Fast coalescent-based computation of local branch support from quartet frequencies

Erfan Sayyari, Siavash Mirarab

University of California, San Diego

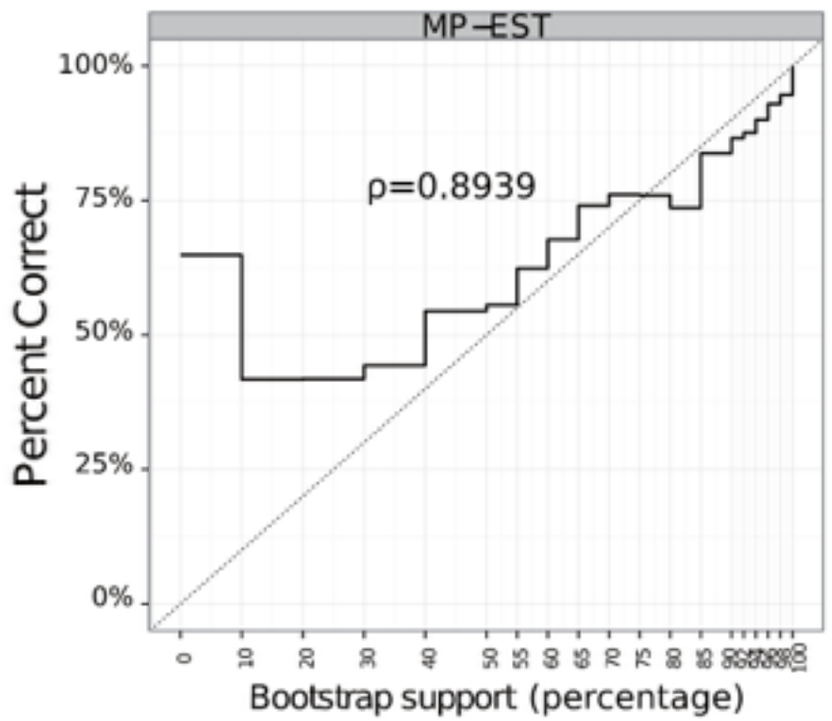
Problem statement

Compute **branch support** support for a **species tree** inferred from a set of gene trees that may disagree with the species tree due to **Incomplete Lineage Sorting (ILS).**

Be **scalable** to very large datasets.

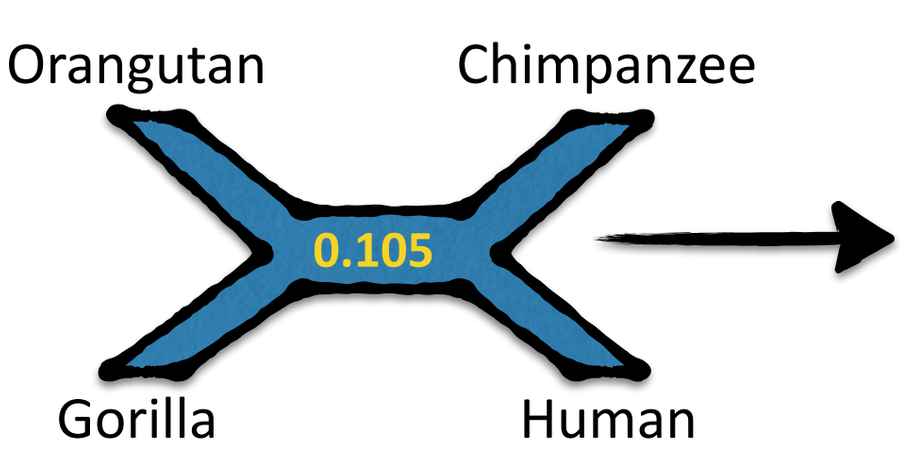
Traditional approach: Multi-locus bootstrapping (MLBS)

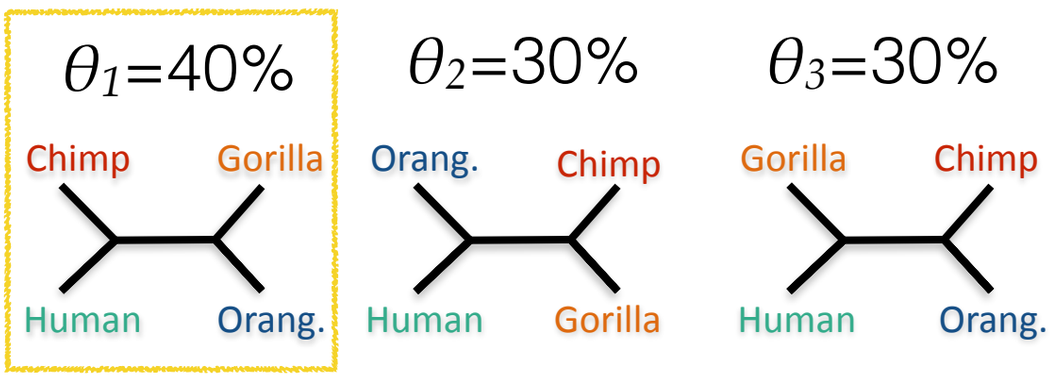
- Slow: requires bootstrapping each gene
- Interpretation is hard (not fully indicative of accuracy)
  [Mirarab et al., Sys bio, 2014; Bayzid et al., PLoS One, 2015]

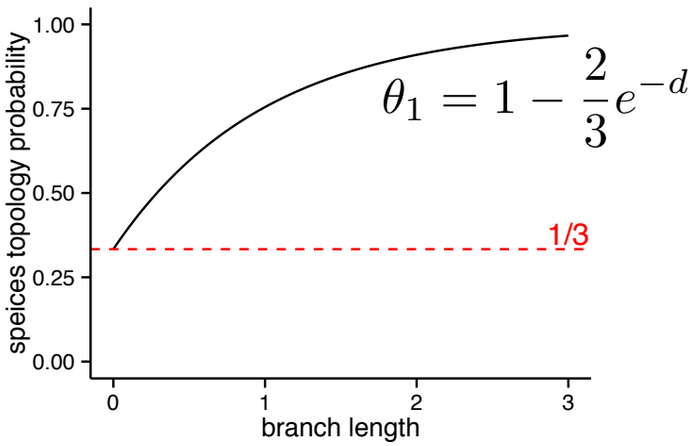


Quartets and Multi-Species Coalescence (MSC)

According to the MSC, for 4 taxa, the unrooted species tree quartet topology has at least 1/3 probability in gene trees (Allman, et al. 2010)

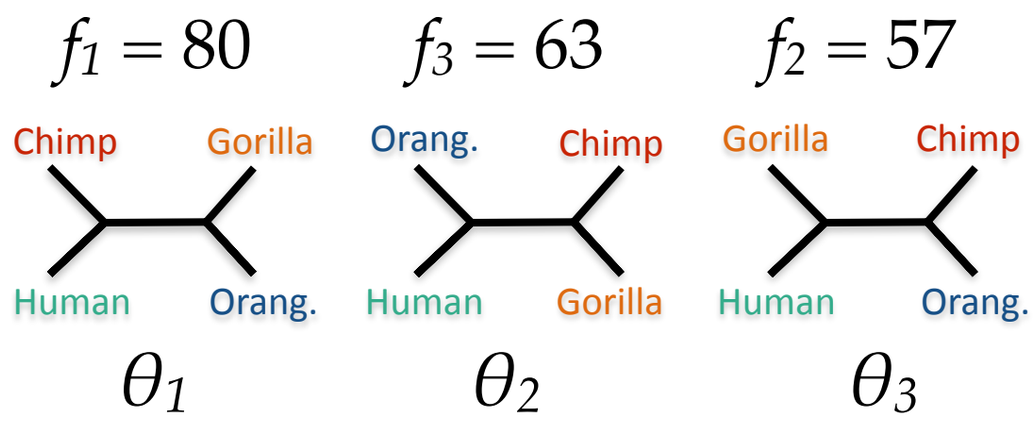






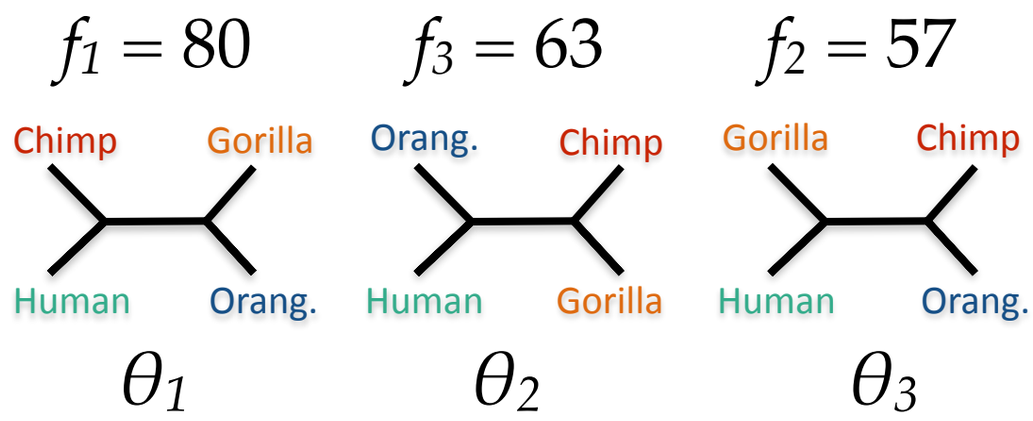
Branch support: single quartet

- Quartet frequencies in gene trees are multinomially distributed



Branch support: single quartet

- Quartet frequencies in gene trees are multinomially distributed



- $\boldsymbol{P}$ **(** quartet tree with frequency f_1 in k gene trees is in the species tree **)**
  =
  $\boldsymbol{P}$ **(** $\theta_1 > 1/3$ **)**

Posterior

$$\Gamma t^{z_1}\left(\frac{1-t}{2}\right)^{n-z_1}$$

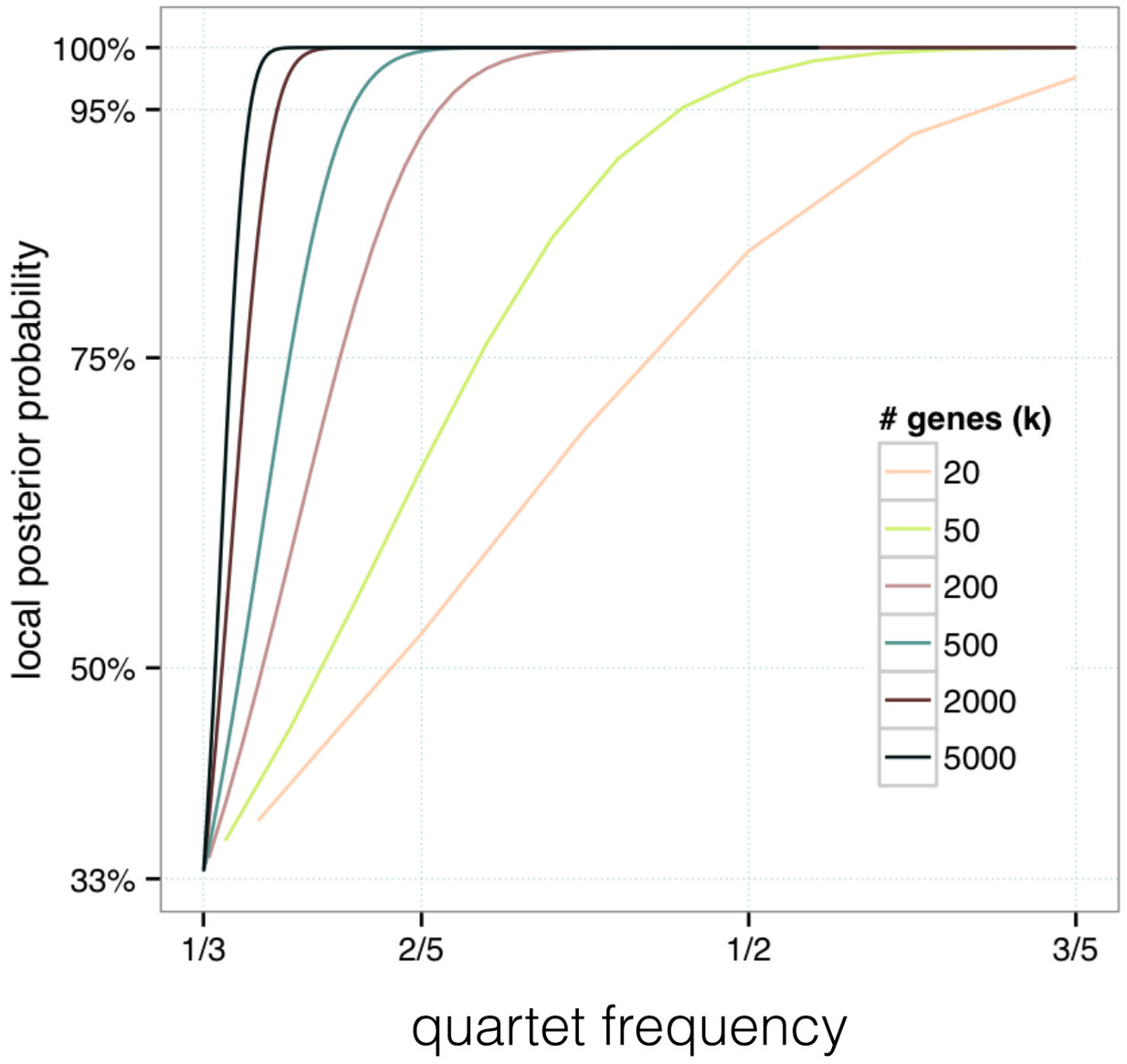
Prior: Yule process become conjugate

$$P\left(\theta_1 > \frac{1}{3}|\bar{Z}=\bar{z}\right) = \frac{\int_{\frac{1}{3}}^{1} P(\bar{Z}=\bar{z}|\theta_1=t) f_{\theta_1}(t)\mathrm{d}t}{P(\bar{Z}=\bar{z})}$$

$$\sum_{j=1}^{3}\int_{\frac{1}{3}}^{1} P(\bar{Z}=\bar{z}|\theta_j=t) f_{\theta_j}(t)\mathrm{d}t$$

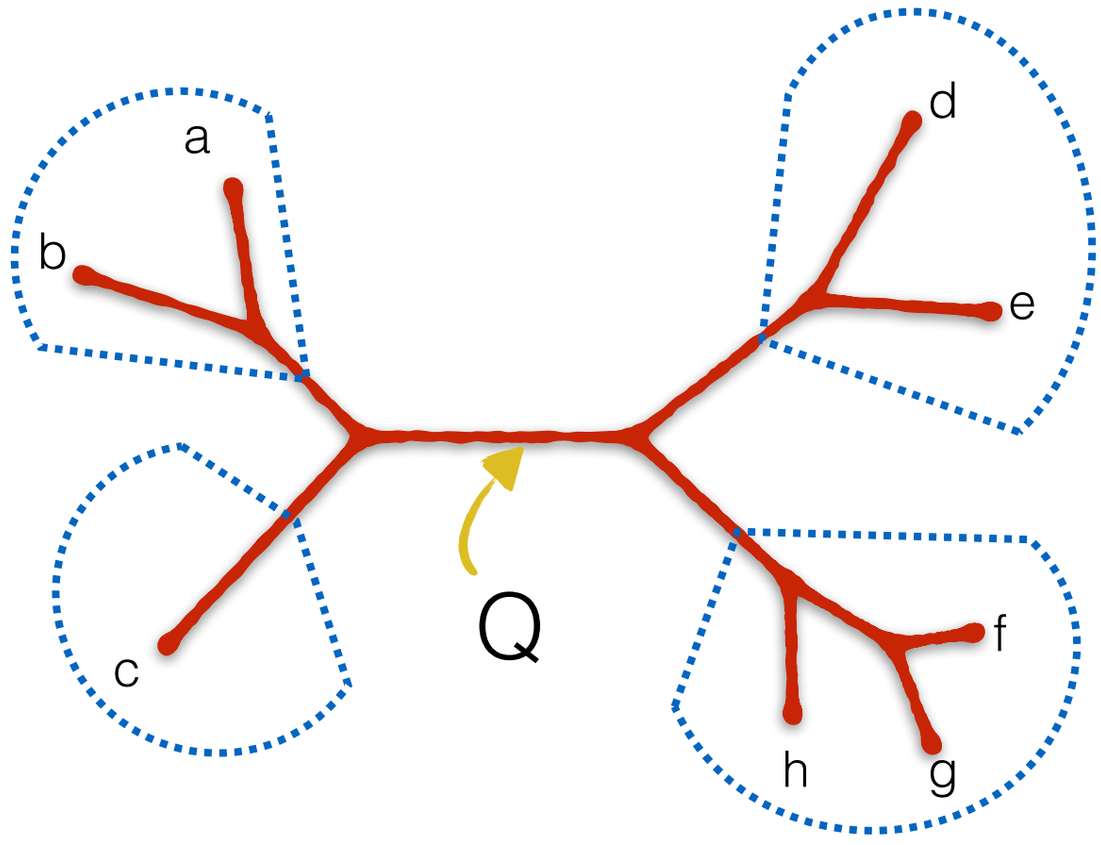
- Easy to calculate
- Depends on the frequency of not just the first topology, but also the frequency of second and third topologies

Local posterior probability versus quartet support



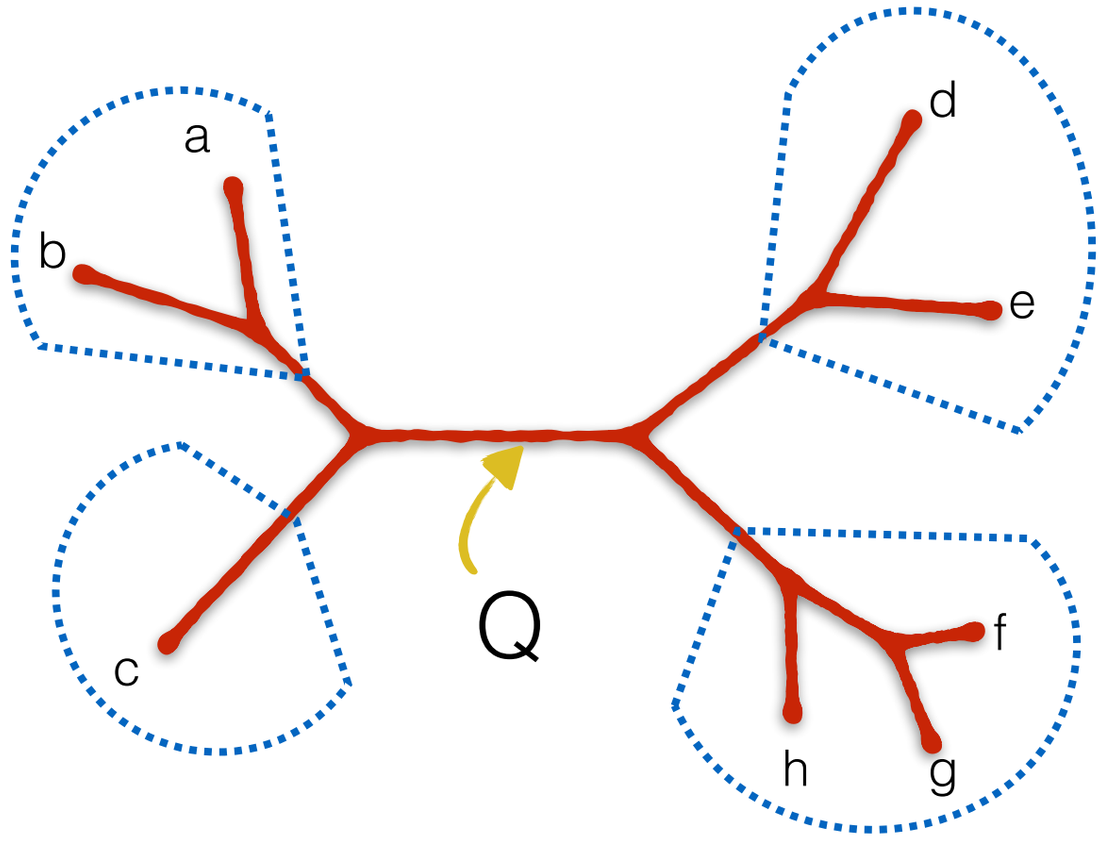
Multiple quartets

- **Locality assumption**: All four clusters around a branch Q are correct
    - Compute support for each branch independently from others



Multiple quartets

- **Locality assumption**: All four clusters around a branch Q are correct
  - Compute support for each branch independently from others
- **Full dependence assumption**: Frequencies of all m quartets around a branch are noisy estimates of a single hidden true probability
  - Simply average quartet frequencies
  - k die tosses, each time reading the results m times with some noise



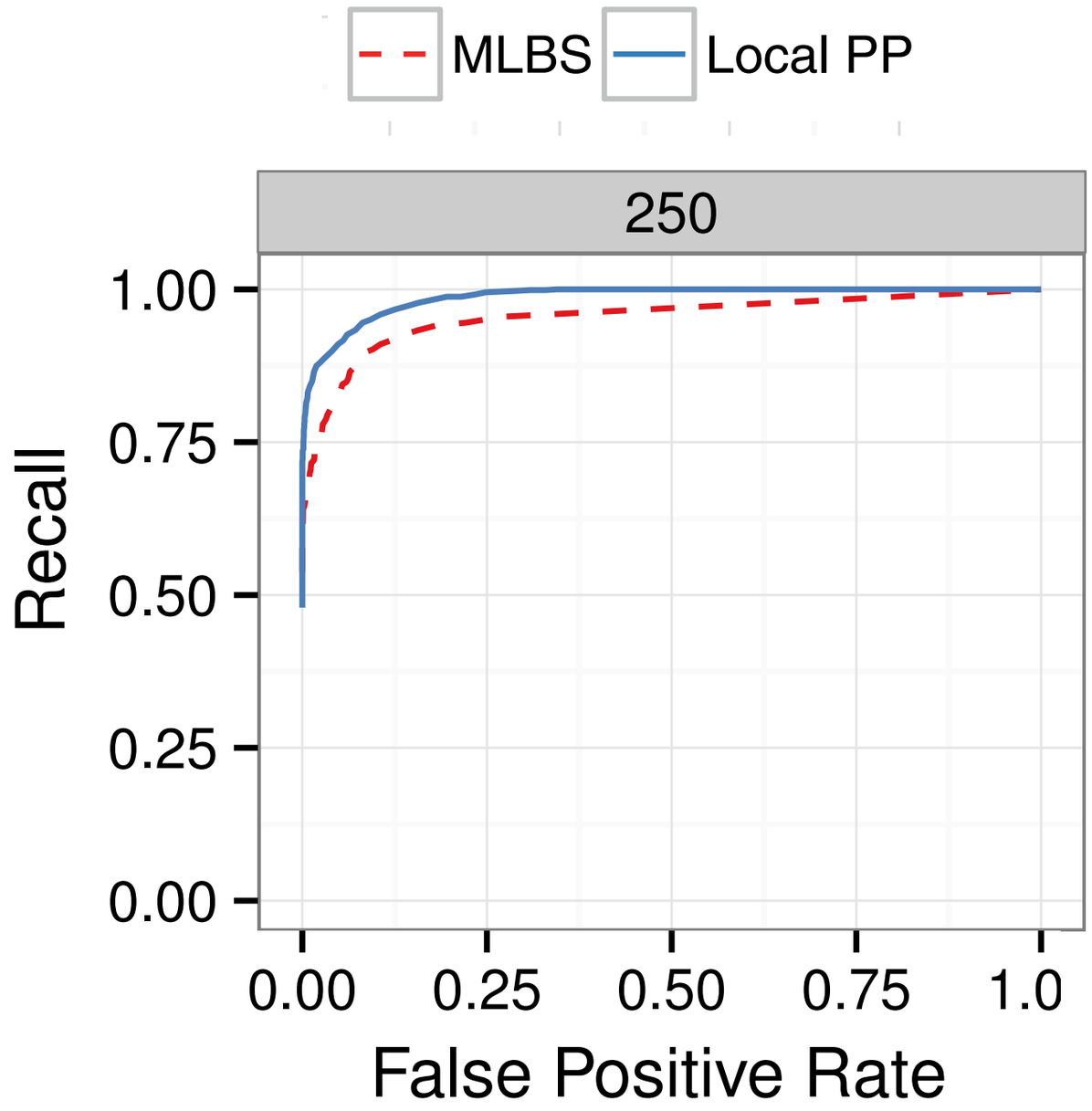
Speed

- We don't need to compute all n choose 4 quartet frequencies

- With algorithmic tricks, we can compute average quartet frequencies around each branch in $O(nk)$

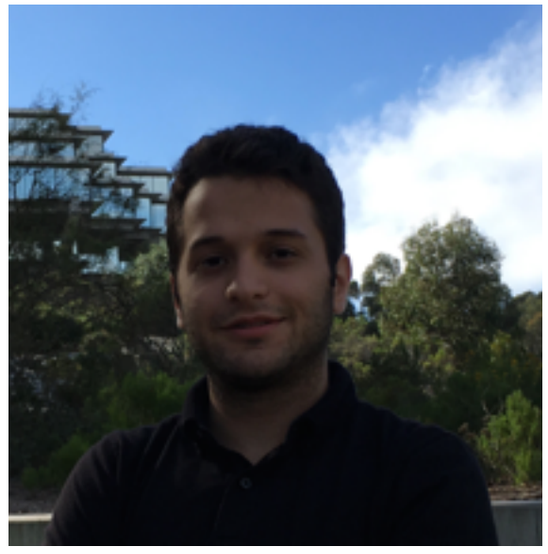
- Couple of minutes for 1000 taxa and 1000 genes

Results (ROC curves)



Avian simulated dataset (48 taxa, 1000 genes)

Erfan Sayyari



- Implemented in ASTRAL
- https://github.com/smirarab/ASTRAL
- Published in

  Erfan Sayyari and Siavash Mirarab. “Fast coalescent-based computation of local branch support from quartet frequencies”. Molecular Biology and Evolution (2016).

  doi:10.1093/molbev/msw079

Conjecture: deficiencies of MLBS relate to increased observed discordance