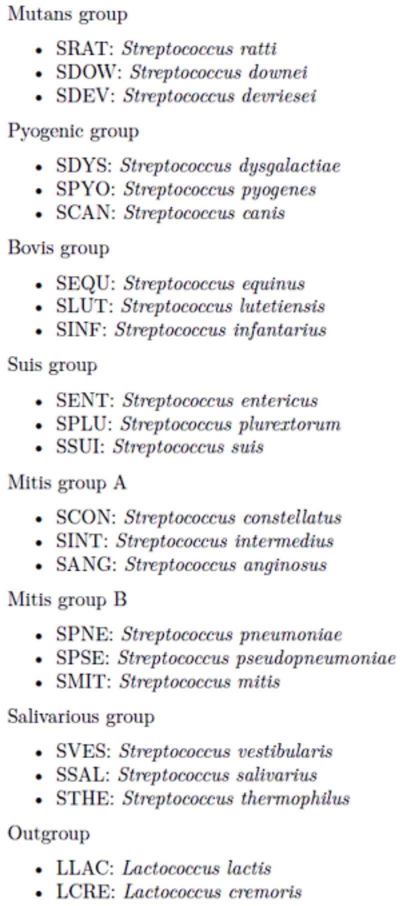
Phylogenetic Pipeline – Outline of Methods and Results

Comparative Genomics Project by Younginn Park

Species Tree

3 species per major groups from the bacterial genus *Streptococcus* were selected from the reference article[1] with species from genus *Lactococcus* that was used as an outgroup for this project.



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*

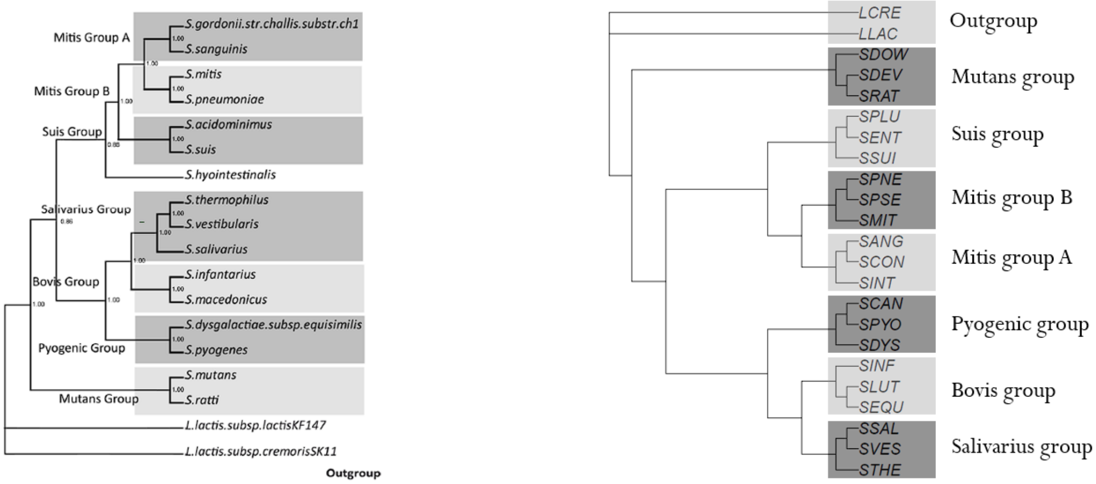
The process of recreating a species tree involves the selection of representative genes, such as *sodA, tuf, rpoB, recN, dnaJ, gyrB, and rnpB*, which are functionally significant and exhibit a balance between conservation and variation across species. These genes serve as the basis for constructing the species tree using a majority consensus approach through DendroPy[2], relying on Maximum Likelihood trees generated by PhyML[3]. To

[1] 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)

[2] https://dendropy.org/

[3] http://www.atgc-montpellier.fr/phyml/binaries.php

refine the analysis, amino acid evolution models are chosen based on the Akaike Information Criterion (AIC) using ProtTest3, with models like WAG and RtREV being selected to capture the evolutionary dynamics of the chosen genes.



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.

The reconstruction of the species tree from the article was successful with visible major clades – first one containing bacteria from the group *Mitis* and *Suis,* and the second one having *Pyogenic, Bovis* and *Salivarius* to share a common ancestor. These clades will serve as a guide for evaluating genome trees obtained in the following methods.

Proteomic sequences were fetched from the NCBI Genome database using the Command Line tool *datasets*[4].

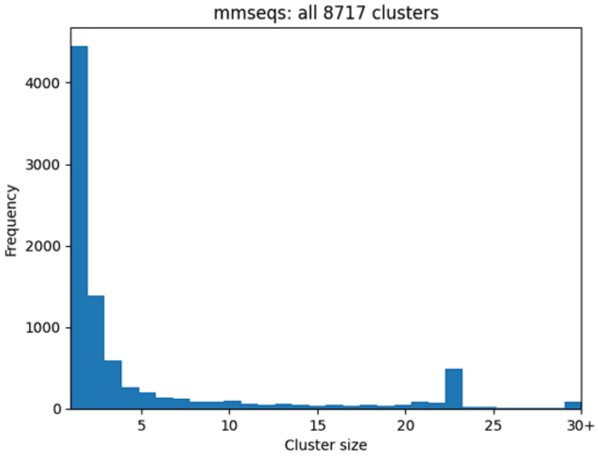
Sequence Clustering

Clustering

Clustering was done using MMseqs[5] with a 30% identity threshold, sequences exhibiting more than 30% similarity were clustered together (a rule of thumb for sequence homology). 8717 clusters were formed.

[4] https://www.ncbi.nlm.nih.gov/datasets/docs/v2/download-and-install/

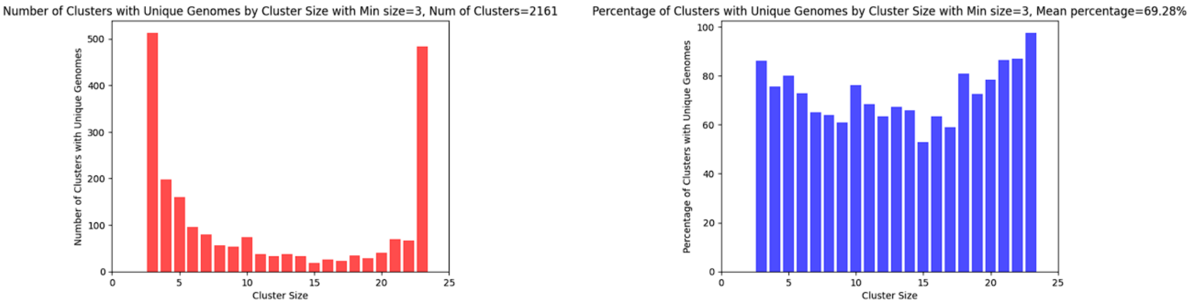
[5] https://anaconda.org/bioconda/mmseqs2



Histograms of cluster sizes

Non-orthologous Clusters (1-1)

Non-orthologous cluster problem was solved by keeping clusters between size 3 and 23 (number of taxa) that had unique genome identifiers, making a 1-1 gene to taxa sets.



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

Multiple Sequence Alignment and Neighbor-Joining

In this project, multiple sequence alignment (MSA) was performed using MAFFT[6] to identify similarities among sequences. The MSA output was then utilized for phylogenetic analysis through the Neighbor-Joining (NJ) method implemented in Biopython[7], employing the BLOSUM62 scoring matrix.

[6] https://mafft.cbrc.jp/alignment/software/windows_without_cygwin.html

[7] https://biopython.org/docs/dev/api/Bio.Phylo.TreeConstruction.html

50 trees with negative branches were discarded from the set of trees without paralogs and 78 trees from the set of trees with paralogs.

Split Support Analysis using Bootstrapping

In this project, bootstrap replicate trees were generated using the Neighbor-Joining method with 100 replicates in Biopython, aiming to assess the robustness of gene trees. The generated trees were subjected to split support analysis using SumTrees, a tool bundled with DendroPy. A criterion for trees with 'good' support was set at an average split support greater or equal to 0.9, which gave room for few individual splits in a tree to have a lower support without the need to discard the whole tree.

A total of 368 out of 2161 trees were excluded due to low branch support, leaving 206 out of 461 trees with a full set of taxa for further analysis, such as majority consensus tree and supertree construction. This stringent approach ensures the inclusion of only robustly supported trees in downstream analyses, contributing to the reliability of the final phylogenetic inferences.

Tree counts for different gene tree sets for each of the following tasks

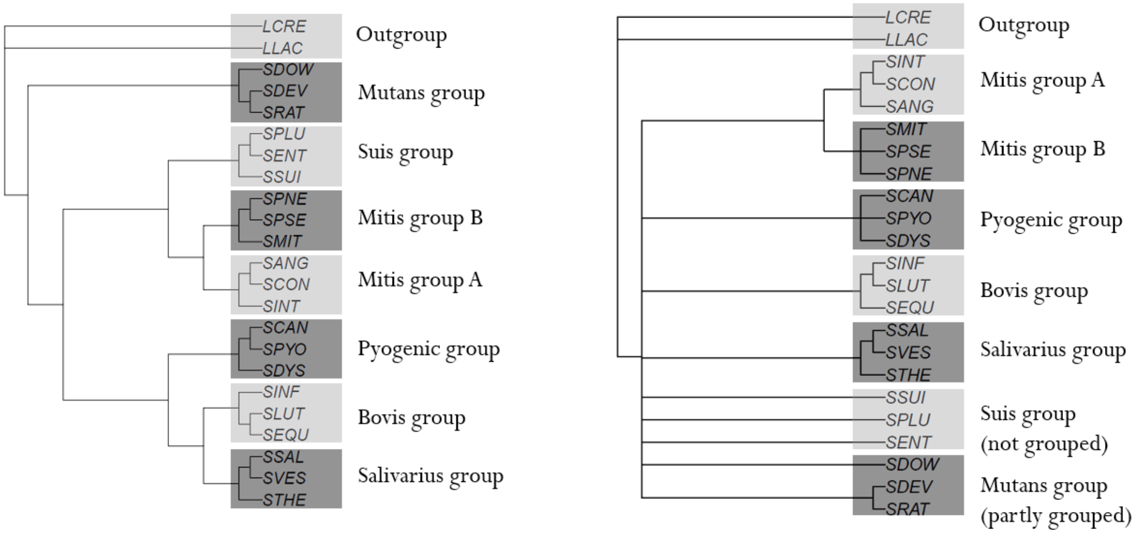
Description of the tree set	*Number of gene trees*
No paralogs, trees size 3 to 23 (for supertree)	2111
No paralogs, only size 23 i.e. 1-1 with taxa set (for consensus tree)[8]	461
No paralogs, trees filtered with bootstrap, size 3 to 23 (for supertree)	1743
No paralogs, filtered with bootstrap, 1-1 with taxa set (for consensus tree)	206
With paralogs, size 3 to 23 (for supertree)	4187

Genome Trees

Majority Consensus Trees

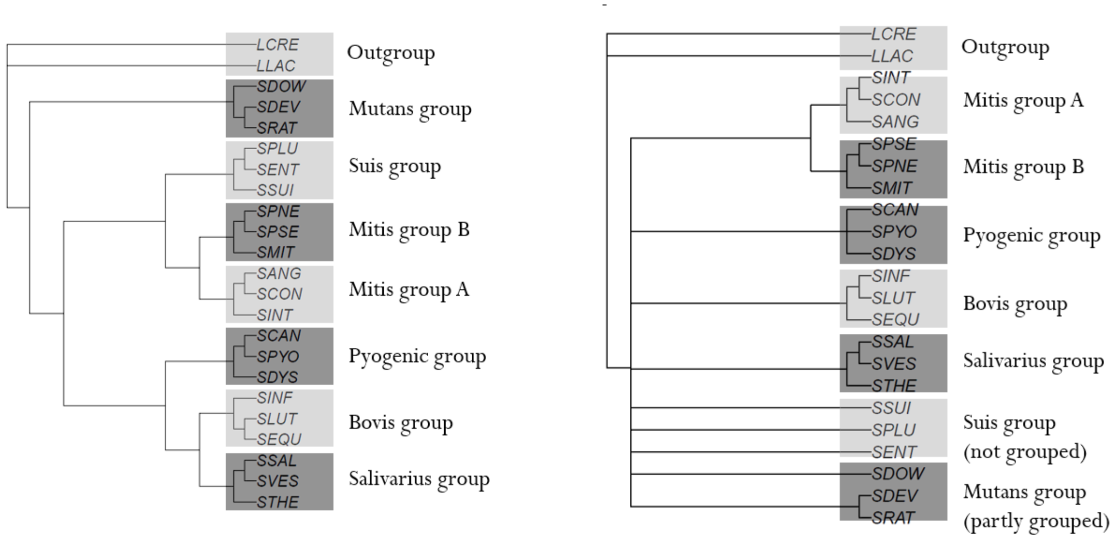
For majority consensus tree inference, a total of 461 gene trees, each with the full set of taxa, were utilized. The implementation involved employing the DendroPy python library, specifically using a 0.5 cutoff (inclusive frequency threshold) for unrooted trees, considering splits in the consensus tree. This process ensured that only splits with a frequency equal to or higher than 0.5 were incorporated. The resulting unrooted consensus tree was then rooted with the specified outgroup.

[8] This set of trees was formed by subsetting the set for the supertree by tree size. Same was done for bootstrap verified trees



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

The majority consensus method failed to group major *Streptococcus* clades. However, it succcessfully grouped bacteria from the two *Mitis* groups together. *Pyogenic, Bovis* and *Salivarius* groups were also formed, however they didn't group into the major clade from the species tree.



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

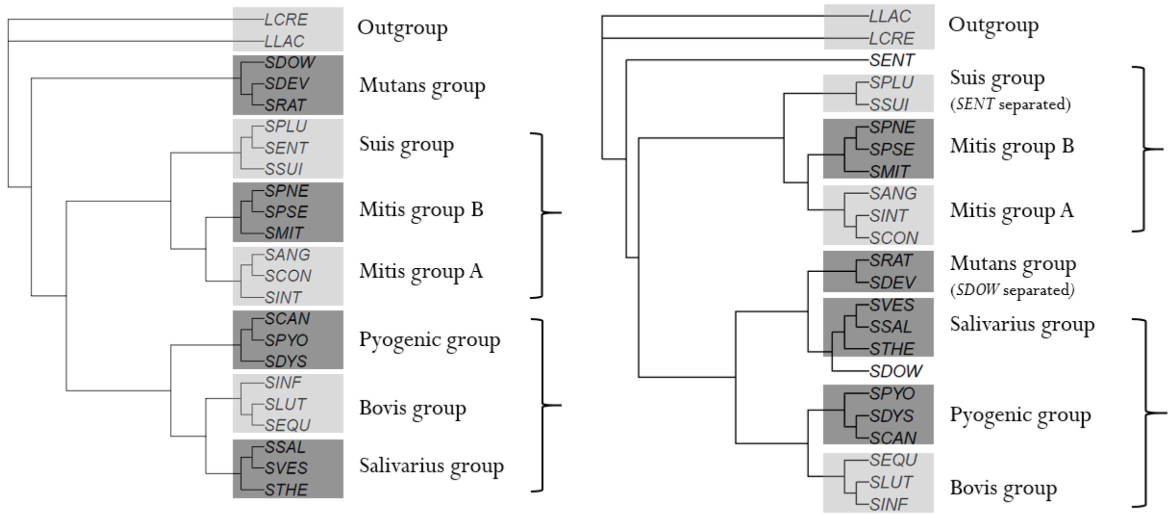
Majority consensus tree constructed using bootstrap verified trees was essentially the same as the one with all gene trees with the exception of the *Mitis group B* which this time resolved the trifurcation observable in the previous genome tree.

Both majority consensus trees formed multifurcations on the *Streptococcus* genus. Conducting an additional analysis using a greedy consensus tree construction with the same set of gene trees and checking split frequencies (supports) for individual branches would be most informative.

Supertrees

In the supertree inference phase, the fasturec[9] tool was employed. The heuristic tree search utilized the -Z option, involving both initial quasi-consensus tree construction and subsequent hill-climbing steps under the duplication-loss cost model. The hill-climbing steps encompassed a mixture of Nearest-Neighbor Interchanges (NNI), Subtree Prune and Regraft (SPR), and Tree Swap (TSW) operations.

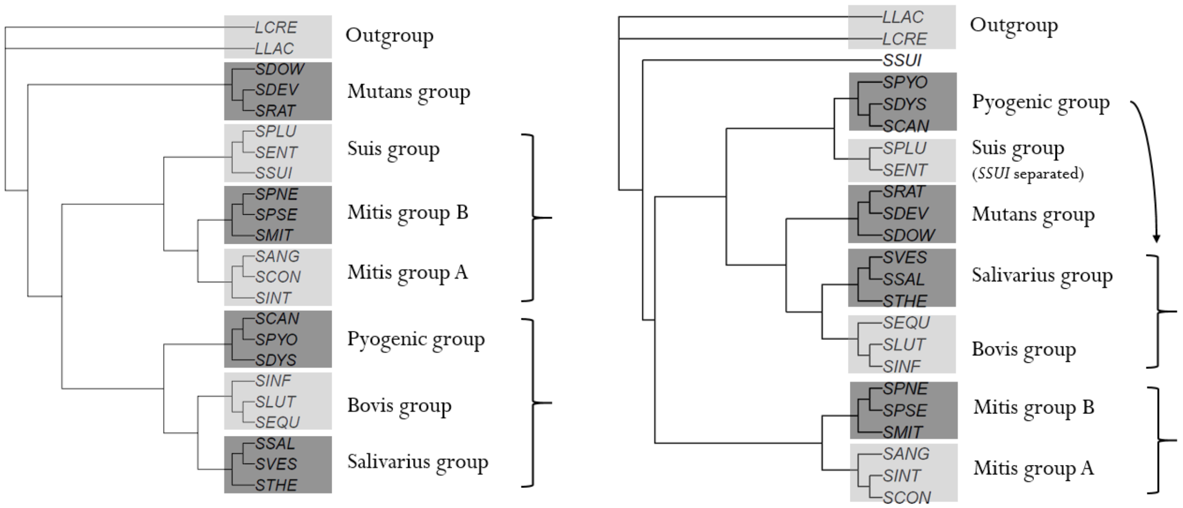
The analysis was performed both with and without considering paralogs to assess their impact on the supertree topology. The resulting supertree was rooted with the specified outgroup, considering gene duplications and losses as well as potential paralogous relationships when such were present.



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

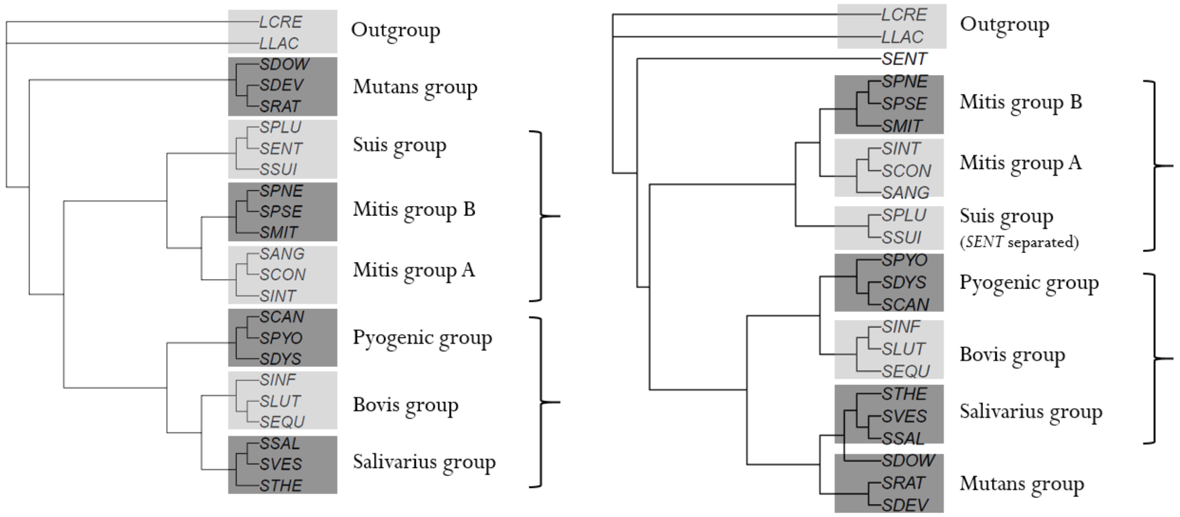
The tree with the best duplication-loss cost was chosen with a value of DL=33899. The supertree formed a binary genome tree, correctly recognizing two major clades – one containing *Mitis groups* and the *Suis group*, and the second one with *Salivarius, Pyogenic* and *Bovis groups* with the *Mutans group* having been joined into this clade. The Robinson Foulds metric on both trees gave a value 19, with most unique splits being the result of the *Mutans group* moving to one of the major clade.

[9] https://bitbucket.org/pgor17/fasturec/src/master/



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

The tree with the best duplication-loss cost was chosen with a value of DL=22114. The supertree constructed using bootstrap verified gene trees was successful in recognizing the *Salivarius-Bovis-Pyogenic* clade with the addition *Mutans group* and *Suis group*, although *Streptococcus Suis*



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

The tree with the best duplication-loss cost was chosen with a value of DL=53205. The supertree method with paralogs returned a tree identical to the one made without paralogs and without bootstrap verification.