Supplemental Materials

Strategy for estimating genome abundance.

The current literature includes a variety of approaches for profiling the abundance of taxa in metagenomes based on mapping reads to reference sequences. We chose the approach used in this study based on the unique characteristics of oral streptococci. Abundance measurements are frequently carried out by mapping to species-specific genes or taxonomically meaningful marker genes (Segata et al., 2012; Sunagawa et al., 2013; Nayfach et al., 2016). By selecting for a subset of genes found only in the taxon of interest, these methods have the benefit of reducing noise caused by cross-mapping of common genes found in related species. However, after correcting the taxonomy of genomes (Fig. S1), we found that some of the oral *Streptococcus* species were distinguished by few to no unique marker genes; the presence of a species-specific set of core genes was contingent on the parameters used for clustering. We therefore chose to profile taxon abundance by mapping to whole genomes (Delmont and Eren, 2018; Eren et al., 2021). A mapping test with computationally generated reads indicated that non-specific mapping could be minimized with our reference genome set by using only the nucleotide positions in the 2nd and 3rd quartiles for coverage to calculate the abundance for a genome (Figs. S5-S7; Tables S10-S13).

Because the number of available cultivar genomes varies greatly between oral *Streptococcus* spp., inclusion of all available genomes would bias the read recruitment in favor of the more densely-sampled taxa. Therefore, following best practices to minimize such bias (Delmont and Eren, 2018; Almeida et al., 2019), we selected a reference genome set that included one representative from each group of genomes that shared a given ANI threshold. Specifically, we selected from among the RefSeq genomes for each species a set of genomes that shared no more than 95% ANI. The mapping tests also indicated that oral *Streptococcus* species could be successfully detected by mapping against the set of genomes that shared no more than 95% ANI; a large fraction of simulated reads mapped, and mapped to the correct species (Table S12), regardless of whether the 95% ANI criterion caused the species to be represented by a single genome or by dozens of genomes.

Test of mapping specificity using computationally-generated short reads indicates that cross-mapping occurs at low levels and generally to closely related species. The accuracy with which a short-read mapping strategy can link a metagenomic sequence with its source species is limited by the degree to which sequences in different target species are equally good matches to the same metagenomic read. In the mapping process, reference genomes from isolated *Streptococcus* spp. strains act as bait to attract reads from the complex mixed population found in the mouth. The number of reads mapped should permit an accurate

estimate of species composition if the short reads from one strain map much more frequently to the genome of the bait strain of that species than to a genome from a different species. To test the accuracy of this expectation, we generated simulated short-read samples and mapped them to the selected set of *Streptococcus* spp. reference genomes. To generate each simulated read set, we computationally generated short reads from a single reference genome so that the simulated short reads covered the template genome to a mean depth of 100x across all nucleotide positions. As templates, we chose type strain genomes that were already in the reference set of genomes to which the reads were mapped; all short reads from these genomes can find a match to their source genome in the reference set but some may map instead to identical regions in other genomes (Table S1). As additional templates, we chose genomes that were not present in the reference set but were from the same species and shared at least 95% ANI with a genome in the reference set, a situation that more closely approximates the expected composition of natural samples from the mouth. As non-*Streptococcus* spp. controls, we also used type strain genomes from other major human oral genera. We mapped the simulated samples to the reference genome set competitively (i.e., each read was compared against a file containing all the reference genomes, so that nucleotide positions across all genomes competed to match the read). In cases where a species was represented by more than one genome, we summed the mean depth of coverage of all the reference genomes belonging to that species and report this species-level coverage value. We calculated two depth of coverage metrics with anvi'o: the average depth of coverage across all nucleotides in the genome (mean depth of coverage) and the average depth of coverage across nucleotide positions in the 2nd and 3rd quartiles when the nucleotides are ranked by their depth of coverage (Q2Q3 mean depth of coverage). Q2Q3 mean depth of coverage has been used as an alternative to mean depth of coverage to reduce bias due to highly conserved sequences that may cause non-specific mapping (Wang and Hong, 2020; Sieradzki et al., 2021; Martínez-Pérez et al., 2022). This resulted in lower values for mean depth of coverage but also lower measured cross-mapping between species relative to when depth of coverage was averaged across the entire genome (Figs. S5, S6; Tables S12, S13).

The results indicated that mapping accurately identified the source species of reads. When the simulated sample used a template that was itself in the reference genome set, 99-100% of aligned reads mapped to the correct species (Fig. S5; Tables S12, S13). When the simulated sample used an oral *Streptococcus* spp. template not in the reference genome set not all reads aligned, but 88-100% of the aligned reads mapped to the correct species. Most cross-mapping was to closely related species, such as from *S. australis* to *S. rubneri*. Reads generated from genomes outside the genus *Streptococcus* generally did not map to the reference genome set (Fig. S5; Tables S12, S13). For all samples simulated from non-*Streptococcus* spp. templates, the percentage

of reads that aligned to the reference genomes was less than 0.5%, which is within roundoff error and indicates that the presence of reads from other genera in metagenomic samples is unlikely to influence the results of mapping to this reference genome set.

Assessing taxon abundance by read mapping to the reference genome set, using Q2Q3 mean depth of coverage (Fig. S5), results in coverage estimates less affected by cross-mapping from related species compared to the depth of coverage assessed from all four quartiles (Fig. S6). However, Q2Q3 mean depth of coverage can underestimate taxon abundance in genomes with unusual coverage patterns. In the tests with simulated reads from a cultivar genome not used in the reference set, designed to mimic a simplified natural sample, mean depth of coverage of *S. mutans*, *S. cristatus*, *S. sanguinis*, and *S. oralis* ranged from 87.9 to 98.9 (Fig. S6, Table S13). However, the Q2Q3 mean depth of coverage was 96.0 and 97.0 for *S. mutans* and *S. cristatus* but 15.3 and 31.5 for *S. oralis* and *S. sanguinis* (Fig. S7, Table S12). For species with one reference genome, like *S. mutans*, and some of the species with multiple reference genomes, like *S. cristatus*, most reads mapped to and evenly covered one reference genome, resulting in similar mean coverage for the 2nd and 3rd quartiles and for all four quartiles (Fig. S7A-B). For the other species with multiple reference genomes, like *S. oralis* and *S. sanguinis*, the low Q2Q3 mean coverage is explained by coverage patterns showing that reads from *S. oralis* mapped to a small number of genomic regions, with high coverage, in each of many genomes (Fig. S7C-D). Due to their small size, these regions would fall within the 4th quartile and be ignored in the Q2Q3 calculation. This pattern may be due to high rates of recombination within *S. oralis*. For other species, reads recruited more evenly across the genome.

Figure S1: A phylogenomic tree based on 205 single-copy core genes (SCGs) indicates many of the mitis group reference genomes have incorrect NCBI species designations. The small text to the right of each node indicates the NCBI taxonomic designation of each genome. The colored labels indicate the revised species designation assigned to the genome. A "◆" indicates that the genome contains a > 99% identity match for the *S. pseudopneumoniae* marker genes and a "◇" indicates that the genome contains a > 99% identity match for the *S. pneumoniae* marker genes. Nodes that delineate species clusters are annotated with blue support values. The scalebar corresponds to a phylogenetic distance of 0.2 nucleotide substitutions per site.

Figure S2: Breadth of coverage also varies for species between oral sites and between strains of the same species within an oral site. The heatmaps show the breadth of coverage (the percentage of nucleotides with a coverage depth of at least 1x) for the oral *Streptococcus* species (A) and individual strains from species with more than one representative genome (B) for each of the metagenomes sampled across nine

oral sites. The values displayed for species with multiple reference genomes are the greatest coverage values, across all the genomes. There are 183 buccal mucosa (BM), 23 keratinized gingiva (KG), 1 hard palate (HP), 220 tongue dorsum (TD), 21 throat (TH), 31 palatine tonsils (PT), 209 supragingival plaque (SUPP), 32 subgingival plaque (SUBP), and 8 saliva (SV) samples. The samples are grouped by site and then ranked by descending number of total reads. The strains are first grouped by species and then ranked by descending mean relative abundance across the site (BM, TD, or SUPP) where they are most abundant. Note that *S. thermophilus* and *S. vestibularis* show consistent breadth of coverage in TD at modest levels (purple values for *S. thermophilus* and red values for *S. vestibularis* in part (A)) despite their low relative abundance in Fig. 2A. This consistent coverage likely results from cross-mapping from the highly abundant *S. salivarius*, whose genome has a relatively high ANIb of 92% with *S. vestibularis* and 89% with *S. thermophilus* (Table S2). In mapping with simulated reads from a *S. salivarius* cultivar genome from outside the reference set, three-fourths of the mapped reads mapped correctly to *S. salivarius*, but a significant fraction cross-mapped to *S. vestibularis* (14%) and *S. thermophilus* (4%) (Table S13, column AE). The fraction of cross-mapping reads is significantly lower (4% and 0.2% respectively) when assessing depth of coverage from nucleotides in the 2nd and 3rd quartiles of coverage (Table S11, column AE).

Figure S3: Breadth of coverage validates site tropisms and indicates how well the sequenced genome matches the gene content of the population in the mouth. The radial heatmap displays the breadth of coverage of the predicted genes from a representative genome from the 30 buccal mucosa, tongue dorsum, and supragingival plaque samples with the most quality-filtered reads. Each radius represents a predicted gene. Each concentric ring represents a metagenomic sample. Genes are black if their breadth of coverage is < 90% and color-coded by site if their breadth of coverage is ≥ 90%. The genes are arranged by breadth of coverage. Representative genomes are shown for species with a mean relative abundance ≥ 3% in at least one site. The genomes displayed here are the genomes from each species with the greatest Q2Q3 mean depth of coverage averaged across all metagenomes and whose species designation at NCBI matched our corrected species designations.

Figure S4: Analysis of phylogeny and analysis of gene content produce congruent results and cluster genomes into the same species-level groups. The phylogenomic tree was constructed using 205 single-copy genes core to the oral streptococci. The pangenomic tree was constructed using the frequencies with which each of the 18,895 genes is present in each genome. Lines connect the end nodes that represent the same genome. Colored boxes indicate species-level clades that contain multiple genomes.

Figure S5: Simulated reads map with a high degree of specificity. For each simulated read sample, the matrix displays the Q2Q3 mean depth of coverage summed across all reference genome with the same species. Mean depth of coverage values are displayed for the ranges 0-100x (A) and 0-20x (B). The reference genome species are arranged by their approximate order in the pangenome. The simulated samples are grouped into reads simulated from streptococci sequences in the reference genome set, streptococci sequences not in the reference genome set, and sequences from other major oral genera. Within the first two groups, the samples are arranged by the order of their species in the pangenome. Gray columns in the heatmap correspond to species for which there was not a RefSeq genome that was not already included in the reference genome set.

Figure S6: More cross-mapping is detected when depth of coverage is averaged across all nucleotide positions. For each simulated read sample, the matrix displays the total mean depth of coverage summed across all reference genome with the same species. The depth of coverage was average across all nucleotide positions. Mean depth of coverage values are displayed for the ranges 0-100 (A) and 0-20 (B). The reference genome species are arranged by their approximate order in the pangenome. The simulated samples are grouped into reads simulated from streptococci sequences in the reference genome set, streptococci sequences not in the reference genome set, and sequences from other major oral genera. Within the first two groups, the samples are arranged by the order of their species in the pangenome. Gray columns in the heatmap correspond to species for which there was not a RefSeq genome that was not already included in the reference genome set.

Figure S7: Gene-level mapping pattern explains anomalously low Q2Q3 mean depth of coverage for some species in the reference genome set. Assessing taxon abundance by read mapping to the reference genome set, using depth of coverage of nucleotide positions in the 2nd and 3rd quartiles of coverage (Fig. S5), results in coverage estimates less affected by cross-mapping from related species compared to the depth of coverage assessed from all four quartiles (Fig. S6). However, Q2Q3 mean coverage can underestimate taxon abundance in genomes with unusual coverage patterns. In tests with simulated reads from a cultivar genome not used in the reference set, the (A) *S. mutans* and (B) *S. cristatus* reference genomes had even coverage. By contrast, the (C) *S. sanguinis* and (D) *S. oralis* reference genomes had high coverage for a small number of genes in each genome. Summing the Q2Q3 mean coverage for each genome within such a taxon then leads to an underestimate of taxon abundance, as most of the coverage is present in Q4. Each line corresponds to the mean depth of coverage across the genes in one reference genome when the genes are ordered by increasing mean depth of coverage.

***S. pneumoniae* marker sequence from Croxen et al. (2018)**

ATGAGTACAAAATATTTATTTATTTACAATGAGATTCGTGAAAAGATTCTTTGTAATAAA
TATACCATGAACGAACAATTGCCTGATGAAATGACATTAGCTAAACAGTTTGCCTGTA
GTCGAATGACGATCAAAAAAGCTTTAGACTTGTTAGTTTCTGAGGGCTTAATTTTTAG
AAAACGTGGGCAGGGAACCTTTGTTCTCTCTCGTGGCAGCTCAAAAAGAAAATTAA
TCGTTCCAGAAAGAGATATCCGGGGACTGACAAAAATATCTGAAGATGCTCATTCTA
CAATTGACTCGAGGATTATTCACTTCAAATTAGAATTTGCAAATGAATTTTTAGCAGAA
AAACTACAGGTCGCTTTGCAGAGTCCAGTTTATAATATTTACCGCCTGCGTATTATTG
ACGGTAAACCTTATGTTCTGGAACAAACTTATATGAGTACCGATGTTATTCCAGGTATT
ACTGAAGATATTTTACAAAAATCGATTTACAATTACATTGAAGGAAAGTTAGGATTGCA
TATTGCCAGTGCTACAAAAATCTTACGAGCTTCTTCTAGTTCAGAAAATGAGCAACAT
TACTTGCAGCTCCTTCCAACGGAACCGGTATTTGAAGTAGAACAAGTGGCTTATTTG
GATAACGGAACTCCGTTTGAGTACTCGATTAGTCGTCATCGCTATGATTTATTTGAAT
TTAATTCTTTTGCATTACGACATTCCTCCTAG

***S. pseudopneumoniae* marker sequence from Croxen et al. (2018)**

ATGTATTACATGAAAAATGAAAATGTTAAGATTTTAATTTGTGAAGATGACTCTTCCGT
TAACAGACTTTTATCCTTAGCAATGGAAGTTGAAGGTTATCATTATGTATCAGTTCGG
ACTGGAGAGGAAGCTTTGCGTCAGATCATTTCGCAATTTCCAGATTTATTATTATTGG
ATTTGGGTTTGCCAGATATGGATGGTAAAGACATTATTGACAAGATTCGTAGCTTTTC
ACAGCTACCTGTTATTGTTGTTAGTGCACGTGGAGAAGAAAGTGACAAGATTGATGC
ACTTGATGCTGGGGCAGATGATTATTTGACGAAACCCTTTAGCATTGATGAGCTTTT
CGCTCGGTTAAGAGTTAGTCTTAGGAGGTCAAAGCAGATTAATCAACAAAGTGACG
GTAATTCTGAAAAATCATCTTTTACTAATGGCTGGCTACATGTTGATTTTTTATCTAATC
GTGTATTTGTTAATAACCAAGAAATTCACTTAACCCCGATTGAGTATAAGTTGCTTTGT
CTTCTATCAGAGAATGTTGATAGAGTGTTGACTTATCGTTTTATTGTCAAGGAAATTT
GGGGATATTATGAGGAAGATTTTTCTGCTTTGAGAGTTTTTGTTAATACATTGCGAAA
AAAAATCGAATTAGGATTGGGTTACTCTAAAATGGTTCAAACTCATATTGGTATCGGTT
ATCGTATGATTAAGATTGAAAATTATGATGACAAATAA

Table S1: Metadata for NCBI RefSeq genomes used in this study. The metadata for each genome sequence includes our species classification (column A), the NCBI species and subspecies classification (B), the strain name (C), whether the strain is the type strain for the species (D), whether our species classification is included in the eHOMD (E), the eHOMD taxon ID of the genome (F), the eHOMD sub-species clade classification of the genome (G), the purpose(s) for which the genome was used (H), the completeness (I) and contamination estimated with CheckM (J), the percent identity for the best match between the genome and the S. pneumoniae (K) and S. pseudopneumoniae marker sequences (L), and the RefSeq assembly accession (M). Additional metadata for each genome from NCBI include the host of the isolate (N), the isolate source (O), the isolation location (P), the BioSample accession (Q), the

BioProject accession (R), the assembly's level of completion (Complete: complete genome assembly, Chromosome: sequence for one or more chromosomes, scaffold - some contigs have been connected to form scaffolds, contigs - no assembly beyond the level of contigs) (S), the assembly size in megabases (Mb) (T), the GC-content expressed as the percent of the sequence (U), the number of scaffolds (V), the number of coding sequences (W), the release date (X), and the FTP address from which the genome was downloaded (Y). The genomes are arranged alphabetically, first by their corrected species designation, and then by their strain.

Table S2: ANI between each genome in the phylogenomic tree. The table displays the ANI, calculated using the BLAST algorithm with pyANI, between every genome used to construct the phylogenomic tree based on 205 SCGs. The genomes are arranged according to their placement in the phylogenomic tree in both directions.

Table S3: HMP metagenomic sample metadata. The metadata for each HMP metagenome includes our sample ID (column A), the NCBI Sequence Read Archive (SRA) accession (B), the full name of the sample site (C), the sample site abbreviation (D), the HMP subject ID of the donor (E) the sex of the subject (F), the total reads in the sample after quality filtering (G), the number of reads that mapped to the reference genome set (H), and the fraction of reads that mapped (I). The metagenomes are ordered by first by site and then by total number of reads.

Table S4: Q2Q3 mean depth of coverage across individual reference genomes for all samples. The reference genomes are ordered by species. The metagenomes are ordered first by site and then by the total number of reads.

Table S5: Breadth of coverage across individual reference genomes for all samples. This table contains the breadth of coverage averaged across all nucleotide positions in each reference genome for every HMP metagenome. The reference genomes are ordered by their order in the pangenome. The metagenomes are ordered first by site and then by the total number of reads.

Table S6: ANI between Pasolli et al. MAGs and reference genomes. The table displays the ANI, calculated using the BLAST algorithm with pyANI, between every putative *Streptococcus* sp. assembled from an oral HMP metagenome and the reference genomes. The reference genomes are arranged according to their placement in the phylogenetic tree. The MAGs are arranged based on the reference genomes they share the greatest ANI.

Table S7: Metadata for Pasolli et al. MAGs. The metadata for each genome includes the genome name (column A), the NCBI Sequence Read Archive (SRA) accession for

the metagenome which the MAG was assembled from (B), the sample site for the metagenome (C), the genome size in Mb (D), the N50 of the genome (E), the number of contigs in the genome (F), the completeness (G) and (H) contamination scores for the genome calculated with CheckM, the species-level name from NCBI of the best hit Pasoli et al. (2019) obtained when they BLASTed the genome sequence (I), and the species level name we assigned the genome (J).

Table S8: Summary of statistical tests. The summary includes the species (column A); the preferred site, buccal mucosa (BM), tongue dorsum (TD), or supragingival plaque (SUPP) (B); the number of metagenomes from each site (C-E); the mean relative abundance in each site (F-H); the standard deviation of the relative abundance in each site (I-K); the chi-squared value from the Kruskal-Wallis test (L); the p-value from the Kruskal-Wallis test (M); the z values from each Dunn's test pairwise comparison (N-P); the unadjusted p-values from each pairwise comparison in the Dunn's test (Q-S); and the Bonferroni-adjusted p-values from each pairwise comparison in the Dunn's test (T-V).

Table S9: Targeted *S. mitis*, *S. oralis*, and *S. infantis* pangenome summary. Each row corresponds to a different gene cluster. The summary includes the gene cluster id (column A); the number of *S. mitis*, *S. oralis*, and *S. infantis* genomes in the pangenome (B-D); the fraction of the genomes from each species in which the gene cluster is present (E-G); the gene cluster category (H); the representative NCBI COGs function (I), accession (J), and category (K); the representative Pfam function (M) and accession (N); and the representative eggNOG function (O) and accession (P).

Table S10: Q2Q3 mean depth of coverage of individual reference genomes for simulated samples. The reference genomes are ordered by their order in the pangenome. The simulated metagenomes are grouped according to the type of genome used as a template for the simulated reads: streptococci in reference genome set, streptococci not in reference genome set, other genera. Within the first two groups, the samples are arranged by the approximate order of their species in the pangenome.

Table S11: Total mean depth of coverage of individual reference genomes for simulated samples. The reference genomes are ordered by their order in the pangenome. The simulated metagenomes are grouped according to the type of genome used as a template for the simulated reads: streptococci in reference genome set, streptococci not in reference genome set, other genera. Within the first two groups, the samples are arranged by the approximate order of their species in the pangenome.

Table S12: Total Q2Q3 mean depth of coverage of species for simulated samples. The species genomes are ordered by their rank in the Fig. S5. The simulated

metagenomes are grouped according to the type of genome used as a template for the simulated reads: streptococci in reference genome set, streptococci not in reference genome set, other genera. Within the first two groups, the samples are arranged by the approximate order of their species in the pangenome.

Table S13: Total total mean depth of coverage of species for simulated samples. The species genomes are ordered by their rank in the Fig. S6. The simulated metagenomes are grouped according to the type of genome used as a template for the simulated reads: streptococci in reference genome set, streptococci not in reference genome set, other genera. Within the first two groups, the samples are arranged by the approximate order of their species in the pangenome.