

RESEARCH

Open Access

Pangenome analysis and virulence profiling of *Streptococcus intermedius*



Dhiraj Sinha^{1,2†}, Xifeng Sun^{1,2†}, Mudra Khare^{1,2}, Michel Drancourt^{2,3}, Didier Raoult^{2,3} and Pierre-Edouard Fournier^{1,2*†}

Abstract

Background: *Streptococcus intermedius*, a member of the *S. anginosus* group, is a commensal bacterium present in the normal microbiota of human mucosal surfaces of the oral, gastrointestinal, and urogenital tracts. However, it has been associated with various infections such as liver and brain abscesses, bacteremia, osteo-articular infections, and endocarditis. Since 2005, high throughput genome sequencing methods enabled understanding the genetic landscape and diversity of bacteria as well as their pathogenic role. Here, in order to determine whether specific virulence genes could be related to specific clinical manifestations, we compared the genomes from 27 *S. intermedius* strains isolated from patients with various types of infections, including 13 that were sequenced in our institute and 14 available in GenBank.

Results: We estimated the theoretical pangenome size to be of 4,020 genes, including 1,355 core genes, 1,054 strain-specific genes and 1,611 accessory genes shared by 2 or more strains. The pangenome analysis demonstrated that the genomic diversity of *S. intermedius* represents an “open” pangenome model. We identified a core virulome of 70 genes and 78 unique virulence markers. The phylogenetic clusters based upon core-genome sequences and SNPs were independent from disease types and sample sources. However, using Principal Component analysis based on presence/ absence of virulence genes, we identified the *sda* histidine kinase, adhesion protein *LAP* and capsular polysaccharide biosynthesis protein *cps4E* as being associated to brain abscess or broncho-pulmonary infection. In contrast, liver and abdominal abscess were associated to presence of the fibronectin binding protein *fbp54* and capsular polysaccharide biosynthesis protein *cap8D* and *cpsB*.

Conclusions: Based on the virulence gene content of 27 *S. intermedius* strains causing various diseases, we identified putative disease-specific genetic profiles discriminating those causing brain abscess or broncho-pulmonary infection from those causing liver and abdominal abscess. These results provide an insight into *S. intermedius* pathogenesis and highlights putative targets in a diagnostic perspective.

Keywords: *Streptococcus intermedius*, *Streptococcus anginosus* group, Infection, Virulence, Comparative genomics, Whole genome sequencing, Pangenome analysis

* Correspondence: pierre-edouard.fournier@univ-amu.fr

[†]Dhiraj Sinha, Xifeng Sun and Pierre-Edouard Fournier participated equally to the work.

¹Aix-Marseille University, IRD, AP-HM, SSA, VITROME, IHU Méditerranée Infection, 19-21 Bd Jean Moulin, 13005 Marseille, France

²IHU Méditerranée Infection, Marseille, France

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

Streptococcus intermedius belongs to the *S. anginosus* group (SAG) that also includes *S. constellatus* and *S. anginosus* [1]. It is part of the normal oral cavity and upper respiratory tract floras, as well as those of the gastrointestinal and female urogenital tracts [2–5]. This bacterium was first described by Guthof in 1956 after being isolated from dental abscesses [6]. *S. intermedius* may also cause human infections, usually monomicrobial, including purulent abscesses of the liver, lungs, psoas, spine and/or central nervous system, and infective endocarditis [7]. Over the years, the role of *S. intermedius* in human infections has increasingly been reported. Patients with invasive *S. intermedius* infections were described to cause significantly longer hospital stays and higher mortality than patients with other *S. anginosus* group infections, suggesting that identifying this species might be important for the management of patients [8].

Various putative virulence factors have been described for *Streptococcus intermedius*, among which the ability to form biofilms to protect itself from antibiotics and the host immune system [9], the production of hydrolytic enzymes, including both glycosaminoglycan-degrading enzymes, such as hyaluronidase and chondroitin sulphate depolymerase, and glycosidases, such as α -N-acetylneuramidase (sialidase), β -D-galactosidase, N-acetyl- β -D-glucosaminidase and N-acetyl- β -D-galactosaminidase, which allow *S. intermedius* to grow on macromolecules found in host tissue [10]; a cytotoxin, intermedilysin (*ILY*), that can directly damage host tissues and immune defense cells and participate in bacterial pathogenicity; and the surface protein antigens I/II that are involved in adhesion to fibronectin and laminin, which is an important step in the pathogenesis of endocarditis and abscess formation [11].

The development of high throughput nucleic acid sequencing technologies has enabled observing variations of the genetic repertoire among strains of a given bacterial species. Our present study analysis aimed at describing the genetic diversity and pathogenesis substratum of *S. intermedius*. Twenty-seven genomic sequences from *S. intermedius* strains, including 13 newly sequenced from our laboratory and 14 from public databases, were used for pan-genomic analysis. Predicted genes were compared among strains to determine the size of the core and dispensable gene pools, the pangenome, the gain/loss of putative virulence determinants, and to identify genomic islands.

Accession numbers

The 13 genome sequences determined in this study were deposited in GenBank and their accession numbers are listed in Table 1.

Materials and methods

Extraction and genome sequencing

The genomic DNA (gDNA) of each studied *S. intermedius* strain was extracted in two steps: a mechanical treatment was first performed using acid-washed glass beads (G4649-500 g Sigma) and a FastPrep BIO 101 instrument (Qbiogene, Strasbourg, France) at maximum speed (6.5) for 90 s. following a 2-hour lysozyme incubation at 37 °C, DNA was extracted using an EZ1 biorobot and the EZ1 DNA Tissue kit (Qiagen, Hilden, Germany). The elution volume was 50 μ L. Genomic DNA was quantified using the Qubit assay (Life technologies, Carlsbad, CA, USA).

The gDNAs were sequenced using a MiSeq sequencer with the Paired-End strategy and the Nextera XT library kit (Illumina, Inc, San Diego, CA, USA). The Paired-End library was prepared using input solutions of 1 ng gDNAs. The gDNAs were fragmented at the tagmentation step. Then, limited cycle PCR amplification (12 cycles) completed the tag adapters and introduced dual-index barcodes. After purification on AMPure beads (Life technologies, Carlsbad, CA, USA), the libraries were normalized according to the Nextera XT protocol (Illumina). Normalized libraries were pooled for sequencing on a MiSeq sequencer (Illumina). Automated cluster generation and paired-end sequencing with dual index reads was performed in a single 39-hour run in a 2 \times 250 bp format. The numbers of paired-end reads were summarized in Table 2. The paired-end reads were trimmed and filtered according to the read qualities.

Genome assembly, annotation and comparison

After sequencing, the obtained reads were assembled using the A5 software [12] with default parameters and then contigs were compared to NCBI using BLASTn to remove contaminations. Then, the online tool Fasta dataset joiner (http://users-birc.au.dk/biopv/php/fabox/fasta_joiner.php) was used to merge sequences into a single molecule. The Mauve software was used for multiple genomic sequence alignment [13]. Genes were annotated using the Prokka software with default parameters [14] in which the similarity e-value cut-off is 0.000001 and the minimum contig size is 200 bp. This pipeline also includes several other tools like Aragorn for tmRNA detection, Barnap to count rRNAs and Prodigal to identify coding sequences. To estimate the mean level of sequence similarity at the genome level among studied strains, we used the OrthoANI [15] and Genome-to-Genome Distance Calculator (GGDC) [16] softwares, with the following respective threshold values of 95–96 and 70 %.

Phylogenetic analysis

A 16 S rRNA-based phylogenetic analysis of the 27 studied *S. intermedius* strains was performed using the

Table 1 Genomic characteristics of the 27 studied *S. intermedius* strains

	Isolation sources	Size (bp)	CDS	tmRNA	tRNAs	rRNA operons	GC%	Avg length CDS (bp)	Coding %	Plasmid	Prophages	Accession number	Country
G1562	Brain abscess	2,052,377	1954	1	50	4	37.3	926	88.173	no	2	UENI000000000.1	France
G1563	Brain abscess	1,890,941	1814	1	50	4	37.7	911	80.54	no	1	UEND000000000.1	France
G1564	Brain abscess	1,887,926	1857	1	50	3	37.7	886	87.192	no	2	UENF000000000.1	France
G1565	Brain abscess	1,972,424	1905	1	50	3	37.6	905	87.4	no	1	UENG000000000.1	France
G1566	Brain abscess	1,897,154	1848	1	56	4	37.6	904	88.077	no	1	UICY000000000.1	France
G1567	Brain abscess	1,897,330	1849	1	50	4	37.6	903	88.033	no	1	UENA000000000.1	France
G1568	Brain abscess	1,898,090	1850	1	51	3	37.6	903	88.051	no	1	UENB000000000.1	France
TYG1620	Brain abscess	2,006,877	1957	1	61	4	37.6	893	87.107	no	3	AP014880.1	Japan
BA1	Brain abscess	1,965,880	1949	1	64	5	37.7	876	86.87	no	4	ANFT000000000.1	US
FDAARGOS_233	Abdominal abscess	1,914,006	1832	1	61	12	37.7	908	86.927	no	1	CF020433.2	US
G1552	Abdominal abscess	1,942,413	1883	1	51	3	37.6	899	87.182	no	1	UENU000000000.1	France
G1553	Abdominal abscess	1,974,268	1907	1	59	4	37.6	909	87.782	no	2	NZ_UZBH000000000.1	France
G1556	Abdominal abscess	1,850,828	1805	1	56	6	37.8	901	87.87	no	1	UENK000000000.1	France
G1557	Abdominal abscess	1,898,297	1834	1	47	4	37.7	905	87.463	no	2	UENH000000000.1	France
B196	Broncho-pulmonary abscess, septic arthritis, osteomyelitis, pyomyositis	1,996,214	1884	1	61	4	37.6	924	87.175	no	1	CF003857.1	Canada
C270	Broncho-pulmonary abscess	1,960,728	1842	1	61	4	37.6	925	86.876	no	1	CF003858.1	Canada
G1554	Broncho-pulmonary abscess	1,886,668	1782	1	54	4	37.8	924	87.317	no	1	UENC000000000.1	France
631_SCON	Broncho-pulmonary abscess	1,968,891	1856	1	55	3	37.8	930	87.666	no	1	JUZ100000000.1	US
KCOM 1545	Endodontic infection	1,908,201	1835	1	32	6	37.6	910	87.514	no	1	CF012718.1	South Korea
F0413	Endodontic infection	1,921,347	1913	1	62	5	37.6	883	87.891	no	2	AFXO000000000.1	US
G1555	Bone abscess	2,003,390	1971	1	55	4	37.5	888	87.382	no	2	UENE000000000.1	France
LC4	Liver abscess	1,914,382	1832	1	41	3	37.8	918	87.889	no	1	PNRP000000000.1	China
30,309	Liver abscess	1,956,646	1864	1	47	3	37.5	920	87.648	no	1	PNRI000000000.1	China
32,811	Liver abscess	1,971,034	1930	1	47	3	37.7	896	87.751	no	1	PNRH000000000.1	China
ATCC 27,335	NA	1,951,449	1871	1	62	4	37.7	908	87.091	no	1	ATFK000000000.1	US
JTH08	NA	1,933,610	1840	1	68	4	37.7	915	87.108	no	1	AP010969.1	Japan
SK54AJKN0	NA	1,910,718	1832	1	29	3	37.6	917	87.941	no	1	AJKN000000000.1	US

NA not available, bp base pairs, CDS coding sequences, prophages Number of prophages

Table 2 Genome sequencing details of the 13 *S. intermedius* strains from our study

Strain	Extraction	Sequencing data	
	DNA concentration ng/μL	Index	Paired end reads
G1552	8.83	4.1	218,834
G1553	5.06	7.45	397,506
G1554	4.66	2.44	130,063
G1555	12.2	5.74	306,282
G1556	11.67	6.83	364,005
G1557	13.37	7.7	410,636
G1562	2.63	4.62	943,724
G1563	1.03	7.73	1,579,852
G1564	0.31	7.51	1,536,166
G1565	0.7	3.91	798,461
G1566	0.65	3.48	712,214
G1567	1.05	5.42	1,107,754
G1568	0.77	6.19	1,265,488

MEGA 7 software [17]. For constructing the phylogenetic tree, the following options were used: Maximum Likelihood method; Kimura 2-parameter model for substitution model; uniform rates among sites; partial deletion option for gaps/missing data; 1000 bootstrap replicates.

Using genomic sequences and the Roary program [18], a clustered heatmap of core genes was drawn on the basis of the presence/absence approach [18]. We also detected SNPs with the snp-sites program [19] from the core genome alignment and drew a phylogenetic tree with CGEwebface [20].

Virulence factor analysis

Virulence-associated genes were detected by comparing studied genomic sequences with the virulence factor database (VFDB) [21] and sequences described in recent publications [22]. The BLASTp search was performed using the threshold scores reported by Olson et al.: 35 % identity and highest scoring pair length of 50 % [22]. Additionally, we reviewed the literature to identify the proteins involved in interactions with the host [10, 23, 24]. A principal component analysis was performed using the XLSTAT program (Data Analysis and Statistical Solution for Microsoft Excel, Addinsoft, Paris, France 2017) in which the Fisher's least significant difference (LSD, $\alpha = 0.005$) and Pearson's correlation coefficients were used, to detect any association of virulence-associated genes with specific clinical conditions.

Core and pan-genome analysis

Get_homologue [25] was used to reveal orthologous genes among *S. intermedius* strains, using the following

parameters: minimal coverage (-C) 40 %, minimum identity (-S) 50 %, minimum e-value (-E) 1e-05. Sequence similarity searches and clustering of coding sequence (CDS) from the 27 genomes were performed using pairwise BLASTp and OrthoMCL algorithms [26]. Sequential inclusion of all possible combinations of up to 27 strains were simulated and fitted by regression analysis [27] of the amount of conserved genes and of strain-specific genes. This allowed to estimate and extrapolate the sizes of core- and pan-genomes. Roary [18] was also used, with default parameters, to confirm the reliability of the obtained pan-genome analysis results (identity percent ≥ 70 %, coverage ≥ 70 %) and to generate the core genome alignment.

Functional classification of orthologous cluster analysis

The Clusters of Orthologous Groups (COGs) database was used to identify gene functions [28] using BLASTP (E-value $1e^{-03}$, coverage 0.7 and identity percent 30 %).

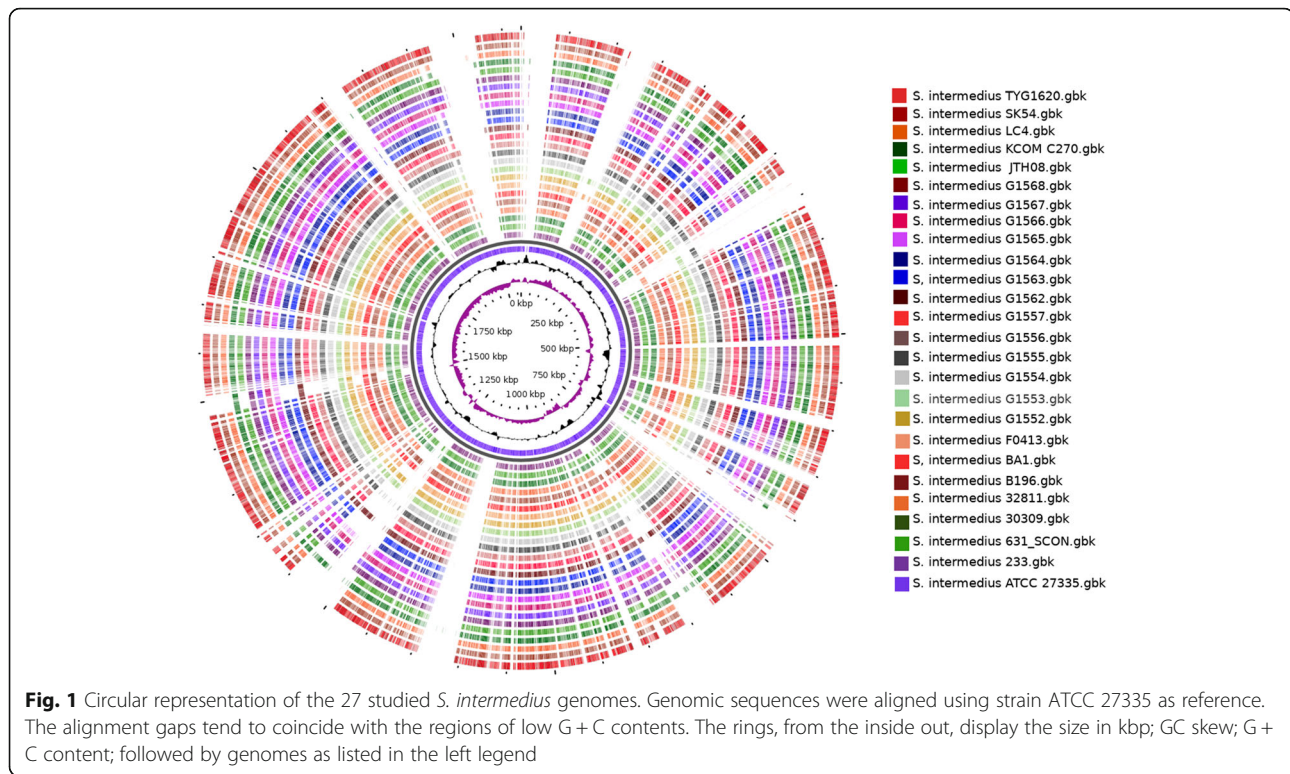
A circular comparison of genomes was obtained using the online GView Server (<https://server.gview.ca/>) with *S. intermedius* strain ATCC 27,335 as reference genome [29]. ResFinder and the ARG-ANNOT database were used to search antibiotic resistance-related markers [30, 31]. The presence of CRISPR repeats and prophages was predicted using the CRISPRFinder [32] and PHAS TER softwares, respectively [33].

Results and discussion

Strain characterization

The 27 studied *S. intermedius* strains originated from China, Canada, South Korea, US, Japan and France. The patients' data was not available for some strains. The 13 French strains (G1552-G1557 and G1562-G1568, Tables 1 and 2) were isolated in our laboratory from patients with various infections (Table 2), from August 2014 to November 2016, on 5 % sheep blood-enriched Columbia agar (BioMérieux) at 37 °C in anaerobic atmosphere. Their identification was confirmed by the high scores (> 2) obtained using MALDI-TOF MS. In addition, 14 *S. intermedius* genome sequences were retrieved from GenBank. The 27 strains were divided into 8 groups according to their isolation source (Table 2). The genome sizes and gene numbers among *S. intermedius* strains were relatively similar, consisting for each strain in a single chromosome but no plasmid was identified in any strains and ranging in size from 1.85Mbp to 2.05Mbp (Table 2).

A schematic view of all 27 studied genomes is provided in (Fig. 1), showing an overall high degree of conservation. The general features of *S. intermedius* genomes are summarized in Table 2. The G + C content of *S. intermedius* ranged from 37.3 to 37.8 % (avg 37.641 %, $n = 27$). All 13 in-house sequenced *S. intermedius*



contained at least 47 tRNA genes, and the number of rRNAs for all strains ranged from 3 to 6. *Streptococcus intermedius* exhibited an average 1870 CDs with a mean length of 907 bp, accounting for 87.3% of the whole genome.

Phylogenetic analysis

The 16 S rRNA-based phylogenetic analysis (Fig. 2), widely used as a gene marker to differentiate *Streptococcus* species [34], demonstrated that all *S. intermedius* strains were grouped in a single cluster that was closely related to *S. anginosus* and *S. constellatus* within the *S. anginosus* group [22] (Fig. 2). In the topology *S. intermedius*, *S. constellatus* and *S. anginosus* strains clustered together with their sub-species. However, the heatmap obtained using Roary [18], based on the core genome, was more discriminatory within the species than the 16 S rRNA-based analysis and identified 3 clusters that were independent from the strain source (Fig. 3).

The three clusters are as follows: strains G1557, G1556, LC4, G1562, SK54, AJKN01, ATCC27335 and JTH08 constituted the first group, strains 30,309, G1563, G1564, 631SC0N and G1554 clustered in the second group while the remaining strains clustered in a third group. There was neither evidence of correlation between strain clusters and their clinical forms, nor between genomic types and the geographical origin of isolates.

To measure the divergence between all studied strains at a deeper level, we also analyzed their phylogenetic relationships on the basis of core genome SNPs, which demonstrated that strains G1562, G1566 and FO413 diverged from other strains and exhibited a higher tendency of recombination. However, again no disease-specific clustering was observed (Fig. 4).

Genomic similarity

Digital DNA-DNA hybridization (dDDH) values ranged from 80.5 to 99.3% between all 27 strains, thus confirming their classification within a single species. This was also cross-validated by the OrthoANI program, which produced pairwise values ranging from 97.78 to 100% which is well above the consensus 95–96% threshold for prokaryotic species demarcation [35]. This corresponded to 100% 16 S rDNA sequence identity across all studied isolates. The above data correlate with a strong degree of genome conservation and synteny.

Functional classification of orthologous cluster

The overall distribution of *S. intermedius* proteins in COG categories was quite similar in all 27 studied strains (Fig. 5). Previous studies of other *Streptococcus* species also suggested that, within a given species, the majority of strains had a similar COG profile [36–38]. Approximately 79.72% of all proteins predicted in all strains were identified in COG superfamilies. The

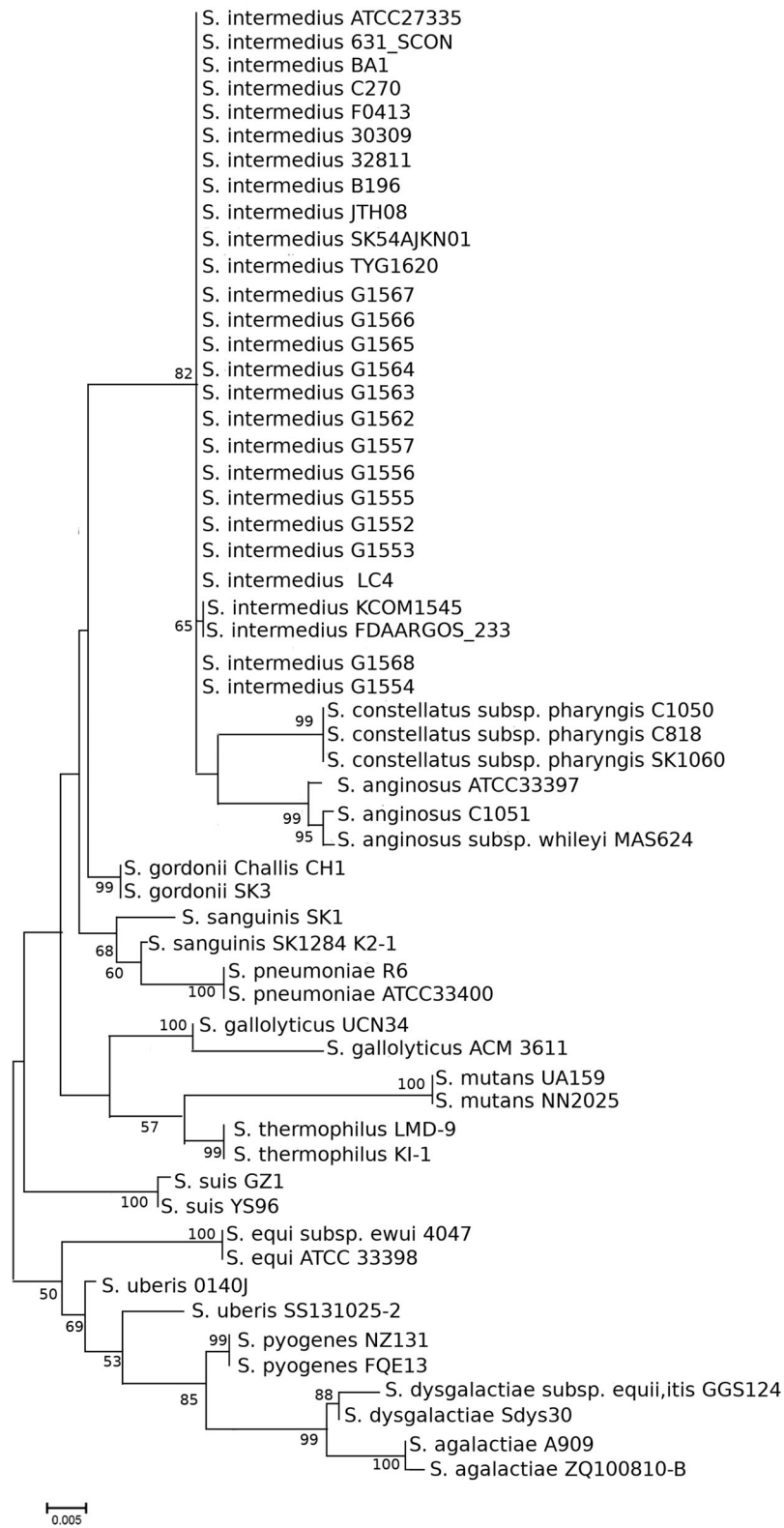
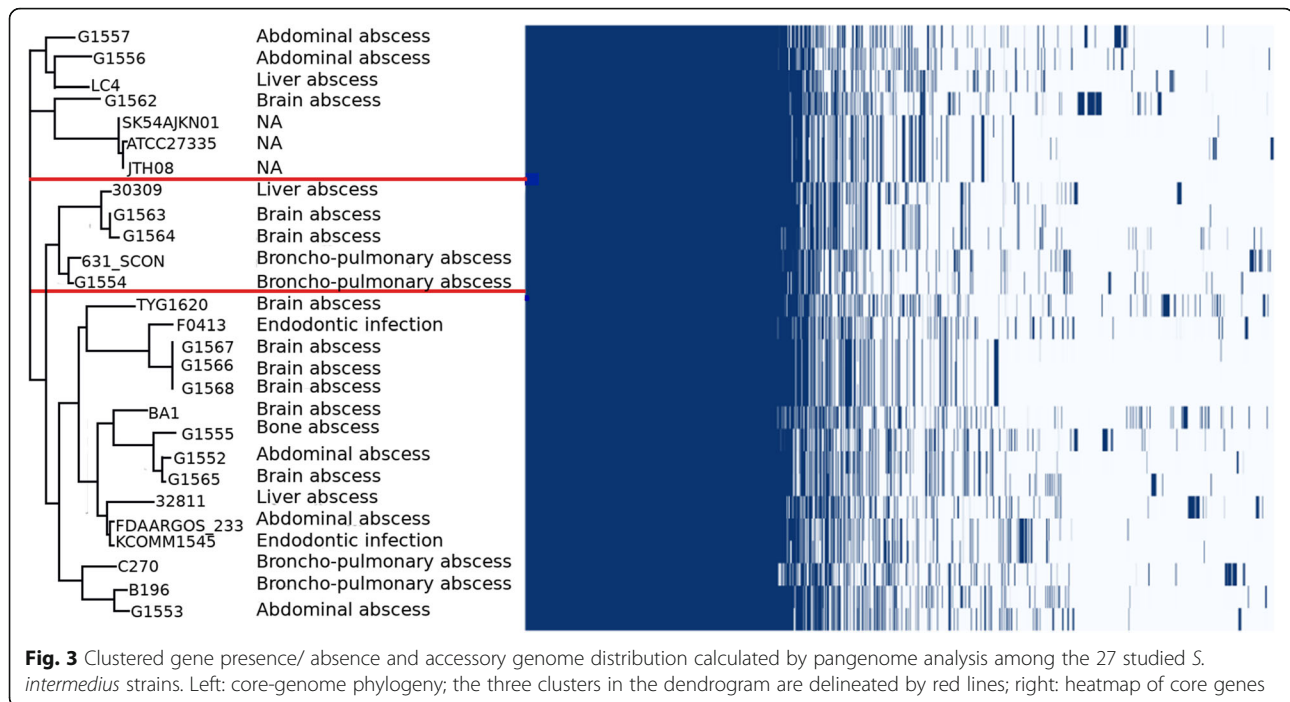


Fig. 2 16S rRNA-based phylogenetic relationships of *S. intermedius* strains using the Maximum Likelihood method with Kimura 2-parameter. The scale bar indicates the evolutionary distance between the sequences determined by a 0.005 substitution per nucleotide position. Numbers at the nodes indicate bootstrap values obtained from 1,000 replicates



proportion of each category fluctuated within a very small range, showing almost similar percentages of distribution in all strains. The most abundant sub-categories were related to carbohydrate transport and metabolism (G) and translation, ribosomal structure and biogenesis (J) like their distribution in core genes.

Less than half of strain-specific genes, but more than 90 % of core genes, had a match in the COGs database. The most abundant functions in core genes were associated with metabolism (Fig. 5a). The overall proportion of metabolic functions in core genes was 32.47 %, whereas that in strain-specific genes was 9.58 %. More specifically, energy production and conversion (C), amino acid transport and metabolism (E), nucleotide transport and metabolism (F), carbohydrate transport and metabolism (G) and coenzyme transport and metabolism (H) were noticeably more abundant in core genes (p -value < 0.01) (Fig. 5b). No mobilome-related functions were detected in *S. intermedius*. The functional category of information storage and processing showed highly different proportions in sub-categories (Fig. 5b). The functions of translation, ribosomal structure and biogenesis (J) were significantly enhanced (p -value < 0.0001) in core genes, whereas the functions of replication, recombination and repair (L) were significantly enhanced (p -value < 0.01) in strain-specific genes. This trend was also observed in other bacteria [35]. In the cellular processing and signaling category, the function of defense mechanisms (V) was more abundant in strain-specific (p -value < 0.05) than in core genes (Fig. 5c).

Pan- and core-genome analyses

The average number of new genes added by a novel genome was 40 when the 27th genome was added (Fig. 6). The exponential decay model shown in Fig. 7a suggests that the number of conserved core genes approached an asymptote with the comparison of 27 genomes. A total of 1,355 core genes were identified in *S. intermedius*. The average proportion and sequence identity of core genes per strain were 72 and 97.79 %, respectively, indicating that core genes in *S. intermedius* are highly conserved and reflecting a low degree of intraspecies genomic variability too. Examination of the functional annotation of these core genes suggests, as expected, that they encode mostly core metabolic processes.

A total of 1,054 strain-specific genes were identified in *S. intermedius* and the average number of strain-specific genes was 39 (Fig. 7b). Among strain-specific genes, 148 genes were found in strain G1562, 107 in strain TYG1620, 105 in strain BA1, 96 in strain 32,811, 82 in strain G1557, 73 in strain C270, 69 in strain 631_SCON, 61 in strain G1555, 41 in strain G1554, 38 each in strains G1565 and F0413, 33 each in strains G1564 and LC4, 27 each in strains G1556 and ATCC27335, 20 in strain 30,309, 16 in strain G1553, 13 in strain G1552, 11 in strain B196, 6 in strain KCOM1545, 5 in strain FDAA RGOS_233 and 1 in strains G1563, G1566, G1567, JTH08, SK54AJKN01, respectively. The size of the pangenome increased steadily without reaching any plateau. The pangenome trend depicted in (Fig. 7b) shows a gradual expansion by addition of new genomes and thus

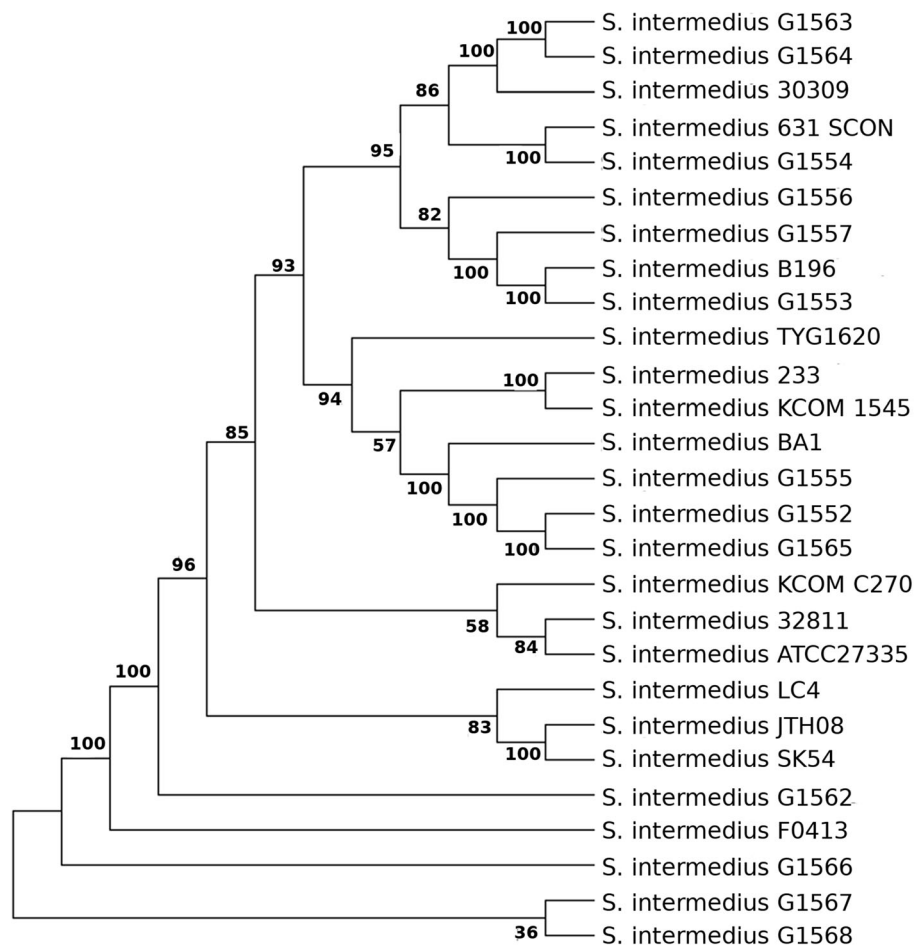


Fig. 4 Phylogenetic tree of *S. intermedius* strains based upon SNPs extracted from the core genome. Sequences were aligned using ClustalW with default parameters and phylogenetic inferences obtained using the Maximum likelihood method within the MEGA, version 7, software. Nodes indicate bootstrap support from 1000 replicates.

the pangenome of *S. intermedius* may be considered as open and indicates a homogenous pattern of genome evolution with similar rates of gene gain/ loss process across the whole population. In addition, a total of 1,611 accessory genes that were shared by two or more strains were identified. Overall, we identified a *S. intermedius* pangenome of 4,020 genes including 1,355 core genes, 1,054 strain-specific genes and 1,611 accessory genes.

Virulence factors and principal component analysis

In the *S. intermedius* pangenome, 252 virulence factors were identified in total. Of these, 70 core virulence factors were shared by all strains and 78 unique virulence factors were present in one strain each (Table 3). Virulence-associated genes present in all studied genomes included homologous virulence genes that contribute to bacterial avoidance of the immune system, such as *ily* which encodes an intermedilysin, the *lmb*, *pspA*, *pavB/pfbB*, *fss3* genes coding surface proteins, the genes coding the polysaccharide capsule (*cps4A*, *cps4B*,

cps4C, *cps4D*, *cps8D*), the auto-inducer LuxS (*luxS*), the binding proteins (*pavA*, *hitC*, *fbpC*, *psaA*, *mntA*, *clpC*, *fss3*), neuraminidase (*nanA*), hyaluronidase (*hysA*), and heat shock protein B (*htpB*), genes from the *sil* locus known to play a role in quorum-sensing and virulence in *S. pyogenes* (*silA*, *silD*, *silE*, *salX*), genes associated with secretion systems (*lem11*, *lem15*, *sdeC*, *ceg32*, *esxA*, *essC*, *lpg2372*, *lirB*), and genes associated with Mg²⁺ transport proteins (*mgtB*, *mgtC*); the response regulator *CsrR* beta-hemolysin gene (*cylG*), laminin-binding surface protein like *Pac* and invasion protein *inlA* were also present in all strains.

Among these core virulence genes, the surface protein antigen I/II that was demonstrated to play a potential role in *S. intermedius* pathogenesis [39], and human fibronectin and laminin that are supposed to bind to this antigenic protein induce IL-18 release from monocytes [39]; genes from the streptococcal invasion locus (*sil*) are related to enhanced virulence in the SAG group and may contribute to the invasive behavior of *S. intermedius*

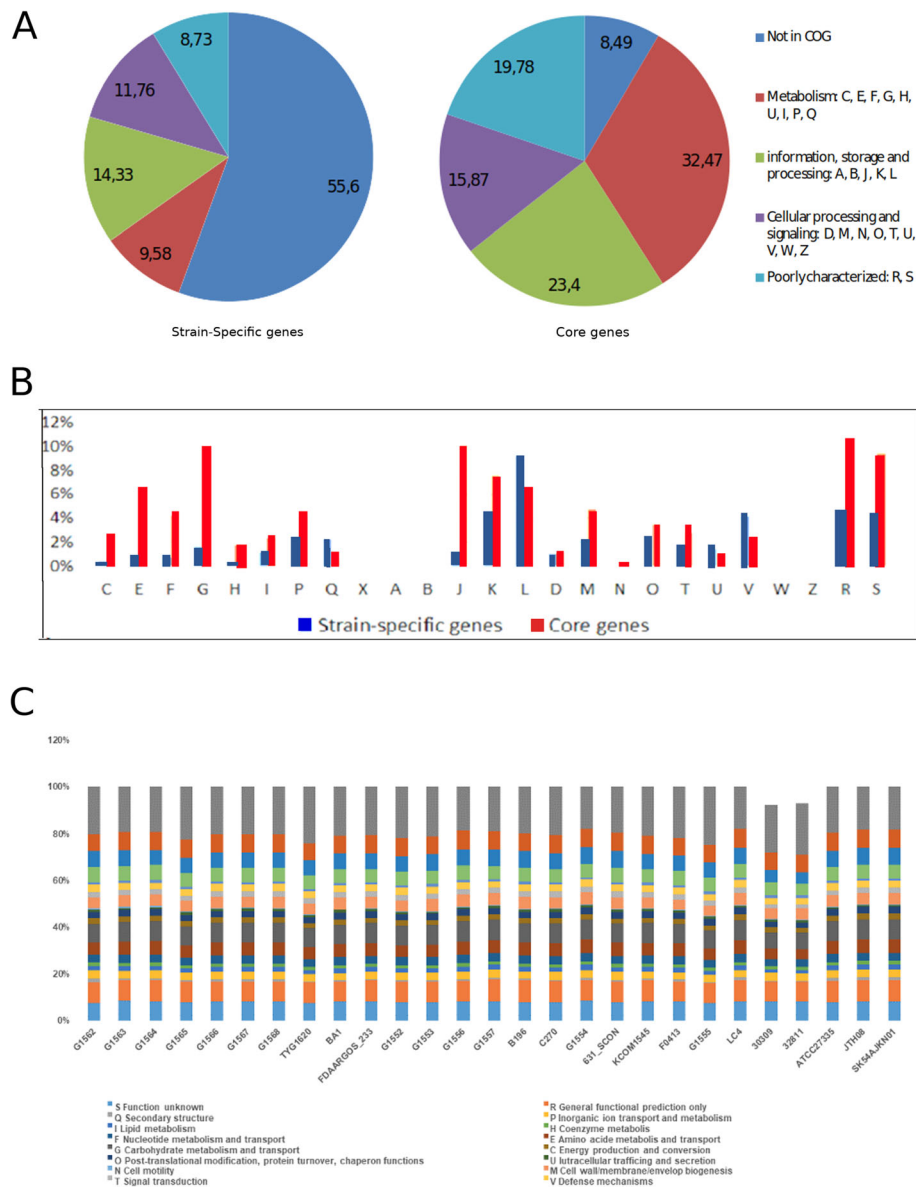


Fig. 5 Differential distribution of COG functional categories in *S. intermedius*: **a** proportion of six classes of functional categories in strain-specific and core genes; **b** functional categories in strain-specific and core genes; **c** functional categories in the 27 *S. intermedius* strains. Category abbreviations are as follows: C, energy production and conversion; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; X, mobilome: prophages, transposons; A, RNA processing and modification; B, chromatin structure and dynamics; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; D, cell cycle control, cell division, chromosomal partitioning; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, posttranslational modification, protein turnover, chaperones; T, signal transduction mechanisms; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; W, extracellular structures; Z, cytoskeleton; R, general function predicted only; S, function unknown

strains; the internalin (*inlA*), likely acquired from *Listeria monocytogenes*, increases the virulence of *S. intermedius* by playing a key role in attachment to host cells [40]; the hyaluronidase (*hysA*) acts in the liquification of tissues and is also involved in biofilm formation, which protects bacteria from host defenses and antibiotics, and plays a role in infection [9]; the *ily*-coded intermedilysin

can directly damage host tissues and immune defense cells, causing human cell death by membrane bleb formation [23]. It has been also reported that intermedilysin helps in invasion and adhesion of bacteria to human liver cells, and in cytotoxicity [41]; the *galE* gene codes galactose which plays a role in biofilm formation and its key residues are essential for epimerase activity [42]; the

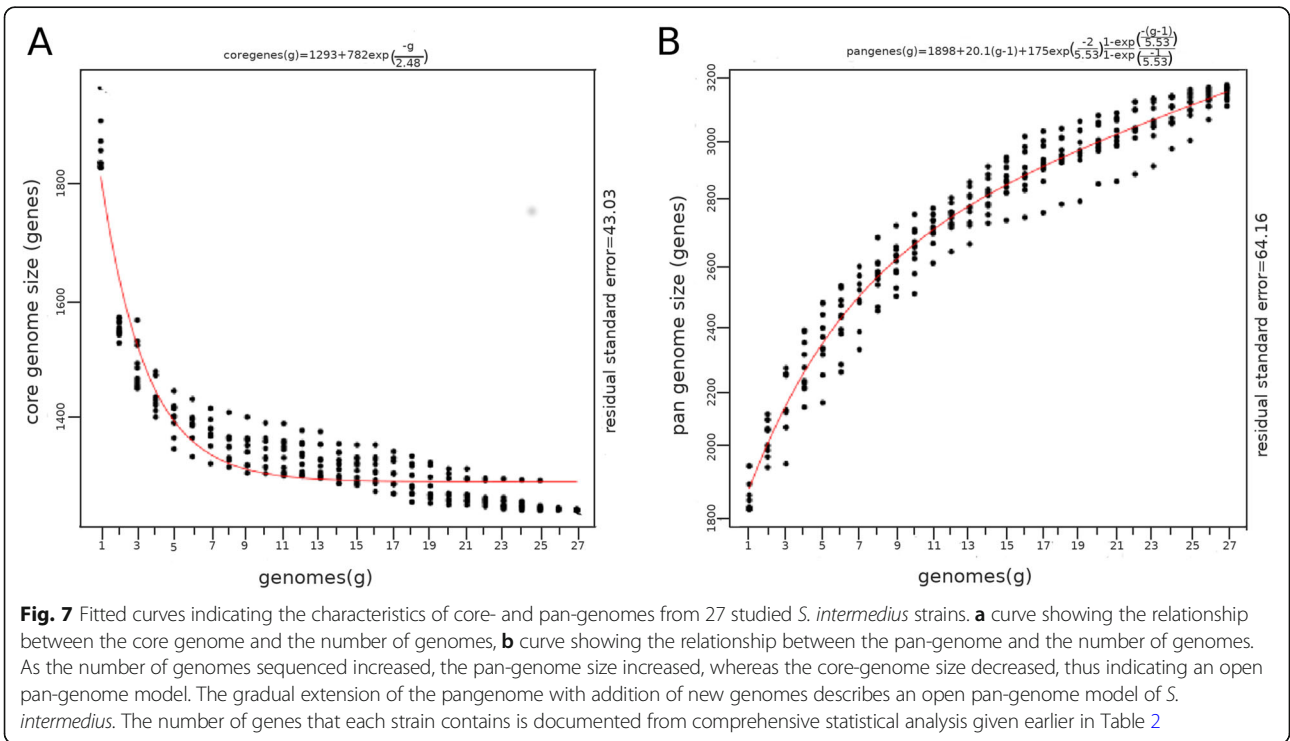
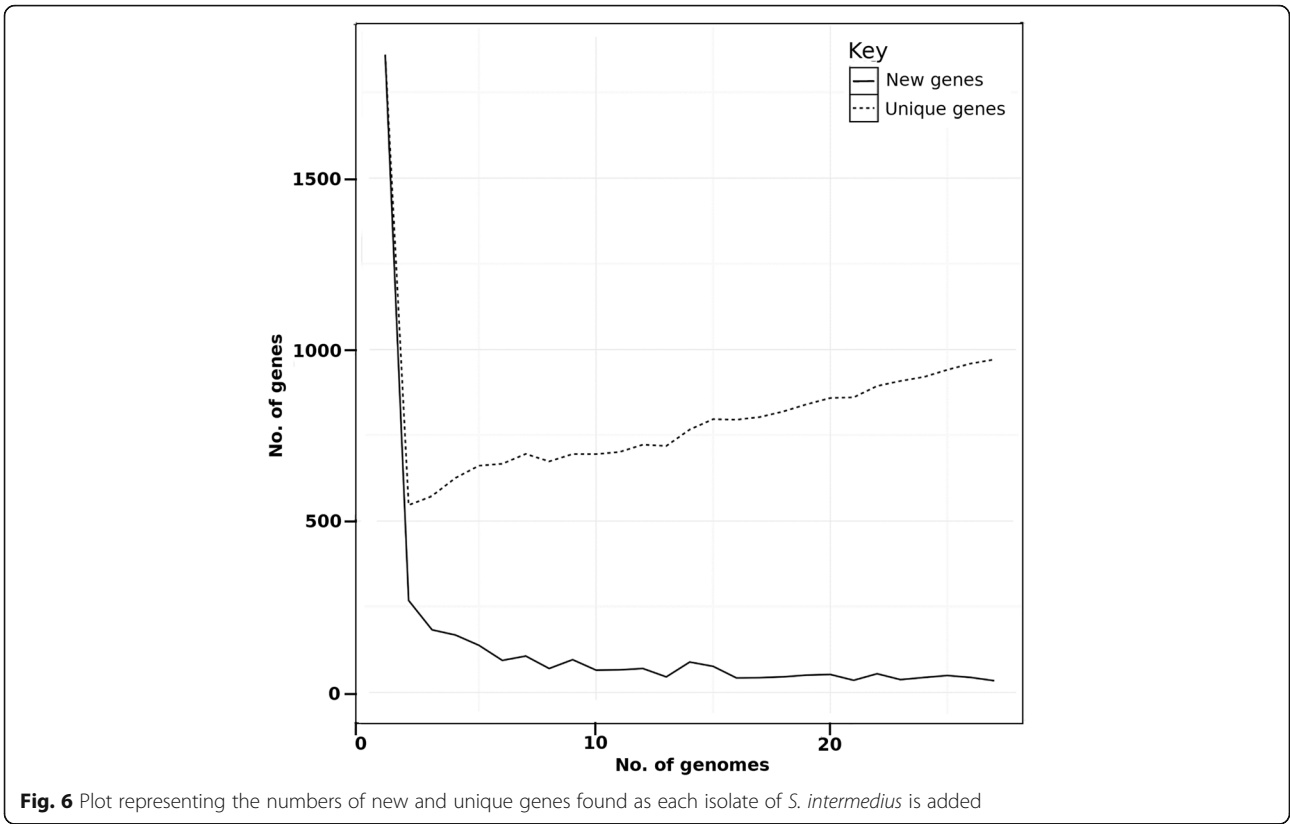


Table 3 Unique Virulence-associated genes detected in the 27 studied *S. intermedius* genomes

Strain	Virulence factor	Amino acid identity% ^a	
G1562	<i>lpg2879</i>	39.286	
	<i>hlyD</i>	46.667	
	<i>sdcA</i>	44.118	
	<i>pvdD</i>	41.509	
	<i>ybtE</i>	40.396	
	<i>SalR</i>	37.879	
	<i>legS2</i>	38.462	
	<i>pieC/lirE</i>	36	
	<i>pvcB</i>	38.776	
	<i>inlK</i>	44.444	
	<i>virB8</i>	38.235	
	<i>sigA</i>	35.294	
	<i>SalK</i>	36.585	
	<i>pvdI</i>	37.092	
	G1563	-	-
	G1564	<i>ctrD</i>	42.308
G1565	<i>aliA</i>	42.5	
	<i>sspH2</i>	38.462	
	<i>espN</i>	35.849	
G1566	-	-	
G1567	-	-	
G1568	-	-	
	<i>ipaJ</i>	38.298	
	<i>FlgG</i>	40.909	
	<i>lpg1803</i>	36.111	
	<i>pilC</i>	39.286	
	<i>brkA</i>	38	
G1552	<i>pilT</i>	35.484	
	<i>lbpA</i>	38.462	
G1553	<i>lpg0365</i>	45.833	
	<i>pvdJ</i>	36.667	
	<i>ravI</i>	42.308	
G1556	<i>vasL</i>	40	
G1554	<i>flgK</i>	35.294	
	<i>vopT</i>	52	
	<i>iraB</i>	46.875	
	<i>hopH</i>	41.935	
G1557	<i>flgI</i>	39.286	
	<i>toxB</i>	37.143	
	<i>SalT</i>	35.294	
	<i>lepA</i>	45.833	
	<i>fcrA</i>	35.938	
	MPN372	35.294	
	<i>bepB</i>	36.364	
	<i>wcbF</i>	35.714	

Table 3 Unique Virulence-associated genes detected in the 27 studied *S. intermedius* genomes (Continued)

Strain	Virulence factor	Amino acid identity% ^a
G1555	<i>lga</i>	35.556
	<i>lpsA</i>	36
	<i>flgJ</i>	36.957
631_SCON	<i>stcE</i>	41.379
	<i>Prt</i>	35
C270	<i>flgD</i>	37.143
	<i>eccE5</i>	36.111
ATCC 27335	<i>lem7</i>	37.736
	EF0818	36.842
JTH08	<i>vpdA</i>	38.462
SK54AJKN01	-	-
	-	-
32811	<i>Tox</i>	37.037
	<i>sipA/sspA</i>	38.235
	<i>fepA</i>	39.13
30309	<i>sdbC</i>	43.243
	<i>lpg2525</i>	50
	<i>flhA</i>	37.838
F0413	<i>legC1</i>	38.235
	<i>tlpB</i>	56.25
	<i>hopZ</i>	47.619
LC4	<i>sidA</i>	42.424
	<i>hifB</i>	45.455
	<i>sipB/sspB</i>	35
KCOM 1545	<i>fliI</i>	36.364
B196	-	-
	-	-
FDAARGOS_233	<i>srtD</i>	37.918
BA1	<i>zmpC</i>	54.839
	<i>recN</i>	42.105
	<i>bsc3</i>	47.826
TYG1620	<i>vscN2</i>	36.364
	<i>fimC</i>	39.13
	<i>mavC</i>	35.593
TYG1620	<i>fliF</i>	40
	<i>tcpB</i>	47.826
	<i>sfbl</i>	37.778
	<i>wcbN</i>	37.778
	<i>pitB</i>	46.429
	<i>lem10</i>	57.143
TYG1620	<i>lapB</i>	35.294
	<i>sopD2</i>	35.484

^aAmino acid identity values were obtained by comparing each gene from each strain to the VFDB database [21]

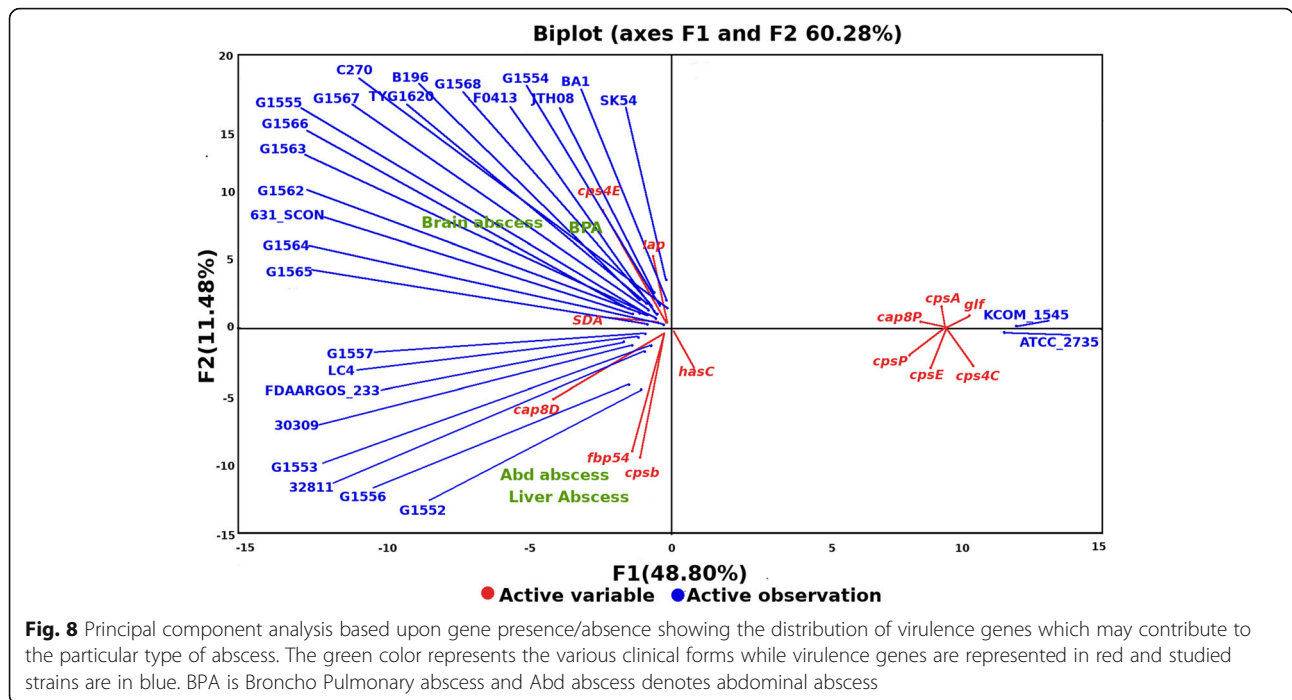
laminin-binding surface protein, homologous to that in *Streptococcus agalactiae* is coded by the *Pac* gene and is essential in binding and invasion of different host surfaces, and is present in almost all group B *Streptococcus* strains causing pneumonia, septicemia and meningitis [43, 44]; *psaA* codes a surface lipoprotein that plays a role in *Streptococcus pneumoniae* systemic infections by interacting with monocytes [45]; we also identified the heat shock protein-coding gene *htbB* that is known in *Legionella pneumophila* to act in adhesion to host fibronectin [46]; the *clpC* gene codes a heat shock protein involved in the invasion of hepatocytes in *Listeria monocytogenes* and has an ATPase activity [47]; ATPase proteins were shown to play a role in the survival and virulence in *Salmonella typhimurium* and *S. aureus* [48]; *clpP* codes an ATP-dependent caseinolytic protease that was proven in *Streptococcus suis* to play a role in colonization and bacterial adaptation to various environmental stresses [49], *pavB* codes a fibronectin-binding protein that mediates bacterial attachment to human epithelial and endothelial cells and also promotes transfer of bacteria to the bloodstream [50, 51]; and *nanaA* codes a highly conserved neuraminidase that also possesses a sialidase activity to catalyze the cleavage of terminal sialic acid residues from glycoconjugates. In *S. pneumoniae*, it promotes biofilm formation and contributes significantly to broncho-pulmonary colonization [52].

Although most of the strains exhibited one to eight unique virulence genes, strains G1562 and BA1 possessed 14 and 10 specific virulence genes, respectively. Eight strains (G1563, G1566, G1567, G1568, BA1, KCOM1545, JTH08, SK54AJKN01) had no strain-specific virulence factor (Table 3).

Among unique virulence genes, *sdca*, *ybtE*, *lpbA*, *SalR*, *salk*, *VopT* are secretory system-associated genes that are involved in iron-mediated transport across cellular membranes. Some of these genes are linked with bacterial growth and act as important anti-inflammatory effectors [42, 53–58]. Among other unique genes, the *pilC* gene is suspected to be essential for secretion and assembly of transcription factor P, important in pilus formation [59] while *pilT* helps in polymerization and depolymerization of pilin [60]. The *brkA* gene inhibits bactericidal activity and protects the bacterium from complement activation products [61]. Other unique genes are linked with bacterial adherence and colonization such as *hopH*, *toxB*, *mpn 372* and *stcE* which contribute significantly to actin organization and bacterial attachment to human surfactant proteins [62–65]. The *iraAB* gene utilizes iron-loaded peptides that promote iron assimilation [66] while *lepA* plays a role in bacterial growth and induces inflammatory response. This gene also plays a key role in pathogenicity in *Pseudomonas aeruginosa* [67]. The *frCA* gene codes a protein

containing receptor domains for immunoglobulins similar to those M-related proteins [68]. Another immunoglobulin-related gene, *aga*, plays a barrier function for mucosal antibodies by cleaving IgA1 [69]. *IpsA* controls transcriptional biogenesis of the cell wall in inositol-derived lipid formation in *Corynebacterium* and *Mycobacterium* species [70]. The *vasL* gene is considered to be component of *vas* genes, associated with the membrane type VI secretion system [71], and *ravL* is presumably activated at low oxygen level and regulates virulence gene expression via *clp* gene [72]. The *lpg0365* codes a lipophosphoglycan that together with other membrane polypeptides, is necessary for *Leishmania* pathogenesis [73]. The *pvdJ* gene is involved in the production of cyclo-dipeptides that may regulate the production of biofilm [74]. In addition, *pvdL* is associated to biosynthesis or uptake of the siderophores pyoverdine and pyochelin that act in the transport of heme and ferrous ions [75], while *pvdD* is involved in the biosynthesis of pyoverdine in *Pseudomonas aeruginosa* [76]. *IpaJ* codes a plasmid antigen involved in demyristoylation of proteins by inducing golgi fragmentation and inhibiting hormone trafficking [77]. *AliA* is associated with nasopharyngeal colonization in *Streptococcus pneumoniae* [78]. The *espN* gene is reported in *Mycobacterium tuberculosis* to play a role in adding an acetyl group to the N-terminus of the *esaT-6* virulence factor [79]. Flagella-related unique genes found in different studied strains include *flgG*, *flgI*, *flgJ* and *flgK* which play a major role in virulence, adhesion and motility. They are mostly involved in flagellum formation and also act as interface with other flagellar proteins [80–83]. The *lnkK* gene was reported in *Listeria monocytogenes* to help avoid autophagy while *virB8* localizes to the inner membrane and is related to the export of alkaline phosphatase to the periplasm [84]. Finally, *sigA* codes a sigma factor linked with galactosidase activity [85].

Using principal component analysis of differentially distributed virulence genes, three distinct clusters were visualized (Fig. 8). A clear separation of virulence genes associated with brain or broncho-pulmonary abscesses (*cps4E*, *sda* and *lap*) from those associated with liver or abdominal abscesses (*cpsB*, *fbp54* and *cap8D*) was observed. The first component which has maximum coverage and represents the largest variation showed that brain abscess-causing strains were associated with genes coding ATP-dependent proteolytic enzymes, which indicates their potential role in abscess formation. Other virulence genes clustered independently, excluding any association with the previous two disease categories. Among virulence genes associated to brain and broncho-pulmonary infections, *sda* codes an histidine kinase that regulates sporulation initiation in *Bacillus subtilis* and mediates the expression of virulence-associated factors [86]; *lap* codes the *Listeria* adhesion protein (*LAP*) that is a host



stress response protein responsible for adhesion and promotion of translocation across monolayers [87]; and *cps4E* codes the capsular polysaccharide biosynthesis protein that was demonstrated in *S. pneumoniae* to prevent phagocytosis by forming an inert shield essential for encapsulation [88, 89].

In *S. pyogenes*, *fbp54* codes a fibronectin-binding protein that acts as an immunogen in humans. The amino acid sequence of *fbp54* in *S. intermedius* is similar to that of *S. pneumoniae*. *cap8D* codes a dehydratase that is essential for the synthesis of the capsule precursor involved in adhesion. It has also been targeted as component for vaccine development [85, 90]; *cpsB* code capsular polysaccharide biosynthesis proteins that are essential for encapsulation in *S. pneumoniae* and are involved in the interaction of bacteria with their environment, notably their host organism [91];

In contrast to the above-mentioned genes, some were not found to be disease-specific. These included *glf*, *cpsE*, *cpsI*, *cpsA*, *cps4C*, *cps8P* and *hasC*. The *glf* gene is involved in the biosynthesis of unusual monosaccharide galactofuranose [92]; *cpsE* codes a glycosyl transferase responsible for the addition of activated sugars to the lipid carriers in the bacterial membrane and are essential for encapsulation in *S. pneumoniae* [93]; *cpsI* is essential for the production of high molecular weight capsular polysaccharides [94]; *cpsA* and *cps8P* are necessary for normal cell wall integrity and composition [95]; *cps4C* codes a polysaccharide tyrosine kinase adaptor protein that plays a key role in the regulation of capsule biosynthesis [96]; finally, *hasC* is involved in biosynthesis of

hyaluronic acid capsule biosynthesis encodes glucose-1-phosphate uridylyltransferase [97].

Resistance-related genes and prophages

The tetracycline resistance gene *tetM* was identified in strains G1552, C270, KCOM1545, G1555, LC4, 30,309 and 32,811 whereas *tet32* was identified in strain 631_SCON (Table 4). The macrolide resistance gene *ermB* was detected in strains G1552, C270, G1555 and 30,309. In other strains, no antibiotic resistance gene was identified.

Table 4 Antimicrobial resistance genes of studied *S. intermedius* strains

Strain name ^a	Resistance gene	Phenotype
G1552	<i>erm(B)</i>	Macrolide resistance
	<i>tet(M)</i>	Tetracycline resistance
C270	<i>erm(B)</i>	Macrolide resistance
	<i>tet(M)</i>	Tetracycline resistance
631_SCON	<i>tet(32)</i>	Tetracycline resistance
KCOM 1545	<i>tet(M)</i>	Tetracycline resistance
G1555	<i>erm(B)</i>	Macrolide resistance
	<i>tet(M)</i>	Tetracycline resistance
LC4	<i>tet(M)</i>	Tetracycline resistance
30309	<i>erm(B)</i>	Macrolide resistance
	<i>tet(M)</i>	Tetracycline resistance
32811	<i>tet(M)</i>	Tetracycline resistance

^aStrains in which no resistance marker was detected were not included

A set of prophage elements was identified in all 27 strains (Table 1). In addition, four prophage-like elements were detected in strain BA1, three in strain TYG1620, and two in strains G1562, G1564, G1553, G1557, F0413 and G1555. The major difference in the genome size between all 27 studied strains of *S. intermedius* resided in the phage numbers and this presence of phages also denotes contribution of horizontal gene transfer in the emergence of this species [98].

CRISPR identification analysis

The search for CRISPR elements showed that 14 of the 27 studied genomes contained CRISPRs. Three of these 14 strains (G1564, G1565, 631_SCON) had more than one CRISPR, for a total of 17 CRISPR modules identified in studied strains. The direct repeat (DR) length in

identified CRISPRs ranged from 24 to 36 bp while there was variation in the number of spacers present within each CRISPR. CRISPRs also differed among strains but the DR regions were similar for a given CRISPR element subtype. Based on the type of cas proteins, the CRISPRs of strains G1562, G1563, G1564, G1556, G1554, 631_SCON, 30,309 were subtype I-C CRISPRs; those of strains FDAARGOS_233 and KCOM1545 were subtype II-A CRISPRs; finally, the CRISPRs of strains G1565, G1552, B196, G1555 and 32811 were subtype II-C CRISPRs [93] (Table 5).

Conclusions

In the present study, we reported 13 new clinical isolates of *S. intermedius* and, based upon a combined approach of pangenomics, core-genomics and virulence profiling

Table 5 CRISPR elements found in studied *S. intermedius* strains

Strain	DR length (nt)	Number of spacers	Spacer Length (nt)	CRISPR length (nt)	CRISPR type	DR consensus
G1562	32	23	33-37	1555	CAS-TypeIC	ATTTCAATCCACGCACCCGCGAAGGGTGCGAC
G1563	32	24	33-38	1632	CAS-TypeIC	ATTTCAATCCACGCACCCGCGAAGGGTGCGAC
G1564	32	17	33-38	1166	CAS-TypeIC	ATTTCAATCCACGCACCCGCGAAGGGTGCGAC
	32	6	33-35	429		ATTTCAATCCACGCACCCGCGAAGGGTGCGAC
G1565	36	33	29-31	2215	CAS-TypeIC	GTTTTACAGTACTTAAATCTTGAGAGTACAAAAAC
	36	9	20-30	628		GTTTTACAGTACTTAAATCTTGAGAGTACAAAAAC
G1566	-					
G1567	-					
G1568	-					
TYG1620	-					
BA1	-					
FDAARGOS_233	36	17	29-30	1155	CAS-TypeIIA	GTTTTAGAGCTGTGCTGTTTCGAATGGTCCAAAAAC
G1552	36	23	29-30	1550	CAS-TypeIC	GTTTTACAGTACTTAAATCTTGAGAGTACAAAAAC
G1553	-					
G1556	32	6	33-35	427	CAS-TypeIC	GTCGCACCCCTTCGCGGGTGCGTGATTGAAAT
G1557	-					
B196	36	19	29-30	1288	CAS-TypeIC	GTTTTGTACTCTCAAGATTTAAGTAACGCAAAAAC
C270	-					
G1554	32	44	33-36	2950	CAS-TypeIC	ATTTCAATCCACGCACCCGCGAAGGGTGCGAC
631_SCON	32	1	33	96	CAS-TypeIC	GTCGCACCCCTTCGCGGGTGCGTGATTGAAAT
	24	1	32	79		ATGTACTTTATTTAAGTGAACACT
KCOM 1545	35	1	31	100	CAS-TypeIIA	GTTTTAGAGCTGTGCTGTTTCGAATGGTCCAAAAAC
F0413	-					
G1555	36	38	29-31	2540	CAS-TypeIC	GTTTTACAGTACTTAAATCTTGAGAGTACAAAAAC
LC4	-					
30309	32	27	34-38	1830	CAS-TypeIC	ATTTCAATCCACGCACCCGCGAAGGGTGCGAC
32811	36	5	30	365	CAS-TypeIC	GTTTTACAGTACTTAAATCTTGAGAGTACAAAAAC
ATCC 27335	-					
JTH08	-					
SK54AJKN01	-					

of 27 strains, attempted identification of disease-specific genetic profiles. The comprehensive analysis revealed a genomic variability across strains within the species, although synteny of the core genome was preserved. Our results highlight the importance of surface proteins like *pavB*, *pspA* and *cps4* (polysaccharide-coding proteins) and the binding proteins *psaA*, *pavA*, which are present in all studied strains, in pathogenesis. PCA results suggests two distinct categories of virulence genes, ATP dependent proteolytic virulence genes *cps4E*, *sda* and *lap* that are associated with brain and broncho pulmonary abscess while capsular polysaccharides protein coding genes *cpsB* and *cps8D* are linked with liver and abdominal abscess formation. The fibronectin binding protein coded by *fbp54* is also showing its connection with liver and abdominal abscess formation. A recent study also attempted to determine the pangenome of *S. intermedius*. [99] The SNP-based phylogenetic tree as well as core gene-based tree showed no clustering related to any disease entity in *S. intermedius* strains. The whole study provides a key genetic framework for assessing and understanding the molecular events contributing to *S. intermedius* pathogenesis. However, due to the limited number of studied strains, validation of the role of these virulence factors will require experimental confirmations.

Authors' contributions

D.S. and X.S. performed the genomic analysis while M.K performs PCA analysis and helps in preparing figures and tables, and M.D, D.Ra., P.E.F. and D.S. wrote the paper and designed the study. All authors reviewed the manuscript. The authors read and approved the final manuscript.

Funding

The study was supported by the Méditerranée Infection foundation, the National Research Agency under the program "Investissements d'avenir", reference ANR-10-IAHU-03 and by Région Provence Alpes Côte d'Azur and European funding FEDER PRIM1.

Availability of data and materials

All Studied sequences are available in GenBank under accession numbers as follows:

UENI000000000.1, UEND000000000.1, UENF000000000.1, UENG000000000.1, UICY000000000.1, UENA000000000.1, UENB000000000.1, AP014880.1, ANFT000000000.1, CP020433.2, UENJ000000000.1, NZ_UZBH000000000.1, UENK000000000.1, UENH000000000.1, CP003857.1, CP003858.1, UENC000000000.1, JUZI000000000.1, CP012718.1, AFXO000000000.1, UENE000000000.1, PNRP000000000.1, PNRI000000000.1, PNRH000000000.1, ATFK000000000.1, AP010969.1, AJKN000000000.1.

Declarations

Ethics approval and consent to participate

The study design was validated by the ethics committee of the institut federatif de recherche 48 under reference 13–035.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest in relation to this research.

Author details

¹Aix-Marseille University, IRD, AP-HM, SSA, VITROME, IHU Méditerranée Infection, 19-21 Bd Jean Moulin, 13005 Marseille, France. ²IHU Méditerranée Infection, Marseille, France. ³Aix-Marseille University, IRD, AP-HM, MEPHI, IHU Méditerranée Infection, Marseille, France.

Received: 10 March 2021 Accepted: 22 June 2021

Published online: 09 July 2021

References

1. Whiley RA, Fraser H, Hardie JM, Beighton D. Phenotypic differentiation of *Streptococcus intermedius*, *Streptococcus constellatus*, and *Streptococcus anginosus* strains within the "Streptococcus milleri group." *J Clin Microbiol.* 1990;28:1497–501.
2. Rabe LK, Winterscheid KK, Hillier SL. Association of viridans group streptococci from pregnant women with bacterial vaginosis and upper genital tract infection. *J Clin Microbiol.* 1988;26:1156–60.
3. Whiley RA, Beighton D, Winstanley TG, Fraser HY, Hardie JM. *Streptococcus intermedius*, *Streptococcus constellatus*, and *Streptococcus anginosus* (the *Streptococcus milleri* group): association with different body sites and clinical infections. *J Clin Microbiol.* 1992;30:243–4.
4. Claridge JE, Attori S, Musher DM, Hebert J, Dunbar S. *Streptococcus intermedius*, *Streptococcus constellatus*, and *Streptococcus anginosus* ("Streptococcus milleri group") are of different clinical importance and are not equally associated with abscess. *Clin Infect Dis.* 2001;32:1511–5.
5. Tran MP, Caldwell-McMillan M, Khalife W, Young VB. *Streptococcus intermedius* causing infective endocarditis and abscesses: a report of three cases and review of the literature. *BMC Infect Dis.* 2008;8:154.
6. Guthof O. [Pathogenic strains of *Streptococcus viridans*; streptococci found in dental abscesses and infiltrates in the region of the oral cavity]. *Zentralbl Bakteriol Orig.* 1956;166:553–64.
7. Jacobs JA, Pietersen HG, Stobberingh EE, Soeters PB. *Streptococcus anginosus*, *Streptococcus constellatus* and *Streptococcus intermedius*. Clinical relevance, hemolytic and serologic characteristics. *Am J Clin Pathol.* 1995;104:547–53.
8. Hasegawa N, Sekizuka T, Sugi Y, Kawakami N, Ogasawara Y, Kato K, et al. Characterization of the Pathogenicity of *Streptococcus intermedius* TYG1620 Isolated from a Human Brain Abscess Based on the Complete Genome Sequence with Transcriptome Analysis and Transposon Mutagenesis in a Murine Subcutaneous Abscess Model. *Infect Immun.* 2017;85:e00886-16.
9. Pecharki D, Petersen FC, Scheie AA. Role of hyaluronidase in *Streptococcus intermedius* biofilm. *Microbiology (Reading, Engl).* 2008;154 Pt 3:932–8.
10. Nagamune H, Whiley RA, Goto T, Inai Y, Maeda T, Hardie JM, et al. Distribution of the intermedilysin gene among the anginosus group streptococci and correlation between intermedilysin production and deep-seated infection with *Streptococcus intermedius*. *J Clin Microbiol.* 2000;38:220–6.
11. Goto T, Nagamune H, Miyazaki A, Kawamura Y, Ohnishi O, Hattori K, et al. Rapid identification of *Streptococcus intermedius* by PCR with the *ily* gene as a species marker gene. *J Med Microbiol.* 2002;51:178–86.
12. Coil D, Jospin G, Darling AE. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics.* 2015;31:587–9.
13. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE.* 2010;5:e11147.
14. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30:2068–9.
15. Lee I, Ouk Kim Y, Park S-C, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol.* 2016;66:1100–3.
16. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics.* 2013;14:60.
17. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol.* 2016;33:1870–4.
18. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics.* 2015;31:3691–3.
19. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom.* 2016;2:e000056.

20. Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS ONE*. 2014;9:e104984.
21. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res*. 2005;33(Database issue): D325-328.
22. Olson AB, Kent H, Sibley CD, Grinwis ME, Mabon P, Ouellette C, et al. Phylogenetic relationship and virulence inference of *Streptococcus Anginosus* Group: curated annotation and whole-genome comparative analysis support distinct species designation. *BMC Genomics*. 2013;14:895.
23. Nagamune H, Ohnishi C, Katsura A, Fushitani K, Whiley RA, Tsuji A, et al. Intermedilysin, a novel cytotoxin specific for human cells secreted by *Streptococcus intermedius* UNS46 isolated from a human liver abscess. *Infect Immun*. 1996;64:3093-100.
24. Mishra AK, Fournier P-E. The role of *Streptococcus intermedius* in brain abscess. *Eur J Clin Microbiol Infect Dis*. 2013;32:477-83.
25. Contreras-Moreira B, Vinuesa R. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol*. 2013;79:7696-701.
26. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13:2178-89.
27. Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc Natl Acad Sci USA*. 2005;102:13950-5.
28. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 2003;4:41.
29. Petkau A, Stuart-Edwards M, Stothard P, Van Domselaar G. Interactive microbial genome visualization with GView. *Bioinformatics*. 2010;26:3125-6.
30. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother*. 2012;67:2640-4.
31. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother*. 2014;58:212-20.
32. Grissa I, Vergnaud G, Pourcel C. CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*. 2008;36(Web Server issue):W145-148.
33. Arndt D, Marcu A, Liang Y, Wishart DS. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Brief Bioinformatics*. 2019; 20:1560-7.
34. Thompson CC, Emmel VE, Fonseca EL, Marin MA, Vicente ACP. Streptococcal taxonomy based on genome sequence analyses. *F1000Res*. 2013;2:67.
35. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol*. 2007;57 Pt 1:81-91.
36. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci USA*. 2001;98:4658-63.
37. Holden MTG, Hauser H, Sanders M, Ngo TH, Cherevach I, Cronin A, et al. Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. *PLoS ONE*. 2009;4:e6072.
38. Xu P, Alves JM, Kitten T, Brown A, Chen Z, Ozaki LS, et al. Genome of the opportunistic pathogen *Streptococcus sanguinis*. *J Bacteriol*. 2007; 189:3166-75.
39. Petersen FC, Pasco S, Ogier J, Klein JP, Assev S, Scheie AA. Expression and functional properties of the *Streptococcus intermedius* surface protein antigen I/II. *Infect Immun*. 2001;69:4647-53.
40. Sawyer RT, Drevets DA, Campbell PA, Potter TA. Internalin A can mediate phagocytosis of *Listeria monocytogenes* by mouse macrophage cell lines. *J Leukoc Biol*. 1996;60:603-10.
41. Sukeno A, Nagamune H, Whiley RA, Jafar SI, Aduse-Opoku J, Ohkura K, et al. Intermedilysin is essential for the invasion of hepatoma HepG2 cells by *Streptococcus intermedius*. *Microbiol Immunol*. 2005;49:681-94.
42. Li C-T, Liao C-T, Du S-C, Hsiao Y-P, Lo H-H, Hsiao Y-M. Functional characterization and transcriptional analysis of galE gene encoding a UDP-galactose 4-epimerase in *Xanthomonas campestris* pv. *campestris*. *Microbiol Res*. 2014;169:441-52.
43. Farley MM. Group B streptococcal disease in nonpregnant adults. *Clin Infect Dis*. 2001;33:556-61.
44. Al Safadi R, Amor S, Hery-Arnaud G, Spellerberg B, Lanotte P, Mereghetti L, et al. Enhanced expression of Imb gene encoding laminin-binding protein in *Streptococcus agalactiae* strains harboring IS1548 in scpB-Imb intergenic region. *PLoS ONE*. 2010;5:e10794.
45. Hu D, Wang D, Liu Y, Liu C, Yu L, Qu Y, et al. Roles of virulence genes (PsaA and CpsA) on the invasion of *Streptococcus pneumoniae* into blood system. *Eur J Med Res*. 2013;18:14.
46. Carkaci D, Højholt K, Nielsen XC, Dargis R, Rasmussen S, Skovgaard O, et al. Genomic characterization, phylogenetic analysis, and identification of virulence factors in *Aerococcus sanguinicola* and *Aerococcus urinae* strains isolated from infection episodes. *Microb Pathog*. 2017;112:327-40.
47. Nair S, Milohanic E, Berche P. ClpC ATPase is required for cell adhesion and invasion of *Listeria monocytogenes*. *Infect Immun*. 2000;68:7061-8.
48. Hensel M, Shea JE, Gleeson C, Jones MD, Dalton E, Holden DW. Simultaneous identification of bacterial virulence genes by negative selection. *Science*. 1995;269:400-3.
49. Abeyta M, Hardy GG, Yother J. Genetic alteration of capsule type but not PspA type affects accessibility of surface-bound complement and surface antigens of *Streptococcus pneumoniae*. *Infect Immun*. 2003;71: 218-25.
50. Pracht D, Elm C, Gerber J, Bergmann S, Rohde M, Seiler M, et al. PavA of *Streptococcus pneumoniae* modulates adherence, invasion, and meningeal inflammation. *Infect Immun*. 2005;73:2680-9.
51. Holmes AR, McNab R, Millsap KW, Rohde M, Hammerschmidt S, Mawdsley JL, et al. The pavA gene of *Streptococcus pneumoniae* encodes a fibronectin-binding protein that is essential for virulence. *Mol Microbiol*. 2001;41:1395-408.
52. Wren JT, Blevins LK, Pang B, Basu Roy A, Oliver MB, Reimche JL, et al. Pneumococcal Neuraminidase A (NanA) Promotes Biofilm Formation and Synergizes with Influenza A Virus in Nasal Colonization and Middle Ear Infection. *Infect Immun*. 2017;85:e01044-16.
53. Papazisi L, Frasca S, Gladd M, Liao X, Yogev D, Geary SJ. GapA and CrmA coexpression is essential for *Mycoplasma gallisepticum* cytoadherence and virulence. *Infect Immun*. 2002;70:6839-45.
54. Buchrieser C, Brosch R, Bach S, Guilyoule A, Carniel E. The high-pathogenicity island of *Yersinia pseudotuberculosis* can be inserted into any of the three chromosomal asn tRNA genes. *Mol Microbiol*. 1998;30:965-78.
55. Arutyunova E, Brooks CL, Beddek A, Mak MW, Schryvers AB, Lemieux MJ. Crystal structure of the N-lobe of lactoferrin binding protein B from *Moraxella bovis*. *Biochem Cell Biol*. 2012;90:351-61.
56. Shappo MOE, Li Q, Lin Z, Hu M, Ren J, Xu Z, et al. SspH2 as anti-inflammatory candidate effector and its contribution in *Salmonella* Enteritidis virulence. *Microbial Pathogenesis*. 2020;142:104041.
57. Li M, Wang C, Feng Y, Pan X, Cheng G, Wang J, et al. Salk/SaIR, a Two-Component Signal Transduction System, Is Essential for Full Virulence of Highly Invasive *Streptococcus suis* Serotype 2. *PLoS ONE*. 2008;3:e2080.
58. Kodama T, Rokuda M, Park K-S, Cantarelli WV, Matsuda S, Iida T, et al. Identification and characterization of VopT, a novel ADP-ribosyltransferase effector protein secreted via the *Vibrio parahaemolyticus* type III secretion system 2. *Cell Microbiol*. 2007;9:2598-609.
59. Tønnum T, Freitag NE, Namork E, Koomey M. Identification and characterization of pilG, a highly conserved pilus-assembly gene in pathogenic *Neisseria*. *Mol Microbiol*. 1995;16:451-64.
60. McCallum M, Benlekbir S, Nguyen S, Tammam S, Rubinstein JL, Burrows LL, et al. Multiple conformations facilitate PilT function in the type IV pilus. *Nat Commun*. 2019;10:5198.
61. Barnes MG, Weiss AA. BrkA protein of *Bordetella pertussis* inhibits the classical pathway of complement after C1 deposition. *Infect Immun*. 2001; 69:3067-72.
62. Dossumbekova A, Prinz C, Mages J, Lang R, Kusters JG, Van Vliet AHM, et al. *Helicobacter pylori* HopH (OipA) and bacterial pathogenicity: genetic and functional genomic analysis of hopH gene polymorphisms. *J Infect Dis*. 2006;194:1346-55.
63. Kannan TR, Provenzano D, Wright JR, Baseman JB. Identification and characterization of human surfactant protein A binding protein of *Mycoplasma pneumoniae*. *Infect Immun*. 2005;73:2828-34.
64. Tozzoli R, Caprioli A, Morabito S. Detection of toxB, a plasmid virulence gene of *Escherichia coli* O157, in enterohemorrhagic and enteropathogenic *E. coli*. *J Clin Microbiol*. 2005;43:4052-6.

65. Grys TE, Siegel MB, Lathem WW, Welch RA. The StcE protease contributes to intimate adherence of enterohemorrhagic *Escherichia coli* O157:H7 to host cells. *Infect Immun*. 2005;73:1295–303.
66. Viswanathan VK, Edelstein PH, Pope CD, Cianciotto NP. The *Legionella pneumophila* iraAB Locus Is Required for Iron Assimilation, Intracellular Infection, and Virulence. *Infection and Immunity*. 2000;68:1069–79.
67. Kida Y, Shimizu T, Kuwano K. Cooperation between LepA and PlcH Contributes to the In Vivo Virulence and Growth of *Pseudomonas aeruginosa* in Mice. *Infect Immun*. 2011;79:211–9.
68. Podbielski A, Kaufhold A, Lütticken R. [The vir-regulon of *Streptococcus pyogenes*: coordinate expression of important virulence factors]. *Immun Infekt*. 1992;20:161–8.
69. Pohliner J, Halter R, Beyreuther K, Meyer TF. Gene structure and extracellular secretion of *Neisseria gonorrhoeae* IgA protease. *Nature*. 1987;325:458–62.
70. Baumgart M, Luder K, Grover S, Gätgens C, Besra GS, Frunzke J. IpsA, a novel LacI-type regulator, is required for inositol-derived lipid formation in *Corynebacteria* and *Mycobacteria*. *BMC Biology*. 2013;11:122.
71. Ishikawa T, Sabharwal D, Bröms J, Milton DL, Sjöstedt A, Uhlin BE, et al. Pathoadaptive conditional regulation of the type VI secretion system in *Vibrio cholerae* O1 strains. *Infect Immun*. 2012;80:575–84.
72. Büttner D, Bonas U. Regulation and secretion of *Xanthomonas* virulence factors. *FEMS Microbiol Rev*. 2010;34:107–33.
73. Elhay M, Kelleher M, Bacic A, McConville MJ, Tolson DL, Pearson TW, et al. Lipophosphoglycan expression and virulence in ricin-resistant variants of *Leishmania major*. *Mol Biochem Parasitol*. 1990;40:255–67.
74. Liu Y, Dai C, Zhou Y, Qiao J, Tang B, Yu W, et al. Pyoverdines are essential for the antibacterial activity of *Pseudomonas chlororaphis* YL-1 under low-iron conditions. *Appl Environ Microbiol*. 2021;87:e02840–20.
75. Ochsner UA, Wilderman PJ, Vasil AI, Vasil ML. GeneChip expression analysis of the iron starvation response in *Pseudomonas aeruginosa*: identification of novel pyoverdine biosynthesis genes. *Mol Microbiol*. 2002;45:1277–87.
76. Merriman TR, Merriman ME, Lamont IL. Nucleotide sequence of pvdD, a pyoverdine biosynthetic gene from *Pseudomonas aeruginosa*: PvdD has similarity to peptide synthetases. *J Bacteriol*. 1995;177:252–8.
77. N B, Tg F, Da P, Jm E, Ba W, As S, et al. Proteolytic elimination of N-myristoyl modifications by the *Shigella* virulence factor IpaJ. *Nature*. 2013;496. <https://doi.org/10.1038/nature12004>.
78. Kerr AR, Adrian PV, Estevão S, de Groot R, Alloing G, Claverys J-P, et al. The Ami-AliA/AlIB Permease of *Streptococcus pneumoniae* Is Involved in Nasopharyngeal Colonization but Not in Invasive Disease. *Infect Immun*. 2004;72:3902–6.
79. Amézquita-López BA, Quiñones B, Lee BG, Chaidez C. Virulence profiling of Shiga toxin-producing *Escherichia coli* recovered from domestic farm animals in Northwestern Mexico. *Front Cell Infect Microbiol*. 2014;4:7.
80. Ran Kim Y, Haeng Rhee J. Flagellar basal body flg operon as a virulence determinant of *Vibrio vulnificus*. *Biochem Biophys Res Commun*. 2003;304:405–10.
81. Wu J-J, Sheu B-S, Huang A-H, Lin S-T, Yang H-B. Characterization of flgK gene and FlgK protein required for *H. pylori* colonization—from cloning to clinical relevance. *World J Gastroenterol*. 2006;12:3989–93.
82. Coloma-Rivero RF, Gómez L, Alvarez F, Saitz W, Del Canto F, Céspedes S, et al. The Role of the Flagellar Protein FlgJ in the Virulence of *Brucella abortus*. *Front Cell Infect Microbiol*. 2020;10:178.
83. Hizukuri Y, Kojima S, Yakushi T, Kawagishi I, Homma M. Systematic Cys mutagenesis of FlgI, the flagellar P-ring component of *Escherichia coli*. *Microbiology (Reading)*. 2008;154 Pt 3:810–7.
84. Thorstenson YR, Zambryski PC. The essential virulence protein VirB8 localizes to the inner membrane of *Agrobacterium tumefaciens*. *J Bacteriol*. 1994;176:1711–7.
85. Cocchiari JL, Gomez MI, Riskey A, Solinga R, Sordelli DO, Lee JC. Molecular characterization of the capsule locus from non-typeable *Staphylococcus aureus*. *Mol Microbiol*. 2006;59:948–60.
86. Rowland SL, Burkholder WF, Cunningham KA, Maciejewski MW, Grossman AD, King GF. Structure and mechanism of action of Sda, an inhibitor of the histidine kinases that regulate initiation of sporulation in *Bacillus subtilis*. *Mol Cell*. 2004;13:689–701.
87. Burkholder KM, Bhunia AK. *Listeria monocytogenes* uses *Listeria* adhesion protein (LAP) to promote bacterial transepithelial translocation and induces expression of LAP receptor Hsp60. *Infect Immun*. 2010;78:5062–73.
88. Hostetter MK. Serotypic variations among virulent pneumococci in deposition and degradation of covalently bound C3b: implications for phagocytosis and antibody production. *J Infect Dis*. 1986;153:682–93.
89. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*. 2001;293:498–506.
90. Kawabata S, Kunitomo E, Terao Y, Nakagawa I, Kikuchi K, Totsuka K, et al. Systemic and Mucosal Immunizations with Fibronectin-Binding Protein FBP54 Induce Protective Immune Responses against *Streptococcus pyogenes* Challenge in Mice. *Infect Immun*. 2001;69:924–30.
91. Skov Sørensen UB, Yao K, Yang Y, Tettelin H, Kilian M. Capsular Polysaccharide Expression in Commensal *Streptococcus* Species: Genetic and Antigenic Similarities to *Streptococcus pneumoniae*. *mBio*. 2016;7:e01844–16.
92. Kleczka B, Lamerz AC, van Zandbergen G, Wiese M. Targeted gene deletion of *leishmania major* UDP-galactopyranose mutase leads to attenuated virulence. *Journal of Biological Chemistry*. 2007;VOL. 282:10498–505.
93. Amonov M, Simbak N, Wan Hassan WMR, Ismail S, A Rahman NI, Clarke SC, et al. Disruption of the cpsE and endA Genes Attenuates *Streptococcus pneumoniae* Virulence: Towards the Development of a Live Attenuated Vaccine Candidate. *Vaccines (Basel)*. 2020;8:187.
94. Thurlow LR, Thomas VC, Hancock LE. Capsular Polysaccharide Production in *Enterococcus faecalis* and Contribution of CpsF to Capsule Serospecificity. *Journal of Bacteriology*. 2009;191:6203–10.
95. Nepal B, Myers R, Lohmar JM, Puel O, Thompson B, Van Cura M, et al. Characterization of the putative polysaccharide synthase CpsA and its effects on the virulence of the human pathogen *Aspergillus fumigatus*. *PLoS One*. 2019;14. <https://doi.org/10.1371/journal.pone.0216092>.
96. Whittall JJ, Morona R, Standish AJ. Topology of *Streptococcus pneumoniae* CpsC, a Polysaccharide Copolymerase and Bacterial Protein Tyrosine Kinase Adaptor Protein. *J Bacteriol*. 2015;197:120–7.
97. Ward PN, Field TR, Ditcham WGF, Maguin E, Leigh JA. Identification and Disruption of Two Discrete Loci Encoding Hyaluronic Acid Capsule Biosynthesis Genes hasA, hasB, and hasC in *Streptococcus uberis*. *Infect Immun*. 2001;69:392–9.
98. Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, Relman DA, et al. Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc Natl Acad Sci U S A*. 2011;108 Suppl 1:4547–53.
99. Issa E, Salloum T, Panossian B, Ayoub D, Abboud E, Tokajian S. Genome Mining and Comparative Analysis of *Streptococcus intermedius* Causing Brain Abscess in a Child. *Pathogens*. 2019;8:22.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

