Wellcome trust advanced course
“Molecular Evolution”

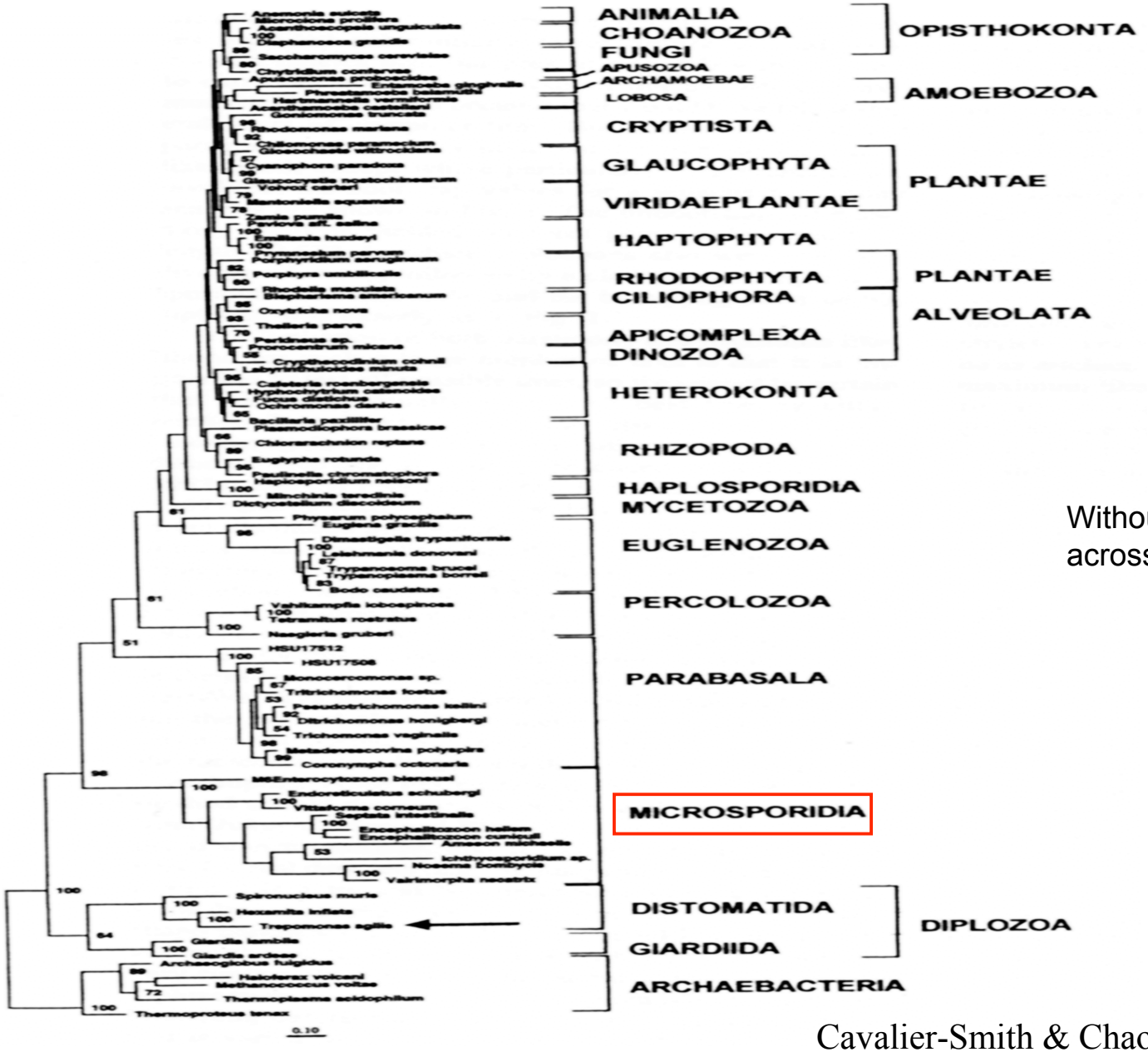
Source of errors, solutions, method comparisons

Manolo Gouy
CNRS - Université de Lyon

Hinxton 29 March - 9 April 2009

The Long Branch Attraction artefact

Phylogenetic analysis of eukaryotic small subunit ribosomal RNA

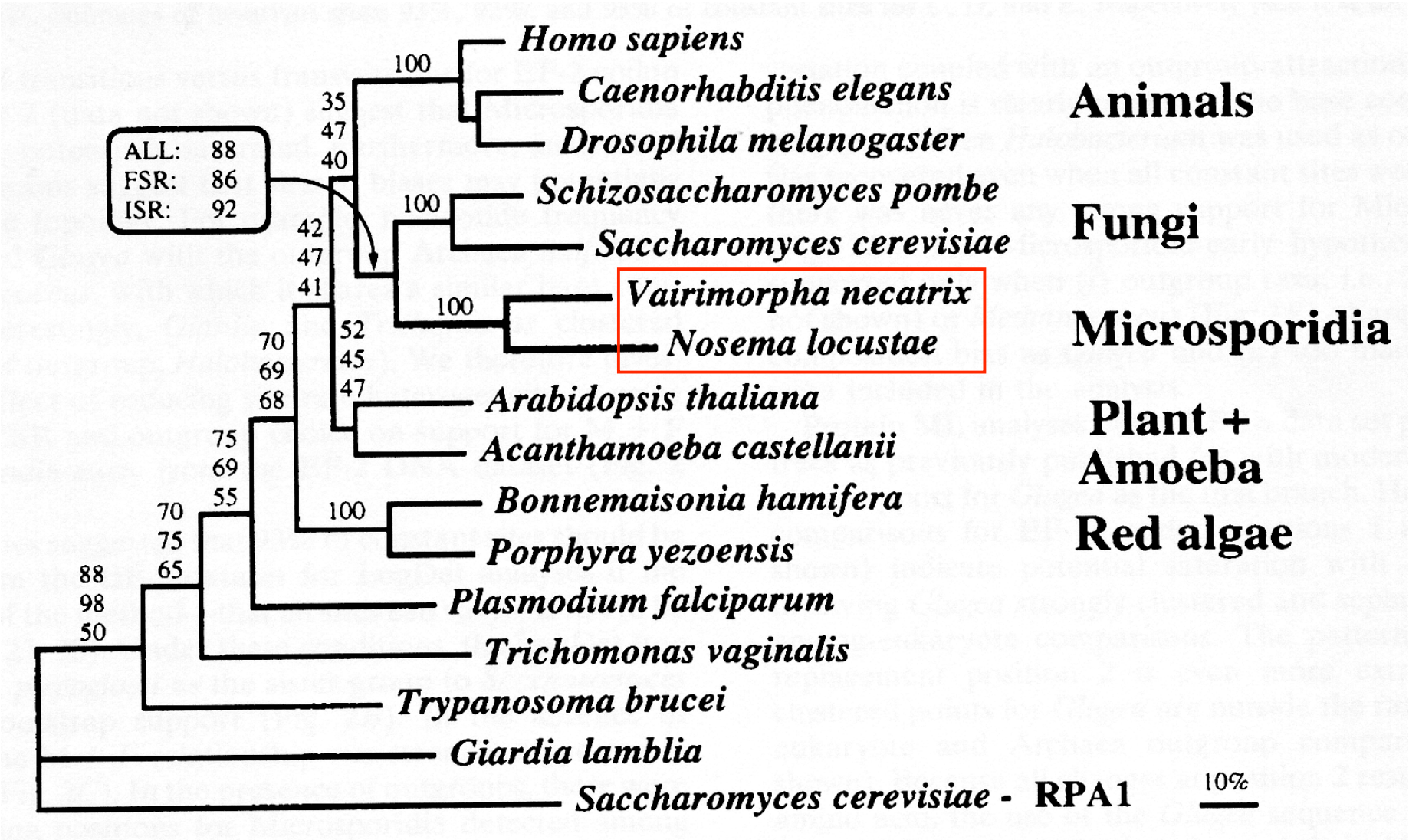


Without accounting for across-sites rate variation

Cavalier-Smith & Chao (1996) J Mol Evol 43:551

Phylogenetic analysis of RNA polymerase II large subunit

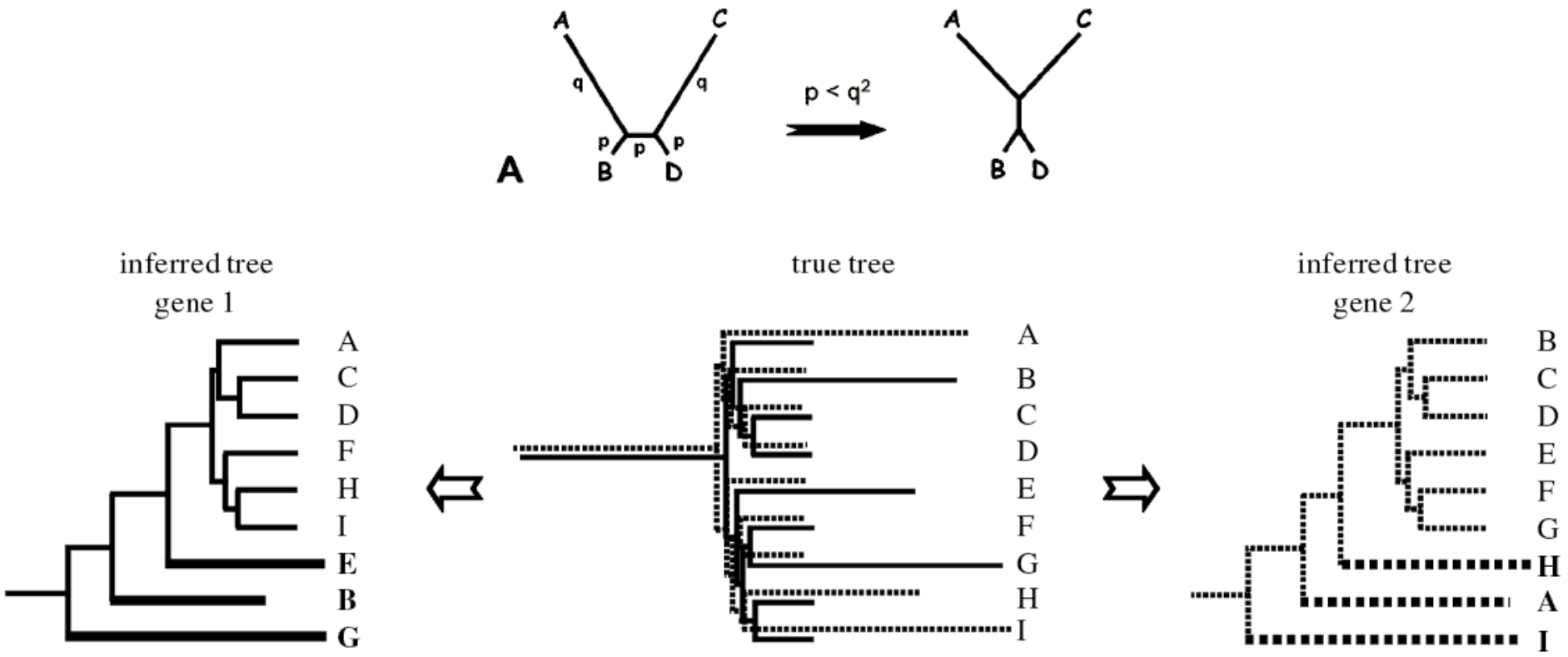
Hirt *et al*. (1999) Proc.Natl.Acad.Sci. USA 96:580



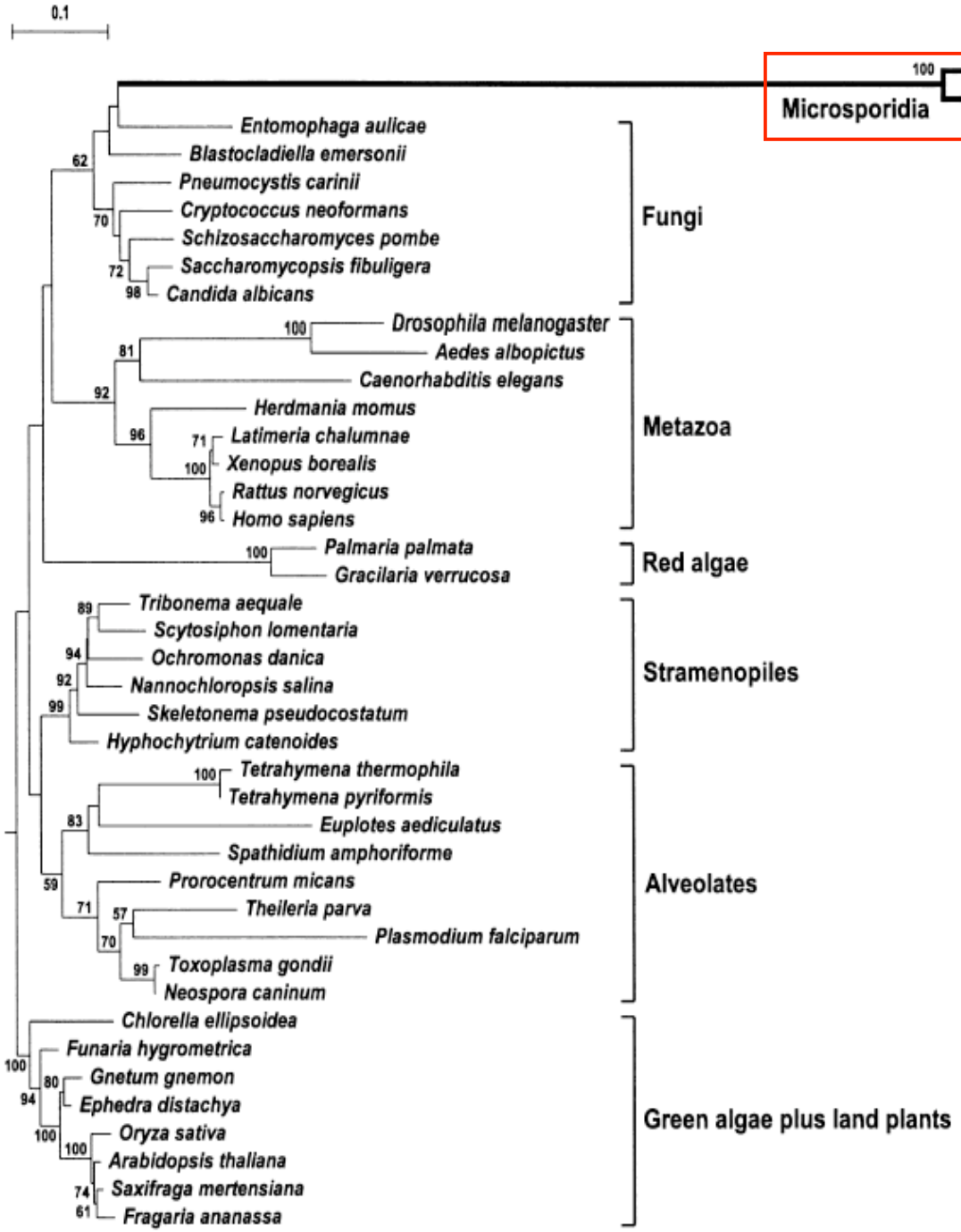
Is it possible to reconcile ribosomal RNAs and RNA polymerases ?

The Long Branch Attraction artifact

[Felsenstein (1978) *Syst Zool* 27:401]



Philippe et al. (2000) Proc. Royal Soc. Lond. B 267:1213.



Distance-based analysis of 42 LSU rRNA sequences from microsporidia and other eukaryotes.

Distances were corrected for site-to-site rate variation.

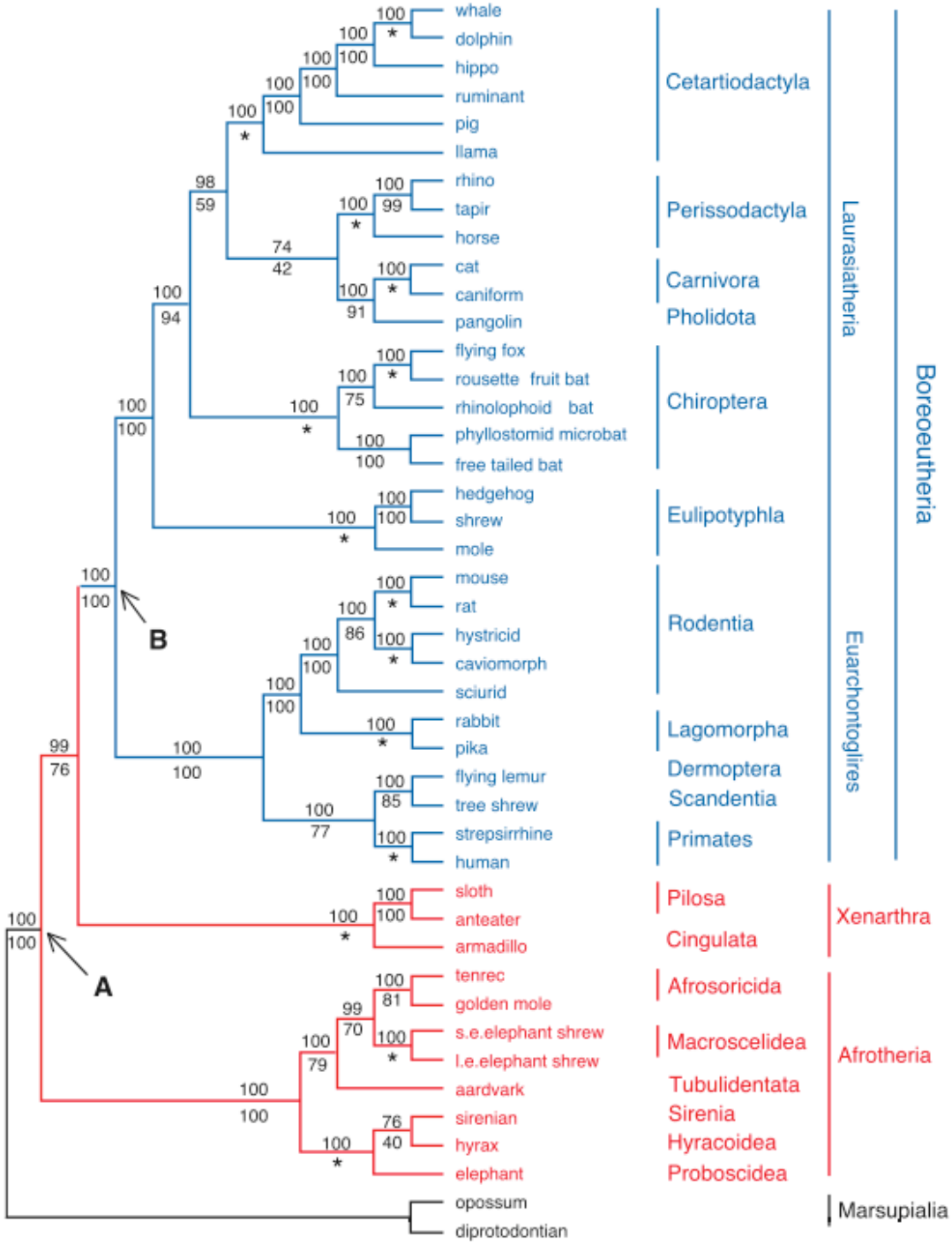
So this analysis uses a more realistic model of molecular evolution.

Van de Peer et al. (2000) Gene 246:1

The effect of the evolutionary model:

more realistic models are better

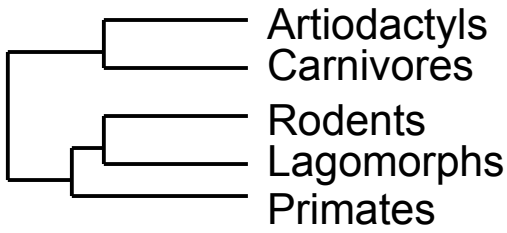
Mammalian phylogeny



Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics

William J. Murphy,[1]* Eduardo Eizirik,[1,2]* Stephen J. O'Brien,[1]†
Ole Madsen,[3] Mark Scally,[4,5] Christophe J. Douady,[4,5]
Emma Teeling,[4,5] Oliver A. Ryder,[6] Michael J. Stanhope,[5,7]
Wilfried W. de Jong,[3,8] Mark S. Springer[4]†

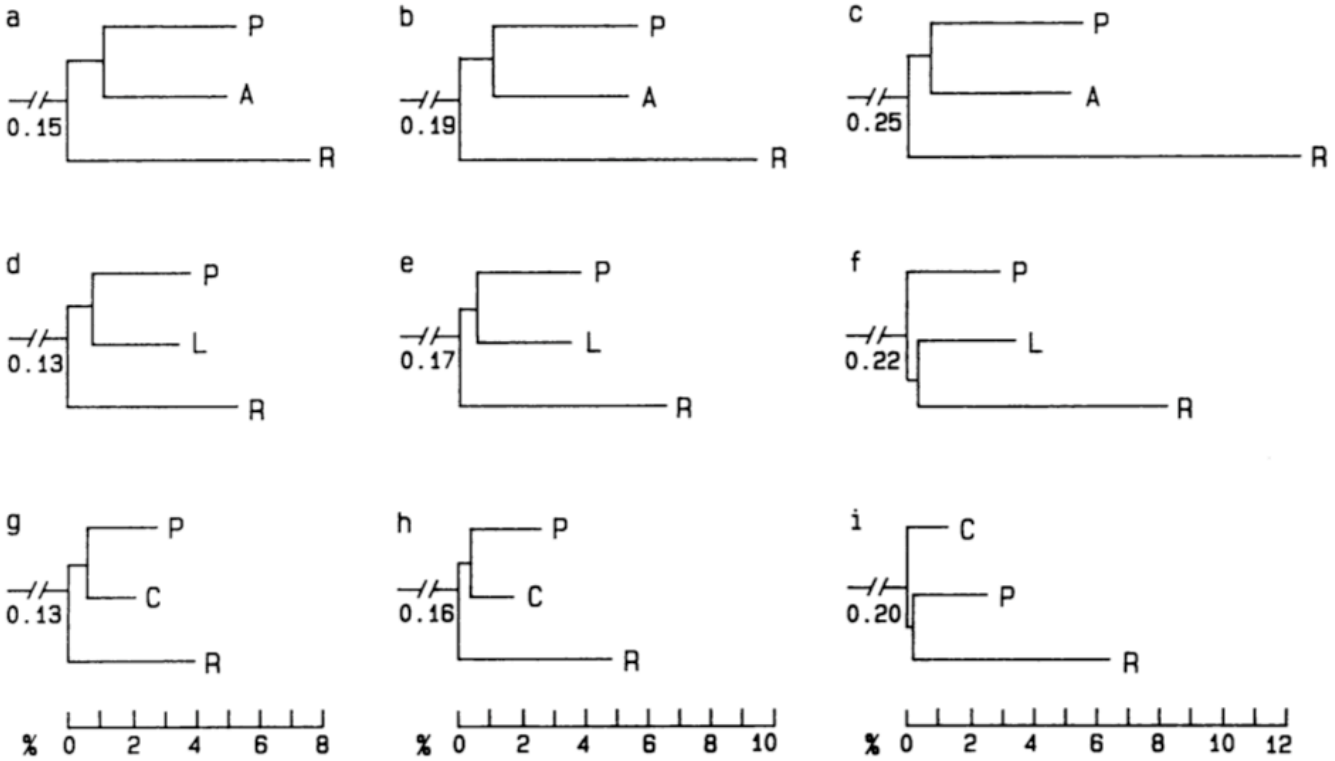
True tree :



no across-site rate variation

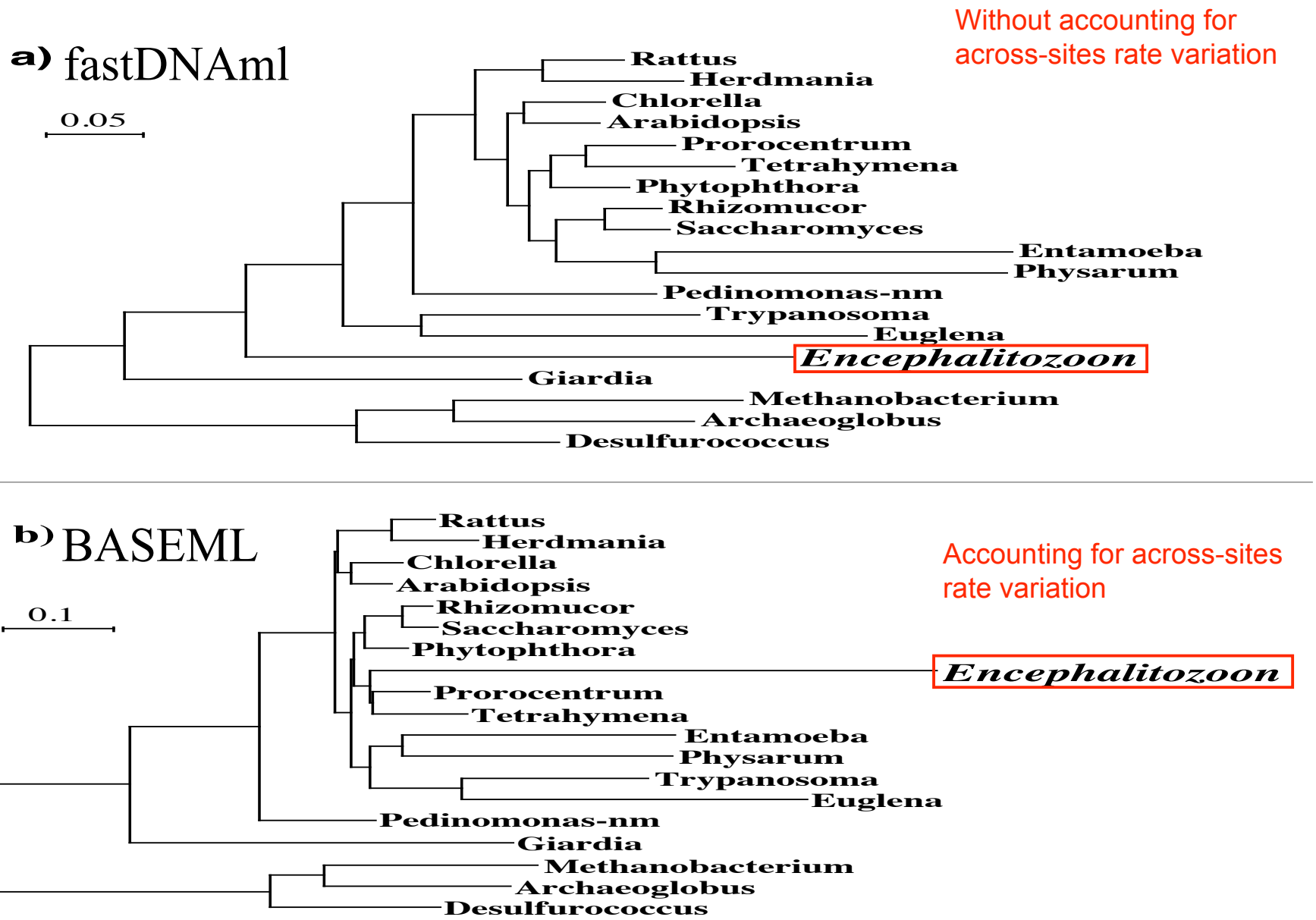
with across-site rate variation

$\alpha = 1$ $\alpha = 0.5$



Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 6703–6707, September 1990

Phylogenetic analysis of LSU rRNA



Peyretaillade et coll. (1998) Nucleic Acids Res 26:3513

The taxon sampling effect

"Molecular phylogeny of the kingdoms Animalia, Plantae, and Fungi"
Gouy & Li (1989)

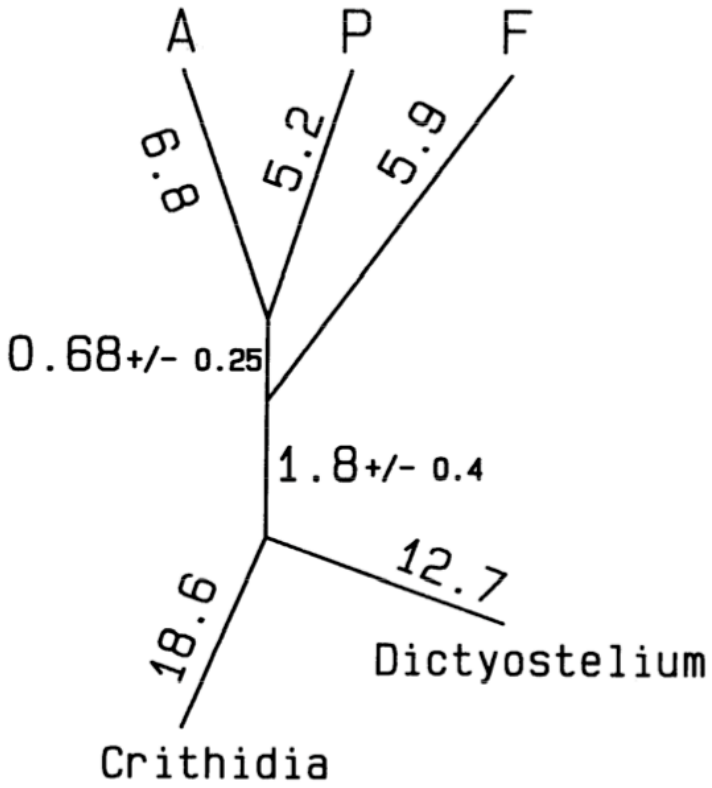


FIG. 2.—Unrooted phylogenetic tree inferred from rRNA sequences. A total of 2,971 sites were analyzed.

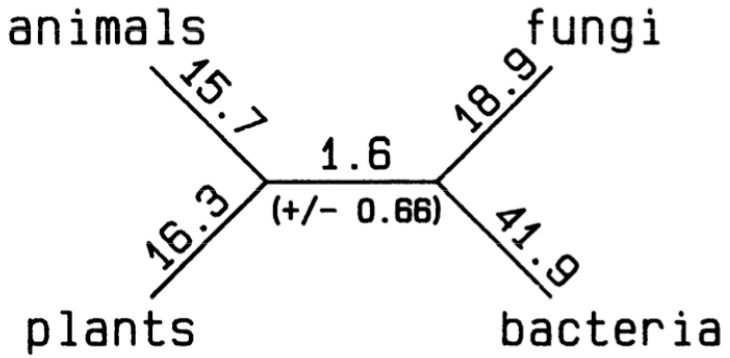
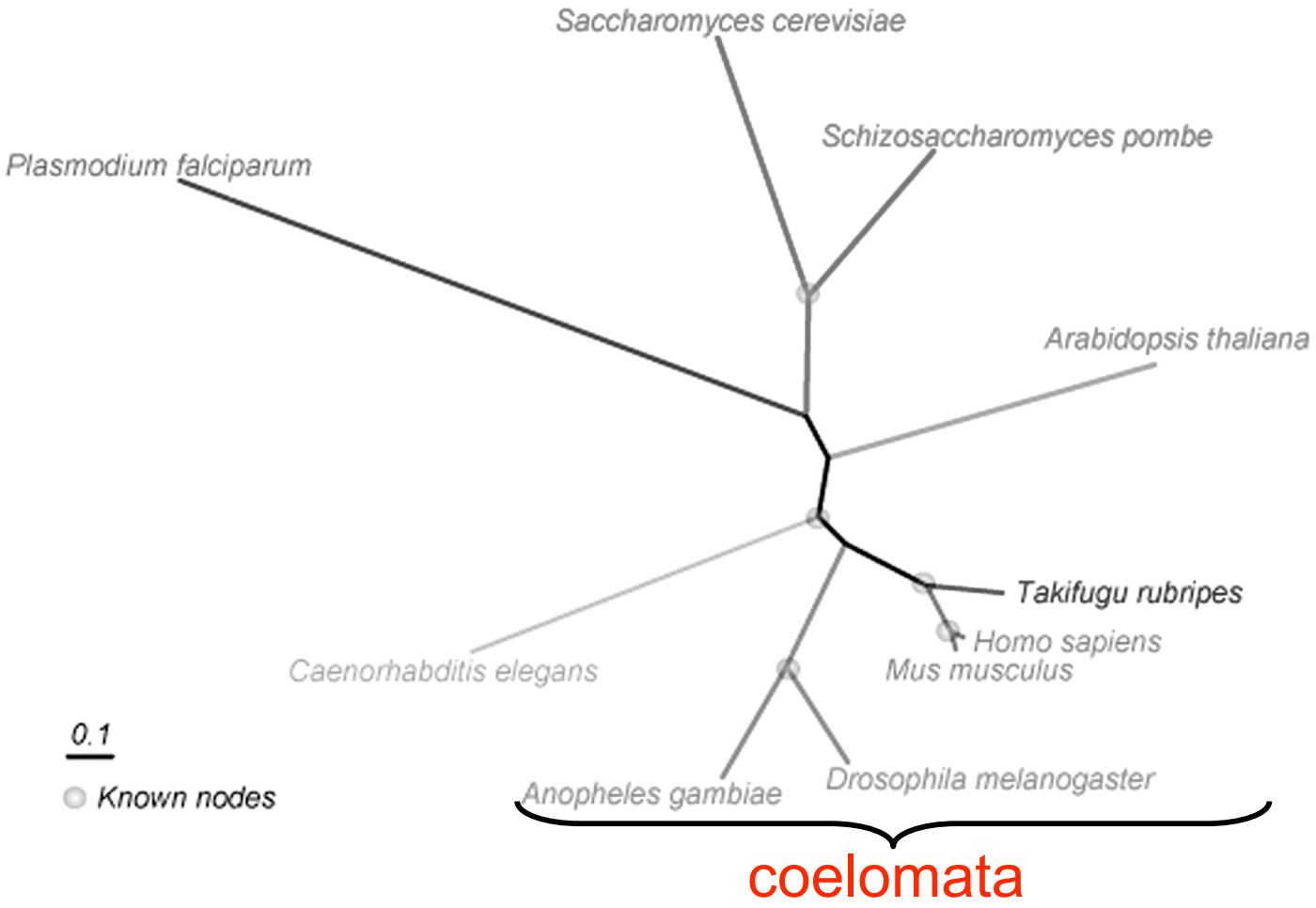


FIG. 4.—Unrooted phylogenetic tree inferred from the pooled protein data set. A total of 1,634 sites were analyzed. Branch lengths are in percent substitutions. The SE of the internal branch length estimate is shown. See table 3 for a description of the data set.

Early analysis with both few genes and few taxa.

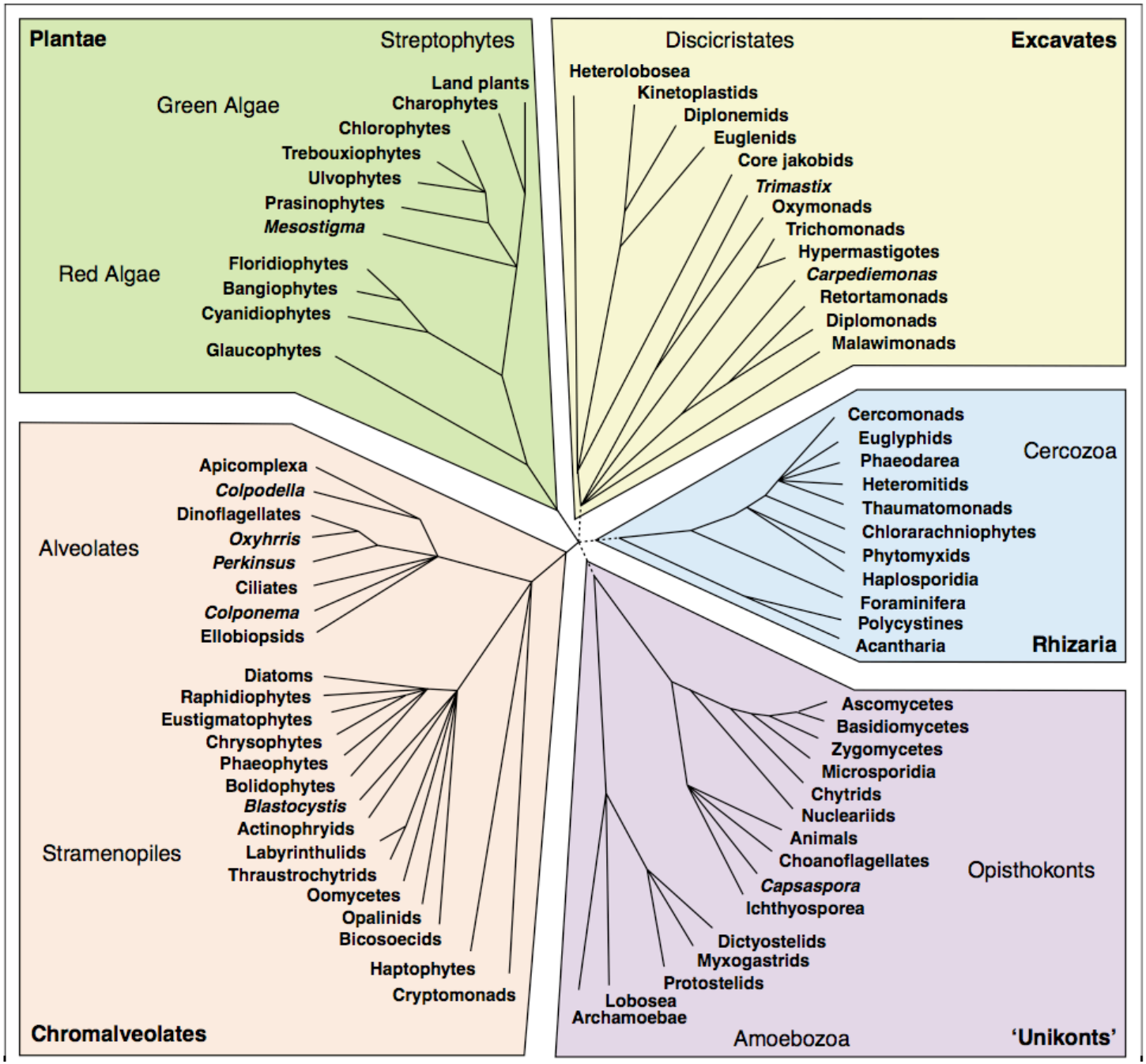
"The Opisthokonta and the Ecdysozoa May Not Be Clades: Stronger Support for the Grouping of Plant and Animal than for Animal and Fungi and Stronger Support for the Coelomata than Ecdysozoa"



Phylogenetic trees from 780 single-gene families from 10 completed genomes and amalgamated into a single supertree. The phylogenetic tree that achieved the best score in 24 of the 26 analyses.

Philip, G. K. et al. Mol Biol Evol 2005 22:1175-1184

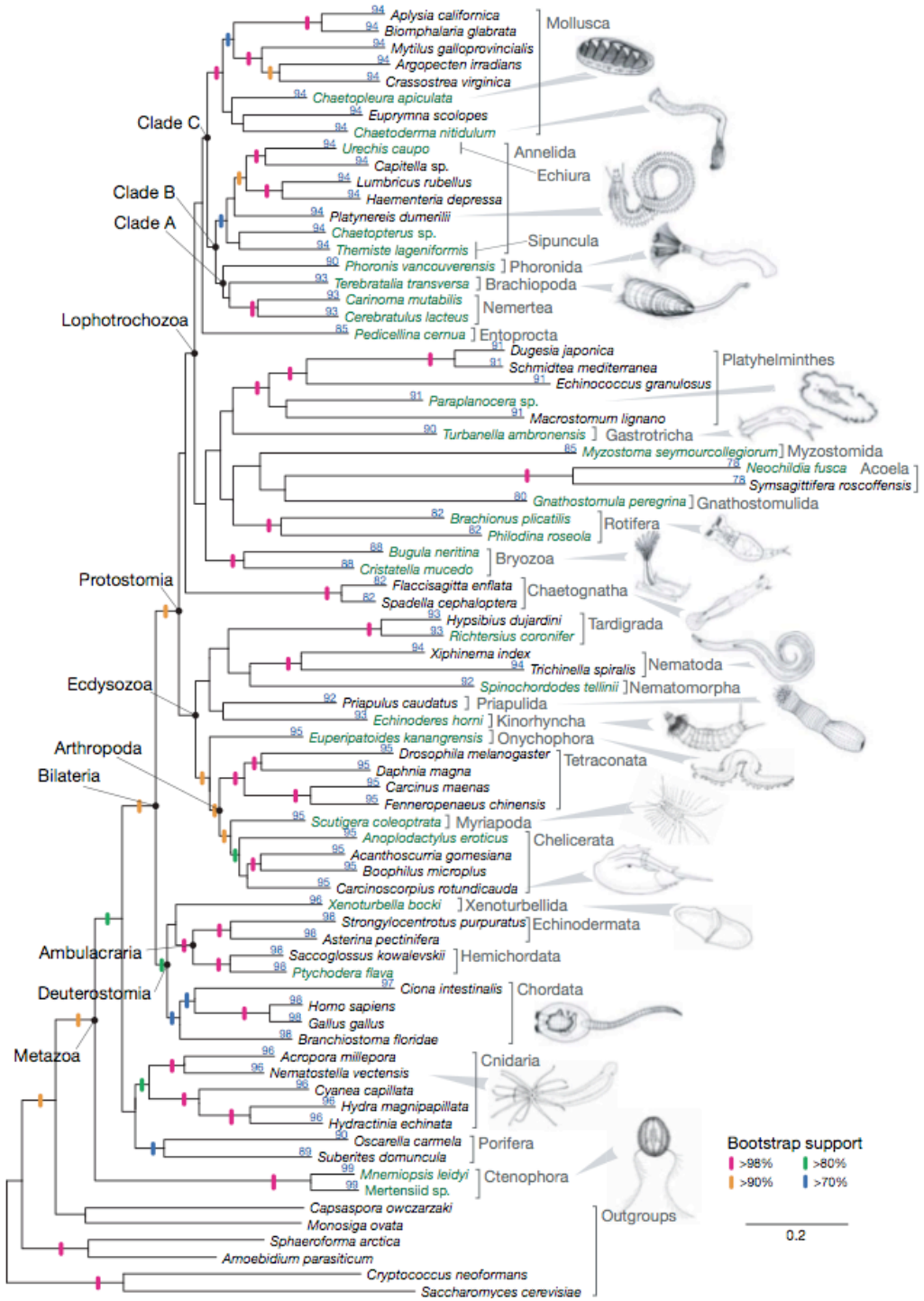
MBE MOLECULAR BIOLOGY AND EVOLUTION



Eukaryotic domain phylogeny.

Emerging consensus for the identification of five super-phyla.

Relationships between them remain very uncertain.



Metazoan phylogeny

- Rejection of the acoelomate, pseudocoelomate, coelomate concept; division lophotrochozoa / ecdysozoa

- Bilataria vs. cnidaria, porifera et ctenophora

- Protostomes vs. Deutérostomes

Vol 452 | 10 April 2008 | doi:10.1038/nature06614 nature

LETTERS

Broad phylogenomic sampling improves resolution of the animal tree of life

Deep phylogeny is difficult. It requires many genes and many taxa. Comb-like trees may be indicative of LBA artefact.

Resolving the phylogeny of a newly sequenced clade is difficult if it is not closely related to something already well placed.

Sampling several members of a new clade is very beneficial: slowly evolving lineages can be identified.

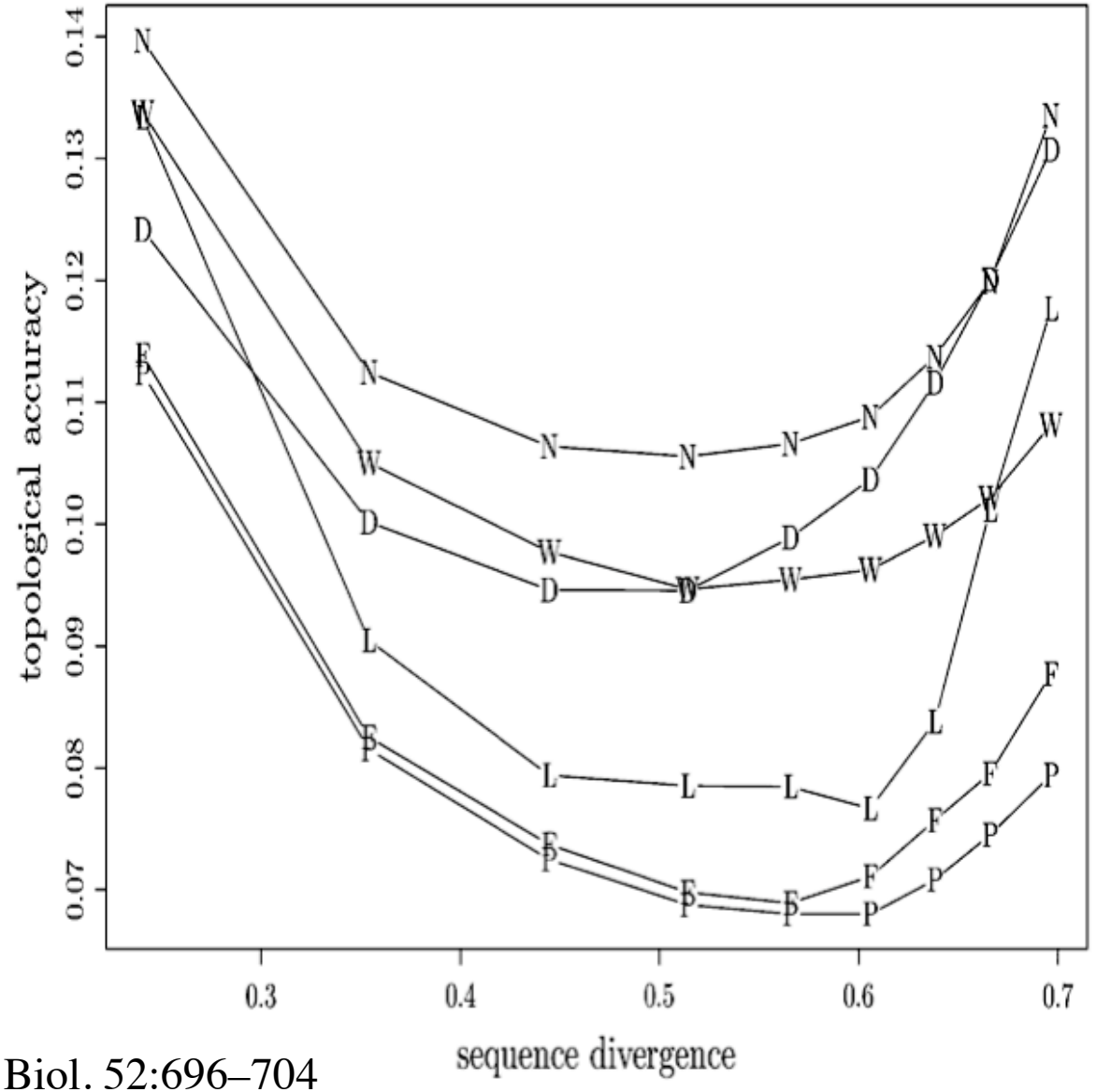
Accounting for across-sites rate variation is necessary in most cases (except for synonymous sites of protein-coding sequences).

In prokaryotes, horizontal transfers can deeply disconnect gene trees and species trees.

Comparison of method performances by sequence and tree simulation experiments

P, PHYML
F, fastDNAml
L, NJML
D, DNAPARS
N, NJ

5000 random trees
40 taxa, 500 bases
no molecular clock
variable sequence divergences
K2P, $\alpha = 2$



Guindon & Gascuel (2003) Syst. Biol. 52:696–704

Comparison of running times for various tree-building algorithms

TABLE 1. Average run times for various methods. The computing times were measured on a 1.8-GHz (1 Go RAM) PC with Linux. For PHYML, the number in parentheses is the average number of refinement stages.

	Simulations		Real data	
Method	40 taxa (500 bp)	100 taxa (500 bp)	218 taxa (4,182 bp)	500 taxa (1,428 bp)
DNADIST+ NJ/BIONJ	0.3 sec	2.3 sec	50 sec	2 min, 19 sec
DNADIST+ Weighbor	1.5 sec	22 sec	4 min, 52 sec	58 min, 40 sec
DNAPARS	0.5 sec	6 sec	4 min, 4 sec	13 min, 12 sec
PAUP*	3 min, 21 sec	1 hr, 4 min		
PAUP*+ NJ	1 min, 10 sec	22 min	10 hr, 50 min	
MrBayes	2 min, 6 sec	32 min, 37 sec		
fastDNAml	1 min, 13 sec	26 min, 31 sec		
NJML	15 sec	6 min, 4 sec		
MetaPIGA	21 sec	3 min, 27 sec	4 hr, 45 min	9 hr, 4 min
MetaPIGA+ NJ	6 sec	23 sec	1 hr, 40 min	3 hr
PHYML	2.7 sec (6.4)	12 sec (8.3)	8 min, 13 sec (15)	11 min, 59 sec (13)

distance < parsimony ~ PHYML << Bayesian < classical ML
NJ DNAPARS PHYML MrBayes fastDNAml,PAUP*

Guindon & Gascuel (2003) Syst. Biol. 52:696–704