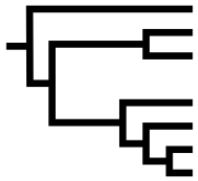


Phylogenetic Pipeline

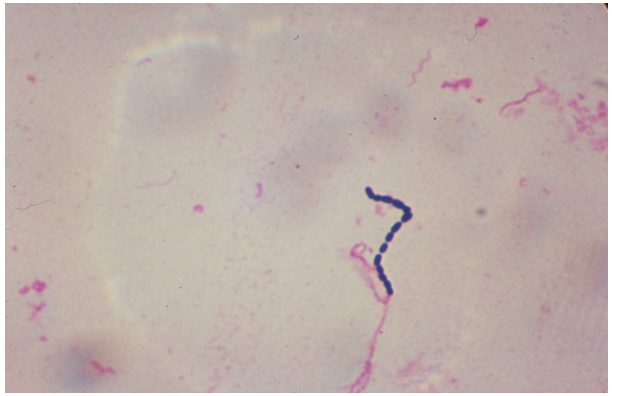
Comparative Genomics

Younginn Park

Introduction

Streptococcus genus

Genus *Streptococcus*



Gram staining on Streptococcus bacteria, http://www.usda.gov/oc/photo/99c0649.jpg

- Gram-positive spherical bacteria
- Clinical significance
  - pathogenic species causing various kinds of infections (*S. pyogenes, S. pneumoniae*)
  - significant in diagnostics, microbial marker for other diseases (J.Yang et al. 2023)
- Dairy industry
  - starter culture to produce fermented milk, yoghurt and cheese (*S. thermophilus*)

- *Lactococcus* – once grouped together with *Streptococcus*, separated in 1984, important e.g. in dairy industries (*Lactococcus lactis*)
  - 2 representatives will be used in this project as outgroup

Reference article

UNIVERSIDAD DE CHILE

Molecular phylogeny and a taxonomic proposal for the genus *Streptococcus*

F. Póntigo[1], M. Moraga[1,2] and S.V. Flores[1]

[1]Laboratorio de Antropología, Departamento de Antropología, Facultad de Ciencias Sociales, Universidad de Chile, Santiago, Chile
[2]Programa de Genética Humana, Instituto de Ciencias Biomédicas, Universidad de Chile, Santiago, Chile

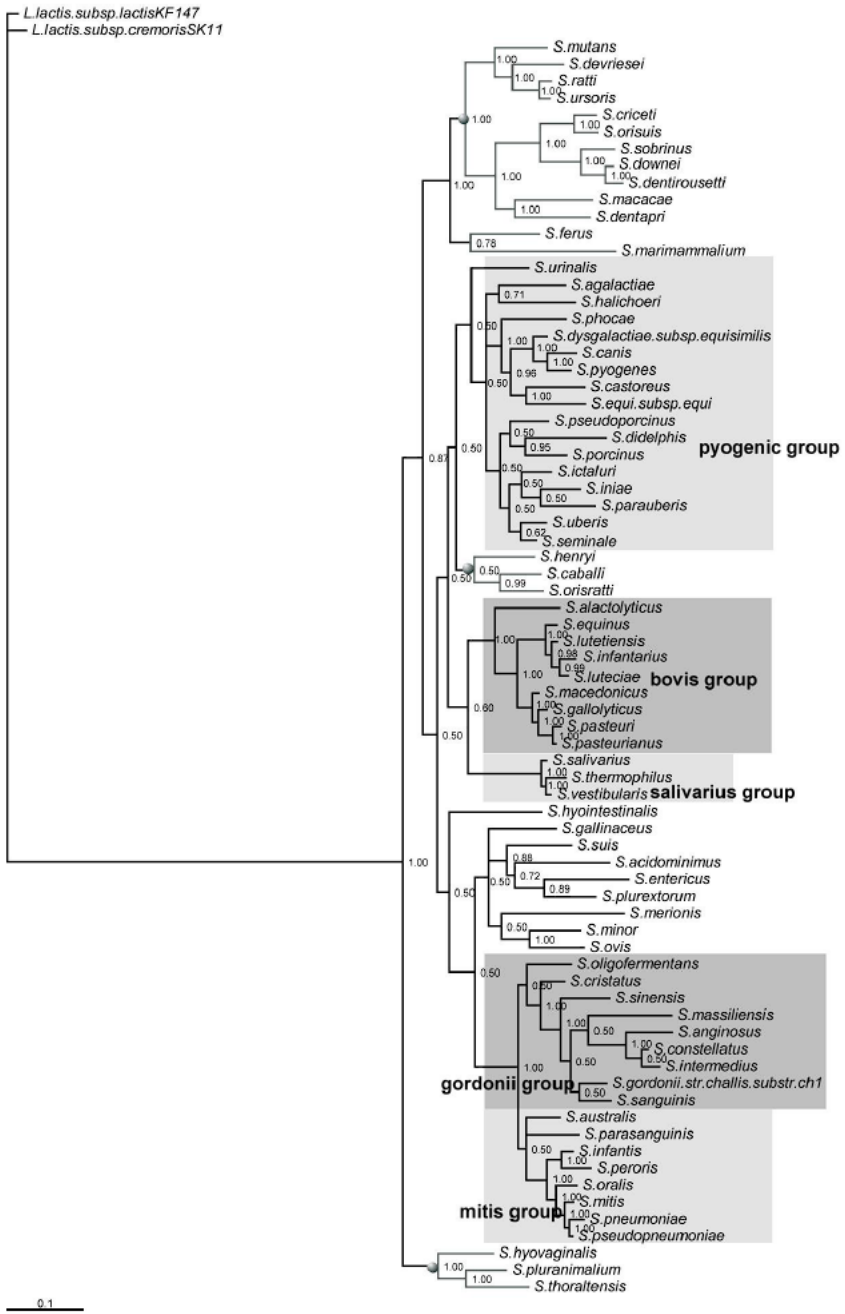
Species tree

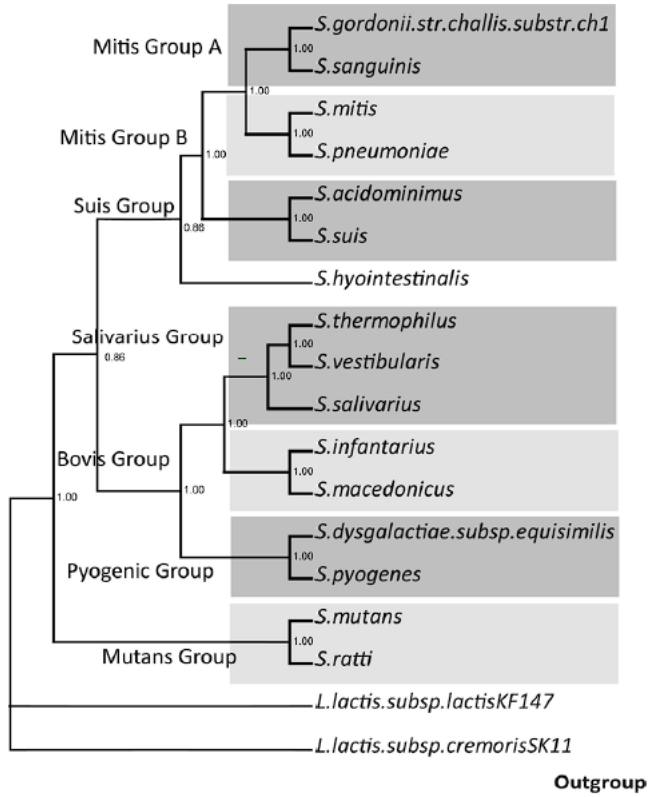
- Selected representative genes – functionally significant, having a balance between conservation and variation across species (*sodA, tuf, rpoB, recN, dnaJ, gyrB, rnpB*)
- Constructed using majority consensus (dendropy) based on Maximum Likelihood trees (PhyML)
- Amino acid evolution models selected according to AIC (Akaike Information Criterion) using ProtTest3 (WAG, RtREV)

F. Póntigo, M. Moraga, S.V. Flores, 2015

Table 1. List of species, accession numbers and size (bp) of each gene fragment.

Species	*sodA*		*tuf*		*rpoB*		*recN*	
	bp	Accession No.	bp	Accession No.	bp	Accession No.	bp	Accession No.
S. mutans	612	AE014133.2	1197	AE014133.2	3105	AP010655.1	1249	EU917289.1
S. agalactiae	609	AE009948.1	1197	AL766847.1	3105	AE009948.1	1249	EU917242.1
S. acidominimus	435	Z95892.1	761	AY266992.1	691	AF535181.1	1249	EU917241.1
S. anginosus	453	FJ712177.1	826	AF276257.1	3105	AF535183.1	1249	EU917248.1
S. alactolyticus	435	AJ297185.1	-	-	680	DQ232445.1	1249	EU917226.1





Majority consensus tree obtained by a Bayesian analysis containing all species (left) and containing two members of each group (above), plus the outgroup.

Selected species

3 representatives from each group:

Table 2. Species of *Streptococcus*, and outgroup species, analyzed in this study.

Mutans group	*S. ratti*, *S. ursoris*, *S. devriesei*, *S. mutans*, *S. macacae*, *S. ferus*, *S. dentapri*, *S. downei*, *S. dentirousetti*, *S. sobrinus*, *S. dentirousetti*, *S. sobrinus*, *S. criceti*, *S. orisuis*, *S. merionis*, *S. caballi*, *S. henryi*, *S. orisratti*, *S. pluranimalium*, *S. thoraltensis*, *S. hyovaginalis*
Pyogenic group	*S. dysgalactiae*, *S. pyogenes*, *S. canis*, *S. castoreus*, *S. equi* subsp *equi*, *S. halichoeri*, *S. phocae*, *S. porcinus*, *S. pseudoporcinus*, *S. didelphis*, *S. uberis*, *S. seminale*, *S. iniae*, *S. ictaluri*, *S. urinalis*, *S. parauberis*, *S. marimammalium*, *S. agalactiae*
Bovis group	*S. equinus*, *S. lutetiensis*, *S. luteciae*, *S. infantarius*, *S. gallolyticus*, *S. pasteurianus*, *S. macedonicus*, *S. alactolyticus*
Suis group	*S. entericus*, *S. plurextorum*, *S. suis*, *S. acidominimus*, *S. minor*, *S. ovis*, *S. gallinaceus*
Mitis group A	*S. constellatus*, *S. intermedius*, *S. anginosus*, *S. massiliensis*, *S. cristatus*, *S. sinensis*, *S. gordonii*, *S. sanguinis*
Mitis group B	*S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, *S. oligofermentans*, *S. infantis*, *S. peroris*, *S. oralis*, *S. australis*, *S. parasanguinis*
Salivarius group	*S. vestibularis*, *S. salivarius*, *S. thermophilus*
Outgroup	*Lactococcus lactis*, *L. cremoris*

F. Póntigo et al. 2015

List of species selected for this project (right)

Mutans group

- SRAT: *Streptococcus ratti*
- SDOW: *Streptococcus downei*
- SDEV: *Streptococcus devriesei*

Pyogenic group

- SDYS: *Streptococcus dysgalactiae*
- SPYO: *Streptococcus pyogenes*
- SCAN: *Streptococcus canis*

Bovis group

- SEQU: *Streptococcus equinus*
- SLUT: *Streptococcus lutetiensis*
- SINF: *Streptococcus infantarius*

Suis group

- SENT: *Streptococcus entericus*
- SPLU: *Streptococcus plurextorum*
- SSUI: *Streptococcus suis*

Mitis group A

- SCON: *Streptococcus constellatus*
- SINT: *Streptococcus intermedius*
- SANG: *Streptococcus anginosus*

Mitis group B

- SPNE: *Streptococcus pneumoniae*
- SPSE: *Streptococcus pseudopneumoniae*
- SMIT: *Streptococcus mitis*

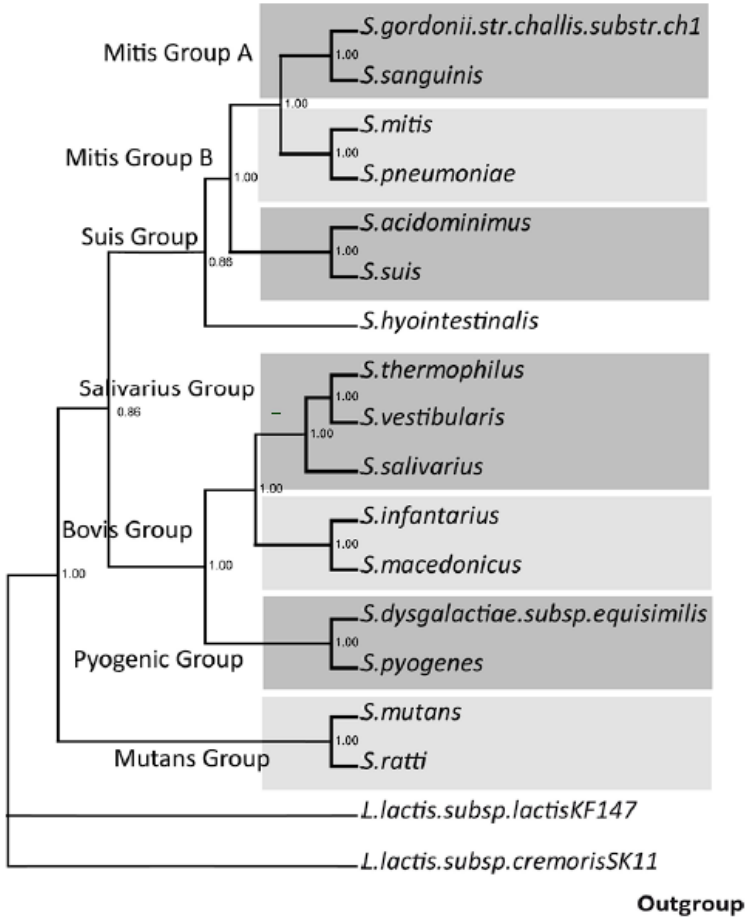
Salivarious group

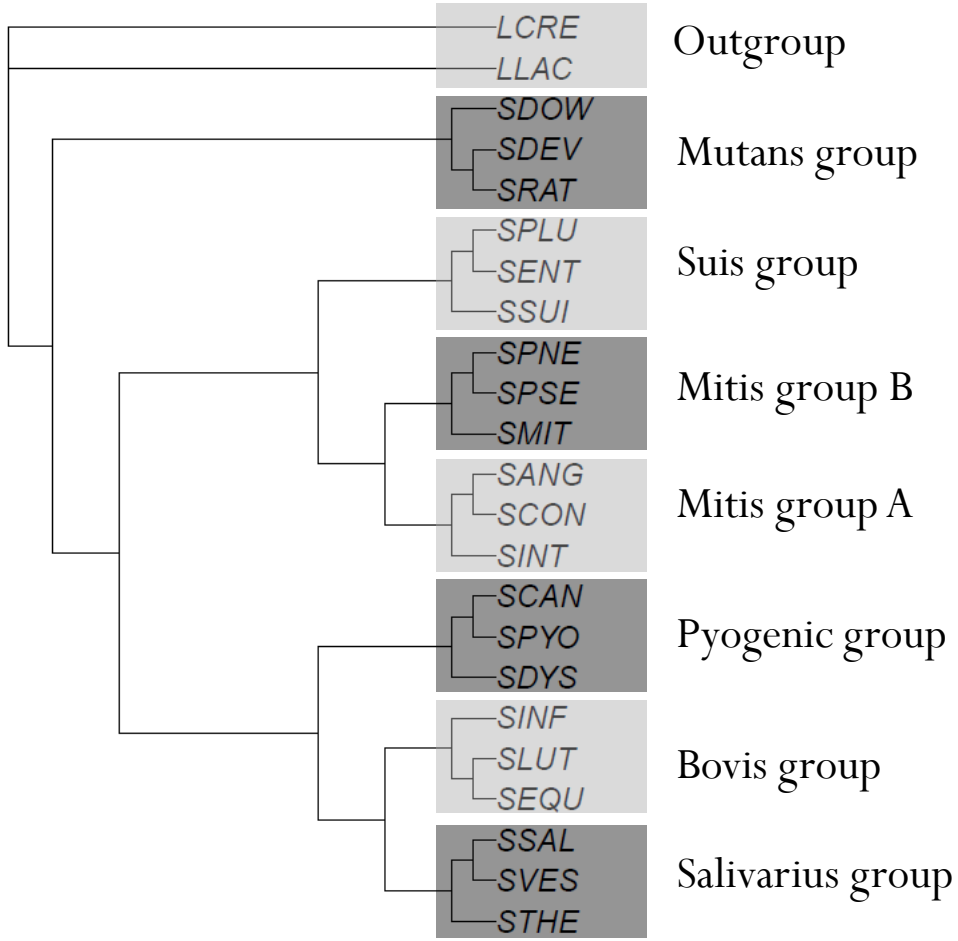
- SVES: *Streptococcus vestibularis*
- SSAL: *Streptococcus salivarius*
- STHE: *Streptococcus thermophilus*

Outgroup

- LLAC: *Lactococcus lactis*
- LCRE: *Lactococcus cremoris*

Species tree reconstruction

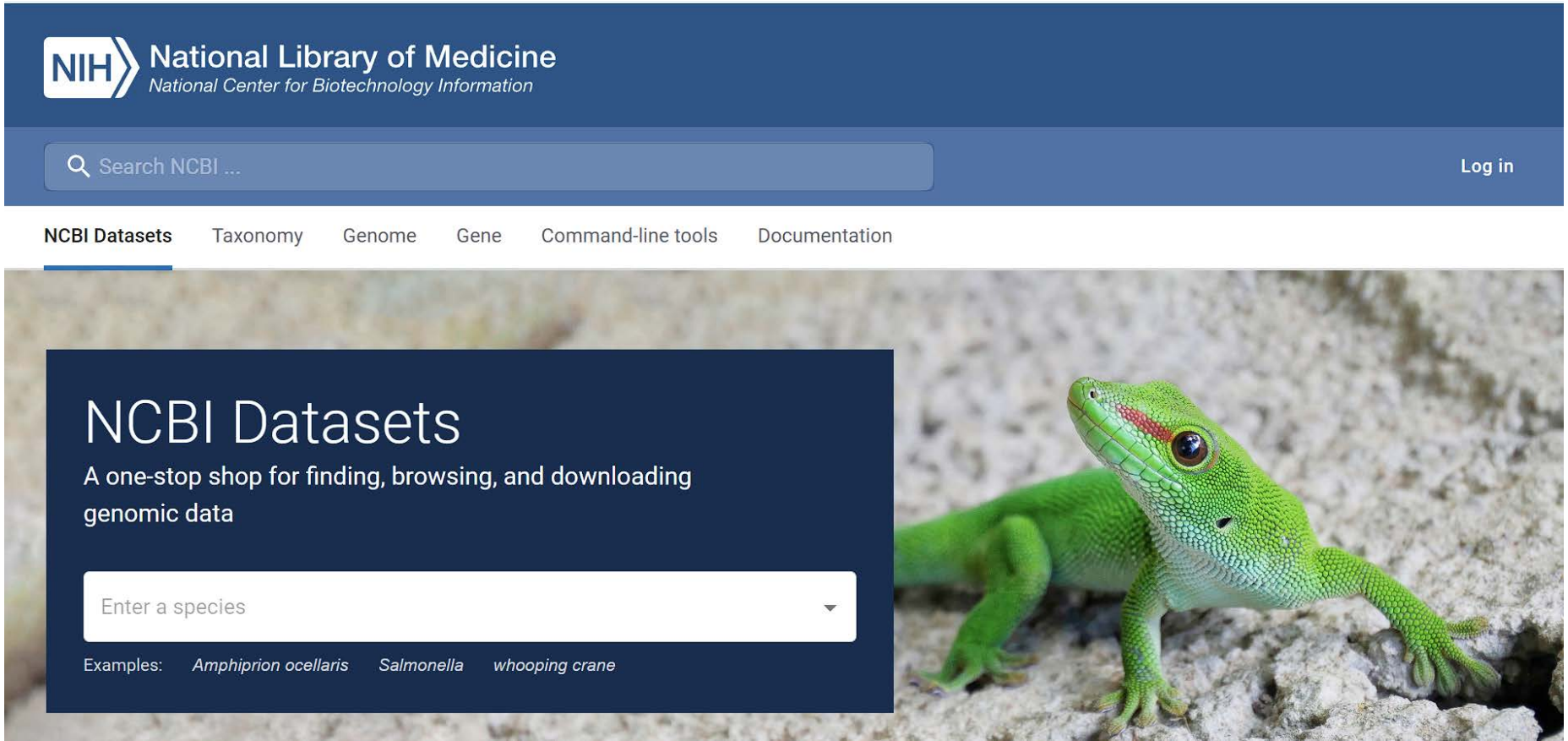




Majority consensus tree obtained by a Bayesian analysis containing two members of each group from the reference study (left) and the reconstructed tree used in this project (right) plus the outgroup.

NCBI Datasets

Proteomes obtained from NCBI Datasets Command Line Tool

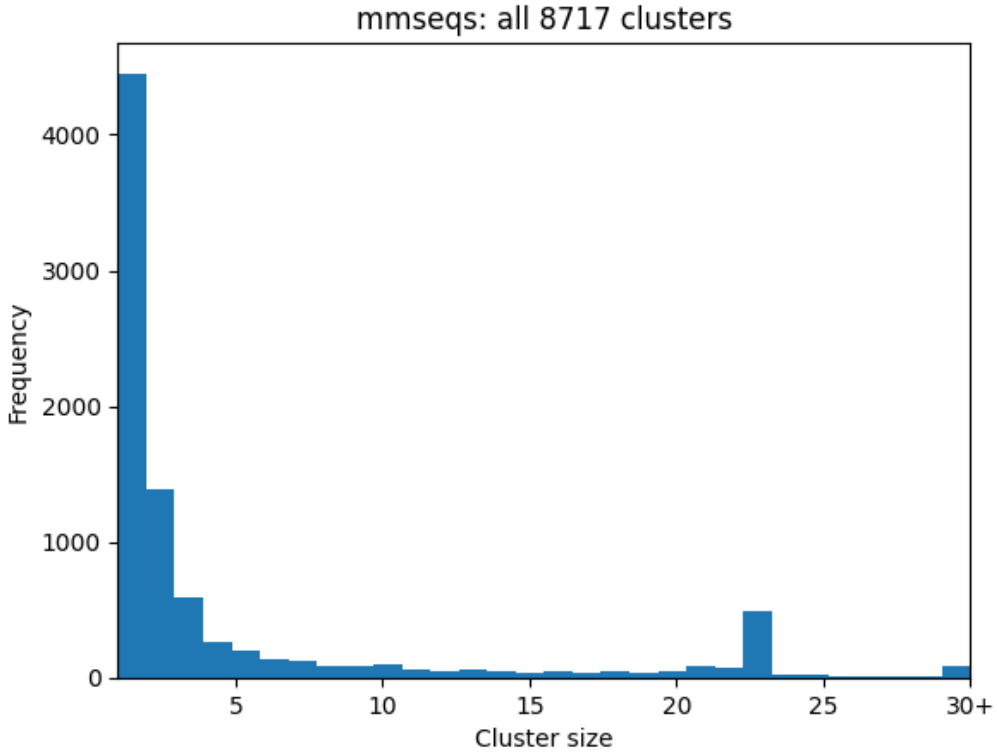


Methods

Clustering and Gene Families

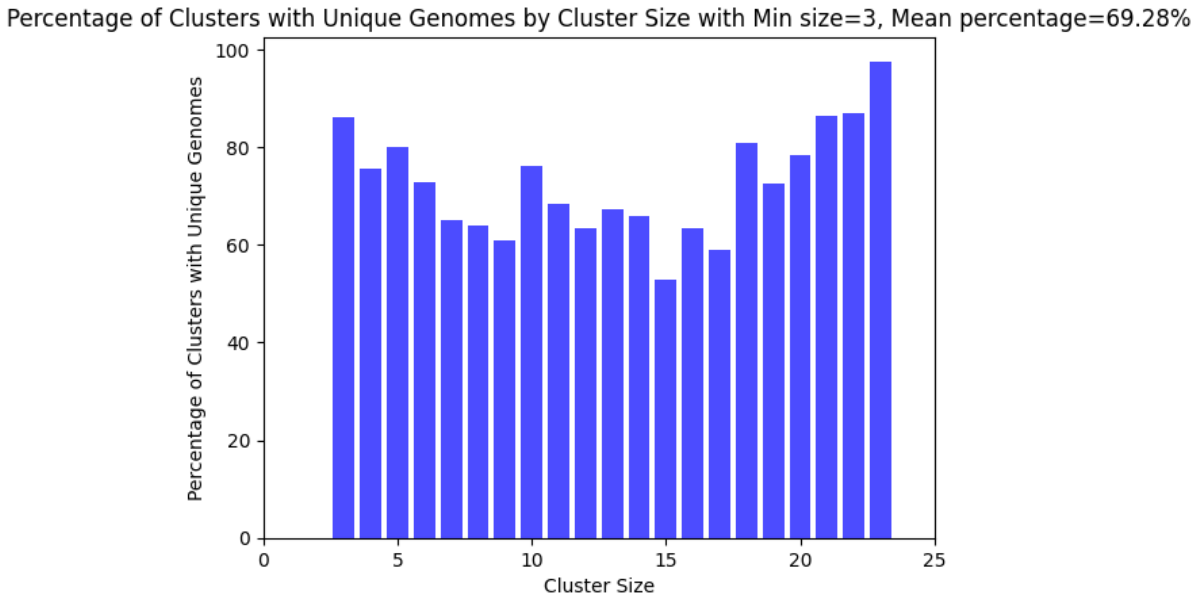
MMseqs

- Mmseqs - clustering based on similarities of k-mers between sequences
- Rule of thumb – „two sequences can be considered homologous if they are more than 30% identical over their entire lengths” (*Pearson 2013, An introduction to sequence similarity "homology" searching*)
- Minimum sequence identity (*--min_seq_id 0.3)*



Histograms of cluster sizes

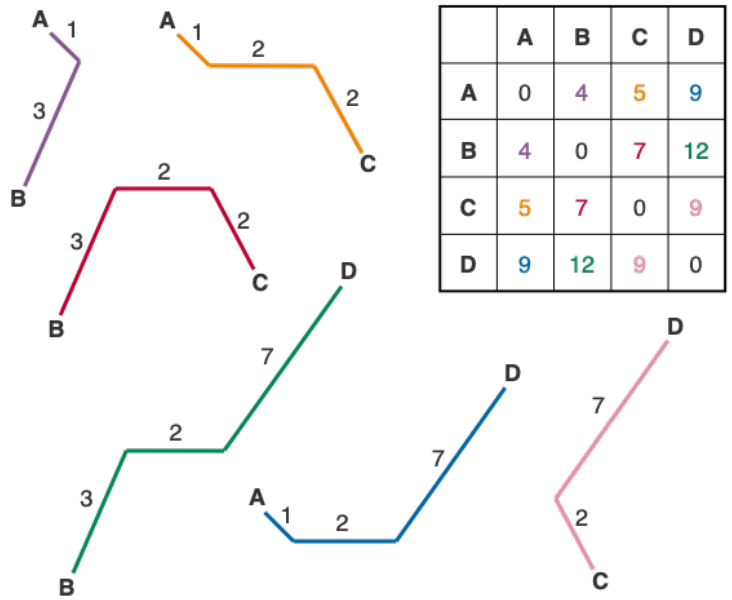
Non-orthologous clusters



Histograms of cluster sizes after removing small clusters (singletons and duplets), clusters bigger than the number of species (>23) and clusters with duplicate genome identifiers (left). Percentage of clusters with desireable size (3-23) that had unique genome identifiers, 1-1 gene sets (right).

Gene trees

- Mafft for Multiple Sequence Alignment
- Neighbor-Joining Method (biopython, BLOSUM62)
  - Joining most closely related taxa based on pairwise distances



	A	B	C	D
A	0	4	5	9
B	4	0	7	12
C	5	7	0	9
D	9	12	9	0

A schematic of the NJ algorithm
https://www.tenderisthebyte.com/blog/2022/08/31/neighbor-joining-trees/

Methods
Genome Trees

Majority Consensus Tree

- Only trees with the full set of taxa could be used (461 gene trees)
- Implementation from DendroPy python library with 0.5 cutoff (inclusive frequency threshold) for unrooted trees (splits)
- Root the resulting tree with the outgroup

Supertree

- fasturec
- Heuristic tree search (-Z) - initial quasi-consensus tree and hill-climbing steps under the duplication-loss cost (mixture of NNI, SPR and TSW)
- With and without paralogs
- Root the resulting tree with the outgroup

Bootstrapping

- Bootstrap replicate trees constructed with Biopython (NJ/100 replicates)
- SumTrees (bundled with DendroPy) for split support analysis of the gene trees
- Trees with high support get to stay for consensus tree and supertree construction
- Generally, splits with at least 0.7 support are considered „well-supported”

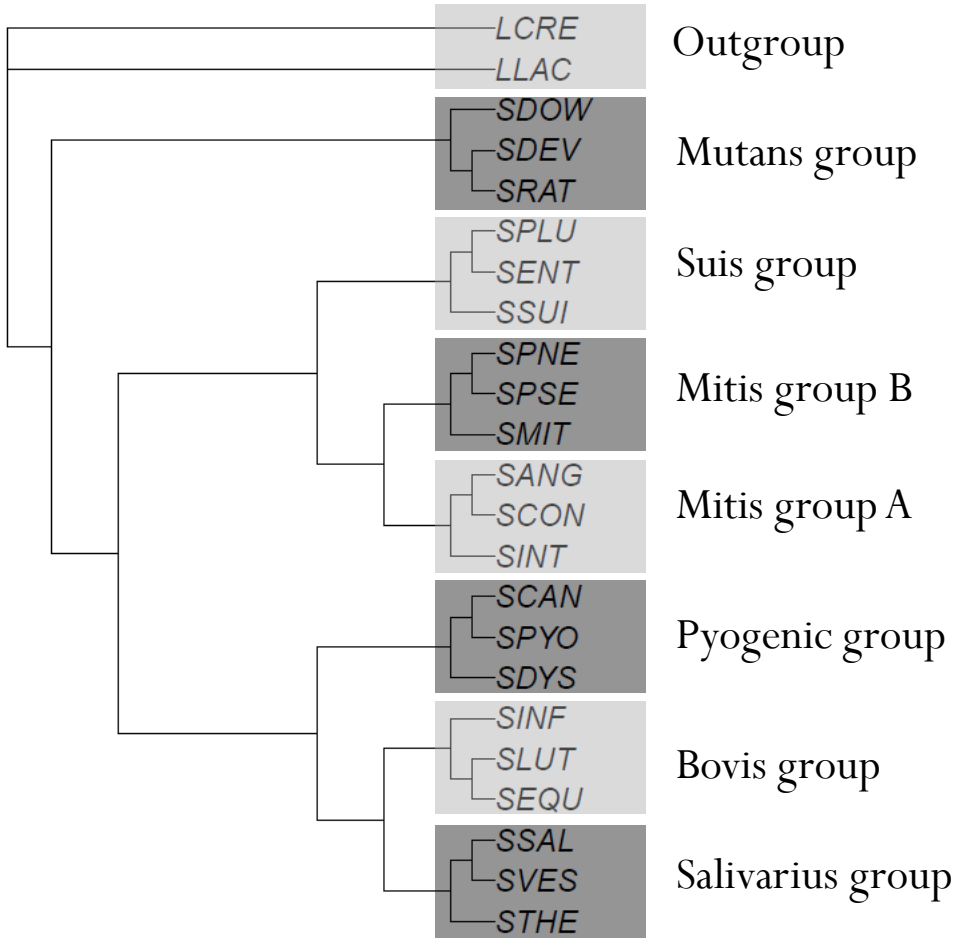
Defining a „good tree”

- Here, the criterion for trees with ‘good’ support – average support for splits in a tree has to be greater or equal 0.9 (gives room for a tree to be chosen even if some of the branches have lower support)
- 368 of 2161 trees in total were removed due to low branch support
- 206 of 461 trees with full set of taxa remained for majority consensus

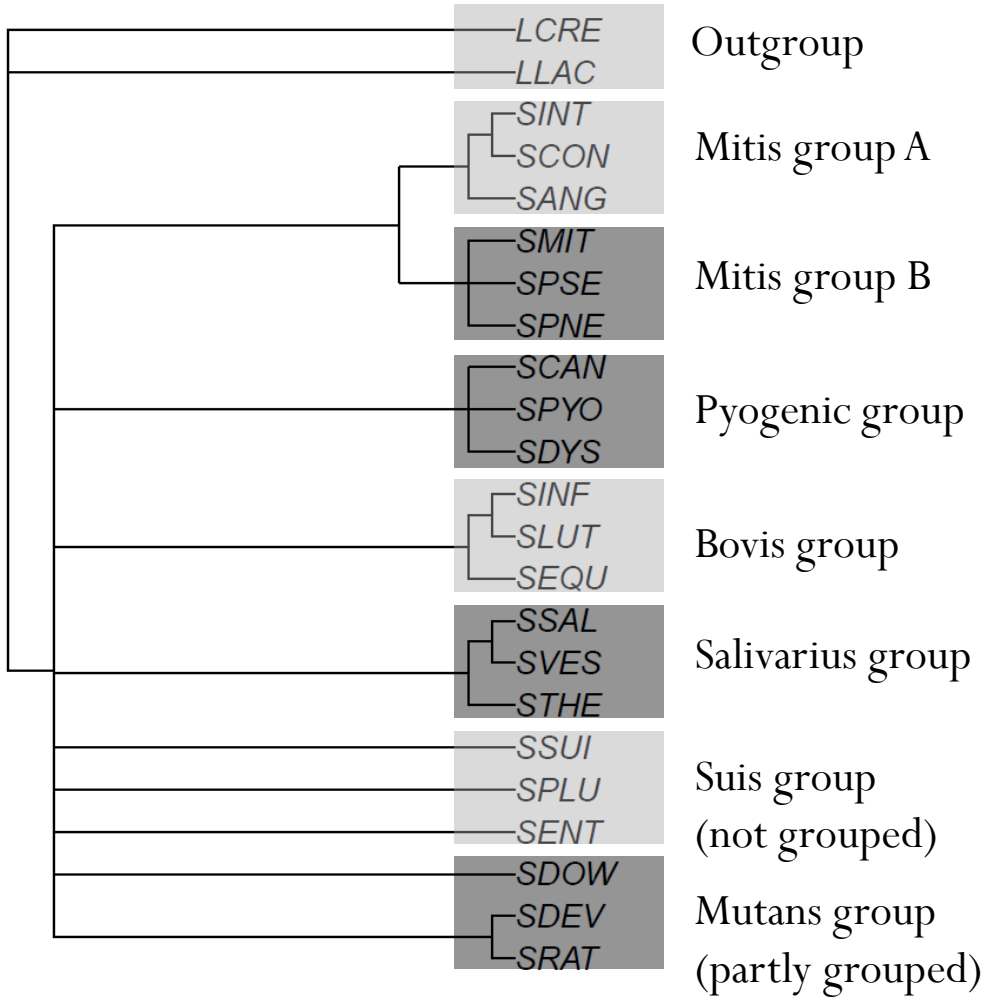
Results and Analysis

Visualizations done using packages ape and TreeDist in R

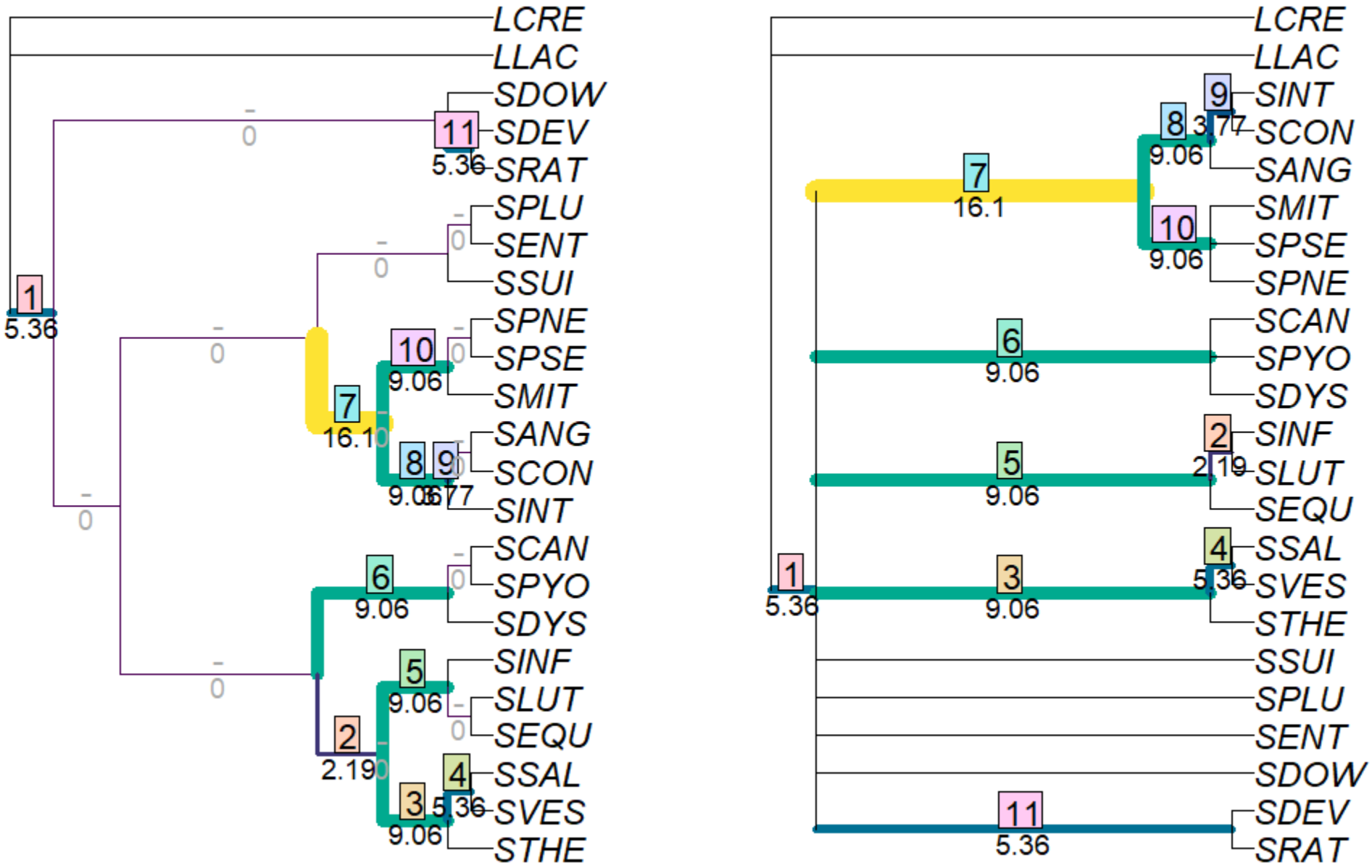
Consensus tree



Unresolved major clades between groups within *Streptococcus* genus (except for 2 Mitis groups)



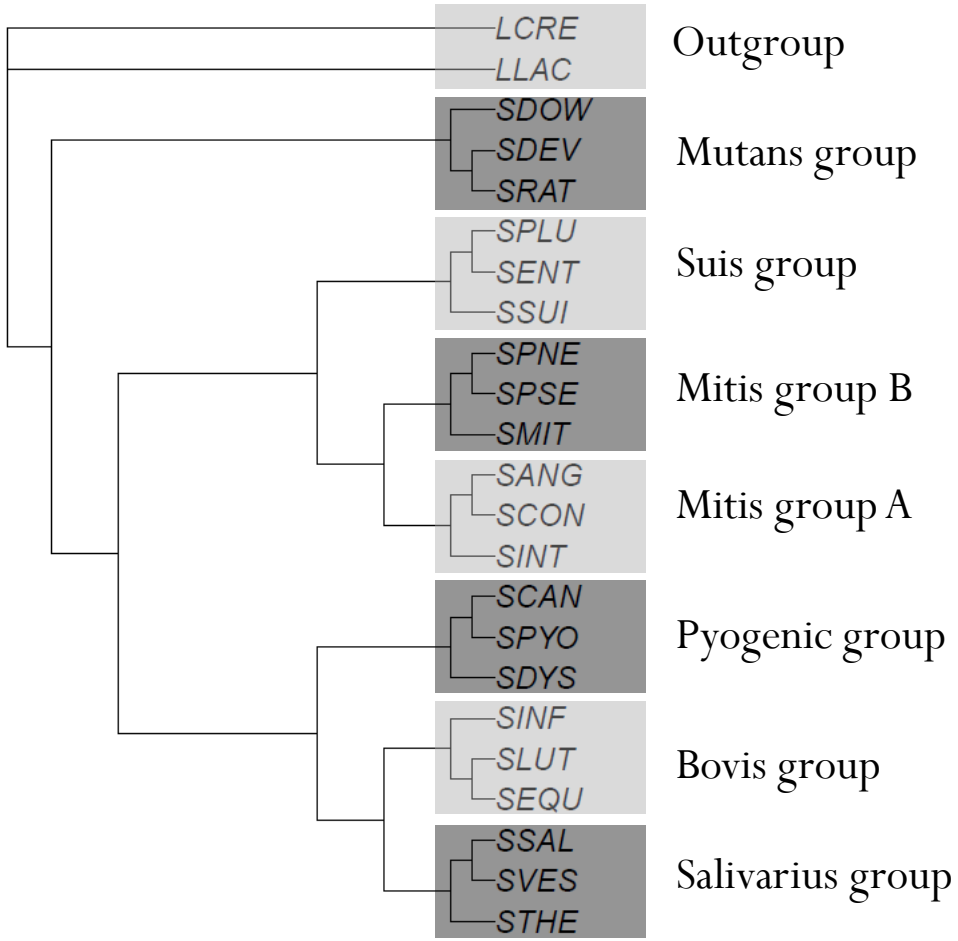
Species tree (left) and majority consensus tree (right) both rooted on the outgroup

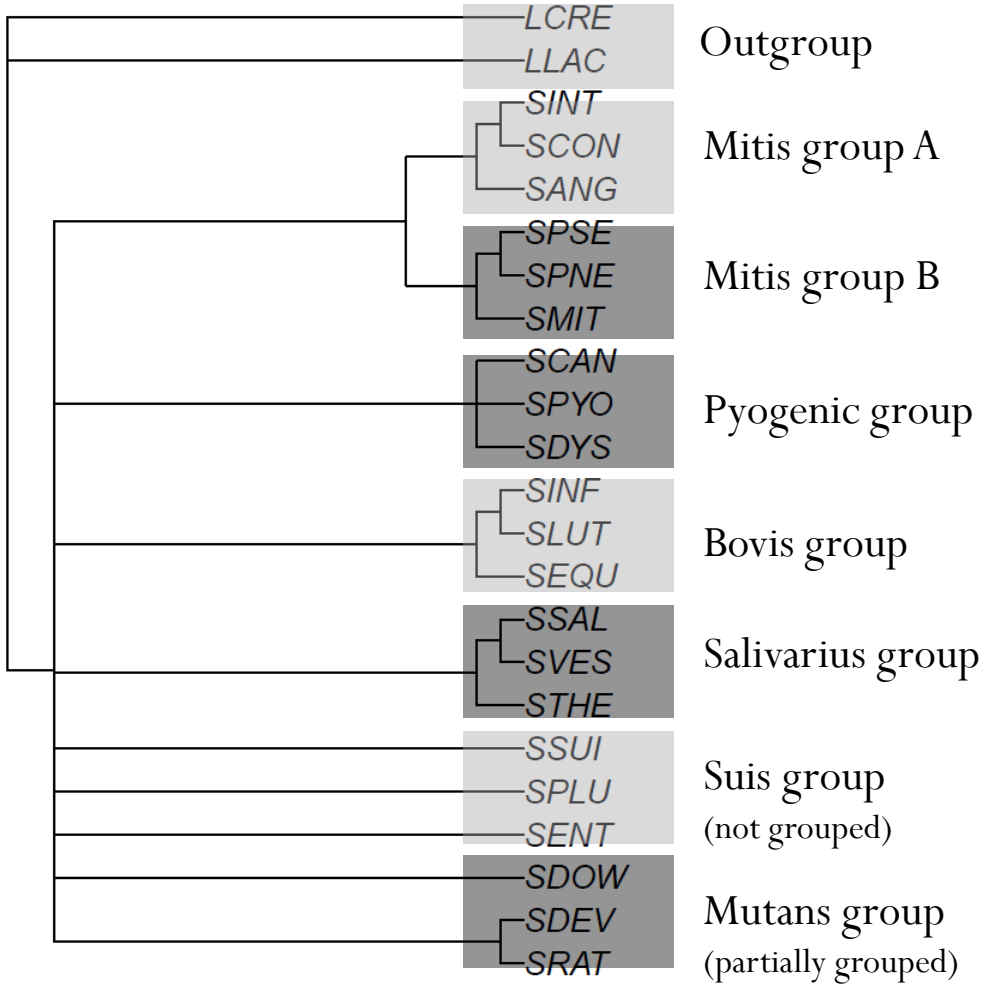


Species tree (left) and majority consensus tree (right) both rooted on the outgroup, with marked **shared** splits

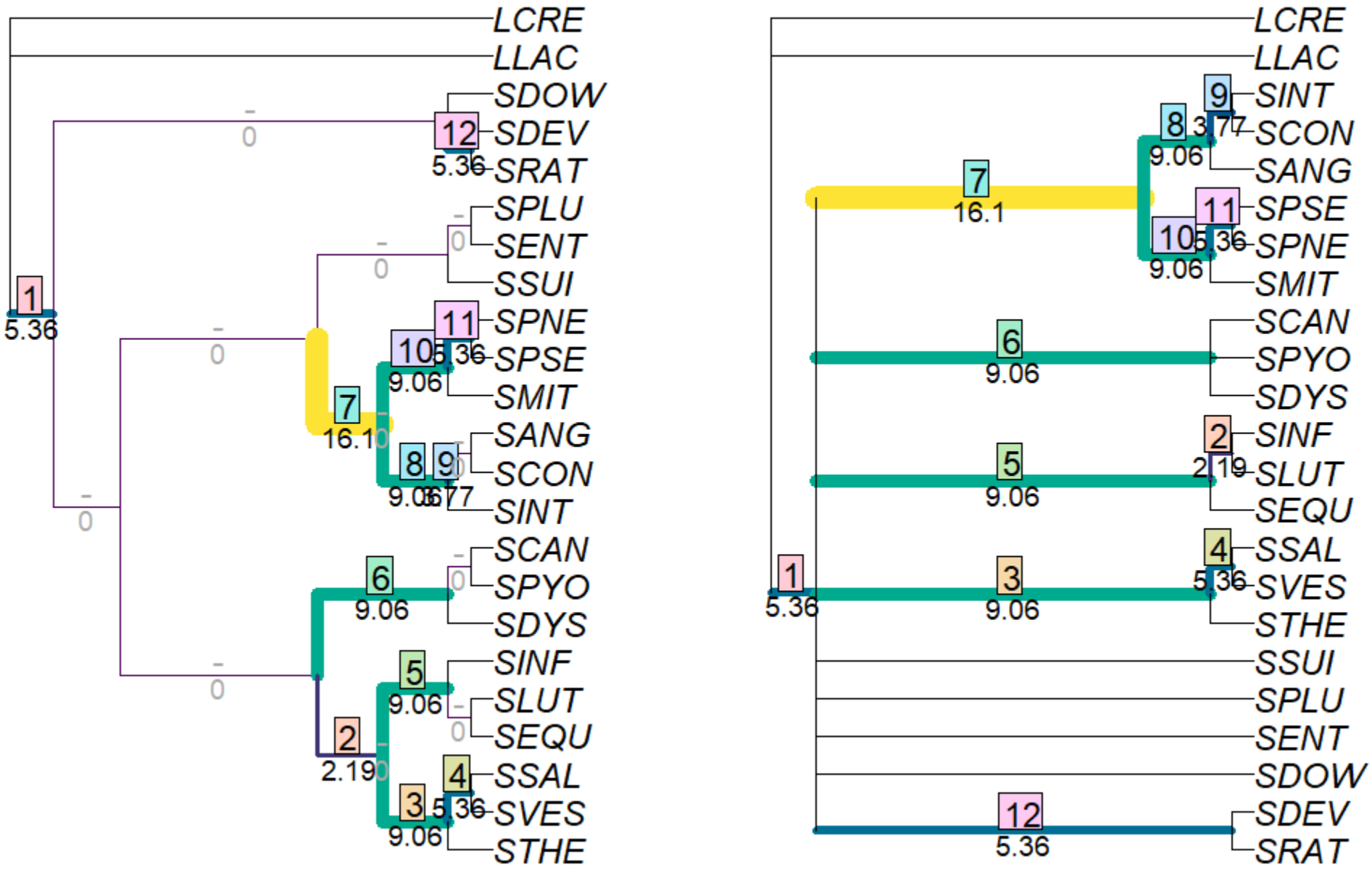
Consensus from bootstrap

Resolved multifurcation in Mitis group B compared to consensus tree without bootstrap



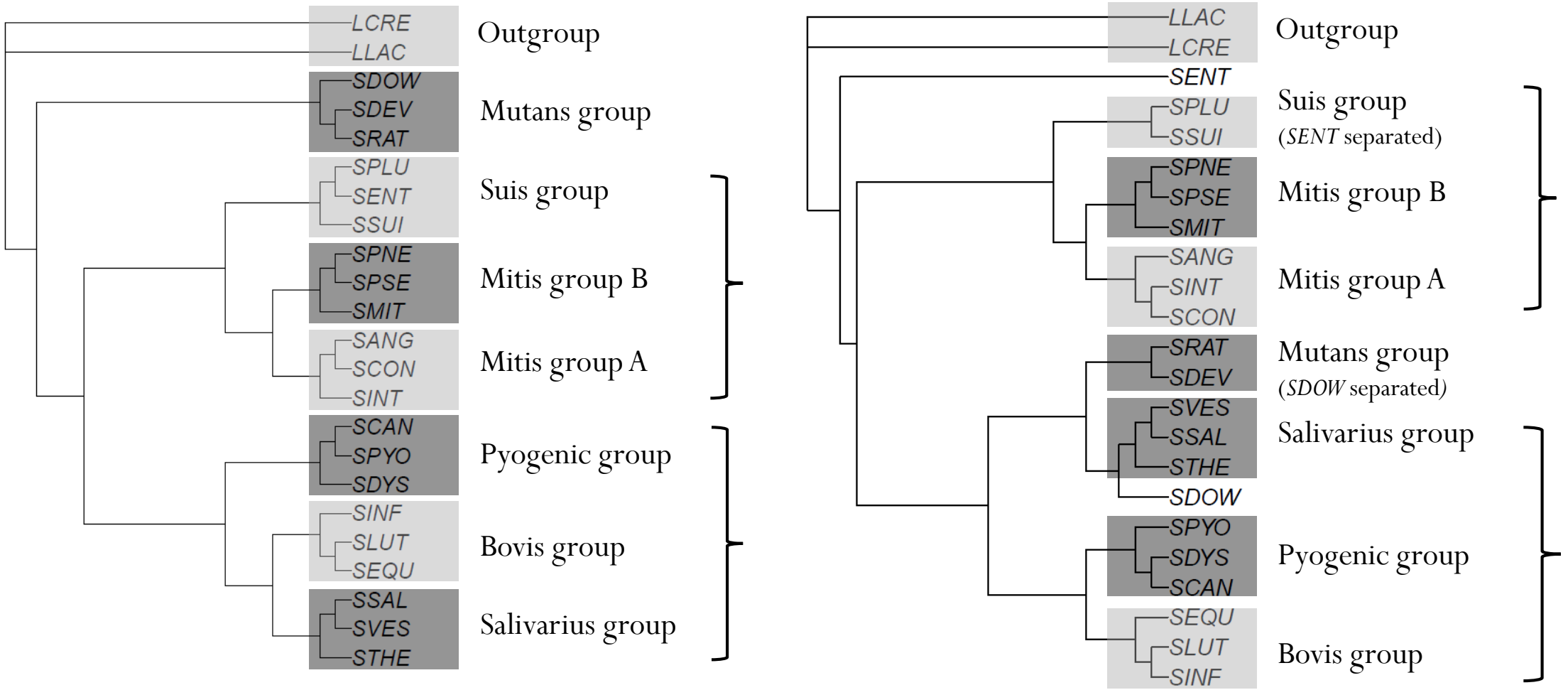


Species tree (left) and majority consensus tree made with bootstrap verified trees (right) both rooted on the otugroup

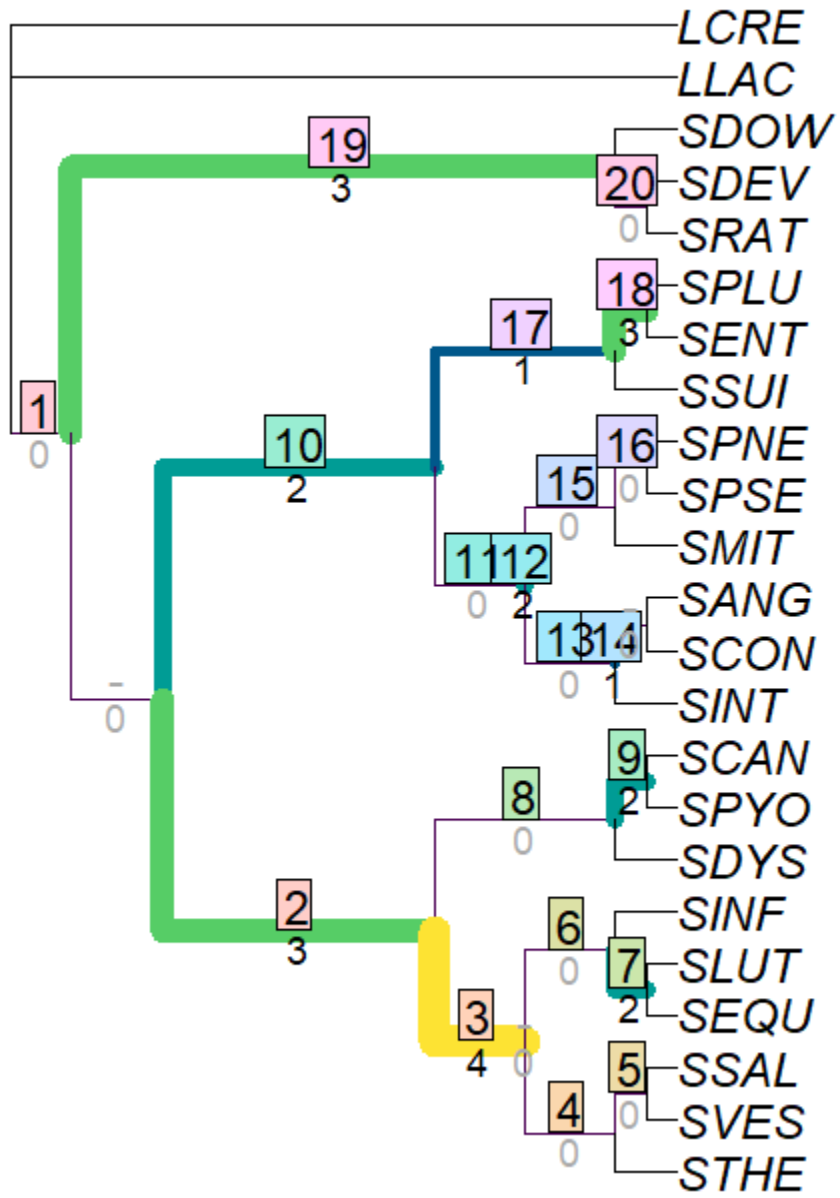


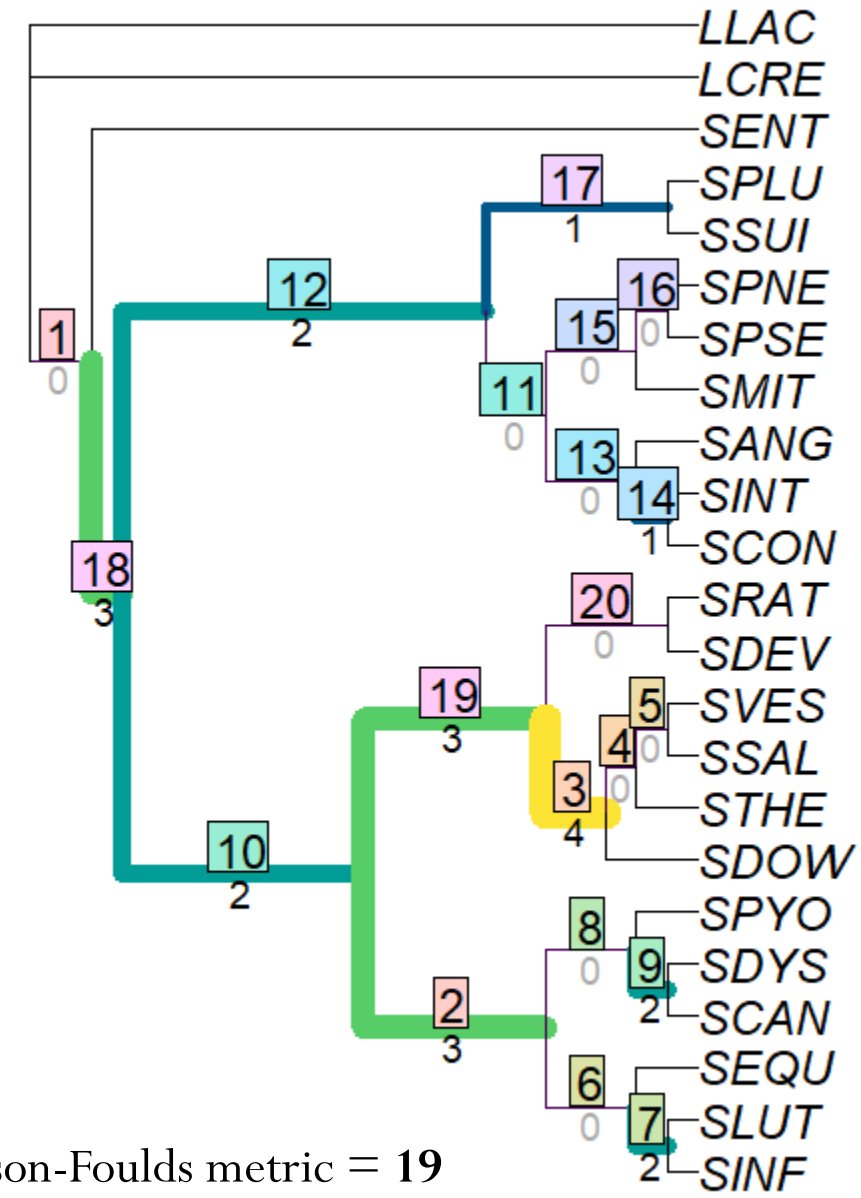
Species tree (left) and majority consensus tree made with bootstrap verified trees (right) both rooted on the outgroup, with marked **shared** splits

Supertree



Species tree (left) and supertree made with gene trees without paralogs (right), rooted on outgroup

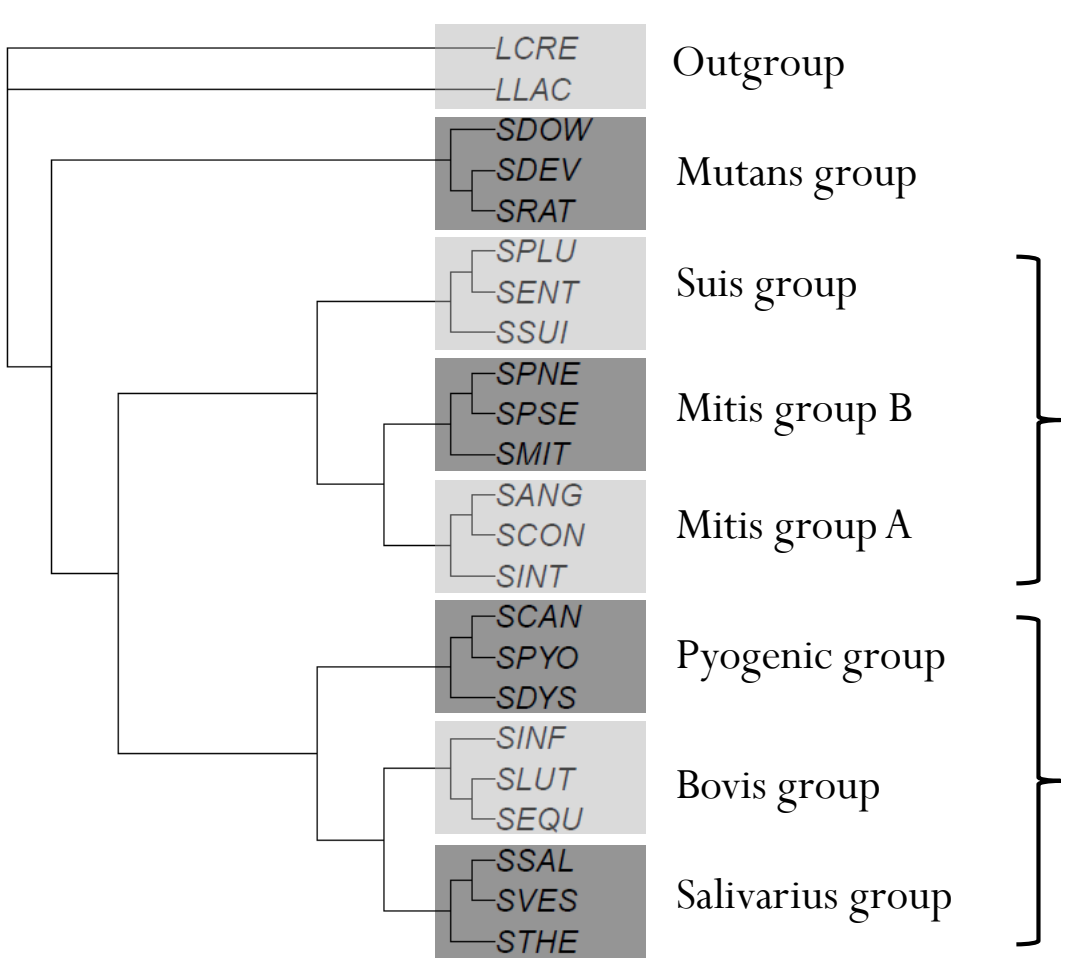


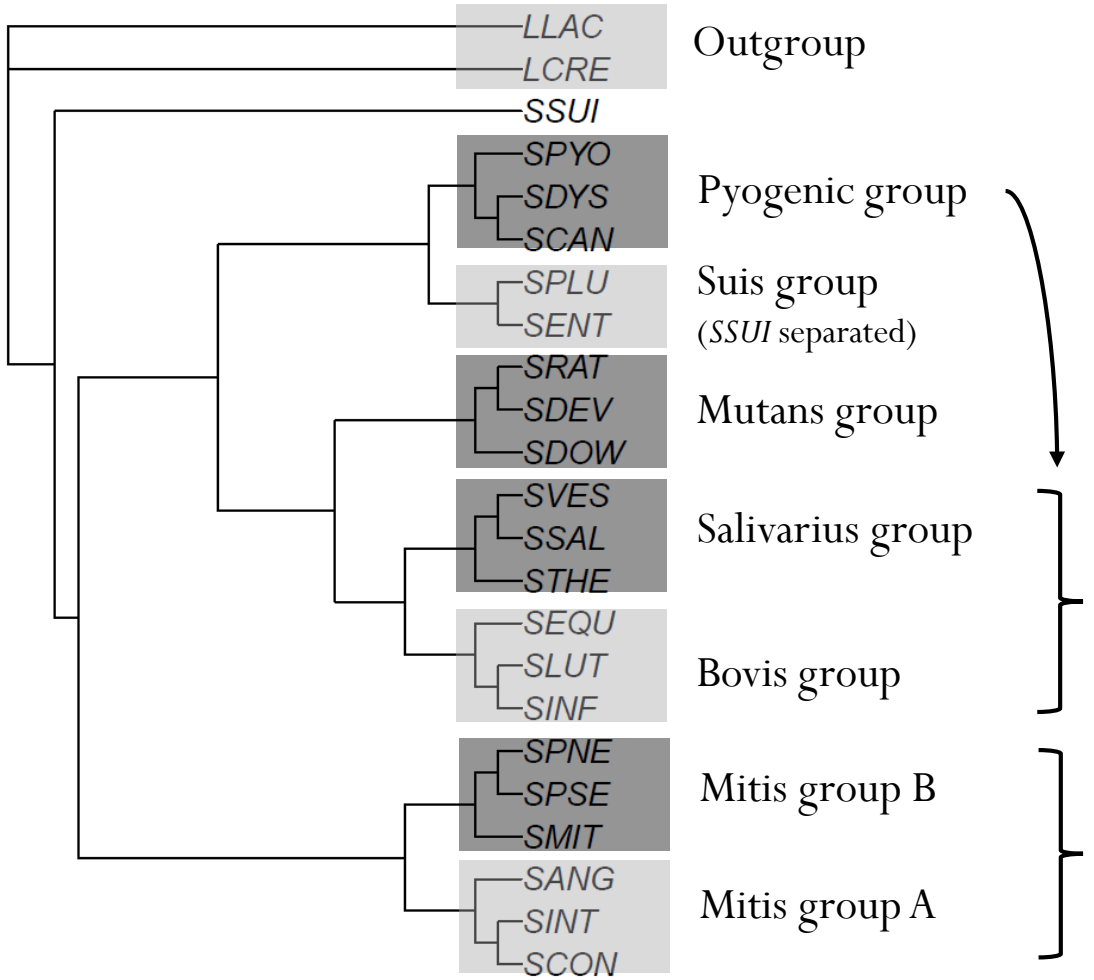


Robinson-Foulds metric = **19**

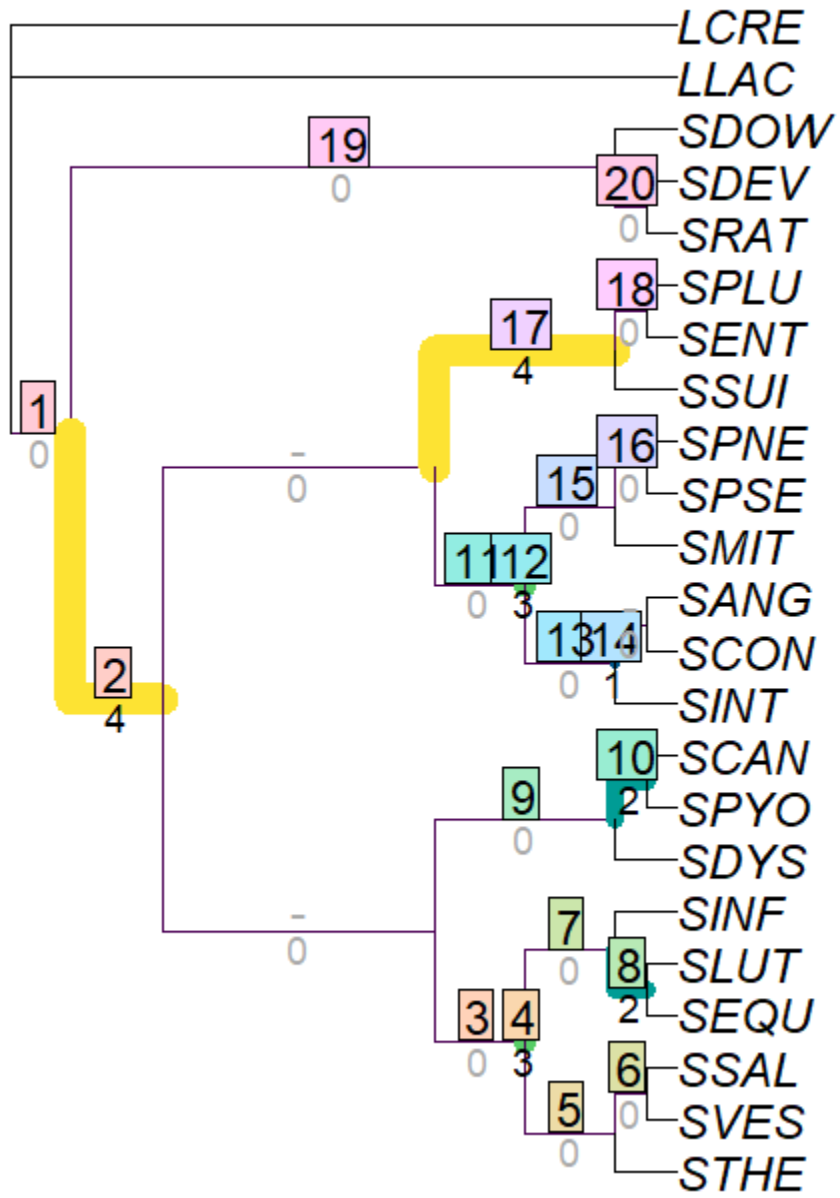
Species tree (left) and majority consensus tree made with bootstrap verified trees, DL=33899 (right) both rooted on the outgroup, with marked **unique** splits for both trees

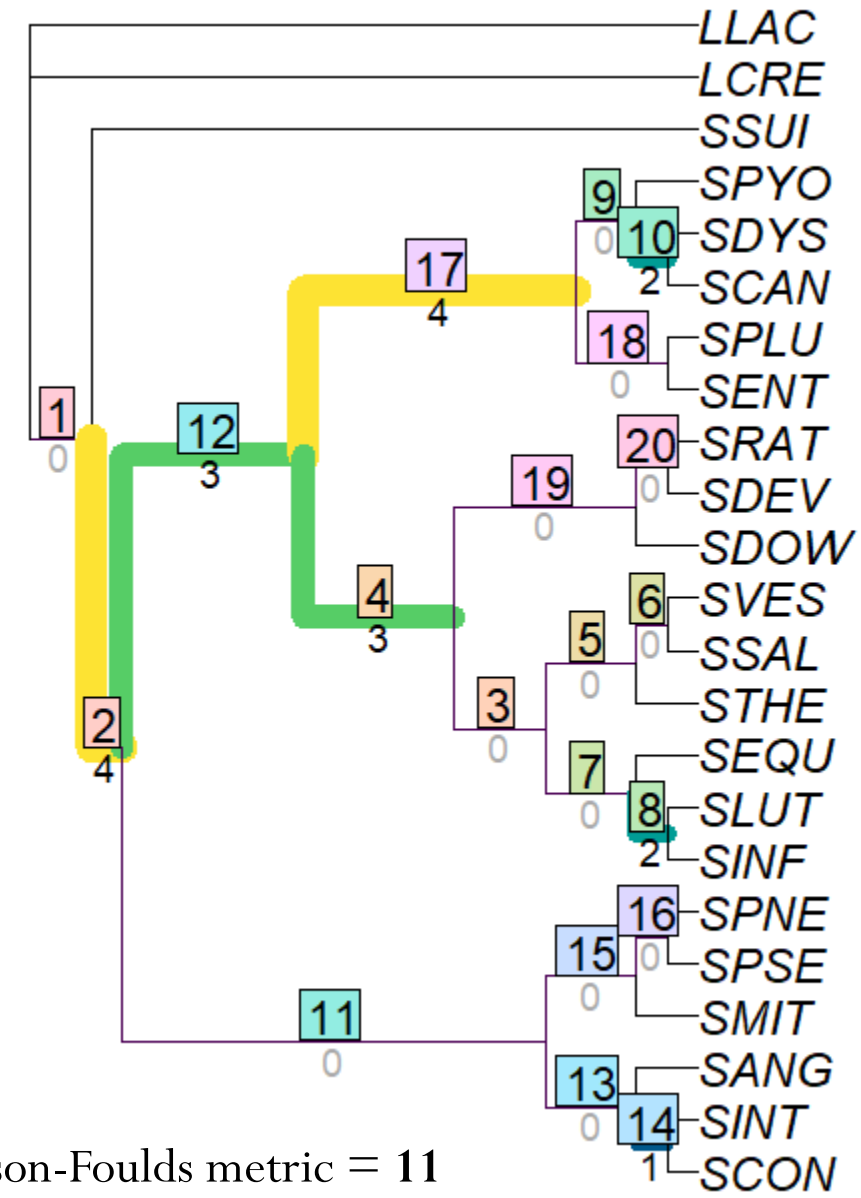
Supertree from bootstrap





Species tree (left) and supertree made with gene trees without paralogs verified with bootstrap (right)

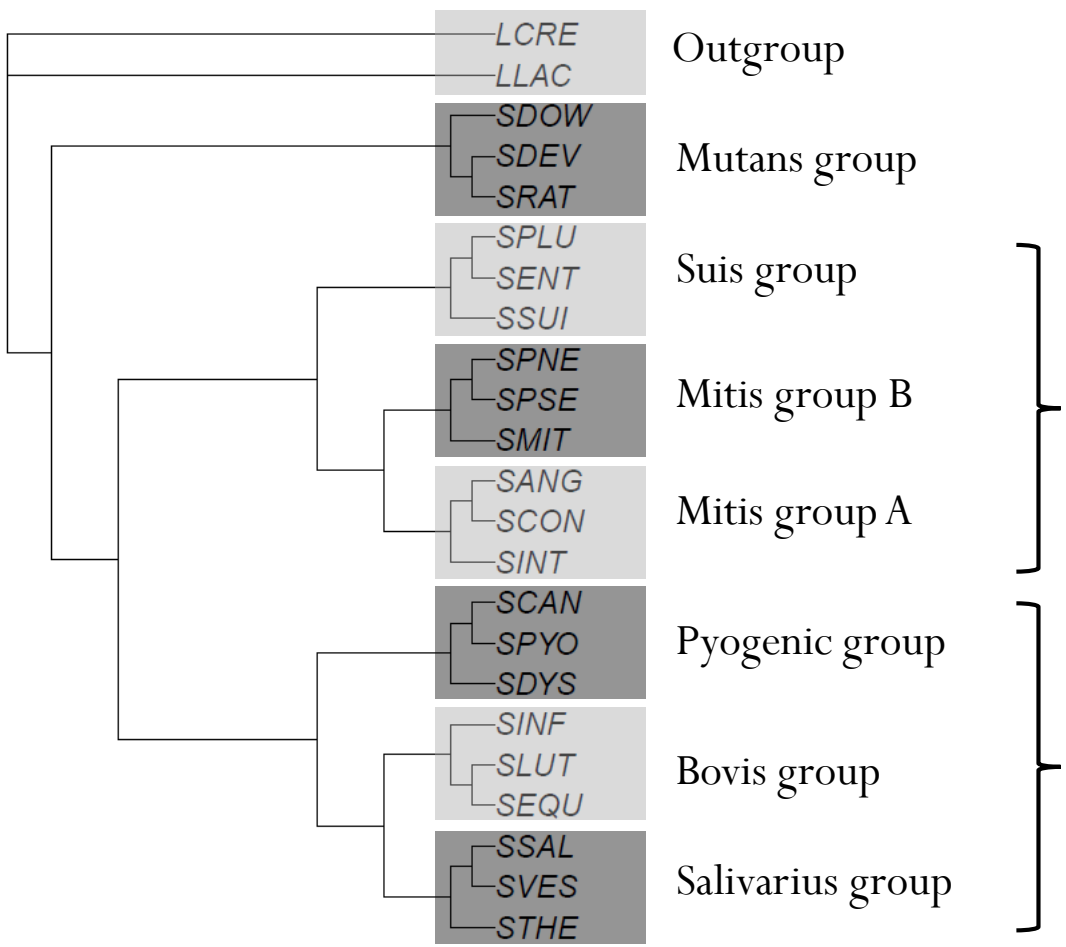


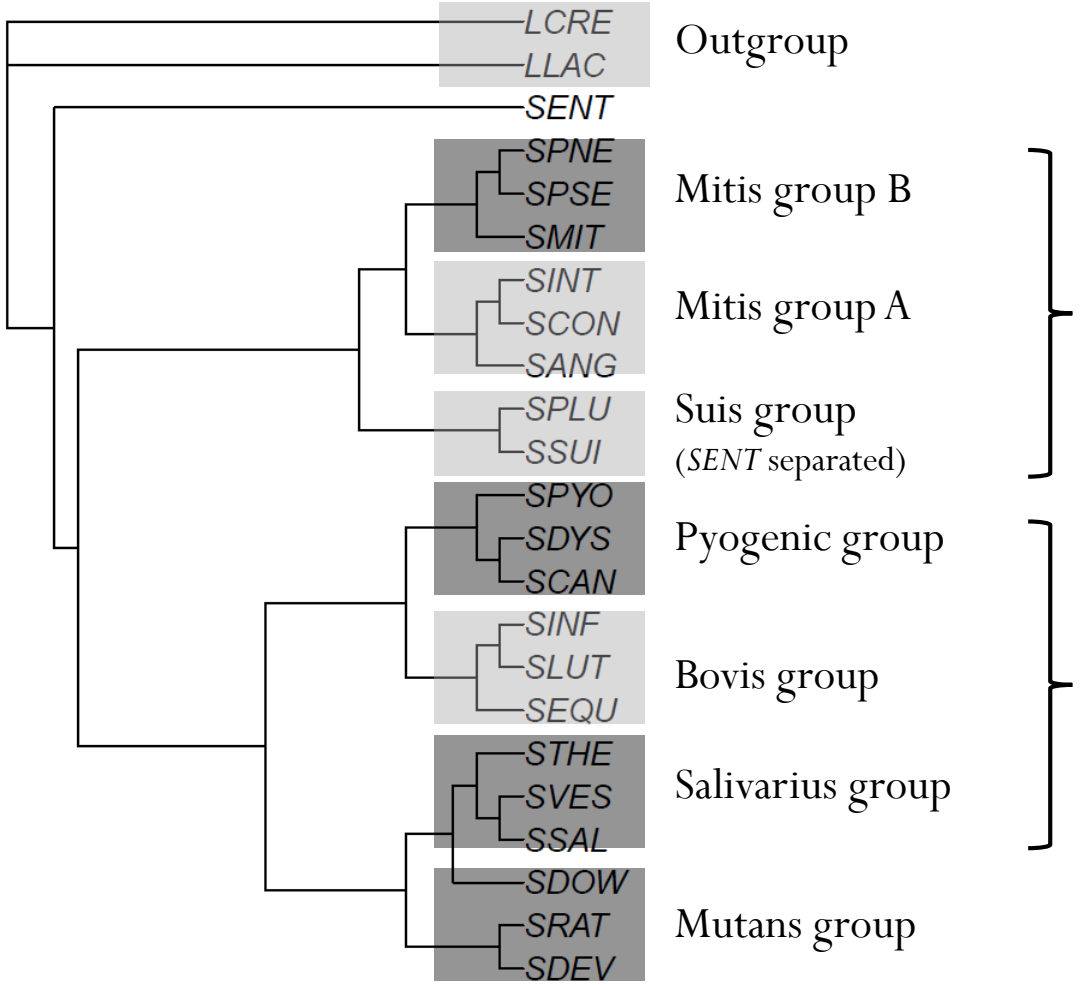


Robinson-Foulds metric = **11**

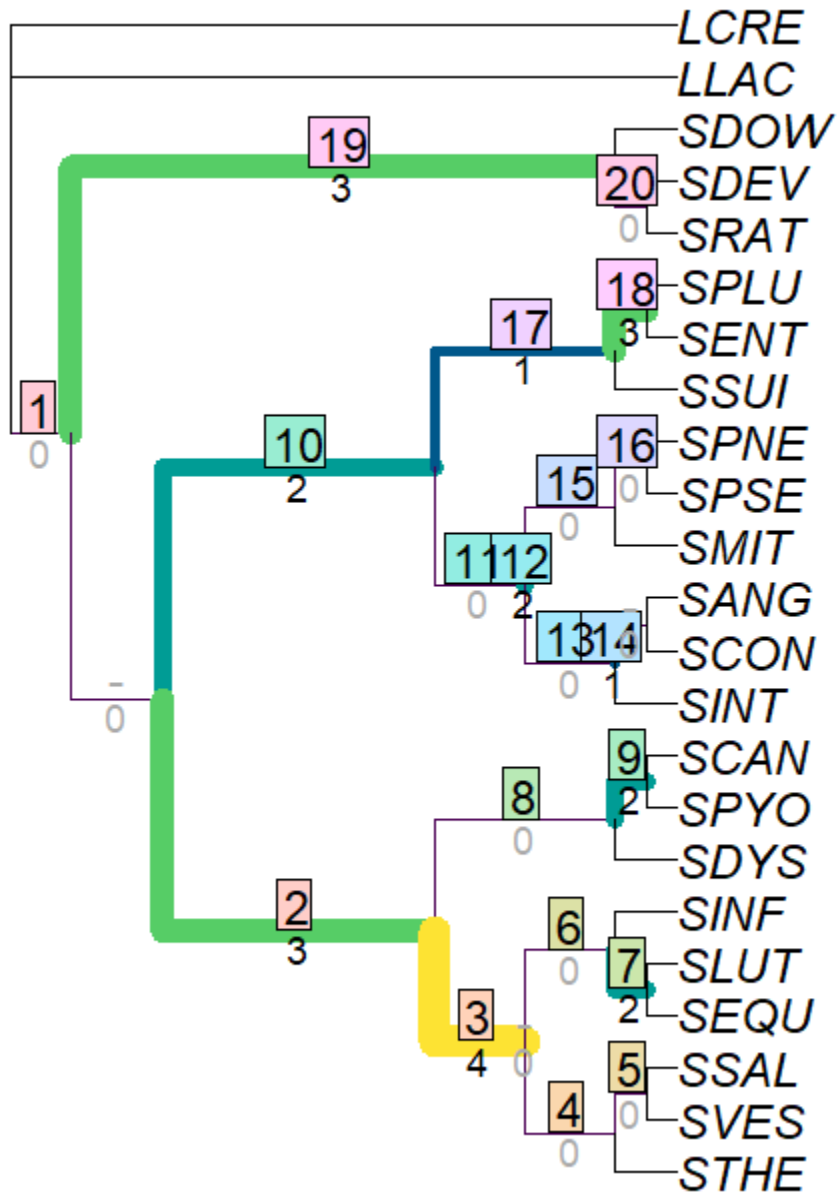
Species tree (left) and supertree made with bootstrap verified trees, DL=22114 (right) both rooted on the outgroup, with marked **unique** splits for both trees

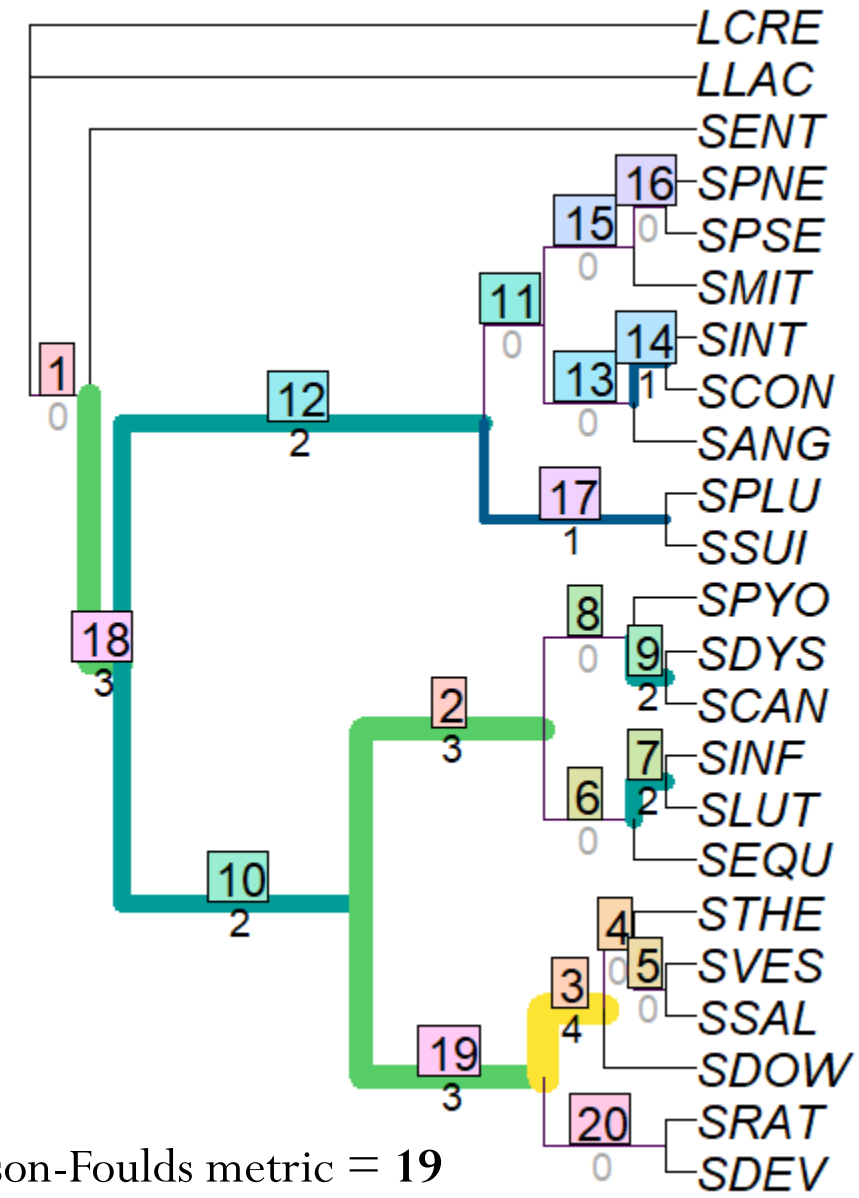
Supertree with paralogs





Species tree (left) and supertree made with gene trees with paralogs (right)





Robinson-Foulds metric = **19**

Species tree (left) and supertree made with paralogs, DL=53205 (right) both rooted on the outgroup, with marked **unique** splits for both trees

Bibliography

- F. Póntigo, M. Moraga, S.V. Flores, Molecular phylogeny and a taxonomic proposal for the genus *Streptococcus,* Genetics and Molecular Research 14 (3): 10905-10918 (2015)
- Facklam R. What happened to the streptococci: overview of taxonomic and nomenclature changes. *Clin Microbiol Rev*. 2002;15(4):613-630. doi:10.1128/CMR.15.4.613-630.2002