Prevalence Analysis of Selected Streptococcus suis Antigens

To quantify the prevalence of five candidate antigens (HP0197, Fnbp, Sao, C5a-peptidase, and Suilysin) within a diverse genomic dataset of Streptococcus suis, a dedicated bioinformatic pipeline, SSUIS-SADE v1.0.0 (Streptococcus suis Antigen Distribution Explorer), was developed.

Dataset Curation and Assembly

A comprehensive set of S. suis assemblies (taxonomy ID 1307) was initially downloaded from the NCBI RefSeq database (n = 402; 1 March 2025). To ensure high quality, this dataset was curated by removing duplicate BioSample IDs and assemblies with a total contig length below 500 kb. This resulted in a final analysis set of 388 genomes, comprising 183 with serotype annotations and 205 untyped isolates. The genomic FASTA files (.fna) were utilized as published without further modification, such as coding sequence (CDS) extraction.

Sequence Database and Antigen Queries

The 388 genomic FASTA sequences were concatenated into a single file (all\_suis\_genomes.fna), which was subsequently indexed to create a searchable nucleotide database using makeblastdb (BLAST+ v2.15.0) [1, 2].

Reference protein sequences for the five full-length antigens were compiled into a query file (query\_antigens.fasta). For a parallel sensitivity analysis, a second query file (query\_antigens\_highlight.fasta) was created, containing only the sequences of conserved domains ("highlight" fragments) from the same antigens.

Homology Search and Prevalence Metrics

The antigen protein queries were aligned against the six-frame translated genomic database using tblastn [1, 2]. Homology searches were performed with an E-value threshold of 1 × 10⁻⁵ and were parallelized across all available CPU cores.

The resulting tabular output (format 6) was processed using a custom script (parse\_prevalence.py) leveraging the pandas [3, 4] and Biopython [5] libraries. An antigen was considered present in a genome if at least one alignment met predefined criteria. For full-length antigens, stringent thresholds of ≥ 70% identity and ≥ 80% query coverage were required. For the highlight-domain sensitivity analysis, more lenient thresholds of ≥ 60% identity and ≥ 50% coverage were applied. These parameters were selected to appropriately balance sensitivity and specificity, consistent with approaches in previous cross-serotype studies of streptococci [6]. Antigen prevalence was calculated as the percentage of positive genomes out of the 388 total genomes analyzed.

Computational Reproducibility

All custom scripts, the Conda environment specification (environment.yml), and a Dockerfile for creating a reproducible container (Ubuntu 22.04, BLAST+, Python) are publicly available in the project's GitHub repository (release v1.0.0). The complete analysis can be executed via the provided Docker container, ensuring full reproducibility:

Bash

docker run --rm -v $(pwd):/work suis-prevalence:1.0 \

    bash run\_suis\_prevalence.sh

    ---

    ### References

    1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. \*J Mol Biol.\* 1990;215(3):403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

    2. Camacho C, Coulouris G, Avagyan V, \*et al.\* BLAST+: architecture and applications. \*BMC Bioinformatics.\* 2009;10:421. https://doi.org/10.1186/1471-2105-10-421

    3. The pandas development team. \*pandas-dev/pandas: Pandas (software).\* Zenodo; 2020. https://doi.org/10.5281/zenodo.3509134

    4. McKinney W. Data structures for statistical computing in Python. In: \*Proceedings of the 9th Python in Science Conference.\* 2010:56-61. https://doi.org/10.25080/Majora-92bf1922-00a

    5. Cock PJA, Antao T, Chang JT, \*et al.\* Biopython: freely available Python tools for computational molecular biology and bioinformatics. \*Bioinformatics.\* 2009;25(11):1422-1423. https://doi.org/10.1093/bioinformatics/btp163

    6. Estrada AA, Gottschalk M, Gebhart C, \*et al.\* Comparative analysis of \*Streptococcus suis\* genomes identifies novel candidate virulence-associated genes in North American isolates. \*Vet Res.\* 2022;53:23. https://doi.org/10.1186/s13567-022-01039-8

압축본   
  
The prevalence of five candidate antigens (HP0197, Fnbp, Sao, C5a-peptidase, and Suilysin) was quantified across 388 curated, high-quality *Streptococcus suis* genomes. A genomic database was constructed from all assemblies and indexed using makeblastdb (BLAST+ v2.15.0) [1, 2].

Reference protein sequences for both full-length antigens and their conserved "highlight" domains were aligned against the nucleotide database using tblastn [1, 2] with an *E*-value threshold of 1 × 10⁻⁵. The tabular results were processed with a custom Python script utilizing the pandas [3, 4] and Biopython [5] libraries.

An antigen was defined as present if a genome contained at least one hit meeting specific criteria: **≥70% identity and ≥80% query coverage** for full-length sequences, or **≥60% identity and ≥50% coverage** for highlight domains. These thresholds were chosen to balance sensitivity and specificity, mirroring previous streptococcal studies [6]. Prevalence was calculated as the percentage of positive genomes. All analysis scripts and a reproducible Docker environment have been made publicly available on GitHub.