would like to thank you for participating in our Species Tree Workshop and we invite you to participate in this short questionnaire. Your answers will not only be used to gain feedback to improve the quality of future workshops, but we would invite your participation as we develop our next book on species tree inference. It has been 10 years since our last book was published now and a lot has changed during this time. The goals of our new book "Species tree inference: a guide to the theoretical and empirical challenges of today and tomorrow" are two-fold. First, we'd like to provide a much-needed update to the collection of methods and ideas included in our first book. Second, we'd like this book to be "forward-looking", in the sense of including consideration of the challenges and issues on the horizon for the fast-moving field of species tree inference.

As an active participant in completing the question, your comments and suggestions will be used as we develop the book. Specifically, your input will be incorporated and acknowledged in our opening chapter of the book to provide a general update on the status of the field, including the breadth of applications and outstanding challenges. Your comments and suggestions will also be used to highlight the motivation and rationale for the topics covered in the book. That is, your input will assure that the book reflects your experiences and will span the interests and concerns of the diverse community that is engaged in species tree inference.

Again, we thank you and look forward to what should be a stimulating and fun workshop!

Best wishes,
Lacey Knowles and Laura Kubatko

Species tree inference:

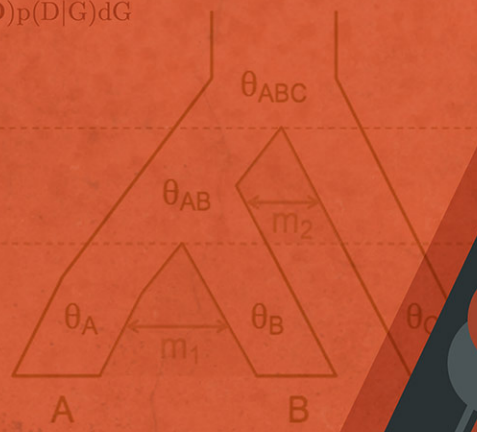
a guide to the theoretical and empirical challenges of today and tomorrow

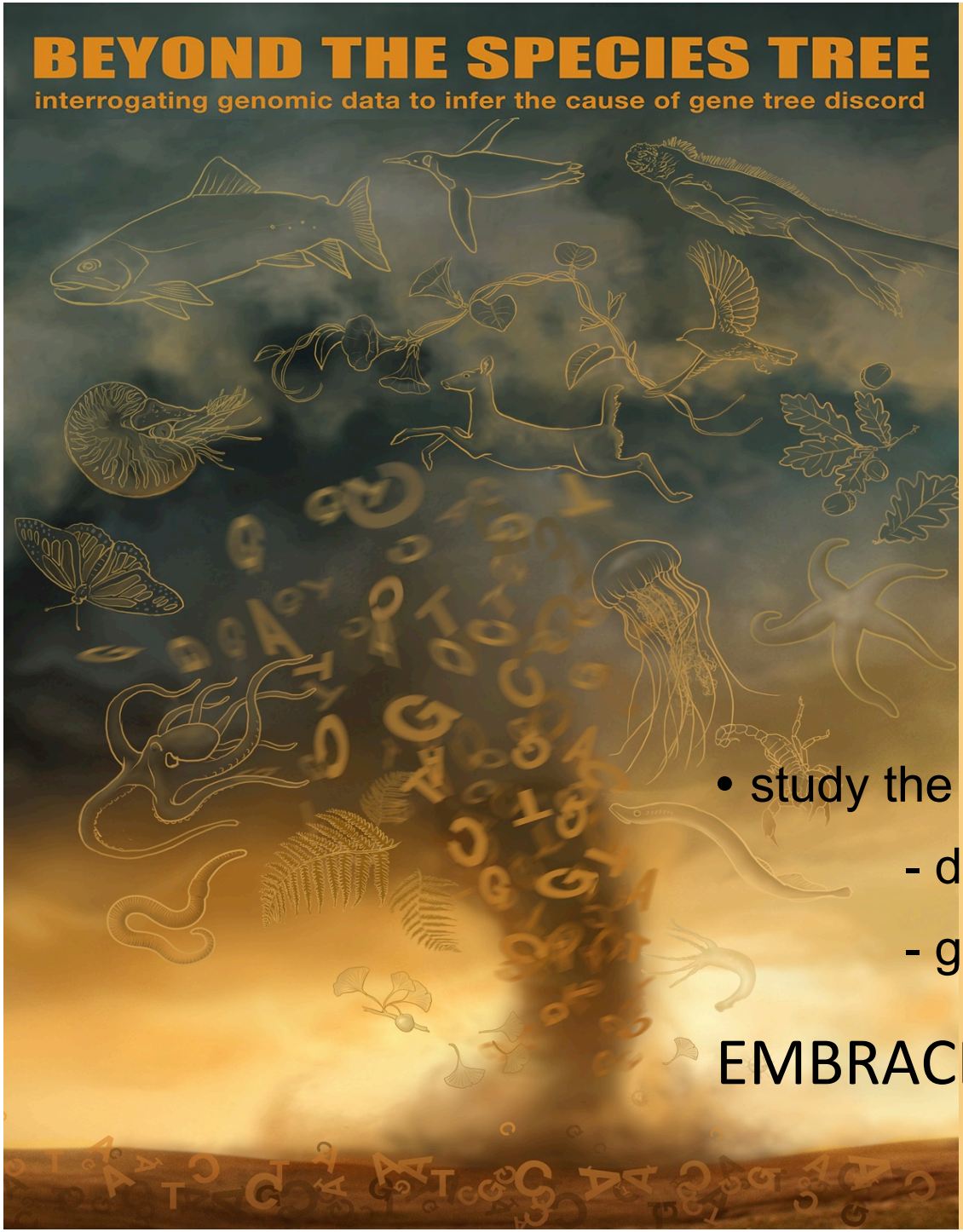
L. S. Kubatko and L. L. Knowles eds.

$$P_{uv}(t) = \sum_{j=v}^{u} e^{-j(j-1)t/2} \frac{(2j-1)(-1)^{j-v}}{v!(j-v)!(v+j-1)} \prod_{y=0}^{j-1} \frac{(v+y)(u-y)}{u+y}$$

$$\mathcal{L}\Big((\mathcal{S},\tau)|D_1,D_2,\ldots,D_L\Big) = \prod_{l=1}^{L}\Big(\sum_{\mathcal{H}}\int_{\mathbf{t}_h}\Big(\prod_{j=1}^{k_l} P(\mathbf{p}_j|(\mathcal{G}_h,\mathbf{t}_h))\Big) f_h(\mathbf{t}_h|(\mathcal{S},\tau))d\mathbf{t}_h\Big)$$

$$p(D|\Theta) = \int_G \mathrm{p(G|\Theta)p(D|G)dG}$$





- study the processes underlying discord
  - diversification history of taxa
  - genome evolution

EMBRACE THE HETEROGENEITY !

Thank you!

- Richie Hodel
  - Postdoctoral fellow

- Laura Kubatko

- Stephen Smith

knowlesl@umich.edu