A Phylogenetic Study of the Streptococcus suis Bacteria

AVIVA ENGLANDER

May 2025

Abstract

Streptococcus suis is a pathogenic species in the streptococcus genus. Because it causes widespread disease transmission between pigs and humans, S. suis has long been a bacteria of interest. S. Suis is connected to pneumonia and meningitis, which can both be serious and at times fatal. Improper antibiotic usage by pig farmers has resulted in antibiotic resistant strains. This study aims to describe the diversity and relatedness of all known S. suis serotypes. Understanding how these serotypes are related provides scientists with insights into how best to treat S. suis infections and how to treat emerging strains which may be genetically close to known serotypes. I used Bayesian and Maximum Likelihood techniques to infer trees from the 16S ribosomal RNA gene data of all known serotypes of S. suis.

1 Introduction

The streptococcus genus contains a wide variety of pathogenic and nonpathogenic bacteria which live on the skin and mucosal surfaces of humans and animals. Streptococcus thermophilus is used in the production of yogurt and dairy products [18], while streptococcus pyogenes have been linked to strep throat, toxic shock syndrome and necrotizing fasciitis[9]. Althought there are many fascinating taxa to study in this genus, my phylogenetic analysis will be focused on streptococcus suis. Streptococcus suis was first linked to Streptococcal meningitis and arthritis in piglets in 1954 [5]. Later on human cases of Streptococcus suis infections were reported in Denmark and the Netherlands [15, 1]. Now streptococcus suis is known worldwide.

S. suis typically infects the respiratory tracts of pigs and can spread to humans through contact with infected animals or consumption of their meat. Unfortunately, due to the indiscriminate use of antibiotics by pig farmers, there are many antibiotic resistant variants [10]. Although S. suis is known to infect other mammals and bird species[4], it is primarily studied in the context of its effect on humans through pig husbandry and pork consumption.

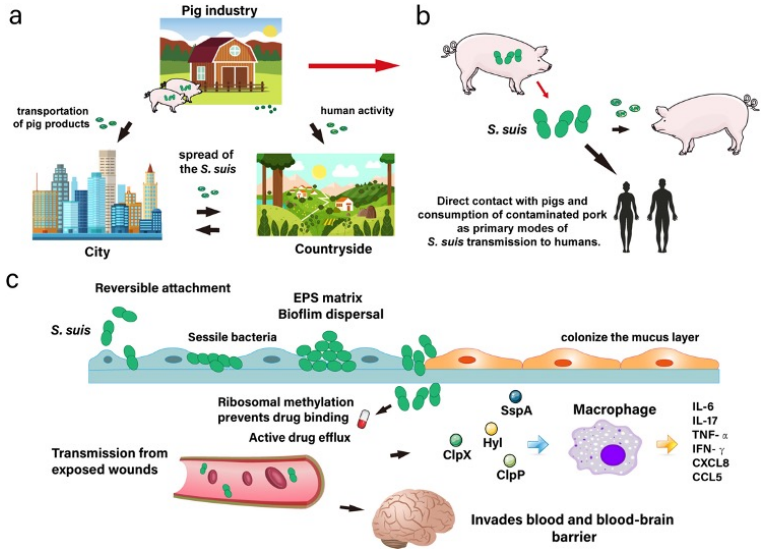

Figure 1: This figure taken from [10] describes the spread of S. suis between pigs and humans

There are over 35 known serotypes of S. suis. A 1987 study on 13 serotypes determined that these serotypes averaged over 80% relatedness confirming that they belong to the same species [8]. Another study [6] using Australian samples of 14 serotypes of S. suis showed that S. suis is more diverse than expected from the earlier study [8]. There have been several studies confirming heterogeneity between various serotypes [2, 12, 13, 7, 16]. One study found that the 16S ribosomal RNA gene was a reliable way to distinguish S. suis from 30 other bacteria species in the Streptococcus genus [3]. This is the same gene used by Chatellier et al [4] in their phylogenetic study of all serotypes of S. suis. They were the first to conduct a study of all S. suis serotypes to gain a better understanding of their diversity. I used the same data set as [4], but different tree inference methods to better understand how related the known serotypes are.

Understanding the relatedness of serotypes helps scientists to develop new therapeutic methods. In addition understanding the diversity of this species may give scientists insight into handling emerging strains which are more related to known serotypes.

2 Methods

2.1 Data

The samples I analyzed were mostly collected from diseased pigs. The rest came from lambs and calves and humans. My data comes from several coun-

tries including Denmark, The Netherlands, The US and Canada. The samples contained the 16S ribosomal RNA gene, a commonly used gene to identify bacteria. I considered all 35 different serotypes of S. suis.

Below is a table describing my samples in greater detail.

Seq Accession Number	Serotype	Source	Origin
AF009475	1	Diseased Pig	Netherlands
AF009476	1/2	Diseased Pig	Netherlands
AF009477	2	Diseased Pig	Netherlands
AF009478	3	Diseased Pig	Denmark
AF009479	4	Diseased Pig	Denmark
AF009480	5	Diseased Pig	Denmark
AF009481	6	Diseased Pig	Denmark
AF009482	7	Diseased Pig	Denmark
AF009483	8	Diseased Pig	Denmark
AF009484	9	Diseased Pig	Denmark
AF009485	10	Diseased Pig	Denmark
AF009486	11	Diseased Pig	Denmark
AF009487	12	Diseased Pig	Denmark
AF009488	13	Diseased Pig	Denmark
AF009489	14	Diseased Human	Netherlands
AF009490	15	Diseased Pig	Netherlands
AF009491	16	Diseased Pig	Denmark
AF009492	17	Clinically Healthy Pig	Canada
AF009493	18	Clinically Healthy Pig	Canada
AF009494	19	Clinically Healthy Pig	Canada
AF009495	20	Diseased Calf	USA
AF009496	21	Clinically Healthy Pig	Canada
AF009497	22	Diseased Pig	Canada
AF009498	23	Diseased Pig	Canada
AF009499	24	Diseased Pig	Canada
AF009500	25	Diseased Pig	Canada
AF009501	26	Diseased Pig	Canada
AF009502	27	Diseased Pig	Canada
AF009503	28	Diseased Pig	Canada
AF009504	29	Diseased Pig	Canada
AF009505	30	Diseased Pig	Canada
AF009506	31	Diseased Calf	Canada
AF009507	32	Diseased Pig	Canada
AF009508	33	Diseased Lamb	Canada
AF009509	34	Diseased Pig	Canada

2.2 Quality Control and Alignment

No quality control was needed for this data because it was already in the fasta format when I pulled it from the NIH genbank database.

I used clustal for my alignment, because I wanted to compare my results to the previous analysis of this data [4] and they used clustal also.

2.3 Phylogenomic Analyses

For the data described in the data subsection, I compared three different methods for tree inference. The first was a distance based approach. The second was a maximum likelihood approach. The third was a bayesian inference method. I used clustal and tcoffee for my alignment so that I could compare them. The distance based approach used the ape toolbox in R[14]. For the Maximum Likelihood approach, I used IQTree [11]. Finally for my bayesian tree I used Mr Bayes [17].

2.4 Maximum Likelihood(ML) Analysis

I used IQTree[11] for my maximum likelihood analysis with the default settings. A benefit of IQTree is that it randomly samples local optima and then picks the best tree from the local optima. It also systematically accounts for missing data by using the terraphast library to report ML trees that are located on a "terrace". This is when we have two or more trees with identical likelihoods, which occurs when we have a lot of missing data. IQtree is also great because it allows for reversible and nonreversible models.

Unfortunately the stochastic approach to optimization results in a lot of computational cost. In addition it can also be very computationally expensive on large datasets. Although IQtree2 does have options to reduce memory usage, it is still computationally costly on large datasets. It also assumes that the appropriate model of sequence evolution is chosen. Due to the stochastic element of the algorithm it also assumes the user will run it several times for optimal results.

I let IQTree run ModelFinder which checked 286 DNA models before settling on TPM3+F+I. TPM3 is a model that allows us to simplify our computations by saying AC=CG, AG=CT, AT=GT and assuming equal base frequencies. +F means we use empirical base frequencies. "This is the default if the model has unequal base freq. In AliSim, if users neither specify base frequencies nor supply an input alignment, AliSim will generate base frequencies from empirical distributions."[11]. In addition the +I means we allow a proportion of invariable sites.

2.5 Bayesian Methods

For the Bayesian inference portion of my analysis, I used Mr Bayes [17]. Mr Bayes is very powerful framework with many parameters and options for specializing our model. It allows model mixtures and users can link and unlink parameters across selected data subsets. It also has a wide variety of models for different types of data including morphological, protein and nucleotide. It is also efficient because when the user changes the substitution model, it only recalculates for the subsets of data affected by the change. In order to escape local optima and speed up converegence it uses Metropolis coupled Markov Chain Monte Carlo.

Mr Bayes by default assumes equal rates of evolution across sites, although that can be changed by the user. The incredible flexibility of Mr Bayes is a double edged sword, whose usefulness depends on who uses it. Similarly to ML models, the substitution model chosen greatly affects results. In addition, the choice of priors can change the results greatly. Someone with more in depth knowledge on a species could take advantage of this to choose better priors and more complex substitution models that reflect known conditions. However in the hands of a non-expert, all of these choices can be overwhelming and the wrong choice can spell disaster. Some other user options include subdividing the data by specifying character sets, and deciding to link or unlink different topologies and branch lengths.

For my bayesian inference I selected serotype 32 as my outgroup, because I knew from my initial distance matrix check that it was distant from the other clusters and serotypes. I also set the following priors and the HKY model with site rates modeled with a discrete model.

```
prset brlenspr=unconstrained:exp(10.0);
prset shapepr=exp(1.0);
prset tratiopr=beta(1.0,1.0);
prset statefreqpr=dirichlet(1.0,1.0,1.0,1.0);
lset nst=2 rates=gamma ngammacat=4;
```

I chose the HKY model, which allows for different nucleotide frequencies for each base, because I hoped that allowing for more flexibility would result in the algorithm finding a better fit. My priors were chosen a bit randomly because I wasn't sure which would make sense for this case.

3 Results

After I did the clustal alignment I made a tree based on the distance matrix in R which I computed using the dist function in ape[14]. Then I noticed that serotype samples 32-34 were very distant from the other serotypes of S. suis. The rest of the serotypes were very close to each other. The 32-34 samples were

all collected in Canada from diseased pigs. This further supports the possibility that they are closely related because they branched off from a common ancestor in Canada.

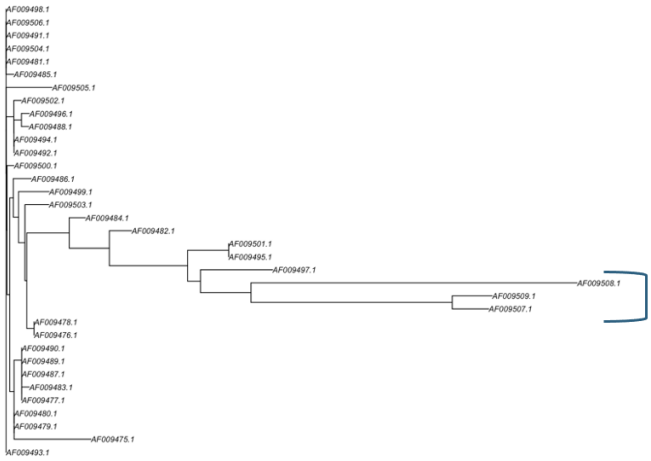

Figure 2: This tree came from my distance matrix computed from my clustal alignment. The bracket shows the 34 and 32 in one clade with 33.

My IQtree analysis results in 4 main clusters. We see the larger cluster where everything is much closer together at the top. Then serotypes 7 and 9 form a clade, these are both samples from Denmark, so their geographic closeness supports the likelihood that they evolved from a nearby common ancestor. The third clade contains serotypes 26 and 20 which are both from North America which also supports their likely genetic closeness. The fourth contains 34 and 32 which we expected because our alignment data showed they were very distant from the other samples along with 33.

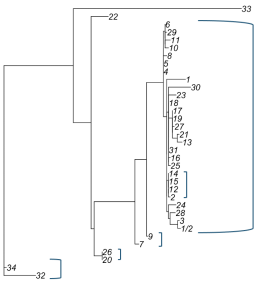

Figure 3: This tree came from my Maximum Likelihood Analysis

When considering the Bayesian analysis the first cluster I note in my figure splits into two clusters where the first cluster (serotypes 15,14,12,2) is mostly collected from the Netherlands. The closeness in location, supports the inference that they should be placed close together in a phylogenetic tree. The bottom cluster of 32-34 which showed up in the distance matrix analysis is still visible. In addition, the second from the bottom cluster (serotypes 26,20,22) are all from North America which further supports that this cluster is most likely closely related and makes sense as a cluster. The rest of the serotypes were very closely related.

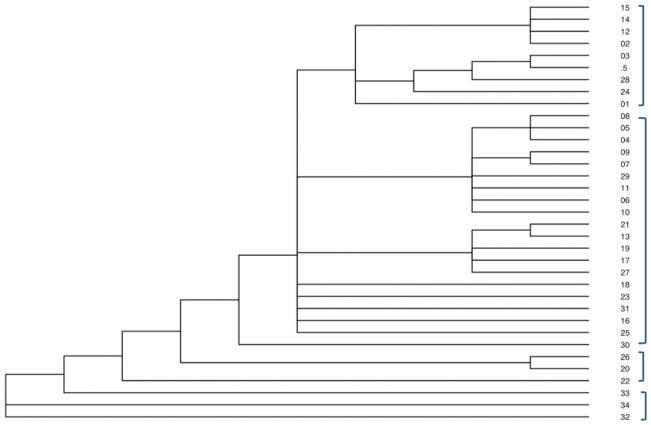

Figure 4: This tree came from my Bayesian Analysis

4 Discussion and Future Work

Comparing my results to the other phylogenetic tree inferred on the same data [4]. My IQtree was closest to theirs and found all of the same major clusters.

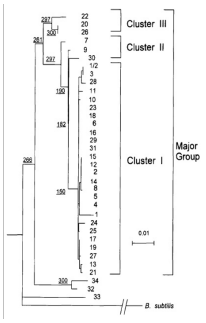

Figure 5: This figure taken from [4] and shows the result of their analysis.

They [4] used a neighborjoining method. The fact that IQtree agrees with their result, makes me think that it is probably correct. IQtree's automatic model selection probably provided a better result than my somewhat random choices in Mr. Bayes.Understanding these clusters may be helpful for extending treatments for similar bacteria serotypes to their neighbors.

In the future, I would like to apply coalescent methods and learn more about bacteria phylogenies, so that I can make more educated guesses for my bayesian analysis. There are so many options for Mr Bayes that I feel I unintentionally constrained myself to a less correct local optima tree. In addition, these bacteria samples were limited to Europe and North America. In future work it would be interesting to compare to Asian samples. China has an incredibly diverse pig population and is also the world's top consumer of pork products. Because of that, there is probably some interesting S. suis evolutionary patterns in Asian bacteria populations.

References

[1] JP Arends and HC Zanen. Meningitis caused by streptococcus suis in humans. *Reviews of infectious diseases*, 10(1):131–137, 1988.

[2] Marc Beaudoin, Josée Harel, Robert Higgins, Marcelo Gottschalk, Michel Frenette, and Janet I Macinnes. Molecular analysis of isolates of streptococcus suis capsular type 2 by restriction-endonuclease-digested dna separated on sds-page and by hybridization with an rdna probe. *Microbiology*, 138(12):2639–2645, 1992.

[3] Robert W Bentley, James A Leigh, and Matthew D Collins. Intrageneric structure of streptococcus based on comparative analysis of small-subunit rrna sequences. *International Journal of Systematic and Evolutionary Microbiology*, 41(4):487–494, 1991.

[4] Sonia Chatellier, Josee Harel, Ying Zhang, Marcelo Gottschalk, Robert Higgins, Luc A Devriese, and Roland Brousseau. Phylogenetic diversity of streptococcus suis strains of various serotypes as revealed by 16s rrna gene sequence comparison. *International Journal of Systematic and Evolutionary Microbiology*, 48(2):581–589, 1998.

[5] HI Field, D Buntain, and JT Done. Studies on piglet mortality. i. streptococcal meningitis and arthritis. 1954.

[6] DJ Hampson, DJ Trott, IL Clarke, CG Mwaniki, and ID Robertson. Population structure of australian isolates of streptococcus suis. *Journal of Clinical Microbiology*, 31(11):2895–2900, 1993.

[7] Josee Harel, Robert Higgins, Marcelo Gottschalk, and Michel Bigras-Poulin. Genomic relatedness among reference strains of different strepto-

coccus suis serotypes. *Canadian Journal of Veterinary Research*, 58(4):259, 1994.

[8] Renate Kilpper-Bälz and Karl Heinz Schleifer. Streptococcus suis sp. nov., nom. rev. *International Journal of Systematic and Evolutionary Microbiology*, 37(2):160–162, 1987.

[9] Christopher N. LaRock and Victor Nizet. Cationic antimicrobial peptide resistance mechanisms of streptococcal pathogens. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1848(11, Part B):3047–3054, 2015. Bacterial Resistance to Antimicrobial Peptides.

[10] Ruoyi Lv, Wenjing Zhang, Zhigang Sun, Xiaohui Si, Hong Dong, and Xiaoye Liu. Current prevalence and therapeutic strategies for porcine ¡i¿streptococcus suis¡/i¿ in china. *Applied and Environmental Microbiology*, 91(3):e02160–24, 2025.

[11] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5):1530–1534, 02 2020.

[12] JD Mogollon, C Pijoan, MP Murtaugh, EL Kaplan, JE Collins, and PP Cleary. Characterization of prototype and clinically defined strains of streptococcus suis by genomic fingerprinting. *Journal of clinical microbiology*, 28(11):2462–2466, 1990.

[13] OGI Okwumabua, Jacque Staats, and MM Chengappa. Detection of genomic heterogeneity in streptococcus suis isolates by dna restriction fragment length polymorphisms of rrna genes (ribotyping). *Journal of Clinical Microbiology*, 33(4):968–972, 1995.

[14] Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2019.

[15] Beate Perch, P Kristjansen, and KN Skadhauge. Group r streptococci pathogenic for man: two cases of meningitis and one fatal case of sepsis. *Acta Pathologica Microbiologica Scandinavica*, 74(1):69–76, 1968.

[16] Hilde E Smith, Martine Rijnsburger, Norbert Stockhofe-Zurwieden, Henk J Wisselink, Uri Vecht, and Mari A Smits. Virulent strains of streptococcus suis serotype 2 and highly virulent strains of streptococcus suis serotype 1 can be recognized by a unique ribotype profile. *Journal of clinical microbiology*, 35(5):1049–1053, 1997.

[17] Okezie Uche-Ikonne, Frank Dondelinger, and Tom Palmer. Software Application Profile: Bayesian estimation of inverse variance weighted and MR-Egger models for two-sample Mendelian randomization studies – mrbayes. *International Journal of Epidemiology*, 50(1):43–49, 2021.

[18] Tingting Zhang, Yan Li, Chunying Yuan, Xiaoce Zhu, Mingyu Wang, Suzhen Yang, and Jian Kong. Cytoprotective effects of the fermented milk by streptococcus thermophilus cgmcc 24468 against ros damage in hacat cells. *Journal of Functional Foods*, 128:106829, 2025.