**Inferring phylogenetic relationships between species of the Staphylococcus genus using single locus and multi-locus data set approaches**

**Introduction**

The genus *Staphylococcus* contains more than 60 species and subspecies. Many of these species lead to high levels of infection among human populations. Others are responsible for agricultural losses within the dairy, swine and poultry industries. Thus, the genus *Staphylococcus* is of interest to both the health and agricultural economic sectors. Since multiple species within this genus are common pathogens in non-human animals it becomes necessary that they be monitored with concern as these animals provide reservoirs for pathogenic bacteria. Host switching is an important mechanism in the evolution of *Staphylococcus*. For example, in *S. aureus*, human-to-poultry and bovine-to-human host switches have been observed. As such, a thorough understanding of species relatedness is a necessity for understanding host-pathogen interactions within this genus. The present study is aimed at determining evolutionary relationships between *Staphylococcus* species. Traditionally 16s rRNA sequences have been used for phylogenetic reconstruction but they have led to poor resolution of phylogenies based on this single gene dataset. This approach will include using a multi-locus dataset to get better hierarchical and evolutionary relationships between species.

**Objectives**

The broader objective of the study is to infer phylogenetic relationships between known species of the *Staphylococcus* genus using 16S rRNA gene sequence data. However, recent reports have revealed that 16S rRNA gene sequences alone are not enough to obtain a superior resolution of evolutionary relationships between different species. This study therefore also aims to analyze the role of multi-locus datasets (16SrRNA, rpoB and dnaJ) to thoroughly explore the phylogenetic signal and provide robust confirmatory evidence for the relationships among *Staphylococcus* species. The analysis will involve multiple sequence alignments of the above-mentioned sequences followed by re-construction of individual phylogenetic trees using this data. This objective will be met by phylogenetic reconstruction using neighbor joining and maximum likelihood methods using PAUP and RaxML.

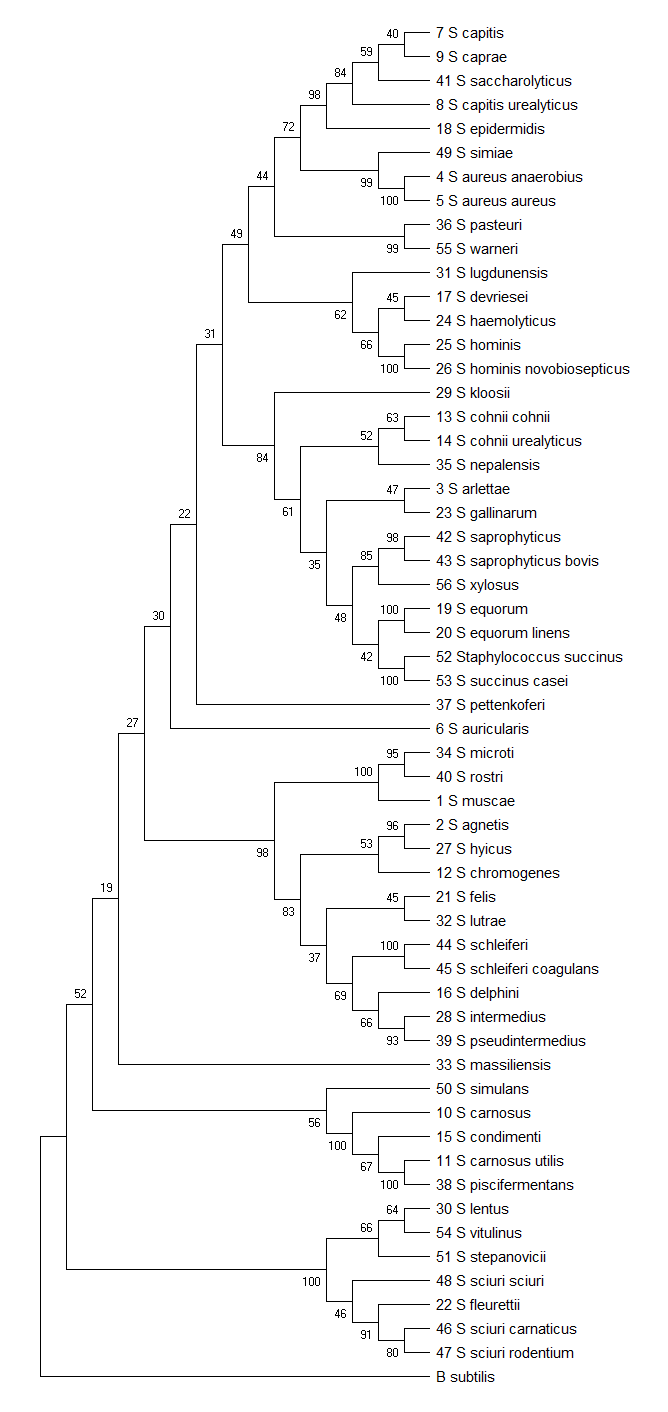
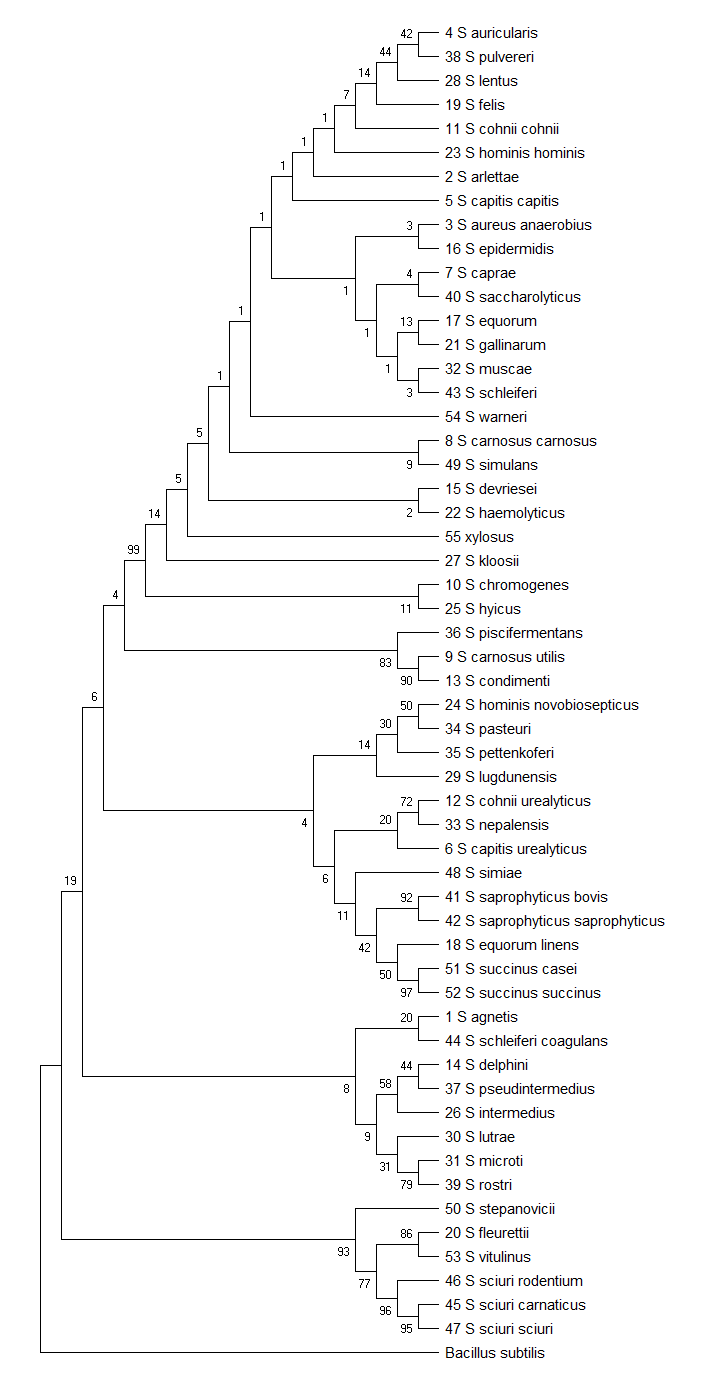
**Methods**

DNA sequences for a total of three genes from 56 staphylococcal species, and one outgroup species (*Bacillus subtilis*) were downloaded from NCBI GenBank. The three loci information included the non-coding 16S rRNA gene sequences, and two protein coding genes: dnaJ, rpoB. The multi-locus dataset was created by concatenating these sequences using SeqNinja in DNAstar. The single locus and the concatenated multi-locus dataset were uploaded on Github. The data was pulled using the command ‘git pull’ using command line. The sequences were aligned using the MAFFT module on High performance computing facility at Iowa State University. The alignments were saved in the default FASTA format. The Fasta alignment files were used for ML tree construction in Raxml-NG

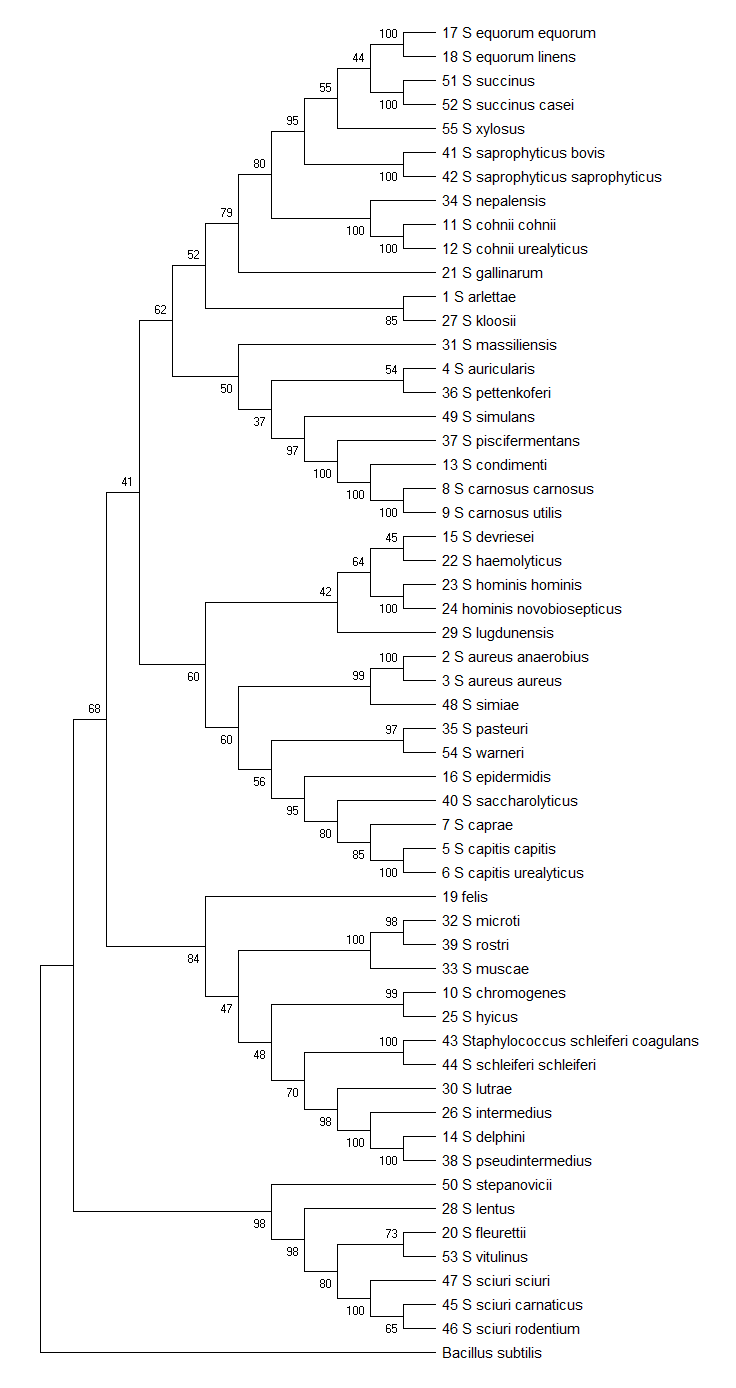
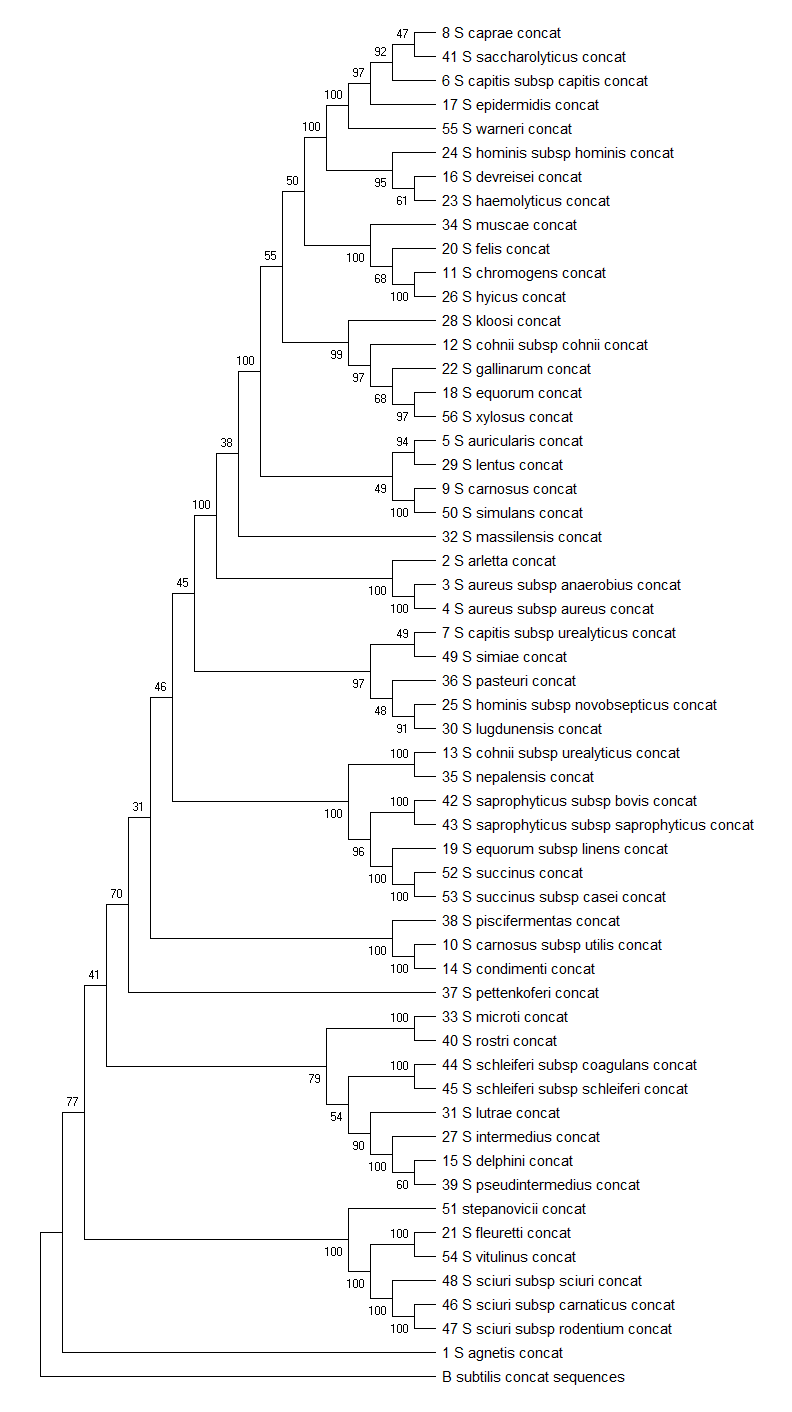
Module PAUP\* was used to construct phylogenetic trees for the single and multi-locus datasets using distance measures. The fasta alignment files were converted to NEXUS format by importing them in PAUP. Bacillus subtilis single gene and concatenated sequences were designated as the outgroup for each of the datasets. The GTR + gamma model was used as nucleotide substitution model and minimum evolution was set as the objective for the analysis. A distance matrix was generated. Neighbor joining trees with branch lengths were generated and saved. Bootstrap analysis with 100 replicates was carried out and 50 % majority rule trees were constructed

Phylogenetic tree construction under the maximum likelihood criterion was carried out using RAxML-NG module on HPC-class. Alignments for the single gene and multi locus dataset were used for the analysis. Model: GTR+FO+G4m was used as the nucleotide substitution model. An ML tree search with 20 distinct starting trees (10 random and 10 parsimony) and a bootstrap analysis (Felsenstein Bootstrap + Transfer Bootstrap) of 50 replicates was carried out for these datasets. The final log likelihood values and AIC and BIC scores were noted and evaluated for each of these datasets.

**Results & Discussion**



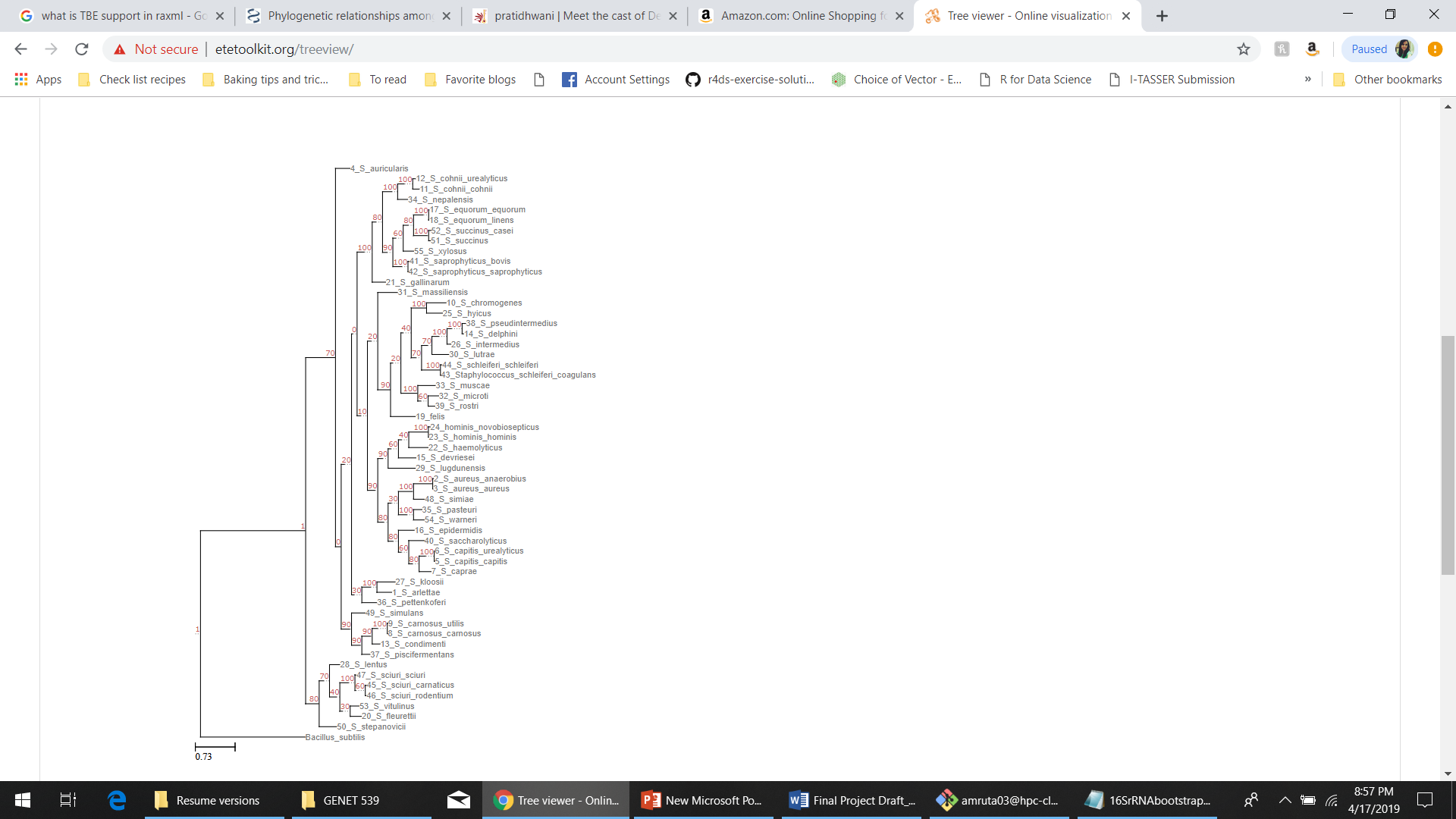
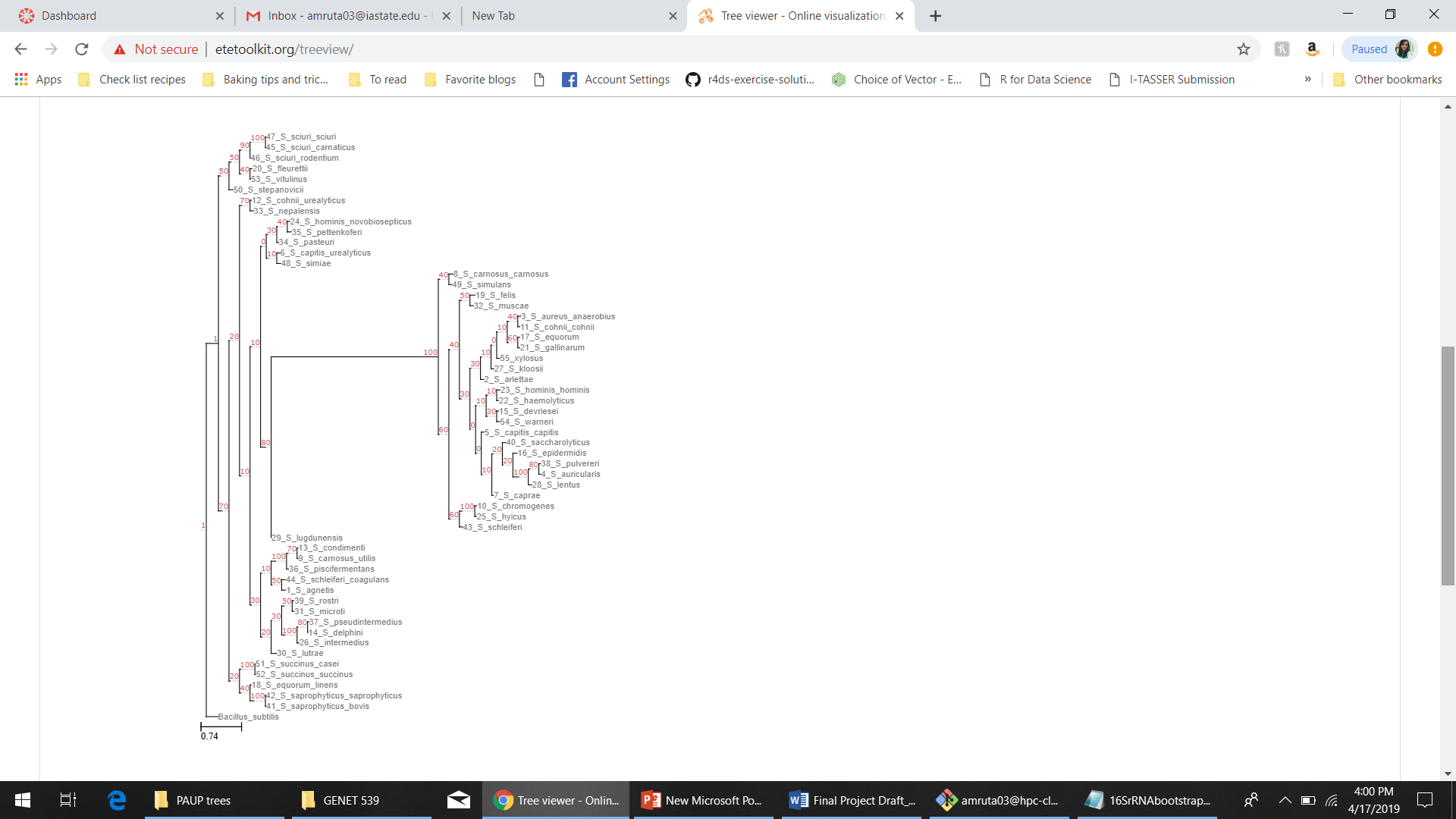
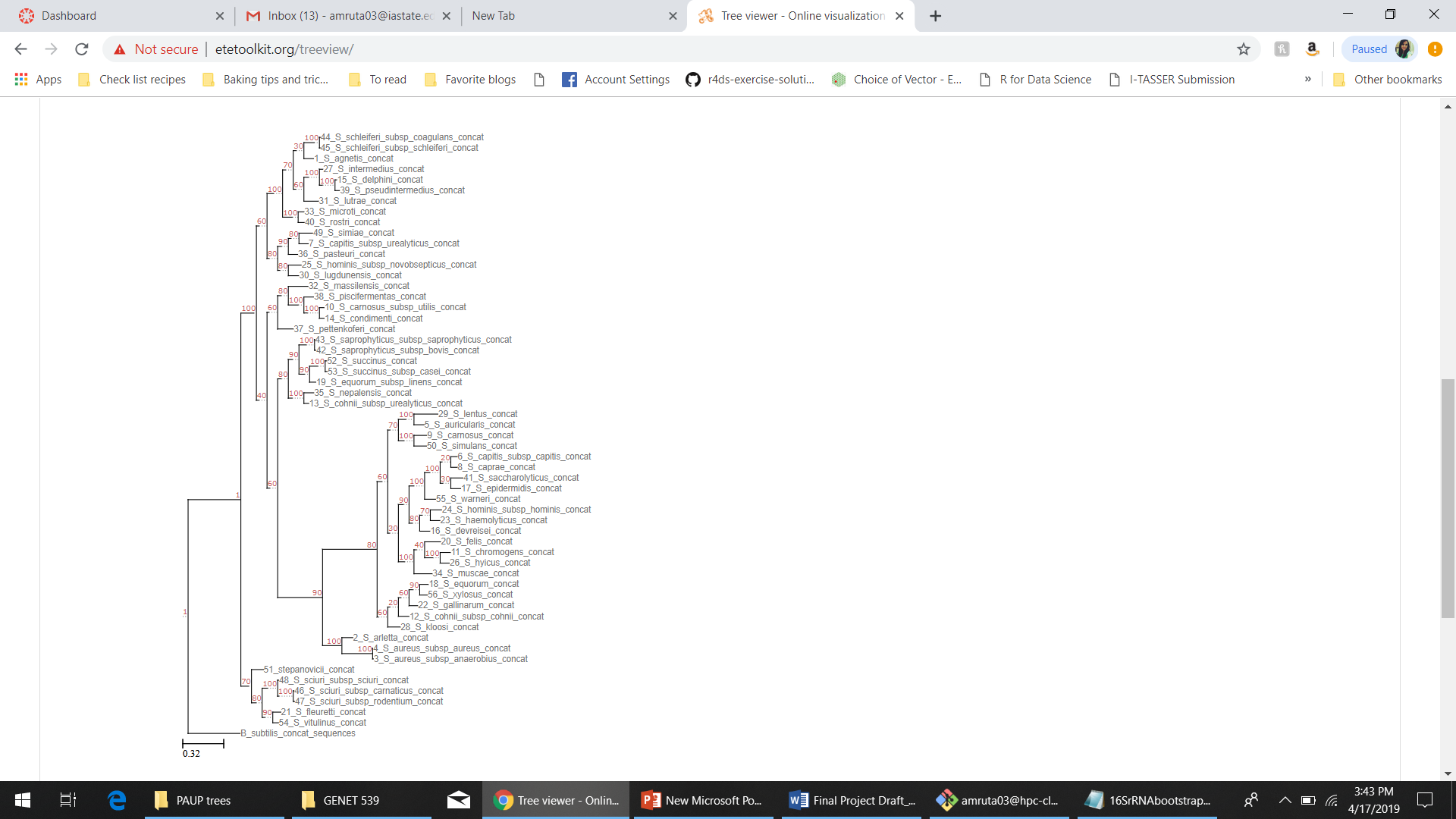
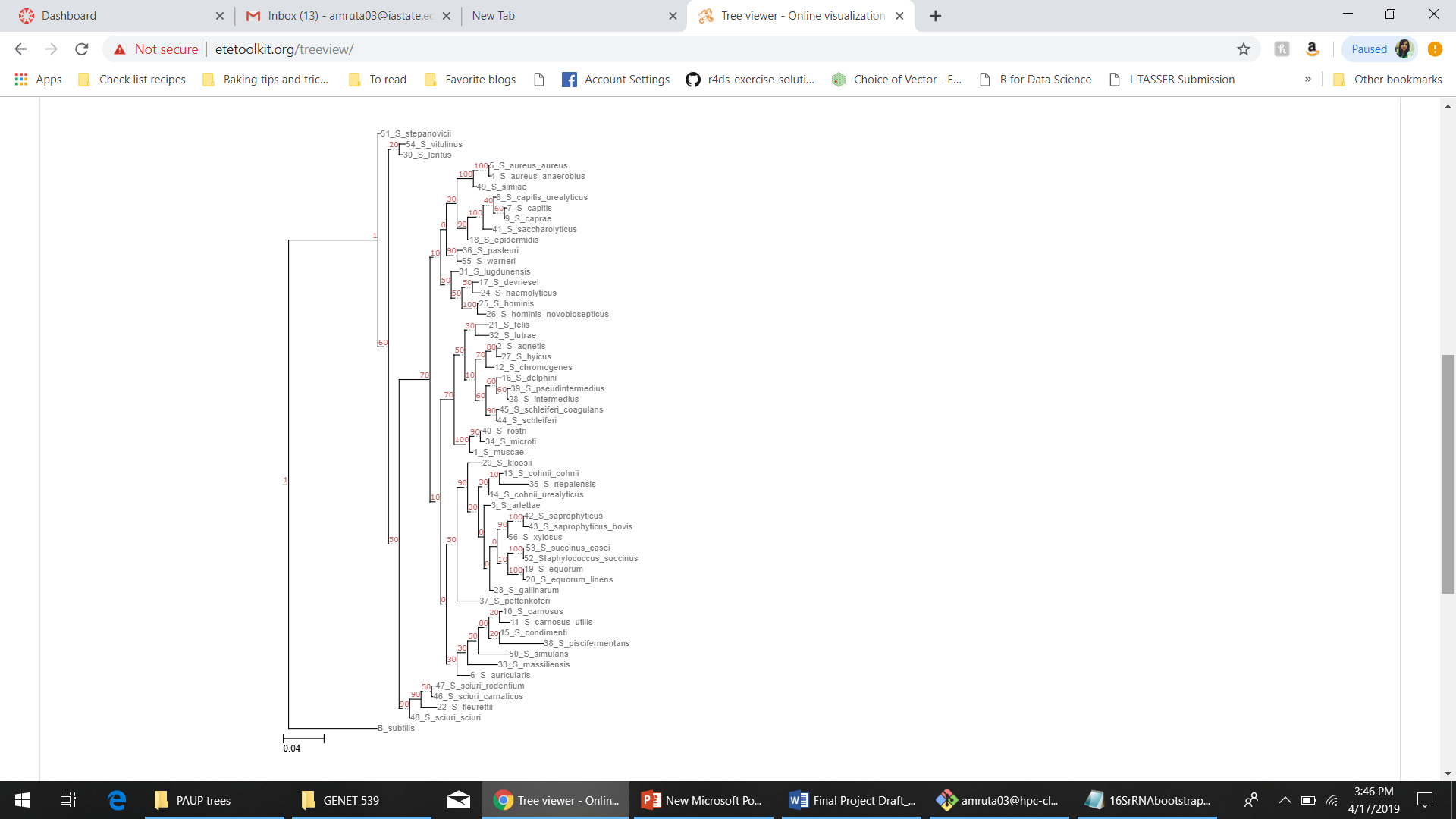
**Phylogenetic trees constructing using Neighbor joining method in PAUP\*. A. 16 SrRNA, B. rpoB,**



**Phylogenetic trees constructing using Neighbor joining method in PAUP\*. C. DnaJ and D. Multilocus dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Log Likelihood | AIC score | BIC score |
| Concatenated dataset | -52286.264 | 104816.38 | 105645.625 |
| 16SrRNA | -6061.650 | 12383.36 | 13006.20 |
| dnaJ | -18107.32 | 36484.68 | 37022.96 |
| rpoB | -21139.621 | 42521.155 | 43281.13 |

**Phylogenetic trees constructing using Maximum likelihood method in RaxML- NG. A. 16 SrRNA, B rpoB, C. DnaJ and D. Multilocus dataset**



The staphylococcus genus is an economically important class of microorganisms since they have both human and animals as hosts. The groupings of the species need revision and hence multi-locus dataset has been used in this study. The neighbor joining and maximum likelihood analysis shows some consensus between higher order relationships among the species. There is overall good support for nodes of higher-level relationships for both neighbor joining and maximum likelihood trees. Evaluation of the ML trees suggests that the concatenated dataset yields better resolution that the single locus datasets as suggested by their log likelihood and AIC scores. (Table 1). Historically Staphylococcus species have been grouped between 4 to 10 different groups. However this analysis which involves the use of a multilocus data has led to resolution to phylogenies to upto 15 different clades. These 15 clades/clusters are also in agreement with the phenotypic properties of the Staphylococcus species. For e.g. the oxidase positive S. sciuri cluster (S. sciuri, S. stepanovicii, S. vitulinus, S. lentus and S. fleurette ) are within the same cluster and are a sister clade to the rest of the oxidative negative species.

**Conclusion**

Previously unreported relationships between Staphylococcus species have been resolved in the present study. This indicates that use of a multi-locus dataset over single loci may be the method of choice for future studies of phylogeny. A comparison of neighbor joining and maximum likelihood analysis of the data revealed that both methods an overall robust inference but maximum likelihood methods work better with concatenated datasets.