Advance Publication by J-STAGE

Japanese Journal of Infectious Diseases

**Comparative genomic features of *Streptococcus canis* based on
the pan-genome orthologous group analysis according to
sequence type**

Jung-Min Kim, Yasuto Fukushima, Haruno Yoshida, Jae-Seok Kim, and
Takashi Takahashi

Received: July 16, 2021. Accepted: September 13, 2021.

Published online: September 30, 2021.

DOI:10.7883/yoken.JJID.2021.533

Advance Publication articles have been accepted by JJID but have not been copyedited
or formatted for publication.

Original Article

Comparative genomic features of *Streptococcus canis* based on the pan-genome orthologous group analysis according to sequence type

Jung-Min Kim¹, Yasuto Fukushima², Haruno Yoshida², Jae-Seok Kim^{1**}, and Takashi Takahashi^{2*}

¹*Department of Laboratory Medicine, Kangdong Sacred Heart Hospital, Hallym University College of Medicine, Yeongdeungpo-gu, Seoul 05355, Republic of Korea*

²*Laboratory of Infectious Diseases, Graduate School of Infection Control Sciences & Ōmura Satoshi Memorial Institute, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan*

Running Title: POG analysis of *S. canis* with WGS

Keywords: *Streptococcus canis*, whole genome sequencing, comparative genomics, pan-genome orthologous group, sequence type

***Corresponding author: Prof. Takashi Takahashi MD, PhD**

Laboratory of Infectious Diseases, Graduate School of Infection Control Sciences & Ōmura Satoshi Memorial Institute, Kitasato University
5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan

Tel: +81-3-5791-6428, Fax: +81-3-5791-6441, E-mail: taka2si@lisci.kitasato-u.ac.jp

****Co-corresponding author: Prof. Jae-Seok Kim MD, PhD**

Department of Laboratory Medicine, Kangdong Sacred Heart Hospital, Hallym University College of Medicine, Yeongdeungpo-gu, Seoul 05355, Republic of Korea
Tel: +82-2-2224-2327, Fax: +82-2-2224-2214, E-mail: jaeseok@hallym.ac.kr

ABSTRACT: Using bacterial pan-genome obtained through whole genome sequencing (WGS), coding DNA sequences (CDSs) can be clustered into pan-genome orthologous groups (POGs). We aimed to investigate comparative genomic features of *Streptococcus canis* based on POG analysis and to determine CDSs specific to prevalent sequence type (ST) 9. Twenty WGS datasets on *S. canis* strains from invasive and non-invasive specimens were retrieved from the National Center for Biotechnology Information Assembly database. Based on the WGS data, we performed comparative genome hybridization (CGH), pan- and core-genome prediction, Venn diagram test with five ST9 strains, and phylogenetic analysis, with ST determination. We compared the CDSs between seven ST9 and 13 non-ST9 strains. We observed genomic diversity based on CGH and Venn diagram. The predicted pan- and core-genomes contained 4,772 and 1,403 genes, respectively. We found five clades consisting of different STs (ST1, ST44/2, ST13/14, ST21/15/41, and ST9) based on phylogenetic tree. There were differences in four pathways (DNA restriction-modification system, DNA-mediated transposition, extracellular region, and response to oxidative stress) regulated by CDSs specific to ST9. Our findings describe genomic diversity in CGH and Venn diagram, pan- and core-genomes, five clades of genomes consisting of different STs, and unique CDS features associated to ST9.

INTRODUCTION

Streptococcus canis, which was first reported in 1986 (1), forms large, smooth, gray-white-colored colonies with strong beta-hemolysis on 5% sheep blood agar plates. According to Lancefield grouping, *S. canis* is classified as group G streptococcus based on cell wall carbohydrate antigenicity. In healthy dogs, this microorganism forms part of the resident microflora of the oropharynx, skin, genitourinary tract, and anus (2). This bacterium is also an emerging pathogen that causes self-limiting dermatitis. However, in some cases, this bacterial infection leads to severe diseases in companion animals, including arthritis, streptococcal toxic shock syndrome, necrotizing fasciitis, septicemia, and pneumonia (3,4). *S. canis* infects humans who have been in close contact with companion animals and can cause either local or systemic diseases (5). In fact, this species was recovered from two Japanese patients with bacteremia who were in close contact with or bitten by companion dogs (6,7). A recent study reported periprosthetic joint infection caused by *S. canis* in a man undergoing elective primary total hip arthroplasty (8): the patient had a pet dog frequently licking his legs. Thus, clinicians should be aware of the possibility that *S. canis*-related invasive zoonotic infections may be underdiagnosed.

Multilocus sequence typing (MLST) is performed to evaluate the clonal spread of strains with genetic similarities, using polymerase chain reaction (PCR) and direct sequencing data or whole genome sequencing (WGS) datasets. Thus, the sequence types (STs) are assessed based on the sequence variations and allelic combination of seven housekeeping genes (allelic profile, *gki-gtr-murI-mutS-recP-xpt-yqiZ*), according to the PubMLST website (<https://pubmlst.org/scanis/>). Recently, STs determined by MLST have been documented in dogs with ulcerative keratitis and in an outbreak of

subclinical mastitis in a dairy herd caused by *S. canis* (9,10). Using the PubMLST Isolate database (https://pubmlst.org/bigsdb?db=pubmlst_scanis_isolates&page=query, a total of twelve blood-origin strains registered on July 8, 2021), we searched for the prevalence ($n = 7$, 53.8%) of blood-origin strains belonging to ST9. Additionally, WGS data of the blood-origin strain (FU149) isolated from a dog with necrotizing soft tissue infection showed that FU149 belonged to ST9 (11). Thus, the prevalent ST responsible for invasiveness seems to be ST9, which may be associated with allele type 1 of *S. canis* M-like protein (SCM), a virulent factor (12).

The complete and draft WGS datasets of invasive and non-invasive *S. canis* strains from animals (11,13,14) and human patients with bacteremia (15,16) from Japan and Korea have already been reported. A total of 20 WGS datasets were registered and published in the National Center for Biotechnology Information (NCBI) Assembly database (<https://www.ncbi.nlm.nih.gov/assembly/?term=Streptococcus%20canis>) on July 15, 2021. It is now common practice to characterize the full gene complement as a “pan-genome” (pan, from the Greek word, meaning whole) of a closely related group of a single bacterial species (e.g., *Streptococcus agalactiae*) using a web-based pipeline (17). After constructing this pan-genome using WGS data, all protein-coding DNA sequences (CDSs) can be clustered into pan-genome orthologous groups (POGs) using the algorithm in the pipeline. However, to the best of our knowledge, there is no description of POG analysis for *S. canis* strains. The purpose of our study was to clarify the comparative genomic features of 20 invasive and non-invasive strains based on POG analysis and to determine the CDSs specific to ST9.

MATERIALS AND METHODS

List of WGS data: Table 1 shows the list of the 20 WGS datasets (including GenBank accession numbers) published in the NCBI database, of which 14 from Japan ($n = 10$) and Korea ($n = 4$) were original. Two sequencing approaches (Illumina/Nanopore sequencing) were conducted for each Korean strain (14). For Illumina sequencing, genomic libraries were prepared using the Nextera DNA Flex library prep kit, and sequencing was performed on the Illumina MiSeq platform with a 2×150 -bp paired-end protocol. We processed the raw reads using fastp v. 0.20.0, with default settings. For Nanopore sequencing, a DNA library was constructed using the rapid sequencing kit (SQK-RAD004). A MinION flow cell (FLO-MIN106; R9.4) was used for sequencing using the MinION software v. 19.12.5 in the standard 48-h sequencing script. Fast5 reads were base called using this MinION software, and the resulting fastq reads were used for assembly by Unicycler. Information on the host, strain collection date, geographic location, and isolation source was retrieved from the corresponding BioSample website. There were three human strains, three blood-origin invasive strains, and six strains with complete WGS datasets. We determined the STs using the MLST database version 2.0 (<https://cge.cbs.dtu.dk/services/MLST/>) (18), except for ST9 of the TA4 strain, which was determined by PCR and direct sequencing. STs were grouped into clonal complexes (CCs), in which related STs were classified as single locus variants differing in only one housekeeping gene. SCM allele typing was conducted on the basis of the variations of SCM amino acid sequences on the phylogenetic tree (19,20). All the seven ST9 strains belonged to allele type 1.

Comparative genome hybridization (CGH): The CGView comparison tool (CCT) application was used for visual comparison of CGH comprising 20 WGS datasets (21). This CCT map consists of several rings showing the reference genome features and the results of basic local alignment search tool (BLAST) comparisons

between the reference and comparison sequences. All 19 WGS datasets were aligned using the complete genome of the HL_77_1 strain (ST1 from Korea) as a reference sequence since this circular genome was constructed for the first time in Korea. Clusters of orthologous gene (COG) functional categories were displayed using different colors based on the NCBI COG database (<https://www.ncbi.nlm.nih.gov/research/cog>). Percent identities were also shown using different colors based on the percent BLAST hits.

Pan- and core-genome prediction, Venn diagram test, and phylogenetic

analysis: The pan- and core-genome prediction curves were analyzed using Roary version 3.11.2 (22) and PanGP version 1.0.1 (23). Roary was run at default parameters using the GFF3 file generated by Prokka version 1.12 (24). PanGP was run using all the traverse sub-sampling algorithms. The numbers of genes in the pan-genome and core-genome were plotted according to the number of genomes sequentially added. Curve fitting of the pan-genome was performed using power-law regression based on Heap's law [$y = A_{\text{pan}}x^{B_{\text{pan}}} + C_{\text{pan}}$]. As previously described (25,26), the y-axis is the pan-genome size, the x-axis is the genome number, and A_{pan} , B_{pan} , and C_{pan} are the fitting parameters. The curve fitting of core-genome also was performed using an exponential regression model [$y = A_{\text{core}}e^{B_{\text{core}}x} + C_{\text{core}}$], where A_{core} , B_{core} , and C_{core} are the fitting parameters.

The OrthoVenn2 program (<https://orthovenn2.bioinfotoolkits.net/home>) with default parameters (E-value 1e-5 and inflation value 1.5) (27) was used to identify orthologous gene clusters in the WGS datasets. A Venn diagram was constructed using selected five representative strains, NCTC 12191^T, TA4, OT1, FU6, and FU149, belonging to ST9. Three blood-origin invasive strains (TA4 and OT1 from two human patients and FU149 from a dog) were included in this diagram.

A phylogenetic tree was constructed based on the pan-genome data obtained from the POG analysis. Alignments of the 20 WGS datasets were performed using Mauve (MegAlign program, DNASTar version 17). Phylogenetic tree analysis was carried out using RAxML8.1, and the tree was visualized using MEGA X program.

CDS characterization: Using OrthoVenn2, 4,143 COGs were obtained from the 20 pan-genome datasets. To obtain specific CDSs based on the POG data, we performed comparative genomic analyses between the following populations: 1) ST9 ($n = 7$) vs. non-ST9 ($n = 13$) strains; 2) invasive ($n = 3$) vs. non-invasive ($n = 17$) strains. The specific orthologous clusters were identified by comparing the differences in the presence or absence of CDSs. When there were two or more orthologous clusters, the p-value was automatically calculated using OrthoVenn2. POG analytical tables [containing cluster number, gene ontology (GO) ID, pathway name, differentially present POG number, and p-value] were used to characterize and analyze the CDSs statistically. Orthologous clusters were obtained using OrthoVenn2 analysis. Cluster number corresponded to orthologous clusters specific to each category (i.e., ST9 or invasive strains). GO ID was assigned using the following website (<http://geneontology.org/docs/ontology-documentation/>), along with ontology description (biological process, molecular function, and cellular component): pathway name referred to the annotation result of GO ID. The POG number indicated the complexity of genes involved in the pathway.

Ethical statement: None to declare since background information (host, strain collection date, geographic location, and isolation source) is available from the NCBI website.

Statistical analysis: We used Fisher's exact probability test (two-sided) to determine significant differences in categorical variables between the two populations.

Statistical significance was set at $P < 0.05$.

RESULTS

Circular representation of *S. canis* genomes using CGH: The circular representation of 20 *S. canis* genomes using CGH is shown in **Fig. 1**. The complete genome of the HL_77_1 strain was selected as a reference, because this strain induced inflammation in the dog ear and belonged to ST1.

The genome of HL_77_1 consisted of a 2,261,128-bp circular chromosome with a 4,146 bp plasmid, a G+C content of 39.6%, and 2,140 CDSs. This genome contained 15 ribosomal RNA and 57 transfer RNA genes. The HL_77_1 strain showed the highest identity with the FMV2238 strain (belonging to the same ST1) and the lowest identity with the FU53 strain (belonging to ST21), suggesting genomic differences among the strains belonging to different STs.

Pan- and core-genomes, Venn diagram, and phylogenetic tree: Pan-genome and core-genome prediction curves obtained from the 20 genomes are shown in **Fig. 2**. The curve for the pan-genome was the least-squares fit of the equation ($y = Ax^B + C$) with the best fit obtained with a correlation $R^2 = 0.999$ for $A = 1,262.23 \pm 5.42$, $B = 0.27$, and $C = 725.26 \pm 15.49$. The curve for the core-genome was the least-squares fit of the equation ($y = Ae^{Bx} + C$), with the best fit obtained with a correlation $R^2 = 0.959$ for $A = 1,083.75 \pm 65.21$, $B = -0.67$, and $C = 1,434.74 \pm 2.02$. The final pan- and core-genomes consisted of 4,772 and 1,403 genes, respectively.

A Venn diagram using selected five representative strains (TA4 and OT1 from human blood, FU149 from dog blood, FU6 from cat pus, and NCTC 12191^T from bovine mastitis) belonging to ST9 is shown in **Fig. 3**. There was a higher number of

strain-unique CDSs in NCTC 12191^T ($n = 60$) than in OT1 ($n = 3$) and FU149 ($n = 9$), suggesting that the type strain might harbor multiform genes.

A phylogenetic tree constructed based on POG analysis data is shown in **Fig. 4**. We found five clades belonging to different STs [clade 1 belonging to ST1, clade 2 belonging to ST44/2, clade 3 belonging to ST13/14 (CC13), clade 4 belonging to ST21/15/41, and clade 5 belonging to ST9].

CDS features specific to ST9: Comparative POG analysis results using the genomes of 20 *S. canis* strains are shown in **Table 2**. There were significant differences in four putative pathways (DNA restriction-modification system, DNA-mediated transposition, extracellular region, and response to oxidative stress) regulated by CDSs specific to ST9, although we found no significant differences in pathways regulated by CDSs specific to invasiveness. The DNA restriction-modification system (ontology biological process) is regulated by the type I restriction-modification system R/S/M proteins (encoded by *hsdR/hsdS/hsdM*), cytosine-specific methyltransferase, and hypothetical protein. The DNA-mediated transposition (ontology biological process) is regulated by the transposase IS1239. The extracellular region (ontology cellular component) is regulated by the agglutinin receptor (encoded by *ssp5*) and T-antigen-like fimbrial structural subunit protein FszC (encoded by *fszC*). The response to oxidative stress (ontology biological process) is regulated by the glutathione peroxidase and putative glutathione peroxidase. There were no significant differences in pathways (ontology molecular function) regulated by CDSs specific to ST9.

DISCUSSION

Of the *S. canis* WGS datasets analyzed, there were six complete sequences,

whereas the remaining data were scaffolds or contigs at the assembly levels. We conducted both Illumina sequencing using Illumina MiSeq and Nanopore sequencing using MinION (14). The hybrid assembly for three strains (HL_77_1, HL_77_2, and HL_98_2, of which two strains resulted in a single chromosome and a plasmid) was performed using Unicycler version 0.4.8 (28) and SeqMan Ngen version 15 was used for the remaining strain (HL_100). In the future, to enhance the quality of comparative genomic analysis, we need to obtain and publish complete sequences in the NCBI database by using short- and long-read sequencing followed by hybrid assembly (using Unicycler).

We found *S. canis* CDS features (DNA restriction-modification system, DNA-mediated transposition, extracellular region, and response to oxidative stress) specific to ST9. All the seven ST9 strains possessed the SCM allele type 1. Furthermore, we assigned 40 strains from each year (2015 and 2017) by random sampling in order, and the relationship between biofilm production ability and other microbiological features was analyzed without bias (29). There was association between the non-producer and allele type 1/ST9. Adjusted Wallace (AW) coefficients with 95% confidence intervals (CIs) were calculated using the Comparing Partitions website (<http://www.comparingpartitions.info>). ST9 appeared to be a good predictor of SCM allele type 1 (AW_{ST9→SCM allele type 1} and 95% CI, 1.000 and 1.000–1.000) (30). A similar association between STs and allele types was found among strains with novel alleles (20). Therefore, the CDS features specific to ST9 seem to be linked with those unique to allele type 1/ST9.

Using a GO browsing tool, QuickGO (<https://www.ebi.ac.uk/QuickGO/>), we searched for the relatedness between DNA restriction-modification system (GO:0009307, ontology biological process) and DNA-mediated transposition

(GO:0006313, ontology biological process), of which both are shown in **Table 2**. The restriction-modification system is involved in DNA modification (GO:0006304) and the subsequent DNA metabolic process (GO:0006259). On the other hand, the transposition is involved in DNA recombination (GO:0006310) and the same DNA metabolic process. On the other hand, there were no significant differences in pathways (ontology molecular function) regulated by CDSs specific to ST9 strains ($n = 7$). If additional WGS datasets from ST9 strains could be obtained in future, the pathways having molecular function (ATP binding, nuclease activity, serine-type endopeptidase activity, metal ion binding, and nucleic acid binding shown in **Table 2**) may become significant. We need to do further investigation by comparative transcriptome and proteome profiling using the ST9 strains, to clarify the validity of relatedness (ontology biological process).

Unfortunately, we could not identify *S. canis* CDS features specific to invasive strains ($n = 3$), although viral genome integration into host DNA and gluconeogenesis pathways were differentially present in one POG for each. On the other hand, genomic features linked to invasiveness of *S. agalactiae* based on the POG analysis, where invasive ($n = 36$) and non-invasive ($n = 30$) strains were used, have recently been reported (31). Additional WGS datasets from invasive strains may provide more information on features unique to invasiveness.

To the best of our knowledge, this is the first report on POG-based genomic features of *S. canis* strains from animals and humans. Our important and novel findings describe the genomic diversity in CGH and Venn diagram, pan- and core-genomes, five clades of genomes consisting of different STs, and unique CDS features associated to ST9. Using the PubMLST Isolate database (225 strains registered, as of July 8, 2021), we confirmed the prevalence ($n = 54$, 23.9%) of ST9 strains, mainly recovered

from ear exudate ($n = 10$), skin and soft tissue exudate ($n = 8$), blood ($n = 7$), urine ($n = 6$), and vaginal exudate ($n = 4$). On the other hand, scarce WGS datasets have been published in the NCBI database. Therefore, we need to obtain and provide a large number of WGS data, especially from ST9 strains with invasiveness, as the next important type of analysis: the significant pathways (molecular function as well as biological process/cellular component) regulated by CDSs specific to ST9/invasiveness may be evident.

Acknowledgments This work was supported in part by the Bio and Medical Technology Development Program of the National Research Foundation funded by the Korean government, Ministry of Science and ICT (2017M3A9E4077232), by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI20C0071) to J-S.K., and by Japan Society for the Promotion of Science KAKENHI (grant numbers: 18K08447 and 21K08513) to T.T.

Conflict of interest None to declare.

REFERENCES

1. Devriese LA, Hommez J, Kilpper-Bälz R, et al. *Streptococcus canis* sp. nov.: a species of group G streptococci from animals. *Int J Syst Bacteriol*. 1986;36:422-5.
2. Devriese LA, Cruz Colque JI, De Herdt P, et al. Identification and composition of the tonsillar and anal enterococcal and streptococcal flora of dogs and cats. *J Appl Bacteriol*. 1992;73:421-5.
3. DeWinter LM, Prescott JF. Relatedness of *Streptococcus canis* from canine streptococcal toxic shock syndrome and necrotizing fasciitis. *Can J Vet Res*. 1999;63:90-5.
4. Lamm CG, Ferguson AC, Lehenbauer TW, et al. Streptococcal infection in dogs: a retrospective study of 393 cases. *Vet Pathol*. 2010;47:387-95.
5. Galpérine T, Cazorla C, Blanchard E, et al. *Streptococcus canis* infections in humans: retrospective study of 54 patients. *J Infect*. 2007;55:23-6.
6. Ohtaki H, Ohkusu K, Ohta H, et al. A case of sepsis caused by *Streptococcus canis* in a dog owner: a first case report of sepsis without dog bite in Japan. *J Infect Chemother*. 2013;19:1206-9.
7. Taniyama D, Abe Y, Sakai T, et al. Human case of bacteremia caused by *Streptococcus canis* sequence type 9 harboring the *scm* gene. *IDCases*. 2017;7:48-52.
8. McGuire A, Krysa N, Mann S. Hair of the Dog? Periprosthetic Joint Infection with *Streptococcus canis*. *Arthroplast Today*. 2021;8:53-6.
9. Enache AE, Mitchell C, Kafarnik C, et al. *Streptococcus canis* multilocus sequence typing in a case series of dogs with ulcerative keratitis. *Vet Ophthalmol*. 2020;23:252-8.

10. Eibl C, Baumgartner M, Urbantke V, et al. An outbreak of subclinical mastitis in a dairy herd caused by a novel *Streptococcus canis* sequence type (ST55). *Animals* (Basel). 2021;11:550.
11. Fukushima Y, Murata Y, Katayama Y, et al. Draft genome sequence of blood-origin *Streptococcus canis* strain FU149, isolated from a dog with necrotizing soft tissue infection. *Microbiol Resour Announc*. 2020;9:e00737-20.
12. Fukushima Y, Tsuyuki Y, Goto M, et al. Novel quinolone nonsusceptible *Streptococcus canis* strains with point mutations in quinolone resistance-determining regions and their related factors. *Jpn J Infect Dis*. 2020;73:242-9.
13. Yoshida H, Fukushima Y, Katayama Y, et al. Draft genome sequences of seven *Streptococcus canis* strains isolated from diseased companion animals in Japan. *Microbiol Resour Announc*. 2020;9:e00123-20.
14. Kim J-S, Sakaguchi S, Fukushima Y, et al. Complete genome sequences of four *Streptococcus canis* strains isolated from dogs in South Korea. *Microbiol Resour Announc*. 2020;9:e00818-20.
15. Yoshida H, Katayama Y, Fukushima Y, et al. Draft genome sequence of *Streptococcus canis* clinical strain TA4, harboring the M-like protein gene and isolated in Japan from a patient with bacteremia. *Genome Announc*. 2018;6:e01469-17.
16. Yoshida H, Katayama Y, Fukushima Y, et al. Draft genome sequence of *Streptococcus canis* clinical strain OT1, isolated from a dog owner with invasive infection without a dog bite in Japan. *Microbiol Resour Announc*. 2019;8:e00770-19.
17. He EM, Chen CW, Guo Y, et al. The genome of serotype VI *Streptococcus*

- agalactiae* serotype VI and comparative analysis. *Gene*. 2017;597:59-65.
18. Larsen MV, Cosentino S, Rasmussen S, et al. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol*. 2012;50:1355-61.
 19. Fukushima Y, Yoshida H, Goto M, et al. Prevalence and diversity of M-like protein (SCM) gene in *Streptococcus canis* isolates from diseased companion animals in Japan: implication of SCM allele. *Vet Microbiol*. 2018;225:120-4.
 20. Fukushima Y, Takahashi T, Goto M, et al. Novel diverse sequences of the *Streptococcus canis* M-like protein (SCM) gene and their prevalence in diseased companion animals: Association of their alleles with sequence types. *J Infect Chemother*. 2020;26:908-15.
 21. Grant JR, Arantes AS, Stothard P. Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics*. 2012;13:202.
 22. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31:3691-3.
 23. Zhao Y, Jia X, Yang J, et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*. 2014;30:1297-9.
 24. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068-9.
 25. Rasko DA, Rosovitz MJ, Myers GS, et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol*. 2008;190:6881-93.
 26. Tettelin H, Riley D, Cattuto C, et al. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*. 2008;11:472-7.
 27. Xu L, Dong Z, Fang L, et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species.

- Nucleic Acids Res. 2019;47:W52-W58.
28. Wick RR, Judd LM, Gorrie CL, et al. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13:e1005595.
29. Fukushima Y, Tsuyuki Y, Goto M, et al. Biofilm production ability and other microbiological features of *Streptococcus canis*. Jpn J Infect Dis. (in press) doi: 10.7883/yoken.JJID.2020.1086.
30. Pinho MD, Foster G, Pomba C, et al. *Streptococcus canis* are a single population infecting multiple animal hosts despite the diversity of the universally present M-like protein SCM. Front Microbiol. 2019;10:631.
31. Takahashi T, Lee S, Kim S. Genomic characteristics of *Streptococcus agalactiae* based on the pan-genome orthologous group analysis according to invasiveness and capsular genotype. J Infect Chemother. 2021;27:814-9.

Figure Legends

Fig. 1. The circular representation of 20 *S. canis* genomes using CGH. The complete genome of the HL_77_1 strain was selected as a reference. Starting from the outer circle, first and fourth colored arrows represent COG functional categories of the HL_77_1. Second and third blue arrows represent CDS. The arrows on the outer four circles indicate forward or reverse strands. Fifth circle indicates the HL_77_1 genome. The following nineteen circles show regions of sequence similarity detected by BLAST comparison conducted among CDS translations from all the genomes of 20 strains (HL_77_1, FMV2238, FU29, TA4, OT1, FSL Z3-227, FU149, FU6, HL_98_2, FU93, B700072, HL_100, FU97, G361, HL_77_2, NCTC 6198, NCTC 12191^T, FU1, FU129, and FU53). In the inner two circles, blue indicates positive GC skew value: yellow indicates negative GC skew value. Green indicates GC ratio higher than average value: violet indicates GC ratio lower than average value.

Fig. 2. Pan-genome and core-genome prediction curves obtained from the 20 genomes. The gene numbers in the pan-genome (blue) and core-genome (red) were plotted according to the number of genomes sequentially added. The pan-genome curve was the least-squares fit of the equation ($y = Ax^B + C$), with the best fit obtained with a correlation $R^2 = 0.999$ for $A = 1,262.23 \pm 5.42$, $B = 0.27$, and $C = 725.26 \pm 15.49$. The core-genome curve was the least-squares fit of the equation ($y = Ae^{Bx} + C$), with the best fit obtained with a correlation $R^2 = 0.959$ for $A = 1,083.75 \pm 65.21$, $B = -0.67$, and $C = 1,434.74 \pm 2.02$. The pan- and core-genomes consisted of 4,772 and 1,403 genes, respectively.

Fig. 3. A Venn diagram using selected five representative strains (TA4/OT1 from human blood, FU149 from dog blood, FU6 from cat pus, and NCTC 12191^T from bovine mastitis) belonging to ST9. NCTC 12191^T might harbor diverse genes.

Fig. 4. A phylogenetic tree constructed based on POG analysis data. We observed five clades belonging to different STs [clade 1 belonging to ST1, clade 2 belonging to ST44/2, clade 3 belonging to ST13/14 (CC13), clade 4 belonging to ST21/15/41, and clade 5 belonging to ST9].

Table 1. Whole genome sequences of *Streptococcus canis* strains published on the NCBI database

Strain	Host	Collection date (year/month/day)	Geographic location	Isolation source	Assembly level	Alias on NCBI assembly	GenBank accession number	ST (CC)	SCM allele type ¹⁾
NCTC 12191 ^T	Bovine	NA	NA	Mastitis	Complete	42912_C01	LR134293.1	ST9	Allele type 1
HL_77_1	Dog	2018/8/1	South Korea: Seoul	Ear	Complete	ASM1099391v2	CP053792.1 (chromosome)/ CP053793.1 (plasmid)	ST1	Allele type 15
HL_100	Dog	2018/8/1	South Korea: Seoul	Urine	Complete	ASM973855v1	CP046521.1	ST13 (CC13)	Allele type 4
HL_98_2	Dog	2018/8/1	South Korea: Seoul	Nasal cavity	Complete	ASM1099386v2	CP053789.1	ST15	Allele type 10
HL_77_2	Dog	2018/8/1	South Korea: Seoul	Ear	Complete	ASM1099384v2	CP053790.1 (chromosome)/ CP053791.1 (plasmid)	ST2	Allele type 2
B700072	Dog	2017/1/4	United Kingdom	Cornea	Complete	5645STDY7765767	LR590625.1	ST13 (CC13)	Allele type 4
FSL Z3-227	Cow	1999	USA: New York	Milk	Contig	ASM26830v2	NZ_AIDX00000000.1/ AIDX00000000.1	ST9	Allele type 1
NCTC 6198	Animal	NA	NA	NA	Contig	42197_B02	NZ_CABEII000000000.1/ CABEII000000000.1	ST9	Allele type 1
FU53	Cat	2017/4/25	Japan: Chiba	Nasal cavity	Contig	ASM1072169v1	NZ_BLKQ000000000.1/ BLKQ000000000.1	ST21	Allele type 10
FU1	Cat	2017/4/1	Japan: Chiba	Pus	Scaffold	ASM981169v1	NZ_BLIS000000000.1/ BLIS000000000.1	ST41	Allele type 11
OT1	Human	2012/5/7	Japan: Gifu	Blood (invasiveness)	Scaffold	ASM654032v2	NZ_BJOW000000000.1/ BJOW000000000.1	ST9	Allele type 1
FU129	Dog	2017/5/30	Japan: Niigata	Pus	Scaffold	ASM981173v1	NZ_BLIU000000000.1/ BLIU000000000.1	ST21	Allele type 10
FU29	Dog	2017/4/15	Japan: Kanagawa	Vagina	Scaffold	ASM1072157v1	NZ_BLKNO000000000.1/ BLKN000000000.1	ST44	Allele type 5
FU93	Dog	2017/5/13	Japan: Chiba	Pus	Scaffold	ASM1072181v1	NZ_BLKPO000000000.1/ BLKP000000000.1	ST21	Allele type 10
FU97	Dog	2017/5/15	Japan: Okayama	Pus	Scaffold	ASM1072183v1	NZ_BLKQ000000000.1/ BLKQ000000000.1	ST14 (CC13)	Allele type 4
TA4	Human	2016/7/15	Japan: Tokyo	Blood (invasiveness)	Scaffold	ASM277055v1	NZ_BEWZ000000000.1/ BEWZ000000000.1	ST9	Allele type 1
FU6	Cat	2017/4/1	Japan: Okayama	Pus	Scaffold	ASM981171v1	NZ_BLIT000000000.1/ BLIT000000000.1	ST9	Allele type 1
G361	Human	2006	Germany: Lower Saxony	Vagina	Contig	ASM223493v1	NZ_NMRV000000000.1/ NMRV000000000.1	ST13 (CC13)	Allele type 4
FU149	Dog	2019/12/3	Japan: Chiba	Blood (invasiveness)	Scaffold	ASM1276785v1	NZ_BLRR000000000.1/ BLRR000000000.1	ST9	Allele type 1
FMV2238	Dog	2002	Portugal: Lisbon	Ear	Contig	2238_02	NZ_UXEP000000000.1/ UXEP000000000.1	ST1	Allele type 15

NA, not available. Gray shading shows the whole genome sequences that we published on the NCBI database.

¹⁾: SCM allele typing was conducted based on our previous manuscripts' typing (19,20).

Table 2. Comparative POG analysis results using genomes of twenty *S. canis*

Cluster number	GO ID	Ontology	Pathway name	Differentially present POG number	P-value
CDS features specific to invasiveness ¹⁾					
cluster3250	GO:0044826	P	Viral genome integration into host DNA	1	NA
cluster4154	GO:0006094	P	Gluconeogenesis	1	NA
CDS features specific to ST9 ²⁾					
cluster1991	GO:0009307	P	DNA restriction-modification system	5	5.74.E-06
cluster2129	GO:0006313	P	DNA-mediated transposition	4	9.95.E-04
cluster1994	GO:0005576	C	Extracellular region	2	2.84.E-03
cluster4098	GO:0006979	P	Response to oxidative stress	2	5.16.E-03
cluster3960	GO:0005886	C	Plasma membrane	1	NA
cluster1993	GO:0009432	P	SOS response	1	NA
cluster2126	GO:0005524	F	ATP binding	1	NA
cluster2157	GO:0006281	P	DNA repair	1	NA
cluster2174	GO:0015970	P	Guanosine tetraphosphate biosynthetic process	1	NA
cluster2304	GO:0006260	P	DNA replication	1	NA
cluster2626	GO:0004518	F	Nuclease activity	1	NA
cluster2690	GO:0006269	P	DNA replication, synthesis of RNA primer	1	NA
cluster2741	GO:0098996	P	Disruption of host cell glycocalyx during viral entry	1	NA
cluster2742	GO:0030435	P	Sporulation resulting in formation of a cellular spore	1	NA
cluster2984	GO:0006259	P	DNA metabolic process	1	NA
cluster3060	GO:0005737	C	Cytoplasm	1	NA
cluster3159	GO:0015074	P	DNA integration	1	NA
cluster3165	GO:0006364	P	rRNA processing	1	NA
cluster3171	GO:0004252	F	Serine-type endopeptidase activity	1	NA
cluster3934	GO:0044826	P	Viral genome integration into host DNA	1	NA
cluster3942	GO:0046872	F	Metal ion binding	1	NA
cluster3945	GO:0008643	P	Carbohydrate transport	1	NA
cluster4090	GO:0003676	F	Nucleic acid binding	1	NA

Ontology "P" indicates biological process, ontology "F" indicates molecular function, and ontology "C" indicates cellular component. NA, not available.

¹⁾: Invasive strains (TA4, OT1, and FU149) vs. other strains (n = 17).

²⁾: ST9 strains (NCTC 12191^T, FSL Z3-227, NCTC 6198, FU6, TA4, OT1, and FU149) vs. other strains (n = 13).

COG functional categories

- RNA processing and modification
- Chromatin structure and dynamics
- Translation, ribosomal structure and biogenesis
- Transcription
- Replication, recombination and repair
- Cell cycle control, cell division, chromosome partitioning
- Post-translational modification, protein turnover, and chaperones
- Cell wall/membrane/envelope biogenesis
- Cell motility
- Inorganic ion transport and metabolism
- Signal transduction mechanisms
- Intracellular trafficking, secretion, and vesicular transport
- Defense mechanisms
- Extracellular structures
- Nuclear structure
- Cytoskeleton
- Energy production and conversion
- Carbohydrate transport and metabolism
- Amino acid transport and metabolism
- Nucleotide transport and metabolism
- Coenzyme transport and metabolism
- Lipid transport and metabolism
- Secondary metabolites biosynthesis, transport, and catabolism
- General function prediction only
- Function unknown
- Unknown COG

BLAST hit

- ≥ 90 % identical
- = 100 % identical
- ≥ 88 % identical
- ≥ 98 % identical
- ≥ 86 % identical
- ≥ 96 % identical
- ≥ 94 % identical
- ≥ 92 % identical
- ≥ 84 % identical
- ≥ 82 % identical
- ≥ 0 % identical

Clade 1 belonging to ST1

Clade 2 belonging to ST44/2

Clade 3 belonging to ST13/14 (CC13)

Clade 4 belonging to ST21/15/41

Clade 5 belonging to ST9