

Phylogenomics

Gene trees/species tree

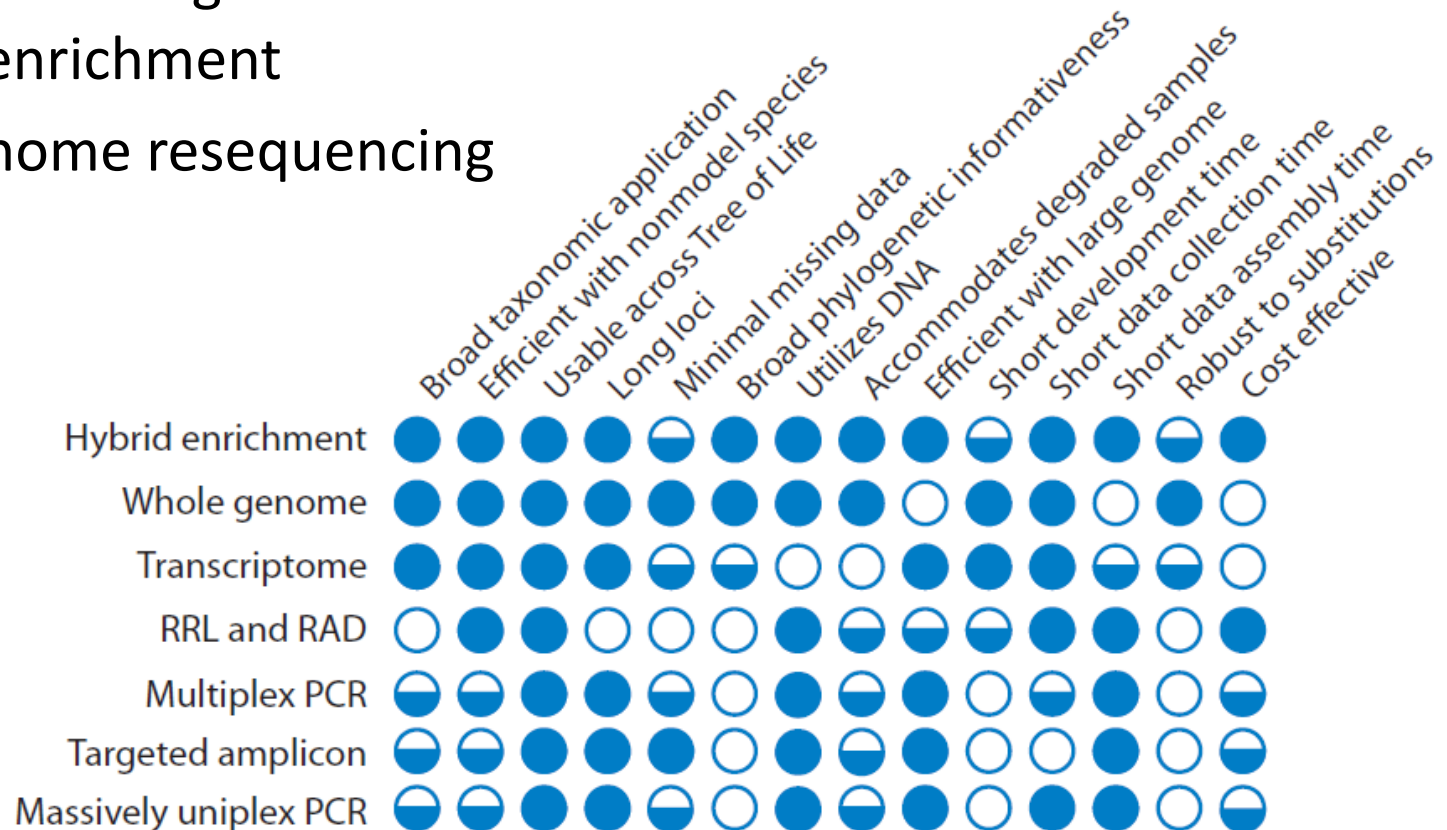
Tomáš Fér
Biosystematics, 2022

Phylogenomics

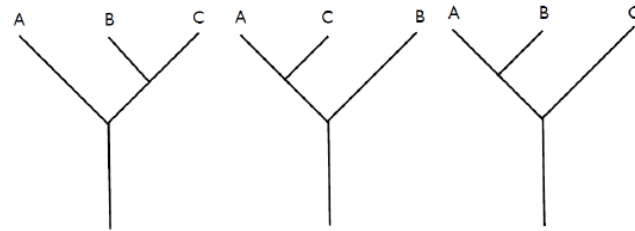
- using whole-genome sequences or large portion of the genome to build a phylogeny
 - whole chloroplast sequences
 - hundreds or thousands of genes
 - transcriptomes
 - target-enrichment (Hyb-Seq)
- gene tree – individual evolutionary history
- species tree – ‘true’ species evolution
- gene tree/species tree (in)congruence

Phylogenomic data sources

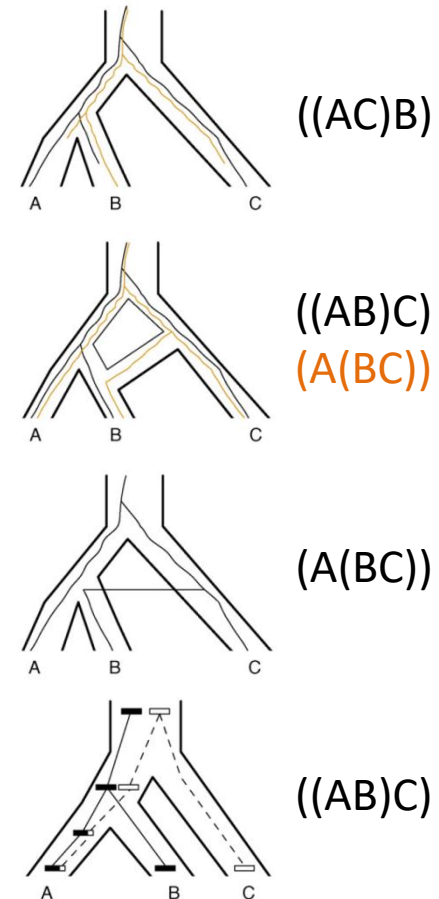
- transcriptomes
- genome skimming
- targeted enrichment
- whole genome resequencing



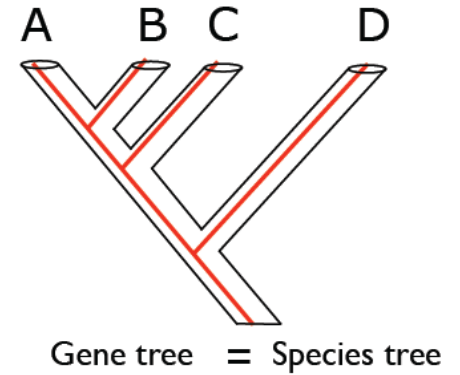
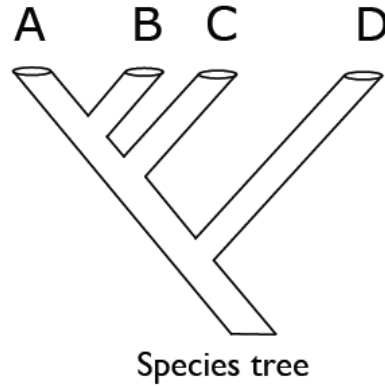
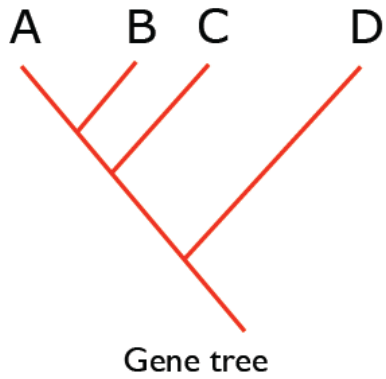
Gene tree incogruence



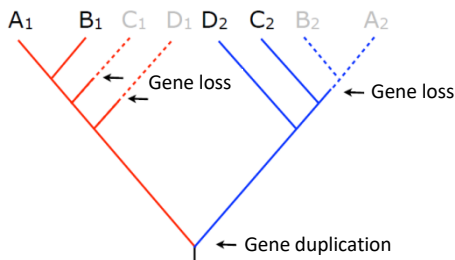
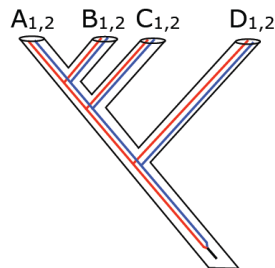
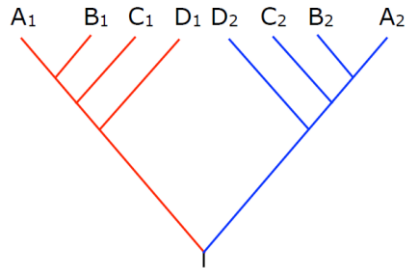
- incomplete lineage sorting (ILS)/deep coalescence
- gene duplications and losses (orthology problem)
- hybridization/polyploidization
 - affects whole genomes
- horizontal gene transfer (HGT)
 - affects small DNA segments
- recombination
 - different histories for neighboring segments in genes



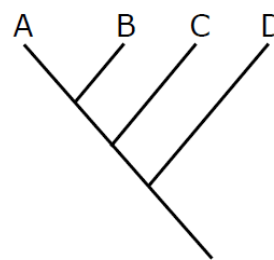
Gene trees vs species tree



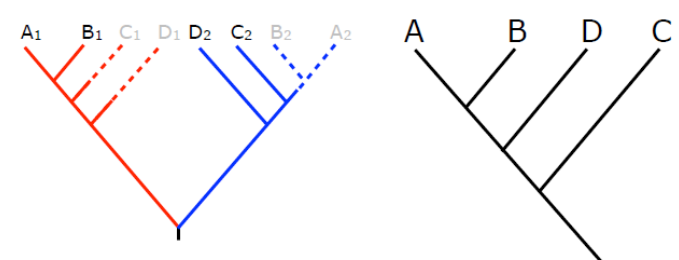
gene duplications and losses



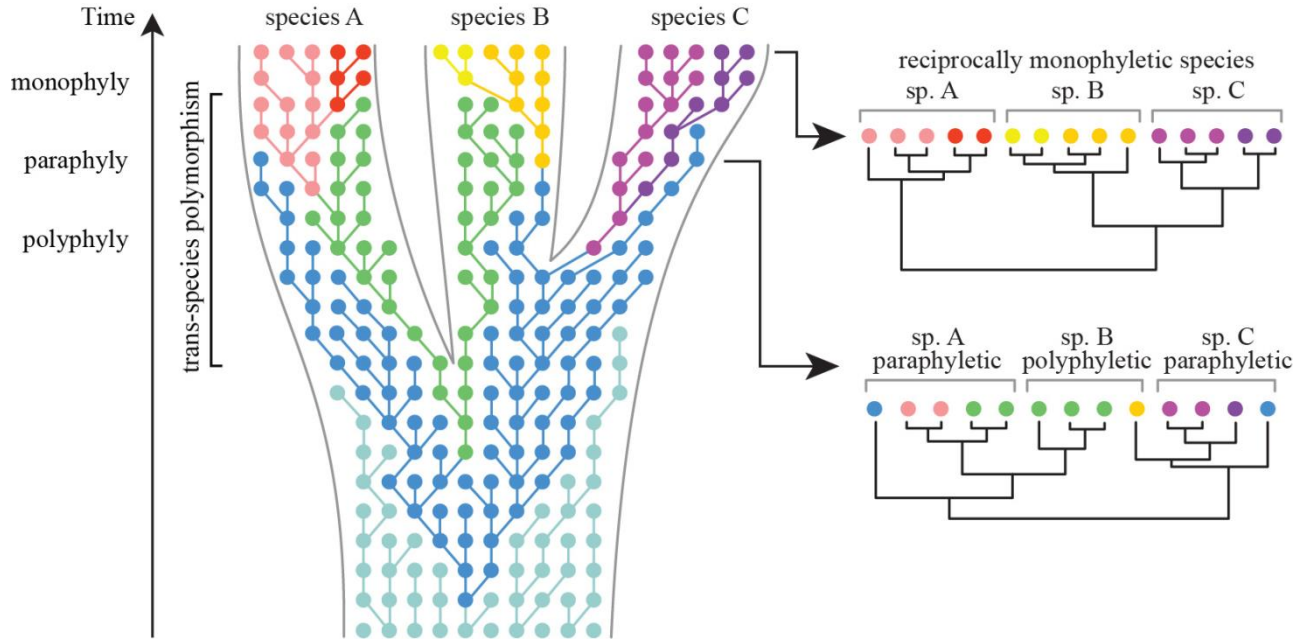
true species tree



inferred species tree

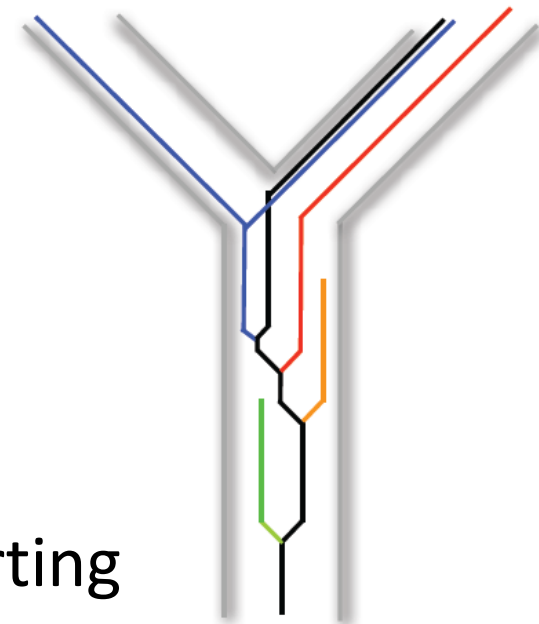


Coalescence processes

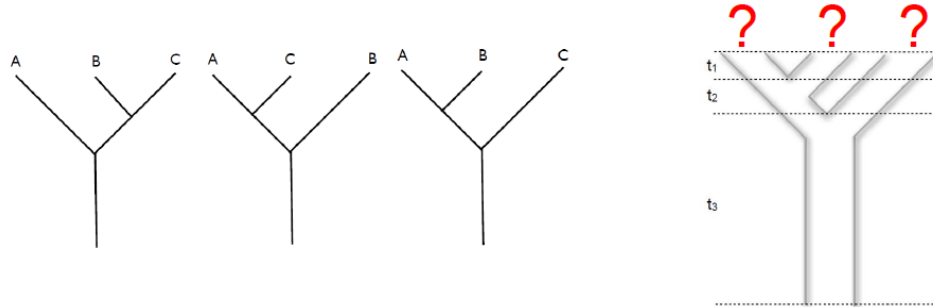


<https://frederikleiaert.wordpress.com/green-algae/dna-based-species-delimitation-in-algae/>

incomplete lineage sorting



Species tree estimation



- **concatenation** (supermatrix) – 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 (not applicable to large datasets)
 - **summary methods**
 - **supertree methods** – MRL (maximum representation using likelihood)
 - MP-EST – maximum likelihood estimation of rooted 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

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

Summary methods

Species tree estimation

require rooted gene trees

- **MP-EST** – maximum pseudo-likelihood approach for estimating species trees
- **STAR** – species tree estimation using average ranks of coalescences

unrooted gene trees

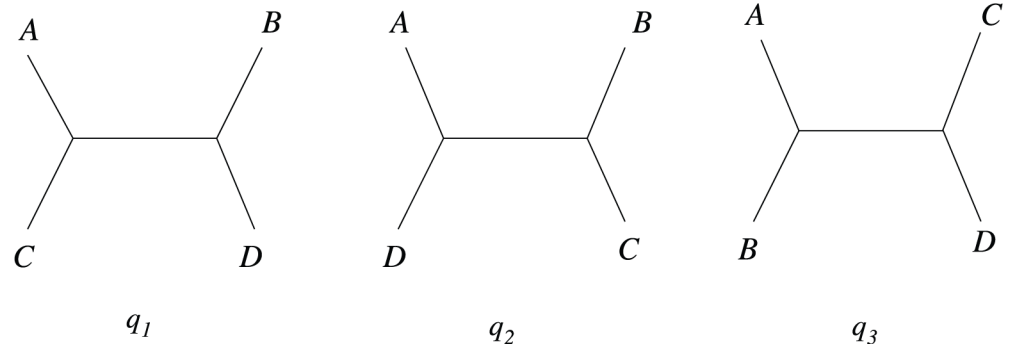
- **STEAC** – species tree estimation using average coalescence times
- **ASTRAL** – Accurate Species Tree Reconstruction ALgorithm
- **ASTRID** – Accurate Species TREes from Internode Distances (reimplementation of NJ_{st} method)

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

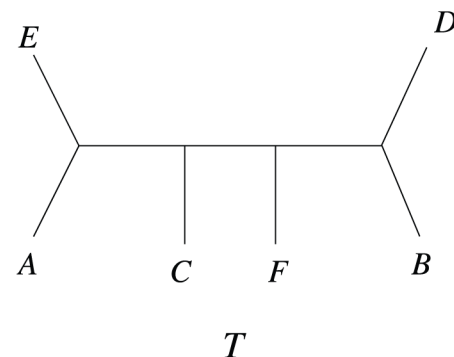
- **SNAPP** – SNP and AFLP Package for Phylogenetic analysis
- **SVDquartets**

Tree reconstruction from quartets

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

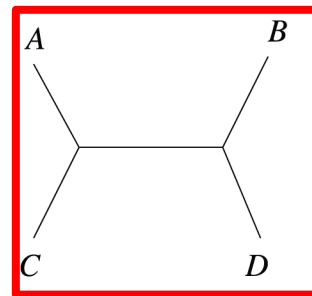


- which *quartet* is consistent with T ?

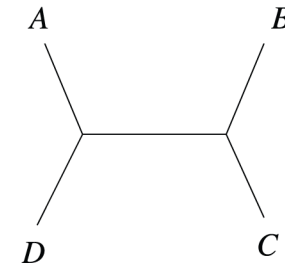


Tree reconstruction from quartets

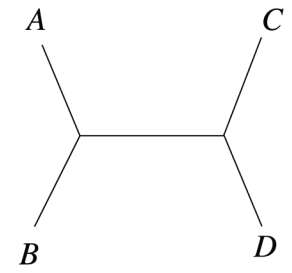
- quartet – unrooted tree over 4 taxa
- three possible quartets
- only one quartet q is consistent with final tree T



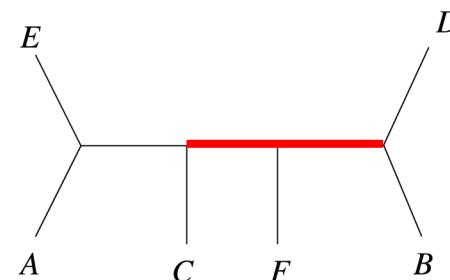
q_1



q_2



q_3



T

ASTRAL

Accurate **S**pecies **T**ree **R**econstruction **A**lgorithm

<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

ASTRAL problems

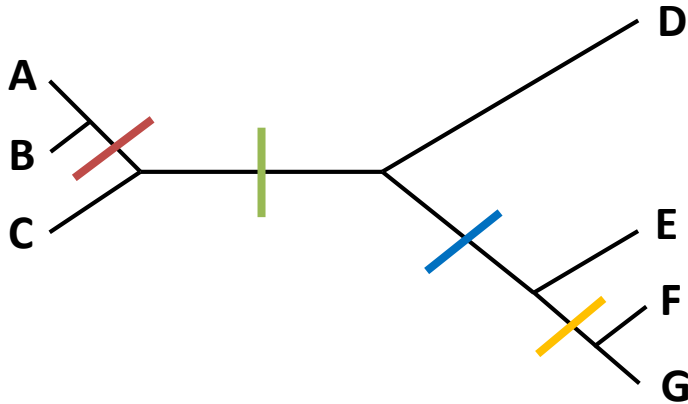
- assumption for statistical consistency
 - randomly distributed sample of gene trees
 - recombination-free
 - reticulation-free
 - error-free
 - orthologous
- in practice: reduced accuracy with low accuracy gene trees
- branch length
 - only for internal branches (unless multiple individuals per species)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

MRL

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

- 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
- similarly MRP – matrix analyzed with parsimony

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
→

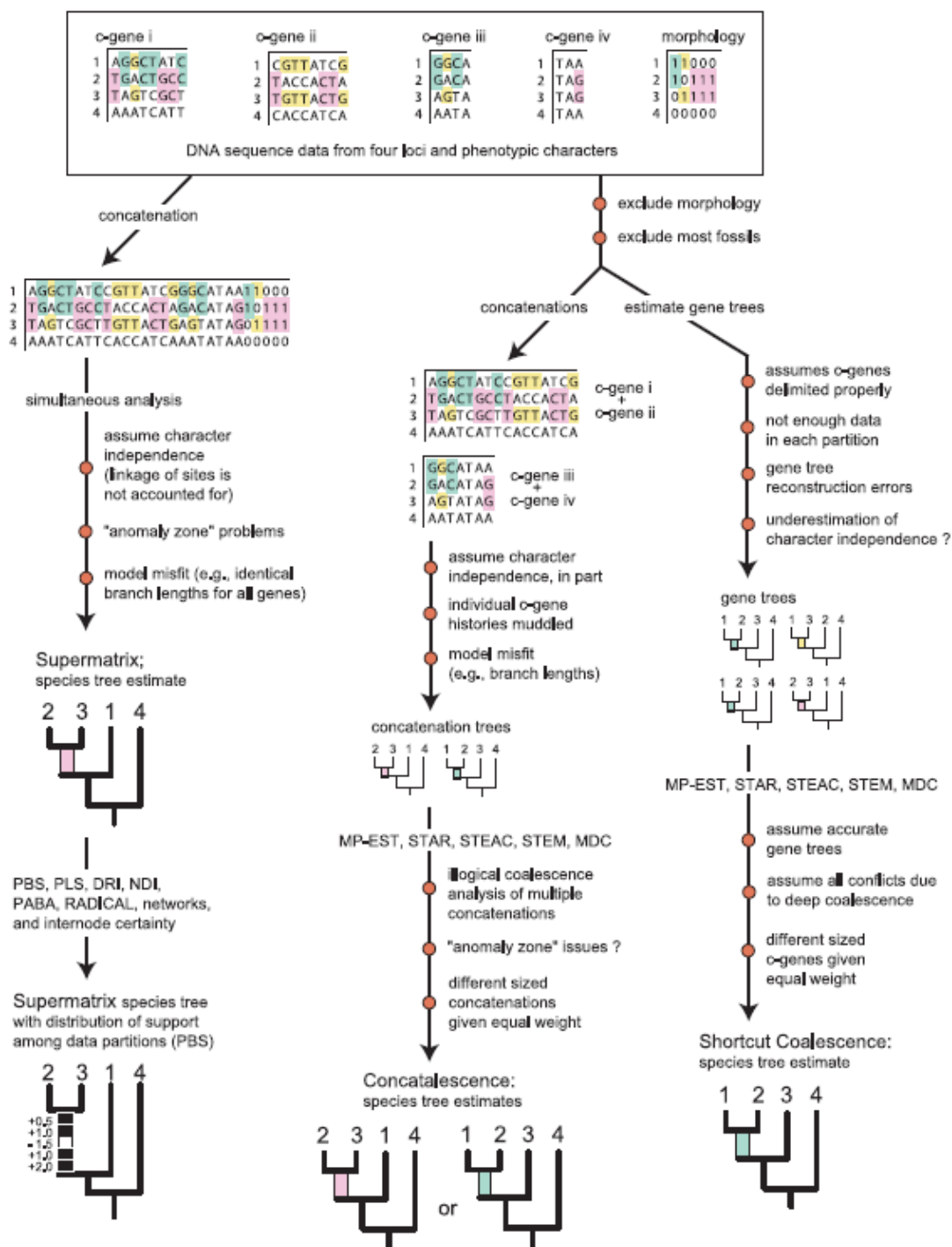
	*	*	*	*	...
A	0	1	1	0	
B	0	1	1	0	
C	1	1	1	0	
D	1	0	1	0	
E	1	0	0	0	
F	1	0	0	1	
G	1	0	0	1	
H	?	?	?	?	

Concatenation vs. coalescence

- concatenation
 - **in favor:** longer datasets allow for hidden support to appear
 - **against:** could be misleading under strong ILS
- coalescence (i.e., “shortcut coalescence” or summary methods)
 - **in favor:** addresses ILS
 - **against:**
 - short genes give poor gene trees (big problem!)
 - definition of coalescence-gene (segments with no internal recombining) debatable
 - concatenating coalescence-genes to longer alignments (“concatalescence”) not recommended?

see also:

Gatesy & Springer (2014): *Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum*. *Molecular Phylogenetics and Evolution* 80: 231–266.



Gatesy & Springer. 2014.
Molec. Phylog. Evol. 80:
231–266.

Filtering datasets

single-copy genes with good properties (no paralogs, low conflicting signal...) – filter out contaminants

- BLAST-based searches
- remove taxa with long branches
- remove poorly aligned regions

alignments

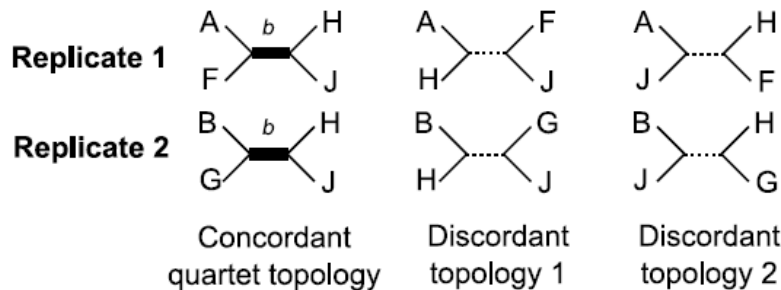
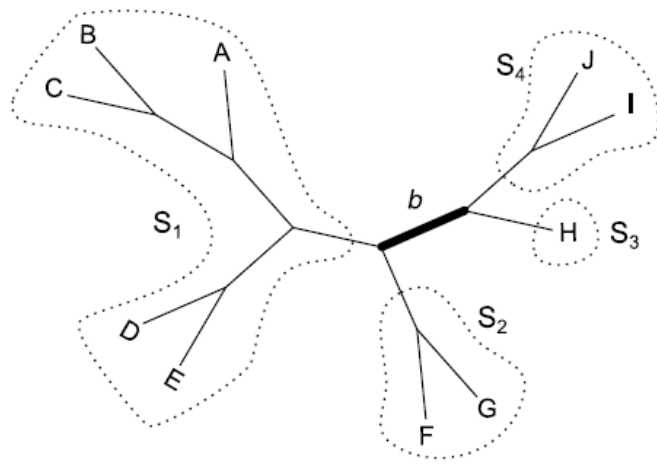
- length – longer better
- missing data – fewer better
- parsimony informative sites – more better
- information content

trees

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

Quartet support

Replacement for bootstrap in phylogenomic studies...



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)

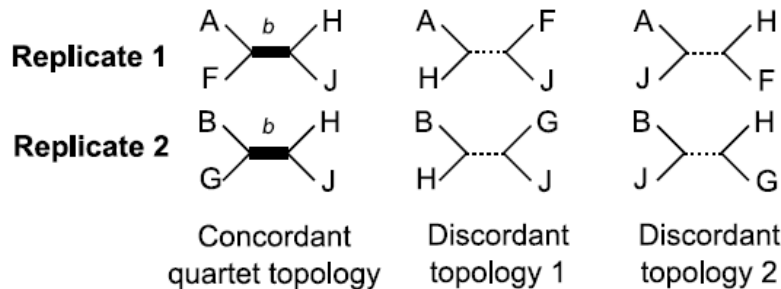
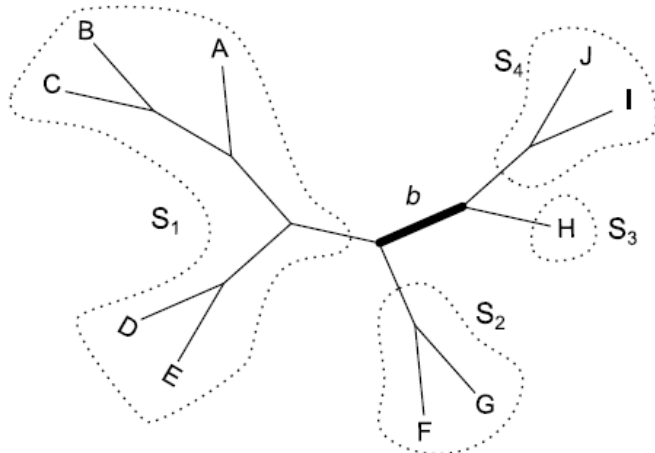
When this taxon is sampled, how often does it produce a concordant topology?
 Examples:
 QF=1 → all concordant
 QF=0.1 → 10% concordant
 QF=0 → none concordant

Quartet Sampling

Replacement for bootstrap in phylogenomic studies...

- 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
- distinguishes strong conflict from weak support

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 quartet inferred over both discordant quartets? 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

Phylotranscriptomic analysis of the origin and early diversification of land plants

Wickett et al., 2014, PNAS

- capstone paper from oneKP project
- **transcriptomes** from 92 streptophyte taxa + 11 **genomes**
- up to **852 nuclear genes**, ~1,700,000 sites
- **69 analyses**
 - missing data filtering
 - supermatrix, supertree, coalescence-based
 - ML, Bayesian
 - partitioned/unpartitioned
 - amino acids, DNA

Taxonomic concepts

- **Streptophytes** – Klebsormidiales, Coleochaetales..., Charales, Zygnematophyceae †
- **Embryophytes (land plants)** – Anthocerotophyta (hornworts), Marchantiophyta (liverworts), Bryophyta (mosses) †
- **Tracheophytes (vascular plants)** – Lycopodiophyta (lycophytes) †
- **Euphyllophytes** – monilophytes (ferns) †
- **Spermatophytes (seed plants)** – Gymnosperms †
- **Angiosperms (flowering plants)** – ANA grade, monocots, magnoliids, eudicots

Introduction

- origin of embryophytes (land plants) – Ordovician (480 Mya)
- innovations – parental protection for embryo, alternation of generations (diploid sporophyte, haploid gametophyte)
- changes in global carbon cycle
- forming terrestrial ecosystems
- series of rapid radiations – most diverse group of extant plants

- main questions
 - which **green algae lineage** is most closely related to embryophytes?
 - what is the **branching order** among the main embryophyte lineages?

Previous studies

- streptophytes monophyletic, but...
- branching order relative to embryophytes uncertain
- shared characters among embryophytes, Charales, Coleochaetales
 - oogamous sexual reproduction
 - apical growth with branching
 - presence of plasmodesmata in gametophyte
 - phragmoplast (microtubules and microfilaments directing formation of cell plate during cytokinesis)
- different relationships recovered
 - Charales sister to embryophyta
 - Coleochaetales/Zygnematophyceae sister to embryophyta
- different relationships of bryophytes, esp. position of hornworts
- position of Gnetales (*Gnetum*, *Welwitschia*, *Ephedra*) within gymnosperms

Methods

- 1KP consortium – transcriptomes
- 2x75- or 2x90-bp reads assembled with SOAPdenovo
- proteins from 25 sequenced plant genomes clustered to gene families (OrthoMCL)
- single-copy families identified, aligned (MAFFT), making profile database (HMMER3)
- transcriptomes translated to AA and searched against 25 genome profiles – most transcript sorted into a single family
- transcriptomes aligned and consensus sequence created
- if the consensus contained more than 5% ambiguities, the taxon/gene combination was excluded (duplication assumed)

Phylogenetic analyses

- 852 gene family files aligned with SATé – both AA and DNA
- RAxML gene trees with 200 bootstrap replicates
 - AA alignments (JTT model)
 - DNA alignments (GTR)
 - codon alignments (in-frame DNA)
 - codon alignments with 3rd position removed
- supermatrix (concatenation) – filtering
 - genes with less than 50% of taxa removed
 - sites with more than 50% of missing characters removed
 - genes not including *Chara* removed
 - taxa on very long branches removed
 - extensive trimming (blastp- and branch-length-based, GBLOCKS to remove poorly aligned positions)

Phylogenetic analyses

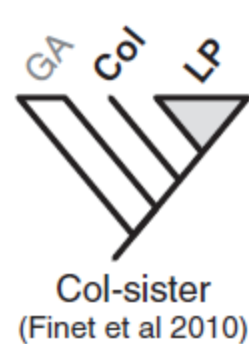
- ML supermatrix – RAxML (GTR for DNA, JTTF for AA), 100 bootstrap
 - unpartitioned
 - partitioned (for codon K-means clustering method used)
- PhyloBayes supermatrix
- coalescent-based analysis (ASTRAL) + multilocus bootstrap
 - all gene trees
 - only gene trees with more than 50% of taxa
 - gene trees after removing sequences with more than 66% gaps
 - gene trees after taxa on long branches removed
 - calculated conflict between species tree and gene trees for each branch
- supertree analysis (Superfine-MRP)

Results

- sequence alignments estimated for 9,610 gene families
- 852 families including at most one gene copy (from at least 24 of the 25 sequenced genomes)
- concatenated untrimmed matrix – 1,701,170 aligned sites
- 69 analyses in total – results highly concordant with ML tree based on 1st and 2nd codon positions
- 3rd codon position – large variation in GC content could lead to model misspecification

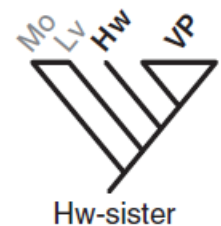
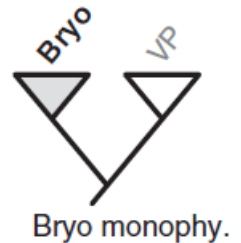
Streptophytic algae and land plants

- Streptophyta monophyletic
- Zygnematophyceae strongly supported as sister lineage of embryophytes – both supermatrix and ASTRAL analyses
- many gene trees with not strong support for hypotheses, small proportion of trees did exhibit well-supported conflict – this is probably due to incomplete lineage sorting of ancestral variation
- phragmoplast – secondary loss in most Zygnematophyceae



Bryophyte relationships

- monophyly of each lineage supported
- liverworts are NOT sister to vascular plants
- 3 alternative hypothesis supported:
 - bryophytes monophyletic in ASTRAL and supertree analyses
 - mosses and liverworts monophyletic
 - hornworts and moss+liverwort clade successively sister to vascular plants in supertree analysis
 - consistent with morphology and development (e.g., pyrenoid shared by hornworts and streptohytic algae)
 - hornworts sister to vascular plants
 - consistent with similarity of gametangia development in hornworts to antheridial/archegonial development in monilophytes

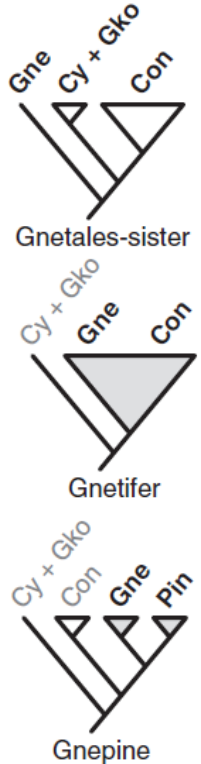


Monilophyte and Lycophyte

- lycophytes and monilophytes are successively sister lineages to the seed plants
- agreement with previous phylogenetic analyses
- resolution of backbone phylogeny of ferns is problematic
- instability in the placement of *Equisetum*

Gymnosperm relationships

- strong monophyly
- Gnetales (*Gnetum*, *Welwitschia*, *Ephedra*) sister to all other lineages only in analyses with all three codon positions
- Gnetales sister to Coniferales – “Gnetifer” hypothesis
 - ASTRAL and supertree analyses
- Gnetales within Coniferales (sister to Pinaceae) – “Gnepine”
 - in supermatrix analyses
 - consistent with previous results
- rapid diversification among Gnetales and two conifer lineages
 - ILS – misleading supermatrix analyses



Angiosperm relationships



Chloranthus

- rapid diversification of flowering plant lineages (Darwin’s “abominable mystery” – resolution of branching remains controversial)
- ANA (*Amborella*-Nymphaeales-Austrobaileyales) grade basal
 - *Amborella* as sister to all other angiosperms
 - Nymphaeales and Austrobaileyales successive sister lineages
- monocots sister to all other
- only PhyloBayes analysis of AA placed magnoliid+Chloranthales sister to eudicot+monocots
- variations in relationships due to
 - model misspecification (simplification)
 - ILS
- increased taxon sampling necessary

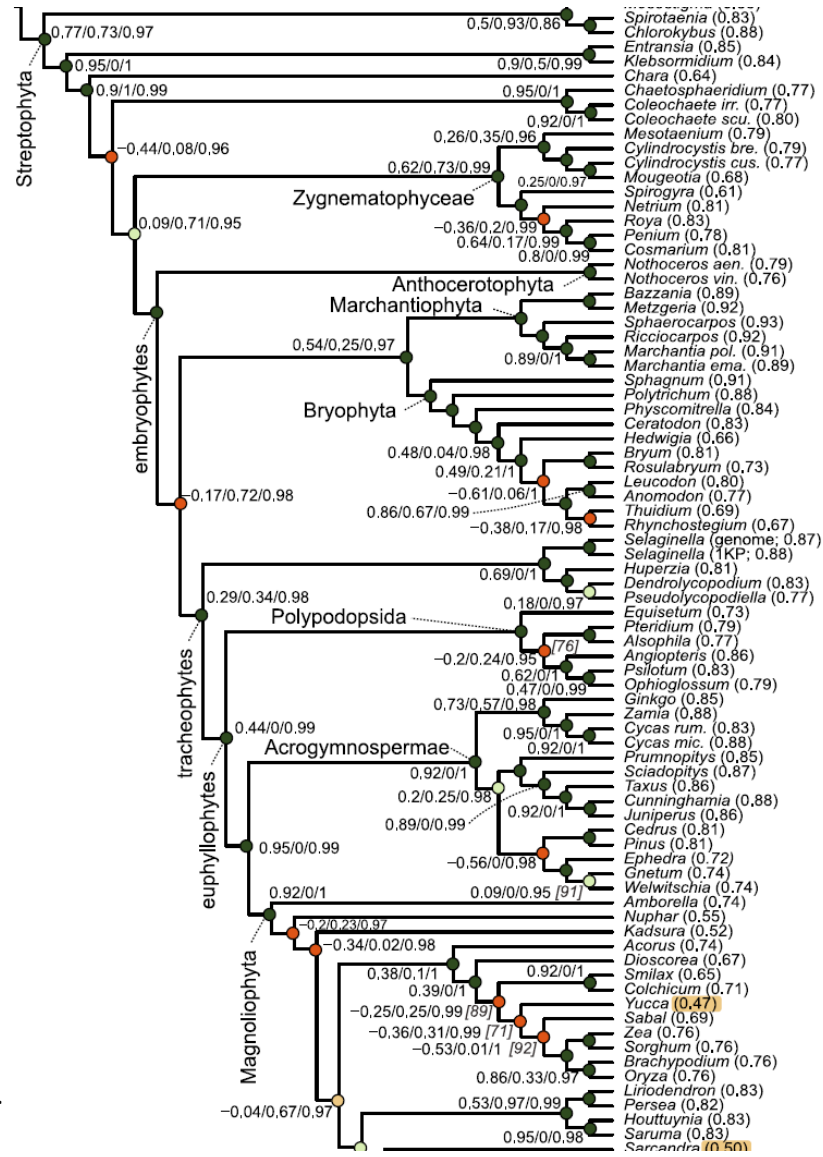


Kadsura
(Austrobaileyales)



Quartet support

Replacement for bootstrap in phylogenomic studies...



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.

Other studies

- Gitzendanner MA et al. (2018): **Plastid phylogenomic analysis of green plants: A billion years of evolutionary history.** *Annals of Botany* 105: 291-301.
- One Thousand Plant Transcriptomes Initiative (2019): **One thousand plant transcriptomes and the phylogenomics of green plants.** *Nature* 574: 679-685
- Zhang J et al. (2020): **The hornwort genome and early land plant evolution.** *Nature Plants* 6: 107-118
- Li FW et al. (2020): ***Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts.** *Nature Plants* 6: 259-272.
- Danyan S et al. (2021): **Large-scale phylogenomic analyses reveal the monophyly of bryophytes and neoproterozoic origin of land plants.** *Molecular Biology and Evolution* 8: 3332-3344.
- Li H-T et al. (2021): **Plastid phylogenomic insights into relationships of all flowering plant families.** *BMC Biology* 19: 232.
- Yang T et al. (2022): **Comparative analyses of 3,654 plastid genomes unravel insights into evolutionary dynamics and phylogenetic discordance of green plants.** *Frontiers in Plant Science* 13: 808156.
- Bowles AMC et al. (2022): **The origin and early evolution of plants.** *Trends in Plant Science*. doi: 10.1016/j.tplants.2022.09.009