Assessment of phylogenetic conflict

Compilation of approaches & methods
(partially based on Joyce et al. 2005)

T. Fér & R. Schmickl
Hyb-Seq Course 2025

Most common sources of phylogenetic conflict

1. Paralogy due to whole genome duplication or gene duplication
2. Hybridization and introgression
3. Deep coalescence, ILS
4. Simultaneous speciation, rapid radiation
5. Methodological artifacts

Conflict assessment

- IQ-TREE - gCF, sCF
- PhyParts
- ASTRAL - alternative hypothese scoring
- BUCKy - concordance/discordance scores
- QuartetSampling - 3 scores per node, 1 score per terminal

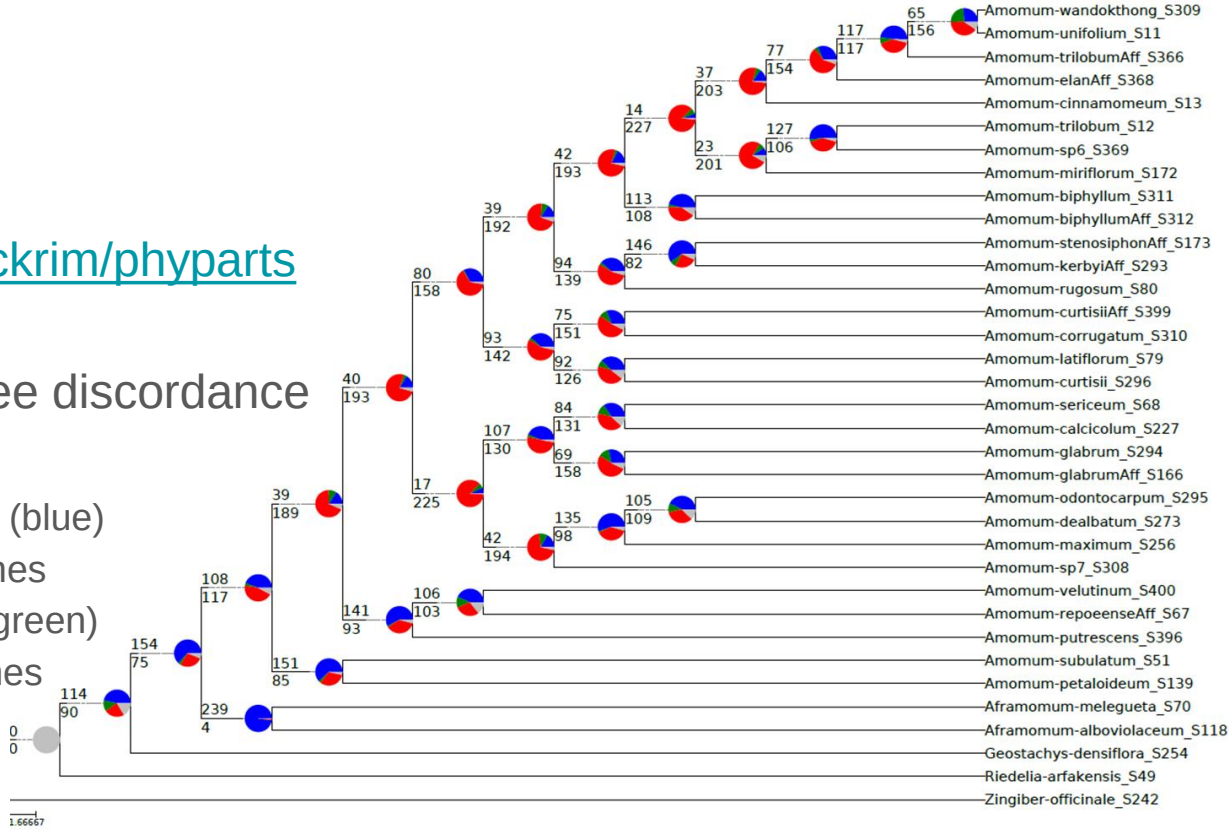
Concordance factors - gCF & sCF

- quantifying genealogical concordance in phylogenomic datasets
- Minh et al. (2020), Lanfear & Hahn (2024)
- implemented in IQ-TREE 2 - for every branch calculates
  - gene concordance factor (gCF) - percentage of “decisive” gene trees containing that branch
  - site concordance factor (sCF) - percentage of decisive alignment sites supporting a branch in the reference tree

https://www.robertlanfear.com/blog/files/concordance_factors.html

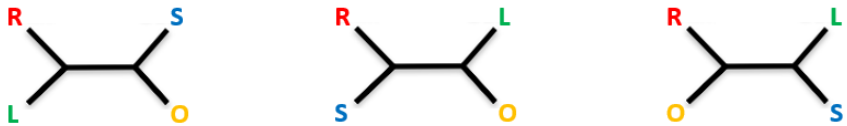
PhyParts

- https://bitbucket.org/blackrim/phyparts
- Smith et al. (2015)
- gene tree vs. species tree discordance
- for every node
  - nr concordant gene trees (blue)
    - value above branches
  - nr discordant trees (red, green)
    - value below branches
- grey – proportion of non-informative genes
- visualization using PhyParts PieCharts (M. Johnson)

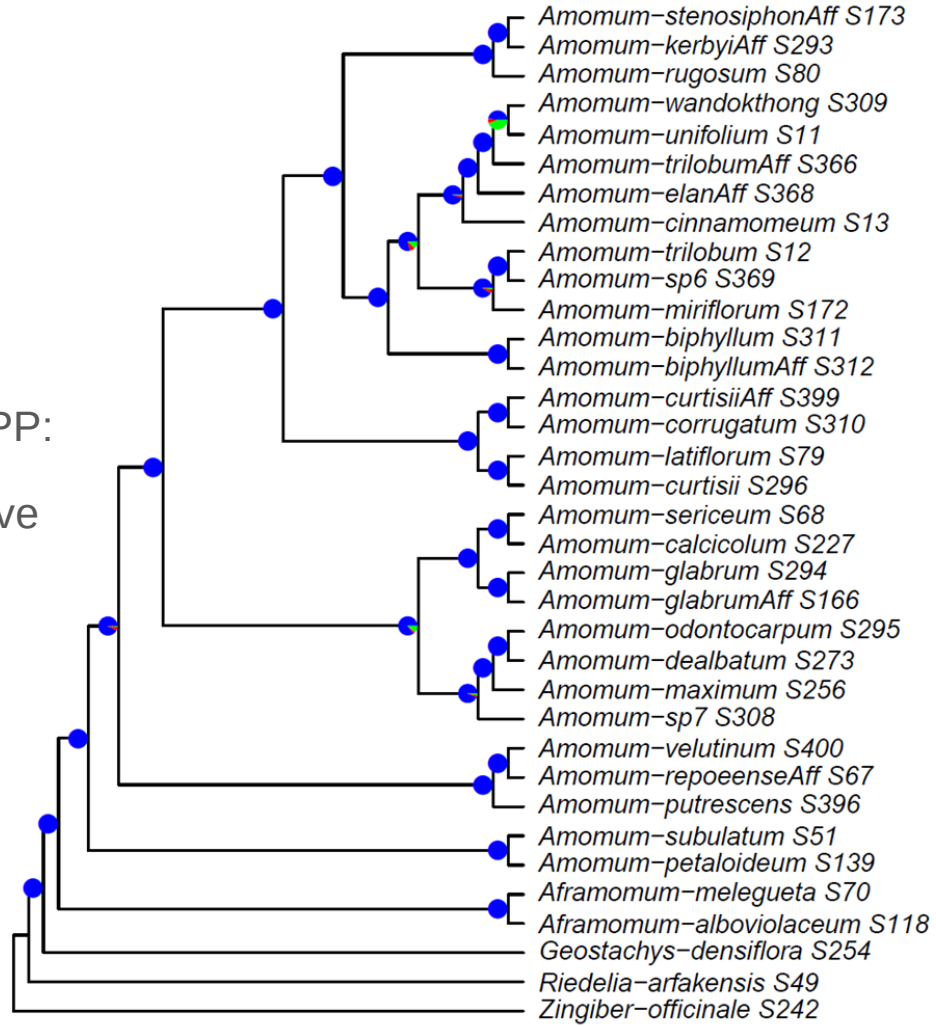


ASTRAL scoring

- *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

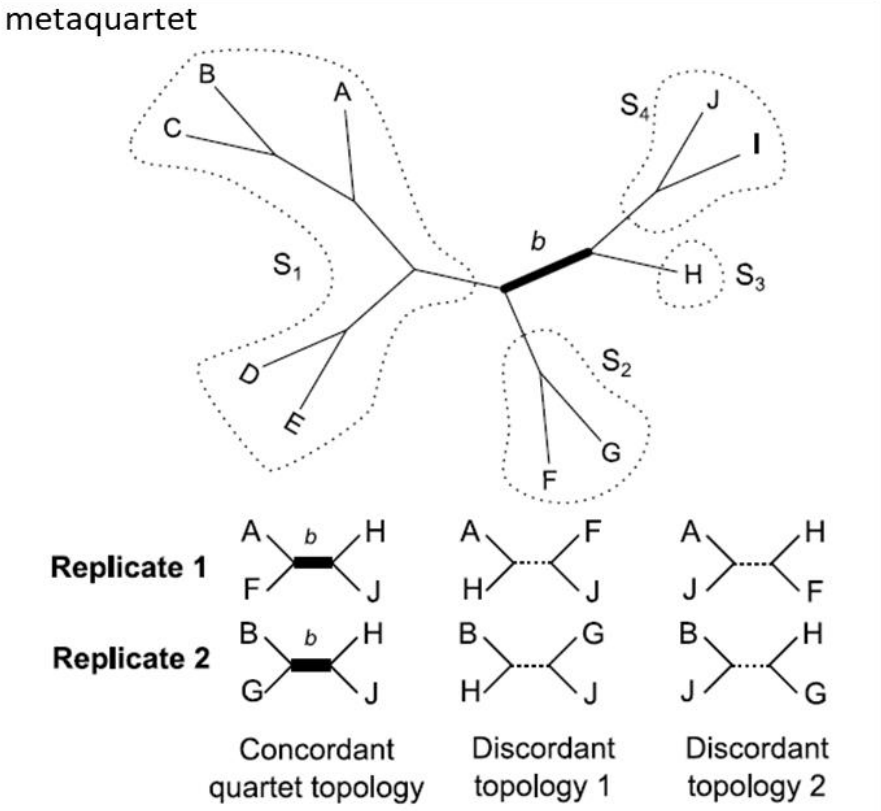


http://tandy.cs.illinois.edu/astral-apro.pdf

Quartet Sampling

- https://github.com/FePhyFoFum/quartetsampling
- Pease et al. (2018)
- evaluates internal branches – likelihood for all three possible phylogenies for the randomly selected quartets spanning particular branch
- quartet-based discordance testing - distinguish conflict from weak support
- takes an existing phylogenetic topology and a molecular dataset

Quartet Sampling



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.

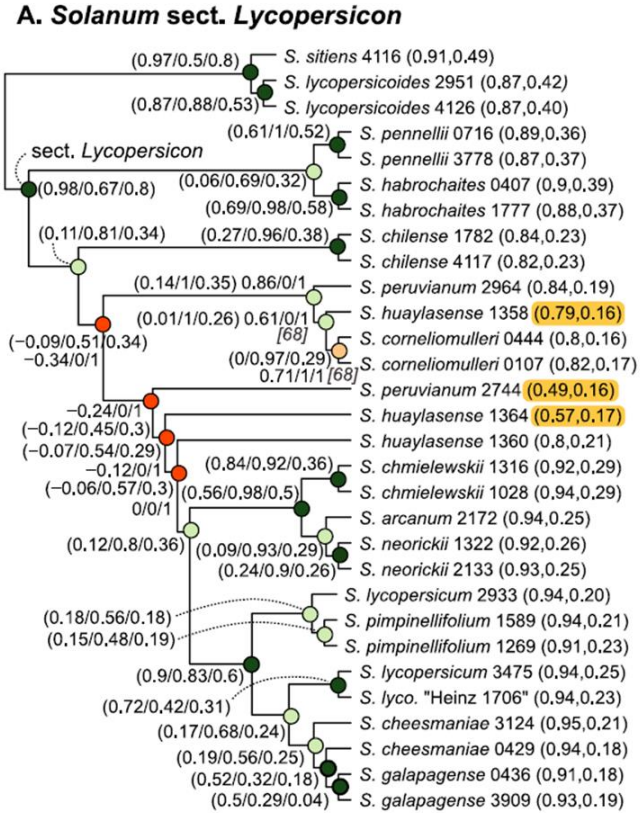
Quartet Differential (QD)
Are discordant #1 and #2 frequencies equal or skewed?
QD=1 → equal #1 and #2
QD=0.3 → skewed
QD=0 → all #1 or #2

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

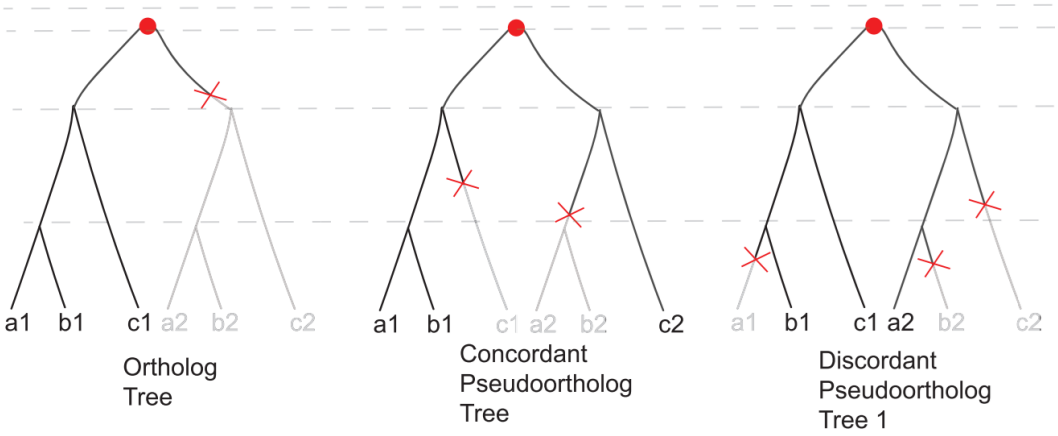
Quartet Sampling



dark green	QC > 0.2
light green	0.2 ≥ QC > 0
orange	0 ≥ QC ≥ −0.05
red	QC < −0.05

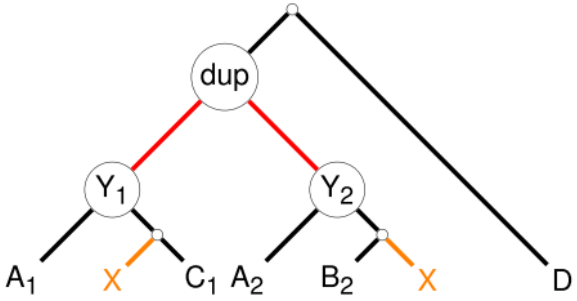
Paralog reconciliation

- paralogy commonly encountered even if "single-copy" loci targeted
- paralogs - gene duplication or WGD followed by divergent evolution, complex pattern of gene duplications & losses
- *hidden paralogs* (pseudo-orthologs) - undetectable (proportion unknown, estimated to be ~10%?), negligible effect on species tree (Smith and Hahn 2022)



Species tree inference (paralog-aware)

- ASTRAL-Pro 3 (Zhang and Mirarab 2022)
  - **ASTRAL** for **P**a**R**alogs and **O**rthologs
  - https://github.com/chaoszhang/ASTER/blob/master/tutorial/astral-pro3.md
  - relatively robust to ILS, GDL and gene tree estimation error
  - https://tandy.cs.illinois.edu/astral-apro.pdf
- FastMulRFS (Molloy and Warnow 2020)
  - https://github.com/ekmolloy/fastmulrfs
  - statistically consistent under a generic model of gene duplication and loss (GDL)
- SpeciesRax (Morel et al. 2022)
  - ML tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss
  - https://github.com/BenoitMorel/GeneRax
- AleRax (Morel et al. 2024)
  - tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss
  - https://github.com/BenoitMorel/AleRax



Gene tree with one duplication and two losses

Approaches for paralogue treatment

remove loci with paralogs - keep single-copy sequences only

mask paralogs - consensus sequences with ambiguities

infer orthologs - keep orthologs from every locus

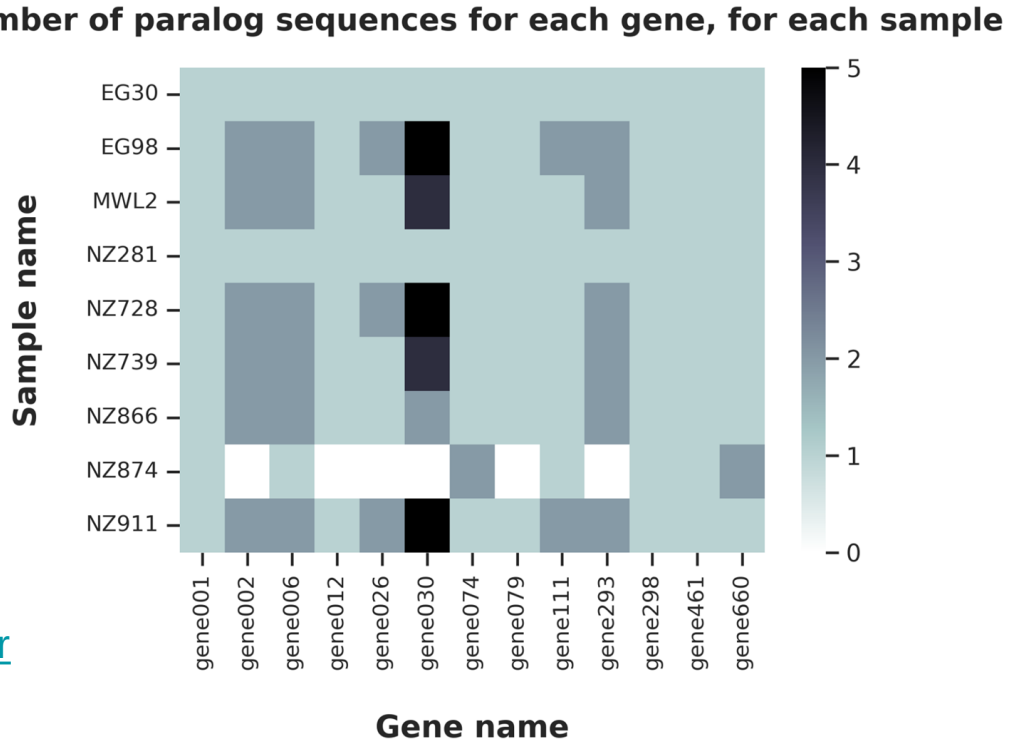
include paralogs - estimate species trees directly with a paralog-aware method

Software dealing with paralogues

- HybPiper - example for an application see Morales-Briones (2021)
- Paralog Wizard (Ufimov et al. 2022)
- PPD (Zhou et al. 2022)
- CAPTUS (Ortiz et al. 2023)
- HybPhaser (Nauheimer et al. 2021)
- ParaGone (Jackson et al. 2023)

HybPiper

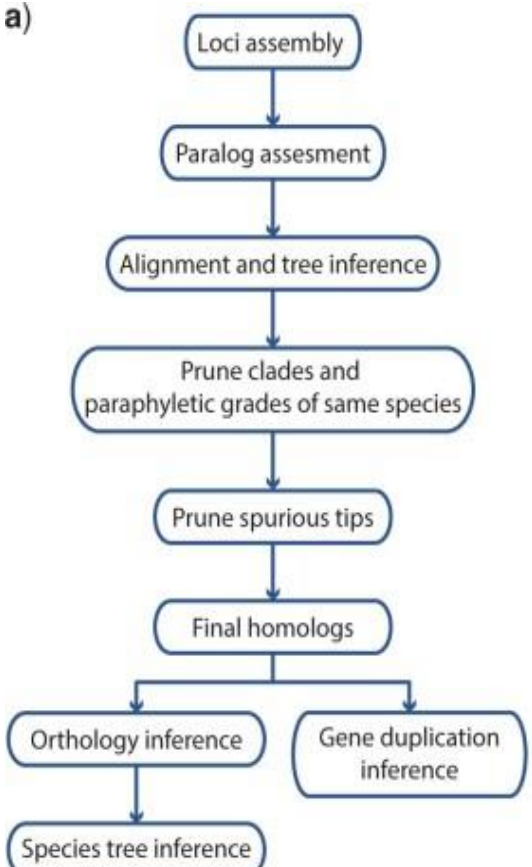
- detects multiple contigs containing long coding sequences - by default at least 75% of the reference sequence
- choice among these competing long contigs by first checking whether one of the contigs has coverage depth that greatly exceeds the others (10x by default). If all competing long contigs have similar depth, the sequence with the greatest percent identity to the reference is chosen.
- https://github.com/mossmatters/HybPiper
- Johnson et al. (2026)

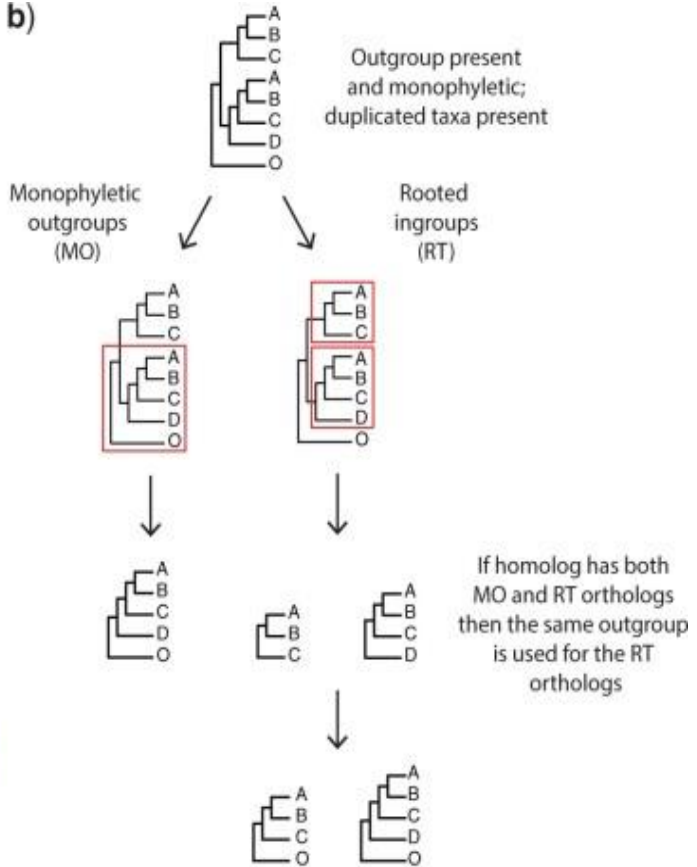


Example application of HybPiper's paralog detection

First, paralogs are detected using HybPiper.

Second, paralog filtering is done via phylogenetic tree building.



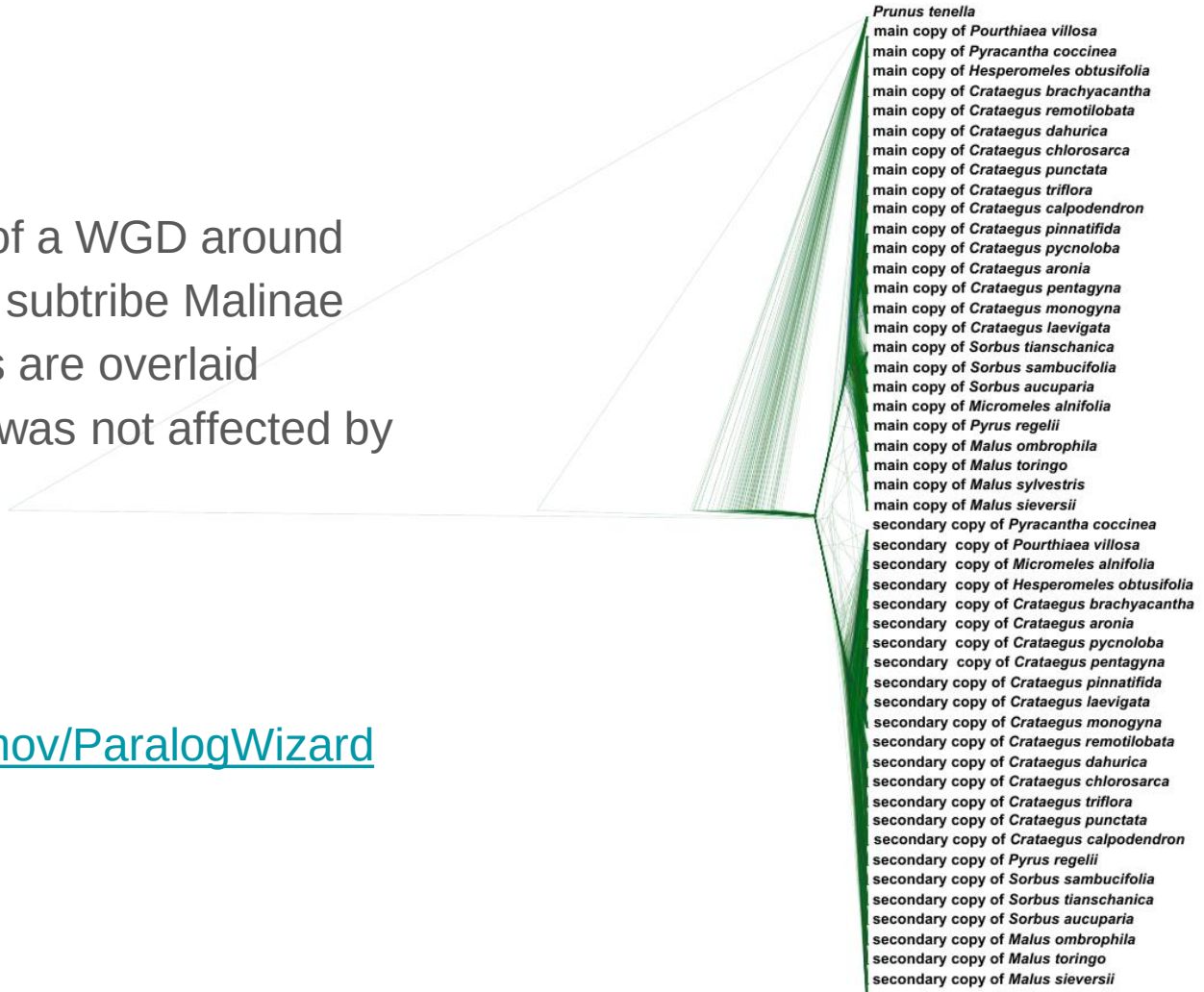


Morales-Briones et al. (2021)

ParalogWizard

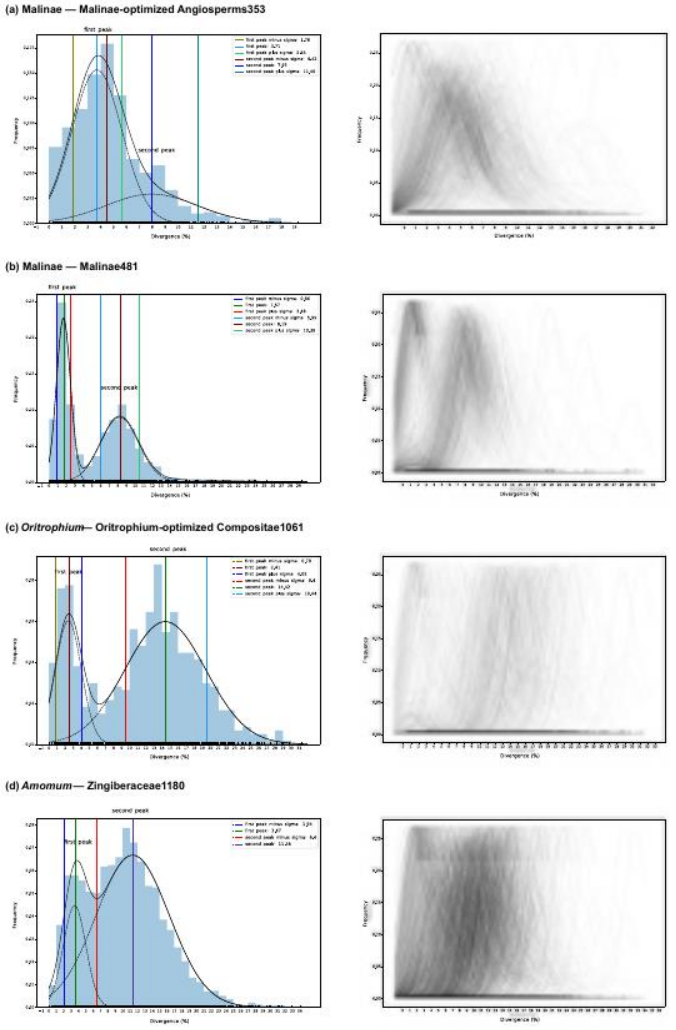
- paralogy as the result of a WGD around the onset of Rosaceae subtribe Malinae
- hundreds of gene trees are overlaid
- note that the outgroup was not affected by the WGD

- Ufimov et al. 2022
- https://github.com/rufimov/ParalogWizard



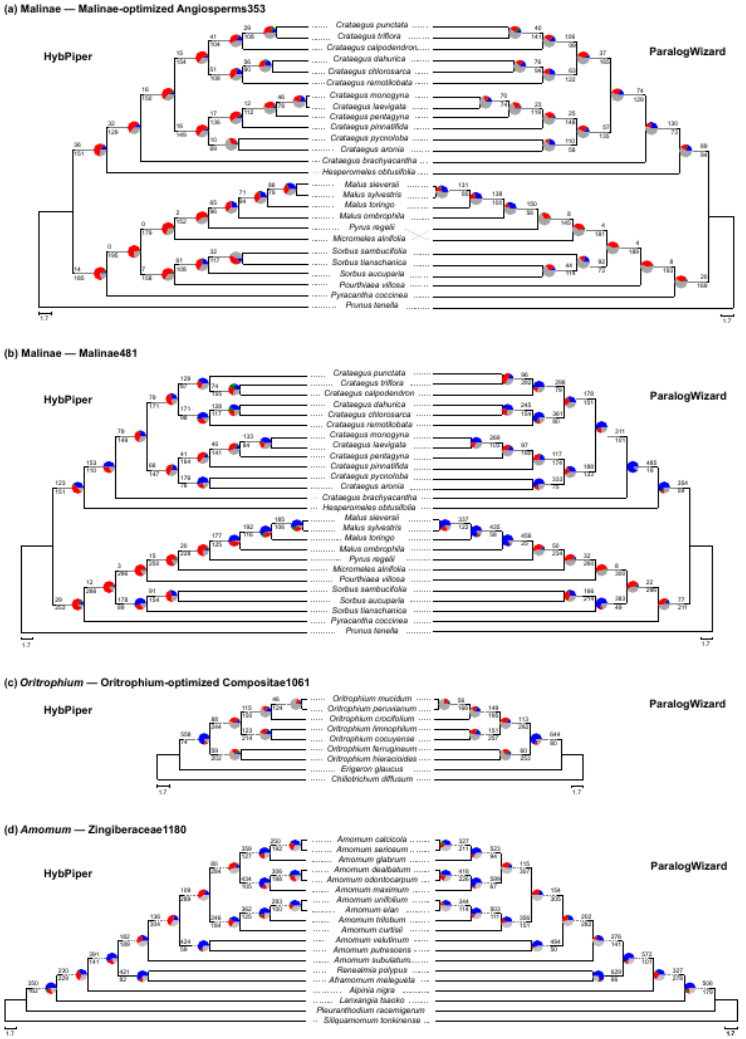
ParalogWizard

Pairwise sequence divergence histograms for different plant groups and probe sets



ParalogWizard

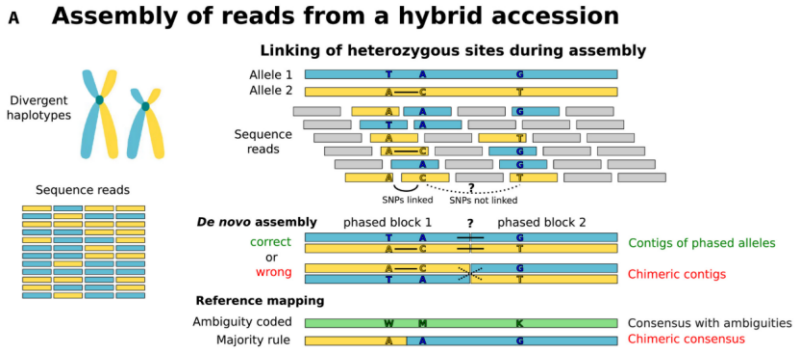
Effect of paralog detection on resolving gene tree conflicts (different plant groups, different probe sets)

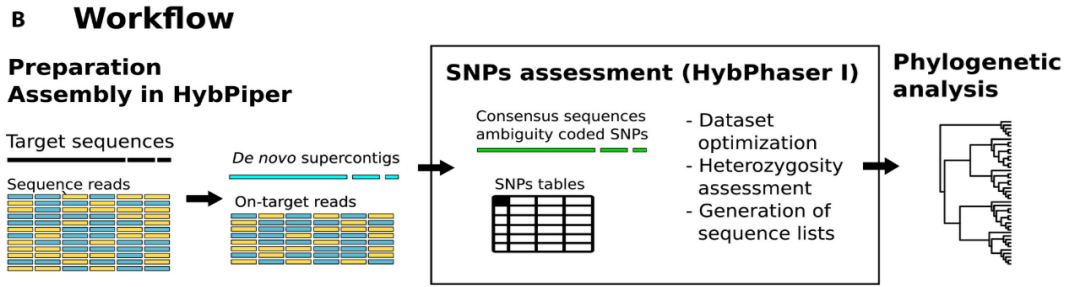


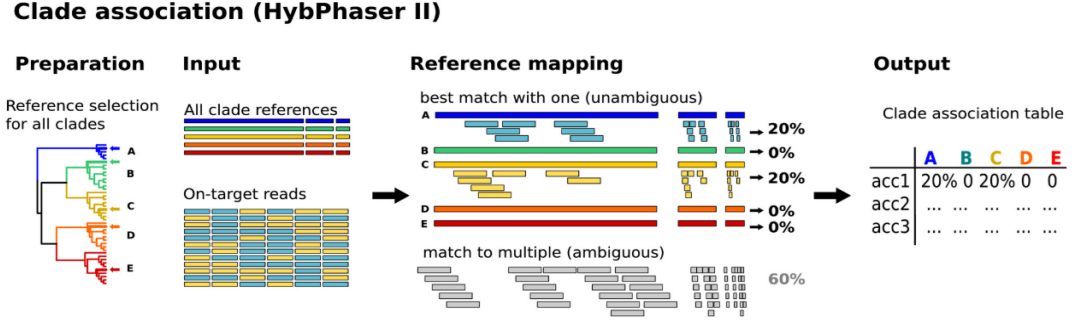
HybPhaser

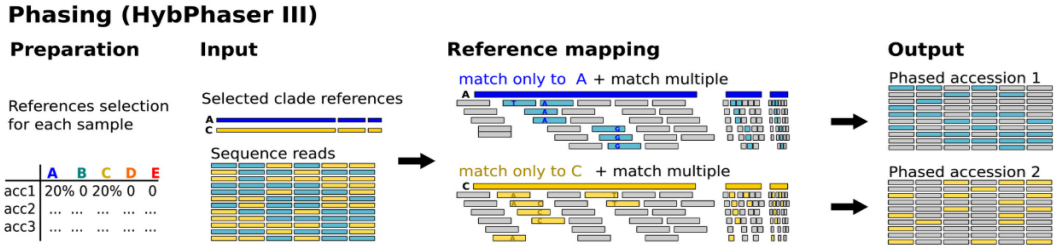
- https://github.com/LarsNauheimer/HybPhaser
- Nauheimer et al. (2021)
- uses HybPiper output
- three phases:
  - (1) assessment of heterozygous sites in assembled sequences to detect putative hybrid accessions
  - (2) creation of a read-to-clade association framework
  - (3) the phasing of read files based on the clade association framework

HybPhaser









[Nauheimer et al. (2021)]

HybPhaser

Sample	LH	AD	perv	maso	dist	khas	dans	graci	bica	ampu	treu	rafl	mira	nort	mapu	rein	camp	grafl	anip	pelt	ceci	raja	bosc	faiz	eppi	voge	veit	glan	maxi	hisp	merr	biak	sibu	vent	mulu	tent	albo	grll	kong	sura	becc	spec	toba	gymn	aris	jamb	CA
angustifolia	97.4	1.51	0.1	0.2	0.2	0.2	0.2	12	0.3	0.1	0.3	0.1	0.2	0.4	0.3	0.2	0.2	0.1	0.2	0.1	0.0	0.2	0.0	0.1	0.0	0.2	0.1	0.1	0.1	1.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.0	0.0	2
ampullaria x gracilis	97.1	1.22	0.2	0.3	0.4	0.3	0.3	7.6	0.6	4.0	0.5	0.2	0.3	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2
kampotiana 1	95.8	1.27	0.1	0.3	0.3	0.3	0.3	0.7	0.4	0.3	0.4	0.5	4.5	0.3	0.3	0.1	0.1	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	1.6	0.1	0.1	0.1	0.0	0.0	0.0	2
rafflesiana x ampullaria 1	95.7	0.98	0.1	0.3	0.3	0.4	0.2	0.9	0.6	3.0	0.6	2.4	0.5	0.2	0.2	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	2
clipeata x ventricosa	95.6	1.16	0.1	0.1	0.1	0.1	0.2	0.2	0.0	0.0	0.2	0.0	0.1	0.2	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.5	0.1	0.3	0.2	0.0	0.3	0.3	0.7	5.0	0.0	0.1	0.1	0.0	0.1	0.2	0.1	0.1	0.0	0.0	0.0	0.0	2
ampullaria x rafflesiana	95.3	1.05	0.0	0.2	0.3	0.2	0.2	0.4	0.5	5.9	0.4	4.7	0.9	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.0	2
izumiae x ventricosa	95.1	1.08	0.0	0.1	0.1	0.1	0.2	0.2	0.0	0.0	0.2	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.2	0.7	4.8	0.0	0.0	0.0	0.1	0.2	0.1	0.2	0.2	0.3	0.4	0.2	1.3	2
gymnamphora 2	94.9	0.42	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.0	0.2	0.0	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.3	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.2	0.4	0.4	5.3	0.4	1.4	2
rafflesiana x ampullaria 2	94.6	1.24	0.0	0.1	0.3	0.2	0.2	0.7	1.2	14	1.3	15	1.7	0.3	0.2	0.1	0.1	0.1	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.1	0.2	0.1	0.1	0.2	0.1	0.0	0.1	0.1	0.0	0.1	0.1	2
lowii x campanulata	94.3	0.87	0.0	0.1	0.1	0.1	0.2	0.1	0.1	0.0	0.1	0.0	0.1	0.1	0.1	0.1	6.9	0.1	0.1	0.0	0.1	0.3	0.1	0.5	0.4	0.5	0.2	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.0	2
bokorensis x ventricosa	93.9	1.12	0.1	0.2	0.3	0.2	0.3	0.6	0.3	0.2	0.3	0.1	0.1	0.4	0.4	0.2	0.2	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.5	0.4	0.5	2.0	0.0	0.0	0.0	0.2	0.3	1.7	0.1	0.1	0.0	0.0	0.0	0.0	2
ampullaria x tobaica	93.8	1.13	0.0	0.2	0.1	0.1	0.2	0.3	0.2	5.4	0.3	0.2	0.2	0.2	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.2	0.2	0.8	0.1	0.4	2.7	0.1	0.4	0.1	2

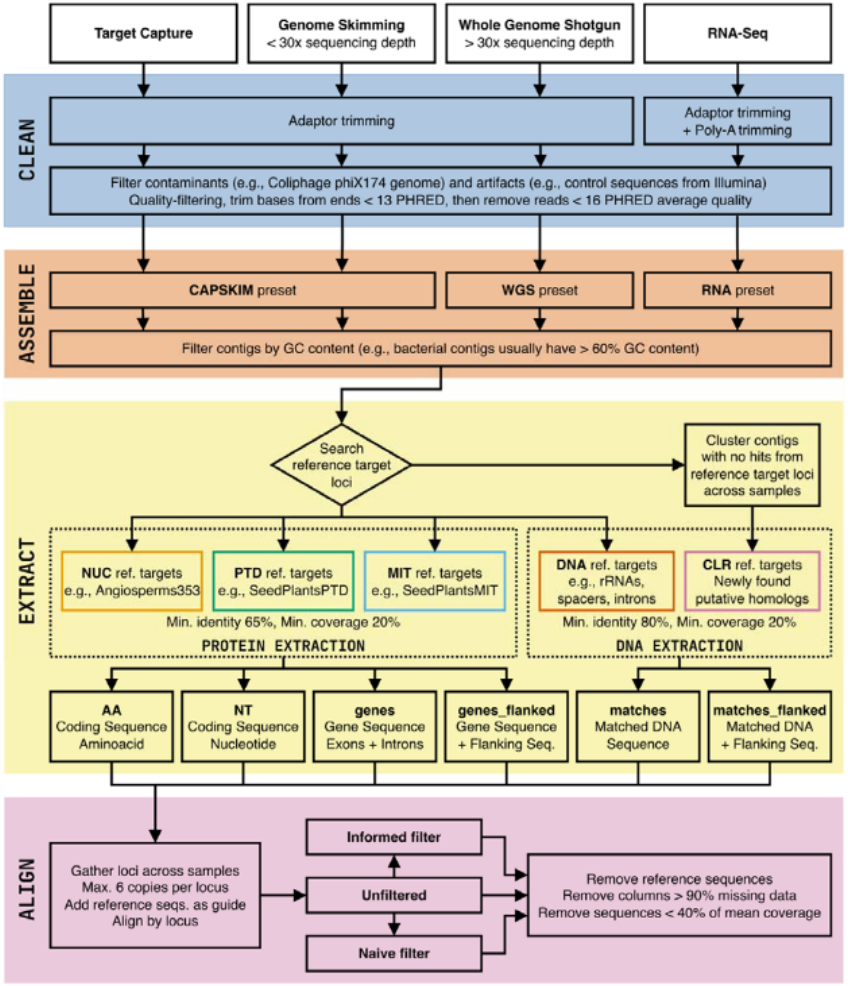
[Nauheimer et al. (2021)]

CAPTUS

By default, CAPTUS collects a maximum of five paralogs per sample and per marker.

Ortiz et al. (2023), Raza et al. (2023)

https://github.com/edgardomortiz/Captus



Hybridization (introgression) = speciation?

- **hybrid speciation** - when hybridization between two distinct species takes place for many generations with fertile offspring; especially when the hybrid species is able to take advantage of a niche that is distinct to that occupied by both parental species
- **introgressive hybridization (introgression)** - was introduced by Anderson & Hubricht (1938), who referred to the introduction of syntenic nucleotide variation by recombination from a donor species into the genome of a recipient species, usually by means of hybridization and backcrossing

Hybridization and phylogenetic inference

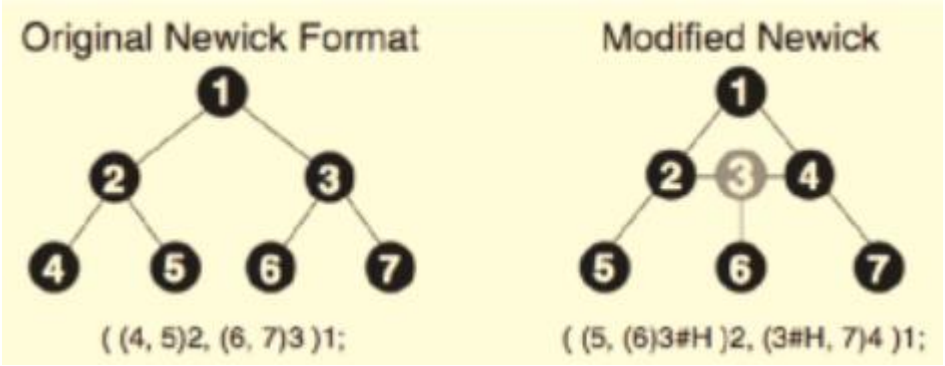
- introgression complicates phylogenetic inference in that **introgression/ introgressed taxa do not follow the assumption of evolution from a common ancestor through a bifurcating process**
- incongruence among single-gene phylogenies is commonly observed

Conflict investigation

- PhyloNet
- PhyloNetworks
- Dsuite
- HyDe
- JML
- QuIBL
- ASTRAL - hard polytomy test

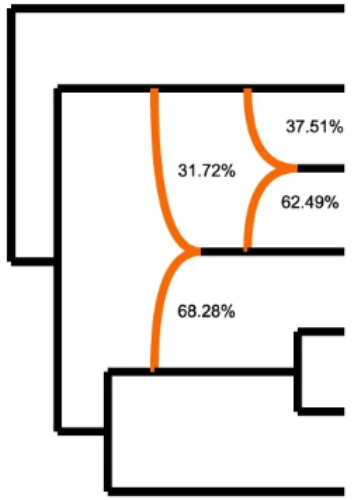
Explicit phylogenetic networks (e.g, PhyloNet)

- evolutionary networks (rooted)
- PhyloNet aims at evaluating networks
- **eNewick format** - a modified Newick format for representing evolutionary networks. The eNewick format is nothing but a collection of trees in the Newick format. In PhyloNet rooting is assumed, since different ways of rooting the same evolutionary network may imply different evolutionary relationships.

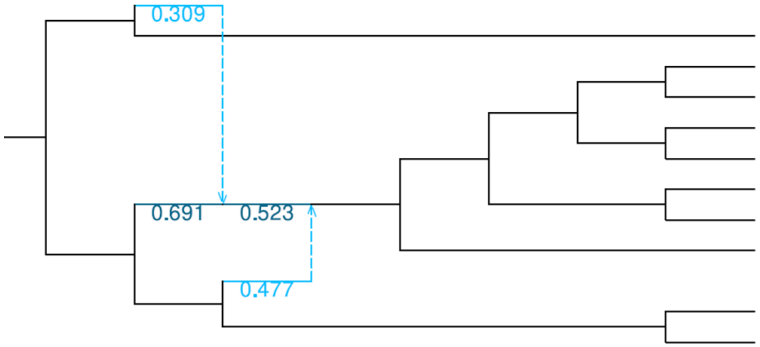


PhyloNet

- [Than et al. (2008)]
- https://phylogenomics.rice.edu/html/phylonet.html
- maximum pseudolikelihood
- selection of best network based on AIC
- plotted using
  - Julia PhyloNetworks package
  - Dendroscope

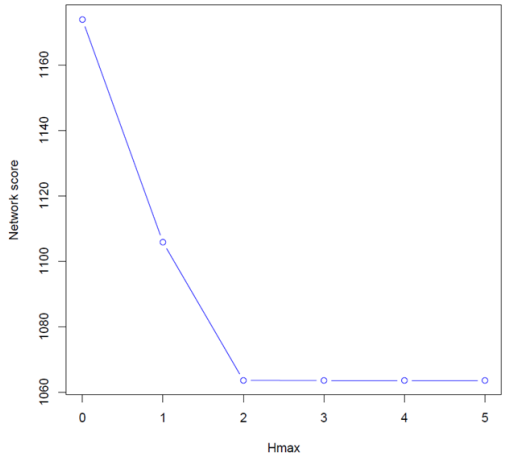


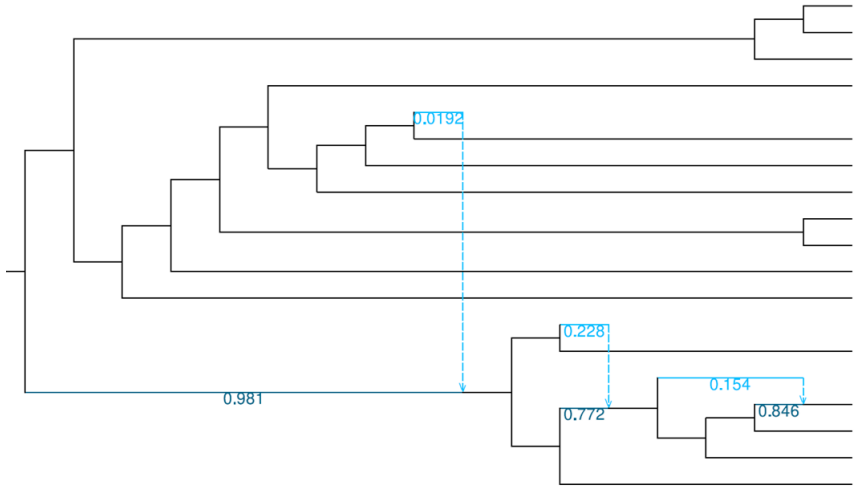
NrHybridizations	NrBranches	NrGeneTrees	logLikelihood	ki	AIC	deltaAIC
4	**21**	**1106**	**-1077277.803**	**1131**	**2156817.606**	**0**
3	23	1106	-1077306.016	1132	2156876.032	58.4261
3	21	1106	-1077318.048	1130	2156896.095	78.4898
3	21	1106	-1077337.985	1130	2156935.97	118.364
0	21	1106	-1077349.118	1127	2156952.236	134.631
1	21	1106	-1077348.539	1128	2156953.077	135.472
2	21	1106	-1077348.583	1129	2156955.166	137.56
2	21	1106	-1077349.405	1129	2156956.811	139.205
1	23	1106	-1077349.273	1130	2156958.546	140.94
4	21	1106	-1077349.233	1131	2156960.467	142.861
5	21	1106	-1077349.892	1132	2156963.784	146.178



SNaQ (PhyloNetworks)

- Species Networks applying Quartets
- Solís-Lemus & Ané (2016)
- PhyloNetworks Julia package (https://github.com/JuliaPhylo/PhyloNetworks.jl)
- best network based on log likelihood increase

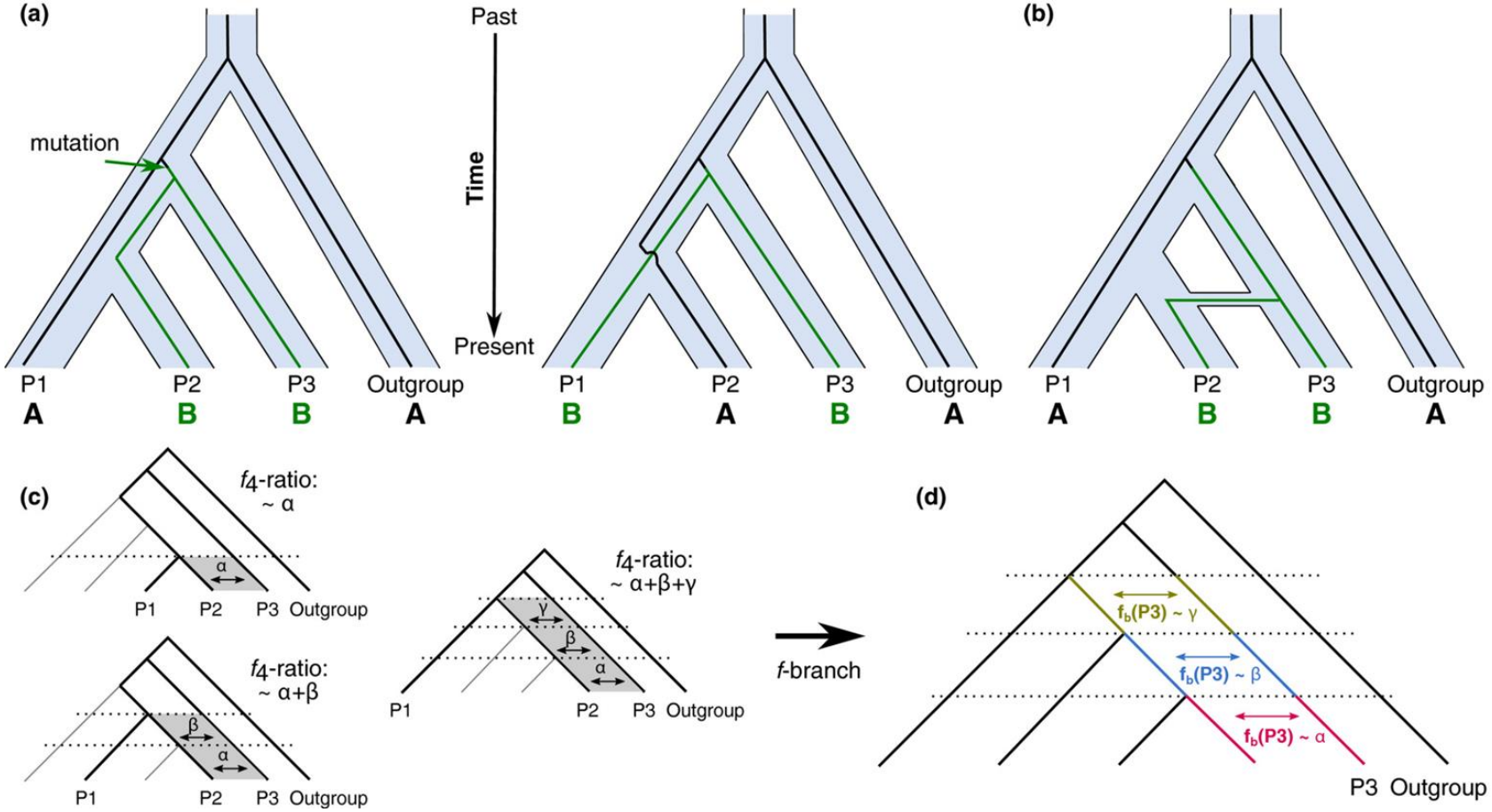




Dsuite

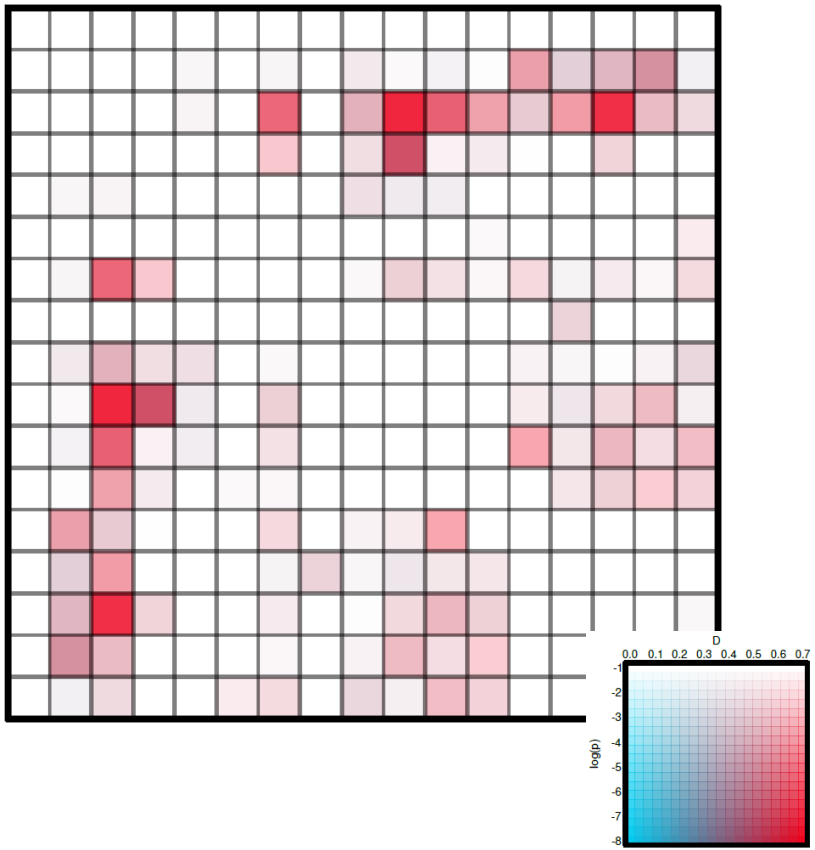
- Malinsky et al. (2021)
- https://github.com/millanek/Dsuite
- admixture/hybridization evidence from variable sites (SNPs) in the genome
- for all possible trios of samples/species (Dtrios command)
  - Patterson's D (also known as ABBA-BABA statistics)
  - estimate of admixture fraction *f* (referred to as the *f*4-ratio)
  - D statistic and f4 ratio belong to a class of methods based on studying correlations of allele frequencies across populations; were developed within a population genetic framework (Patterson et al., 2012)
  - the use of the *D* and f_4-ratio statistics involves fitting a simple explicit phylogenetic tree model to a quartet of populations or species
- D and f statistics are calculated for branches on a tree that relates the samples/species (Fbranch command)

Dsuite

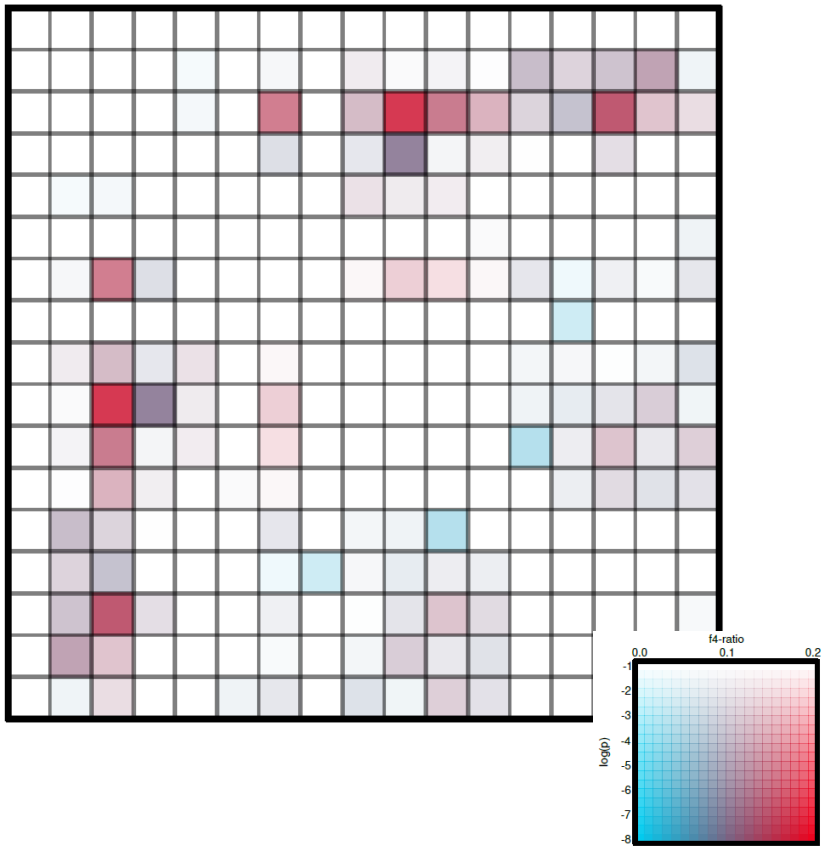

(a)
mutation
Past
Time
Present
P1
P2
P3
Outgroup
A
B
B
A
(b)
B
A
B
A
A
B
B
A
(c)
f4-ratio:
~ α
α
f4-ratio:
~ α+β
β
α
f4-ratio:
~ α+β+γ
γ
β
α
f-branch
(d)
fb(P3) ~ γ
fb(P3) ~ β
fb(P3) ~ α

Dsuite

BBAA

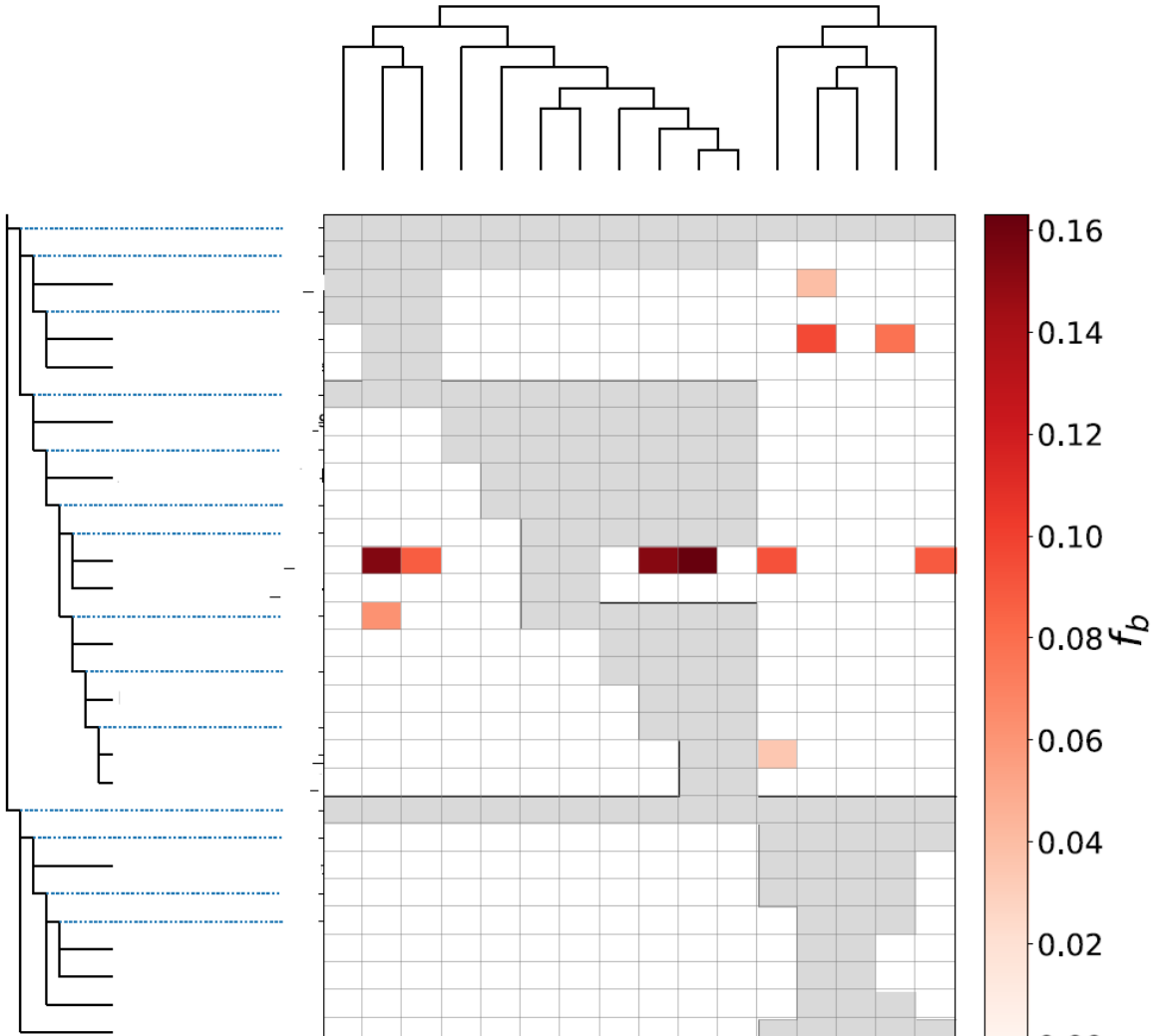


f4 ratio



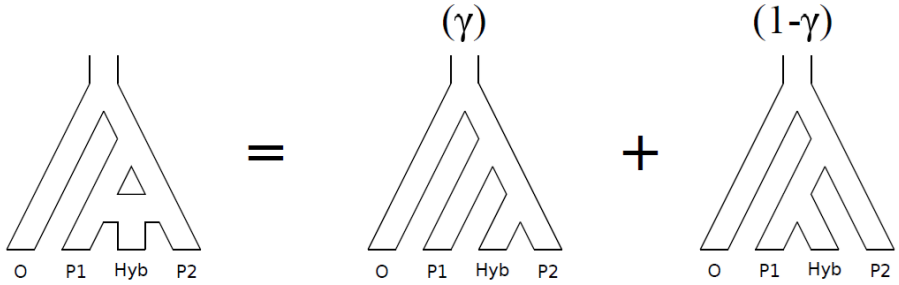
Dsuite

gene flow



HyDe

- Python package for genome-scale **Hy**bridization **De**tection
- https://github.com/pblischak/HyDe; Blischak et al. (2018)
- detecting hybridization using *phylogenetic invariants* arising under the coalescent model with ILS and hybridization
- test for admixture on rooted, four- or five-taxon trees using genome-wide SNP data ("ABBA-BABA"-like methods, *D*-statistic)
- *phylogenetic invariants* - functions of phylogenetic model parameters (e.g. site pattern probabilities) whose expected difference is always 0



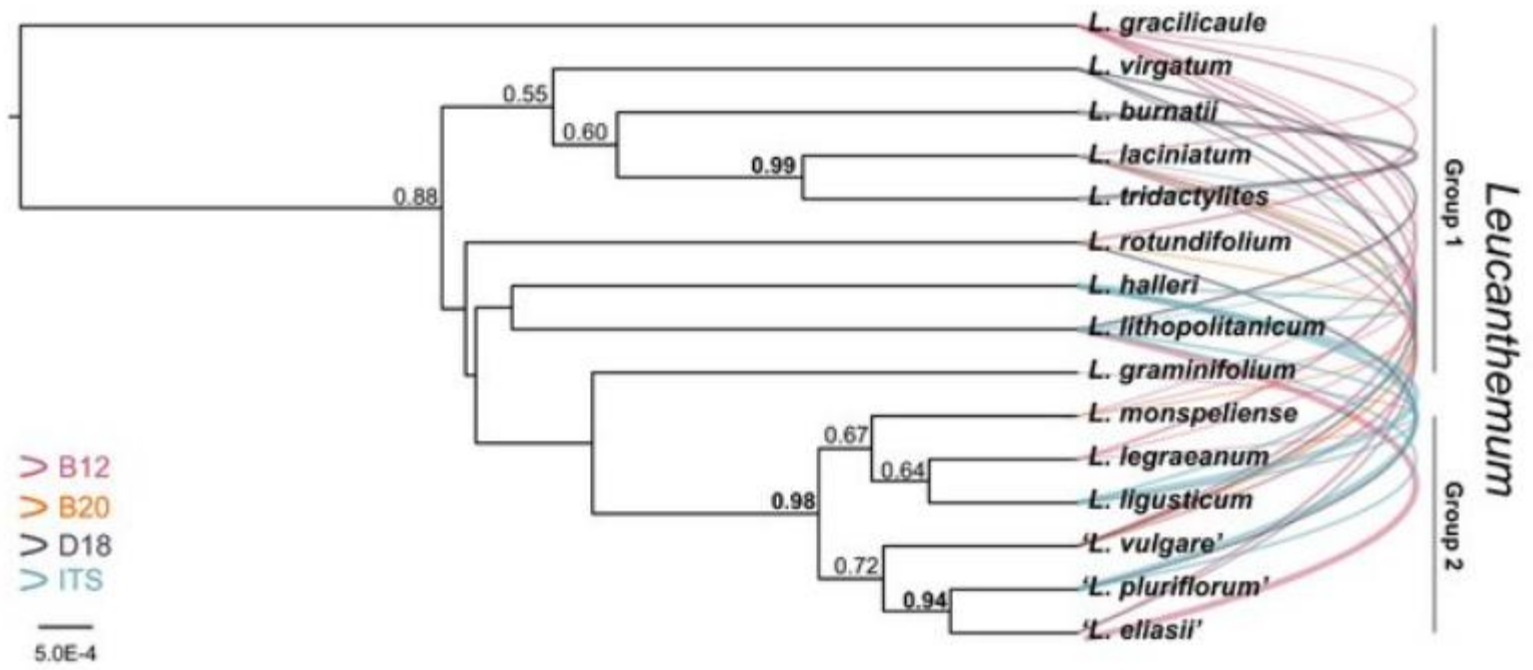
hybrid

(sp1, sp2, sp3)	
Zscore	8.179342
Pvalue	2.22e-16
Gamma	0.4594077

JML

- testing hybridization using species trees
- test whether the minimum distance between sequences of two species is smaller than expected under a scenario without hybridization
    - observed distance compared to simulated values
    - if the distance is smaller than 95% of the simulations than ILS cannot explain the data and hypothesis of hybridization can be accepted
- https://github.com/simjoly/jml; Joly et al. (2009), Joly (2011)
- input is *BEAST species tree file (MCMC sample)

JML



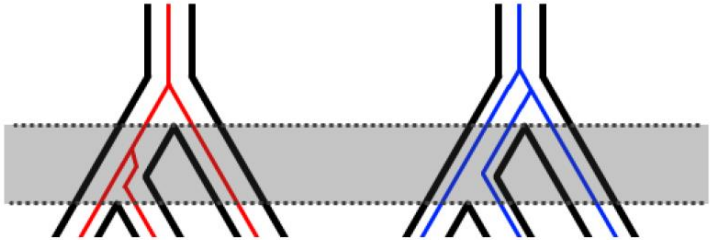
Wagner et al. 2019

QuIBL

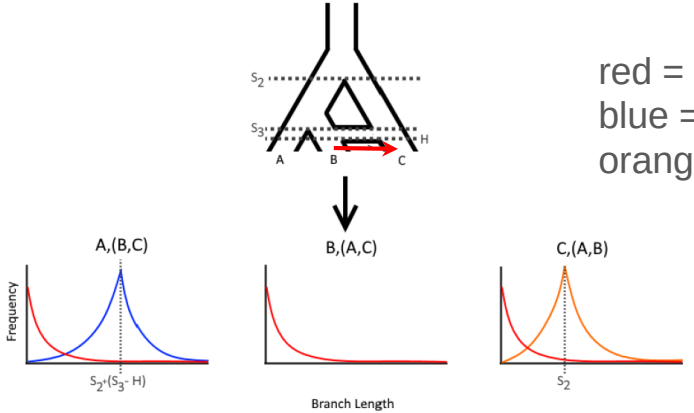
- **Qu**antifying **I**ntrogression via **B**ranch **L**engths
- https://github.com/miriammiyagi/QuIBL; Edelman et al. (2019)
- distinguish between ILS and introgression - based on the distribution of internal branch lengths among loci for a given three-taxon subtree, conditional on its topology
- in the *absence of introgression* - internal branch lengths of triplet topologies discordant with the species tree (due to ILS) expected to be exponentially distributed
- if *introgression has occurred* - their distribution should have that same exponential component plus an additional component with a non-zero mode (corresponding to the time between the introgression event and the most recent common ancestor of all three species)
- compare likelihood of the two scenarios - BIC test with a cutoff of $\Delta BIC > 10$

QuIBL

- *absence of introgression* - internal branch lengths of triplet topologies discordant with the species tree (ILS) expected to be exponentially distributed
- *introgression* - their distribution should have that same exponential component plus an additional component with a non-zero mode (corresponding to the time between the introgression event and the most recent common ancestor of all three species)
- compare likelihood of the two scenarios - BIC test with a cutoff of $\Delta BIC > 10$



window where coalescence occur



red = ILS
blue = introgression
orange = speciation

QuIBL

Table S13: QuIBL results

triplet	outgroup	C1	C2	topology ILS proportion	topology non-ILS proportion	numTrees	BIC ILS+Introgression	BIC ILS only	dBIC	total non-ILS
Hdem_HeraRef_Hhim	HeraRef	0	2.057531	0.793465	0.206535	71	-361.505298	-370.06882	8.563526	0.002618569
Hdem_HeraRef_Hhim	Hhim	0	15.227301	0.983312	0.016688	61	-294.136955	-284.72458	-9.412373	0.00018178
Hdem_HeraRef_Hhim	Hdem	0	1.125482	0.010098	0.989902	5469	-29597.94544	-25645.684	-3952.261131	0.966745364
Hdem_HeraRef_Htel	HeraRef	0	0.359318	0.197908	0.802092	1746	-12345.2966	-12113.84	-231.456445	0.250080827
Hdem_HeraRef_Htel	Hdem	0	0.443302	0.136761	0.863239	2816	-19898.97183	-19384.341	-514.630632	0.434085897
Hdem_HeraRef_Htel	Htel	0	0.487047	0.278998	0.721002	1039	-7396.028943	-7289.867	-106.161934	0.133771621

triplet: The three-taxon subset considered. Species abbreviations separated by underscores.
Outgroup: Species inferred to be the outgroup in the triplet gene tree topology tested.
Cx: Inferred species tree branch length for (1) the ILS case and (2) the non-ILS case. The ILS case is forced to be 0, as all lineages must be in the same population.
Topology proportions: Inferred mixture proportion for the ILS and non-ILS distributions. These values sum to 1.
numTrees: Frequency of the topology in the sample.
BICx: Raw BIC values for each model
dBIC: difference in BIC value between the models. dBIC < -10 implies that the ILS+introgression model is a better fit for the data.
total non-ILS: topology non-ILS proportion * (numTrees/total trees in sample). This value represents the genome-wide introgression fraction

ILS, deep coalescence

- stochastic inheritance of ancestral alleles after split (due to genetic drift)
- allele sorting depends on
  - population size
  - time (number of generations) since lineage split
- especially problematic in species with large population sizes

Simultaneous speciation, rapid radiation

- *simultaneous speciation* (separation into more than two lineages) - **hard polytomy** - multifurcations
- if multifurcations forced to be represented as bifurcations - random pattern in gene trees
- *rapid radiations* - bifurcations after short time (few mutations, short branches) - **soft polytomies** (lack of information)