THE ANNALS OF DENTISTRY
University of Malaya
ORIGINAL ARTICLE

Genome Analysis of *Streptococcus gordonii* SK12

Khairuldin AM[1], Ibrahim IK[1], Wakiyuddin SB[1], Wenning Z[1,2], Lesley AO[3], Nicholas SJ[3,4]*, Siew WC[1,2]*

[1] *Faculty of Dentistry, University of Malaya, Kuala Lumpur, Malaysia*
[2] *Genome Informatics Research Laboratory (GIRG), High Impact Research Building (HIR) Building, University of Malaya, Kuala Lumpur, Malaysia*
[3] *Centre for Oral Health Research, School of Dental Sciences, Newcastle University, United Kingdom*
[4] *Genome Solutions Sdn Bhd, Innovation Incubator UM, Research Management & Innovation Complex, University of Malaya, Kuala Lumpur, Malaysia*

ABSTRACT

The gram-positive, mesophilic and non-motile coccus *Streptococcus gordonii* is an important causative agent of infective endocarditis (IE). This pioneer species of dental plaque also causes bacteraemia in immune-supressed patients. In this study, we analysed the genome of a representative strain, *Streptococcus gordonii* SK12 that was originally isolated from the oral cavity. To gain a better understanding of the biology, virulence and phylogeny, of this potentially pathogenic organism, high-throughput Illumina HiSeq *technology* and different bioinformatics approaches were performed. Genome assembly of SK12 was performed using CLC Genomic Workbench 5.1.5 while RAST annotation revealed the key genomic features. The assembled draft genome of *Streptococcus gordonii* SK12 consists of 27 contigs, with a genome size of 2,145,851 bp and a G+C content of 40.63%. Phylogenetic inferences have confirmed that SK12 is closely related to the widely studied strain *Streptococcus gordonii* Challis. Interestingly, we predicted 118 potential virulence genes in SK12 genome which may contribute to bacterial pathogenicity in infective endocarditis. We also discovered an intact prophage which might be recently integrated into the SK12 genome. Examination of genes present in genomic islands revealed that this oral strain might has potential to acquire new phenotypes/traits including strong defence system, bacitracin resistance and collateral detergent sensitivity. This detailed analysis of *S. gordonii* SK12 further improves our understanding of the genetic make-up of *S. gordonii* as a whole and may help to elucidate how this species is able to transition between living as an oral commensal and potentially causing the life-threatening condition infective endocarditis.

Keywords: Dental biofilm, dental plaque, genome analysis, genome annotation, genome assembly, oral bacterial, strain SK12, *Streptococcus gordonii.*

INTRODUCTION

Infective endocarditis (IE) is a disease that affects patients with vascular abnormalities (1). This disease most frequently occurs in patients with intra-cardiac devices, prosthetic heart valves or unpaired cyanotic congenital heart disease. It is characterised as microbial infection of cardiac valves (the endocardium) due to bacteraemia or fungimia (2). The prominent genera associated with IE include

Streptococcus, Staphylococcus and *Enterococcus* (2). *Streptococcus gordonii* are among several species of streptococci that are common agents of IE (3, 4). More recently, *S. gordonii* have been reported to be opportunistic agents causing bacteraemia in immuno-compromised patients. It has been reported that *Streptococcus gordonii* contributes to approximately 40% of cases which involved neutropenia cancer patients (5). It is also one of the pioneer species of bacteria that initiate dental plaque formation (3). It is thought that the ability to colonize early in dental plaque is related to the large number of cell surface adhesin proteins that are produced by *S. gordonii* and mediate adhesion both to salivary pellicle coating the tooth surface and to other oral bacteria (1). The gram-positive, mesophilic and non-motile *S. gordonii* grow in pairs or bead like chains. They belong to the mitis group oral streptococci which are generally considered commensals in the human oral cavity. *S. gordonii* can enter the bloodstream via inflamed gums or other oral tissues, or even just during daily toothbrushing. Normally, bacteraemia is only transient, but occasionally the presence of bacteria in the bloodstream can eventually cause IE. In this study, we sequenced and characterised the *S. gordonii* SK12 genome using a variety of bioinformatics approaches. We aimed to yield a better understanding on the oral bacteria biology, genetics, and pathogenicity in order to further target and combat IE.

MATERIALS AND METHODS

1. Bacterial DNA extraction and sequencing.

   S. gordonii (SK12) was isolated from the oral cavity of a volunteer in Denmark by Kilian and colleagues (6). The genomic DNA was extracted as previously described (7). The SK12 strain was sequenced using Illumina HiSeq 2000 platform (8).

2. Data pre-processing and genome assembly.

   To ensure high quality genomic sequence data, raw sequencing reads generated by Illumina HiSeq machine were quality-checked using PRINSEQ lite version 0.20.3. CLC Genomic Workbench 5.1.5 was utilized to remove the adaptor sequences, low quality reads and sequences. The pre-processed reads were again imported into CLC workbench 5.1.5 for assembly into contigs and scaffolds.

3. Genome annotation

   To predict genes and their functions, we annotated the assembled genome using Rapid Annotation Using Subsystem Technology (RAST) pipeline (9). Using the assembled genome sequence as an input file to the RAST server, the functional elements such as protein-coding genes, tRNA genes and rRNAswere predicted. The functions of the protein-coding genes were also predicted using the RAST subsystem technology.

4. Phylogenetic analysis

   The phylogenetic analysis was performed to determine the taxonomic position and infer the phylogenetic relationship between SK12 and its closely related *Streptococcus* strains/species. 16S rRNA sequences from other *Streptococcus* species/strains were extracted via RNAMMER program (10). Next, the 16S rRNA sequences were aligned using MAFFT software (11). Lastly, the 16S rRNA-based phylogenetic tree was generated using Molecular Evolutionary Genetic Analysis version 5 (MEGA5) software with 1000 bootstraps (12).

5. Prophage and Genomic Island analysis

   PHAST (PHage Search Tool) (13) was used to identify the putative prophages in the genome of SK12. On the other hand, the putative Genomic Islands (cluster of genes of probable horizontal origin) were predicted using the Island Viewer online tool (14) .

6. Virulence factors analysis

   To identify putative virulence genes in the sequenced genome of SK12, BLAST searches were performed on the RAST-predicted protein-coding genes against Virulence Factor Database (VFDB) that stores manually curated known virulence genes from literature (15). The putative virulence genes were predicted based on the protein sequence homology. The virulence profile of SK12 was represented or visualize in a heat map generated using in-house scripts.

RESULTS AND DISCUSSION

Genome characteristics

The assembled genome of *S. gordonii* SK12 consists of 27 contigs, with a contig N50 of 226,260 bp. The size of this sequenced genome was approximately 2,145,851 bp with a G+C content of 40.63% which is similar with the average G+C content of the published *S. gordonii* genomes (16).

Figure 1 describes the subsystem distribution statistics of *S. gordonii* SK12 based on the RAST genome annotation. RAST predicted 2,097 coding sequences (CDSs) and 56 rRNA/tRNAs in the SK12 genome. RAST functional annotation analysis predicted that most of these genes are likely to be involved in basic functions such as those associated with carbohydrates (238 genes), amino acid and derivatives (204 genes), co-factors, vitamins, prosthetic group, pigment (87 genes), DNA metabolism (78 genes), membrane transport (60 genes), RNA metabolism (99 genes). No genes were predicted to be in the functional category of photosynthesis and nitrogen metabolism.

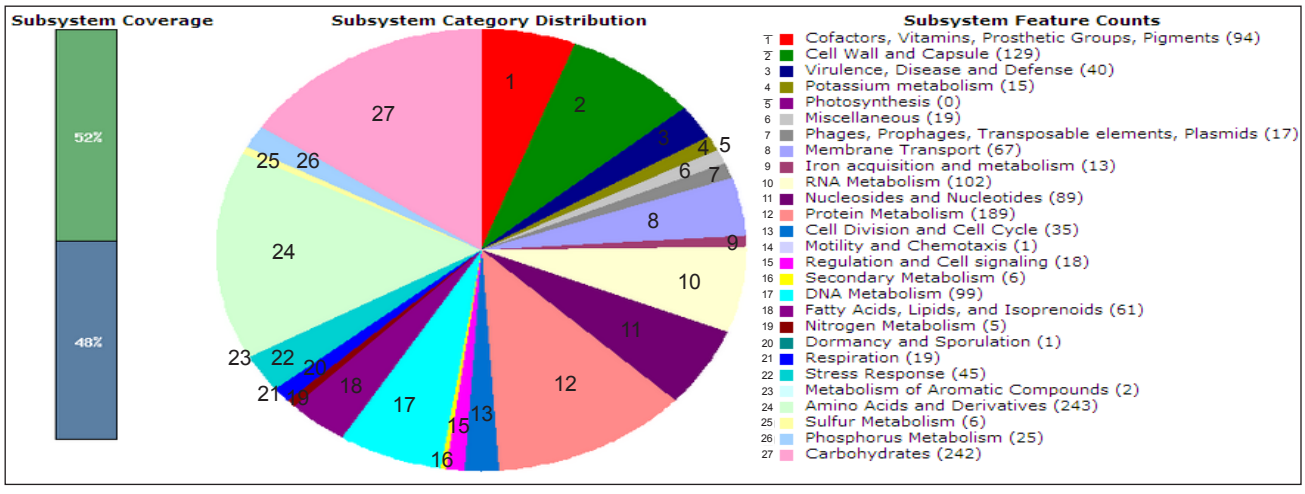

Figure 1: RAST functional analysis. Different functional categories/features were represented by different colors. Numbers in brackets represented the number of genes in the functional category.

TAXONOMIC CLASSIFICATION

To identify the taxonomic position of *S. gordonii* SK12, we reconstructed a neighbour joining phylogenetic tree using 16S RNA gene sequences. The candidates of the tree include 24 species from The National Centre for Biotechnology Information (NCBI). The size of the 16s rRNA was 1,500 base pair and all the sequences were compiled into a file. The sequences were aligned using a Multiple Sequence Alignment Program (MAFFT) (11) and the phylogenetic tree was generated using Molecular Evolutionary Genetics Analysis (MEGA) 6.0 software (12). In general, *S. gordonii* SK12 was clustered into a big clade which included all the *Streptococcus* species (Figure 2). Our data showed that *S. gordonii* Challis was the closest neighbour of our strain SK12.

To further confirm relat ionships between these species, we constructed a more robust core-genome SNP-based tree. To construct this phylogenetic tree, all genome sequences that we used in this analysis were uploaded into the Pan-Genome Sequence Analysis (PANSEQ) for the alignment and SNP identification. The identified SNPs in the core or conserved genomic region among all genomes were extracted and aligned. The SNPs were used to gennerate a phylogenetic tree (Figure 3). Our data clearly showed that *S. gordonii* SK12 was also included in a big clade with the rest of the *Streptococcus* species and the closest neighbour was *S. gordonii* strain Challis substrate CH1. This result was consistent with the result obtained from the 16S gene-based phylogenetic tree.

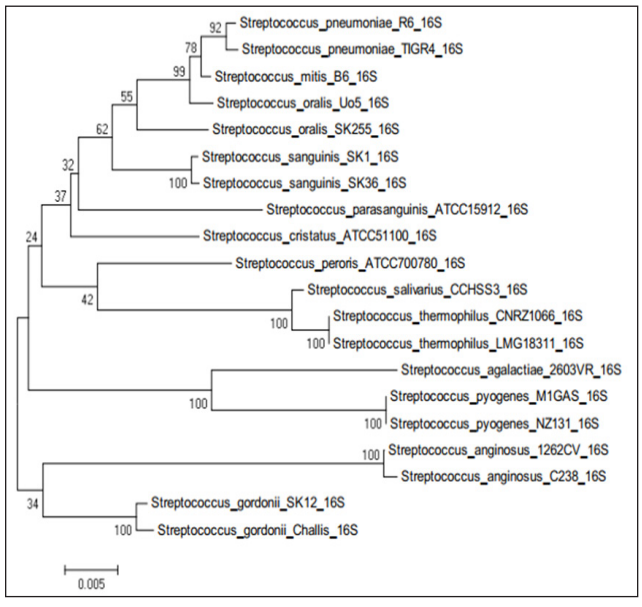

Figure 2: 16S RNA-based phylogenetic tree of oral streptococci. *S. gordonii* SK12 was closely related to *S. gordonii* Challis.

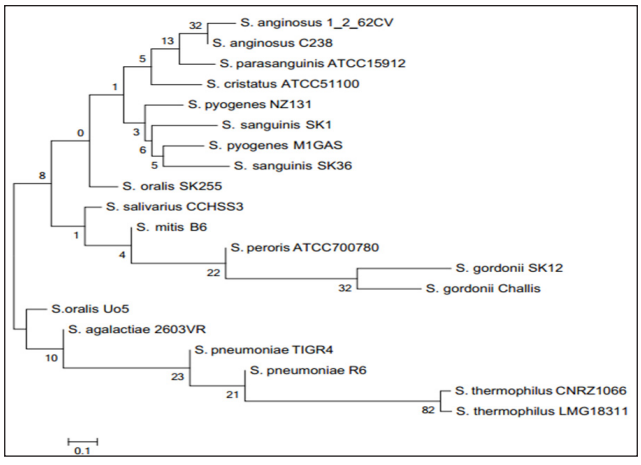

Figure 3: Core-genome SNP-based phylogenetic tree. The closest bacterial species to SK12 was *S. gordonii* Challis.

GENOMIC ISLAND ANALYSIS

Genomic Islands (GIs) are the parts of the genome that have likely arisen from horizontal gene transfer. Horizontal gene transfer is very important in the evolution of bacteria and can influence traits such as antibiotics drug resistance, symbiosis, fitness, and adaption to different environments (17). GIs are usually characterised by their large size (mostly more than 10 Kb), their frequent association with tRNA-encoding genes and a different G+C content compared with the rest of the genome (17). Genomic Islands are postulated to have important roles in bacterial adaptation and pathogenicity as well as other important functions. Identifying the Genomic islands in our strain will provide more insights into the new capability of this bacterium which might be acquired through horizontal gene transfer (18).

Using bioinformatics approach, we found 10 putative GIs in the genome of SK12 (Table 1). Of these, GI 1 contains the gene encoding the bacitracin ABC transporter Bcr. *Bacillus subtilis* cells carrying *bcr* genes which are responsible for bacitracin resistance and collateral detergent sensitivity. These complex mechanisms consist of three main components which are two hydrophobic proteins BcrB and BcrC which presumably form a diffusion channel and two-identical ATP-binding subunits, BcrA. The hydrophobic protein BcrC mediates partial bacitracin resistance by binding to the antibiotic and is responsible for the collateral detergent sensitivity by binding to the detergent, which may induce the disruption of the neighbouring membrane. It has been shown that resistance to bacitracin is closely related to detergent sensitivity. The higher the degree of resistance to bacitracin the more sensitivity it is to detergent. The key point of this process appears to

Table 1: 10 predicted genomic islands in the SK12 genome.

GI	Start	End	Size	GC content (%)	Selected genes in GI
1	800714	804785	4071	33.9	A hypothetical protein, conserved uncharacterized protein, bacitracin ABC transporter, permease protein, ABC transporter, ATP-binding protein, TetR/AcrR family transcriptional regulator
2	864815	872741	7926	37	4 hypothetical proteins, UPF0325 protein YaeH, Adenylosuccinate lyase (EC 4.3.2.2), Transcriptional regulator, TetR family, Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1), Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1), Cadmium efflux system accessory protein
3	1390814	1396804	5990	35.3	6 hypothetical proteins
4	1470629	1475713	5084	35.2	6 hypothetical proteins
5	1528931	1534181	5250	33.8	Transcriptional regulator in cluster with unspecified monosaccharide ABC transport system, M16 family peptidase, Zinc protease, a hypothetical protein, DNA recombination and repair protein RecF
6	1545008	1552372	7364	32.1	Chromosome (plasmid) partitioning protein ParB, Serine protease, DegP/HtrA, do-like (EC 3.4.21.-), LSU m3Psi1915 methyltransferase RlmH, a hypothetical protein, Histidine kinase of the competence regulon

					ComD, Response regulator of the competence regulon ComE, GTP-binding and nucleic acid-binding protein YchF, Peptidyl-tRNA hydrolase (EC 3.1.1.29), Transcription-repair coupling factor
					Phosphoglycolate phosphatase (EC 3.1.3.18), SSU ribosomal protein S10p (S20e), LSU ribosomal protein L3p (L3e), LSU ribosomal protein L4p (L1e), LSU ribosomal protein L23p (L23Ae), LSU ribosomal protein L2p (L8e), SSU ribosomal protein S19p (S15e), LSU ribosomal protein L22p (L17e), SSU ribosomal protein S3p (S3e), LSU ribosomal protein L16p (L10e), LSU ribosomal protein L29p (L35e), SSU ribosomal protein S17p (S11e), LSU ribosomal protein L14p (L23e), LSU ribosomal protein L24p (L26e), LSU ribosomal protein L5p (L11e), SSU ribosomal protein S14p (S29e), zinc-dependent, SSU ribosomal protein S8p (S15Ae), LSU ribosomal protein L6p (L9e), LSU ribosomal protein L18p (L5e), SSU ribosomal protein S5p (S2e), LSU ribosomal protein L30p (L7e), LSU ribosomal protein L15p (L27Ae), Preprotein translocase secY subunit (TC 3.A.5.1.1), Adenylate kinase (EC 2.7.4.3), Translation initiation factor, SSU ribosomal protein S13p (S18e),
7	1579495	1599433	19938	39.3	SSU ribosomal protein S11p (S14e), DNA-directed RNA polymerase alpha subunit (EC 2.7.7.6), LSU ribosomal protein L17p, 5 hypothetical proteins, 2 phage proteins, DNA replication protein, phage-associated, DNA replication protein DnaC
8	1771689	1777547	5858	31.8	CRISPR-associated protein Cas7, CRISPR-associated protein Cas2, CRISPR-associated protein Cas1, CRISPR-associated protein, Csn1 family
9	1879499	1884488	4989	33.3	8 hypothetical proteins
10	2093277	2104069	10792	45	2 phage proteins, Phage transcriptional regulator, Cro/CI family, 2 hypothetical proteins, Phage excisionase, Substrate-specific component PdxU of predicted pyridoxine ECF transporter

be the membrane protein BcrC, which itself alone can provide resistance to bacitracin and simultaneously render the strain sensitive to detergent. The role of BcrA is to provide energy for the transport of bacitracin across the cell membrane, but the presence of BcrA does not contribute to detergent sensitivity (19). The presence of this gene in the horizontally transferred GI1 suggests that SK12 might have acquired the capability for bacitracin resistance and collateral detergent sensitivity.

In GI 2, a gene encoding for a transcriptional regulator of the TetR family and a cadmium efflux system accessory protein were discovered. Transcriptional regulators of the TetR family are involved in the regulation of multidrug efflux pump expression, pathways for biosynthesis of antibiotics, responses to osmotic stress and toxic chemical, control of catabolic pathways, and differentiation processes and pathogenicity (20). Cadmium efflux system accessory protein has also been found in *Listeria monocytogenes* Lm_1889. The putative

function of this protein (based on the predicted molecular functions) is in DNA binding and transcriptional regulation.

In GI 8, a gene encoding CRISPR-associated proteins was detected. There were 8 types of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) associated proteins. CRISPR is believed to participate in the defence against virus in bacteria and archaea. These systems have been found in the genomes of approximately 40% of sequenced bacteria and 90% of sequenced archaea (21). CRISPRs consist of identical repeated DNA sequences (repeats), interspaced by highly variable sequences referred to as spacers. The spacers originate from either phages or plasmids. CRISPR-associated (*cas*) genes encode conserved proteins that together with CRISPRs make-up the CRISPR/Cas system, responsible for defending the prokaryotic cell against invaders. CRISPR-mediated resistance involves three stages: (i) CRISPR-Adaptation, the invader DNA is encountered by the CRISPR/Cas machinery and an invader-derived short DNA fragment is incorporated in the CRISPR array. (ii) CRISPR-Expression, the CRISPR array is transcribed and the transcripts are processed by Cas proteins. (iii) CRISPR-Interference, the invaders' nucleic acid is recognized by complementarity to the crRNA and neutralized. An application of the CRISPR/Cas system is the immunization of industry-relevant prokaryotes (or eukaryotes) against mobile-genetic invasion (22).The presence of these CRISPR-associated genes in the genomic island might suggest that SK12 has capability to prevent invasion of foreign DNA through the CRISPR/Cas system.

VIRULENCE FACTORS

Virulence factors are genes that increase the severity of infections. According to the heat map (Figure 4), we predicted the total number of virulence factors of *S. gordonii* SK12 is 118. No one virulence gene is specific to SK12 as indicated in the heat map (Figure 4). The virulence factors that have homologues in all the other bacteria analyzed are listed in Table 2.

A previous study showed that fibronectin-binding proteins such as *fbp54* and *pavA*, allowing the bacteria to adhere to host cells (23). In invasive streptococci, fibronectin-binding proteins help to trigger endothelial cells to uptake the streptococci via Rac1-dependent phagocytosis which follows the classical endocytic pathway with lysosomal

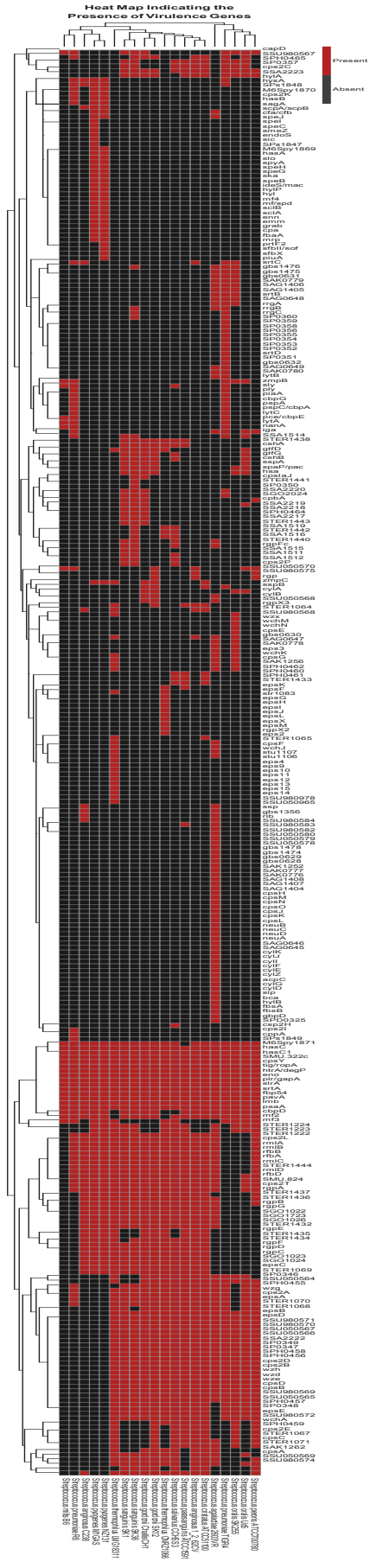

Figure 4: A heat map showing the virulence gene profiles across different species or strains.

Table 2: List of predicted Virulence factors in the SK12 genome.

VIRULENCE FACTOR	FUNCTION
SPs1849	Glucose pyrophosphorylase
M6Spy1871	UTP-glucose-1-phosphate uridyltransferase
HasC	UTP-glucose-1-phosphate uridyltransferase
	putative UDP-glucose pyrophosphorylase
	UDP-glucose pyrophosphorylase
HasC1	UTP-glucose-1-phosphate uridyltransferase 1
SMU.322c	Glucose-1-phosphate uridyltransferase
CpsY	Transcriptional regulator CpsY
tig/ropA	Trigger factor, PPIase
	FKBP-type peptidyl-prdyl cis-trans isomerise
htrA/degP	Serine protease
	Serine peptidase HtrA
	Hypothetical protein
	Trypsin domain protein
	Endopeptidase degP
	Protease Do
	Putative serine protease
	Endopeptidase
Eno	Phosphopyruvate hydratase
	Enolase, putative
	Enolase
plr/gapA	Glyceraldehyde-3-phosphate dehydrogenase
	Glyceraldehyde-3-phosphatendehydrogenase, type 1
	Glyceraldehyde-3-phosphate dehydrogenase, plasmin receptor
	Glyceraldehyde-3-phosphate dehydrogenase, putative
SlrA	Peptidyl-prolyl cis-trans isomerase, cyclophilin-type
	Peptidyl-prolyl cis-trans isomerase
SrtA	Sortase
	Hypothetical protein
	Sortase, putative
	Sortase, truncated
	Putative fimbrial associated sortase-like protein
	Putative fimbrial associated sortase
	Sortase family protein
	Sortase A
	Putative sortase
fbp54	Fibronectin-binding protein A
	Putative fibronectin/fibrinogen-binding protein
	Fibronectin/fibrinogen-binding protein
	Fibronectin-binding protein
	Fibrinogen-binding protein
	Putative fibronectin-binding protein-like protein A
	Putative fibrinogen-binding protein-like protein A
PavA	Adherence and virulence protein A
	Fibronectin/fibrinogen binding protein
	Fibronectin-binding protein-like protein A
	Fibronectin-binding protein A, putative
	Hypothetical protein

destination (23). Additionally, fibronectin binding proteins initiate the process that contribute to deep tissue tropism and lead to invasion of bacteria into vascular endoepithelial lining. Hence, we suggest that in *S. gordonii* SK12 fibrinogen binding proteins may mediate adherence to multiple tissues and play major roles in the pathogenesis of septic arthritis, endocarditis and other infectious diseases (23).

In addition, *cpsY* transcriptional.regulator predicted in SK12 is believed to be associated with systemic infection and is required for survival in neutrophils but not in macrophages. It has been shown that *cpsY* knockout strains have growth defects when cultured in vitro in human plasma. It is also found to regulate the methionine metabolic

pathway in addition to contribute in systemic infection. Furthermore, *cpsY* is believed to be essential for the bacterial invasion and survival in whole blood due to systemic dissemination (24).

Another virulence factor that we found in SK12 is sortase, *srtA*. In *Streptococcus sanguinis*, *SrtA* has been annotated as a putative housekeeping sortase which involved in covalent attachment of the majority of substrates (25). However, sortases of *S. sanguinis* has been found to have a modest effect in competitive colonization at the onset of IE.

PROPHAGE ANALYSIS

As mobile DNA elements, phage DNA is a vector for lateral gene transfer between bacteria (26). The integration of phages into the bacterial genome can bring in new genes to the bacteria which may affect phenotypes such as drug resistance and virulence. We found one putative intact or complete prophage region in the sequenced genome of SK12 (Figure 5). The genomic size of this prophage is about 36564bp starting from 1597011-1633575bp. The prophage is located in contig 8 and has a G+C content of 41.25% which is slightly higher than the average G+C content of the SK12 genome (27).To identify the origin of this prophage, we extracted the prophage sequence and BLASTed it against NCBI nucleotide database. This analysis showed that this putative intact prophage sequence was highly similar to *Streptococcus* phage PH15 with a 93% of sequence identity, suggesting that it has close relationship with this known prophage. From the PHAST prediction results, a few genes were predicted in the prophage (Figure 5 & 6). For instance, a gene encoding for terminase protein was predicted in this intact prophage. This gene exhibits a terminase activity that binds the lambda DNA and proheads and packages the DNA into the prohead. Terminase also has endonuclease activity and cleaves the phage DNA at a specific site known as cos, so that the single genome lengths are packaged into each phage head (28).

Another gene predicted to encode for head protein phage is also present in the predicted prophage. Head proteins markedly contribute to immunological memory to the phage and consist of highly antigenic outer capsid protein and major capsid protein (16). A tail phage protein is also present in the predicted prophage and has a structurally well conserved dodecameric portal at the capsid. The portal plays critical roles in head assembly, genome packaging, neck/tail attachment, and genome ejection. In addition, phage-like protein and hypothetical protein phage are also present. Hypothetical protein phage does not have a specific function and phage-like proteins help the bacteria to survive in harsh environments and to wait for next opportunity to affect other new bacteria.

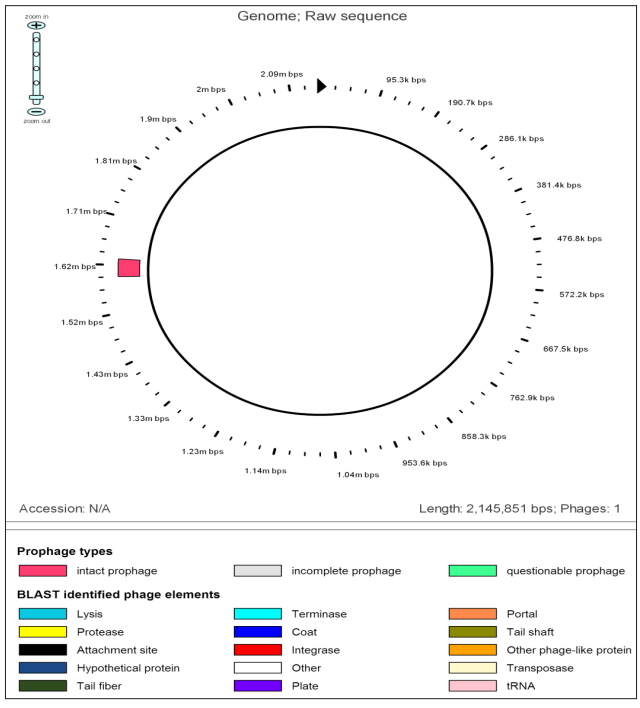

Figure 5: Bacterial genome map. An intact prophage was predicted in the SK12 genome by the PHAST software.

CONCLUSION

Here we report a new genome sequence of *S. gordonii* SK12. As expected, this clinically-derived strain has genome size and G+C content which are consistent with other *S. gordonii* published strains. Our phylogenomic analysis confirms that this strain is indeed *S. gordonii* as it is closely related to the well-studied *S. gordonii* Challis. Through the acquisition of genes through horizontal transfer, SK12 has obtained genes that might give this potential pathogen the ability for bacitracin resistance and collateral detergent sensitivity through ABC transporter. Moreover, this strain might have strong immune/defense system to prevent the invasion of foreign DNA, supported by the presence of multiple CRISPR associated genes in the genomic island. Besides that, SK12 has numerous virulence genes which may explain how this apparent commensal colonizer of the oral cavity is able to cause serious diseases in some circumstances. The addition of this

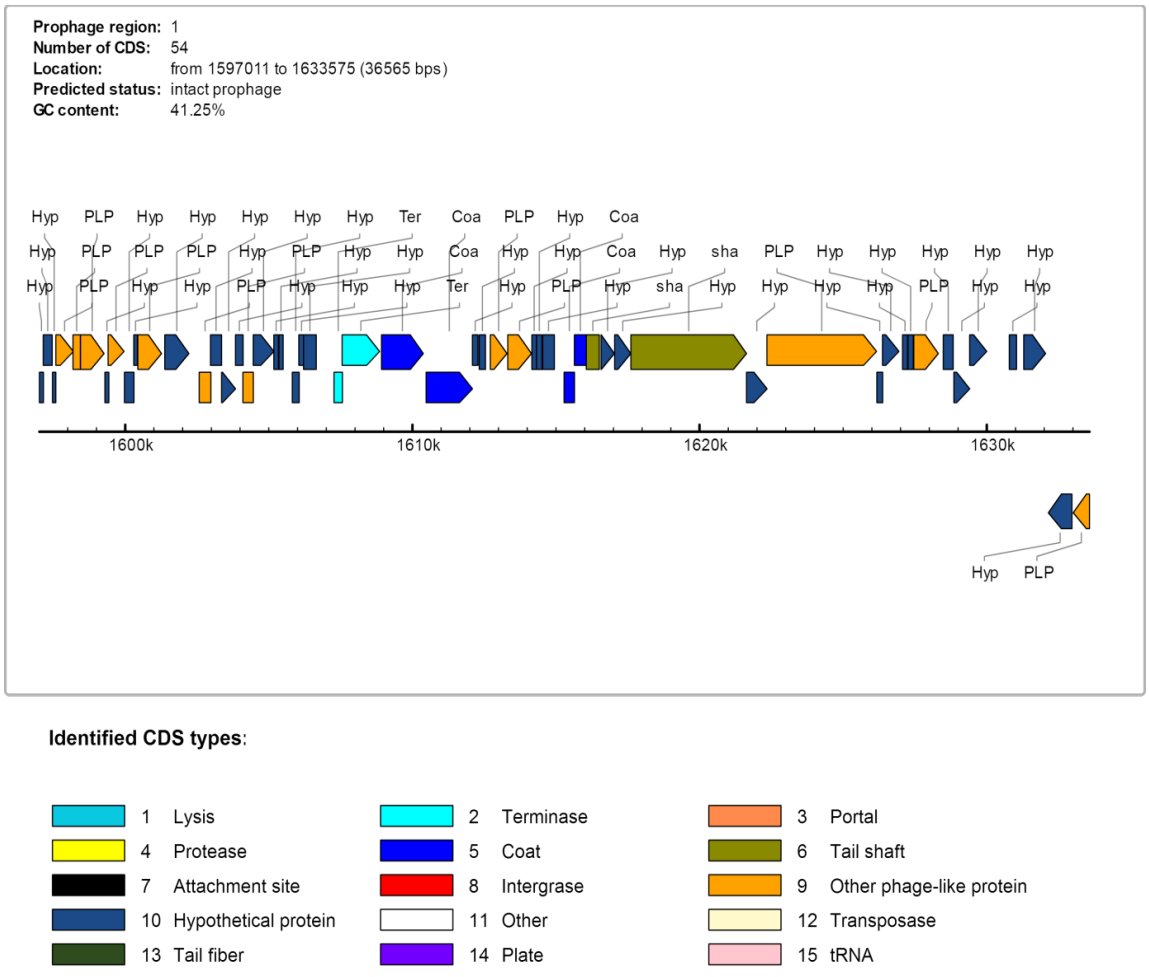

Figure 6: A detailed prophage structure view (linear). Many phage-related genes were predicted in this prophage.

new genome sequence will be useful for the future comparative analysis of *S. gordonii* species which will ultimately lead to a much better understanding of the biology, genetics, pathogenicity, and phylogeny of this important oral microorganism. New insights from our study may contribute to better clinical management of the streptococcal associated diseases.

REFERENCES

1. Tleyjeh IM *et al*. A systematic review of population-based studies of infective endocarditis. CHEST Journal. 2007; 132(3): 1025-1035.
2. Baddour LM *et al*. Infective endocarditis diagnosis, antimicrobial therapy, and management of complications: a statement for healthcare professionals from the committee on rheumatic fever, endocarditis, and Kawasaki disease, council on cardiovascular disease in the young, and the councils on clinical cardiology, stroke, and cardiovascular surgery and anesthesia, American Heart Association: endorsed by the Infectious Diseases Society of America. Circulation, 2005; 111(23): e394-e434.
3. Loo C, Corliss D, and Ganeshkumar N. *Streptococcus gordonii* biofilm formation: identification of genes that code for biofilm phenotypes. Journal of Bacteriology. 2000; 182(5): 1374-1382.
4. Wood AJ and Durack DT. Prevention of infective endocarditis. New England Journal of Medicine. 1995; 332(1): 38-44.
5. Gonzalez-Barca E *et al*. Prospective study of 288 episodes of bacteremia in neutropenic cancer patients in a single institution. European Journal of Clinical Microbiology and Infectious Diseases. 1996; 15(4): 291-296.

6. Kilian M, Mikkelsen L and Henrichsen J. Taxonomic study of viridans streptococci: description of *Streptococcus gordonii* sp. nov. and emended descriptions of *Streptococcus sanguis* (White and Niven 1946), *Streptococcus oralis* (Bridge and Sneath 1982), and *Streptococcus mitis* (Andrewes and Horder 1906). International Journal of Systematic and Evolutionary Microbiology. 1989; 39(4): 471-484.
7. Zheng W *et al*. StreptoBase: An Oral Streptococcus mitis Group Genomic Resource and Analysis Platform. PloS one, 2016; 11(5): e0151908.
8. Patel RK and Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PloS one. 2012; 7(2): e30619.
9. Aziz RK *et al*. The RAST Server: rapid annotations using subsystems technology. BMC genomics. 2008; 9(1): 75.
10. Lagesen K *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research. 2007; 35(9): 3100-3108.
11. Katoh K *et al*. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002; 30(14): 3059-3066.
12. Kumar S, Tamura K and Nei M. MEGA: molecular evolutionary genetics analysis software for microcomputers. Computer applications in the biosciences: CABIOS. 1994; 10(2): 189-191.
13. Zhou Y *et al*. PHAST: a fast phage search tool. Nucleic acids research. 2011: p. gkr485.
14. Langille MG and Brinkman FS. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics. 2009; 25(5): 664-665.
15. Chen L, *et al*. VFDB: a reference database for bacterial virulence factors. Nucleic acids research. 2005; 33(suppl 1): D325-D328.
16. Rouillard JM. and Gulari E. OligoArrayDb: pangenomic oligonucleotide microarray probe sets database. Nucleic acids research. 2009; 37(suppl 1): D938-D941.
17. Navarre WW *et al*. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science. 2006; 313(5784): 236-238.
18. Dobrindt U *et al*. Genomic islands in pathogenic and environmental microorganisms. Nature Reviews Microbiology. 2004; 2(5): 414-424.
19. Podlesek Z *et al*. The role of the bacitracin ABC transporter in bacitracin resistance and collateral detergent sensitivity. FEMS microbiology letters. 2000; 188(1): 103-106.
20. Ramos JL *et al*. The TetR family of transcriptional repressors. Microbiology and Molecular Biology Reviews. 2005; 69(2): 326-356.
21. Grissa I, Vergnaud G and Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC bioinformatics. 2007; 8(1): 172.
22. Al-Attar S *et al*. Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biological chemistry. 2011; 392(4): 277-289.
23. Amelung S *et al*. The FbaB-type fibronectin-binding protein of *Streptococcus pyogenes* promotes specific invasion into endothelial cells. Cellular microbiology. 2011; 13(8): 1200-1211.
24. Allen JP and Neely MN. The *Streptococcus iniae* transcriptional regulator CpsY is required for protection from neutrophil-mediated killing and proper growth in vitro. Infection and immunity. 2011; 79(11): 4638-4648.
25. Turner LS *et al*. Comprehensive evaluation of *Streptococcus sanguinis* cell wall-anchored proteins in early infective endocarditis. Infection and immunity, 2009; 77(11): 4966-4975.
26. Davidson A *et al*. Isolation and characterization of mutations in the bacteriophage lambda terminase genes. Journal of bacteriology. 1991; 173(16): 5086-5096.
27. Dąbrowska K *et al*. Immunogenicity studies of proteins forming the T4 phage head surface. Journal of virology. 2014; 88(21): 12551-12557.
28. Padilla-Sanchez V *et al*. Structure–function analysis of the DNA translocating portal of the bacteriophage T4 packaging machine. Journal of molecular biology. 2014; 426(5): 1019-1038.

Corresponding author:

Siew Woh Choo
Nicholas S Jakubovics
Email: lchoo@um.edu.my
nick.jakubovics@newcastle.ac.uk