The Distribution of Sialic Acid Binding Proteins in *Streptococcus oralis* subsp. *oralis*

Undergraduate Research Thesis

Presented in Partial Fulfillment of the Requirements for Graduation "with Honors Research Distinction in Microbiology" in the Undergraduate Colleges of The Ohio State University

by

Gabriella Matheny

The Ohio State University

May 2021

Project Advisor: Dr. Samantha King, Department of Pediatrics

Contents

Abstract

Streptococcus oralis subsp. *oralis*, a commensal of the human oral cavity, can enter the bloodstream and cause sub-acute infective endocarditis (IE). The dogma is IE develops when bacteria bind platelets on damaged heart valves via sialic acid. Sialic acid binding adhesins found in *S. oralis* subsp. *oralis* are Associated with Sialic acid Adhesion A (AsaA) and Fimbria-Associated Protein 1 (Fap1). Both bind sialic acid via Sialic acid binding immunoglobulin-like lectin (Siglec)-like domains within their non-repeat regions but consist mainly of different repeats. This project's purpose was to determine the distribution, genomic context, and diversity of AsaA and Fap1 with the long-term goal of defining biological consequences of encoding them. Of the 61 *S. oralis* subsp. *oralis* strains used, 12 (20%) encode *asaA*, 35 (57%) encode *fap1*, and 14 (23%) encode neither. No strain contained both *asaA* and *fap1*, even though they are found in distinct but conserved loci. In the *asaA* locus, *asaA* is always accompanied by *zmpB*, but some *asaA* negative strains also contain *zmpB*. It is likely *asaA* and *zmpB* were acquired through the same gene transfer event, as *zmpB* is more conserved in *asaA* positive strains than in *asaA* negative strains. Like the *asaA* locus, the *fap1* locus also has variation in the genes present. However, no correlation exists between genes in the highly variable region downstream of the *fap1* locus and the presence or absence of *fap1*. Alleles of each adhesin are highly conserved and the Siglec-like and Unique domains share predicted structural similarity. Yet, the Siglec-like domains of AsaA and Fap1 share low sequence identity, suggesting AsaA and Fap1 may bind different sialic acid containing glycans. The mechanism of how the IE-isolated strains lacking *asaA* and *fap1* cause IE is unknown. They may encode a novel sialic acid adhesin, bind platelets through a sialic acid-independent mechanism, or cause IE by binding other host components. This study furthered the understanding of the mechanisms *S. oralis*

subsp. *oralis* use to cause IE and provided information that may be valuable when establishing therapeutic targets for treatment.

Introduction

Infective endocarditis (IE) is caused by bacteria adhering to the heart valve. IE cases fall into one of two categories. Acute IE affects previously healthy heart valves, has a serious and rapid onset, and is usually caused by staphylococci. Subacute IE affects previously damaged heart valves, has a slow and progressive onset, and is usually caused by oral streptococci (1, 2). Gram positive cocci, including staphylococci and streptococci, cause 79–90% of all IE cases (3). Without any treatment, IE is lethal. Even with antibiotics and surgical treatment, the average in-hospital mortality is as high as 20%, and after one-year mortality is as high as 40% (4, 5). Depending on the region, the incidence of IE ranges from 1.4 to 12.7 cases per 100,000 people per year (4).

The exact steps in sub-acute IE development are poorly understood but a general mechanism is known. First, bacteria gain access to the bloodstream either through the oral cavity or upper respiratory tract (4). After gaining access to the bloodstream, the bacteria adhere to platelet-fibrin deposits on either mechanically injured or inflamed heart valve surfaces (4, 6). It is unclear if the bacteria bind directly to platelet-fibrin deposits already at the heart valve site or if the bacteria bind to platelets in circulation, which then bring the bacteria to the heart valve (1). Regardless of the exact process, the binding of bacteria to platelets is seen as an integral step.

Some bacteria bind platelets via platelet surface glycoproteins (6). A common terminal carbohydrate of platelet glycoproteins are sialic acids, which are α-keto acids that share a common nine-carbon backbone found on multiple cellular structures (7, 8). Until recently, all known sialic acid-binding adhesins found within Gram-positive cocci were serine-rich repeat proteins (SRRPs) (2, 9). SRRPs are commonly found within oral streptococci, which cause 17% of IE cases (3, 9). These adhesins are found within genomic islands that include genes necessary for SRRP glycosylation and export. The seven core genes present include two glycosyltransferase genes (*gtfA* and *gtfB*) necessary for SRRP glycosylation to ensure protein stability and five genes (*secA2, secY2, asp1–3*) that form a non-canonical Sec translocase, SecY2A2, for SRRP export (1, 9). Species and strain-specific variations exists within the SRRP genomic island. This variation includes the presence of additional glycosyltransferases or additional SecY2A2 accessory proteins (9).

Scanning electron microscopy shows SRRPs form fimbriae-like structures that extent outward from the bacterial surface, allowing for adhesion to host receptors (9). Most SRRPs consist of a N-terminal signal sequence and export-targeting region, a short serine-rich repeat region, a non-repeat binding region, a second larger serine-rich repeat region, and a C-terminal LPxTG cell wall anchoring motif (1, 5, 9). Some SRRPs have a non-repeat region before the first serine-rich repeat region (9). The non-repeat binding regions of SRRPs have a modular organization, meaning the region consists of different domains, allowing this adhesin family to have a broad range of binding targets, including sialic acid on platelet glycoproteins (5).

SRRPs that bind sialic acid do so via a Sialic acid-binding immunoglobulin-like lectin (Siglec)-like domain (5, 7). While different Siglec-like domain-containing SRRPs have different sialic acid linkages as targets, most bind sialic acid through a semi-conserved YTRY sequence motif within the Siglec-like domain (1, 9, 10). Studies show the arginine residue of the YTRY motif is essential for sialic acid binding on platelets, yet some SRRPs bind sialic acid without the arginine (1, 5, 11, 12). Following the Siglec-like domain is a Unique domain, which is thought to affect the Siglec-like domain stability (5, 13). Examples of sialic acid binding SRRPs in oral streptococci are GspB and Hsa from *S. gordonii* strains M99 and DL1, respectively, and SrpA from *S. sanguinis* strain SK36 (6, 7, 9). In animal models of IE, *S. gordonii* strains with mutated *gspB* or *hsa* have reduced vegetation and bacterial load, suggesting the respective adhesin contributes to IE severity (14, 15). A *S. sanguinis* SK36 strain with a *srpA* mutation does not have reduced virulence in animal models even though SrpA *in vitro* has similar binding properties as GspB and Hsa, suggesting the strain has another virulence factor contributing to adhesion (6, 7, 16).

A fourth sialic acid binding SRRP is Fimbriae-Associated Protein 1 (Fap1) from *S. oralis*. The adhesin shares the name with the SRRP found in *S. parasanguinis*, but the two have very little sequence similarity (1). *S. oralis* Fap1 follows the established SRRP structure and appears in the SRRP genomic island, however its genomic island encodes two glycosyltransferase family 8 proteins and two additional SecY2A2 accessory components (Asp4 and Asp5) (1, 9). *S. oralis* is a common and important member of the commensal oral microbiota and has been frequently isolated from IE cases (17, 18, 19). The species is highly variable, usually multiple genotypes are present within the same individual and rarely do unrelated individuals share the same genotype (19). In 2016, the Mitis group of *Streptococcus* was reclassified using whole genome phylogenetic analysis, defining *S. oralis* into three subspecies: *oralis, tigurinus,* and *dentisani* (20). ATCC 10557, the strain in which Fap1 was originally identified, was classified as a *S. oralis* subsp. *oralis* strain (1).

The established paradigm of IE formation was bacteria bind to sialic acid on platelet glycoproteins via SRRPs. However, some SRRP negative *S. oralis* subsp. *oralis* IE isolates could still bind sialic acid. It was discovered that these strains contain the novel sialic acid binding adhesin, Associated with Sialic acid Adhesion A (AsaA). Other IE-causing species contain AsaA orthologs and a rabbit model of *S. oralis* subsp. *oralis* strain IE12 showed that an *asaA* mutation reduces virulence, suggesting AsaA is a virulence factor involved in causing IE in multiple species (2).

AsaA consists of an N-terminal signal sequence and export-targeting region, a non-repeat region, a repeat region of varying number of DUF1542 domains, and a C-terminal LPxTG cell wall-anchoring motif. The non-repeat region contains a found in various architectures (FIVAR) domain and two Siglec-like and Unique domains (2). The function of the FIVAR domain is unknown, but it is also found in Embp of *Staphylococcus aureus* and Ebh of *Staphylococcus epidermidis* (21, 22). Embp and Ebh, which each contain over 50 FIVAR domains, bind fibronectin, an extracellular host matrix component, so it is unlikely AsaA binds fibronectin as it only has a single copy of the FIVAR domain (2, 21, 22). After the FIVAR domain are the Siglec-like and Unique domains. The first Siglec-like domain contains a non-canonical YTRY motif (GTRY) theorized to be required for sialic acid binding in AsaA as it is required in SRRPs (1, 2, 6). The YTRY binding motif was recently revised to be ϕTRX, where ϕ represents W, F, or Y, while X represents Y, T, G, H, or K (12). Even with the broader range of possible binding motifs, the second Siglec-like domain lacks any version of the ϕTRX motif. However, the second Siglec-like domain may still bind sialic acid through interactions involving flexible loops as that is how the Siglec-like domains of SK1, the SRRP in *S. sanguinis*, are proposed to bind as they also lack the conserved YTRY motif (2, 12, 13). The discovery that IE can be caused by SRRP-negative *S. oralis* subsp. *oralis* strains that bind sialic acid via a novel protein shifted the paradigm that all Streptococci bind sialic acid using SRRPs.

Published data indicate that *S. oralis* subsp. *oralis* encodes two sialic-acid binding proteins, AsaA and Fap1. Because of their important role in causing IE, these adhesins may be good targets for IE therapies. Strains with mutant SRRP genes or mutant *asaA* have reduced virulence as they no longer express a sialic acid adhesin (2, 6). Therefore, blocking binding of the adhesins to their ligand may also prevent IE formation. However, the prevalence of *asaA* and *fap1* among different *S. oralis* subsp. *oralis* strains and whether strains contain both or neither gene was unknown. Furthermore, it was unknown if AsaA and Fap1 are functionally equivalent or provide different properties for the bacteria. In this study, we will define the distribution, genomic context, and diversity of the genes encoding sialic acid-binding proteins in *S. oralis* subsp. *oralis*.

Methods

Selecting *S. oralis* subsp. *oralis* strains.

A concatenated protein-based maximum-likelihood phylogenetic tree with 1,000 Bootstrap replicates was created by Dr. Arturo Vera-Ponce de León using the protein sequences predicted by the 92 core genes of the *S. oralis* genome (23, 24, 25, 26, 27). A total of 133 genomes annotated as *S. oralis* as of July 2020 in NCBI were downloaded from the NCBI "RefSeq assembly" to obtain the predicted protein sequences for the phylogenetic tree (28).

We used this tree as a starting point to select *S. oralis* subsp. *oralis* strains. In 2016, the *S. oralis* species was redefined and divided into three subspecies: *oralis*, *dentisani*, and *tigurinus* (20). Strains previously identified as one of the three subspecies were used as markers for where each subspecies fell on the tree. Strains now classified as *S. mitis* or isolated from a non-human source were not further considered. Strains that showed no genetic diversity on the tree (separated by a vertical line) and confirmed to be duplicates by cross checking the strain names in the literature and culture collections were also removed. Lastly, genome sequences containing over 300 contigs were removed as their genomes would be difficult to work with as the loci of interest in this study were split over multiple contigs and proper reconstruction would be almost impossible. Following these criteria, 61 *S. oralis* subsp. *oralis* strains were selected. A maximum-likelihood phylogenetic tree with 1,000 Bootstrap replicates was created by Dr. Arturo Vera-Ponce de León using these 61 strains along with *S. mitis* strain NCTC 12261 (GCF_000148585.2) as an outgroup. He used a total of 940 single copy clusters of orthologous genes to form the tree.

Determining presence of AsaA and Fap1.

Align Sequences Translated Basic Local Alignment Search Tool (tBLASTn) was used to search for AsaA and Fap1 in the 61 *S. oralis* subsp. *oralis* genomes using AsaA QLL99049.1 from *S. oralis* subsp. *oralis* IE12 (SN51445) and Fap1 WP_000466185.1 from *S. oralis* subsp. *oralis* ATCC 10557 (1, 2, 29). These genomes were accessed through NCBI "Nucleotide RefSeq." The presence of each adhesin was verified by confirming the previously defined domains using Signal P 5.0, Pfam, and HHpred (30, 31, 32, 33). tBLASTn was also used to search for SecA2 as a potential marker for the presence of other SRRPs. The SecA2 sequence used was WP_000489782.1 from *Streptococcus pneumoniae* (25).

Characterizing the proteins within the AsaA and Fap1 loci.

Pfam, and HHpred were used to identify the other proteins found in each locus (31, 32, 33). CLUSTAL Omega was used to create multiple sequence alignments, construct phylogenetic trees, and calculate percent identity of these proteins found in the *asaA* and *fap1* loci (34).

MEGA X was used to create phylogenetic trees of *zmp*, using the minimum evolution algorithms with 1000 Bootstrap replicates (35). Following the methodology used in Dr. Mogens Kilian's *zmp* studies, the N-terminal region of *zmp* genes were excluded due to high variability. This resulted in alignment of *zmp* sequences of around 3,600 nucleotides (36). To identify the Zmp protein present within the *asaA* locus, 321 *zmp* sequences were used, 24 *zmp* sequences coming from strains used in this study and 297 *zmp* sequences from *Streptococcus* and *Gemella* strains sent by Dr. Mogens Kilian (36). The 24 *zmpB* sequences from this study were also used to determine the diversity of *zmpB* alleles present in the *asaA* locus.

Boundaries of the Siglec-like and Unique domains of AsaA and Fap1 were established using SWISS-MODEL and mTm-align. SWISS-MODEL was used to create the models of the non-repeat regions of AsaA and Fap1 and mTM-align defined the domain boundaries, using GspB as a template (PDB: 3QC5 and 6EF7) (2, 6, 37, 38, 39). The FIVAR domain of AsaA was also defined using SWISS-MODEL to make a model of the N-terminus of the non-repeat region and mTm-align to define the domain boundaries, using Embp (extracellular matrix binding protein) as a template (PDB: 6GV8) (21, 37, 38, 39).

A potential *asaA* donor was searched for using Standard Nucleotide BLAST (BLASTn). The DNA sequences used were from *S. oralis* subsp. *oralis* IE12 (SN51445) and included *asaA* along with its adjacent gene, *zmpB*, or just *asaA* and *zmpB* separately (28, 29). MEGA X was used to create a minimum evolution phylogenetic tree with 1000 Bootstrap replicates of the non-repeat regions of *asaA* (35). Only the non-repeat region was used as the *asaA* genes are usually split over two contigs in the repeat region.

Results

AsaA and Fap1 are mutually exclusive.

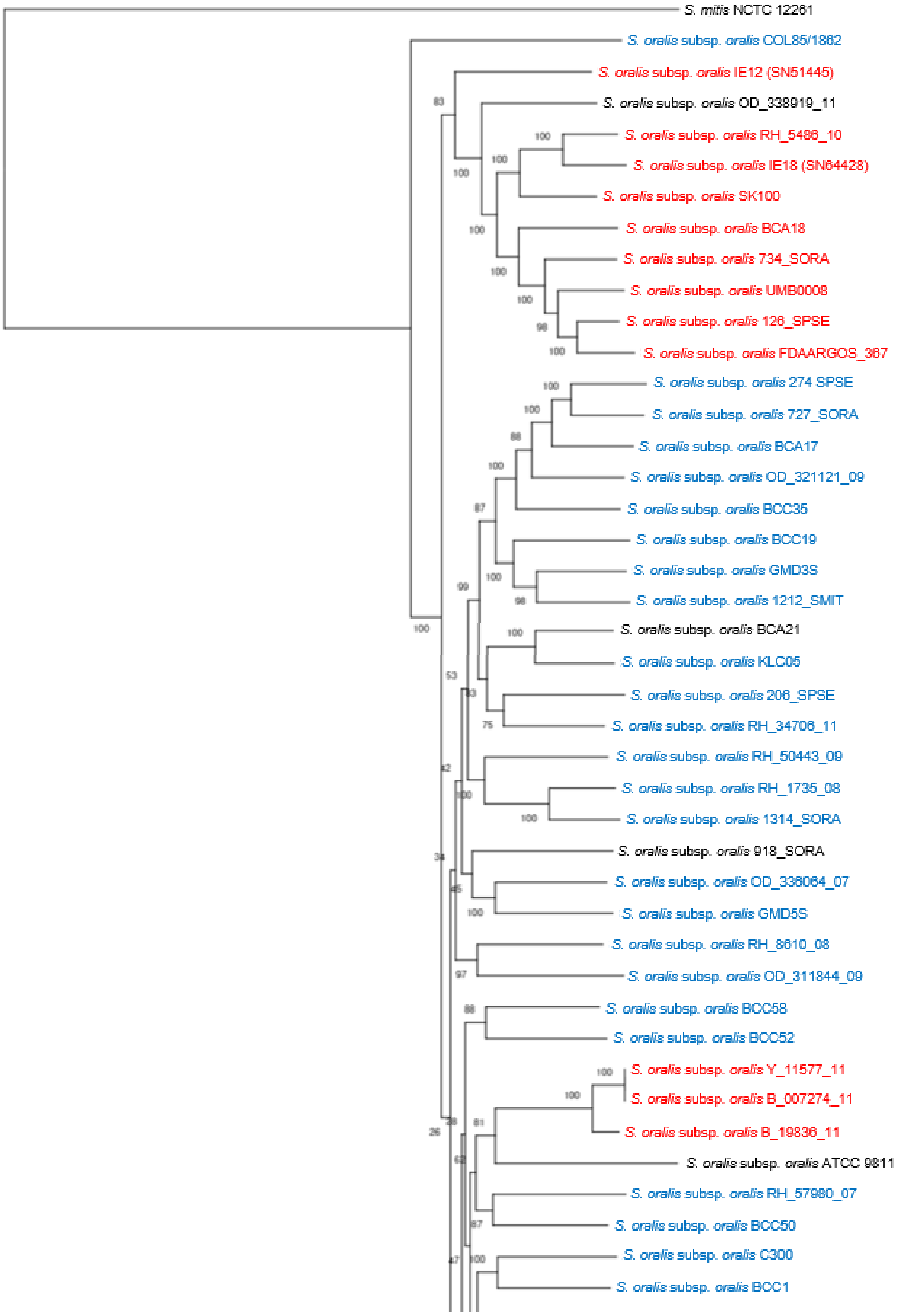
S. oralis subsp. *oralis* strains used in previous studies either contained *asaA* or *fap1*, but it is unknown if strains can have both or if the genes are mutually exclusive (1, 2). Of the 61 *S. oralis* subsp. *oralis* strains used in this study, 12 (20%) contained *asaA*, 35 (57%) contained *fap1*, and 14 (23%) contained neither (Table 1). Therefore, 77% of isolates contained a known sialic acid-binding protein. No strain contained both *asaA* and *fap1*, supporting the hypothesis that these adhesins are mutually exclusive. However, definitive conclusions cannot be made as the sequenced strains may not be an accurate representation of the entire subspecies. For the 22 strains isolated from blood, 77% contained a known adhesin, 7 (32%) containing *asaA* and 10 (45%) containing *fap1*. As of now, it is unknown why *asaA* and *fap1* are mutually exclusive.

A phylogenetic tree of the 61 strains was constructed with *S. mitis* NCTC 12261 as an outgroup by Dr. Arturo Vera-Ponce de León to visualize the distribution of *asaA* and *fap1* (Fig. 1). The *asaA* positive strains have a limited distribution as they fall in two separate locations on the phylogenetic tree. This suggests *asaA* was recently introduced, so a potential donor strain was looked for using BLASTn and *asaA* with and without its adjacent gene, *zmpB*, from *S. oralis* subsp. *oralis* IE12. A donor was not found, but it is possible the donor has not been sequenced. The *fap1* positive strains are more dispersed on the tree with no distinct patterns.

Table 1: Strain table of all *S. oralis* subsp. *oralis* strains used in this study

Strain name	GenBank Assembly Accession No.	*Homo sapiens* isolation site	*asaA*	*fap1*	*secA2*
FDAARGOS_367	GCA_002386345.1	blood	+	-	-
Y_11577_11	GCA_002096435.1	blood, infective endocarditis	+	-	-
B_007274_11	GCA_002096195.1	blood, infective endocarditis	+	-	-
RH_5486_10	GCA_002096205.1	blood, infective endocarditis	+	-	-
B_19836_11	GCA_002096605.1	blood, infective endocarditis	+	-	-
IE12 (SN51445)	GCA_013488045.1	blood, infective endocarditis	+	-	-
IE18 (SN64428)	GCA_013592015.1	blood, infective endocarditis	+	-	-
734_SORA	GCA_001074085.1	bronchioalveolar lavage	+	-	-
UMB0008	GCA_002860885.1	catheter	+	-	-
BCA18	GCA_003942285.1	dental plaque	+	-	-
SK100	GCA_000257475.1	oral cavity	+	-	-
126_SPSE	GCA_001068925.1	wound	+	-	-
OD_311844-09	GCA_002096595.1	blood, infective endocarditis	-	+	+
RH_50443_09	GCA_002096455.1	blood, infective endocarditis	-	+	+
OD_321121_09	GCA_002096575.1	blood, infective endocarditis	-	+	+
RH_57980_07	GCA_002096445.1	blood, infective endocarditis	-	+	+
RH_1735_08	GCA_002096535.1	blood, infective endocarditis	-	+	+
RH_34706_11	GCA_002096525.1	blood, infective endocarditis	-	+	+
OD_332610_07	GCA_002096515.1	blood, infective endocarditis	-	+	+
OD_336064_07	GCA_002096495.1	blood, infective endocarditis	-	+	+
RH_8610_08	GCA_002096375.1	blood, infective endocarditis	-	+	+
ATCC 10557	GCA_013488065.1	blood, infective endocarditis	-	+	+
206_SPSE	GCA_001071755.1	bronchioalveolar lavage	-	+	+
727_SORA	GCA_001074055.1	bronchioalveolar lavage	-	+	+
1212_SMIT	GCA_001070335.1	bronchioalveolar lavage	-	+	+
274 SPSE	GCA_001072035.1	bronchioalveolar lavage	-	+	+
KLC05	GCA_003942785.1	dental plaque	-	+	+
BCA17	GCA_003944395.1	dental plaque	-	+	+
BCC02	GCA_003942985.1	dental plaque	-	+	+
BCC11	GCA_003943875.1	dental plaque	-	+	+
BCC19	GCA_003943835.1	dental plaque	-	+	+
BCC35	GCA_003943805.1	dental plaque	-	+	+
BCC38	GCA_003942675.1	dental plaque	-	+	+
BCC50	GCA_003942635.1	dental plaque	-	+	+
BCC52	GCA_003942875.1	dental plaque	-	+	+
BCC58	GCA_003942805.1	dental plaque	-	+	+
BCC63	GCA_003943825.1	dental plaque	-	+	+
GMD3S	GCA_000298695.2	lower right subgingival plaque	-	+	+
GMD5S	GCA_000298715.2	lower right subgingival plaque	-	+	+
Uo5	GCA_000253155.1	nasal swab	-	+	+
ATCC 35037T	GCA_900637025.1	oral cavity	-	+	+

ATCC 49296	GCA_000185265.1	oral cavity	-	+	+
COL85/1862	GCA_000959945.1	oral cavity	-	+	+
SK143	GCA_000722845.1	oral cavity	-	+	+
SK610	GCA_000257455.1	throat	-	+	+
C300	GCA_000187645.1	upper respiratory tract	-	+	+
1314_SORA	GCA_001069675.1	urine	-	+	+
SC15-3744	GCA_001885595.1	bacterial infectious disease	-	-	-
DD30	GCA_001579095.1	blood	-	-	-
OD_338919_11	GCA_002096175.1	blood, infective endocarditis	-	-	-
S.MIT/ORALIS-351	GCA_001983955.1	blood, infective endocarditis	-	-	-
IE6 (SN31376)	GCA_017154175.1	blood, infective endocarditis	-	-	-
IE17 (SN63707)	GCA_017154185.1	blood, infective endocarditis	-	-	-
918_SORA	GCA_001075675.1	bronchioalveolar lavage	-	-	-
JPIIBV3	GCA_001588645.1	dental plaque	-	-	-
BCA21	GCA_003942565.1	dental plaque	-	-	-
NU39	GCA_004127235.1	middle ear effusion fluid	-	-	-
OP51	GCA_000959975.1	oral cavity	-	-	-
ATCC 9811	GCA_006175905.1	oral cavity	-	-	-
DD27	GCA_001579025.1	oropharynx	-	-	-
201_SPSE	GCA_001071715.1	sputum	-	-	-

S. mitis NCTC 12261
S. oralis subsp. oralis COL85/1862
S. oralis subsp. oralis IE12 (SN51445)
83
S. oralis subsp. oralis OD_338919_11
100
S. oralis subsp. oralis RH_5486_10
100
100
S. oralis subsp. oralis IE18 (SN64428)
S. oralis subsp. oralis SK100
100
S. oralis subsp. oralis BCA18
S. oralis subsp. oralis 734_SORA
100
S. oralis subsp. oralis UMB0008
100
S. oralis subsp. oralis 126_SPSE
98
100
S. oralis subsp. oralis FDAARGOS_367
100
S. oralis subsp. oralis 274 SPSE
100
S. oralis subsp. oralis 727_SORA
88
S. oralis subsp. oralis BCA17
100
S. oralis subsp. oralis OD_321121_09
87
S. oralis subsp. oralis BCC35
S. oralis subsp. oralis BCC19
99
S. oralis subsp. oralis GMD3S
100
98
S. oralis subsp. oralis 1212_SMIT
100
100
S. oralis subsp. oralis BCA21
S. oralis subsp. oralis KLC05
53
83
S. oralis subsp. oralis 206_SPSE
75
S. oralis subsp. oralis RH_34706_11
S. oralis subsp. oralis RH_50443_09
42
100
S. oralis subsp. oralis RH_1735_08
100
S. oralis subsp. oralis 1314_SORA
S. oralis subsp. oralis 918_SORA
34
S. oralis subsp. oralis OD_336064_07
45
100
S. oralis subsp. oralis GMD5S
S. oralis subsp. oralis RH_8810_08
97
S. oralis subsp. oralis OD_311844_09
88
S. oralis subsp. oralis BCC58
S. oralis subsp. oralis BCC52
100
S. oralis subsp. oralis Y_11577_11
100
S. oralis subsp. oralis B_007274_11
28
81
26
S. oralis subsp. oralis B_19836_11
62
S. oralis subsp. oralis ATCC 9811
S. oralis subsp. oralis RH_57980_07
87
S. oralis subsp. oralis BCC50
47
100
S. oralis subsp. oralis C300
S. oralis subsp. oralis BCC1

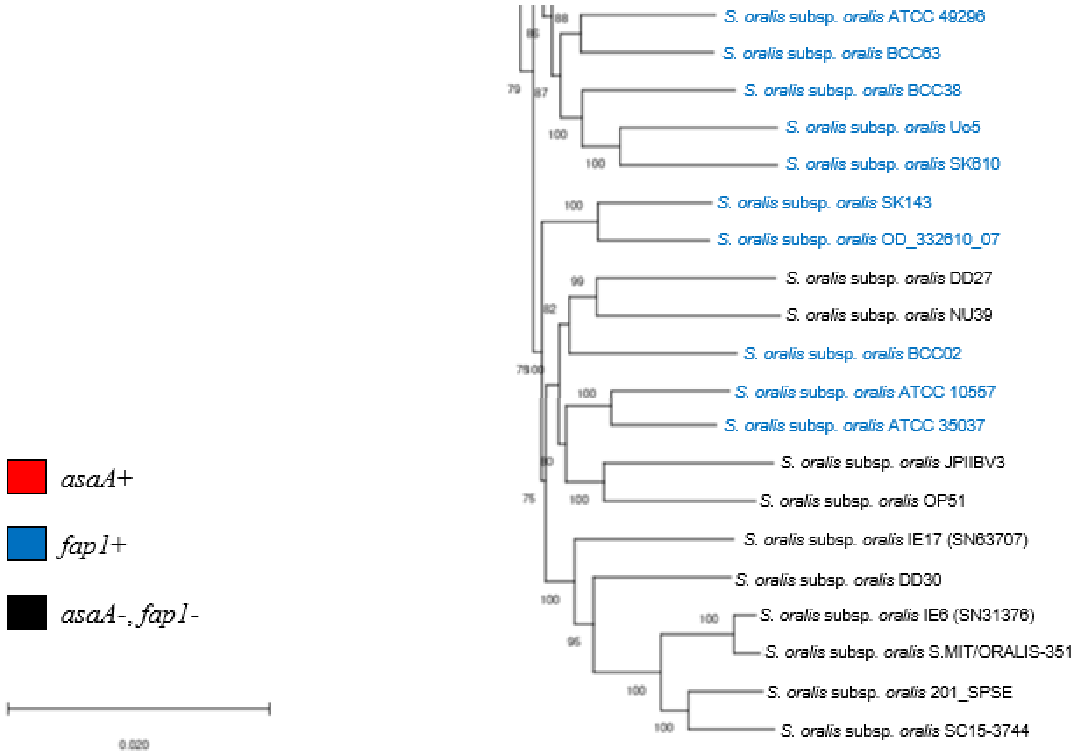

asaA+
fap1+
asaA-, fap1-
0.020
S. oralis subsp. oralis ATCC 49296
S. oralis subsp. oralis BCC63
S. oralis subsp. oralis BCC38
S. oralis subsp. oralis Uo5
S. oralis subsp. oralis SK610
S. oralis subsp. oralis SK143
S. oralis subsp. oralis OD_332610_07
S. oralis subsp. oralis DD27
S. oralis subsp. oralis NU39
S. oralis subsp. oralis BCC02
S. oralis subsp. oralis ATCC 10557
S. oralis subsp. oralis ATCC 35037
S. oralis subsp. oralis JPIIBV3
S. oralis subsp. oralis OP51
S. oralis subsp. oralis IE17 (SN63707)
S. oralis subsp. oralis DD30
S. oralis subsp. oralis IE6 (SN31376)
S. oralis subsp. oralis S.MIT/ORALIS-351
S. oralis subsp. oralis 201_SPSE
S. oralis subsp. oralis SC15-3744
88
86
79
87
100
100
100
99
82
79
100
100
80
75
100
100
100
95
100
100
100

Figure 1: *Streptococcus oralis* subsp. *oralis* strains

A maximum-likelihood phylogenetic tree with 1,000 Bootstrap replicates created by Dr. Arturo Vera-Ponce de León using 61 isolates determined to be *S. oralis* subsp. *oralis* along with *S. mitis* strain NCTC 12261 as an outgroup. He used a total of 940 single copy clusters of orthologues genes to form the tree. Those in red are *asaA* positive, in blue are *fap1* positive, and in black are *asaA* and *fap1* negative.

AsaA and Fap1 are encoded in different loci.

A potential explanation for *asaA* and *fap1* being mutually exclusive is the genes exist in the same genomic location. This is not the case as examination of the flanking genes of *asaA* and *fap1* revealed, while the genomic location of each gene is conserved, the genes exist at different genomic sites (Fig. 2). While the genomic locations of *asaA* and *fap1* are conserved, some genetic variation exists at each locus. For the *asaA* locus, the conserved flanking genes are predicted to encode a TIGR01440 family protein, a member of the DUF436 super family, and PabB, aminodeoxychorismate synthase component I (Fig 2B, Fig. 3) (40).

In all *asaA* positive strains used in this study, *asaA* was adjacent to a gene encoding a predicted zinc metalloprotease (Zmp) (Fig. 3A). This gene was also identified in this genomic location of 12 *asaA* negative strains (Fig. 3Bi). Four Zmp proteins can be encoded by oral streptococci (IgA1 protease, ZmpB, ZmpC, and ZmpD) and based on the predicted active site of HETTH, the Zmp encoded within the *asaA* locus was either ZmpB or ZmpD. Previous reports state ZmpD is not found in *S. oralis*, so the protein was most likely ZmpB. This was confirmed by constructing a phylogenetic tree using the *zmp* sequences from strains used in this study and *zmp* sequences provided by Dr. Mogens Kilian (Fig. 4) (36). All 24 *zmp* sequences fell within the *zmpB* paraphyletic clade. Phylogenetic analysis of the *zmpB* sequences from strains used in this study showed *zmpB* is more conserved in *asaA* positive strains than in *asaA* negative strains (Fig. 5). This was confirmed by looking at the shared amino acid identity, which was 99.31 to 100% in *asaA* positive strains vs 58.05 to 100% in *asaA* negative strains. This conservation excludes the N terminus of the proteins as it has high variability (36). Another gene present in the some *asaA* negative strains is a gene predicted to encode an AAA-family ATPase, suggesting other horizontal gene transfer events have happened within this locus (Fig. 3Bii).

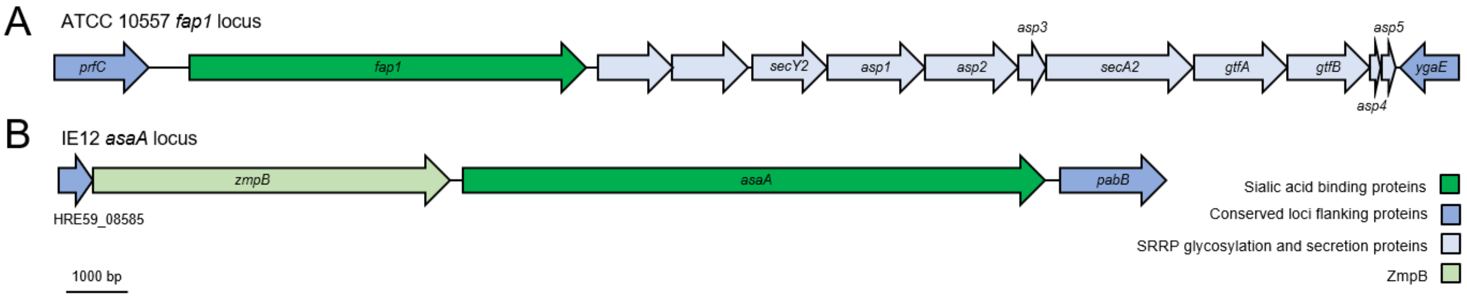


Figure 2: *asaA* and *fap1* loci

Schematic representing the genomic arrangement of the (A) *Streptococcus oralis* subsp. *oralis* ATCC 10557 *fap1* locus and the (B) *Streptococcus oralis* subsp. *oralis* IE12 *asaA* locus. Open reading frames predicted within each locus are shown by arrows.

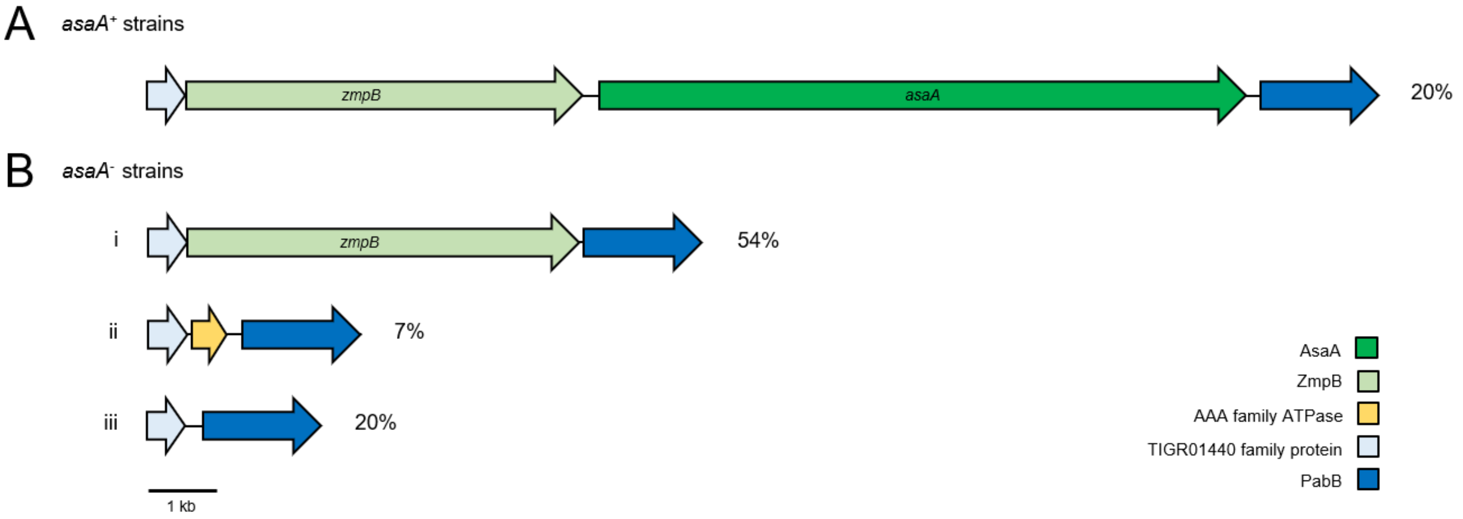


Figure 3: Variation in the *asaA* locus

Schematic representing the genomic arrangement of the *asaA* locus for (A) *asaA*+ strains and (B) *asaA*- strains. Open reading frames predicted within the locus are shown by arrows. Percentages represent the frequency of each genomic arrangement (n = 61).

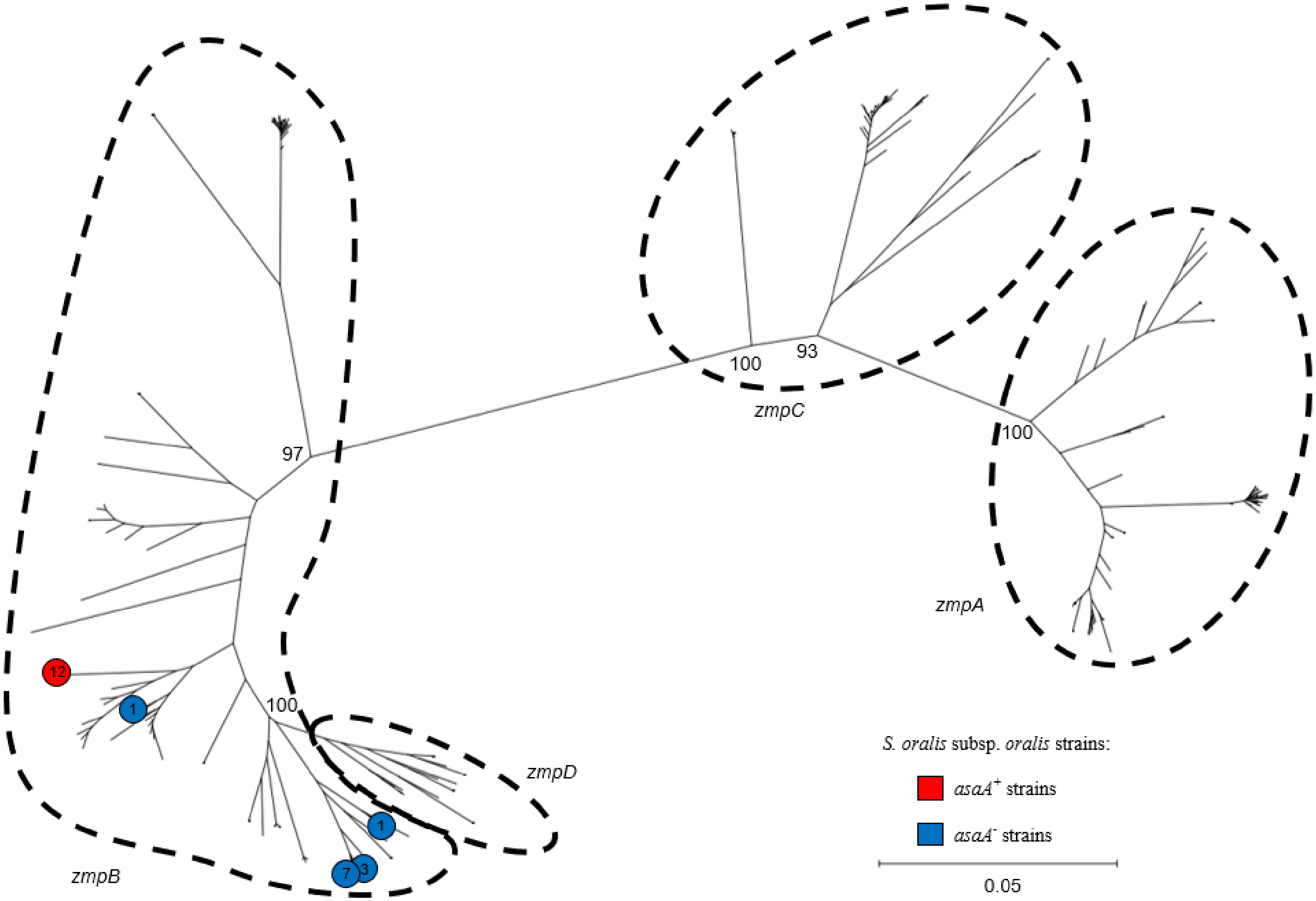


Figure 4: Phylogenetic tree of *zmp* genes

A maximum-likelihood phylogenetic tree with 1,000 Bootstrap replicates created by Dr. Arturo Vera-Ponce de León using 297 *zmp* sequences from *Streptococcus* and *Gemella* strains provided by Dr. Mogens Kilian and 24 *zmp* sequences from strains used in this study. The numbers in the circles correspond to the number of *zmp* sequences of the 24 *zmp* positive strains from this study they represent.

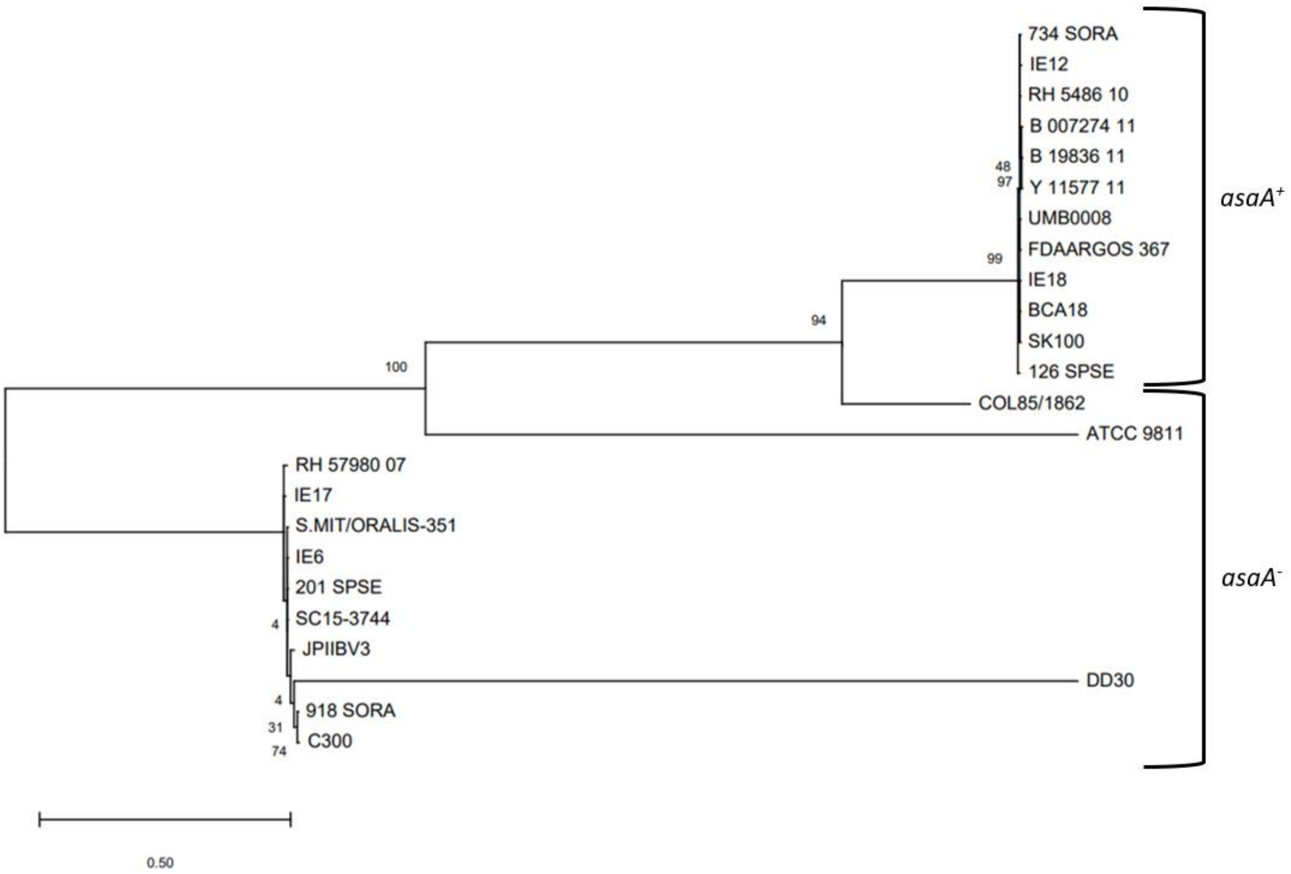


Figure 5: Phylogenetic tree of *zmpB* gene sequences

A maximum-likelihood phylogenetic tree with 1,000 Bootstrap replicates created using *zmp* sequences from the 24 *zmpB* positive *S. oralis* subsp. *oralis* strains used in this study.

High variability exists around the *fap1* locus. The conserved genes flanking the *fap1* loci are predicted to encode peptide chain release factor 3 (PrfC) and an aromatic acid exporter family protein (YgaE) (Fig. 2A). In *fap1* positive strains, adjacent to *prfC*, starts the *fap1* genomic island, containing *fap1* and the machinery dedicated for its glycosylation and secretion, which is consistent with the previously described *fap1* genomic island (1). After the genomic island in most *fap1* positive strains or starting directly after *prfC* in *fap1* negative strains is a variable region, differing in number and identity of proteins (Table 2). For the strains used in this study, 21 different genes can be present within the variable region. Between 0 and 10 of these 21 genes are present within a single strain, existing in 31 different combinations, which appear in up to five different strains. No apparent correlation exists between the presence of *fap1* and the genes encoded in the downstream variable region.

The presence of the genes in the downstream variable region were looked for in *S. oralis* subsp. *oralis* strains ATCC 35037, ATCC 10557, and DD27. ATCC 35037 and ATCC 10557 are *fap1* positive strains without a variable region. DD27 is a *fap1* negative strain with only one gene in the variable region. These three strains did not have any of the variable region genes elsewhere in their genomes, suggesting the downstream variable regions genes were acquired through horizontal gene transfer events rather than through genomic rearrangement events.

Table 2: Predicted open reading frames of the variable region downstream of the *fap1* locus

The table shows the presence of each predicted open reading frame downstream of the *fap1* locus for each strain. Highlighted in blue are the proteins encoded by the *fap1* locus flanking genes: peptide chain release factor 3 (PrfC) and an aromatic acid exporter family protein (YgaE). Highlighted in green are the proteins encoded by the *fap1* genomic island. The numbers going across indicates the order of the proteins encoded by the genes in the variable region of that strain.

S. oralis subsp. *oralis* strain	PrfC	Fap1 Genomic Island	dehydrogenase	DUF600 family protein	Hypothetical Protein 1	DUF600 family protein	Immunity protein 47; pfam15573	Hypothetical Protein 2	Suppressor of fused protein	Hypothetical Protein 3	Hypothetical Protein 4	beta-carotene 15,15'-monooxygenase	Hypothetical Protein 5	Hypothetical Protein 6	DUF600 family protein	DUF1851 domain-containing protein	Hypothetical Protein 7	Hypothetical Protein 8	Hypothetical Protein 9	Hypothetical Protein 10	Immunity protein 22	DUF1851 domain-containing protein	IS200/IS605 family transposase	YgaE
BCC58	+	+		1	2	3	4	5		6	7	8								9		10		+
OD_332610_07	+	+		1	2	3	4		5						6							7		+
BCC02	+	+		1	2	3	4			5	6	7	8											+
OD_311844_09	+	+		1	2	3	4			5	6	7	8											+
RH_8610_08	+	+		1	2	3			4			5			6					7		8		+
Uo5	+	+		1	2	3				4	5	6	7											+
BCC50	+	+		1	2	3				4	5	6	7											+
BCC38	+	+		1	2	3				4	5	6	7											+
BCC63	+	+		1	2	3				4	5	6								7		8		+
206_SPSE	+	+		1			2	3	4						5		6							+
OD_336064_07	+	+		1					2									3	4			5		+
GMD3S	+	+		1					2									3	4			5		+
BCC52	+	+		1					2									3	4			5		+
BCC11	+	+		1					2									3	4			5		+

ATCC 49296	+	+		1						2	3	4	5											+
OD_321121_09	+	+		1																	2	3		+
BCC19	+	+		1																	2	3		+
BCC35	+	+		1																	2	3		+
RH_50443_09	+	+		1																		2		+
SK143	+	+		1																		2		+
1212_SMIT	+	+		1																		2		+
RH_57980_07	+	+		1																				+
SK610	+	+				1	2	3		4	5	6	7											+
COL85/1862	+	+				1	2		3						4							5		+
RH_1735_08	+	+				1			2									3	4			5		+
GMD5S	+	+				1			2									3	4			5		+
1314_SORA	+	+				1			2									3	4			5		+
RH_34706_11	+	+				1															2	3		+
727_SORA	+	+				1															2	3		+
274_SPSE	+	+				1															2	3		+
KLC05	+	+				1															2	3		+
BCA17	+	+				1															2	3		+
ATCC 35037	+	+																						+
C300	+	+																						+
ATCC 10557	+	+																						+
RH_5486_10	+	-	1	2	3	4	5	6	7						8		9					10		+
IE18	+	-		1	2	3	4	5	6						7	8	9					10		+
OD_338919_11	+	-		1	2	3	4	5	6						7		8					9		+
ATCC 9811	+	-		1	2	3	4	5	6						7							8		+
IE12	+	-		1	2	3	4		5						6							7		+
FDAARGOS_367	+	-		1	2	3																4		+
126_SPSE	+	-		1	2	3																4		+
734_SORA	+	-		1	2	3																4		+
UMB0008	+	-		1	2	3																4		+
BCA18	+	-		1	2	3																4		+

DD30	+	-		1			2																	+
IE17	+	-		1										2			3					4		+
IE6	+	-		1												2	3					4		+
201_SPSE	+	-		1													2					3		+
S.MIT/ORALIS-351	+	-		1													2					3		+
SC15-3744	+	-		1													2					3		+
918_SORA	+	-		1																		2		+
BCA21	+	-		1																		2		+
Y_11577_11	+	-				1	2	3	4						5		6					7		+
B_007274_11	+	-				1	2	3	4						5		6					7		+
B_19836_11	+	-				1	2	3	4						5		6					7		+
SK100	+	-				1	2	3	4						5		6					7		+
OP51	+	-								1	2	3	4											+
NU39	+	-								1	2	3	4											+
JPIIBV3	+	-								1	2	3	4											+
DD27	+	-																					1	+

Diversity exists within AsaA and Fap1.

AsaA and Fap1 may have different biological activities as they have different domains. These domain structures have been described previously and were confirmed for the adhesins encoded by strains in this study (1, 2) (Fig. 6). Here we defined the diversity of these domains across all AsaA and Fap1 proteins present in the *S. oralis* subsp. *oralis* strains used in this study.

Twelve strains encode AsaA, and the previously described domains were confirmed in eleven of these strains. In the twelfth strain, SK100, the non-repeat region of *asaA* is divided over two contigs, preventing confirmation that the full non-repeat region is conserved. However, the amino acid sequence obtained is 100% identical to another strain, BCA18, AsaA non-repeat region. Overall, the non-repeat region of the twelve AsaA proteins share 98.2 to 100% amino acid identity. Within the non-repeat region, AsaA has two Siglec-like and Unique domains. The first Siglec-like domains share 100% amino acid identity, meaning all contain the non-canonical YTRY motif (GTRY) containing the arginine residue proposed to be required for sialic acid binding in AsaA (2). The second Siglec-like domain does not have any YTRY motif but is still highly conserved, ranging from 97.16 to 100% amino acid identity (Fig. 7A).

This high conservation could be the result of *asaA* being acquired through a horizontal gene transfer from a single donor. To determine if the distribution of *asaA* positive strains match the diversity of the *asaA* genes, a phylogenetic tree of the non-repeat region of *asaA* genes from strains used in this study was created (Fig. 8). The repeat regions were excluded because they are usually split over two contigs in the repeat region. The *asaA* positive strains and the diversity of *asaA* genes do not have the same distribution, which further supports intraspecies gene transfer events occurred (Fig. 1, Fig 8).

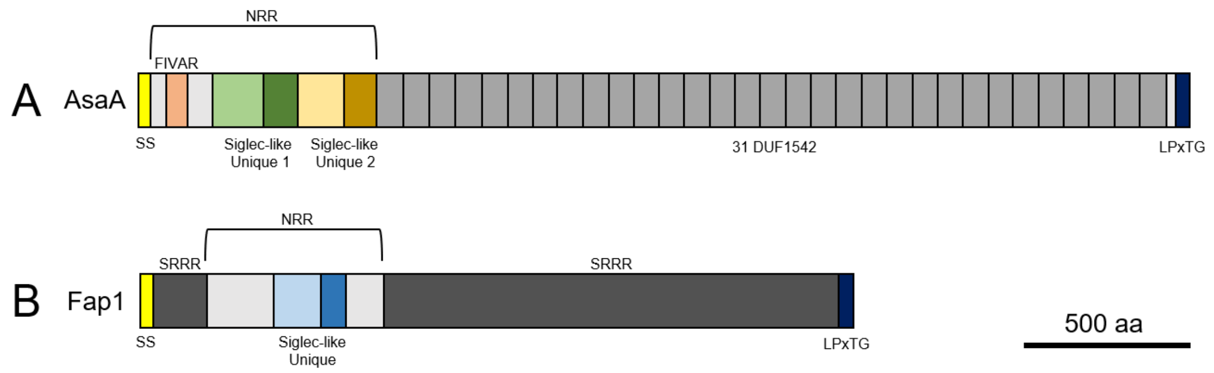


Figure 6: Predicted domain structure of *S. oralis* subsp. *oralis* sialic acid binding adhesins.

(A) A schematic of the predicted protein structure of AsaA from *S. oralis* subsp. *oralis* IE12. AsaA has a secretion signal (SS), a FIVAR domain and two Siglec-like and Unique domains within its non-repeat region (NRR), 31 DUF1542 domains, and a LPxTG cell wall binding motif. (B) A schematic of the predicted protein structure of Fap1 from *S. oralis* subsp. *oralis* ATCC 10557. Fap1 has a secretion signal (SS), a short serine-rich repeat region (SRRR), one Siglec-like and Unique domains within its non-repeat region (NRR), a second, larger SRRR, and a LPxTG cell wall binding motif.

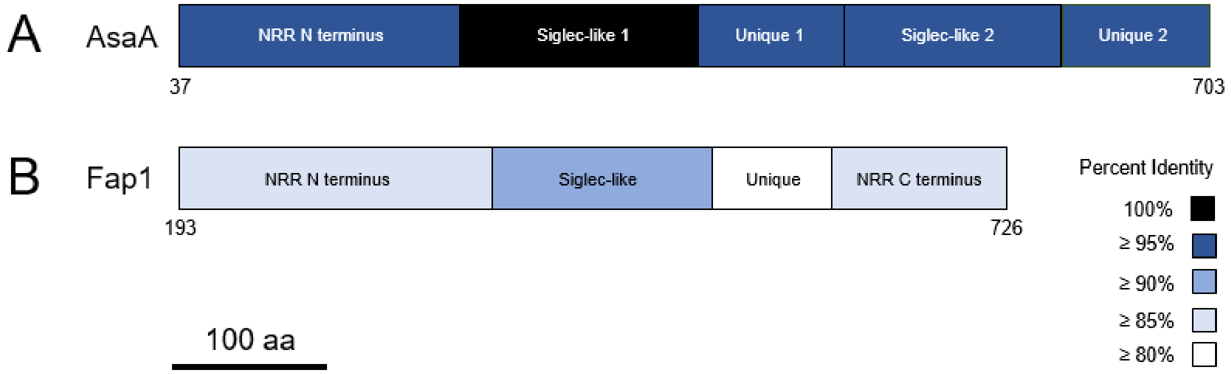


Figure 7: Percent amino acid identity within the non-repeat region of each adhesin.

(A) The percent identity of each domain within the non-repeat region (NRR) for all twelve AsaA proteins used in this study. Amino acid number indicated references the position within the full-length protein of AsaA from *S. oralis* subsp. *oralis* IE12. (B) The percent identity of each domain within the NRR for 34 Fap1 proteins used in this study. Amino acid number indicated references the position within the full-length protein of Fap1 from *S. oralis* subsp. *oralis* ATCC 10557. The 35th Fap1 protein (from strain BCC63) is omitted from the figure as its NRR most likely underwent a recombination event, resulting in a different Siglec-like domain being present.

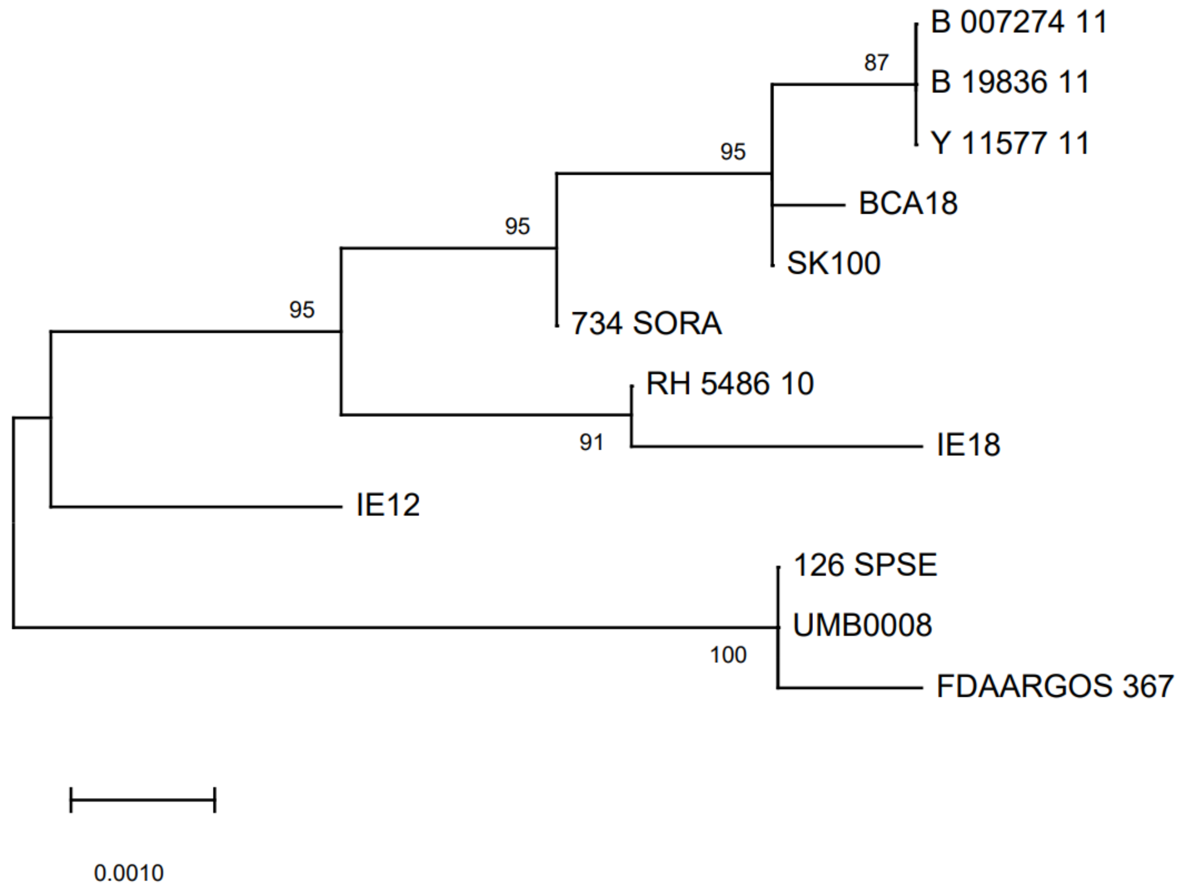


Figure 8: Phylogenetic tree of the non-repeat region of *asaA* gene sequences

A maximum-likelihood phylogenetic tree with 1,000 Bootstrap replicates created using the non-repeat regions of the twelve *asaA* gene sequences.

Thirty-five strains encode Fap1, and the previously described domains were identified in all of them (1). Overall, the non-repeat region of the thirty-five Fap1 proteins share 74.67 to 100% amino acid identity. Within the non-repeat region, Fap1 has one Siglec-like domain and one Unique domain. For thirty-four of the Fap1 proteins, the Siglec-like domain is very conserved, ranging from 90.85 to 100% amino acid identity (Fig. 7B). In the thirty-fifth strain, BCC63, most likely a recombination event occurred resulting in a divergent amino acid sequence from position 380 to 553 (Fig. 9). This region shares either 45.24 or 45.83% amino acid identity with the same region within the other Fap1 proteins. As most of this region is within the Siglec-like domain, it is likely the Siglec-like domain present within BCC63's Fap1 has a different binding specificity. Further BLASTn searches did not find a potential donor for this recombined Siglec-like domain. Even so, all the present Siglec-like domains contain a non-canonical YTRY motif (WTRY) containing the arginine residue required for sialic acid binding on platelets (1).

The Siglec-like and Unique domains of AsaA and Fap1 have high predicted structural similarity (Fig. 10). However, the Siglec-like domains of AsaA and Fap1 share low amino acid identity ($\leq 26\%$). The low sequence identity may allow AsaA and Fap1 to bind different sialic acid linkages or sialic acid containing glycans.

```
ATCC 10557    GTQNQANSNLSERASVGVQSQYQASESARETVKESQPSKELKAVDFSTESLPQSQSGRVK    252
BCC63         GTQNQANSNLSERASVGVQSQYQASESARETVKESQPSKELKAVDFSTESLPQSQSGRVK    252
              ************************************************************

ATCC 10557    NEGVTAESSLTMTSVALTEKQSEEKRKKLEALSAEIGQFLAQAQGLPNSDEAIAKASLAK    312
BCC63         NEGVTAESSLTMTSVALTEKQSEEKRKKLEALSAEIGQFLTQAQGLPNSDEAIAKASLAK    312
              ****************************************:*******************

ATCC 10557    NEIAEALKGEVSDLAPVLQKATEARNSIANAVLRANSGPRDSRNGQALTKASNTASFRAA    372
BCC63         NEIAEALKGEASDLATILQKATEVRNSIANAVLRANSGLRDSRNGQALTQASNTASFRAA    372
              **********.**** :******.*************** **********:*********

ATCC 10557    RDTEKPELQKITVTGGAVLEGQKFKIYREENFSATIEFTDNSGRIEHAKFVPTAVPAAYP    432
BCC63         RDTEKPERTRIVTNEGATVDGQLIRVYREERFEATFEFTDNSGRIEHARVEKYPTVALVP    432
              *******  :*... **.::** :::****.*.**:************:.    . *  *

ATCC 10557    ATSTVVSFTT----SNGQSISMIVPTNKLAKDGNATASNPFTVSITGSVGKNQAVNSLWT    488
BCC63         RGKTVDTITSSNVKNDVHTITTTVPTEKFGQDGNATSTNPFKVTASGSISKSIQAGGLWT    492
                .** ::*:    .: ::*:  ***:*:.:*****::***.*: :**:.*.  ...***

ATCC 10557    RYVFTYDQEGNFSGNTTDVGLVKDLTANPAAIQFEVHAQSEKYEPAINAEVNRNFTLTAN    548
BCC63         RYIYTYDQANNYNGNDVDTN-KKSVTENPAAIQFVAVAQTEKYTAVAKGNSSQTLTLNSG    551
              **::**** .*:.** .*..  *.:* ******* . **:***  . :.: .:.:**.:.

ATCC 10557    SGTVSVGEASQYITNATGTPELPTTGITKGTRTTYTWKSGTNTNLSAGRHTLTAVVTYPD    608
BCC63         QTTISVGEASQYITNAAGTPELPTTGITPGTQTTYTWKSGTNTNLSAGRHTLTAVVTYPD    611
              . *:************:*********** **:****************************

ATCC 10557    GSTDEIDVSFTVRPQTPRIENQYLNEKGGLSNQAITVDGVAPGGTVTLTIAGETFTKQAT    668
BCC63         GSTDEVEIPIEVRPQTPRIEERFLNEKGGLTNQAITVDGVTPGGTVTLTIAGETFTKQAT    671
              *****::: : *********:::*******:*********:*******************

ATCC_10557    GSSTSVTFTATDLKKVYDRNGGRLPSGPVTASTTVDGLVSDVFNGQITPNQASISVSN       726
BCC63         GSSTSVTFTANELKKVYDRNGGRLPSGPVTASTTVNGLVSDVFNGQITPNQASISVSN       729
              **********.:***********************:*********************
```

Figure 9: Fap1 non-repeat region amino acid alignment

Amino acid alignment of the non-repeat region of Fap1 from *S. oralis* subsp. *oralis* strains ATCC 10557 and BCC63. ATCC 10557 is used to establish the amino acid number position within the full-length Fap1 protein and to determine where a recombination event may have occurred within Fap1 of BCC63, shown by the grey highlight. Red font shows the Siglec-like domain. Red highlighting shows the conserved, non-canonical YTRY motif. Blue font shows the Unique domain.

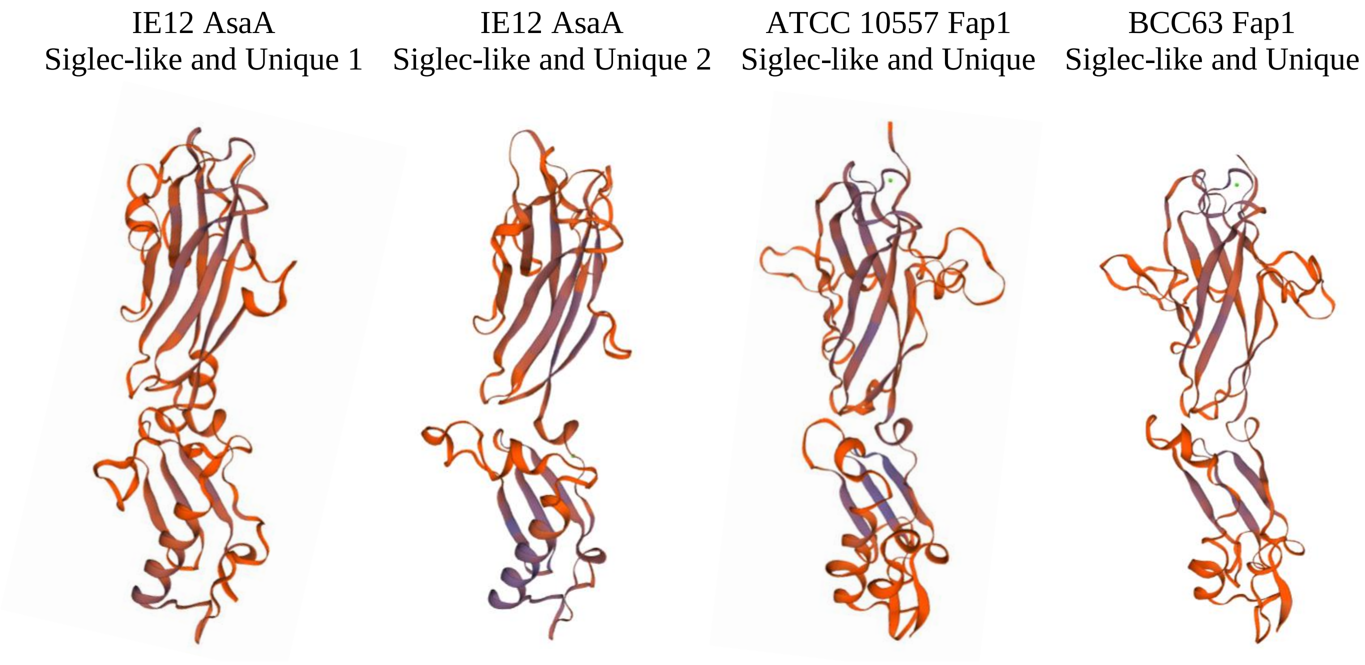


Figure 10: Comparison of the structural predictions of the Siglec-like and Unique domains of AsaA and Fap1 The predicted tertiary structure of the two Siglec-like and Unique domains of AsaA from *S. oralis* subsp. *oralis* IE12 and the Siglec-like and Unique domain of Fap1 from *S. oralis* subsp. *oralis* ATCC 10557 and BCC63 are shown for comparison. The predicted structures were made using SWISS-MODEL.

Some strains lack a known sialic acid adhesin.

While a majority of *S. oralis* subsp. *oralis* strains encode a sialic acid adhesin, 14 (23%) do not contain *asaA* or *fap1*. Of these 14 strains, 5 (36%) were isolated from IE, showing these adhesins are not essential for disease. It is possible these strains encode another sialic acid adhesin. Except for AsaA, all described sialic acid adhesins of *Streptococcus* are Siglec-like containing SRRPs. Because Siglec-like domains have structural similarity but low sequence identity, the distribution of *secA2*, an essential component of the SRRP secretion system, was examined (1, 9). All 14 *S. oralis* subsp. *oralis* strains lacking *asaA* and *fap1* also lacked *secA2*, meaning these is no indication of other sialic acid adhesins.

Discussion

These data suggest *asaA* and *fap1* are mutually exclusive as none of the 61 *S. oralis* subsp. *oralis* strains used in this study had both *asaA* and *fap1*. However, there is no discernable reason as to why *asaA* and *fap1* are mutually exclusive since they are encoded in different loci. A possible explanation is encoding both *asaA* and *fap1* is detrimental. It is possible strains only need one sialic acid adhesin and expressing two sialic acid adhesins is an unnecessary energy expense. An alternative possibility is the proteins interfere with each other, preventing them from functioning properly and being able to bind sialic acid.

Another reason *asaA* and *fap1* may be mutually exclusive is they bind different sialic acid containing glycans found in the oral cavity since their Siglec-like domains share low amino acid sequence identity. If AsaA and Fap1 do in fact bind different sialic acid containing glycans, the selection for each adhesin could be driven by its ability to bind components present in the strain's microbial niche within the human oral cavity and upper respiratory tract, the main habitat of *S. oralis* subsp. *oralis* (7, 8, 17, 19). While the sialic acid variants present on platelet glycoproteins are the only confirmed sialic acid AsaA and Fap1 bind, unpublished data from experiments done by Dr. Meztlli Gaytán shows AsaA and Fap1 bind sialic acid on saliva (1, 2). However, encoding both adhesins would potentially allow the strain to bind multiple sialic acid containing glycans which may be advantageous. Therefore, other characteristics of the strain may prevent it from surviving in multiple niches and there may be selection against encoding an adhesin that binds a sialic acid containing glycan not found in the niche.

It could also be due to chance *asaA* and *fap1* appear to be mutually exclusive. AsaA is highly conserved within *S. oralis* subsp. *oralis*, suggesting the *asaA* gene may have been acquired through a single interspecies horizontal gene transfer event and spread through subsequent intraspecies horizontal gene transfers. This is supported by the fact the three strains, B_007274_11, B_19836_11, and Y_11577_11, which form a single clade on the *S. oralis* subsp. *oralis* phylogenetic tree separate from the larger group of *asaA* positive strains, do not form their own outgroup on the *asaA* gene phylogenetic tree. These *asaA* genes are more closely related to other *S. oralis* subsp. *oralis asaA* genes within the larger group and were probably acquired through an intraspecies gene transfer event, followed by natural strain divergence.

The acquisition of *asaA* from a donor species is further supported by the fact the adjacent ZmpB is more conserved in *asaA* positive strains than in *asaA* negative strains. These two genes may have been transferred together, however BLASTn searches could not find a potential donor species for *asaA*, with or without *zmpB*. It is possible the donor species has simply not been sequenced as many different and diverse microbes inhabit the oral cavity (17). Since it is likely *asaA* was acquired through a horizontal gene transfer, it is possible the original strain that acquired *asaA* was *fap1* negative and subsequent *asaA* intraspecies transfers just happened to introduce the gene into *fap1* negative strains. A further possibility is the acquisition of *asaA* drove strains to lose *fap1*, but there is no evidence if *asaA* positive strains were ever *fap1* positive as no correlation exist between presence of *asaA* and genes found downstream of the *fap1* locus.

ZmpB, a zinc metalloprotease, being present in the *asaA* locus for all *asaA* positive strains is a point of intrigue as ZmpB has been suggested to play an unidentified but important role in Mitis group *Streptococcus* virulence, which includes *S. oralis* subsp. *oralis*. The zinc metalloprotease is found in all strains of the Mitis and Salivarius groups, except for in *S. thermophilus*, the only species not associated with humans. All Zmps have been shown to be crucial in pneumococcal pathogenicity within a mouse model, but the main virulence was assigned to IgA1 protease and ZmpB (36). ZmpB function is unknown, but it has been shown to contribute to inflammation in the lower respiratory tract by increasing tumor necrosis factor alpha (TNF-α) levels, a proinflammatory cytokine. Increased TNF-α levels can be harmful, impairing lung tissue integrity and enhancing dissemination of bacteria into the bloodstream (41). It is unclear how increasing TNF-α levels connects to the selection of *asaA* and *zmpB* in the oral cavity, but other unknown functions of ZmpB could contribute to *asaA* and *zmpB* selection.

Fap1 is more diverse than AsaA and the origin of the genomic island encoding Fap1 is unclear. The variable region downstream of the *fap1* locus is a hot spot of recombination, as shown by the high diversity of genes present. These genes were most likely acquired through horizontal gene transfer events as three *S. oralis* subsp. *oralis* strains with either no genes or one gene present in their variable region do not have variable region genes elsewhere in their genome. Therefore, the *fap1* genomic island could have been acquired through an interspecies gene transfer event and from there spread throughout the subspecies. Another possibility is the *fap1* genomic island was present in the progenitor of *S. oralis* subsp. *oralis* and negative strains have lost the genomic island. The high level of recombination downstream of the *fap1* locus could support either hypothesis.

Among the genes present in the variable region downstream of the *fap1* locus are multiple DUF600 family proteins. DUF600 family proteins can be found in three locations within the variable region downstream of the *fap1* locus. For the most part, the DUF600 family proteins have higher percent amino acid identity at each position than with DUF600 family proteins found at another location. It is possible these DUF600 family proteins act as antitoxins. In *Staphylococcus aureus*, the DUF600 containing antitoxin EsaG binds and neutralizes the nuclease toxin, EsaD, during the biosynthesis of EsaD (42, 43). When EsaD is secreted by the type VII protein secretion system, EsaG is shaved off as EsaG is only found within the *S. aureus* cytoplasm. *S. aureus* strains without EsaD still encode at least two copies of EsaG-like proteins to protect themselves from EsaD secreted from *esaD* positive strains (42, 43). It is possible the DUF600 family proteins found within the *fap1* locus may also act as protective antitoxins against nucleases.

A major difference within the non-repeat regions of AsaA and Fap1 is AsaA has two Siglec-like domains. The first Siglec-like domain contains the YTRY motif while the second Siglec-like domain does not. However, the second Siglec-like domain is predicted to form a similar structure which suggest the domain may bind sialic acid (2). It is theorized the predicted flexible loops of the Siglec-like domain play a role in mediating binding for those lacking the YTRY motif (2, 12, 13). The potentially two functioning Siglec-like domains may widen the number of sialic acid containing glycans AsaA can bind at a time. It is also possible the two Siglec-like domains could bind different sialic acid containing glycans, broadening AsaA specificity.

AsaA and Fap1 also have other domains present in their non-repeat regions. AsaA has a FIVAR domain present in the N-terminal region of its non-repeat region. This FIVAR domain is found in some CAZymes (Carbohydrate-active enzymes) and in a putative carbohydrate-binding module (CBM) found in an endo-beta-N-acetyl-glucosaminidase (EndoS) from *Streptococcus pyogenes* (2). The putative CBM in EndoS has structural homology with CBM62, which binds galactose-containing structures (PDB: 4NUZ) (2, 44). However, AsaA is not involved with binding β-1,4-linked galactose on platelets (2). Fap1 potentially contains a CBM44 domain in the C-terminus of its non-repeat region (PDB: 2C26) (1, 45). However, when using a more recent database, structural similarity to this domain was not identified again (1, 46). Fap1 is associated with binding β-1,4-linked galactose on platelet glycoproteins after removal of the terminal sialic acid by neuraminidase, which is secreted by *S. oralis* subsp. *oralis* (1). Unpublished data from experiments done by Dr. Meztlli Gaytán shows AsaA and Fap1 are associated with binding β-1,4-linked galactose on saliva, which could potentially be mediated by the FIVAR and CBM44 domains, respectively.

The repeat regions of AsaA and Fap1 are also different. AsaA contains DUF1542 repeats, which are predicted to form α-helical structures (2, 47). Fap1 contains a serine rich repeat region that is heavily glycosylated (1, 9). For other SRRPs, this glycosylation is essential for proper protein function (48). Therefore, it can be assumed that glycosylation is required for proper Fap1 function. These different repeat regions most likely impact the biology of the strains that contain them. However, they likely have the same baseline function of forming an appendage to extend the non-repeat region away from the cell surface to bind sialic acid (1, 2).

Some strains isolated from IE lack AsaA and Fap1, meaning that AsaA and Fap1 are not essential to cause IE. These strains also lack SecA2, an essential component of the SRRP secretion system, meaning there is no indication of other SRRPs in their genomes (1, 2). Therefore, it is possible these strains bind sialic acid with a novel adhesin. A similar method to how AsaA was identified could be applied to AsaA and Fap1 negative strains (2). First, it needs to be determined if these negative strains bind platelets, and if so, if the binding is via sialic acid. This could be addressed using platelet adherence assays that involve the removal of and competitive binding to sialic acid. If the negative strains do bind sialic acid, then the third sialic acid adhesin would need to be identified. Both Fap1 and AsaA are attached to the cell wall through a LPxTG motif by SrtA (1, 2, 49). AsaA was found by looking at LPxTG proteins found in sialic acid binding, Fap1 negative strains (2). Comparative analysis could be done between the LPxTG proteins found in Fap1 positive strains, AsaA positive strains, and negative strains to find LPxTG proteins only found in negative strains. The domains of these candidate proteins will then need to be defined to determine if they have a sialic acid binding domain.

However, it is possible AsaA and Fap1 negative strains do not bind sialic acid on platelets and instead bind platelets through a novel receptor. The same comparative analysis described previously would work to find candidate proteins, but other domains would need to be defined to determine what platelet structures these proteins bind. Another possibility is these strains do not bind platelets at all and instead cause IE through a novel mechanism. Many host components are present at the damage heart valve site, including endothelial cells, immune cells, and extracellular matrices. It is possible these strains bind a receptor on one of these components, and therefore, they don’t bind sialic acid nor platelets. These scenarios would challenge the paradigm that sialic acid binding on platelet glycoproteins is essential for causing IE.

AsaA and Fap1 are two sialic acid adhesins found within *S. oralis* subsp. *oralis* and are mutually exclusive even though they are found in different but conserved loci. AsaA was most likely acquired through a single horizontal gene transfer event. Fap1 is less conserved than AsaA likely because it is older within *S. oralis* subsp. *oralis*, and therefore had more time to diversify. Strains lacking AsaA and Fap1 may cause IE through a novel mechanism that does not involve sialic acid or platelets, which would challenge the dogma that sialic acid binding on platelets is necessary for Mitis group Streptococci to cause IE.

References

1. Singh AK, Woodiga SA, Grau MA, King SJ. *Streptococcus oralis* Neuraminidase Modulates Adherence to Multiple Carbohydrates on Platelets. Infect Immun. 2017 Feb 23;85(3):e00774-16. doi: 10.1128/IAI.00774-16. PMID: 27993975; PMCID: PMC5328485.

2. Gaytán MO, Singh AK, Woodiga SA, Patel SA, An SS, Vera-Ponce de León A, McGrath S, Miller AR, Bush JM, van der Linden M, Magrini V, Wilson RK, Kitten T, King SJ. A novel sialic acid-binding adhesin present in multiple species contributes to the pathogenesis of Infective endocarditis. PLoS Pathog. 2021 Jan 19;17(1):e1009222. doi: 10.1371/journal.ppat.1009222. PMID: 33465168; PMCID: PMC7846122.

3. Andersen MH, Holle SLK, Klein CF, Bruun NE, Arpi M, Bundgaard H, Tønder N, Iversen KK. Risk for infective endocarditis in bacteremia with Gram positive cocci. Infection. 2020 Dec;48(6):905-912. doi: 10.1007/s15010-020-01504-6. Epub 2020 Aug 25. PMID: 32844380.

4. Werdan K, Dietz S, Löffler B, Niemann S, Bushnaq H, Silber RE, Peters G, Müller-Werdan U. Mechanisms of infective endocarditis: pathogen-host interaction and risk states. Nat Rev Cardiol. 2014 Jan;11(1):35-50. doi: 10.1038/nrcardio.2013.174. Epub 2013 Nov 19. PMID: 24247105.

5. Vogkou CT, Vlachogiannis NI, Palaiodimos L, Kousoulis AA. The causative agents in infective endocarditis: a systematic review comprising 33,214 cases. Eur J Clin Microbiol Infect Dis. 2016 Aug;35(8):1227-45. doi: 10.1007/s10096-016-2660-6. Epub 2016 May 11. PMID: 27170145.

6. Pyburn TM, Bensing BA, Xiong YQ, Melancon BJ, Tomasiak TM, Ward NJ, Yankovskaya V, Oliver KM, Cecchini G, Sulikowski GA, Tyska MJ, Sullam PM, Iverson TM. A structural model for binding of the serine-rich repeat adhesin GspB to host carbohydrate receptors. PLoS Pathog. 2011 Jul;7(7):e1002112. doi: 10.1371/journal.ppat.1002112. Epub 2011 Jul 7. Erratum in: PLoS Pathog. 2012 December; 8(12): 10.1371/annotation/9bb4b6f9-d220-4c81-bdbc-ebdb33e6d892. PMID: 21765814; PMCID: PMC3131266.

7. Deng L, Bensing BA, Thamadilok S, Yu H, Lau K, Chen X, Ruhl S, Sullam PM, Varki A. Oral streptococci utilize a Siglec-like domain of serine-rich repeat adhesins to preferentially target platelet sialoglycans in human blood. PLoS Pathog. 2014 Dec 4;10(12):e1004540.

8. Varki A. Sialic acids in human health and disease. Trends Mol Med. 2008 Aug;14(8):351-60. doi: 10.1016/j.molmed.2008.06.002. Epub 2008 Jul 6. PMID: 18606570; PMCID: PMC2553044.

9. Lizcano A, Sanchez CJ, Orihuela CJ. A role for glycosylated serine-rich repeat proteins in gram-positive bacterial pathogenesis. Mol Oral Microbiol. 2012 Aug;27(4):257-69. doi: 10.1111/j.2041-1014.2012.00653.x. Epub 2012 Jun 11. PMID: 22759311; PMCID: PMC3390760.

10. Varki A. Sialic acids as ligands in recognition phenomena. FASEB J. 1997 Mar;11(4):248-55. doi: 10.1096/fasebj.11.4.9068613. PMID: 9068613.

11. Ronis A, Brockman K, Singh AK, Gaytán MO, Wong A, McGrath S, Owen CD, Magrini V, Wilson RK, van der Linden M, King SJ. *Streptococcus oralis* subsp. *dentisani* Produces Monolateral Serine-Rich Repeat Protein Fibrils, One of Which Contributes to Saliva Binding via Sialic Acid. Infect Immun. 2019 Sep 19;87(10):e00406-19. doi: 10.1128/IAI.00406-19. PMID: 31308084; PMCID: PMC6759294.

12. Stubbs HE, Bensing BA, Yamakawa I, Sharma P, Yu H, Chen X, Sullam PM, Iverson TM. Tandem sialoglycan-binding modules in a Streptococcus sanguinis serine-rich repeat adhesin create target dependent avidity effects. J Biol Chem. 2020 Oct 23;295(43):14737-14749. doi: 10.1074/jbc.RA120.014177. Epub 2020 Aug 20. PMID: 32820052; PMCID: PMC7586212.

13. Bensing BA, Khedri Z, Deng L, Yu H, Prakobphol A, Fisher SJ, Chen X, Iverson TM, Varki A, Sullam PM. Novel aspects of sialoglycan recognition by the Siglec-like domains of streptococcal SRR glycoproteins. Glycobiology. 2016 Nov;26(11):1222-1234. doi: 10.1093/glycob/cww042. Epub 2016 Apr 1. PMID: 27037304; PMCID: PMC6086536.

14. Xiong YQ, Bensing BA, Bayer AS, Chambers HF, Sullam PM. Role of the serine-rich surface glycoprotein GspB of Streptococcus gordonii in the pathogenesis of infective endocarditis. Microb Pathog. 2008 Oct;45(4):297-301. doi: 10.1016/j.micpath.2008.06.004. Epub 2008 Jul 5. PMID: 18656529; PMCID: PMC2574613.

15. Takahashi Y, Takashima E, Shimazu K, Yagishita H, Aoba T, Konishi K. Contribution of sialic acid-binding adhesin to pathogenesis of experimental endocarditis caused by *Streptococcus gordonii* DL1. Infect Immun. 2006 Jan;74(1):740-3. doi: 10.1128/IAI.74.1.740-743.2006. PMID: 16369032; PMCID: PMC1346603.

16. Turner LS, Kanamoto T, Unoki T, Munro CL, Wu H, Kitten T. Comprehensive evaluation of *Streptococcus sanguinis* cell wall-anchored proteins in early infective endocarditis. Infect Immun. 2009 Nov;77(11):4966-75. doi: 10.1128/IAI.00760-09. Epub 2009 Aug 24. PMID: 19703977; PMCID: PMC2772543.

17. Abranches J, Zeng L, Kajfasz JK, Palmer SR, Chakraborty B, Wen ZT, Richards VP, Brady LJ, Lemos JA. Biology of Oral Streptococci. Microbiol Spectr. 2018 Oct;6(5):10.1128/microbiolspec.GPP3-0042-2018. doi: 10.1128/microbiolspec.GPP3-0042-2018. PMID: 30338752; PMCID: PMC6287261.

18. Douglas CW, Heath J, Hampton KK, Preston FE. Identity of viridans streptococci isolated from cases of infective endocarditis. J Med Microbiol. 1993 Sep;39(3):179-82. doi: 10.1099/00222615-39-3-179. PMID: 8366515.

19. Do T, Jolley KA, Maiden MCJ, Gilbert SC, Clark D, Wade WG, Beighton D. Population structure of *Streptococcus oralis*. Microbiology (Reading). 2009 Aug;155(Pt 8):2593-2602. doi: 10.1099/mic.0.027284-0. Epub 2009 May 7. PMID: 19423627; PMCID: PMC2885674.

20. Jensen A, Scholz CFP, Kilian M. Re-evaluation of the taxonomy of the Mitis group of the genus *Streptococcus* based on whole genome phylogenetic analyses, and proposed reclassification of *Streptococcus dentisani* as *Streptococcus oralis* subsp. *dentisani* comb. nov., *Streptococcus tigurinus* as *Streptococcus oralis* subsp. *tigurinus* comb. nov., and *Streptococcus oligofermentans* as a later synonym of *Streptococcus cristatus*. Int J Syst Evol Microbiol. 2016 Nov;66(11):4803-4820. doi: 10.1099/ijsem.0.001433. Epub 2016 Aug 17. PMID: 27534397.

21. Büttner H, Perbandt M, Kohler T, Kikhney A, Wolters M, Christner M, Heise M, Wilde J, Weißelberg S, Both A, Betzel C, Hammerschmidt S, Svergun D, Aepfelbacher M, Rohde H. A Giant Extracellular Matrix Binding Protein of *Staphylococcus epidermidis* Binds Surface-Immobilized Fibronectin via a Novel Mechanism. mBio. 2020 Oct 20;11(5):e01612-20. doi: 10.1128/mBio.01612-20. PMID: 33082256; PMCID: PMC7587433.

22. Cheng AG, Missiakas D, Schneewind O. The giant protein Ebh is a determinant of *Staphylococcus aureus* cell size and complement resistance. J Bacteriol. 2014 Mar;196(5):971-81. doi: 10.1128/JB.01366-13. Epub 2013 Dec 20. PMID: 24363342; PMCID: PMC3957702.

23. Contreras-Moreira, B., and Vinuesa, P. (2013). GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. Appl. Environ. Microbiol. 79, 7696–7701. doi:10.1128/AEM.02411-13.

24. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589. doi:10.1038/nmeth.4285.

25. Li, L. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13, 2178–2189. doi:10.1101/gr.1224503.

26. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534. doi:10.1093/molbev/msaa015.

27. Vera-Ponce de León, A., Jahnes, B. C., Duan, J., Camuy-Vélez, L. A., and Sabree, Z. L. (2020). Cultivable, Host-Specific Bacteroidetes Symbionts Exhibit Diverse Polysaccharolytic Strategies. Appl. Environ. Microbiol. 86. doi:10.1128/AEM.00091-20. doi:10.1128/AEM.00091-20.

28. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018 Jan 4;46(D1):D8-D13. doi: 10.1093/nar/gkx1095. PMID: 29140470; PMCID: PMC5753372.

29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

30. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019 Apr;37(4):420-423. doi: 10.1038/s41587-019-0036-z. Epub 2019 Feb 18. PMID: 30778233.

31. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427-D432. doi: 10.1093/nar/gky995. PMID: 30357350; PMCID: PMC6324024.

32. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018 Jul 20;430(15):2237-2243. doi: 10.1016/j.jmb.2017.12.007. Epub 2017 Dec 16. PMID: 29258817.

33. Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinformatics. 2020 Dec;72(1):e108. doi: 10.1002/cpbi.108. PMID: 33315308.

34. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019; 47(W1):W636–W41. https://doi.org/10.1093/nar/gkz268 PMID: 30976793

35. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018 Jun 1;35(6):1547-1549. doi: 10.1093/molbev/msy096. PMID: 29722887; PMCID: PMC5967553.

36. Bek-Thomsen M, Poulsen K, Kilian M. Occurrence and evolution of the paralogous zinc metalloproteases IgA1 protease, ZmpB, ZmpC, and ZmpD in *Streptococcus pneumoniae* and related commensal species. mBio. 2012; 3(5):e00303-12.

37. Dong R, Pan S, Peng Z, Zhang Y, Yang J. (2018). mTM-align: a server for fast protein structure database search and multiple protein structure alignment, Nucleic Acids Research, 46: W380–W386

38. Dong R, Peng Z, Zhang Y, Yang J. (2018). mTM-align: an algorithm for fast and accurate multiple protein structure alignment, Bioinformatics, 34: 1719-1725

39. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018; 46(W1):W296–W303. https://doi.org/10.1093/nar/gky427 PMID: 29788355

40. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020 Jan 8;48(D1):D265-D268. doi: 10.1093/nar/gkz991. PMID: 31777944; PMCID: PMC6943070.

41. Blue CE, Paterson GK, Kerr AR, Bergé M, Claverys JP, Mitchell TJ. ZmpB, a novel virulence factor of *Streptococcus pneumoniae* that induces tumor necrosis factor alpha production in the respiratory tract. Infect Immun. 2003 Sep;71(9):4925-35. doi: 10.1128/iai.71.9.4925-4935.2003. PMID: 12933834; PMCID: PMC187332.

42. Cao Z, Casabona MG, Kneuper H, Chalmers JD, Palmer T. The type VII secretion system of *Staphylococcus aureus* secretes a nuclease toxin that targets competitor bacteria. Nat Microbiol. 2016 Oct 10;2:16183. doi: 10.1038/nmicrobiol.2016.183. PMID: 27723728; PMCID: PMC5325307.

43. Aly KA, Anderson M, Ohr RJ, Missiakas D. Isolation of a Membrane Protein Complex for Type VII Secretion in *Staphylococcus aureus*. J Bacteriol. 2017 Oct 31;199(23):e00482-17. doi: 10.1128/JB.00482-17. PMID: 28874412; PMCID: PMC5686593.

44. Trastoy B, Lomino JV, Pierce BG, Carter LG, Günther S, Giddens JP, Snyder GA, Weiss TM, Weng Z, Wang LX, Sundberg EJ. Crystal structure of *Streptococcus pyogenes* EndoS, an immunomodulatory endoglycosidase specific for human IgG antibodies. Proc Natl Acad Sci U S A. 2014 May 6;111(18):6714-9. doi: 10.1073/pnas.1322908111. Epub 2014 Apr 21. PMID: 24753590; PMCID: PMC4020096.

45. Najmudin S, Guerreiro CI, Carvalho AL, Prates JA, Correia MA, Alves VD, Ferreira LM, Romão MJ, Gilbert HJ, Bolam DN, Fontes CM. Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. J Biol Chem. 2006 Mar 31;281(13):8815-28. doi: 10.1074/jbc.M510559200. Epub 2005 Nov 28. PMID: 16314409.

46. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015 Jun;10(6):845-58. doi: 10.1038/nprot.2015.053. Epub 2015 May 7. PMID: 25950237; PMCID: PMC5298202.

47. Lin IH, Hsu MT, Chang CH. Protein domain repetition is enriched in Streptococcal cell-surface proteins. Genomics. 2012 Dec;100(6):370-9. doi: 10.1016/j.ygeno.2012.08.001. Epub 2012 Aug 22. PMID: 22921469.

48. Zhou M, Wu H. Glycosylation and biogenesis of a family of serine-rich bacterial adhesins. Microbiology (Reading). 2009 Feb;155(Pt 2):317-327. doi: 10.1099/mic.0.025221-0. PMID: 19202081.

49. Paterson GK, Mitchell TJ. The biology of Gram-positive sortase enzymes. Trends Microbiol. 2004 Feb;12(2):89-95. doi: 10.1016/j.tim.2003.12.007. PMID: 15036325.

Acknowledgments

I want to thank everyone who helped me in completing my thesis. Dr. Samantha King mentored me extensively during the years I have worked in her lab and with her guidance, I have been able to improve my practical and computational skills. Dr. Meztlli Gaytán has also been infallible as she assisted me through many steps of my research. Both Dr. King and Dr. Gaytán have been extremely helpful throughout my in person and remote career in the King Lab.

I would also like to thank Dr. Arturo Vera-Ponce de León for creating phylogenetic trees used in this thesis and Dr. Mogens Kilian for sharing his *zmp* data.

Thank you to my thesis committee, Dr. Matthew Anderson and Dr. Natacha Ruiz. Dr. Ruiz also helped extensively is coordinating my research classes, which allowed me to work in the King Lab.

Lastly, thank you to the Undergraduate Research Office for providing funding through the Autumn 2020 Undergraduate Research Scholarship and an opportunity to present during the 2021 Virtual Spring Undergraduate Research Festival.