Reviews in
Computational Biology

Methodological Challenges in the Pursuit of the Tree of Life

Christophe Dessimoz

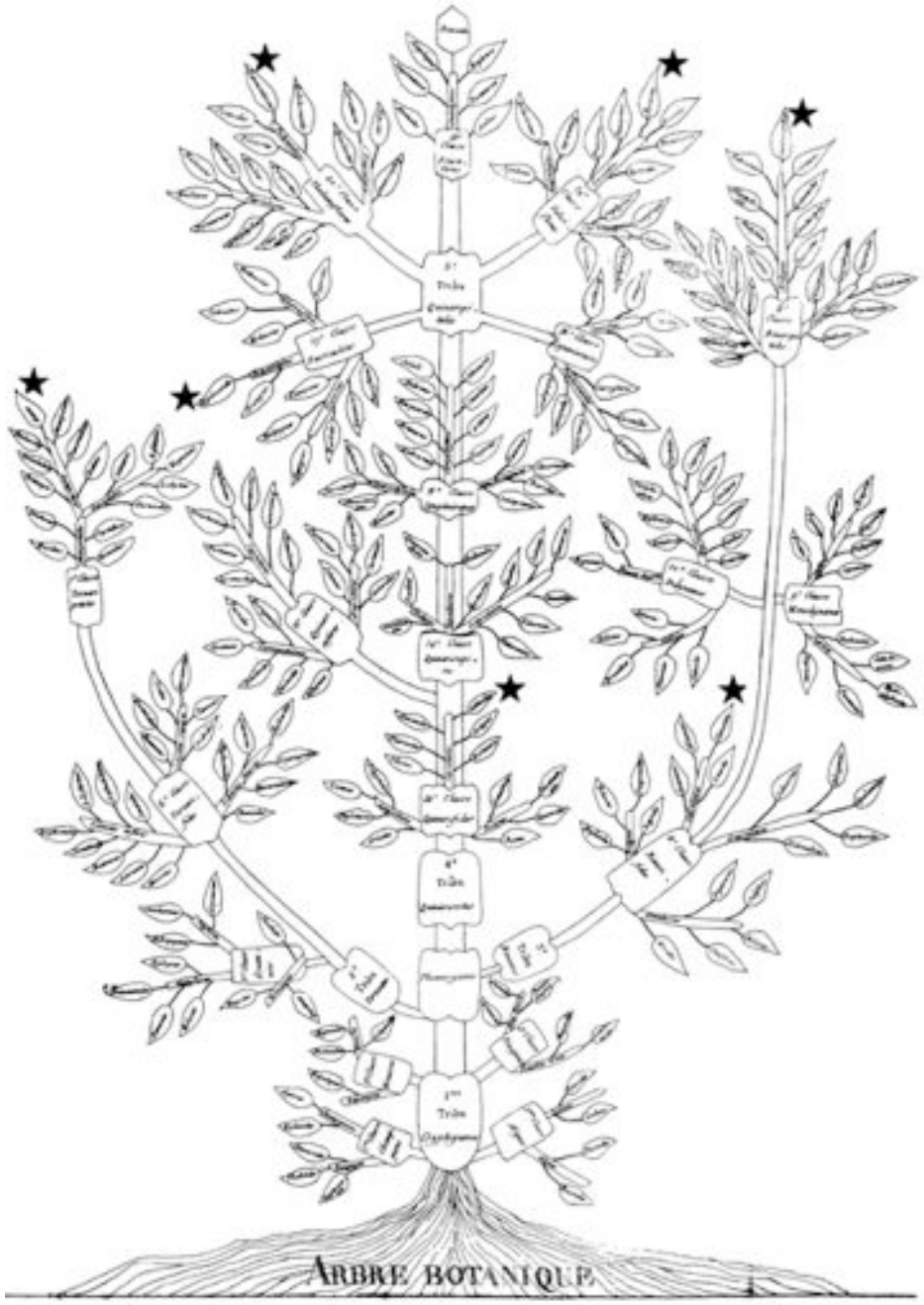
March 7th, 2011

ETH Zürich inf | Informatik
Computer Science

Outline

- Introduction
- Mature methods: supermatrix, supertree
- Emerging methods: species-tree
- Outlook

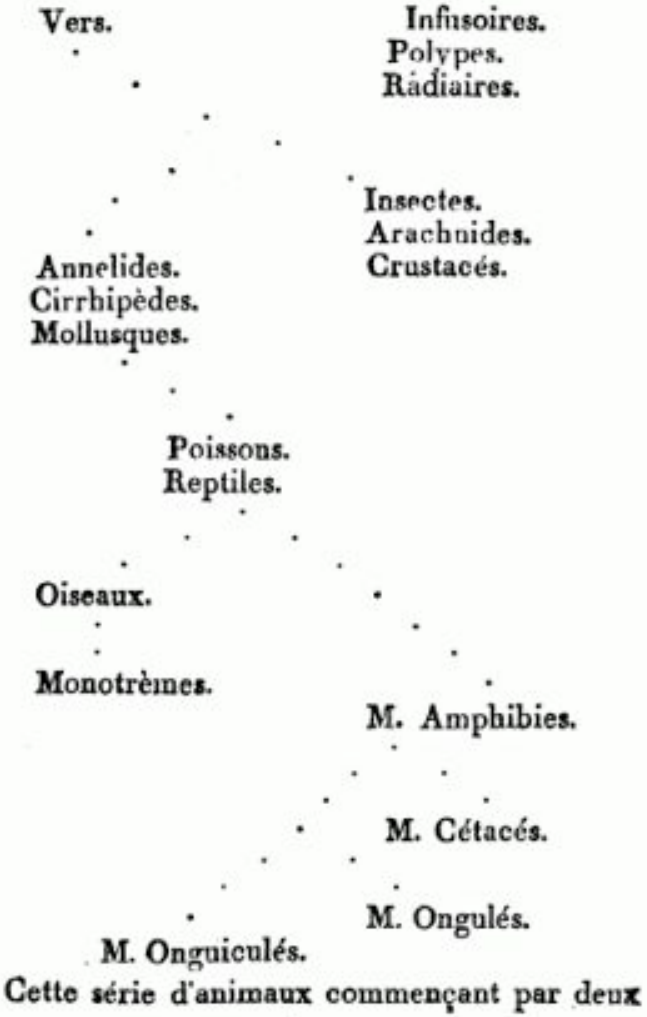
Augustin Augier,
Arbre Botanique
(1801)



Lamarck, Philosophie Zoologique , 1809

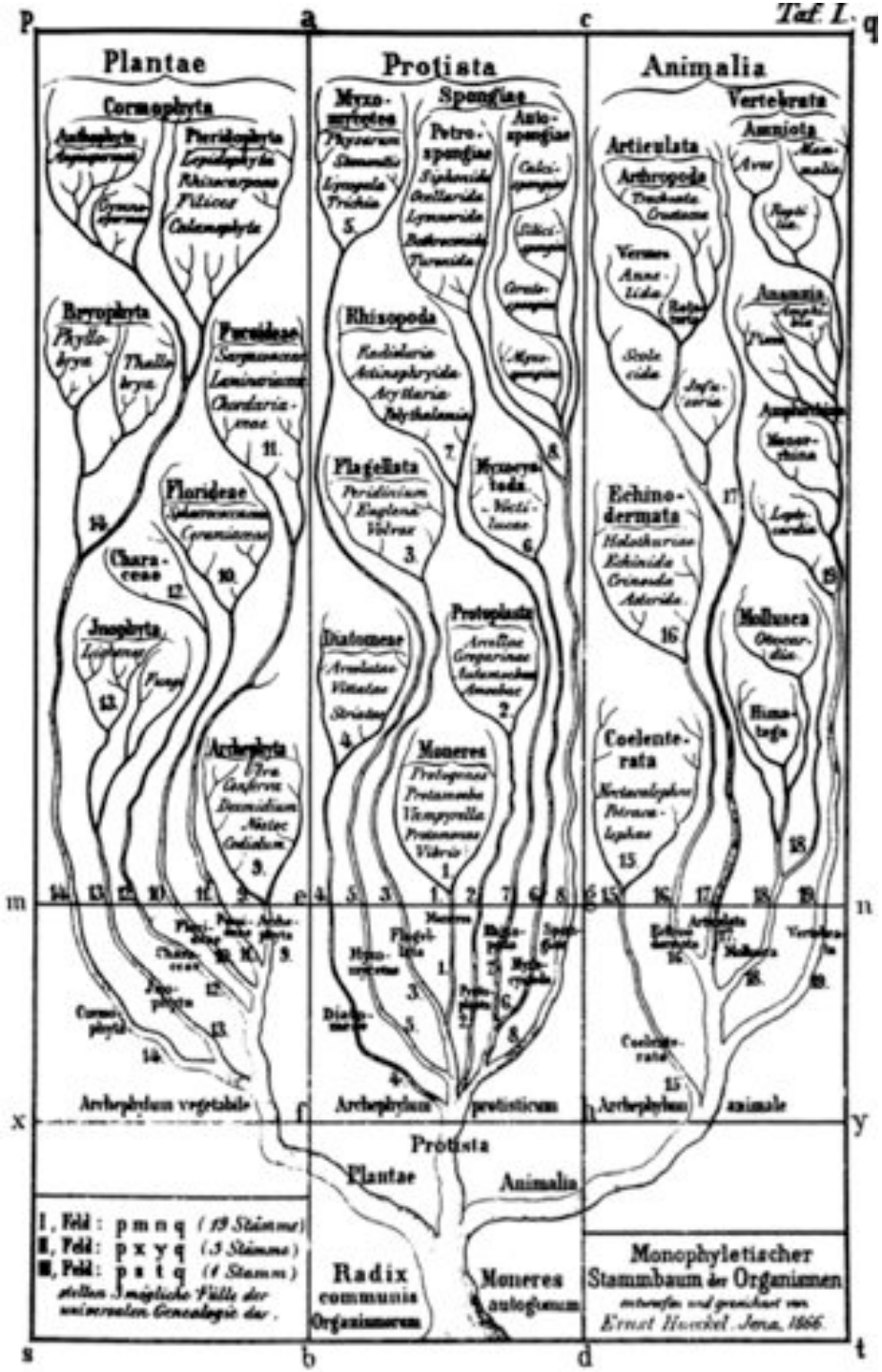
TABLEAU

Servant à montrer l'origine des différens animaux.



Cette série d'animaux commençant par deux

Darwin, Notebook B, 1837

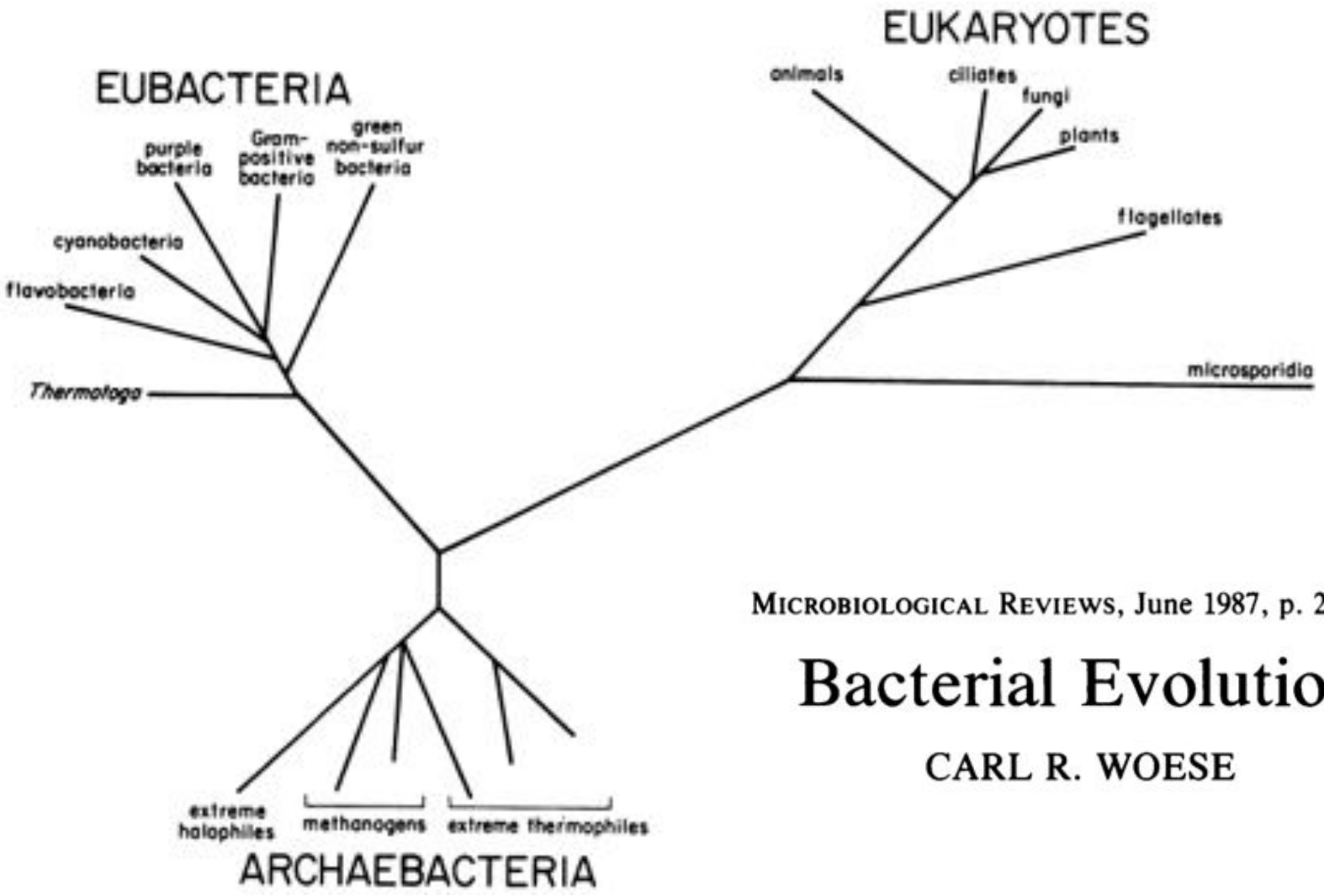


Monophyletischer Stammbaum der Organismen

entworfen und gezeichnet von

Ernst Haeckel. Jena, 1866.

16S rRNA was used by Woese (1987) to group early life forms into three kingdoms

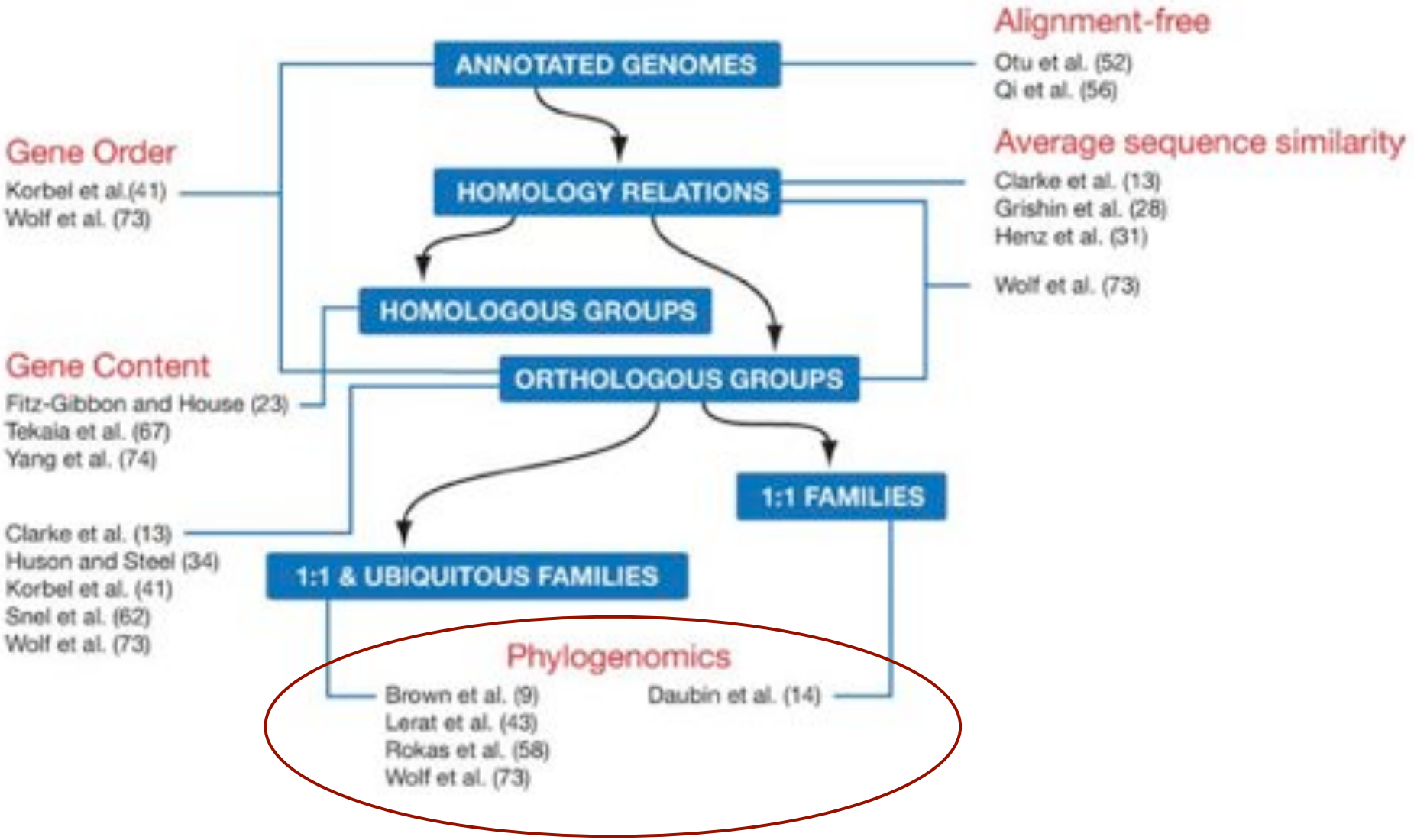


MICROBIOLOGICAL REVIEWS, June 1987, p. 221–271

Bacterial Evolution

CARL R. WOESE

Genomic Era



Snel et al. Genome trees and the nature of genome evolution. Annu Rev Microbiol (2005) vol. 59 pp. 191-209

Evolution

Ending incongruence

Henry Gee

Recovering the true evolutionary history of any group of organisms has seemed impossible. The availability of large amounts of genomic data promises an era in which the uncertainties are better constrained.

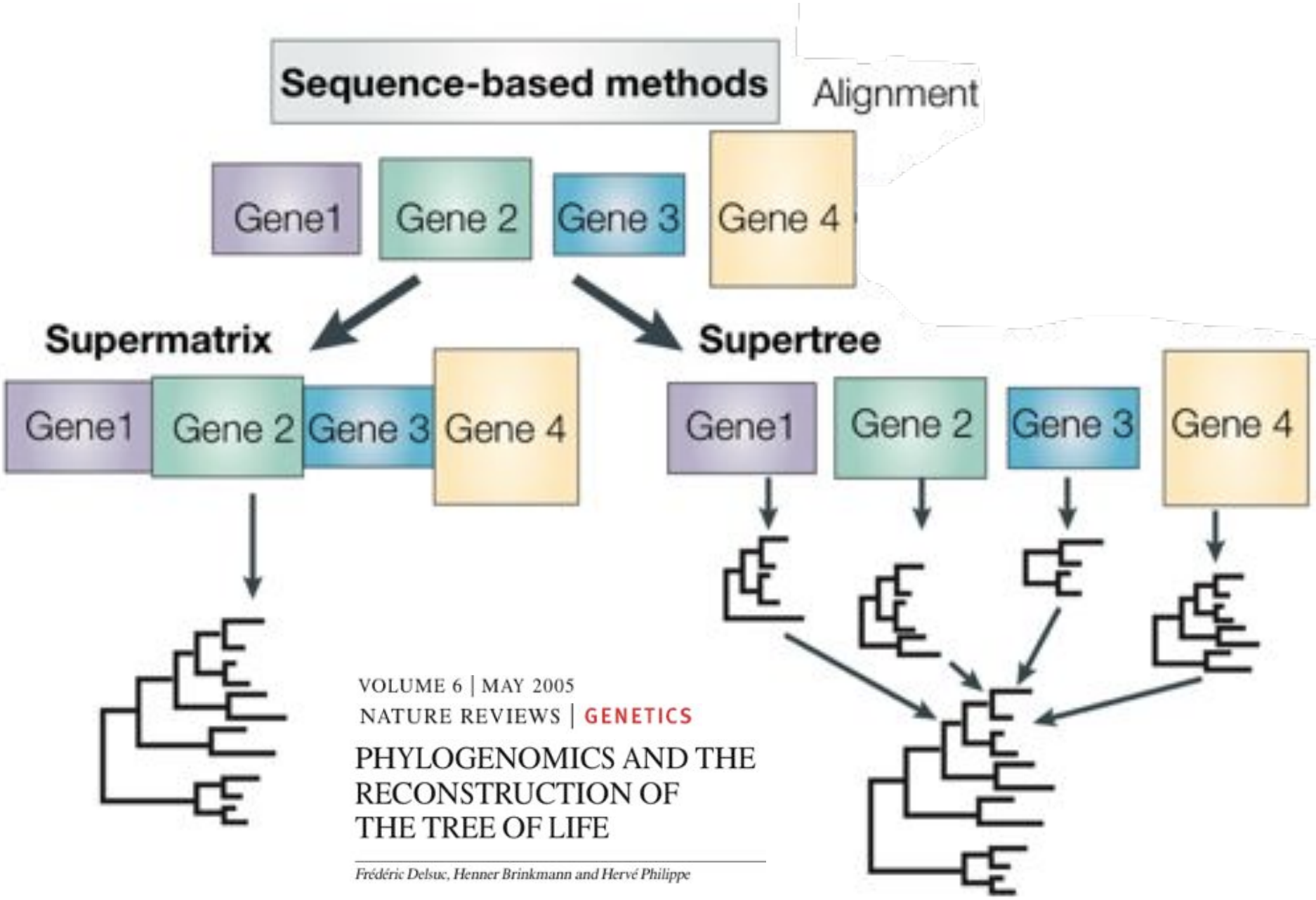
The careful reader of this issue will come across a picture that, at first sight, has a startling message. Those impatient to see it should turn to the report from Rokas *et al.* (A. Rokas, B. L. Williams, N. King & S. B. Carroll *Nature* 425, 798–804; 2003). The authors' Fig. 4, on page 801, is the object of interest. What, you might ask yourself, is so remarkable about a phylogeny — an evolutionary tree — of seven species of yeast of the genus *Saccharomyces*? Closer inspection, however, will show that the authors are making an unprecedented how should the information from characters be 'weighted'? When technology allowed examination of the sequences of genes, rather than features of anatomy or physiology, the feeling was that truth would be simpler to reach. Genes are a direct expression of inheritance, and not signals refracted through the distorting lens of an organism's physical or biochemical characteristics.

But genes are not immune to external influences: like any feature of anatomy, they have histories that can confuse as well as enlighten. The result has been several decades

In this review, I will show that

Molecular data is not the bottleneck.
Modeling and Analysis are.

PART I
Established Methods: Supermatrix and Supertree



VOLUME 6 | MAY 2005
NATURE REVIEWS | GENETICS

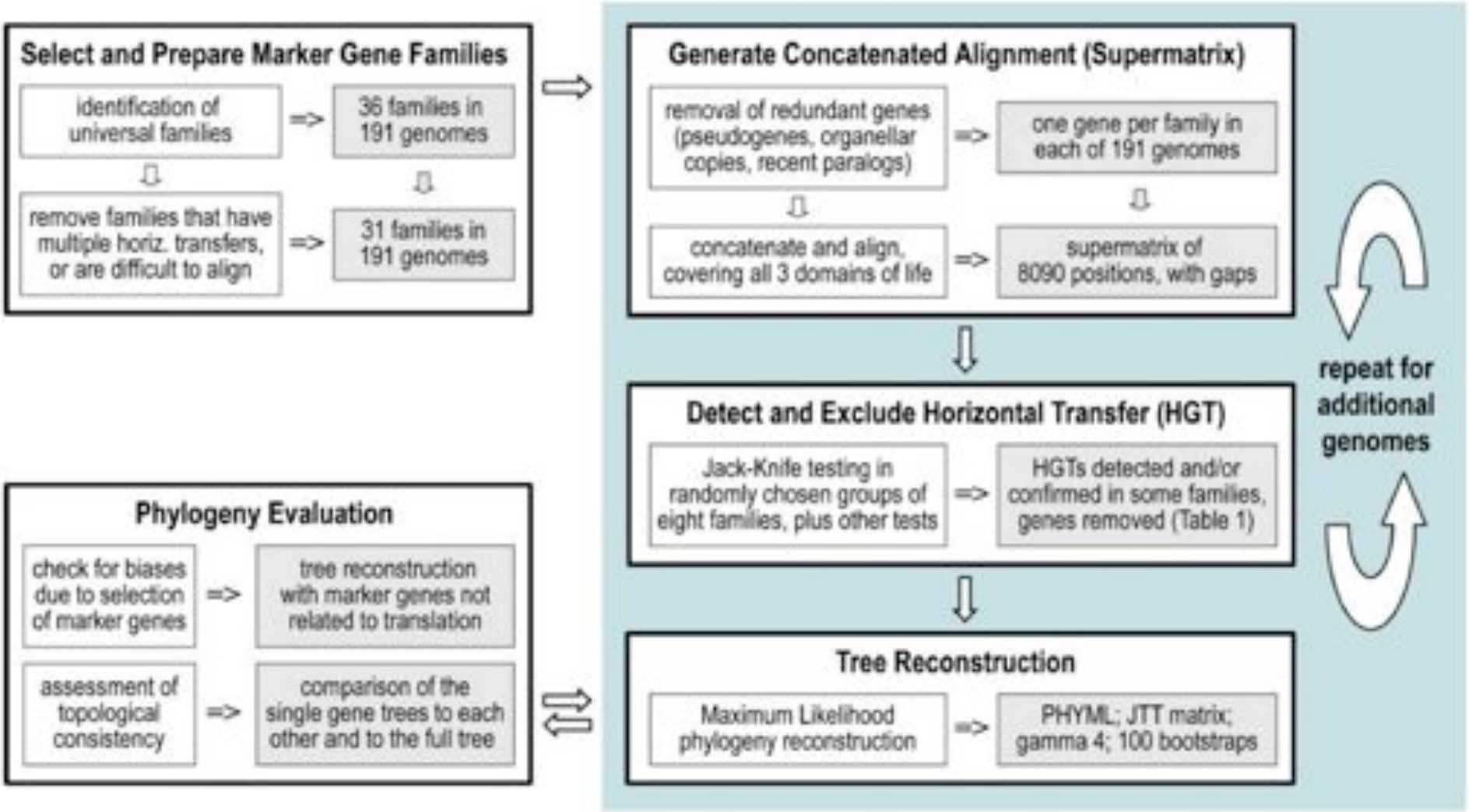
PHYLOGENOMICS AND THE RECONSTRUCTION OF THE TREE OF LIFE

Frédéric Delsuc, Henner Brinkmann and Hervé Philippe

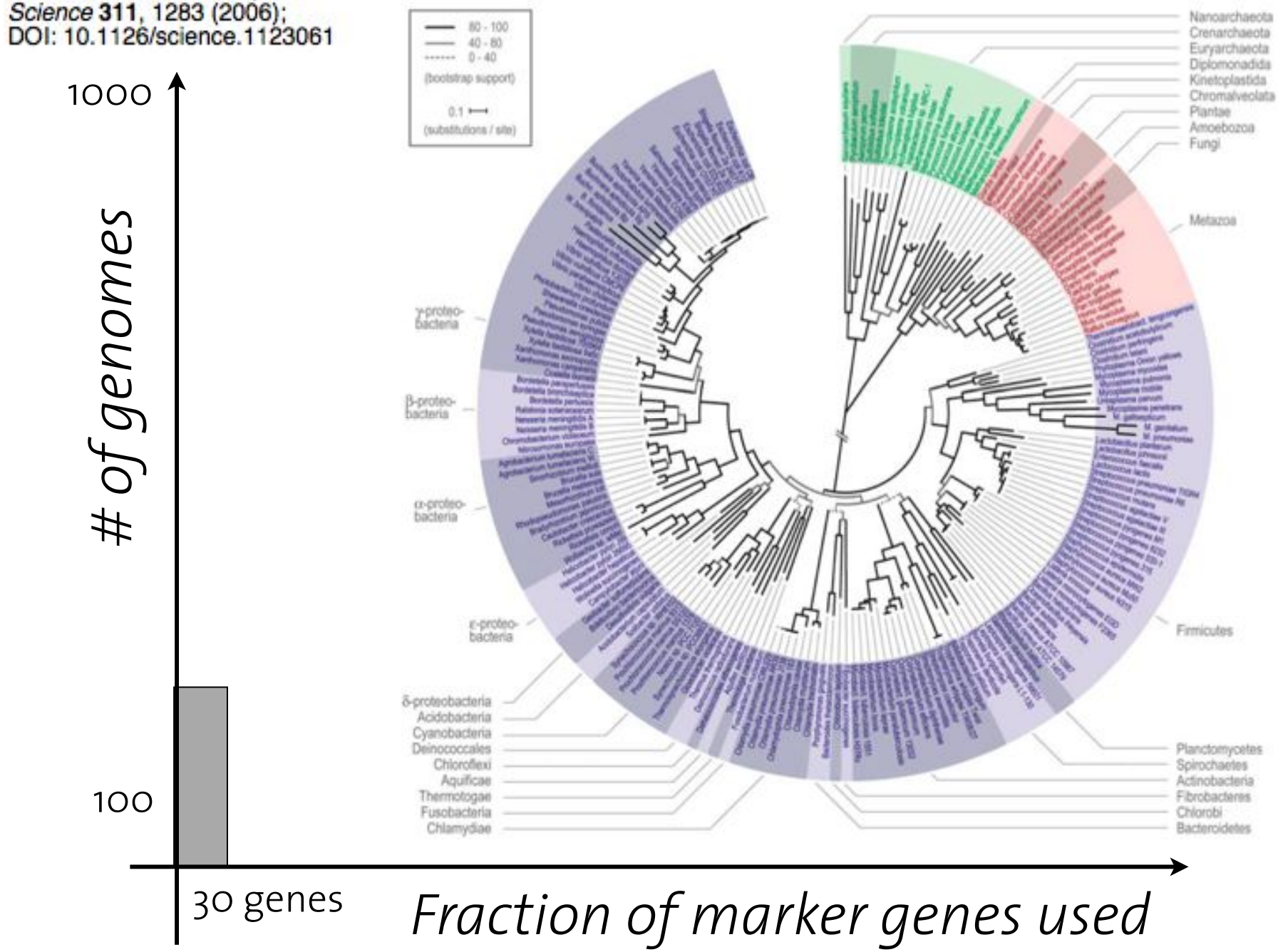
Toward Automatic Reconstruction of a Highly Resolved Tree of Life

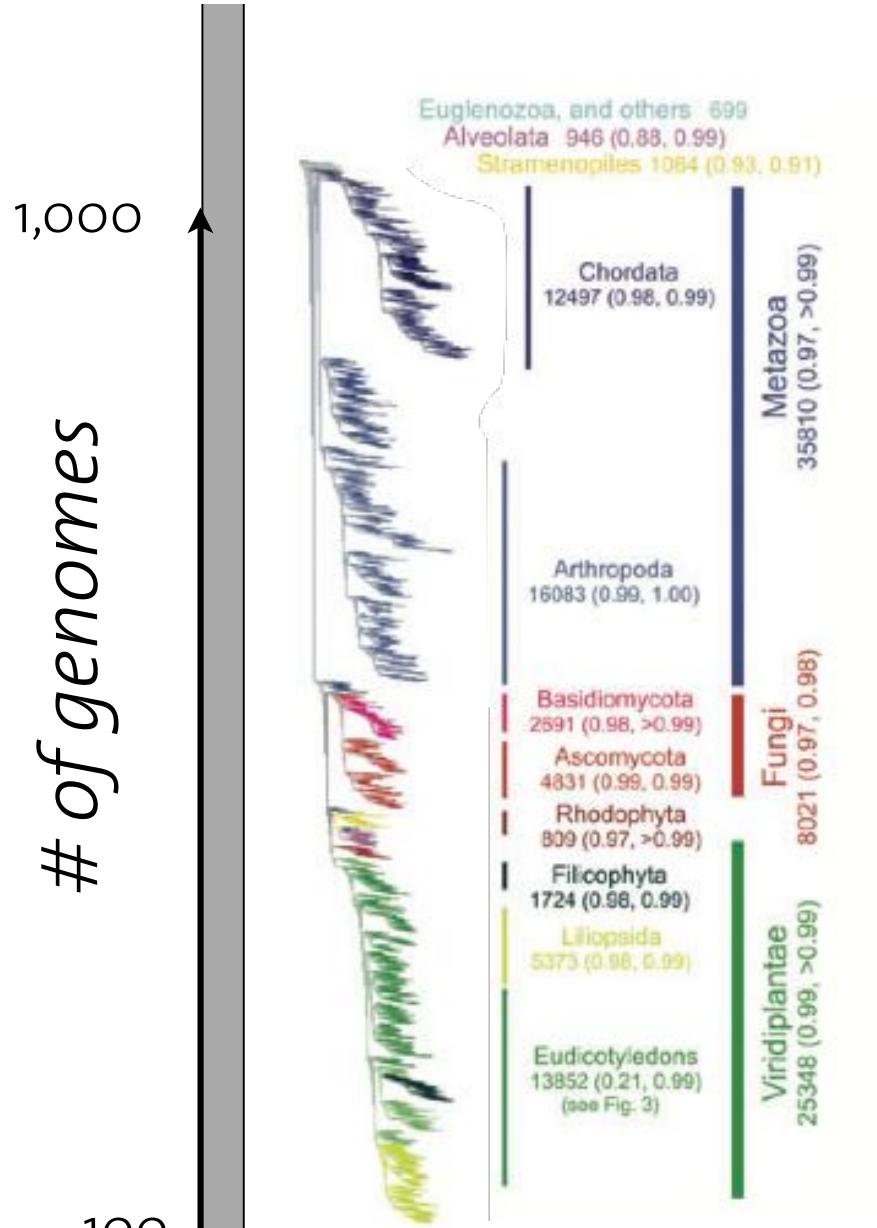
Francesca D. Ciccarelli, *et al.*

Science **311**, 1283 (2006);
DOI: 10.1126/science.1123061



Toward Automatic Reconstruction of a Highly Resolved Tree of Life
Francesca D. Ciccarelli, *et al.*
Science **311**, 1283 (2006);
DOI: 10.1126/science.1123061





25 (2009) 211–230

Cladistics

Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups

Pablo A. Goloboff[a,*], Santiago A. Catalano[b], J. Marcos Mirande[b], Claudia A. Szumik[a], J. Salvador Arias[a], Mari Källersjö[c] and James S. Farris[d]

Tree searches, identical for molecular and combined data sets, ran in parallel on three computers (totalling 16 processors and 96 GB RAM), examining for each data set $\sim 7.5 \times 10^{14}$ rearrangements in $\sim$ 2.5 months' processor-time.

1,000

100

# of genomes

13 genes

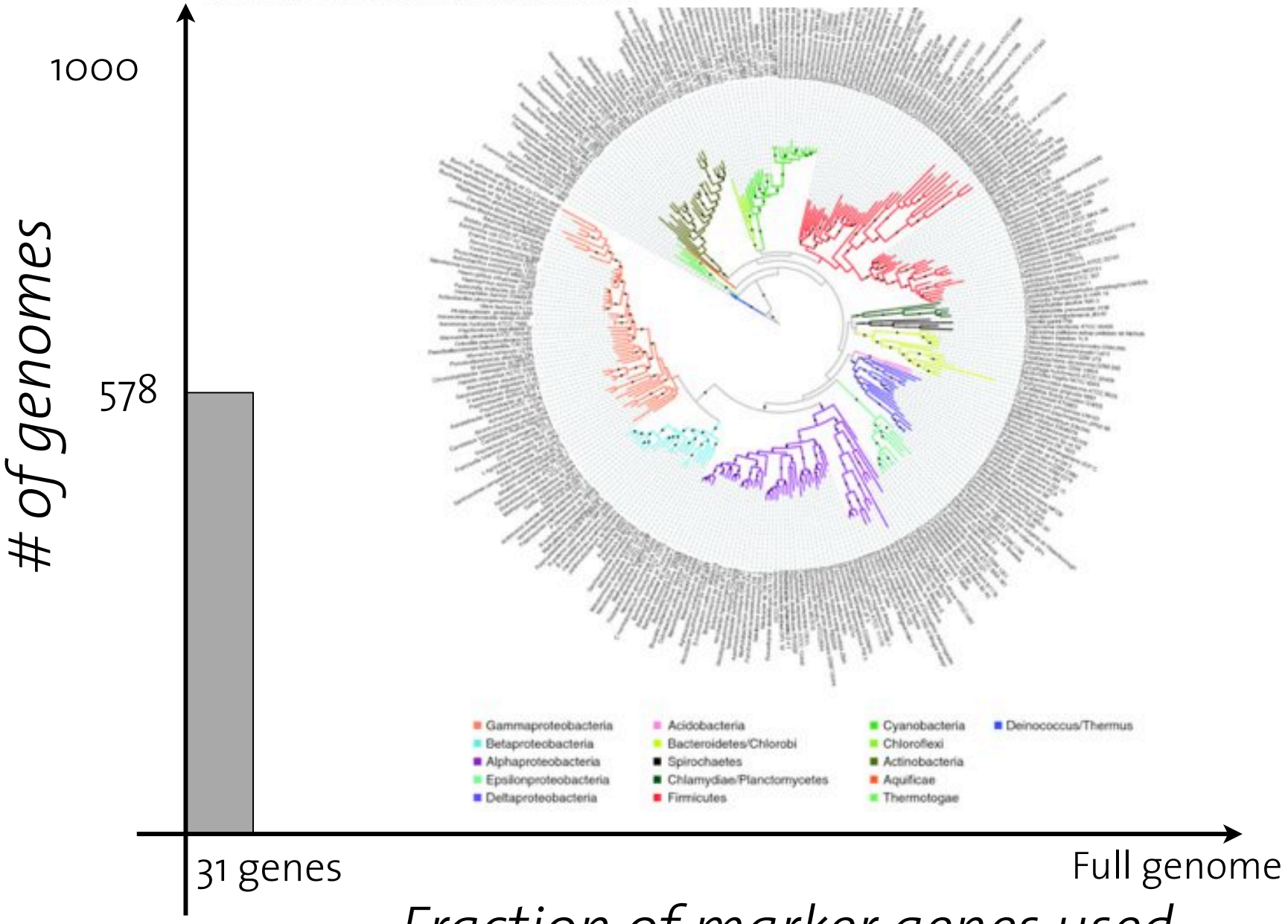
Full genome

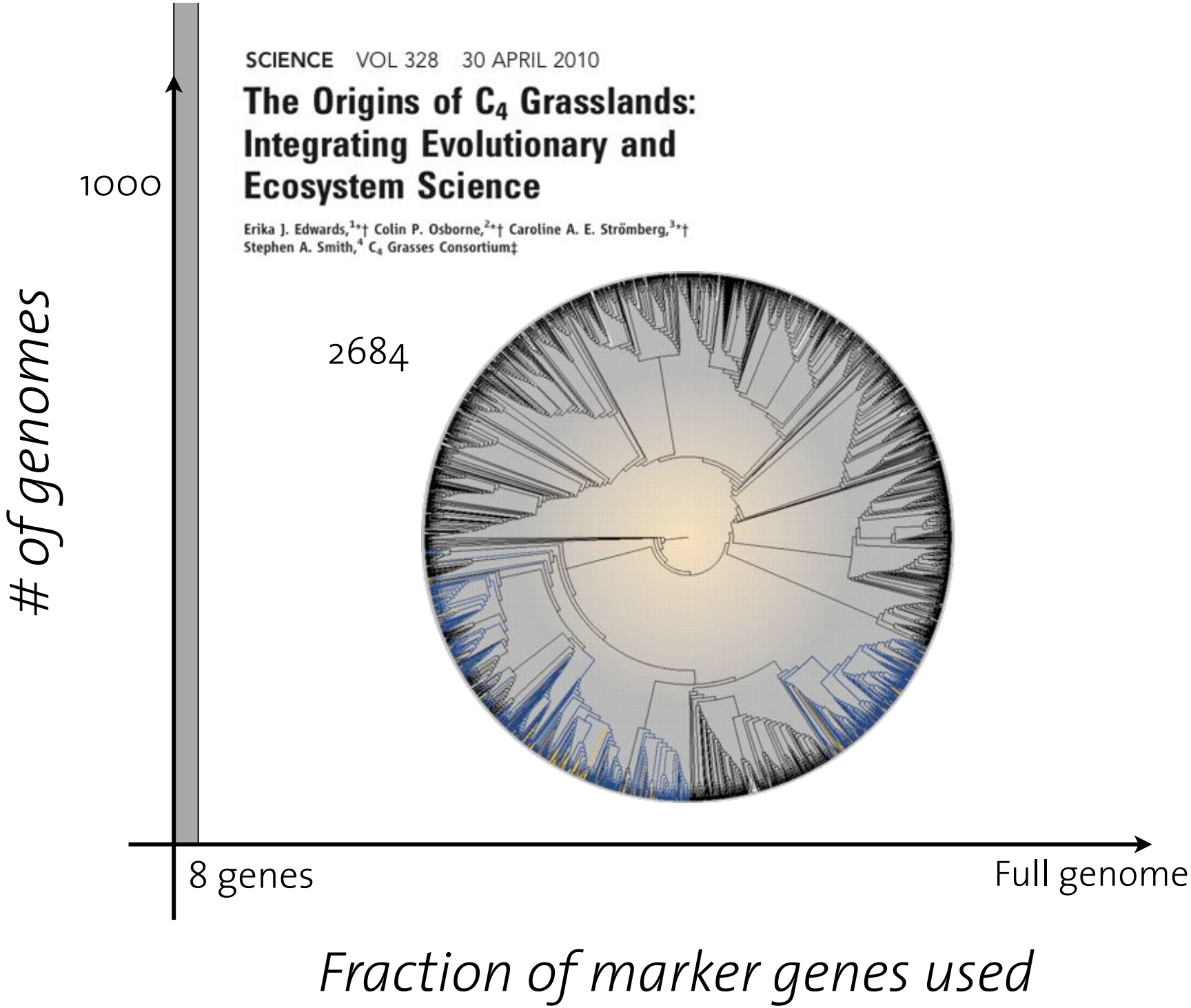
Fraction of marker gene used

Genome ***Biology*** 2008, **9**:R151

A simple, fast, and accurate method of phylogenomic inference

Martin Wu* and Jonathan A Eisen*†‡




SCIENCE VOL 328 30 APRIL 2010
The Origins of C4 Grasslands: Integrating Evolutionary and Ecosystem Science
Erika J. Edwards,[1*†] Colin P. Osborne,[2*†] Caroline A. E. Strömberg,[3*†]
Stephen A. Smith,[4] C4 Grasses Consortium‡
1000
2684
# of genomes
8 genes
Full genome
Fraction of marker genes used

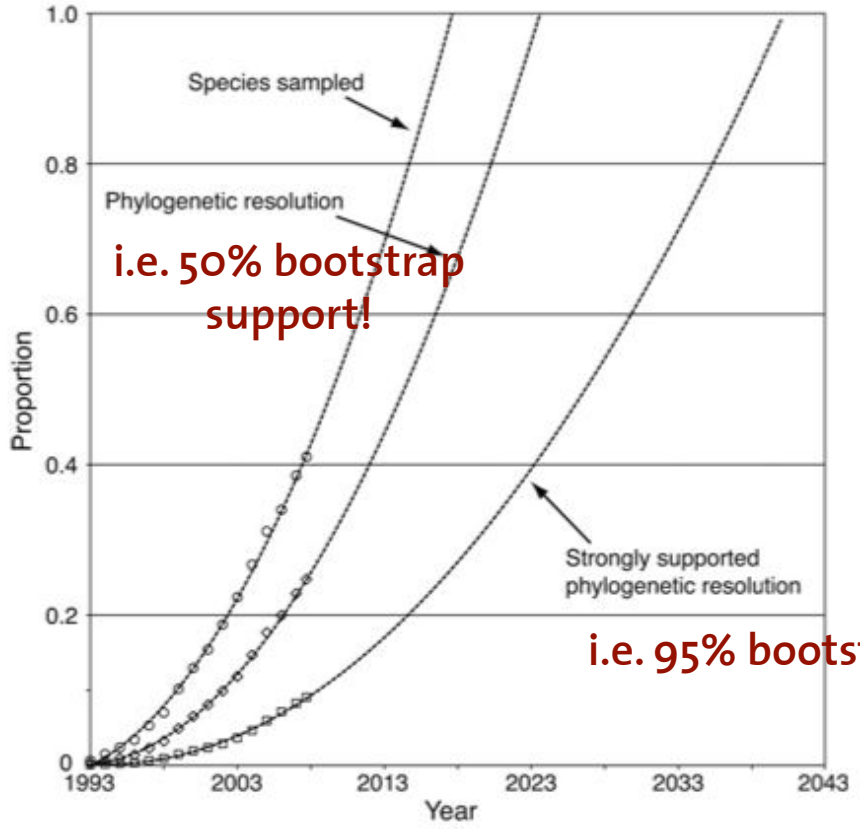
Thomson and Shaffer *BMC Biology* 2010, **8**:19
http://www.biomedcentral.com/1741-7007/8/19

BMC Biology

CORRESPONDENCE Open Access

Rapid progress on the vertebrate tree of life

Robert C Thomson*, H Bradley Shaffer



i.e. 50% bootstrap support!

i.e. 95% bootstrap support!

But!

Actually, use only small fraction of data.

Genome ***Biology*** 2006, **7:**118

Opinion

The tree of one percent

Tal Dagan and William Martin

Why?

Gene tree ≠ Species tree

Opinion TRENDS in Genetics Vol.22 No.4 April 2006

Phylogenomics: the beginning of incongruence?

Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc and Hervé Philippe

Canadian Institute for Advanced Research, Centre Robert-Cedergren, Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec, Canada, H3C3J7

The incongruence between two phylogenies can be the result of: (i) violations of the orthology assumption generated by mechanisms such as gene duplication, horizontal gene transfer or lineage sorting [4]; (ii) stochastic error related to the length of the genes; and (iii) systematic error leading to tree reconstruction artifacts generated by the presence of a nonphylogenetic signal in the data. Adopting a genome-scale approach

Gene tree ≠ Species tree

- **Gene duplication (paralogs)**
- **Lateral gene transfer (xenologs)**
- **Endosymbiosis** (e.g. Delusc et al. 2005)
- **Hybridization** (Hallström & Janke 2008)
- **Incomplete lineage sorting (aka deep coalescence)**

Jeffroy et al. 2006
McInerney et al. 2008
Edwards 2009

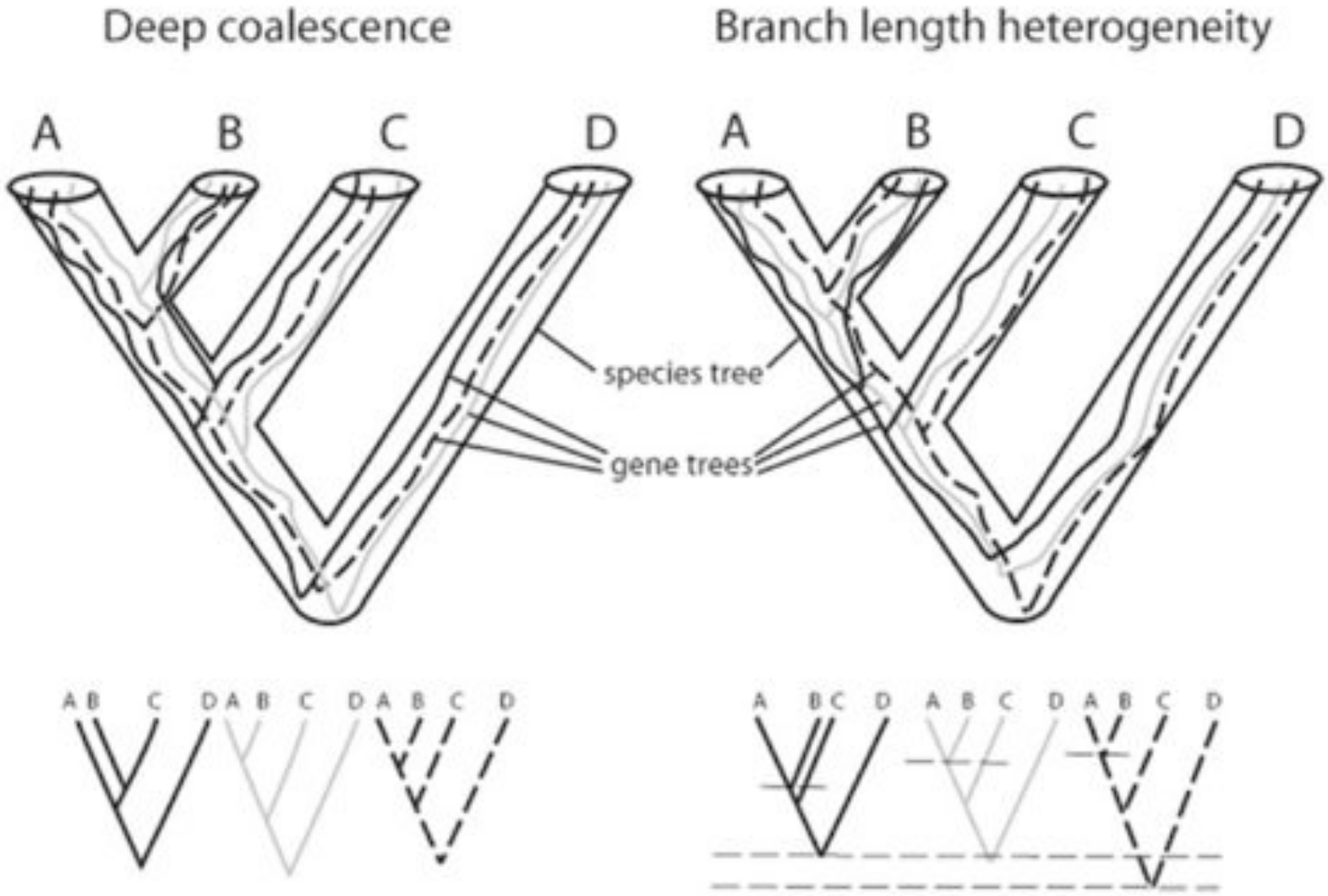
Systematic Errors

- **Branch-length heterogeneity** (Matsen & Steel 2007, Edwards 2009)
- **Nucleotide composition heterogeneity across species** (Hasegawa & Hashimoto 1993, Jeffroy et al. 2006)
- **Missing data** (Hartmann & Vision 2008)
- **In general:** ***model violations***

IS A NEW AND GENERAL THEORY OF MOLECULAR SYSTEMATICS EMERGING?

Scott V. Edwards[1,2]

EVOLUTION *JANUARY 2009*

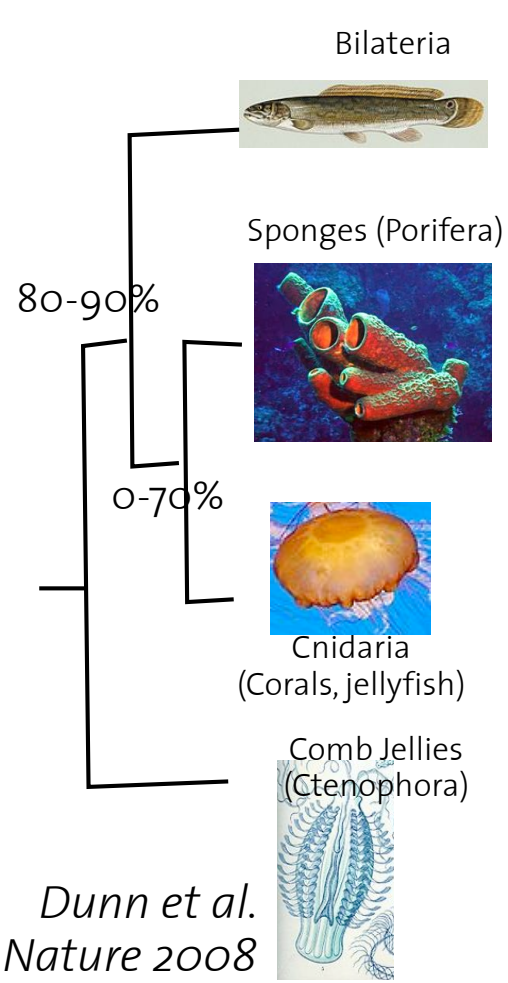


Systematic error can result in overconfidence

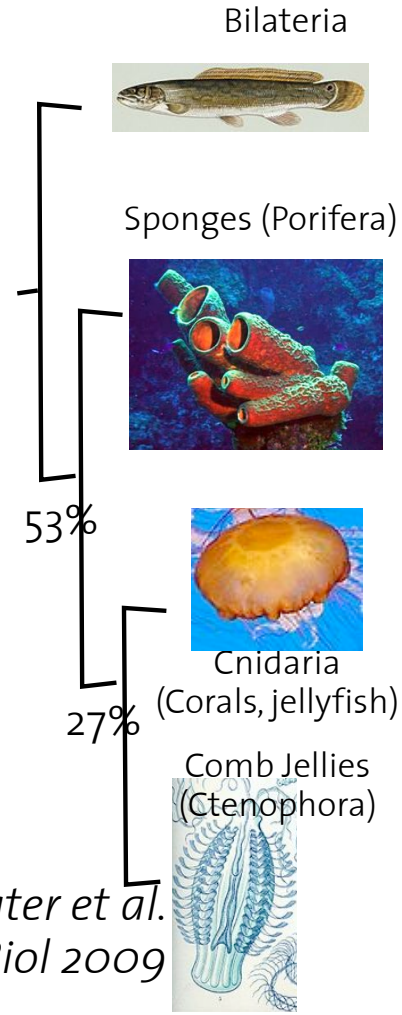
e.g. Cladistics 26 (2010) 444–452

Unringing a bell: metazoan phylogenomics and the partition bootstrap

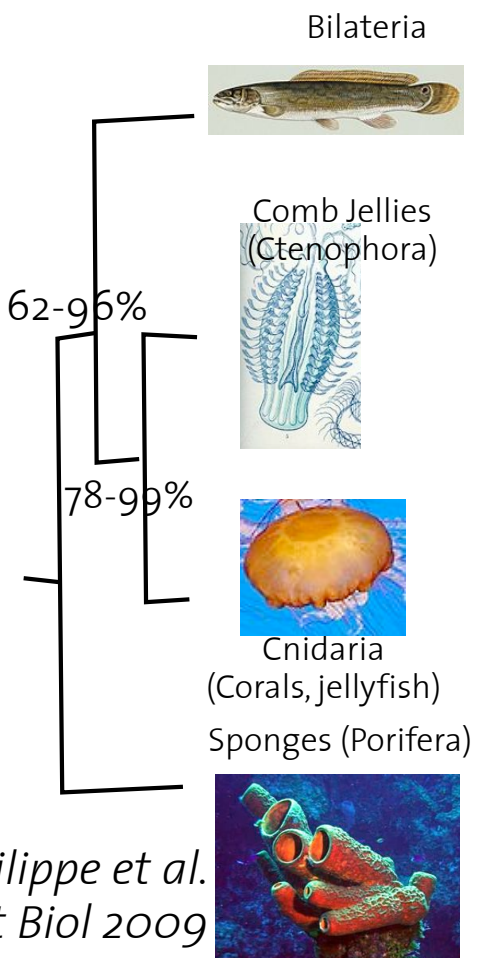
Mark E. Siddall*



Dunn et al. Nature 2008



Schierwater et al. PLoS Biol 2009



Philippe et al. Current Biol 2009

All photos from Wikipedia

PART II
Emerging Methods:
Species-Tree Inference Methods

Two main classes

- **Methods modeling specific processes ("mechanistic")**
  - Deep coalescence
  - Gene duplication (Arvestad et al. 2003, 2009)
  - LGT (that is for another review!)
  - Rate variation among markers (Pupko et al. 2002)
- **Process agnostic ("empirical")**

Modeling Coalescent

Genetics **164:** 1645–1656 (August 2003)

Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci

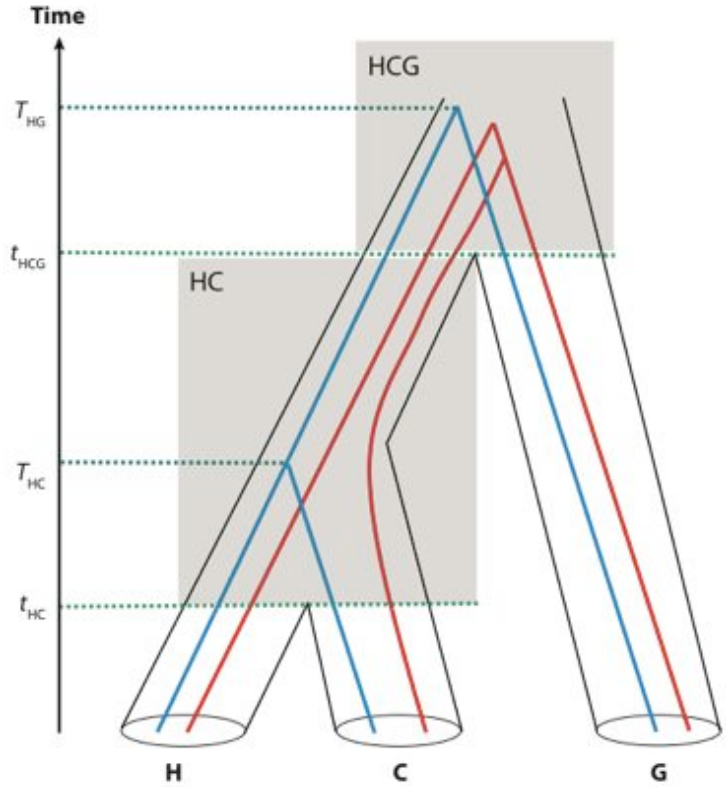
Bruce Rannala* and Ziheng Yang†,1

$$f(\Theta, G|D) \propto f(D|G)\ f(G|\Theta)\ f(\Theta).$$

Model Parameters — Gene Trees — Sequence alignment

$$f(G|\Theta) = \prod_i f(G_i|\Theta) = \prod_i f(T_i, \mathbf{t}_i|\Theta).$$

$$f(T_{HC}) = \frac{1}{2N_e} e^{-(T_{HC} - t_{HC})/(2N_e)}$$



Rannala & Yang, Annu Rev Genomics Hum Genet 2008

Methods

BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 21 2008, pages 2542–2543
doi:10.1093/bioinformatics/btn484

Phylogenetics

BEST: Bayesian estimation of species trees under the coalescent model

Liang Liu

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 7 2009, pages 971–973
doi:10.1093/bioinformatics/btp079

Phylogenetics

STEM: species tree estimation using maximum likelihood for gene trees under coalescence

Laura S. Kubatko[1,*], Bryan C. Carstens[2] and L. Lacey Knowles[3]

(summary statistics)

Syst. Biol. 58(5):468–477, 2009

Estimating Species Phylogenies Using Coalescence Times among Sequences

LIANG LIU[1,*], LILI YU[2], DENNIS K. PEARL[3], AND SCOTT V. EDWARDS[1]

Mol. Biol. Evol. 27(3):570–580. 2010

Bayesian Inference of Species Trees from Multilocus Data

Joseph Heled*[,1] and Alexei J. Drummond[1,2,3]

also see review of Liu et al 2009

Process agnostic

Mol. Biol. Evol. 24(2):412–426. 2007

Bayesian Estimation of Concordance among Gene Trees

Cécile Ané,† Bret Larget,*† David A. Baum,† Stacey D. Smith,‡ and Antonis Rokas§*

$$P\{M|X\} \propto P\{M\} \prod_{i=1}^{G} \mathrm{P}\{T_i|X_i\}.$$

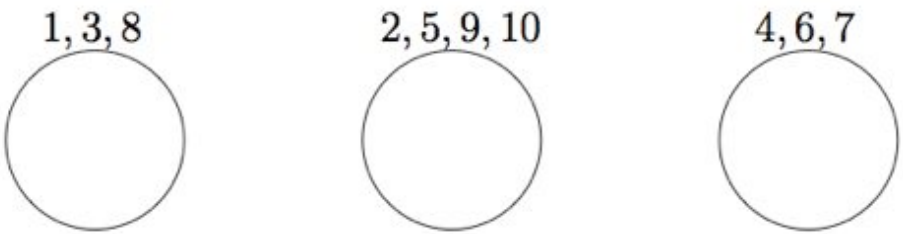
Gene-to-tree map | All Sequence alignments | Tree of gene i

- Independent tree inference for each gene (relatively efficient!)
- Number of different trees modeled as Dirichlet process

Dirichlet Process a.k.a. Chinese Restaurant Process

1. The first customer always chooses the first table.
2. The nth customer chooses the first unoccupied table with probability $\frac{\alpha}{n-1+\alpha}$, and an occupied table with probability $\frac{c}{n-1+\alpha}$, where c is the number of people sitting at that table.

e.g.



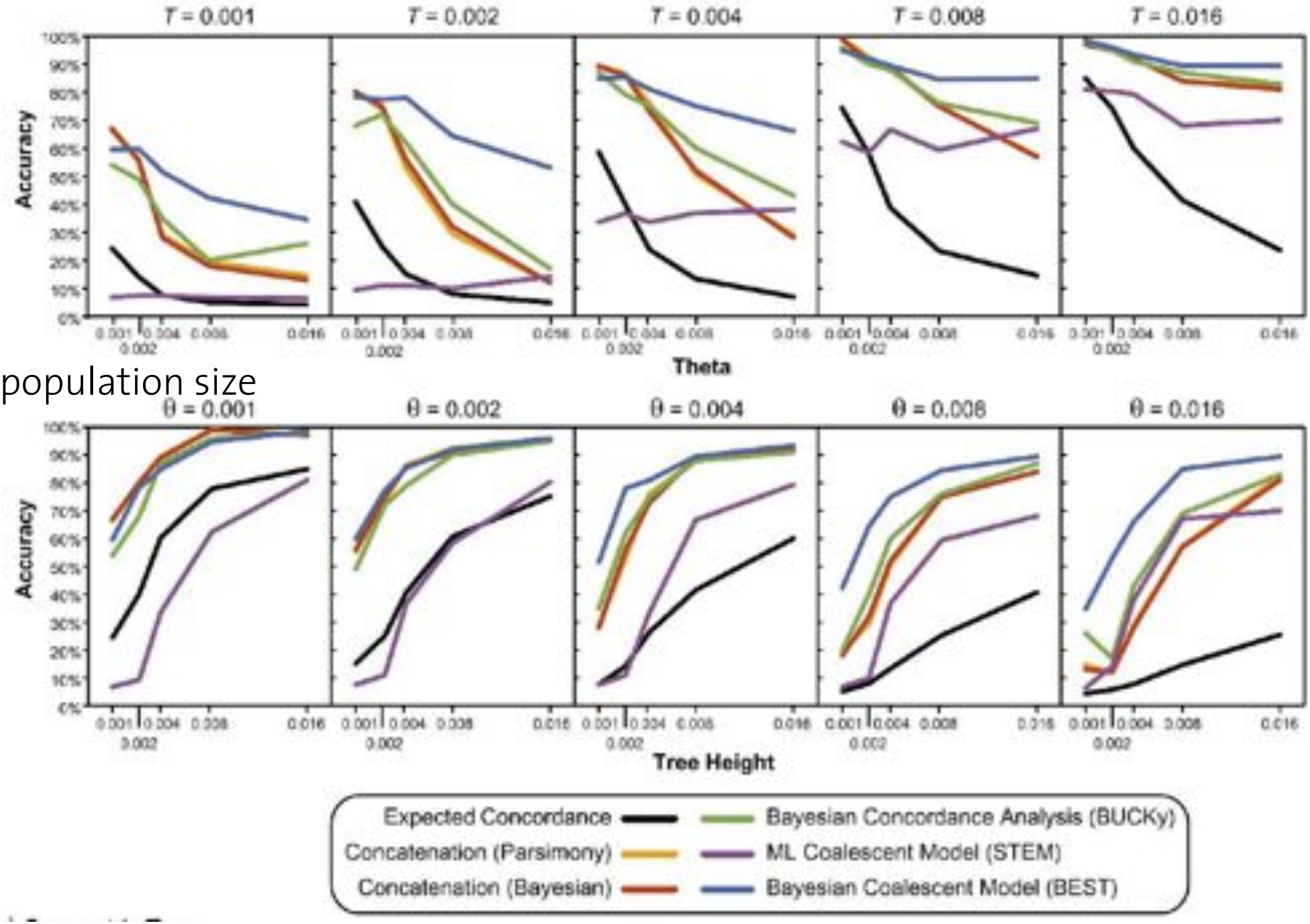
http://www.cs.princeton.edu/courses/archive/fall07/cos597C/scribe/20070921.pdf

Evaluation with simulated data

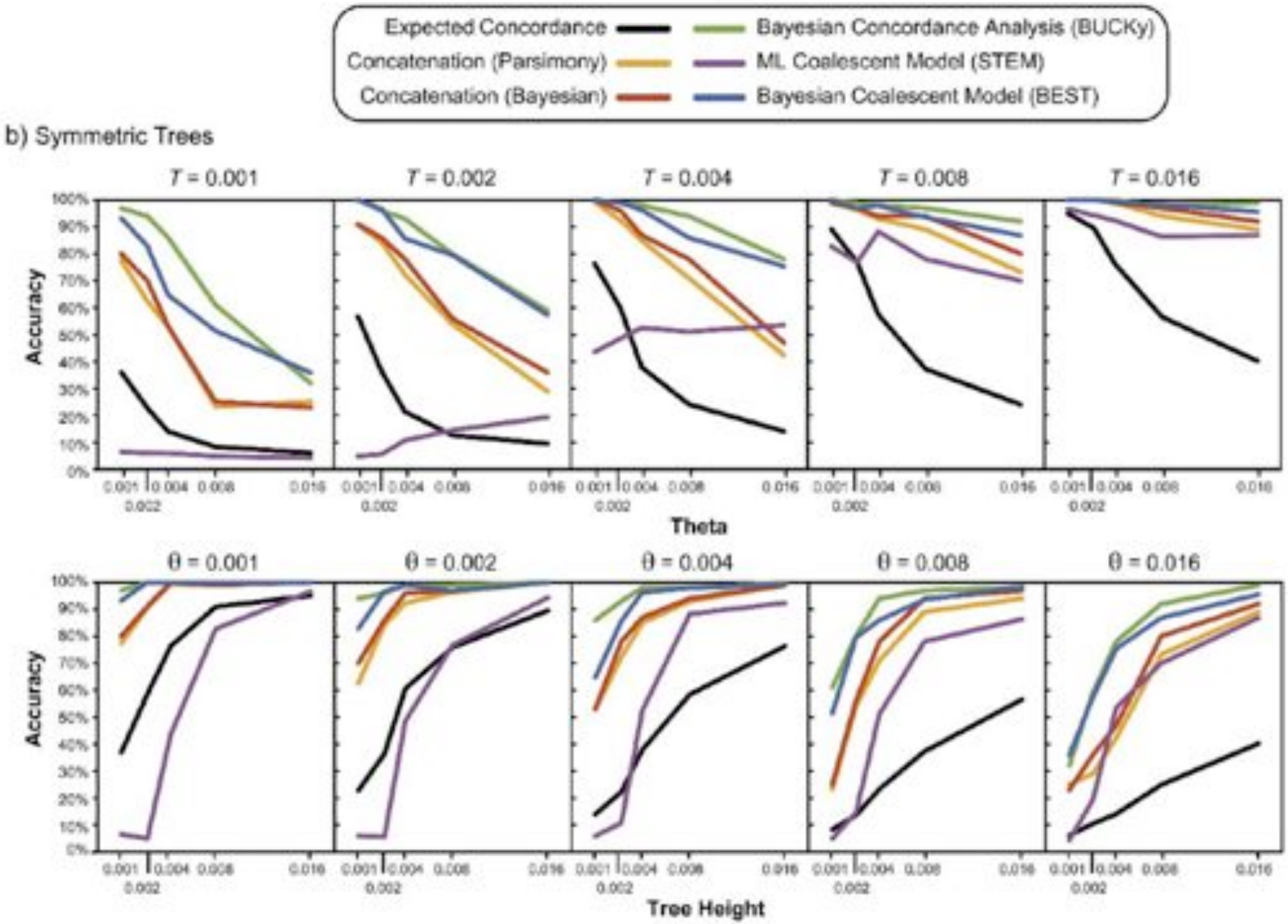
Leaché & Rannala, Syst Biol 2010

tree length

population size



Leaché & Rannala, Syst Biol 2010



Evaluation with empirical data

Syst. Biol. 58(5):489–500, 2009

Species Trees from Highly Incongruent Gene Trees in Rice

KAREN A. CRANSTON[1,2,*], BONNIE HURWITZ[1,3], DOREEN WARE[3,4], LINCOLN STEIN[3,5], AND ROD A. WING[6,7,8]

- "Note that the concordance factors in the BCA tree are much more conservative than the posterior probabilities in the topology estimated from the concatenated alignment"
- "Taking into account the incongruence between gene trees does not drastically change our overall view of rice phylogeny, but it does give a more varied picture of the support across the tree."
- "The BCA method is robust to the prior probability on gene tree incongruence (the α parameter)"
- "[The 6-species, 162 genes Bayesian analysis] had not yet reached stationarity after 1.6 billion iterations." (2 months on 96 CPU cores)

Outlook

- **Bottleneck is methods, not data**
- **Need methods able to deal with different gene histories**
- **Efficiency needs to be improved**
  ("*The largest data set yet tested with these species tree methods is yeast, with 106 loci in 8 species*" Cranston 2009)