Class discussion on 03/09/2016

Xi Z., Liu L., Davis C.C (2015). The impact of missing data on species tree estimation. *Molecular Biology Evolution*, Advanced access published November 20, 2015.

Motivation

Species tree estimation datasets are unlikely to include molecular sequence data for every species and every gene under investigation. Missing data could be

- randomly distributed, e.g., insufficient genome coverage during sequencing experiments
- biased, e.g., rapidly evolving genes or species can fail to be recognized by universal primers

Simulation Study (100 replicates)

(1) Define three species trees from the same 17-taxon tree topology by varying rates of ILS

(2) Generate 50 to 2000 gene trees from species trees under the multi-species coalescent mode (PhyBase `sim.coaltree.sp`)

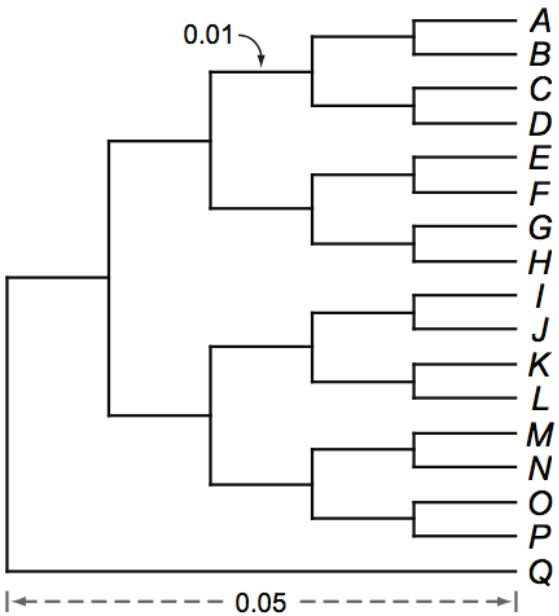
(3) Generate molecular sequence data with 1000 base paris were generated from the gene trees (Seq-Gen w/ Jukes-Cantor)

(4) Remove 35%, 53%, and 70% of the data with respect to the following missing data conditions:

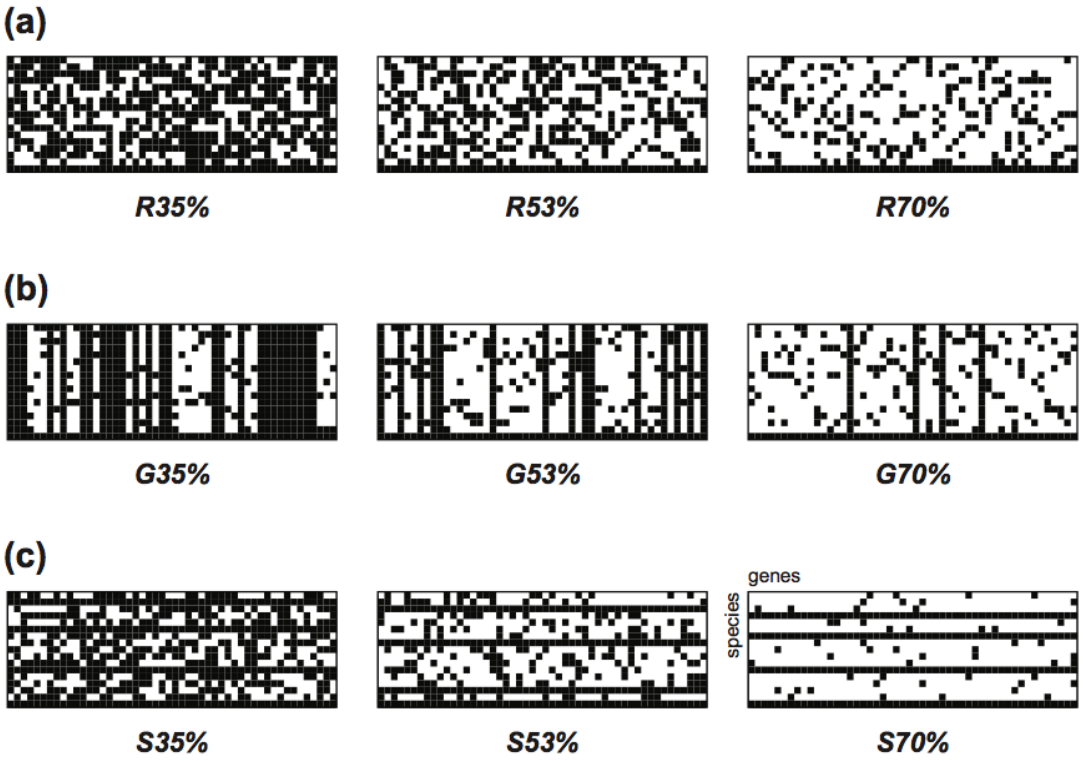
- (*R*) randomly distributed across all genes of the ingroup species
- (*G*) concentrated in a random subset of ingroup genes
- (*S*) concentrated in a random subset of ingroup species

Accuracy, specifically, the topological difference between the true and estimated species, is given by the normalized Robinson-Foulds distance.

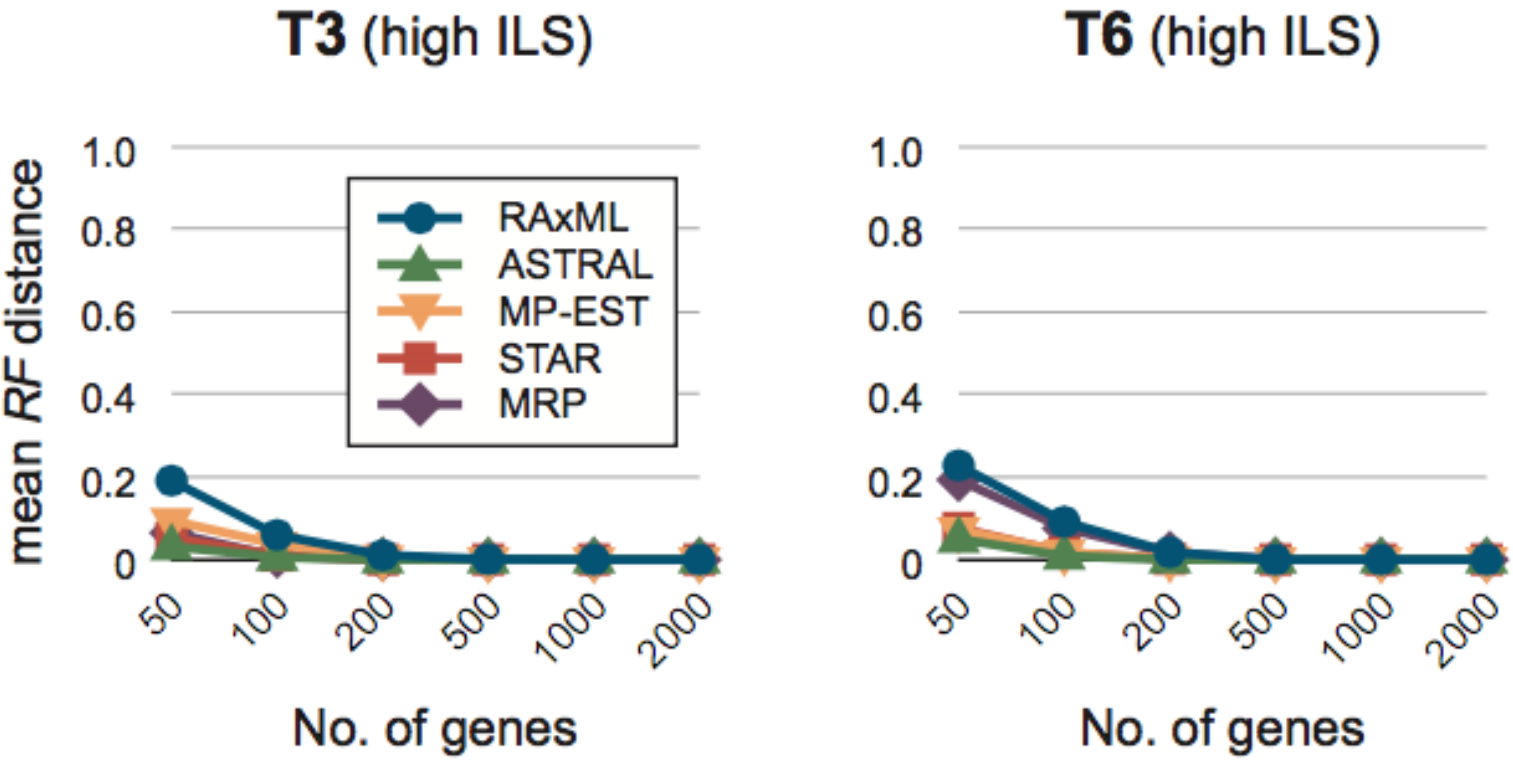
Topology: Q is the outgroup, Alignment: white denotes missing data



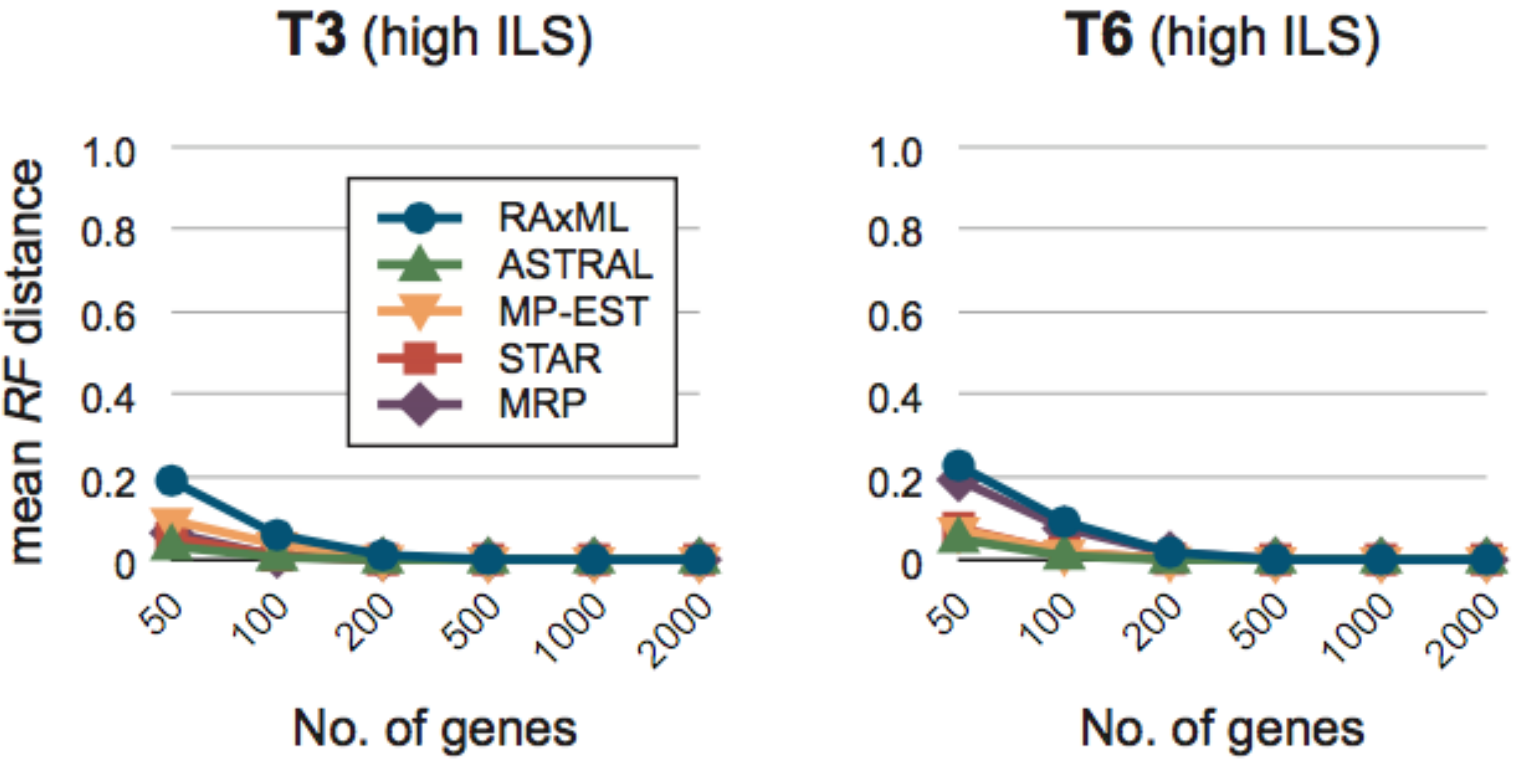
T1 (θ = 0.001, no ILS)
T2 (θ = 0.01, low ILS)
T3 (θ = 0.1, high ILS)



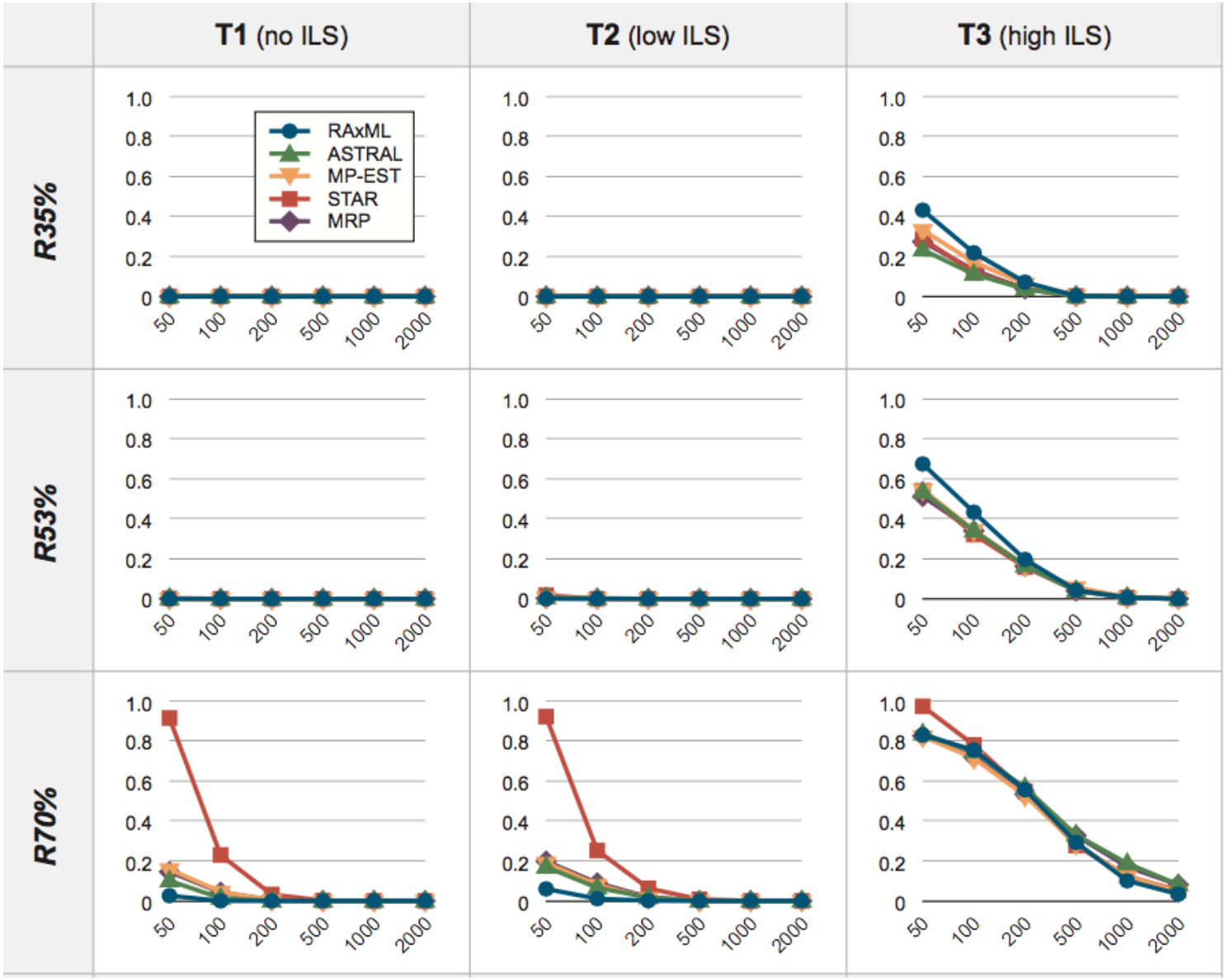
Effect of incomplete lineage sorting with NO missing data



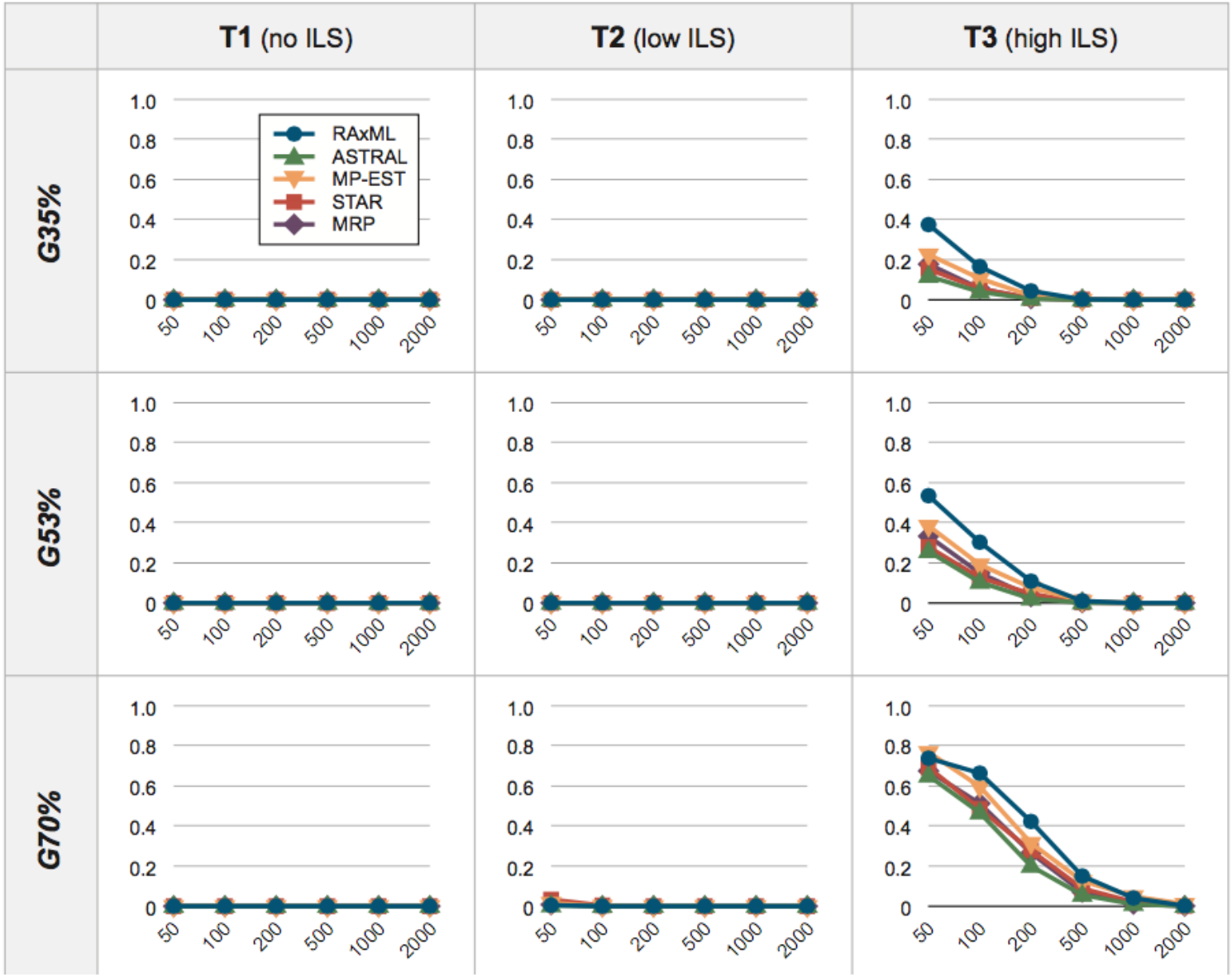
Effect of incomplete lineage sorting with NO missing data



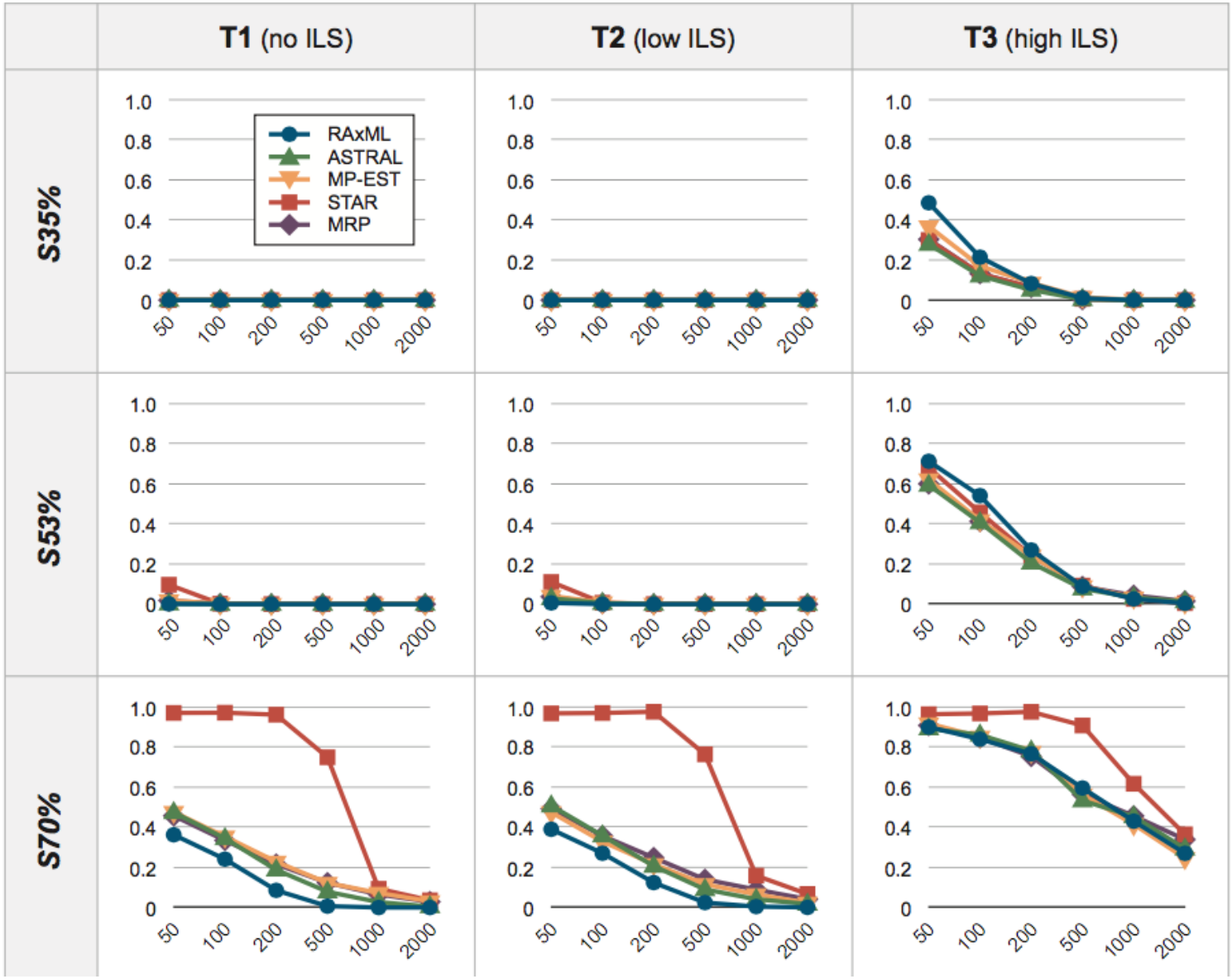
Missing data is randomly distributed across genes & species



Missing data is concentrated in a random subset of genes



Missing data is concentrated in a random subset of species



Remarks

This papers examines a diverse set of small species trees by varying rates of ILS and rates of evolution across species. Future analyses could include

- a species tree with more than 25 taxa
- including the outgroup in simulations of missing data
- alignments simulated under more complicated models of evolution
  - Kimura or GTR
  - insertions/deletions - INDELible
  - free energy - RNAsim
- missing data during the alignment estimation phase