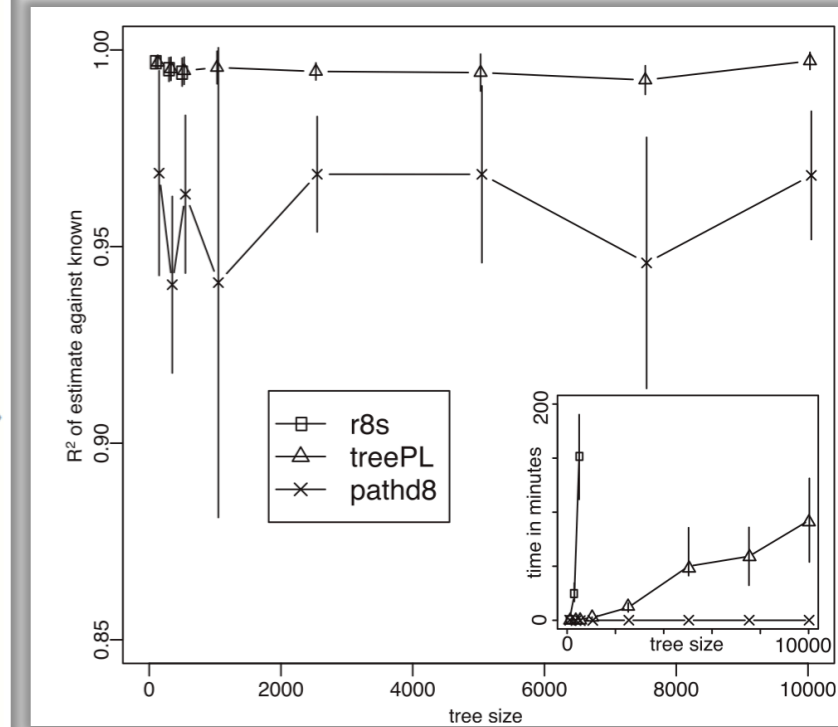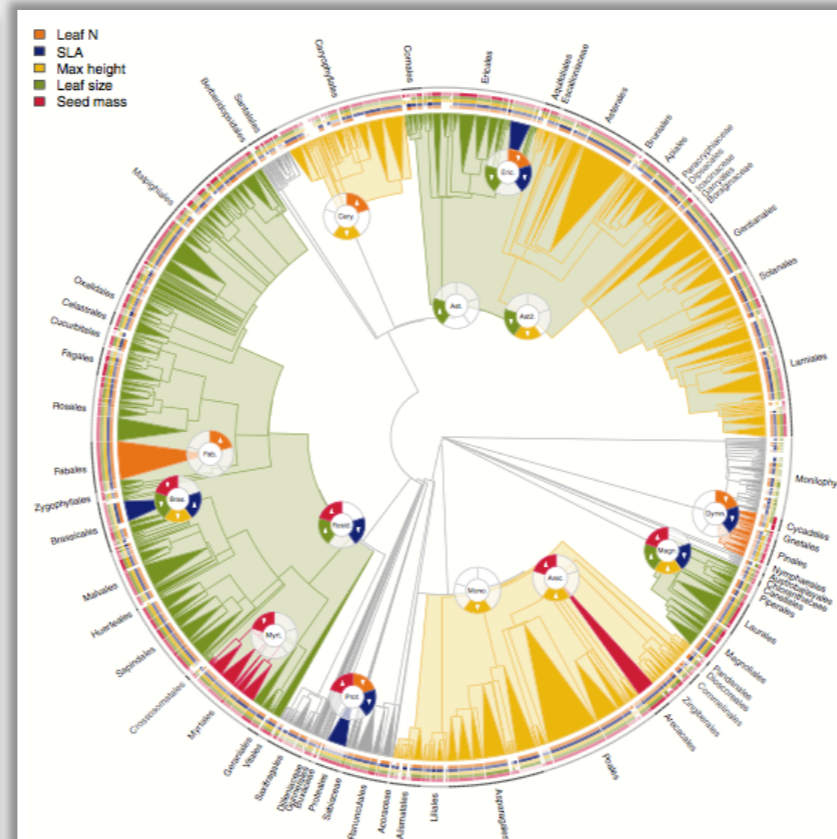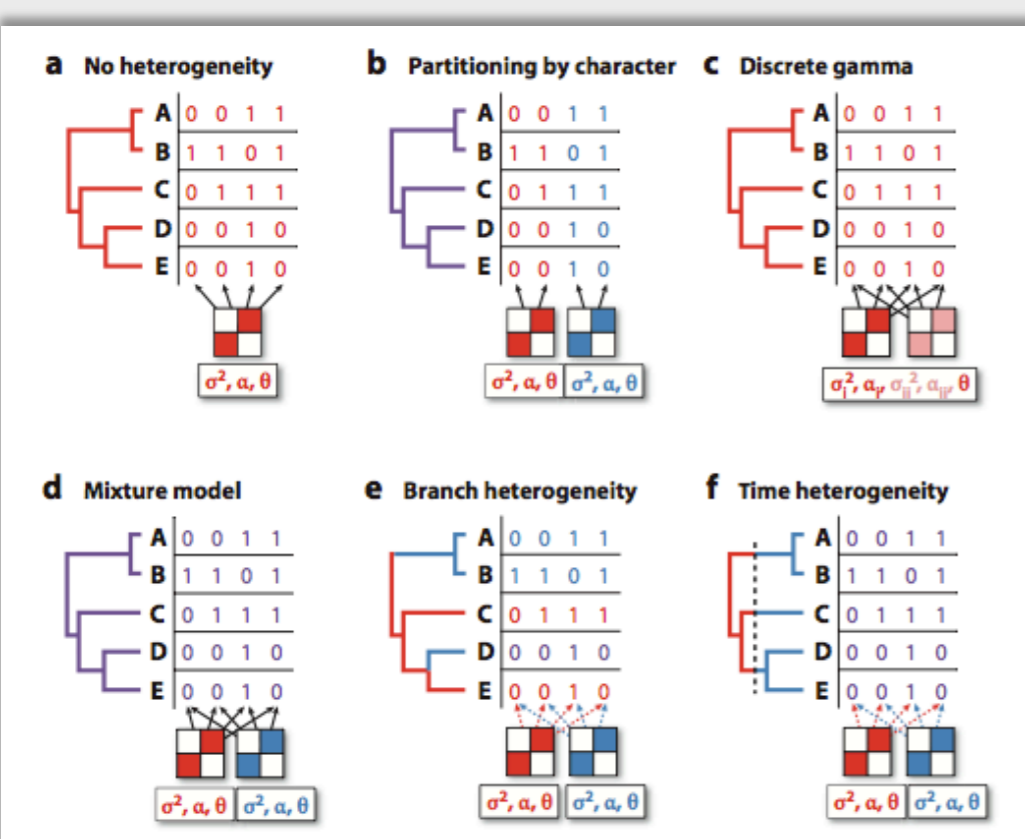# Developing and applying tools to answer outstanding biological questions using phylogenies, II

Brian O'Meara
brianomeara.info
UT Knoxville

Teaching

	2014-15	2013-14	2012-13	2011-12	2010-11	2009-10
Speciation	2	1				1
Macroevolution	1	1	1	1	1	1
Core: Evolution*	1	1	1	1	1	1
HOFF*	2	2	2	2		
Bio130: Biodiversity		1		1		
Seminar					1	

- High performance computing for phylogenetics, NIMBioS tutorial.   Organizer
- Species delimitation course, Gothenburg, Sweden.   Instructor
- eFlower summer school, Vienna, Austria.   Instructor
- Evolutionary quantitative genetics, NESCent.   Instructor
- Markov processes, Lausanne, Switzerland.   Instructor
- Computing in the cloud, NIMBioS tutorial.   Co-organizer
- Evolutionary quantitative genetics, NIMBioS tutorial.   Instructor
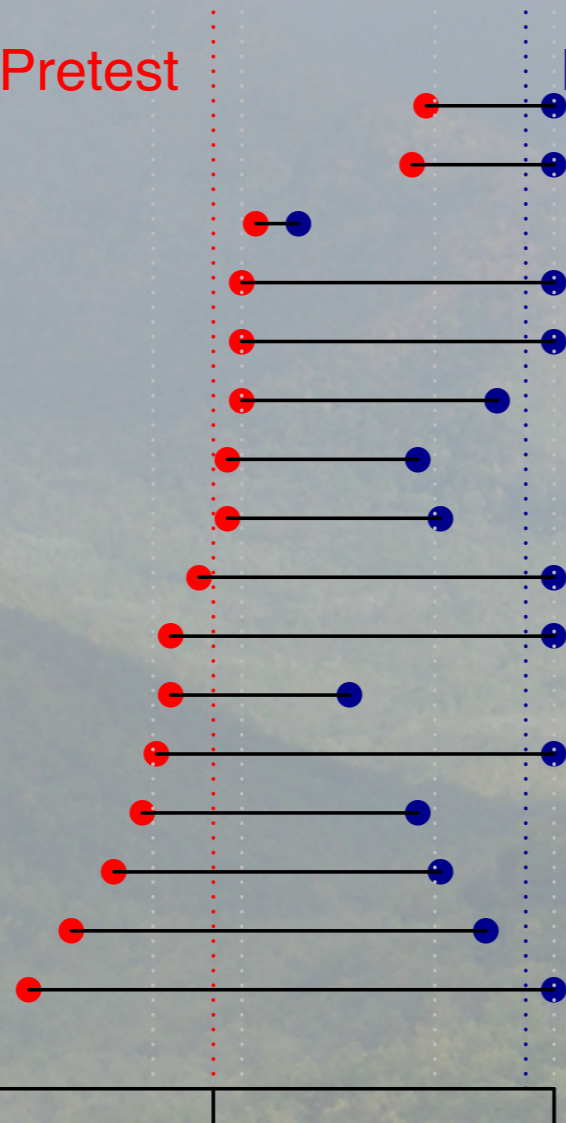
* Team taught courses

Teaching

Assessment

Innovation

Teaching

- Bio130: Greatest contribution to learning: "The open dialogue we had in every lecture"

- Macroevolution: "I found this class wonderfully stimulating and I really appreciated that Prof. O'Meara would stop in the middle of his lecture to answer any and all questions without it bothering him"

- Core: "The subject is a difficult one, but O'Meara communicated complex topics in a very understandable way. In short, he is a great teacher with many great qualities (organized, attentive, intelligent but relatable)"

People

Students
Sam Borstein
Jenn Bosco
Katie Massana
Orlando Schwery

People

Students
Sam Borstein
Jenn Bosco
Katie Massana
Orlando Schwery

Sarah Depper

Conflict
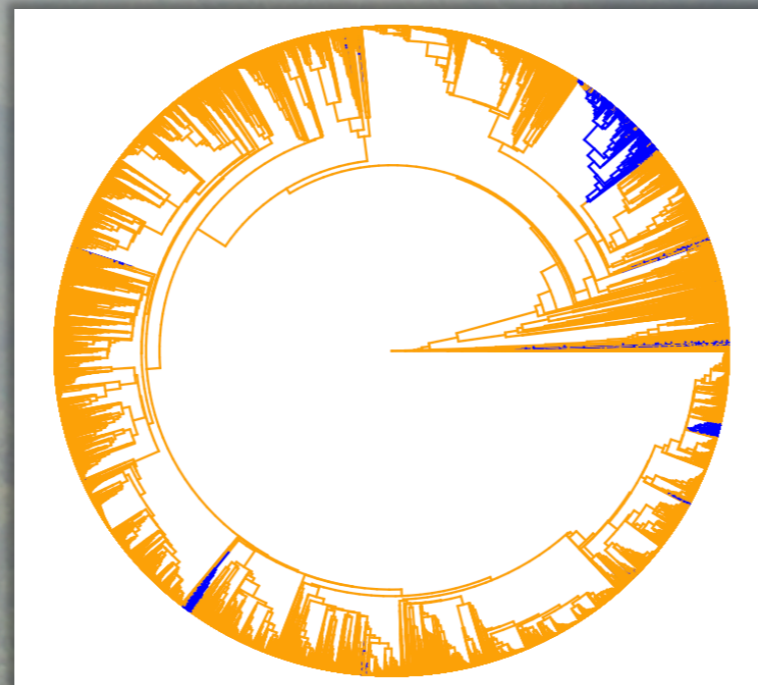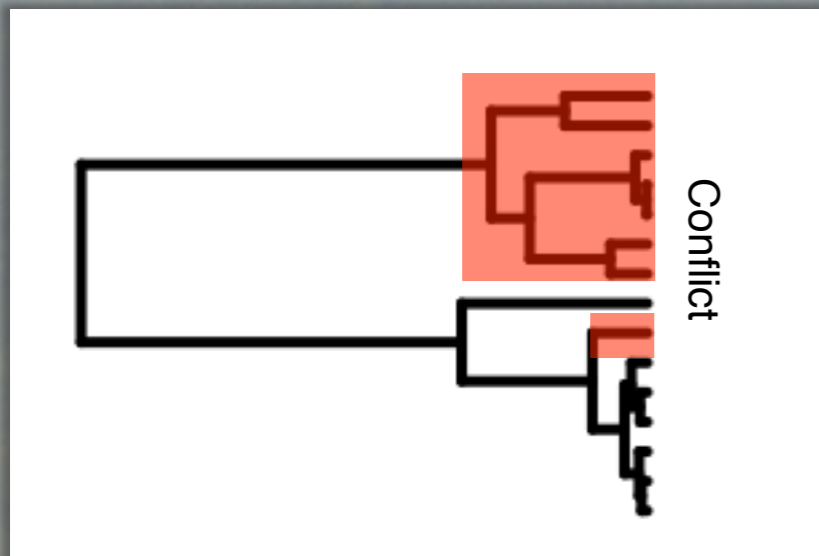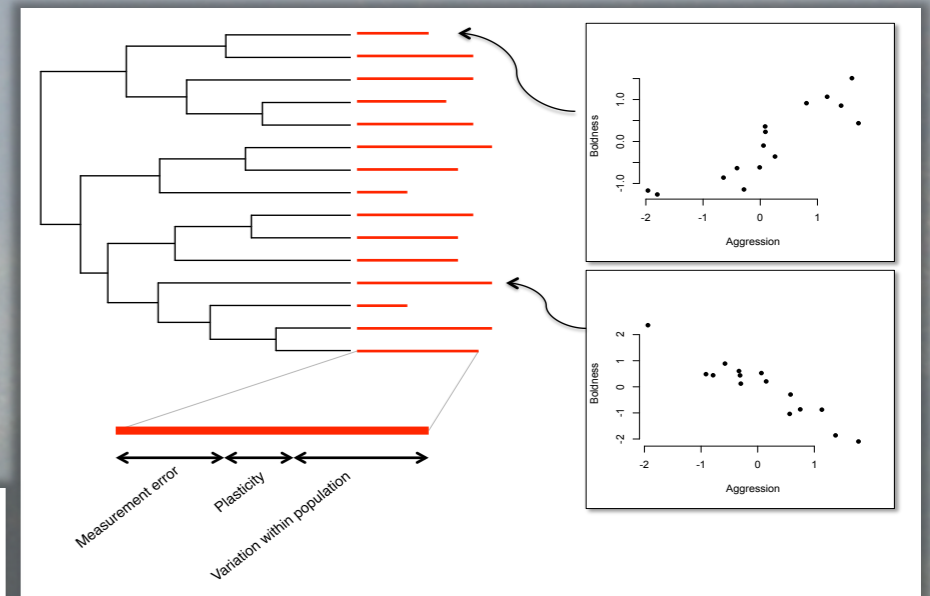
People

Students
Sam Borstein
Jenn Bosco
Katie Massana
Orlando Schwery

Postdocs
Hugo Alamillo
Barb Banbury
Jeremy Beaulieu*
JJ Chai*
Nathan Jackson
Tony Jhwueng*
Sandy Kawano*
Michelle Lawing*
Ryan Martin*
Nick Matzke*

* = NIMBioS postdocs I co-mentored, about 19% of all NIMBioS postdocs

EEB Network,
w/out O'Meara
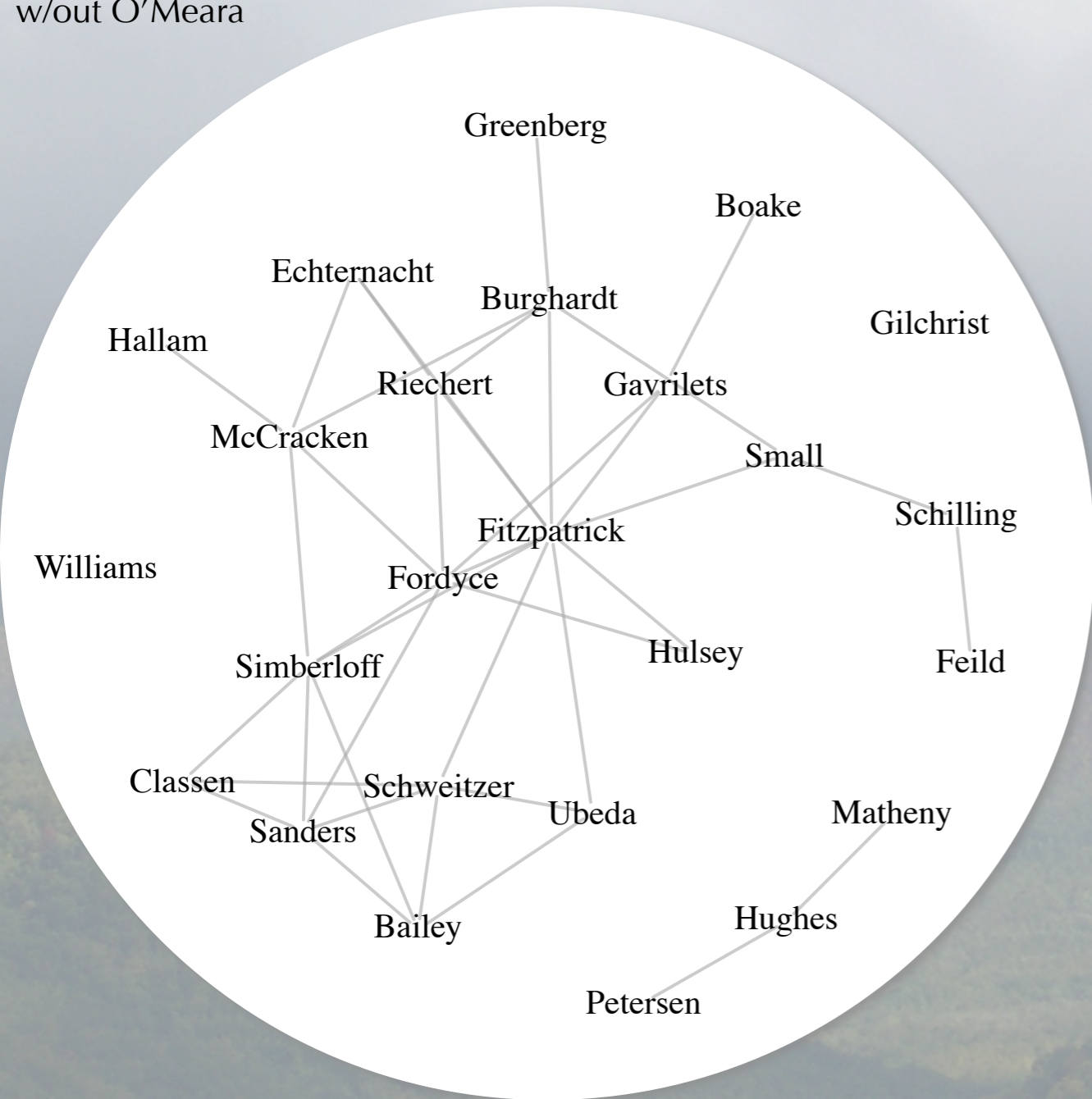
Greenberg

Boake

Echternacht
Burghardt
Gilchrist

Hallam
Riechert
Gavrilets

McCracken
Small

Schilling

Williams
Fitzpatrick

Fordyce

Hulsey
Feild

Simberloff

Classen
Schweitzer
Ubeda
Matheny

Sanders

Bailey
Hughes

Petersen

Networks includes all EEB faculty from 2008 – present who
are connected to at least one other member on a network

EEB Network, w/out O'Meara

EEB Network, with O'Meara
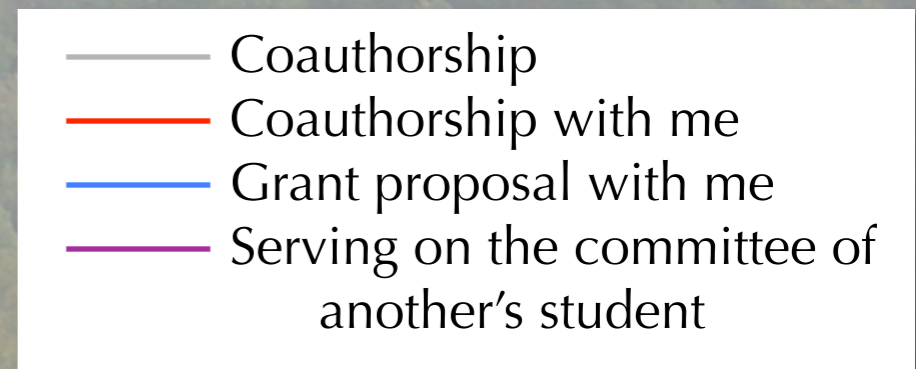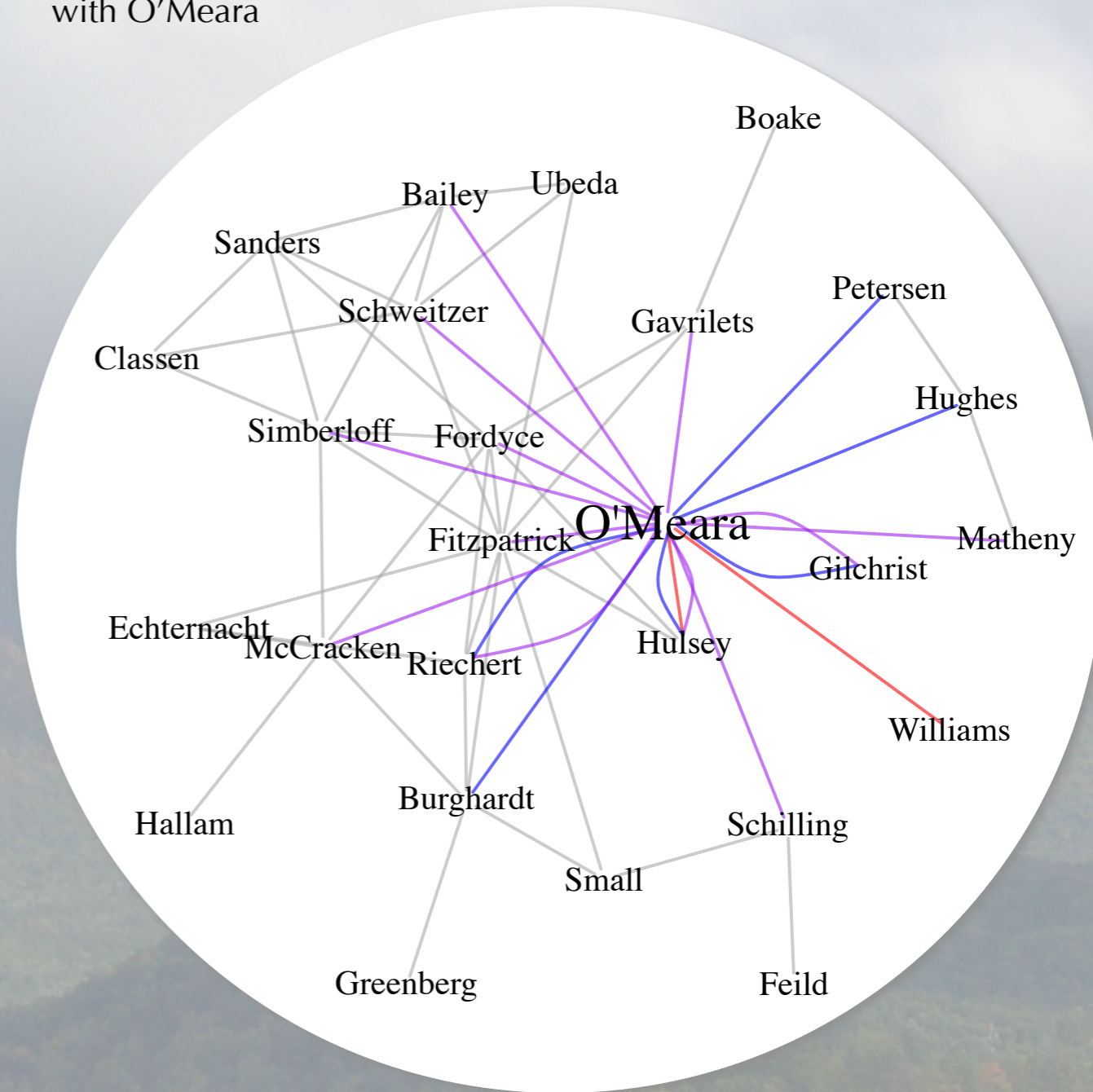
Legend:
- Coauthorship
- Coauthorship with me
- Grant proposal with me
- Serving on the committee of another's student

Networks includes all EEB faculty from 2008 – present who are connected to at least one other member on a network

Service & Outreach

- Co-Organizer, Evolution 2014 (2000 participants)
- Lightning talk organizer, Evolution 2013
- Phylotastic Leadership Team (hackathons)
- iEvoBio Leadership (satellite conference)
- Elected member of Society of Systematic Biologists council

http://www.comicvine.com/forums/battles-7/akatsuki-vs-jedi-council-1451424/

Service & Outreach

- Co-Organizer, Evolution 2014 (2000 participants)
- Lightning talk organizer, Evolution 2013
- Phylotastic Leadership Team (hackathons)
- iEvoBio Leadership (satellite conference)
- Elected member of Society of Systematic Biologists council
- Darwin Day Tennessee advisor
- Applications Editor, *Methods in Ecology and Evolution*
- Reviewer for *Science*, NSF, *Systematic Biology*, and numerous other journals
- Maintainer of CRAN Task View for Phylogenetics in R
- EEB Dept. web committee, head search committee, undergrad affairs, grad admissions, Dean's Advisory Council, many grad student committees in four departments

Service & Outreach

- Co-organized Women in Science seminar series in first year at UTK, got funds from Haines-Morris and buy-in from multiple departments

Service & Outreach

- Co-organized Women in Science seminar series in first year at UTK, got funds from Haines-Morris and buy-in from multiple departments
- Highlighted twice in *Science Careers* for work helping grad students prepare for NSF proposals during federal shutdown

Service & Outreach

- Co-organized Women in Science seminar series in first year at UTK, got funds from Haines-Morris and buy-in from multiple departments
- Highlighted twice in *Science Careers* for work helping grad students prepare for NSF proposals during federal shutdown

Fortunately, there are workarounds. Start at the Web site of Brian O'Meara, an assistant professor at the University of Tennessee, Knoxville, in the Department of Ecology and Evolutionary Biology. O'Meara has located and stashed away some NSF forms likely to be in frequent demand: NSF's Grant Proposal Guide; the program solicitation for Directorate for Biological Sciences Doctoral Dissertation Improvement Grants (Due date: 10 October); and the program solicitation for the crosscutting Graduate Research Fellowship Program, which has due dates in early November.

Email Article
Email Editor
Discuss in Forum
Related Articles
Print Article
Free Newsletter
BOOKMARK

Government Shutdown

The Latest Shutdown Information for NIH- and NSF-Funded Researchers (UPDATE)

By Jim Austin
September 30, 2013

Do you have new shutdown-related information that's relevant to NSF- or NIH-funded investigators or those seeking funding from the major agencies? Send us an email.

On Friday, 27 September, we described the likely impact on scientists funded by the National Institutes of Health (NIH) and the National Science Foundation (NSF) if the government entered a period of partial shutdown. That indeed came to pass the following Monday, at midnight. Here's what we've learned since then:

10/06/2013

- More on accessing information from NSF's Web site--this via Twitter from Brian O'Meara, who, as we mentioned on Wednesday, had archived some critical files on his Web site. O'Meara tweets,

  Internet Archive to the rescue: http://t.co/MSvmF5A6gi .@SciCareerEditor #shutdown .@NSF

  — Brian O'Meara (@omearabrian) October 5, 2013

  You may need to be patient because it's slow, but that should get you to most of what you need from NSF. (I've been using the Wayback Machine forever, so I'm surprised I didn't think of that earlier. Thanks Brian!)

10/04/2013

Service & Outreach

- Co-organized Women in Science seminar series in first year at UTK, got funds from Haines-Morris and buy-in from multiple departments
- Highlighted twice in *Science Careers* for work helping grad students prepare for NSF proposals during federal shutdown

Fortunately, there are workarounds. Start at the Web site of Brian O'Meara, an assistant professor at the University of Tennessee, Knoxville, in the Department of Ecology and Evolutionary Biology. O'Meara has located and stashed away some NSF forms likely to be in frequent demand: NSF's Grant Proposal Guide; the progr...

date:

whicl...

10/06/2013

- More on accessing information from NSF's Web site--this via Twitter from Brian O'Meara, who, as we mentioned on Wednesday, had archived some critical files on his Web site. O'Meara tweets,

Internet Archive to the rescue: http://t.co/MSvmF5A6gi .@SciCareerEditor #shutdown .@NSF

— Brian O'Meara (@omearabrian) October 5, 2013

You may need to be patient because it's slow, but that should get you to most of what you need from NSF. (I've been using the Wayback Machine forever, so I'm surprised I didn't think of that earlier. Thanks Brian!)

been using the Wayback Machine forever, so I'm surprised I didn't think of that earlier. Thanks Brian!)

Service & Outreach

- Co-organized Women in Science seminar series in first year at UTK, got funds from Haines-Morris and buy-in from multiple departments
- Highlighted twice in *Science Careers* for work helping grad students prepare for NSF proposals during federal shutdown
- A representative from SSB on joint *American Society of Naturalists*, *Society for the Study of Evolution*, and *Society of Systematic Biologists* group to help gather data for proposal to improve representation of women in those societies
- Co-organizer of symposium at Evolution 2015 on "Women in Science: Pairing Advances in the Theory and Application of Phylogenetic Methods"
- Outreach via social media and web (i.e., Google "Akaike weight")

Research: Acknowledgments

Research: Acknowledgments

NSF (Co-PI): $141K

NSF (PI): $340K

EOL (Co-PI): $50K

NESCent: 2 WG

iPlant: $369K

NIMBioS: 3 WG + colleagues

Google $10K

NSF (PI): $520K
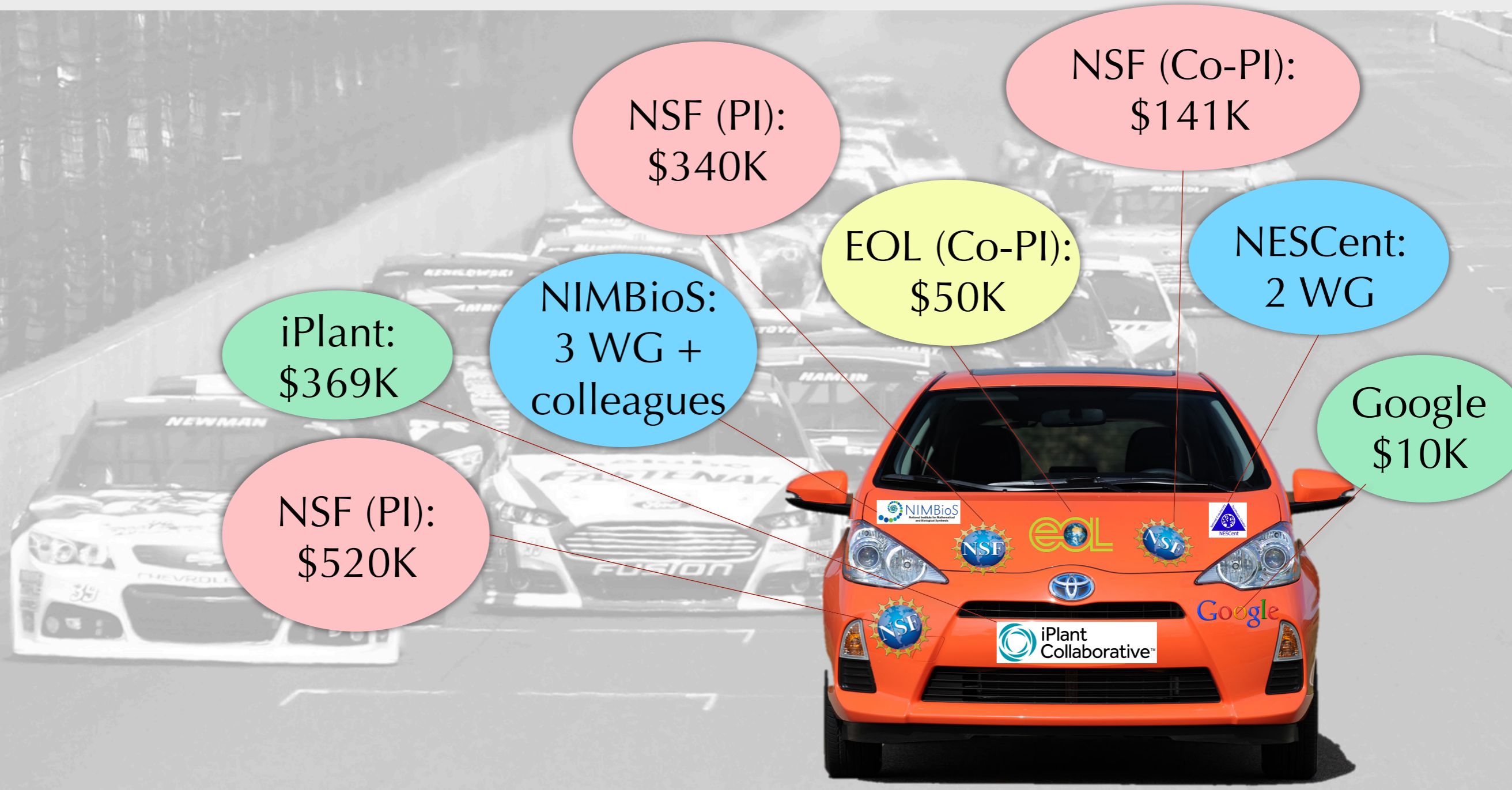
Continuous trait evolution

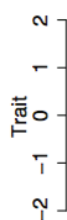384   B. C. O'Meara and J. M. Beaulieu
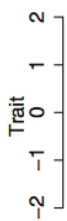
(a) Brownian motion
(b) OU, one sigma, one theta, one alpha
(c) OU, one sigma, two theta, one alpha
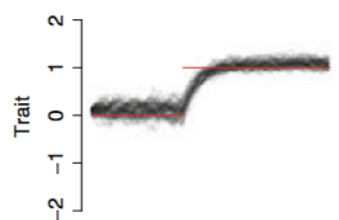(d) OU, two sigma, two theta, one alpha
(e) OU, one sigma, one theta, two alpha
(f) OU, one sigma, two theta, two alpha
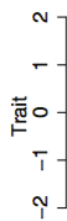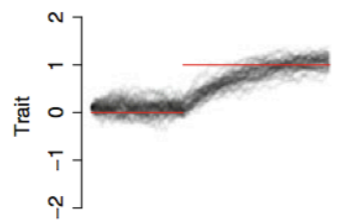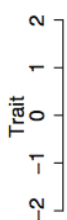
Table 2. The fit of alternative models of genome size evolution in monocots. The best model, based on ΔAIC and Akaike weights, was the OU_{MVA}, which estimated a separate θ, α, and σ^2 for woody and herbaceous monocot lineages.

Model	-lnL	AIC	ΔAIC	w_i
BM1	−227.6	459.1	156.2	<0.01
BMS	−203.0	412.0	109.1	<0.01
OU1	−160.0	326.1	23.2	<0.01
OU_M	−159.2	326.5	23.6	<0.01
OU_{MV}	−147.3	304.6	1.7	0.290
OU_{MA}	−159.2	328.5	25.6	<0.01
OU_{MVA}	−145.3	302.9	0.0	0.678

Cited 249 times

Cited 63 times

Continuous trait evolution, ongoing

Brownian motion on phylogenetic networks, not trees
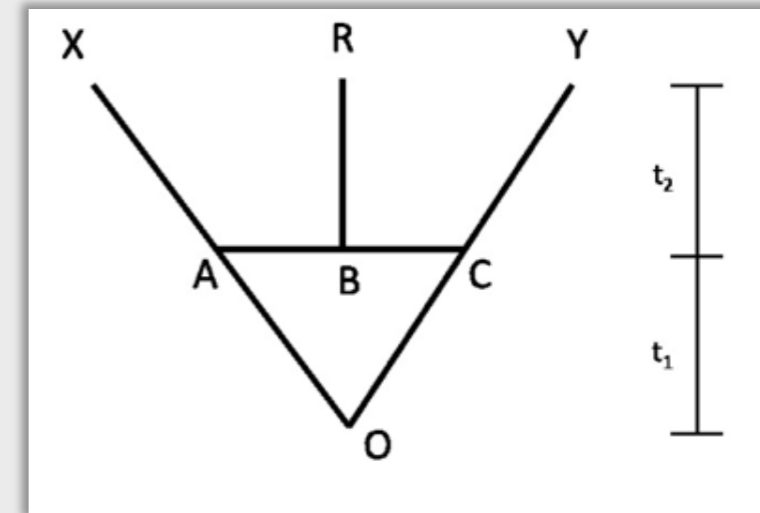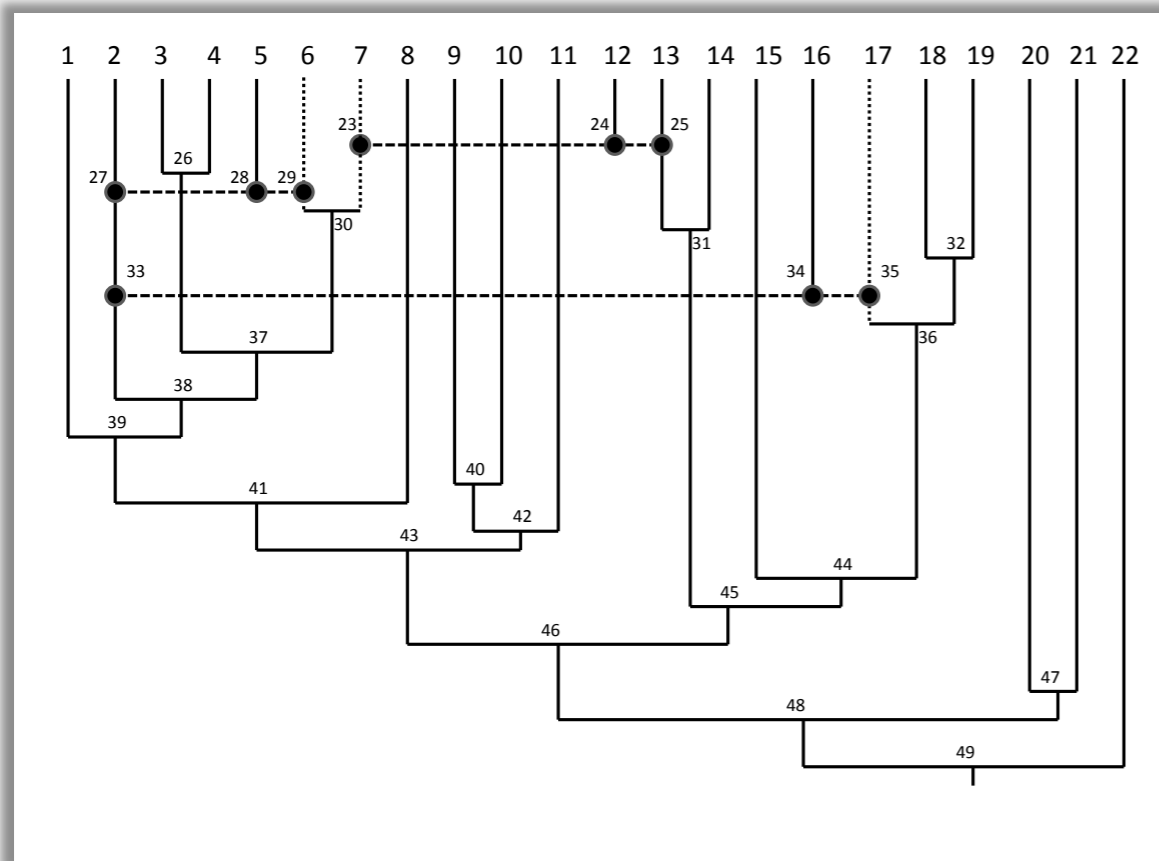
$$V_R = \begin{array}{c} \\ X \\ R_1 \\ Y \\ R_2 \\ Z \end{array} \begin{array}{ccccc} X & R_1 & Y & R_2 & Z \\ \left[\begin{array}{ccccc} 5\sigma^2 & 2\beta_1\sigma^2 & 0 & 0 & 0 \\ 2\beta_1\sigma^2 & (3+4\beta_1^2)\sigma^2 + v_{1,H} & 2\beta_1\sigma^2 & 2\beta_1\beta_2\sigma^2 & \beta_1\sigma^2 \\ 0 & 2\beta_1\sigma^2 & 5\sigma^2 & 3\beta_2\sigma^2 & \sigma^2 \\ 0 & 2\beta_1\beta_2\sigma^2 & 3\beta_2\sigma^2 & (2+5\beta_1^2)\sigma^2 + v_{2,H} & 3\beta_2\sigma^2 \\ 0 & \beta_1\sigma^2 & \sigma^2 & 3\beta_2\sigma^2 & 5\sigma^2 \end{array} \right] \end{array}$$

Jhwueng & O'Meara, in revision

Species delimitation
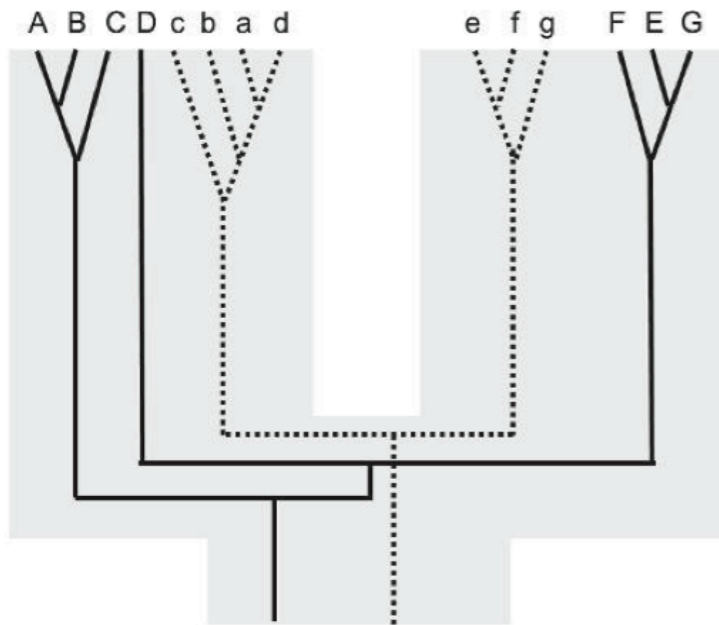
Total cost = (1-*wt*) • GTP cost + *wt* • structure

Total cost = 0.5 • (1 + 0) + 0.5 • (1.009 + 0)

Total cost = 1.0045

structure

$$((A,B),C) = \frac{((a,d),b)}{((a,d),c)} \atop {((a,b),c) \atop ((b,d),c)}$$

$((e,f),g) \neq ((E,G),F)$

$((1/0.33143)-1)/2 = \mathbf{1.009}$

$((1/1)-1)/2 = \mathbf{0}$

gene tree parsimony (GTP)

1 duplication (solid) + **0** duplications (dotted)

New Heuristic Methods for Joint Species Delimitation and Species Tree Inference

BRIAN C. O'MEARA*

Department of Ecology & Evolutionary Biology, University of Tennessee, 569 Dabney Hall, Knoxville, TN 37996-1610, USA;
**Correspondence to be sent to: Department of Ecology & Evolutionary Biology, University of Tennessee, 569 Dabney Hall, Knoxville, TN 37996-1610, USA; E-mail:bomeara@utk.edu.*

Abstract.—Species delimitation and species tree inference are difficult problems in cases of recent divergence, especially when different loci have different histories. This paper quantifies the difficulty of jointly finding the division of samples to species and estimating a species tree without constraining the possible assignments a priori. It introduces a parametric and a nonparametric method, including new heuristic search strategies, to do this delimitation and tree inference using individual gene trees as input. The new methods were evaluated using thousands of simulations and 4 empirical data sets. These analyses suggest that the new methods, especially the nonparametric one, may provide useful insights for systematists working at the species level with molecular data. However, they still often return incorrect results. [Brownie; gene tree parsimony; gene tree species tree; speciation; species delimitation.]

Two of the main goals of systematics are dividing the diversity of life into species and discovering the phylogenetic relationships of these species. Both can be difficult to achieve. Processes such as lineage sorting, introgression, and undetected gene duplication may cause gene trees to disagree with the true tree of species, potentially obscuring the species tree signal (Fitch 1970; Goodman et al. 1979; Avise 1983; Tajima 1983; Pamilo and Nei 1988; Doyle 1992; Hudson 1992; Maddison 1997). For species delimitation, a systematist must choose both a species concept and a criterion to apply this species concept to data. Even if speciation itself is effectively instantaneous, the time required for sufficient evolutionary changes to appear to allow 2 distinct lineages to be recognized will not be (De Queiroz 2007). This causes delimitation of species to be difficult.

These two questions are biologically linked but rarely methodologically coupled. If intervals between speciation events were long enough that all species were monophyletic for all their genes, once the species were correctly delimited, any species could be adequately represented by a single individual on a phylogeny. In reality, putatively independently evolving lineages are often not monophyletic (Funk and Omland 2003). The phylogeny of species, unless they are defined under a strict genealogical species concept (GSC; Baum and Shaw 1995; Hudson and Coyne 2002), will have an assortment of independent evolutionary lineages, which will probably include paraphyly for at least some of their genes. Here, I attempt to unite these two questions as the more general one of jointly inferring the species boundaries and the species tree. I calculate the computational complexity of the problem, develop and implement methods for addressing it, and perform simulations and analyses across hundreds of parameter combinations to evaluate the feasibility. I also analyse 4 empirical data sets, *Drosophila* (Machado et al. 2002; Machado and Hey 2003), *Manacus* (Passeriformes) (Brumfield et al. 2008), *Lactarius* fungi (Nuytinck and Verbeken 2007), and *Melanoplus* grasshoppers (Carstens and Knowles 2007), to evaluate the performance of the new methods.

MATERIALS AND METHODS

Problem Definition

Given a set of sequences from multiple individuals, the general problem is to allocate those individuals into putative species and estimate the species tree. This solution, the species tree with assignment of samples to species, is termed the "delimited species tree." Optimally, a method will assign species and estimate the species tree correctly, in a statistically and computationally efficient manner. An estimate of the delimited species tree may differ from the true delimited species tree in topological error and/or through assignment of individuals to the wrong species. The latter might happen by merging 2 species that should be 1, splitting 1 true species into 2, having an individual of 1 species assigned to a different species, or a complex mixture of these. This is a more difficult problem than is typically addressed in DNA barcoding approaches (Hebert et al. 2003; Tautz et al. 2003), where 1 or more unknown individuals are assigned to existing species (Manel et al. 2005; Matz and Nielsen 2005; Abdo and Golding 2007; Zhang et al. 2008).

Most methods in systematics work on restrictions of this general problem, such as assuming that assignments to species are known (Nielsen and Wakeley 2001; Carstens and Knowles 2007; Edwards et al. 2007; Liu and Pearl 2007) or assuming that the gene tree matches the species tree (Hebert et al. 2003; Pons et al. 2006). I follow the approach of several recent authors (Pons et al. 2006; Carstens and Knowles 2007; Knowles and Carstens 2007; Mossel and Roch 2007; Kubatko et al. 2009), in restricting the problem using estimated gene trees as input rather than by integrating across a set of possible gene trees. The restricted problem still makes

59

Cited 135 times

Hidden rates

woody,slow → herb,slow

woody,med → herb,med

woody,fast ⇄ herb,fast

corHMM: Analysis of binary character evolution

Fits a hidden rates model that allows different transition rate classes on different portions of a phylogeny by treating rate classes as hidden states in a Markov process and various other functions for evaluating models of binary character evolution.

Version:	1.15
Depends:	ape, nloptr
Imports:	expm, numDeriv, corpcor, phangorn, rgenoud, parallel
Published:	2014-09-18
Author:	Jeremy M. Beaulieu, Jeffrey C. Oliver, Brian O'Meara

Chapter 16
Hidden Markov Models for Studying the Evolution of Binary Morphological Characters

Jeremy M. Beaulieu and Brian C. O'Meara

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

REPEATED EVOLUTION OF TRICELLULAR (AND BICELLULAR) POLLEN

Empirical papers

PHYLOGENETIC RELATIONSHIPS AND CHARACTER EVOLUTION ANALYSIS OF SAXIFRAGALES USING A SUPERMATRIX APPROACH[1]

Douglas E. Soltis[2,3,7], Mark E. Mort[4], Maribeth Latvis[2], Evgeny V. Mavrodiev[3], Brian C. O'Meara[5], Matthew W. Pennell[6], Daniel J. McGlinn[6], Jonathan M. Eastman[6],

---

Mitochondrial genome primers for Lake Malawi cichlids

C. DARRIN HULSEY, BENJAMIN P. KECK, HUGO ALAMILLO and BRIAN C. O'MEARA
Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA

---

100 YEARS | **Journal of Ecology**

SPECIAL FEATURE
THE TREE OF LIFE IN ECOSYSTEMS

Functional distinctiveness of major plant lineages

William K. Cornwell[1,2], Mark Westoby[3], Daniel S. Falster[3], Richard G. FitzJohn[3,4], Brian C. O'Meara[5], Matthew W. Pennell[6], Daniel J. McGlinn[6], Jonathan M. Eastman[6], Angela T. Moles[2], Peter B. Reich[8], David C. Tank[7], Ian J. Wright[3], Lonnie Aarssen[9], Jeremy M. Beaulieu[10], Robert M. Kooyman[3,11], Michelle R. Leishman[3], Eliot T. Miller[12], Ülo Niinemets[13], Jacek Oleksyn[14], Alejandro Ordonez[15], Dana L. Royer[16], Stephen A. Smith[17], Peter F. Stevens[12], Laura Warman[2,18], Peter Wilf[19] and Amy E. Zanne[20,21*]

Summary

1. Plant traits vary widely across species and underpin differences in ecological strategy. Despite centuries of interest, the contributions of different evolutionary lineages to modern-day functional diversity remain poorly quantified.

2. Expanding data bases of plant traits plus rapidly improving phylogenies enable for the first time a data-driven global picture of plant functional diversity across the major clades of higher plants. We mapped five traits relevant to metabolism, resource competition and reproductive strategy onto a phylogeny across 48324 vascular plant species world-wide, along with climate and biogeographic data. Using a novel metric, we test whether major plant lineages are functionally distinctive. We then highlight the trait–lineage combinations that are most functionally distinctive within the present-day spread of ecological strategies.

3. For some trait–clade combinations, knowing the clade of a species conveys little information to neo- and palaeo-ecologists. In other trait–clade combinations, the clade identity can be highly revealing, especially informative clade–trait combinations include Proteaceae, which is highly distinctive, representing the global slow extreme of the leaf economic spectrum. Magnoliidae and Rosidae contribute large leaf sizes and seed masses and have distinctively warm, wet climatic distributions.

*Correspondence author. E-mail: aezanne@gmail.com

EvoDevo

RESEARCH   Open Access

Developmental evolution of flowering plant pollen tube cell walls: callose synthase (CalS) gene expression

Jason M Abercromb

Abstract

Background: A nu...
early step involved...
distance transport...
and septae (callose...
12 CalS gene famil...
orthologues are pre...
origin of CalS5 or w...

Results: We investi...
(gymnosperms) and...
determined the nea...
(Nymphaeales). Sen...
but strong expressi...
Nymphaeales and...
genomes indicated...
Amborella. Among...
but a novel CalS pa...

Conclusion: The fir...
and model system...
callose plugs. The c...
plants and was the c...
of extant angiosper...

Keywords:

Habitat use affects morphological diversification in dragon lizards

D. C. COLLAR*, J. A. SCHULTE II†, B. C. O'MEARA‡ & J. B. LOSOS*
*Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA
†Department of Biology, Clarkson University, Potsdam, NY, USA
‡Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, USA

Keywords:
Agamidae;
Brownian motio...
ecomorphology;
Iguania;
locomotion;
phylogenetic co...

LETTER

Three keys to the radiation of angiosperms into freezing environments

Amy E. Zanne[1,2], David C. Tank[3,4], William K. Cornwell[5,6], Jonathan M. Eastman[3,4], Stephen A. Smith[7], Richard G. FitzJohn[8,9], Daniel J. McGlinn[10], Brian C. O'Meara[11], Angela T. Moles[6], Peter B. Reich[12,13], Dana L. Royer[14], Douglas E. Soltis[15,16,17], Peter F. Stevens[18], Mark Westoby[5], Ian J. Wright[5], Lonnie Aarssen[19], Robert I. Bertin[20], Andre Calaminus[15], Rafaël Govaerts[21], Frank Hemmings[6], Michelle R. Leishman[5], Jacek Oleksyn[12,22], Pamela S. Soltis[16,17], Nathan G. Swenson[23], Laura Warman[6,24] & Jeremy M. Beaulieu[25]

Early flowering plants are thought to have been woody species restricted to warm habitats[1–3]. This lineage has since radiated into almost every climate, with manifold growth forms[4]. As angiosperms spread and climate changed, they evolved mechanisms to cope with episodic freezing. To explore the evolution of traits underpinning the ability to persist in freezing conditions, we assembled a large species-level database of growth habit (woody or herbaceous; 49,064 species), as well as leaf phenology (evergreen or deciduous), diameter of hydraulic conduits (that is, xylem vessels and tracheids) and climate occupancies (exposure to freezing). To model the evolution of species' traits and climate occupancies, we combined these data with an unparalleled dated molecular phylogeny (32,223 species) for land plants. Here we show that woody clades successfully moved into freezing-prone environments by either possessing transport networks of small safe conduits[5] and/or shutting down hydraulic function by dropping leaves during freezing. Herbaceous species largely avoided freezing periods by senescing cheaply constructed aboveground tissue. Growth habit has long been considered labile[6], but we find that growth habit was less labile than climate occupancy. Additionally, freezing environments were largely filled by lineages that had already become herbs or, when remaining woody, already had small conduits (that is, the trait evolved before the climate occupancy). By contrast, most deciduous woody lineages had an evolutionary shift to seasonally shedding their leaves only after exposure to freezing (that is, the climate occupancy evolved before the leaves). For angiosperms to inhabit novel cold environments they had to gain new structural and functional trait solutions; our results suggest that many of these solutions were probably acquired before their foray into the cold.

Flowering plants (angiosperms) today grow in a vast range of environmental conditions, with this breadth probably related to their diverse morphology and physiology[7]. However, early angiosperms are generally thought to have been woody and restricted to warm understorey habitats[1–3]. Debate continues about these assertions, in part because of the paucity of fossils and uncertainty in reconstructing habits for these first representatives[8–11]. Nevertheless, greater mechanical strength of woody tissue would have made extended lifespans possible at a height necessary to compete for light[12,13]. A major challenge resulting from increased stature is that hydraulic systems must deliver water at tension

to greater heights: as path lengths increase so too does resistance[5]. Among extant strategies, the most efficient method of water delivery is through large-diameter water-conducting conduits (that is, vessels and tracheids) within xylem[5].

Early in angiosperm evolution they probably evolved larger conduits for water transport, especially compared with their gymnosperm cousins[14]. Although efficient in delivering water, these larger cells would have impeded angiosperm colonization of regions characterized by episodic freezing[4,15], as the propensity for freezing-induced embolisms (air bubbles produced during freeze/thaw events that block hydraulic pathways)

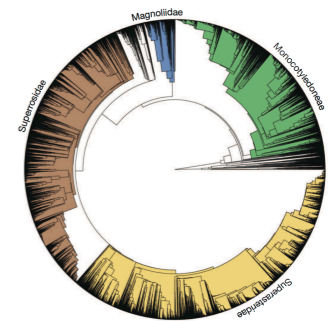**Figure 1 | Time-calibrated maximum-likelihood estimate of the molecular phylogeny for 31,749 species of seed plants.** The four major angiosperm lineages discussed in the text are highlighted: Monocotyledoneae (green), Magnoliidae (blue), Superrosidae (brown) and Superasteridae (yellow). Non-seed plant outgroups (that is, bryophytes, lycophytes and monilophytes) were removed for the purposes of visualization.

Scaling methods for large trees

Left paper

treePL: divergence time estimation using penalized likelihood for large phylogenies

Stephen A. Smith[1,*] and Brian C. O'Meara[2]

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, 48109 and [2]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, 37996 USA

Associate Editor: David Posada

ABSTRACT

Summary: Ever larger phylogenies are being constructed due to the explosion of genetic data and development of high-performance phylogenetic reconstruction algorithms. However, most methods for calculating divergence times are limited to datasets that are orders of magnitude smaller than recently published large phylogenies. Here, we present an algorithm and implementation of a divergence time method using penalized likelihood that can handle datasets of thousands of taxa. We implement a method that combines the standard derivative-based optimization with a stochastic simulated annealing approach to overcome optimization challenges. We compare this approach with existing software including r8s, PATHd8 and BEAST.

Availability: Source code, example files, binaries and documentation for treePL are available at https://github.com/blackrim/treePL.

Contact: eebsmith@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 14, 2012; revised on July 31, 2012; accepted on August 1, 2012

1 INTRODUCTION

Divergence times make phylogenetic hypotheses easier to interpret in light of other information on geology, biogeography and co-diversification. A number of methods exist for transforming branch lengths to be proportional to time. However, many of these methods do not scale well for phylogenies with thousands of taxa. The number of species included in published phylogenies has exploded (Goloboff *et al.*, 2009; Smith *et al.*, 2009). With the increase in the size of phylogenies, there is the need to develop divergence time methods that are capable of handling larger datasets. Here, we present a method, implemented in the software treePL, for calculating divergence time estimates using penalized likelihood (Sanderson, 2002) on large phylogenies. Penalized likelihood uses a semi-parametric approach that allows for different rates on different branches but has a smoothing parameter, set using cross-validation, that affects how much rate differences over the tree are penalized. Our approach attempts to overcome the problem of local minima, a problem amplified by including more taxa.

*To whom correspondence should be addressed.

2 METHODS

2.1 Penalized likelihood

The optimality criterion used in this article is the penalized likelihood framework described in Sanderson (2002). Note that though this is framed as a likelihood method, it is possible to interpret this as a maximum *a posteriori* estimate in a Bayesian context given particular priors (Thorne and Kishino, 2005). It is also possible, following Sanderson's modifications to r8s, to use a related penalty that uses the differences of log rates rather than differences of untransformed rates (the additive penalty). This has been implemented in treePL, but analyses below use the original additive penalty.

2.2 The algorithm

Divergence time estimation, and penalized likelihood especially, presents a number of optimization challenges. One challenge is the large number of parameters and the ratio of free parameters to observations. This can be dampened by a large penalty function, but still presents a parametric optimization problem. The second challenge is the presence of barriers placed by both the user and the hierarchical nature of phylogenies. We present a first step to overcoming these problems with extremely large phylogenies. Our algorithm includes two main phases: a greedy hill-climbing phase and a stochastic phase. This combination is meant to both speed optimization and avoid local optima.

The greedy phase of the optimization step consists of two gradient-based optimization rounds. The first round of optimization uses gradient values from an exact gradient calculator with math derived from Sanderson (2002). The second round of optimization uses auto-differentiation (Gay, 2005; Griewank *et al.*, 1996). Auto-differentiation, because it has the same complexity as the original function, often continues hill climbing after the exact gradient calculator fails. However, we have found, empirically, that alone it is not as efficient and requires more rounds and restarts than combining exact gradients and auto-differentiation.

We found that a stochastic phase, in this case a partial simulated annealing procedure, is helpful in easing the parameters away from local optima. We consider it partial as the step does not consist of a full simulated annealing run. Instead, the phase consists of thousands of stochastic optimization generations followed by intermittent greedy phases. Although this additional step may lengthen convergence time, the result will often improve.

The performance of optimization methods can vary dramatically by dataset. We have provided a 'priming' step that will run through each optimization option and report those that show the greatest difference in starting and stopping likelihood scores. Although this can help, researchers should replicate analyses and experiment with settings to assure convergence. We have also provided a 'wiggle' option that will report divergence times that are within two log likelihood units of the best likelihood, suggested by Edwards (1992) as a confidence window. This allows for the visualization of nodes that are particularly uncertain, but

Right paper

The iPlant collaborative: cyberinfrastructure for plant biology

Stephen A. Goff[1]*, Matthew Vaughn[2], Sheldon McKay[1], Eric Lyons[1], Ann E. Stapleton[3,4], Damian Gessler[1], Naim Matasci[1], Liya Wang[5], Matthew Hanlon[2], Andrew Lenards[1], Andy Muir[1], Nirav Merchant[1], Sonya Lowry[1], Stephen Mock[2], Matthew Helmke[1], Adam Kubach[2], Martha Narro[1], Nicole Hopkins[1], David Micklos[6], Uwe Hilgert[6], Michael Gonzales[2], Chris Jordan[2], Edwin Skidmore[1], Rion Dooley[2], John Cazes[2], Robert McLay[2], Zhenyuan Lu[5], Shiran Pasternak[5], Lars Koesterke[2], William H. Piel[7], Ruth Grene[8], Christos Noutsos[5], Karla Gendler[2], Xin Feng[5,9], Chunlao Tang[5], Monica Lent[1], Seung-Jin Kim[1], Kristian Kvilekval[10], B. S. Manjunath[10,27], Val Tannen[11], Alexandros Stamatakis[12], Michael Sanderson[13], Stephen M. Welch[14], Karen A. Cranston[15], Pamela Soltis[16], Doug Soltis[17], Brian O'Meara[18], Cecile Ane[19,20], Tom Brutnell[21], Daniel J. Kleibenstein[22], Jeffery W. White[23], James Leebens-Mack[24], Michael J. Donoghue[25], Edgar P. Spalding[26], Todd J. Vision[28], Christopher R. Myers[32], David Lowenthal[29], Brian J. Enquist[13], Brad Boyle[13], Ali Akoglu[30], Greg Andrews[29], Sudha Ram[31], Doreen Ware[5], Lincoln Stein[5,9] and Dan Stanzione[2]

[1] BIO5 Institute, University of Arizona, Tucson, AZ, USA
[2] Texas Advanced Computer Center, University of Texas, Austin, TX, USA
[3] Department of Biology, University of North Carolina, Wilmington, NC, USA
[4] Department of Marine Sciences, University of North Carolina, Wilmington, NC, USA
[5] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[6] DNA Learning Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[7] Yale Peabody Museum, Yale University, New Haven, CT, USA
[8] Department of Plant Pathology, Physiology and Weed Science, Virginia Tech University, Blacksburg, VA, USA
[9] Ontario Center for Cancer Research, Toronto, ON, Canada
[10] Center for Bio-image Informatics, University of California, Santa Barbara, CA, USA
[11] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA
[12] Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[13] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA
[14] Department of Agronomy, Kansas State University, Manhattan, KS, USA
[15] National Evolutionary Synthesis Center (NESCent), Durham, NC, USA
[16] Florida Museum of Natural History, University of Florida, Gainesville, FL, USA
[17] Department of Biology, University of Florida, Gainesville, FL, USA
[18] Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, USA
[19] Department of Statistics, University of Wisconsin, Madison, WI, USA
[20] Department of Botany, University of Wisconsin, Madison, WI, USA
[21] Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY, USA
[22] Department of Plant Sciences, University of California, Davis, CA, USA
[23] Arid-Land Agricultural Research Center, United States Department of Agriculture-Agricultural Research Service, Maricopa, AZ, USA
[24] Department of Plant Biology, University of Georgia, Athens, GA, USA
[25] Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA
[26] Department of Botany, University of Wisconsin, Madison, WI, USA
[27] Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA
[28] Department of Biology, University of North Carolina, Chapel Hill, NC, USA
[29] Department of Computer Science, University of Arizona, Tucson, AZ, USA
[30] Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA
[31] Eller School of Business, University of Arizona, Tucson, AZ, USA
[32] Department of Physics, Cornell University, Ithaca, NY, USA

Edited by:
Gane Ka-Shu Wong, University of Alberta, Canada

Reviewed by:
Michael Deyholos, University of Alberta, Canada
Sean W. Graham, University of British Columbia, Canada

*Correspondence:
Stephen A. Goff, iPlant Collaborative, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA.
e-mail: sgoff@iplantcollaborative.org

The iPlant Collaborative (iPlant) is a United States National Science Foundation (NSF) funded project that aims to create an innovative, comprehensive, and foundational cyberinfrastructure in support of plant biology research (PSCIC, 2006). iPlant is developing cyberinfrastructure that uniquely enables scientists throughout the diverse fields that comprise plant biology to address Grand Challenges in new ways, to stimulate and facilitate cross-disciplinary research, to promote biology and computer science research interactions, and to train the next generation of scientists on the use of cyberinfrastructure in research and education. Meeting humanity's projected demands for agricultural and forest products and the expectation that natural ecosystems be managed sustainably will require synergies from the application of information technologies. The iPlant cyberinfrastructure design is based on an unprecedented period of research community input, and leverages developments in high-performance computing, data storage, and cyberinfrastructure for the physical sciences. iPlant is an open-source project with application programming interfaces that allow the community to extend the infrastructure to meet its needs. iPlant is sponsoring community-driven workshops addressing specific

Realistic substitution models
Work with Mike Gilchrist and JJ Chai
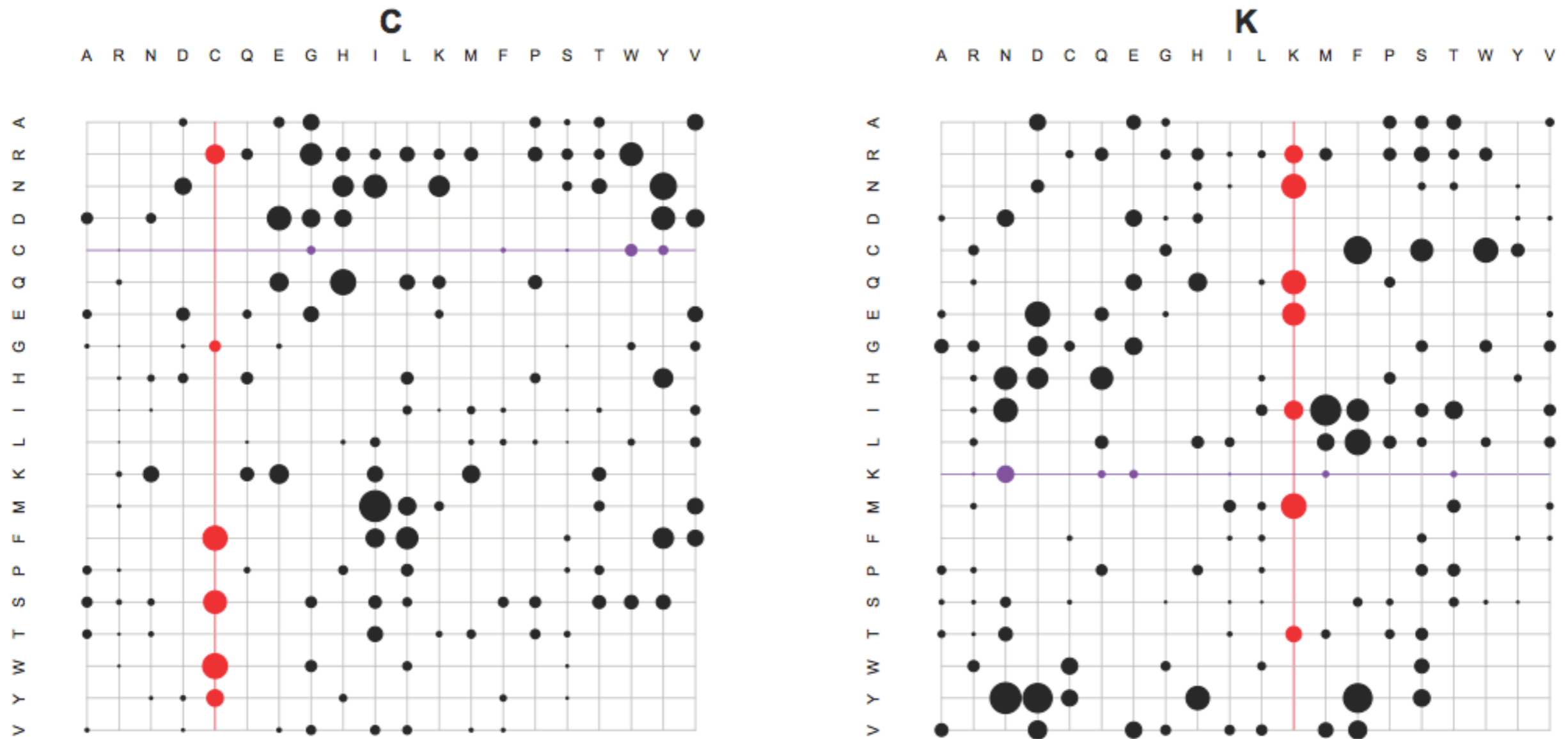
Figure 3: Instantaneous substitution rates from amino acid on row to amino acid in column under our preliminary model. The left panel is when cysteine is the optimal amino acid, the right is when lysine is the optimal amino acid. These are just two of the substitution matrices that are present in a given SAC_{AA} model. Bubble area (not diameter) shows the relative substitution rate from the amino acid listed on the left to the amino acid listed above. Red highlights the substitutions moving towards the optimal amino acid, purple those moving away from the optimal amino acid.

Large scale syntheses of the field

ANNUAL REVIEWS **Further**
Click here for quick links to Annual Reviews content online, including:
• Other articles in this volume
• Top cited articles
• Top downloaded articles
• Our comprehensive search

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

Keywords

methods, continuous-time Markov Chain, multivariate normal, birth-death, tree stretching

Abstract

There are many methods for making evolutionary inferences from phylogenetic trees. Many of these can be divided into three main classes of models: continuous-time Markov chain models with finite state space (CTMC-FSS), multivariate normal models, and birth-death models. Numerous approaches are just restrictions of more general models to focus on particular questions or kinds of data. Methods can be further modified with the addition of tree-stretching algorithms. The recent realization of the effect of correlated trait evolution with diversification rates represents an important advance that is slowly revolutionizing the field. Increased attention to model adequacy may lead to future methodological improvements.

267

ENCYCLOPEDIAS OF THE NATURAL WORLD

ENCYCLOPEDIA OF

THEORETICAL ECOLOGY

EDITED BY ALAN HASTINGS AND LOUIS GROSS

by adaptation in a fixed (nonplastic) genotype. Under this scenario, individuals having genetic variation for phenotypic plasticity will be favored. Whether the population will adapt by increasing its capacity for phenotypic plasticity ultimately depends on the costs and physiological limits associated with the plastic phenotype. Finally, it is worth noting that already in 1881 (in a letter to Karl Semper) Charles Darwin, clearly before his time, was able to envision the possibility of plastic phenotypes evolving in response to environmental changes: "I speculate whether a species very liable to repeated and great changes of conditions might not assume a fluctuating condition ready to be adapted to either condition."

SEE ALSO THE FOLLOWING ARTICLES

Environmental Heterogeneity and Plants / Evolution of Dispersal / Integrated Whole Organism Physiology / Mutation, Selection and Genetic Drift / Quantitative Genetics

FURTHER READING

Agrawal, A. A. 2001. Phenotypic plasticity in the interactions amd evolution of species. *Science* 294: 321–326.
deWitt, T. J., and S. M. Scheiner, eds. 2004. *Phenotypic plasticity: functional and conceptual approaches*. New York: Oxford University Press.
Miner, B. G., S. E. Sultan, S. G. Morgan, D. K. Padilla, and R. A. Releya. 2005. Ecological consequences of phenotypic plasticity. *Trends in Ecology and Evolution* 20: 685–692.

Pigliucci, M. 2001. *Phenotypic plasticity: beyond nature and nurture.* Syntheses in Ecology and Evolution. Baltimore, MD: Johns Hopkins University Press.
Pigliucci, M. 2005. Evolution of phenotypic plasticity: where are we going now? *Trends in Ecology and Evolution* 20: 481–486.
Pfennig, D. W., M. A. Wund, E. C. Snell-Rood, T. Cruickshank, C. D. Schlichting, and P. Moczek. 2010. Phenotypic plasticity's impact on diversification and speciation. *Trends in Ecology and Evolution* 25: 459–467.
Price, T. D., A. Qvarnström, and D. E. Irwin. 2003. The role of phenotypic plasticity in driving genetic evolution. *Proceedings of the Royal Society London B: Biological Sciences* 270: 1433–1440.
Tollrian, R., and C. D. Harvell, eds. 1998. *The ecology and evolution of inducible defenses.* Princeton: Princeton University Press.
West-Eberhard, M-J. 2003. *Developmental plasticity and evolution.* New York: Oxford University Press.

PHYLOGENETIC RECONSTRUCTION

BRIAN C. O'MEARA

University of Tennessee, Knoxville

A phylogeny is a depiction of the evolutionary history of a set of organisms. Typically, this is a branching diagram showing relationships between species, but phylogenies can be drawn for individual genes, for populations, or for other entities.

WHAT DO PHYLOGENIES MEAN?

A phylogeny represents a history of populations. Take the example of Figure 1. Starting at the bottom (root) of the tree, one population splits into two. The population on the left speciates again, but one of the descendant species eventually goes extinct without leaving any descendants. Various other processes occur: population sizes (width of the tree's branches) vary, speciation happens through a gradual rather than instant reduction of gene flow, populations develop and lose subdivision, one species forms as a hybrid of two other species, a few genes introgress from one species to another, and so forth. The history of genes evolving within these populations may be even more complex, with selective sweeps, ancestral polymorphisms persisting across speciation events, gene copies being duplicated and lost within the genome, and recombination shuffling histories within and between genes. All of this complex history is typically summarized by a figure like that of Figure 1, with most of the complex history abstracted away to leave only a simplified history of populations.

It is important to interpret phylogenies correctly. Chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*)

Much more...

BMC Research Notes

DATA NOTE | Open Access

Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis

Arlin Stoltzfus[1*], Brian O'Meara[2], Jamie Whitacre[3], Ross Mounce[4], Emily L Gillespie[5], Sudhir Kumar[6], Dan F Rosauer[7] and Rutger A Vos[8]

Abstract

Dwueng-Chwuan Jhwueng, Snehalata Huzurbazar, Brian C. O'Meara and Liang Liu*

Investigating the performance of AIC in selecting phylogenetic models

Abstract: The popular likelihood-based model selection criterion, Akaike's Information Criterion (AIC), is a breakthrough mathematical result derived from information theory. AIC is an approximation to Kullback-Leibler (KL) divergence with the derivation relying on the assumption that the likelihood function has finite second derivatives. However, for phylogenetic estimation, given that tree space is discrete with respect to tree topology, the assumption of a continuous likelihood function with finite second derivatives is violated. In this paper, we investigate the relationship between the expected log likelihood of a candidate model, and the expected KL divergence in the context of phylogenetic tree estimation. We find that given the tree topology, AIC is an unbiased estimator of the expected KL divergence. However, when the tree topology is unknown, AIC tends to underestimate the expected KL divergence for phylogenetic models. Simulation results suggest that the degree of underestimation varies across phylogenetic models so that even for large sample sizes, the bias of AIC can result in selecting a wrong model. As the choice of phylogenetic models is essential for statistical phylogenetic inference, it is important to improve the accuracy of model selection criteria in the context of phylogenetics.

Keywords: AIC; Kullback-Leibler divergence; model selection; phylogenetics.

DOI 10.1515/sagmb-2013-0048

1 Introduction

Probabilistic models are fundamental to statistical phylogenetic inference (Johnson and Omland, 2004; Sullivan and Joyce, 2005; Kelchner, 2009). A phylogenetic model assumes that the evolution of molecular sequences follows a substitution process along the branches of a phylogenetic tree. The random process of nucleotide substitutions over time is described probabilistically by a substitution model; over the years many such substitution models have been developed. The parameters in a phylogenetic model include the branch lengths and topology of the phylogenetic tree, as well as the parameters in the substitution model (Bos and Posada, 2005).

One of the major goals of phylogenetic model selection is to select a good substitution model for estimating phylogenetic trees from sequence data (Shapiro et al., 2006). Since statistical approaches for phylogenetic inference are based on particular models, model choice may significantly affect the resulting estimates of the phylogenetic parameters (Buckley and Cunningham, 2002; Posada and Buckley, 2004). Standard model selection criteria have been introduced for selecting phylogenetic models, but the biggest challenge in

*Corresponding author: Liang Liu, Department of Statistics and Institute of Bioinformatics, University of Georgia, 101 Cedar Street, Athens, GA 30606 USA, Phone: +1-706-542-3309, Fax: +1-706-542-3391, e-mail: lliu@uga.edu
Dwueng-Chwuan Jhwueng: Department of Statistics, Feng-Chia University, Taichung, Taiwan 40724, USA;
Snehalata Huzurbazar: Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709, USA; Department of Statistics, University of Wyoming, Laramie, WY 82071, USA; and Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
Brian C. O'Meara: Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA

Ecology and Evolution

Open Access

Reol: R interface to the Encyclopedia of Life

Barbara L. Banbury & Brian C. O'Meara

Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee

Abstract

The Encyclopedia of Life is a website that hosts information about life on Earth. Its mission is to increase awareness and understanding of living nature through a freely accessible digital source. Information is publicly available through graphical webpages (browser interface) or through an application programming interface (API). We developed *Reol*, an open-source package for the R environment, which downloads data from the EOL API, searches for and extracts specific information, and builds tables with quantitative data and/or hierarchical classifications. We provide a detailed description how *Reol* can be used as a bridge between the R environment and the EOL API to extract quantitative or hierarchical content. It will be particularly useful for researchers who want information about taxonomic groups of interest (for example: how much

Introduction

The Encyclopedia of Life (h...) cated to preserving informa... developed in 2007 to host... however, it has expanded... resources such as museum... scientists, about all taxon... rankings). It brings togethe... scientists, educators, studen...

Page content comes in ... media, common names, ... built from multiple sources... tions that have agreed to ... org/content_partners) and ... ute hierarchical classificat...

LETTER

Morphogenera, monophyly, and macroevolution

Jablonski and Finarelli (1) suggest that morphogenera, even when they are nonmonophyletic, serve as good representatives for large-scale evolutionary studies. We feel there are two issues that warrant further discussion. First, the test used to evaluate the effect of using nonmonophyletic groups for macroevolutionary studies was not conservative and thus does not provide strong evidence about the impact of nonmonophyly on evolutionary studies. Their test examined whether the median trait value for the species in a nonmonophyletic genus correlated with the median trait value for those same species plus the additional species needed to make the set monophyletic. Correlation of the median value of a set of measurements with the median

value of a superset of those measurements is to be expected (as is briefly mentioned in the methods section of ref. 1). To demonstrate this, we correlated the median values of *ln* body mass from random sets of 3, 7, and 15 mammal species and 0–30 additional species for 45 simulated genera, using data from ref. 2 (Fig. 1A). Even when the median of 3 species is correlated with those 3 plus 15 additional species, the correlation is significant. Moreover, slightly better than random assignment of species to genera increases the expected correlation. To show this, we compare the results of figure 3A from ref. 1 with simulations where 45 genera of 3 species each were created from species chosen randomly from (i) the same family (Fig. 1B) and (ii) the same order (Fig. 1C) (data from refs. 2 and 3). Correlations between morphogenus median and the smallest clade containing the morphogenus species were significant (P < 0.0001). Thus, the observed correlations presented in ref. 1 are not surprising.

Second, ref. 1 also appears agnostic as to whether the morphogenera should be used together with a phylogeny when
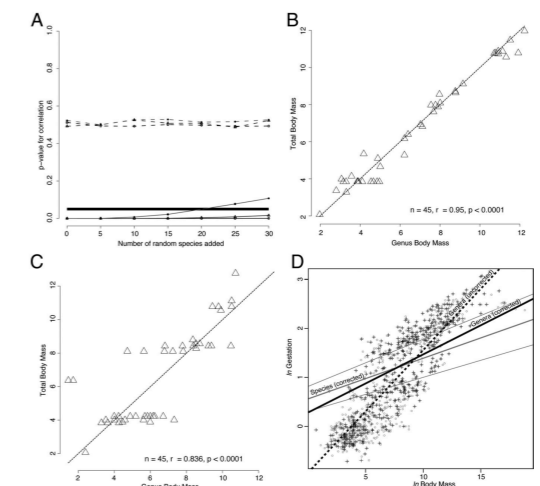

Fig. 1. Correlation analyses of morphogenera and comparative analysis of body mass and gestation. (A) Significance (Spearman correlation, as in ref. 1) of correlation of means of *ln* body mass of a set of X species with a set of X + Y species (solid lines), where X is 3 (closed circles), 7 (triangles), or 15 (open circles) and Y is 0, 5, 10, 15, 20, 25, 30, where the larger set includes all the species in the smaller set. Dotted lines correspond to correlating a set X with a set Z + Y species, where the size of Z and X are the same, but the species differ. The dark solid line represents P = 0.05. For each point, 2,000 replicates, each with 45 randomly generated sets, were used. The masses came from ref. 2, pruned down to the set of species also in the tree of ref. 3. (B and C) Plots of *ln* body mass 45 simulated morphogenera, as in figure 3A of ref. 1, where genera consist of 3 species chosen at random from within mammal families (B) or orders (C) (species chosen without replacement within genera but with replacement between genera, and genera simulated until 45 nonmonophyletic ones were recovered). The x axis shows the median value for the morphogenus, the y axis shows the median for the smallest clade containing all of the species (based on the tree of ref. 3). No distinction between polyphyletic or paraphyletic genera was attempted (open triangle used for both). Otherwise, plot and analysis as in figure 3A of ref. 1. (D) Generalized least-squares analyses of *ln* gestation period against *ln* body mass for all mammal species and genera. The thick black line shows the fit of these variables for all genera. The medium grey line shows the fit for all mammal species, and the light grey lines show the fit without correcting for the phylogeny. The thick dashed line shows the fit for the genera without correcting for the phylogeny. Grey dots represent data using species-level values and black crosses represent generic-level data.