HybPhyloMaker

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

Tomáš Fér

Dept. of Botany, Charles University, Prague

June 2023

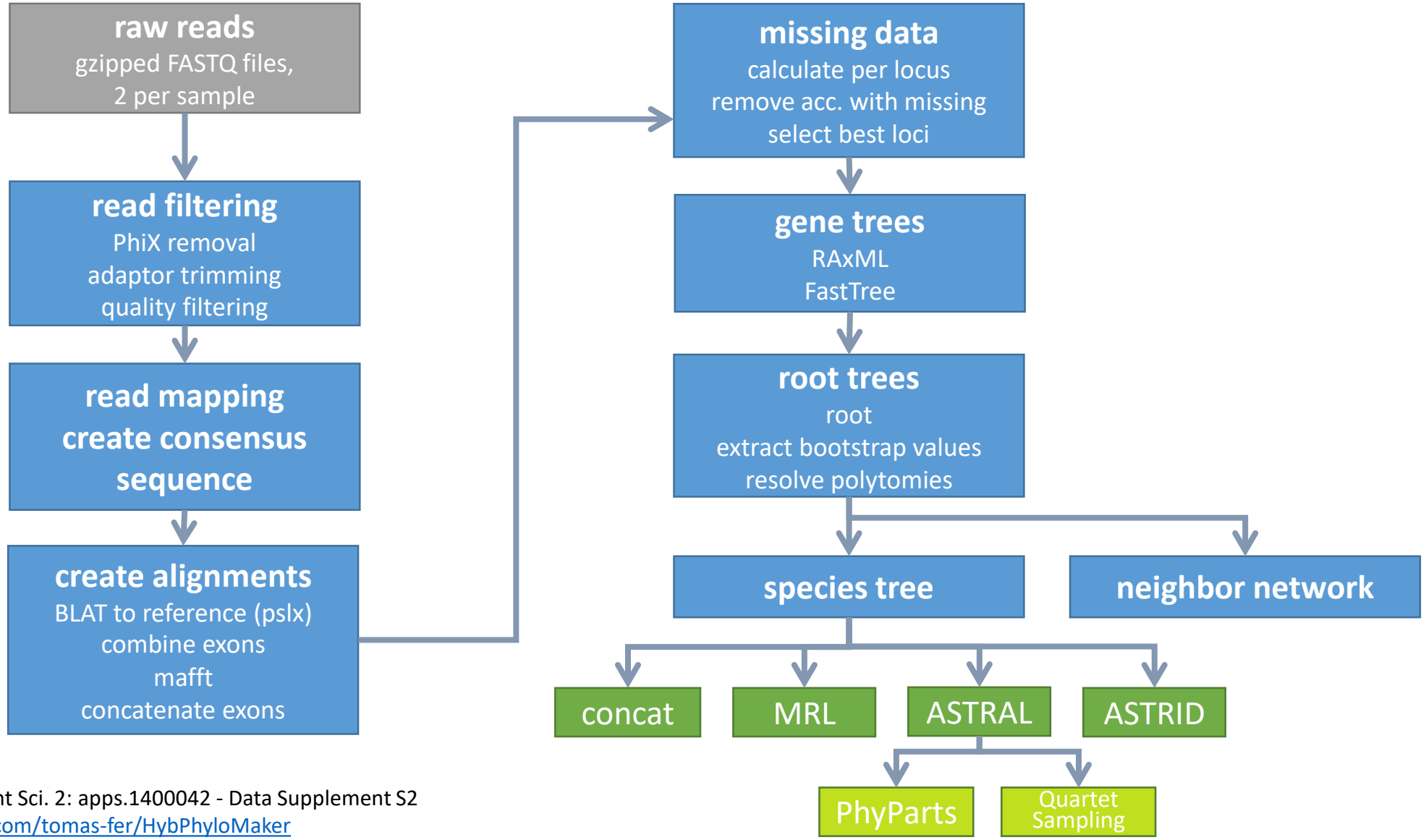
Hyb-Seq data analysis software

- 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 pipeline



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 – script 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

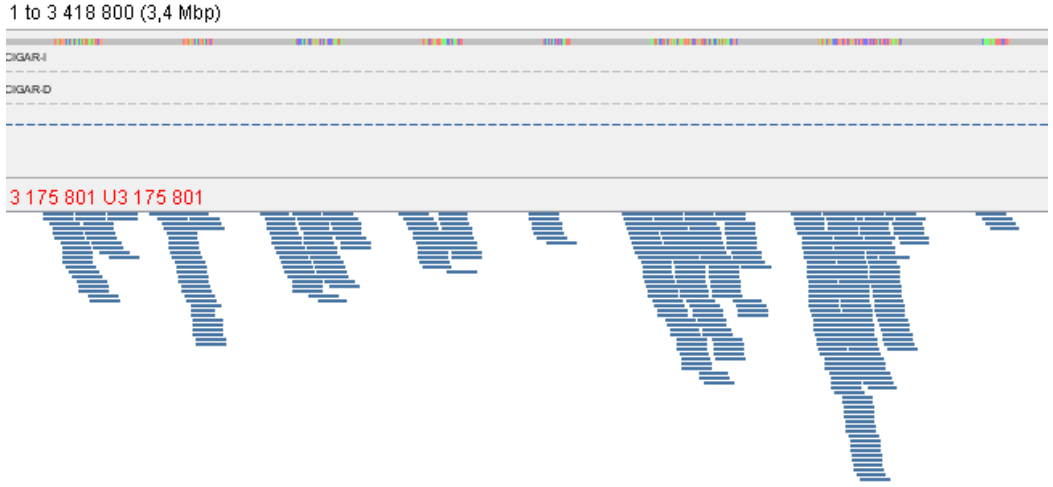
Read mapping to ‘pseudoreference’

parallelized (one job per sample)

- 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`

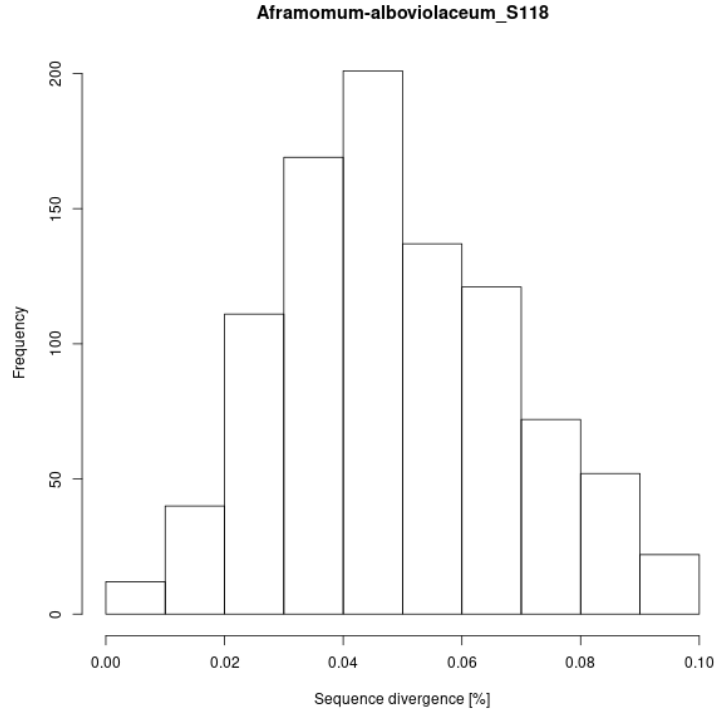


locus	1	1	1	1	1	10014	10014	10014	10014	10014	10046	10046	10046	10046
exon	1	3	7	8	9	1	2	3	5	6	1	2	3	4
Aframomum-alboviolaceum_S118	83.74	48.75	6.40	39.16	86.82	2.93	36.91	30.99	39.27	1.69	38.23	28.54	79.42	15.32
Aframomum-melegueta_S70	73.10	76.80	2.40	48.78	80.74	3.71	47.43	17.43	26.08	2.74	28.80	34.48	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

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

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

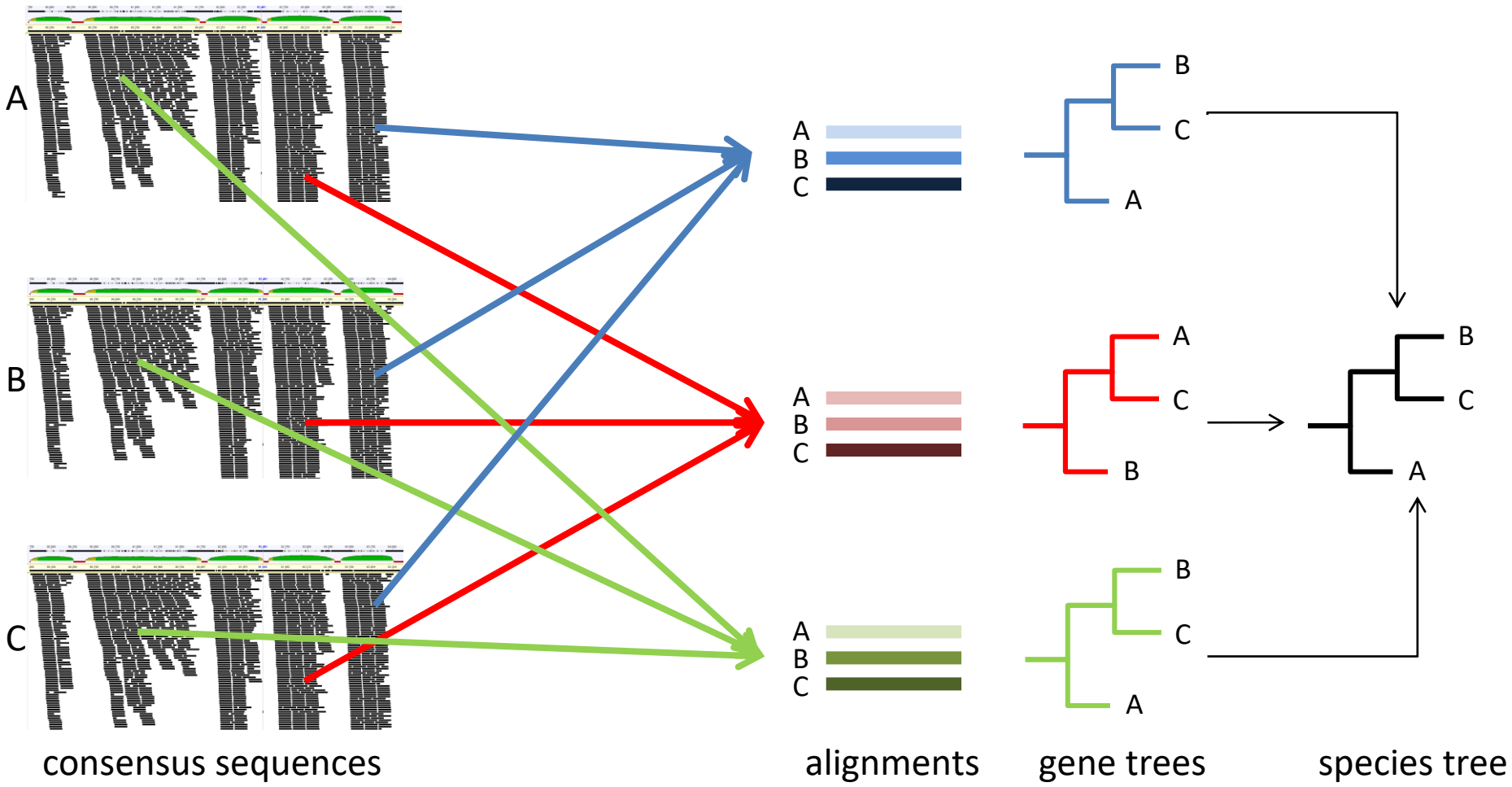


psLayout version 3

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_Strelitzia-nicolai	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_Strelitzia-nicolai	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_Strelitzia-nicolai	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_Strelitzia-nicolai	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_Strelitzia-nicolai	453	3	453	Assembly_10373_Contig_2_463	463	0	450	1	450,	3,	0, atttgttggtactcacgagtacttggcccctgagatcatcaag…

…

Read processing



Alignment building

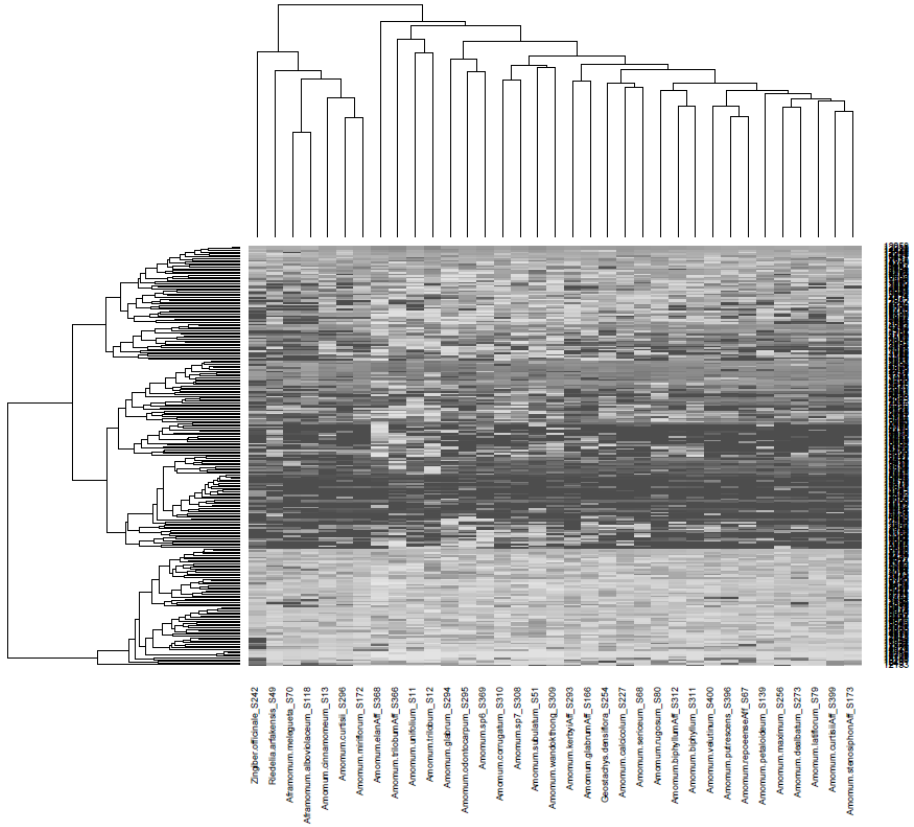
- selecting best hits from PSLX files ‘assembled_exons_to_fastas.py’ (Weitemier et al. 2014)
- MAFFT alignment of exons
- concatenate exons to loci (AMAS - https://github.com/marekborowiec/AMAS)

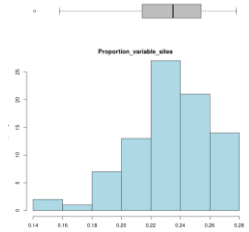
- `exons/60mafft`
- `exons/70concatenated_exon_alignments`

Missing data filtering

- 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 without 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)

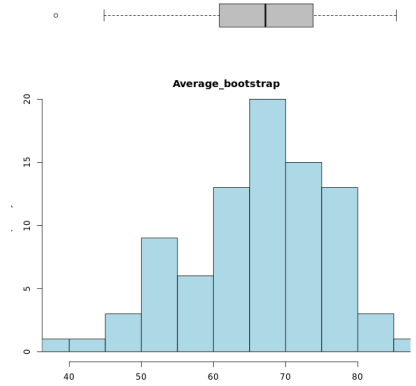




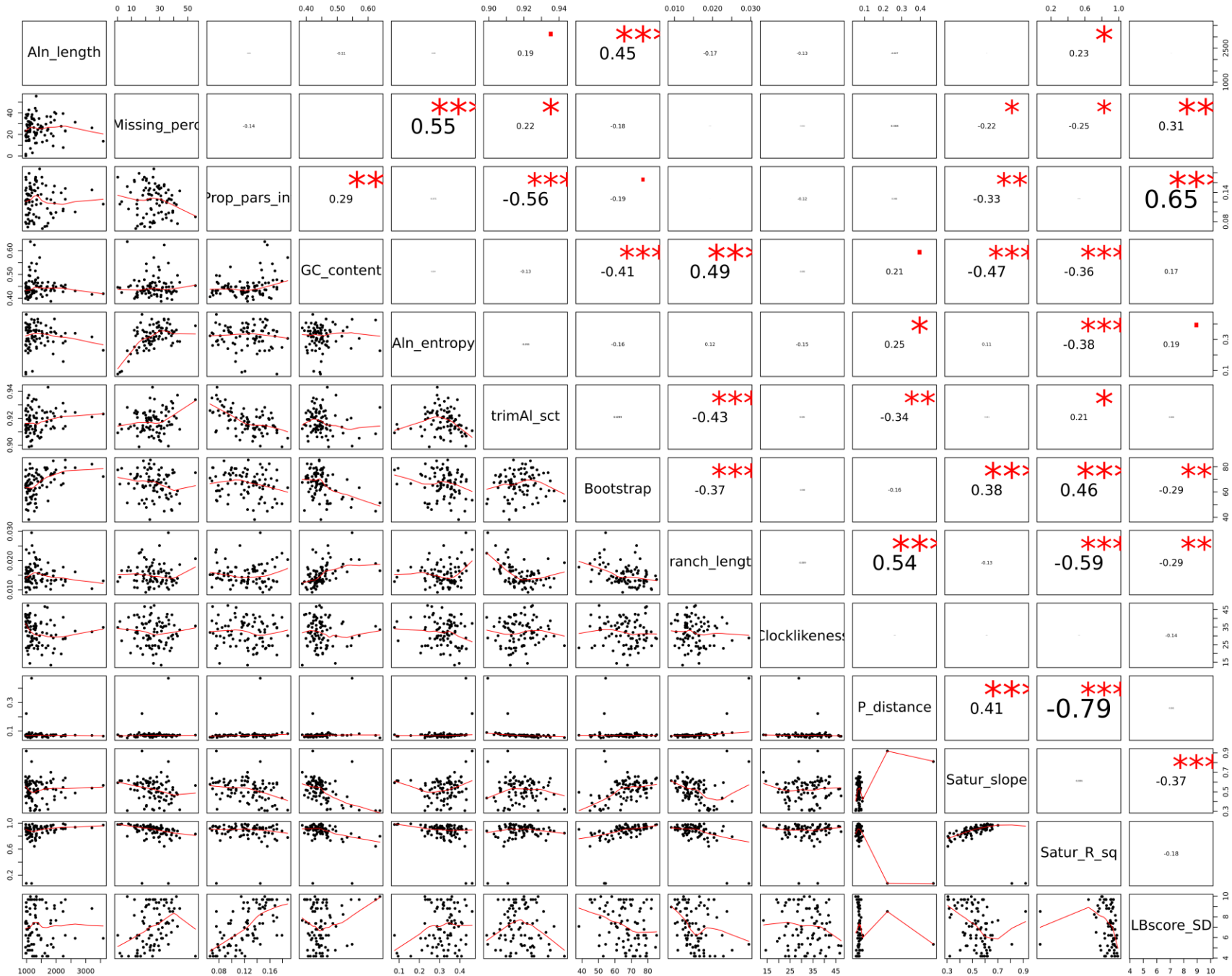
Gene tree building

- FastTree – standard or with bootstrapping
- RAxML
  - 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
  - `exons/72treesMISSINGPERCENT_SPECIESPRESENCE/RAxML/species_trees`


Aln_length
Missing_perc
Prop_pars_in
GC_content
Aln_entropy
trimAl_sct
Bootstrap
ranch_lengt
Clocklikenes
P_distance
Satur_slope
Satur_R_sq
LBscore_SD

Species tree building etc.

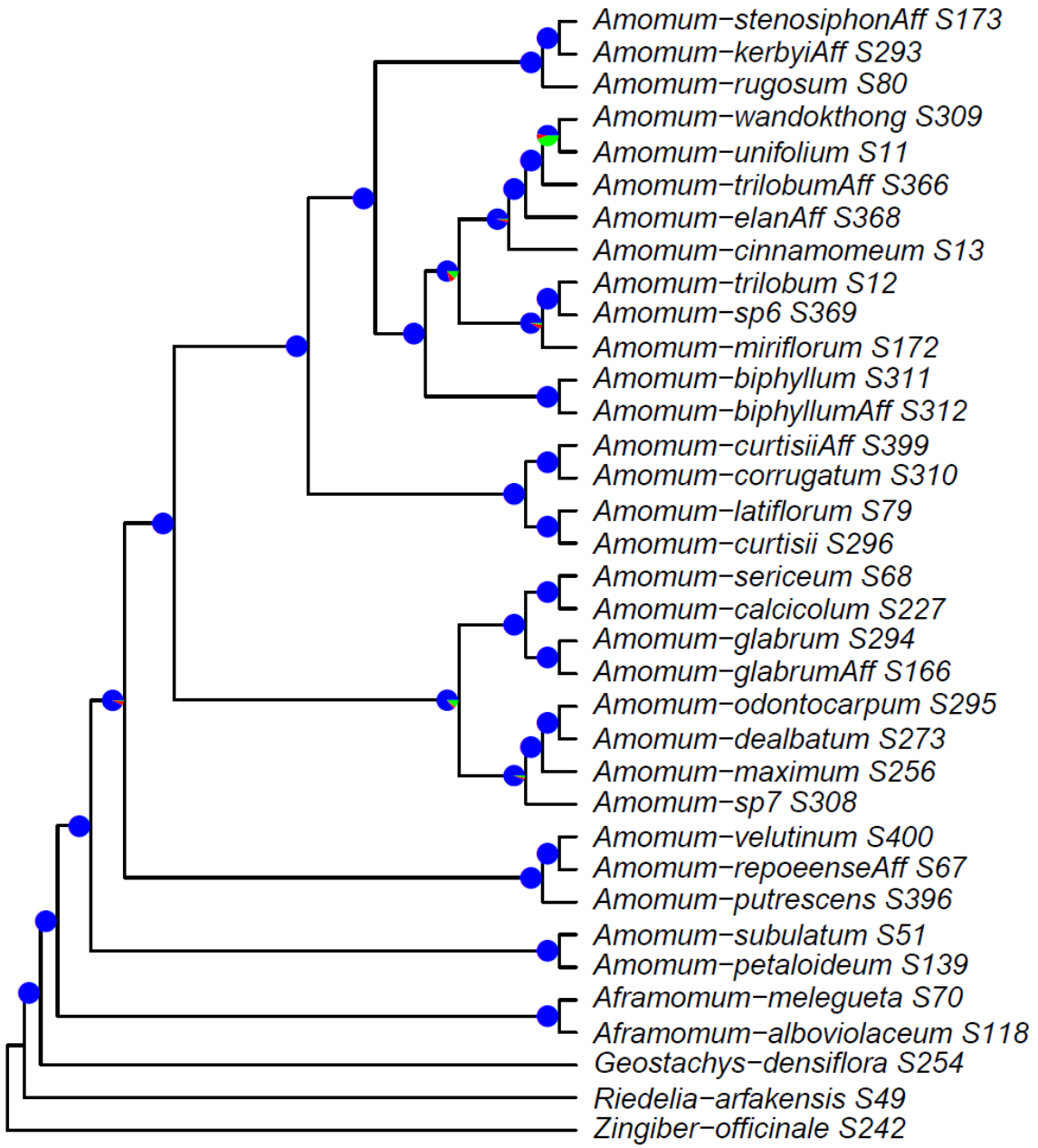
- ASTRAL
- ASTRID
- MRL (maximum representation with likelihood)
- concatenation FastTree
- concatenation ExaML (fully partitioned – PartitionFinder)
- BUCKy (Bayesian concordance analysis)

- neighbour network
- SuperQ network (supernetwork from quartets)
- quartet sampling

- PhyParts

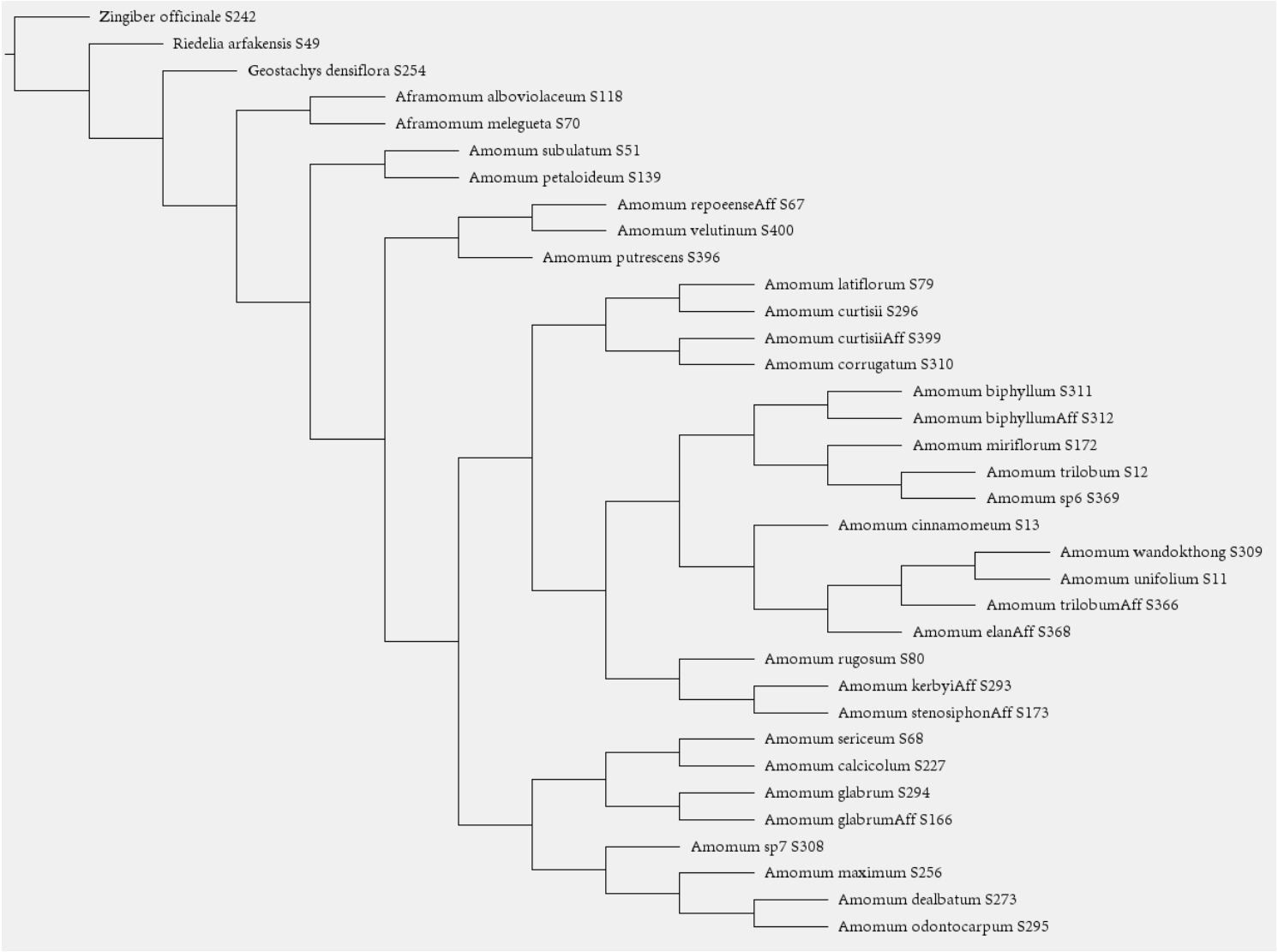
ASTRAL tree

- standard
- LPP, MLBS
- combination of trees
  - main tree
  - greedy consensus
  - bootstrap support
- ‘-t 4’ quartet scoring



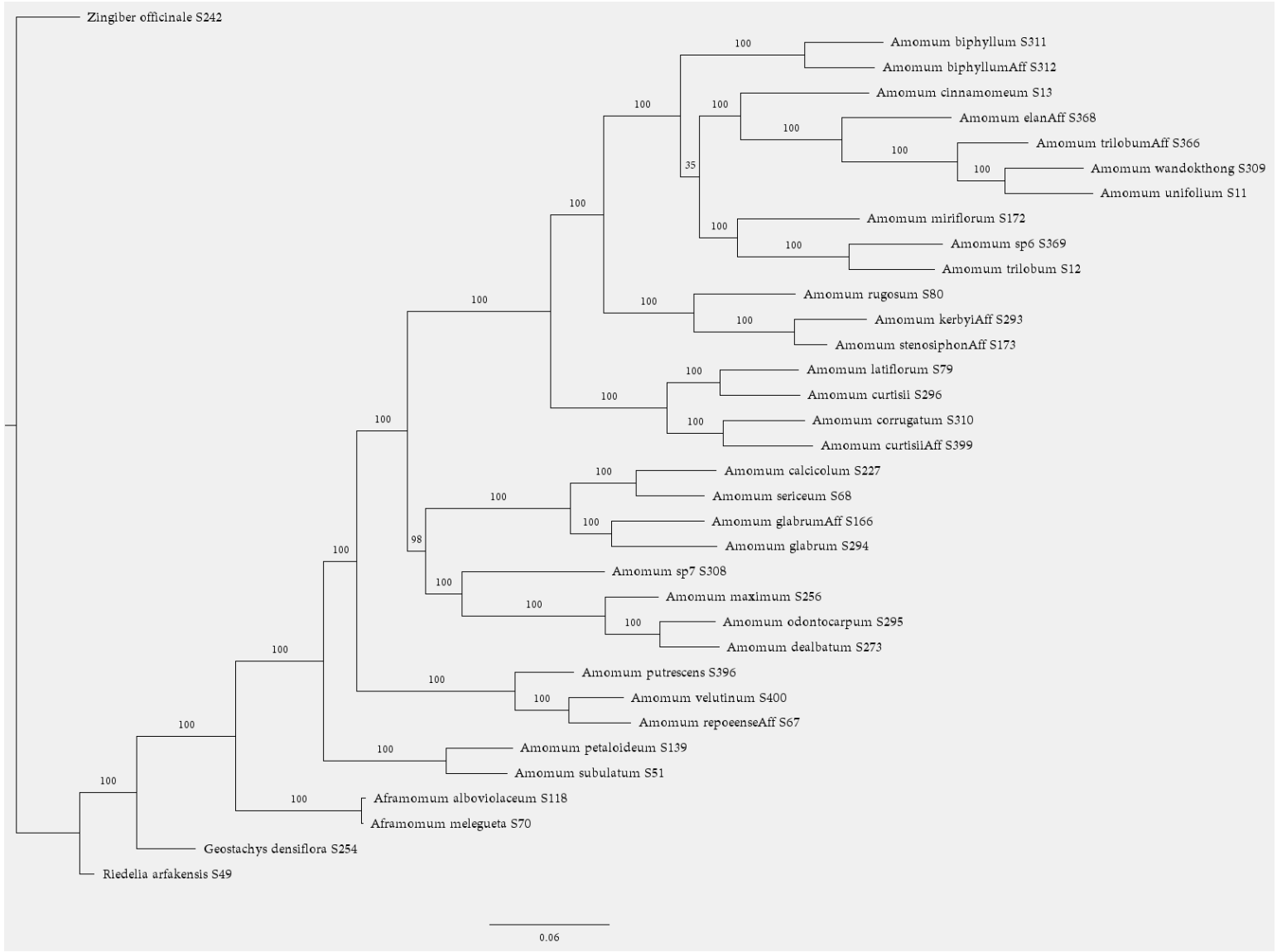
ASTRID tree

- standard
- MLBS
- combination of trees



MRL tree

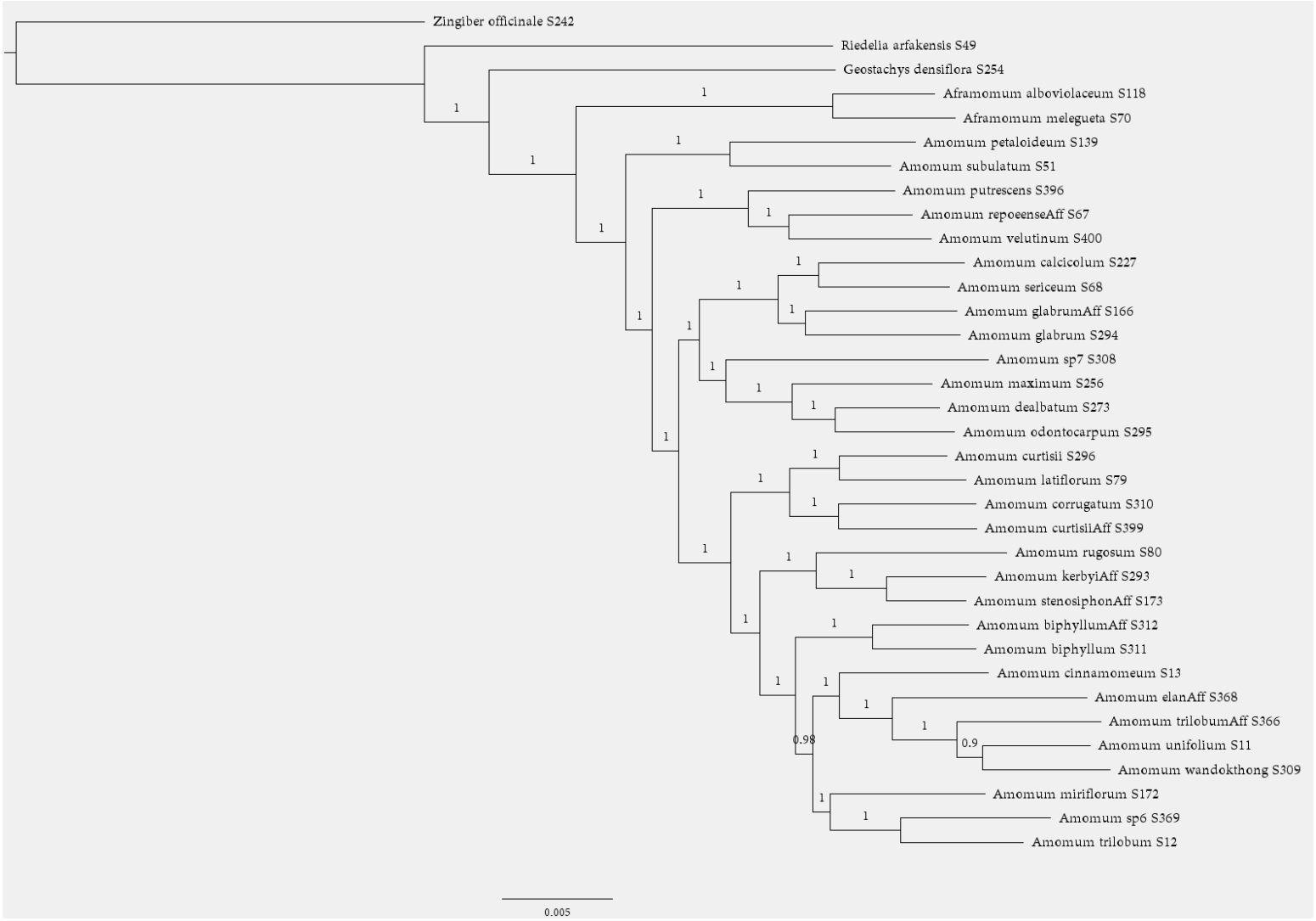
- BINGAMMA model
- RAxML with standard bootstrapping



Concatenated tree

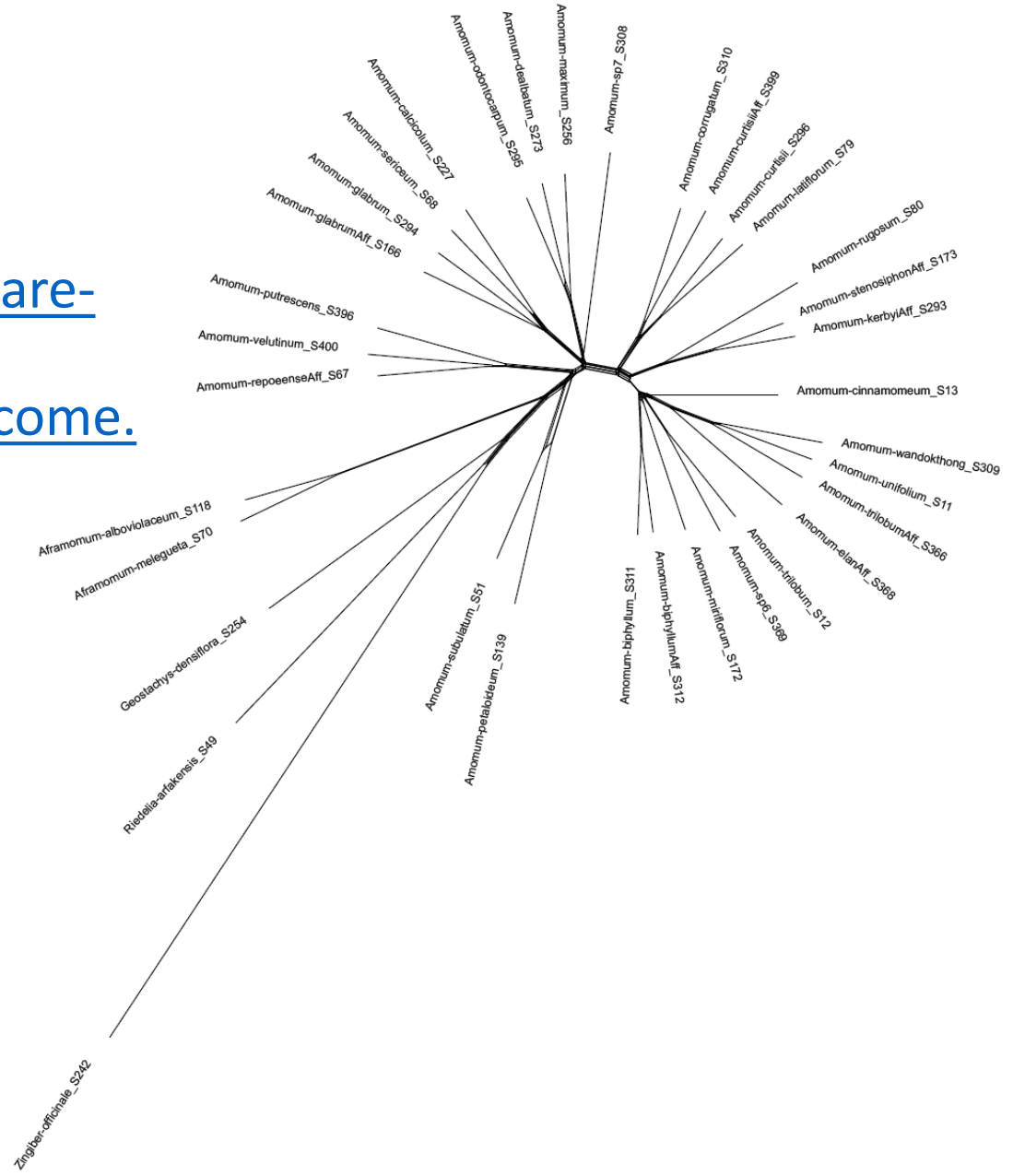
- FastTree or ExaML
- missing % table

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

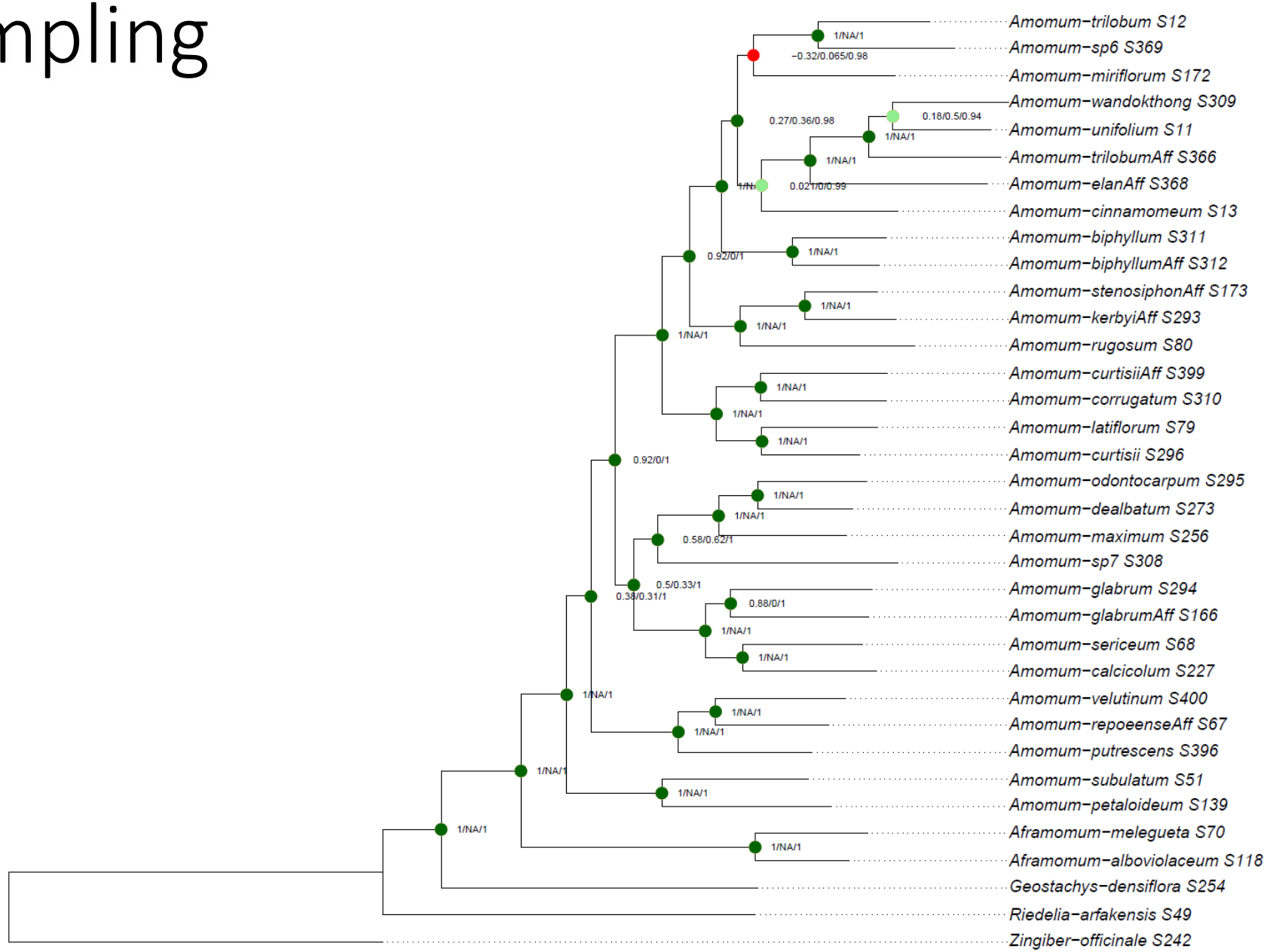


Neighbour network

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



Quartet sampling



PhyParts

- gene tree vs. species tree discordance
- for every node
  - nr concordant gene trees (blue)
  - nr discordant trees (red, green)