HybPhyloMaker

https://github.com/tomas-fer/HybPhyloMaker

Tomáš Fér & Roswitha Schmickl

Dept. of Botany, Charles University, Prague

June 2026

Hyb-Seq data analysis software

- PHYLUCE
  - software for UCE (and general) phylogenomics
  - UCE – ultraconserved elements (http://ultraconserved.org)
  - Faircloth (2016): *PHYLUCE is a software package for the analysis of conserved genomic loci*. Bioinformatics 32:786-788.
  - https://github.com/faircloth-lab/phyluce
- HybPiper
  - Johnson et al. (2016): HybPiper: *Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment*. Applications in Plant Sciences 4(7): 1600016
  - https://github.com/mossmatters/HybPiper
  - allows analysis of intronic regions, putative paralog flagging
  - de novo assembly of each locus
- **HybPhyloMaker**
  - Fér & Schmickl (2018): *HybPhyloMaker: target enrichment data analysis from raw reads to species trees*. Evolutionary Bioinformatics 14: 1-9.
  - https://github.com/tomas-fer/HybPhyloMaker
  - complete solution from raw reads to species trees
  - mapping to the reference

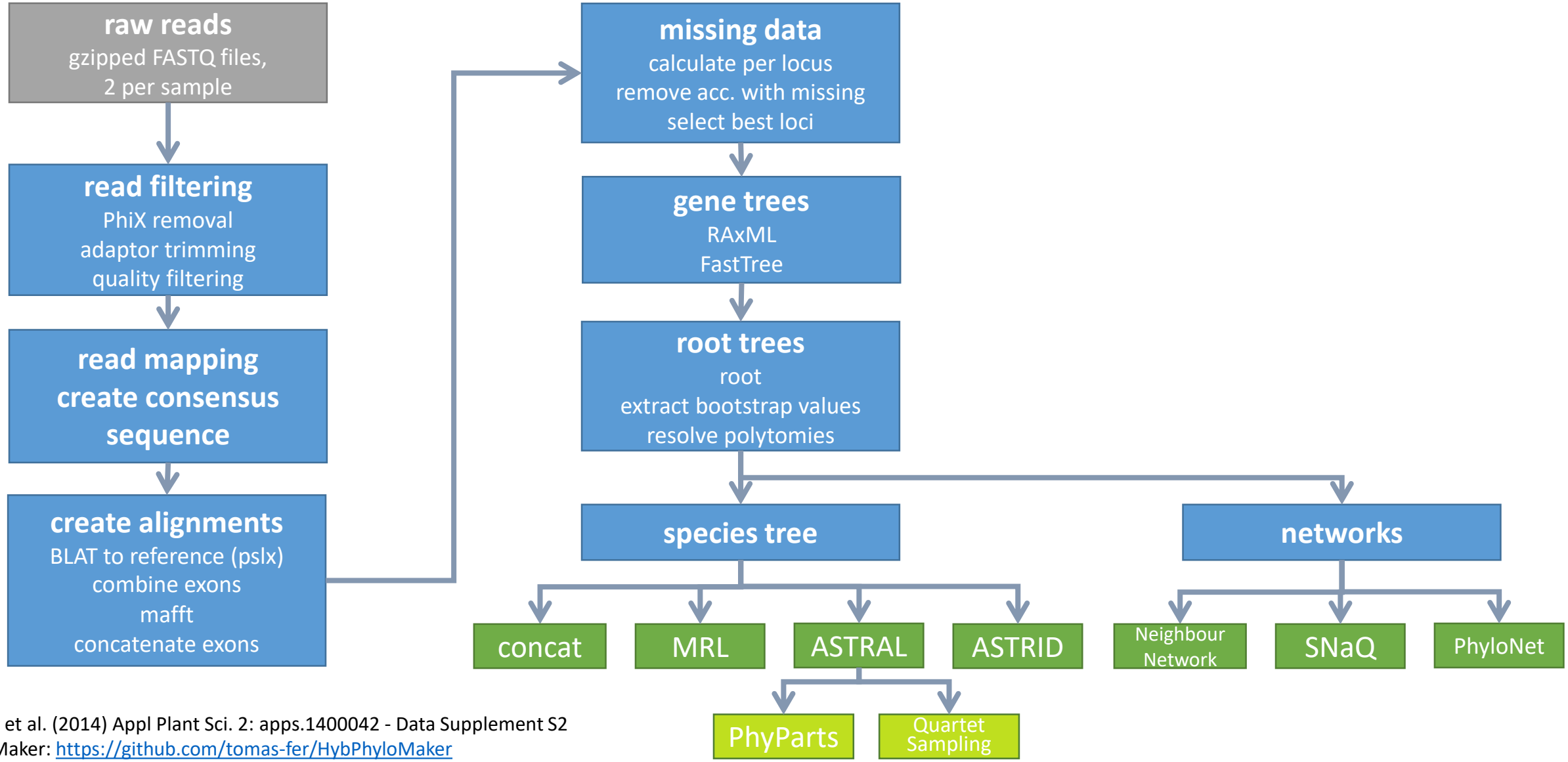
Hyb-Seq data analysis software 2

- aTRAM
  - automated Target Restricted Assembly Method
  - Allen et al. (2015): *aTRAM - automated target restricted assembly method a fast method for assembling loci across divergent taxa from next-generation sequencing data*. BMC Bioinformatics 16:98
  - https://github.com/juliema/aTRAM
- SECAPR
  - Andermann et al. (2018): *SECAPR—a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments*. PeerJ 6:e5175
  - https://github.com/mossmatters/HybPiper
  - de novo assembly of reference, reference based assembly, allele phasing
- reads2trees
  - Heyduk et al. (2016): *Phylogenomic analyses of species relationships in the genus* Sabal *(Arecaceae) using targeted sequence capture*. Biological Journal of the Linnean Society 117:106–120
  - https://github.com/kheyduk/reads2trees
  - de novo assembly approach

Hyb-Seq data analysis software 3

- CAPTUS
  - toolkit for the assembly of phylogenomic datasets from HTS data
  - Ortiz et al. (2023): *A novel phylogenomics pipeline reveals complex pattern of reticulate evolution in* Cucurbitales. bioRxiv. https://doi.org/10.1101/2023.10.27.564367
  - https://github.com/edgardomortiz/Captus
  - Hyb-Seq, genome skimming, RNA-seq, Whole Genome Sequencing
- HybSuite
  - an integrated pipeline for hybrid capture phylogenomics from reads to trees
  - Liu et. al. (2026): *HybSuite: An integrated pipeline for hybrid capture phylogenomics from reads to trees*. Applications in Plant Sciences 14: e70059.
  - https://github.com/Yuxuanliu-HZAU/HybSuite

Hyb-Seq data analysis pipeline (HybPhyloMaker)



Weitemier et al. (2014) Appl Plant Sci. 2: apps.1400042 - Data Supplement S2
HybPhyloMaker: https://github.com/tomas-fer/HybPhyloMaker

Raw reads filtering (script 1)

parallelized (one job per sample – scripts 1a and 1a2)

- PhiX removal
  - ssDNA of phi X 174 bacteriophage
  - balance base pattern of the genome (95% belongs to coding genes)
  - spike-in control for alignment calculations and quantification efficiency
- trimming (Trimmomatic) – adaptor & low quality
  - ILLUMINACLIP:../NEBNext-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:36
    - remove adapters (ILLUMINACLIP:NEBNext-PE.fa:2:30:10)
    - remove leading low quality or N bases (below quality 20) (LEADING:20)
    - remove trailing low quality or N bases (below quality 20) (TRAILING:20)
    - scan the read with a 5-base wide sliding window, cutting when the average quality per base drops below 20 (SLIDINGWINDOW:5:20)
    - drop reads below the 36 bases long (MINLEN:36)
- duplicate removal (fastuniq – https://sourceforge.net/projects/fastuniq/)
- `20filtered` folder created
  - paired/unpaired fastq.gz files with/without duplicates
  - reads_summary.txt

Sample no.	Genus	Species	Nr. of pairs	Nr. of reads	Nr. of reads without PhiX	Both surviving	Nr. reads after quality trimming	% quality trimmed reads	Nr. reads without duplicates	% duplicates
S118	Aframomum	alboviolaceum	225732	451464	451464	225194	450926	0.12	450240	0.16
S70	Aframomum	melegueta	269023	538046	538046	268346	537369	0.13	536179	0.23
S311	Amomum	biphyllum	53169	106338	106338	53059	106228	0.11	106198	0.03
S312	Amomum	biphyllumAff	53509	107018	107018	53384	106893	0.12	106855	0.04
S227	Amomum	calcicolum	53821	107642	107642	53687	107508	0.13	107446	0.06
S13	Amomum	cinnamomeum	69817	139634	139634	69656	139473	0.12	139353	0.09

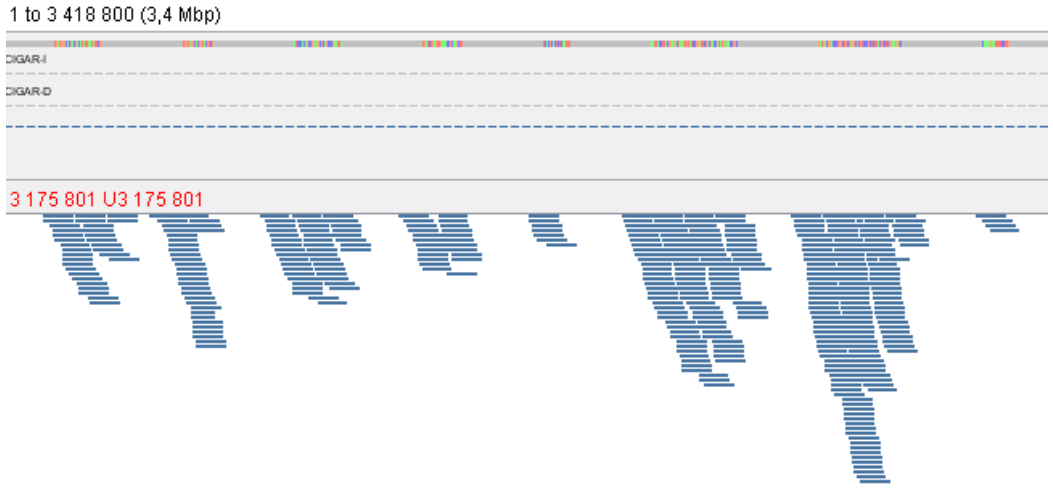
Read mapping to ‘pseudoreference’ (script 2)

parallelized (one job per sample – scripts 2a and 2a2)

- bowtie2 or BWA
- consensus call
  - kindel (https://github.com/bede/kindel)
  - ConsensusFixer (https://github.com/cbg-ethz/ConsensusFixer)
- coverage (Picard tools – https://broadinstitute.github.io/picard/)

Sample no.	Genus	Species	Total nr. reads	Nr. paired reads	Nr. forward unpaired reads	Nr. reverse unpaired reads	Nr. mapped reads	Percentage of mapped reads
S118	Aframomum	alboviolaceum	454178	224851	359	179	268577	59.134
S70	Aframomum	melegueta	541303	267751	419	258	348524	64.386
S312	Amomum	biphyllumAff	107795	53365	84	41	59887	55.556
S311	Amomum	biphyllum	107079	53044	63	47	58503	54.635
S227	Amomum	calcicolum	108334	53656	84	50	63800	58.891
S13	Amomum	cinnamomeum	140561	69596	117	44	85107	60.548
S310	Amomum	corrugatum	102971	51005	78	46	58085	56.409

- `exons/21mapped` – indexed/sorted BAM files + coverage summary tables
- `exons/30consensus`



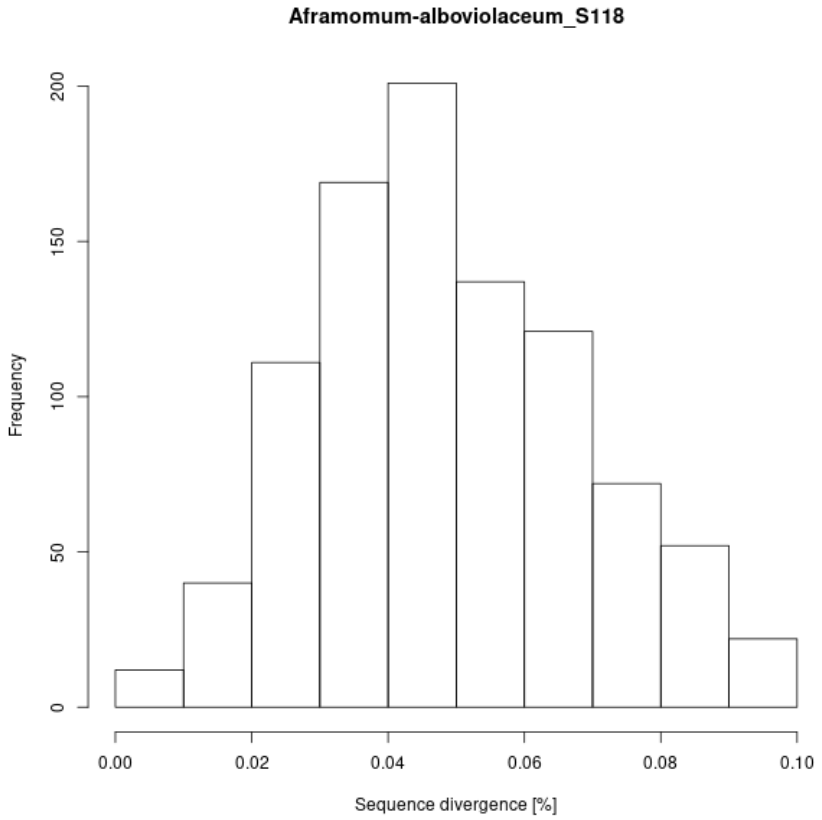
locus	1	1	1	1	1	10014	10014	10014	10014	10014	10046	10046	10046	10046
exon	1	3	7	8	9	1	2	3	5	6	1	2	3	4
Aframomum-alboviolaceum_S118	83.74	48.75	6.40	39.16	86.82	2.93	36.91	30.99	39.27	1.69	38.23	28.54	79.42	15.32
Aframomum-melegueta_S70	73.10	76.80	2.40	48.78	80.74	3.71	47.43	17.43	26.08	2.74	28.80	34.48	73.32	13.32
Amomum-biphyllumAff_S312	20.67	18.01	2.19	6.98	22.32	0.93	23.76	16.83	13.40	0.00	13.96	9.78	24.26	7.16
Amomum-biphyllum_S311	15.51	9.26	1.12	7.61	12.67	0.00	16.16	9.96	13.80	0.00	8.31	6.78	18.00	1.81
Amomum-calcicolum_S227	18.61	10.05	2.01	7.97	24.15	1.04	27.68	18.27	18.11	0.00	7.48	12.50	13.04	2.77
Amomum-cinnamomeum_S13	21.92	15.57	0.41	9.35	29.53	0.00	26.54	10.26	13.93	0.96	8.65	10.86	19.09	8.81
Amomum-corrugatum_S310	22.52	12.31	0.33	12.04	18.68	0.00	26.55	13.73	26.43	0.00	8.24	8.74	24.02	2.49
Amomum-curtisiiAff_S399	18.12	13.46	1.36	9.85	25.11	1.86	37.86	22.79	19.19	1.65	20.82	10.07	25.19	4.78
Amomum-curtisii_S296	40.81	32.03	3.48	21.95	56.51	0.30	65.19	19.69	47.88	2.56	15.89	21.98	40.00	7.32
Amomum-dealbatum_S273	27.23	13.74	2.06	13.96	20.03	0.34	13.21	11.82	19.81	0.00	15.37	11.68	29.63	8.01
Amomum-elanAff_S368	7.34	8.93	0.00	7.14	14.55	0.53	19.70	7.85	7.42	0.38	3.86	7.80	7.89	3.42
Amomum-glabrumAff_S166	17.31	10.31	1.28	7.48	15.97	0.61	14.47	6.88	10.88	1.55	9.78	8.33	22.13	7.25

PSLX files (script 3)

- BLAT (BLAST-like alignment tool)
  - search consensus against exon references (~probes)

- `exons/40contigs`
- `exons/50pslx`

- PSLX = **P**attern **S**pace **L**ayout E**x**tended

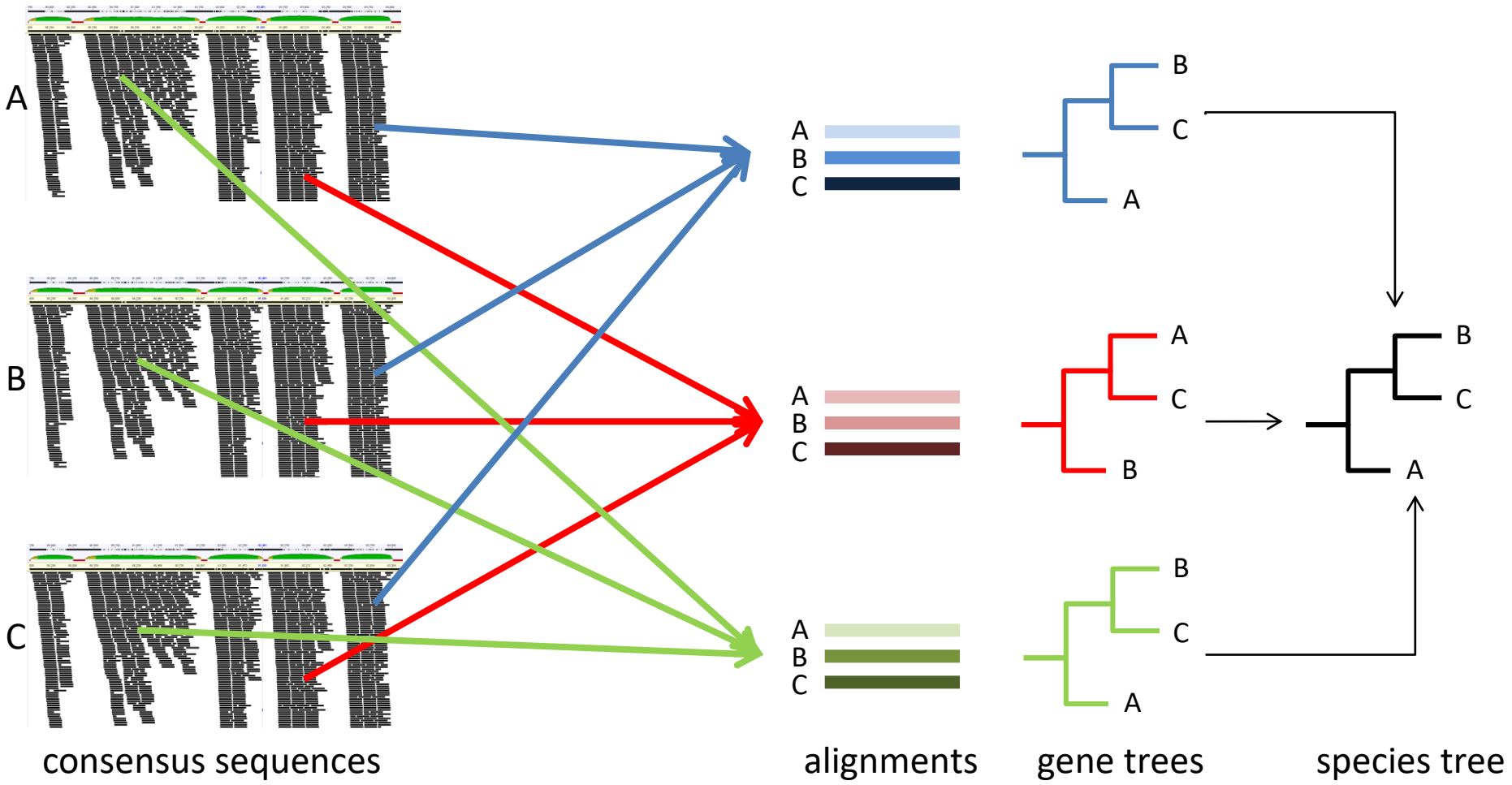


psLayout version 3

match	mis-match	rep. match	N's	Q gap count	Q gap bases	T gap count	T gap bases	strand	Q name	Q size	Q start	Q end	T name	T size	T start	T end	block count	blockSizes	qStarts	tStarts
77	8	0	0	0	0	0	0	+	Contig1_Aframomum-alboviolaceum	1024	81	166	Assembly_15219_Contig_2_286	286	116	201	1	85,	81,	116, tttgaagagcttaacatggacctattcaggaagtgtatggag…
206	15	0	0	0	0	0	0	+	Contig3_Aframomum-alboviolaceum	263	42	263	Assembly_1033_Contig_3_235	235	12	233	1	221,	42,	12, aagttggaacgccgaactgttgcttaccatgaagcaggtcat…
437	48	0	0	0	0	0	0	+	Contig4_Aframomum-alboviolaceum	519	33	518	Assembly_10372_Contig_1_516	516	31	516	1	485,	33,	31, tgagagtggcaagagtagcatgtgcagggctagtacgagca…
143	14	0	0	0	0	0	0	+	Contig6_Aframomum-alboviolaceum	498	133	290	Assembly_10373_Contig_3_215	215	20	177	1	157,	133,	20, cctccttgctttggaatacctgcatatgcttggcatcatctatc…
376	25	0	49	0	0	0	0	+	Contig8_Aframomum-alboviolaceum	453	3	453	Assembly_10373_Contig_2_463	463	0	450	1	450,	3,	0, atttgttggtactcacgagtacttggcccctgagatcatcaag…

…

Read processing



Alignment building (script 4)

- selecting best hits from PSLX files ‘assembled_exons_to_fastas.py’ (Weitemier et al. 2014)
- MAFFT alignment of exons
- concatenate exons to loci (AMAS – https://github.com/marekborowiec/AMAS)

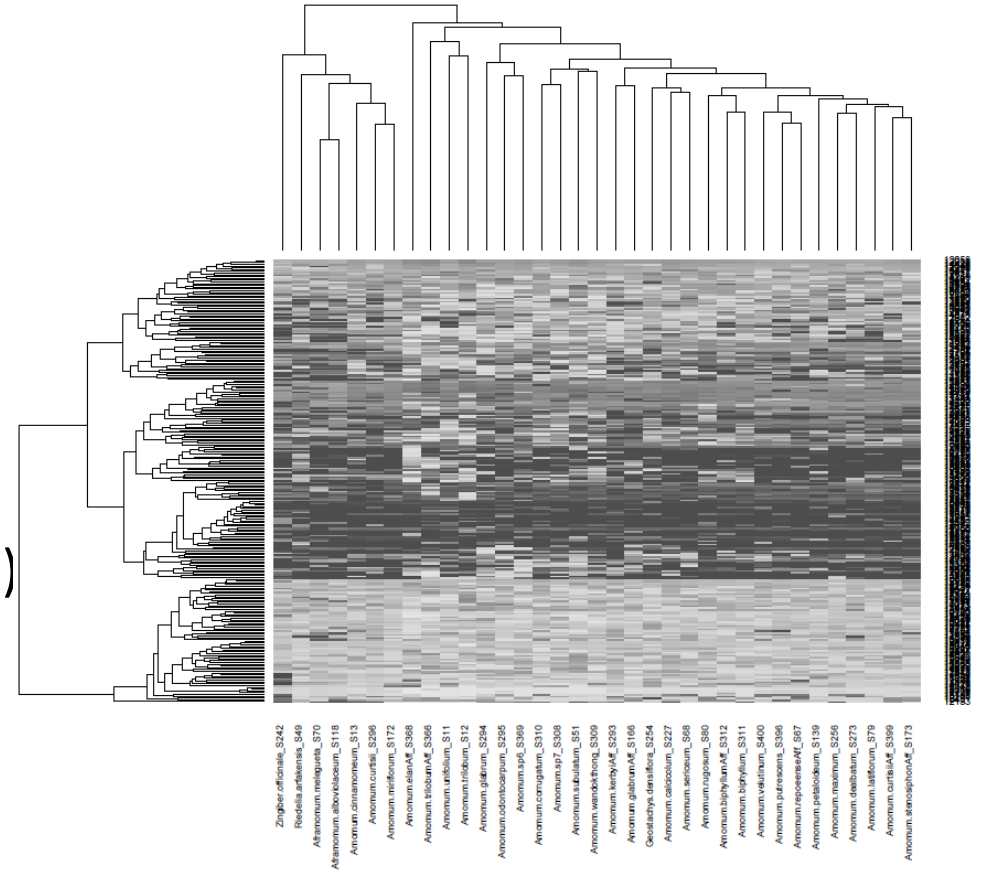
- `exons/60mafft`
- `exons/70concatenated_exon_alignments`

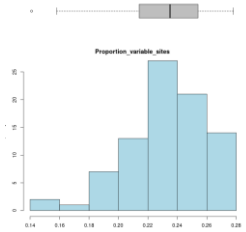
```
>Aframomum-alboviolaceum_S118_contigs.fas
AGTTCTCCAAAGAGAAAAGGAGCATTCTTCCTGTTGGGGACATCACTGAGACTATTCCTGATGAATTGGCAGAGATTGCT
ATACTTGAGGAACCAGAACATCTTACATGGTACCATCATGGTCGGCAATGGAAAACTAAATTCCGAAGTGTGATAGGTGT
>Aframomum-melegueta_S70_contigs.fas
AGTTCTCCAAAGAGAAAAGGAGCATTCTTCCTGTTGGGGACATCACCGAGACTATTCCTGATGAATTGGCAGAGATTGCT
ATACTTGAGGAACCAGAACATCTTACATGGTACCATCATGGTCGGCAATGGAAAACTAAATTCCGAAGTGTGATAGGTGT
>Amomum-biphyllumAff_S312_contigs.fas
AGTTCTCCAAAGAGAAAAGGAGCATTCTTCCTGTTGGGGACATCACTGAGACTATTCCTGATGAATTGGCAGAGATTGCT
ATACTTGAGGAACCAGAACATCTTACATGGTACCATCATGGTCGGCAATGGAAAACTAAATTCCGAAGTGTGATAGGTGT
>Amomum-biphyllum_S311_contigs.fas
AGTTCTCCAAAGAGAAAAGGAGCATTCTTCCTGTTGGGGACATCACTGAGACTATTCCTGATGAATTGGCAGAGATTGCT
ATACTTGAGGAACCAGAACATCTTACATGGTACCATCATGGTCGGCAATGGAAAACTAAATTCCGAAGTGTGATAGGTGT
```

Missing data filtering (script 5)

- samples with more than a certain percentage of missing data per gene are deleted (MISSINGPERCENT=)
- genes with more than the specified percentage of samples per gene are kept (SPECIESPRESENCE=)

- `exons/71selectedMISSINGPERCENT`
  - `deleted_aboveMISSINGPERCENT`
    - alignments with deleted samples
  - list of selected genes
  - missing data overview
  - histograms for selected properties (e.g., aln length, missing %, prop. variable sites…)
  - gene vs. taxon heatmap of missing % (R)

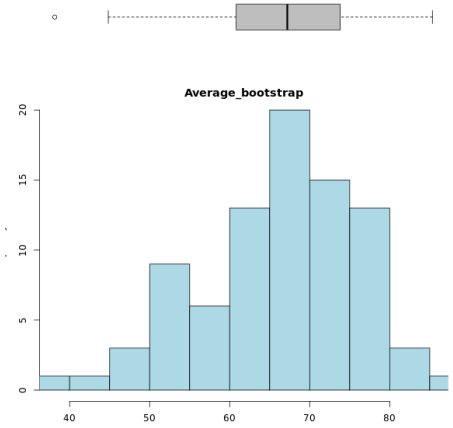




Gene tree building (script 6)

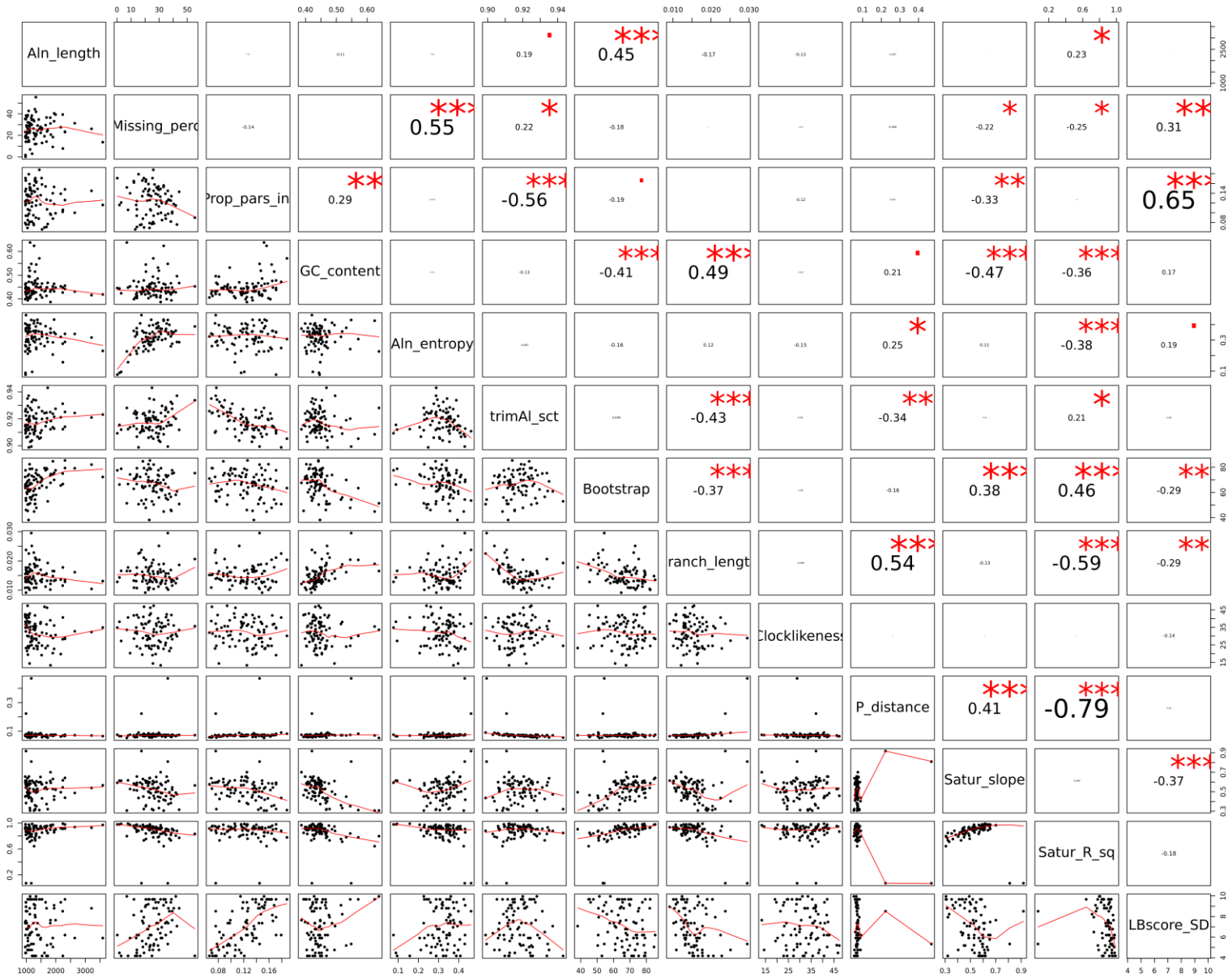
- FastTree – standard or with bootstrapping
- RAxML (parallelized – scripts 6a and 6a2)
  - rapid or standard bootstrap, bootstopping
  - GTRGAMMA or GTRCAT model
  - partitioning – none, per exon, per codon (in case of frame-corrected data)
  - parallelized (one or several genes per job)

- `exons/72treesMISSINGPERCENT_SPECIESPRESENCE/RAxML`
  - gene trees + logs
  - tree statistics (e.g., average BS, average branch length…)
  - histograms
  - gene/alignment properties correlations



- root with outgroup and combine gene trees into a single file (script 7)
  - `exons/72treesMISSINGPERCENT_SPECIESPRESENCE/RAxML/species_trees`

Gene/alignment properties correlations



Species tree building etc.

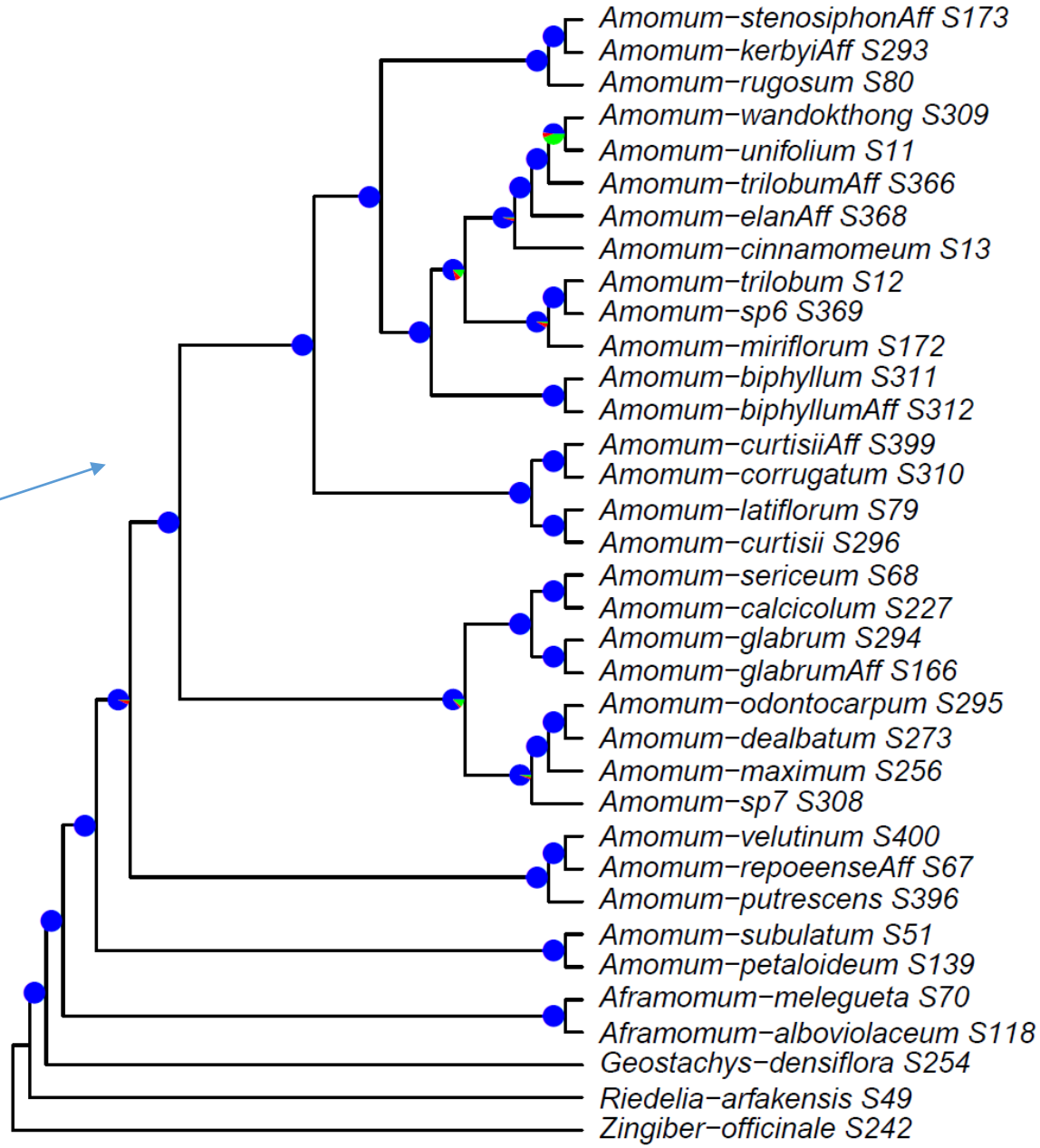
- ASTRAL (ASTRAL-III, with -t 4, ASTRAL-IV) – scripts 8a and 8a2
- ASTRID – script 8b
- MRL (maximum representation with likelihood) – script 8c
- concatenation FastTree – script 8e
- concatenation ExaML (fully partitioned – PartitionFinder) – script 8f
- BUCKy (Bayesian concordance analysis) – script 8g

- neighbour-network – script 8h
- SuperQ network (supernetwork from quartets) – script 8j
- SNaQ – script 8l
- PhyloNet – scripts 8m and 8m2 (parallelized)

- quartet sampling – script 8k
- PhyParts – script 11
- similarity SNP heatmap – script 8n
- treePL (divergence dating) – script 14
- combine trees – script 15

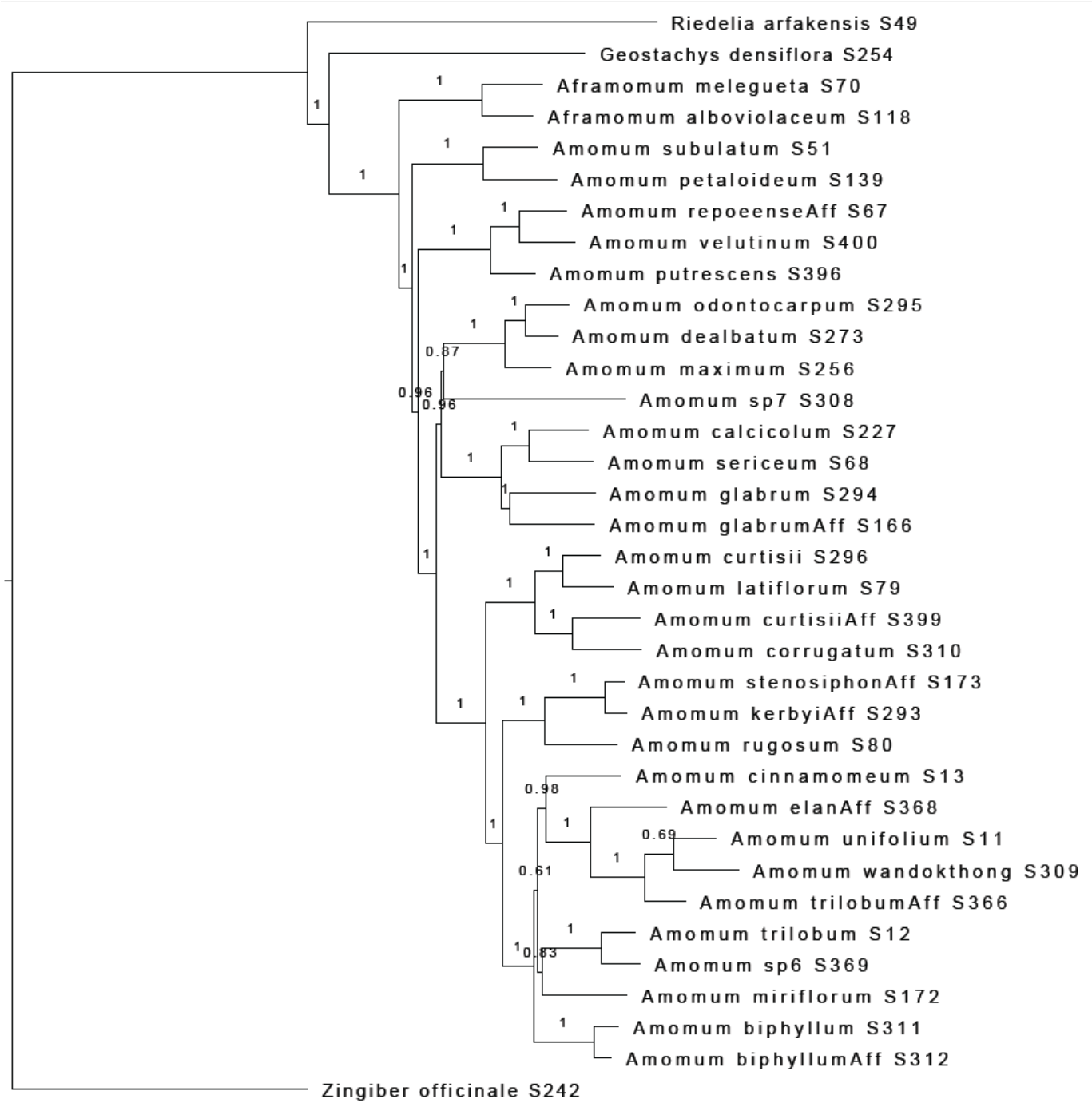
ASTRAL tree

- standard
- local posterior probability (LPP)
- multi-locus bootstrap support (MLBS)
- combination of trees (p4)
  - main tree
  - greedy consensus
  - bootstrap support
- ‘-t 4’ quartet scoring
  - three local posterior probabilities
    - one for the main topology
    - one for each of the two alternatives



ASTRAL-IV tree

- with branch lengths in substitution-per-site units
- LPP

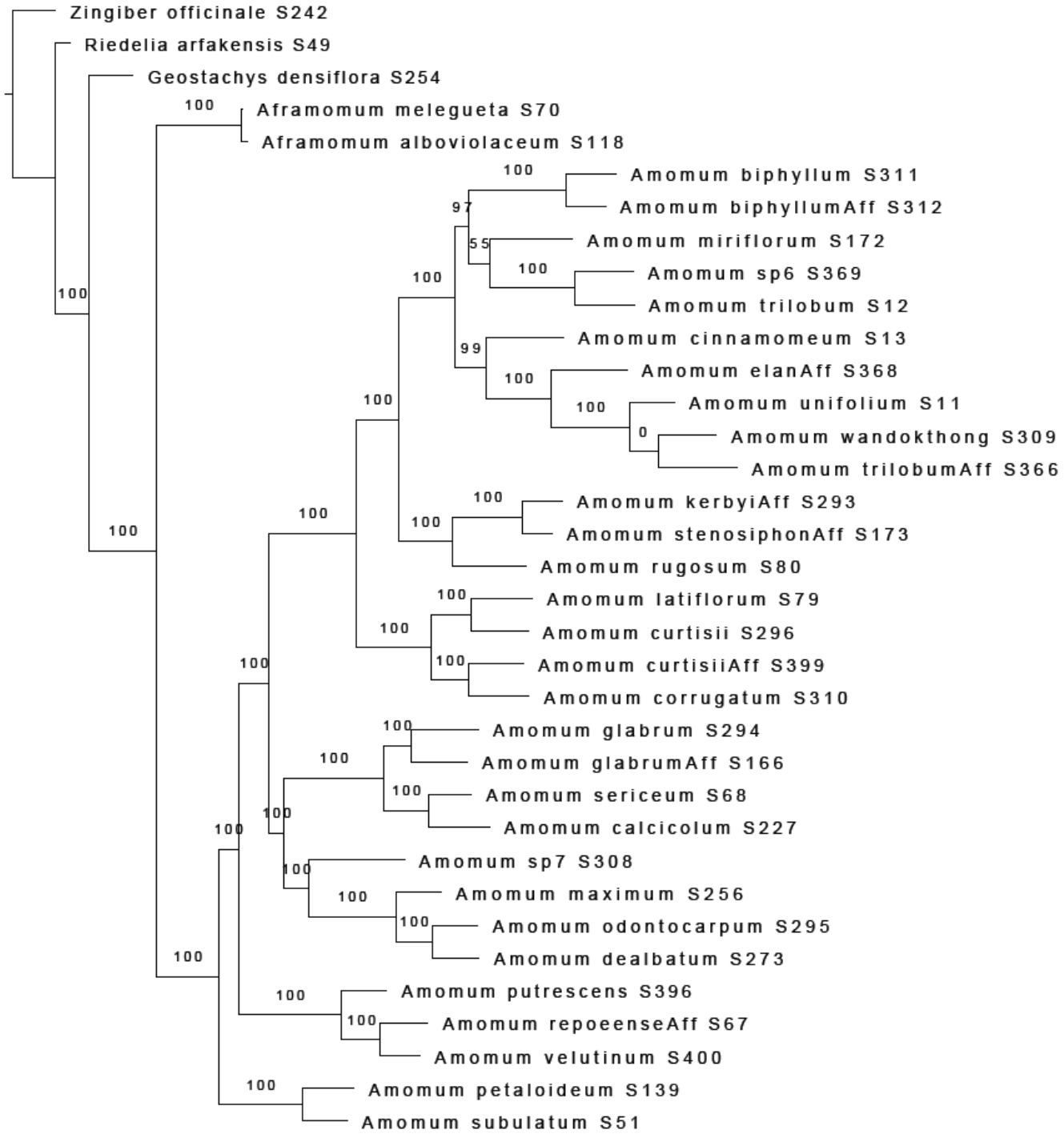


ASTRID tree

- standard
- multi-locus bootstrap support (MLBS)
- combination of trees
  - MLBS
  - consensus

MRL tree

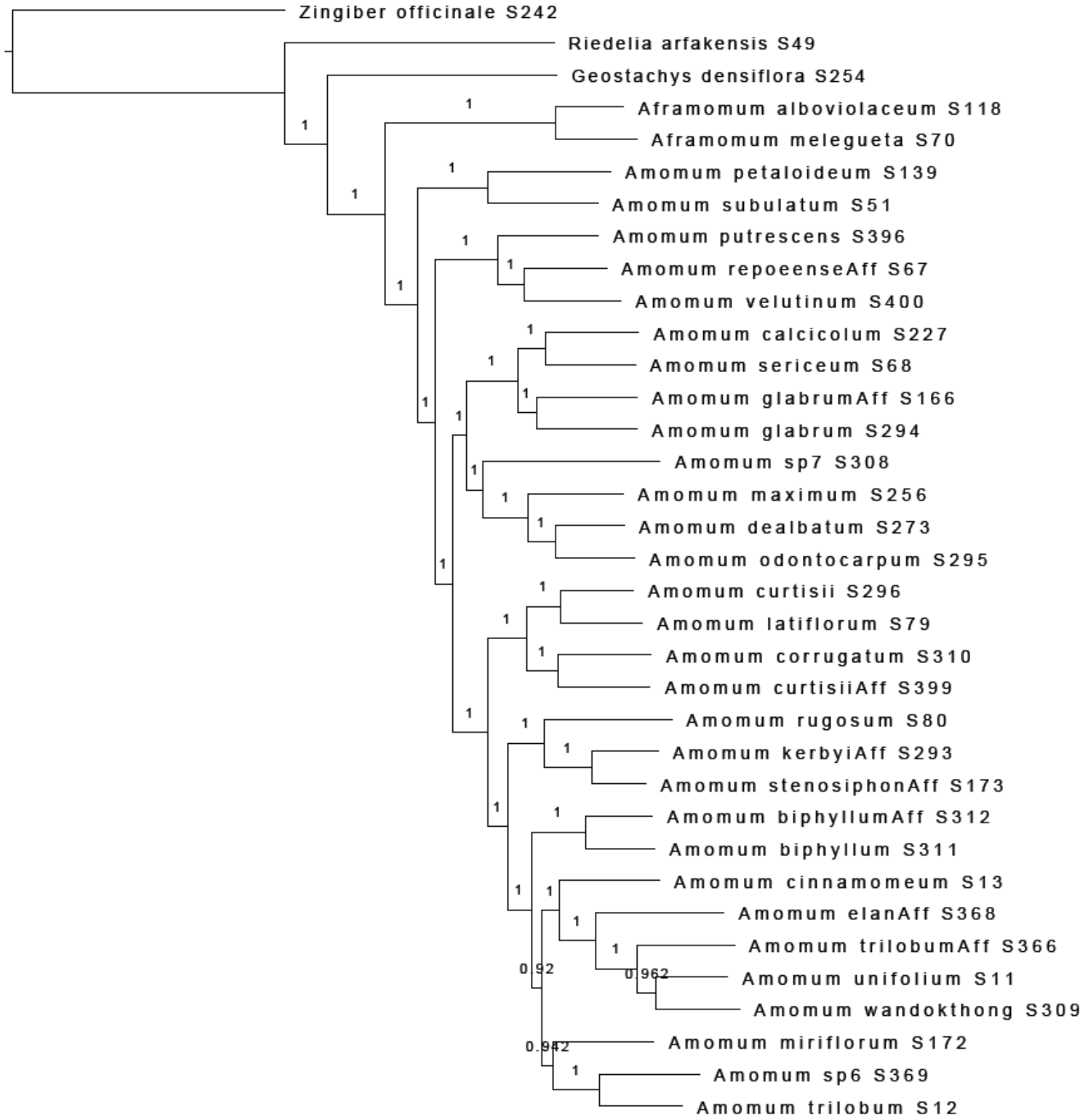
- BINGAMMA model
- RAxML with standard bootstrapping



Concatenated tree

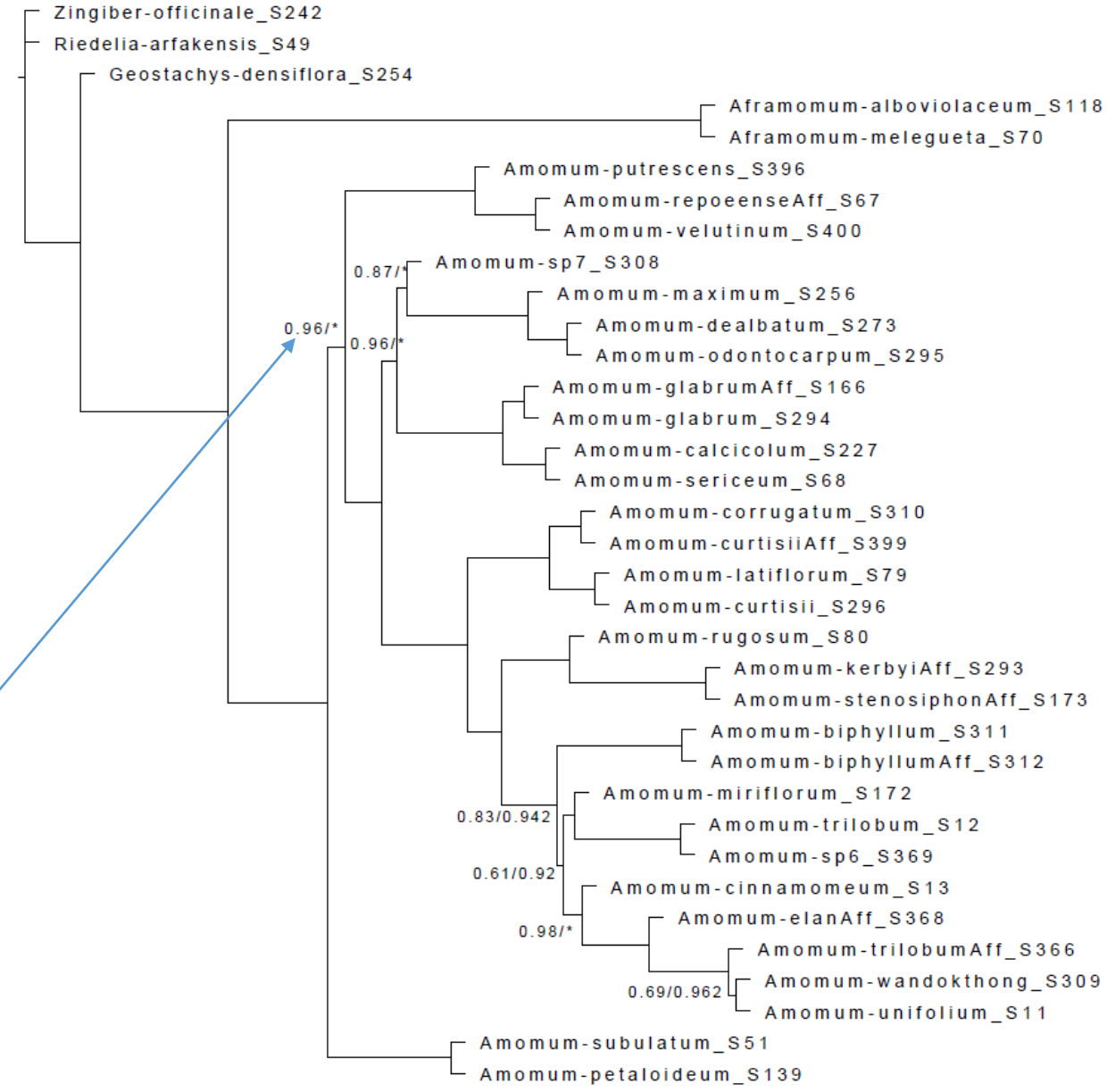
- FastTree or ExaML
- missing % table

Aframomum-alboviolaceum_S118	3.34
Aframomum-melegueta_S70	3.65
Amomum-biphyllumAff_S312	6.48
Amomum-biphyllum_S311	6.26
Amomum-calcicolum_S227	6.54
Amomum-cinnamomeum_S13	5.42
Amomum-corrugatum_S310	7.00
Amomum-curtisiiAff_S399	5.01
Amomum-curtisii_S296	4.47
Amomum-dealbatum_S273	4.50
Amomum-elanAff_S368	10.28
Amomum-glabrumAff_S166	8.00
Amomum-glabrum_S294	8.44
Amomum-kerbyiAff_S293	5.23
Amomum-latiflorum_S79	4.54



Combination of BS values from multiple trees

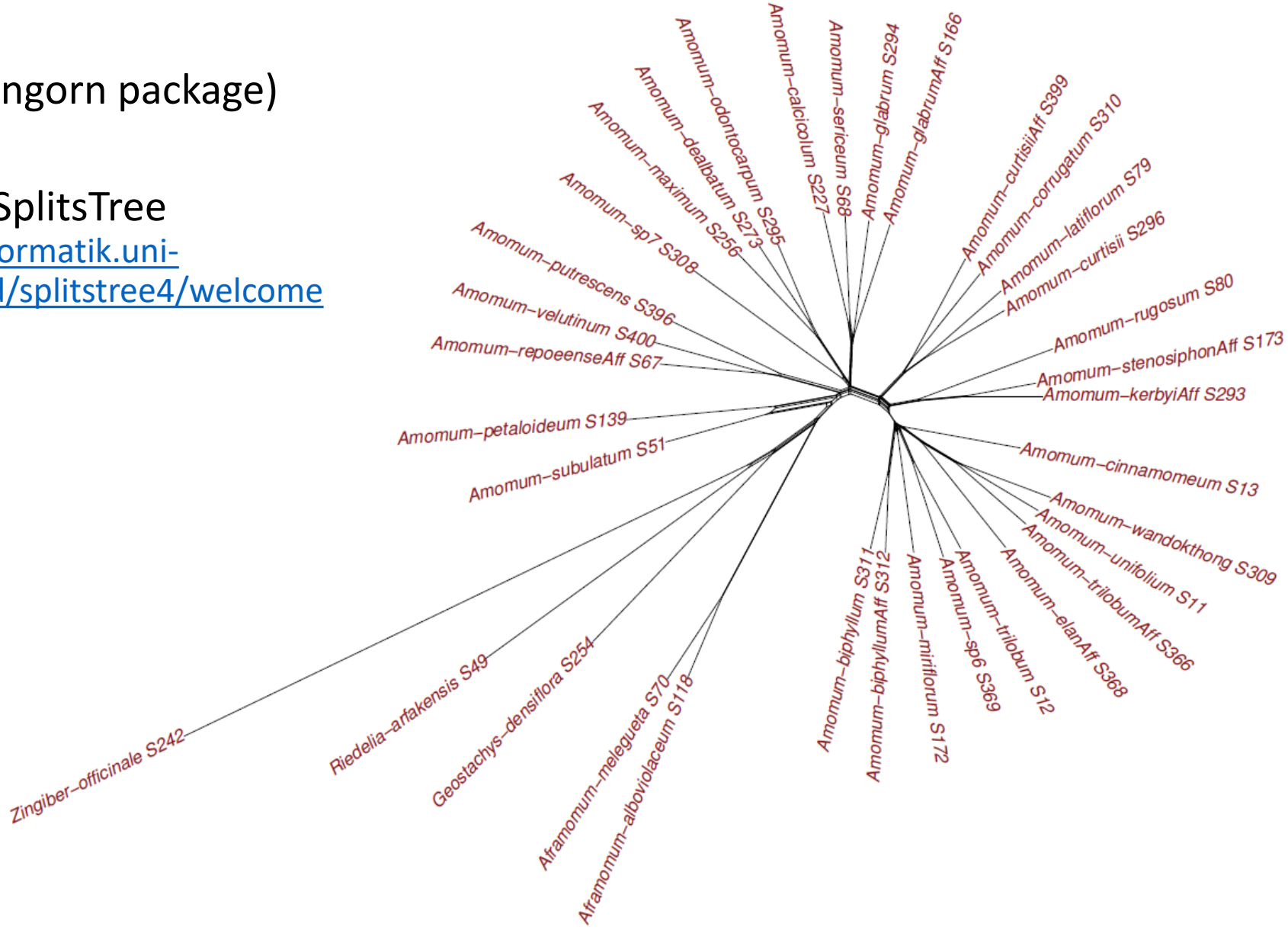
- ASTRAL, ASTRAL-IV, ASTRID, MRL, FastTree or ExaML
- two values per node
- rounding to nr. of decimals (prec=)
- full support marked with '*'
- empty nodes – full support in both trees
- '-' – this group does not exist in the particular tree (= different topology)



0.96 in ASTRAL
1 in FastTree

Neighbour network

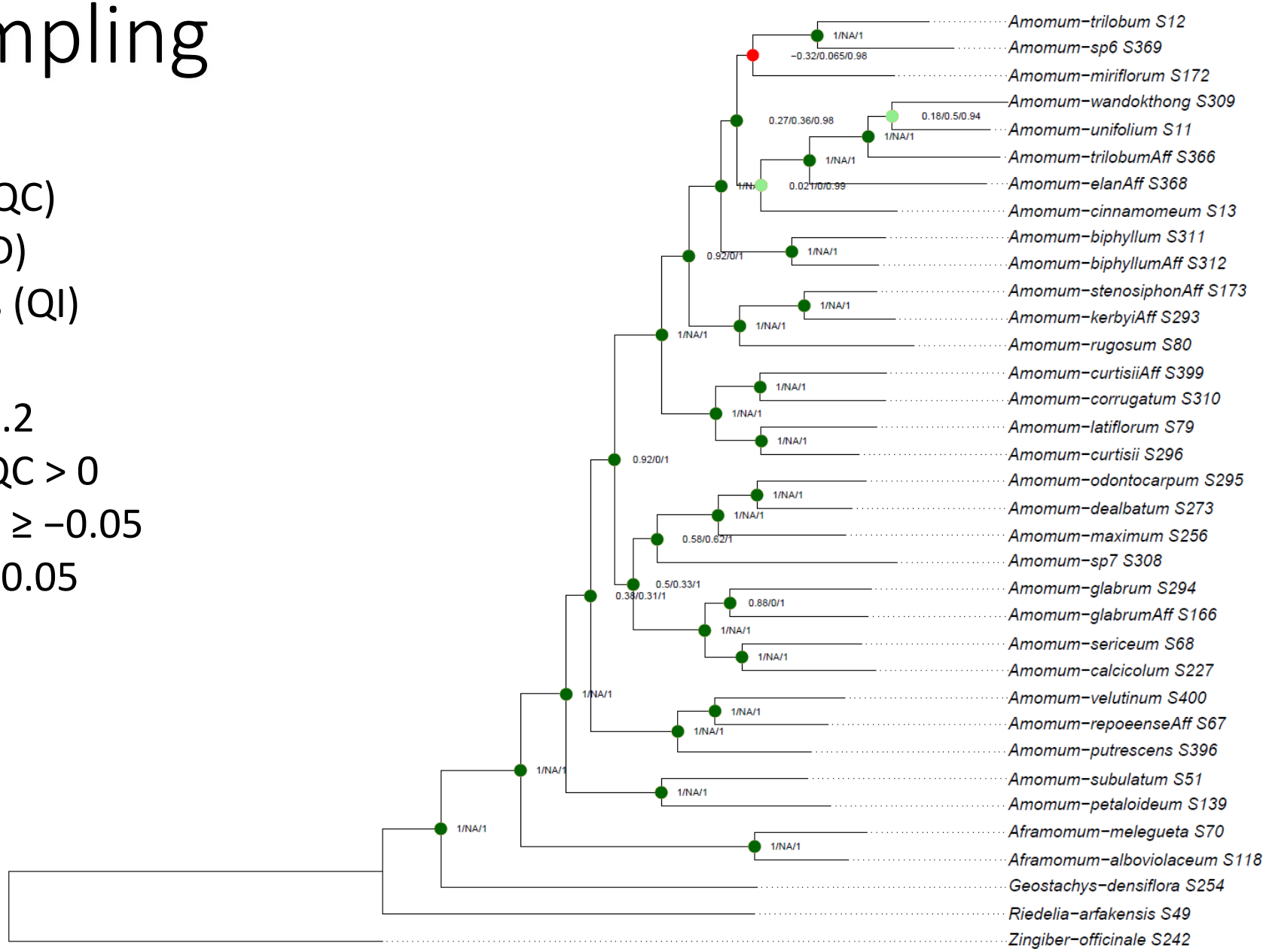
- computed in R (phangorn package)
- plotting to PDF
- or to be opened in SplitsTree (https://software-ab.informatik.uni-tuebingen.de/download/splitstree4/welcome.html)



Quartet sampling

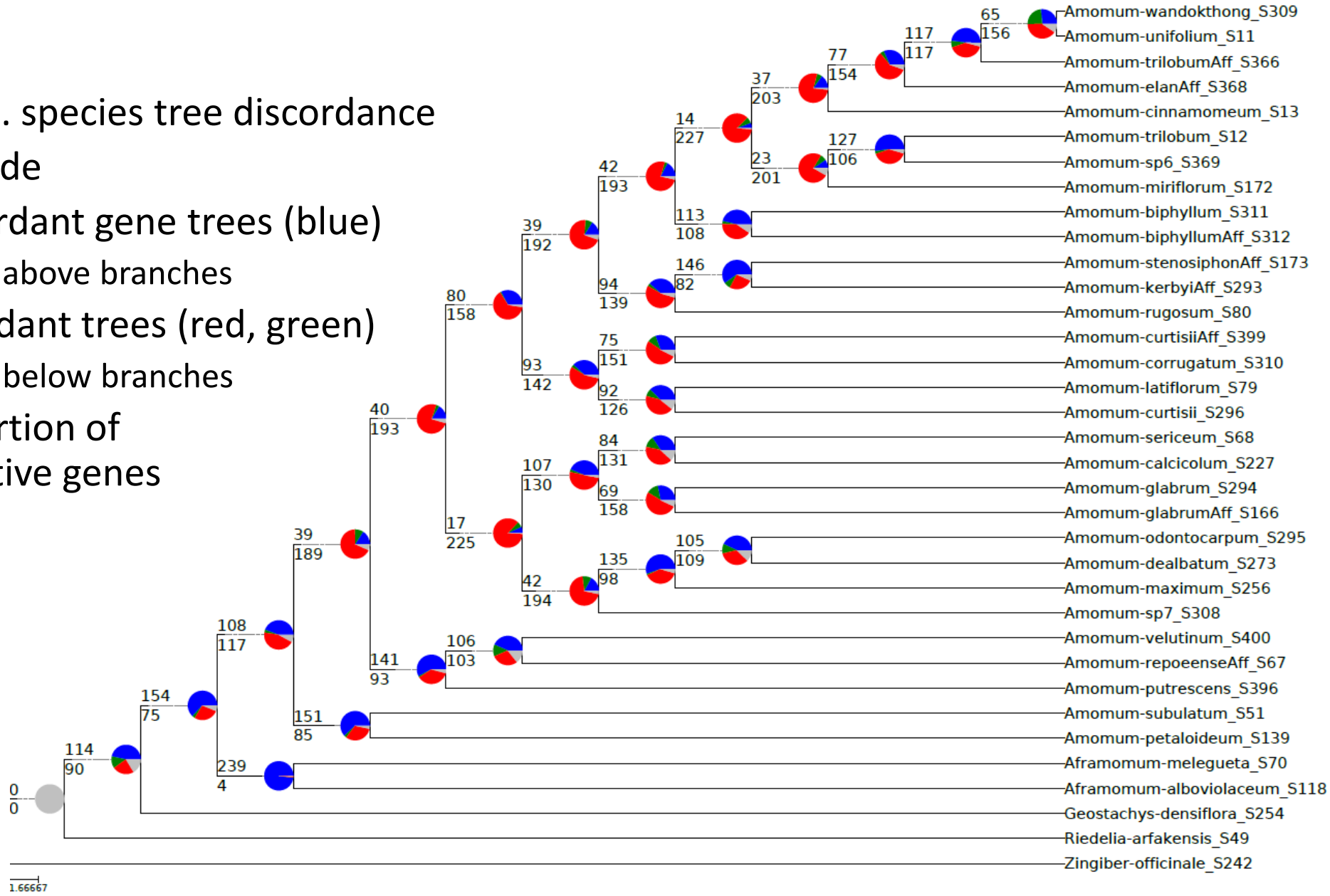
quartet concordance (QC)
quartet differential (QD)
quartet informativness (QI)

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



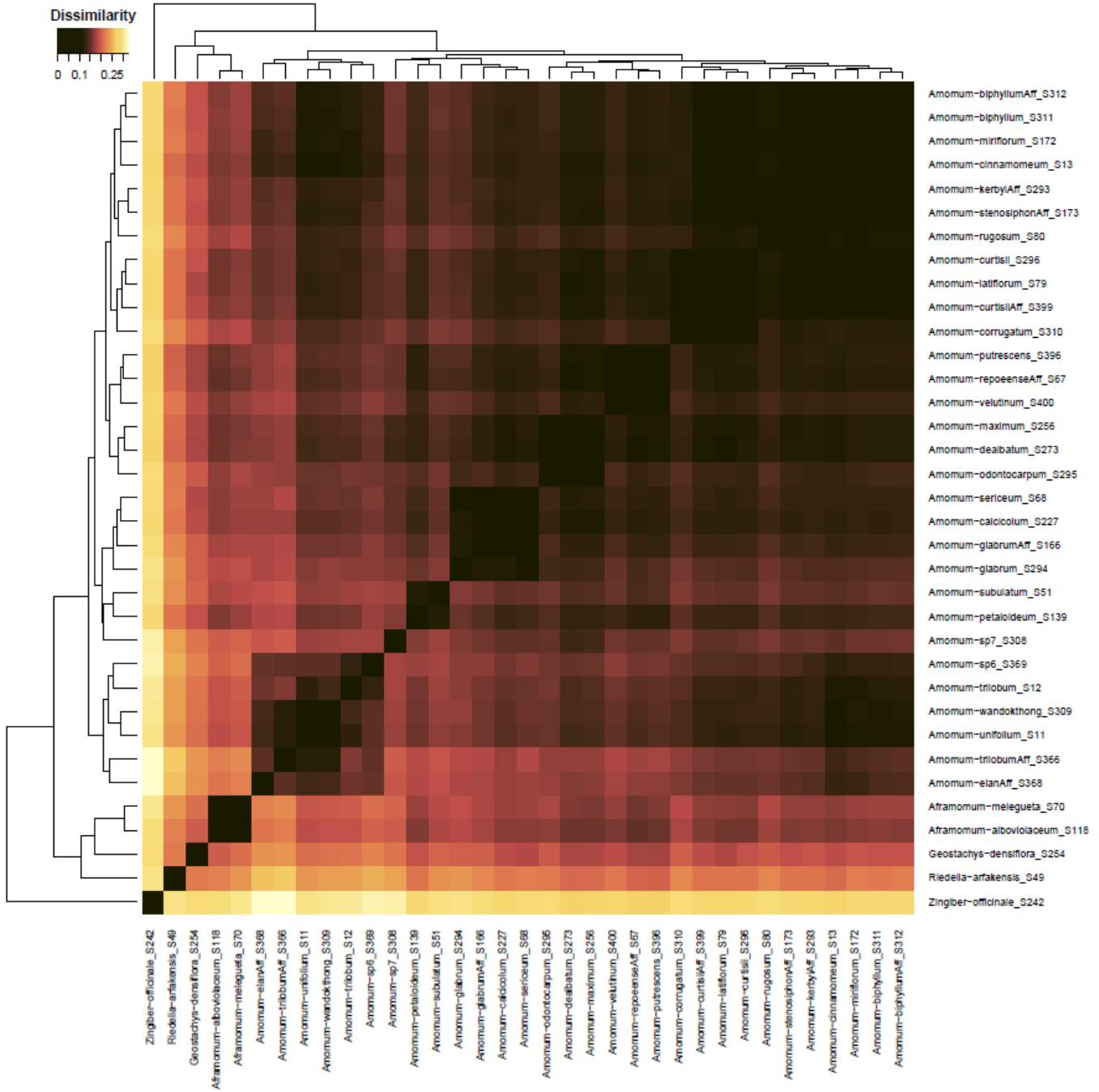
PhyParts

- 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



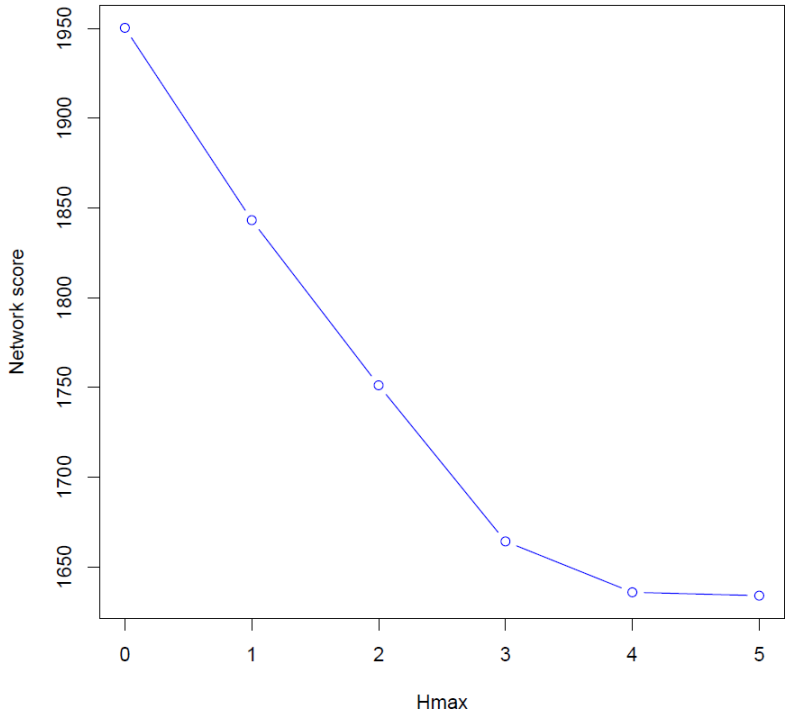
SNP (dis)similarity matrix

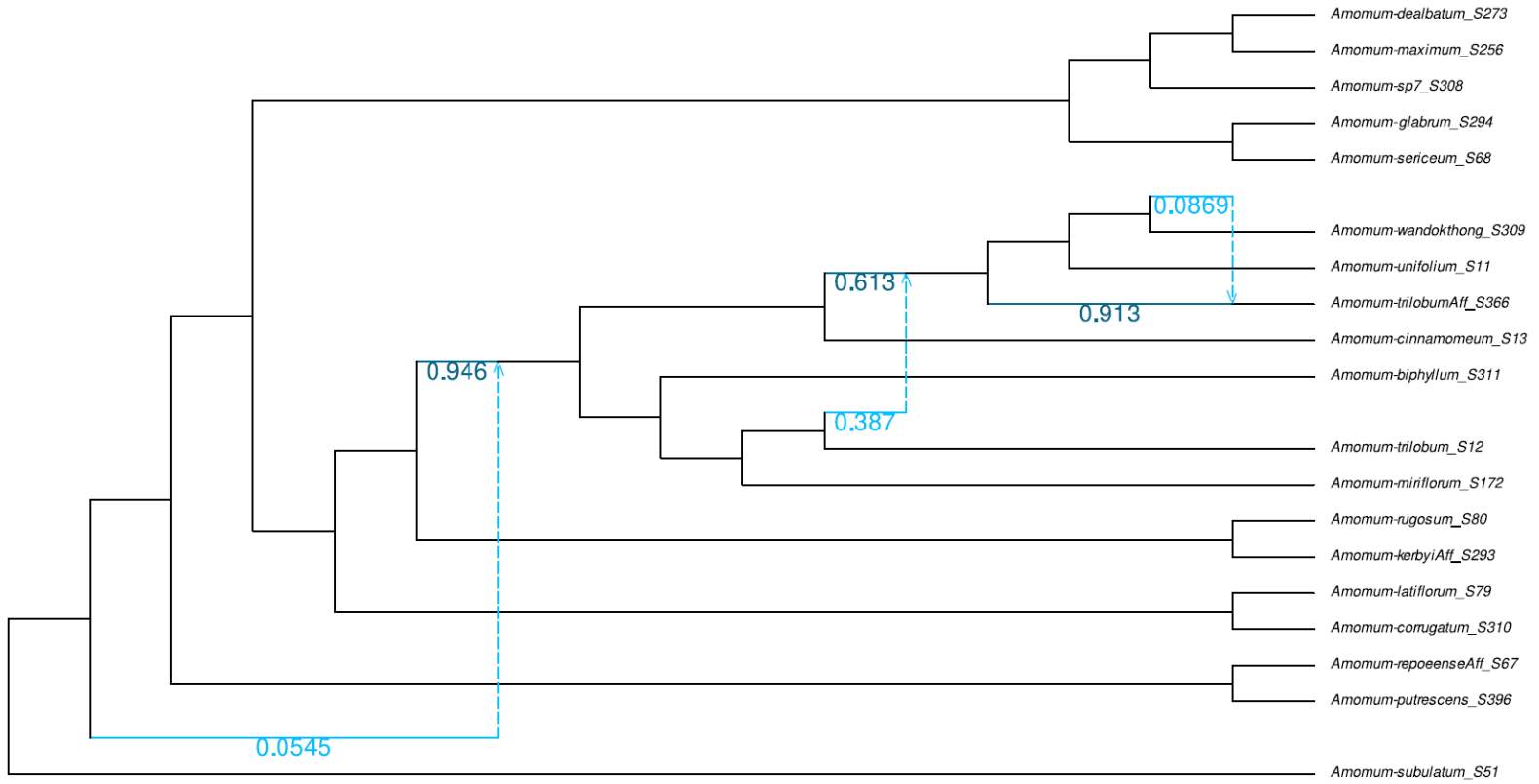
- pair-wise distances derived from filtered variable sites (SNPs) – simple-matching coefficient
- all variable sites with less than, e.g., 20% missing data



SNaQ

- **S**pecies **N**etworks **a**pplying **Q**uartets
- Solís-Lemus & Ané (2016)
- PhyloNetworks Julia package (https://github.com/JuliaPhylo/PhyloNetworks.jl)
- best network based on log likelihood increase



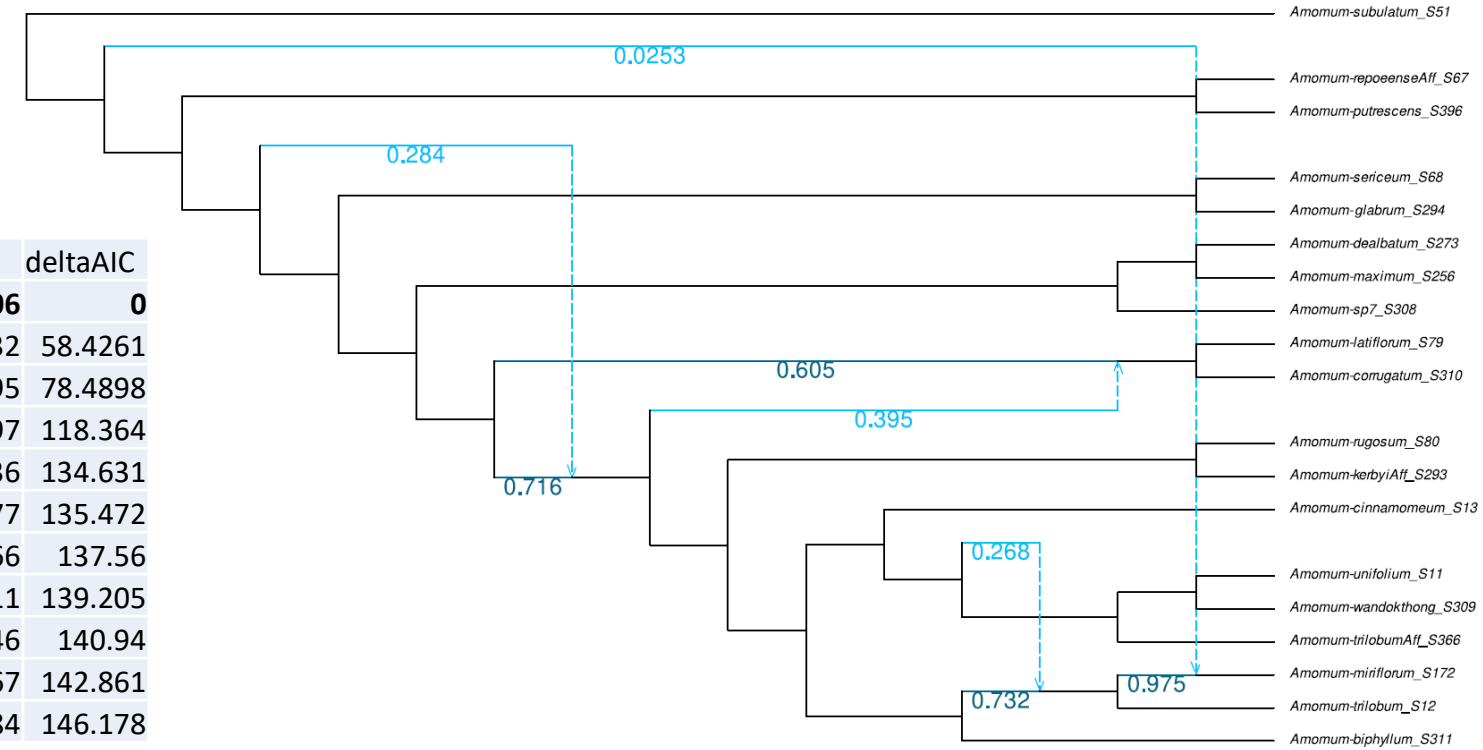


PhyloNet (MPL)

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

PhyloNet_summary.txt

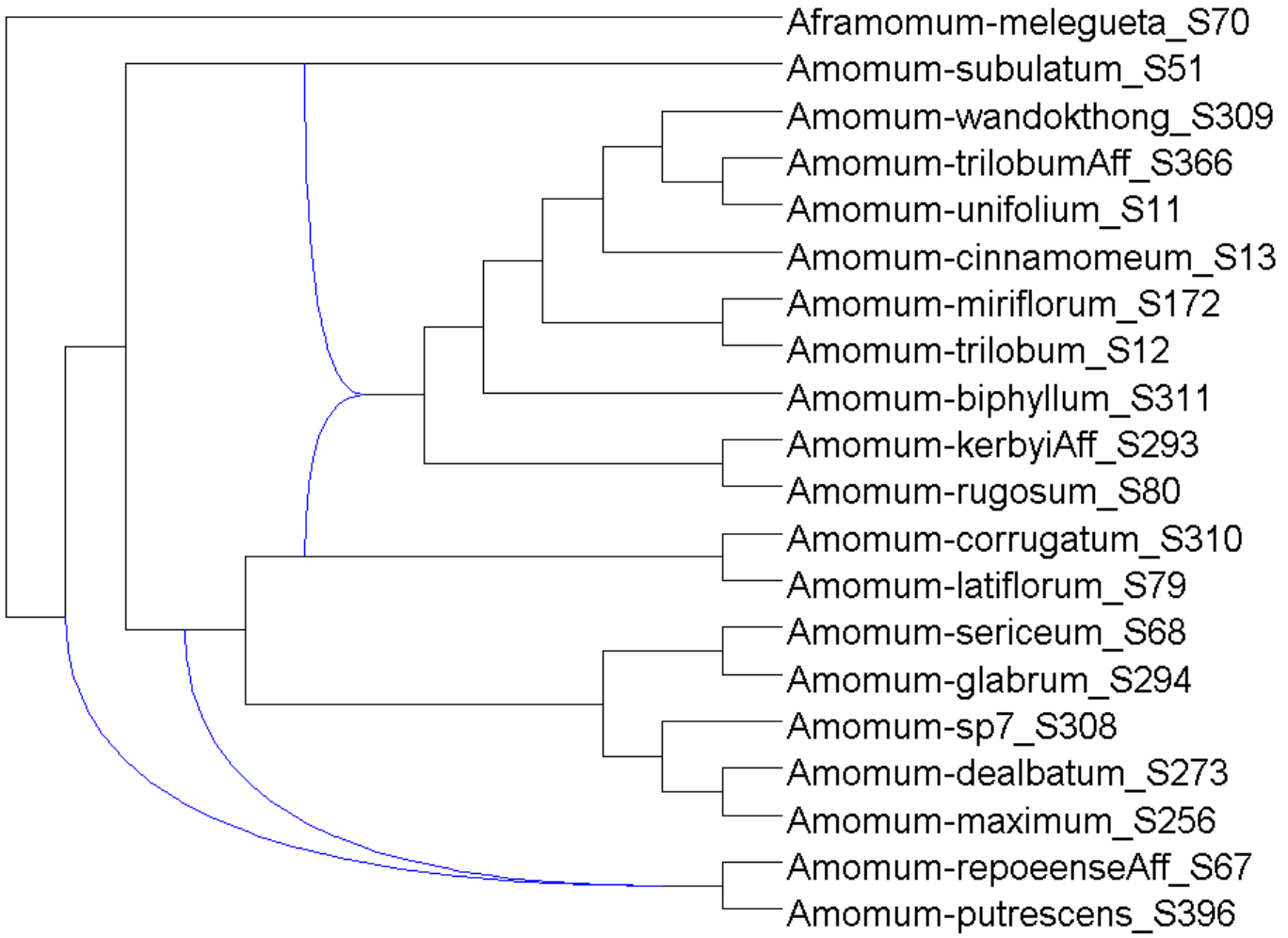
NrHybridizations	NrBranches	NrGeneTrees	logLikelihood	ki	AIC	deltaAIC
4	**21**	**1106**	**-1077277.803**	**1131**	**2156817.606**	**0**
3	23	1106	-1077306.016	1132	2156876.032	58.4261
3	21	1106	-1077318.048	1130	2156896.095	78.4898
3	21	1106	-1077337.985	1130	2156935.97	118.364
0	21	1106	-1077349.118	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



Networks visualization – Dendroscope

- Huson & Scornavacca (2012)
- https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/dendroscope/

- PhyloNet_nr_Dendroscope.networks



extended Newick tree format

for Dendroscope:

((((Amomum-subulatum_S51,(((Amomum-kerbyiAff_S293,Amomum-rugosum_S80),(((Amomum-trilobum_S12,Amomum-miriflorum_S172),((Amomum-wandokthong_S309,(Amomum-trilobumAff_S366,Amomum-unifolium_S11)),Amomum-cinnamomeum_S13)),Amomum-biphyllum_S311)))**#H1**),((((Amomum-corrugatum_S310,Amomum-latiflorum_S79),**#H1**),((Amomum-sp7_S308,(Amomum-dealbatum_S273,Amomum-maximum_S256)),(Amomum-sericeum_S68,Amomum-glabrum_S294))),((Amomum-putrescens_S396,Amomum-repoeenseAff_S67))**#H2**)),**#H2**),Aframomum-melegueta_S70);

Hybrid edge number:branch length::gamma coefficient

((((Amomum-subulatum_S51:1.0,(((Amomum-kerbyiAff_S293:1.0,Amomum-rugosum_S80:1.0):0.485492090949261 6,(((Amomum-trilobum_S12:1.0,Amomum-miriflorum_S172:1.0):0.1134162621213768 3,((Amomum-wandokthong_S309:1.0,(Amomum-trilobumAff_S366:1.0,Amomum-unifolium_S11:1.0):0.0011774181844964955):0.768172053690023,Amomum-cinnamomeum_S13:1.0):0.0832803567939911 5):0.04542614957407823 5,Amomum-biphyllum_S311:1.0):0.303840166051184 2):0.35776517567940297)**#H1:0.0011774181844964955::0.14063145854111692**):0.0011774181844964955,((((Amomum-corrugatum_S310:1.0,Amomum-latiflorum_S79:1.0):0.488183572431556 2,**#H1:0.0011774181844964955::0.859368541458883 1**):0.8496132507646107,((Amomum-sp7_S308:1.0,(Amomum-dealbatum_S273:1.0,Amomum-maximum_S256:1.0):0.769108054873450 6):0.2035712231448452 4,(Amomum-sericeum_S68:1.0,Amomum-glabrum_S294:1.0):0.828556363440200 2):0.026322906906597957):0.12568589532238 64,((Amomum-putrescens_S396:1.0,Amomum-repoeenseAff_S67:1.0):1.049980502183750 7)**#H2:0.18260130897169333::0.6632639438671224**):0.25009483188794623):0.589354722406496 1,**#H2:0.13383557483940117::0.33673605613287 76**):5.911656966691677,Aframomum-melegueta_S70:1.0);