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.

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 VGS species, including *Streptococcus mitis* and *Streptococcus oralis* (7). 16S rRNA gene sequencing, multilocus sequence analysis and other genotyping schemes (5-10), and GyrB typing (11) are commonly used methods for molecular VGS identification but are not generally employed in clinical laboratories.

Retrospective studies using molecular approaches have determined that in addition to being present as oral commensals in healthy individuals, *S. mitis* and *S. oralis* stand out among VGS as major causative agents of VGS bacteremia and IE (8, 12-15)**.** *S. mitis* and *S. oralis* are members of the mitis group streptococci, a subgroup within the VGS that is closely related to the major human pathogen *Streptococcus pneumoniae* (16, 17). A recent study in oncology

patients demonstrated that *S. mitis* (58%) and *S. oralis* (19%) were the most frequently identified species in VGS infections over ~1.5 years (8). While *S. mitis* and *S. oralis* are a significant burden on immunocompromised patients, the mechanisms of virulence within these species have not been fully elucidated. More specifically, it is not known whether all members of these species have equal pathogenic potential, or whether some strains have a higher propensity for causing invasive disease than others.

In this study, we aimed to investigate the mechanisms of invasiveness among VGS by characterizing species diversity of presumptive VGS obtained clinically from bacteremia and endocarditis patients. Clinical isolate genomes obtained in this study were supplemented by existing genome sequences with curated metadata in public databases and genome sequences of oral isolates collected from healthy volunteers. Our results support metagenomic sequence binning as a high-resolution tool for differentiating species within the VGS grouping. After a preliminary analysis in which we determined species designations for clinical isolates that were diagnosed as VGS, we focused our study on *S. oralis* as a prominent cause of invasive infection. Within a large sample of *S. oralis* isolates we found high levels of diversity and strikingly high recombination rates. We used multiple genome-wide association study (GWAS) methods to test the hypothesis that specific genetic variations are associated with invasive infection (when compared to commensal isolates) among *S. oralis* isolates. We discovered a SNP in a previously uncharacterized gene that is significantly enriched in invasive isolates compared to non-invasive isolates. The contribution of this novel locus to growth with human serum and in a simulated infective endocarditis vegetation model (18) was assessed, although we were unable to identify a phenotype for a gene knockout or for either

allele of the significant variant under the conditions tested. This work has 1) increased the genomic information available for VGS strains, 2) described population structure and large-scale homologous recombination within the *S. oralis* species, and 3) provided evidence that the propensity for virulence in *S. oralis* is at least in part genetically determined.

Results

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

VGS isolates from Dallas, Texas and Jackson, Mississippi areas were collected from clinically confirmed VGS bacteremia and infective endocarditis patients. Isolates were initially characterized using either VITEK®2 or MicroScan platforms. A total of 66 clinical isolates were successfully sub-cultured in the laboratory, however two isolates were incorrectly identified as *Streptococcus* spp.: one *Enterococcus faecalis*, and one *Aerococcus urinae* strain, resulting in 64 presumptive VGS strains collected (Table S1). To compare isolates causing invasive disease against commensal isolates, healthy volunteers were recruited for oral swab collection on the University of Texas at Dallas campus, for a total of 81 VGS isolates (Table S1).

As has previously been reported, clinical methods of VGS species identification (VITEK®2 and MicroScan) were not effective for delineating closely related species, particularly within the mitis group; however, they were almost always correct in identifying an isolate as a part of the VGS (Table S1). The taxonomy of the clinical and oral isolates was analyzed by 16S rRNA gene sequencing, GyrB typing (19), and analysis of Illumina whole genome sequencing (WGS) data using Kraken (20). Using these three different methods of species identification, we classified isolates into VGS groups using the taxonomy described by Facklam (21). All six of

the major VGS groups (1) were represented in our sample of 63 isolates, with two isolates not fitting into any of these six groups (Figure 1A). 16S rRNA sequencing was frequently unable to provide resolution to the species level but was often able to identify which group the isolate belonged to. GyrB typing (19) and Kraken (20) were generally in agreement, particularly when it comes to distinguishing *S. oralis* and *S. mitis* (Figure 1A).

One downside of species identification with Kraken is it seems unable to distinguish between anginosus group isolates: *S. anginosus* and *S. constellatus* are species within the anginosus group (also known as the milleri group) of the VGS and are recognized as abscess-causing bacteria, with more recent data suggesting their emergence as uropathogens (22-24). Kraken identified 19 *S. anginosus* isolates, and via GyrB typing a total of 20 anginosus group isolates were identified (12 *S. anginosus* and 8 *S. constellatus*). These data suggest that *S. mitis* and *S. oralis* are accurately speciated with either Kraken or GyrB typing, yet anginosus group species may be better identified by GyrB typing.

One drawback to species identification with GyrB typing is the current scheme is limited in the species it can distinguish (Figure 1A). To further assess the relative functionality of these two methods we inferred a phylogenetic tree from GyrB sequences and mapped species as defined by Kraken onto the tree (Figure 1B). The added benefit of a phylogeny of GyrB sequences is the visualization of additional species not included in the typing scheme (e.g. species in the bovis group; Figure 1B) which when combined with Kraken output allowed us to confidently assign species to most isolates. There was only a single instance where it seems that Kraken may have mis-identified the species: an isolate identified by Kraken as *S. cristatus*

(a member of the mitis group) appears to cluster with *S. sanguinis* isolates (of the sanguinis group) on the GyrB phylogeny (Figure 1B). Additionally, combining data from Kraken and GyrB typing showed that 2 out of the 3isolates identified as *S. sp. oral taxon 431* by Kraken were identified as *S. infantis* (a member of the mitis group) by GyrB typing, and all 3 cluster together on the GyrB phylogeny, indicating that *S. infantis* is likely the correct designation (Figure 1B).

Overall, our data show that extremely closely related *S. mitis* and *S. oralis* can be speciated accurately via either Kraken or GyrB typing; however, anginosus group isolates are better distinguished by GyrB typing. For datasets suspected to contain a diverse sample of different VGS species, species identification using Kraken with whole genome sequence data appears the most robust. GyrB typing is an accurate method for making distinctions within species subsets, and a phylogeny of GyrB sequences may also provide additional information. Our results support previous assertions that 16S rRNA sequencing is not an effective method for distinguishing the closely related species of the VGS (1).

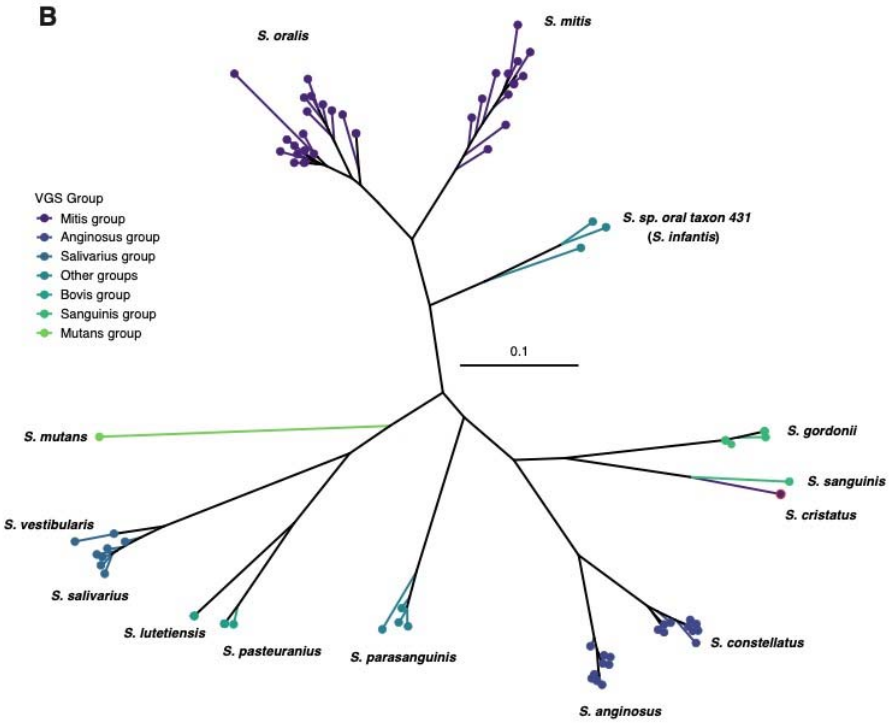
A

Species identified by different methods

Species	Kraken	gyrB	16S
Anginosus group			7
S. anginosus	19	12	6
S. constellatus	0	8	0
S. intermedius	0	*na*[*]	3
Bovis group			0
S. gallolyticus	1	*na*	0
S. lutetiensis	2	*na*	2
S. pasteuranius	3	*na*	3
Mitis group			26
S. mitis	13	13	1
S. oralis	20	21	3
S. cristatus	1	*na*	0
S. infantis	0	2	1
Mutans group			0
S. mutans	1	*na*	1
Salivarius group			8
S. salivarius	6	0	0
S. vestibularis	2	0	0
Sanguinis group			0
S. sanguinis	1	0	1
S. gordonii	5	*na*	5
Other groups			0
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

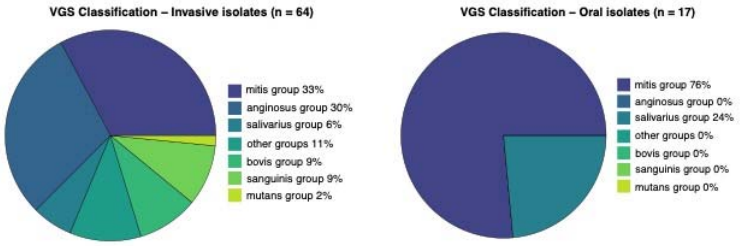


Figure 1: Species identification of blood and oral commensal viridans group streptococci (VGS). A) Three different methods for species identification were used (Kraken, GyrB typing & 16S rRNA sequencing) providing differing results. **B)** Phylogenetic tree of all 81 isolates inferred from an alignment 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.

When we separated our sample into isolates from blood cultures and those from the mouths of healthy persons, we saw that invasive isolates spanned all 6 groups while isolates from oral

isolates were less diverse and only originated from the mitis and salivarius groups (Figure 1C). The lack of diversity in the oral isolates is due to the use of Mitis-Salivarius Agar which is selective for the *S. mitis* and *S. salivarius* species. These species produce different colony morphologies allowing for preferential selection of these targeted species. It has been reported that *S. mitis* is the predominant species found in healthy oral microbiomes (25-30) with other commonly found VGS species including *S. oralis* (27, 28), *S. sanguinis* (28, 29), and *S. salivarius* (29).

High diversity among commensal and invasive *S. oralis* strains

S. oralis was the predominant single species (20/81 isolates) in our sample, and so to identify possible genetic variants associated with invasive infection, we gathered a larger sample of *S. oralis*. We obtained all isolates labeled as *S. oralis* and *S. mitis* from NCBI (see Methods for inclusion criteria) and used Kraken, a core genome phylogeny, and accessory gene content to confirm *S. oralis* isolates and identify mislabeled isolates. We found that a core genome alignment produced by Roary (31) effectively delineates *S. oralis* from *S. mitis*, as do patterns of accessory content (Figure S1). We were able to identify 11 isolates mis-labeled as *S. mitis* within NCBI databases, which we identified as *S. oralis* using these methods (Table S2). We ended up with a total of 108 *S. oralis* isolates: 57 oral commensals and 51 from invasive infections (including bacteremia and IE) that we refer to as “oral” and “blood” isolates, respectively (Table S2). After assembly and annotation of all genomes, we performed a pangenome analysis with Roary and found that our sample has a core genome of only 801 genes – with an average genome size of 1898 genes. This means that a significant proportion of genes (57%) encoded by an individual isolate are variable accessory genes (Figure 2A).

The accessory gene content in our sample of *S. oralis* isolates is highly diverse, with 88% of all genes found in our sample at a frequency of 14% or less, and > 19,000 genes identified in the pan genome (Figure 2A). This high level of gene content diversity is mirrored in the core genome phylogeny, which has long terminal branch lengths indicating high variability between core genome sequences (Figure 2B). Isolates from commensal and invasive sources are interspersed on the phylogeny and do not generally form monophyletic clades, nor are the commensal or invasive phenotypes associated any sub-clades within the phylogeny (Figure 2B).

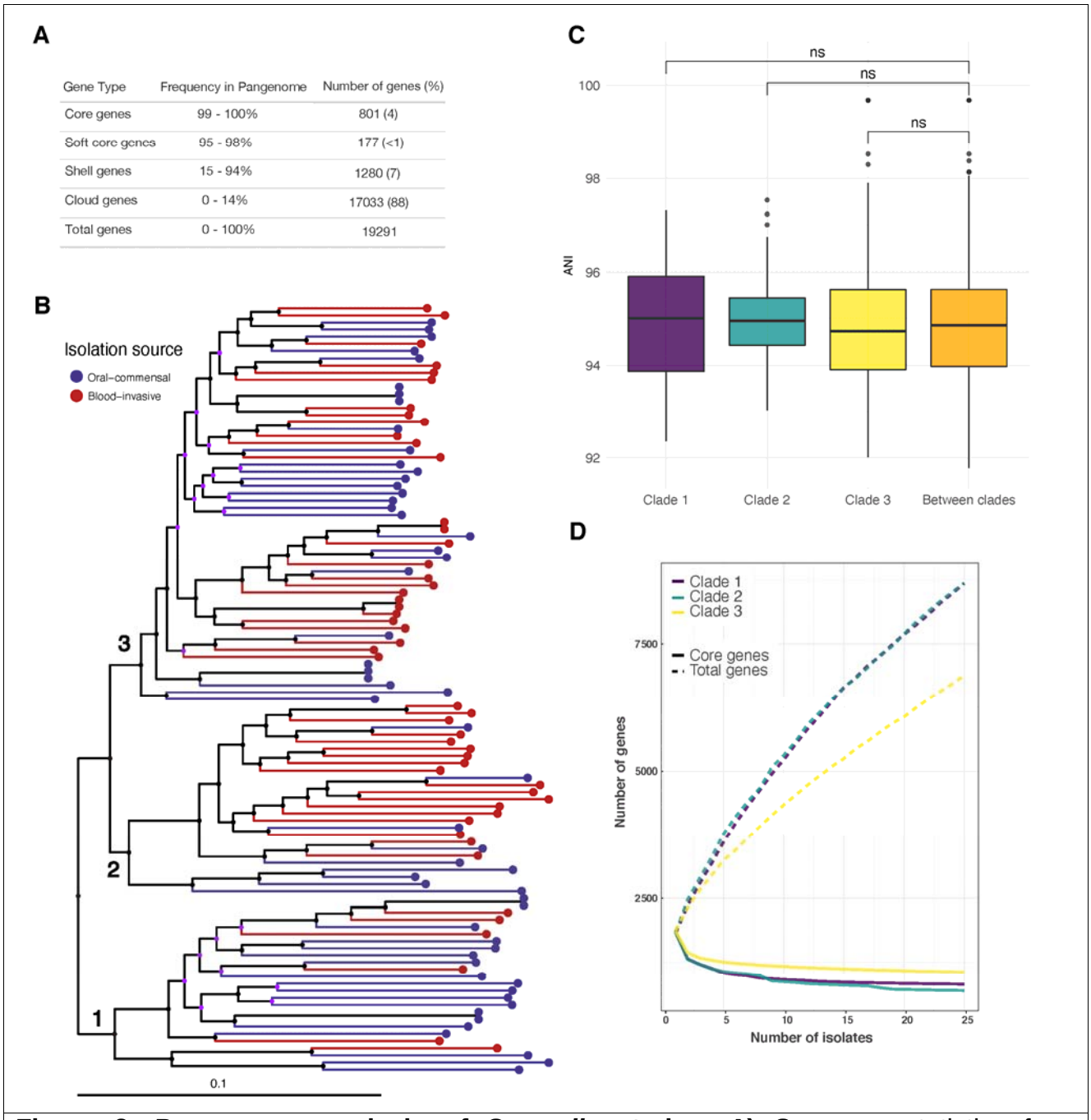


Figure 2: Pangenome analysis of *S. oralis* strains. A) Summary statistics from pangenome analysis of all *S. oralis* GWAS isolates (n=108). **B)** Core genome phylogeny of *S. oralis* isolates inferred using RAxML and midpoint rooted. Tips colored by isolation source (oral-commensal in blue, blood-invasive in red). 108 isolates total, 57 isolated from the oral cavity and 51 isolated from invasive infections. Three major clades are labeled by number 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 three major clades of *S. oralis* phylogeny show low sequence conservation (~95% ANI) between closely related strains. ANI values compared within/between clades do not indicate higher sequence conservation within clades (Mann-Whitney U test with Benjamini-Hochberg

correction, ns; not significant with *P*-values > 0.05). **D)** Accumulation and rarefaction curves for three clades. All samples were repeatedly sub-sampled to the size of the smallest sample (Clade 1, n=25) and the median core/total gene values plotted for 100 iterations.

We identified three sub-clades within our core genome phylogeny (Figure 2B) that we initially hypothesized to correspond to the three previously described sub-species of *S. oralis* (i.e. *subsp. oralis, subsp. tigurinus* and *subsp. dentisani*). However, a comparison between clade designation and sub-species as annotated on NCBI identified no correlation between the two classifications (Table S2). Average nucleotide identity (ANI) values for core genome sequences within each clade are not significantly higher than values between clades (Figure 2C). This indicates that - at least at the nucleotide level - strains within a clade are no more alike than strains of different clades. We did note differences in pangenome structure between clades, in particular that clade 3 seems to have a more closed pangenome (larger core genome, smaller accessory genome) when compared to clades 1 and 2 (Figure 2D). Given that we identified no barriers to genetic exchange between clades (Figure 2B) this suggests that clades 1 and 2 may have access to more diverse partners for horizontal gene transfer (HGT).

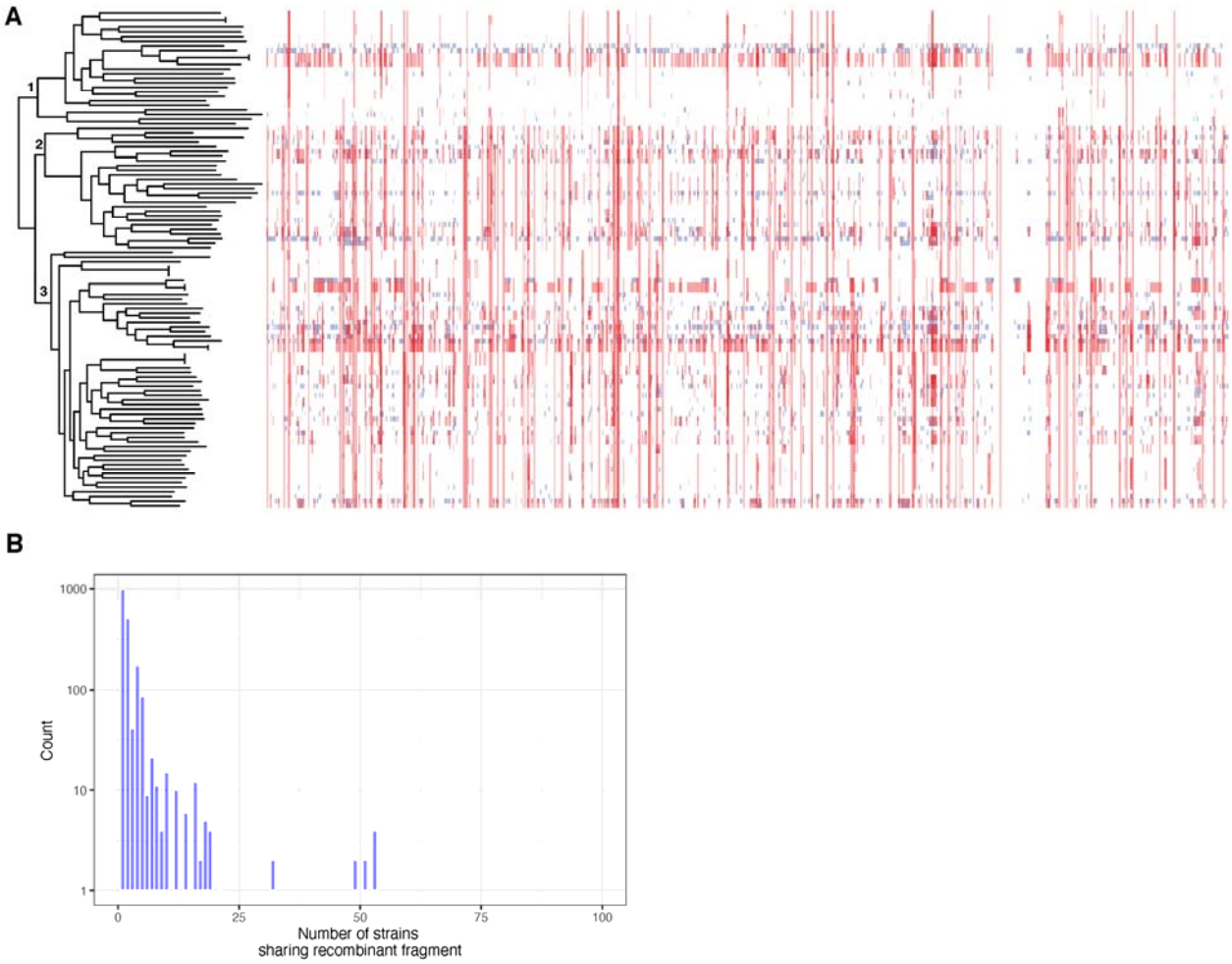


Figure 3: *S. oralis* strains are exceptionally recombinogenic. A) Recombination tracts predicted by Gubbins in the core genome of *S. oralis* isolates plotted alongside the core genome phylogeny. Red tracts represent recombination within the sample and blue tracts represent recombination with isolates outside this sample. 99.9% of the core genome has been affected by recombination. Visualization performed with Phandango. Numbers indicate *S. oralis* clades. **B)** Histogram showing the number of strains in our sample that share a given recombinant fragment.

High levels of recombination among *S. oralis* isolates

Viridans group streptococci (VGS) are known for being naturally competent (32) and for evolving rapidly via widespread homologous recombination (33, 34). We characterized the signatures of recombination in the core genome of *S. oralis* using Gubbins (35) and identified

extreme amounts of recombination: 99.9% of the core genome is within a predicted recombinant fragment in at least one isolate (Figure 3A). Additionally, we noted that recombinant fragments are not shared across multiple isolates but are usually present in only a handful of isolates (Figure 3B). Using ClonalFrameML (36) we calculated the *r/m* value - the ratio of SNPs imported via recombination (r) to those introduced randomly (m) – and found that with an *r/m* of 5.77, genetic diversity in our sample is ~6x more likely to be introduced via recombination. This is slightly less than the notoriously recombinogenic *S. pneumoniae* [*r/m* = ~7;(33)] but much higher than other IE causing bacteria such as *S. aureus* [*r/m* = <1;(37)].

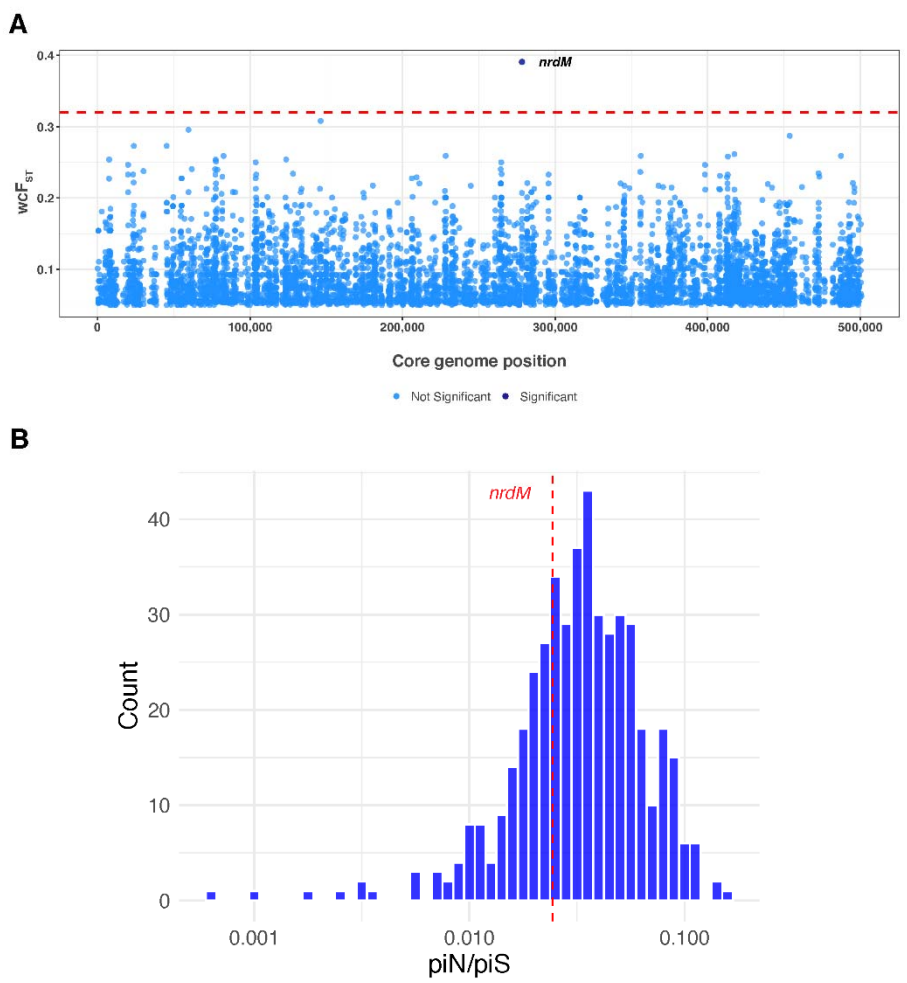


Figure 4: A Single core genome variant is associated with invasiveness in *S. oralis*

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).

Genome wide association study reveals variant associated with invasiveness

With a large sample balanced between our phenotypes of interest, we used multiple genome-wide association study (GWAS) methods to identify genetic variants associated with invasive disease, which included isolates from both IE and bacteremia patients. We started by looking for associations between accessory genes and invasiveness using Scoary (38), which yielded no significant results. This is perhaps not surprising given that accessory gene content is so diverse in the sample and individual accessory genes are thus unlikely to be shared by a large portion of isolates (Figure 1A).

To identify core genome variants associated with invasiveness, we started with a F_{ST} outlier analysis, which delineates allele frequency differences between sub-populations and identifies variants with extreme measures of differentiation. We defined sub-populations of our sample as being of "oral" or "blood" source and used vcflib (https://github.com/vcflib/vcflib) to calculate Weir and Cockerham's F_{ST} (wcF_{ST}, abbreviated to F_{ST}) for bi-allelic SNPs (n=67,026) in the core genome alignment (Figure 4A). We identified a single SNP with a significant F_{ST} value, indicating significant allele frequency differences between oral and blood sub-populations of *S. oralis* (Figure 4A). The SNP of interest lies within a gene originally annotated as a homologue of *nrdD*, an anaerobic ribonucleoside triphosphate reductase present in many streptococci,

involved in synthesis of deoxyribonucleotides under anaerobic conditions (39). However, further inspection of the sequence revealed that the canonical *nrdD* gene was annotated separately in our *S. oralis* isolates, and that the novel protein was shorter than *nrdD*. The novel protein does share some sequence features with *nrdD* including an ATP cone domain, and so we will refer to the novel locus as *nrdM*. Given the amount of recombination present in our sample (Figure 3A) we validated the results of our F_{ST} outlier analysis using GWAS methods specifically designed for use with microbial genomes. We used two programs: treeWAS (40), which corrects for the presence of recombination, and BugWAS (41), which identifies lineage effects and controls for population structure. Results from both of these tools replicated the results of our F_{ST} outlier analysis with both methods returning a single variant that was significantly associated with the phenotype: the same SNP in *nrdM* (Figure S2).

The SNP of interest in *nrdM* is a synonymous mutation (I78I), and mapping the nucleotide alleles of the mutation (C233T) onto the core genome phylogeny shows interspersal of alleles on the tree, with some structure in clades 1 and 2 (Figure S3). Most interesting is the striking association between isolates from invasive infections and the *nrdM* allele (Figure S3). A multiple sequence alignment of NrdM from our sample of *S. oralis* strains reveals high protein sequence conservation (Figure S4), and so to determine whether *nrdM* is more or less conserved than other core genes we calculated the average $\pi N/\pi S$ (piN/piS) values for each core gene (n=801). The distribution of gene-wise values of piN/piS indicate the majority of genes are evolving under relatively strong purifying selection, including *nrdM*, which had a piN/piS value slightly below the mean (Figure 4B). This is not surprising given that

recombination strengthens the efficiency of selection by enabling rapid removal of deleterious mutations (42).

Selection on *nrdM* variant

Given the strong association between the *nrdM* variant and invasiveness, we wondered if this SNP is under positive selection in our sample. One method for identifying positive selection is 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.

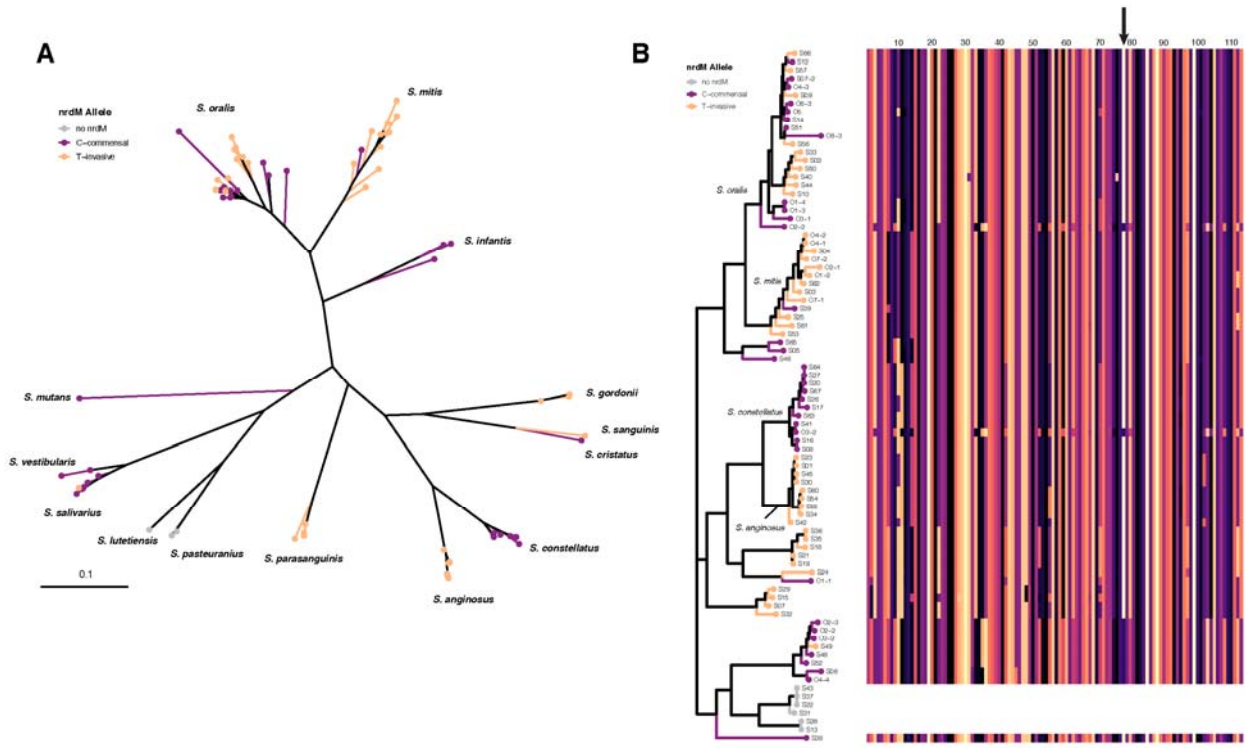


Figure 5: A) Homologous *nrdM* genes identified in all newly sequenced VGS isolates with alleles from our variant of interest plotted on a phylogeny made from GyrB sequences. Only 2 species in our sample lacked *nrdM* (*S. lutetiensis* and *S. pasteuranius*) both of which are in the bovis group and are 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*.

***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. Notably, *S. constellatus* isolates (n = 8) all have the *nrdM* variant

associated with commensalism, while all *S. anginosus* isolates (n = 11) have the variant associated with invasiveness (Figure 5A). 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).

Table 1: Strains used in this study.

Organism	Strain	Description	Ref
Mitis group streptococci	1643 (SM43)	Wild-type infective endocarditis isolate	(18)
	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	(18)
	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

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*.

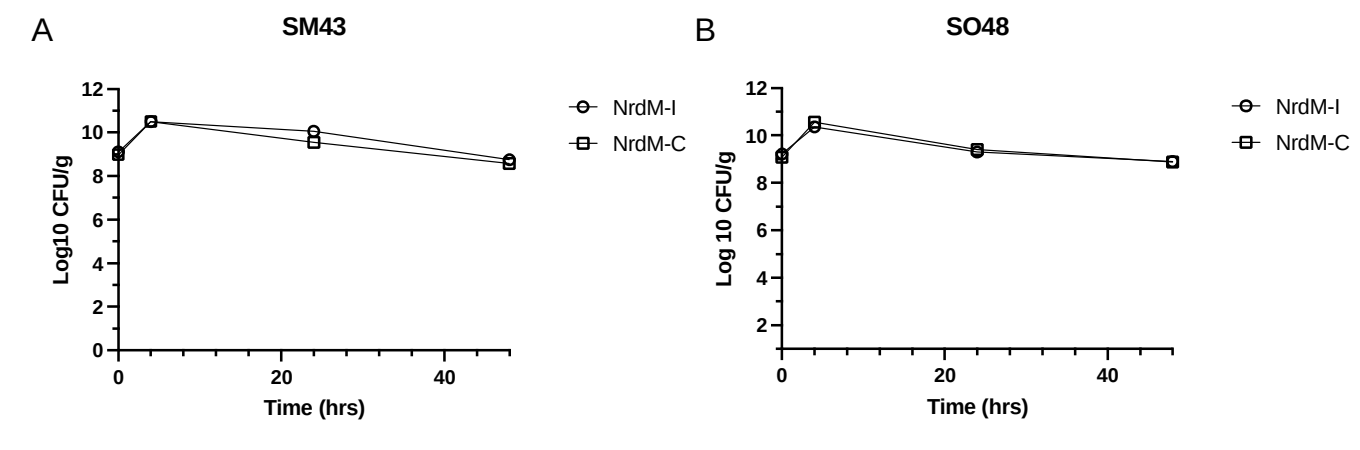


Figure 6: Simulated infective endocarditis vegetation model of *nrdM* allele swapped strains. Survival of either **A)** SM43 or **B)** SO48 harboring either the infectious *nrdM*-I allele or the commensal *nrdM*-C allele during growth in vegetations. Biological triplicates were performed for each strain 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.

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) (18). 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 our 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* do not correspond to sub-clades we identified in a phylogeny inferred from core genome sequences (Table S2). We did see evidence of a slightly different pangenome structure in one of the clades, but no evidence of genomic incompatibility between clades (Figure 2D). 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 multiple sub-clades and diverse isolates (Figure S3). This finding contributes to our increasing recognition that synonymous variation can have sizeable fitness effects on bacteria (51, 52). 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 (53). 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 (54). 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 (55, 56). 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, 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 (*43*).

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 (57). 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 Kraken (20, 58). Raw data was quality checked and trimmed using FastQC v0.11.8 (59) 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 (60). Assemblies were checked for quality using Quast v5.0.2 (61) 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 (62) 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 (63) and visualized in R with ggtree (64). 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). 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 (65) and visualized in R with ggtree (64).

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 (66). 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 (67). piN/piS values were calculated for all pairwise combinations for each *S. oralis* core gene were calculated using Egglib (68), then the average piN/piS value for each gene was calculated.

Rarefaction & accumulation plots

Rarefaction and accumulation curves for the *S. oralis* clades were calculated from Roary gene presence absence files. Briefly, separate pangenome analyses were performed as described above for each individual clade 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 (69) 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.

***nrdM* allele swaps**

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

Growth curves

Growth curves in THB were performed in biological duplicate. Wild-type and Δ*nrdM* strains were grown overnight as described, then diluted to OD_{600nm} 0.05 in approximately 12 mL THB. OD_{600nm} was monitored every hour using a Thermo Scientific Genesys 30 spectrophotometer. For growth curves in the presence of human serum, biological triplicate overnight cultures were grown in streptococcal defined medium (45, 46, 70) and diluted to OD_{600nm} 0.1 in defined medium supplemented with 5% v/v human serum (Sigma-Aldrich). OD_{600nm} was monitored at 3, 6 and 24 hrs as described above.

Simulated Infective Endocarditis Vegetation Model (SIEVM)

Strains were inoculated from freezer stocks into 5 mL Mueller Hinton Broth (MHB) (BD Bacto) and incubated overnight at 37˚C and 5% CO_2. 1 mL was expanded into 100 mL pre-warmed MHB and incubated overnight as described above. The SIEVM was set up and performed as

previously described (18) using human blood products from the American Red Cross (UTD IRB 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 (71) combining SMRT long reads generated in this study and Illumina reads previously generated for SO48 (accession PRJNA354070;(72)).

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. Doern CD, Carey-Ann BD. 2010. It's Not Easy Being Green: the Viridans Group Streptococci, with a Focus on Pediatric Clinical Manifestations. *Journal of Clinical Microbiology* 48:3829-3835.
2. Gudiol C, Bodro M, Simonetti A, Tubau F, González-Barca E, Cisnal M, Domingo-Domenech E, Jiménez L, Carratalà J. 2013. Changing aetiology, clinical features, antimicrobial resistance, and outcomes of bloodstream infection in neutropenic cancer patients. *Clinical Microbiology and Infection* 19:474-479.
3. Marín M, Gudiol C, Garcia-Vidal C, Ardanuy C, Carratalà J. 2014. Bloodstream infections in patients with solid tumors: Epidemiology, antibiotic therapy, and outcomes in 528 episodes in a single cancer center. *Medicine* 93:143-149.
4. Murdoch DR, Corey GR, Hoen B, Miró JM, Fowler VG, Bayer AS, Karchmer AW, Olaison L, Pappas PA, Moreillon P, Chambers ST, Chu VH, Falcó V, Holland DJ, Jones P, Klein JL, Raymond NJ, Read KM, Tripodi MF, Utili R, Wang A, Woods CW, Cabell CH, International Collaboration on Endocarditis-Prospective Cohort Study I. 2009. Clinical presentation, etiology, and outcome of infective endocarditis in the 21st century: the International Collaboration on Endocarditis-Prospective Cohort Study. *Archives of Internal Medicine* 169:463-473.
5. Teles C, Smith A, Ramage G, Lang S. 2011. Identification of clinically relevant viridans group streptococci by phenotypic and genotypic analysis. *European Journal of Clinical*

Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology 30:243-250.

6. Hoshino T, Fujiwara T, Kilian M. 2005. Use of phylogenetic and phenotypic analyses to identify nonhemolytic streptococci isolated from bacteremic patients. *Journal of Clinical Microbiology* 43:6073-6085.
7. Isaksson J, Rasmussen M, Nilson B, Stadler LS, Kurland S, Olaison L, Ek E, Herrmann B. 2015. Comparison of species identification of endocarditis associated viridans streptococci using rnpB genotyping and 2 MALDI-TOF systems. *Diagnostic Microbiology and Infectious Disease* 81:240-245.
8. Shelburne SA, Sahasrabhojane P, Saldana M, Yao H, Su X, Horstmann N, Thompson E, Flores AR. 2014. *Streptococcus mitis* strains causing severe clinical disease in cancer patients. *Emerging Infectious Diseases* 20:762-771.
9. Bishop CJ, Aanensen DM, Jordan GE, Kilian M, Hanage WP, Spratt BG. 2009. Assigning strains to bacterial species via the internet. *BMC biology* 7:3.
10. Velsko IM, Perez MS, Richards VP. 2019. Resolving Phylogenetic Relationships for *Streptococcus mitis* and *Streptococcus oralis* through Core- and Pan-Genome Analyses. *Genome Biol Evol* 11:1077-1087.
11. Galloway-Peña J, Sahasrabhojane P, Tarrand J, Han XY, Shelburne SA. 2014. GyrB Polymorphisms Accurately Assign Invasive Viridans Group Streptococcal Species. *Journal of Clinical Microbiology* 52:2905-2912.
12. Sahasrabhojane P, Galloway-Peña J, Velazquez L, Saldaña M, Horstmann N, Tarrand J, Shelburne SA. 2014. Species-level assessment of the molecular basis of fluoroquinolone resistance among viridans group streptococci causing bacteraemia in cancer patients. *International Journal of Antimicrobial Agents* 43:558-562.
13. Kitten T, Munro CL, Zollar NQ, Lee SP, Patel RD. 2012. Oral streptococcal bacteremia in hospitalized patients: taxonomic identification and clinical characterization. *Journal of Clinical Microbiology* 50:1039-1042.
14. Chun S, Huh HJ, Lee NY. 2015. Species-specific difference in antimicrobial susceptibility among viridans group streptococci. *Annals of Laboratory Medicine* 35:205-211.
15. Kim SL, Gordon SM, Shrestha NK. 2018. Distribution of streptococcal groups causing infective endocarditis: a descriptive study. *Diagnostic Microbiology and Infectious Disease* 91:269-272.
16. Kilian M, Riley DR, Jensen A, Brüggemann H, Tettelin H. 2014. Parallel Evolution of *Streptococcus pneumoniae* and *Streptococcus miti*s to Pathogenic and Mutualistic Lifestyles. *mBio* 5:e01490-14.
17. Kilian M, Poulsen K, Blomqvist T, Håvarstein LS, Bek-Thomsen M, Tettelin H, Sørensen UBS. 2008. Evolution of *Streptococcus pneumoniae* and Its Close Commensal Relatives. *PLoS ONE* 3:e2683.
18. Akins RL, Katz BD, Monahan C, Alexander D. 2015. Characterization of high-level daptomycin resistance in viridans group streptococci developed upon in vitro exposure to daptomycin. *Antimicrobial Agents and Chemotherapy* 59:2102-2112.
19. Galloway-Peña J, Sahasrabhojane P, Tarrand J, Han XY, Shelburne SA. 2014. GyrB polymorphisms accurately assign invasive viridans group streptococcal species. *Journal of clinical microbiology* 52:2905-12.

20. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20:257.
21. Facklam R. 2002. What happened to the streptococci: Overview of taxonomic and nomenclature changes. *Clinical Microbiology Reviews* 15:613-630.
22. Doern CD, Burnham CAD. 2010. It's not easy being green: The Viridans Group Streptococci, with a focus on pediatric clinical manifestations. *Journal of Clinical Microbiology* 48:3829-3835.
23. Furuichi M, Horikoshi Y. 2018. Sites of infection associated with *Streptococcus anginosus* group among children. *J Infect Chemother* 24:99-102.
24. Pearce MM, Hilt EE, Rosenfeld AB, Zilliox MJ, Thomas-White K, Fok C, Kliethermes S, Schreckenberger PC, Brubaker L, Gai X, Wolfe AJ. 2014. The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *mBio* 5:e01283-14.
25. Pearce C, Bowden GH, Evans M, Fitzsimmons SP, Johnson J, Sheridan MJ, Wientzen R, Cole MFY. Identification of pioneer viridans streptococci in the oral cavity of human neonates. *Journal of Medical Microbiology* 42:67-72.
26. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the Normal Bacterial Flora of the Oral Cavity. *Journal of Clinical Microbiology* 43:5721-5732.
27. Mager DL, Ximenez-Fyvie LA, Haffajee AD, Socransky SS. 2003. Distribution of selected bacterial species on intraoral surfaces. *Journal of Clinical Periodontology* 30:644-654.
28. Peterson SN, Snesrud E, Liu J, Ong AC, Kilian M, Schork NJ, Bretz W. 2013. The Dental Plaque Microbiome in Health and Disease. *PLOS ONE* 8:e58487.
29. Caselli E, Fabbri C, D'Accolti M, Soffritti I, Bassi C, Mazzacane S, Franchi M. 2020. Defining the oral microbiome by whole-genome sequencing and resistome analysis: the complexity of the healthy picture. *BMC Microbiology* 20:120.
30. Al-hebshi NN, Nasher AT, Maryoud MY, Homeida HE, Chen T, Idris AM, Johnson NW. 2017. Inflammatory bacteriome featuring *Fusobacterium nucleatum* and *Pseudomonas aeruginosa* identified in association with oral squamous cell carcinoma. *Scientific Reports* 7:1834.
31. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691-3693.
32. Mitchell J. 2011. *Streptococcus mitis*: Walking the line between commensalism and pathogenesis. *Molecular Oral Microbiology* 26:89-98.
33. Chaguza C, Cornick JE, Everett DB. 2015. Mechanisms and impact of genetic recombination in the evolution of *Streptococcus pneumoniae*. *Computational and Structural Biotechnology Journal* 13:241-247.
34. Andam CP, Hanage WP. 2015. Mechanisms of genome evolution of *Streptococcus*. *Infection, Genetics and Evolution* 33:334-342.
35. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* 43:e15.
36. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS computational biology* 11:e1004041.

37. Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, Bowden R, Auton A, Votintseva A, Larner-Svensson H, Charlesworth J, Golubchik T, Ip CLC, Godwin H, Fung R, Peto TEA, Walker AS, Crook DW, Wilson DJ. 2014. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. *Nature Communications* 5:3956.
38. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology* 17:238.
39. Rhodes DV, Crump KE, Makhlynets O, Snyder M, Ge X, Xu P, Stubbe J, Kitten T. 2014. Genetic Characterization and Role in Virulence of the Ribonucleotide Reductases of *Streptococcus sanguinis*. *Journal of Biological Chemistry* 289:6273-6287.
40. Collins C, Didelot X. 2018. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS computational biology* 14:e1005958.
41. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, Woodford N, Smith EG, Ismail N, Llewelyn MJ, Peto TE, Crook DW, McVean G, Walker AS, Wilson DJ. 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology* 1:16041.
42. Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. *Genetical Research* 67:159-174.
43. Mortimer TD, Weber AM, Pepperell CS. 2018. Signatures of Selection at Drug Resistance Loci in *Mycobacterium tuberculosis*. *mSystems* 3:e00108-17.
44. Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* 4.
45. Joyce LR, Guan Z, Palmer KL. 2019. Phosphatidylcholine Biosynthesis in Mitis Group Streptococci via Host Metabolite Scavenging. *Journal of bacteriology* 201:e00495-19.
46. Joyce LR, Guan Z, Palmer KL. 2021. *Streptococcus pneumoniae*, *S. pyogenes* and *S. agalactiae* membrane phospholipid remodelling in response to human serum. *Microbiology (Reading)* 167.
47. Azarian T, Martinez PP, Arnold BJ, Qiu X, Grant LR, Corander J, Fraser C, Croucher NJ, Hammitt LL, Reid R, Santosham M, Weatherholtz RC, Bentley SD, O'Brien KL, Lipsitch M, Hanage WP. 2020. Frequency-dependent selection can forecast evolution in *Streptococcus pneumoniae*. *PLOS Biology* 18:e3000878.
48. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, Fung R, Godwin H, Knox K, Votintseva A, Everitt RG, Street T, Cule M, Ip CLC, Didelot X, Peto TEA, Harding RM, Wilson DJ, Crook DW, Bowden R. 2013. Within-Host Evolution of *Staphylococcus aureus* during Asymptomatic Carriage. *PLOS ONE* 8:e61319.
49. Muthukumarasamy U, Preusse M, Kordes A, Koska M, Schniederjans M, Khaledi A, Häussler S. 2020. Single-Nucleotide Polymorphism-Based Genetic Diversity Analysis of Clinical *Pseudomonas aeruginosa* Isolates. *Genome Biology and Evolution* 12:396-406.
50. Bohr LL, Youngblom MA, Eldholm V, Pepperell CS. 2021. Genome reorganization during emergence of host-associated *Mycobacterium abscessus*. *Microbial Genomics* 7:000706.
51. Bailey SF, Alonso Morales LA, Kassen R. 2021. Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments. *Genome Biology and Evolution* 13:evab141.

52. Bailey SF, Hinz A, Kassen R. 2014. Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. *Nature Communications* 5:4076.
53. Paik S, Senty L, Das S, Noe JC, Munro CL, Kitten T. 2005. Identification of Virulence Determinants for Endocarditis in *Streptococcus sanguinis* by Signature-Tagged Mutagenesis. *Infection and Immunity* 73:6064-6074.
54. Rasmussen LH, Højholt K, Dargis R, Christensen JJ, Skovgaard O, Justesen US, Rosenvinge FS, Moser C, Lukjancenko O, Rasmussen S, Nielsen XCYR. In silico assessment of virulence factors in strains of *Streptococcus oralis* and *Streptococcus mitis* isolated from patients with Infective Endocarditis. *Journal of Medical Microbiology* 66:1316-1323.
55. Baker SP, Nulton TJ, Kitten T. 2019. Genomic, Phenotypic, and Virulence Analysis of *Streptococcus sanguinis* Oral and Infective-Endocarditis Isolates. *Infection and Immunity* 87:e00703-18.
56. Iversen KH, Rasmussen LH, Al-Nakeeb K, Armenteros JJA, Jensen CS, Dargis R, Lukjancenko O, Justesen US, Moser C, Rosenvinge FS, Nielsen XC, Christensen JJ, Rasmussen S. 2020. Similar genomic patterns of clinical infective endocarditis and oral isolates of *Streptococcus sanguinis* and *Streptococcus gordonii*. *Scientific Reports* 10:2728.
57. Galkiewicz JP, Kellogg CA. 2008. Cross-kingdom amplification using bacteria-specific primers: complications for studies of coral microbial ecology. *Appl Environ Microbiol* 74:7828-31.
58. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15:R46.
59. Andrews S. 2010. FastQC: A Quality Control tool for High Throughput Sequence Data, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
60. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19:455-477.
61. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)* 29:1072-1075.
62. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* doi:10.1093/bioinformatics/btu153:btu153.
63. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* 5:e9490.
64. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8:28-36.
65. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312-1313.
66. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. 2018. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 34:292-293.

67. Lee I, Ouk Kim Y, Park S-C, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology* 66:1100-1103.
68. De Mita S, Siol M. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genetics* 13:27.
69. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* 2.
70. van de Rijn I, Kessler RE. 1980. Growth characteristics of group A streptococci in a new chemically defined medium. *Infection and Immunity* 27:444-448.
71. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology* 13:e1005595.
72. Adams HM, Joyce LR, Guan Z, Akins RL, Palmer KL. 2017. *Streptococcus mitis* and *S. oralis* Lack a Requirement for CdsA, the Enzyme Required for Synthesis of Major Membrane Phospholipids in Bacteria. *Antimicrobial agents and chemotherapy* 61:e02552-16.

Supplementary figures

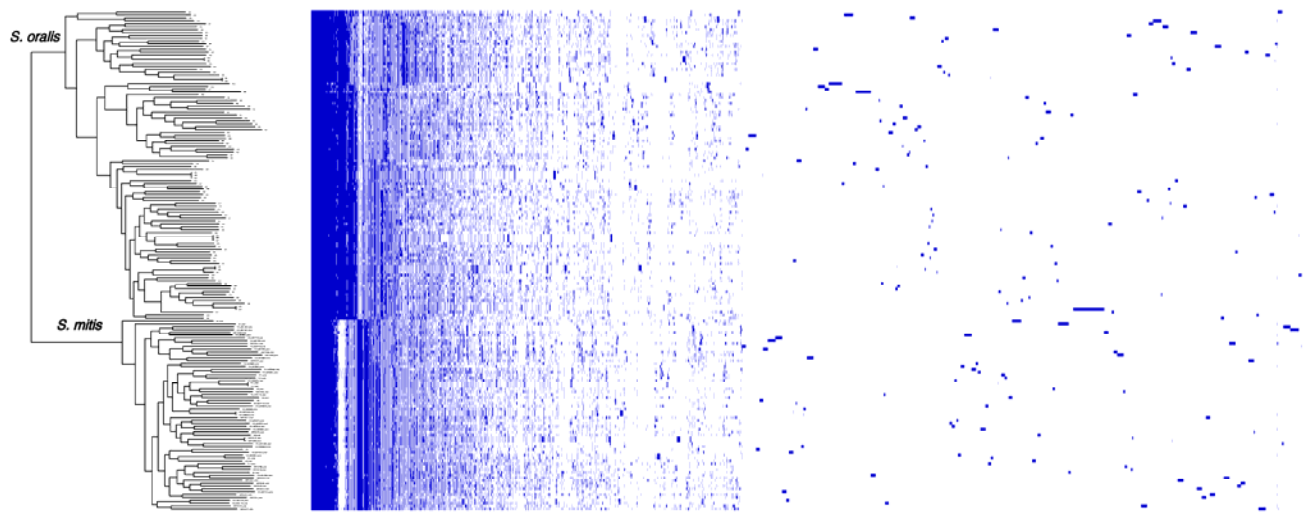


Figure S1: Core genome phylogeny of all *S. oralis* and *S. mitis* isolates (from newly sequenced samples and from NCBI) alongside a matrix of gene content. Clear delineation between *S. oralis* and *S. mitis* can be seen both in the core genome sequence as well as in core and accessory gene content differences between the species. Pangenome analysis performed using Roary and visualization with Phandango.

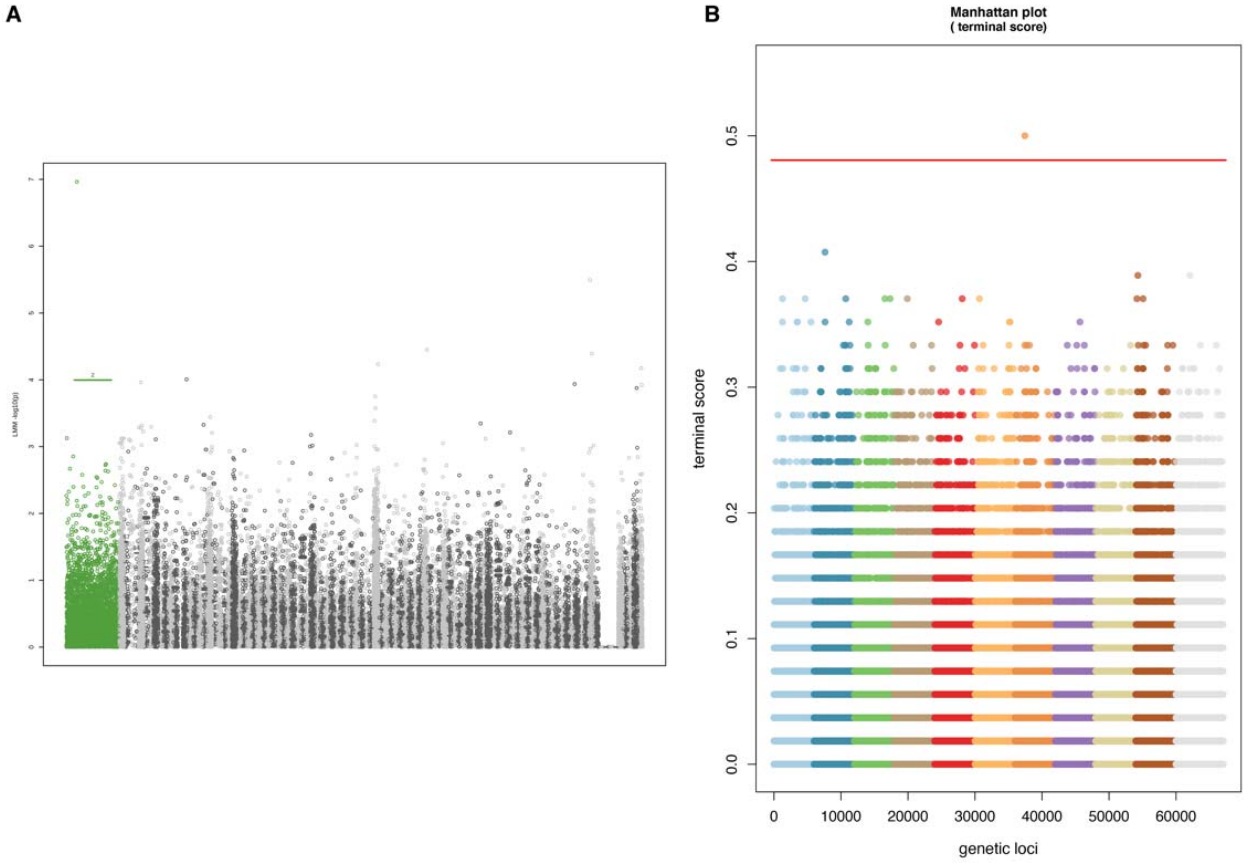


Figure S2: GWAS results confirming results of F_{ST} outlier analysis. BugWAS (left) and TreeWAS (right) results each showing a single significant variant associated with invasiveness: the same variant in *nrdM* identified in the F_{ST} outlier analysis.

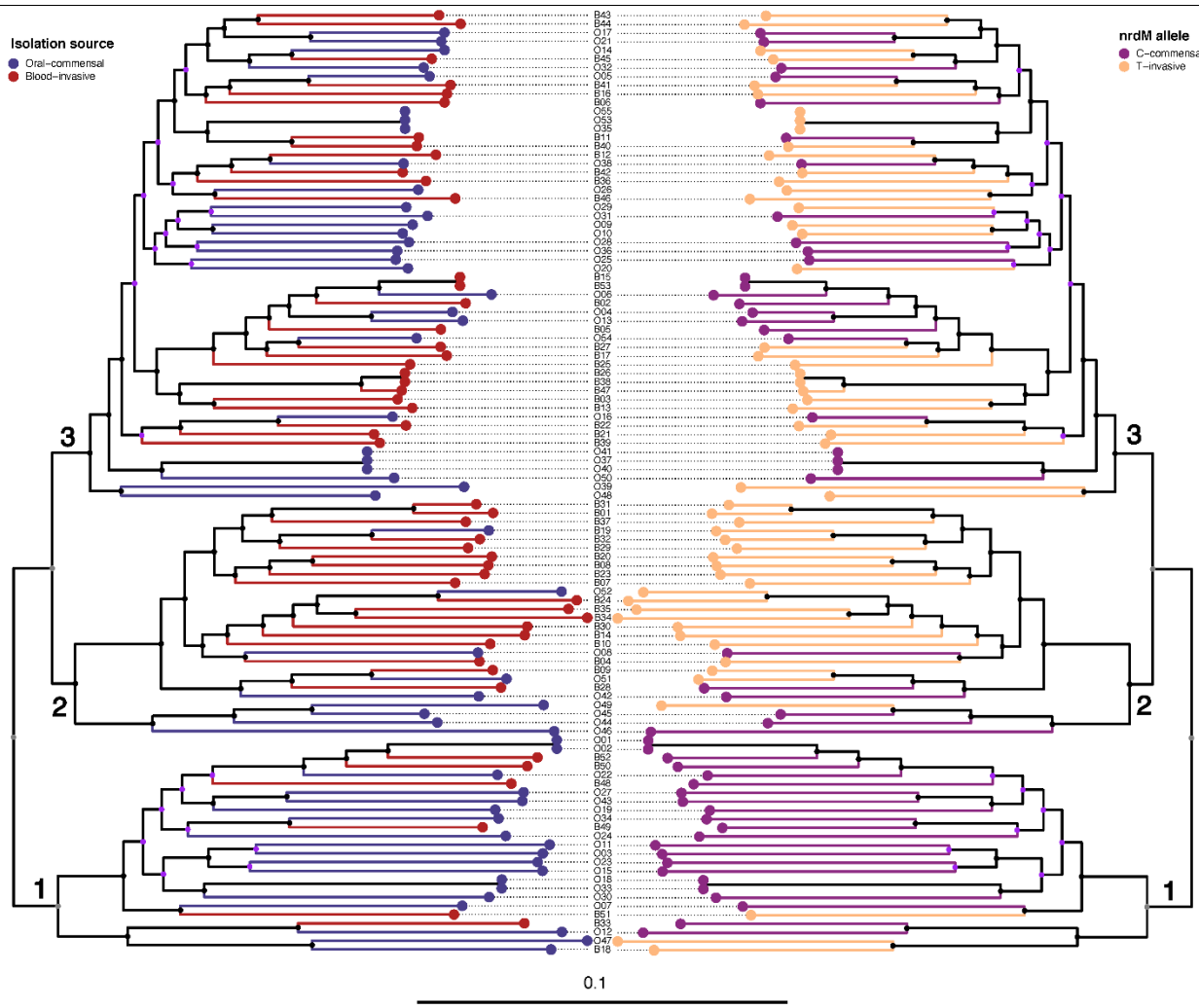


Figure S3: Comparison between isolation source (oral commensal or blood/invasive infection) and *nrdM* allele (C or T) shown on mirrored core genome phylogenies. Tree tips connected by dotted lines have convergent genotype-phenotype combinations (*i.e.* the *nrdM* allele associated with invasiveness in an isolate from an invasive infection).

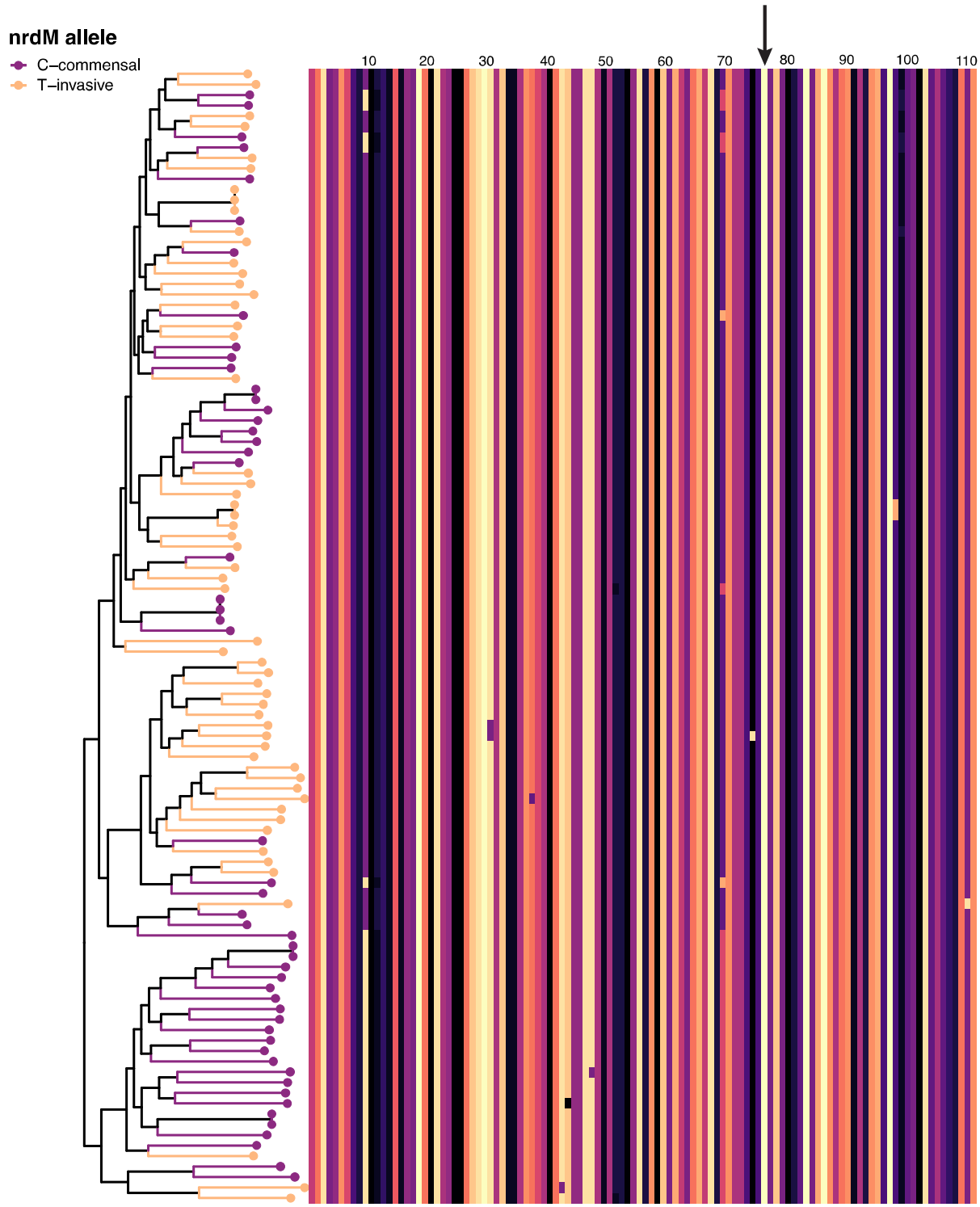


Figure S4: Multiple sequence alignment (MSA) of NrdM plotted next to the *S. oralis* core genome phylogeny where tips have been colored by *nrdM* allele. Scale on top of the MSA represents length in amino acids, the position of the synonymous mutation of interest (I78I) indicated with an arrow.

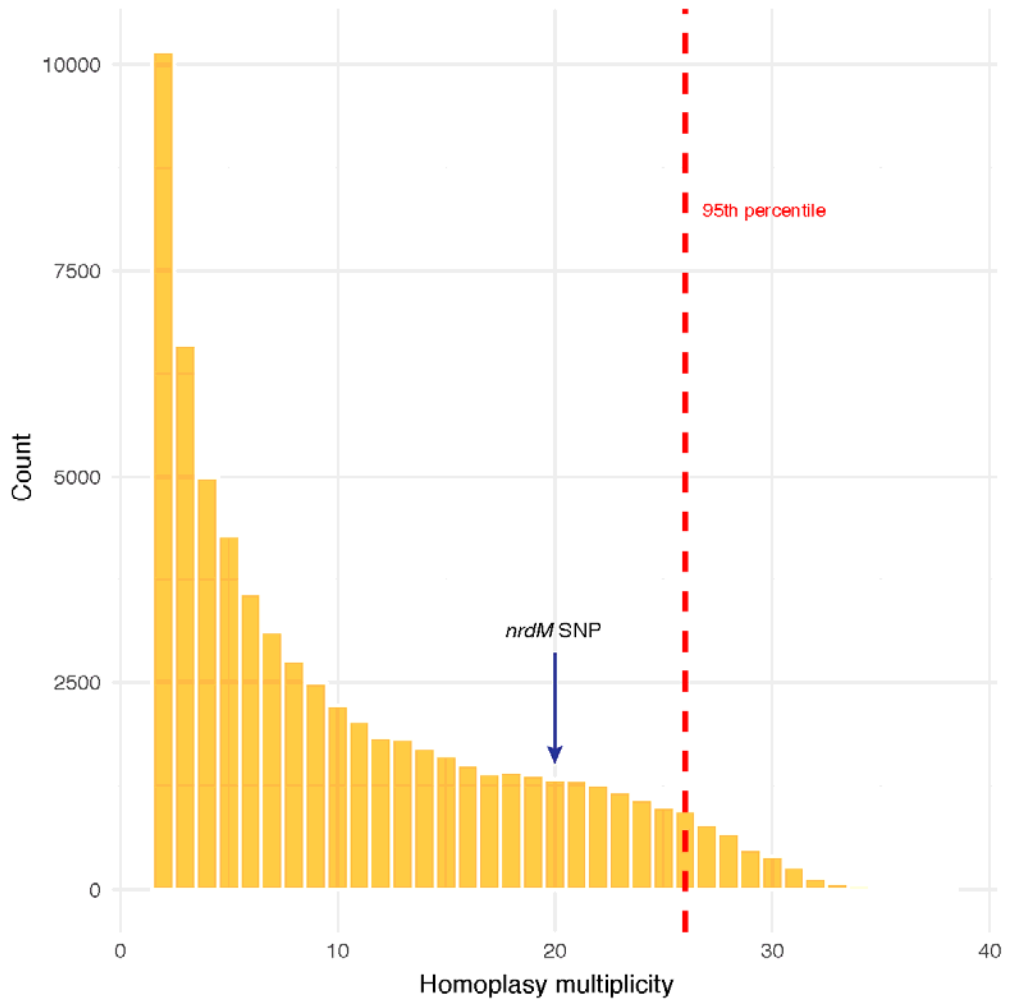


Figure S5: Histogram of mutation multiplicity for all homoplastic mutations on the *S. oralis* core genome phylogeny. Red dotted line shows 95th percentile cutoff and the multiplicity of the *nrdM* SNP (20) indicated by the black arrow.

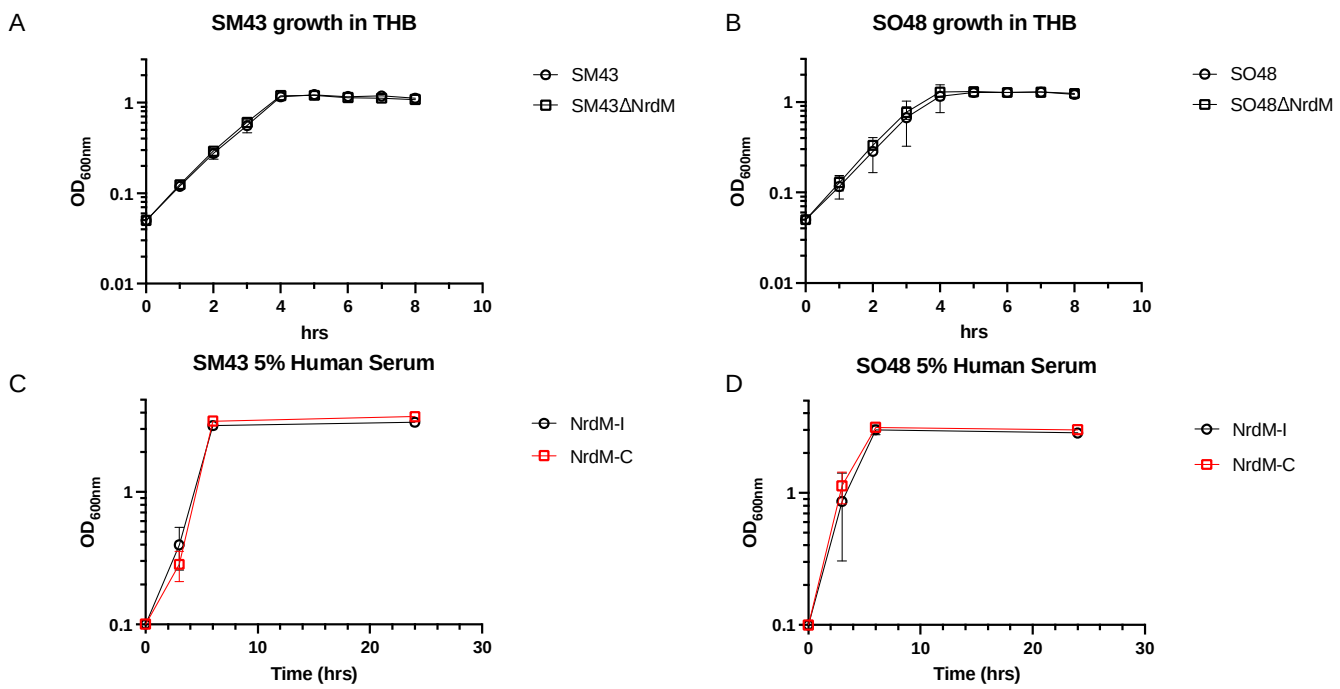


Figure S6: Growth curves of mutant strains. Growth curves of wildtype and Δ*nrdM* strains of A) SM43 and B) SO48 performed in Todd-Hewitt Broth, in biological duplicate. OD600 readings were performed every hour for 8 hrs. Mean and SD indicate. Growth of NrdM-I and NrdM-C alleles in C) SM43 and D) SO48 strain backgrounds when grown in streptococcal defined medium supplemented 5% v/v human serum. Serum growth was performed in biological triplicate and OD600 readings were taken at 0, 3, 6, and 24 hrs. Mean and SEM indicated.

Table S1: Newly sequenced VGS isolates

Table S2: *S. oralis* GWAS Strains

Table S3: Primers used in this study