Fast and Accurate Methods for Phylogenomic Analyses

Jimmy Yang and Tandy Warnow

University of Texas at Austin

Oct 8, 2011

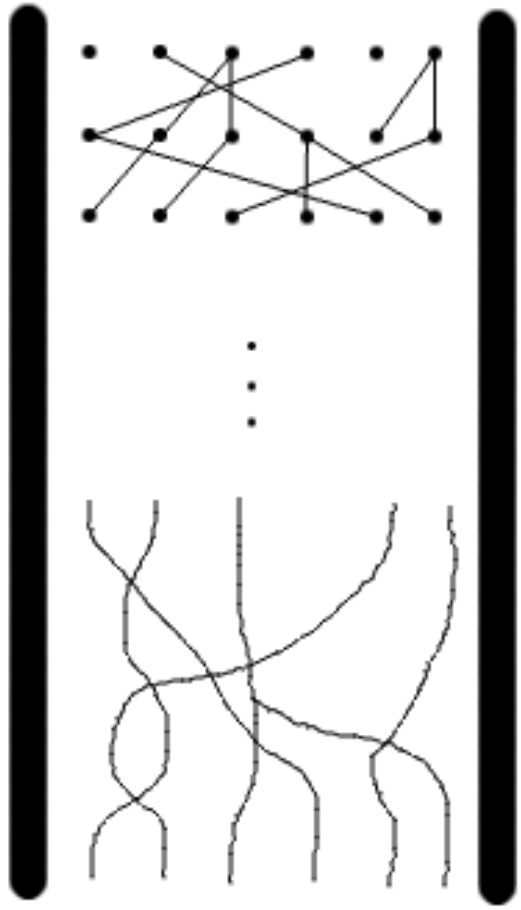
Phylogenomic Analyses

- Input: set of estimated gene alignments and/or trees
- Output: species tree

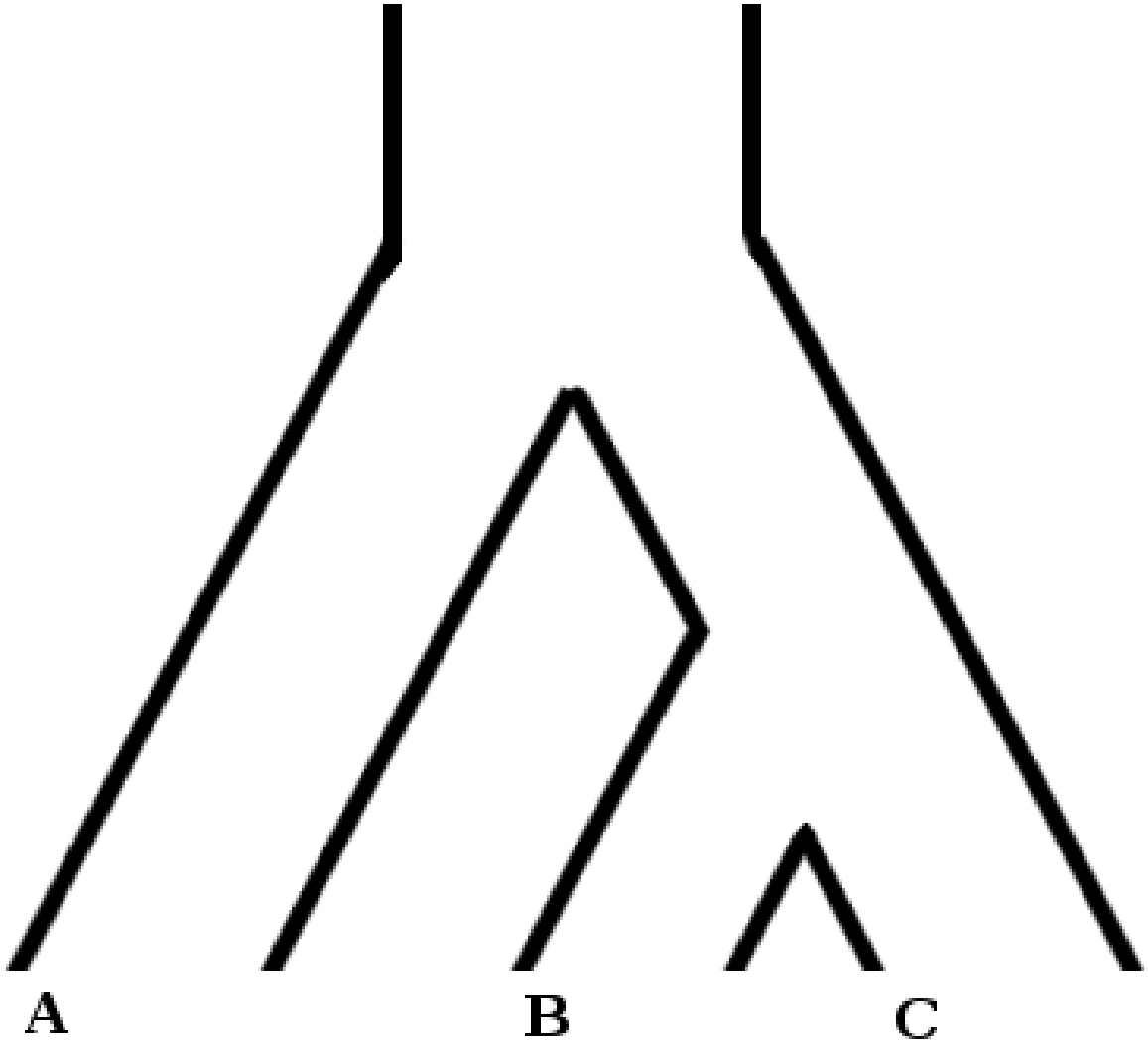
Species Trees / Gene Trees Discordance

- Gene trees differ from species trees in biological data
- Incomplete lineage sorting (ILS) commonly studied under the coalescent model
- Other causes: gene duplications and losses, horizontal gene transfer (HGT), hybridization, recombination, etc.

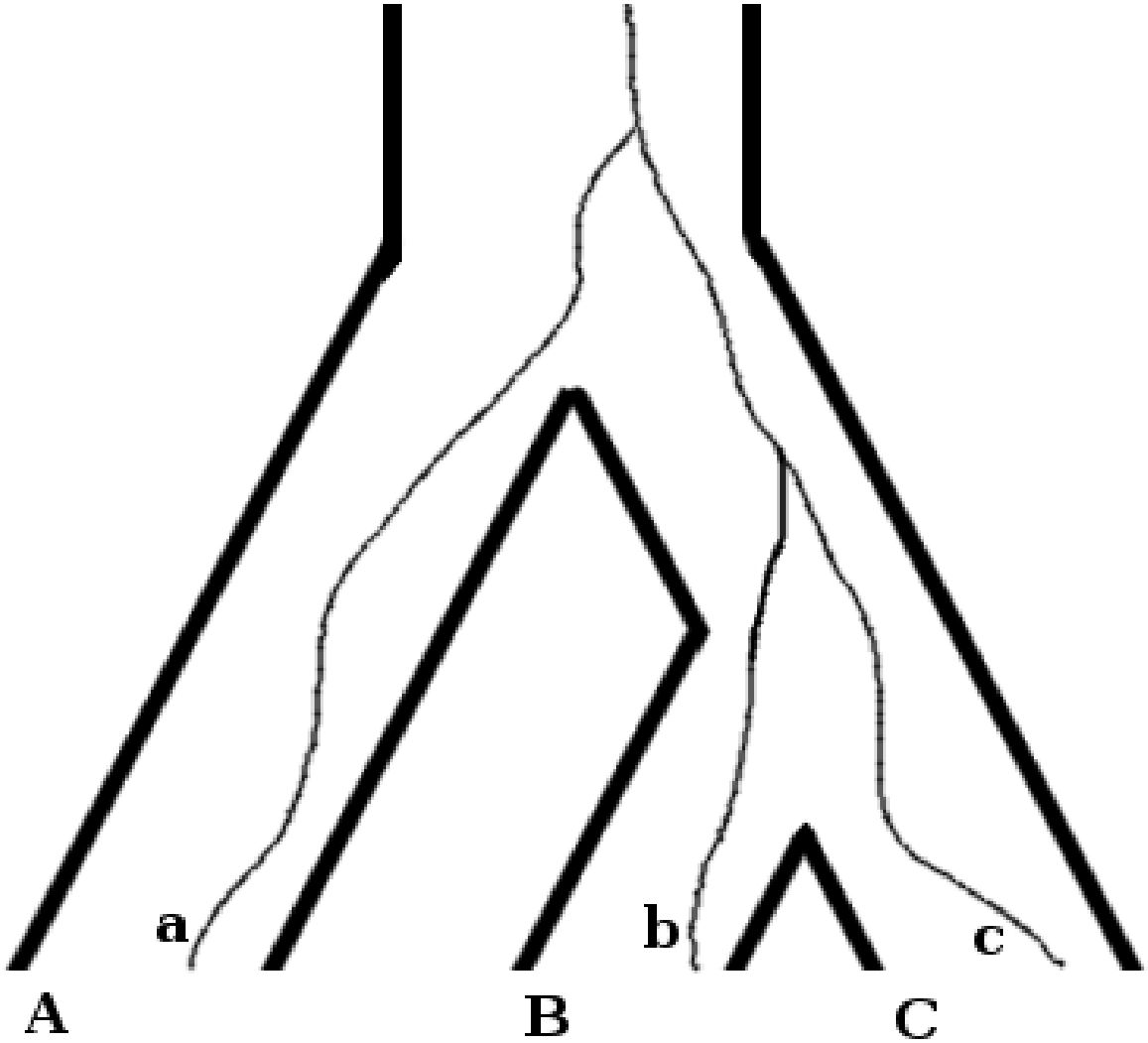
Coalescent Model



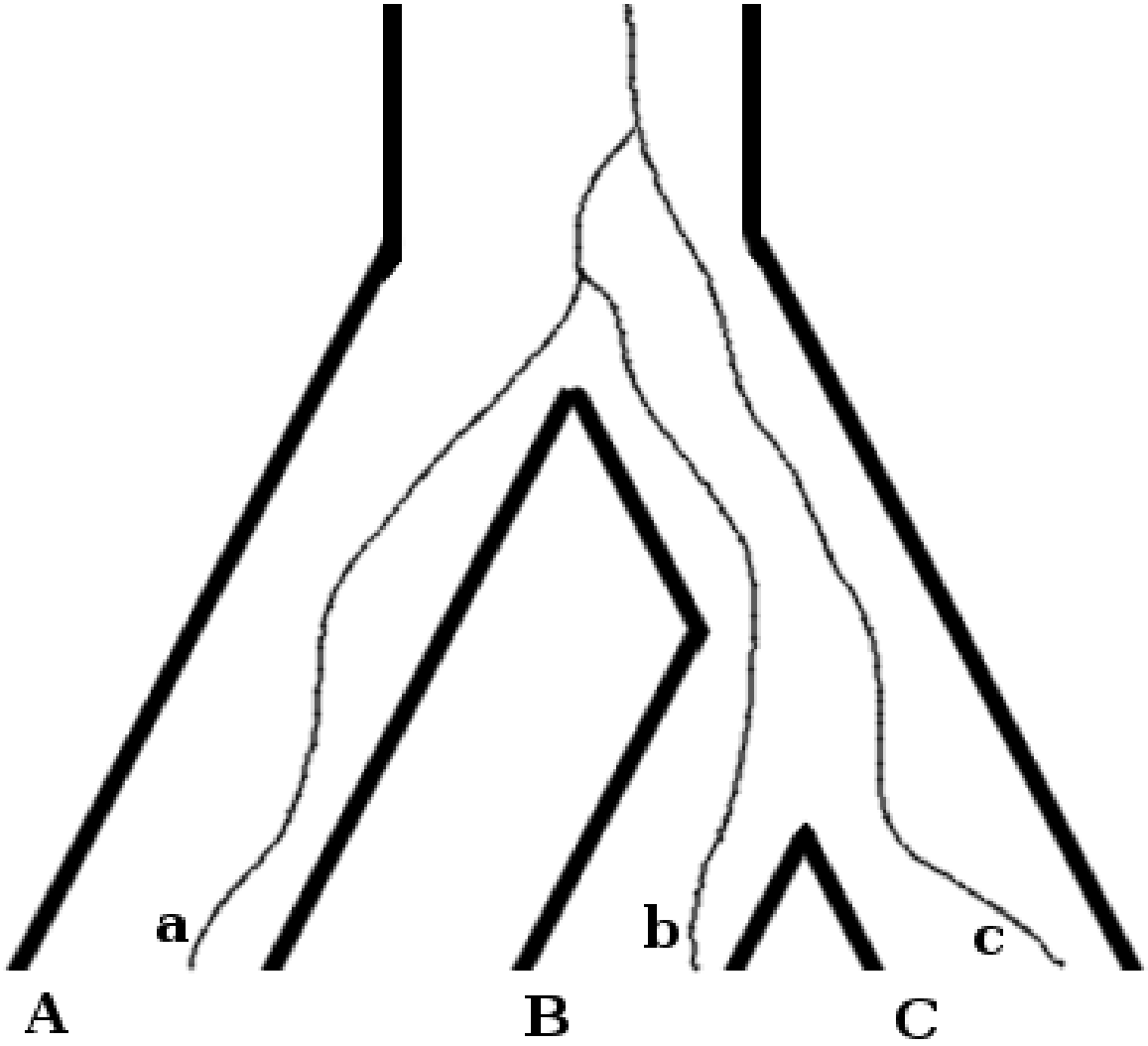
Multispecies Coalescent Model



Multispecies Coalescent Model



Multispecies Coalescent Model



Questions

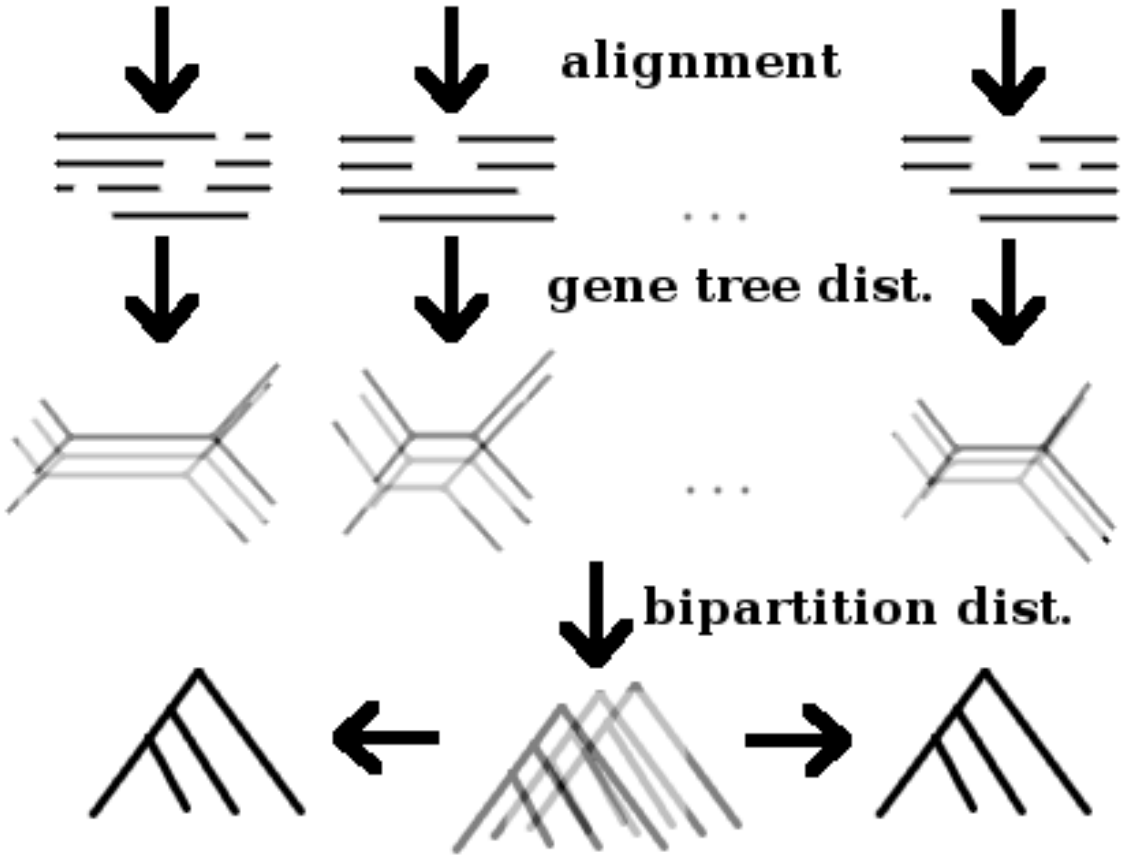
- Which methods produce the most accurate species trees? How do these methods scale (in terms of computational requirements) with the number of taxa?
- Can we improve species tree estimations by considering gene tree estimation error? For example, as in Yu, Warnow, Nakhleh (RECOMB 2011), by contracting low support edges in estimated gene trees, or as in BUCKy (Ané et al.), by using gene tree distributions?
- Are there fast methods with accuracy competitive with the most promising statistical methods (e.g., BUCKy, *BEAST, BEST)?

BUCKy

Ané et al., MBE 2007, and Larget et al., Bioinformatics 2010.

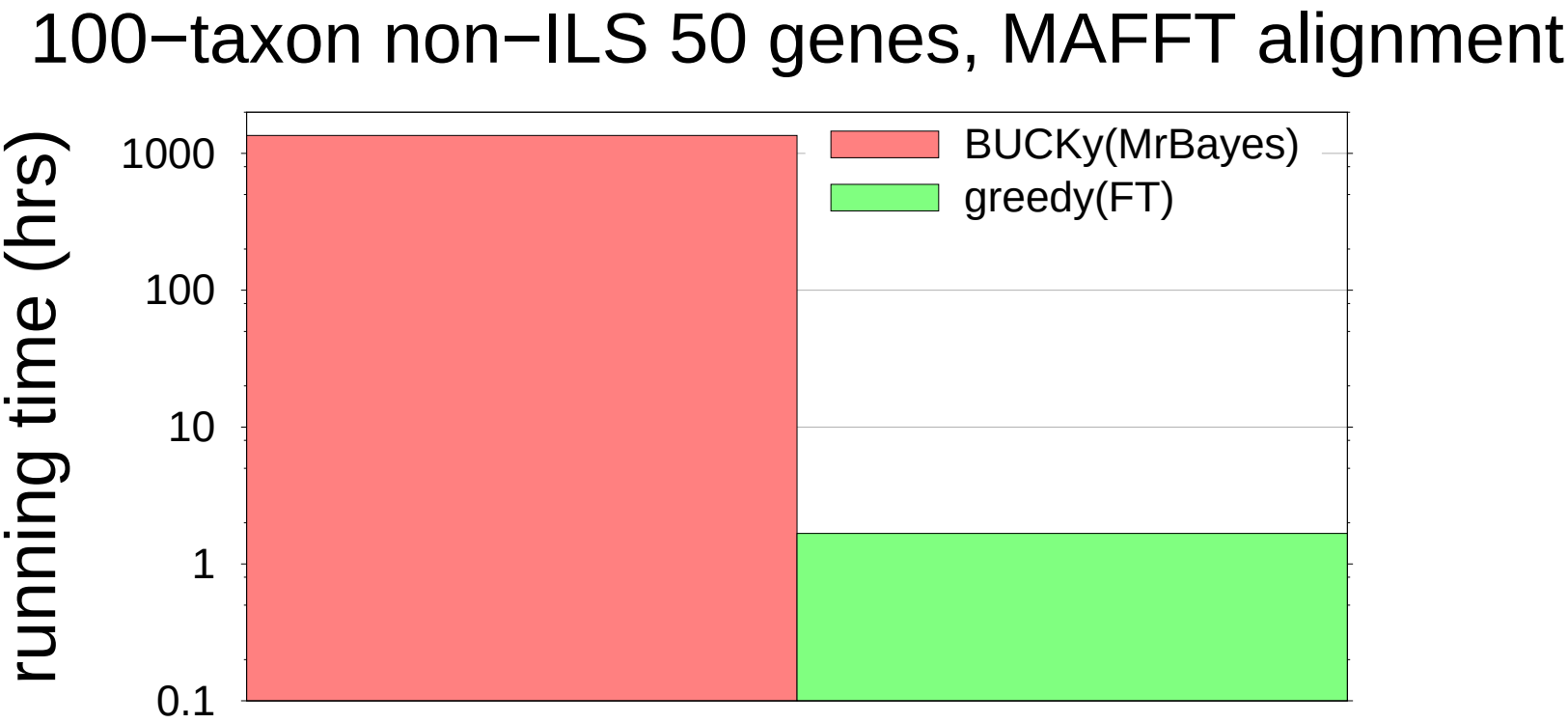
- BUCKy-pop/con, takes gene tree distributions as input, uses *concordance factors* on quartets to compute the population tree and concordance factors on clades to compute the concordance tree.
- BUCKy-pop is statistically consistent under ILS.
- BUCKy-con is not statistically consistent under ILS.

BUCKy(MrBayes) Analysis



1. MAFFT
2. MrBayes
3. BUCKy (concordance and population tree)

BUCKy(MrBayes) vs. Greedy



Memory usage:

- BUCKy: 34-234 GB
- greedy: $<$ 9 MB

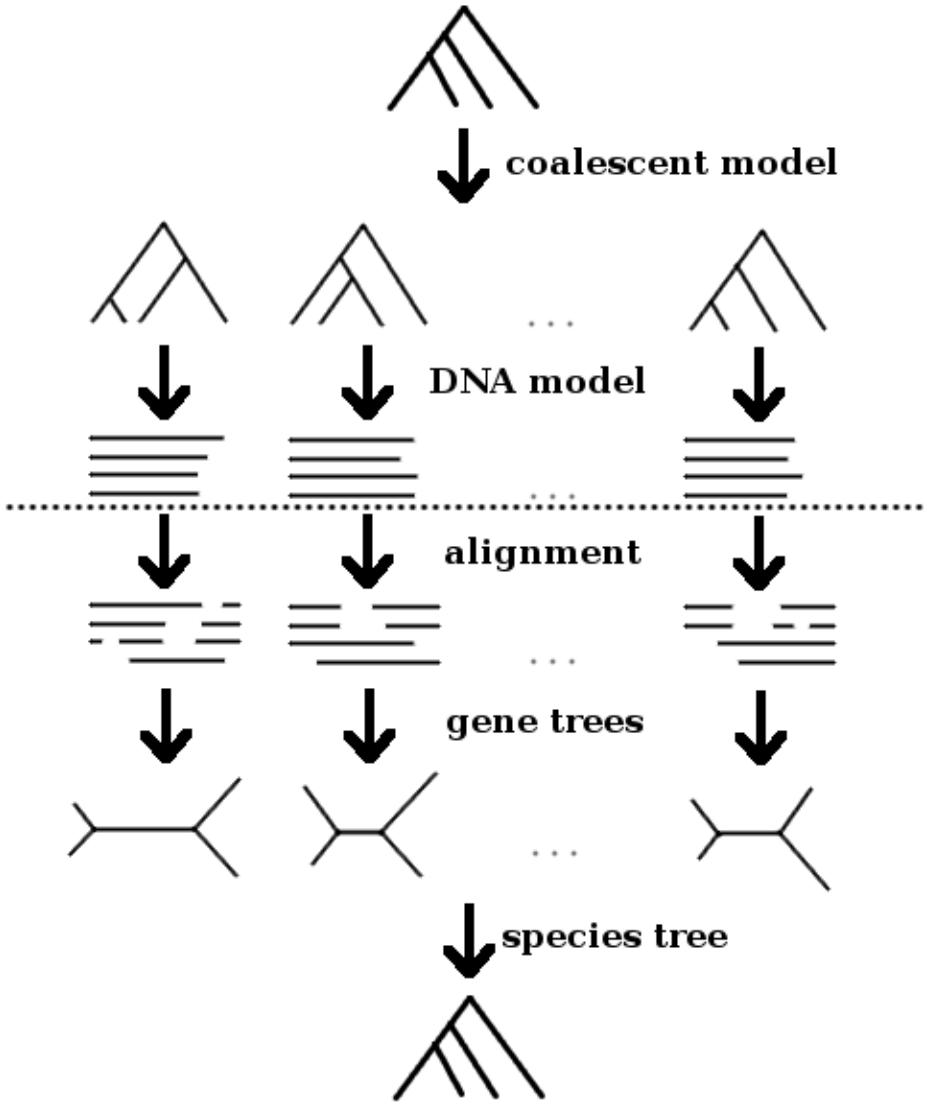
Using MrBayes to estimate gene tree distributions

- Computational issues:
  - Long running times
  - Convergence to stationarity
  - Large numbers of sampled gene trees makes BUCKy slow and memory-intensive
- Alternatives to "proper" MrBayes analysis
  - non-converged distributions
  - sparse MrBayes samples
  - replacing MrBayes with other methods (e.g., bootstrap trees using RAxML)

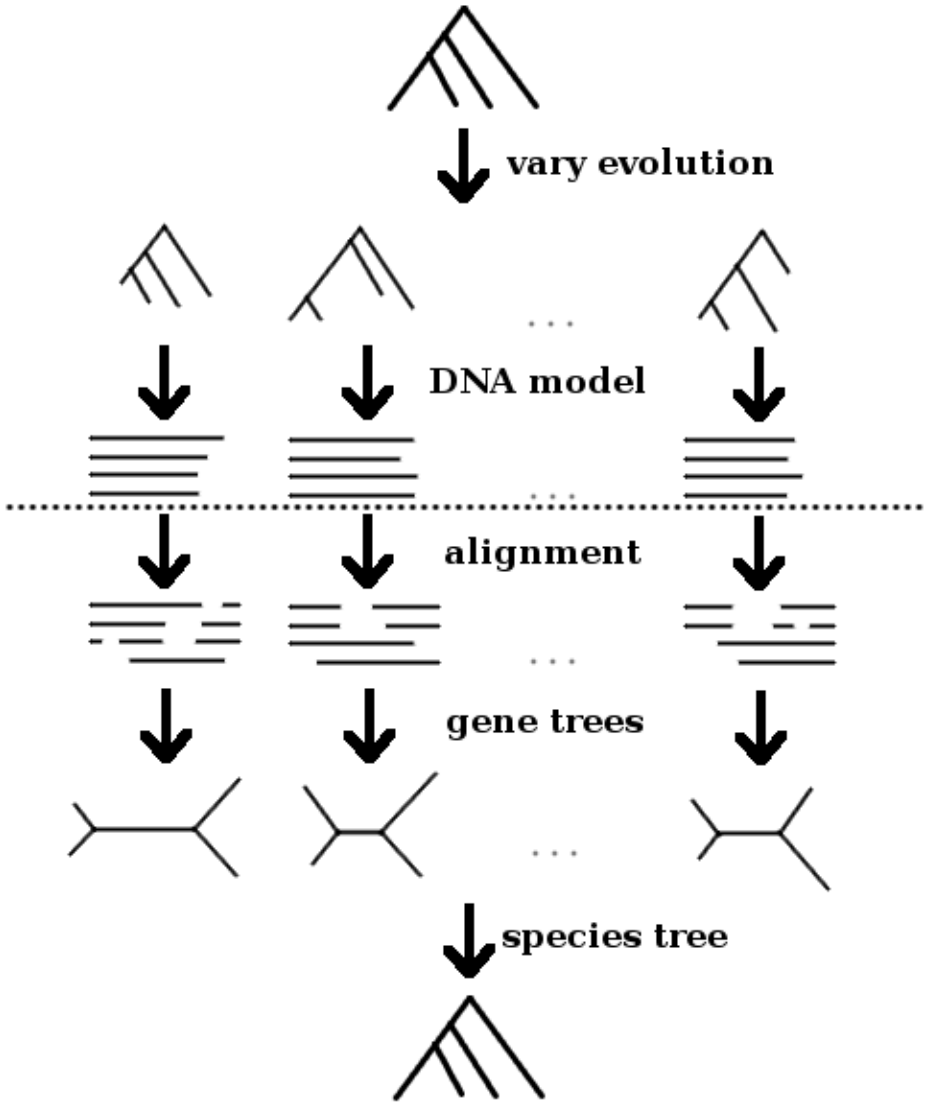
Other methods

- GLASS, distance-based (statistically consistent)
- Phylonet and iGTP for MDC
- iGTP for duplication and duplication/loss
- Greedy consensus

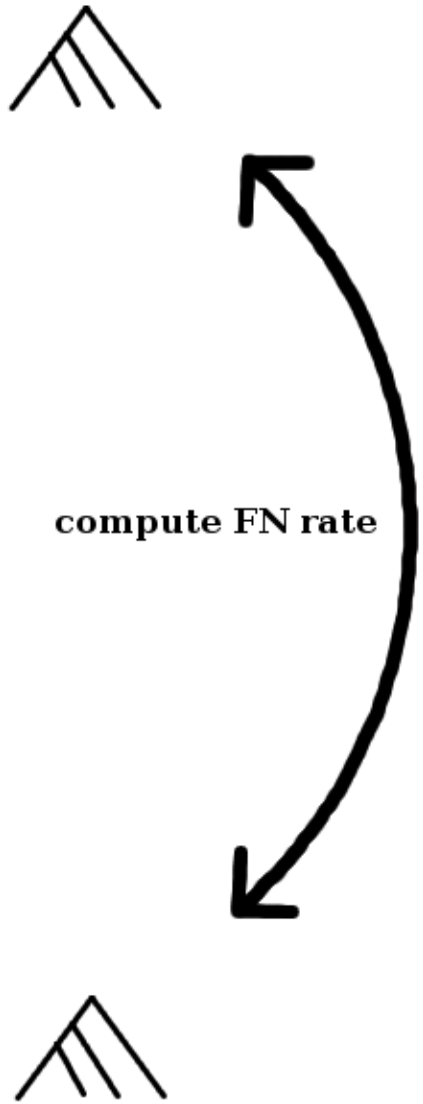
Simulation Study



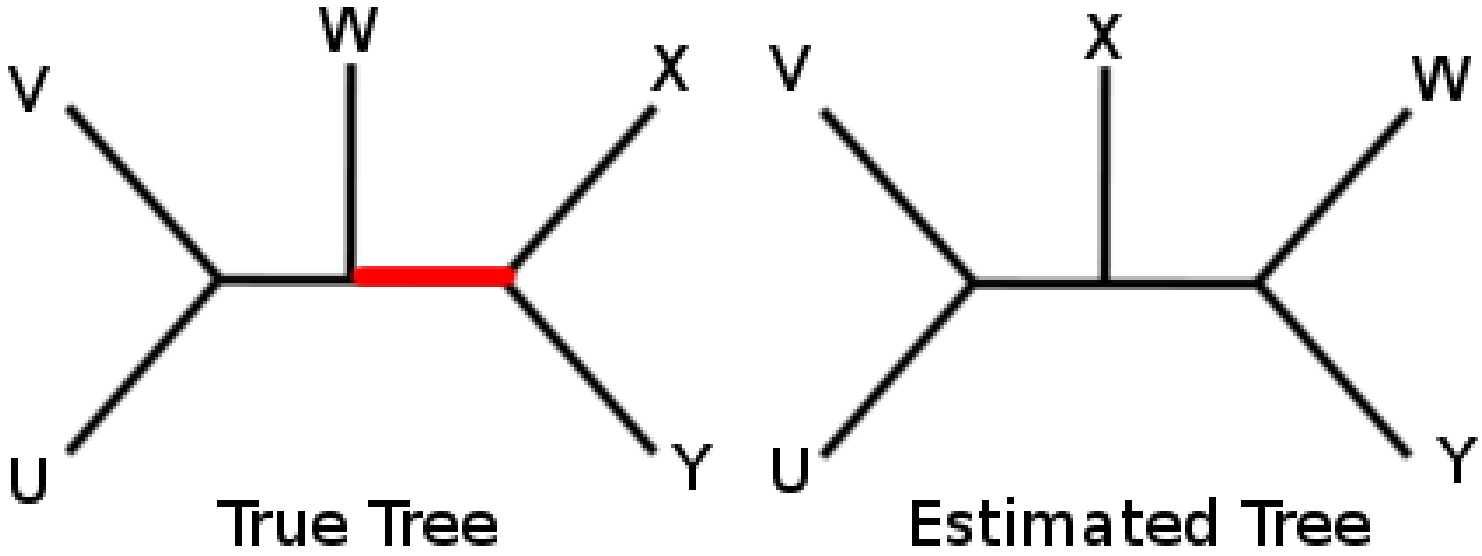
Simulation Study



Simulation Study

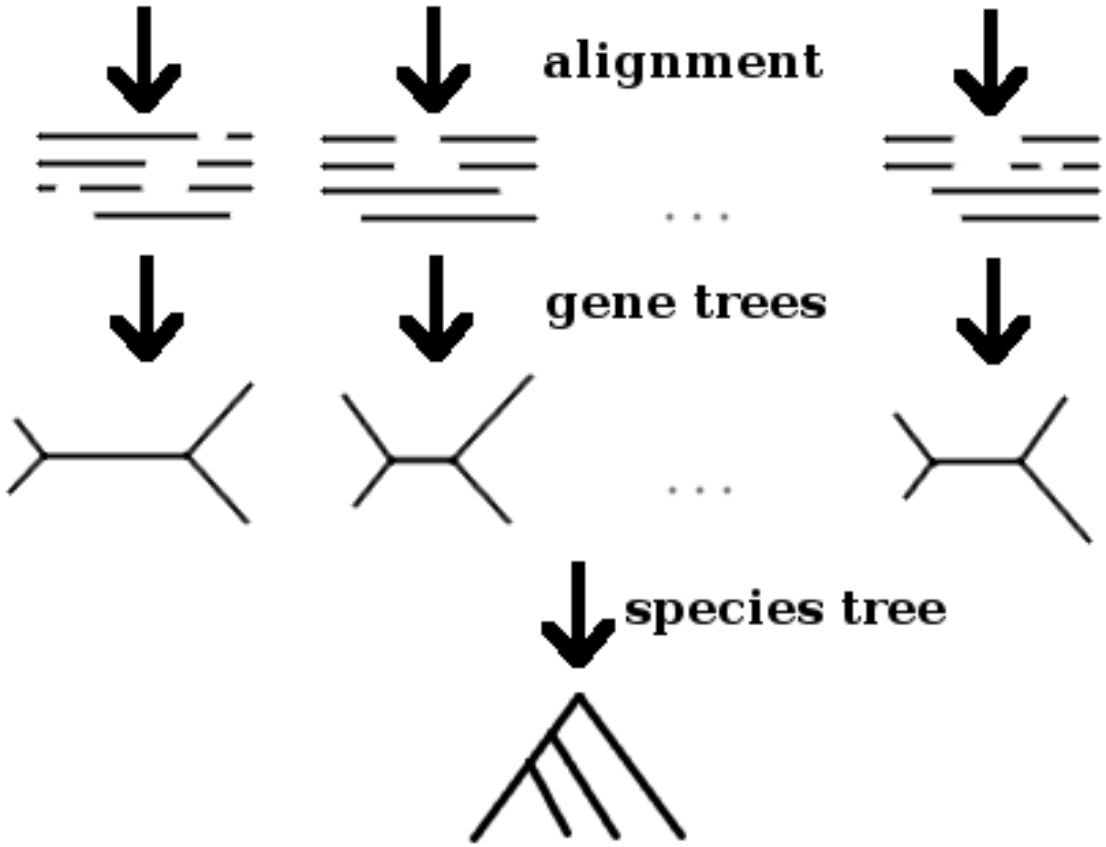


Comparing Trees



- False Negative: edge in the true tree missing from the estimated tree
- FN rate (missing branch rate): 50%

Methods

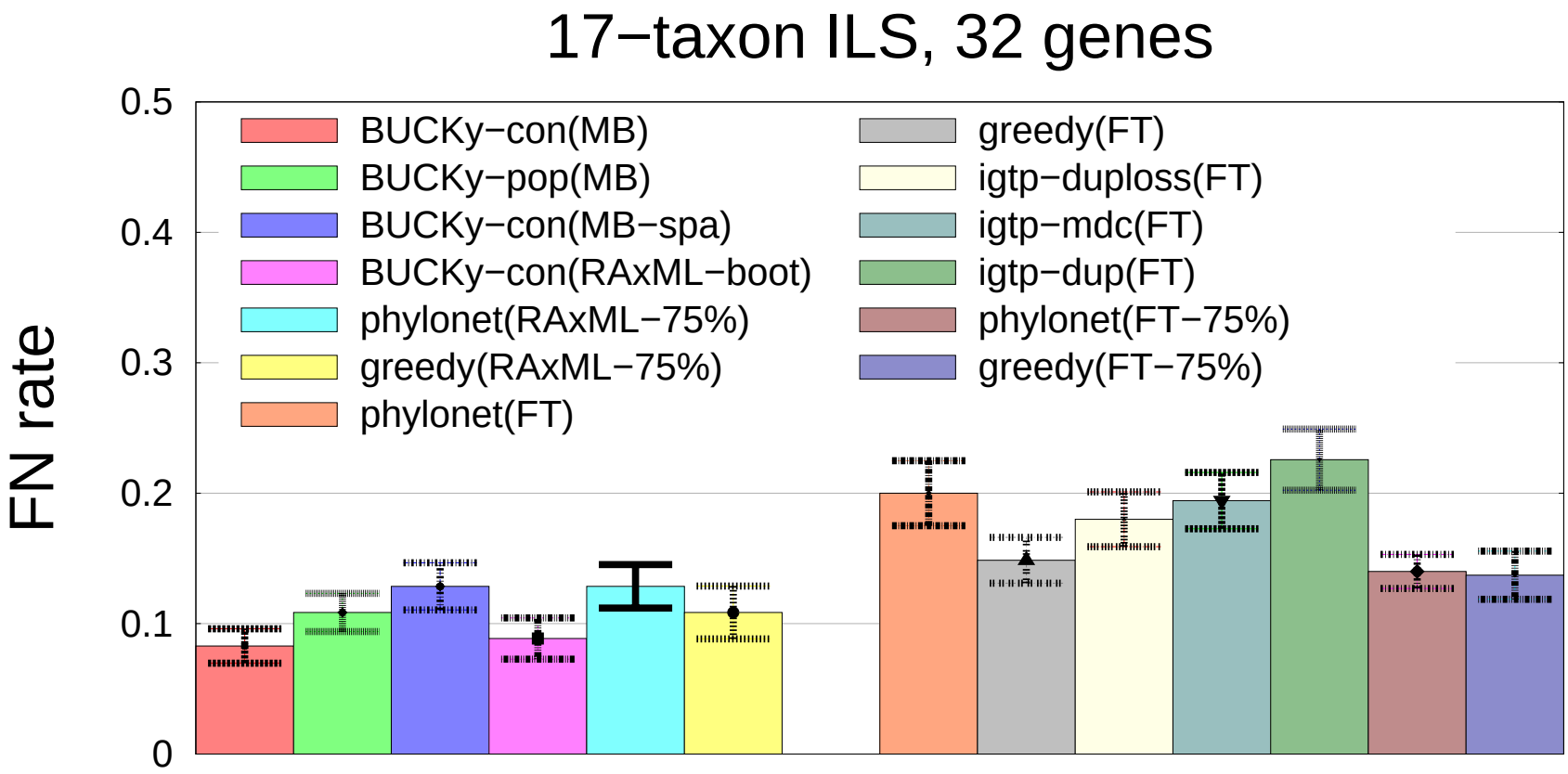


1. MAFFT
2. RAxML, FastTree, MrBayes
3. BUCKy, PhyloNet, iGTP, greedy consensus, GLASS

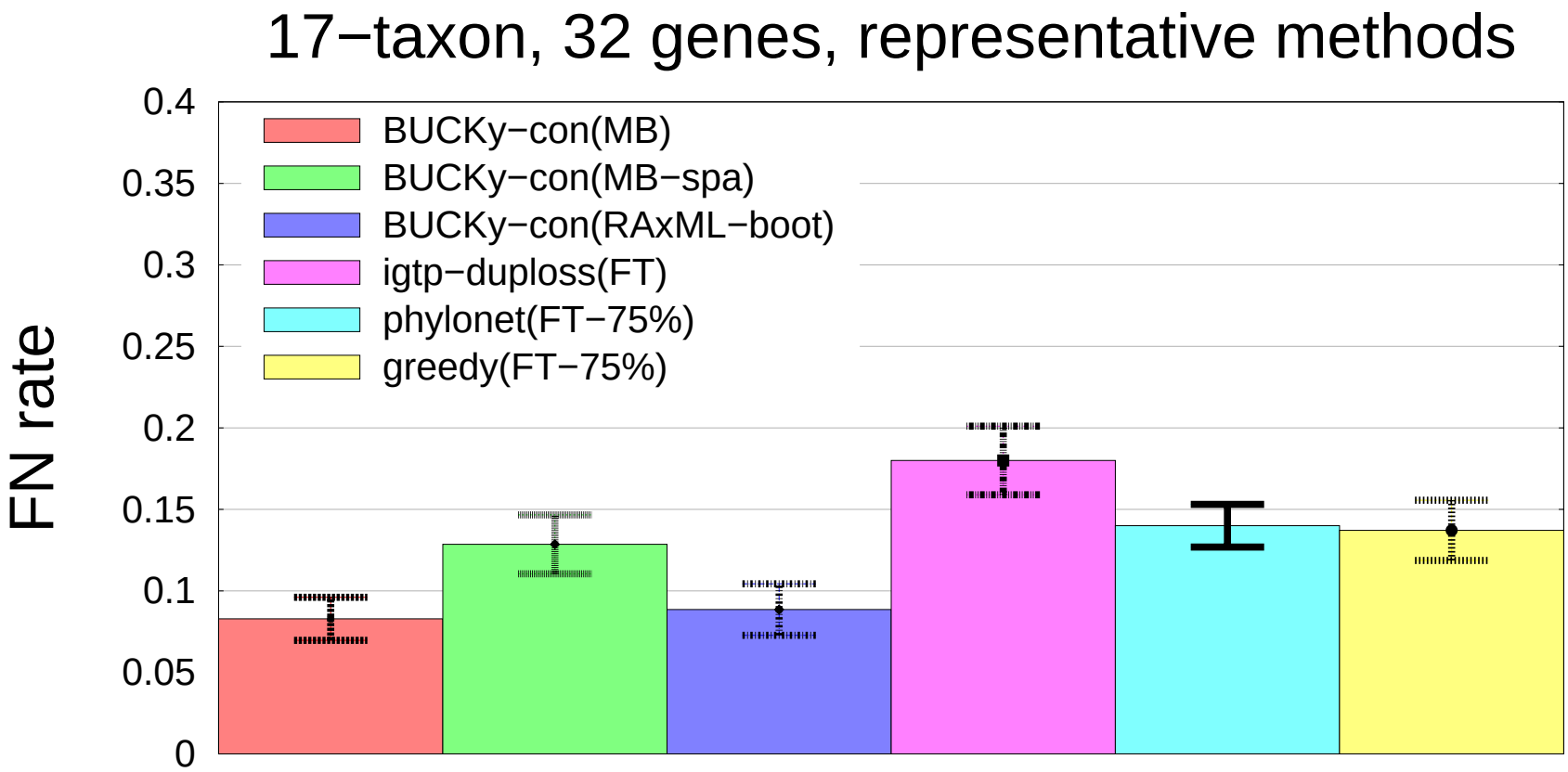
Simulation Parameters

	previous studies	this study
number of taxa	4-20	17-500
number of genes	≤ 100	25-50
evolution model	JC, HKY	GTR + Γ + Indels
cause of discord	ILS, HGT	none, ILS

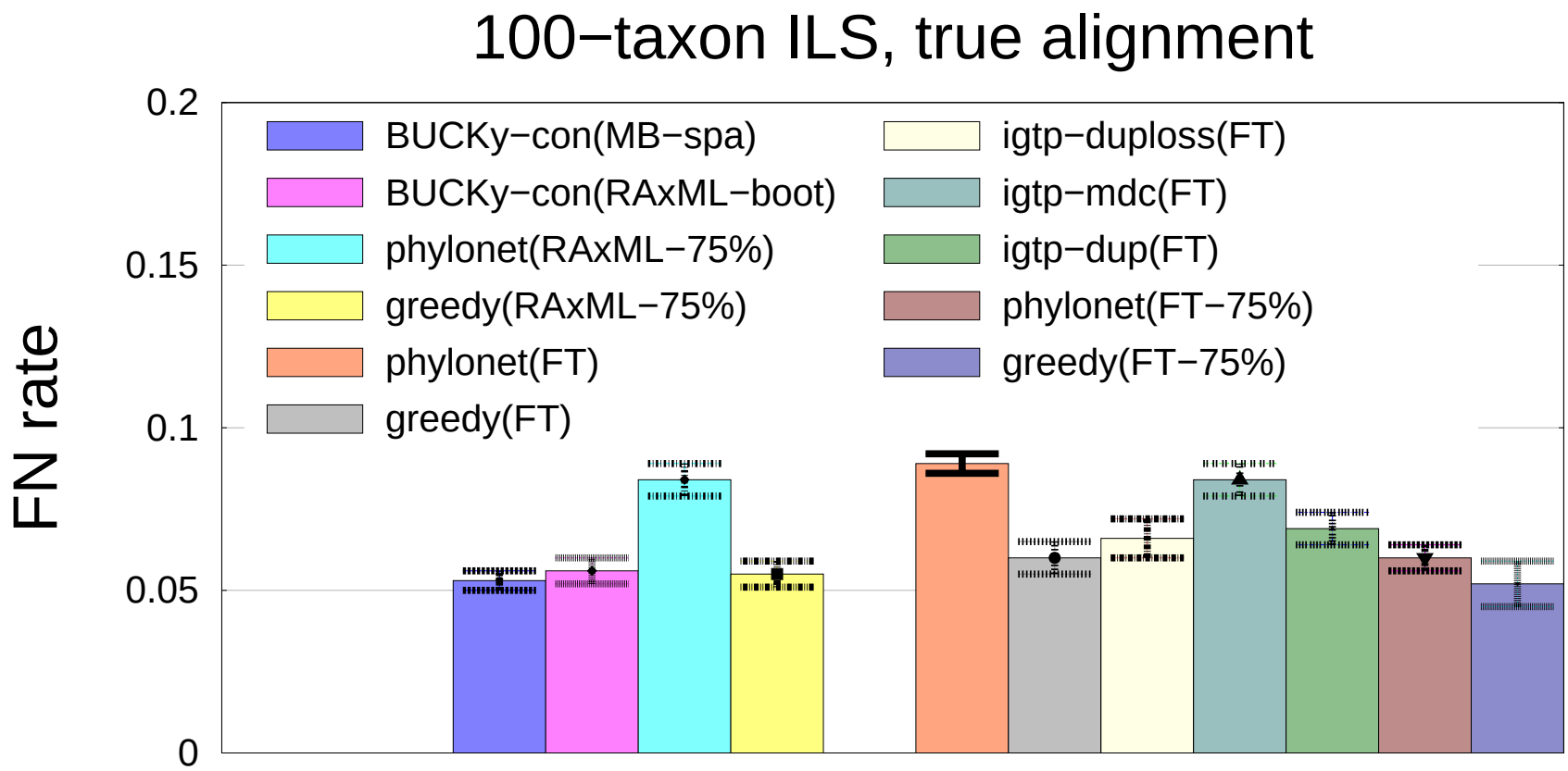
Results on 17-taxon datasets, all methods



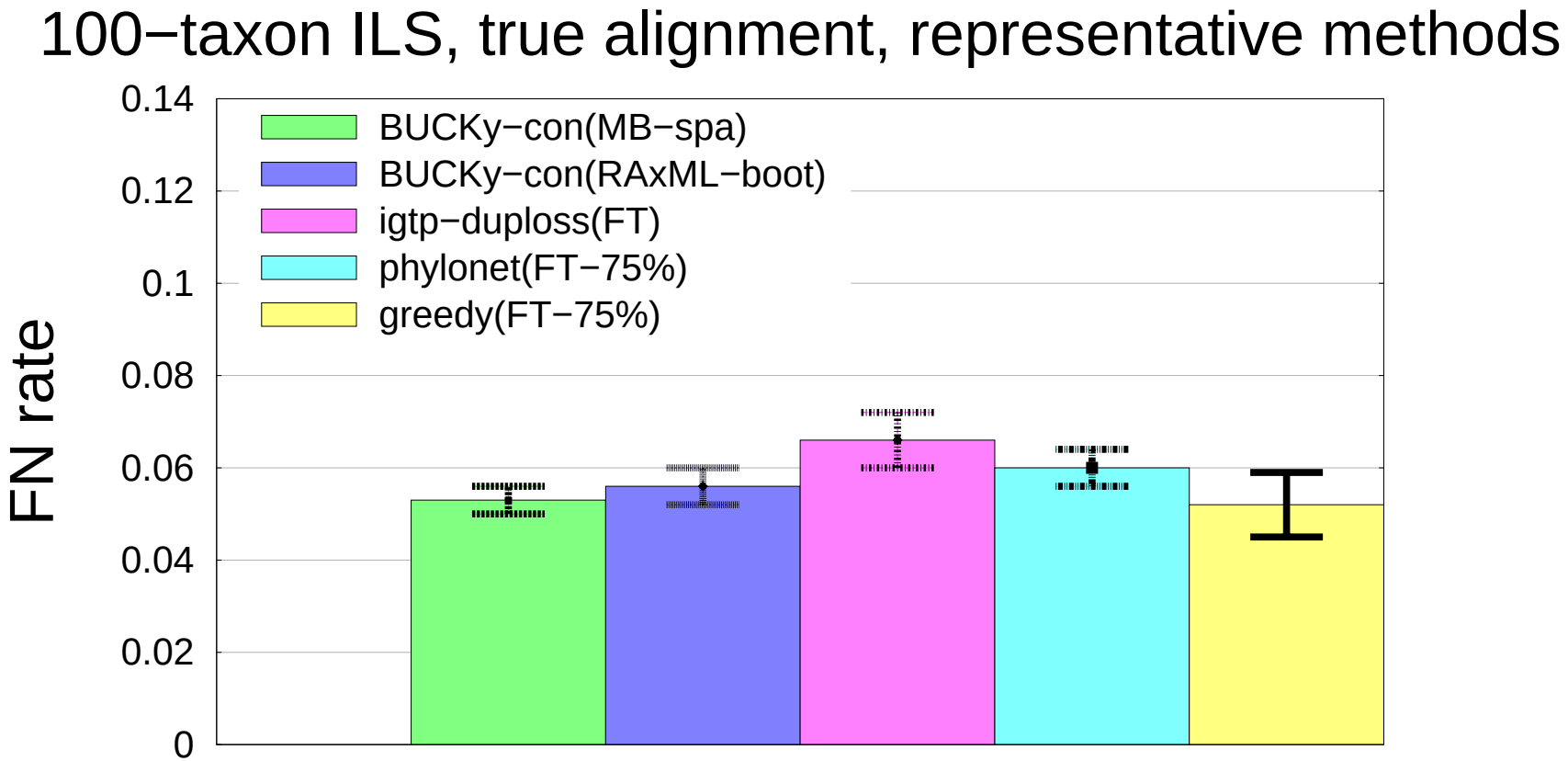
Results on 17-taxon datasets, representative methods



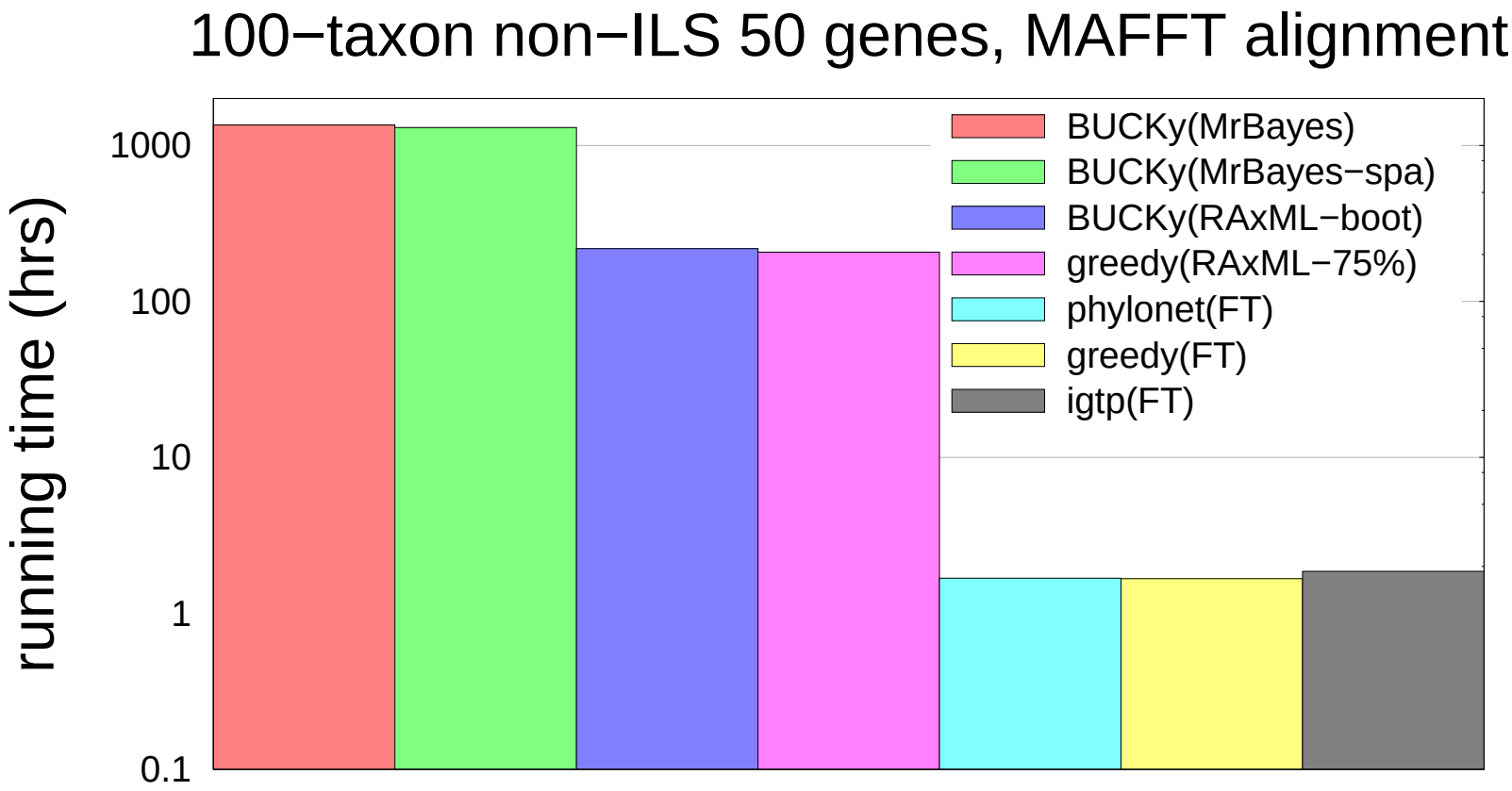
Results on 100-taxon datasets, all methods



Results on 100-taxon datasets, representative methods



Computational Requirements



Memory usage:

- BUCKy: 34-234 GB
- PhyloNet, GLASS, iGTP, greedy: < 9 MB

Findings

- Accounting for gene tree estimation error improves methods
- MrBayes is expensive to run correctly - even on 17-taxon inputs. Using other methods to estimate the gene tree distribution does not reduce accuracy for BUCKy very much.
- Some fast methods (e.g., Greedy(FT)) have accuracy close to that of BUCKy-con(MrBayes)
- BUCKy-con more accurate than BUCKy-pop
- iGTP-duploss more accurate than iGTP-MDC
- GLASS fast but not competitive with other methods

Observations:

- Statistical guarantees are often not predictive of performance on finite data
- Performance on large datasets can be different than on small datasets

Open Questions:

- Why is Greedy so accurate?
- How well do other methods (e.g., *BEAST) perform?
- How do methods perform on incomplete gene trees?
- How do methods perform when gene tree incongruence is due to other factors than ILS?

Acknowledgements

- Jimmy Yang
- Luay Nakhleh and Yun Yu for datasets
- Steve Evans for statistical testing advice
- NSF
- Guggenheim Foundation