Assessing the limits of phylogenomics: can too much data be a bad thing?

Olaf R. P. Bininda-Emonds

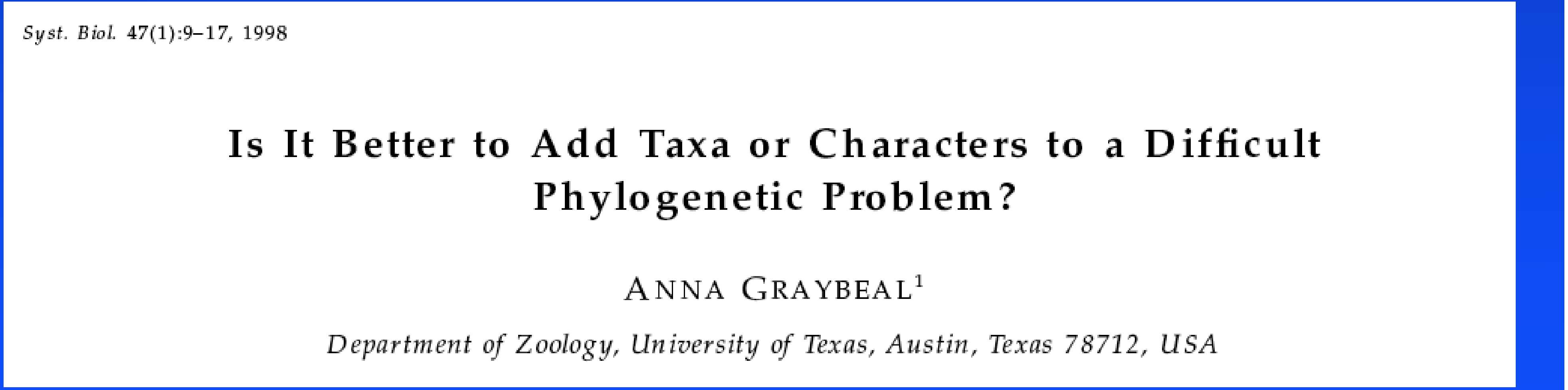
Cral von Ossietzky Universität Oldenburg

Outline

- a historical perspective: then vs. now
  - then → problems with limited data
  - now → problems with too much data???
- a brief look at noise
- effects of large problem sizes
  - number of taxa
  - number of characters
- conclusions

Some (not-so-distant) history

- early molecular phylogenetic studies faced problems of limited data
  - taxa → four-taxon problem and long-branch attraction
  - characters → single-gene phylogenies (at best!)



Syst. Biol. 47(1):9–17, 1998

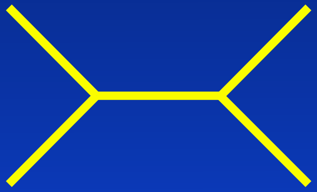
Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem?

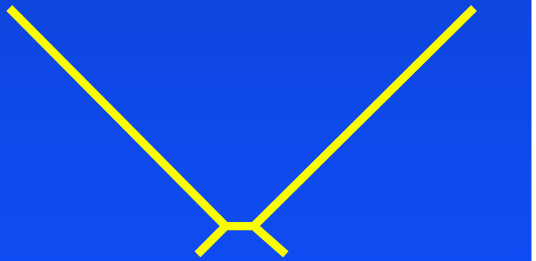
ANNA GRAYBEAL[1]

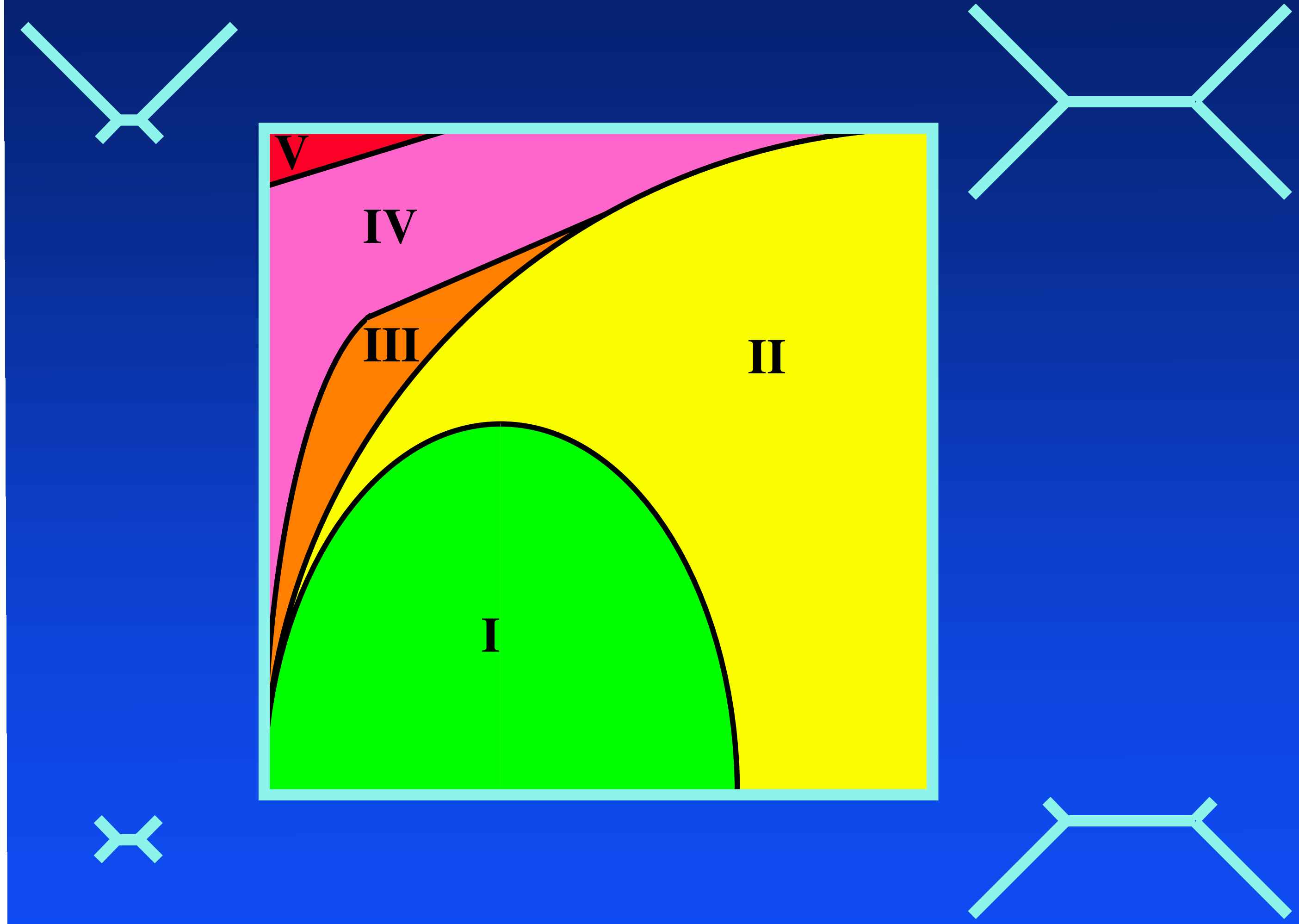
Department of Zoology, University of Texas, Austin, Texas 78712, USA

The four-taxon problem

- four-taxon trees can be very difficult to reconstruct accurately (Hillis *et al*., 1994)
  - requires 1000s of nucleotides
  - even infinite amounts of data may not work in some instances

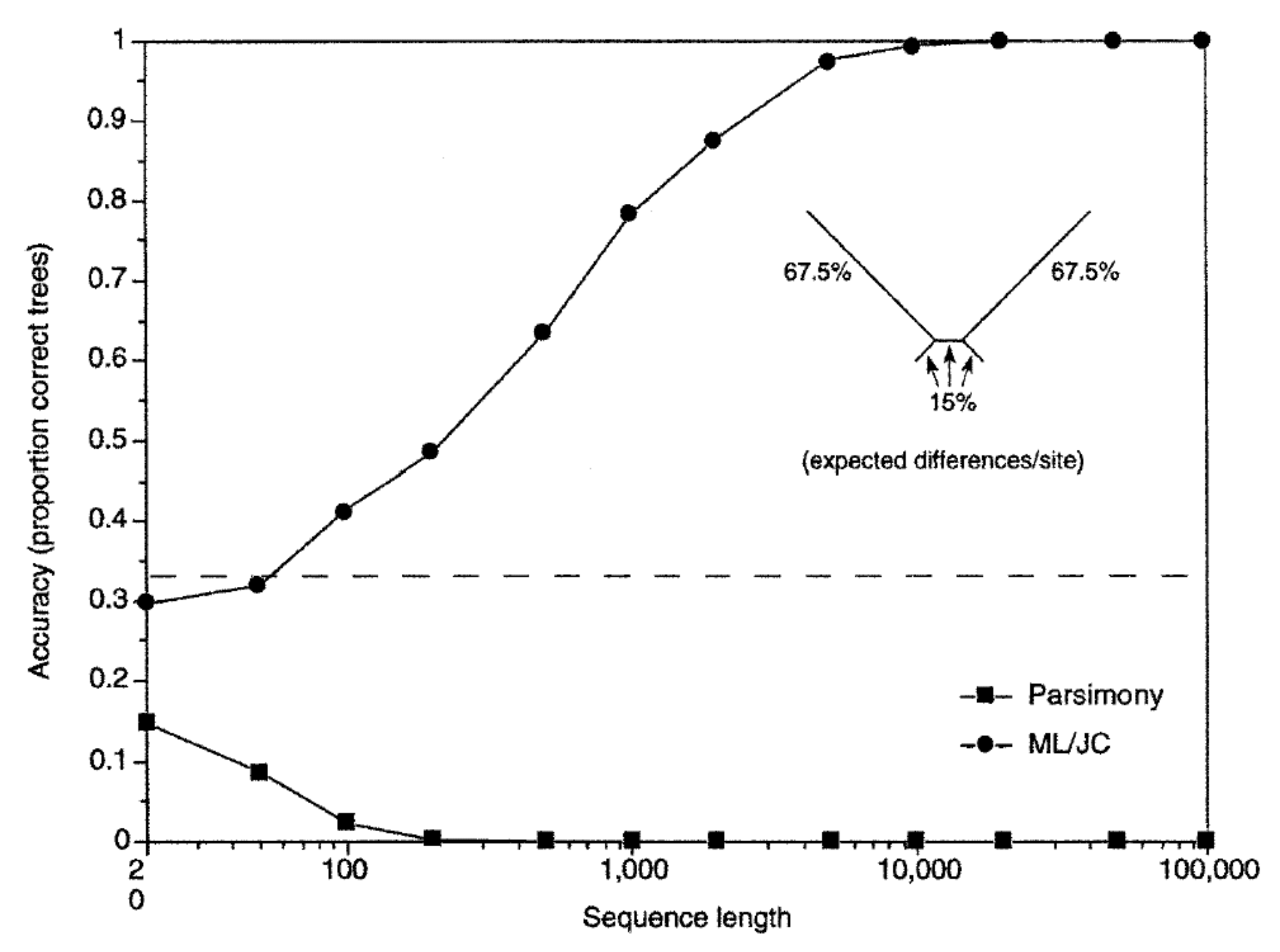




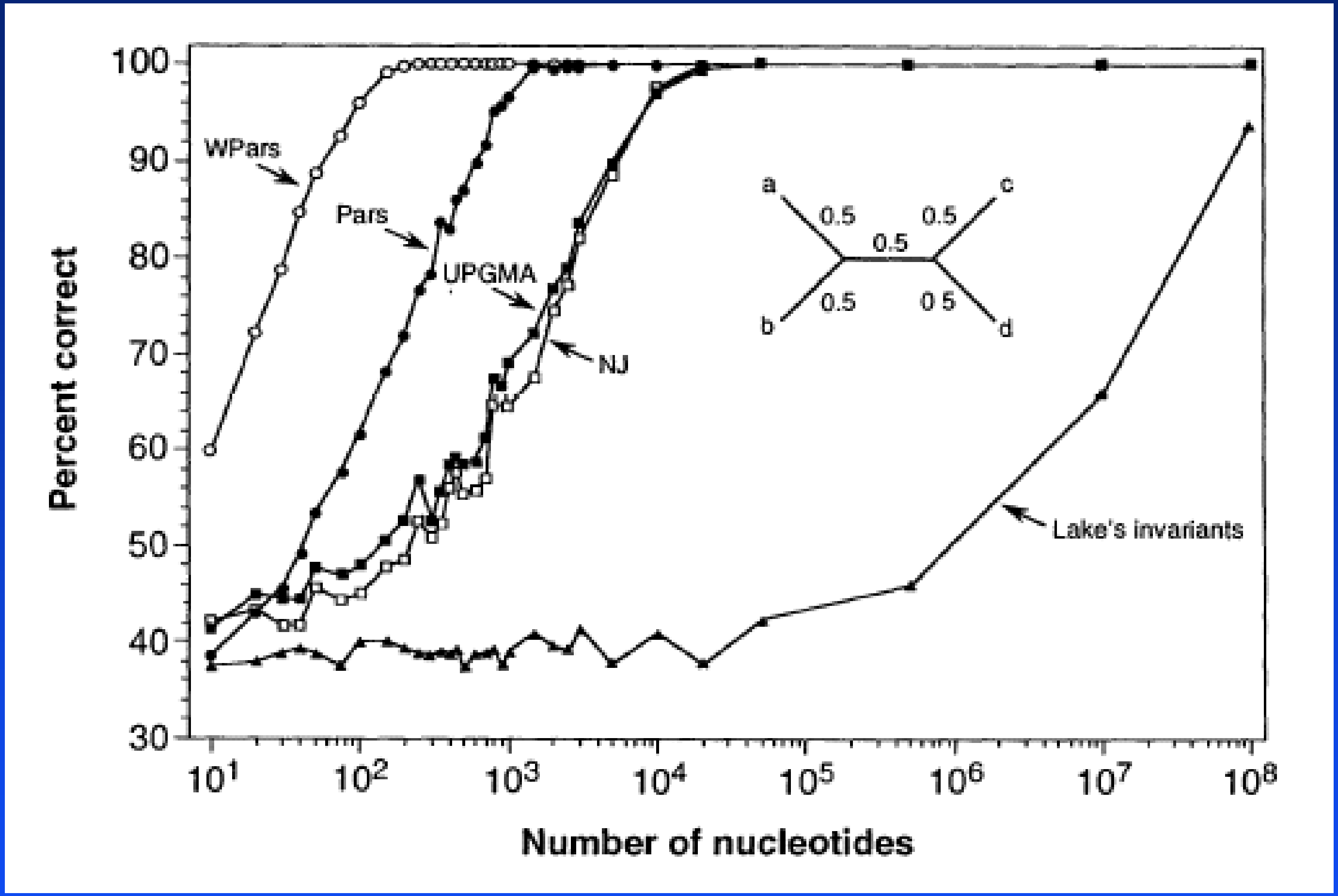


- I: most methods perform well
- II: methods require lots of data or higher weighting of more slowly evolving characters
- III–V: most methods perform increasingly poorly or are positively misleading

- from Huelsenbeck and Hillis (1993)

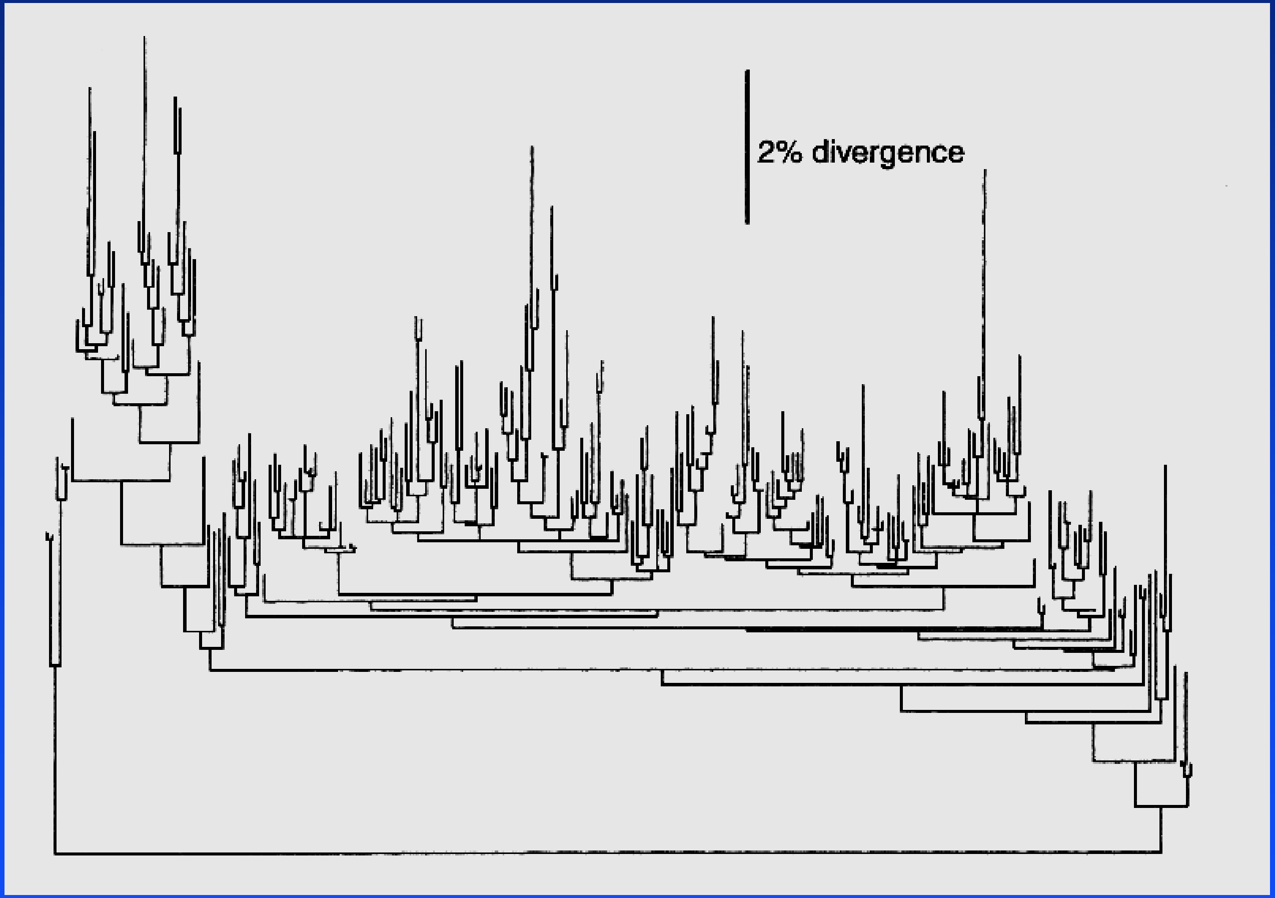


- from Swofford *et al*. (2001)



- from Hillis *et al*. (1994)

Angiosperm phylogeny — 18S rDNA



- 228 species
- complete sequences (1855 bp)

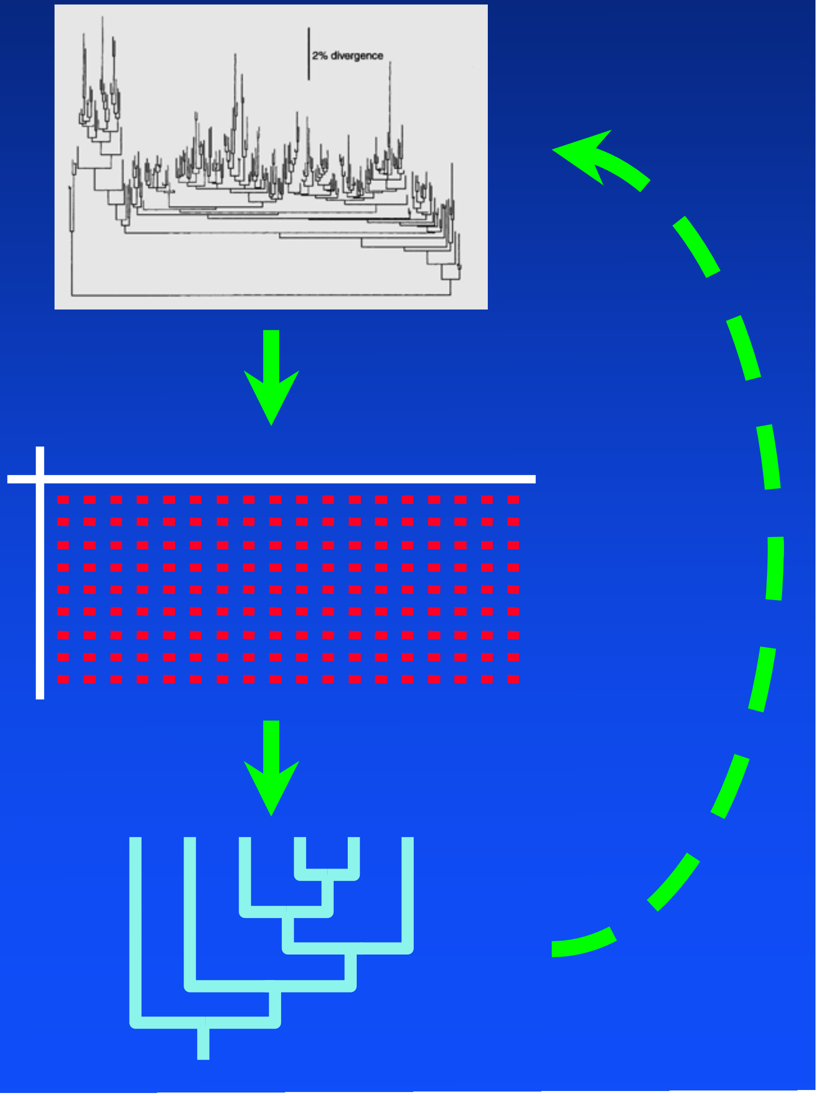
- from Hillis (1996) (data from Soltis *et al*., 1997)

Parametric bootstrapping

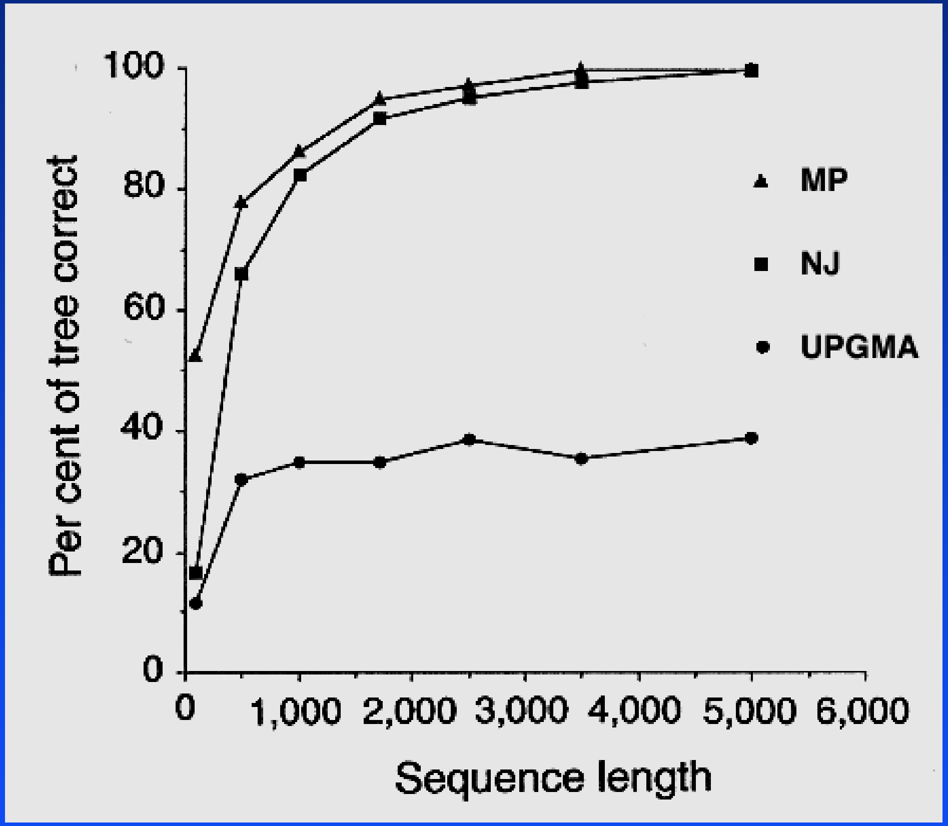
- simulate data according to a specified model of evolution down a model tree

  e.g., in Hillis (1996), the model tree was the inferred phylogeny of Soltis *et al*. (1997)

- analyze data to obtain an estimated tree
- compare model and estimated trees



How accurate are “large” phylogenies?

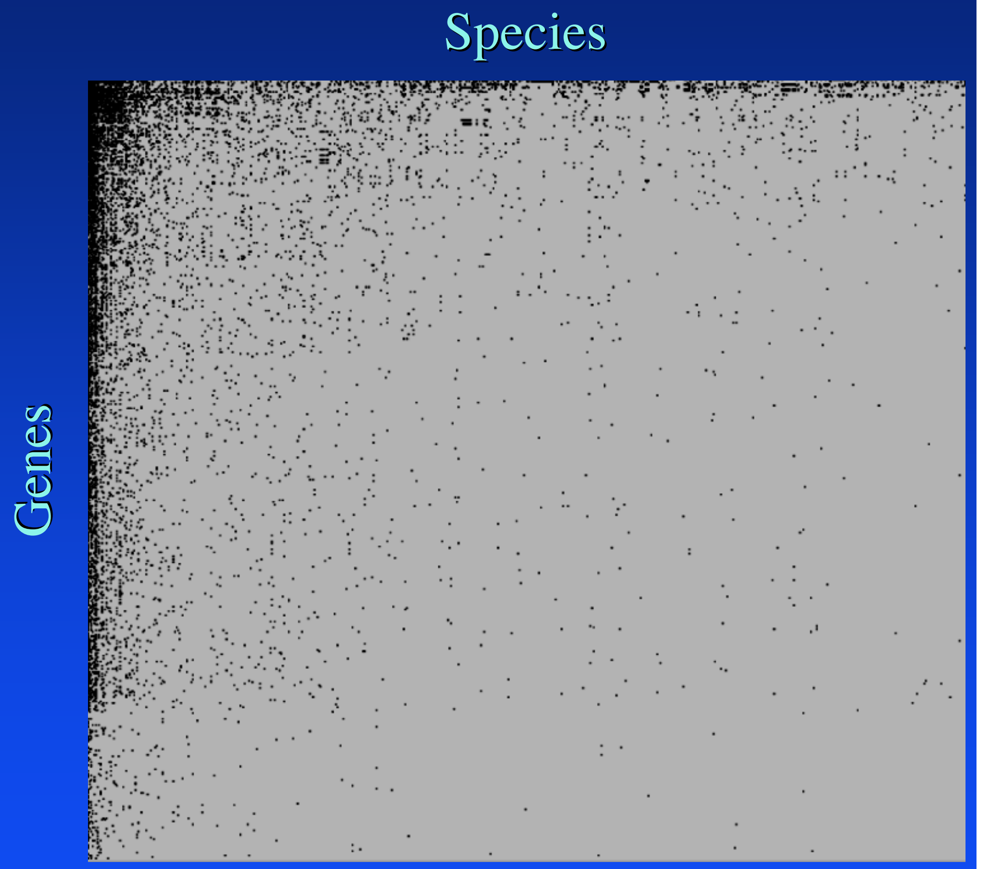


- stunning answer:
  - about 1.2 × 10^{502} possible solutions for 228 species
  - > 95% accuracy with only 5000 nucleotides but **without** branch swapping (MP or NJ)

- from Hillis (1996)

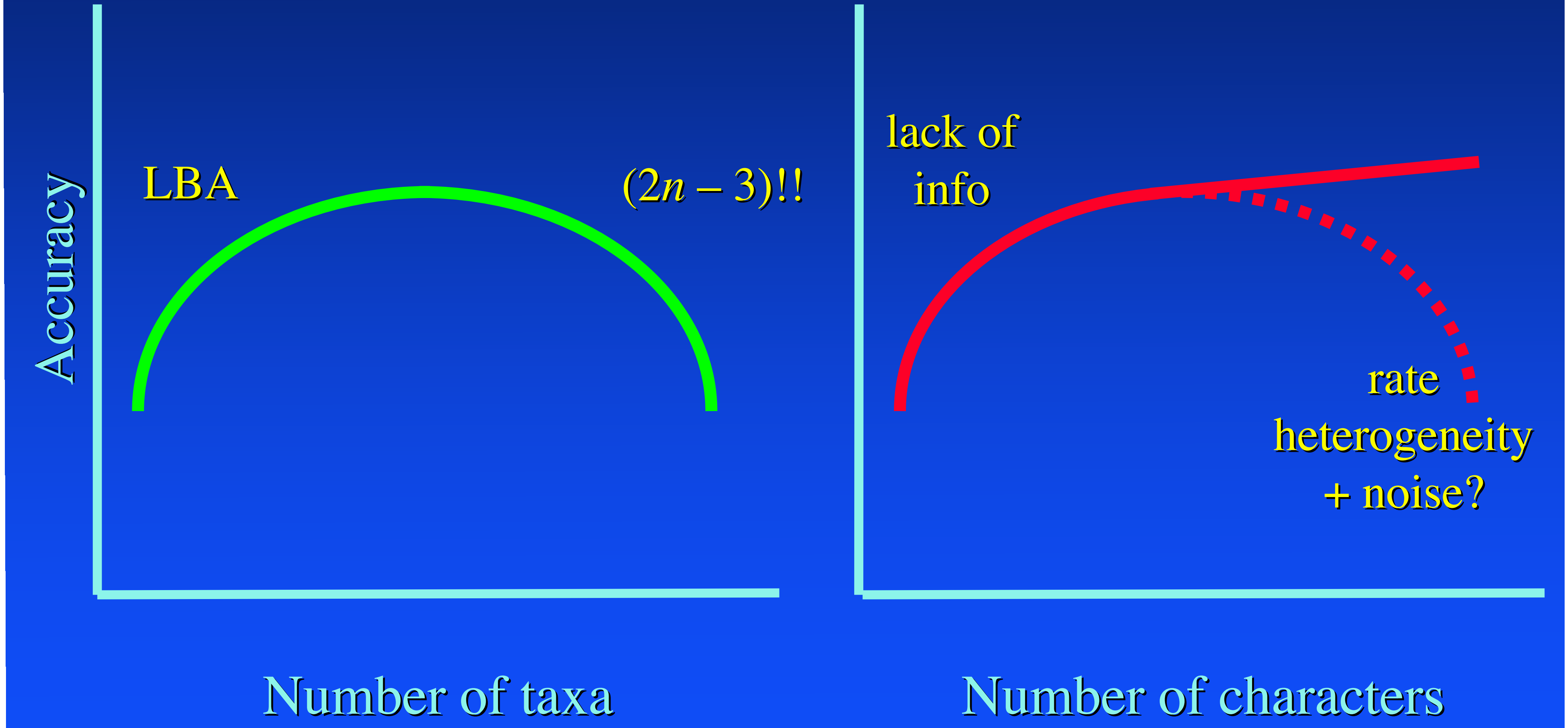
Phylogenomics: promise and perils

- data are increasingly not the limiting factor
  - large numbers of taxa
  - huge number of characters
- but too much data could also be a problem …
  - lot of attention paid to taxa, much less to characters



- “data availability matrix” for green plants (from Sanderson and Driskell, 2003)

Expectations

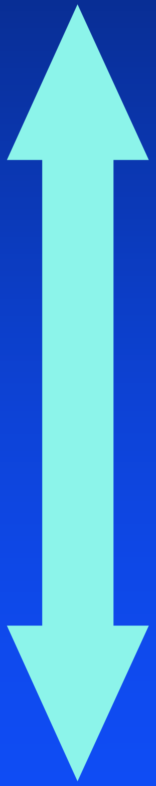


Phylogenetic noise

- no clear definition
  - ≈ anything that is not "phylogenetic signal"
  - commonly viewed as fast-evolving, highly saturated sites

Dealing with noise

ignore it (default)



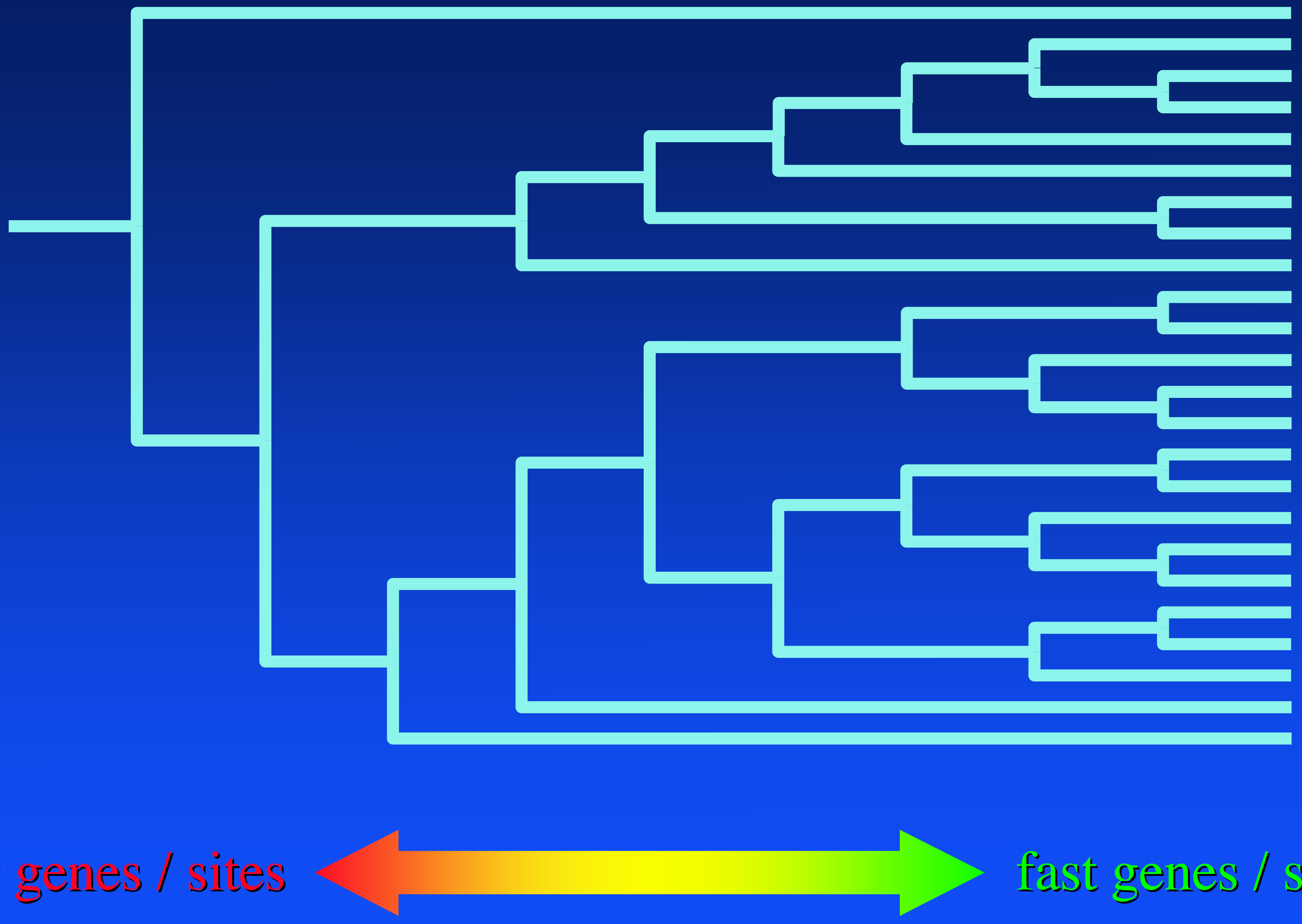
work with it

- gene- or codon-partitioned models
- gamma distribution
- recoding (RY, AA, redundancy)

remove it

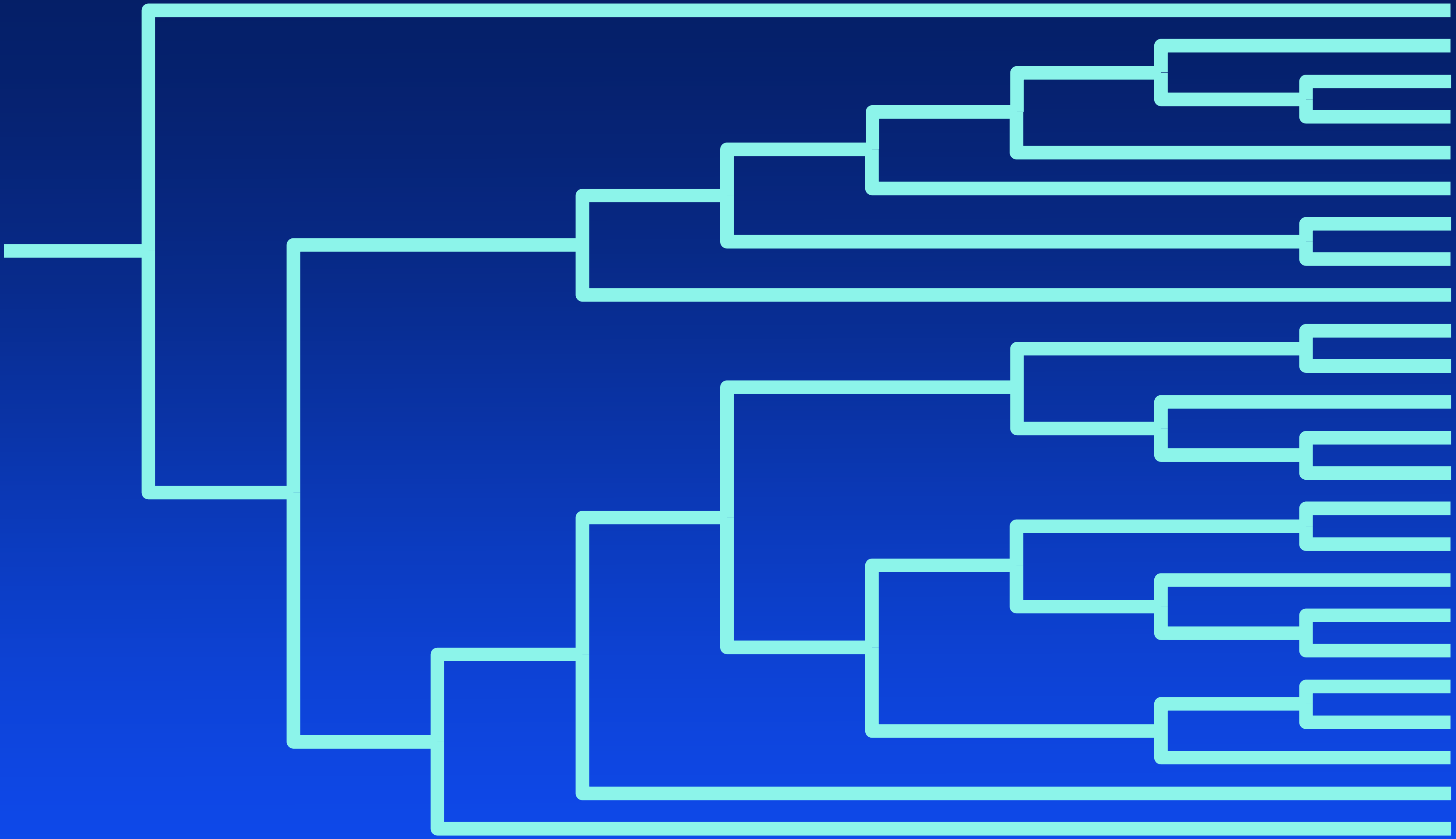
Noise can be signal!

- globally
  - majority of signal in *rbc*L phylogeny of 2538 angiosperm species was coming from 3rd codon positions (Källersjö *et al*. 1999)
- locally
  - any negative effects of “noise” only manifest themselves going towards the root of the tree …



slow genes / sites

fast genes / sites



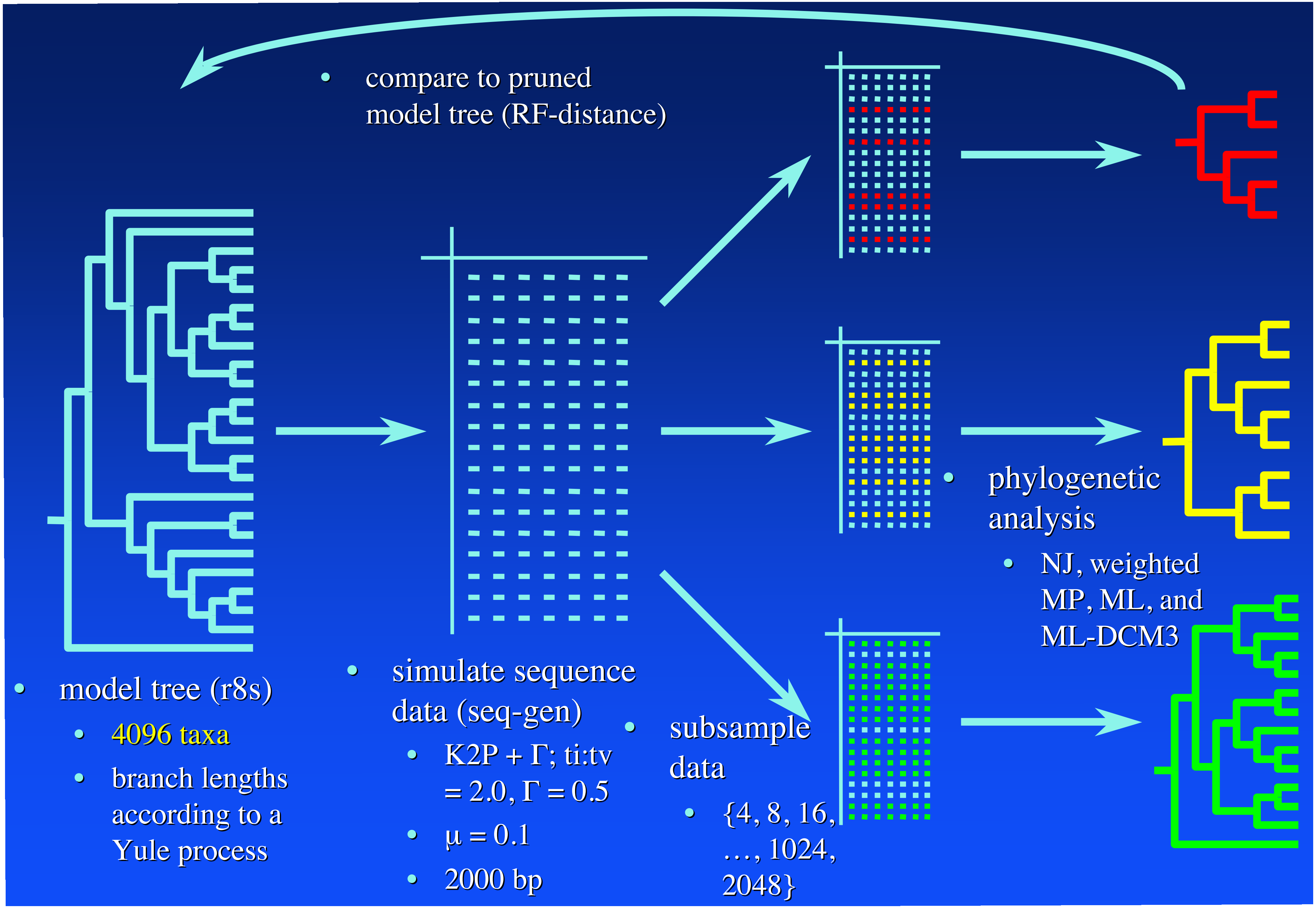
misleading (?) ← fast genes / sites → informative

informative ← slow genes / sites → uninformative

Too many characters?

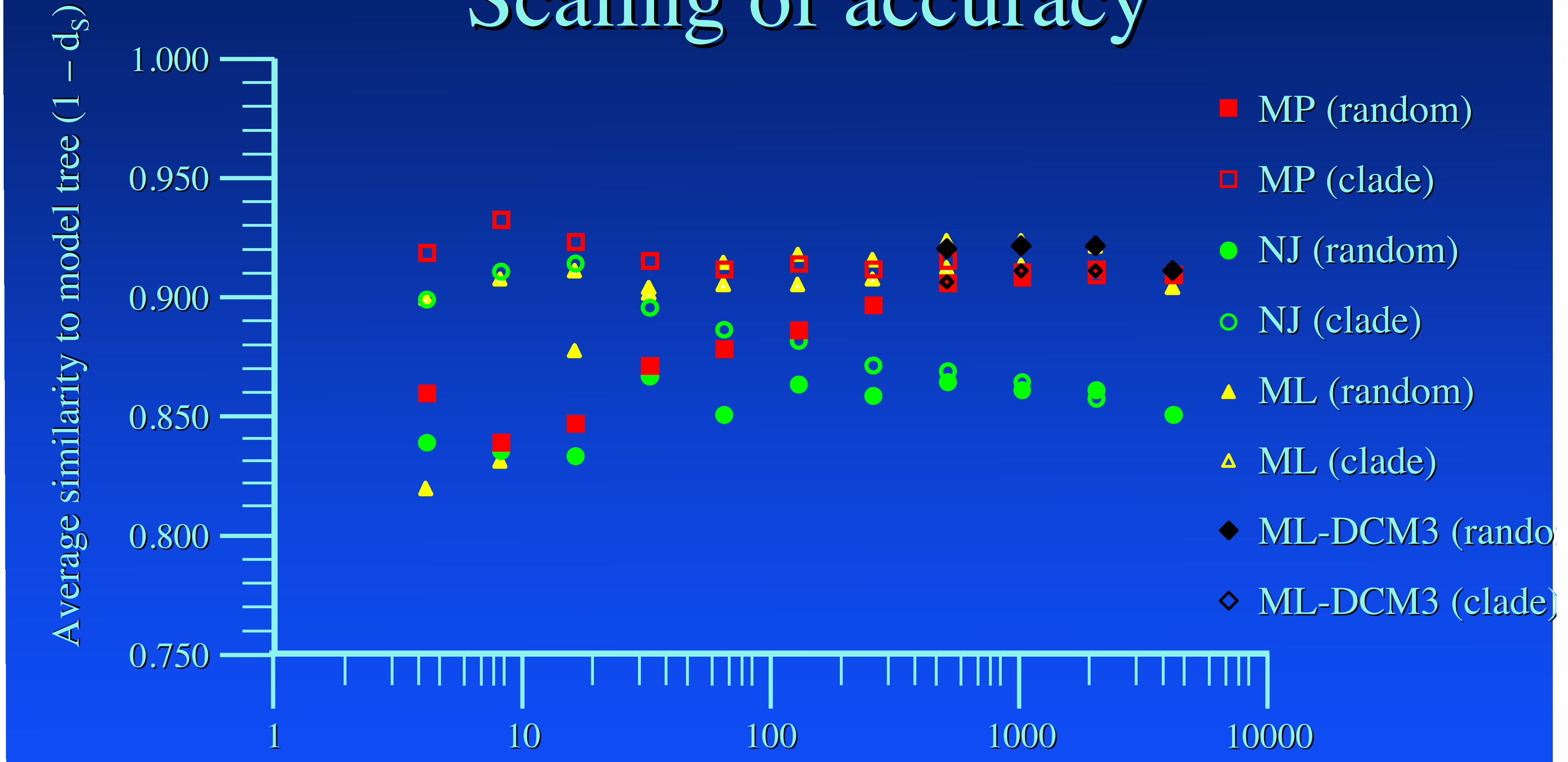
- phylogenomic data sets include many genes, often with very different rates
  - rate heterogeneity needed for full resolution across tree
  - but fast genes should generate noise at deeper levels
- problem: how to gain resolution from fast genes but control for noise?
  - or is this even a problem???

Increasing numbers of taxa



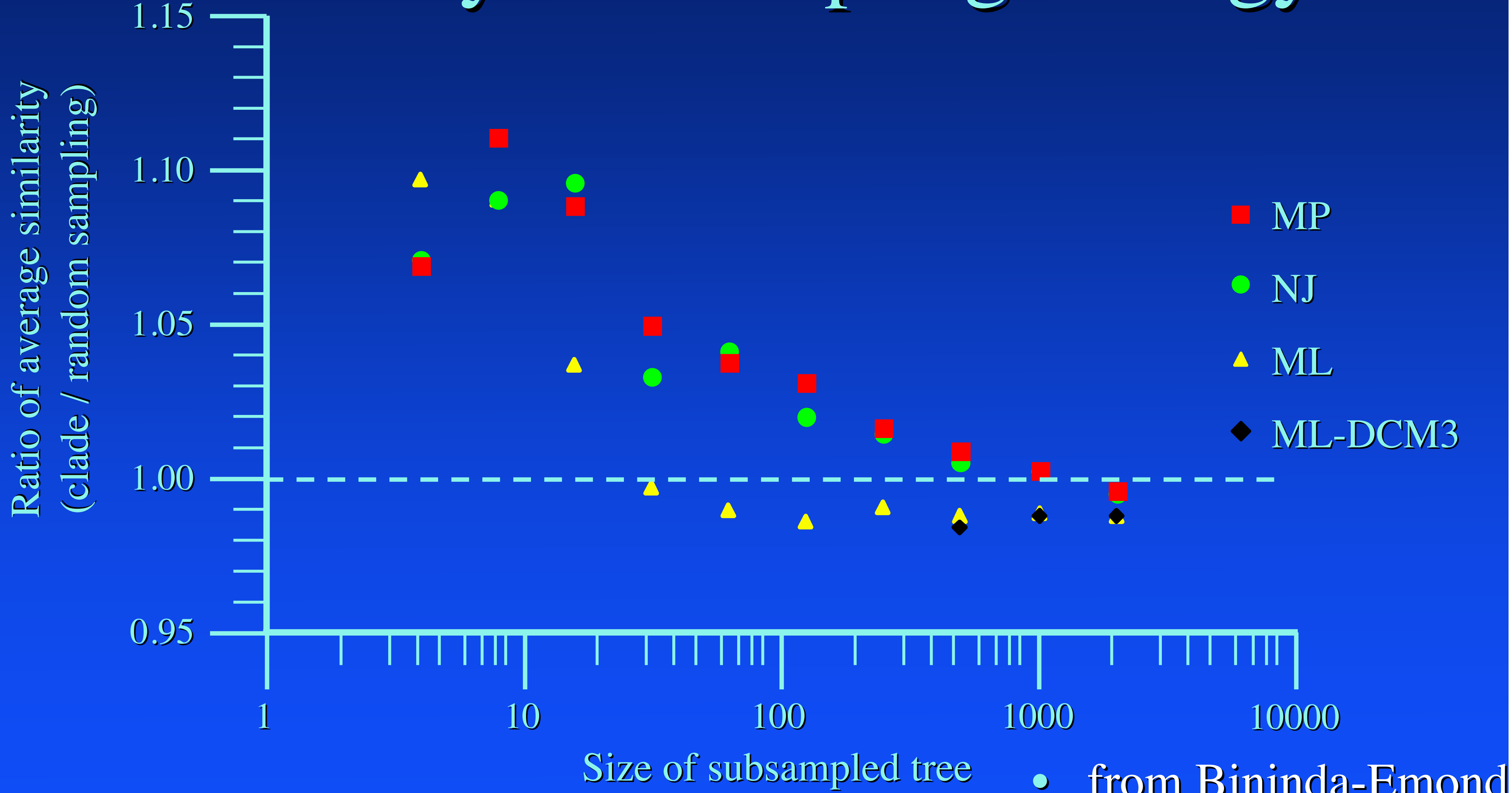
- model tree (r8s)
  - 4096 taxa
  - branch lengths according to a Yule process
- simulate sequence data (seq-gen)
  - K2P + Γ; ti:tv = 2.0, Γ = 0.5
  - μ = 0.1
  - 2000 bp
- subsample data
  - {4, 8, 16, …, 1024, 2048}
- phylogenetic analysis
  - NJ, weighted MP, ML, and ML-DCM3
- compare to pruned model tree (RF-distance)

Scaling of accuracy



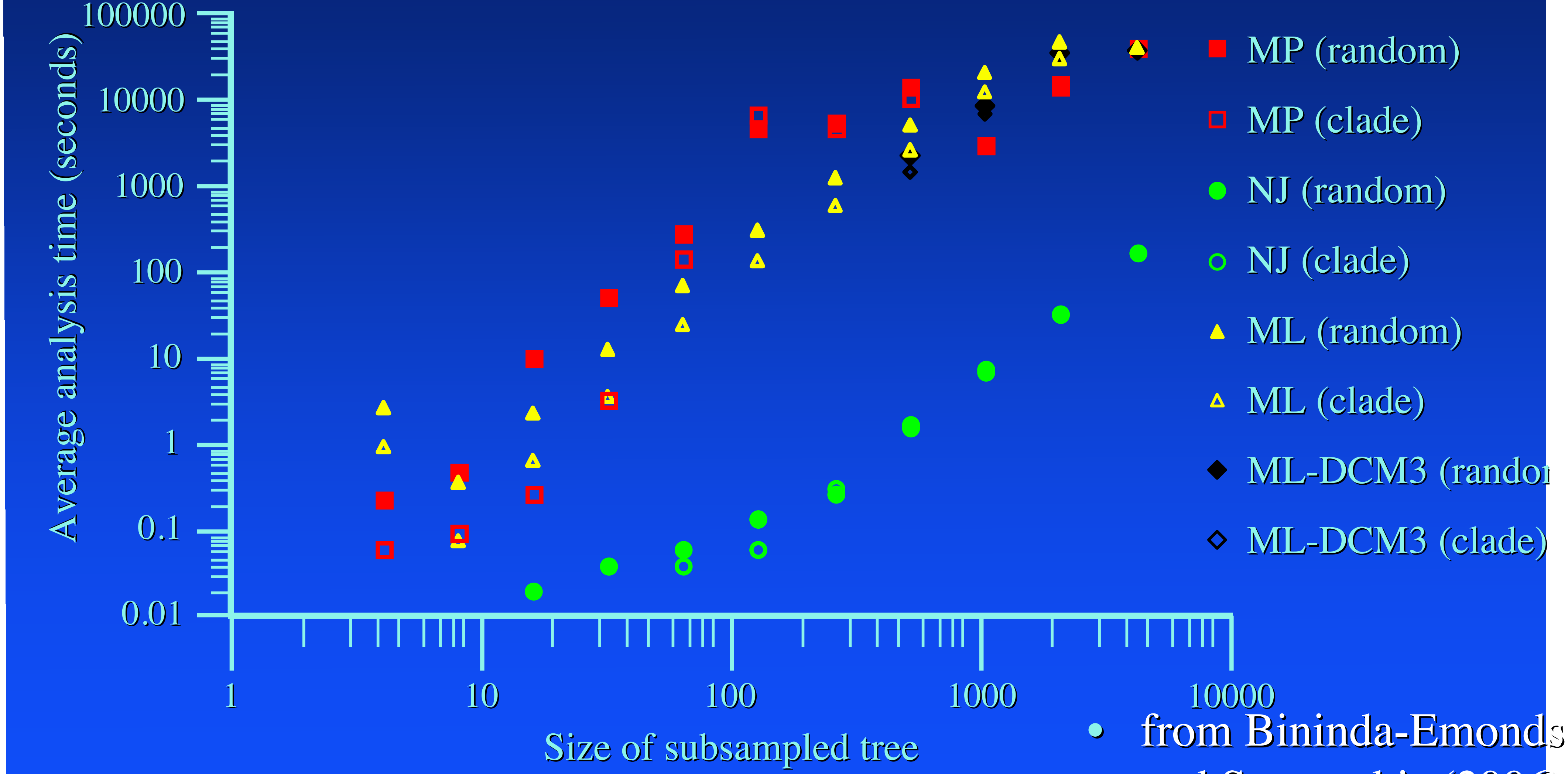
- from Bininda-Emonds and Stamatakis (2006)

Accuracy and sampling strategy



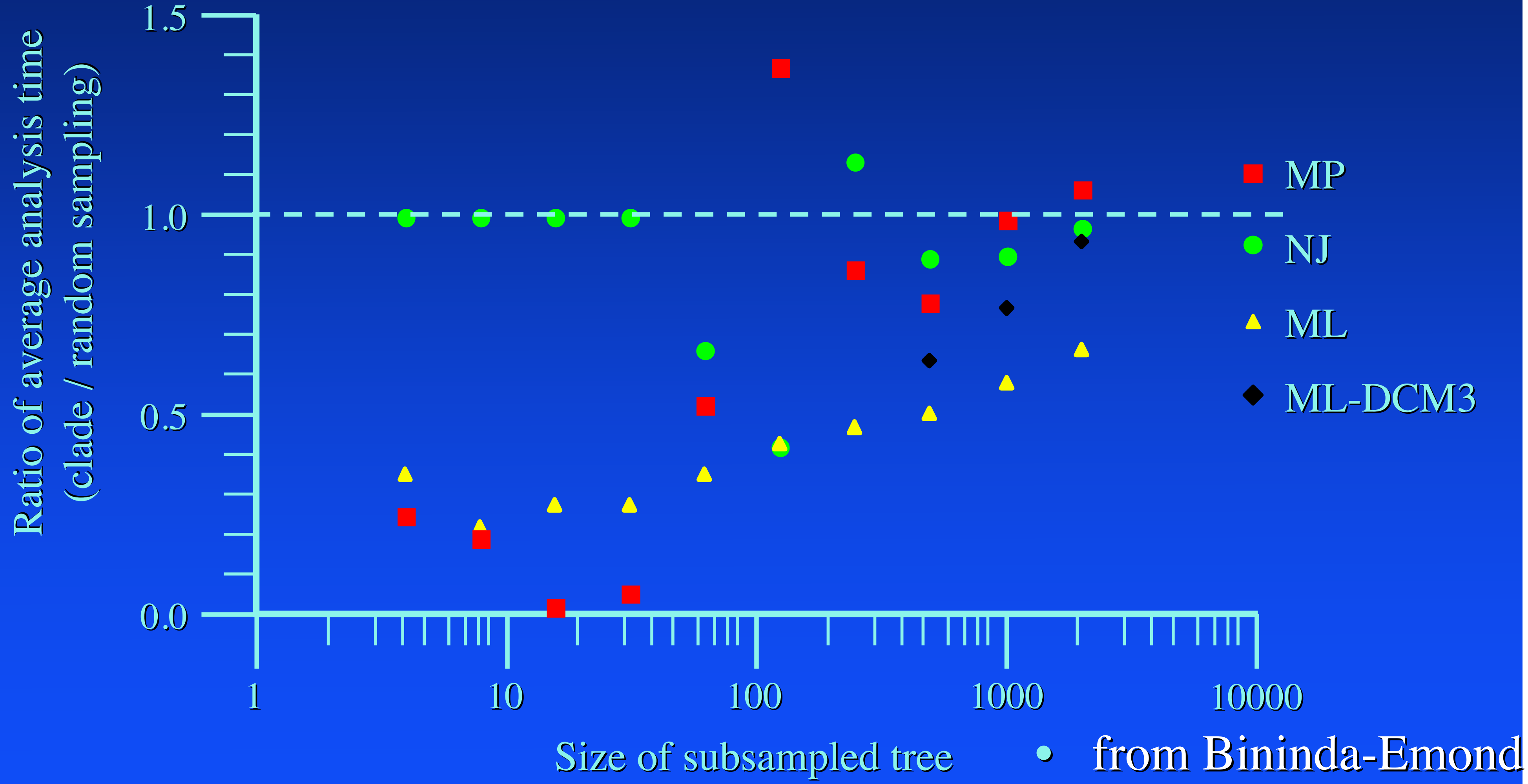
- from Bininda-Emonds and Stamatakis (2006)

Scaling of analysis time



- from Bininda-Emonds and Stamatakis (2006)

Analysis time and sampling strategy



- from Bininda-Emond
and Stamatakis (2006

Conclusions – large taxon problems

- seemingly no drop-off in accuracy up to “moderate” problem sizes
  - important to have complete sampling
  - unanswered question: how complete is complete enough?
- inherent trade-off between time and accuracy
  - complete (= compact?) sampling
  - parallelization
  - new, faster heuristics (including divide-and-conquer approaches)

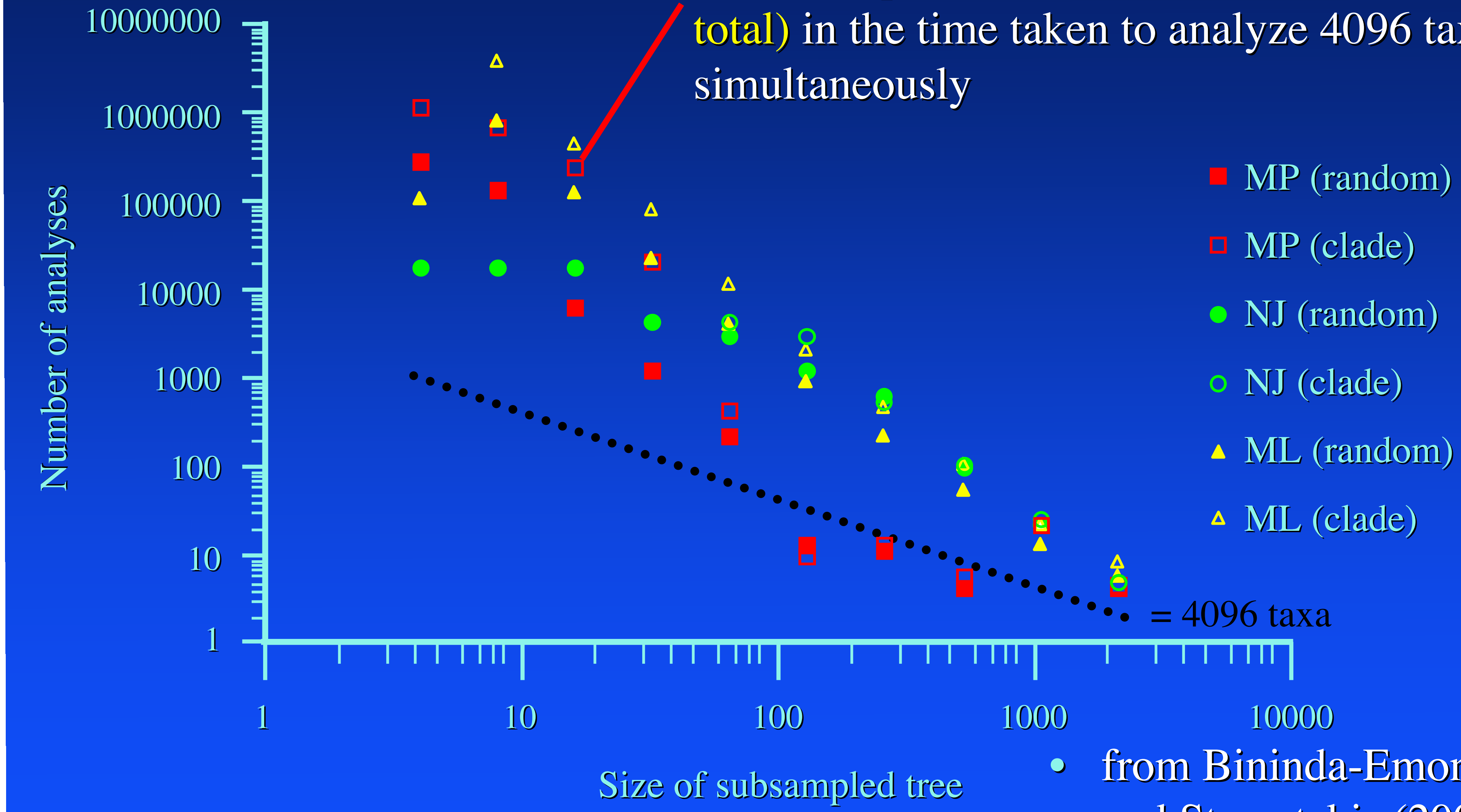
Does divide-and-conquer work?

- it should / could:
  - tremendous speed gain to analyzing many, smaller problems:

$$\text{time} \sum_{1}^{n} x \ll \text{time } nx$$

  - accuracy ~flat with respect to problem size

- e.g., can run ~250 000 MP analyses of 16 clade-sampled taxa (≈ 4 000 000 taxa in total) in the time taken to analyze 4096 taxa simultaneously



- from Bininda-Emonds and Stamatakis (2006)

Does divide-and-conquer work?

- it should / could:
  - tremendous speed gain to analyzing many, smaller problems:

$$\text{time} \sum_{1}^{n} x \ll \text{time } nx$$

  - accuracy ~flat with respect to problem size
- but these potential savings aren’t realized in full empirically …

Analyses of full 4096-taxon data set

Method	Accuracy ($1 – d_S$)	Time taken (seconds)
NJ	0.857	193
MP	0.917	69 392
ML-DCM3	0.921	195 371
ML (“standard hill climbing”)	0.923	303 450

1.55x

- from Bininda-Emonds and Stamatakis (2006)

Analyses of full data set

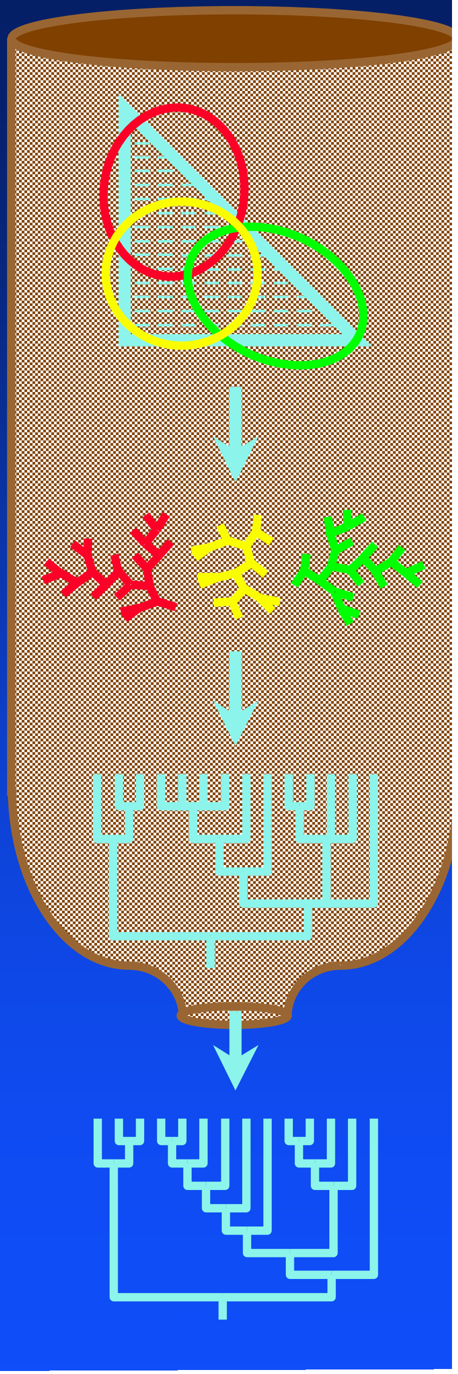
Method	Accuracy (1 – d_S)	Time taken (seconds)
NJ	0.857	193
MP	0.917	69 392
ML (“fast hill climbing”)	0.912	38 737
ML-DCM3	0.921	195 371
ML (“standard hill climbing”)	0.923	303 450

5.04x

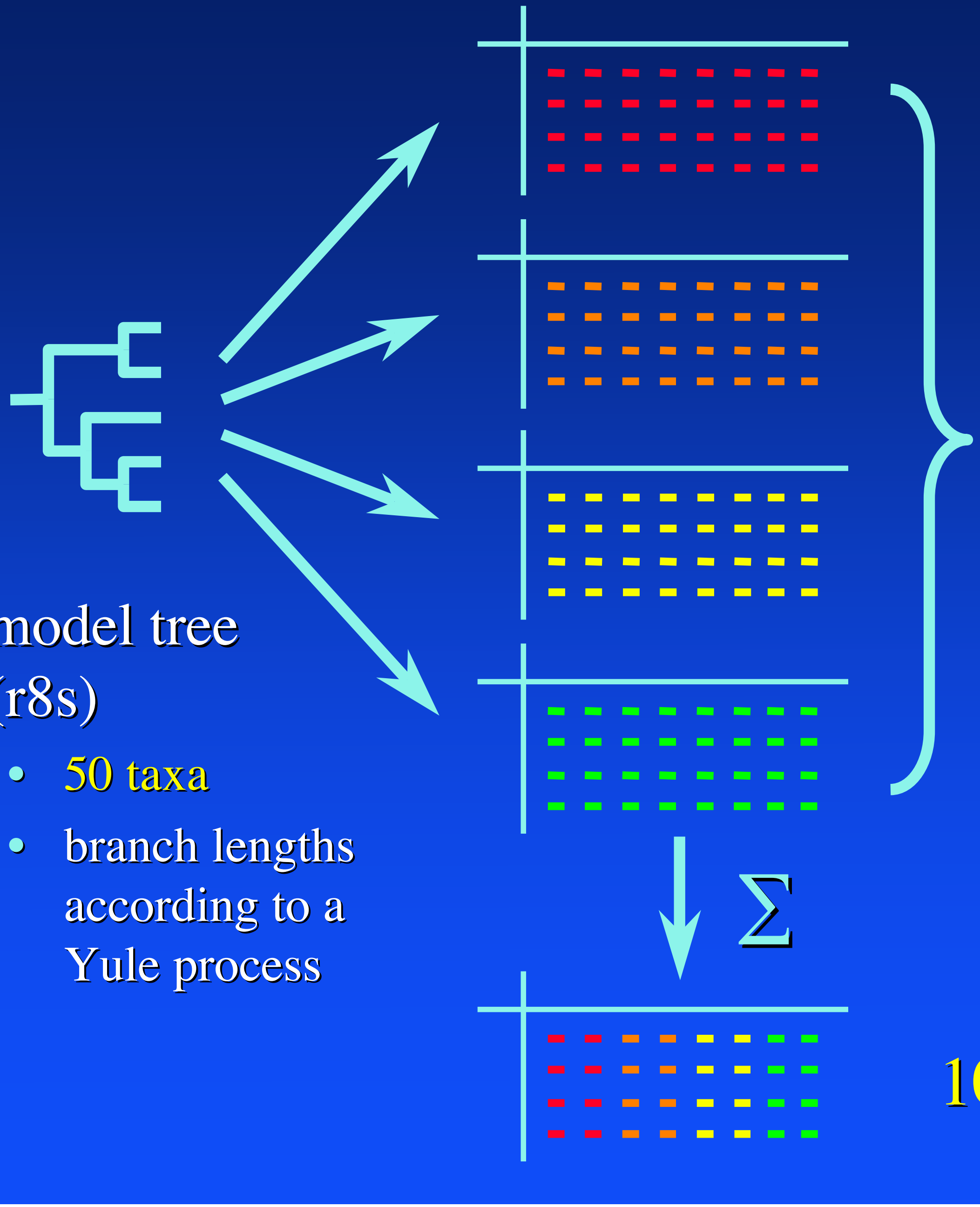
- from Bininda-Emonds and Stamatakis (2006)

What’s the problem?

- bottleneck remains terminal global optimization step
  - any excessive branch swapping will slow it down
  - but branching swapping crucial for accuracy
- therefore, key is to provide as accurate of a starting tree as possible
  - NB: accuracy ≠ resolution
  - could serve as a constraint tree (at least of well supported nodes)



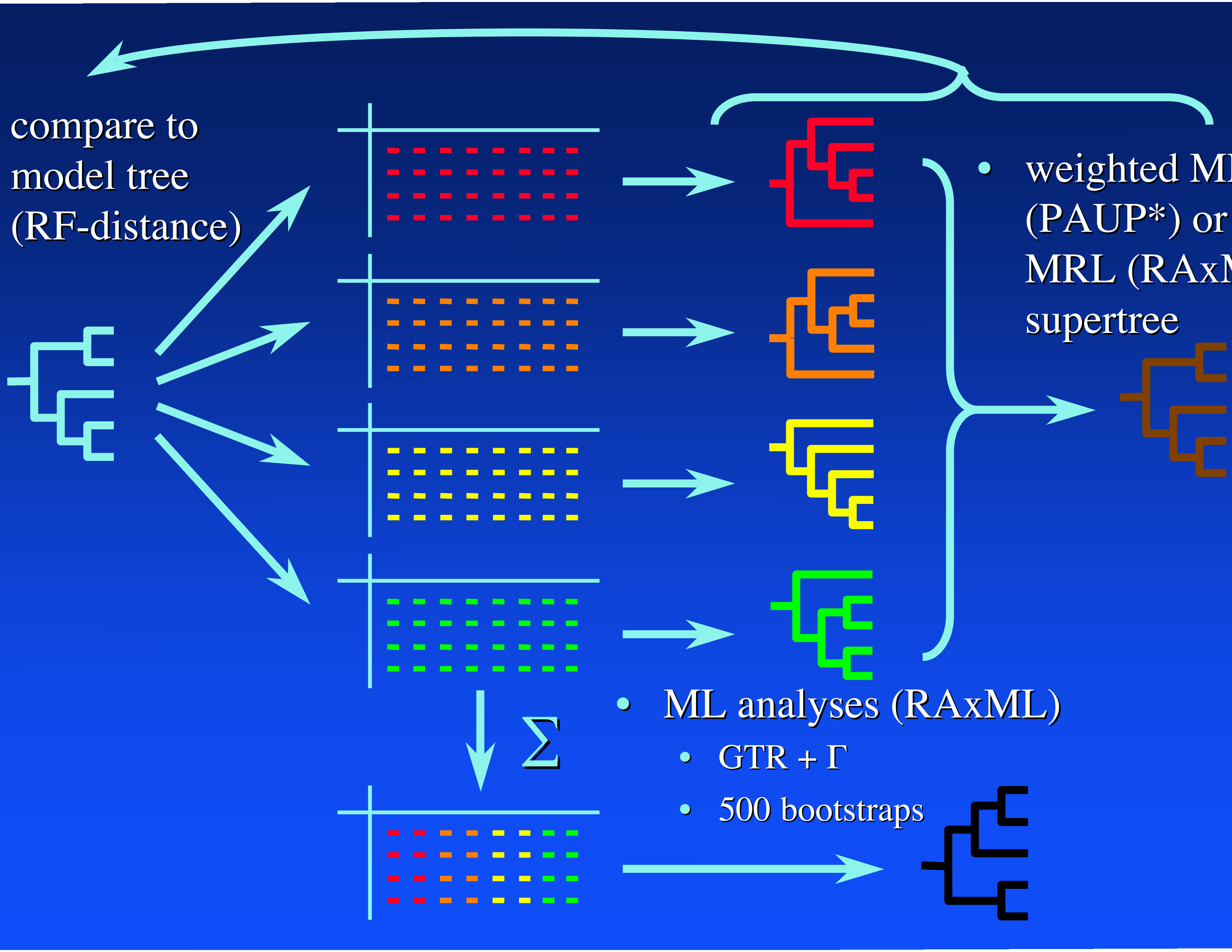
Increasing amount of characters



- model tree (r8s)
  - 50 taxa
  - branch lengths according to a Yule process

- simulate sequence data (seq-gen)
  - GTR with no Γ; parameters bounded, but set randomly between partitions
  - 500 bp per partition
  - μ = {0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1.0, 2.5, 5.0, 10,0}

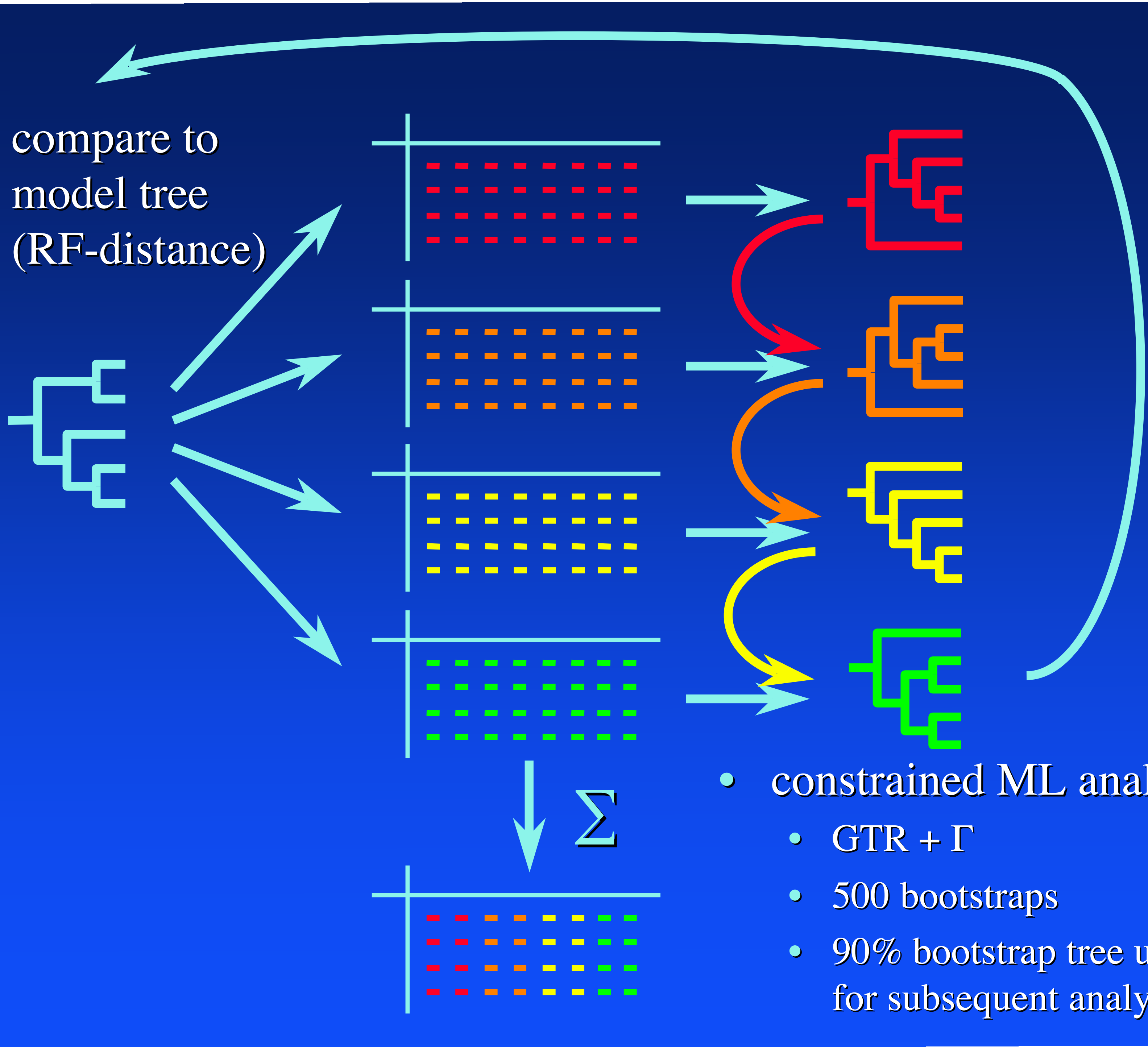
1000-fold range in μ



- compare to model tree (RF-distance)

- weighted MRP (PAUP*) or MRL (RAxML) supertree

- ML analyses (RAxML)
  - GTR + Γ
  - 500 bootstraps



- compare to model tree (RF-distance)

- constrained ML analyses (RAxML)
  - GTR + Γ
  - 500 bootstraps
  - 90% bootstrap tree used as constraint for subsequent analysis

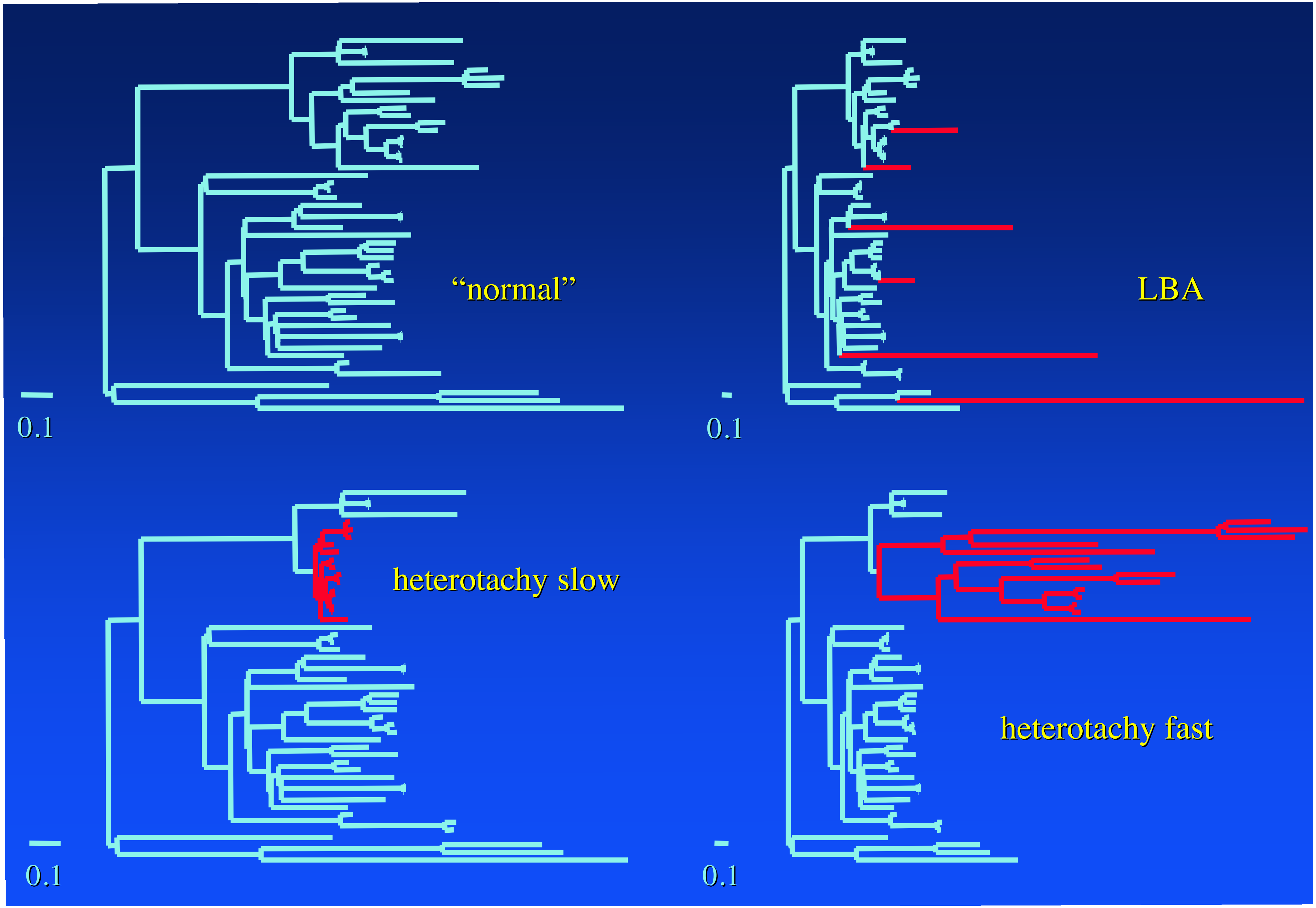
Variations

More rate heterogeneity

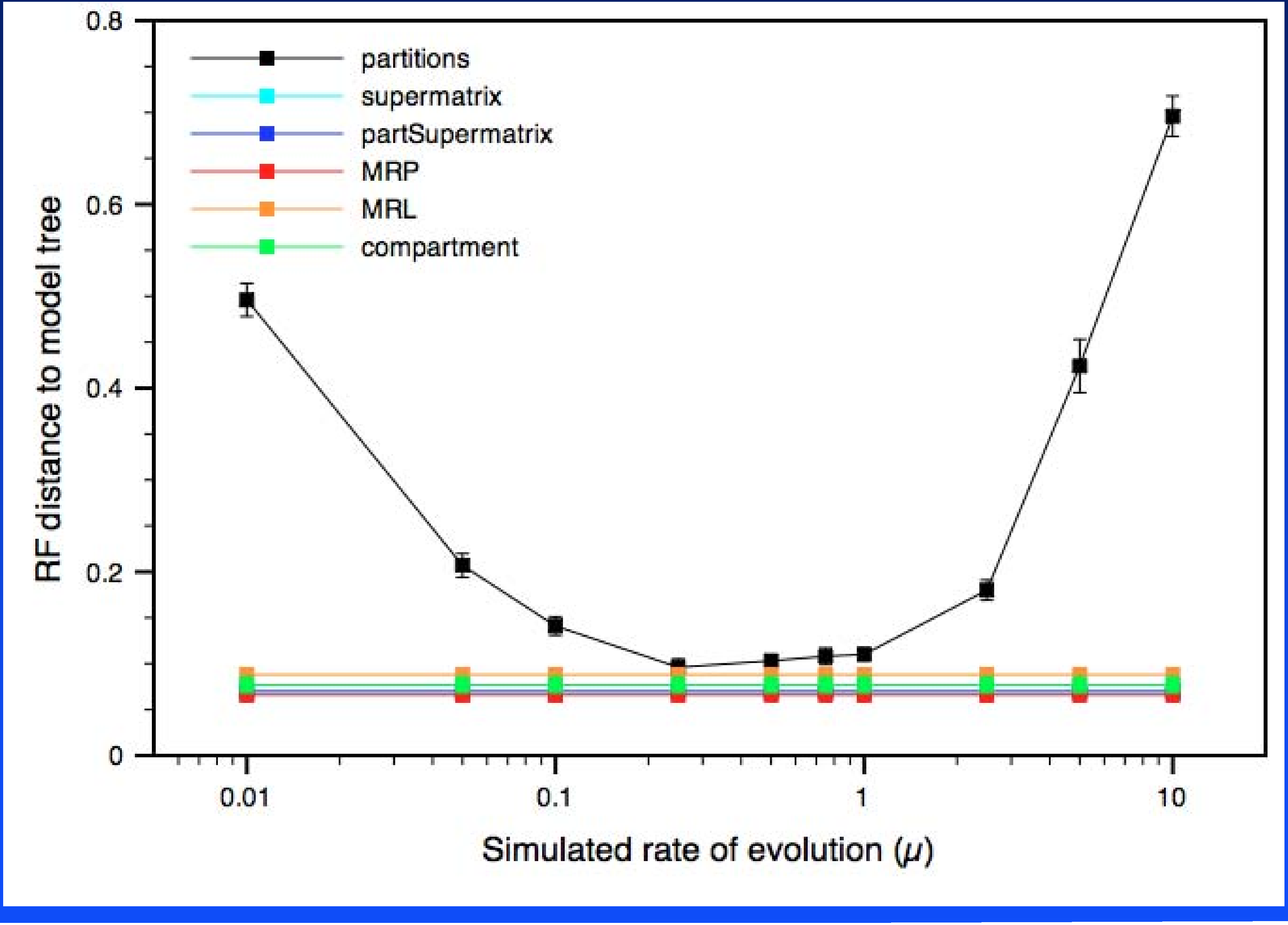
- branch lengths on model tree altered to simulate:
  - LBA → terminal branch lengths of five taxa increased by 10x
  - heterotachy → all branch lengths within one clade of 10^+ taxa increased / decreased by 5x

Taxon sampling

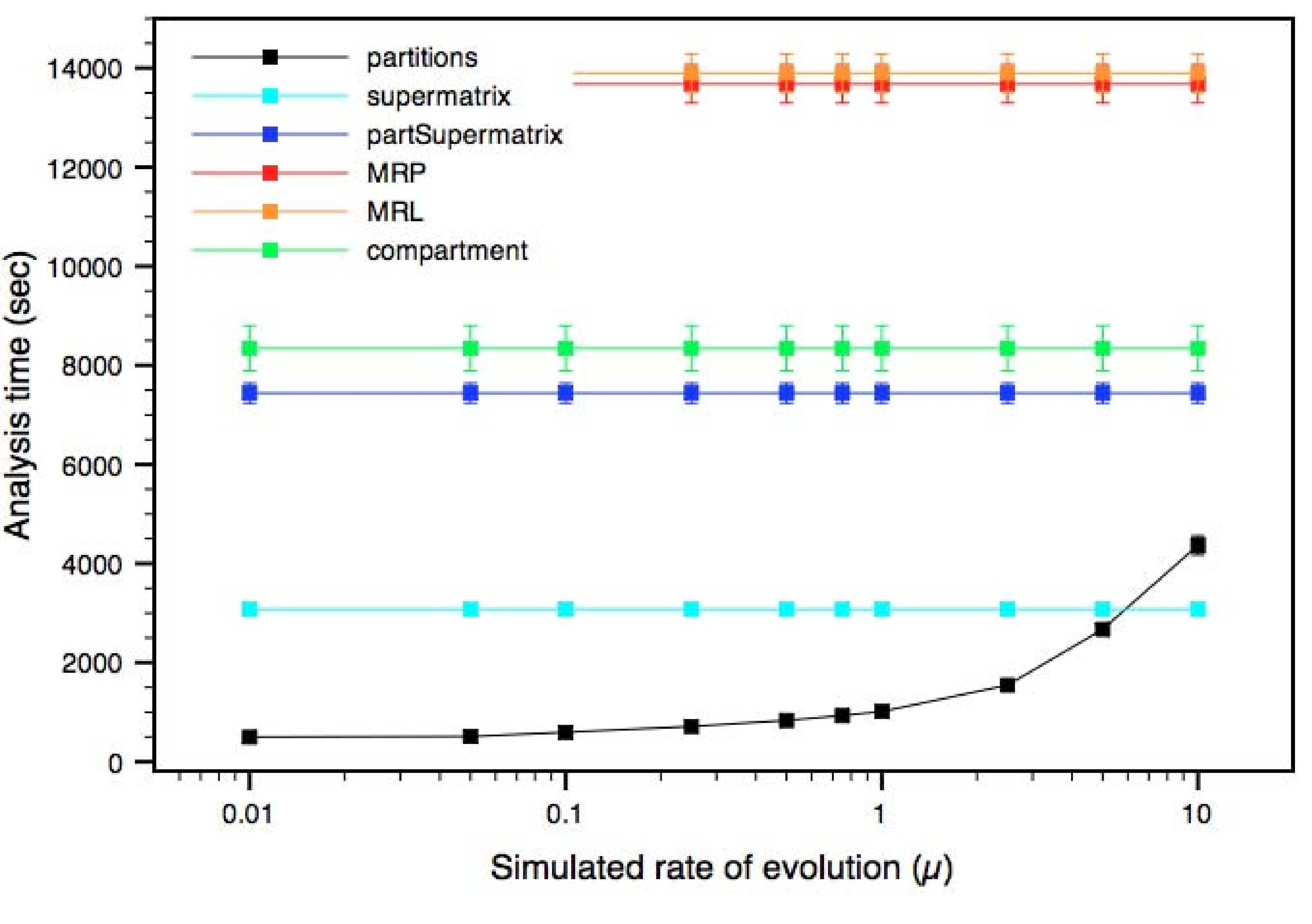
- taxa deleted from 40% of partitions chosen at random


"normal"
0.1
LBA
0.1
heterotachy slow
0.1
heterotachy fast
0.1

“normal” model tree; no taxon deletion



“normal” model tree; no taxon deletion



Taxon deletion

Normal model tree

- MRP > supermatrix > partSupermatrix > compartment > MRL
- (all methods very good ($d_S \leq 0.088$) and better than analysis of any single partition)

Normal model tree with taxon deletion

- partSupermatrix: +2.8%
- supermatrix: –18.1%
- MRP: –43.7%
- MRL: –44.5%
- compartment: –97.5% (0.151)
- slower partitions: + change
- faster partitions: – change

Variations

LBA

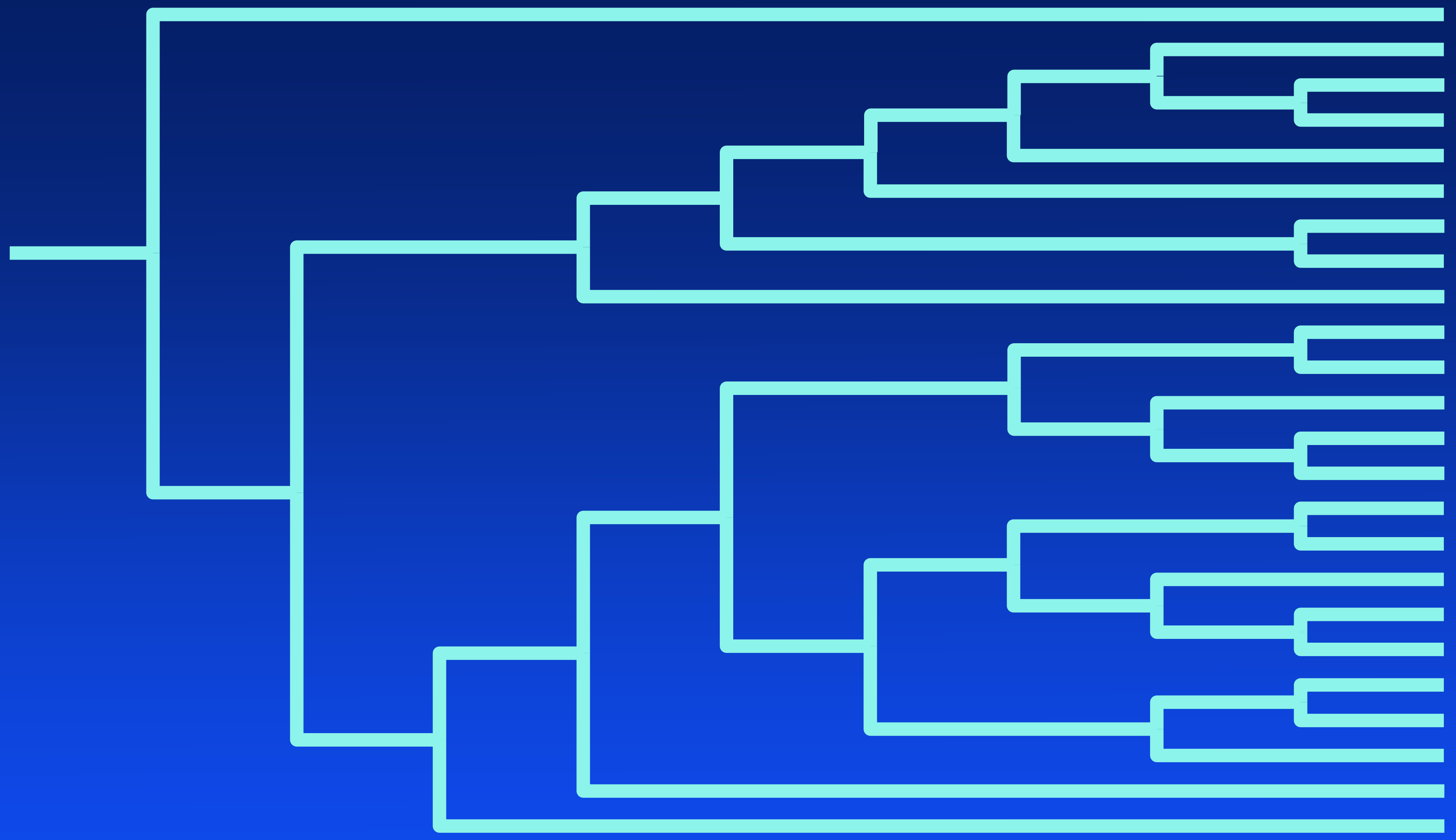
- always increased analysis times
- decreased accuracy, especially for non-supermatrix methods

Heterotachy slow

- all methods relatively static (accuracy and analysis times)

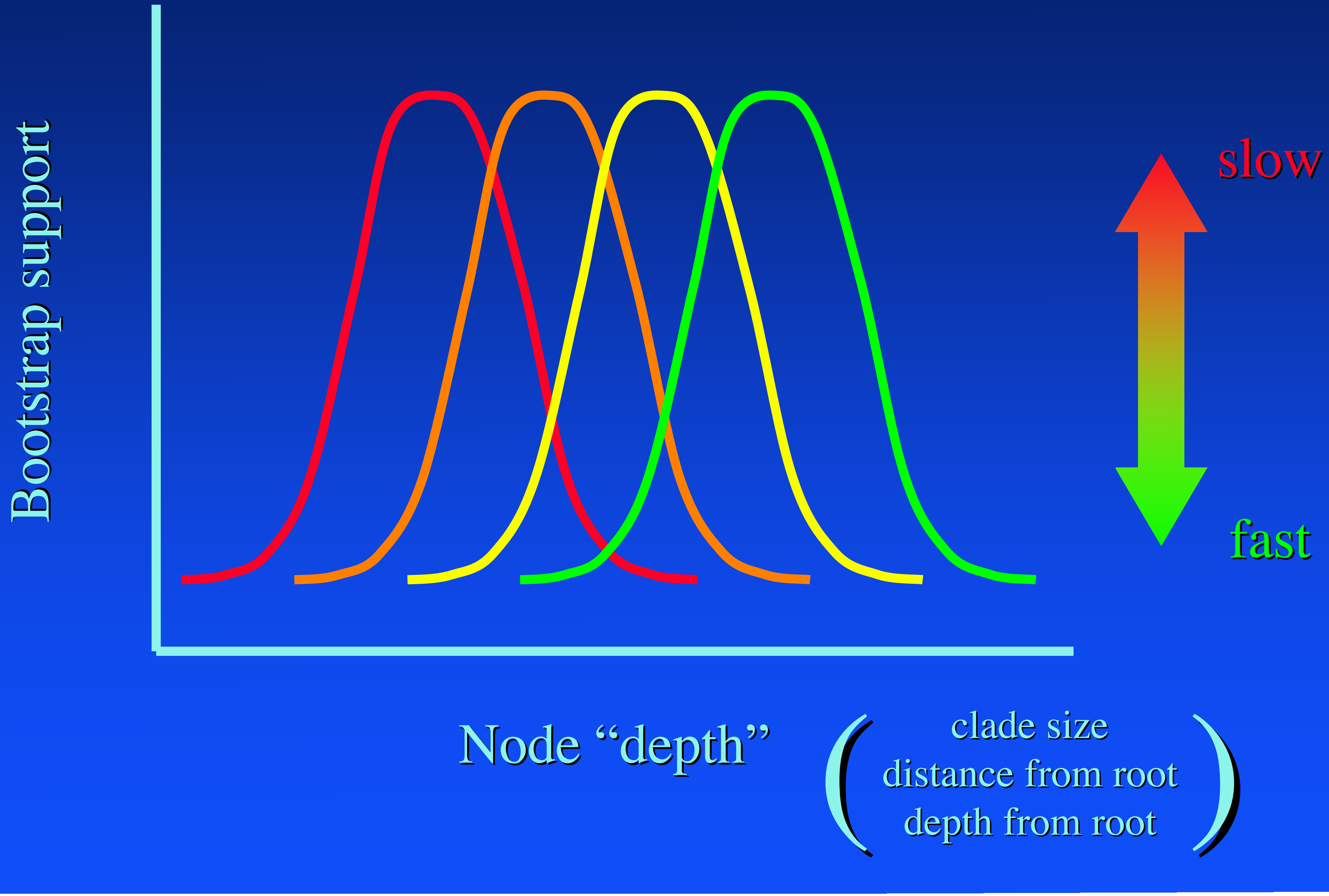
Heterotachy fast

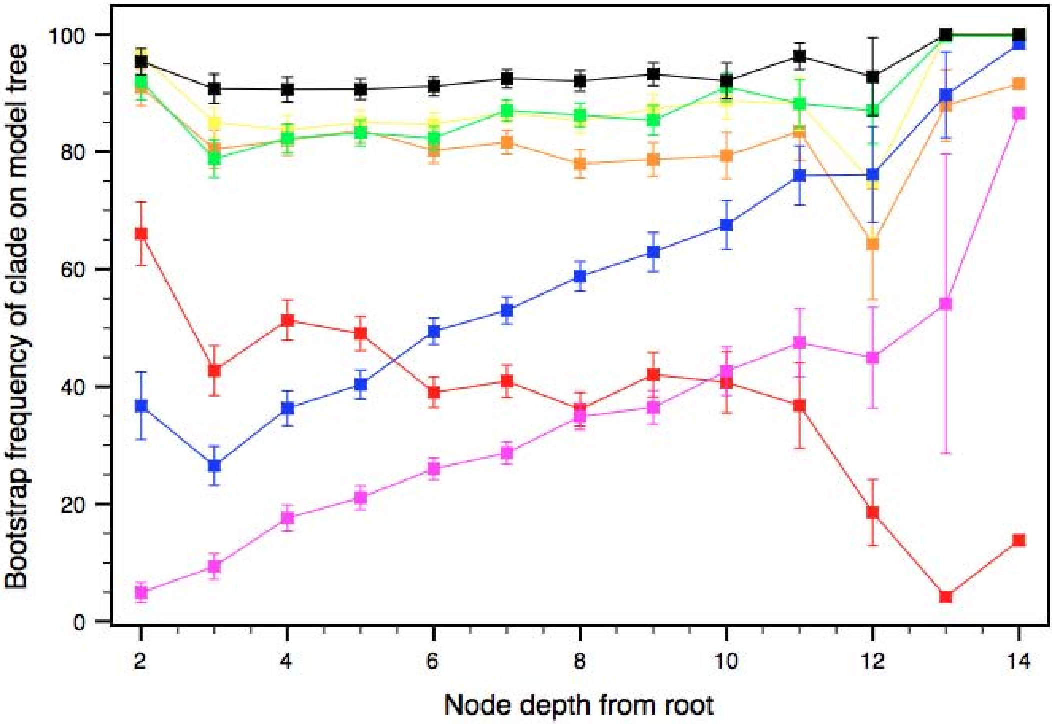
- always increased analysis times
- decreased accuracy of most methods, but especially so with taxon deletion

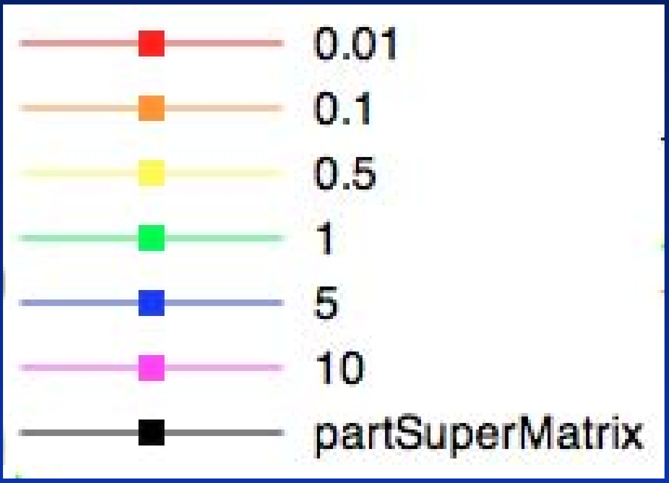


slow genes / sites

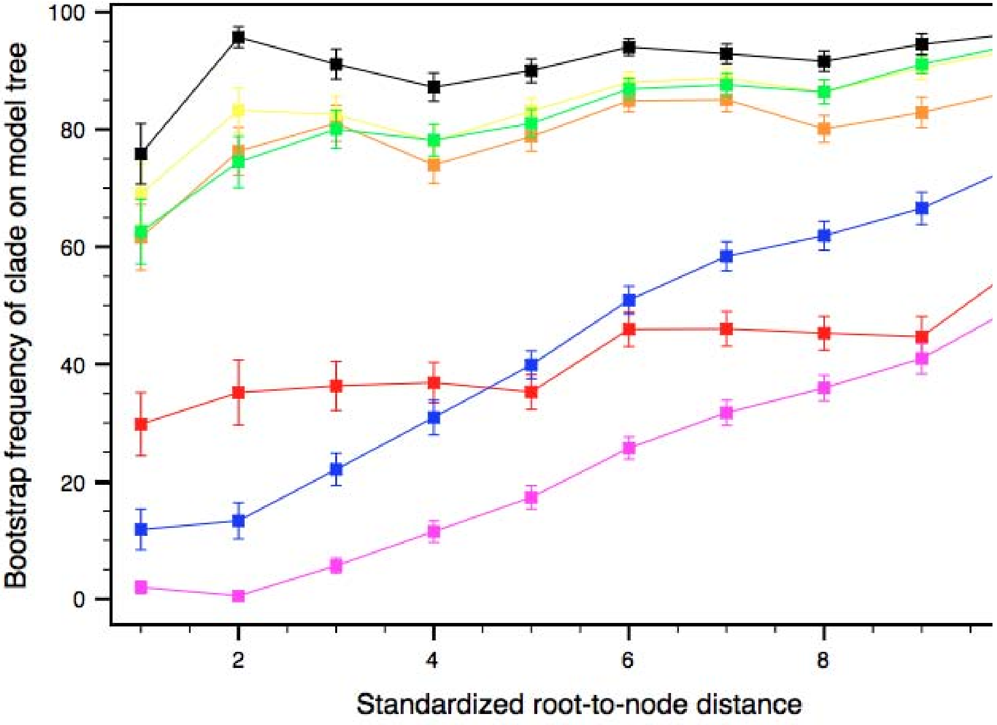
fast genes / sites

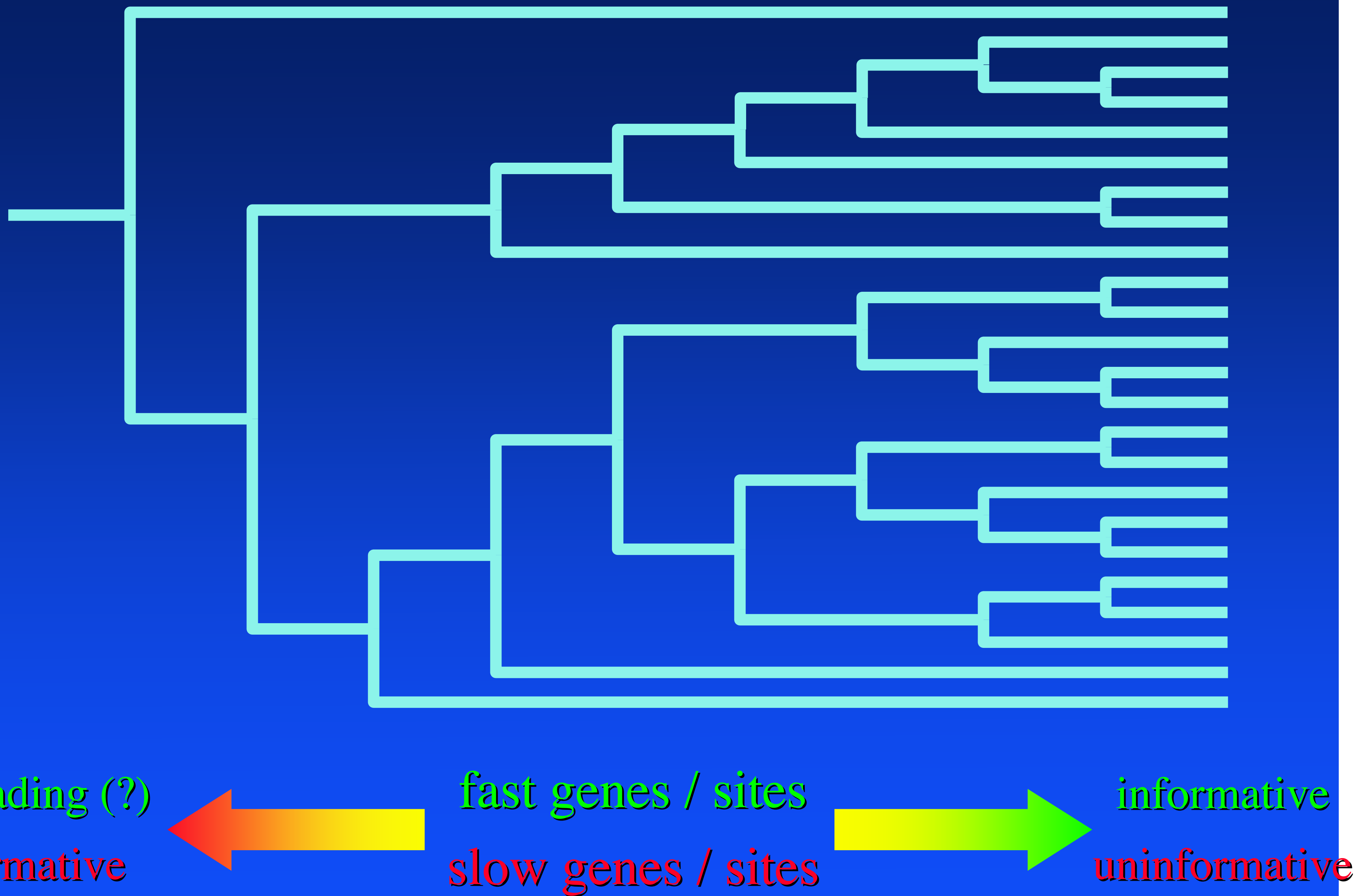

Bootstrap support
slow
fast
Node “depth”
clade size
distance from root
depth from root





“normal” model tree;
no taxon deletion





misleading (?)

informative

fast genes / sites

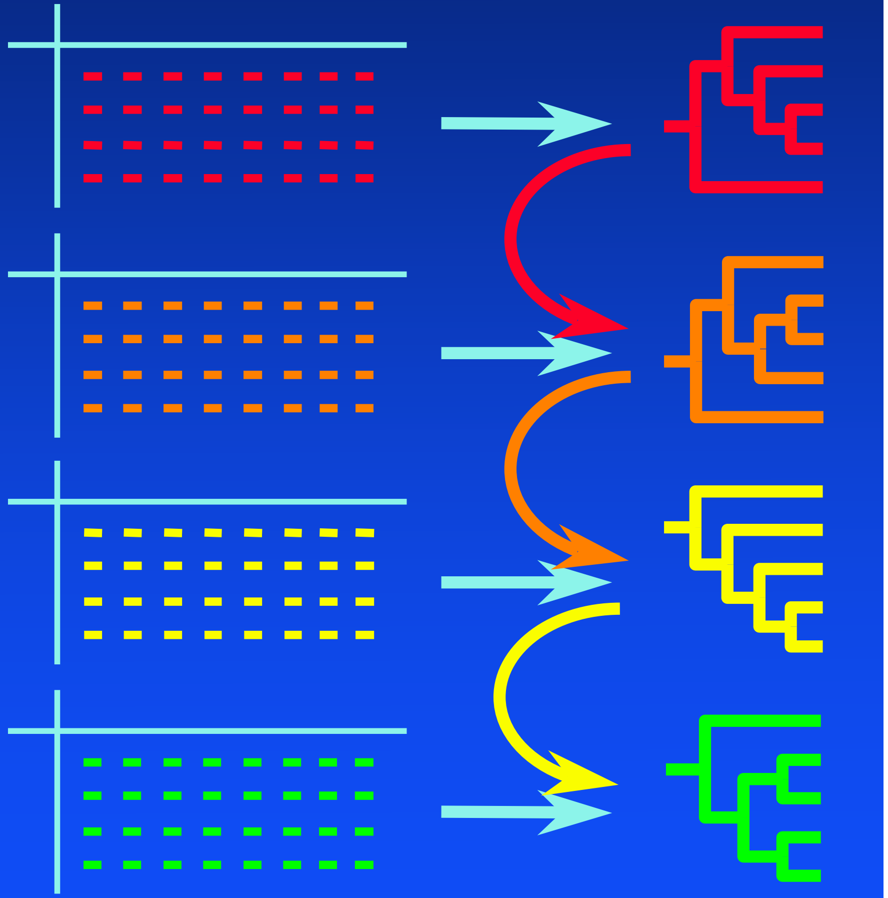
slow genes / sites

informative

uninformative

Compartment analyses

- slow ≠ old → slow = rare
- shouldn’t affect analyses in principle
- hard constraints too hard?
  - “constraint trees” more suited as Bayesian priors???



Conclusions – large character problems

- supermatrix methods gave most stable results (d_S usually < 0.10)
  - unpartitioned analyses on a par with partitioned ones …
  - … and definitely faster
- remaining methods could outperform supermatrix ones, but more variable
  - but d_S always still < 0.17
  - MRP always better than MRL

Conclusions – large character problems

- missing data
  - decrease accuracy of all "global" methods, but increase that of slowest rate partitions
  - improve supertree analysis times, but decreases that of all other global methods
- rate changes
  - rate slowdowns neutral WRT accuracy and analysis times
  - rate speedups always increase analysis times and tend to decrease accuracy in combination with missing data; lba more problematic than fast heterotachy

Take home message: bigger is better!

- is noise often random and not misleading???
- best results when data sets as complete as possible:
  - all species within focal clade
  - as few missing data as possible

- yields both increased accuracy and decreased running times
  - with no evidence of computational constraints

With thanks to …

Alexis Stamatakis

Nikos Alachiotis

(Heidelberg Institute for Theoretical Studies)