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

Amruta Bapat [1]

[1]*Department of Agronomy, Iowa State University, Ames, IA USA 50014*

Correspondence to Amruta Bapat : amruta03@iastate.edu

Abstract

***Staphylococcus* phylogeny studied previously have not included all identified taxa and have mostly relied on 16SrRNA sequences for phylogenetic reconstruction. This paper explores the possibility of using two protein coding genes – dnaJ and rpoB in addition to 16SrRNA for phylogenetic inference. The paper also assesses the efficacy of using a multi-locus dataset in further refining the clusters of Staphylococcus taxa. The phylogeny of ~ 56 taxa of the *Staphylococcus* genus was estimated using neighbor joining and maximum likelihood phylogenetic approaches. Regardless of the method used it was observed that the *Staphylococcus* phylogeny requires refinement. The use of dnaJ sequences resulted in identifying previously unreported relationships among *Staphylococcus* taxa. Although the use of a multi-locus dataset has been suggested to provide robust phylogenetic inference, the present study did not find an advantage of using such a dataset over single gene loci.**

Background

The genus *Staphylococcus* contains more than 55 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.[1] Thus, the genus *Staphylococcus* is of interest to both health and agricultural economic sectors. Since multiple species within this genus are common pathogens in 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* [2]. For example, in *S. aureus*, human-to-poultry and bovine-to-human host switches have been observed. The present study is aimed at

determining evolutionary relationships between *Staphylococcus* species. Traditionally 16s rRNA sequences have been used as the gold standard for phylogenetic reconstruction.[3] This gene is well conserved with some variations and is present in all bacteria. However, this gene has been found to have some limitations when closely related isolates are considered. Hence the use of other genes such as dnaJ and rpoB has been proposed in clinical microbiology to overcome limitations of the 16SrRNA sequences.[4][5]. The rpoB gene codes for the beta subunit of RNA polymerase and is sufficiently conserved to be used as a molecular clock. The dnaJ gene codes for a chaperone protein. This gene is also present in most bacteria and is well conserved. The present study evaluates the use of other sequence information besides 16SrRNA sequences as well as the use of a multi-locus data set to infer *Staphylococcus* phylogenetic relationships.

Objectives

The broader objective of the study is to refine phylogenetic relationships between all known species of the *Staphylococcus* genus. The clinical identification of isolates of *Staphylococcus* genus has been done using tests for novobiocin resistance/susceptibility , presence/absence of coagulase and presence/absence of oxidase activity [6]. In order to build refined relationships between the 56 known *Staphylococcus* taxa, the study also aims to analyze conserved genes which can be used to infer robust phylogenetic signals. In addition to this, the efficacy of a multi-locus dataset (concatenated 16SrRNA, rpoB and dnaJ sequences) will help 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-NG respectively. The two approaches – neighbor joining and maximum likelihood methods have been taken to meet the project objectives. While the neighbor joining uses a clustering algorithm that clusters taxa based on genetic distances, it is a fast method which can provide a good phylogenetic inference with the right set of parameters. Maximum likelihood on the other hand applies a model of sequence evolution and is well suited for sequence data but is computationally intensive. [7]

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 data (16SrRNA, DnaJ and rpoB sequences) and the concatenated multi-locus dataset were uploaded to Github. 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.

Module PAUP* was used to construct phylogenetic trees for the single and multi-locus datasets using distance measures. The fasta alignment files were imported to NEXUS format in PAUP*. *Bacillus subtilis* 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. The .tre files were visualized using FigTree v1.4.4.

Color	*Novobiocin*	*Coagulase*	*Oxidase*
Blue	Resistant	Negative	Positive
Green	Susceptible	Negative	Negative
Orange	Resistant	Negative	Negative
Purple	Susceptible	Positive	Negative
Red	Susceptible	Variable	Negative

Table 1: The different clades in the tree diagrams were highlighted using the following scheme

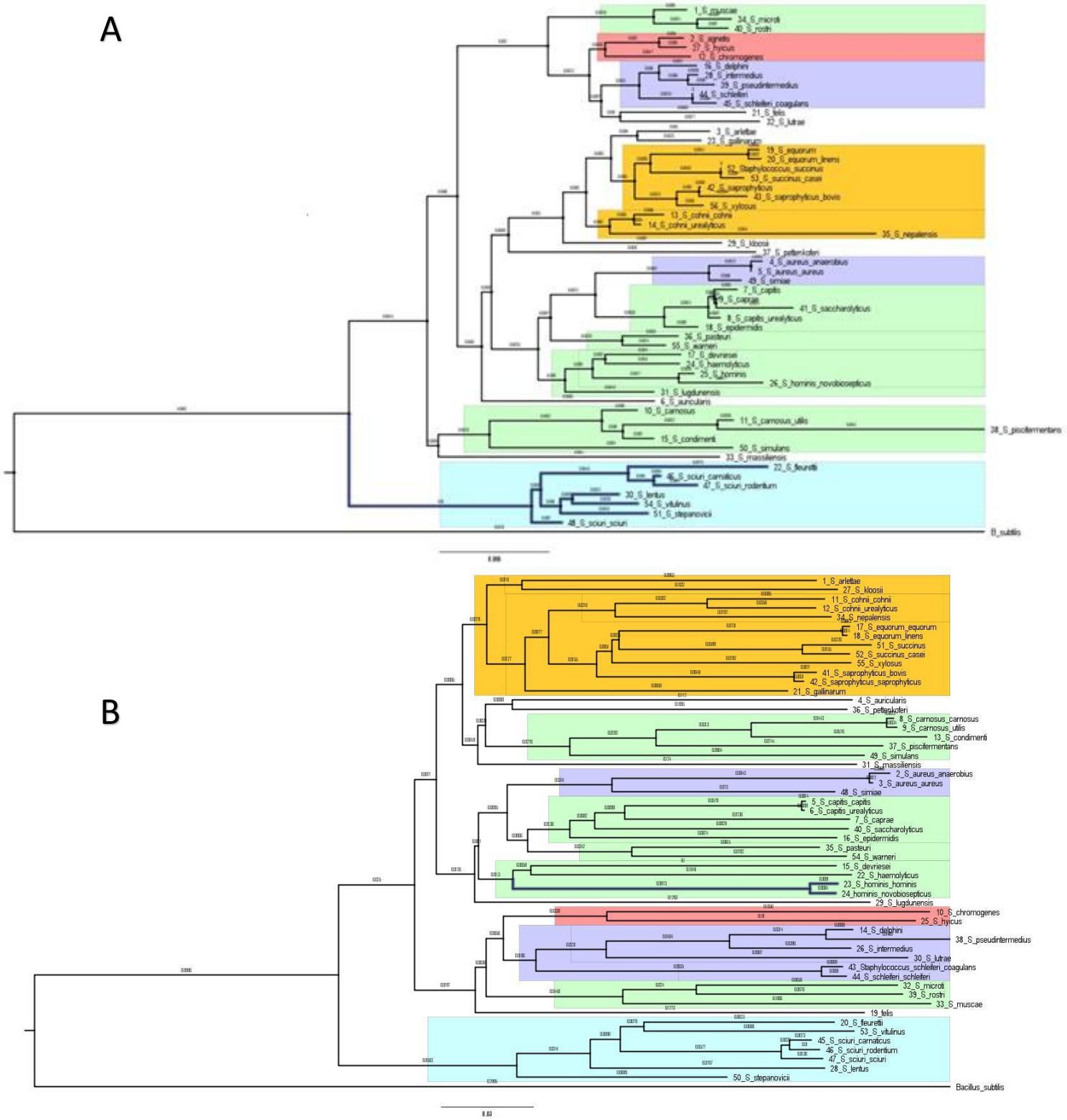
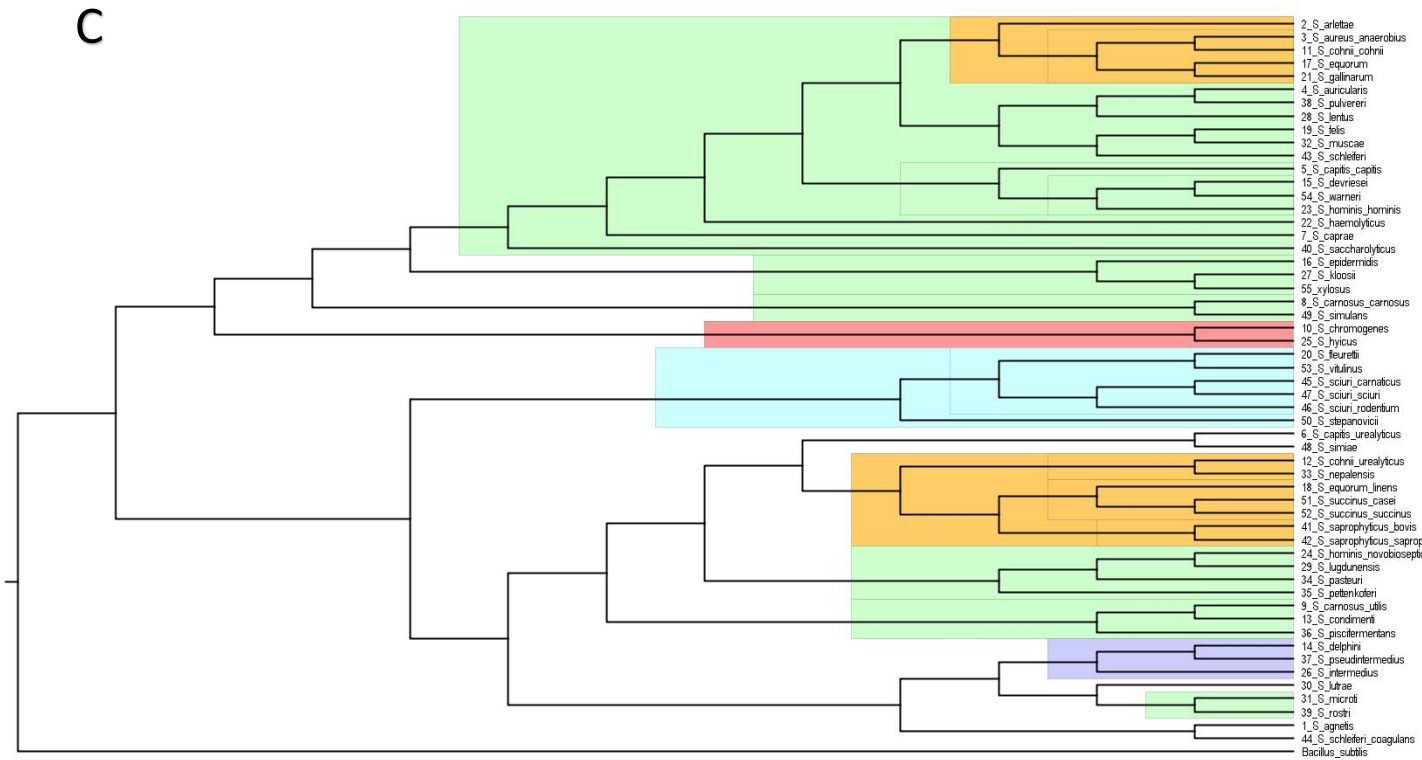


Fig 1A & 1B represent Neighbor joining trees for 16SrRNA and DnaJ single locus data



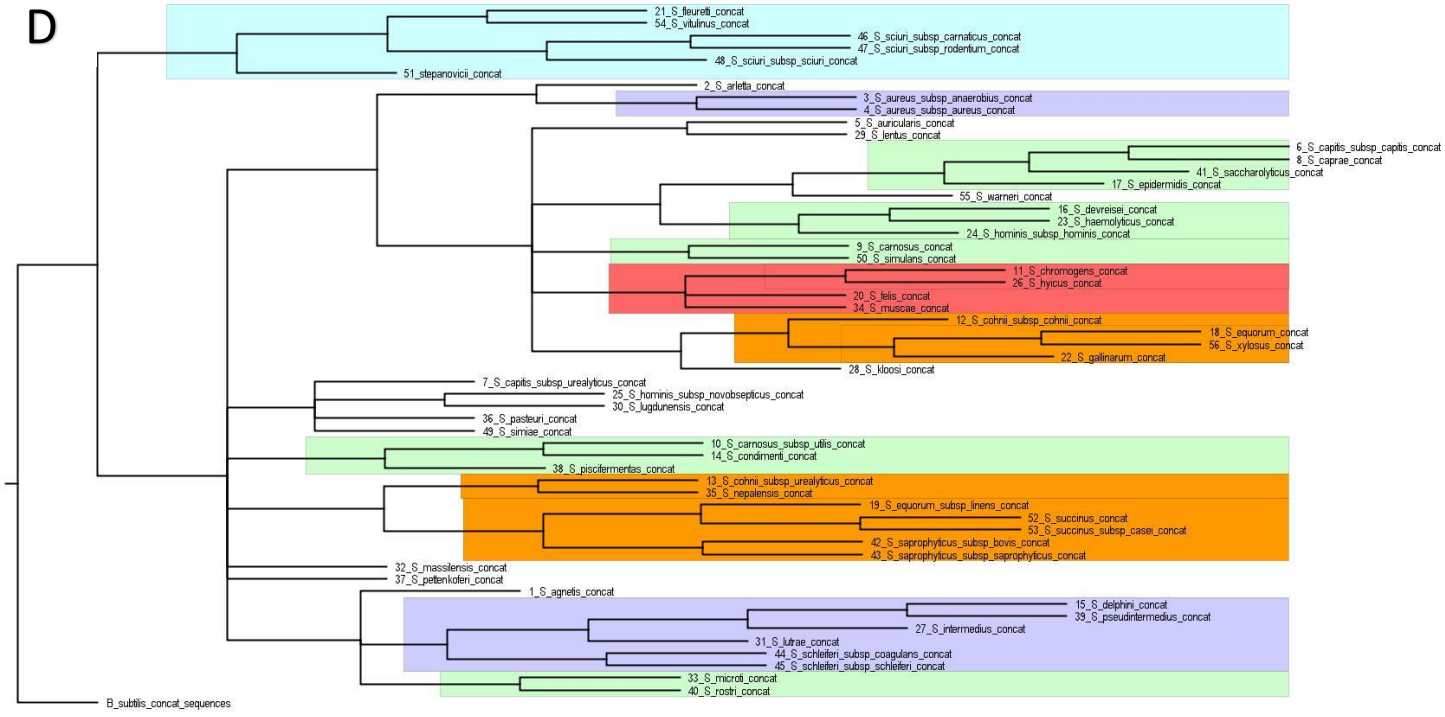


Fig 1C & 1D represent Neighbor joining trees for rpoB single and multilocus dataset

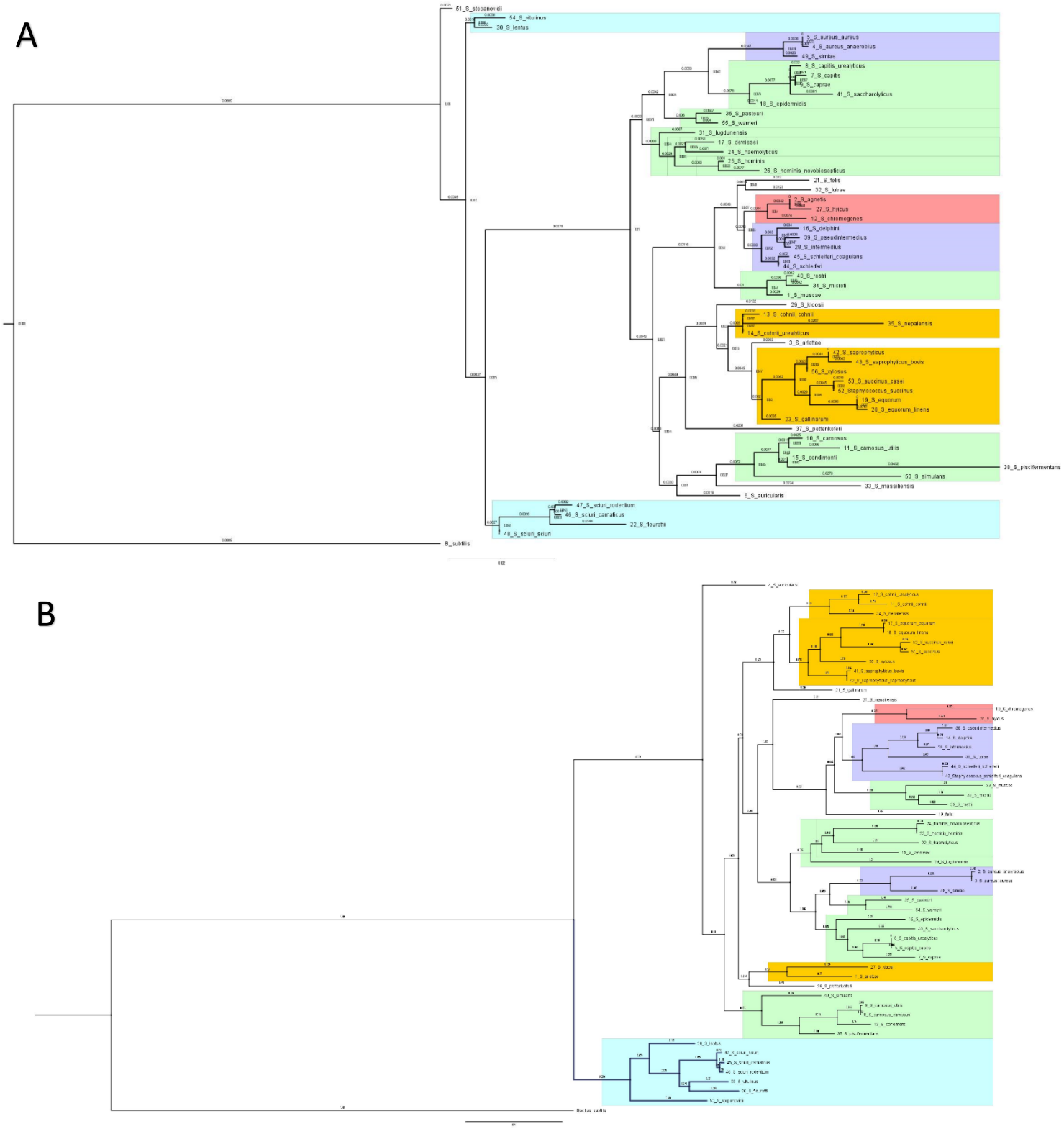
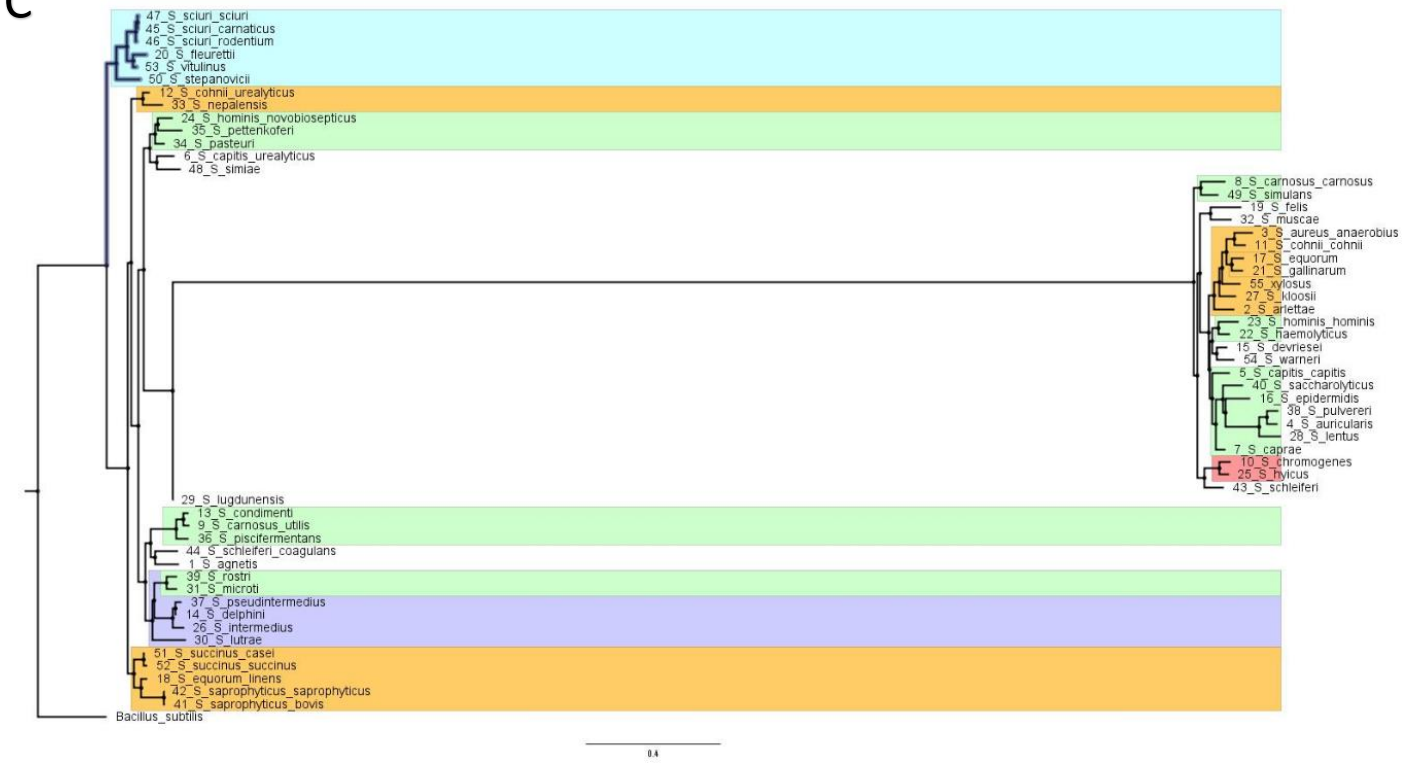


Fig2A & 2B represent Maximum likelihood trees for 16SrRNA and DnaJ single locus dataset

C



D

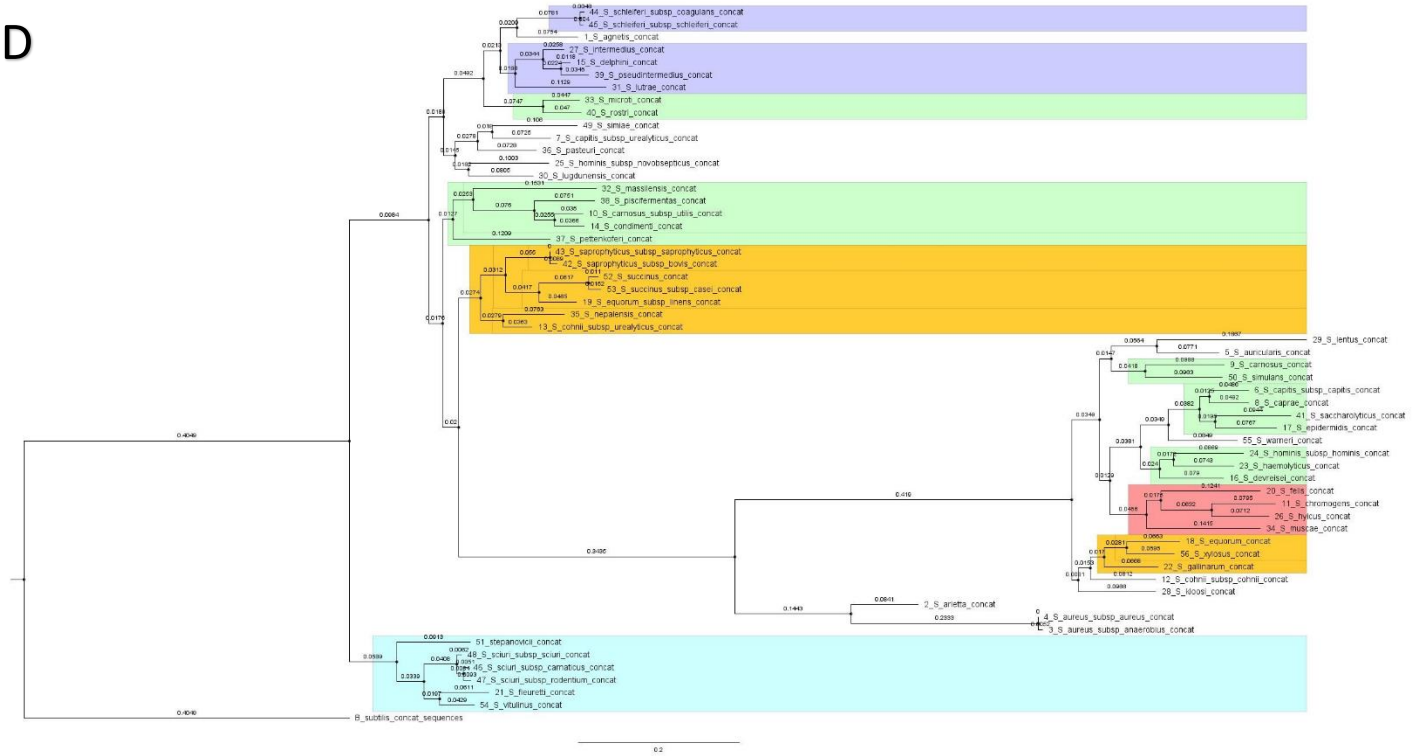


Fig 2C & 2D represent Maximum likelihood trees for rpoB single locus and multi-locus dataset

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

Table2: Log Likelihood and AIC scores for the best Models in RAxML-NG maximum likelihood analysis

Results & Discussion

The staphylococcus genus is an economically important class of microorganisms since they have both human and animals as hosts. The genus has been classified broadly into 5 species groups based on their phenotypic characteristics – novobiocin resistance/susceptible, oxidase positive/negative and coagulase positive/negative. The present study resulted in a refined classification of these 5 groups into 15 clusters. In accordance with previous studies these analyses identified the Sciuri group (*S. sciuri, S. stepanovicii, S. vitulinus, S. lentus and S. fleurette)* containing the novobiocin-resistant, oxidase positive species as the sister group to all other *Staphylococcus species*. This cluster group also contains the relatively recently discovered *S. stepanovicii*.

Phylogenetic analysis of single gene datasets revealed that 16SrRNA and dnaJ resolved more or less similar clades. However, the use of protein coding dnaJ sequences resulted in previously

unreported relationships between a few taxa. For instance, Arlettae-kloosi group comprising *S. arlettae* and *S. kloosi* species and Muscae group comprising *S. muscae, S.microti* and *S. rostri* were resolved with dnaJ sequences.

RpoB sequences reported relationships inconsistent with classification obtained by using 16SrRNA and dnaJ sequences as well as did not agree with phenotypic grouping of *Staphylococcus* species. For this reason, the rpoB dataset was not included in the multi-locus dataset analysis.

Comparing the neighbor joining and maximum likelihood methods of phylogenetic analysis revealed that there was some consensus between higher order relationships among the species. For e.g. for 16SrRNA sequences, the resolution of most clades was consistent between the two approaches except for the Sciuri group which was resolved further into two clusters with the ML approach (Fig 1A and 2A) . For DnaJ dataset the Hycius group consisted of only *S. hyicus* and *S. chromogenes* species for both approaches as opposed to the inclusion of *S. agnetis* and *S.felis* in this group for the 16SrRNA dataset. The inclusion of *S. muscae* in the Hyicus group for the multi-locus dataset was an interesting observation and suggests further evaluation of taxa relationships at the molecular level (Fig. 1D and 2D). There is overall good support for nodes of higher-level relationships for both neighbor joining and maximum likelihood trees (Supplementary information)

Some reports have suggested that using a multi-locus data set provides better resolution of phylogenetic trees of microbial studies involving many taxa. The present study however did not find an advantage of using the concatenated multilocus dataset of 16SrRNA and DnaJ sequences.

Conclusion

Regardless of the method used it was observed that the *Staphylococcus* phylogeny requires refinement. The use of dnaJ sequences resulted in identifying previously unreported relationships among Staphylococcus taxa. Although the use of a multi-locus dataset has been suggested to provide robust phylogenetic inference, the present study did not find an advantage of using such a dataset over single gene loci.

References

[1] O. Sakwinska, M. Giddey, M. Moreillon, D. Morisset, A. Waldvogel, and P. Moreillon, "Staphylococcus aureus host range and human-bovine host shift," *Appl. Environ. Microbiol.*, vol. 77, no. 17, pp. 5908–5915, 2011.

[2] D. Balasubramanian, L. Harper, B. Shopsin, and V. J. Torres, "Staphylococcus aureus pathogenesis in diverse host environments," no. October 2016, pp. 1–13, 2017.

[3] J. E. Clarridge and C. Alerts, "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases," *Clin. Microbiol. Rev.*, vol. 17, no. 4, pp. 840–862, 2004.

[4] R. J. Case, Y. Boucher, I. Dahllöf, C. Holmström, W. F. Doolittle, and S. Kjelleberg, "Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies," *Appl. Environ. Microbiol.*, vol. 73, no. 1, pp. 278–288, 2007.

[5] R. P. Lamers, G. Muthukrishnan, T. A. Castoe, S. Tafur, A. M. Cole, and C. L. Parkinson, "Phylogenetic relationships among Staphylococcus species and refinement of cluster groups based on multilocus data," *BMC Evol. Biol.*, vol. 12, no. 1, p. 171, 2012.

[6] B. Ghebremedhin, F. Layer, W. König, and B. König, "Genetic classification and distinguishing of Staphylococcus species based on different partial gap, 16S rRNA, hsp60, rpoB, sodA, and tuf gene sequences," *J. Clin. Microbiol.*, vol. 46, no. 3, pp. 1019–1025, 2008.

[7] Z. Yang and B. Rannala, "Molecular phylogenetics: principles and practice," *Nat. Rev.*

Genet., vol. 13, no. 5, pp. 303–314, 2012.