Gene trees vs species tree

Tomáš Fér

2025

some parts based on

http://tandy.cs.illinois.edu/astral-tutorial.pdf (Siavash Mirarab)

http://tandy.cs.illinois.edu/astrid-ssb-v2.pdf (Pranjal Vachaspati)

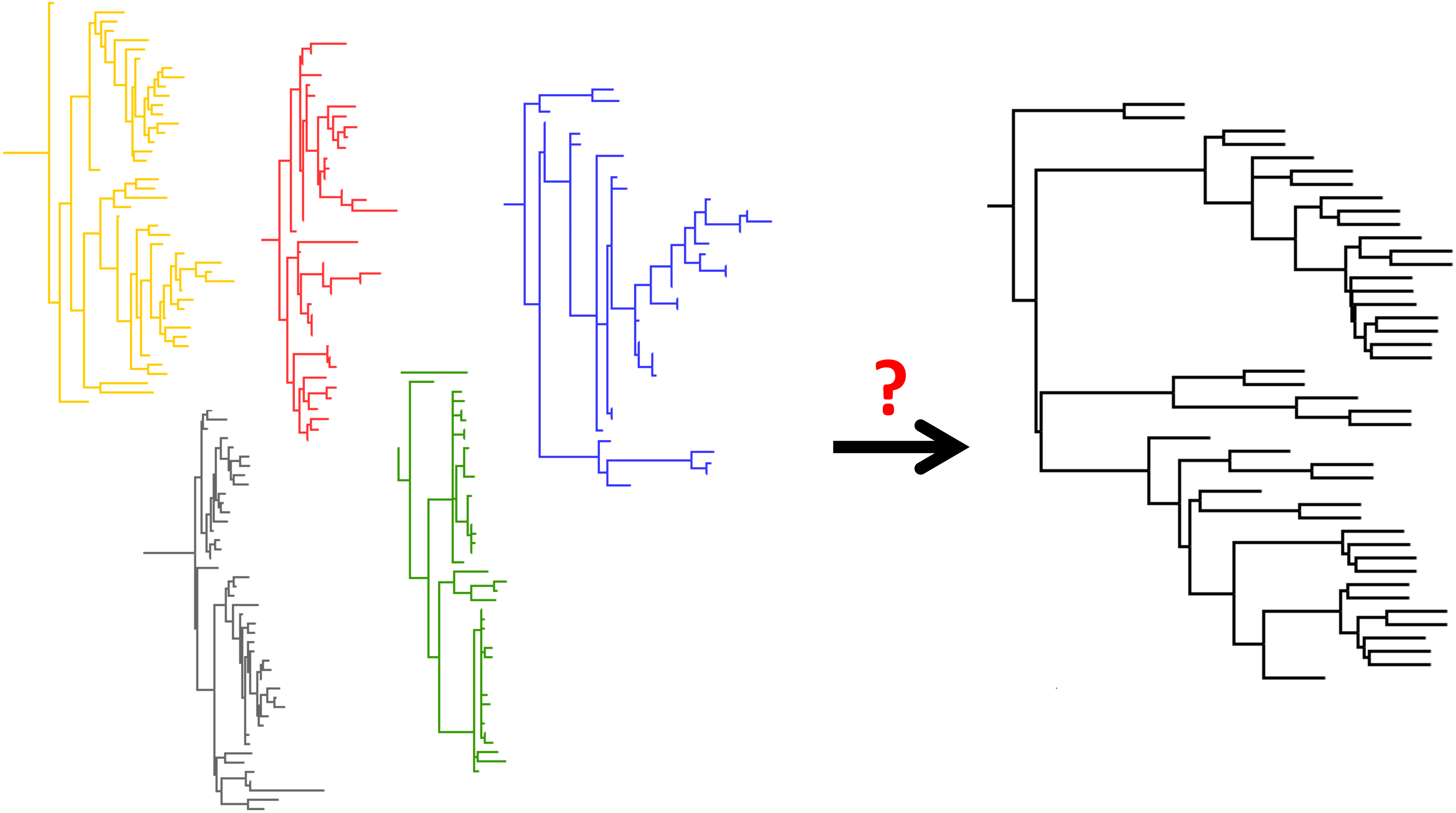
Outline

- gene trees and gene tree-species tree incongruence
- species tree estimation approaches
- multispecies coalescence (MSC)
  - coestimation (*BEAST)
  - summary methods
- ASTRAL
  - quartets under MSC
  - input/output
  - assumptions/problems
  - multiple individuals per species
  - filtering data
  - localPP vs. MLBS
  - branch lengths
  - tree scoring
  - polytomy testing
- ASTRID
  - principle, accuracy
- MRL
- Quartet Sampling
  - principle, applications
- Gene selection/filtering
  - calculating alignment/tree characteristics

Practicals

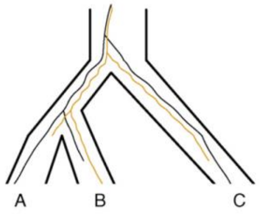
- ASTRAL
- ASTRID
- MRL
- Quartet Sampling
- alignment/tree characteristics (AMAS/R)

Species tree from gene trees

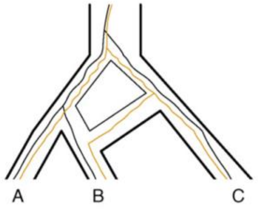


Gene tree incogruence

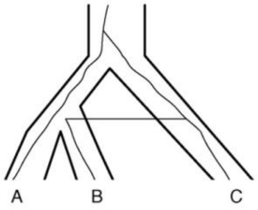
- incomplete lineage sorting (ILS)/deep coalescence
- gene duplications and losses (GDL)
  - orthology problem
- hybridization/(allo)polyploidization
  - affects whole genomes
- horizontal gene transfer (HGT)
  - affects small DNA segments
- recombination
  - different histories for neighboring segments in genes



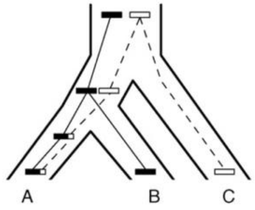
((AC)B)



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



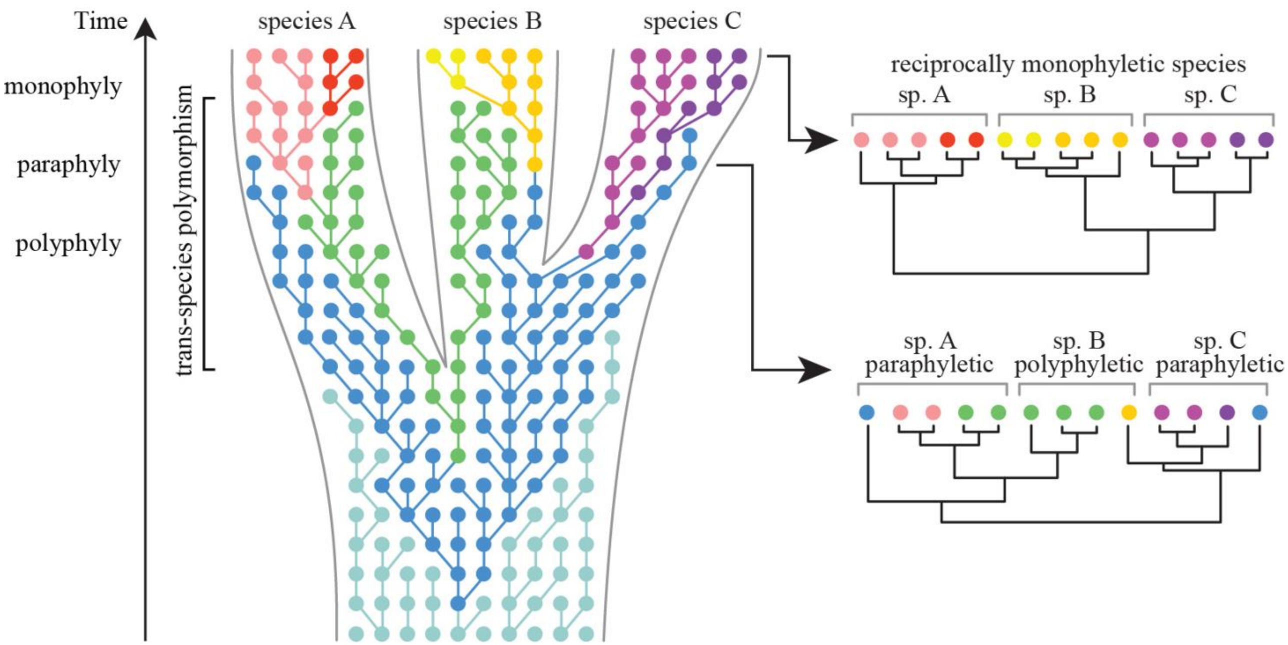
(A(BC))



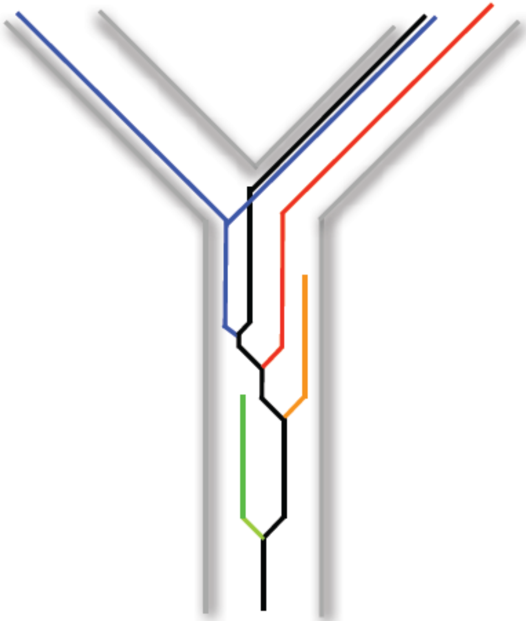
((AB)C)
((AC)B)

Degnan & Rosenberg, 2009

Coalescence processes

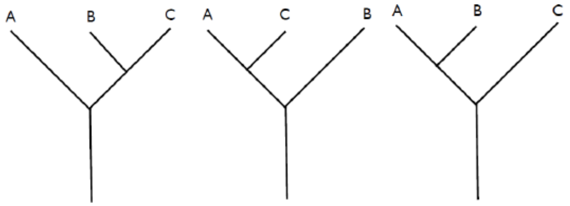


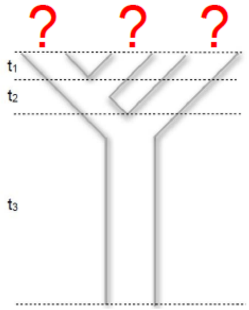
https://frederikleliaert.wordpress.com/green-algae/dna-based-species-delimitation-in-algae/



incomplete lineage sorting

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

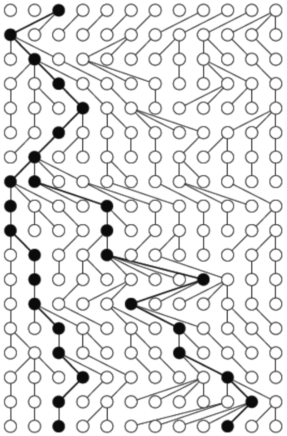
- **single** partition model
  - the whole alignment analyzed with the same parameters
  - statistically inconsistent

- **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

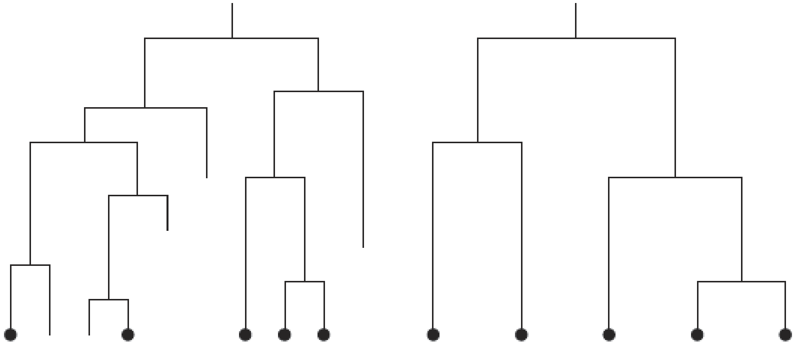
Multispecies coalescent

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

coalescence



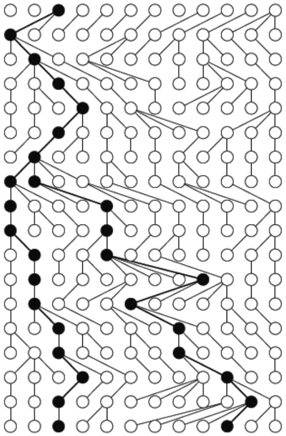
birth-death model



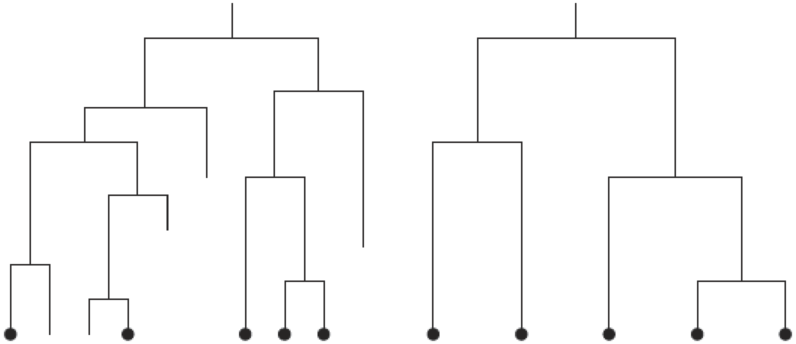
Multispecies coalescent

- 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

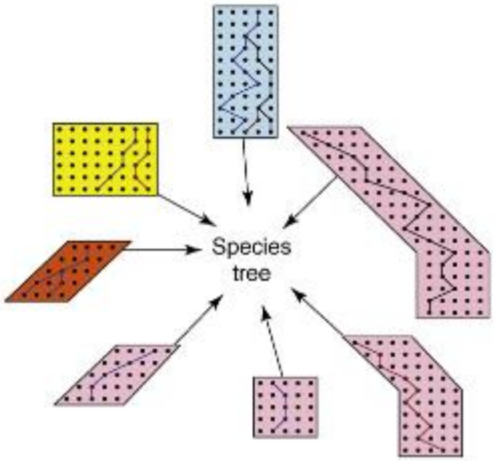


birth-death model



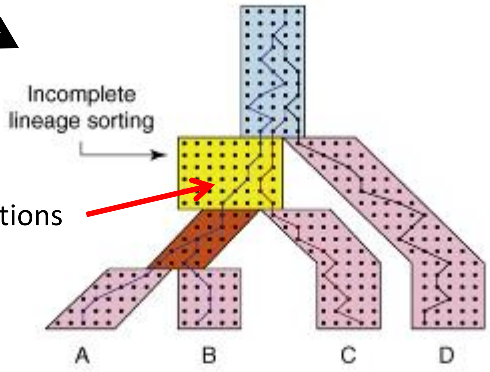
Multispecies coalescent

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

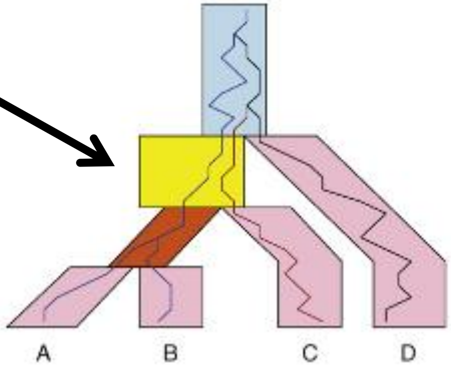


the coalescent in several populations
pink pops – only one lineage (gene copy)

populations arranged by evolutionary relationships



fail to coalesce within populations

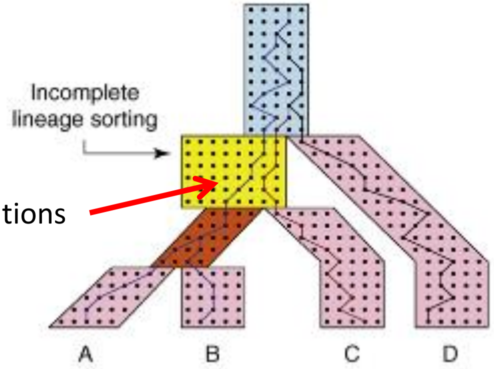


gene tree ((AB)(CD)) in a species tree ((AB)C)D)

Degnan & Rosenberg, 2009

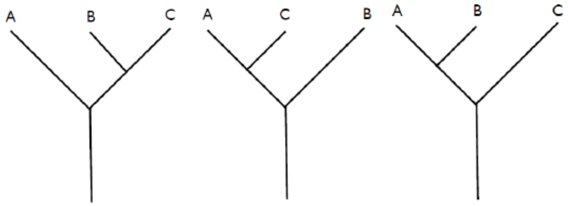
Multispecies coalescent

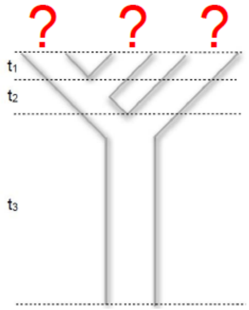
- (incomplete) lineage sorting
  - particular types of genealogical pattern
  - process explaining gene tree discordance
  - failure of lineages in a population to coalesce



fail to coalesce within populations

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

*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)

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

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 (Zimmermann et al., 2014)

Summary methods

Species tree estimation

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

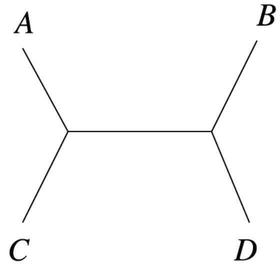
- STEAC – **s**pecies **t**ree **e**stimation using **a**verage **c**oalescence times
- ASTRAL – **A**ccurate **S**pecies **T**ree **R**econstruction **AL**gorithm
- ASTRID – **A**ccurate **S**pecies **TR**ees from **I**nternode **D**istances (reimplementation of NJ$_{st}$ method)

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

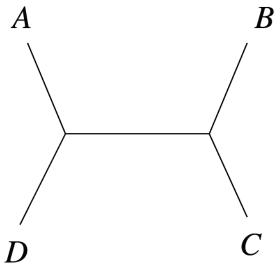
- SNAPP – SNP and AFLP Package for Phylogenetic analysis
- SVDquartets

Tree reconstruction from quartets

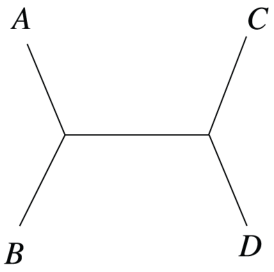
- quartet – unrooted tree over 4 taxa
- three possible quartets
- only one quartet *q* is consistent with final tree **T**



q_1

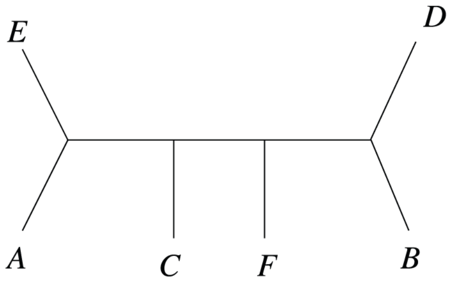


q_2



q_3

- which *quartet* is consistent with *T* ?

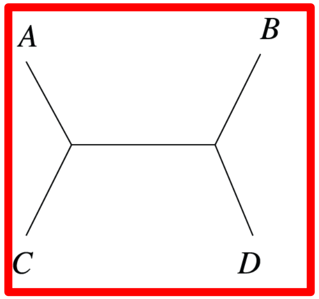


T

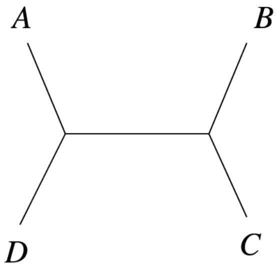
Reaz et al. (2015): *Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach*. PLoS ONE 9, e104008.

Tree reconstruction from quartets

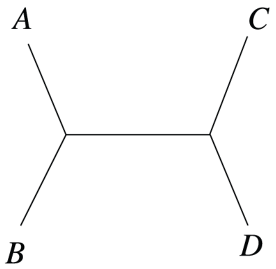
- quartet – unrooted tree over 4 taxa
- three possible quartets
- only one quartet *q* is consistent with final tree **T**



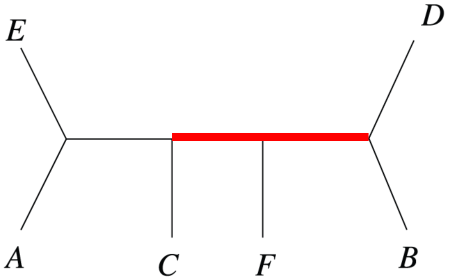
q_1



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 versions

- ASTRAL-I (<v. 4.3.7)
- ASTRAL-II (<v. 5.1.0)
  - improved the accuracy at the expense of running time
  - can handle polytomies
- ASTRAL-III (>v. 5.1.1)
  - changed the search space again for a better running time versus accuracy trade-off
  - improved running time for unresolved trees
- ASTRAL-IV (within ASTER)
  - scalable alternative
  - computes terminal and internal branch length in substitution-per-site units

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

ASTRAL problems

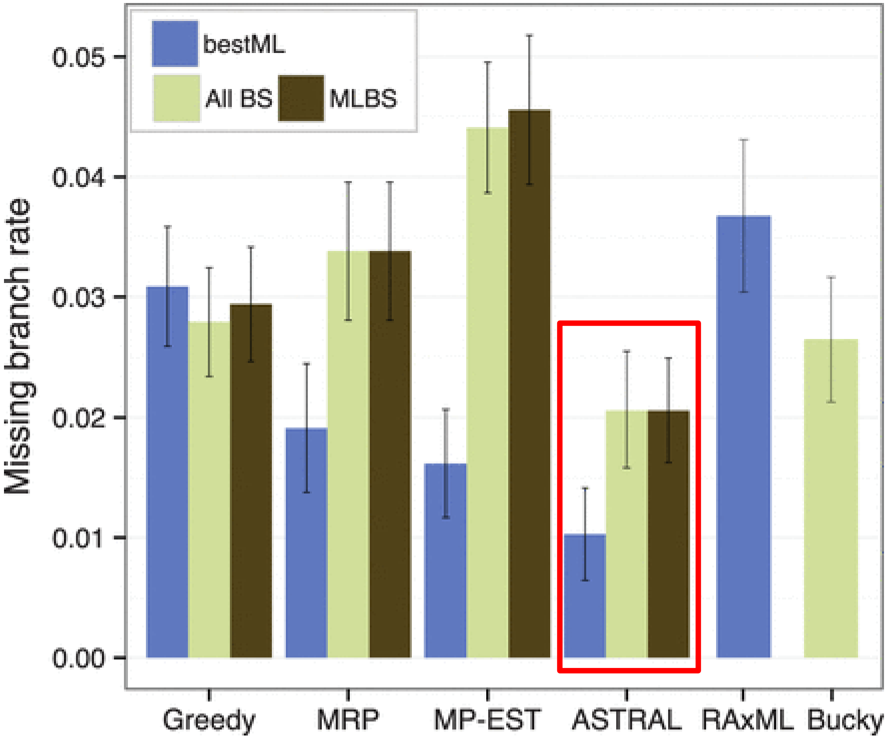
- assumption for statistical consistency
  - randomly distributed sample of gene trees
    - recombination-free
    - reticulation-free
    - error-free
    - orthologous

- in practice: reduced accuracy with low accuracy gene trees

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

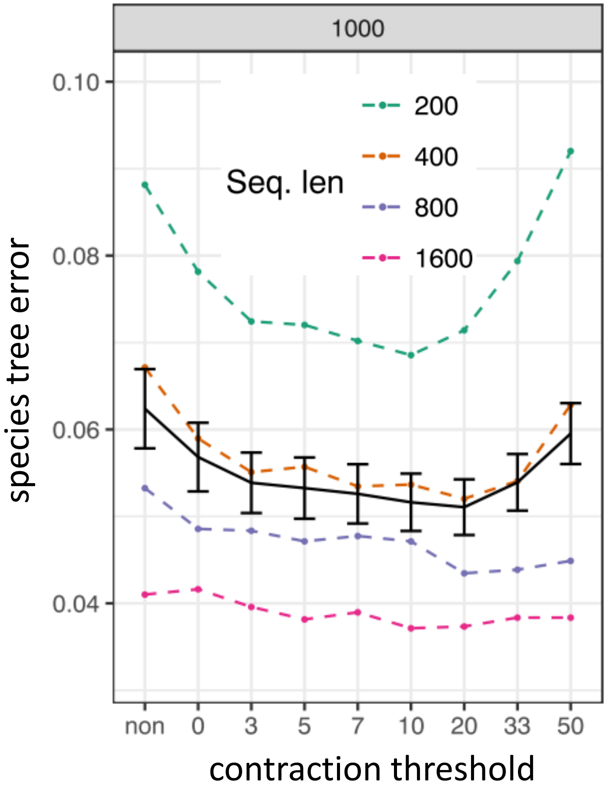
How input gene trees influence ASTRAL

BS and MLBS degrade accuracy compared to simple ML gene trees



Mirarab et al. 2014, Bioinformatics

contraction of branches with very low support helps

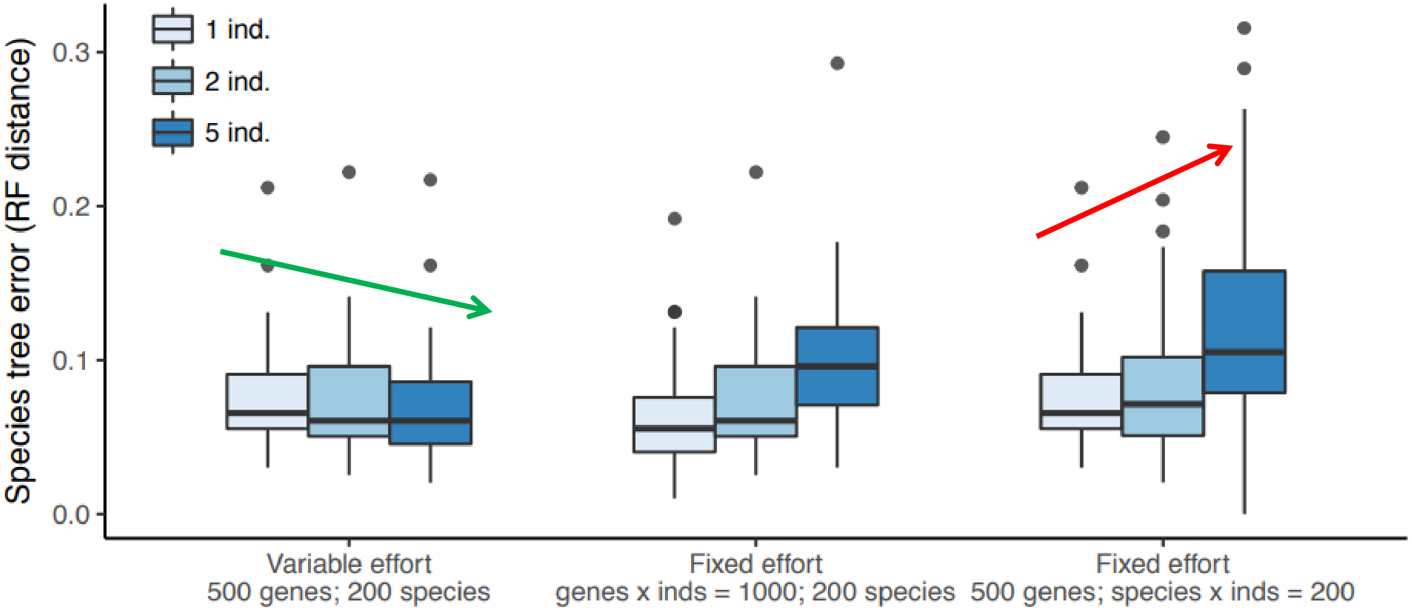


Zhang et al. 2018, BMC Bioinformatics

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

Multiple individuals per species

- sampling multiple individuals – extra signal
- individual can be non-monophyletic in gene trees (i.e., in recently diversified groups)



only helpful with variable (i.e., increased) sequencing effort, not with fixed (i.e., when ‘nr. genes × nr. individuals’ remains the same)

Rabiee, Sayyari, Mirarab 2019, MPE

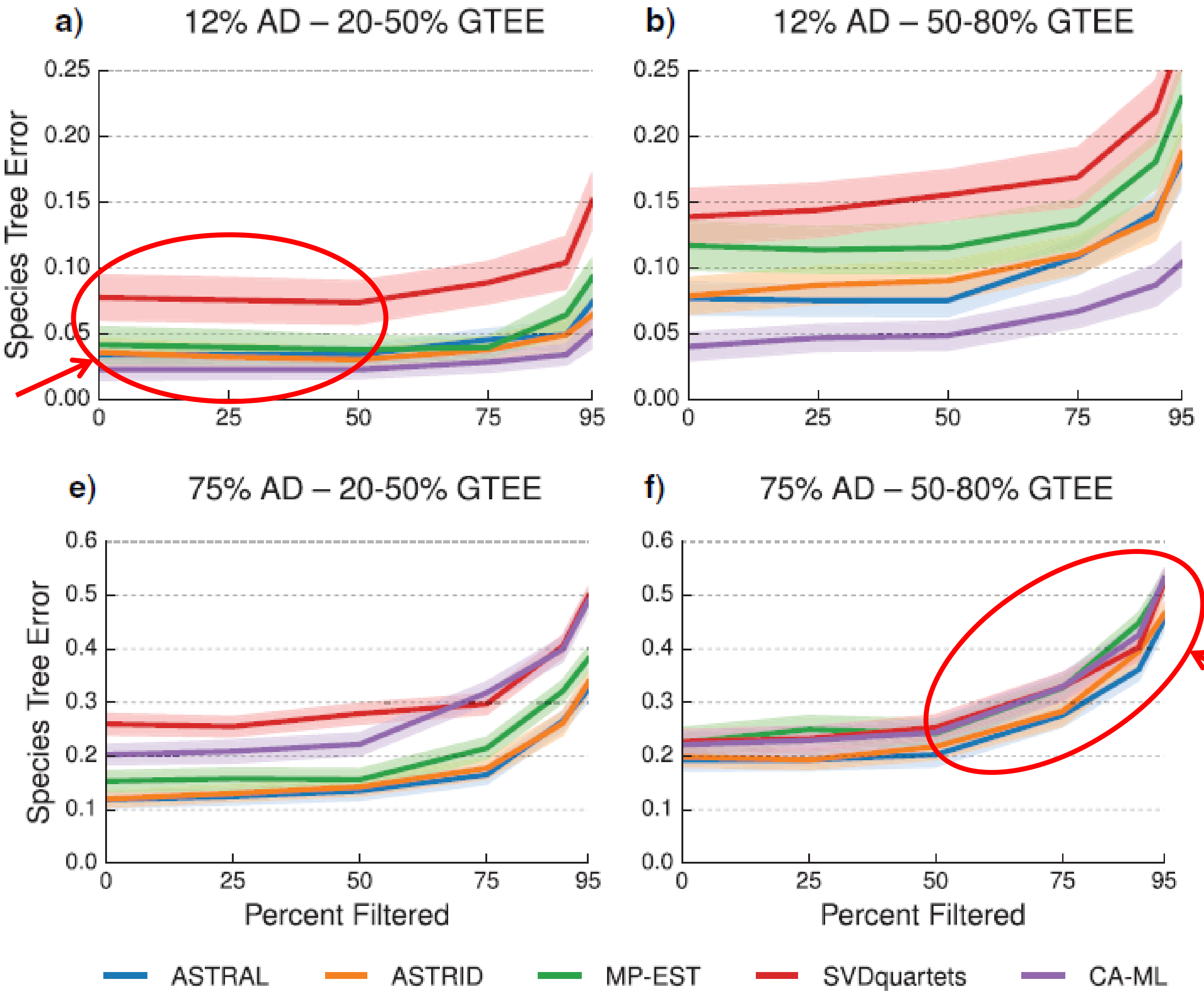
http://tandy.cs.illinois.edu/astral-apro.pdf

Filtering?

- genes based on missing species?
    - generally not beneficial (Moloy & Warnow 2018)
- genes based on gene tree estimation error (GTEE)
    - depends on condition (Moloy & Warnow 2018)
- filtering fragmentary sequences (and keeping gene)
    - often beneficial (Sayyari, Whitfield, Mirarab, 2017)

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

Gene filtering – missing species



low/moderate ILS

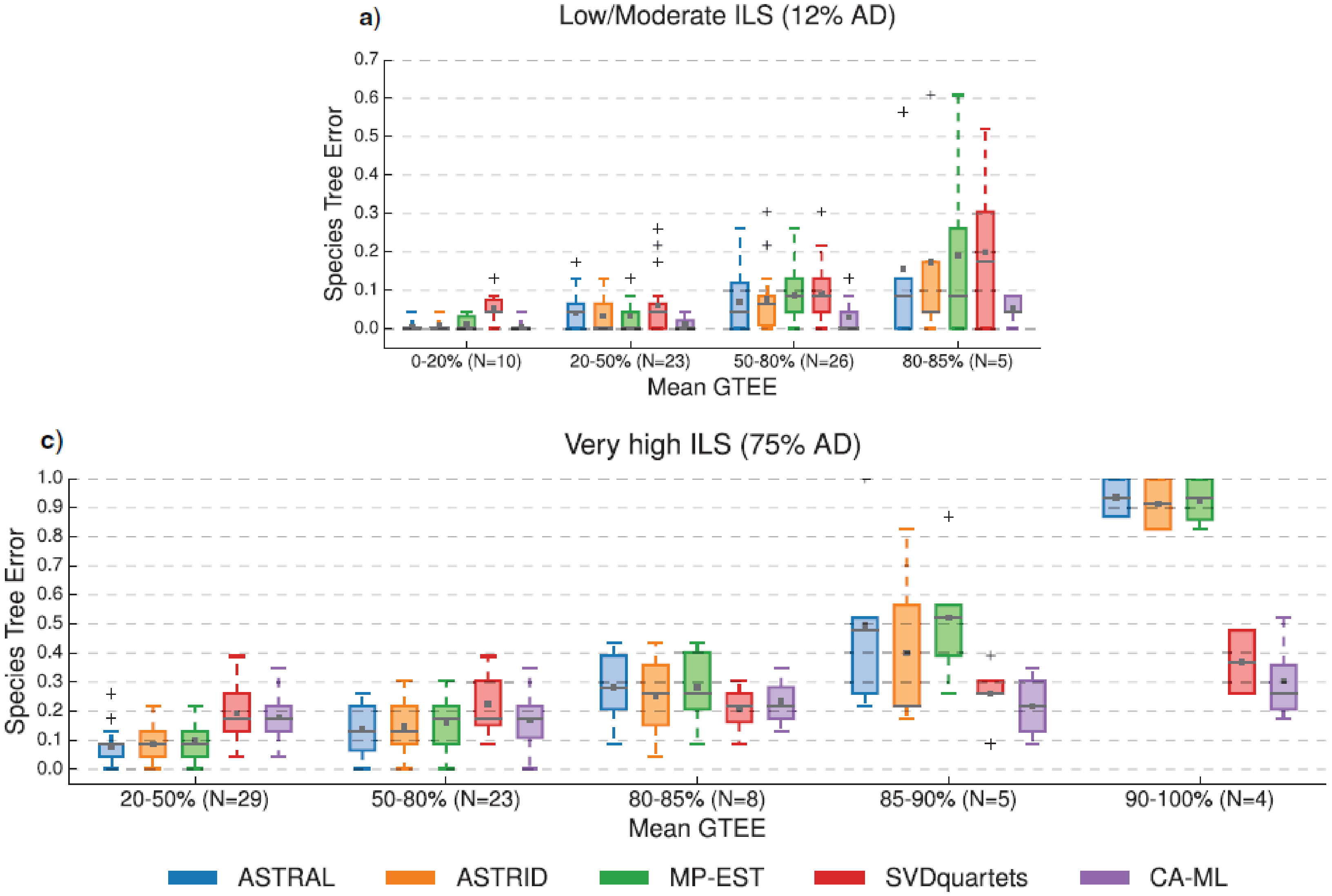
high ILS

not beneficial

or even worth

Moloy & Warnow, 2018, Syst. Biol.

Gene filtering – gene tree estimation error

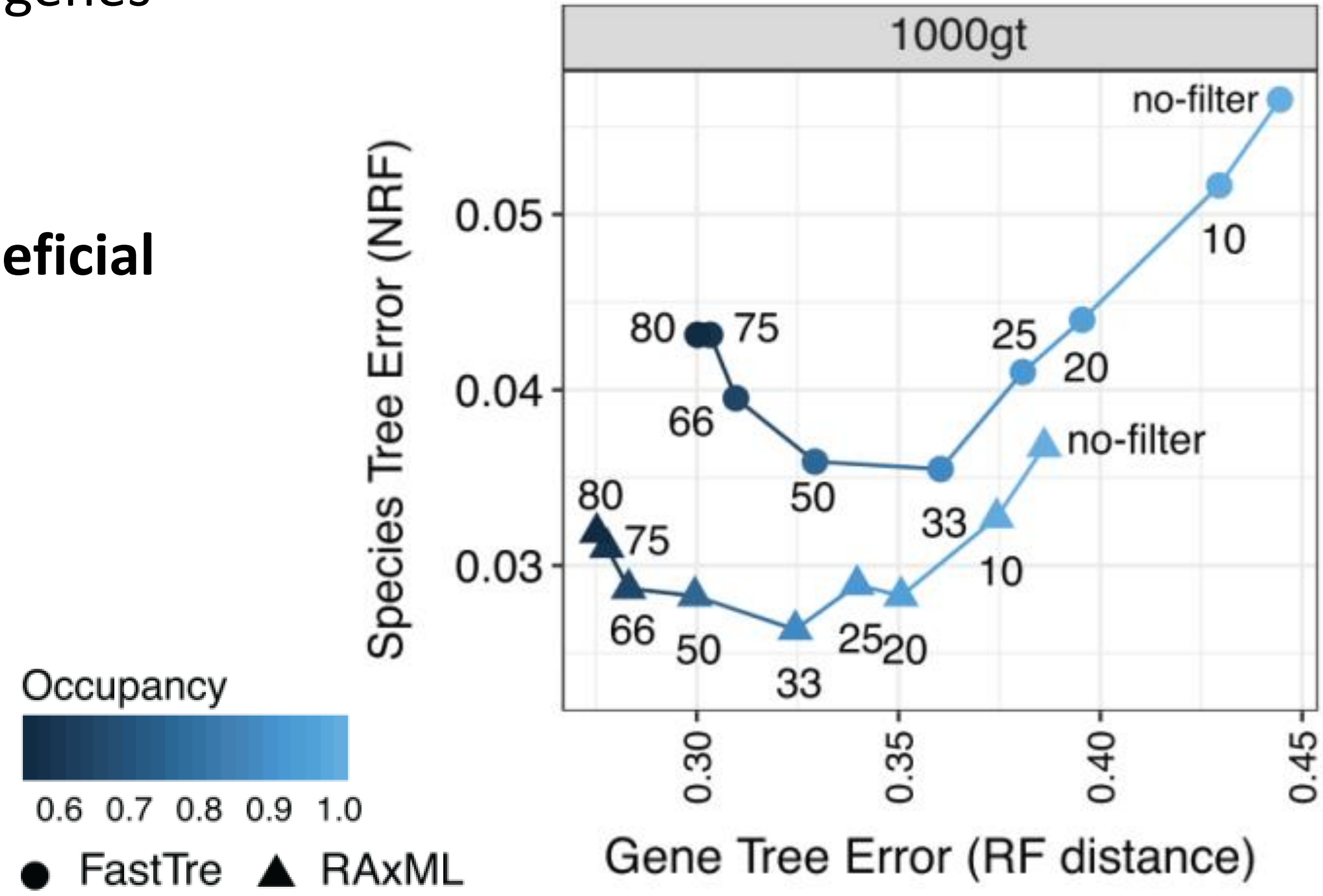


Moloy & Warnow, 2018, Syst. Biol.

Filtering fragmentary sequences

- remove fragmentary data from genes
- BUT keep genes

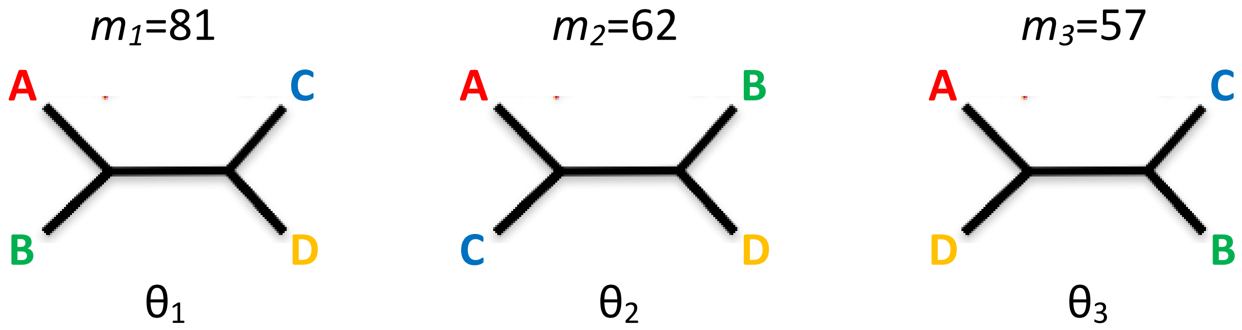
- **often beneficial**



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

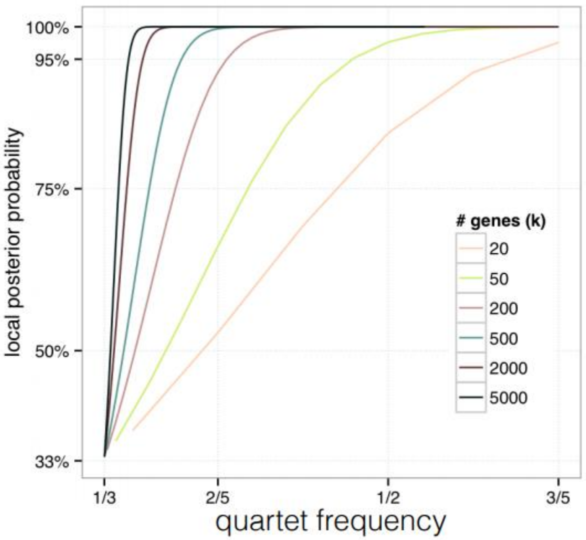
Local posterior probability

- quartet frequencies follow a multinomial distribution



- ***P*** (gene tree seen m_1/m times = species tree) = ***P***($\theta_1 > 1/3$)
  - 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

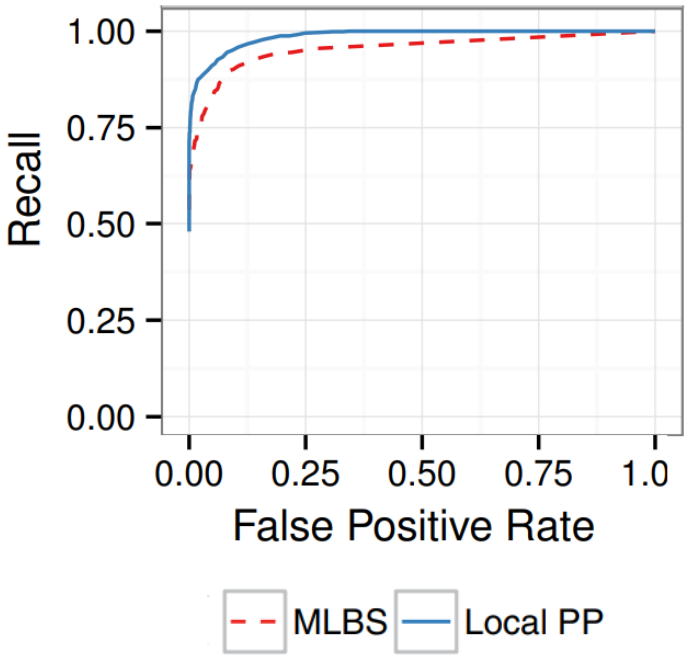


http://tandy.cs.illinois.edu/astral-apro.pdf

localPP vs. MLBS

- LPP = local posterior probability
- MLBS = multi-locus bootstrap (Seo, 2008)
  - requires bootstrap replicate trees for each gene

- localPP is more accurate than bootstrapping
- calculating localPP is also ca. 100× faster

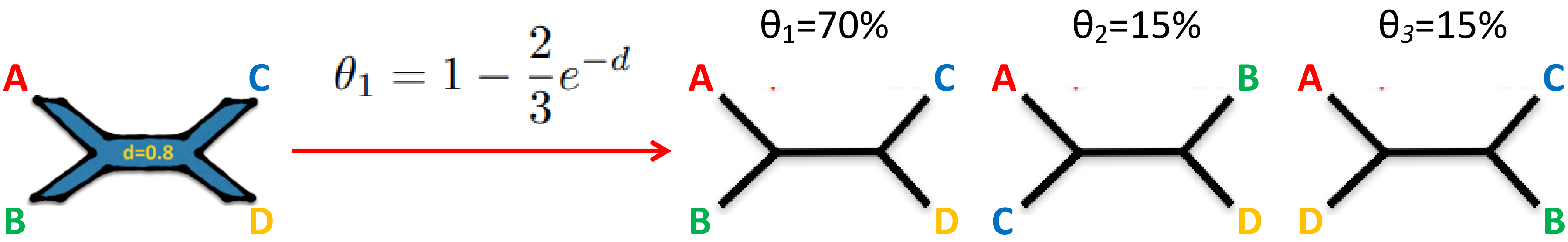


http://tandy.cs.illinois.edu/astral-apro.pdf

Sayyari & Mirarab, 2016, MBE

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

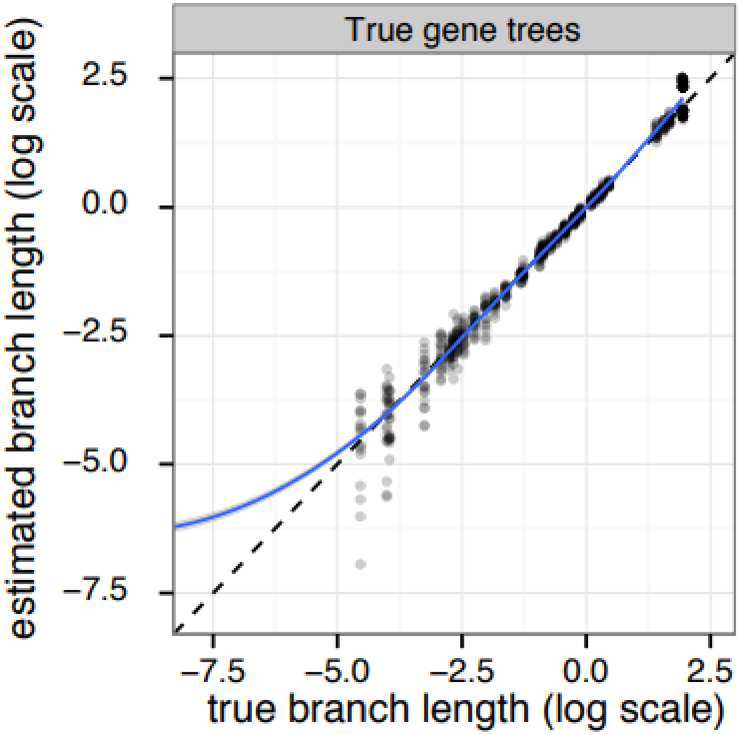


$$\theta_1 = 1 - \frac{2}{3}e^{-d}$$

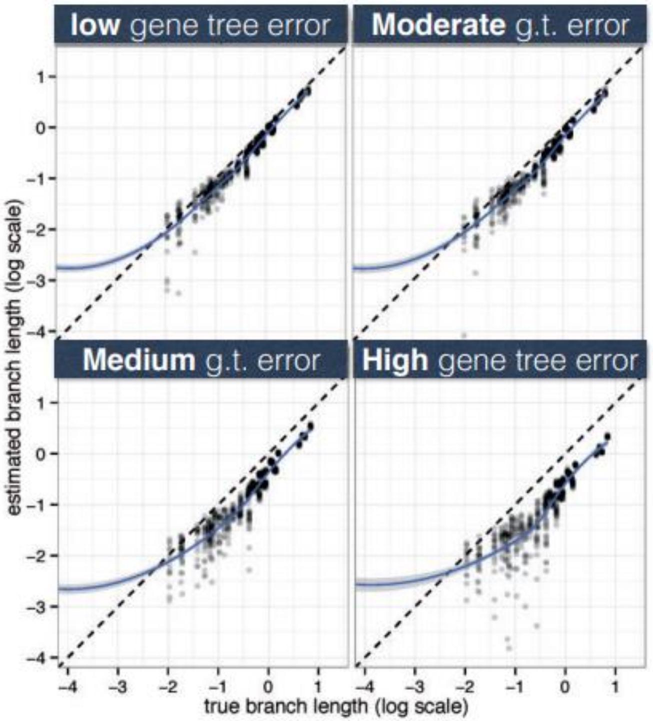
Sayyari & Mirarab, 2016, MBE

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

Branch lengths (BL) accurracy



correct BL estimate with *true* gene trees



underestimated BL estimate with *error-prone* gene trees

Sayyari & Mirarab, 2016, MBE

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

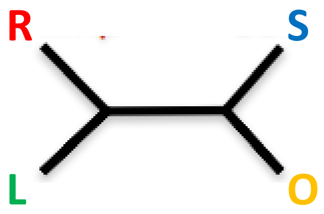
ASTRAL problems

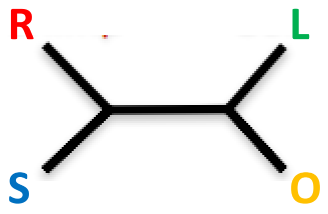
- assumption for statistical consistency
  - randomly distributed sample of gene trees
    - recombination-free
    - reticulation-free
    - error-free
    - orthologous
- in practice: reduced accuracy with low accuracy gene trees
- branch length
  - only for internal branches (unless multiple individuals 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

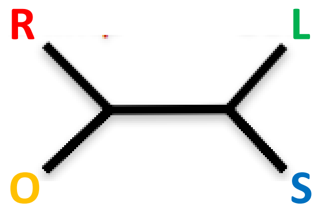
http://tandy.cs.illinois.edu/astral-apro.pdf

ASTRAL tree scoring

- *quartet support* (-t 1) – percentage of quartets that agrees with the branch (measuring the amount of gene tree conflict)
- *alternative posteriors* (-t 4) – three localPP: (1) main topology (RL|SO), (2) first alternative (RS|LO), (3) second alternative (RO|LS)







- *full annotation* (-t 2)
- *alternative quartet topologies* (-t 8) – quartet support for the main and alternative topologies
- *polytomy test* (-t 10) – runs a test if the hypothesis (branch is a polytomy) could be rejected

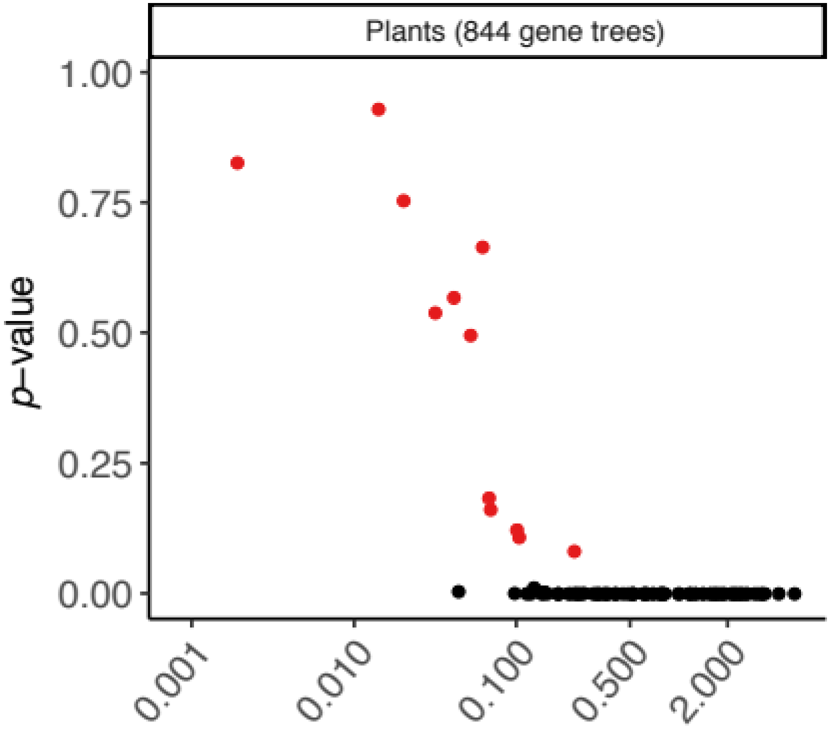
Hard polytomy test

- soft polytomy – unresolved relationships in an estimated tree
- hard polytomy – multifurcation in the true tree

- null hypothesis – branch has zero length and should be removed to create a polytomy
- try to reject null hypothesis
  - reject (i.e., $p<0.05$) – branch does exist (is supported by the data)
  - fail to reject – branch can be replaced by polytomy
- in ASTRAL with '-t 10'
- Sayyari, Erfan, Mirarab (2018): *Testing for polytomies in phylogenetic species trees using quartet frequencies*. Genes 9: 132

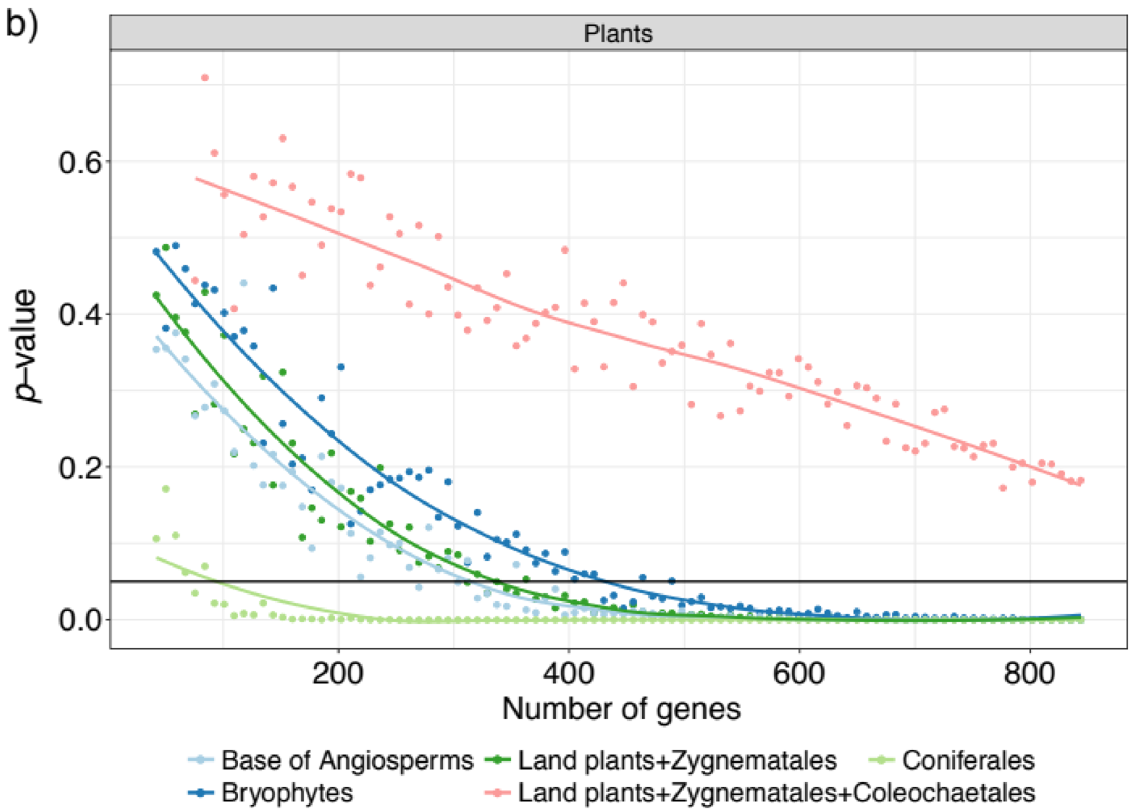
Polytomy vs. branch length

- *x*-axis – branch length in log CU (coalescence units)
- *y*-axis – polytomy test p-value
- points with p < 0.05 in black
- longer branches are usually supported



Sayyari, Erfan, Mirarab, 2018, Genes

Polytomy tests – plant dataset



- polytomy hypotheses rejected with increasing number of genes
- correct relationship between *Chara* and Coleochaetales remains hard to resolve, may be with more data?

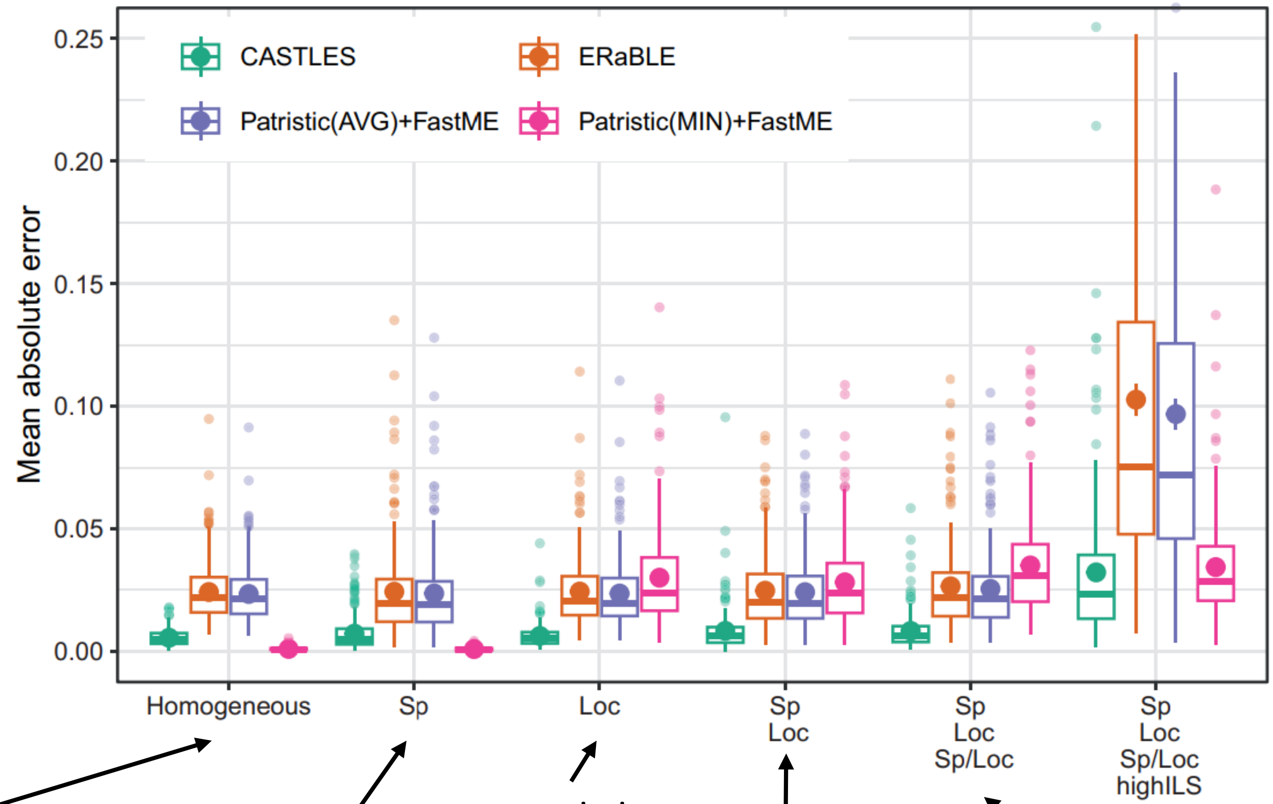
Sayyari, Erfan, Mirarab, 2018, Genes

ASTRAL-IV

https://github.com/chaoszhang/ASTER/blob/master/tutorial/astral4.md

- scalable alternative to ASTRAL – large datasets, multi-individual, gene trees with missing taxa
- integrates CASTLES-II algorithm (Tabatabaee et al. 2023) – computes terminal and internal branch lengths in expected substitution-per-site units
  - **C**oalescent-**A**ware **S**pecies **T**ree **L**ength **E**stimation in **S**ubstitution-unit
  - addresses gene tree heterogeneity due to incomplete lineage sorting (ILS), as modeled by the multi-species coalescent (MSC) model
  - does not assume strict molecular clock
  - superior accuracy compared to alternatives
- part of ASTER❋ (https://github.com/chaoszhang/ASTER) – **A**ccurate **S**pecies **T**ree **E**stimato**R**

Accuracy of CASTEL algorithm



strict molecular clock, no rate variation

variation across species tree branches

rate variation only across genes

rate variation both across species and across genes

rate variations specific to each branch of each gene tree = heterotachy

Tabatabaee et al. 2023

ASTRAL practical

https://github.com/smirarab/ASTRAL

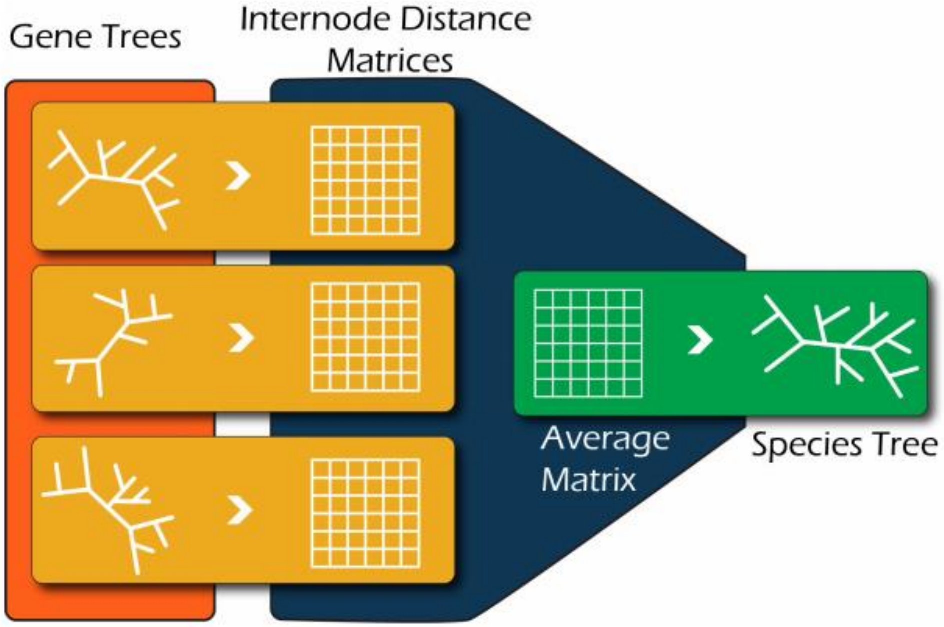
- **help**: `java -jar astral.5.7.3.jar`
- **simple run**: `java -jar astral.5.7.3.jar -i input.trees -o output.tre`
- **scoring existing tree**: `java -jar astral.5.7.3.jar –q existing.tre -i input.trees -o output.tre`
- **branch annotation**: `java -jar astral.5.7.3.jar –q existing.tre -i input.trees –t 2 -o output.tre`
- **polytomy test**: `java -jar astral.5.7.3.jar –q existing.tre -i input.trees –t 10 -o output.tre`
- **multilocus bootstrapping**: `java -jar astral.5.7.3.jar -i input.trees –b bootstrapTreeList -o output.tre` (bootstrapTreeList is a text file containing file path of gene tree bootstrap files, one per line, i.e., file full of paths of files full of trees)

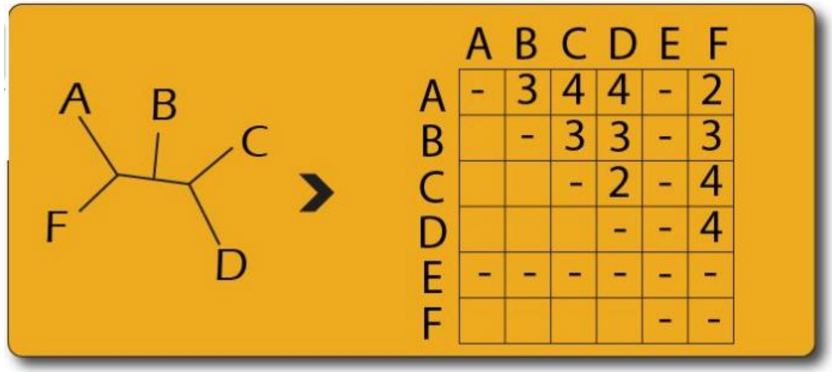
ASTRID

Accurate **S**pecies **TR**ees from **I**nternode **D**istances

github.com/pranjalv123/ASTRID

- species tree computed from internode distance matrices
- reimplementation of NJst method (Liu & Yu, 2011)
- uses FastME-2 (Lefort, Desper & Gascuel 2015)
- fast, can scale to extremely large datasets





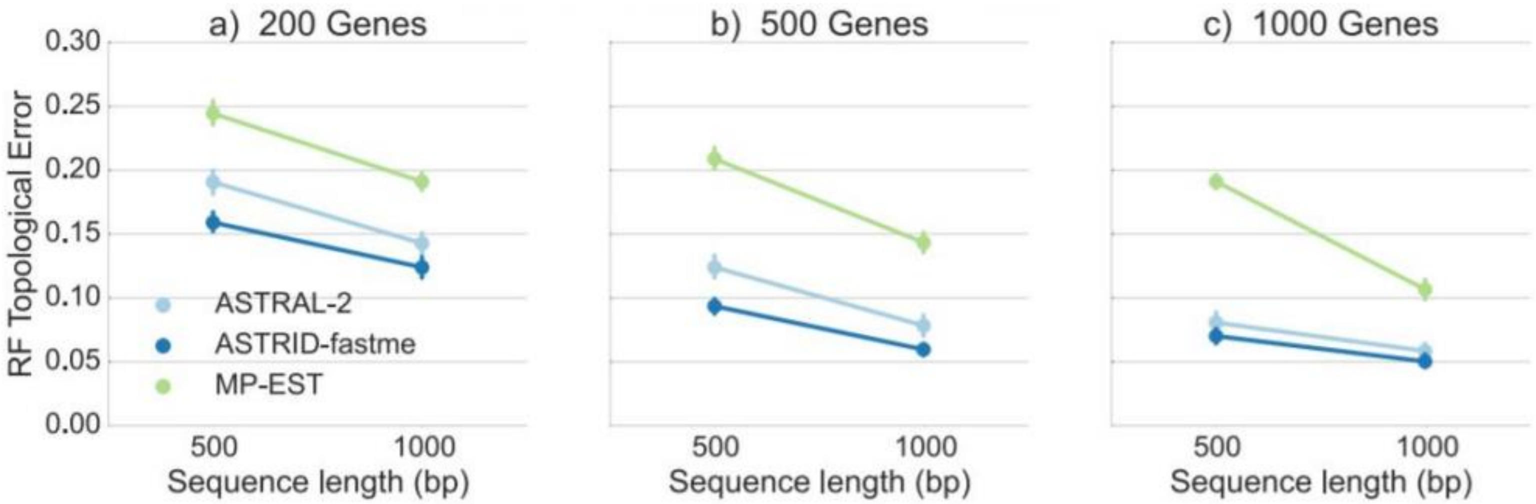
	A	B	C	D	E	F
A	-	3	4	4	-	2
B		-	3	3	-	3
C			-	2	-	4
D				-	-	4
E	-	-	-	-	-	-
F					-	-

Pranjal Vachaspati

http://tandy.cs.illinois.edu/astrid-ssb-v2.pdf

ASTRID accuracy

- high accuracy on datasets with ILS



- 47-taxon simulated dataset, high ILS

http://tandy.cs.illinois.edu/astrid-ssb-v2.pdf

ASTRID practical

- download ASTRID binary
  - https://github.com/pranjalv123/ASTRID/releases
  - Win version: https://github.com/pranjalv123/ASTRID-1/releases

- run over your trees
  - `ASTRID -i input.trees -o output.tre`

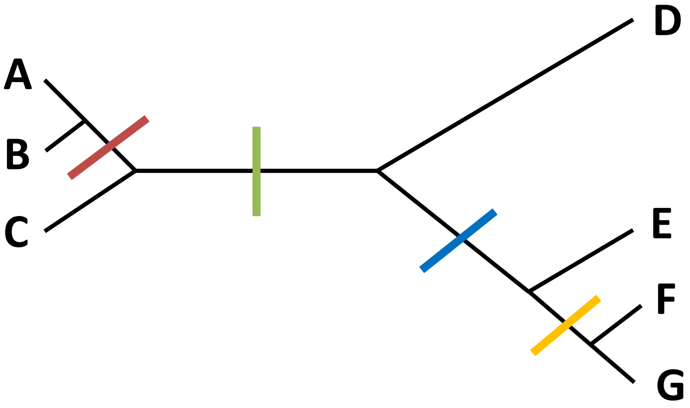
- no support values unless you do MLBS
  - `ASTRID -i input.trees –b bootstrapTreeList -o output.tre`

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 ‘BINCAT’ 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	?	?	?	?	

MRL

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

- download mrp.jar from https://github.com/smirarab/mrpmatrix
- run on your trees
  - `java -jar mrp.jar trees.nwk MRLmatrix.phylip -randomize`
- analyze MRL matrix is using RAxML as ‘BINGAMMA’ model
  - `raxmlHPC-PTHREADS -T nrCores -f a -s MRLmatrix.phylip -n MRLresult -m BINGAMMA -p 1234 -x 1234 -N nrBSreps`

Quartet Sampling (QS)

Replacement for bootstrap in phylogenomic studies…

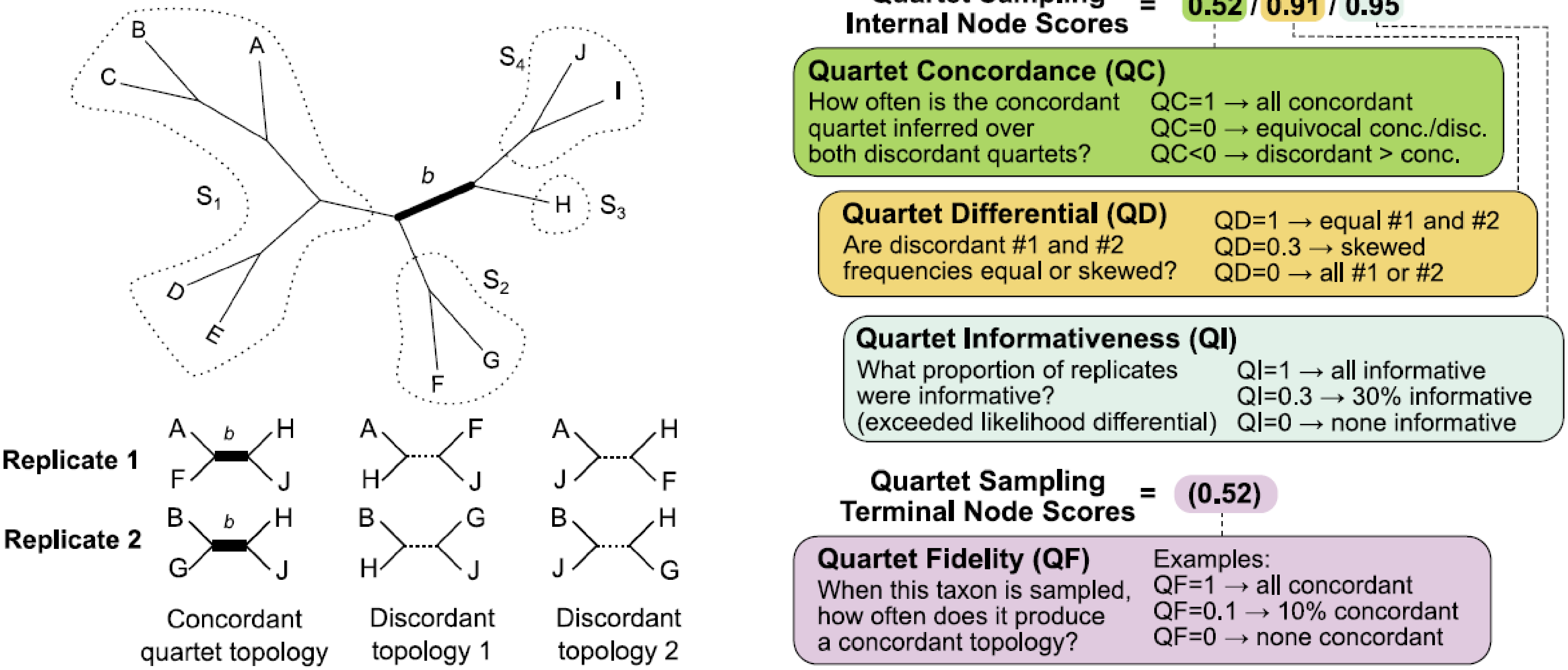
- quartet-based evaluation system
- synthetizes several phylogenetic and genomic analytical approaches
- discordance testing
- distinguishes strong conflict from weak support
- three different scores per branch
  - Quartet Concordance (QC)
  - Quartet Differential (QD)
  - Quartet Informativeness (QI)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

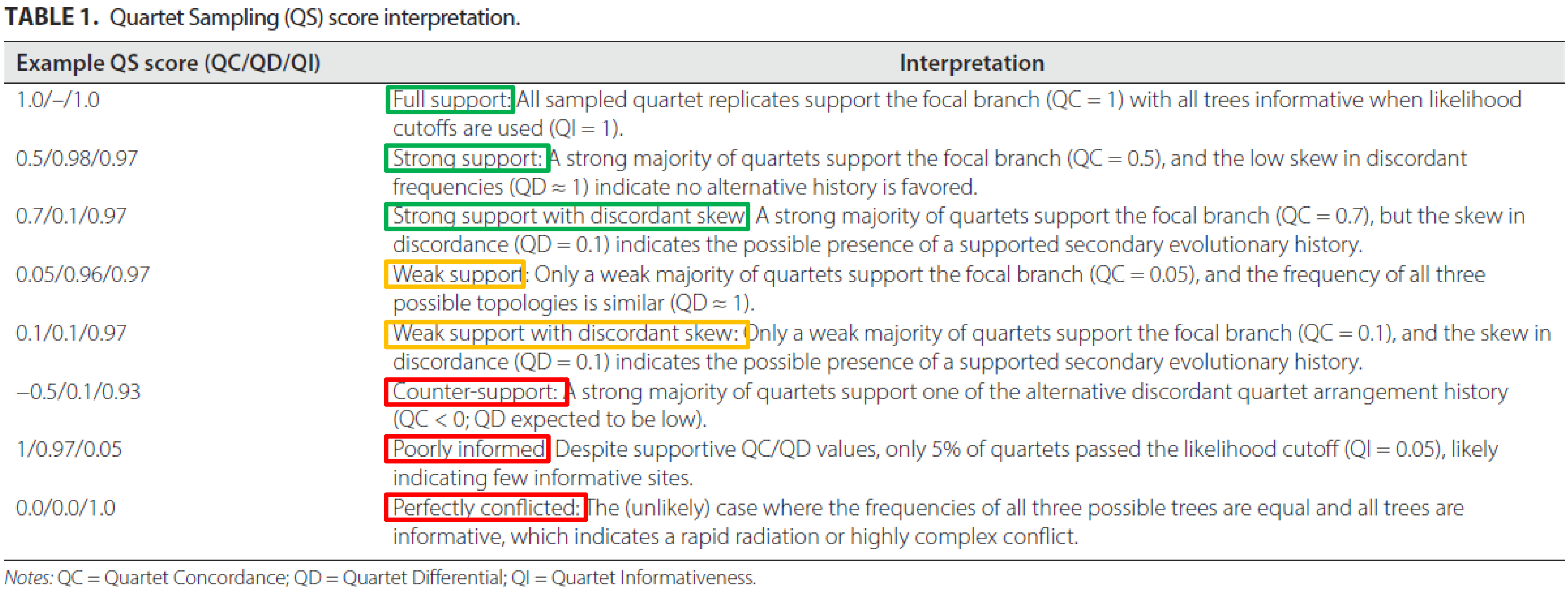


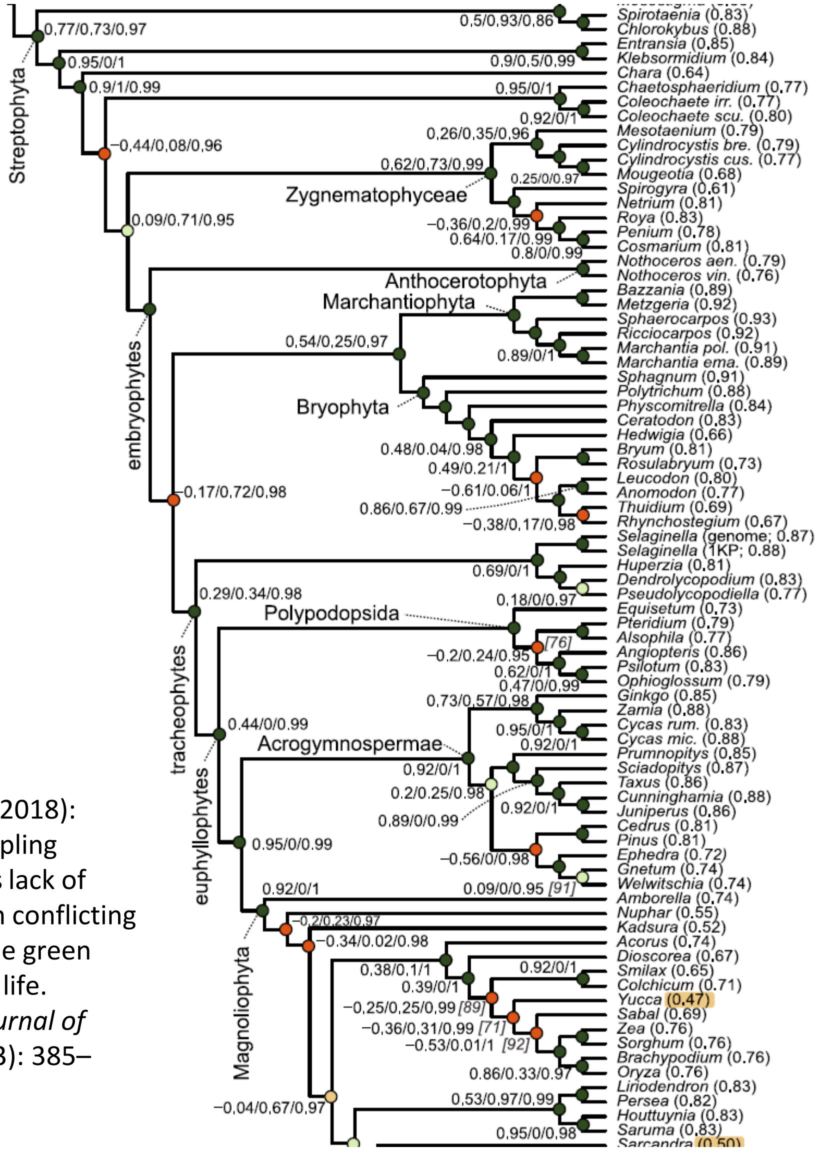
TABLE 1. Quartet Sampling (QS) score interpretation.

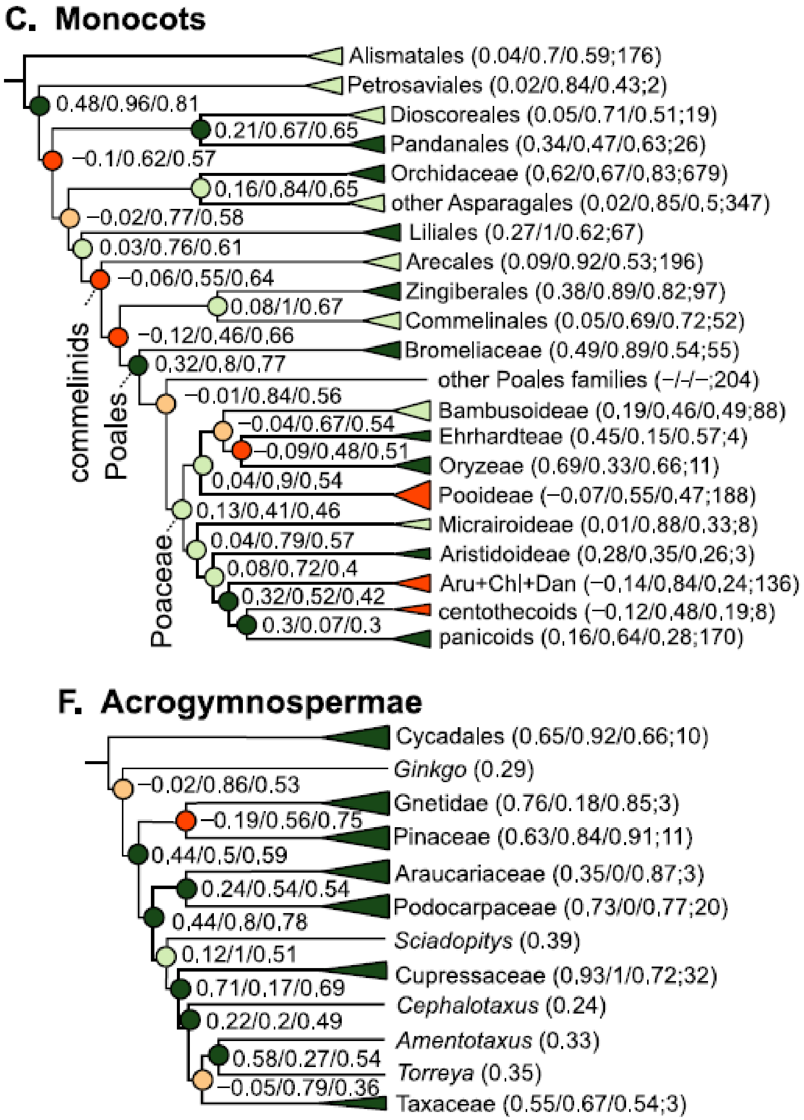
Example QS score (QC/QD/QI)	Interpretation
1.0/–/1.0	Full support: All sampled quartet replicates support the focal branch (QC = 1) with all trees informative when likelihood cutoffs are used (QI = 1).
0.5/0.98/0.97	Strong support: A strong majority of quartets support the focal branch (QC = 0.5), and 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.

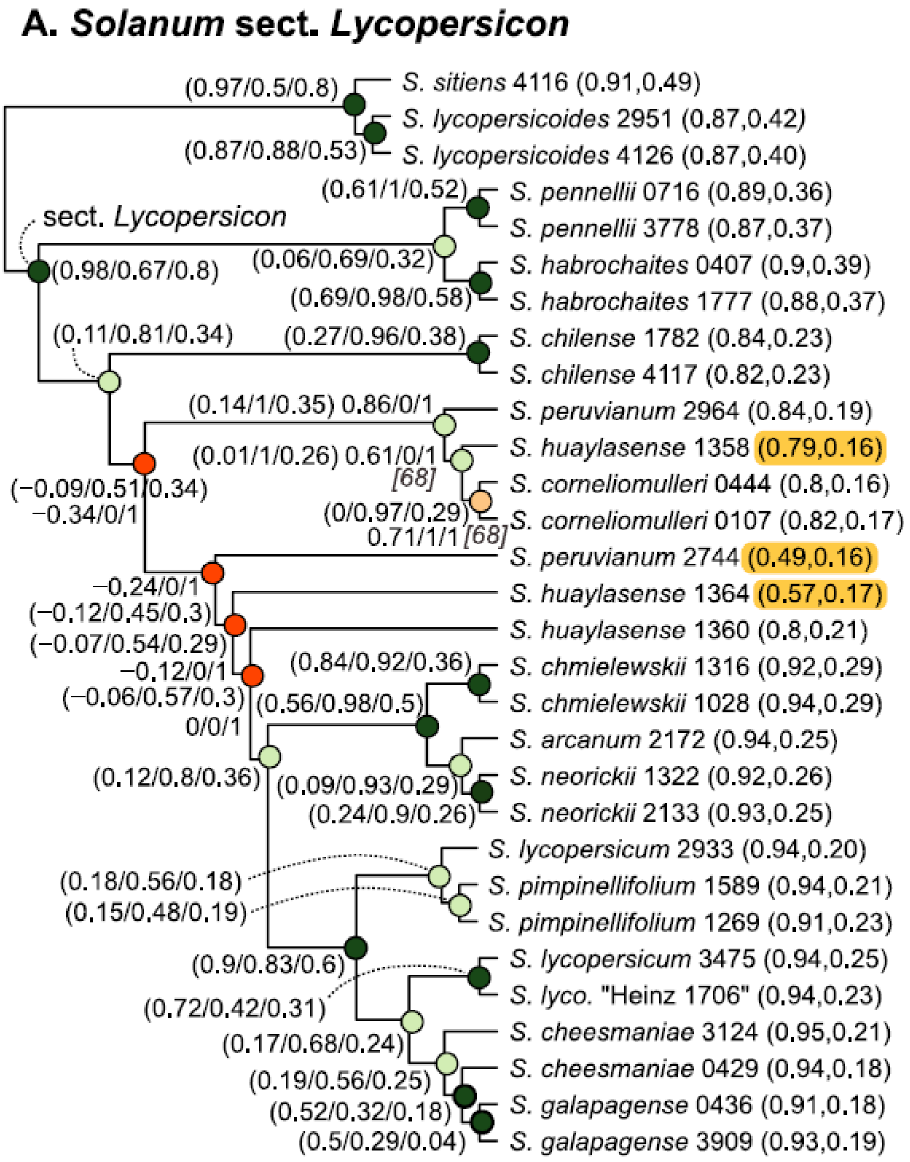
Quartet Sampling – land plants





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*

Pease et al. (2016)

Quartet Sampling

https://github.com/FePhyFoFum/quartetsampling

- requirements
  - Python 3.x (Numpy and Scipy for some options)
  - RAxML-ng 0.9.0 (or RAxML 8.1+, *PAUP, IQ-TREE 1.6.x)

- NEWICK phylogeny
- alignment in Relaxed Phylip Format – DNA or AA
- RAxML-style partitions file (optional)

-
```
python3 quartet_sampling.py –tree TREE.nwk –align
ALIGNMENT.phy –reps 100 –threads 4 –lnlike 2
```

Quartet Sampling

https://github.com/FePhyFoFum/quartetsampling

output tables

- **RESULT.node.scores.csv** (scores and other info per branch/terminal)
- **RESULT.node.counts.csv** (counts of concordant and discordant quartet arrangements)

output trees

- **RESULT.labeled.tre.freq/qc/qd/qu** (Newick tree with the each internal branch labeled with frequency of concordant replicates or QC/QD/QI scores)
- **RESULT.labeled.tre.figtree** (FigTree format phylogeny that contains all QS scores and a “score” field with QC/QD/QI for internal branches)

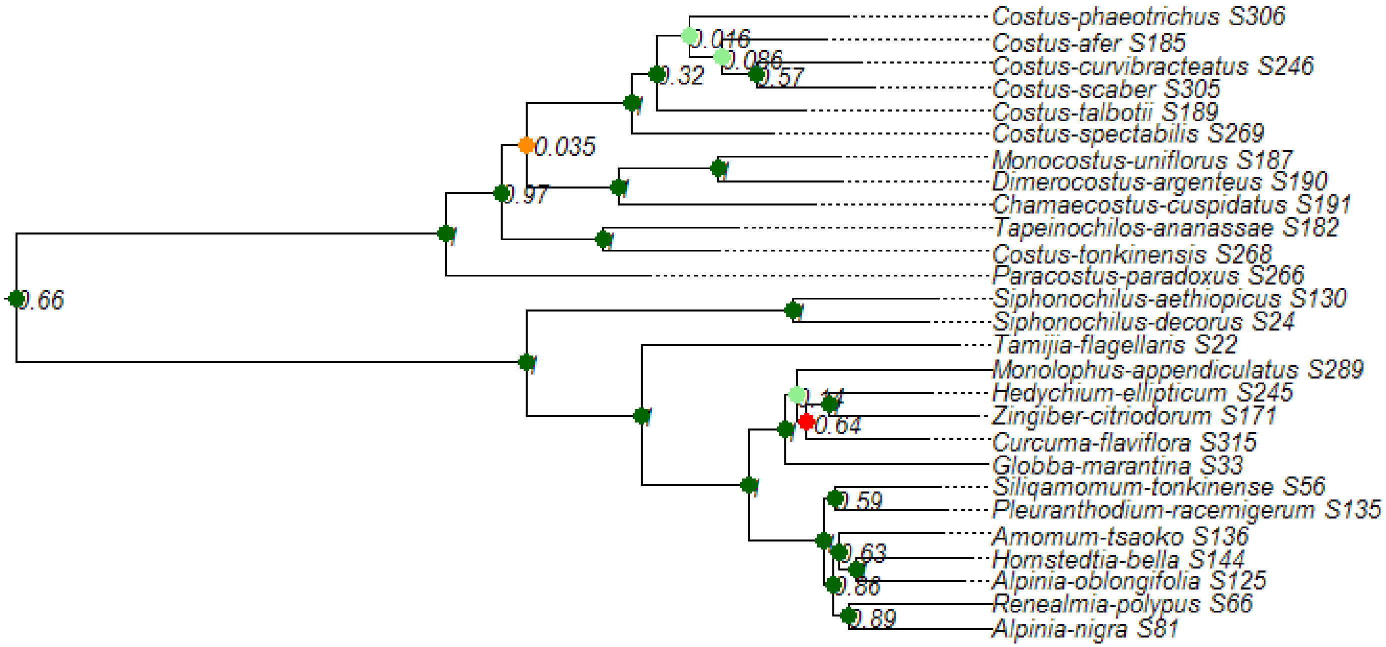
Quartet Sampling

https://github.com/FePhyFoFum/quartetsampling

modify and plot tree using a custom script (bash/R)

https://github.com/tomas-fer/scripts/blob/master/quartetsampling.r

- RESULT.labeled.tre.qc – tree with 'qc' values (used later for coloring nodes)
- RESULT.labeled.tre.figtree – tree with all values (used for plotting the tree and three scores



Quartet Sampling

https://github.com/FePhyFoFum/quartetsampling

- start interactive job on MetaCentrum

  -
    ```
    qsub -I -l select=1:ncpus=4:mem=4gb
    ```
- go to your directory (with species tree and concatenated alignment)
- download quartet sampling from GitHub

  -
    ```
    git clone https://github.com/FePhyFoFum/quartetsampling.git
    ```
- enable modules

  -
    ```
    module add python-3.6.2-gcc
    ```

  -
    ```
    module add raxml-ng-8
    ```
- run this to see options

  -
    ```
    python3 quartetsampling/pysrc/quartet_sampling.py –h
    ```

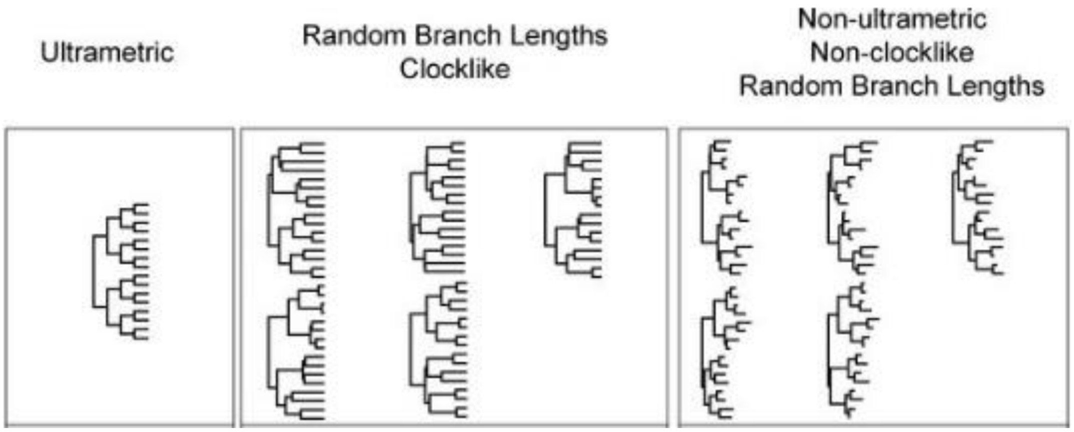
-
  ```
  quartetsampling/pysrc/quartet_sampling.py --tree Astral.tre --align concatenated.phylip --reps 100 --threads 4 --lnlike 2
  ```
- process the tree with the instructions in `https://github.com/tomas-fer/scripts/blob/master/quartetsampling.r`

Gene selection/filtering

- missing data
    - bases per accession (…remove accession)
    - species per gene (…remove gene)
    - alignment improvement
- variability/informativness
    - information content
    - nr. or % variable/parsimony informative sites
    - GC content
    - evolutionary rate (slow/fast genes)
- tree characteristics
    - overall bootstrap support (higher BS = stronger phylogenetic signal)
    - clocklikeness (a measure how close to ultrametric a tree is: the algorithm finds a root that minimizes the coefficient of variation in root to tip distances and returns that value; a lower value is more clock-like, an ultrametric tree has a score of 0)
    - long branch score (standard deviation from the taxon-specific long branch score defined by Struck, 2014)
- saturation (simple linear regression on uncorrected p-distances against inferred distances, i.e., branch length - slope and R^2; higher values mean lower saturation potential)

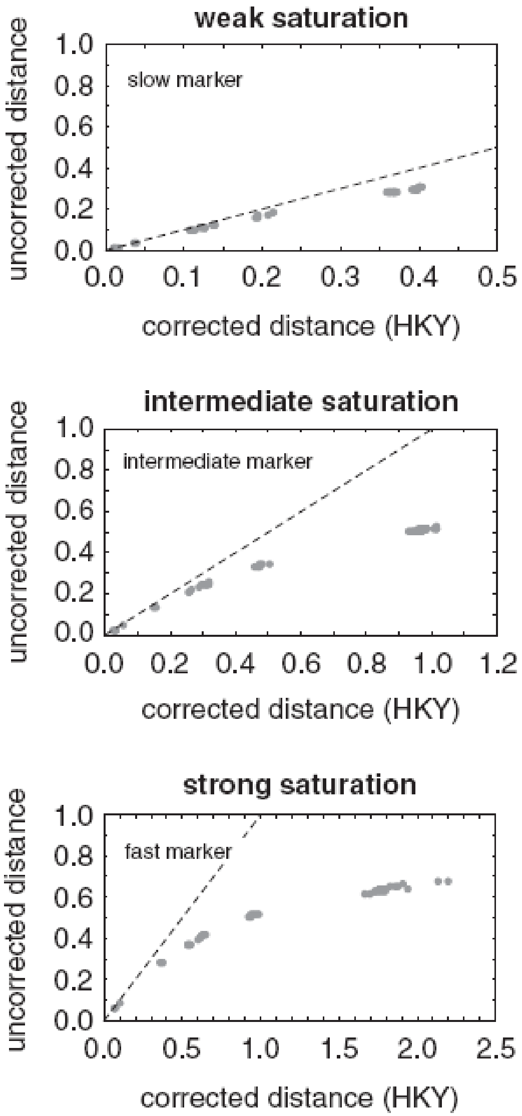
Clocklikeness

- a measure how close to ultrametric a tree is
- the algorithm finds a root that minimizes the coefficient of variation in root to tip distances and returns that value
- a lower value is more clock-like
- an ultrametric tree has a score of 0



Saturation

- simple linear regression on
  - uncorrected p-distances
  - inferred distances, i.e., branch length
- slope and R^2 is reported
- higher values mean lower saturation potential

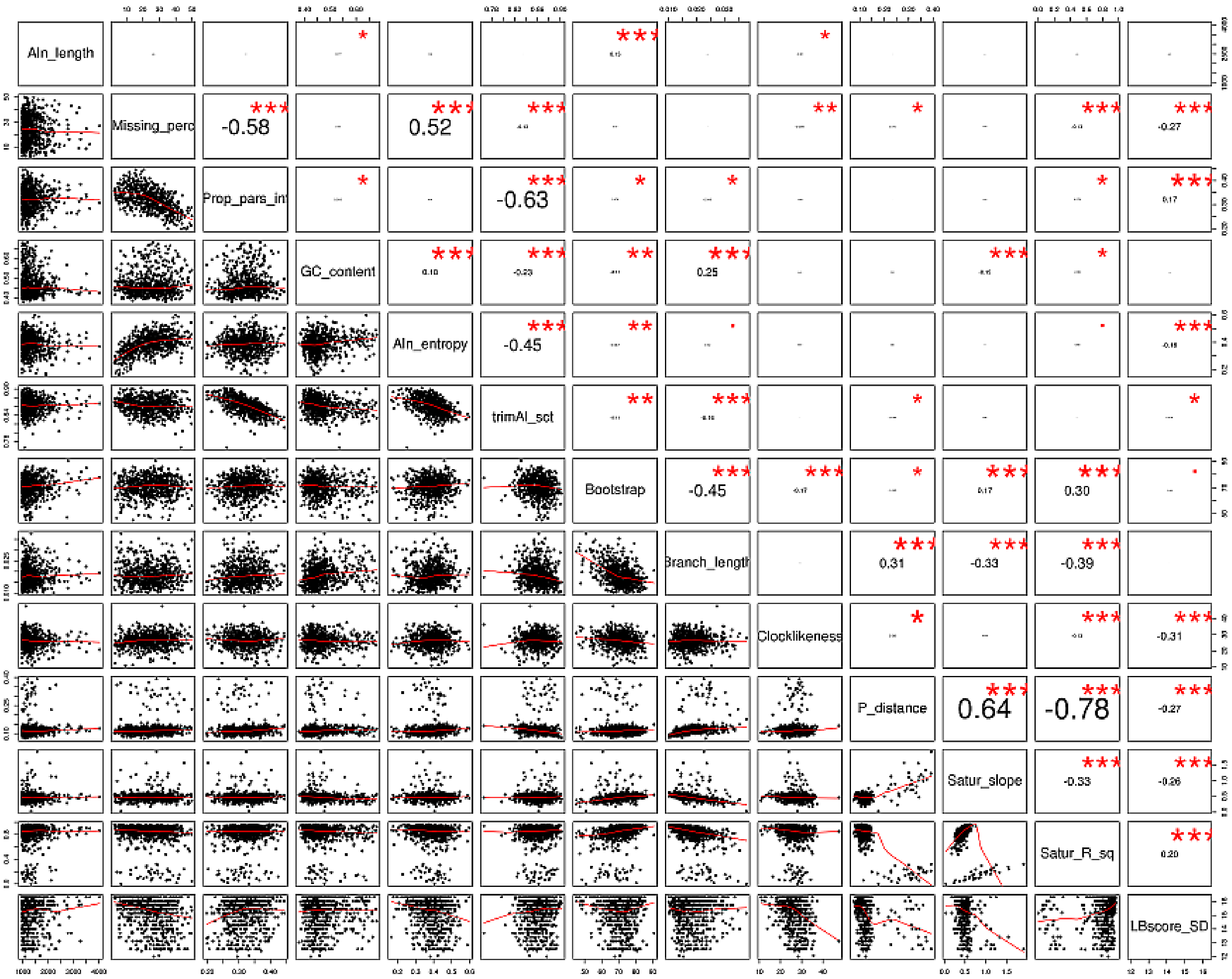


How to calculate alignment/tree characteristics?

- **AMAS** (https://github.com/marekborowiec/AMAS)
  - concat, convert, summary, remove, translate
  - `python3 AMAS.py summary -f fasta -d dna -i *.fasta`

- **R scripts** (e.g., https://github.com/marekborowiec/good_genes)
  - average bootstrap
  - average branch length
  - saturation (slope, R^2)

Relationships among alignment/tree characteristics



Thank you…

dragon blood tree (*Dracaena cinnabari*), Socotra

References

http://tandy.cs.illinois.edu/astral-tutorial.pdf (Siavash Mirarab)

http://tandy.cs.illinois.edu/astrid-ssb-v2.pdf (Pranjal Vachaspati)

Degnan JH & Rosenberg NA. 2009. *Gene tree discordance, phylogenetic inference and the multispecies coalescent*. Trends in Ecology & Evolution 24, 332-340.

Drummond AJ & Bouckaert RR. 2015. *Bayesian evolutionary analysis with BEAST*. Cambridge University Press.

Lefort V et al. 2015. FastME 2.0: *A comprehensive, accurate, and fast distance-based phylogeny Inference program*. Molecular Biology and Evolution 32, 2798-2800.

Liu L & Yu L. 2011. *Estimating species trees from unrooted gene trees*. Systematic Biology 60, 661-667.

Mirarab S et al. 2014. *ASTRAL: genome-scale coalescent-based species tree estimation*. Bioinformatics 30, i541-i548.

Moloy EK & Warnow T. 2018. *To include or not to include: the impact of gene filtering on species tree estimation methods*. Systematic Biology 67, 285-303.

Nguyen N et al. 2012. MRL and SuperFine+MRL: new supertree methods. Algorithms for Molecular Biology 7, 3.

Pease JB 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.

Rabiee M et al. 2019. *Multi-allele species reconstruction using ASTRAL*. Molecular Phylogenetics & Evolution 130, 286-296.

Reaz R et al. 2014. *Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach*. PLoS ONE 9, e104008.

Sayyari E & Mirarab S. 2016. *Fast coalescent-based computation of local branch support from quartet frequencies.* Molecular Biology and Evolution 33, 1654-1668.

Sayyari E et al. 2017. *Fragmentary gene sequences negatively impact gene tree and species tree reconstruction*. Molecular Biology and Evolution 34, 3279-3291.

Sayyari E et al. 2018. *Testing for polytomies in phylogenetic species trees using quartet frequencies*. Genes 9, 132.

Seo TK. 2008. *Calculating bootstrap probabilities of phylogeny using multilocus sequence data*. Molecular Biology and Evolution 25, 960-971.

Tabatabaee Y et al. 2023. *Phylogenomic branch length estimation using quartets.* Bioinformatics 39, Supplement_1, i185-i193

Zhang Ch et al. 2018. *ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees*. BMC Bioinformatics 19, 153.

Zimmermann T et al. 2014. *BBCA: Improving the scalability of *BEAST using random binning*. BMC Genomics 15(Suppl 6), S11.