Comparative genomics of *Streptococcus oralis* identifies large scale homologous recombination and a genetic variant associated with infection

Luke R. Joyce[1]*, Madison A. Youngblom[2,3]*, Harshini Cormaty[1], Evelyn Gartstein[1], Katie E. Barber[4], Ronda L. Akins[5], Caitlin S. Pepperell[3,6#], and Kelli L. Palmer[1#]

*Authors contributed equally, listed in alphabetical order

[1] Department of Biological Sciences, The University of Texas at Dallas, Richardson, TX, 75080

[2] Microbiology Doctoral Training Program, University of Wisconsin-Madison

[3] Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, 53706

[4] Department of Pharmacy Practice, University of Mississippi School of Pharmacy, The University of Mississippi, Jackson, MS, 38677

[5] Methodist Charlton Medical Center, Dallas, TX, 75237

[6] Department of Medicine (Infectious Diseases), School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, 53706

Running title: Genomic comparison of *Streptococcus oralis.*

- Corresponding authors

Kelli L. Palmer: Kelli.Palmer@UTDallas.edu

Caitlin S. Pepperell: cspepper@medicine.wisc.edu

Abstract

The viridans group streptococci (VGS) are a large consortium of commensal streptococci that colonize the human body. Many species within this group are opportunistic pathogens causing bacteremia and infective endocarditis (IE), yet little is known about why some strains cause invasive disease. Identification of virulence determinants is complicated by the difficulty of distinguishing between the closely related species of this group. Here, we analyzed genomic data from VGS isolated from patient blood cultures with invasive infections and from oral swabs of healthy volunteers and determined the best performing methods for species identification. Using whole-genome sequence data, we characterized the population structure of a diverse sample of *Streptococcus oralis* isolates and found evidence of frequent recombination. We used multiple genome-wide association study tools to identify candidate determinants of invasiveness. These tools gave consistent results, leading to the discovery of a single synonymous single nucleotide polymorphism (SNP) that was significantly associated with invasiveness. This SNP is within a previously undescribed gene that is conserved across the majority of VGS species. Using growth in the presence of human serum and a simulated infective endocarditis vegetation model, we were unable to identify a phenotype for the enriched allele in laboratory assays, suggesting a phenotype may be specific to natural infection. These data highlight the power of analyzing natural populations for gaining insight into pathogenicity, particularly for organisms with complex population structures like the VGS.

Importance

The viridians group streptococci (VGS) are a large collection of closely related commensal streptococci, with many being opportunistic pathogens causing invasive diseases such as

bacteremia and infective endocarditis. Little is known about virulence determinants in these species, and there is a distinct lack of genomic information available for the VGS. In this study, we collected VGS isolates from invasive infections and healthy volunteers and performed whole genome sequencing for a suite of downstream analyses. We focused on a diverse sample of *Streptococcus oralis* genomes and identified high rates of recombination in the population as well as a single genome variant highly enriched in invasive isolates. The variant lies within a previously uncharacterized gene, *nrdM*, which shares homology with the anaerobic ribonucleoside triphosphate reductase, *nrdD*, and is highly conserved among VGS. This work increases our knowledge of VGS genomics and indicates that differences in virulence potential among *S. oralis* isolates is, at least in part, genetically determined.

Introduction

The viridans group streptococci (VGS) comprise a diverse collection of alpha and non-hemolytic streptococci that inhabit the oral cavity and gastrointestinal and genitourinary tracts of healthy humans (*1*). VGS are also associated with invasive disease, particularly in immunocompromised hosts, and are estimated to cause ~23% of Gram-positive bacteremia (*2*, *3*) and ~17% of infective endocarditis (IE) cases (*4*). Bacterial determinants of invasiveness among the VGS are not well understood. Research in this area is hampered by the fact that the specific species of VGS causing bacteremia and IE are infrequently determined in a clinical context due to lack of resolution of existing diagnostic microbiological tools (*5*, *6*). VITEK®2 and MicroScan allow for general assignment of isolates to VGS, and assignment of a limited number of VGS species specifically. Even a relatively newer technique in clinical diagnostics, matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS), fails to resolve certain

70 VGS species, including *Streptococcus mitis* and *Streptococcus oralis* (*7*). 16S rRNA gene
71 sequencing, multilocus sequence analysis and other genotyping schemes (*5–10*), and GyrB
72 typing (*11*) are commonly used methods for molecular VGS identification but are not generally
73 employed in clinical laboratories.
74
75 Retrospective studies using molecular approaches have determined that in addition to being
76 present as oral commensals in healthy individuals, *S. mitis* and *S. oralis* stand out among VGS
77 as major causative agents of VGS bacteremia and IE (*8*, *12–15*)**.** *S. mitis* and *S. oralis* are
78 members of the mitis group streptococci, a subgroup within the VGS that is closely related to the
79 major human pathogen *Streptococcus pneumoniae* (*16*, *17*). A recent study in oncology patients
80 demonstrated that *S. mitis* (58%) and *S. oralis* (19%) were the most frequently identified species
81 in VGS infections over ~1.5 years (*8*). While *S. mitis* and *S. oralis* are a significant burden on
82 immunocompromised patients, the mechanisms of virulence within these species have not been
83 fully elucidated. More specifically, it is not known whether all members of these species have
84 equal pathogenic potential, or whether some strains have a higher propensity for causing
85 invasive disease than others.
86
87 In this study, we aimed to investigate the mechanisms of invasiveness among VGS by
88 characterizing species diversity of presumptive VGS obtained clinically from bacteremia and
89 endocarditis patients. Clinical isolate genomes obtained in this study were supplemented by
90 existing genome sequences with curated metadata in public databases and genome sequences
91 of oral isolates collected from healthy volunteers. Our results support metagenomic sequence
92 binning as a high-resolution tool for differentiating species within the VGS grouping. After a

93 preliminary analysis in which we determined species designations for clinical isolates that were
94 diagnosed as VGS, we focused our study on *S. oralis* as a prominent cause of invasive infection.
95 Within a large sample of *S. oralis* isolates from all three described sub-species (*18*) (*subsp.*
96 *dentisani*, *subsp. tigurinus*, and *subsp. oralis*) we found high levels of diversity and strikingly high
97 recombination rates. We used multiple genome-wide association study (GWAS) methods to test
98 the hypothesis that specific genetic variations are associated with invasive infection (when
99 compared to commensal isolates) among *S. oralis* isolates. We discovered a SNP in a previously
100 uncharacterized gene that is significantly enriched in invasive isolates compared to non-invasive
101 isolates. The contribution of this novel locus to growth with human serum and in a simulated
102 infective endocarditis vegetation model (*19*) was assessed, although we were unable to identify
103 a phenotype for a gene knockout or for either allele of the significant variant under the conditions
104 tested. This work has 1) increased the genomic information available for VGS strains, 2)
105 described population structure and large-scale homologous recombination within the *S. oralis*
106 species, and 3) provided evidence that the propensity for virulence in *S. oralis* is at least in part
107 genetically determined.
108

109 **Results**

110 **GyrB typing and Kraken efficiently speciate *S. mitis* and *S. oralis.***

111 VGS isolates from Dallas, Texas and Jackson, Mississippi areas were collected from clinically
112 confirmed VGS bacteremia and infective endocarditis patients. Isolates were initially
113 characterized using either VITEK®2 or MicroScan platforms. A total of 66 clinical isolates were
114 successfully sub-cultured in the laboratory, however two isolates were incorrectly identified as
115 *Streptococcus* spp.: one *Enterococcus faecalis*, and one *Aerococcus urinae* strain, resulting in

116 64 presumptive VGS strains collected (Table S1). To compare isolates causing invasive disease
117 against commensal isolates, healthy volunteers were recruited for oral swab collection on the
118 University of Texas at Dallas campus, for a total of 81 VGS isolates (Table S1).
119
120 As has previously been reported, clinical methods of VGS species identification (VITEK®2 and
121 MicroScan) were not effective for delineating closely related species, particularly within the mitis
122 group; however, they were almost always correct in identifying an isolate as a part of the VGS
123 (Table S1). The taxonomy of the clinical and oral isolates was analyzed by 16S rRNA gene
124 sequencing, GyrB typing (*11*), and analysis of Illumina whole genome sequencing (WGS) data
125 using Kraken (*20*). While we expected that WGS would provide the clearest results, we also
126 wanted to assess the utility of the GyrB typing scheme for a diverse set of VGS species, as this
127 method is cheaper, faster and more feasible for researchers without computational experience
128 and resources. Using these three different methods of species identification, we classified
129 isolates into VGS groups using the taxonomy described by Facklam (*21*). All six of the major
130 VGS groups (*1*) were represented in our sample of 81 isolates, with two isolates not fitting into
131 any of these six groups (Figure 1A). 16S rRNA sequencing was frequently unable to provide
132 resolution to the species level but was often able to identify which group the isolate belonged to.
133 GyrB typing (*11*) and Kraken (*20*) were both effective in distinguishing *S. oralis* and *S. mitis*,
134 however GyrB typing is not generally accurate for other VGS groups (Figure 1A).
135
136 One downside of species identification with Kraken is it seems unable to distinguish between
137 anginosus group isolates: *S. anginosus* and *S. constellatus* are species within the anginosus
138 group (also known as the milleri group) of the VGS and are recognized as abscess-causing

139 bacteria, with more recent data suggesting their emergence as uropathogens (*1*, *22*, *23*). Kraken
140 identified 19 *S. anginosus* isolates, and via GyrB typing a total of 20 anginosus group isolates
141 were identified (12 *S. anginosus* and 8 *S. constellatus*). These data suggest that *S. mitis* and *S.*
142 *oralis* are accurately speciated with either Kraken or GyrB typing, yet anginosus group species
143 may be better identified by GyrB typing.
144
145 One drawback to species identification with GyrB typing is the current scheme is limited in the
146 species it can distinguish (Figure 1A). To further assess the relative functionality of these two
147 methods we inferred a phylogenetic tree from GyrB sequences and mapped species as defined
148 by Kraken onto the tree (Figure 1B). The added benefit of a phylogeny of GyrB sequences is the
149 visualization of additional species not included in the typing scheme (e.g. species in the bovis
150 group; Figure 1B) which when combined with Kraken output allowed us to confidently assign
151 species to most isolates. There was only a single instance where Kraken may have mis-identified
152 the species: an isolate identified by Kraken as *S. cristatus* (a member of the mitis group) appears
153 to cluster with *S. sanguinis* isolates (of the sanguinis group) on the GyrB phylogeny (Figure 1B).
154 Additionally, combining data from Kraken and GyrB typing showed that 2 out of the 3 isolates
155 identified as *S. sp. oral taxon 431* by Kraken were identified as *S. infantis* (a member of the mitis
156 group) by GyrB typing, and all 3 cluster together on the GyrB phylogeny, indicating that *S. infantis*
157 is likely the correct designation (Figure 1B).
158
159 Overall, our data show that closely related *S. mitis* and *S. oralis* can be speciated accurately via
160 either Kraken or GyrB typing; however, anginosus group isolates are better distinguished by
161 GyrB typing. For datasets suspected to contain a diverse sample of different VGS species,

162 species identification using Kraken with whole genome sequence data appears the most robust.
163 GyrB typing is an accurate method for making distinctions within species subsets, and a
164 phylogeny of GyrB sequences may also provide additional information. Our results support
165 previous assertions that 16S rRNA sequencing is not an effective method for distinguishing the
166 closely related species of the VGS (*1*).
167
168 When we separated our sample into isolates from blood cultures and those from the mouths of
169 healthy persons, we saw that invasive isolates spanned all 6 groups while isolates from oral
170 isolates were less diverse and only originated from the mitis and salivarius groups (Figure 1C).
171 The lack of diversity in the oral isolates is due to the use of Mitis-Salivarius Agar which inhibits
172 Gram-negative and most Gram-positive bacterial growth due to the presence of inhibitory
173 nutrients; however, *S. mitis, S. salivarius,* and enterococci will still grow and produce different
174 colony morphologies to allow preferential selection (*24*). It has been reported that *S. mitis* is the
175 predominant species found in healthy oral microbiomes (*25–30*) with other commonly found VGS
176 species including *S. oralis* (*27*, *28*)*, S. sanguinis* (*28*, *29*)*,* and *S. salivarius* (*29*)*.*
177
178 **High diversity among commensal and invasive *S. oralis* strains**
179 *S. oralis* was the predominant single species (20/81 isolates) in our sample, and so to identify
180 possible genetic variants associated with invasive infection, we gathered a larger sample of *S.*
181 *oralis*. We obtained all isolates labeled as *S. oralis* and *S. mitis* from NCBI (see Methods for
182 inclusion criteria) and used Kraken, a core genome phylogeny, and accessory gene content to
183 confirm *S. oralis* isolates and identify mislabeled isolates. We found that a core genome
184 alignment produced by Roary (*31*) effectively delineates *S. oralis* from *S. mitis*, as do patterns

185 of accessory gene content (Figure S1). We were able to identify 11 isolates mis-labeled as *S.*
186 *mitis* within NCBI databases, which we identified as *S. oralis* using these methods (Table S2).
187 We ended up with a total of 108 *S. oralis* isolates: 57 oral commensals and 51 from invasive
188 infections (including bacteremia and IE) that we refer to as "oral" and "blood" isolates,
189 respectively (Table S2). After assembly and annotation of all genomes, we performed a
190 pangenome analysis with Roary and found that our sample has a core genome of only 801
191 genes – with an average genome size of 1898 genes. This means that a significant proportion
192 of genes (57%) encoded by an individual isolate are variable accessory genes (Figure 2A). The
193 accessory gene content in our sample of *S. oralis* isolates is highly diverse, with 88% of all genes
194 found in our sample at a frequency of 14% or less, and > 19,000 genes identified in the pan
195 genome (Figure 2A). This high level of gene content diversity is mirrored in the core genome
196 phylogeny, which has long terminal branch lengths indicating high variability between core
197 genome sequences (Figure 2B). Isolates from commensal and invasive sources are
198 interspersed on the phylogeny and do not generally form monophyletic clades, nor are the
199 commensal or invasive phenotypes associated any sub-clades within the phylogeny (Figure 2B).
200
201 We identified three sub-clades within our core genome phylogeny (Figure 2B) that we found to
202 correspond with the three previously described sub-species of *S. oralis*: *subsp. dentisani, subsp.*
203 *tigurinus,* and *subsp. oralis* (*18*) by cross-referencing our tree with the subspecies of some of
204 the publicly available isolates in our dataset (Table S2). Average nucleotide identity (ANI) values
205 for core genome sequences within each subspecies show that *subsp. oralis* is the most
206 conserved, followed by *subsp. tigurinus* and *subsp. dentisani* (Figure 2C). *Subsp. dentisani* is
207 unique in that it has ANI values resembling those measured between isolates of different

208 subspecies – this indicates that levels of diversity within this subspecies are similar to those
209 found between subspecies (Figure 2C). Following the trends we identified in core genome
210 diversity levels, the pangenome of *subsp. oralis* isolates appears more conserved (more core
211 genes, fewer accessory genes) than either *subsp. tigurinus* or *subsp. dentisani* (Figure 2D).
212 Higher core genome and pangenome diversity levels among *subsp. tigurinus* and *subsp.*
213 *dentisani* isolates could indicate that these isolates have access to more diverse partners for
214 horizontal gene transfer (HGT).
215

216 **High levels of recombination among *S. oralis* isolates**

217 Viridans group streptococci (VGS) are known for being naturally competent (*32*) and for evolving
218 rapidly via widespread homologous recombination (*33*, *34*). We characterized the signatures of
219 recombination in the core genome of *S. oralis* using Gubbins (*35*) and identified extreme
220 amounts of recombination: 99.9% of the core genome is within a predicted recombinant fragment
221 in at least one isolate (Figure 3A). Additionally, we noted that recombinant fragments are not
222 shared across multiple isolates but are usually present in only a handful of isolates (Figure 3B).
223 Using ClonalFrameML (*36*) we calculated the *r/m* value - the ratio of SNPs imported via
224 recombination (r) to those introduced randomly (m) – and found that with an *r/m* of 5.77, genetic
225 diversity in our sample is ~6x more likely to be introduced via recombination. This is slightly less
226 than the notoriously recombinogenic *S. pneumoniae* [*r/m* = ~7; (*33*)] but much higher than other
227 IE causing bacteria such as *S. aureus* [*r/m* = <1; (*37*)].
228

229 **Genome wide association study reveals variant associated with invasiveness**

230 With a large sample balanced between our phenotypes of interest, we used multiple genome-
231 wide association study (GWAS) methods to identify genetic variants associated with invasive
232 disease, which included isolates from both IE and bacteremia patients. We started by looking for
233 associations between accessory genes and invasiveness using Scoary (*38*), which yielded no
234 significant results. This is perhaps not surprising given that accessory gene content is so diverse
235 in the sample and individual accessory genes are thus unlikely to be shared by a large proportion
236 of isolates (Figure 1A).
237
238 To identify core genome variants associated with invasiveness, we started with a F_{ST} outlier
239 analysis, which delineates allele frequency differences between sub-populations and identifies
240 variants with extreme measures of differentiation. We defined sub-populations of our sample as
241 being of "oral" or "blood" source and used vcflib (https://github.com/vcflib/vcflib) to calculate Weir
242 and Cockerham's F_{ST} (wcF_{ST}, abbreviated to F_{ST}) for bi-allelic SNPs (n=67,026) in the core
243 genome alignment (Figure 4A). We identified a single SNP with a significant F_{ST} value, indicating
244 significant allele frequency differences between oral and blood sub-populations of *S. oralis*
245 (Figure 4A). The SNP of interest lies within a gene originally annotated as a homologue of *nrdD*,
246 an anaerobic ribonucleoside triphosphate reductase present in many streptococci, involved in
247 synthesis of deoxyribonucleotides under anaerobic conditions (*39*). However, further inspection
248 of the sequence revealed that the canonical *nrdD* gene was annotated separately in our *S. oralis*
249 isolates, and that the novel protein was shorter than *nrdD*. The novel protein does share some
250 sequence features with *nrdD* including an ATP cone domain, and so we will refer to the novel
251 locus as *nrdM*. Given the amount of recombination present in our sample (Figure 3A) we
252 validated the results of our F_{ST} outlier analysis using GWAS methods specifically designed for

253 use with microbial genomes. We used two programs: treeWAS (*40*), which corrects for the
254 presence of recombination, and BugWAS (*41*), which identifies lineage effects and controls for
255 population structure. Results from both of these tools replicated the results of our F_{ST} outlier
256 analysis with both methods returning a single variant that was significantly associated with the
257 phenotype: the same SNP in *nrdM* (Figure S2).
258
259 The SNP of interest in *nrdM* is a synonymous mutation (I78I), and mapping the nucleotide alleles
260 of the mutation (C233T) onto the core genome phylogeny shows interspersal of alleles on the
261 tree, with some structure in *subsp. dentisani* and *subsp. tigurinus* (Figure S3). Most interesting
262 is the striking association between isolates from invasive infections and the *nrdM* allele: out of
263 51 invasive isolates in our sample, 39 encode the *nrdM* SNP associated with invasiveness
264 (Figure S3; Table S2). A multiple sequence alignment of NrdM from our sample of *S. oralis* strains
265 reveals high protein sequence conservation (Figure S4), and so to determine whether *nrdM* is
266 more or less conserved than other core genes we calculated the average $\pi N/\pi S$ (piN/piS) values
267 for each core gene (n=801). The distribution of gene-wise values of piN/piS indicate the majority
268 of genes are evolving under relatively strong purifying selection, including *nrdM*, which had a
269 piN/piS value slightly below the mean (Figure 4B). This is not surprising given that recombination
270 strengthens the efficiency of selection by enabling rapid removal of deleterious mutations (*42*).
271
272 **Selection on *nrdM* variant**
273 Given the strong association between the *nrdM* variant and invasiveness, we wondered if this
274 SNP is under positive selection in our sample. One method for identifying positive selection is
275 the identification of homoplastic mutations, *i.e*. mutations that arose more than once on the

phylogeny, which we have used previously to screen for drug resistance loci in *Mycobacterium tuberculosis* (*43*). We used TreeTime (*44*) to identify homoplastic mutations in our sample and found numerous homoplastic mutations – this is expected since intergenomic recombination can produce homoplasies by lateral transfer of sequence variants. The variant in *nrdM* arose 20 times on the phylogeny, which was in the 85th percentile of mutation multiplicity (Figure S5). This finding, along with gene-wise piN/piS values (Figure 4B), suggests that selection pressures are similar at this locus to others in the genome; however, further analysis with a larger sample size could elucidate more subtle signs of selection.

***nrdM* is conserved among VGS species**

In order to investigate the presence of *nrdM* homologues in other *Streptococcus* spp. we used Blast to search annotated genes in all newly sequenced isolates and found *nrdM* to be present in all VGS species in our original sample, except for two bovis group species: *S. lutetiensis* and *S. pasteuranius* (Figure 5A). Additionally, the position of the variant associated with invasiveness in *S. oralis* is conserved, and we were able to identify variation in *nrdM* alleles between different species. As in *S. oralis*, NrdM is highly conserved at the protein level across the VGS species in our sample, indicating the gene may have a conserved function across multiple species (Figure 5B).

Deletion of *nrdM* does not affect *in vitro* growth

To investigate the impact the *nrdM* orphan allele has on growth in the two infective endocarditis strains *S. oralis* 1648 (SO48) and mitis group *Streptococcus* 1643 (SM43), we deleted *nrdM*, locus tags MP387_03665 and FD735_06230, respectively, from each strain, via homologous

recombination (Table 1). SO48 possesses the *nrdM*-C ("C" = commensal) allele, and SM43 possesses the *nrdM*-I ("I" = invasive) allele. Knockout mutants were confirmed via Sanger sequencing. No difference in growth was observed for knockout mutants compared to wild-type in Todd-Hewitt broth (THB) (Figure S6). Further, *nrdM* allele swaps from SO48 and SM43 were placed back into the genome of SO48Δ*nrdM* and SM43Δ*nrdM* to generate Δ*nrdM*::*nrdM*-I or Δ*nrdM*::*nrdM*-C strains in each strain background (Table 1). Since the presence of human serum has been shown to impact mitis group streptococci physiology (*45*, *46*), we assessed the impact of *nrdM* during growth with human serum by performing growth curves of allele replacement strains in chemically defined medium supplemented with 5% v/v human serum (Figure S6). No difference in growth profile was observed between alleles in either strain background. Thus, the *nrdM* locus does not impact growth in rich laboratory medium or chemically defined medium supplemented with 5% v/v human serum *in vitro*.

nrdM has no impact on growth in a simulated infective endocarditis vegetation model

Finally, the allele swapped strains were investigated using the pharmacological *in vitro* simulated infective endocarditis vegetation model (SIEVM). We reasoned that if *nrdM*-I confers enhanced fitness in simulated endocardial vegetations, we would observe significantly different vegetation CFUs for SO48Δ*nrdM*::*nrdM*-I versus SO48Δ*nrdM*::*nrdM*-C. The SM43 strain was similarly tested, to assess the effect of *nrdM* allele swapping in a closely related but different (i.e. non-*oralis*) genetic background. Strains were inoculated into vegetation clots at ~ 1 x10^8 CFU, incubated in the chemostat model with Mueller Hinton Broth (MHB) in parallel for 48hrs (see Material and Methods). At 4, 24, and 48 h post inoculation, 4 clots from each model were removed for CFU enumeration (Figure 6). No significant difference was observed during growth

in the SIEVM between *nrdM* allele swap in either strain background. Ultimately, these data and the growth curve data together show that under these *in vitro* conditions, *nrdM* is not essential, and no phenotype could be assigned to either of the *nrdM* alleles.

Discussion

The viridans group streptococci (VGS) are a large collection of closely related *Streptococcus* spp. that inhabit the oral cavity and gastrointestinal and genitourinary tracts of humans as commensals, but can invade other tissues to cause severe diseases such as bacteremia and infective endocarditis (*1*). Our understanding of the pathogenesis of VGS disease is limited by lack of knowledge surrounding the genetic and environmental conditions that facilitate a switch from commensalism to pathogen. Additionally, clinical detection and identification of bacterial infections is critical for patient care and recovery, especially in neutropenic and immunocompromised patients. A major etiological agent of disease in this patient demographic are the VGS, especially *S. mitis* and *S. oralis.* Yet, clinical methods of speciation are still largely inaccurate, and even newer methods like MALDI-TOF MS struggle to differentiate between closely related species such as *S. oralis* and *S. mitis* (*7*). In this study, we collected and whole genome-sequenced a variety of invasive and non-invasive VGS isolates, compared bioinformatic methods for delineating closely related species, and identified a SNP that may predispose certain isolates of *S. oralis* to invasive disease.

Complex population structure of *S. oralis*

We characterized the population structure of 81 *S. oralis* isolated from healthy oral microbiota and from blood, and found high levels of diversity, both in the core (Figure 2B-C) and accessory

genomes (Figure 2A,D). We found that using our methods, previously described sub-species of *S. oralis* corresponded to sub-clades we identified in a phylogeny inferred from core genome sequences (Table S2). We also identified *subsp. oralis* as being the least diverse of the three subspecies, with respect to both gene content and sequence variation in the core genome (Figure 2). We assert that recombination is likely the primary mechanism for generating and maintaining diversity in this species, as 99.9% of the *S. oralis* core genome has been affected by recombination (Figure 3). The rarity of a given recombinant fragment in our sample also indicates that *S. oralis* isolates are likely participating in HGT with diverse species – made possible by the fact that they inhabit complex communities. We observed a mean piN/piS value across *S. oralis* core genes of 0.039 which is an order of magnitude lower than comparator species such as *S. aureus* for which values of 0.32 (*47*) and 0.55 (*48*) have been reported, and lung-colonizing *P. aeruginosa* from cystic-fibrosis patients that have a mean of 0.14. (*49*). This indicates that *S. oralis* core genes are under strong purifying selection. We recently identified a similar phenomenon among environmental isolates of *Mycobacterium abscessus* that are also highly recombinogenic and manifest high amounts of synonymous diversity in their core genomes (*50*). This suggests that like *M. abscessus*, *S. oralis* inhabits environments alongside diverse microbial species where high rates of recombination enable ready acquisition of novel genetic material and rapid removal of deleterious mutations.

Switching from commensal to pathogen

Our identification of extremely high recombination rates within *S. oralis* points to the unique genomic features characterizing this VGS species. High rates of recombination allow the species to adapt to fluctuating environments encountered within the human body, and potentially enable

invasion of new niches, including pathogenic niches. We used three different GWAS methods that replicably identified a strong association between a synonymous SNP within an undescribed protein, NrdM, and invasive disease (Figure 4, Figure S2). The association was robust across diverse isolates of all three subspecies (Figure S3). We also found that *nrdM* was conserved among *S. oralis* isolates (Figure S4) and is one of only ~440 genes conserved across all VGS species in our sample indicating that it is likely to play an important, and perhaps a similar, role across species. Utilizing *in vitro* growth in the presence of human serum and a simulated infective endocarditis vegetation model, we compared *nrdM* knockout strains and allele-swapped invasive and commensal isolate alleles in two strain backgrounds. No phenotype was observed under the tested conditions, suggesting that if NrdM is contributing to invasive disease, it requires different conditions (possibly specific to the *in vivo* environment) for a potential phenotype to be observed. NrdD, which shares the same ATPase cone domain as NrdM, is required for anaerobic growth in *S. sanguinis* (*39*), and, notably, disruption of this gene results in attenuated virulence (*51*). Long thought to have neutral fitness effects, synonymous mutations are increasingly recognized as having significant effects on bacterial fitness, for example by impacting gene expression and protein folding (*52–56*). We believe that further study of this synonymous variant could reveal fitness effects and that adaptation within *nrdM* may affect a multitude of phenotypes that aid in transition from the oral cavity to the bloodstream, although further investigation is necessary.

A major hurdle to successful genomic analyses in VGS is the correct classification of the different species. It is possible that incorrect species identification has contributed to the lack of clear results from studies looking at the genetic determinants of virulence in VGS species. One such

study by Rasmussen *et. al.* used a sample of both *S. mitis* and *S. oralis* to search for known virulence factors, which they found at varying frequencies, indicating that genetic differences may be responsible for variability in virulence among strains (*57*). Similar to our sample of *S. oralis* isolates (Figure 2B), studies of the closely related VGS *S. sanguinis* and *S. gordonii* were unable to identify phylogenetic patterns based on invasive disease (*58*, *59*). These studies, however, were unable to identify specific genomic variants associated with invasiveness. What our work and the work of others indicates, is that virulence properties differ between VGS species, despite being so closely related and often causing similar disease. This is further illustrated by the fact that although *nrdM* is conserved in *S. mitis,* the association between the invasive allele identified in *S. oralis* and pathogenicity in *S. mitis* is not significant (data not shown). Going forward, analyses using larger samples of well-defined, individual species will provide better resolution for identifying the genetic determinants of virulence in the VGS.

Materials and methods

Collection of clinical Isolates

Clinical strains obtained from routine blood cultures and identified as viridans group streptococci were stored in the microbiology laboratory at each hospital (Methodist Health System [MHS] and University of Mississippi Medical Center [UMMC]). Respective site investigators reviewed patient clinical information to confirm invasive infection (i.e., bacteremia or endocarditis), then de-identified specimens and provided blinded isolates to the University of Texas at Dallas (UTD) laboratory for study. Specimens were not included for further analysis if the isolate was deemed a contaminant and did not require antibiotic therapy as determined by the treating physician. Site

investigators obtained approval from respective Institutional Review Boards (MHS UTD IRB 18-121 and UMMC 2018-0068).

Processing of clinical isolates

Clinical isolates were struck on Mitis-Salivarius Agar (MSA) (BD Bacto and incubated overnight in 37˚C and 5% CO_2. MSA plates were observed for homogenous colony morphology, and a single colony was inoculated into 10 mL THB for overnight growth at 37˚C and 5% CO_2. If more than one colony morphology was identified on MSA plates, broth cultures were made from each colony morphology. Overnight cultures were stored at -80°C in 25% glycerol. The remaining culture volume was pelleted at 4,280 x *g* in a Sorvall RC6+ floor centrifuge and genomic DNA was extracted using Qiagen DNeasy Blood and Tissue Kit per manufacturer protocols, with minor modifications as described in (*6*).

Collection of oral swab samples

Oral swabs were obtained from healthy adult volunteers at the UTD campus (UTD IRB 17-170). The following exclusion criteria were applied: previous history of bacteremia or endocarditis; recent antibiotic exposure (prior 30 days); history of periodontal disease; personal or family history of immunocompromise. No participants were excluded based on these criteria. The volunteer was asked to rinse their mouth with sterile saline, and then self-swab their teeth and tongue with a sterile swab (Puritan). Swabs were stored at 4˚C until processing.

Processing of oral swabs

Oral swabs were processed as described above. Briefly, swabs were struck onto MSA plates and grown overnight. After incubation, the MSA plates were observed and colony morphologies consistent with *Streptococcus mitis* and *S. oralis* were selected for overnight growth in THB. In addition, approximately 3 random colonies of different morphologies were also selected for overnight growth. Cultures were processed as described above.

16S rRNA sequence analysis and GyrB typing

PCR reactions were performed using Taq polymerase (New England Biolabs) with primer sequences in Table S3. 16S rRNA genes were amplified using universal primers 8F and 1492R (*60*). The DNA Gyrase B gene (*gyrB*) was amplified using previously reported primers (Table S3). PCR reactions were analyzed by agarose gel electrophoresis and purified using the GeneJET PCR purification kit (Thermo Fisher) per the manufacturer protocols. Products were sequenced at the Massachusetts General Hospital DNA Core. 16S rRNA sequences were trimmed using Geneious R11 (https://www.geneious.com) allowing a maximum of 10 low quality bases and 6 ambiguities. Trimmed sequences were used as queries for NCBI Blastn against the 16S ribosomal RNA sequences (Bacteria and Archaea) database and species were assigned only when the forward and reverse sequencing reactions had the same top Blastn result (Table S1). GyrB nucleotide sequences were translated, and amino acid sequences pairwise aligned to *Streptococcus mitis* ATCC 49456 GyrB (locus tag SM12261_0755). Amino acid variations were identified using the method of Galloway-Peña et al. (*11*).

Ilumina sequencing

Sequencing was performed at The University of Texas at Dallas Genome Core using Illumina Nextseq 500 platform with a mid-output 300 cycle of 2 x 75 bp paired-end reads for clinical isolates S1-S31, or 2 x 150 bp paired-end reads for all other isolates.

Genome assembly & annotation

Using the raw sequencing data from newly sequenced clinical and oral isolates (Table S1), species identification was additionally performed using Kraken2 (*20*) Raw data was quality checked and trimmed using FastQC v0.11.8 (*61*) and TrimGalore v0.6.4, (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore), respectively. Contigs were assembled using SPAdes v3.13.0 with default parameters (*62*). Assemblies were checked for quality using Quast v5.0.2 (*63*) filtering out contigs shorter than 500bp or with coverage lower than 5x, as well as confirming all assemblies had a N50 > 50,000bp. Contigs were annotated using Prokka v1.13.3 (*64*) before a pangenome analysis was performed with Roary v3.12.0 using a blastp identity threshold of 75% (*31*). Using a nucleotide sequence alignment of GyrB (as clustered by Roary) a phylogenetic tree was made using FastTree v2.1.9 (*65*) and visualized in R with ggtree (*66*). GyrB typing was confirmed using these sequences with a custom script (code available at https://github.com/myoungblom/VGS_GWAS.git) assigning species based on the scheme outlined by Galloway-Peña et al. (*11*). *nrdM* sequences from all newly sequenced VGS isolates were identified with Blastp using the *S. oralis* NrdM amino acid sequence as the query sequence.

***S. oralis* genome collection for GWAS analyses**

From our sample of clinical and oral isolates, *S. oralis* made up the largest part of our sample so we decided to proceed with analyses of just this species. To create a dataset large enough for a powerful genome-wide association study (GWAS) we identified all *S. oralis* and *S. mitis* isolates from NCBI (NCBI SRA & Assembly databases accessed June 2019) with the proper metadata indicating they were isolated from the mouth (e.g oral cavity, dental plaque, dental biofilm, etc) or from blood (e.g. infective endocarditis, bloodstream infection, blood, etc). We assumed isolates uploaded to NCBI with various "oral" sources were all commensal and all those from "blood" were from an invasive infection (Table S2). We chose to start with both *S. oralis* and *S. mitis* because these species are so closely related, they are often mistaken for each other and uploaded to NCBI under the wrong species (Table S2) as has previously been reported (*10*). We pulled out the true *S. oralis* isolates using Kraken and GyrB typing as described above. Samples for which raw sequence data was available were assembled as described above, and then annotated with the remainder of assemblies downloaded from NCBI (Table S2). Note that since the time of data collection, three of the assemblies used in this dataset have been suppressed (Table S2). We then performed a pangenome analysis on the *S. oralis* sample as described above (using a Blastp identity threshold of 95% and Prank to align core genes) and a phylogenetic tree was inferred from the resulting core genome alignment using RAxML v8.2.3 (*67*) and visualized in R with ggtree (*66*).

Recombination analyses

We identified recombinant fragments in the *S. oralis* core genome using Gubbins v2.4.1 with default parameters (*35*). Recombinant fragments were visualized alongside the core genome

phylogeny using Phandango (*68*). We used ClonalFrameML v1.11 (*36*) with default parameters to estimate *r/m*.

Population genetics statistics

Pairwise average nucleotide identity (ANI) values of *S. oralis* core genome sequences were calculated with OrthoANI (*69*). piN/piS values were calculated for all pairwise combinations for each *S. oralis* core gene were calculated using Egglib (*70*), then the average piN/piS value for each gene was calculated.

Rarefaction & accumulation plots

Rarefaction and accumulation curves for the *S. oralis* subspecies were calculated from Roary gene presence absence files. Briefly, separate pangenome analyses were performed as described above for each individual subspecies and then each dataset was iteratively subsampled to the size of the smallest dataset (n=25) and the median number of core and total genes was plotted from all iterations.

GWAS

We first queried for genetic associations with the 'invasive' phenotype in our dataset by identifying accessory gene content significantly associated with invasiveness using Scoary v1.16.6 (*38*). We then performed a preliminary GWAS of core genome variants using an F_{ST} outlier analysis. Briefly, a VCF file containing all core genome variants was made using SnpSites v2.0.3 (*71*) and reformatted using a custom script (code available https://github.com/myoungblom/VGS_GWAS.git). Then we calculated Weir and Cockerham's

F_{ST} for bi-allelic SNPs using vcflib (https://github.com/vcflib/vcflib). Using a custom script (code available https://github.com/myoungblom/VGS_GWAS.git) we permuted the phenotypes in this analysis 100x and used the maximum F_{ST} value observed in the null distribution as a cutoff to identify significant F_{ST} outliers. To validate the results of our F_{ST} outlier analysis we also used two GWAS programs designed specifically for use with microbial genomes: treeWAS (*40*) that corrects for the presence of recombination, and BugWAS (*41*) that identifies lineage effects and controls for population structure. TreeWAS was run using the recombination-adjusted phylogenetic tree made with Gubbins (see above) using 10x the number of SNPs in the core genome for the parameter "n.snps.sim". BugWAS was run using default parameters.

Homoplasy analysis

Homoplasy analysis was performed using TreeTime v0.9.0-b.2 (*44*) with default parameters.

Deletion of *nrdM*

Knockout constructs of *nrdM* in SM43 (Locus ID FD735_06230) and SO48 (Locus ID MP387_03665) were generated as previously described (*45*). Briefly, linear constructs were generated by amplifying ~2 kb regions upstream and downstream of *nrdM* using Phusion polymerase (Thermo Fisher) using primers in Table S3. SOEing PCR was used to stitch fragments together and the amplified product was assessed via agarose gel electrophoresis. Gel extraction was performed using the QIAQuick Gel Extraction kit (Qiagen). Linear constructs were transformed by natural transformation as described in (*45*). Transformation plates were incubated overnight, and putative transformant colonies were screened via PCR for the *nrdM* deletion.

547

548 ***nrdM* allele swaps**

549 Allele swap strains were generated using the same strategy as the deletion, except the linear
550 construct contained either the *nrdM*-I or *nrdM*-C allele coupled with the flanking regions for the
551 respective strain. The linear construct was transformed into SM43Δ*nrdM* and SO48Δ*nrdM.* The
552 allele swap region in transformants was amplified using primers in Table S3. Products were
553 sequenced for validation of the allele swap (Massachusetts General Hospital DNA Core).

554

555 **Growth curves**

556 Growth curves in THB were performed in biological duplicate. Wild-type and Δ*nrdM* strains were
557 grown overnight as described, then diluted to OD_{600nm} 0.05 in approximately 12 mL THB. OD_{600nm}
558 was monitored every hour using a Thermo Scientific Genesys 30 spectrophotometer. For growth
559 curves in the presence of human serum, biological triplicate overnight cultures were grown in
560 streptococcal defined medium (*45*, *46*, *72*) and diluted to OD_{600nm} 0.1 in defined medium
561 supplemented with 5% v/v human serum (Sigma-Aldrich). OD_{600nm} was monitored at 3, 6 and 24
562 hrs as described above.

563

564 **Simulated Infective Endocarditis Vegetation Model (SIEVM)**

565 Strains were inoculated from freezer stocks into 5 mL Mueller Hinton Broth (MHB) (BD Bacto)
566 and incubated overnight at 37˚C and 5% CO_2. 1 mL was expanded into 100 mL pre-warmed
567 MHB and incubated overnight as described above. The SIEVM was set up and performed as
568 previously described (*19*) using human blood products from the American Red Cross (UTD IRB
569 19MR0160). Briefly, 500 µL of pooled human cryoprecipitate (American Red Cross), 50 µL ~2

TIU/mL aprotinin (Sigma-Aldrich), ~100,000 human platelets (American Red Cross), and 10^8 CFU/g bacteria were combined in a sterile Eppendorf tube and vortexed. Sterile monofilament line was positioned before 100 µL of ~2 KU/mL high activity bovine thrombin (Sigma-Aldrich) was added to congeal the vegetation. Vegetations were placed into the glass apparatus in a 37˚C water bath, and MHB was pumped through the model at a pre-calibrated rate of 0.4 mL/min. Four vegetations were removed at designated time points for each strain, weighed, removed from the monofilament line and placed in 1.25% Trypsin solution (Sigma-Aldrich) in sterile screw cap microcentrifuge tubes (Fisher Scientific) with ~5-8 2.7 mm glass beads (BioSpec). Clots were homogenized for 10-15 min horizontally on a vortex before serial dilution and plating on THB agar plates for enumeration. CFU/g was calculated by multiplying the observed CFU/mL by the net weight of the vegetation. SIEVM was performed in biological triplicate for each strain, with four vegetations per time point per strain.

S. oralis 1648 hybrid genome assembly

Pacific Biosciences single molecule real time (SMRT) sequencing was performed by the Johns Hopkins Genome Core. The SO48 whole genome was assembled using the Unicyler assembly pipeline (*73*) combining SMRT long reads generated in this study and Illumina reads previously generated for SO48 (accession PRJNA354070; (*74*)).

Accession numbers

The SO48 whole genome sequence generated in this study has been deposited in Genbank under the accession number CP094226. Genome constructs and Illumina and SMRT sequence

reads generated in this study have been deposited in the Sequence Read Archive under the BioProject accession PRJNA817585, see Table S1.

Acknowledgements

This work was supported by R21 AI130666 from the National Institutes of Health and the Cecil H. and Ida Green Chair in Systems Biology Science to K.L.P., and by the National Science Foundation Graduate Research Fellowship to M.A.Y under grant number DGE-1747503. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. C. D. Doern, C.-A. D. Burnham, It's not easy being green: the viridans group streptococci, with a focus on pediatric clinical manifestations. *J. Clin. Microbiol.* **48**, 3829–3835 (2010).

2. C. Gudiol, M. Bodro, A. Simonetti, F. Tubau, E. González-Barca, M. Cisnal, E. Domingo-Domenech, L. Jiménez, J. Carratalà, Changing aetiology, clinical features, antimicrobial resistance, and outcomes of bloodstream infection in neutropenic cancer patients. *Clin Microbiol Infect*. **19**, 474–479 (2013).

3. M. Marín, C. Gudiol, C. Garcia-Vidal, C. Ardanuy, J. Carratalà, Bloodstream infections in patients with solid tumors: epidemiology, antibiotic therapy, and outcomes in 528 episodes in a single cancer center. *Medicine (Baltimore)*. **93**, 143–149 (2014).

4. D. R. Murdoch, G. R. Corey, B. Hoen, J. M. Miró, V. G. Fowler, A. S. Bayer, A. W. Karchmer, L. Olaison, P. A. Pappas, P. Moreillon, S. T. Chambers, V. H. Chu, V. Falcó, D. J. Holland, P. Jones, J. L. Klein, N. J. Raymond, K. M. Read, M. F. Tripodi, R. Utili, A. Wang, C. W. Woods, C. H. Cabell, International Collaboration on Endocarditis-Prospective Cohort Study (ICE-PCS) Investigators, Clinical presentation, etiology, and outcome of infective endocarditis in the 21st century: the International Collaboration on Endocarditis-Prospective Cohort Study. *Arch Intern Med*. **169**, 463–473 (2009).

5. C. Teles, A. Smith, G. Ramage, S. Lang, Identification of clinically relevant viridans group streptococci by phenotypic and genotypic analysis. *Eur J Clin Microbiol Infect Dis*. **30**, 243–250 (2011).

621 6. T. Hoshino, T. Fujiwara, M. Kilian, Use of phylogenetic and phenotypic analyses to identify
622 nonhemolytic streptococci isolated from bacteremic patients. *J Clin Microbiol*. **43**, 6073–6085
623 (2005).

624 7. J. Isaksson, M. Rasmussen, B. Nilson, L. S. Stadler, S. Kurland, L. Olaison, E. Ek, B. Herrmann,
625 Comparison of species identification of endocarditis associated viridans streptococci using rnpB
626 genotyping and 2 MALDI-TOF systems. *Diagn Microbiol Infect Dis*. **81**, 240–245 (2015).

627 8. S. A. Shelburne, P. Sahasrabhojane, M. Saldana, H. Yao, X. Su, N. Horstmann, E. Thompson, A. R.
628 Flores, Streptococcus mitis Strains Causing Severe Clinical Disease in Cancer Patients. *Emerg Infect*
629 *Dis*. **20**, 762–771 (2014).

630 9. C. J. Bishop, D. M. Aanensen, G. E. Jordan, M. Kilian, W. P. Hanage, B. G. Spratt, Assigning strains to
631 bacterial species via the internet. *BMC Biol*. **7**, 3 (2009).

632 10. I. M. Velsko, M. S. Perez, V. P. Richards, Resolving Phylogenetic Relationships for Streptococcus
633 mitis and Streptococcus oralis through Core- and Pan-Genome Analyses. *Genome Biology and*
634 *Evolution*. **11**, 1077–1087 (2019).

635 11. J. Galloway-Peña, P. Sahasrabhojane, J. Tarrand, X. Y. Han, S. A. Shelburne, GyrB Polymorphisms
636 Accurately Assign Invasive Viridans Group Streptococcal Species. *J Clin Microbiol*. **52**, 2905–2912
637 (2014).

638 12. P. Sahasrabhojane, J. Galloway-Peña, L. Velazquez, M. Saldaña, N. Horstmann, J. Tarrand, S. A.
639 Shelburne, Species-level assessment of the molecular basis of fluoroquinolone resistance among
640 viridans group streptococci causing bacteraemia in cancer patients. *Int J Antimicrob Agents*. **43**,
641 558–562 (2014).

642 13. T. Kitten, C. L. Munro, N. Q. Zollar, S. P. Lee, R. D. Patel, Oral streptococcal bacteremia in
643 hospitalized patients: taxonomic identification and clinical characterization. *J Clin Microbiol*. **50**,
644 1039–1042 (2012).

645 14. S. Chun, H. J. Huh, N. Y. Lee, Species-specific difference in antimicrobial susceptibility among
646 viridans group streptococci. *Ann Lab Med*. **35**, 205–211 (2015).

647 15. S. L. Kim, S. M. Gordon, N. K. Shrestha, Distribution of streptococcal groups causing infective
648 endocarditis: a descriptive study. *Diagnostic Microbiology and Infectious Disease*. **91**, 269–272
649 (2018).

650 16. M. Kilian, D. R. Riley, A. Jensen, H. Brüggemann, H. Tettelin, Parallel evolution of Streptococcus
651 pneumoniae and Streptococcus mitis to pathogenic and mutualistic lifestyles. *mBio*. **5**, e01490-
652 01414 (2014).

653 17. M. Kilian, K. Poulsen, T. Blomqvist, L. S. Håvarstein, M. Bek-Thomsen, H. Tettelin, U. B. S. Sørensen,
654 Evolution of Streptococcus pneumoniae and Its Close Commensal Relatives. *PLoS One*. **3**, e2683
655 (2008).

656 18. A. Jensen, C. F. P. Scholz, M. 2016 Kilian, Re-evaluation of the taxonomy of the Mitis group of the
657 genus Streptococcus based on whole genome phylogenetic analyses, and proposed
658 reclassification of Streptococcus dentisani as Streptococcus oralis subsp. dentisani comb. nov.,
659 Streptococcus tigurinus as Streptococcus oralis subsp. tigurinus comb. nov., and Streptococcus
660 oligofermentans as a later synonym of Streptococcus cristatus. *International Journal of Systematic*
661 *and Evolutionary Microbiology*. **66**, 4803–4820.

662 19. R. L. Akins, B. D. Katz, C. Monahan, D. Alexander, Characterization of High-Level Daptomycin
663 Resistance in Viridans Group Streptococci Developed upon In Vitro Exposure to Daptomycin.
664 *Antimicrob Agents Chemother*. **59**, 2102–2112 (2015).

665 20. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biology*.
666 **20**, 257 (2019).

667 21. R. Facklam, What happened to the streptococci: overview of taxonomic and nomenclature
668 changes. *Clin Microbiol Rev*. **15**, 613–630 (2002).

669 22. M. Furuichi, Y. Horikoshi, Sites of infection associated with Streptococcus anginosus group among
670 children. *J Infect Chemother*. **24**, 99–102 (2018).

671 23. M. M. Pearce, E. E. Hilt, A. B. Rosenfeld, M. J. Zilliox, K. Thomas-White, C. Fok, S. Kliethermes, P. C.
672 Schreckenberger, L. Brubaker, X. Gai, A. J. Wolfe, The female urinary microbiome: a comparison of
673 women with and without urgency urinary incontinence. *mBio*. **5**, e01283-01214 (2014).

674 24. G. H. Chapman, The isolation and testing of fecal streptococci. *Am J Dig Dis*. **13**, 105–107 (1946).

675 25. C. Pearce, G. H. Bowden, M. Evans, S. P. Fitzsimmons, J. Johnson, M. J. Sheridan, R. Wientzen, M. F.
676 Y. 1995 Cole, Identification of pioneer viridans streptococci in the oral cavity of human neonates.
677 *Journal of Medical Microbiology*. **42**, 67–72.

678 26. J. A. Aas, B. J. Paster, L. N. Stokes, I. Olsen, F. E. Dewhirst, Defining the Normal Bacterial Flora of the
679 Oral Cavity. *Journal of Clinical Microbiology*. **43**, 5721–5732 (2005).

680 27. D. L. Mager, L. A. Ximenez-Fyvie, A. D. Haffajee, S. S. Socransky, Distribution of selected bacterial
681 species on intraoral surfaces. *Journal of Clinical Periodontology*. **30**, 644–654 (2003).

682 28. S. N. Peterson, E. Snesrud, J. Liu, A. C. Ong, M. Kilian, N. J. Schork, W. Bretz, The Dental Plaque
683 Microbiome in Health and Disease. *PLOS ONE*. **8**, e58487 (2013).

684 29. E. Caselli, C. Fabbri, M. D'Accolti, I. Soffritti, C. Bassi, S. Mazzacane, M. Franchi, Defining the oral
685 microbiome by whole-genome sequencing and resistome analysis: the complexity of the healthy
686 picture. *BMC Microbiology*. **20**, 120 (2020).

687 30. N. N. Al-hebshi, A. T. Nasher, M. Y. Maryoud, H. E. Homeida, T. Chen, A. M. Idris, N. W. Johnson,
688 Inflammatory bacteriome featuring Fusobacterium nucleatum and Pseudomonas aeruginosa
689 identified in association with oral squamous cell carcinoma. *Sci Rep*. **7**, 1834 (2017).

690 31. A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes, D. Falush, J.
691 A. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. **31**,
692 3691–3693 (2015).

693 32. J. Mitchell, Streptococcus mitis: walking the line between commensalism and pathogenesis. *Mol*
694 *Oral Microbiol*. **26**, 89–98 (2011).

695 33. C. Chaguza, J. E. Cornick, D. B. Everett, Mechanisms and impact of genetic recombination in the
696 evolution of Streptococcus pneumoniae. *Computational and Structural Biotechnology Journal*. **13**,
697 241–247 (2015).

698 34. C. P. Andam, W. P. Hanage, Mechanisms of genome evolution of Streptococcus. *Infection, Genetics*
699 *and Evolution*. **33**, 334–342 (2015).

700 35. N. J. Croucher, A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, S. R.
701 Harris, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome
702 sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).

703 36. X. Didelot, D. J. Wilson, ClonalFrameML: efficient inference of recombination in whole bacterial
704 genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).

705 37. R. G. Everitt, X. Didelot, E. M. Batty, R. R. Miller, K. Knox, B. C. Young, R. Bowden, A. Auton, A.
706 Votintseva, H. Larner-Svensson, J. Charlesworth, T. Golubchik, C. L. C. Ip, H. Godwin, R. Fung, T. E.
707 A. Peto, A. S. Walker, D. W. Crook, D. J. Wilson, Mobile elements drive recombination hotspots in
708 the core genome of Staphylococcus aureus. *Nat Commun*. **5**, 3956 (2014).

709 38. O. Brynildsrud, J. Bohlin, L. Scheffer, V. Eldholm, Rapid scoring of genes in microbial pan-genome-
710 wide association studies with Scoary. *Genome Biology*. **17**, 238 (2016).

711 39. D. V. Rhodes, K. E. Crump, O. Makhlynets, M. Snyder, X. Ge, P. Xu, J. Stubbe, T. Kitten, Genetic
712 Characterization and Role in Virulence of the Ribonucleotide Reductases of Streptococcus
713 sanguinis. *J. Biol. Chem.* **289**, 6273–6287 (2014).

714 40. C. Collins, X. Didelot, A phylogenetic method to perform genome-wide association studies in
715 microbes that accounts for population structure and recombination. *PLoS Comput. Biol.* **14**,
716 e1005958 (2018).

717 41. S. G. Earle, C.-H. Wu, J. Charlesworth, N. Stoesser, N. C. Gordon, T. M. Walker, C. C. A. Spencer, Z.
718 Iqbal, D. A. Clifton, K. L. Hopkins, N. Woodford, E. G. Smith, N. Ismail, M. J. Llewelyn, T. E. Peto, D.
719 W. Crook, G. McVean, A. S. Walker, D. J. Wilson, Identifying lineage effects when controlling for
720 population structure improves power in bacterial association studies. *Nature Microbiology*. **1**,
721 16041 (2016).

722 42. M. Nordborg, B. Charlesworth, D. Charlesworth, The effect of recombination on background
723 selection. *Genet Res*. **67**, 159–174 (1996).

724 43. T. D. Mortimer, A. M. Weber, C. S. Pepperell, Signatures of Selection at Drug Resistance Loci in
725 Mycobacterium tuberculosis. *mSystems*. **3**, e00108-17 (2018).

726 44. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus*
727 *Evol*. **4** (2018), doi:10.1093/ve/vex042.

728 45. L. R. Joyce, Z. Guan, K. L. Palmer, Phosphatidylcholine Biosynthesis in Mitis Group Streptococci via
729 Host Metabolite Scavenging. *J Bacteriol*. **201**, e00495-19 (2019).

730 46. L. R. Joyce, Z. Guan, K. L. Palmer, Streptococcus pneumoniae, S. pyogenes and S. agalactiae
731 membrane phospholipid remodelling in response to human serum. *Microbiology (Reading)*. **167**,
732 001048 (2021).

733 47. T. Azarian, P. P. Martinez, B. J. Arnold, X. Qiu, L. R. Grant, J. Corander, C. Fraser, N. J. Croucher, L. L.
734 Hammitt, R. Reid, M. Santosham, R. C. Weatherholtz, S. D. Bentley, K. L. O'Brien, M. Lipsitch, W. P.
735 Hanage, Frequency-dependent selection can forecast evolution in Streptococcus pneumoniae.
736 *PLOS Biology*. **18**, e3000878 (2020).

737 48. T. Golubchik, E. M. Batty, R. R. Miller, H. Farr, B. C. Young, H. Larner-Svensson, R. Fung, H. Godwin,
738 K. Knox, A. Votintseva, R. G. Everitt, T. Street, M. Cule, C. L. C. Ip, X. Didelot, T. E. A. Peto, R. M.
739 Harding, D. J. Wilson, D. W. Crook, R. Bowden, Within-Host Evolution of Staphylococcus aureus
740 during Asymptomatic Carriage. *PLOS ONE*. **8**, e61319 (2013).

741 49. U. Muthukumarasamy, M. Preusse, A. Kordes, M. Koska, M. Schniederjans, A. Khaledi, S. Häussler,
742 Single-Nucleotide Polymorphism-Based Genetic Diversity Analysis of Clinical Pseudomonas
743 aeruginosa Isolates. *Genome Biology and Evolution*. **12**, 396–406 (2020).

744 50. L. L. Bohr, M. A. Youngblom, V. Eldholm, C. S. Pepperell, Genome reorganization during emergence
745 of host-associated Mycobacterium abscessus. *Microbial Genomics*. **7** (2021),
746 doi:10.1099/mgen.0.000706.

747 51. S. Paik, L. Senty, S. Das, J. C. Noe, C. L. Munro, T. Kitten, Identification of Virulence Determinants for
748 Endocarditis in Streptococcus sanguinis by Signature-Tagged Mutagenesis. *Infection and*
749 *Immunity*. **73**, 6064–6074 (2005).

750 52. S. F. Bailey, A. Hinz, R. Kassen, Adaptive synonymous mutations in an experimentally evolved
751 Pseudomonas fluorescens population. *Nat Commun*. **5**, 4076 (2014).

752 53. S. F. Bailey, L. A. Alonso Morales, R. Kassen, Effects of Synonymous Mutations beyond Codon Bias:
753 The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments.
754 *Genome Biology and Evolution*. **13**, evab141 (2021).

755 54. A. Ballard, S. Bieniek, D. B. Carlini, The fitness consequences of synonymous mutations in
756 Escherichia coli: Experimental evidence for a pleiotropic effect of translational selection. *Gene*.
757 **694**, 111–120 (2019).

758 55. J. Kristofich, A. B. Morgenthaler, W. R. Kinney, C. C. Ebmeier, D. J. Snyder, W. M. Old, V. S. Cooper,
759 S. D. Copley, Synonymous mutations make dramatic contributions to fitness when growth is
760 limited by a weak-link enzyme. *PLOS Genetics*. **14**, e1007615 (2018).

761 56. E. Lundin, P.-C. Tang, L. Guy, J. Näsvall, D. I. Andersson, Experimental Determination and Prediction
762 of the Fitness Effects of Random Point Mutations in the Biosynthetic Enzyme HisA. *Molecular*
763 *Biology and Evolution*. **35**, 704–718 (2018).

764 57. L. H. Rasmussen, K. Højholt, R. Dargis, J. J. Christensen, O. Skovgaard, U. S. Justesen, F. S.
765 Rosenvinge, C. Moser, O. Lukjancenko, S. Rasmussen, X. C. Y. 2017 Nielsen, In silico assessment of
766 virulence factors in strains of Streptococcus oralis and Streptococcus mitis isolated from patients
767 with Infective Endocarditis. *Journal of Medical Microbiology*. **66**, 1316–1323.

768 58. S. P. Baker, T. J. Nulton, T. Kitten, Genomic, Phenotypic, and Virulence Analysis of Streptococcus
769 sanguinis Oral and Infective-Endocarditis Isolates. *Infection and Immunity*. **87**, e00703-18 (2019).

770 59. K. H. Iversen, L. H. Rasmussen, K. Al-Nakeeb, J. J. A. Armenteros, C. S. Jensen, R. Dargis, O.
771 Lukjancenko, U. S. Justesen, C. Moser, F. S. Rosenvinge, X. C. Nielsen, J. J. Christensen, S.
772 Rasmussen, Similar genomic patterns of clinical infective endocarditis and oral isolates of
773 Streptococcus sanguinis and Streptococcus gordonii. *Sci Rep*. **10**, 2728 (2020).

774 60. J. P. Galkiewicz, C. A. Kellogg, Cross-kingdom amplification using bacteria-specific primers:
775 complications for studies of coral microbial ecology. *Appl Environ Microbiol*. **74**, 7828–7831
776 (2008).

777 61. S. Andrews, FastQC: A Quality Control tool for High Throughput Sequence Data (2010).

778 62. A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I.
779 Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A.
780 Alekseyev, P. A. Pevzner, SPAdes: A New Genome Assembly Algorithm and Its Applications to
781 Single-Cell Sequencing. *J Comput Biol*. **19**, 455–477 (2012).

782 63. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome
783 assemblies. *Bioinformatics*. **29**, 1072–1075 (2013).

784 64. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, btu153 (2014).

785 65. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2 – Approximately Maximum-Likelihood Trees for
786 Large Alignments. *PLOS ONE*. **5**, e9490 (2010).

787 66. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: an r package for visualization and
788 annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol*
789 *Evol*. **8**, 28–36 (2017).

790 67. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
791 phylogenies. *Bioinformatics*. **30**, 1312–1313 (2014).

792 68. J. Hadfield, N. J. Croucher, R. J. Goater, K. Abudahab, D. M. Aanensen, S. R. Harris, Phandango: an
793 interactive viewer for bacterial population genomics. *Bioinformatics*. **34**, 292–293 (2018).

794 69. I. Lee, Y. Ouk Kim, S.-C. Park, J. Chun, OrthoANI: An improved algorithm and software for
795 calculating average nucleotide identity. *Int J Syst Evol Microbiol*. **66**, 1100–1103 (2016).

796 70. S. De Mita, M. Siol, EggLib: processing, analysis and simulation tools for population genetics and
797 genomics. *BMC Genetics*. **13**, 27 (2012).

798 71. A. J. Page, B. Taylor, A. J. Delaney, J. Soares, T. Seemann, J. A. Keane, S. R. Harris, SNP-sites: rapid
799 efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*. **2** (2016),
800 doi:10.1099/mgen.0.000056.

801 72. I. van de Rijn, R. E. Kessler, Growth characteristics of group A streptococci in a new chemically
802 defined medium. *Infect Immun*. **27**, 444–448 (1980).

803 73. R. R. Wick, L. M. Judd, C. L. Gorrie, K. E. Holt, Unicycler: Resolving bacterial genome assemblies
804 from short and long sequencing reads. *PLoS Comput Biol*. **13**, e1005595 (2017).

805 74. H. M. Adams, L. R. Joyce, Z. Guan, R. L. Akins, K. L. Palmer, Streptococcus mitis and S. oralis Lack a
806 Requirement for CdsA, the Enzyme Required for Synthesis of Major Membrane Phospholipids in
807 Bacteria. *Antimicrob Agents Chemother*. **61**, e02552-16 (2017).

808

809

Table and Figures

Table 1: Strains used in this study.

Organism	Strain	Description	Ref
Mitis group streptococci	1643 (SM43)	Wild-type infective endocarditis isolate	(19)
	SM43ΔNrdM	SM43 with clean deletion of *nrdM*, FD735_06230	This work
	SM43:NrdM-I	SM43 with SM43 *nrdM* allelic replacement into SM43Δ*nrdM*	This work
	SM43:NrdM-C	SM43 with SO48 *nrdM* allelic replacement into SM43Δ*nrdM*	This work
S. oralis	1648 (SO48)	Wild-type infective endocarditis isolate	(19)
	SO48ΔNrdM	SO48 with clean deletion of *nrdM*, MP387_03665	This work
	SO48:NrdM-I	SM43 *nrdM* allelic replacement into SO48Δ*nrdM*	This work
	SO48:NrdM-C	SO48 *nrdM* allelic replacement into SO48Δ*nrdM*	This work

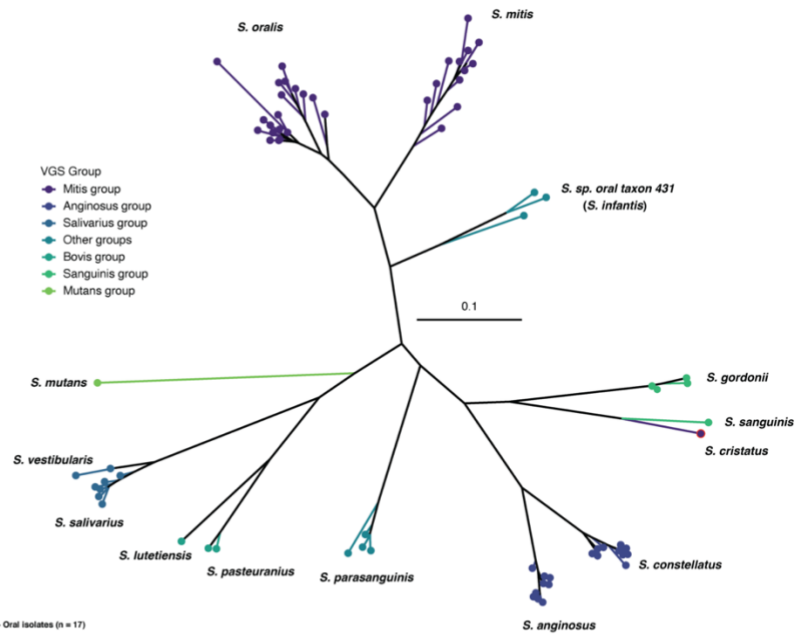
A

Group	Species	Kraken	gyrB	16S
Anginosus group	no species identified			7
	S. anginosus	19	12	6
	S. constellatus	0	8	0
	S. intermedius	0	na*	3
Bovis group	*S. gallolyticus*	1	na	0
	S. lutetiensis	2	na	2
	S. pasteuranius	3	na	3
Mitis group	no species identified			26
	S. mitis	13	13	1
	S. oralis	20	21	3
	S. cristatus	1	na	0
	S. infantis	0	2	1
Mutans group	*S. mutans*	1	na	1
Salivarius group	no species identified			8
	S. salivarius	6	0	0
	S. vestibularis	2	0	0
	S. sanguinis	1	0	1
	S. gordonii	5	na	5
Other	*S. parasanguinis*	4	0	2
	S. sp. oral taxon 431†	3	na	0
	undetermined	0	25	12

* na = gyrB typing scheme does not include this species

† likely *S. infantis*

B



C

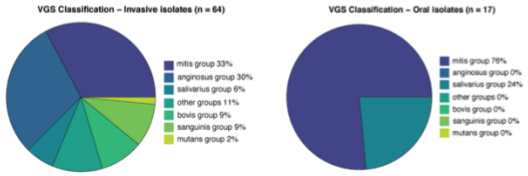


Figure 1: Species identification of blood and oral commensal viridans group streptococci (VGS). A) Three different methods for species identification were used (Kraken, GyrB typing & 16S rRNA sequencing) providing differing results. “na” values in the gyrB column indicate species not included in this typing scheme. “No species identified” indicates that the isolate could only be identified to the group level. **B)** Phylogenetic tree of all 81 isolates inferred from an alignment of GyrB sequences and identified by VGS group and species as assigned by Kraken. The isolate which appears to have been mis-identified by Kraken is outlined in red. **C)** Invasive (left) and oral (right) isolates by VGS group according to species identification performed by Kraken.

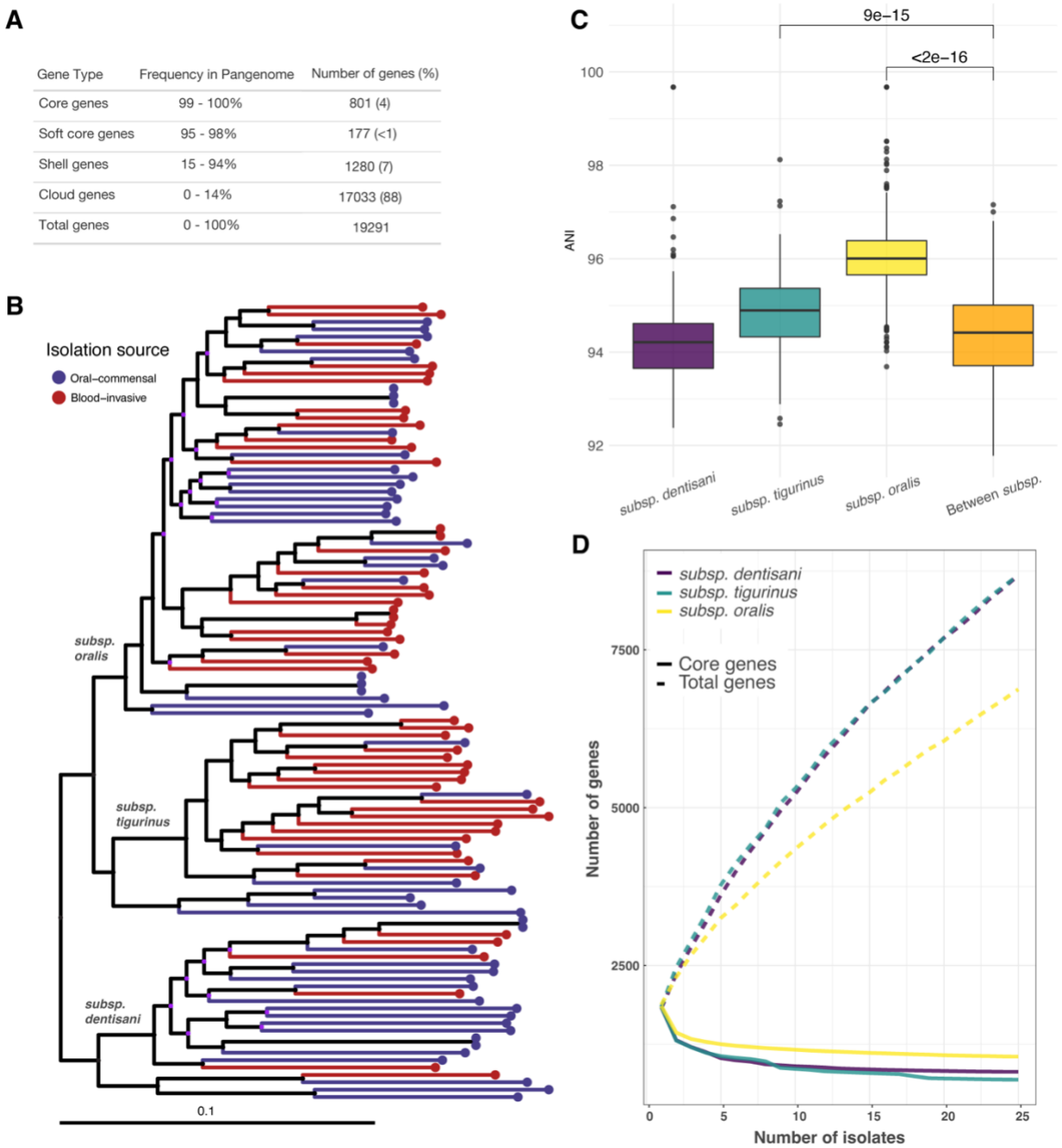


Figure 2: Pangenome analysis of *S. oralis* strains. A) Summary statistics from pangenome analysis of all *S. oralis* GWAS isolates (n=108). **B)** Core genome phylogeny of *S. oralis* isolates inferred using RAxML and midpoint rooted. Tips colored by isolation source (oral-commensal in blue, blood-invasive in red). 108 isolates total, 57 isolated from the oral cavity and 51 isolated from invasive infections. Three subspecies are labeled and scale bar given in SNPs per site. Nodes colored in purple represent bootstrap values < 50. **C)** Average nucleotide identity (ANI) calculated within and between the three subspecies of *S. oralis* phylogeny show relatively low sequence conservation (~94-96% ANI) between even strains of the same subspecies. ANI values within *subsp. oralis* and *subsp. tigurinus* show significantly higher sequence conservation within compared to between subspecies (Mann-Whitney U test with Benjamini-Hochberg correction). **D)** Accumulation and rarefaction curves for three subspecies. All samples were

repeatedly sub-sampled to the size of the smallest sample (*subsp. dentisani,* n=25) and the median core and total gene values plotted for 100 iterations.

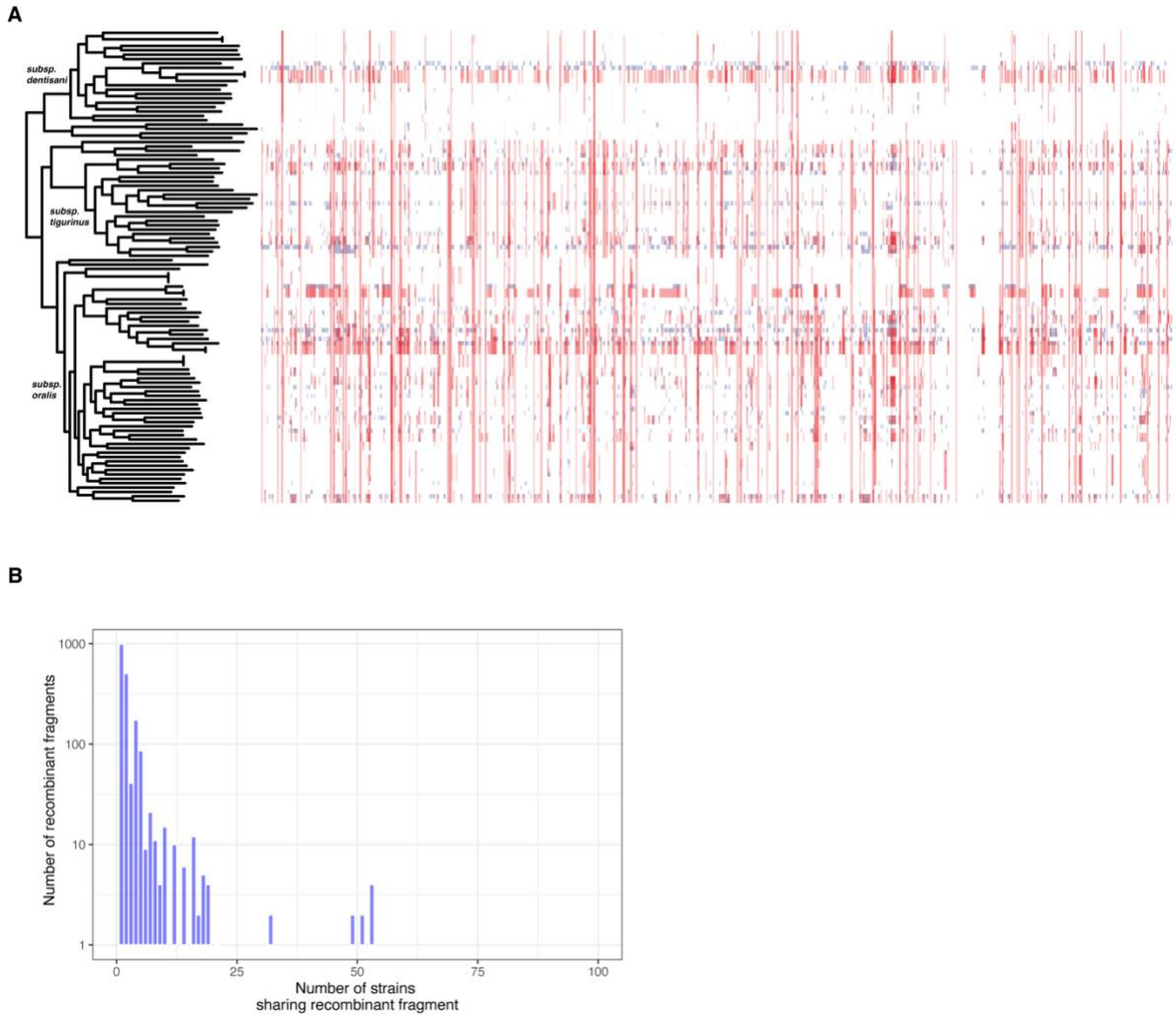


Figure 3: *S. oralis* strains are exceptionally recombinogenic. A) Recombination tracts predicted by Gubbins in the core genome of *S. oralis* isolates plotted alongside the core genome phylogeny. Red tracts represent recombination within the sample and blue tracts represent recombination with isolates outside this sample. 99.9% of the core genome has been affected by recombination. Visualization performed with Phandango. **B)** Histogram showing the number of strains in our sample that share a given recombinant fragment. The x-axis is the number of strains that share a recombinant fragment (with a maximum of 108 isolates, the size of our sample) and the y-axis is the number of fragments identified by Gubbins that are shared by that number of isolates.

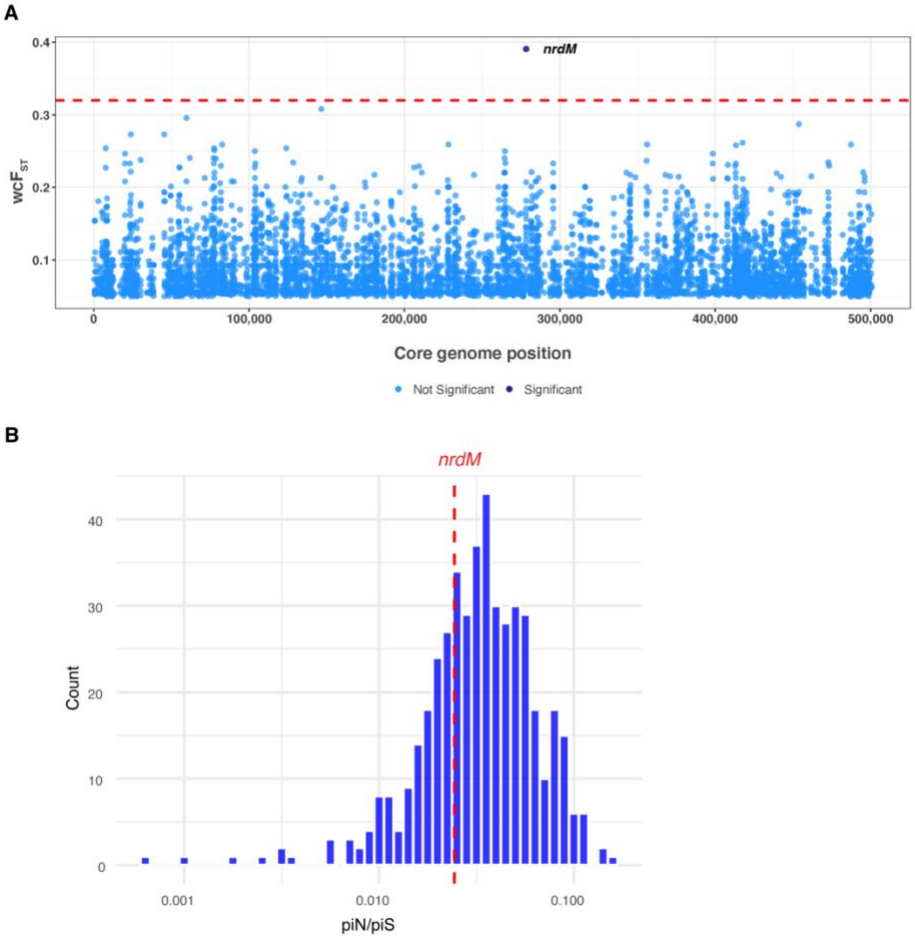


Figure 4: A Single core genome variant is associated with invasiveness in *S. oralis* isolates. A) Weir and Cockerham's F_{ST} (wcF_{ST}) values calculated for each core genome variant (n=67,026) and plotted against the core genome position. Significance threshold (red dotted line) estimated by taking the maximum wcFst value from 100 random permutations of phenotypes. Variants with non-significant F_{ST} values shown in light blue, a single variant with a significant F_{ST} value shown in dark blue. **B)** Pairwise piN/piS values calculated and averaged for each core gene and plotted as a histogram. Average piN/piS value across all genes is 0.039, red dotted line represents the average piN/piS value for *nrdM* (0.024).

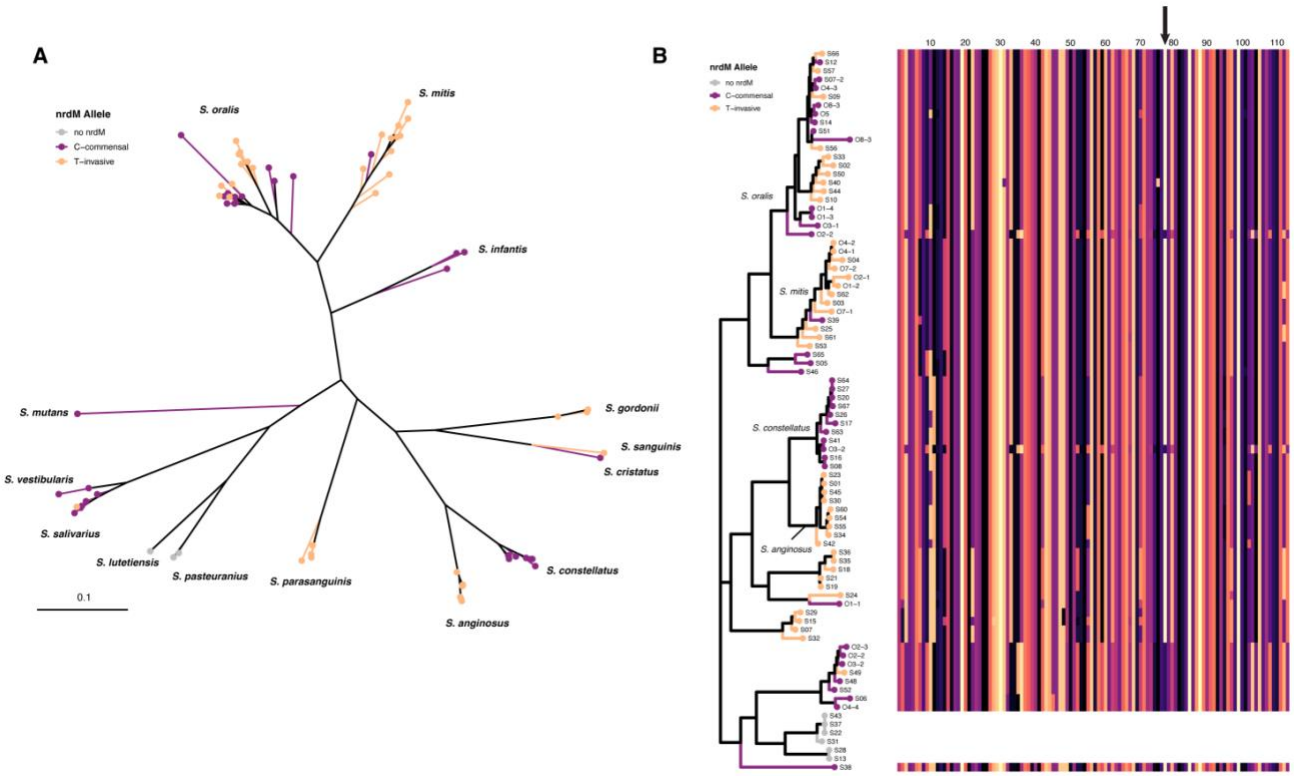


Figure 5: A) Homologous *nrdM* genes identified in all newly sequenced VGS isolates with alleles from our variant of interest plotted on a phylogeny made from GyrB sequences. Only 2 species in our sample lacked *nrdM* (*S. lutetiensis* and *S. pasteuranius*) both of which are in the bovis group and are shown in grey. All other species contained *nrdM* and were conserved at the position of interest. **B)** Multiple sequence alignment (MSA) of NrdM plotted next to the *gyrB* phylogeny (same tree as in A) where tips have been colored by *nrdM* allele. Scale on top of the MSA represents length in amino acids, the position of the synonymous mutation of interest (I78I) indicated with an arrow. Clade labels shown for the most common species: *S. oralis, S. mitis, S. anginosus* and *S. constellatus*.

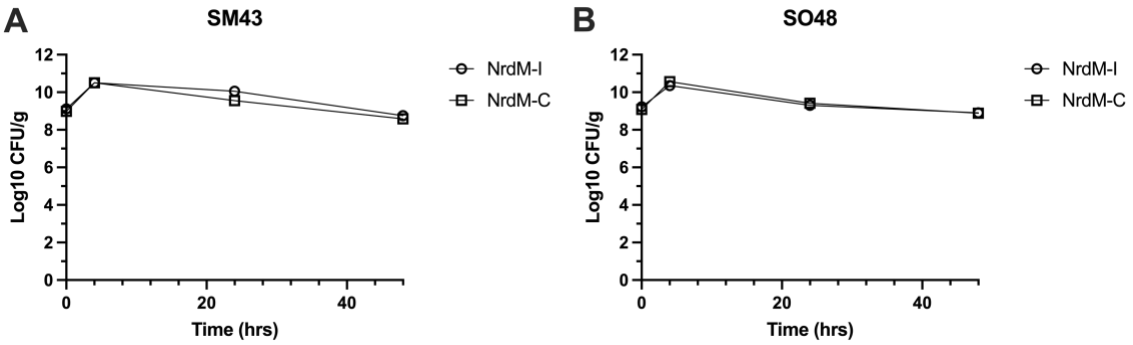


Figure 6: Simulated infective endocarditis vegetation model of *nrdM* allele swapped strains. Survival of either **A)** SM43 or **B)** SO48 harboring either the infectious *nrdM*-I allele or the commensal *nrdM*-C allele during growth in vegetations. Biological triplicates were performed for each strain with four technical replicates per time point. Mean and SEM indicated. No significant difference was observed under these conditions. Two-way ANOVA with Sidak's multiple comparisons test.