Species tree estimation from phylogenomic datasets

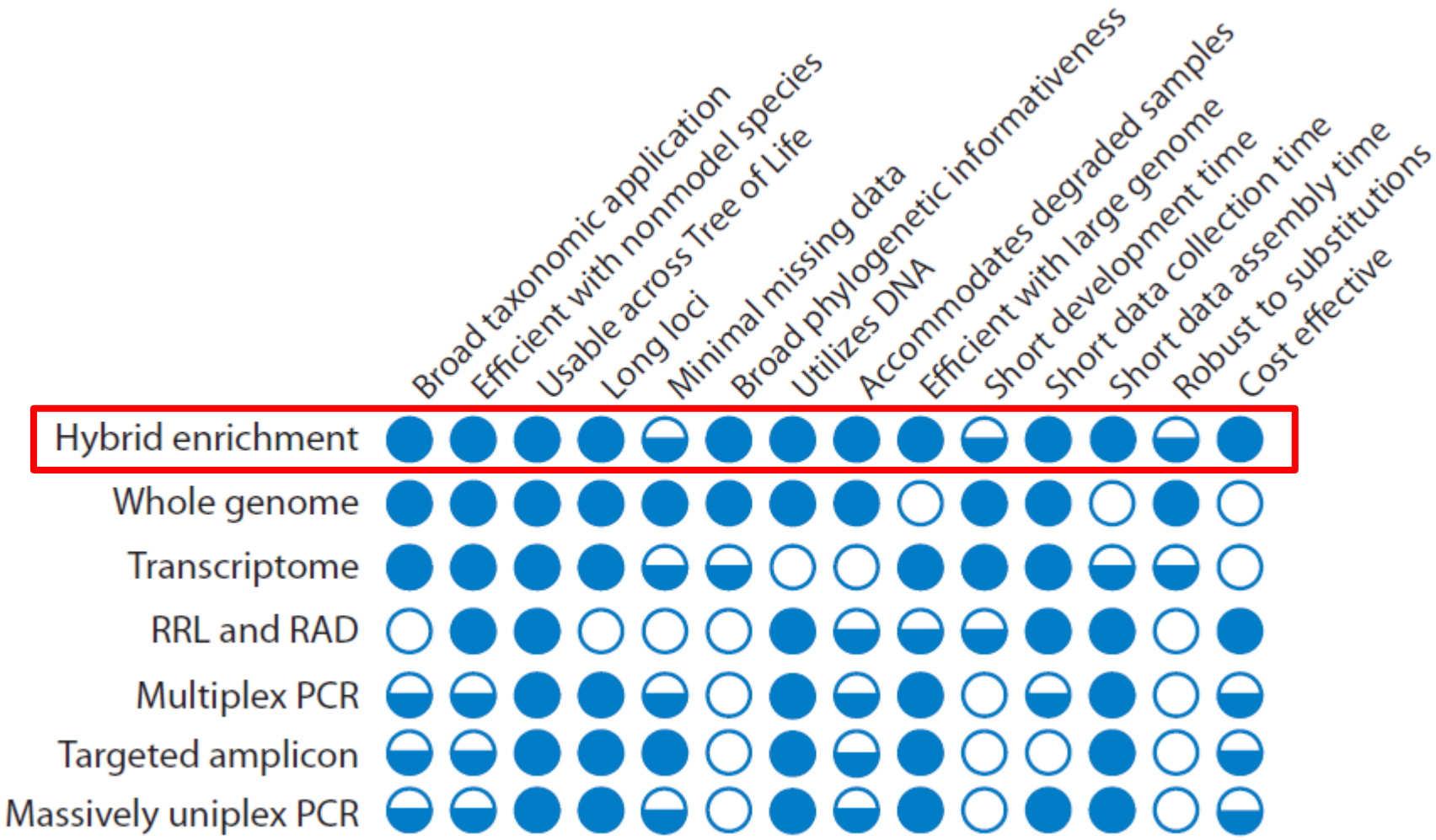
Tomáš Fér

Phylogenomics

- using whole-genome sequences or large portion of the genome to build a phylogeny
  - whole chloroplast sequences
  - hundreds or thousands of genes

- gene tree – individual evolutionary history
- species tree – ‘true’ species evolution
- gene tree/species tree

Different phylogenomic approaches

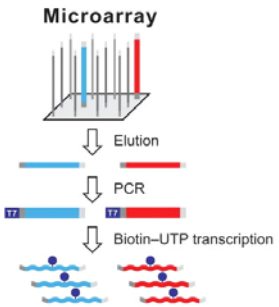


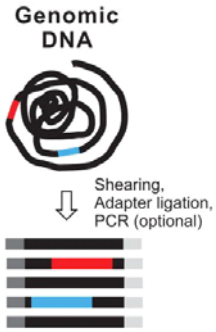
Lemmon E.M. & Lemmon A.R. (2013): *High-throughput genomic data in systematics and phylogenetics*. Annu. Rev. Ecol. Evol. Syst, 44, 99–121.

Hyb-Seq overview

probe design

bait design and synthesis (MYcroarray MYbaits)

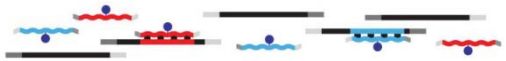




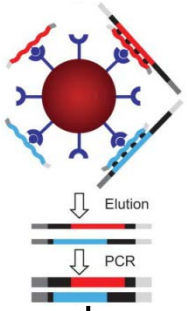
DNA extraction

sonication

library preparation



solution hybridization



Illumina MiSeq 2x150 PE

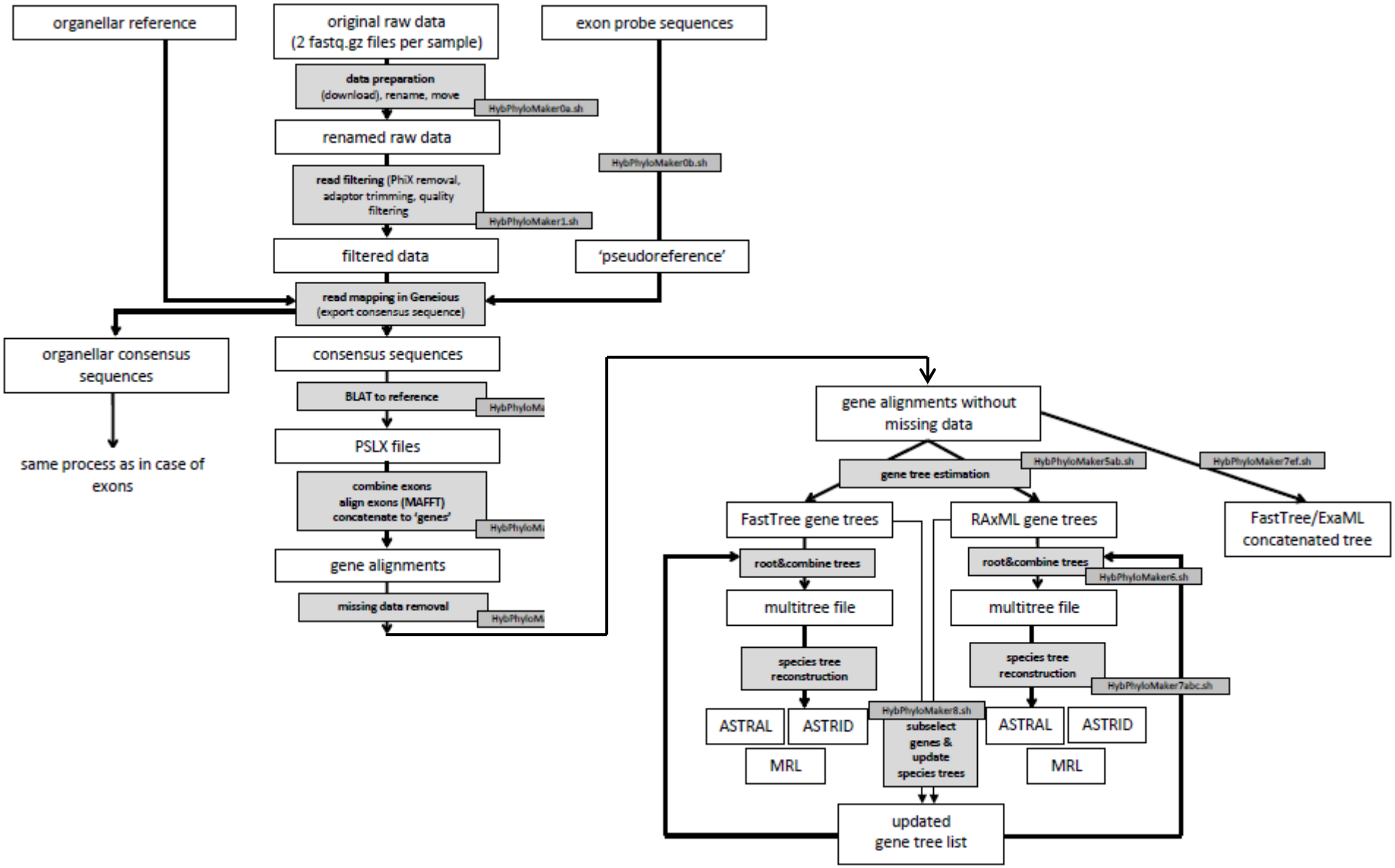
sequencing

data analysis

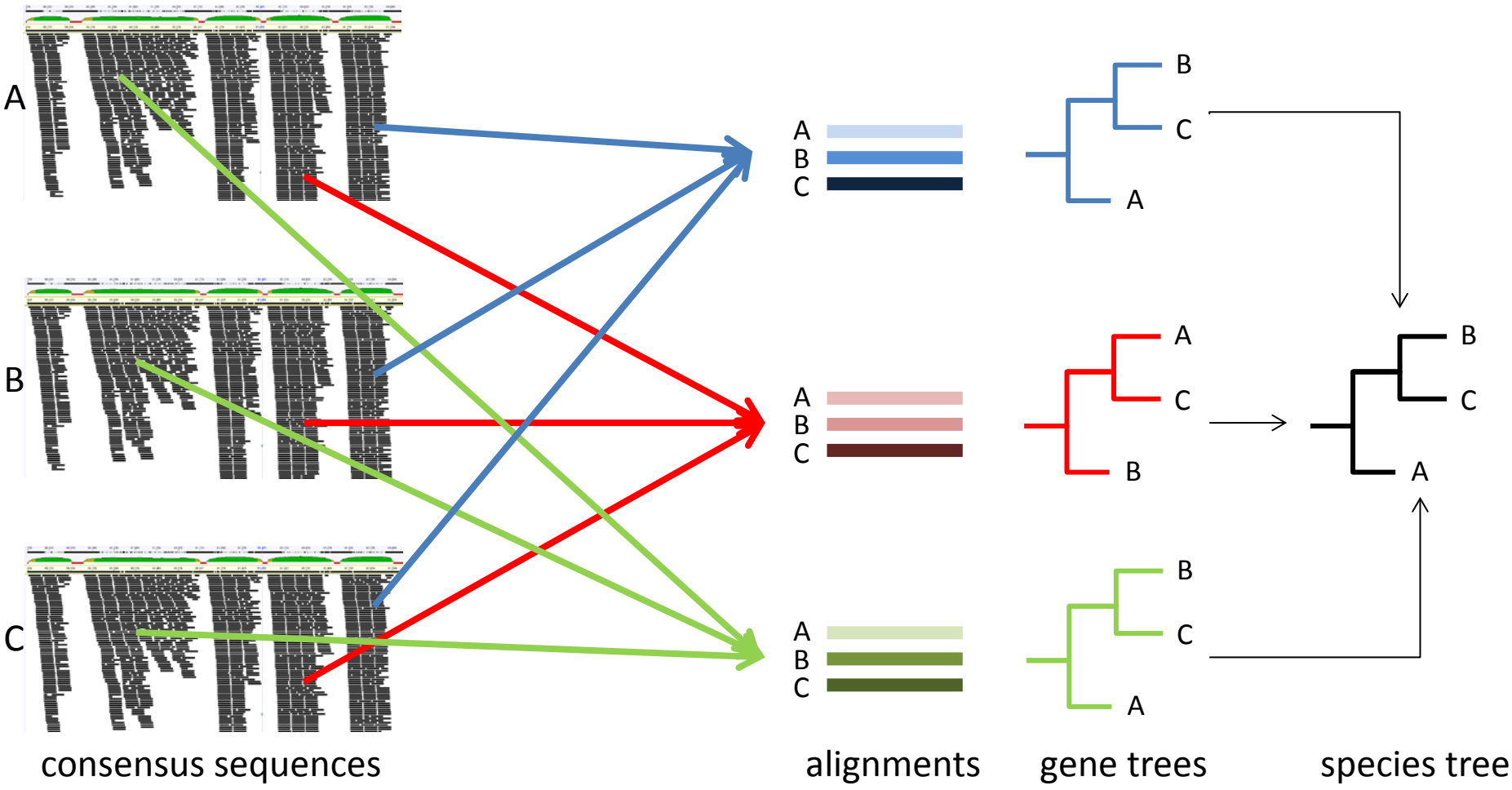
Gnirke et al. 2009

HybPhyloMaker

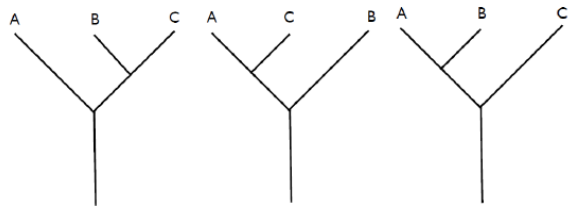
https://github.com/tomas-fer/HybPhyloMaker

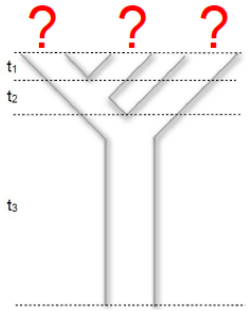


Read processing



Species tree estimation





- **concatenation** – good unless strong ILS
  - single partition model (e.g., MP)
  - multiple partitions model (ML or Bayesian)
- **consensual methods** using MP – minimizes deep coalescences (MDC)
- multispecies coalescence (all incongruences due to differences in coalescence processes, no hybridization)
  - **coestimation** of gene trees and species tree – *BEAST – Bayesian analysis
  - **summary methods**
    - supertree methods – MRL (maximum representation with likelihood)
    - MP-EST – maximum pseudo-likelihood for estimating species tree
    - ASTRAL, ASTRID, STAR, STEAC – very fast and accurate
- Bayesian concordance analysis (BUCKy) – quartet-based Bayesian species tree estimation – uses concordance factor to build dominant history

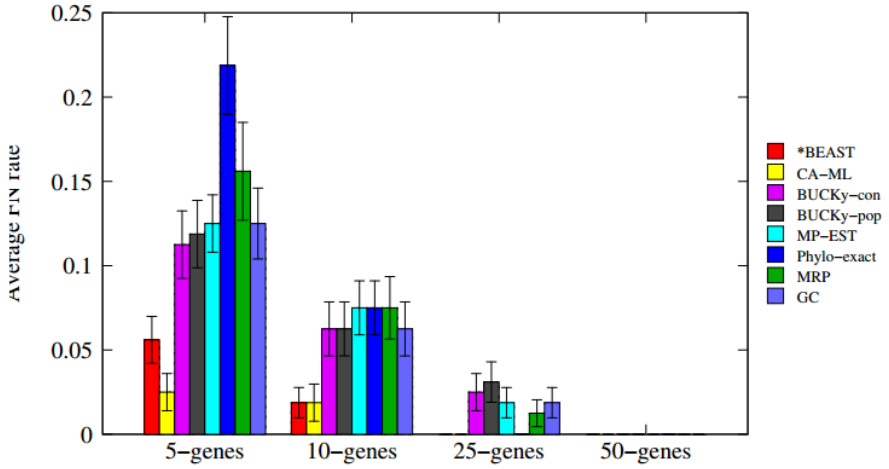
Summary methods

Species tree estimation

- supertree
  - MRP – maximum representation using parsimony
  - MRL – maximum representation using likelihood
- MP-EST – maximum pseudo-likelihood approach for estimating species trees
- STEAC – species tree estimation using average coalescence times
- STAR – species tree estimation using average ranks of coalescences
- ASTRAL – Accurate Species Tree Reconstruction ALgorithm
- ASTRID – Accurate Species TRees from Internode Distances (reimplementation of NJ_{st} method)

Methods comparison

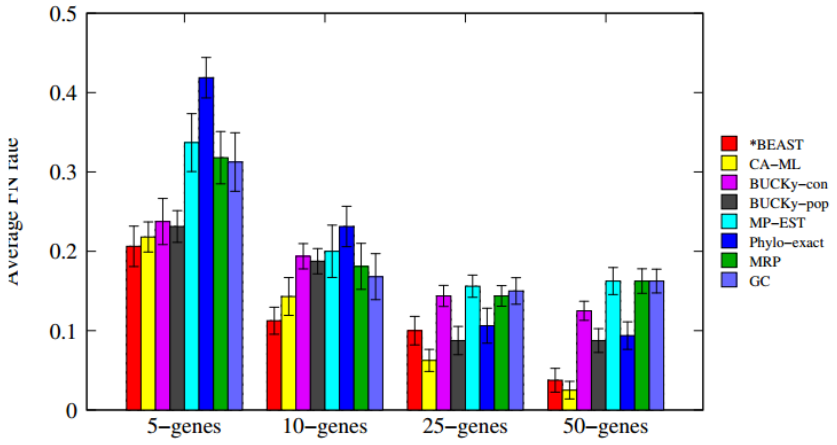
Results on 11-taxon datasets with weak ILS



*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: concatenated analysis) most accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

Results on 11-taxon datasets with strongILS

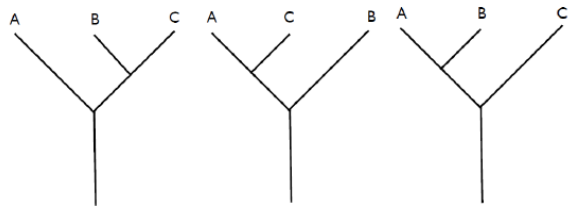


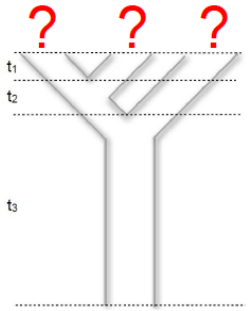
*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: (concatenated analysis) also very accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

T. Warnow, The University of Texas
https://www.cs.utexas.edu/users/tandy/394C-nov20-2013.pdf

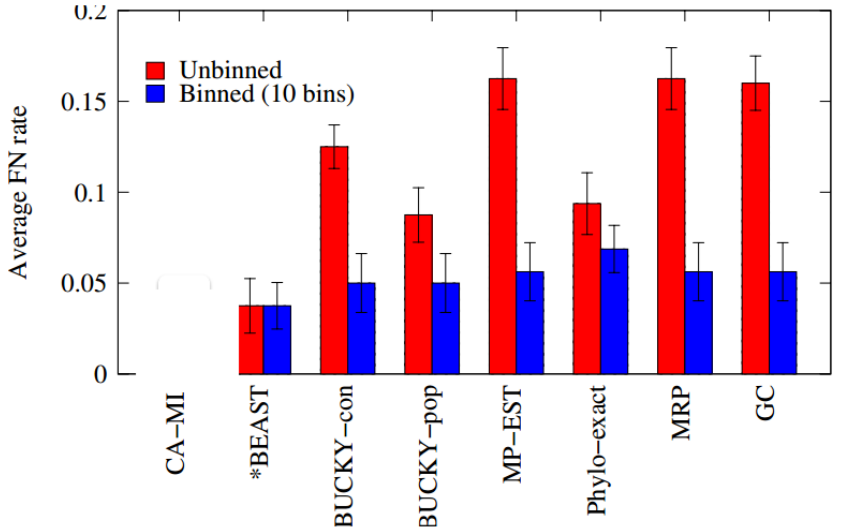
Species tree estimation





- wrong species tree if poor gene trees
  - shorter alignments usually give poorly supported trees
- improve gene trees
  - collapse unsupported branches
  - **binning** – assign gene to bins
    – create supergene alignments
    - **naïve binning** – random
    - **statistical binning**
      – no incompatibility among gene trees in the same set

11-taxon strongILS with 50 genes



T. Warnow, The University of Texas
https://www.cs.utexas.edu/users/tandy/394C-nov20-2013.pdf

ASTRAL

Accurate Species Tree Reconstruction ALgorithm

- unrooted gene trees
- species tree that agrees with the largest number of quartet trees induced by the set of gene trees
- weighting all three alternative quartet topologies according to their relative frequencies within gene trees
  - much more frequent topology – trees without this topology are penalized
  - similar frequencies (i.e., close to 0.33) – the quartet has little impact to optimization
- final species tree with
  - local posterior probability that the branch is in the species tree
  - the length of internal branches in coalescent units

MRL

Maximum Representation with Likelihood

- supertree methods – estimates species tree on full taxon sets from sets of smaller trees (i.e., with missing species)
- encodes a set of gene trees by a large randomized matrix
- each edge (branch) in each gene tree
  - ‘0’ for the taxa that are on one side of the edge
  - ‘1’ for the taxa on the other side
  - ‘?’ for all the remaining taxa (i.e., the ones that do not appear in the tree)
- MRL matrix is analyzed using heuristics for a symmetric 2-state Maximum Likelihood
  - in RAxML as ‘BINCAT’ model

‘Good genes’

genes with good properties (no paralogs, low conflicting signal…)

alignments

- length – longer better
- missing data – fewer better
- parsimony informative sites – more better
- information content

trees

- average bootstrap support – higher better
- average branch length – higher means faster gene
- saturation – correlation between p-distances and tree distance