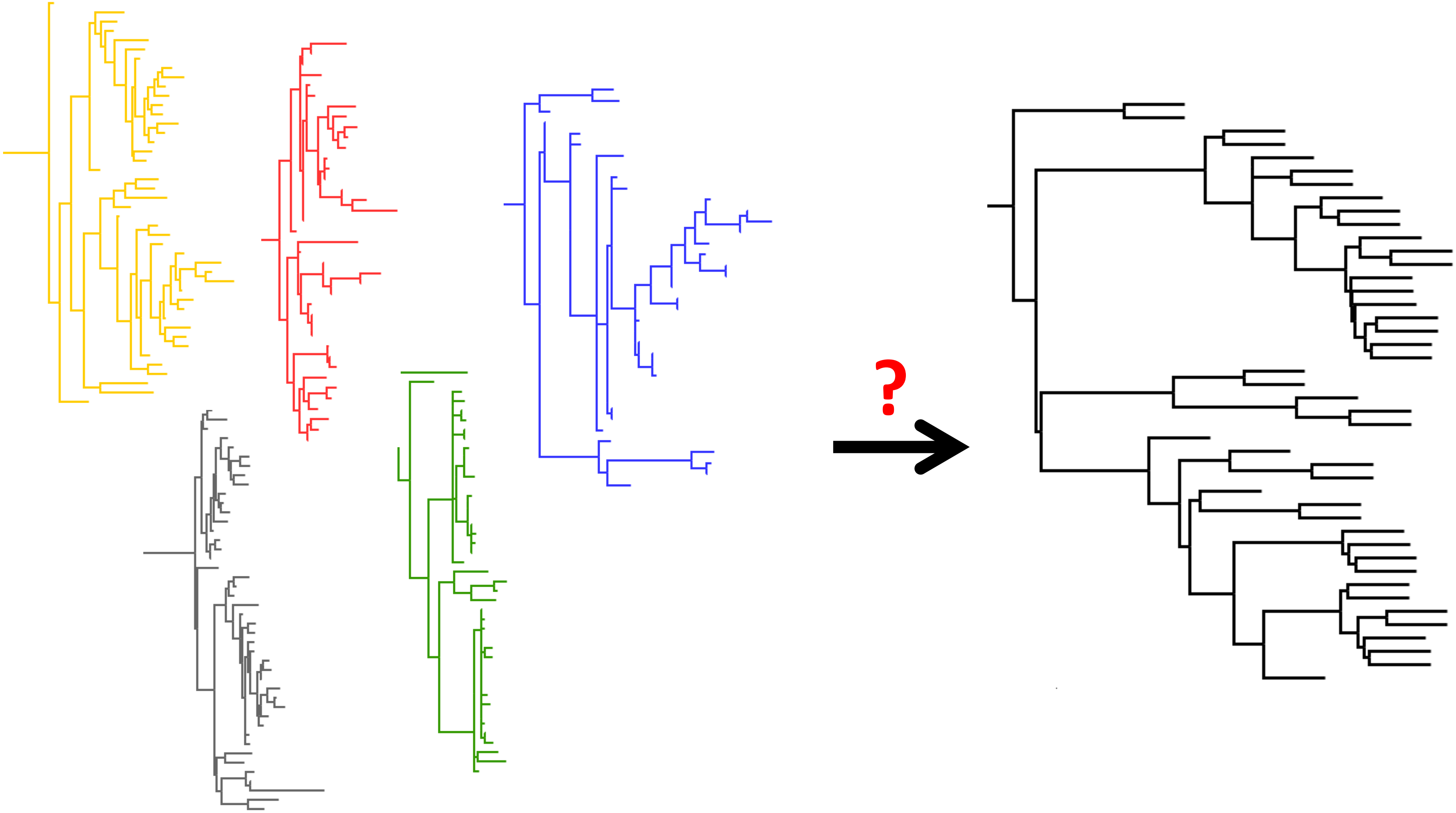
Gene trees vs species tree

Phylogenetic methods

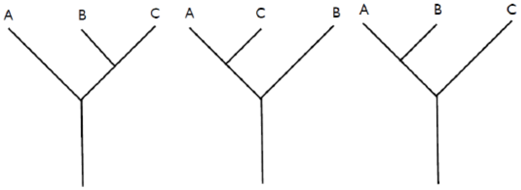
Tomáš Fér

2024

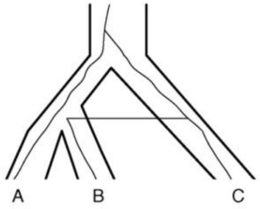
Species tree from gene trees



Incongruencies among loci: gene trees vs species tree

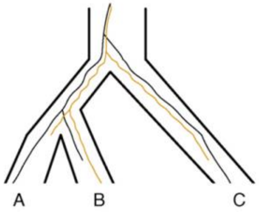


- incomplete lineage sorting (ILS)



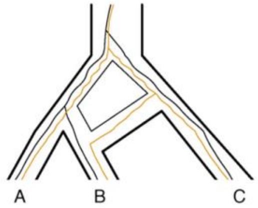
(A(BC))

- horizontal gene transfer (HGT)
  - affects small DNA segments
- gene duplication and loss (GDL)
  - orthology problem



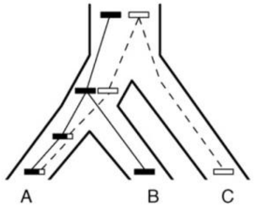
((AC)B)

- hybridization
  - affects whole genomes



((AB)C)
(A(BC))

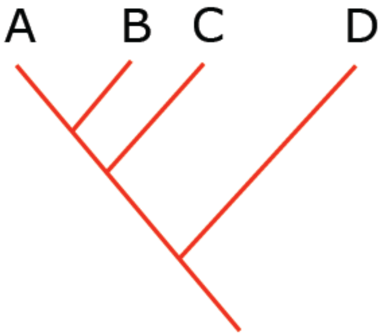
- recombination
  - different histories for neighboring segments in genes



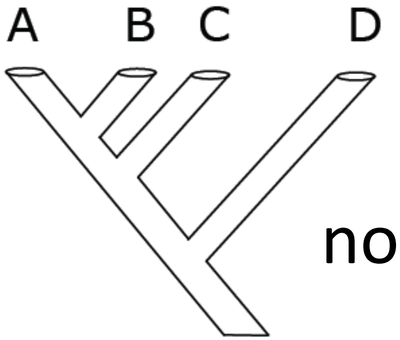
((AB)C)
((AC)B)

Degnan & Rosenberg, 2009

Gene duplications and losses

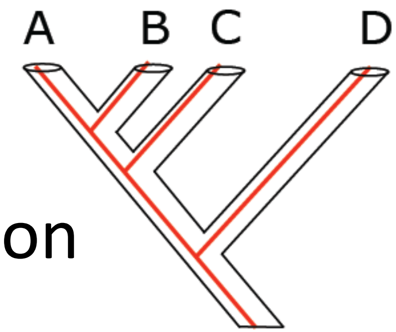


Gene tree

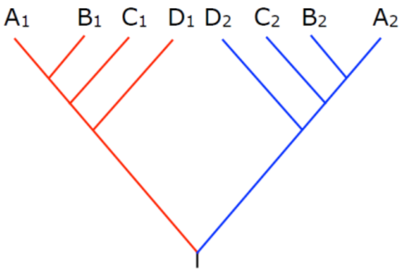


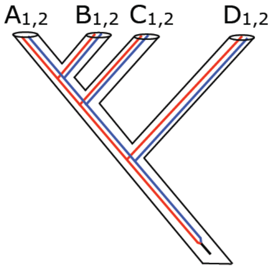
Species tree

no gene duplication

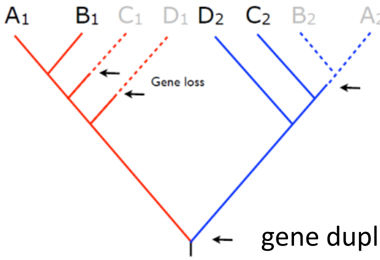


Gene tree = Species tree





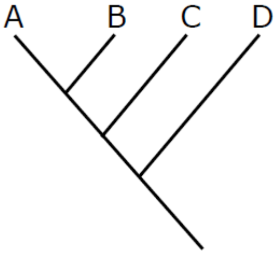
gene duplications and losses (GDL)

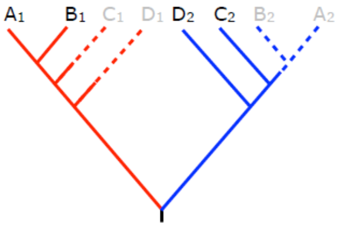


gene loss

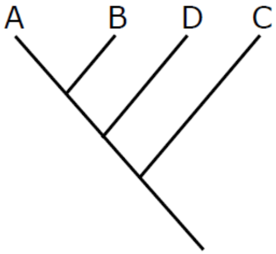
gene duplication

true species tree





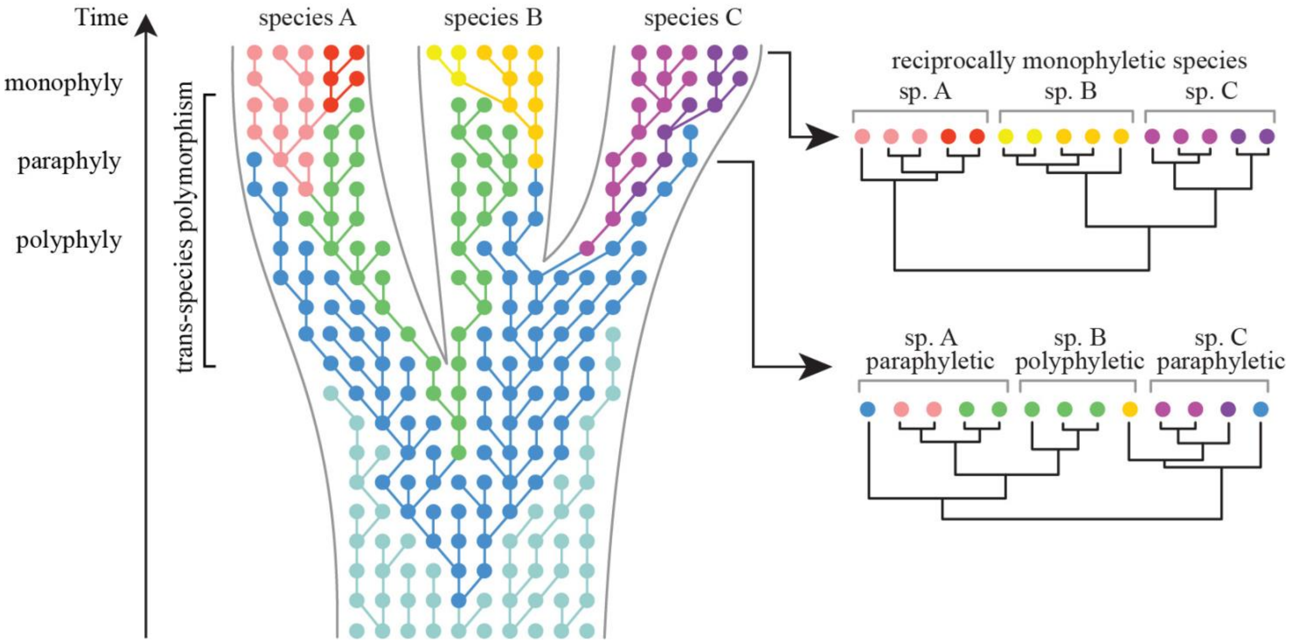
inferred species tree



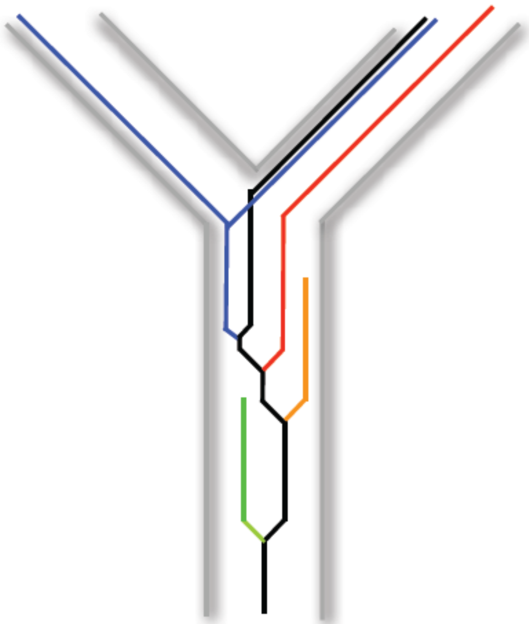
M. Popp, Oslo

Incomplete lineage sorting

Coalescence processes

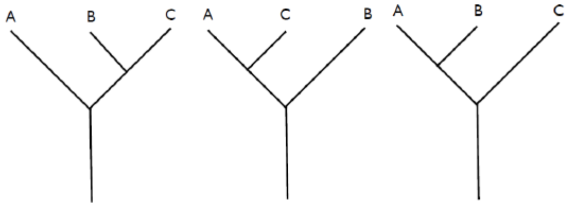


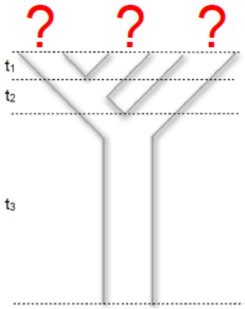
https://frederikleliaert.wordpress.com/green-algae/dna-based-species-delimitation-in-algae/



M. Popp, Oslo

Species tree estimation





- **concatenation**

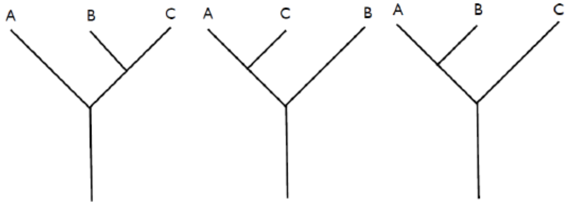
- multispecies coalescence
  - *BEAST (**coestimation** of gene trees and species tree)
  - **summary** methods (combining gene trees)

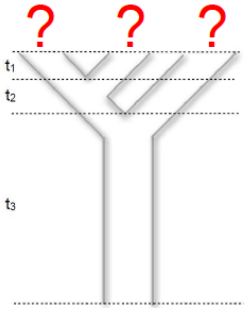
- **supertree** methods
  - MRL (maximum representation using likelihood)

- Bayesian **concordance** analysis (BUCKy)
  - quartet-based Bayesian species tree estimation

- **site-based** methods
  - SNAPP, SVDquartets

Species tree estimation





- **concatenation**

- multispecies coalescence
  - *BEAST (**coestimation** of gene trees and species tree)
  - **summary** methods (combining gene trees)

- **supertree** methods
  - MRL (maximum representation using likelihood)

- Bayesian **concordance** analysis (BUCKy)
  - quartet-based Bayesian species tree estimation

- **site-based** methods
  - SNAPP, SVDquartets

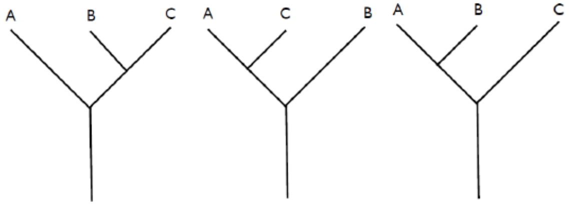
Concatenation

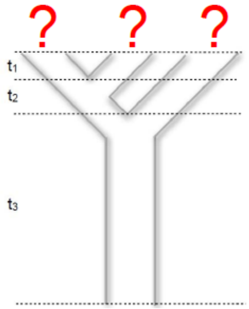
- put all the loci after each other (superalignment, supermatrix)
- very good accuracy under low ILS model conditions
- i.e., good approach unless strong ILS

- **single** partition model
  - the whole alignment analyzed with the same parameters
  - statistically inconsistent

- **multiple** partitions model (ML or Bayesian)
  - each alignment (or even codon position) analyzed with separate parameters
  - best partitioning scheme by, e.g., PartitionFinder or ModeltestNG or IQtree
  - fully partitioned analysis
  - maximum likelihood (CA-ML) – RAxML-ng, ExaML
  - or Bayesian inference – MrBayes, ExaBayes

Species tree estimation





- **concatenation**

- multispecies coalescence
  - *BEAST (**coestimation** of gene trees and species tree)
  - **summary** methods (combining gene trees)

- **supertree** methods
  - MRL (maximum representation using likelihood)

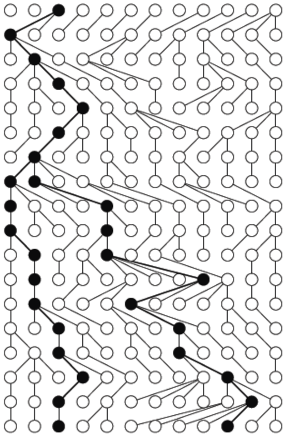
- Bayesian **concordance** analysis (BUCKy)
  - quartet-based Bayesian species tree estimation

- **site-based** methods
  - SNAPP, SVDquartets

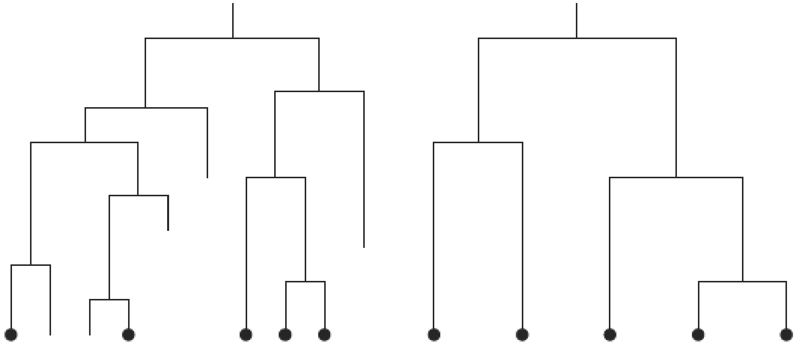
Multispecies coalescent

- coalescent model applied to gene trees in a species tree
  - combines coalescent and birth-death models

coalescence



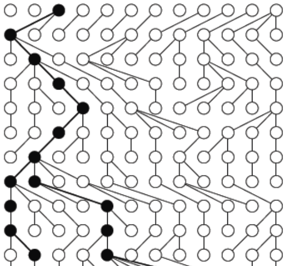
birth-death model



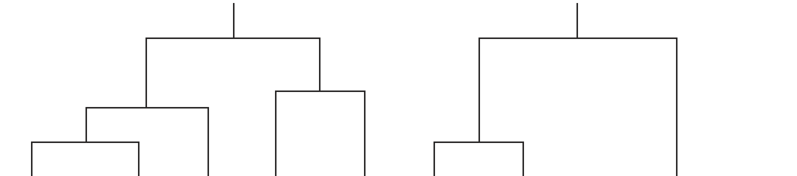
Multispecies coalescent

- used to assemble separate coalescent processes occurring in populations connected by an evolutionary tree
  - coalescent tree distribution (probability of sharing common ancestor t generations back)
  - birth-death model with stochastic rate of birth and death
  - describes probability of gene tree(s) within a species tree

coalescence

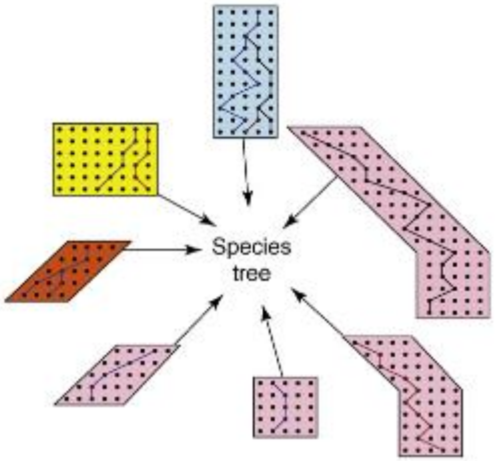


birth-death model



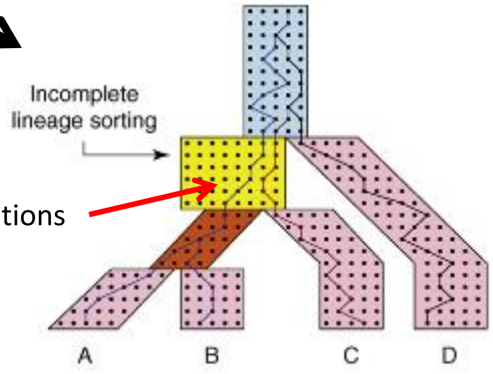
Multispecies coalescent

dot	individual gene copy
row	generation
lines	connection to gene ancestors in previous generations

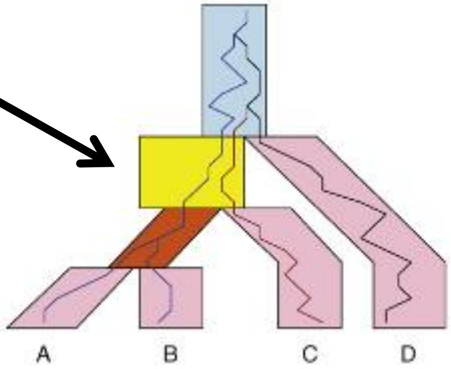


the coalescent in several populations
pink pops – only one lineage (gene copy)

populations arranged by evolutionary relationships



fail to coalesce within populations

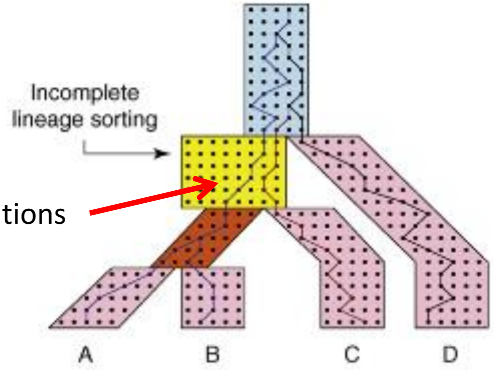


gene tree ((AB)(CD)) in a species tree ((AB)C)D)

Degnan & Rosenberg, 2009

Multispecies coalescent

- (incomplete) lineage sorting
  - particular types of genealogical pattern
  - process explaining gene tree discordance
  - failure of lineages in a population to coalesce



fail to coalesce within populations

*BEAST

STAR-BEAST = **S**pecies **T**ree **A**ncestral **R**econstruction

- Bayesian framework for species tree reconstruction
- assumptions
  - no recombination within locus
  - free recombination between loci
  - no hybridization
  - each sample mapped to appropriate species

prior distribution

likelihood

posterior distribution

$$f(\theta|D) = \frac{\Pr(D|\theta)f(\theta)}{\Pr(D)}$$

marginal likelihood

original Bayesian theorem

gene tree likelihood

probability of the species tree S given the data (D)

prior on species tree

multispecies coalescent likelihood (prior on gene tree given species tree)

$$f(\mathrm{g}, S|D) = \frac{f(S)}{\Pr(D)} \prod_{i=1}^{m} \Pr(D_i|g_i) f(g_i|S).$$

gene tree

marginal likelihood

Drummond & Bouckaert, 2015

*BEAST

STAR-BEAST = **S**pecies **T**ree **A**ncestral **R**econstruction

- co-estimates gene trees and species tree
- most accurate species tree method
- computationally intensive

- not suitable for large datasets, i.e.
  - no more than ~50 loci
  - no more than ~20-30 species

- BBCA – divide-and-conquer technique (Zimmerman et al., 2014)

Summary methods

Estimate each gene tree independently -> summarize them

require rooted gene trees

- MP-EST – **m**aximum **p**seudo-likelihood approach for **e**stimating **s**pecies **t**rees
- STAR – **s**pecies **t**ree estimation using **a**verage **r**anks of coalescences

unrooted gene trees

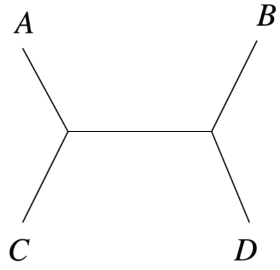
- STEAC – **s**pecies **t**ree **e**stimation using **a**verage **c**oalescence times
- ASTRAL – **A**ccurate **S**pecies **T**ree **R**econstruction **AL**gorithm
- ASTRID – **A**ccurate **S**pecies **TR**ees from **I**nternode **D**istances (reimplementation of NJ_{st} method)

site-based methods (estimate species trees from the distribution on site pattern within unlinked loci)

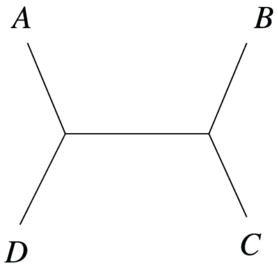
- SNAPP – SNP and AFLP Package for Phylogenetic analysis
- SVDquartets

Tree reconstruction from quartets

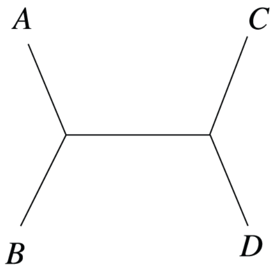
- quartet – unrooted tree over 4 taxa
- three possible quartets
- only one quartet *q* is consistent with final tree **T**



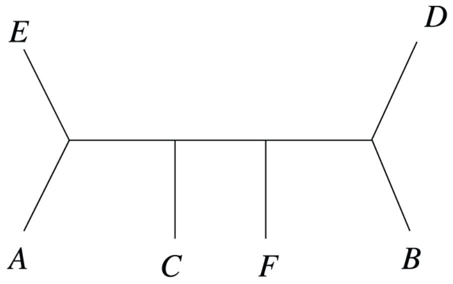
q_1



q_2



q_3

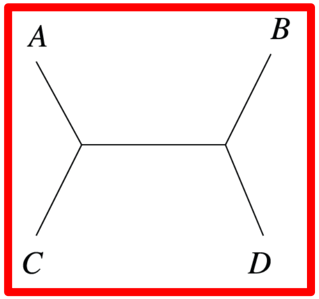


T

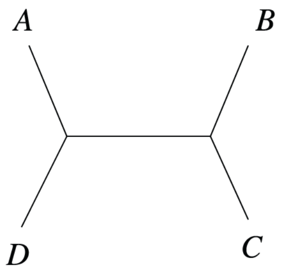
Reaz et al. (2015): *Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach*. PLoS ONE 9, e104008.

Tree reconstruction from quartets

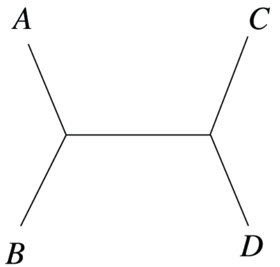
- quartet – unrooted tree over 4 taxa
- three possible quartets
- only one quartet *q* is consistent with final tree **T**



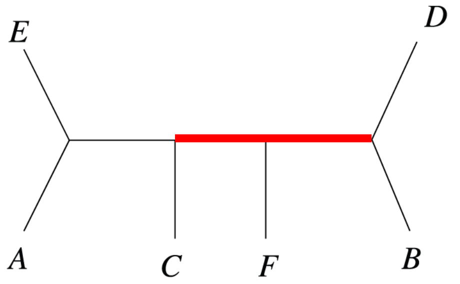
q_1



q_2



q_3



T

Reaz et al. (2015): *Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach*. PLoS ONE 9, e104008.

ASTRAL

Accurate **S**pecies **T**ree **R**econstruction **Al**gorithm
https://github.com/smirarab/ASTRAL

- 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

Siavash Mirarab

Unrooted quartets under MSC model

- **for a quartet (4 species) –** the most probable unrooted quartet tree (among the gene trees) is the unrooted species tree topology

- **for 5 or more species –** the unrooted species tree topology can be different from the most probable gene tree (called “anomaly zone”)
  - break gene trees into quartets of species
  - find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees (NP-hard optimization problem)
  - statistically consistent under the multispecies coalescent model with error-free input
  - solved by dynamic programming – ASTRAL

http://tandy.cs.illinois.edu/astral-apro.pdf

ASTRAL input/output

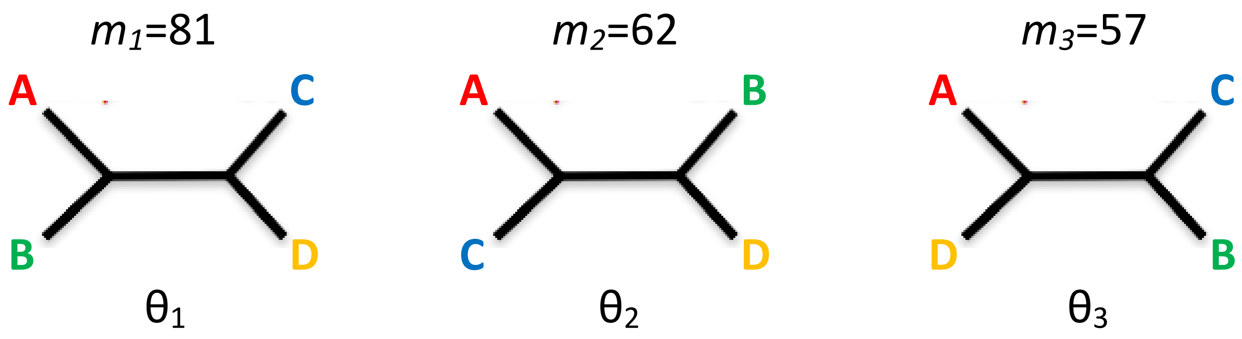
- input – unrooted gene trees
  - missing data allowed
  - polytomies allowed
  - multiple alleles per species allowed

- output – estimated unrooted species tree
  - **branch lengths** in coalescent units (on internal branches)
  - measure **of branch support** (LPP, local posterior probability)

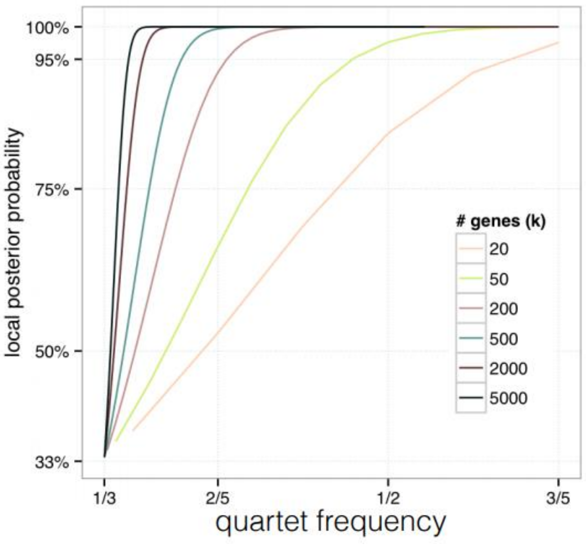
http://tandy.cs.illinois.edu/astral-apro.pdf

Local posterior probability

- quartet frequencies follow a multinomial distribution



- ***P*** (gene tree seen m_1/m times = species tree) = ***P***(θ_1 > 1/3)
  - possible to solve analytically
  - resulting measure is localPP
  - for n>4 – averaging quartet scores
- more accurate and faster than multi-locus bootstrap (MLBS; Seo 2008)

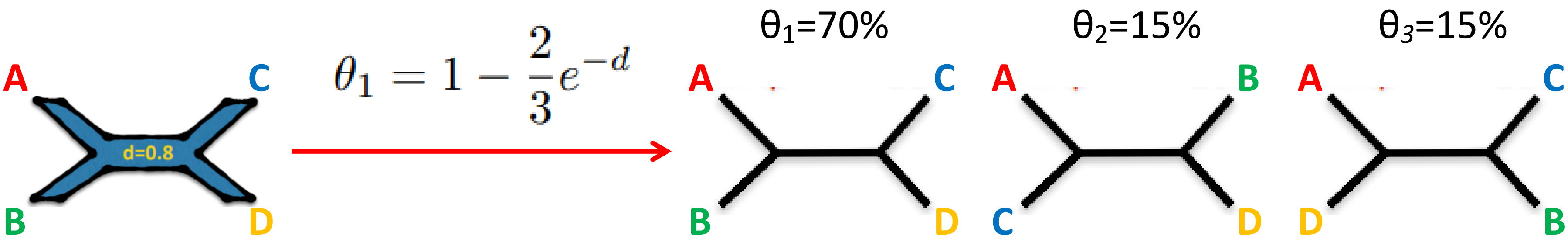


increased number of genes = increased support
decreased discordance = increased support

http://tandy.cs.illinois.edu/astral-apro.pdf

Branch length of ASTRAL trees

- branch length in coalescence units = the level of discordance
- for a single quartet (i.e., *n*=4) – reverse the discordance formula to get multilocus estimate
- for n>4 – average frequencies around the branch



$$\theta_1 = 1 - \frac{2}{3}e^{-d}$$

Sayyari & Mirarab, 2016, MBE

http://tandy.cs.illinois.edu/astral-apro.pdf

ASTRAL problems

- assumption for statistical consistency
  - randomly distributed sample of gene trees
    - recombination-free
    - reticulation-free
    - error-free (i.e., topology correctly estimated)
    - orthologous
- in practice: reduced accuracy with low accuracy gene trees
- branch length
  - only for internal branches (unless multiple individuals per species)
  - in coalescent units, i.e., “true value” is a function of population size and generation time
- local posterior probability (LPP)
  - better than MLBS (empirically) but based on many assumptions

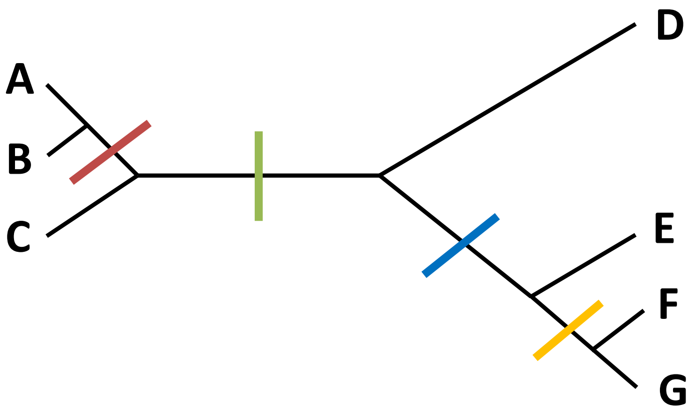
http://tandy.cs.illinois.edu/astral-apro.pdf

MRL

Maximum **R**epresentation with **L**ikelihood; Nguyen et al. 2012

- supertree method – 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
  - using mrp.jar; https://github.com/smirarab/mrpmatrix
- 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 ‘BINGAMMA’ model

MRL binary matrix



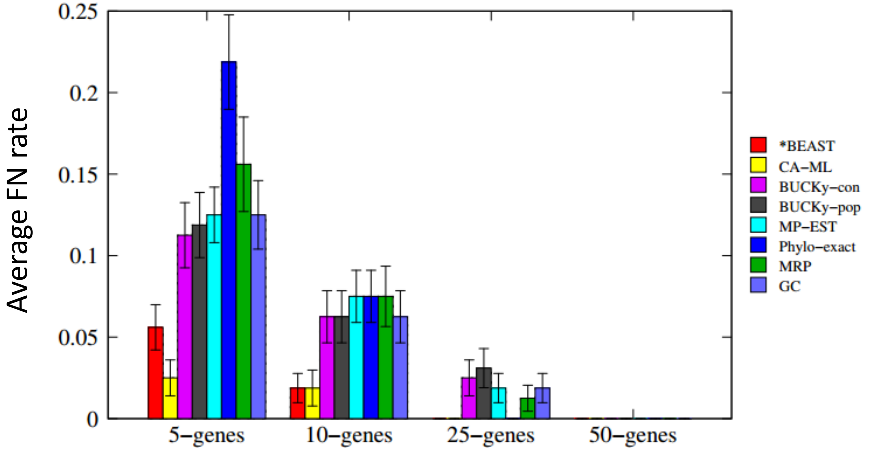
	*	*	*	*	…
A	1	1	1	1	
B	1	1	1	1	
C	0	1	1	1	
D	0	0	1	1	
E	0	0	0	1	
F	0	0	0	0	
G	0	0	0	0	
H	?	?	?	?	

randomization →

	*	*	*	*	…
A	0	1	1	0	
B	0	1	1	0	
C	1	1	1	0	
D	1	0	1	0	
E	1	0	0	0	
F	1	0	0	1	
G	1	0	0	1	
H	?	?	?	?	

Methods comparison

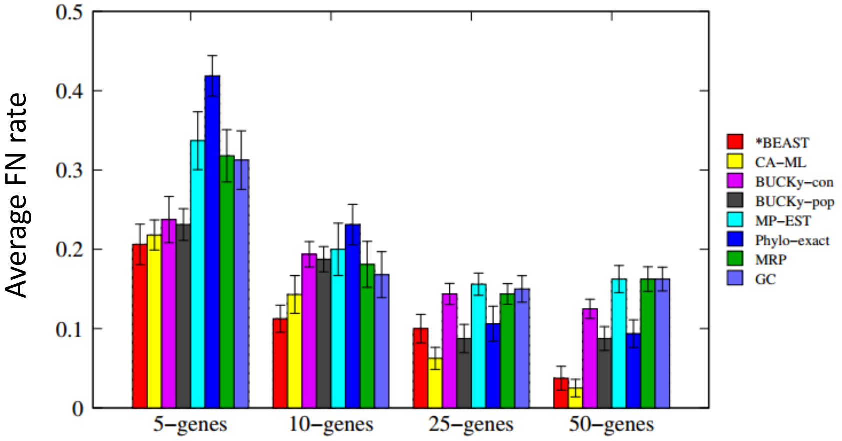
Results on 11-taxon datasets with weak ILS



*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

Quartet Sampling (QS)

Replacement for bootstrap in phylogenomic studies…

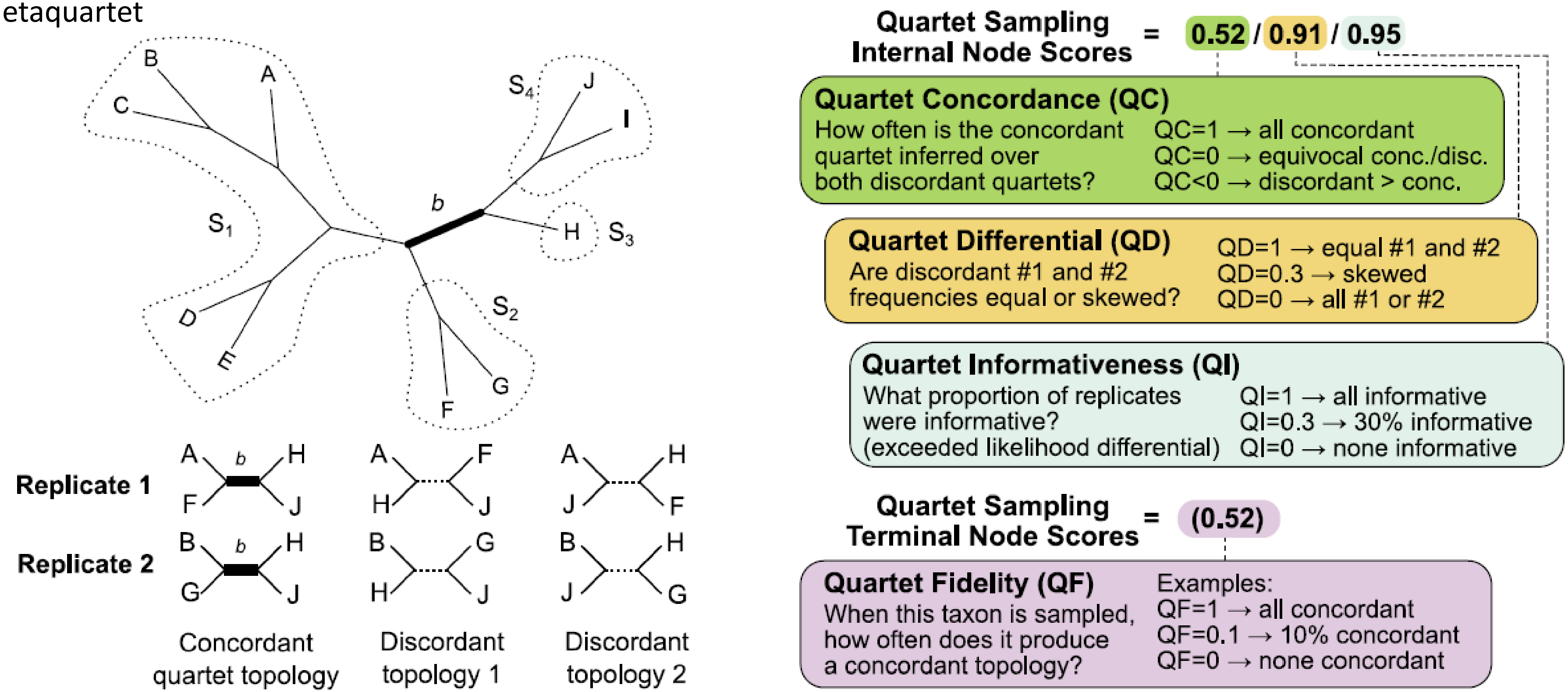
- quartet-based evaluation system
- synthetizes several phylogenetic and genomic analytical approaches
- discordance testing
- distinguishes strong conflict from weak support
- three different scores per branch
  - Quartet Concordance (QC)
  - Quartet Differential (QD)
  - Quartet Informativeness (QI)
- terminal node score
  - Quartet Fidelity (QF)

Pease et al. (2018): Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany* 105(3): 385–403.

Quartet Sampling

- takes an existing phylogenetic topology and a molecular dataset
- evaluates internal branches – likelihood for all three possible phylogenies for the randomly selected quartets spanning particular branch

metaquartet



Quartet Sampling Internal Node Scores = 0.52 / 0.91 / 0.95

Quartet Concordance (QC)
How often is the concordant quartet inferred over both discordant quartets?
QC=1 → all concordant
QC=0 → equivocal conc./disc.
QC<0 → discordant > conc.

Quartet Differential (QD)
Are discordant #1 and #2 frequencies equal or skewed?
QD=1 → equal #1 and #2
QD=0.3 → skewed
QD=0 → all #1 or #2

Quartet Informativeness (QI)
What proportion of replicates were informative? (exceeded likelihood differential)
QI=1 → all informative
QI=0.3 → 30% informative
QI=0 → none informative

Quartet Sampling Terminal Node Scores = (0.52)

Quartet Fidelity (QF)
When this taxon is sampled, how often does it produce a concordant topology?
Examples:
QF=1 → all concordant
QF=0.1 → 10% concordant
QF=0 → none concordant

Pease et al. (2018): Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany* 105(3): 385–403.

Quartet Sampling

Replacement for bootstrap in phylogenomic studies…

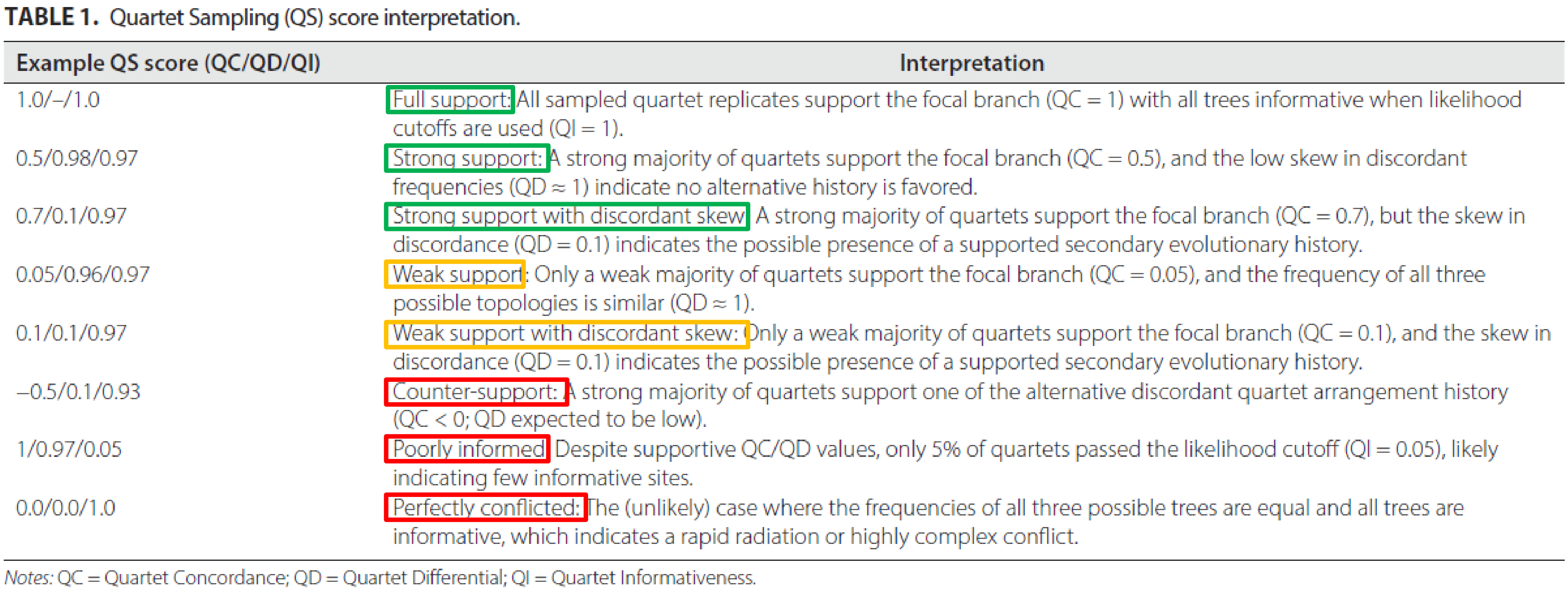


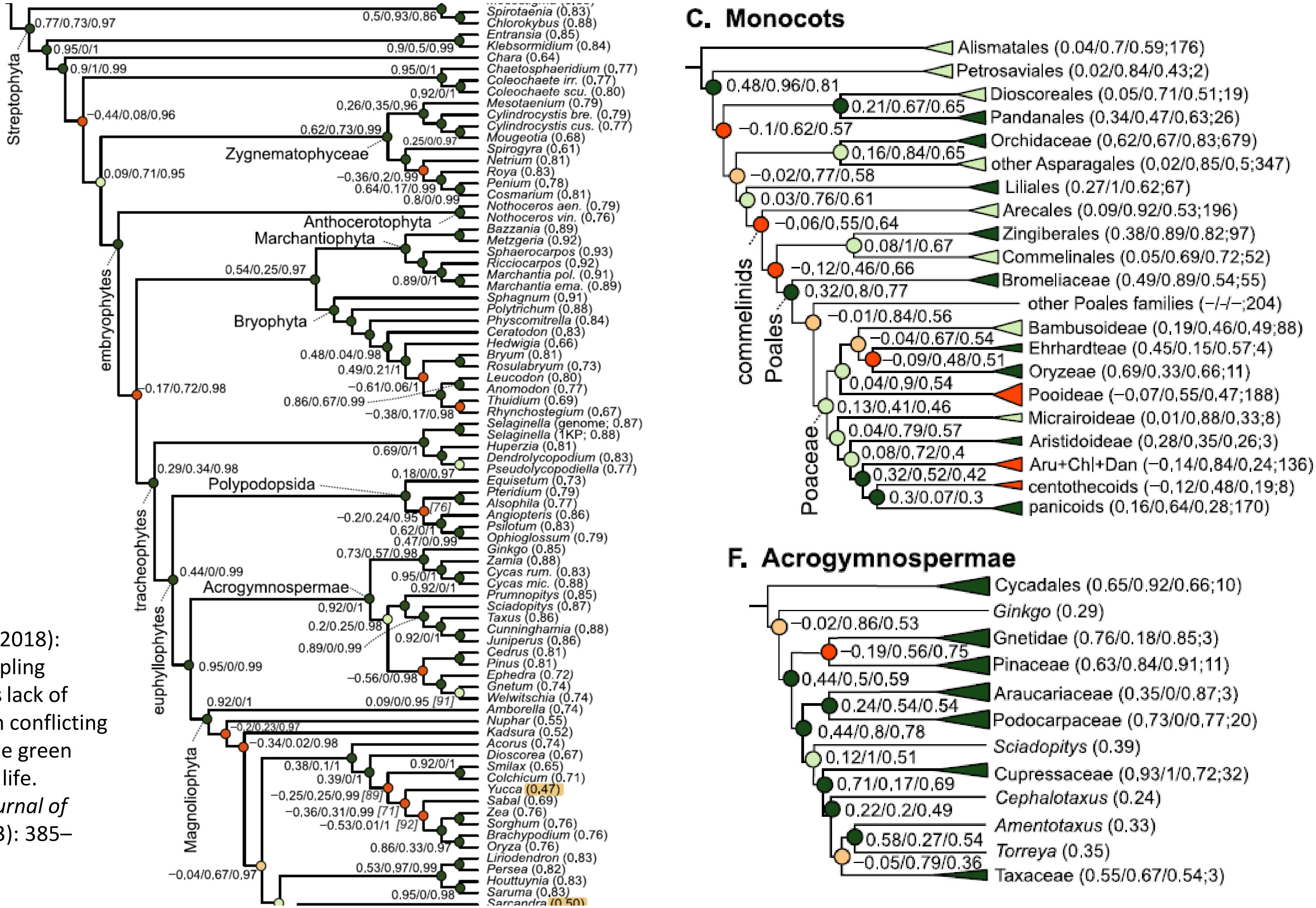
TABLE 1. Quartet Sampling (QS) score interpretation.

Example QS score (QC/QD/QI)	Interpretation
1.0/–/1.0	Full support: All sampled quartet replicates support the focal branch (QC = 1) with all trees informative when likelihood cutoffs are used (QI = 1).
0.5/0.98/0.97	Strong support: A strong majority of quartets support the focal branch (QC = 0.5), and the low skew in discordant frequencies (QD ≈ 1) indicate no alternative history is favored.
0.7/0.1/0.97	Strong support with discordant skew: A strong majority of quartets support the focal branch (QC = 0.7), but the skew in discordance (QD = 0.1) indicates the possible presence of a supported secondary evolutionary history.
0.05/0.96/0.97	Weak support: Only a weak majority of quartets support the focal branch (QC = 0.05), and the frequency of all three possible topologies is similar (QD ≈ 1).
0.1/0.1/0.97	Weak support with discordant skew: Only a weak majority of quartets support the focal branch (QC = 0.1), and the skew in discordance (QD = 0.1) indicates the possible presence of a supported secondary evolutionary history.
–0.5/0.1/0.93	Counter-support: A strong majority of quartets support one of the alternative discordant quartet arrangement history (QC < 0; QD expected to be low).
1/0.97/0.05	Poorly informed: Despite supportive QC/QD values, only 5% of quartets passed the likelihood cutoff (QI = 0.05), likely indicating few informative sites.
0.0/0.0/1.0	Perfectly conflicted: The (unlikely) case where the frequencies of all three possible trees are equal and all trees are informative, which indicates a rapid radiation or highly complex conflict.

Notes: QC = Quartet Concordance; QD = Quartet Differential; QI = Quartet Informativeness.

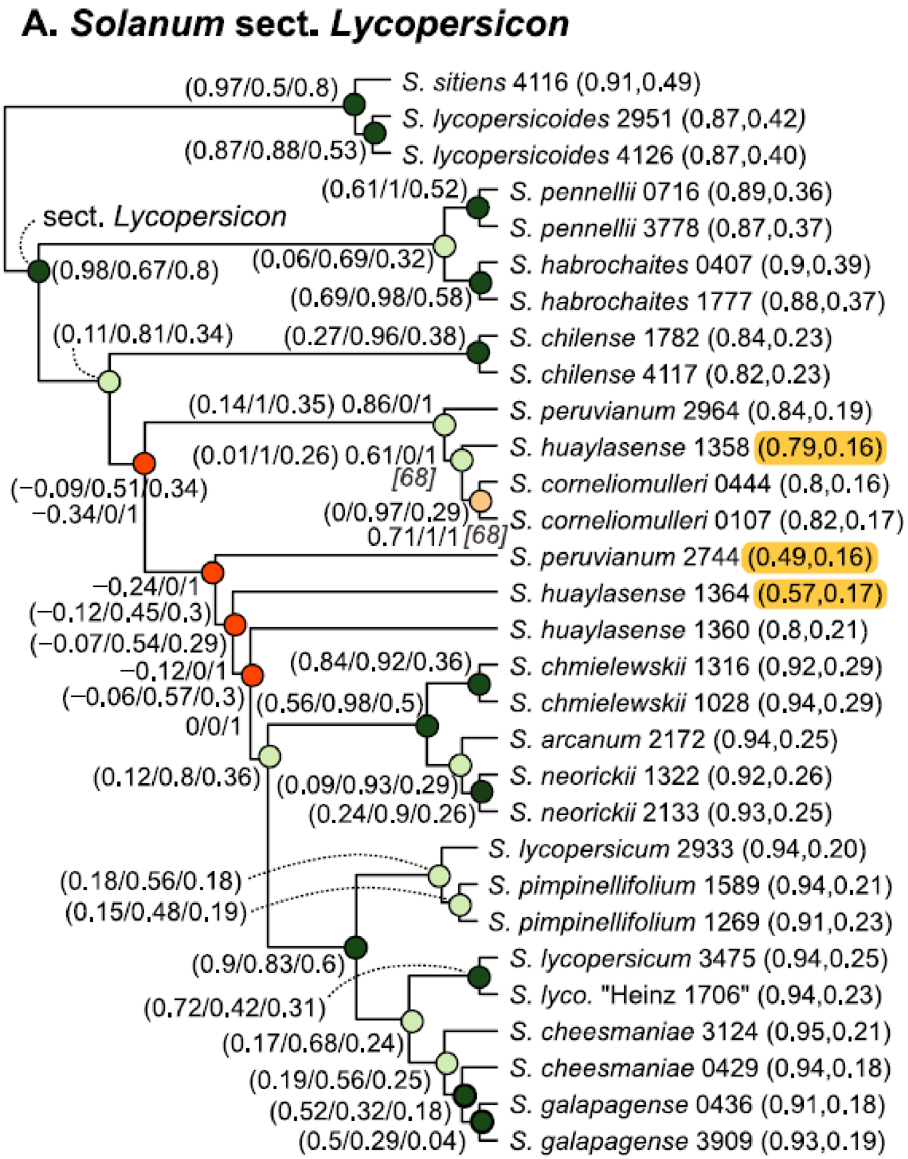
Pease et al. (2018): Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany* 105(3): 385–403.

Quartet Sampling – land plants



Pease et al. (2018): Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany* 105(3): 385–403.

Quartet Sampling – generic level


A. Solanum sect. Lycopersicon
(0.97/0.5/0.8)
S. sitiens 4116 (0.91,0.49)
S. lycopersicoides 2951 (0.87,0.42)
(0.87/0.88/0.53)
S. lycopersicoides 4126 (0.87,0.40)
(0.61/1/0.52)
S. pennellii 0716 (0.89,0.36)
sect. Lycopersicon
S. pennellii 3778 (0.87,0.37)
(0.06/0.69/0.32)
(0.98/0.67/0.8)
S. habrochaites 0407 (0.9,0.39)
(0.69/0.98/0.58)
S. habrochaites 1777 (0.88,0.37)
(0.11/0.81/0.34)
(0.27/0.96/0.38)
S. chilense 1782 (0.84,0.23)
S. chilense 4117 (0.82,0.23)
(0.14/1/0.35) 0.86/0/1
S. peruvianum 2964 (0.84,0.19)
(0.01/1/0.26) 0.61/0/1
S. huaylasense 1358 (0.79,0.16)
[68]
(−0.09/0.51/0.34)
−0.34/0/1
S. corneliomulleri 0444 (0.8,0.16)
(0/0.97/0.29)
S. corneliomulleri 0107 (0.82,0.17)
0.71/1/1 [68]
S. peruvianum 2744 (0.49,0.16)
−0.24/0/1
(−0.12/0.45/0.3)
S. huaylasense 1364 (0.57,0.17)
(−0.07/0.54/0.29)
S. huaylasense 1360 (0.8,0.21)
−0.12/0/1
(0.84/0.92/0.36)
S. chmielewskii 1316 (0.92,0.29)
(−0.06/0.57/0.3)
(0.56/0.98/0.5)
S. chmielewskii 1028 (0.94,0.29)
0/0/1
S. arcanum 2172 (0.94,0.25)
(0.12/0.8/0.36)
(0.09/0.93/0.29)
S. neorickii 1322 (0.92,0.26)
(0.24/0.9/0.26)
S. neorickii 2133 (0.93,0.25)
S. lycopersicum 2933 (0.94,0.20)
(0.18/0.56/0.18)
S. pimpinellifolium 1589 (0.94,0.21)
(0.15/0.48/0.19)
S. pimpinellifolium 1269 (0.91,0.23)
S. lycopersicum 3475 (0.94,0.25)
(0.9/0.83/0.6)
S. lyco. "Heinz 1706" (0.94,0.23)
(0.72/0.42/0.31)
S. cheesmaniae 3124 (0.95,0.21)
(0.17/0.68/0.24)
S. cheesmaniae 0429 (0.94,0.18)
(0.19/0.56/0.25)
S. galapagense 0436 (0.91,0.18)
(0.52/0.32/0.18)
(0.5/0.29/0.04)
S. galapagense 3909 (0.93,0.19)

Pease et al. (2016)

References

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.

Drummond, A.J. & Bouckaert, R.R., 2015. Bayesian Evolutionary Analysis with BEAST. Cambridge University Press.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., S. Swenson, M., Warnow, T., 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548

Sayyari, E., Mirarab, S., 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Mol. Biol. Evol. 33, 1654–68.

ASTRAL presentation: http://tandy.cs.illinois.edu/astral-apro.pdf

Nguyen, N., Mirarab, S., Warnow, T., 2012. MRL and SuperFine+MRL: New supertree methods. Algorithms Mol. Biol. 7, 3.

Pease, J.B. et al., 2018. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. American Journal of Botany 105, 385–403.