PNAS nexus

PNAS Nexus, 2022, **1**, 1–16
https://doi.org/10.1093/pnasnexus/pgac239
Advance access publication date: 18 October 2022
Research Report

Differential network analysis of oral microbiome metatranscriptomes identifies community scale metabolic restructuring in dental caries

Josh L. Espinoza[a,b,c], Manolito Torralba[c], Pamela Leong[d], Richard Saffery[d], Michelle Bockmann[e], Claire Kuelbs[b], Suren Singh[c], Toby Hughes[e], Jeffrey M. Craig[d,f], Karen E. Nelson[b,c] and Chris L. Dupont[a,b,*]

[a]Department of Environment and Sustainability, J. Craig Venter Institute, La Jolla, CA 92037, USA
[b]Department of Human Biology and Genomic Medicine, J. Craig Venter Institute, La Jolla, CA 92037, USA
[c]Department of Human Biology and Genomic Medicine, J. Craig Venter Institute, Rockville, MD 20850, USA
[d]Epigenetics, Murdoch Children's Research Institute and Department of Paediatrics, The University of Melbourne, Parkville, VIC 3052, Australia
[e]Adelaide Dental School, The University of Adelaide, Adelaide, SA 5005, Australia
[f]IMPACT Strategic Research Centre, Deakin University School of Medicine, Geelong, VIC 3220, Australia
*To whom correspondence should be addressed: Email: cdupont@jcvi.org
Edited By: David Brenner

Abstract

Dental caries is a microbial disease and the most common chronic health condition, affecting nearly 3.5 billion people worldwide. In this study, we used a multiomics approach to characterize the supragingival plaque microbiome of 91 Australian children, generating 658 bacterial and 189 viral metagenome-assembled genomes with transcriptional profiling and gene-expression network analysis. We developed a reproducible pipeline for clustering sample-specific genomes to integrate metagenomics and metatranscriptomics analyses regardless of biosample overlap. We introduce novel feature engineering and compositionally-aware ensemble network frameworks while demonstrating their utility for investigating regime shifts associated with caries dysbiosis. These methods can be applied when differential abundance modeling does not capture statistical enrichments or the results from such analysis are not adequate for providing deeper insight into disease. We identified which organisms and metabolic pathways were central in a coexpression network as well as how these networks were rewired between caries and caries-free phenotypes. Our findings provide evidence of a core bacterial microbiome that was transcriptionally active in the supragingival plaque of all participants regardless of phenotype, but also show highly diagnostic changes in the ways that organisms interact. Specifically, many organisms exhibit high connectedness with central carbon metabolism to *Cardiobacterium* and this shift serves a bridge between phenotypes. Our evidence supports the hypothesis that caries is a multifactorial ecological disease.

Keywords: dental caries, feature engineering, machine-learning, machine-learning, oral microbiome

Significance Statement:

Using metagenomics and metatranscriptomics, this study characterized 658 bacterial and 189 viral metagenome-assembled genomes (MAGs) from the supragingival plaque oral microbiome to investigate dental caries in 91 children. We developed methodologies for species-level clustering to characterize biologically accurate MAGs across nonoverlapping samples. We also developed novel feature engineering and network analysis techniques, which can be used to gain deeper insight into microbial diseases than differential abundance methods alone. With these new techniques, we identified regime shifts between caries and caries-free microbiomes where certain taxa switch their interactions with other organisms and metabolic pathways. Our study provides evidence for the hypothesis that caries is a multifactorial ecological disease and contains generalizable methods for microbiome research.

Introduction

Oral diseases such as dental caries are a critical concern for public health. Untreated tooth decay is the most common chronic health condition and affects nearly 3.5 billion people worldwide (1). Tooth decay is most common in younger individuals with a prevalence of 20% of children aged 5 to 11 and 13% in adolescents aged 12 to 19, with low-income children being twice as likely to have cavities (2). In most low- and middle-income countries, the prevalence of oral diseases increases with urbanization and inadequate access to medical treatment (3). In high-income countries, dental treatment averages 5% of total health expenditure and 20% of out-of-pocket health expenditure (4) making the disease a socioeconomic issue for all. This silent epidemic has long been coupled with the rise of civilization as early evidence from the Pleistocene era sug-

Competing Interest: The authors declare no competing interest.
Received: January 21, 2022. **Accepted:** October 14, 2022

gests that agriculture and the exploitation of starchy plant foods have burdened mankind with carious lesions since early prehistory (5).

In the modern "ecological plaque hypothesis," oral diseases arise from environmental perturbations leading to a shift in the endogenous microbial community (6) where the selection of pathogenic bacteria is coupled with the environment and any species with germane traits can contribute to pathogenesis (7). The evidence for this theory is largely from the advent of next generation sequencing (NGS) technologies and the ability to sequence uncultivated organisms. In the context of carious lesions, the environmental perturbation arises from persistent consumption of dietary sugars leading to a decrease in pH and, when sustained, shifts the population to a more aciduric and cariogenic microbial community, which degrades the enamel (8, 9). Changing environments (e.g. a substantial increase in acidity) can destabilize previously stable microbial communities reconfiguring them into new stability domains, referred to as regime shifts (10), such as a cariogenic microbiome (11). Furthermore, this regime shift of the oral microbiome towards a cariogenic state is the result of an environment partially created by the bacteria themselves creating a complex feed-forward loop. This complex feed-forward loop makes it difficult to diagnose the exact cause of each case and the development of therapeutics for severe cases.

Understanding the roles of microbes in the context of caries-related dysbiosis (from the perspective of human health and not microbial stability) is nontrivial and is often explored using association networks. Association networks such as coexpression (transcriptomics) or coabundance (genomics) are powerful frameworks for investigating inferred biological interactions by grouping biological features, such as genes or microbes, with related metabolism (12) or complementary ecological niches (13). Despite widescale usage, many approaches do not address NGS compositionality and this is a major concern because noncompositionally-aware metrics (e.g. correlation) are known to yield spurious associations with no biological meaning (14). Although packages such as WGCNA (15) introduced intuitive and clever ways for analyzing fully connected weighted gene association networks, they do not directly support compositionally-aware association metrics such as proportionality (16, 14, 17); thus, the findings using such methods are based on a statistical fallacy. However, awareness of compositional data analysis (CoDA) has increasingly made its way from geology to bioinformatics (16, 14, 18, 19) with many advancements in the context of network analysis (20).

Association networks are often applied to individual organisms in controlled settings (21) and extending these concepts to ecosystems introduces many challenges that arise from the complexity of the data where the exact abundances of biological features are often unknown a priori and the number of features increase by several orders of magnitude. Furthermore, as systems biology deals with interactions amongst biological features, the number of pairwise interactions scale quadratically. The vast number of variables in a microbiome not only makes hypothesis testing precarious but can also lead to statistical artifacts in downstream analysis due to the "curse-of-dimensionality" making interpretation exceedingly difficult (22). Many dimensionality-reduction methods such as PCA, [N]MDS, t-SNE (23), or UMAP (24) lose accessibility to original biological features rendering interpretation limited and unintuitive. The genome-resolved hierarchical complexity of microbiomes results in dynamic distributions of expression or abundance influenced by other microbes and latent environmental variables not accounted for by the experimental design. These community-level datasets require representations of the data that account for these abstractions and group genes within their genome-resolved structure; that is, explainable biological feature engineering.

The balance between biological accuracy and analytical practicality is a constant theme when pairing metagenomic assemblies with metatranscriptomic sequencing. On one hand, consensus assemblies and binning of metagenome-assembled genomes (MAGs) make it feasible to cross-reference genomic features across samples while providing a relatively dense counts table. Though not inherently sparse, sparsity in NGS technologies is common and a major hurdle in CoDA (25, 26). This approach often produces user-friendly data structures, but the resulting MAGs will likely be composites of multiple *bona fide* microbial genomes of highly similar strains resulting in redundant and contaminated MAGs that lack true biological interpretation. On the other hand, assembling samples individually and binning genomes will produce more biologically accurate MAGs, but mapping to these samples produces inherently sparse matrices that have an extremely large number of features, which is problematic for statistical analysis (22), mostly filled with zeros because the *one-to-one* mapping of reads to highly redundant genomic features. When using paired metagenomics and metatranscriptomics to investigate dysbiotic systems, it is imperative to address these pitfalls by leveraging compositionally-aware network methodologies simultaneously with natural hierarchies inherent in the data.

Results

MAGs for bacteria and viruses

Metagenomes from 88 Australian children in this study were evaluated and analyzed previously (27, 28) but substantial improvements in assembly, binning, and quality assessment methodologies warranted revisitation and reanalysis. Further, this reanalysis from a multiomics perspective provides a unique opportunity to couple phylogeny with metabolic function in the context of health and disease; the metatranscriptomic patient cohort is described in Table 1. Our reanalysis of combined metagenomics and metatranscriptomics, we isolated 658 bacterial MAGs, 179 DNA viruses, and 10 RNA viruses after quality assessment (Fig. 1, Table S2). For clarity, DNA and RNA viruses refer to viruses derived from metagenomics or metatranscriptomic sequencing, respectively. These bacterial MAGs clustered (95% ANI where ANI refers to Average Nucleotide Identity) into 135 unique species-level clusters (SLC) representing 49 hitherto unclassified species with 26 of which classified as *Patescibacteria* candidate phyla radiation (CPR; 6 *Gracilibacteria/SR1*, 43 *Saccharibacteria*) and a total of 69 CPR MAGs collectively. Of the non-CPR SLCs, we identified 31 *Bacteroidota*, 22 *Proteobacteria*, 21 *Actinobacteriota*, 23 *Firmicutes*, 8 *Fusobacteriota*, and 4 *Campylobacterota* (Tables 2, and Tables S2 and S3). The DNA and RNA viruses clustered into 137 and 5 unique SLCs, respectively. Most of the DNA viruses were classified as *Caudovirales*, of unknown species, associated with the human oral (42 SLCs), gut (41 SLCs), human respiratory (1 SLC), and nonspecific environments (1 SLC). Aside from these unknown species, we also identified several *Caudovirales* phages for *Arthrobacter* (7 SLCs), *Streptococcus* (4 SLCs), *Klebsiella*, *Haemophilus*, *Pasteurella*, *Pseudomonas*, and *Burkholderia*. Other than *Caudovirales*, we identified *Streptococcus* satellite phages (2 SLCs), unclassified CRESS-DNA *Parvovirus* associated with the human gut (2 SLCs), and an unclassified virus associated with the human oral environment. Most of the RNA viruses were *Escherichia* phages (4 SLCs) designated as Qbeta BZ1, MS2, and BZ13 strains but we also uncovered a novel virus with no close taxonomic

Table 1. Metatranscriptomics cohort sample metadata.

Metadata	Caries	Caries-free
Age (μ, σ^2, min, max)	(8.09, 2.7, 5.5, 10.9)	(7.82, 2.21, 5.4, 10.8)
Sex (female)	19	31
Sex (male)	17	20
Center (CBRG)	21	28
Center (MCRI)	15	23
Total samples	36	51

Overview of sample size for cohort with respect to phenotype and several metadata.

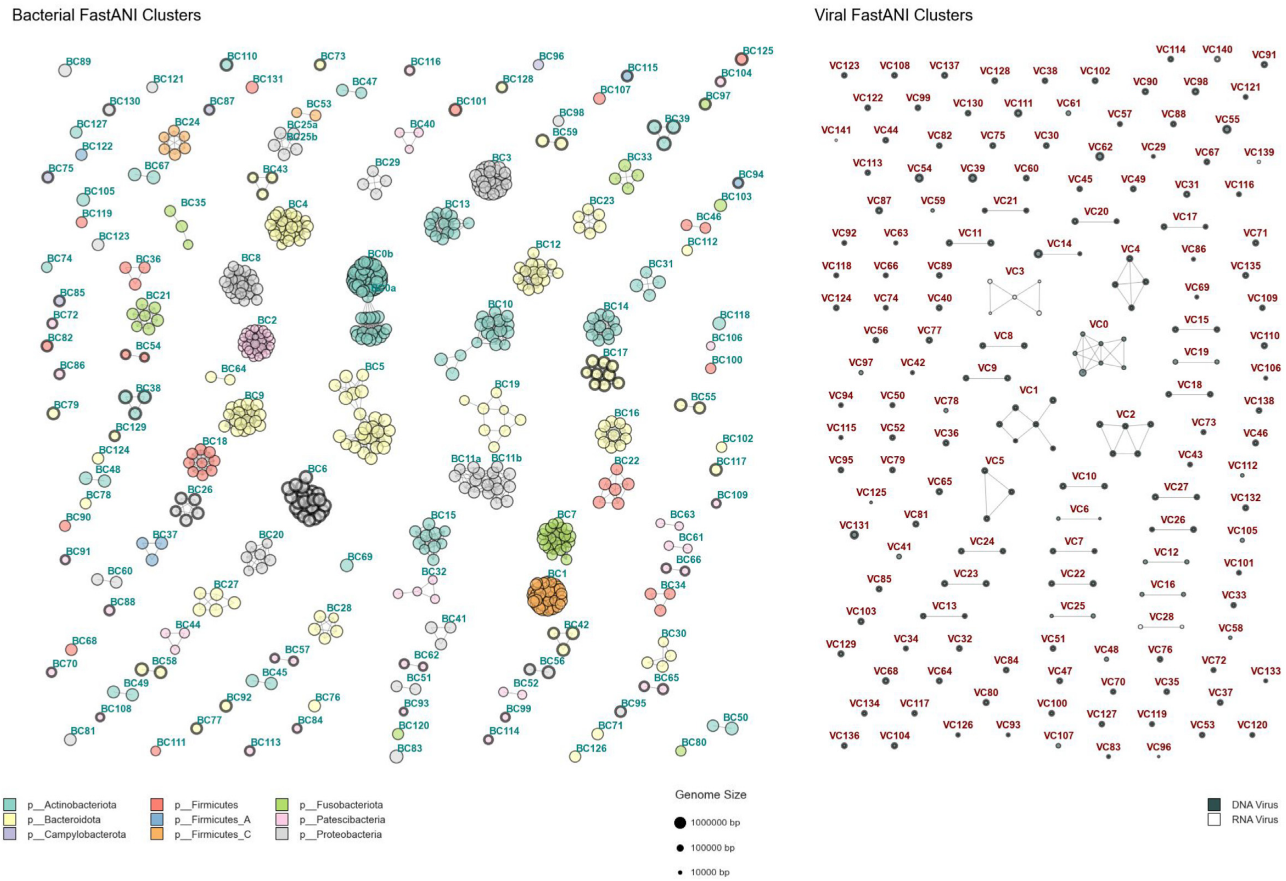

Fig. 1. Network of *FastANI* clusters for bacterial and viral MAGs. Network with edge weights corresponding to ANI and nodes representing MAGs. (Left) Bacterial MAGs colored by phylum. (Right) Viral MAGs colored by either DNA or RNA virus type. Thicker edge weights indicate novel species not found in *GTDB-Tk* or *CheckV* for bacterial and viral *FastANI* clusters, respectively.

Table 2. Bacterial SLCs and MAGs with respect to phylum.

Bacterial phyla	SLCs	MAGs	Novel species (SLCs)
p__Actinobacteriota	21	153	3
p__Bacteroidota	31	160	13
p__Campylobacterota	4	4	3
p__Firmicutes	16	35	4
p__Firmicutes_A	4	6	2
p__Firmicutes_C	3	62	0
p__Fusobacteriota	8	40	1
p__Patescibacteria	26	69	18
p__Proteobacteria	22	129	5
Total	135	658	49

The number of bacterial MAGs, SLCs, and novel species with respect to phyla.

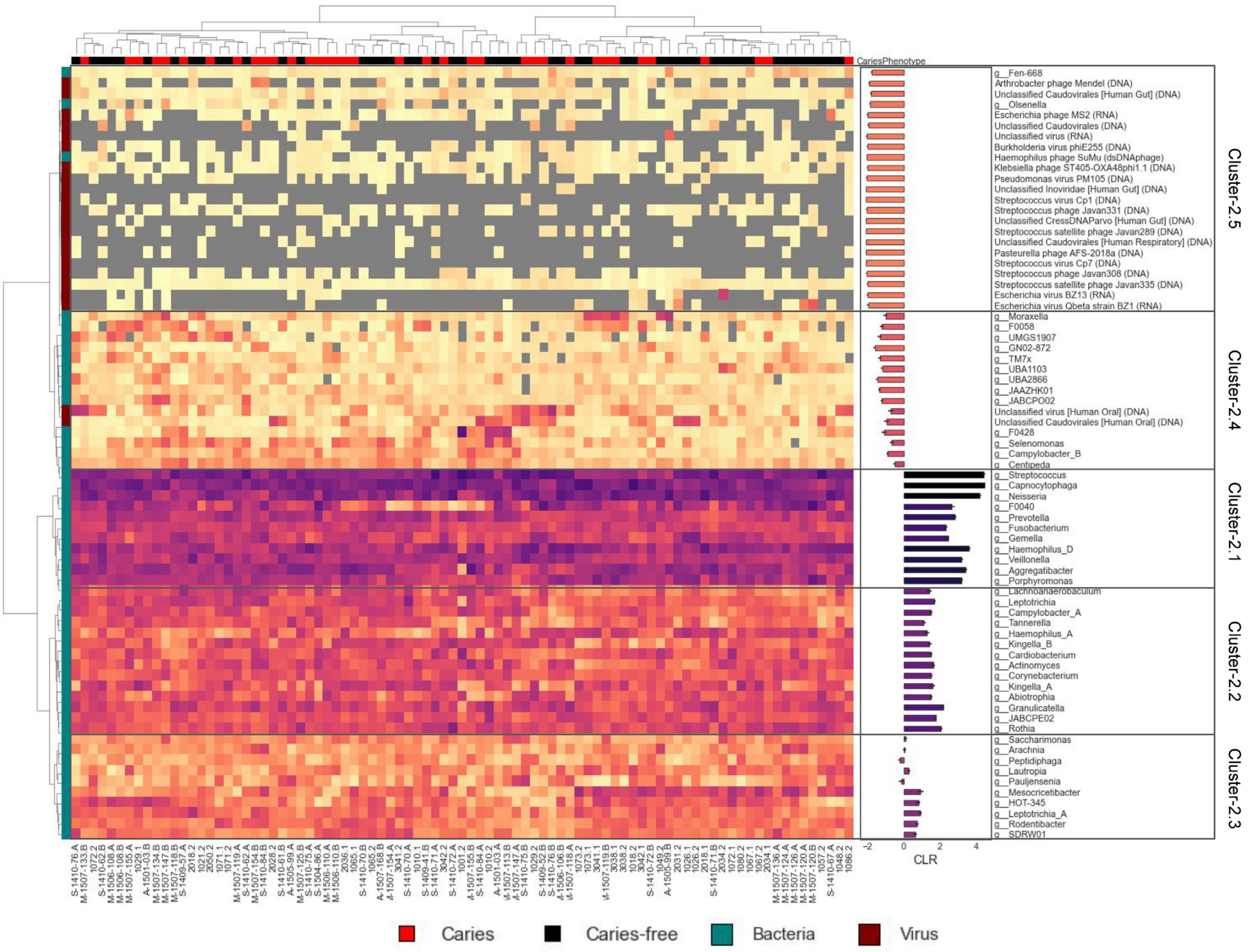

Fig. 2. Taxonomic expression. Center log-ratio (CLR) transformed abundances of taxa from metatranscriptomics. Row colors represent bacterial (teal) or viral (maroon) MAGs while the column colors represent caries (red) or caries-free (black) phenotypes. Clustering was performed using Euclidean distance and Ward linkage.

classification. Only high-confidence viruses based on strict *CheckV* thresholds were considered for analysis to reduce false positives and increase interpretability. The bacterial and viral SLCs contained 64 and 113 singleton clusters; individual genomes that did not share 95% ANI with any other organisms in the dataset. No archaea, eukaryote, or novel bacterial genera beyond the CPR were detected.

Our metapantranscriptomics approach collapsed 1,248,783 ORFs into 255,737 SLC-specific orthogroups (247,943 bacterial and 7,794 viral) reducing the dimensionality by 80% with minimal loss in information content. By using SLC-specific orthogroups, we were able to maintain a "bag-of-genomes" paradigm, opposed to that of a "bag-of-genes" and preserving natural hierarchical structures inherent in ecology.

We observed only one taxonomic database discrepancy, *M-1507–144.A__MAXBIN2__bin.008* was classified by *GTDB-Tk* as a novel *Tannerella* but this MAG clustered in BC14 with 10 *Peptidiphaga sp000466175* with high confidence (>95% ANI via *FastANI*), which suggests an update to the GCF_003033925.1 reference taxonomy in NCBI. This result is further strengthened by the *CheckM* basal classification of the *Actinobacteria* phylum.

Relative taxonomic expression and abundance

Clustering using genome-resolved gene expression grouped subjects into five distinct clusters (Fig. 2) but these expression patterns were not able to discriminate samples based on the presence or absence of caries. We also measured the silhouette scores using Aitchison distance against caries status and observed average scores close to zero ($|x_{\text{silhouette}}| < 0.003$) for bacterial and viral microbiomes in both metatranscriptomics and metagenomics datasets indicating minimal phenotype partitioning capacity using individual features.

As these expression patterns are in CLR, values close to 0 can be considered basal community-level expression and close to the geometric mean of expression values for the microbiome. We observed a core bacterial supragingival plaque microbiome at the genus level as almost every genus is transcriptionally active in every sample (Clusters-2.1 to 2.4, see the "Methods" section for naming scheme), regardless of phenotype, but this is not the case for either DNA or RNA viruses. Most of the viruses were detected with low expression and grouped in Cluster-2.5. In Cluster-2.5, there are a few DNA viruses that are detected in almost every sample including *Streptococcus satellite phage Jaavan335*, unclassified *Caudovirales* associated with human gut, *Burkholderia phage phiE255*, *Haemophilus phage SuMu*, and *Klebsiella phage ST405-OXA48phi1*

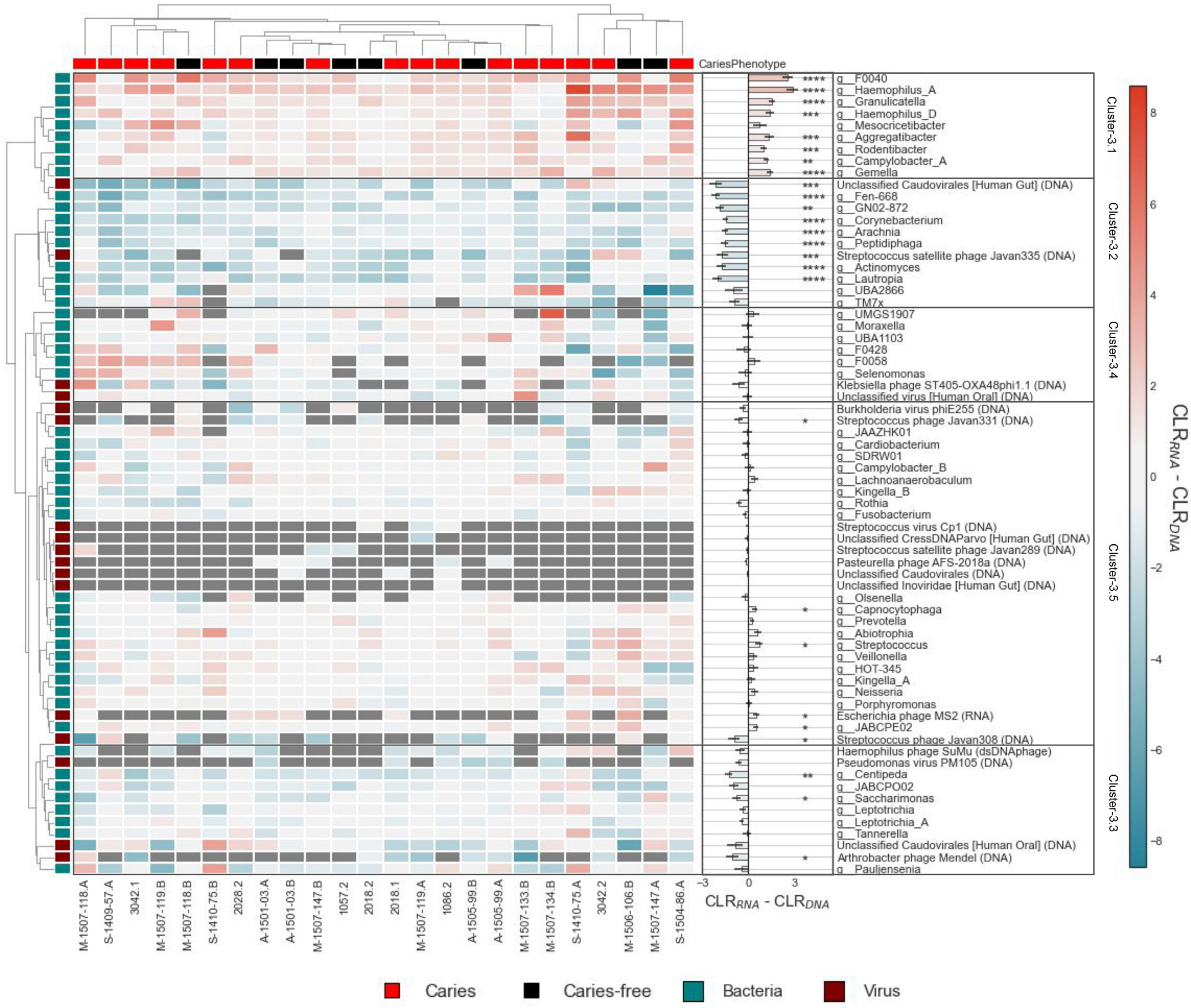

Fig. 3. Taxonomic expression:abundance ratios. Difference in CLR between metatranscriptomics expression (RNA) and metagenomics abundance (DNA) where values indicate higher expression relative to abundance and vice versa.

with *Escherichia phage MS2* as the only high prevalence RNA virus. Cluster-2.4 contains an unclassified human oral DNA virus and an unclassified oral *Caudovirales* that are transcriptionally active in every sample at modest levels on par with many bacteria in the microbiome. Cluster-2.1, the cluster with highest overall transcript abundance, we observe mostly genera from *Bacteroidota*, *Proteobacteria*, *Firmicutes*(C), and *Fusobacteriota*. The most transcriptionally active genera in Cluster-2.1 are *Capnocytophaga*, *Streptococcus*, *Neisseria*, *Haemophilus_D*, *Aggregibacter*, *Porphyromonas*, *and Veillonella* with modest expression from other bacteria. Cluster-2.2, with CLRs of 1 and 2, includes genera that appear in downstream analysis including *Cardiobacterium*, *Corynebacterium*, and *Tannerella*. Cluster-2.3 is the cluster with baseline transcriptional activity (i.e. CLR close to 0), which contains all the *Saccharibacteria* and *SR1* CPR clade.

We also investigated the RNA:DNA ratios from the 26 overlapping metatranscriptomic and metagenomic samples (Fig. 3). Based on clustering of RNA:DNA ratios, five distinct groupings were observed (Clusters-3.1 to 5). The most prominent findings are from Clusters-3.1 and 2 where taxa have highest and lowest RNA:DNA ratios, respectively. Cluster-3.1, the most transcriptionally active, included *Haemophilus_A* and *Alloprevotella*, the most active genera, as well as *Aggregibacter*, *Gemella*, and *Campylobacter_A*. Cluster-3.2 has a more uniform distribution of low RNA:DNA ratios and contained *Gracilibacteria*, *Saccharibacteria*, *Streptococcus satellite phage Javan335*, and an unclassified *Caudovirales* phage associated with the human gut. Clustering of these RNA:DNA ratios also did not differentiate subjects based on phenotype.

In terms of alpha diversity, we did not observe any difference in bacterial richness for metatranscriptomics (x = 131 SLCs) or metagenomics (x = 134 SLCs) datasets between caries and caries-free microbiomes (Mann–Whitney $P > 0.05$). For viral richness, we did not observe difference in the metatranscriptomics dataset (x = 26 SLC) but observed a slight enrichment in viral richness in the caries microbiome ($x_{\text{Caries}} = 34.5$ SLCs, $x_{\text{Caries-free}} = 30$ SLCs; Mann–Whitney $P = 0.026$).

We implemented differential expression analysis between caries and caries-free cohorts at the taxonomic level (SLC expression) and PGFC level (engineered taxonomy-functional composite features). We did not observe any statistically significant components, neither SLCs nor PGFCs, using compositionally-aware methods such as ANCOM and ALDEx2. However, the lack of clear taxonomic or functional differences between the cohorts suggests interactions between variables is important, illustrating the need of differential networks to interrogate the caries and caries-free microbial systems.

Phenotype-specific coexpression networks reveal unique taxonomic and metabolic characteristics

As the caries phenotype is a multifactorial disease (7), the most natural approach for investigating associations would be through network analysis as such methodologies are useful for modeling complex systems with unknown structure. To be specific, the true structure (if one exists) of microbial interactions within each phenotype in our dataset is unknown a priori. Therefore, we must infer the structure of each network using data-driven approaches. Using an ensemble approach, we computed compositionally-aware coexpression networks, with PGFCs, an engineered feature based on taxonomy and functional potential (see the "Methods" section), as nodes ($N_{nodes} = 2{,}478$) and *rho* proportionalities as edge weights ($N_{edges} = 3{,}069{,}003$), for caries and caries-free microbiomes (PSCN$_{\text{Caries}}$ and PSCN$_{\text{Caries-free,}}$ respectively). The total connectivity of the PSCN$_{\text{Caries}}$ was 301,163.9 *k* with PSCN$_{\text{Caries-free}}$ $\sim$7% lower (279,832.1 *k*, Table S6). Unsupervised clustering of the PSCNs sorted by median connectivity revealed clusters heterogeneous with respect to taxonomy, and a sharp drop off in connectivity at 250 *k* (Fig. 4A and B, Tables S5 and S7). In this high-connectivity range, there are 12 PSCN$_{\text{Caries}}$ clusters (749 PGFCs) and 8 PSCN$_{\text{Caries-free}}$ clusters (555 PGFCs), which will be referred to as high-connectivity PSCN clusters (HCPC).

One approach in computing homogeneity is via normalized entropy and, in this context, can be interpreted as cluster homogeneity where low entropy translates to a cluster being dominated by a single taxa (more homogenous) and high entropy as taxa being evenly distributed within a cluster (more heterogeneous). The most highly connected cluster in both PSCNs is Cluster-1 (HCPC-4A.1 and HCPC-4B.1), which is the second largest cluster in each network and one of the most heterogenous with respect to taxonomy. We observed a modest trend that HCPCs in the caries microbiome have higher taxonomic homogeneity than the caries-free microbiome. The caries HCPCs tend to have lower normalized entropy than the caries-free HCPCs especially compared to when considering all clusters; though, the number of observations was not sufficient for this statistical analysis and these results will not be further explored.

Despite the caries microbiome HCPCs being slightly more homogenous, the highest connectivity PSCN$_{\text{Caries}}$ cluster (HCPC-4A.1) is one of the most heterogenous clusters in the system. The majority of HCPC-4A.1 connectivity (82%) is from *Veillonella*, *Streptococcus*, *Granulicatella*, and *Kingella_B*. The remaining caries HCPCs are enriched in other bacteria including *Streptococcus*, *Capnocytophaga*, *Haemophilus_D*, *Neisseria*, *Cardiobacterium*, and *Aggregatibacter*. The highest connectivity cluster in PSCN$_{\text{Caries-free}}$ is HCPC-4B.1 whose connectivity is primarily from *Streptococcus sanguinis*, *Veillonella parvula_A*, and *Granulicatella adiacens*. The remaining caries-free HCPCs are enriched in *Neisseria*, *Capnocytophaga*, *Fusobacterium*, *Haemophilus_D*, and *Prevotella*. We observed a substantial overlap in high-connectivity genera but the cluster membership of these genera is phenotype-specific and these configurations may provide key insight into how a system, whether caries or caries-free, stabilizes.

The second highest connectivity HCPCs in both caries (HCPC-4A.32) and caries-free (HCPC-4B.22) microbiomes are homogenous for *Streptococcus* and *Neisseria*, respectively. Connectivity from HCPC-4A.32 is mainly derived from *S. sanguinis* (98%) while connectivity from HCPC-4B.22 is 100% attributable to a novel *Neisseria* species (BC6). We also observed HCPC-4A.39 as another homogenous caries HCPC for *Capnocytophaga sputigena.* In both PSCNs, carbohydrate metabolism (27% PSCN$_{\text{Caries}}$, 36% PSCN$_{\text{Caries-free}}$) and cofactor/vitamin biosynthesis (10.7% PSCN$_{\text{Caries}}$, 8.2% PSCN$_{\text{Caries-free}}$) are attributable to most of the connectivity (Fig. S2). Glycolysis, gluconeogenesis, and pentose phosphate are heterogenous amongst the HCPCs regardless of phenotype. The citric acid cycle was responsible for the majority of the carbohydrate connectivity in PSCN$_{\text{Caries}}$ HCPC-4A.29; a heterogenous cluster enriched in *Neisseria* and *Prevotella*. Several cofactor and vitamin metabolic pathways were common amongst the HCPCs.

Community detection algorithms such as Louvain (29) and, its updated successor, Leiden (30) have been used to investigate the structure of large and complex networks. The former has been used to study various biological networks (31, 32, 33, 34) while Leiden is new and sparingly applied to biological systems, it addresses defects associated with Louvain. As these algorithms are stochastic, we utilized an iterative version of the Leiden community detection algorithm to investigate how these phenotype-specific HCPCs are structured and how the HCPCs partition into tightly connected high-confidence communities in an induced graph. The caries HCPCs naturally partition into Communities-4C.I-III while the caries-free HCPCs partition into Communities-4D.I-II (Fig. 4C and D, Table S5). Leiden communities revealed similar coexpression of two complementary configurations in both PSCNs: (1) majority *Bacteroidota* (PSCN$_{\text{Caries}}$ Community-4C.I and PSCN$_{\text{Caries-free}}$ Community-4D.I); and (2) majority *Firmicutes* via *Streptococcus* (PSCN$_{\text{Caries}}$ Community-4C.II and PSCN$_{\text{Caries-free}}$ Community-4D.II) (Fig. 4E and F).

Community-4C.I and 4D.I have high overlap in taxonomic membership, but they also have several unique taxa that may provide insight into phenotype-specific system states. Interestingly, no *Neisseria* were observed in PSCN$_{\text{Caries}}$ Community-4C.I but high *Neisseria* genus-level membership was observed in the complementary PSCN$_{\text{Caries-free}}$ Community-4D.I (Fig. 4E). However, in PSCN$_{\text{Caries}}$, we observed high *Neisseria* genus-level membership in Community-4C.III coexpressed with more *Neisseria* and *Haemophilus* (Fig. 4E). *Neisseria* are highly connected in both PSCNs but their community membership, the taxa they are interacting with, is phenotype specific. More specifically, *Neisseria* appears to shift from high coexpression with several *Bacteroidota* species in the caries-free cohort to other *Neisseria* and *Haemophilus* in the caries cohort. The connectivity of *Haemophilus_D parainfluenzae* and an unclassified *Neisseria* (BC6) is relatively high in PSCN$_{\text{Caries}}$ Community-4C.III and these taxa are completely absent from the *Neisseria* enriched community in PSCN$_{\text{Caries-free}}$.

In terms of metabolism, drug resistance is only observed in PSCN$_{\text{Caries}}$ Community 4C.II, specifically *S. sanguinis* beta-Lactam resistance (Fig. S3). Both caries and caries-free microbiomes lack arginine, proline, and lipid metabolism in the *Bacteriodota*-centric communities (PSCN$_{\text{Caries}}$ Community-4C.I and PSCN$_{\text{Caries-free}}$ Community-4D.I) but provide these pathways in the *Firmicutes*-centric communities (PSCN$_{\text{Caries}}$ Community-4C.II and PSCN$_{\text{Caries-free}}$ Community-4D.II). Conversely, these *Firmicutes*-centric communities lack sulfur metabolism which appears to be provided by *Bacteriodota*-centric community. Central carbohydrate metabolism connectivity is much higher in the *Bacteriodota*-

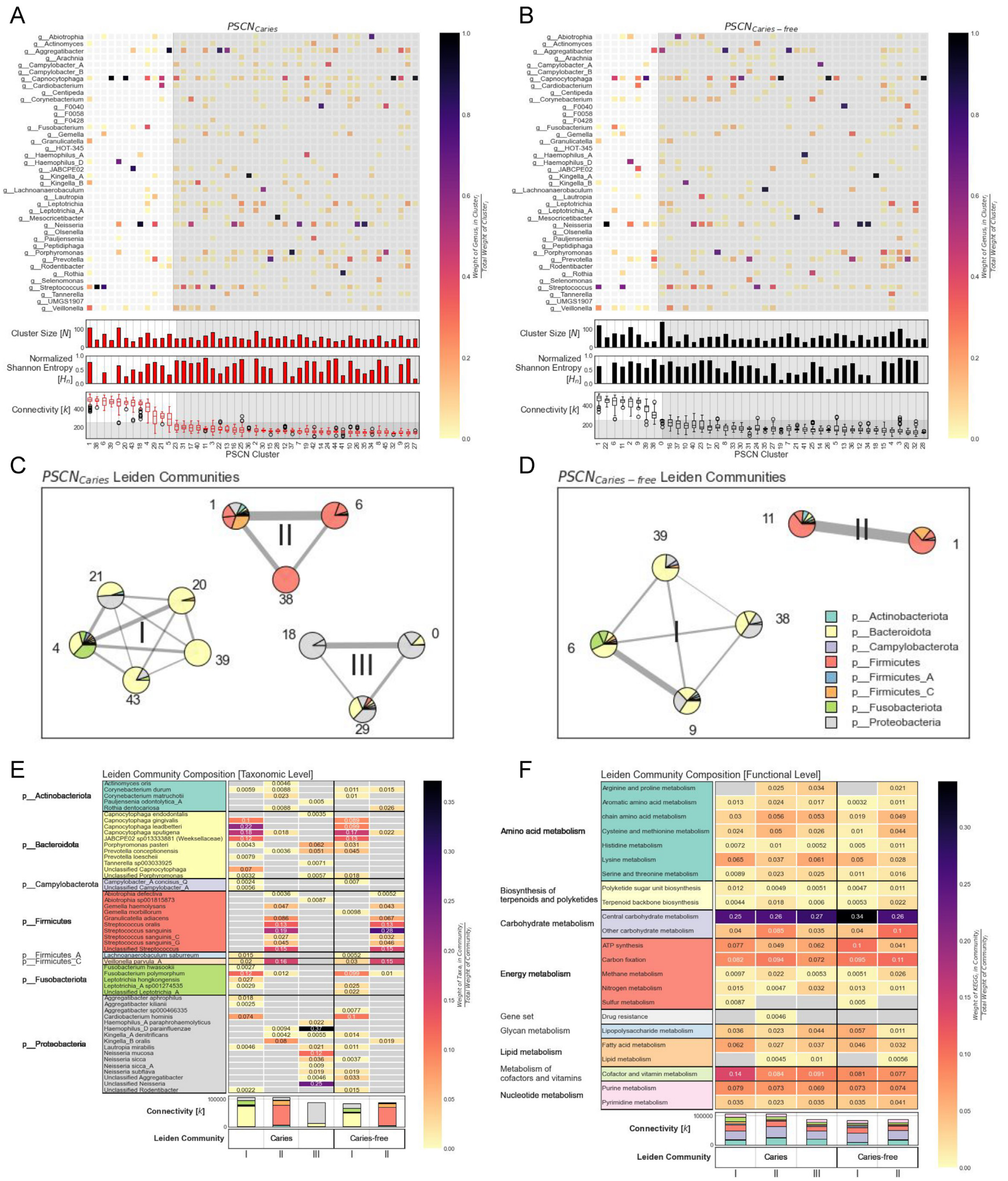

Fig. 4. Connectivity-based community detection in PSCNs. Heatmap of clustered PSCNs for (A) caries and (B) caries-free phenotypes sorted by median cluster connectivity [*k*] in box-plot below with threshold for high-connectivity clusters set at 250 *k* in both PSCNs. Each *i*, *j* value in the heatmap represents the weight of genus *i* in cluster *j* divided by the total weight of cluster *j*; that is, the weighted proportion of each genus in each cluster. Leiden community detection algorithm applied to high-connectivity PSCN clusters for (C) caries and (D) caries-free phenotypes. Roman numerals indicate PSCN-specific Leiden communities for reference. Pie charts indicate proportion of genus weight in each Leiden community and colored by phyla. Clustering was performed using the distance version of *rho* proportionality and Ward linkage. Heatmaps of Leiden Community connectivity (C and D) relative to taxonomy (E) and KEGG functional pathways (F) showing the connectivity of each grouping relative to the total connectivity in the community.

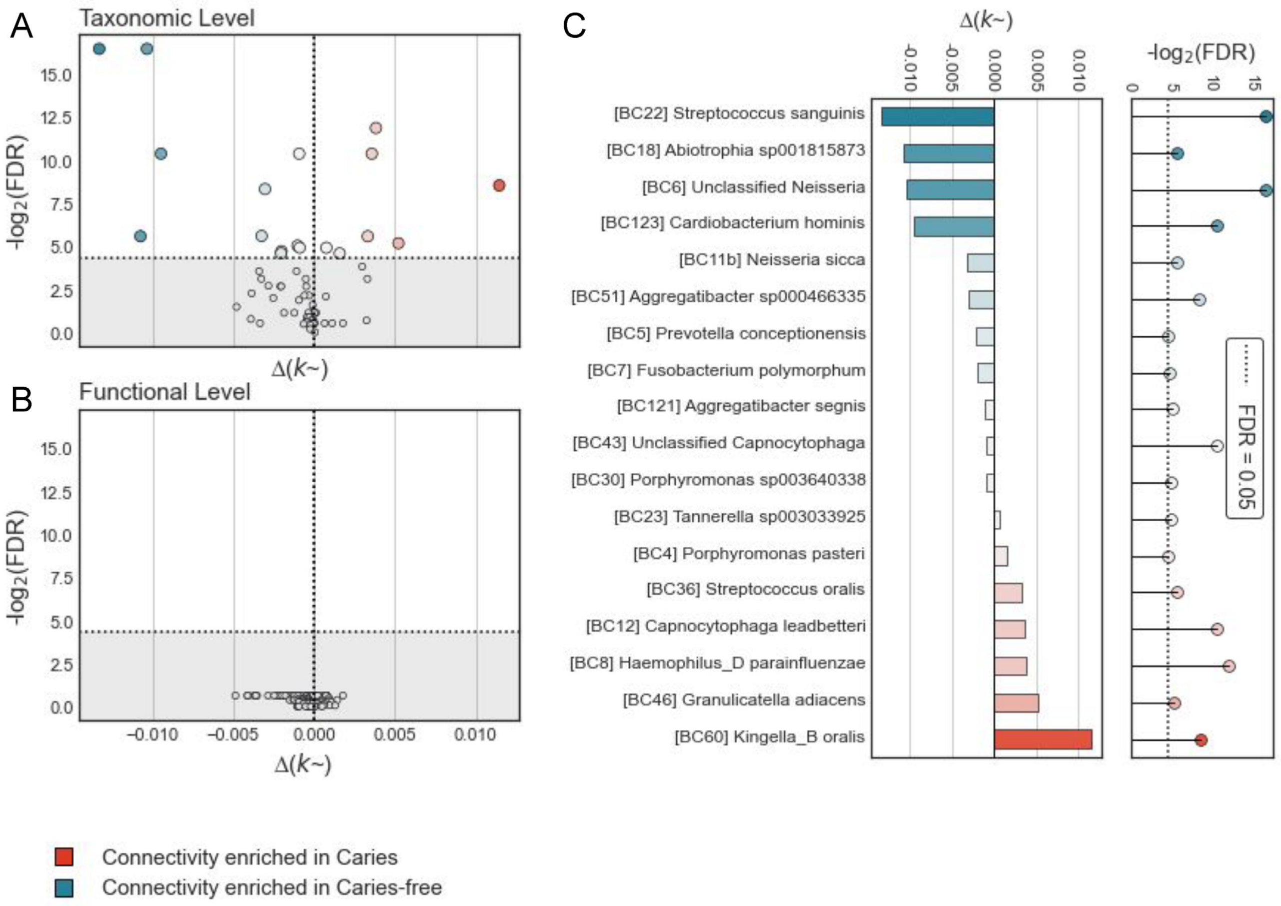

Fig. 5. Comparing PSCNs with respect to taxonomic or functional levels. Volcano plots of Leiden community PGFCs from Fig. 4 showing change in scaled connectivity $[\Delta k^{\sim}]$ and -$\log_2$(FDR) with respect to (A) taxonomic and (B) functional PGFC levels. (C) Sorted barchart of taxa with statistically different connectivities between caries and caries-free PSCNs. FDRs computed using Wilcoxon signed-rank test followed by Benjamini/Hochberg multiple hypothesis correction. Red represents an enrichment in connectivity in the caries PSCN with blue represents an enrichment in connectivity in the caries-free PSCN.

centric $\text{PSCN}_{\text{Caries-free}}$ Community-4D.I relative to all of the other communities, which may suggest that the taxa and central carbohydrate mechanisms in this community promote a healthy oral microbiome.

We compared the scaled connectivity of different PGFC groupings between caries and caries-free PSCNs using the union of PGFCs in caries and caries-free HCPCs. We observed statistically significant differential connectivity when grouping PGFCs by taxonomic level ($N = 7$ PGFCs enriched in $\text{PSCN}_{\text{Caries}}$ and $N = 11$ PGFCS enriched in $\text{PSCN}_{\text{Caries-free}}$) and none when grouped by functional level (Fig. 5). Although, the connectivity of high-level metabolic functional profiles is similar for both PSCNs, the taxa responsible for these driving functions are unique to the phenotype. The taxa with enriched connectivity in $\text{PSCN}_{\text{Caries}}$ were *Kingella_B oralis* trailed by *G.*, *Haemophilus_D parainfluenzae*, *Capnocytophaga leadbetteri*, and *Streptococcus oralis.* The taxa with greatest enriched connectivity in $\text{PSCN}_{\text{Caries-free}}$ were *S. sanguinis*, *Abiotrophia sp001815873*, an unclassified *Neisseria* (BC6), and *Cardiobacterium hominis.* Although unclassified *Neisseria* and *Abiotrophia. sp001815873* are enriched in $\text{PSCN}_{\text{Caries-free}}$, they are not present in the caries-free communities (Fig. 4E) because they were not in the caries-free HCPCs. This discrepancy in membership suggests that connectivities of these taxa, though enriched, were masked by other high-connectivity taxa in $\text{PSCN}_{\text{Caries-free}}$.

Differential coexpression networks suggests community scale metabolic restructuring through *C. hominis*

Differential coexpression networks (DCNs) reveal changes in connectivity between a reference (caries-free) and treatment (caries) network. As ensemble PSCNs are the building blocks of DCNs, our DCNs provide the same benefits with respect to outlier resistance. Previous approaches have used DCNs but did not use compositionally-aware association metrics or ensemble networks (21, 35). While differential abundance/expression analyses can be useful in identifying feature enrichment (e.g. OTU, MAG, ORF, gene, etc.), each method has their own caveats in assumptions about the data distributions [well characterized in ref. (36) with the establishment of reference frames] and provide no information regarding differences in pairwise interactions; an essential perspective when studying diseases resulting from dysbiosis. Using the $\text{PSCN}_{\text{Caries-free}}$ as a reference network and $\text{PSCN}_{\text{Caries}}$ as the treatment network, we were able to construct a DCN using the 875 PGFCs from the community detection analysis for seamless cross-referencing between PSCNs and the DCN. In the DCN, differential connectivity (denoted as $\Delta k^{\sim}$) is positive and negative when a connectivity is enriched in the caries and caries-free microbiomes, respectively. Unsupervised clustering of the DCN revealed six clusters (Fig. 6, Tables S5 and S6), of which there were three high-

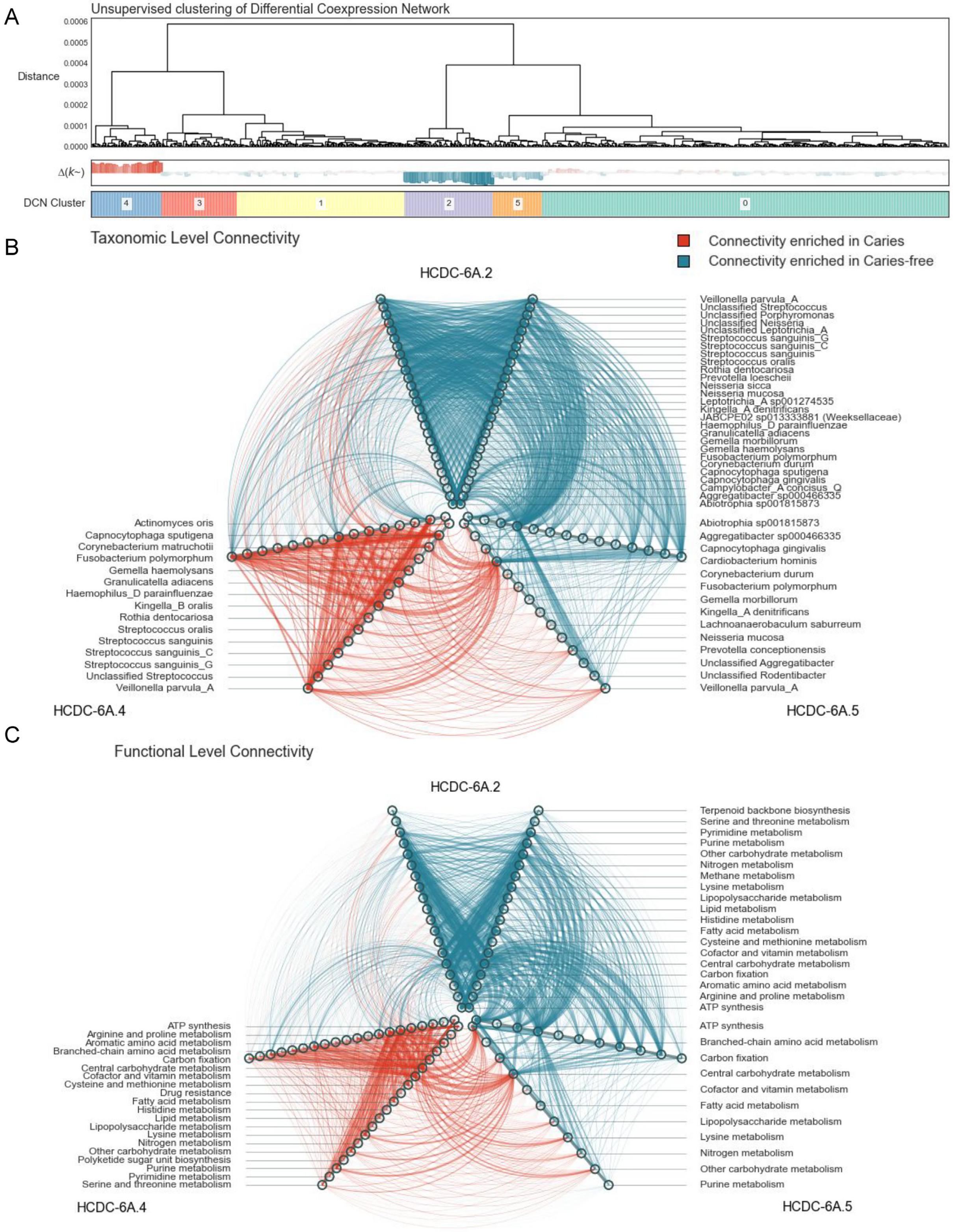

Fig. 6. Differential network analysis between caries and caries-free PSCNs. (A) Hierarchical clustering of DCN using Leiden community PGFCs from Fig. 4. Barchart shows the differential connectivity [$\Delta k^{\sim}$] for PGFC nodes with positive (red) values indicating higher scaled connectivity in caries PSCN and negative (blue) values indicating higher scaled connectivity in caries-free PSCN. Colored panel on bottom shows DCN clusters sorted by the number of PGFCs in cluster with the largest cluster being 0 and the smallest being 5. (B) Shows hive plot of taxonomic categories for DCN(Cluster-4), DCN(Cluster-2), and DCN(Cluster-5) with red and blue edges following the scheme in (A). (C) Shows the same hive plot in (B) but grouping PGFCs by higher-level KEGG categories instead of taxonomic categories.

connectivity DCN clusters (HCDC), each being diagnostic of phenotype; HCDC-6A.4 had enriched connectivity in the caries microbiome while HCDCs-6A.2 and 5 had enriched connectivity in the caries-free microbiome. For the only HCDC with connectivity enriched in the caries microbiome (HCDC-6A.4), the differential connectivity was primarily from *C. sputigena*, *Kingella_B oralis, Vellonella parvula_A*, *S. sanguinis*, *S. oralis*, and several species of unclassified *Streptococci* (Fig. S4A) via carbohydrate and cofactor/vitamin metabolism (Fig. S4B).

HCDC-6A.4 included 43% of all taxa within the DCN. HCDCs with enriched connectivity in caries-free microbiome contained a broader range of microbes. However, most of these taxa were in HCDC-6A.2 with more than 77% of the taxa in the DCN, which was not the case for HCDC-6A.5 with 40% of the taxa. In HCDC-6A.2, most of the differential connectivity was attributable to *S. sanguinis*, *Abiotrophia. sp001815873*, an unclassified *Neisseria* (BC6), *Rothia dentocariosa*, and several *Fusobacteriota* via carbohydrate metabolism, ATP synthesis, carbon fixation, and cofactor/vitamin biosynthesis (Fig. S4B). While HCDC-6A.2 is heterogenous in terms of taxa membership, HCDC-6A.5 is fairly homogenous with most of the connectivity from *C. hominis* via carbohydrate metabolism. *Veillonella parvula_A* and *Fusobacterium polymorphum* are the only taxa that had membership in all HCDCs. As *Veillonella. parvula_A* had high differential connectivity in both caries and caries-free phenotypes through different metabolic pathways, this finding may provide insight into dysbiosis. Several other taxa including *Haemophilus_D parainfluenzae* had membership in the HCDC-6A.4 and with at least one HCDC with negative differential connectivity (Fig. S4A).

Comparing set membership between HCDCs revealed key metabolic differences between microbiomes. HCDC-6A.4 exclusively had 14 KEGG modules with the most notable including pentose phosphate pathway, phosphate acetyltransferase-acetate kinase, beta-Lactam resistance, several cofactor/vitamin pathways (Table S5). HCDC-6A.2 had 19 KEGG modules not in HCDC-6A.4, which included many carbohydrate metabolic, reductive pentose phosphate cycle, and dissimilatory nitrate reduction pathways. HCDC-6A.5 only had four exclusive KEGG modules, including citrate cycle, fumarate reductase, and Raetz pathway with citrate cycle, and fumarate reductase metabolism from *C. hominis*.

Hive plots are a network visualization framework that groups nodes with respect to predefined axes. In this case, grouping PGFCs by taxa or KEGG categories for higher-level nodes and HCDCs for axes. The hive structure visualizes both intra- and inter-cluster differential connectivity clearly revealing hub nodes connecting clusters (Fig. 6B and C). In the context of this DCN, *C. hominis* was a link between the highest differential connectivity HCDCs for caries (HCDC-6A.4) and caries-free (HCDC.6A.2) microbiomes even though each cluster's intra-cluster connectivity is sign specific. HCDC-6A.4 had very low connectivity to HCDC-6A.2 but both have high connectivity to HCDC-6A.5 primarily via *C. hominis*. However, positive differential connectivity from HCDC-6A.5 to HCDC-6A.4 was mainly from *C. hominis* carbohydrate metabolism and ATP synthesis from other bacterial species. In the connection between *C. hominis* and HCDC-6A.2, we observed many more taxa, also at greater differential connectivity magnitude, primarily through *S. sanguinis*, *Abiotrophia. sp001815873*, and an unclassified *Neisseria* (BC6) with a long tail of taxa with negative differential connectivity. In this latter case, the highly negative differential connectivity from *C. hominis* to HCDC-6A.2 is spread out across many metabolic pathways and is not disproportionally weighted at carbohydrate and ATP synthesis suggesting *C. hominis* may have a holistic relationship in a caries-free microbiomes while also playing a potentially nonbeneficial role in caries microbiomes. We expanded *C. hominis* carbohydrate metabolism and ATP synthesis modules out in a separate DCN (Fig. 7, Table 3, and Tables S5 and S7).

After removing low-connectivity edges (Fig. S5), this DCN revealed five Leiden communities, denoted as Communities-7.I-V, with the two largest communities being Community-7.I and Community-7.II. Consistent with our previous hive networks, we observed a community with connectivity primarily enriched in the caries microbiome (Community-7.I) and several communities with connectivity almost exclusively enriched in caries-free microbiome (Communities-7.II to V). The exception to the latter is *C. hominis* pyruvate oxidation and *C. sputigena* polyketide sugar unit biosynthesis connectivity enriched in caries microbiome in Community-7.III. Community-7.II, the largest of the negative differential connectivity communities, was far less complex than Community-7.I and has *C. hominis* pentose phosphate as highly central nodes. There were three other small communities with negative differential connectivity and the most interesting of these is Community-7.IV as *C. hominis* glucose to UDP-glucose conversion is connected to mostly to *Veillonella. parvula_A* cofactor/vitamin metabolism but also *Neisseria* nitrogen metabolism/methane metabolism and *Corynebacterium durum* carbohydrate metabolism.

The most complex and informative community is Community-7.I, which is primarily composed of positive differential connectivity edges, those enriched in the caries microbiome. The negative differential connectivity edges are primarily from ATP synthesis of *Veillonella. parvula_A* and *Leptotrichia_A sp001274535*. Said nodes only have negative differential connectivity edges, which suggest they are influential to the rest of the community in a caries-free microbiome and this may provide insight into community-scale restructuring in the caries microbiome. Also, worthy of note, the only nodes with both positive and negative differential connectivity edges are from *C. hominis* supporting the hypothesis that *C. hominis* is an essential player in the transition from caries-free to caries phenotypes and vice versa. However, the most striking feature of this community is that *C. hominis* citrate cycle and fumarate reductase are highly centralized suggesting a shift in carbohydrate metabolism from pentose phosphate cycle to citrate acid cycle in the caries microbiome. We also observed various types of carbohydrate metabolism in Community-7.I with positive differential connectivity from several other organisms (Table 3).

Predictive models applied to caries diagnosis

Feature selection and predictive modeling was implemented to further evaluate PGFC features that were indicative of caries diagnosis. In particular, the *Clairvoyance* feature selection algorithm (37) that has been previously evaluated on identifying diagnostic genes related to antibiotic resistance (37) and multimodal associations related to childhood undernutrition (38) was used to identify PGFCs that were able to accurately discriminate caries individuals from caries-free individuals. To allow for seamless interpretation with the network analysis, the set of 212 PGFCs from the DCN were used as input into the feature selection algorithm and this was implemented for PGFCs represented as MCR and as CLR transformed abundances to yield two separate feature sets. This mixed feature architecture allowed for a novel type stacking ensemble where each base classifier uses a specific feature set and feature representation (e.g. MCR and CLR values, simultaneously) leveraging the strength of each measurement in the ability to predict caries phenotype. The MCR feature set included 36 PGFCs, the

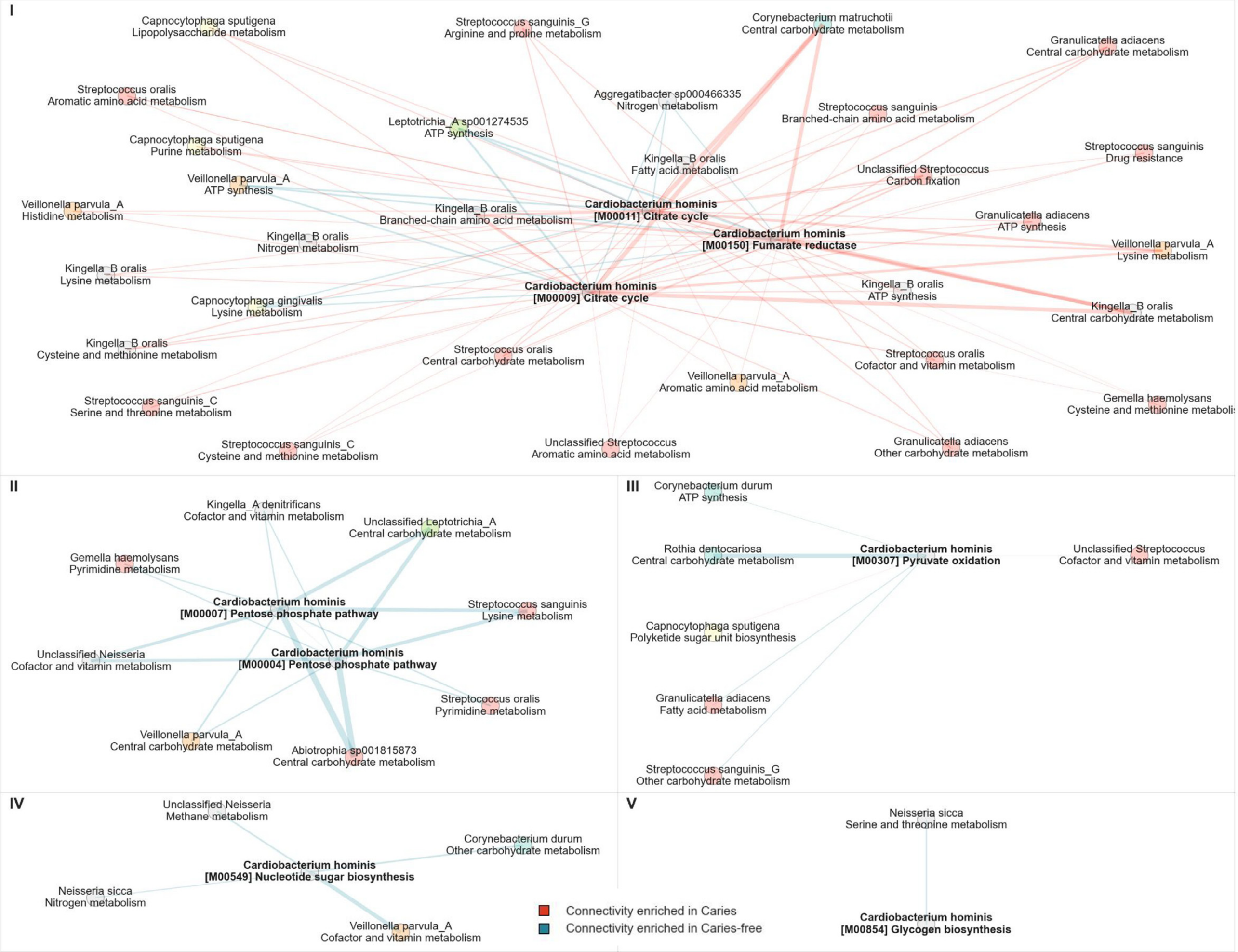
I
Capnocytophaga sputigena
Lipopolysaccharide metabolism
Streptococcus sanguinis_G
Arginine and proline metabolism
Corynebacterium matruchotii
Central carbohydrate metabolism
Granulicatella adiacens
Central carbohydrate metabolism
Streptococcus oralis
Aromatic amino acid metabolism
Aggregatibacter sp000466335
Nitrogen metabolism
Streptococcus sanguinis
Branched-chain amino acid metabolism
Leptotrichia_A sp001274535
ATP synthesis
Capnocytophaga sputigena
Purine metabolism
Streptococcus sanguinis
Drug resistance
Kingella_B oralis
Fatty acid metabolism
Unclassified Streptococcus
Carbon fixation
Veillonella parvula_A
ATP synthesis
Veillonella parvula_A
Histidine metabolism
Kingella_B oralis
Branched-chain amino acid metabolism
Cardiobacterium hominis
[M00011] Citrate cycle
Granulicatella adiacens
ATP synthesis
Kingella_B oralis
Nitrogen metabolism
Cardiobacterium hominis
[M00150] Fumarate reductase
Veillonella parvula_A
Lysine metabolism
Kingella_B oralis
Lysine metabolism
Cardiobacterium hominis
[M00009] Citrate cycle
Kingella_B oralis
ATP synthesis
Capnocytophaga gingivalis
Lysine metabolism
Kingella_B oralis
Central carbohydrate metabolism
Kingella_B oralis
Cysteine and methionine metabolism
Streptococcus oralis
Central carbohydrate metabolism
Streptococcus oralis
Cofactor and vitamin metabolism
Veillonella parvula_A
Aromatic amino acid metabolism
Streptococcus sanguinis_C
Serine and threonine metabolism
Gemella haemolysans
Cysteine and methionine metaboli
Streptococcus sanguinis_C
Cysteine and methionine metabolism
Unclassified Streptococcus
Aromatic amino acid metabolism
Granulicatella adiacens
Other carbohydrate metabolism
II
Kingella_A denitrificans
Cofactor and vitamin metabolism
Unclassified Leptotrichia_A
Central carbohydrate metabolism
Gemella haemolysans
Pyrimidine metabolism
Cardiobacterium hominis
[M00007] Pentose phosphate pathway
Streptococcus sanguinis
Lysine metabolism
Unclassified Neisseria
Cofactor and vitamin metabolism
Cardiobacterium hominis
[M00004] Pentose phosphate pathway
Streptococcus oralis
Pyrimidine metabolism
Veillonella parvula_A
Central carbohydrate metabolism
Abiotrophia sp001815873
Central carbohydrate metabolism
III
Corynebacterium durum
ATP synthesis
Rothia dentocariosa
Central carbohydrate metabolism
Cardiobacterium hominis
[M00307] Pyruvate oxidation
Unclassified Streptococcus
Cofactor and vitamin metabolism
Capnocytophaga sputigena
Polyketide sugar unit biosynthesis
Granulicatella adiacens
Fatty acid metabolism
Streptococcus sanguinis_G
Other carbohydrate metabolism
IV
Unclassified Neisseria
Methane metabolism
Corynebacterium durum
Other carbohydrate metabolism
Cardiobacterium hominis
[M00549] Nucleotide sugar biosynthesis
Neisseria sicca
Nitrogen metabolism
Veillonella parvula_A
Cofactor and vitamin metabolism
🟥 Connectivity enriched in Caries
🟦 Connectivity enriched in Caries-free
V
Neisseria sicca
Serine and threonine metabolism
Cardiobacterium hominis
[M00854] Glycogen biosynthesis

Fig. 7. *Cardiobacterium hominis* carbohydrate and ATP synthesis metabolism centric DCN. DCN of PGFCs from DCN(Cluster-4), DCN(Cluster-2), and DCN(Cluster-5) grouped by either (1) taxonomy and higher-level KEGG categories for non-*C. hominis* PGFCs; and (2) *C. hominis* KEGG modules related to carbohydrate and ATP synthesis metabolism. Edge weights indicate differential connectivity with positive (red) edges indicating higher scaled connectivity in caries PSCN and negative (blue) edges indicating higher scaled connectivity in caries-free PSCN. Roman numerals indicate connected components in DCN (i.e. isolated subgraphs within the larger graph).

CLR feature set included 27 PGFCs, and 11 PGFCs were shared between both base models (52 unique PGFCs) (Fig. S6B). The PGFCs selected via feature selection included *Cardibacterium hominis* pentose phosphate and TCA cycle as some of the highest weighted features that were able to discriminate caries phenotypes (Table S9). The stacking ensemble classification model was able to predict unobserved twin groupings with an accuracy of >96.5% (see the "Methods" section for cross-validation). In this context, accuracy can be interpreted as the reliability of a feature set to be sufficient in diagnosing caries. This is in stark contrast to predictive modeling using the 212 PGFCs from the DCN without feature selection, which yielded a baseline classification accuracy of only 58.8%.

Discussion

This study provides evidence of a core bacterial microbiome and a personalized viral microbiome that is transcriptionally active within the supragingival plaque of this cohort of Australian children regardless of collection center, age, or sex. This core microbiome supports the ecological plaque hypothesis that environmental conditions influence the metabolism of existing microbes nudging the community into a cariogenic configuration, rather than it being associated with extensive gain or loss of taxa. As the oral community is able to shift the collective metabolism to adapt to a cariogenic environment, the reverse must also be true given the prevalence of this core community. The implications of such a finding are both therapeutic and diagnostic. The specific abundances of taxa or even transcripts are not diagnostic, but a panel of transcripts and their associations are highly diagnostic and could be used as remote dentistry as demonstrated by the predictive model's high accuracy in diagnosing caries phenotype. Similarly, probiotics that revert community associations might be powerful therapeutics. Characterizing the interactions between microbes and their additive metabolism is expected to provide a deeper insight into what it means metabolically to have a cariogenic oral environment and, also important, a caries-free environment. One objective of the study was to determine if an untargeted tooth swab of both cariogenic and noncariogenic communities combined with sequencing and in-silico analysis could predict the signals diagnostic of phenotype with an accuracy >96.% using 52 unique biological features. This type of "take-at-home" assay augments current dental prophylaxis, which is dependent upon in-person visitation, and increases patient equity.

Table 3. Carbohydrate metabolism and ATP synthesis nodes in DCN Leiden Communities.

PGFC	Species	Category	Description	HCDC	Community
BC123|M00150	*C. hominis*	ATP	Fumarate reductase	5	I
BC46|M00157	*G. adiacens*	ATP	F-type ATPase	4	I
BC60|M00144	*Kingella_B oralis*	ATP	NADH:quinone oxidoreductase	4	I
BC21|M00159	*Leptotrichia_A sp001274535*	ATP	V-type ATPase	2	I
BC1|M00153	*Veillonella. parvula_A*	ATP	Cytochrome bd ubiquinol oxidase	2	I
BC123|M00009	*C. hominis*	CCM	Citrate cycle (TCA cycle, Krebs cycle)	5	I
BC123|M00011	*C. hominis*	CCM	Citrate cycle, 2-oxoglutarate ≥ oxaloacetate	5	I
BC0b|M00003	*Corynebacterium matruchotii*	CCM	Gluconeogenesis, oxaloacetate ≥ fructose-6P	4	I
BC0b|M00001	*C. matruchotii*	CCM	Glycolysis (EMP), glucose ≥ pyruvate	4	I
BC0b|M00002	*C. matruchotii*	CCM	Glycolysis	4	I
BC46|M00007	*G. adiacens*	CCM	Pentose phosphate pathway, nonoxidative phase	4	I
BC60|M00011	*Kingella_B oralis*	CCM	Citrate cycle, 2-oxoglutarate ≥ oxaloacetate	4	I
BC60|M00003	*Kingella_B oralis*	CCM	Gluconeogenesis, oxaloacetate ≥ fructose-6P	4	I
BC60|M00001	*Kingella_B oralis*	CCM	Glycolysis (EMP), glucose ≥ pyruvate	4	I
BC60|M00002	*Kingella_B oralis*	CCM	Glycolysis	4	I
BC60|M00307	*Kingella_B oralis*	CCM	Pyruvate oxidation, pyruvate ≥ acetyl-CoA	4	I
BC60|M00308	*Kingella_B oralis*	CCM	Semi-phosphorylative (EDP), gluconate ≥ glycerate-3P	4	I
BC36|M00005	*S. oralis*	CCM	PRPP biosynthesis, ribose 5P ≥ PRPP	4	I
BC46|M00854	*G. adiacens*	OCM	Glycogen biosynthesis, glucose-1P ≥ glycogen/starch	4	I
BC18|M00157	*Abiotrophia. sp001815873*	ATP	F-type ATPase	2	II
BC18|M00159	*Abiotrophia. sp001815873*	ATP	V-type ATPase	2	II
BC123|M00004	*C. hominis*	CCM	Pentose phosphate pathway	5	II
BC123|M00007	*C. hominis*	CCM	Pentose phosphate pathway, non-oxidative phase, fructose 6P ≥ ribose 5P	5	II
BC1|M00003	*Veillonella. parvula_A*	CCM	Gluconeogenesis, oxaloacetate ≥ fructose-6P	5	II
BC1|M00001	*Veillonella. parvula_A*	CCM	Glycolysis (Embden–Meyerhof pathway), glucose ≥ pyruvate	5	II
BC1|M00002	*Veillonella. parvula_A*	CCM	Glycolysis, core module involving three-carbon compounds	5	II
BC123|M00854	*C. hominis*	OCM	Glycogen biosynthesis, glucose-1P ≥ glycogen/starch	5	III
BC123|M00307	*C. hominis*	CCM	Pyruvate oxidation, pyruvate ≥ acetyl-CoA	5	IV
BC123|M00549	*C. hominis*	OCM	Nucleotide sugar biosynthesis, glucose ≥ UDP-glucose	5	V

Category refers to KEGG Level 3 metabolic category while description refers to KEGG module description. Community refers to Leiden communities for DCN. Acronyms: ATP—ATP Synthesis, CCM—Central carbohydrate metabolism, OCM—Other carbohydrate metabolism.

The complexity of this study required many novel methods to be developed for identifying mechanisms involved in caries-related dysbiosis. With a scope only considering single taxa expression patterns (e.g. unsupervised clustering of samples and differential expression analysis), we would have not been able to identify any distinguishing features of caries or caries-free phenotypes. Our development of novel association network methodologies built on the fundamental network concepts inspired by WGCNA such as implementing fully connected, undirected, and weighted networks that can be clustered via hierarchical clustering. Our approach augments legacy methods by leveraging feature engineering to reduce dimensionality and exploitation of natural biological ontologies, the use of proportionality instead of correlation, consensus Leiden community detection, and a novel ensemble network framework to build distributions of edge weights rather than singular point estimates. Only through the inferred interactions between PGFCs were we able to notice trends in the data that could describe caries-related regime shifts. We averted the intractability of naive "bag-of-genes" representations of the data (i.e. ORFs as individual units) by flexible feature engineering methodologies to simulate an in-silico reconstruction of the microbial community grouped by taxonomy and function; a "bag-of-genomes" approach. Analyzing caries-related dysbiosis using differential networks in PGFC-space rather than gene-space allowed for rapid and computationally economical methods for packing and unpacking biological hierarchies to explore regime shifts; bridging the gap between machine intelligence and biological insight. Although the feature engineering methods in this study are largely dependent on curated metabolic pathways, they were designed to be generalizable to custom databases and unsupervised paradigms such as hierarchical clustering paired with gene ontology analysis.

Exploring the oral microbiome from unique vantage points through analyzing networks specific to a phenotype, the comparison of connectivity profiles, and the differentials between networks provided insight into not only dysbiotic regime shifts but also maintenance of caries-free status. Even high-level network statistics are biologically relevant in the context of caries-related dysbiosis. For instance, the caries PSCN had substantially greater total connectivity than the caries-free PSCN, which can be interpreted as a higher number of interacting microbes and diverse metabolic pathways. This enrichment in inferred interactions within a diseased community relative to a nondiseased community has been observed in other forms of dysbiosis in the human gut microbiome such as inflammatory bowel disease and obesity (39). This finding is especially relevant considering the microbial richness does not differ between caries and caries-free microbiomes. The larger total connectivity and number of high-connectivity coexpression clusters in caries microbiome suggests that there are more microbial and metabolic interactions occurring in carious systems. Similarly, the fewer number of high-connectivity clusters in a caries-free microbiome suggests that the caries-free phenotype has a microbiome dominated by a few key taxa and metabolic pathways.

Neisseria appears to be a key player with high connectivity in the supragingival plaque oral microbiome regardless of caries phenotype. Previous research has observed *Neisseria* as highly abundant in both caries and caries-free microbiomes (40) but, to our knowledge, this study is the first to report this trend in the context of network connectivity (Figs. 4 and 6) and RNA:DNA (Fig. 3). Although the connectivity of *Neisseria* is comparable in both microbiomes, the high connectivity in the caries microbiome is masked by a plethora of other highly connected genera and is ranked higher in the caries-free microbiome as a result of fewer high connectivity genera (Fig. 4A and B). However, we observed different microbial communities interacting with *Neisseria* when comparing between caries and caries-free microbiomes. In particular, several species of *Neisseria* were interacting with members of *Bacteroidota* in the caries-free microbiome and shifts to interactions with *Haemophilus_D parainfluenzae* and fellow *Neisseria* [mostly an unclassified *Neisseria* (BC6) and *Neisseria mucosa*] in the caries microbiome (Fig. 4E). This is interesting because several species of *Neisseria* had enriched connectivity in the caries-free microbiome and *Haemophilus_D parainfluenzae* had enriched connectivity in the caries microbiome (Fig. 5A and C). Although, *Neisseria* and *Haemophilus parainfluenzae* are both common in the oral cavity of caries-free individuals from the perspective of abundance (41, 42, 43), their interactions with other coexpressed microbes known to be associated with infections in humans [e.g. *Prevotella conceptionensis* from Community-C.II (44, 45)], may be indicative of caries dysbiosis. Many of the organisms discussed in this research have not been exhaustively characterized in the context of dental caries from an ecological perspective which presents an opportunity for future co-culture experiments.

The ability to collapse and expand PGFCs in these abstract network spaces can be used to identify unanticipated players with uncharacterized interactions relevant to maintaining either caries-free or caries microbiomes. For instance, when comparing PSCNs *C. hominis* is revealed to be one of the microbes with the highest enrichment in connectivity in the caries-free microbiome (Fig. 5A and C). However, the narrative is more complex when partitioning the PGFCs by differentially connected clusters (Fig. 6) and collapsing PGFCs by taxa-specific higher order KEGG categories. In a hive network layout, *C. hominis* emerges as a hub not only in the caries-free microbiome but also in the caries microbiome as it constitutes the majority of the differential connectivity within HCDC.5 primarily through ATP synthesis and carbohydrate metabolism. With *C. hominis* ATP synthesis and carbohydrate metabolism as a focal point, we were able to expand our focus to more specific KEGG modules while retaining high-level KEGG categories for the other microbes in the network to avoid the infamous and uninformative hairball plots (46) of overly complex networks (Fig. 7).

This hierarchical network, further validated through predictive modeling, implicates *C. hominis* as a nexus between caries-free and caries dysbiotic states through a switch from pentose phosphate to TCA cycle carbohydrate metabolism. Previous metabolic research confirms that both the TCA cycle and the pentose phosphate pathway function within the supragingival plaque in vivo and glycolytic activation caused an increase in pentose phosphate activity (47). These findings suggest that *C. hominis* mediated pentose phosphate pathway metabolism promotes a caries-free microbiome with the support of *S. sanguinis* lysine metabolism, *Abiotrophia. sp001815873* ATP synthesis, and *Neisseria* cofactor metabolism (Community-7.II). This hypothesis agrees with previous research as *S. sanguinis* and *Abiotrophia* have been known to co-occur in caries-free children (48) while *Neisseria*, as mentioned previously, has been associated with beneficial oral health. The simplicity of interactions enriched in the caries-free microbiome, Communities-7.II-V, agree with our theory that fewer taxa with more defined metabolisms are indicative of stable and healthy oral communities; thus, opening the door for potential probiotics, engineered microbial communities, and therapeutics for oral health and resilience.

The evidence for *C. hominis* TCA cycle and its association with caries dysbiosis is more complex in Community-7.I, which has considerably more taxa and metabolic pathways than communities that include the pentose phosphate pathway. However, this agrees with our earlier finding that caries-related regime shifts include more high-connectivity interactions without an increase in microbial richness; that is, greater total connectivity with the same core microbiome. Previous research has shown that the caries microbiome has the potential to metabolize more diverse sugar sources than the caries-free microbiome (27), which supports the notion that metabolism associated with dysbiotic caries communities is more complex than healthy communities without dental caries and, therefore, higher total network connectivity. In Communities-7.I-V, *C. hominis* is the only microbe that has connectivities enriched in caries and in caries-free microbiomes which supports our hypothesis of turncoat behavior in regards to oral health. *Cardiobacterium hominis* TCA cycle had enriched connectivity to carbohydrate metabolism from *Kingella oralis* (49), *S. oralis* (48), and *C. matruchotii* (40) in the caries microbiome, which have previously been statistically associated with caries dysbiosis in children.

Diseases stemming from microbial dysbiosis are often complex and difficult to investigate due to computational limitations and human interpretation. Our research addresses a critical limitation in paired metagenomics and metatranscriptomics across multiple samples/subjects: that is, how to have biologically accurate assemblies not biased by coassembled chimeras while also producing overlapping features (e.g. SLC, SLC-specific orthogroups). Furthermore, the analytical methodology (e.g. feature engineering and network analysis) employed in this study, though developed for the oral microbiome, can be generalized to other diseases, environments, and modalities (e.g. clinical measurements, metabolomics, proteomics). This research demonstrates how investigating microbiomes from different vantage points can provide insight into microbial ecosystems and their relevance in health and disease.

Methods

Sample collection

The study design has been described previously for this BioProject (PRJNA383868) in sister studies (27, 50, 13). In particular for the metatranscriptomics cohort, dental plaque samples were collected from participants of the University of Adelaide Craniofacial Biology Research Group Tooth Emergence and Oral Health Study (CBRG) ($n = 52$), and the Murdoch Children's Research Institute (MCRI) Peri/Postnatal Epigenetic Twins Study (PETS) ($n = 39$). Human research with PETS subjects was approved by the Royal Children's Hospital Human Research Ethics Committee (#3174), and the CBRG cohort was approved by The University of Adelaide Human Research Ethics Committee (#H-2013–097). Research at the J. Craig Venter Institute was approved by the JCVI Institutional Review Board (#2013–182). All research was performed according to the listed institutions guidelines and informed consent was obtained from all participants' parent and/or legal guardians. Inclusion criteria included 5 to 11-y-old twins whose parents consented to this portion of the study. Our protocol samples the supragingival plaque of all teeth in the oral cavity during sample collection regardless of whether a tooth is suspected of containing a cavity. Although this yields a mixture of caries and caries-free communities, it provides a powerful opportunity to develop diagnostic "at-home" tests where samples would be collected by patients. Our quality-controlled cohort consists of 36 caries and 51 caries-free samples sampled in this method. Please refer to *Supplemental Methods* for detailed descriptions on study design and sample collection implemented in this study.

Bioinformatics and data analysis

Please refer to *Supplemental Methods* for detailed descriptions on computational and analytical methodologies implemented in this study. Schematics for metagenomics workflows and sample metadata are detailed in Fig. S1 and Table S1, respectively. Metagenomic and metatranscriptomic workflows were performed using early versions of the *VEBA* software suite (51).

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Acronyms

MAG—Metagenome-assembled genome
SLC—Species-level cluster
CPR—Candidate phyla radiation
PSCN—Phenotype-specific coexpression network
PGFC—Phylogenomic functional category
HCPC—High-connectivity PSCN cluster
DCN—Differential coexpression network
HCDC—High-connectivity DCN cluster
LFOCV—Leave family out cross-validation

Funding

The research in this publication was supported by the National Institute of Dental and Craniofacial Research of the National Institutes of Health under award number R01DE019665 to KEN, and 1R01AI170111-01, NSF OCE-1558453, NSF OCE-2049299, and P01AI118687 to CLD. PETS was supported by grants from the Australian National Health and Medical Research Council (grant numbers 437015 and 607358 to J.M.C. and R.S.), the Bonnie Babes Foundation (grant number BBF20704 to J.M.C.), the Financial Markets Foundation for Children (grant no. 032-2007 to J.M.C.), and the Victorian Government's Operational Infrastructure Support Program. CBRG was supported by grants from the Australian National Health and Medical Research Council (grant numbers 349448 and 1006294 to T.H.) and the Financial Markets Foundation for Children (grant no. 223-2009 to T.H.).

Authors' Contributions

Conceptualization: J.L.E., C.L.D., K.E.N., T.H., and J.M.C.; Data curation: J.L.E.; Formal analysis: J.L.E.; Funding acquisition: K.E.N., T.H., R.S., and J.M.C., Sample collection: M.B., T.H., R.S., and J.M.C.; DNA and RNA extractions: M.T. and C.K.; Methodology: J.L.E. and C.L.D.; Project administration: C.L.D. and K.E.N., Resources: C.L.D., K.E.N., and M.T.; Software: J.L.E.; Supervision: C.L.D., K.E.N., and S.S.; Validation: J.L.E. and C.L.D.; Visualization: J.L.E.; Writing—original draft preparation: J.L.E. and C.L.D.; and Writing—review and editing: J.L.E., C.L.D., and P.L.

Data Availability

Metatranscriptomes and metagenomes were deposited in NCBI under BioProject PRJNA383868. Counts tables (ORF, orthogroup, MAG), genomes, gene models, and annotations are available on FigShare (doi: 10.6084/m9.figshare.18180614). Reproducible methods for feature engineering and network analysis are available at https://github.com/jolespin/ensemble_networkx.

References

1. James SL, *et al.* 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 392: 1789–1858.
2. Dye BA, Li X, Beltran-Aguilar ED. 2012. Selected oral health indicators in the United States, 2005–2008. NCHS Data Brief. 96: 1–8.
3. Peres MA, *et al.* 2019. Oral diseases: a global public health challenge. Lancet. 394:249–260.

4. OECD. 2017. Health at a Glance 2017: OECD Indicators. Paris: OECD Publishing. https://doi.org/10.1787/health_glance-2017-en
5. Humphrey LT, *et al.* 2014. Earliest evidence for caries and exploitation of starchy plant foods in Pleistocene hunter–gatherers from Morocco. Proc Natl Acad Sci USA. 111(3):954–959.
6. Marsh PD. 1994. Microbial ecology of dental plaque and its significance in health and disease. Adv Dent Res. 8(2):263–271.
7. Takahashi N, Nyvad B. 2011. The role of bacteria in the caries process: ecological perspectives. J Dent Res. 90(3):294–303.
8. Kleinberg I. 2002. A mixed-bacteria ecological approach to understanding the role of the oral bacteria in dental caries causation: an alternative to Streptococcus mutans and the specific-plaque hypothesis. Crit Rev Oral Biol Med. 13(2):108–125.
9. Marsh PD, Bradshaw DJ. 1997. Physiological approaches to the control of oral biofilms. Adv Dent Res. 11:176–185.
10. Folke C, *et al.* 2004. Regime shifts, resilience, and biodiversity in ecosystem management. Annu Rev Ecol Evol Syst. 35: 557–581.
11. Nyvad B, Takahashi N. 2020. Integrated hypothesis of dental caries and periodontal diseases. J Oral Microbiol. 12:1710953.
12. Smith SR, *et al.* 2016. Transcriptional orchestration of the global cellular response of a model pennate diatom to diel light cycling under iron limitation. PLoS Genet. 12(12):e1006490.
13. Gomez A, *et al.* 2017. Host genetic control of the oral microbiome in health and disease. Cell Host Microbe. 22(3):269–278.e3.
14. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. 2015. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol. 11(3):e1004075.
15. Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9(1):559.
16. Erb I, Notredame C. 2016. How should we measure proportionality on relative gene expression data? Theory Biosci. 135(1–2):21–36.
17. Quinn TP, Richardson MF, Lovell D, Crowley TM. 2017. Propr: an R-package for identifying proportionally abundant features using compositional data analysis. Sci Rep. 7(1):1–9.
18. Morton JT, *et al.* 2017. Balance trees reveal microbial niche differentiation. mSystems. 2(1):e00162–e00116.
19. Quinn TP, Erb I, Richardson MF, Crowley TM. 2018. Understanding sequencing data as compositions: an outlook and review. Bioinformatics. 34(16):2870–2878.
20. Espinoza JL, Shah N, Singh S, Nelson KE, Dupont CL. 2020. Applications of weighted association networks applied to compositional data in biology. Environ Microbiol. 22(8) :3020–3038.
21. Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S. 2007. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome. 18(6–7): 463–472.
22. Altman N, Krzywinski M. 2018. The curse(s) of dimensionality. Nat Methods. 15(6):399–400.
23. Van Der Maaten L. 2014. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res. 15:1–21.
24. McInnes L, Healy J, Melville J. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. https://doi.org/10.48550/arXiv.1802.03426 [accessed date 9/10/2019]
25. Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 10(12):1200–1202.
26. Silverman JD, Roche K, Mukherjee S, David LA. 2020. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J. 18: 2789–2798.
27. Espinoza JL, *et al.* 2018. Supragingival plaque microbiome ecology and functional potential in the context of health and disease. mBio. 9(6):e01631.
28. Shaiber A, Eren AM. 2019. Composite metagenome-assembled genomes reduce the quality of public genome repositories. mBio. 10(3):e00725–19.
29. Blondel VD, Guillaume J.-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008(10):P10008.
30. Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 9:1–12.
31. Jackson MA, *et al.* 2018. Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. PeerJ. 6(2):e4303.
32. Wilson SJ, Wilkins AD, Lin CH, Lua RC, Lichtarge O. 2017. Discovery of functional and disease pathways by community detection in protein–protein interaction networks. Pac Symp Biocomput. 22:336–347.
33. XH Y, M R. 2021. Multi-label classification and label dependence in in silico toxicity prediction. Toxicol In Vitro. 74:105157.
34. Zheng F, Zhang S, Churas C, Pratt D, Bahar I, Ideker T. 2021. HiDeF: identifying persistent structures in multiscale ‘omics data. Genome Biol. 22(1):1–15.
35. Hsu C-L, Juan H-F, Huang H-C. 2015. Functional analysis and characterization of differential coexpression networks. Sci Rep. 5(1):13295.
36. Morton JT, *et al.* 2019. Establishing microbial composition measurement standards with reference frames. Nat Commun. 10(1):1–11.
37. Espinoza JL, *et al.* 2021. Predicting antimicrobial mechanism-of-action from transcriptomes: a generalizable explainable artificial intelligence approach. PLoS Comput Biol. 17(3):e1008857.
38. Nabwera HM, *et al.* 2021. Interactions between fecal gut microbiome, enteric pathogens, and energy regulating hormones among acutely malnourished rural Gambian children. EBioMedicine. 73:103644.
39. Chen L, *et al.* 2020. Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. Nat Commun. 11(1):1–12.
40. Yang X, He L, Yan S, Chen X, Que G. 2021. The impact of caries status on supragingival plaque and salivary microbiome in children with mixed dentition: a cross-sectional survey. BMC Oral Health. 21(1):1–13.
41. BJ K *et al.* 2008. Pyrosequencing analysis of the oral microflora of healthy adults. J Dent Res. 87(11):1016–1020.
42. Liljemark WF, *et al.* 1984. Distribution of oral Haemophilus species in dental plaque from a large adult population. Infect Immun. 46(3):778–786.
43. Zaura E, Keijser BJ, Huse SM, Crielaard W. 2009. Defining the healthy “core microbiome” of oral microbial communities. BMC Microbiol. 9:259.
44. Amat S, Lantz H, Munyaka PM, Willing BP. 2020. Prevotella in pigs: the positive and negative associations with production and health. Microorganisms 8: 1–27.
45. N UT *et al.* 2018. Performance of mass spectrometric identification of clinical Prevotella species using the VITEK MS system: a prospective multi-center study. Anaerobe. 54:205–209.
46. Krzywinski M, Birol I, Jones SJ, Marra MA. 2012. Hive plots—rational approach to visualizing networks. Brief Bioinform. 13(5):627–644.

47. Takahashi N, Washio J, Mayanagi G. 2010. Metabolomics of supragingival plaque and oral bacteria. J Dent Res. 89(12):1383–1388.
48. Kanasi E, *et al.* 2010. Clonal analysis of the microbiota of severe early childhood caries. Caries Res. 44(5):485.
49. Cherkasov SV, *et al.* 2019. Oral microbiomes in children with asthma and dental caries. Oral Dis. 25(3):898–910.
50. Freire M, *et al.* 2020. Longitudinal study of oral microbiome variation in twins. Sci Rep. 10(1):7954.
51. Espinoza JL, Dupont CL. 2022. VEBA: a modular end-to-end suite for in silico recovery, clustering, and analysis of prokaryotic, microeukaryotic, and viral genomes from metagenomes. BMC Bioinformatics. 23(1):1–36.