Supplemental Material:

Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules

Authors: Baker, J.L.[1,*], Morton, J.T.[2]., Dinis, M.[3], Alverez, R.[3], Tran, N.C.[3], Knight, R.[4,5,6,7], Edlund, A.[1,5,*]

[1] Genomic Medicine Group
J. Craig Venter Institute
4120 Capricorn Lane
La Jolla, CA 92037

[2]Systems Biology Group
Flatiron Institute
162 5th Avenue
New York, NY 10010

[3]Section of Pediatric Dentistry
UCLA School of Dentistry
10833 Le Conte Ave.
Los Angeles, CA 90095-1668

[4]Center for Microbiome Innovation
University of California at San Diego
La Jolla, CA 92023

[5]Department of Pediatrics
University of California at San Diego
La Jolla, CA 92023

[6]Department of Computer Science and Engineering
University of California at San Diego
9500 Gilman Drive
La Jolla, CA 92093

[7]Department of Bioengineering
University of California at San Diego
9500 Gilman Drive
La Jolla, CA 92093

*Corresponding Authors: JLB: jobaker@jcvi.org, AE: aedlund@jcvi.org

ORCIDs: JLB: 0000-0001-5378-322X, AE: 0000-0002-3394-4804

SUPPLEMENTAL METHODS

Study Design. Subjects were included in the study if the subject was 3 years old or older, in good general health according to a medical history and clinical judgment of the clinical investigator, and had at least 12 teeth. Subjects were excluded from the study if they had generalized rampant dental caries, chronic systemic disease, or medical conditions that would influence the ability to participate in the proposed study (i.e., cancer treatment, HIV, rheumatic conditions, history of oral candidiasis). Subjects were also excluded it they had open sores or ulceration in the mouth, radiation therapy to the head and neck region of the body, significantly reduced saliva production or had been treated by anti-inflammatory or antibiotic therapy in the past 6 months. Ethnic origin was mixed for healthy subjects (Hispanic, Asian, Caucasian, Caucasian/Asian), while children with caries were of Hispanic origin. For the latter group, no other ethnic group enrolled despite several attempts to identify interested families/participants. Children with both primary and mixed dentition stages were included (caries group: 18 children with mixed dentition and 6 with primary dentition; healthy group: 19 children with mixed dentition and 6 with primary dentition). To further enable classification of health status (caries and healthy), a comprehensive oral examination of each subject was performed as described below. Subjects were dichotomized into two groups: caries free (dmft/DMFT = 0) and caries active (subjects with ≥2 active dentin lesions). If the subject qualified for the study, (s)he was to abstain from oral hygiene activity, and eating and drinking for 2 hours prior to saliva collection in the morning. An overview of the subjects and associated metadata is provided in Supplemental Table S1.

Oral examination and study groups. The exam was performed by a single calibrated pediatric dental resident (RA), using a standard dental mirror, illuminated by artificial light. The visual inspection was aided by tactile inspection with a community periodontal index (CPI) probe when necessary. Radiographs (bitewings) were taken to determine the depth of carious lesions. The number of teeth present was recorded and their dental caries status was recorded using decayed

(d), missing due to decay (m), or filled (f) teeth in primary and permanent dentitions (dmft/DMFT), according to the criteria proposed by the World Health Organization (1997) (Organization 1971). Duplicate examinations were performed on 5 randomly selected subjects to assess intra-examiner reliability. Subjects were dichotomized into two groups: caries free (CF; dmft/DMFT=0) and caries active (CA; subjects with ≥2 active dentin lesions). The gingival health condition of each subject was assessed using the Gingival Index (GI) (Loe 1967). GI data was published previously (Aleti et al. 2019). Additionally, parent/guardian of each participant completed a survey regarding oral health regimen.

Radiographic assessment. Bitewing radiographs were analyzed on the XDR Imaging Software (Los Angeles, CA). Lesion depth was determined with the measuring tool, and categorized as follows: E1 (radiolucency extends to outer half of enamel), E2 (radiolucency may extend to the dentinoenamel junction), D1 (radiolucency extends to the outer one-third of dentin), D2 (radiolucency extends into the middle one third of dentin), and D3 (radiolucency extends into the inner one third of dentin)(Anusavice 2005). To calculate the depth of lesion score, the following scores were assigned to each lesion depth: E1 = 1, E2 = 2, D1 = 3, D2 = 4, and D3 = 5, afterwards a total depth score was calculated for each subject.

Saliva collection. Unstimulated saliva was collected between 8:00-11:00am for the salivary immunological marker analysis. Subjects were asked to abstain from oral hygiene activity, and eating and drinking for two hours prior to collection. Before collection, subjects were instructed to rinse with water to remove all saliva from the mouth. In this study, unstimulated saliva was collected for salivary immunological marker analysis, while stimulated saliva (by chewing on sterile parafilm) was collected for Illumina sequencing (to dilute and amount of human DNA and material present). 2 ml of unstimulated saliva was collected from subjects by drooling/spitting directly into a 50mL Falcon conical tube (Fisher Scientific, Pittsburg PA) at regular intervals for a

period of 5-20 minutes. Saliva samples were immediately placed on ice and protease inhibitor cocktail (Sigma, MO, USA) was added at a ratio of 100uL per 1mL of saliva to avoid protein degradation. Then saliva samples were processed by centrifugation at 6,000 x g for 5 min at 4°C, and the supernatants were transferred to cryotubes. The samples were immediately frozen in liquid nitrogen and stored at -80 °C until analysis. 2 ml of stimulated saliva was collected immediately following collection of unstimulated saliva.

Sequencing read quality control. As reported in Aleti et al., 2019, raw Illumina reads were subjected to quality filtering and barcode trimming using KneadData v0.5.4 (available at https://bitbucket.org/biobakery/kneaddata) by employing trimmomatic settings of 4-base wide sliding window, with average quality per base >20 and minimum length 90 bp. Reads mapping to the human genome were also removed. KneadData quality control information is provided in Supplemental Table S1.

Assembly and binning of MAGs. metaSPAdes was utilized to *de novo* assemble metagenomes from the quality-filtered Illumina reads (Nurk et al. 2017). A separate assembly was performed for each sample, as opposed to a co-assembly assembly of all samples, an alternative approach used by some studies. The pros and cons of a co-assembly versus individual assemblies have been discussed previously (Pasolli et al. 2019). The resulting assemblies were binned using the MetaWRAP pipeline v1.1.5 (Uritskiy et al. 2018). The MetaWRAP initial_binning module used Maxbin2 (Wu et al. 2016), Metabat2 (Kang et al. 2019), and Concoct (Alneberg et al. 2014). Subsequently, the bin_refinement module was used to construct the best final bin set by comparing and selecting the most complete and least contaminated results of the 3 binning tools. The bin_reassembly module was then used to reassemble the refined bin set to further improve the final bins. The quality control cutoffs for all MetaWRAP modules were >50% completeness

and <10% contamination, which are the cutoffs for Medium-Quality Draft Metagenome-Assembled Genomes as described by the Genome Standards Consortium (Bowers et al. 2017). This approach generated 527 metagenome-assembled genomes (MAGs) that were at least of medium quality. Generally speaking, the samples with deeper sequencing provided more bins meeting this threshold, with sample SC33, with 249 million non-human reads yielding 69 bins, while sample SC26, with 1.4 million non-human reads yielding just 3 bins. The metaWRAP classify_bins and quant_bins modules were used to respectively obtain a taxonomy estimate and to provide the quantity of each bin in the form of 'genome copies per million reads'.

Dereplicating to species-level genome bins. As a result of the individual assemblies, many of the 527 MAGs were likely to represent redundant species across samples. fastANI (Jain et al. 2018) was used to compare the ANI of all 527 MAGs and generate a distance matrix that was used to dereplicate the MAGS into SGBs using a cutoff of 95%ANI. This distance matrix was visualized using Cytoscape (Shannon et al. 2003) (Figure 1A). Although a topic of some debate, 95%ANI has been used by several recent landmark studies as the cutoff for genomes representing the same species (Jain et al. 2018). To taxonomically identify species-level genome bins (SGBs), Mash v2.1 (Ondov et al. 2016) was used to query all 527 MAGs against the entire RefSeq database with a Mash distance cutoff of 5 (corresponding to a 95% average nucleotide identity (ANI)). All MAGs with a RefSeq hit with a Mash distance of <5 were assigned the species name of that hit. The fact that Mash distance <5 and fastANI ANI>95% aligned almost perfectly served as a useful internal control. There were several rare occasions in this dataset where SGBs, as defined by the 95% ANI distance matrix, included MAGs that best matched different (although closely related) RefSeq references. Whether this indicates that 95% ANI is not stringent enough (e.g. these should in fact be classified as multiple species) or too stringent (e.g. they should all be classified as the same species) is a debate beyond the scope of this work. This approach yielded 90 known species-level-genome bins (kSGBs), representing 399 MAGs with

≥95% ANI to a RefSeq genome (based on Mash), and 60 unknown SGBs (uSGBs), representing 128 MAGs (Figure 1A-G), with no genome in RefSeq with an ANI ≥95% (Figure 1A-G). MAG pie charts (Figure 4B-C) were created using Microsoft Excel and MAG quality violin chart were created using Graph Pad Prism. Statistical significance was determined using a Tukey's multiple comparisons post-test following a one-way ANOVA using Graph Pad Prism.

Determining taxonomy of unknown SGBs (uSGBs). Here, a strategy for classifying uSGBs into genus-level genome bins (GGBs), which have a 85%-95% ANI to a GenBank genome, and family-level genome bins (FGBs), which have no match ≥85% ANI to a GenBank genome, was employed, similar to the method described in Pasolli et al., 2019. The predicted family for each MAG was first inferred using the CheckM (Parks et al. 2015), Kraken (Wood and Salzberg 2014), and classify_bins tools from within the MetaWRAP pipeline. Next, because there are publicly available, and in many cases described, genomes in GenBank that do not appear in the RefSeq database used by Mash, each uSGB was compared against all GenBank genomes in its predicted family using fastANI. This process reassigned 18 uSGBs (rSGBs), representing 31 MAGs, to kSGBs, as they had ≥95% ANI match in GenBank (Supplemental Table S2). For the remaining "true" uSGBs, 20 uSGBs, representing 48 MAGs, that had 85%-95% ANI match to a GenBank genome were termed genus-level genome bins (GGBs), as the genus can be assigned with a fair amount of confidence, while the species appears to be not previously described. The final 22 bins, representing 49 MAGs, had no matching reference in GenBank with an ANI ≥85%. These were termed family-level genome bins (FGBs), as the family or higher-level taxa can be inferred, but the MAGs likely represent novel genera. When uSGBs contained multiple MAGs, the MAG with the best quality score according to the formula (completion – (2x contamination)) was used to find the best hit.

Phylogenetic placement of uSGBs. Anvi'o (Eren et al. 2015) was used to determine the phylogeny of the Saccharibacteria kSGBs and uSGBs. The Anvi'o Snakemake (Koster and Rahmann 2012) phylogenomics workflow utilizes muscle (Edgar 2004), trimAl (Capella-Gutierrez et al. 2009), and IQ-TREE (Nguyen et al. 2015). PhyloPhlAn2 (Pasolli et al. 2019) was used to determine the phylogeny of Clostridiales and Bacteroidales uSGBs. The following parameters were used: --diversity medium –accurate. The following PhyloPhlAn2 external tools were used: diamond (Buchfink et al. 2015), mafft (Katoh and Standley 2013), trimAl (Capella-Gutierrez et al. 2009), fasttree (Price et al. 2009), and RAxML (Stamatakis 2014). Resulting phylogenetic trees were visualized using iToL 4(Letunic and Bork 2019). Anvi'o was utilized to examine Saccharibacteria phylogeny because its pipeline was more amenable to the highly reduced number of single copy marker genes found in Saccharibacteria genomes. PhyloPhlAn2 was used to examine Bacteroidales and Clostridiales phylogeny because, compared to Anvi'o, it requires less preprocessing of the reference and sample genomes prior to actual phylogenetic analysis, making the pipeline more appropriate for these queries that have hundreds or thousands of genomes. Labels on the Bacteroidales and Clostridiales trees in Figure 2A-B indicate the genus of the GGB, and the family and most-closely related genus of the FGBs, as determined by fastANI against all Bacteroidales or Clostridiatles in GenBank. The Saccharibacteria clades were based upon the phylogeny and reference genomes published in (Figure 2C)(McLean et al. 2020). Differences in the phylogeny reported here compared to (McLean et al. 2020) are likely to result from a different set of marker genes used to construct the phylogeny.

Pangeomics and functional enrichment. Individual assembled genomes were annotated with Anvi'o (COG Functions) and eggnog-mapper v2 (Huerta-Cepas et al. 2017; Huerta-Cepas et al. 2019). Anvi'o was utilized to perform pangenomics analysis (Delmont and Eren 2018). This pipeline also utilized HMMSEARCH (Eddy 2011) and INFERNAL (Nawrocki and Eddy 2013). tRNAs were predicted using Anvi'o and 16S rRNA genes were predicted using the cmsearch

function of INFERNAL. 16S rRNA sequences were identified using the HOMD 16S rRNA Sequence Identification page of the expanded Human Oral Microbiome Database (eHOMD, http://www.homd.org/, (Escapa et al. 2018). To examine strain-level differences between caries and health-associated isolates, SGBs with at least 4 representatives from healthy subjects and 4 representatives from caries subjects were considered. 13 SGBs met these criteria: *Atopobium sp. ICM42b, Eubacterium sulci, Haemophilus parainfluenzae, Candidatus Lachnospiraceae FGB2, Mogibacterium diversum, Megasphaera micronuciformis, Prevotella histicola, Prevotella ICM33, Porphymonadaceae bacterium KA00676, Prevotella pallens, Prevotella salivae, Peptostreptococcus GGB1,* and *Candidatus Solobacterium GGB1.* Pangenomes of these SGBs and significantly enriched functional pathways between caries and health-associated isolates were examined using Anvi'o. There were no significantly enriched gene clusters in caries compared to health, or vice versa, within these pangenomes. Meanwhile, a pangenome of the Saccharibacteria genomes and several reference strains was examined, and there was a large amount of functional pathway enrichment between Saccharibacteria clades (Figure 3B-C, Supplemental Fig.S3A-B). KEGG KO Functional pathway occurrence frequency across the Saccharibacteria clades was exported using Anvi'o, and used to create the metabolic network in Supplemental Fig. S5B using KEGG Mapper (https://www.genome.jp/kegg/mapper.html).

Inference of actively replicating taxa. iRep was utilized (Brown et al. 2016) to estimate which taxa identified in the metagenomics analysis were alive and metabolically active, and to compare this data between health- and caries-associated microbiomes. iRep infers replication rates based upon differential sequencing coverage of genomic regions with respect to the origin of replication (Brown et al. 2016). iRep requires 75% completeness, <175 contigs per 1mb, and looks at contigs > 5000bp with a > 5 average coverage. iRep was able to compute replication rates for 183 of 527 MAGs. The replication rate of all bacteria was not significantly different among the taxa derived from the saliva of healthy children compared to the saliva from the children with caries

(Supplemental_Fig_S6.pdf). At the individual SGB level, there were only 8 SGBs with sufficient (at least 3) MAGs, which passed the requirements for iRep, from both the healthy and the caries groups. None of these 8 SGBs had statistically different rates of replication between the healthy and caries groups (Supplemental_Fig_S6.pdf).

Genome-based taxonomic abundance analysis. Overall, the taxonomy of the assembled MAGs largely reflected the taxonomy of the communities predicted by MetaPhAn2. One notable exception was the lack of Streptococci among the assembled bins. Difficulty assembling quality Streptococcal genomes from metagenomics datasets has been noted previously (Espinoza et al. 2018), and is thought to occur because the high promiscuity of *Streptococcus* k-mers (due to high intra-genera diversity within Streptococcus, particularly from oral samples). Assembly biases such as this are examples that support use of unassembled reads for taxonomic abundance profiling and similar analyses, as performed by MetaPhlAn2 and the downstream tools in this study. However, the taxa in the MetaPhlAn2 database do not match exactly with the taxa of the genomes assembled from these microbiomes, and it was difficult to determine whether any of the novel taxa assembled here are detected by MetaPhlAn2. To examine the abundances of the taxa with assembled genomes in this study, a BWA index was created using a database consisting of the best quality (completion – (2 x contamination)) genome from each SGB in this study. The post-QC sequencing reads were mapped to the database using BWA-MEM (Li et al. 2009; Li 2014). This was used to create taxonomic abundance table, which was analyzed by DEICODE (Supplemental Fig. S6C). Many of the overall trends were preserved compared to the DEICODE analysis of the unassembled reads, and several novel taxa (*Candidatus Nanosynbacter GGB3* and *GGB4*, and *Candidatus Gracilibacteria FGB1*) appeared to be associated with caries (Supplemental Fig. S6C). However, the authors recommend utilizing unassembled reads for abundances purposes to avoid biases introduced by the assembly process. A significant

difference in beta divsersity between the healthy and caries groups was observed by PERMANOVA performed using the QIIME2 diversity plugin.

Read-based taxonomy. Filtered reads were then analyzed using MetaPhlAn2 v2.7.5 (Truong et al. 2015) to determine relative abundances of taxa. The predicted total number of reads for each sample was multiplied by the relative abundances of each taxa within the sample to obtain an estimated number of reads of each taxa. The resulting OTU table was utilized for downstream diversity and correlation analysis. Unlike 16S sequencing, metagenomic sequencing detects viruses and eukaryotes in addition to bacteria. 12 viruses were detected in this study, including several human herpesviruses and several bacteriophage (Supplemental Fig. S4A,B,D). The viruses were detected at relatively low frequency and did not appear to be significant drivers of beta diversity in this study group (Figure 4A). The fungal pathogen, *Candida albicans* is known to be involved in pathogenesis in many cases of dental caries (reviewed in (Pereira et al. 2018)), therefore it was surprising that it was not detected by MetaPhlAn2 in this study. Mapping Illumina reads directly to the *C. albicans* genome indicated the presence of *C. albicans* in the samples, but the number of reads was small, and thus any fungal pathogens present in the study group were likely to be below the threshold of detection employed in taxonomic quantification by MetaPhlAn2 (data not shown). This is likely in part due to the use of DNA extraction methods designed for bacteria, not fungi.

Diversity analyses. The taxonomic abundance table (i.e. OTU table) generated from MetaPhlAn2 was used as input for QIIME2 (Bolyen et al. 2019). The QIIME2 diversity plugin was used to calculate alpha diversity (within sample diversity) (Bolyen et al. 2019). The QIIME2 plugin, DEICODE (Martino et al. 2019), was used to calculate beta diversity with feature loadings. DEICODE utilizes matrix completion and robust Aitchison principal components analysis (PCA), providing several advantages over other tools, including the ability to accurately handle sparse

datasets (e.g. in most microbial communities, most taxa are not present in a majority of samples), scale invariance (negating the need for rarefaction) and preservation of feature loadings (i.e. which taxa are driving the differences in PCA ordination space)(Martino et al. 2019). The resulting ordination was visualized using the QIIME2 plugin Emperor (Vazquez-Baeza et al. 2013). Taxa ranks driving differences in ordination space along axis 2 (the axis with the most difference in disease status) were visualized using Qurro . A significant difference in beta diversity between the caries and healthy groups was observed with a PERMANOVA using the QIIME2 diversity plugin.

Taxa associated with disease. The OTU table generated by MetaPhlAn2 was used as input for Songbird (Morton et al. 2019b), to rank species association with disease status. Songbird utilizes reference frames to rank the association of taxa with a given metadata category, and alleviates many of the issues caused by the compositional nature of sequencing data. Because this method is sensitive to sparsity, only species observed in at least 10 samples and having over 10,000 total predicted counts were analyzed. The following parameters were used in the songbird multinomial script: number of random test examples: 5, epochs: 50,000, batch size: 8, differential-prior: 1, learning rate: 0.001. Taxa ranks were visualized using Qurro. Log ratios of *Prevotella* to *Rothia, Haemophilus,* or *Neisseria* were extrapolated using Qurro, and graphed using Graph Pad Prism. Significance was determined using a Welch's *t*-test between groups, performed by Graph Pad Prism.

Functional profiling and diversity. HUMAnN2 (Franzosa et al. 2018) was used to provide abundance information about the functional pathways present in the metagenomes, which was also stratified by species. QIIME2, DEICODE, Songbird, Emperor, and Qurro were utilized to analyze the unstratified functional pathway data in a manner similar to that described for the taxonomic abundance information as described above. Significant differences in alpha and beta

diversity between the healthy and caries groups was observed with a Kruskal-Wallis test and a PERMANOVA, respectively, using the QIIME2 diversity plugin. Contributional diversity, as plotted in Figure 5C, was examined across the 69 core pathways that were present in each sample, and had at least 3 taxa contributing to the pathway. Contributional alpha and beta diversity of these pathways were calculated using QIIME2.

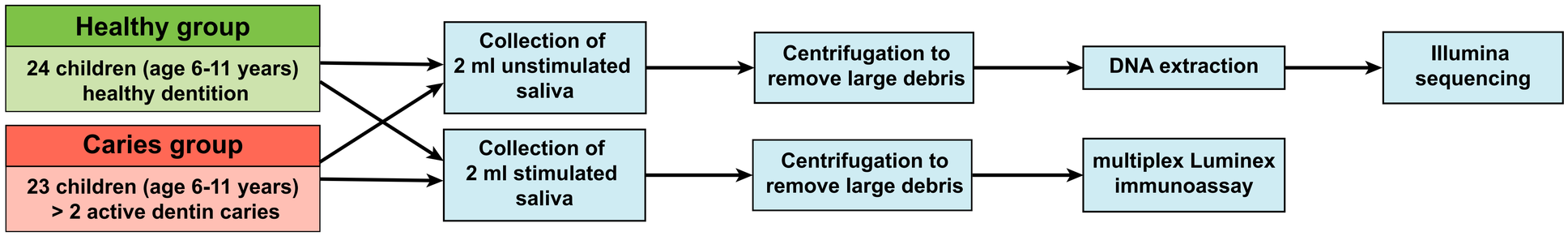
Species-immunological marker co-occurrences. The co-occurrences of species and immunological markers was estimated using neural networks via mmvec (Morton et al. 2019a). mmvec uses neural networks for estimating microbe-metabolite interactions through their co-occurrence probabilities. The following parameters were used: number of testing examples: 5, minimum feature count: 10, epochs: 1000, batch size: 3, latent dim: 3, input prior: 1, output prior: 1, learning rate 0.001. The QIIME2 Emperor plugin was used to visualize the resulting ordination.

SUPPLEMENTAL FIGURES

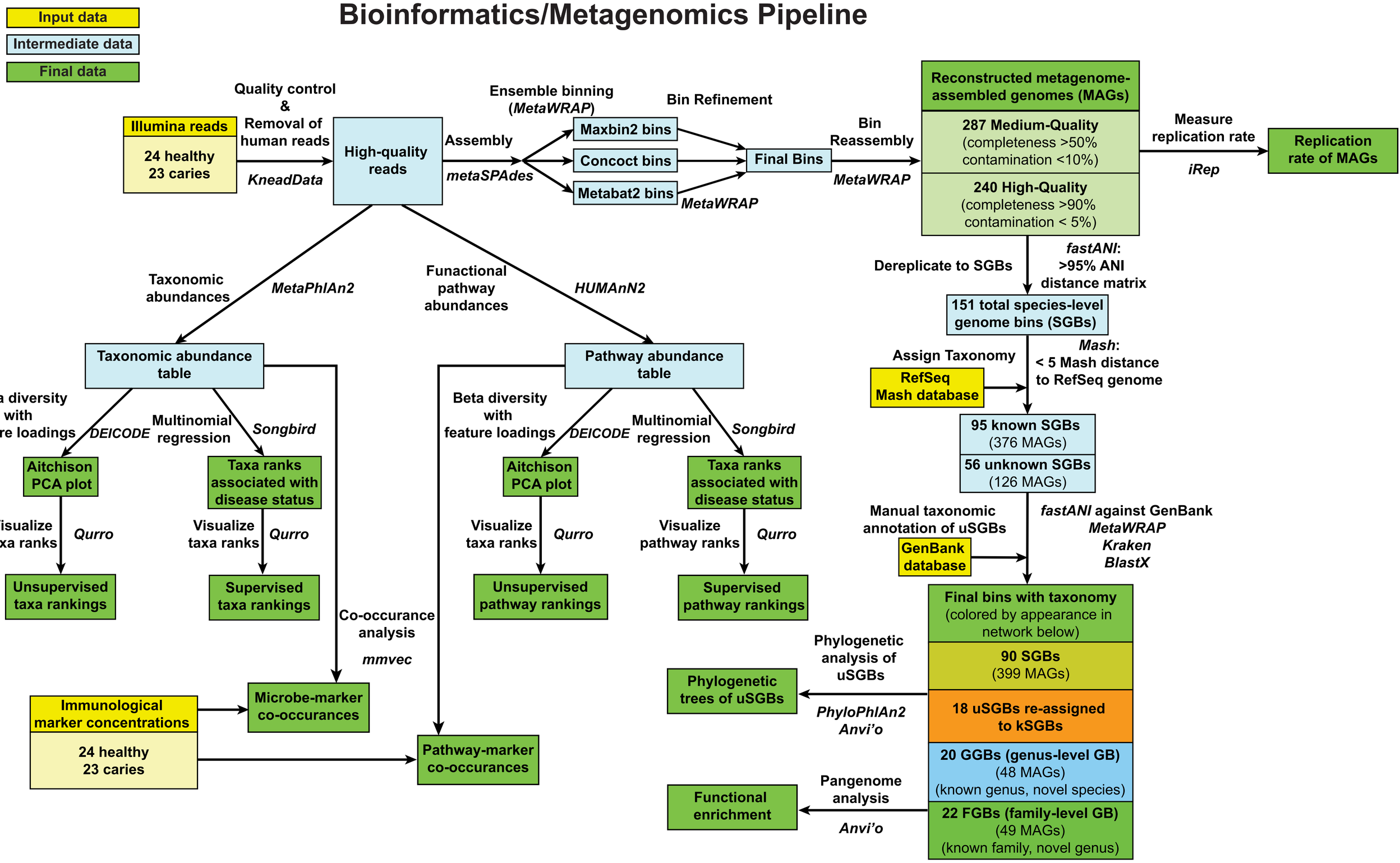
Figure S1

Summary of study design

A



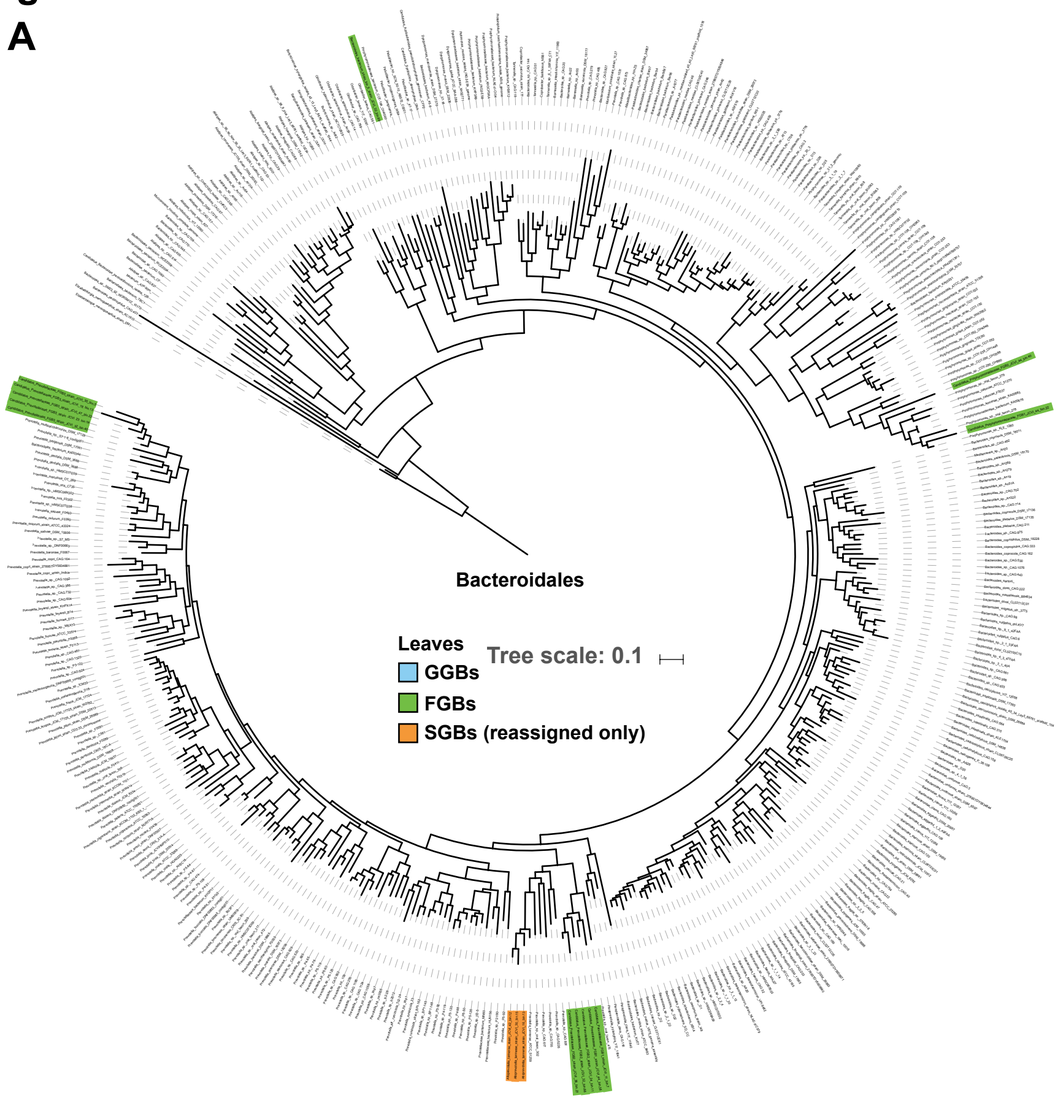
B



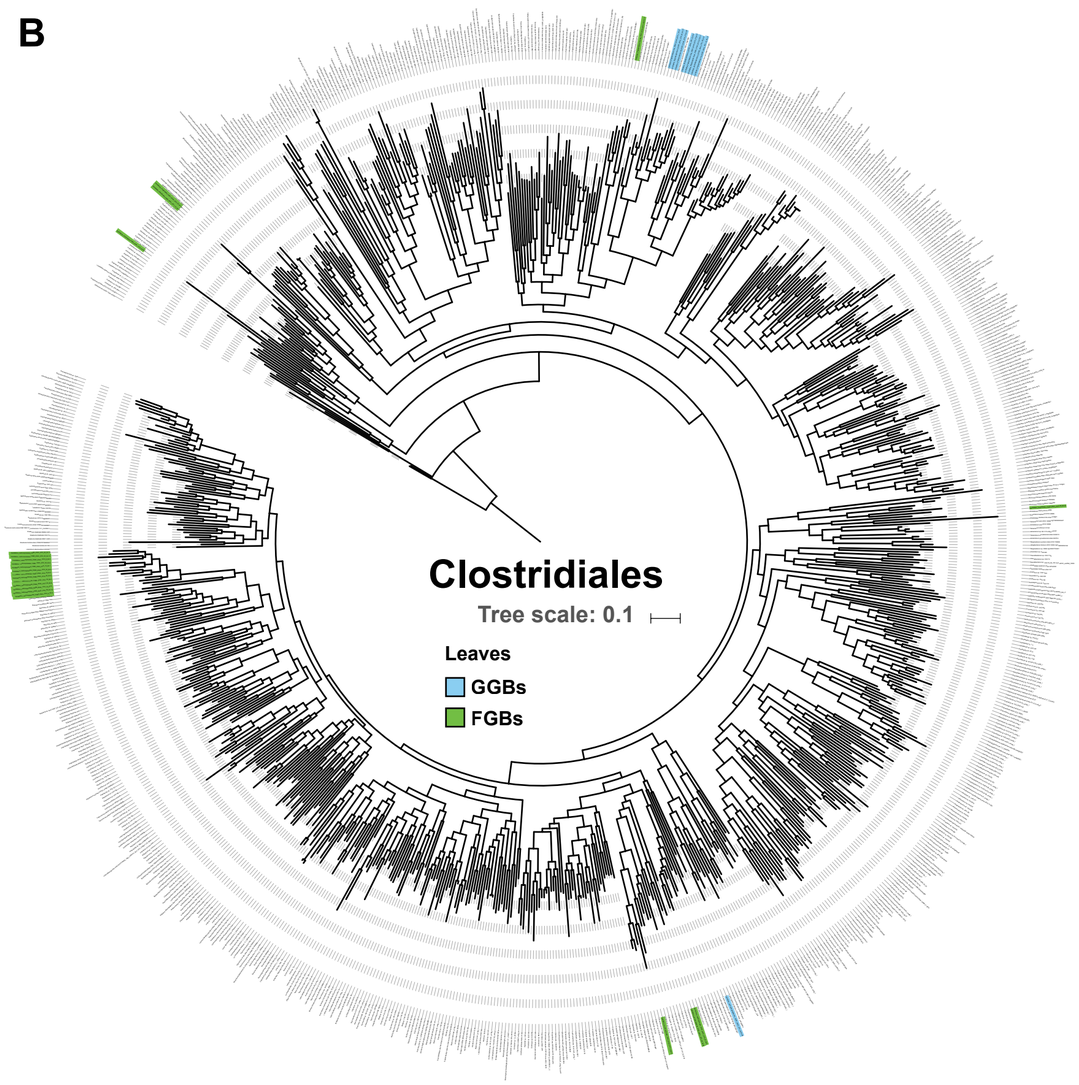
Supplemental Fig. S1: Overview of study design and bioinformatics methods. (A) Flow chart illustrating the steps taken to get from clinical specimen to bioinformatics data. **(B)** Flow chart illustrating the computational methodology utilized in this study. Input data is in yellow boxes, intermediate data is in blue boxes, and final data is in green boxes. For each step, the tool(s) or package(s) used are provided in italics. The 'Final bins with taxonomy' box is color-coded to match the metagenome-assembled genome (MAG) average nucleotide (ANI) network in Fig. 1.

Figure S2

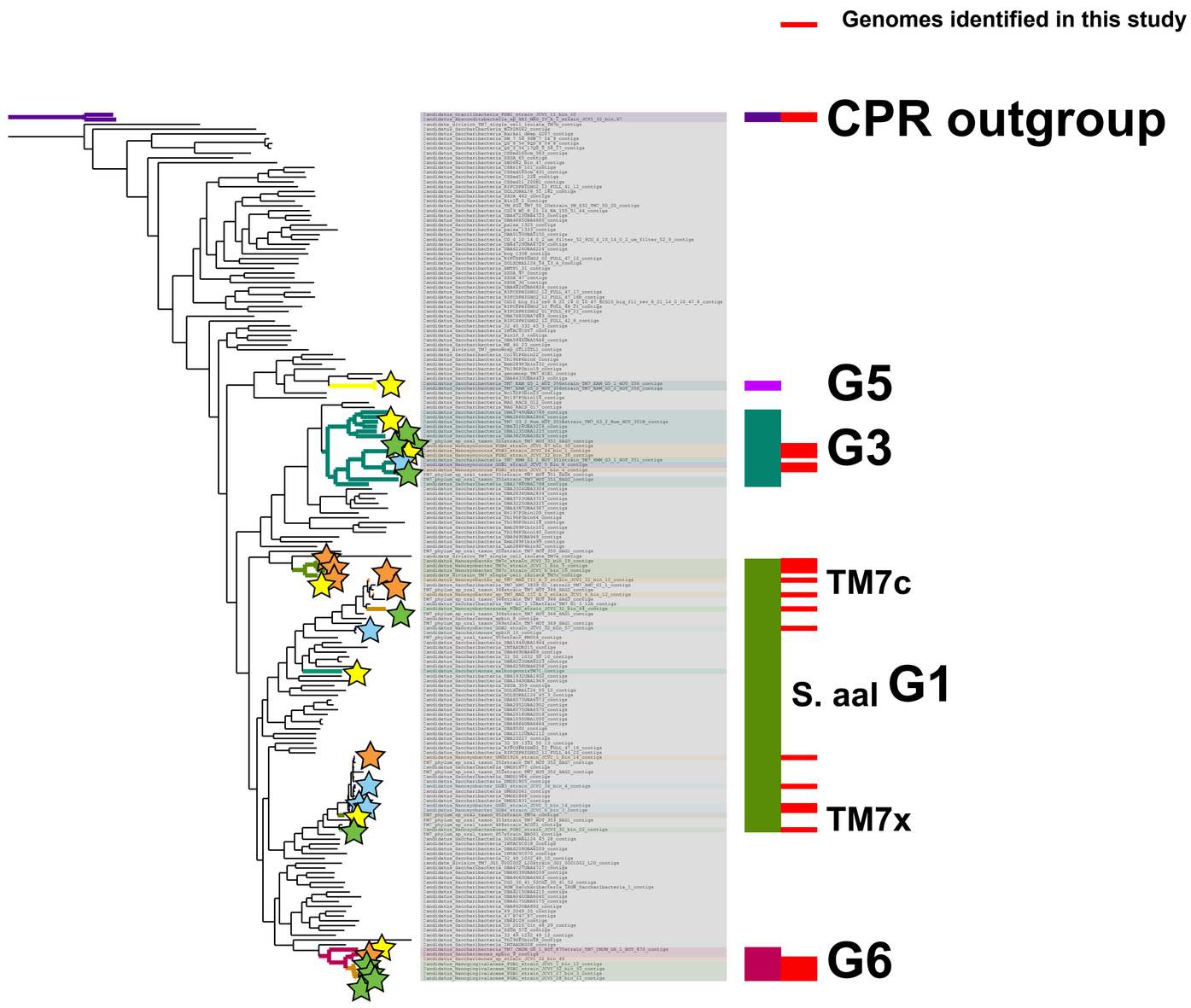
A



B



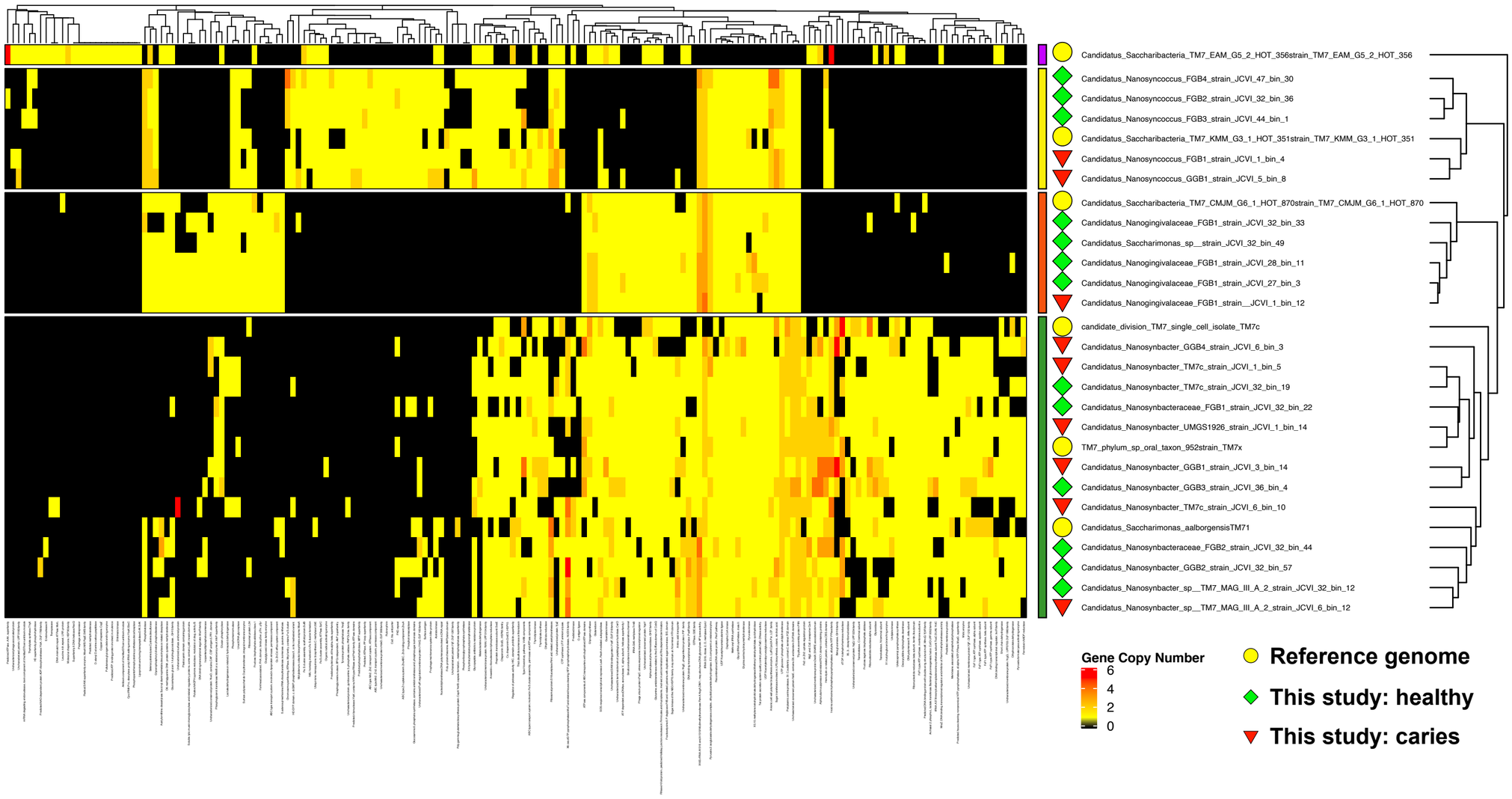
C



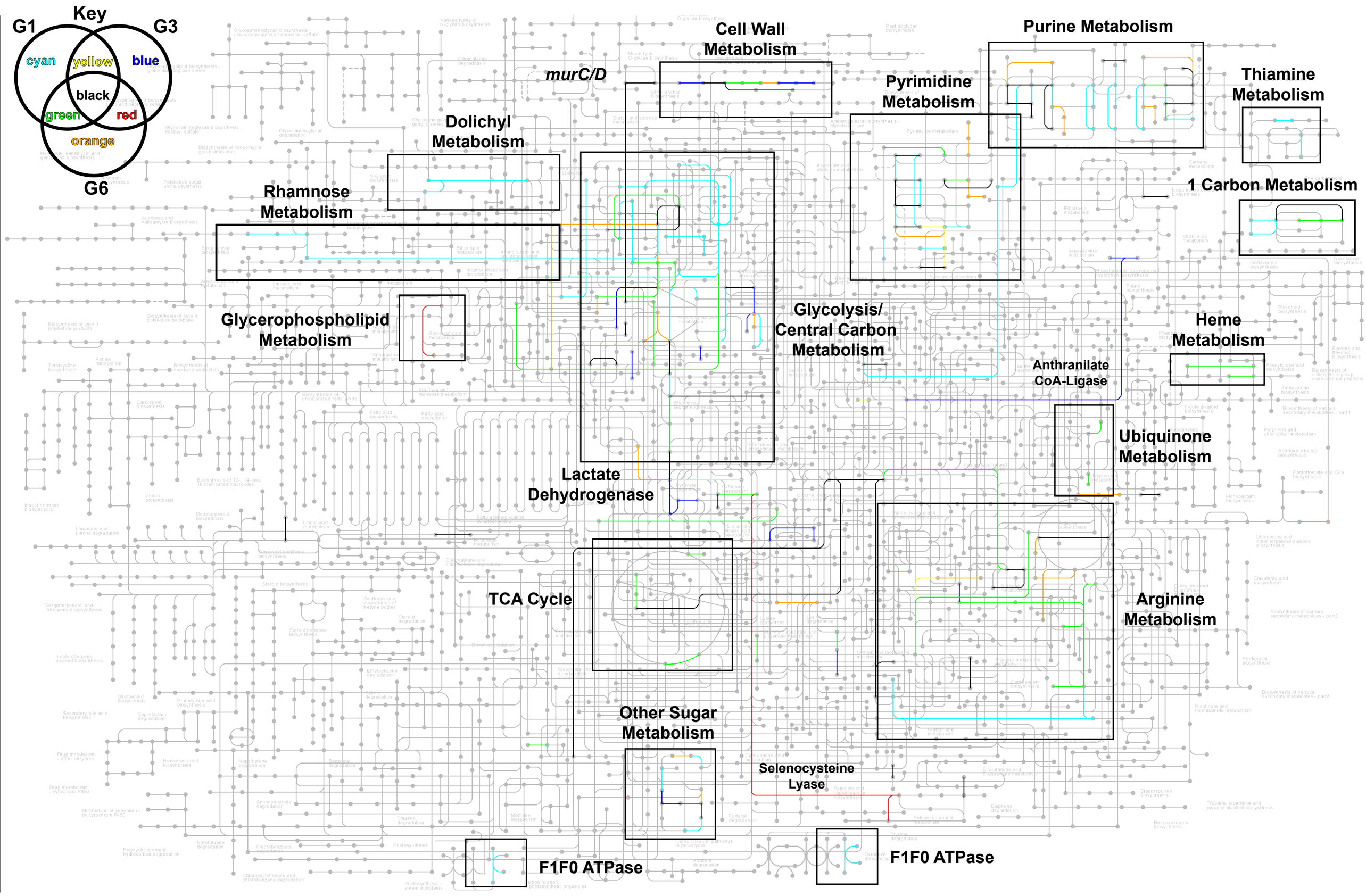
Supplemental Fig. S2: Phylogenetic placement of unknown species-level genome bins (uSGBs) within Bacteroidales, Clostridiales, and Saccharibacteria. Bacteroidales and Clostridiales trees were generated using PhyloPhlAn2 (Pasolli et al. 2019) and the Saccharibacteria tree was generated using Anvi'o. All trees were visualized using iToL (Letunic and Bork 2019). **(A) Bacteroidales (B) Clostridiales** and **(C) Saccharibacteria**

Figure S3

A

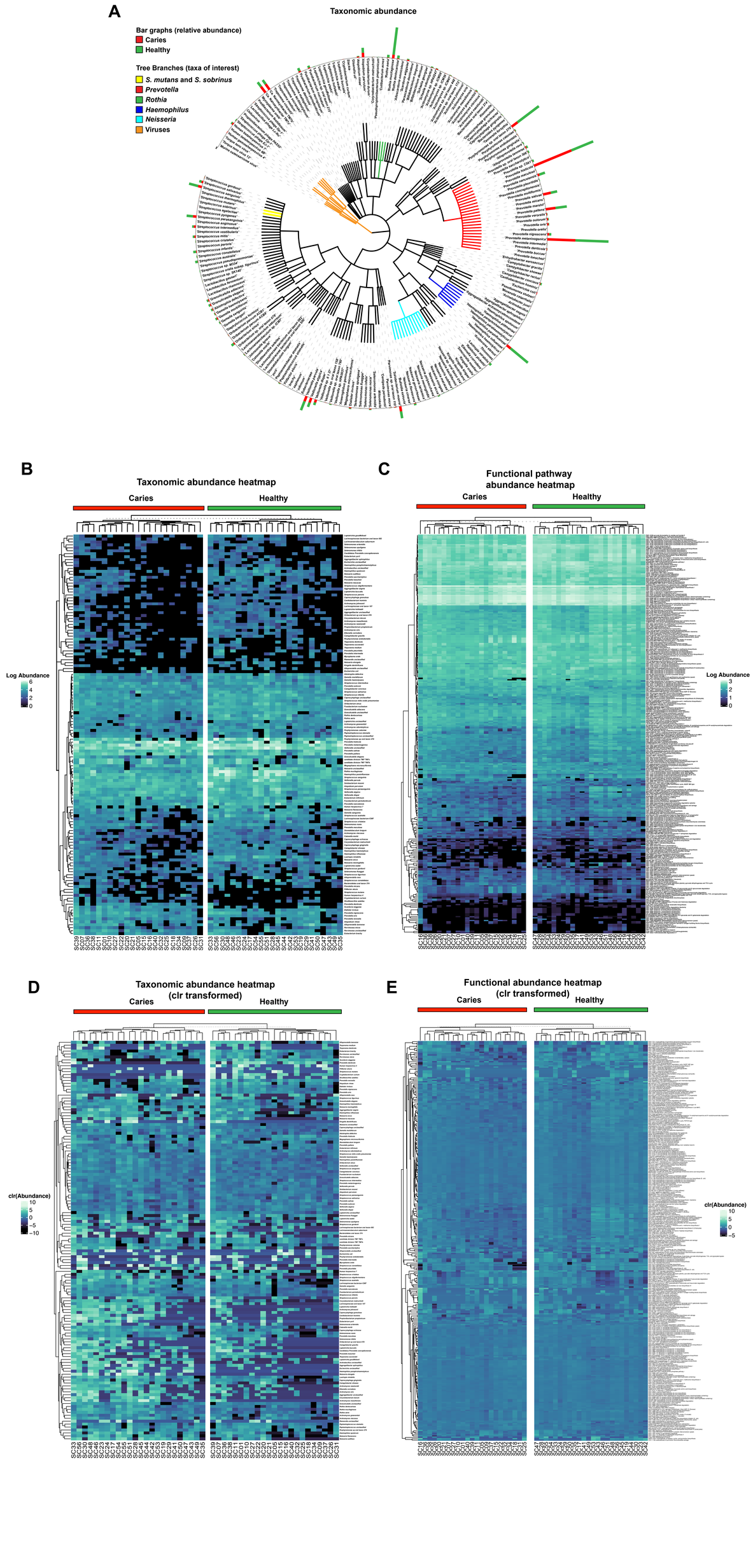

Candidatus_Saccharibacteria_TM7_EAM_G5_2_HOT_356strain_TM7_EAM_G5_2_HOT_356
Candidatus_Nanosyncoccus_FGB4_strain_JCVI_47_bin_30
Candidatus_Nanosyncoccus_FGB2_strain_JCVI_32_bin_36
Candidatus_Nanosyncoccus_FGB3_strain_JCVI_44_bin_1
Candidatus_Saccharibacteria_TM7_KMM_G3_1_HOT_351strain_TM7_KMM_G3_1_HOT_351
Candidatus_Nanosyncoccus_FGB1_strain_JCVI_1_bin_4
Candidatus_Nanosyncoccus_GGB1_strain_JCVI_5_bin_8
Candidatus_Saccharibacteria_TM7_CMJM_G6_1_HOT_870strain_TM7_CMJM_G6_1_HOT_870
Candidatus_Nanogingivalaceae_FGB1_strain_JCVI_32_bin_33
Candidatus_Saccharimonas_sp__strain_JCVI_32_bin_49
Candidatus_Nanogingivalaceae_FGB1_strain_JCVI_28_bin_11
Candidatus_Nanogingivalaceae_FGB1_strain_JCVI_27_bin_3
Candidatus_Nanogingivalaceae_FGB1_strain__JCVI_1_bin_12
candidate_division_TM7_single_cell_isolate_TM7c
Candidatus_Nanosynbacter_GGB4_strain_JCVI_6_bin_3
Candidatus_Nanosynbacter_TM7c_strain_JCVI_1_bin_5
Candidatus_Nanosynbacter_TM7c_strain_JCVI_32_bin_19
Candidatus_Nanosynbacteraceae_FGB1_strain_JCVI_32_bin_22
Candidatus_Nanosynbacter_UMGS1926_strain_JCVI_1_bin_14
TM7_phylum_sp_oral_taxon_952strain_TM7x
Candidatus_Nanosynbacter_GGB1_strain_JCVI_3_bin_14
Candidatus_Nanosynbacter_GGB3_strain_JCVI_36_bin_4
Candidatus_Nanosynbacter_TM7c_strain_JCVI_6_bin_10
Candidatus_Saccharimonas_aalborgensisTM71
Candidatus_Nanosynbacteraceae_FGB2_strain_JCVI_32_bin_44
Candidatus_Nanosynbacter_GGB2_strain_JCVI_32_bin_57
Candidatus_Nanosynbacter_sp__TM7_MAG_III_A_2_strain_JCVI_32_bin_12
Candidatus_Nanosynbacter_sp__TM7_MAG_III_A_2_strain_JCVI_6_bin_12
Gene Copy Number
6
4
2
0
Reference genome
This study: healthy
This study: caries

B


Key
G1
G3
G6
cyan
yellow
blue
black
green
red
orange
Cell Wall Metabolism
murC/D
Dolichyl Metabolism
Rhamnose Metabolism
Glycerophospholipid Metabolism
Glycolysis/ Central Carbon Metabolism
Lactate Dehydrogenase
TCA Cycle
Other Sugar Metabolism
Selenocysteine Lyase
F1F0 ATPase
F1F0 ATPase
Purine Metabolism
Pyrimidine Metabolism
Thiamine Metabolism
1 Carbon Metabolism
Heme Metabolism
Anthranilate CoA-Ligase
Ubiquinone Metabolism
Arginine Metabolism

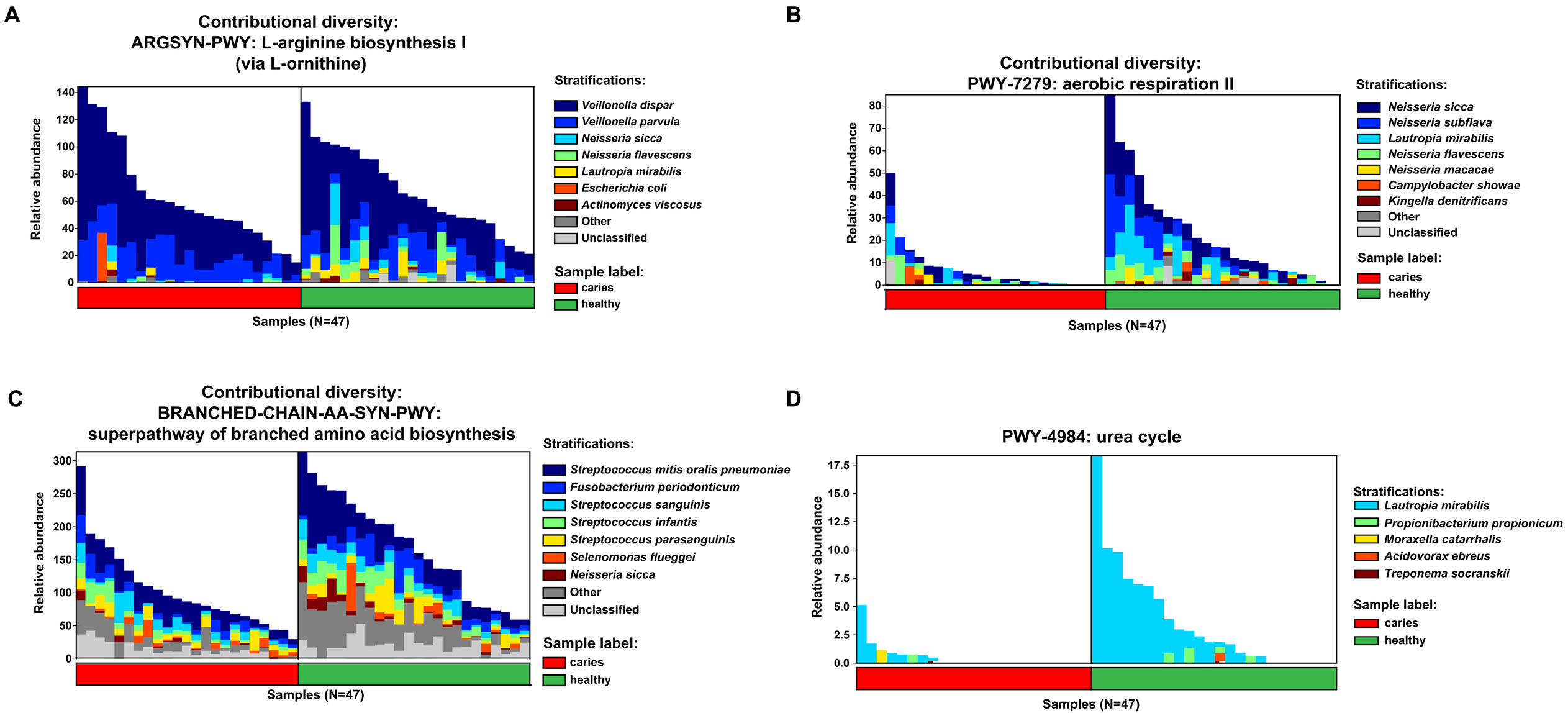
Supplemental Fig. S3: Fucntional pathways encoded by oral Saccharibacteria clades G1, G3, and G6. (A) Differences in encoded COG functions pathways across human-associated Saccharibacteria. Heatmap illustrating the presence of various genes across 4 major clades of human-associated Saccharibacteria. Rows and columns were clustered using the Jaccard distance and the "complete" clustering method. Only COG functions that were significantly different between clades are shown (adjusted q-value < 0.05). **(B) Metabolic circuit indicating the metabolic pathways encoded by Saccharibacteria clades G1, G3, and G6.** KEGG KO pathways present in the Saccharibacteria clades, using the TM7x, TM7_KMM_G3_HOT351, and TM7_G6_1_HOT870 genomes as representatives.

Figure S4


A
Taxonomic abundance
Bar graphs (relative abundance)
Caries
Healthy
Tree Branches (taxa of interest)
S. mutans and S. sobrinus
Prevotella
Rothia
Haemophilus
Neisseria
Viruses
B
Taxonomic abundance heatmap
Caries
Healthy
Log Abundance
6
4
2
0
C
Functional pathway abundance heatmap
Caries
Healthy
Log Abundance
3
2
1
0
D
Taxonomic abundance heatmap (clr transformed)
Caries
Healthy
clr(Abundance)
10
5
0
−5
−10
E
Functional abundance heatmap (clr transformed)
Caries
Healthy
clr(Abundance)
10
5
0
−5

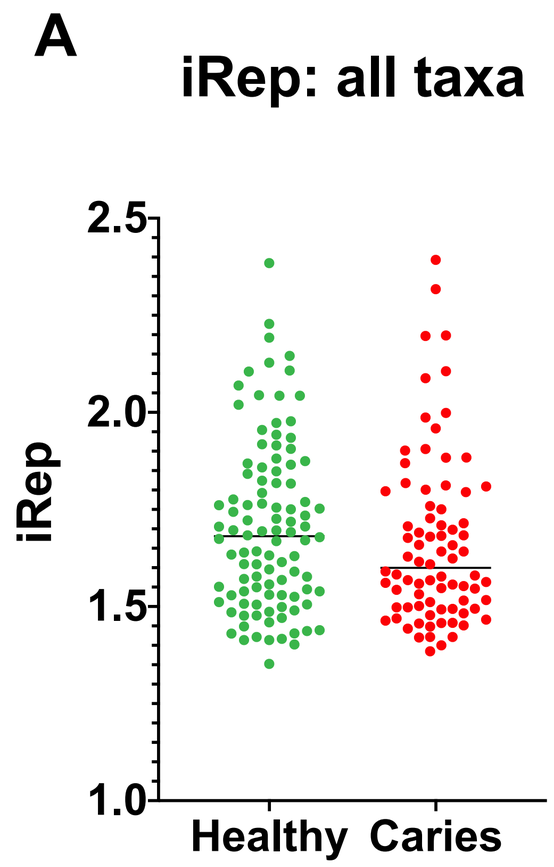
Supplemental Fig. S4: (A) Species abundance. Phylogenetic tree illustrating the species present across the saliva metagenomes. The relative abundance of each taxa is represented by the bar graph at the end of each leaf, with the relative abundance in the healthy group in blue and the caries group in red. Taxa of interest are highlighted with colored leaves on the tree: *Streptococcus mutans* and *Streptococcus sobrinus* = yellow; *Prevotella* spp.= red; and *Rothia* spp. = green. **(B) Heatmap illustrating species-level taxonomic abundances in the caries versus healthy groups.** Plot was generated using the ComplexHeatmaps R package (Gu et al. 2016). Abundances were determined using MetaPhAn2 (Truong et al. 2015). Rows and columns were clustered using a Bray-Curtis distance matrix and the "complete" clustering method. Only taxa appearing in at least 10 samples are included in this figure. **(C) Heatmap illustrating species-level functional pathway abundances in the caries versus healthy groups.** Plot was generated using the ComplexHeatmaps R package (Gu et al. 2016). Abundances were determined using HumanN2 (Franzosa et al. 2018). Rows and columns were clustered using a Bray-Curtis distance matrix and the "complete" clustering method. Only pathways included in at least 10 samples are included in this figure. **(D) Heatmap illustrating centered log ratio (clr)-transformed species-level taxonomic abundances in the caries versus healthy groups.** Same data as panel B except the data is clr-transformed (as is performed in DEICODE). **(E) Heatmap illustrating clr-transformed functional pathway abundances in the caries versus healthy groups.** Same data as panel C except the data is clr-transformed (as is performed in DEICODE).

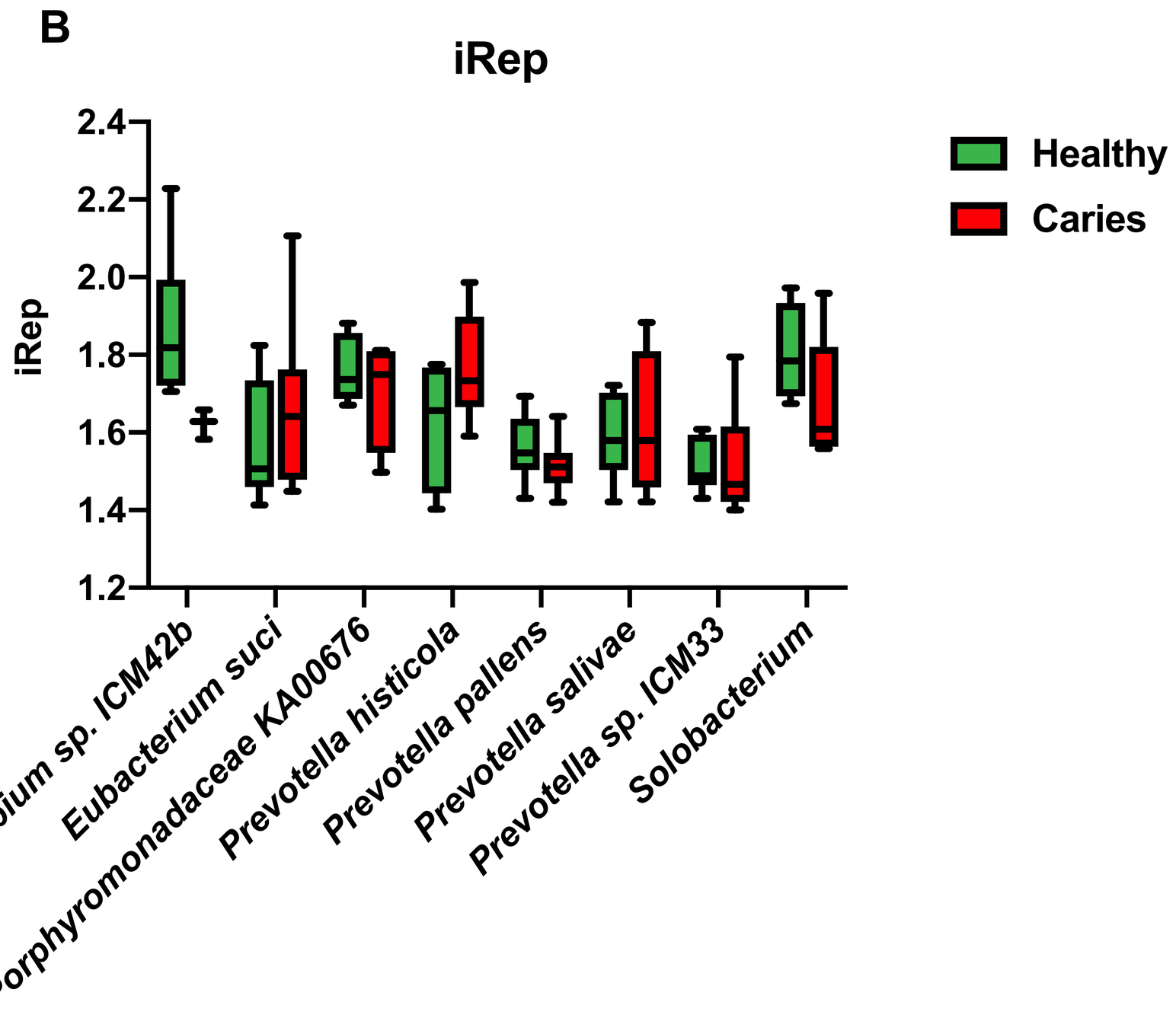
Figure S5


A
Contributional diversity:
ARGSYN-PWY: L-arginine biosynthesis I
(via L-ornithine)
Relative abundance
Samples (N=47)
Stratifications:
Veillonella dispar
Veillonella parvula
Neisseria sicca
Neisseria flavescens
Lautropia mirabilis
Escherichia coli
Actinomyces viscosus
Other
Unclassified
Sample label:
caries
healthy
B
Contributional diversity:
PWY-7279: aerobic respiration II
Relative abundance
Samples (N=47)
Stratifications:
Neisseria sicca
Neisseria subflava
Lautropia mirabilis
Neisseria flavescens
Neisseria macacae
Campylobacter showae
Kingella denitrificans
Other
Unclassified
Sample label:
caries
healthy
C
Contributional diversity:
BRANCHED-CHAIN-AA-SYN-PWY:
superpathway of branched amino acid biosynthesis
Relative abundance
Samples (N=47)
Stratifications:
Streptococcus mitis oralis pneumoniae
Fusobacterium periodonticum
Streptococcus sanguinis
Streptococcus infantis
Streptococcus parasanguinis
Selenomonas flueggei
Neisseria sicca
Other
Unclassified
Sample label:
caries
healthy
D
PWY-4984: urea cycle
Relative abundance
Samples (N=47)
Stratifications:
Lautropia mirabilis
Propionibacterium propionicum
Moraxella catarrhalis
Acidovorax ebreus
Treponema socranskii
Sample label:
caries
healthy

Supplemental Fig. S5: Contributional diversity of functional pathways of interest in caries versus health. Stacked bar chart illustrating the relative abundance and contributional diversity of arginine biosynthesis across the samples. **(A)** Arginine synthesis via L-ornithine **(B)** Aerobic respiration **(C)** Branched-chain amino acid biosynthesis **(D)** Urea cycle.

Figure S6





Permanova (Caries vs. Healthy): p = 0.011

Supplemental Fig. S6: iRep and genome-based taxonomic abundance. Replication rates of all taxa (A) and most abundant taxonomic groups (B). Replication rates were determined using iRep (Brown et al. 2016). Replication rates were not significantly different between healthy and caries groups for any taxonomic groups shown here for overall (A). **(C) Beta diversity of taxa using quantification of assembled genome bins.** 3D PCA plot generated using DEICODE (robust Aitchison PCA) (Martino et al. 2019). Data points represent individual subjects and are colored with a gradient to visualize DMFT score, indicating severity of dental caries. Feature loadings (i.e. taxa driving differences in ordination space) are illustrated by the vectors, which are labeled with the cognate species name.

SUPPLEMENTAL REFERENCES

Aleti G, Baker JL, Tang X, Alvarez R, Dinis M, Tran NC, Melnik AV, Zhong C, Ernst M, Dorrestein PC et al. 2019. Identification of the Bacterial Biosynthetic Gene Clusters of the Oral Microbiome Illuminates the Unexplored Social Language of Bacteria during Health and Disease. *MBio* **10**.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144-1146.

Anusavice KJ. 2005. Present and future approaches for the control of caries. *J Dent Educ* **69**: 538-554.

Bolyen E Rideout JR Dillon MR Bokulich NA Abnet CC Al-Ghalith GA Alexander H Alm EJ Arumugam M Asnicar F et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852-857.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**: 725-731.

Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* **34**: 1256-1263.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59-60.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* **6**: e4320.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**: e1319.

Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. 2018. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. *mSystems* **3**.

Espinoza JL, Harkins DM, Torralba M, Gomez A, Highlander SK, Jones MB, Leong P, Saffery R, Bockmann M, Kuelbs C et al. 2018. Supragingival Plaque Microbiome Ecology and Functional Potential in the Context of Health and Disease. *MBio* **9**.

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N et al. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**: 962-968.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**: 2847-2849.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**: 2115-2122.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**: D309-D314.

Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**: 5114.

Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**: e7359.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.

Koster J, Rahmann S. 2012. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**: 2520-2522.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**: W256-W259.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**: 2843-2851.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Loe H. 1967. The Gingival Index, the Plaque Index and the Retention Index Systems. *J Periodontol* **38**: Suppl:610-616.

Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, Zengler K. 2019. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **4**.

McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, Hendrickson EL, Wrighton K, Shi W, He X. 2020. Acquisition and Adaptation of Ultra-small Parasitic Reduced Genome Bacteria to Mammalian Hosts. *Cell Rep* **32**: 107939.

Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y et al. 2019a. Learning representations of microbe-metabolite interactions. *Nat Methods* **16**: 1306-1314.

Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K, Knight R. 2019b. Establishing microbial composition measurement standards with reference frames. *Nat Commun* **10**: 2719.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933-2935.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268-274.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**: 824-834.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**: 132.

Organization WH. 1971. *Oral health surveys: basic methods*. World Health Organization.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043-1055.

Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P et al. 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**: 649-662 e620.

Pereira D, Seneviratne CJ, Koga-Ito CY, Samaranayake LP. 2018. Is the oral fungal pathogen Candida albicans a cariogen? *Oral Dis* **24**: 518-526.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641-1650.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498-2504.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-1313.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**: 902-903.

Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**: 158.

Vazquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* **2**: 16.

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**: R46.

Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**: 605-607.