Evaluating the ecological hypothesis: Early life salivary microbiome assembly predicts dental caries in a longitudinal case-control study

Freida Anne Blostein (blostein@umich.edu)
University of Michigan School of Public Health https://orcid.org/0000-0002-5734-4969
Deesha Bhaumik
University of Michigan School of Public Health
Elyse Davis
University of Michigan School of Public Health
Elizabeth Salzman
University of Michigan School of Public Health
Kerby Shedden
University of Michigan School of Public Health
Melissa Duhaime
University of Michigan
Kelly M Bakulski
University of Michigan School of Public Health
Daniel W McNeil
West Virginia University
Mary L Marazita
University of Pittsburgh School of Dental Medicine
Betsy Foxman
University of Michigan School of Public Health https://orcid.org/0000-0001-6682-238X

Research Article

Keywords: Oral microbiome, early childhood, ecological hypothesis, early childhood caries, 16S rRNA gene, whole genome shotgun metagenomics

Posted Date: February 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-848589/v1

Abstract

Background

Early childhood caries (ECC) – dental caries (cavities) occurring in primary teeth up to age 6-years - is a prevalent childhood oral disease that includes a microbial etiology. *Streptococcus mutans* was previously considered a primary cause, but recent research promotes the ecologic hypothesis, in which a dysbiosis in the oral microbial community leads to caries. In this incident density sampled case control study of 189 children followed from 2-months to 5-years, we use the salivary bacteriome to 1) prospectively test the ecological hypothesis in salivary bacteriome communities and 2) identify co-occurring salivary bacterial communities predicting future ECC.

Results

Supervised classification of future ECC case status using salivary samples from age 12-months using bacteriome-wide data (AUC-ROC 0.78 95% CI: (0.71–0.85)) improves prediction of future ECC status over using *S. mutans* amplicon abundance alone (AUC-ROC 0.44 (0.35–0.53)). Dirichlet multinomial community state typing and co-occurrence network analysis identified similar robust and replicable groups of co-occurring taxa. Mean relative abundance of a *Haemophilus parainfluenzae/Neisseria/Fusobacterium periodonticum* group was lower in future ECC cases (0.14) than controls (0.23, *P* value < 0.001) in pre-incident visits, positively correlated with saliva pH (Pearson rho = 0.31, *P* value < 0.01) and reduced in individuals who had acquired *S. mutans* by the next study visit (0.12) versus those who did not (0.19, *P* value < 0.01). In a subset of whole genome shotgun sequenced samples, case plaque had higher abundances of antibiotic production and resistance gene orthologs, including a major facilitator superfamily multidrug resistance transporter (MFS DHA2 family P_{BH} value = $1.9*10^{-28}$), lantibiotic transport system permease protein (P_{BH} value = $6.0*10^{-6}$) and bacitracin synthase I (P_{BH} value = $5.6*10^{-6}$). The oxidative phosphorylation KEGG pathway was enriched in case plaque (P_{BH} value = $1.2*10^{-8}$), while the ABC transporter pathway was depleted (P_{BH} value = $3.6*10^{-3}$).

Conclusions

Early-life bacterial interactions predisposed children to ECC, supporting a time-dependent interpretation of the ecological hypothesis. Bacterial communities which assemble before 12-months of age can promote or inhibit an ecological succession to *S. mutans* dominance and cariogenesis. Intragenera competitions and intergenera cooperation between oral taxa may shape the emergence of these communities, providing points for preventive interventions.

Background

In 2015–2016, 21% of US children aged 2–5 years showed evidence of early childhood caries (ECC), i.e. at least one primary tooth with one or more decayed, missing or filled tooth surfaces (1, 2). ECC can be painful, may negatively impacts self-esteem, and is a strong predictor of future oral health problems (3, 4). Microbial digestion of carbohydrates to acids which erode tooth enamel is the proximate cause (5–7). Acid-producing bacteria, particularly *Streptococcus mutans* (*S. mutans*), are frequently associated with ECC (5, 8). No single bacteria species, however, has been conclusively identified as a necessary and sufficient cause of ECC across human populations (5, 8, 9). Recent research emphasizes the ecologic hypothesis, which posits that overall shifts in the composition, structure, functional potential of the oral microbial community leads to dental decay (5, 10). However, few studies of ECC have prospectively and directly tested the ecologic hypothesis.

To assess the bacterial community in saliva and plaque samples, 16S rRNA gene amplicon sequencing simultaneously measures the bacterial taxa present (11–15). However, common methods for analyzing 16S rRNA gene data fail to capture the spirit of the ecological hypothesis. Estimating the effect of each identified taxa as an independent predictor ignores how bacteria interact to affect risk, which is a key component of the ecological hypothesis (5, 16). Diversity metrics, such as alpha and beta diversity, conveniently and efficiently summarize information across all measured taxa, but findings on associations between diversity metrics and cariogenesis are mixed (17–21). The lack of consistency may be attributed to differences in study design, conduct and analysis, but also may reflect the inherent limitations of diversity metrics. These metrics ignore taxonomic, ecologic, and functional differences between bacteria which can impact disease processes such as cariogenesis (22). *Common methods for analyzing 16S rRNA gene data do not adequately encapsulate the ecologic hypothesis.*

Microbial communities and ecologies are dynamic, and early childhood is a susceptible life-period for short- and long-term oral microbial community assembly. The oral microbiome is acquired after birth and influenced by environmental factors (11, 12, 23). Very few studies have prospectively tested the effect of oral microbial community assembly on ECC risk. A 2019 Australian study of 134 children followed for 5 years noted a shift in salivary microbiome composition at 39 and 48.6 months of age associated with future ECC (14). Microbial taxa, including *Streptococcus sobrinus* and *Scardovia wiggsiae*, were identified as potential biomarkers of ECC onset. The percentage of *S. mutans* in saliva was the best prospective predictor of future ECC (13, 14). The authors concluded, however, that the magnitude of change in the salivary bacteriome was inadequate to differentiate between health and disease at clinical levels. A smaller 2020 study of 56 children aged 1–3 years followed for 2-years demonstrated that the early life salivary bacteriome could prospectively classify future ECC onset (area under the receiver operating curve = 0.71) and identified several taxa that may serve as biomarkers of ECC (15). These studies prospectively link community-wide shifts in the early-life salivary microbiome to ECC. However, they did not evaluate how co-occurrence or functional interactions between taxa influence ECC risk. *Few longitudinal cohorts have explicitly evaluated the ecological hypothesis and ECC, and none have identified co-occurring groups of oral bacteria that influence ECC risk.*

To understand the influence of oral microbial community assembly on future oral health, explicit tests of the ecological hypothesis and identification of influential microbial communities is required. We used a longitudinal cohort of children to: 1) prospectively test the ecological hypothesis in salivary bacteriome communities and 2) identify co-occurring salivary bacterial communities influencing the risk of future ECC. We performed 16S rRNA gene amplicon sequencing on 855 longitudinal saliva samples from 99 children with ECC and 90 incidence-density sampled control children followed from 2-months to 5-years of age. We show that bacteriome-wide taxonomic information at 12-months of age better classifies future ECC status than *S. mutans* amplicon abundance alone. We identify robust and replicable communities of co-occurring bacteria using unsupervised clustering techniques, including a protective community of *Neisseria/Haemophilus parainfluenzae/Fusobacterium periodonticum* which was less abundant in future ECC cases. Finally, we comment on ecological and functional interactions that may shape the assembly of these communities using clinical data and functional potential measurements from a subcohort with shotgun metagenonomic sequencing data.

Results

Description of cohort

We analyzed 99 ECC case children and 90 matched control children who were free of dental lesions at the age of their matched case's diagnosis. Of the 189 children, 169 were White and 20 were bi- or multi-racial, 100 were from West Virginian and 89 from Pennsylvanian, and 97 male and 92 female. None of these characteristics differed between cases and controls (Additional file 1). Among the 99 ECC case children, the youngest age of diagnosis was 12-months, with a mean age of diagnosis of 38 months. We sequenced the V4 16S rRNA gene region in saliva samples from the visit corresponding to ECC diagnosis (incident visit) and all preceding visits (non-incident visits) for case and control children (Fig. 1, Figure S1-2 in Additional File 2, Additional File 3). From the 855 saliva samples across all incident and non-incident visits, we identified 3194 amplicon sequence variants (ASVs). We labeled ASVs that did not classify to the species level with ASV numbers. Alpha diversity of the salivary microbiome increased as children aged. Alpha diversity was inconsistently associated with future ECC diagnosis across visits (Table 1).

Table 1
Associations between future early childhood caries and salivary microbiome measures among non-incident children from Appalachia

	~ 2 month visit[a]			~ 12 month visit[a]			~ 24 month visit[a]		
	Case N = 92	Control N = 85	p	Case N = 84	Control N = 76	p	Case N = 70	Control N = 65	p
Shannon index (mean (sd))	2.05 (0.53)	1.85 (0.54)	0.01	2.87 (0.45)	3.01 (0.37)	0.04	3.49 (0.34)	3.52 (0.38)	0.71
Chao1 (mean (sd))	31.8 (12.3)	27.8 (10.0)	0.02	57.4 (16.8)	62.7 (14.6)	0.04	86.6 (20.1)	88.1 (19.2)	0.66
S. mutans abundance (mean relative abundance (sd))	0.00 (0.00)	0.00 (0.00)		0.00 (0.00)	0.00 (0.00)	0.28	0.01 (0.02)	0.00 (0.00)	0.41
S. mutans ASV detected (N (%))			1.000			0.12			<0.001
Yes	90 (97.8%)	83 (97.6%)		78 (92.9%)	75 (98.7%)		51 (72.9%)	64 (98.5%)	
No	2 (2.17%)	2 (2.35%)		6 (7.14%)	1 (1.32%)		19 (27.1%)	1 (1.5%)	
Community state type (N (%))[b]			0.28			<0.001			<0.001
Streptococcus ASV1 *dominated w/ Gemella* ASV2	44 (48.4%)	51 (60.7%)		3 (3.7%)	1 (1.3%)		0 (0.0%)	0 (0.0%)	
Gemella ASV2 - H. parainfluenzae - Neisseria ASV9	3 (3.30%)	1 (1.19%)		28 (34.6%)	53 (69.7%)		0 (0.00%)	6 (9.23%)	
H. parainfluenzae - Neisseria ASV9	0 (0.00%)	0 (0.00%)		2 (2.5%)	7 (9.2%)		24 (34.8%)	43 (66.2%)	
Streptoccous ASV1 *dominated w/ G. elegans*	43 (47.3%)	30 (35.7%)		12 (14.8%)	6 (7.9%)		0 (0.0%)	0 (0.0%)	
Streptococcus ASV8 - *Neisseria* ASV12	1 (1.10%)	2 (2.38%)		34 (42.0%)	9 (11.8%)		15 (21.7%)	6 (9.2%)	
Neisseria ASV12 – *Veillonella* ASV5	0 (0.00%)	0 (0.00%)		2 (2.47%)	0 (0.0%)		31 (44.3%)	10 (15.4%)	
Network modules (summed module relative abundance mean[c] (sd))									
Streptococcus ASV1 *& Neisseria* ASV12 *network*	0.03 (0.06)	0.04 (0.08)	0.15	0.30 (0.15)	0.28 (0.13)	0.26	0.50 (0.15)	0.51 (0.16)	0.80
Haemophilus parainfluenzae & Neisseria ASV9 *network*	0.42 (0.27)	0.31 (0.25)	0.006	0.14 (0.13)	0.23 (0.13)	<0.001	0.14 (0.09)	0.18 (0.09)	0.01
Veillonella ASV5 *& Streptococcus* ASV8 *network*	0.09 (0.09)	0.10 (0.12)	0.33	0.21 (0.19)	0.13 (0.12)	0.001	0.13 (0.13)	0.08 (0.10)	0.01
Veillonella sp._HMT_780 & Granulicatella elegans network	0.27 (0.17)	0.37 (0.21)	<0.001	0.30 (0.17)	0.31 (0.17)	0.62	0.15 (0.14)	0.17 (0.12)	0.25
Gemella ASV2 *& Leptotrichia shahii network*	0.20 (0.20)	0.18 (0.18)	0.39	0.06 (0.05)	0.05 (0.03)	0.32	0.08 (0.06)	0.05 (0.04)	0.01

[a]Includes duplicate records for 1 child selected as a control at 36 months and a case at 60 months, and 1 child selected as a control at both 36 and 60 months. Excludes samples from children diagnosed as a case at that visit and their corresponding risk-set controls (N = 6 at 12-months, N = 37 at 24-months).

[b]Each sample is assigned to the community state type for which it had the highest posterior probability, samples with no community state type with posterior probability > 80% were label as unassigned. 1 control sample and 1 case sample were unassigned at 3-month visit, 3 case samples unassigned at 12-month visit.

[c]Summed module relative abundance calculated by summing the relative abundance of all ASVs assigned to the network module together, ranges from 0–1

S. mutans did not associate with future ECC diagnosis before 24-months of age, but was elevated in cases at the visit of first ECC diagnosis

A single ASV identified as *S. mutans.* We validated the identity of this ASV using BLAST and shotgun metagenomic sequencing data (Additional File 4; Figure S3 Additional File 2). At the 2- and 12-month visits, *S. mutans* was rare and not associated with future ECC diagnosis (Table 1). By the 24-month visit, *S. mutans* was more prevalent in future cases (Table 1; *P* value < 0.001). *S. mutans* prevalence and abundance was elevated in cases at the visit of ECC diagnosis: 13 of 19 ECC cases diagnosed at 24-months had *S. mutans* at the 24-month visit vs 2 of 18 matched controls (Additional File 5; *P* value = 0.001).

At 12- and 24-months of age, supervised random forest using the salivary bacteriome is more accurate at predicting ECC case status than S. mutans alone

To test the ecological hypothesis, we used a 5-repeat, 10-fold cross-validated random forest classifier of future ECC status. Separate classifiers were built for the 12- and 24-month visits. Samples from cases first diagnosed with ECC at those visits (and their matched controls) were excluded. We compared classification performance using only *S. mutans* to using 273 ASVs present in > 5% of samples (see Methods) as features. Classification using only *S. mutans* was poor at both the 12-month (AUC (95% CI): 0.44, (0.35–0.53)) and 24-month visits (AUC (95% CI): 0.55, (0.45–0.65)) (Fig. 2A).

Using 273 ASVs more accurately classified future ECC status at the 12-month (AUC (95% CI): 0.78, (0.71–0.85)) and 24-month visits (AUC (95% CI): 0.72, (0.63–0.81)) (Fig. 2A). Important features were consistent across visits and included *Neisseria*, *Rothia*, and *Prevotella* ASVs (Fig. 2B&C). Two *Streptococcus* ASVs were important features but were not *S. mutans*; *Streptococcus* ASV8 was likely *Streptococcus salivarius*, and while *Streptococcus ASV14* was closely related to *Streptococcus lactarius/peroris* (Additional file 4; Figure S3 in additional file 2).

Unsupervised clustering techniques identify similar groups of co-occurring taxa, which associate with ECC

Next, we attempted to identify ecologically meaningful groups of co-occurring taxa. To do so, we used two different unsupervised clustering techniques.

Using Dirichlet multinomial community state typing, we identified 6 community state types (CSTs) (Figure S4, additional file 2). CSTs were named after the ASVs defining their separation. Each sample was assigned to a single CST. As children aged, they transitioned from *Streptococcus*-dominated to more diverse CSTs (Fig. 3A, Figure S5 in additional file 2). *Hemophilus parainfluenzae* and *Neisseria* ASV9 CSTs were less prevalent in cases while *Streptococcus* ASV8, *Neisseria* ASV12 and *Veillonella* ASV5 CSTs were more prevalent (Table 1, additional file 5).

Using weighted co-occurrence network analysis, we identified five network modules of co-occurring ASVs. Network modules were named after the top two most abundant ASVs in the network (Fig. 3B; Figures S6-7 additional file 2; additional file 6). A *Haemophilus parainfluenzae* and *Neisseria* ASV9 network module was more abundant in controls; a *Veillonella* ASV5 and *Streptococcus* ASV8 network module was more abundant in cases (Fig. 3B; Table 1). Important ASVs for defining network modules and CSTs overlapped (Additional file 2 Figure S8).

The networks identified highly central taxa, such as *Fusobacterium periodonticum*, the central taxa in the *Haemophilus parainfluenzae/ Neisseria* ASV9 network (Fig. 2C). *Fusobacterium periodonticum* peaked in abundance at 12-months of age and was more abundant in controls (Fig. 4). In contrast, *Fusobacterium nucleatum subsp animalis* and *S. mutans*, closely associated members of the *Streptococcus* ASV1-*Neisseria* ASV12 network module, increased in abundance after 12-months of age and were more abundant in cases (Fig. 4; Figure S8 in additional file 2).

The ECC-associated communities identified through unsupervised clustering were robust to varying hyperparameters. In the CST analysis, we varied the number of k CSTs (k = 4 vs 5 vs 6, additional file 7). In the network analysis, we varied the normalization transform function (Hellinger vs central-log, Figure S9-10).

Communities identified through unsupervised clustering are reproducible in an external cohort

To examine the reproducibility of these bacterial communities, we performed the same analytic pipeline (see Methods) on publicly available 16S rRNA gene sequencing data from longitudinal saliva samples of similarly aged children (*Holgerson et al.*; PRJEB35824; (11)). We were unable to obtain access to metadata for these samples.

A *Haemophilus parainfluenzae/Neisseria perflava* network module with central taxa *Fusobacterium periodonticum* was also identified in the Holgerson *et al.* sample (additional file 2 Figures S11-12). The *Neisseria* ASV9 amplicon from our cohort was closely related to the *Neiseria perflava* amplicon from the Holgerson *et al.* cohort (additional file 2 Figure S13A).

A similar *Veillonella dispar/Streptococcus/Prevotella* network module was also identified in the Holgerson *et al.* sample. The *Veillonella* ASV5 amplicon from our cohort was closely related to the *Veillonella dispar* amplicon from the Holgerson *et al.* cohort (additional file 2 Figure S13B).

Early-life bacterial communities are associated with concurrent salivary pH, future S. mutans prevalence, and primary teeth count

We tested if bacterial communities from our unsupervised clustering associated with etiologically relevant variables in our cohort. Mean salivary pH was higher in samples in CSTs characterized by *H. parainfluenzae* and *Neisseria* ASV9 (12-month mean: 6.78; 24-month mean: 6.71) than in those characterized by *Streptococcus* ASV8, *Neisseria* ASV12 and *Veillonella* ASV5 (12-month: 6.54, Wilcoxon *P* value = 0.05; 24-month: 6.63, Wilcoxon *P* value = 0.03). Abundance of the *H. parainfluenzae/Neisseria* ASV9 network module was correlated with increasing salivary pH (12-month rho = 0.31, *P* value = 0.002; 24-month rho = 0.28; *P* value = < 0.001).

Children with communities characterized by *Streptococcus* ASV8, *Neisseria* ASV12, and *Veillonella* ASV5 at the 12- and 24-month visits were more likely to have *S. mutans* detected at their next visit than children with communities characterized by *Haemophilus parainfluenzae* and *Neisseria* ASV9 (percent with *S. mutans* at 24-months: 34% vs 9%, Fisher's exact *P* value = 0.03; at 36-months: 46% vs 21%, *P* value < 0.01). Children who acquired *S. mutans* by their next visit

had lower abundances of the *Haemophilus parainfluenzae/Neisseria* ASV9 network and higher abundances of the *Veillonella* ASV5-*Streptococcus* ASV8 network than children who did not go on to have *S. mutans* (additional file 2 Figure S14).

The average number of primary teeth present was higher in children assigned to CSTs which were more prevalent at later ages and the relative abundance of the *Streptococcus* ASV1-*Neisseria* ASV12 network correlated with the number of primary teeth present at the 12- and 24-month visits. This was not true for the *H. parainfluenzae/Neisseria* ASV9 network (Additional file 2 Figure S15). For Pennsylvania children, the approximate age at first tooth emergence was available but was not associated with CST nor network modules.

Whole-genome shotgun metagenomics of incident samples revealed functional differences between incident case- and control-samples

We tested for differences in the community composition and functional potential of cases and controls using saliva and plaque samples from the visit of case ECC diagnosis for 15 cases and 15 matched controls. Among others, *Scardovia wiggsiae, Prevotella histicola, Veillonella dispar, Streptococcus mutans and Streptococcus salivarius* were more abundant in case than matched control saliva and plaque samples at the time of diagnosis (Additional file 2 Figure S16; Additional file 8). *Prevotella salivae* was more abundant in case than matched control saliva but not plaque samples (Benjamini-Hochberg P_{BH} value < 0.05). The fungal genus *Candida* was only present in case plaque samples.

Cases and controls differed in the abundance of gene orthologs (Fig. 5). Associations with case status were stronger in plaque than saliva. Gene orthologs related to antibiotic production and resistance were more abundant in case plaque, including a major facilitator superfamily multidrug resistance transporter (P_{BH} value = $1.9*10^{-28}$) and lantibiotic transport system permease protein (P_{BH} value = $6.0*10^{-6}$) (Additional file 9). The oxidative phosphorylation KEGG pathway was enriched in case plaque (P_{BH} value = $1.2*10^{-8}$), while the ABC transporter pathway was depleted (P_{BH} value = $3.6*10^{-3}$, Additional file 10).

Functional differences sometimes reflected taxonomic differences. All the case-associated gene orthologs annotating to oxidative phosphorylation were found only in *Candida* (Additional file 11 and Figure S17 in additional file 2).

Discussion

Results of our analysis of 99 ECC cases and 90 incidence density matched children supports the ecologic hypothesis for ECC. We showed that bacteriome-wide information classified future ECC status better *S. mutans* alone. We expanded on previous work by identifying replicable groups of co-occurring bacteria, which may represent true ecological interactions. We showed that these groups associate with concurrent salivary pH, future *S. mutans* acquisition and future ECC diagnosis, suggesting an ecological succession to cariogenesis. By incorporating shotgun metagenomic sequencing data, we identified functional mechanisms for ecological interactions between bacteria, including pathways related to antibiotic production and resistance. Together, these observations suggest early-life bacterial interactions during a susceptible life period can predispose individuals to ECC.

Our findings on salivary bacteriome assembly and association with ECC fit within the previous literature. We observed a well-documented succession from *Streptococcus*-dominated, low-diversity communities to more diverse communities by 24-months of age with stabilization thereafter (11, 12, 24, 25). As in cross-sectional dental research, we found an association between *S. mutans* and ECC at the time of ECC diagnosis (8). Like previous prospective studies of ECC, we found evidence for an association between early life salivary bacteriome composition and future ECC (14, 15). We were able to distinguish ECC cases from controls more accurately and at an earlier age than reported by Dashper *et al.*, while the AUC-ROC for our 12-month random forest (0.78) is close to that of Grier *et al.* (0.71) (14, 15). While *S. mutans* was elevated in cases at diagnosis, we found that the prospective predictive ability of the salivary bacteriome was superior to that of *S. mutans* alone at 12- and 24-months. This supports a time-dependent interpretation of the ecological hypothesis, in which dysbiosis in the oral microbial community precedes salivary *S. mutans* detection, a marker of late-stage cariogenesis. Our findings highlight the first two years of life as a susceptible period for assembly of a cariogenic oral microbial community.

Unlike previous work, we identified specific ECC-associated bacterial communities using unsupervised clustering techniques. These unsupervised techniques better encapsulate the ecological hypothesis than diversity metrics, which may be too coarse to summarize finer level differences in communities (22). In our cohort, alpha diversity was weakly and inconsistently associated with future ECC status, echoing previous mixed findings (17–21). In contrast, groups of taxa from unsupervised clustering techniques were strongly and prospectively associated with ECC: a *Haemophilus parainfluenzae, Neisseria*, and *Fusobacterium periodonticum* community was depleted in cases while a *Prevotella, Streptococcus* and *Veillonella* community was more abundant. These bacterial communities were consistent across clustering methods, reproducible in an external cohort (11), and in line with previous work. A 2014 study of longitudinal tongue samples found strong correlations between abundances of *Haemophilus parainfluenzae* and *Neisseria subflava* (26). A cross-sectional 2017 analysis of young adult saliva found similar *Haemophilus parainfluenzae/ Neisseria subflava* and *Veillonella/Prevotella* clusters (27). A 2021 co-occurrence analysis of cross-sectional adult saliva samples found a similar network module of *Prevotella salivae, Veillonella atypica* and *Streptococcus salivarius* (28). Our unsupervised clusters were also separated by genetically distinct sequence variants of *Neisseria*, *Veillonella* and *Fusobacterium*. Intragenera competition and strain-specific impacts on biofilm formation have been observed for *Neisseria* (29, 30), *Veillonella* (31, 32) and *Fusobacterium* species in previous work (33–35). Groups of ECC-associated taxa identified from our unsupervised clustering may reflect ecological roles and interactions, including intergenera cooperation and intragenera competition.

We also tested how these bacterial communities were associated with etiologically relevant variables. The *Streptococcus* ASV1- *Neisseria* ASV12 network was correlated with the number of primary teeth present. Members of this network, including *S. mutans* and *Streptococcus sobrinus*, are known to have a preference for hard oral surfaces, increasing in abundance after tooth emergence (12). The protective *Haemophilus parainfluenzae, Neisseria*, and *Fusobacterium periodonticum* network was correlated with salivary pH and inversely associated with future *S. mutans* detection. In a recent *in vitro* study, *Neisseria* was positively correlated with salivary pH (36). The associations of our identified bacterial communities with etiologically relevant variables provides

further evidence that the communities are biologically meaningful. Moreover, these associations link the early-life oral environment to an ecological succession towards cariogenesis.

Importantly, we identified differences in the functional potential of microbial communities at the time of ECC case diagnosis using shotgun metagenomic sequencing. ECC cases and controls differed in the abundance of gene orthologs annotating to KEGG pathways related to antibiotic production and resistance. This finding is supported by previous analyses of shotgun metagenomic sequencing and dental caries (37, 38). In our study, case-enriched gene orthologs included those involving competition between related species, such as bacteriocin exporters (39) and lantibiotic production (40). These functions may represent mechanisms for the co-occurrences and potential interactions we observed between oral taxa.

The tissue of measurement for assessment of the oral microbiome is an important consideration. We performed 16S rRNA gene sequencing on longitudinal saliva samples, and shotgun metagenomic sequencing on a subsample of cross-sectional plaque and saliva samples. Plaque, not saliva, is the most proximate tissue in cariogenesis. However, plaque is difficult to collect from edentulous children, has a low biomass, and is unlikely to be used as a prognostic marker in a clinical setting. Thus, the predictive power of the early-life salivary microbiome is of practical, clinical interest. Saliva is also a composite tissue and washes over many oral surfaces with different microbial communities (7, 41–43). Therefore, differences in saliva bacteriome composition between cases and controls could reflect differences in the bacterial abundance of oral surfaces rather than changes in only the salivary bacteriome composition. Consequently, the co-occurrence patterns we identified may reflect niche-sharing of oral surfaces rather than cooperation between taxa. Although ECC-associated communities did not associate with proxies for soft-to-hard tissue ratio (tooth number; age at first tooth emergence) we cannot conclusively rule out this explanation. While having both saliva and plaque samples in the shotgun metagenomic sequencing subsample is a strength, our analysis is limited by not including longitudinal plaque samples.

Sequencing methods influence the inferences that are possible from the data. The V4 region of the 16S rRNA gene is limited in ability to resolve fine-level taxonomic differences. This could affect the identification and measurement of *Streptococcus* amplicons in our dataset. We validated the identity of *Streptococcus* amplicons using BLAST and shotgun metagenomic data, but nondifferential exposure misclassification of *S. mutans* prevalence is possible. The 16S rRNA gene also does not measure virus, eukaryotes, or interspecies functional variation. Without longitudinal shotgun metagenomic data, we cannot comment on virome or mycobiome assembly, or longitudinal changes in functional potential. Both 16S rRNA gene and shotgun metagenomic sequencing data is inherently compositional. We instituted transformations to address compositionality but did not have absolute abundance data. While our use of both 16S rRNA gene and shotgun metagenomic sequencing is a strength, our measurement of the oral microbiome is limited by these characteristics of sequencing methods.

Our study design is observational, so causality cannot be conclusively proved. However, our exposure measurements precede our outcome, fulfilling a key causal requirement. Our study population was primarily White children from northern and north central Appalachia. Although some of the unsupervised clusters from our cohort were replicable in the Swedish Holgerson *et al.* cohort, microbial communities can differ by geography, race, and ethnicity. Thus, the generalizability of our findings may also be limited. Further studies in additional populations, incorporating shotgun metagenomic sequencing, quantification of absolute bacterial load, and site-specific measures of oral bacterial communities are warranted.

Conclusions

We found that the early-life salivary microbiome associated with risk of ECC before *S. mutans* could be detected, supporting a time-dependent interpretation of the ecological hypothesis. Our analysis is strengthened by a longitudinal design, balanced case-control ratios, incorporation of both amplicon and shotgun metagenomic sequencing, and replication analyses. Our observations on the suitability of diversity measures vs other clustering techniques to detect fine scale differences are applicable in other microbial contexts. Our findings on ecological succession and bacterial interactions in early life may also be generalizable to other systems of microbiome development. Overall, our analyses support a developmental interpretation of the ecological hypothesis, and raise the possibility that intragenera cooperation, intergenera competition, and ecological successions in early life can predispose children to ECC.

Methods

Study cohort We used data from the Center for Oral Health Research in Appalachia 2 study (COHRA2) (44). COHRA2 recruited White, pregnant women between 2011 and 2015 from Pennsylvania and West Virginia. Healthy women who were in the 12th to 29th week of pregnancy, of European descent, over 18 years of old, fluent in English, and with a singleton pregnancy were eligible for inclusion. Women and their babies were followed longitudinally through the early years of the baby's life. Women were excluded if they had tuberculosis, were immunocompromised, thought they might soon leave the general regions of West Virginia or southwestern Pennsylvania, or did not have a reliable telephone contact. Mother-child pairs also were excluded from the study if the child was delivered before the 35th week of pregnancy or if the mother or child developed a serious medical condition.

Participants completed in-person visits when the child was 2-months and 12-months old, then yearly thereafter. Mother-child pairs from the Pennsylvania site had additional in-person visits at birth and when the child's first primary tooth erupted. At in-person visits mothers and children underwent a comprehensive dental assessment by a trained and calibrated dental professionals (training and calibration described in detail in Neiswanger *et al* (44)); participants were asked not to eat or drink for 2 hours prior to the examination. The examination included caries assessment via the PhenX Toolkit Dental Caries Experience Prevalence Protocol (http://www.phenxtoolkit.org/, protocol number 080300) which allows for the calculation of the decayed, missing, and filled tooth count to be calculated either including or excluding white spots. The dental examination also included collection of microbial samples from saliva, plaque and gingival swabs using OMNIgene Discover kits (OM-501 or 505 DNA Genotek); only saliva and plaque samples were used in this analysis. Saliva was collected via swabs for children too young to spit into a collection tube and via spitting otherwise. Pooled plaque samples were taken with a Stimudent or curette from three intact tooth surfaces (8-buccal, 24-buccal, 31-occlusal or nearby surfaces if these were not intact). Plaque is also taken from tooth surfaces with

untreated dental lesions. Salivary pH was also measured at visits where the child was old enough to spit (most by 12-months, all by 24-months), the Dentobuff strip was used to measure salivary pH in early visits.

A 30–45 minute telephone interview was administered to the mothers at approximately 6-month intervals to capture sociodemographic and behavioral data.

Sampling & case definition For this analysis we selected 99 children who had any dental lesions, including white spots (d1mft), at or prior to the 60-month visit as early childhood caries (ECC) cases. The visit in which a child was first identified as having a dental lesion was the incident-visit for that child. We then selected a similar number of children who were free of dental lesions at the same visit as the cases to serve as incidence-density sampled controls (n = 90). Incidence density sampling does not preclude the reselection of a control as a case at later time points; controls can also be selected as controls for multiple cases (Fig. 1) (45). In this analysis, one control was later selected as a case and one control was selected as a control twice (n = 92 control records). Duplicate records of the case/control and control/control children were not used in the supervised random forest, the case/control was only included as a case and the control/control was only included as a control once. In both unsupervised clustering techniques, we did not include duplicate records from these individuals when performing initial clustering but did include them when graphing and testing associations between identified clusters and variables of interest (i.e., in Table 1, Figs. 1–3). The number of total unique individuals in the analysis was 189, with 191 unique person-records (Additional file 3).

All available saliva samples from cases and controls, up to and including the incident-visit saliva sample, were pulled for 16S rRNA amplicon sequencing (Fig. 1). Selected individuals occasionally missed visits, did not have a saliva sample available or had a saliva sample which failed 16S amplicon quality control (Additional file 3). Additionally, we randomly selected a subcohort of 15 cases presenting with enamel lesions at or after the 36-month visit and 15 corresponding controls. Plaque and saliva samples from the visit of case diagnosis for these 30 individuals were submitted for shotgun metagenomic sequencing.

Laboratory and bioinformatics pipeline for 16S rRNA amplicon metagenomic sequencing Bacterial DNA was extracted from aliquots of saliva. Library preparation and sequencing of the 16S rRNA V4 amplicon was performed by the Michigan Microbial Systems Molecular Biology Laboratory using previously validated protocols (46). DNA extraction was performed using the Eppedorf EpMotion liquid handling system following the Qiagen MagAttract PowerMicrobiome kit protocol. The V4 variable region was amplified from extracted DNA using barcoded dual-index primers and sequenced on the Illumnia MiSeq platform using the MiSeq Reagent Kit V2 500 cycles. Each plate of samples was submitted with a positive mock community control, a DNA extraction kit control, and a negative water control (Additional file 2 Figures S3-4). Reads were processed to amplicon sequence variants (ASVs) using DADA2 (version 1.14.1) (47) and the Human Oral Microbiome Database (HOMD) version 15.2 (48). To identify contaminants, we used the R package decontam (version 1.8.0) (49). We filtered out samples with less than 1000 reads (n = 6 samples lost). Diversity metrics were calculated using the estimate_richness function from the R package phyloseq all ASVs. However, to limit the number of features used in supervised and unsupervised learning, we instituted a prevalence-abundance ASV filter. ASVs which were present in less than 5% of all samples *and* which represented less than 5% of all sequences in the samples in which they were present were excluded from the analytic subset for supervised random forest and unsupervised clustering techniques (m = 273 ASVs in analytic subset). ASVs were not collapsed at the genus or species level.

Random forest We used the 12- and 24-month visits as inputs for the random forest as these visits preserved a large subset of pre-incident samples. Only non-incident cases and matched controls were used in the random forest: individuals with incident-visit saliva samples were excluded (6 individuals who were identified as cases or controls at the 12-month visit and 37 individuals identified at the 24- month visit were excluded, total sample size of n = 158 and n = 133). Hellinger transformed ASV counts from the 273 ASVs in our analysis subset were used in the random forest. Using the train function in the R package caret (50), we ran 5 repeats of 10-fold cross validated random forest machine algorithms with 500 trees. We allowed the mtry parameter (number of parameters randomly sampled as candidates at each tree split) to be tuned from a choice of 2, 136, or 271 using the receiver operating characteristic curve; for both the 12-month and 24-month all taxa random forest an mtry parameter of 2 was selected. For the *S. mutans* only random forests at 12- and 24-months, only the abundance value of the ASV which identified as S. mutans was used as a feature in the random forests; therefore, the mtry parameter was 1. Area under the receiver operating curve and other evaluation statistics were calculated using the R package MLeval (51).

Dirichlet multinomial community state typing We used the R package DirichletMultinomial to cluster samples into community state types (CSTs) using Dirichlet multinomial mixture models (52). We fit ten Dirichlet multinomial models, using as input the count matrix of the 855 samples by 273 ASVs in the analytic subset and varying the number of Dirichlet components (i.e., CSTs) from 1 to 10. We calculated the Laplace measure of fit for each model and plotted against k, identifying $k = 6$ as the best model. We varied $k=\{4, 5\}$ as a sensitivity analysis. Samples were assigned to the single k CST for which they had the highest posterior probability of membership; if a sample assigned to no CST at a posterior probability > 80%, the sample was not assigned to any CST.

Weighted co-occurrence networks We used the R package WGCNA to build a signed weighted network of ASVs using Hellinger transformed count matrix of 855 samples and 273 ASVs (53). As a sensitivity analysis, we used the center-log ratio transformed count matrix instead. The soft thresholding power of the signed network was selected to maximize the R^2 of the model fit while preserving the mean connectivity of the network using the pickSoftThreshold function in WCGNA. We used a dynamic tree cut and the cutreeDynamic function in WCGNA to identify network modules or clusters using a minimum module size of 5 and a deep split value of 4, with the aim of producing more fine-grained clusters. Intramodular connectivity statistics were calculated for each ASV using the intramodularConnectivity function. Finally, per-sample module relative abundances were calculated by summing the relative abundances of all ASVs belonging to the same module.

Replication cohort We performed the exact same bioinformatics and analytic pipeline on publicly available V3-V4 16S rRNA gene data from the Holgerson cohort (PRJEB35824) (11), as we did to the COHRA2 samples. This cohort was also composed of sequential salivary samples from similarly aged children, the prevalence of ECC was 10% by 60 months of age. The laboratory methods for these samples are described in Holgerson *et al.* (11). All the bioinformatics parameters and steps were the same as described above, with the exception that decontam was not used to identify potential contaminants as the publicly available data did not include DNA quantitation data. Since we could not obtain access to any metadata characteristics of these samples, including ECC

status, the random forest models could not be run. For the community state typing, 6 Dirichlet components yielded the best model fit. To compare the relatedness of the amplicon sequence variants assigned to various network modules across the COHRA2 and Holgerson cohort, we performed multiple sequence alignment of the amplicons using the R packages msa, using the ClustalW algorithim (54). We computed pairwise distances from the DNA sequences using the r function dist.dml from the r package phangorn (55), using the JC69 model. We created a neighbor joining tree using the phangorn function NJ, then fit a generalized time-reversible with gamma rate maximum likelihood tree using the neighbor joining tree as a starting point. We obtained 100 bootstrap values for the tree using bootstrap.pml and plotted the tree using ggtree (56) and collapsed branches present in < 50 of the bootstrapped trees.

Laboratory and bioinformatics pipeline for shotgun metagenomic sequencing DNA was extracted from plaque and saliva samples using the Zymobiomics miniprep kit according to the manufacturer's instructions. Isolated DNA was quantified by Qubit. DNA libraries were prepared using the Illumina Nextera XT library preparation kit according to the manufacturer's protocol. Library quantity and quality was assessed with Qubit (ThermoFisher) and Tapestation (Agilent Technologies, CA, USA). Libraries were then sequenced on Illumina HiSeq platform 2x150bp. Quality filtering and adapter trimming were performed using Trimmomatic and the Nextera PE adapters. Host DNA was removed using bowtie2 and the GRCh38 index. Trimmed, cleaned and decontaminated reads were processed through both the Humann3 short-read profiling pipeline (57) and the SqueezeMeta assembly-based pipeline (version 1.4.0) (58). Plaque and saliva samples were run separately through the assembly pipeline. Briefly, assembly was done using Megahit, ORFs were predicted using Prodigal, and similarity searches against GenBank, eggnog and KEGG were conducted using Diamond. Read mapping against contigs was performed using Bowtie2. Binning was done using MaxBin2 and Metabat2 and bins were combined using DAS Tool. To test for differential abundance of KEGG orthologs and taxa abundance estimated from contigs, we used DESeq2, first filtering out KEGG or taxa with fewer than 500 reads from the testing subset. We tested for enrichment in KEGG pathways using gene set enrichment analysis and the R package fgsea separately on plaque and saliva samples. We used the package SQMTools to extract functional and taxonomic subsets of interest, such as the KEGG orthologs which annotated to oxidative phosphorylation. To test correlations between 16S rRNA gene amplicon sequence variants and abundances of taxa from whole genome sequencing, we used a partial spearman correlation while controlling for incident visit and case status.

Abbreviations

ECC
early childhood caries
ASV
amplicon sequence variant
16S rRNA
16S ribosomal RNA
CST
community state type

Declarations

Ethics approval and consent to participate The study has IRB approval from the University of Pittsburgh and West Virginia University. All potential participants have the study explained to them in detail and are sent copies of the consent forms before their initial appointments. At the first visit, the study is explained again, questions are answered, and the women sign consent forms prior to any research assessments.

Consent for publication Not applicable

Availability of data and materials The 16S rRNA gene amplicon sequencing data and shotgun metagenomic sequencing data from the COHRA2 study is publicly available at the PRJNA752888 repository. The 16S rRNA gene amplicon sequencing data from the Holgerson *et al.* replication cohort is publicly available at the PRJEB35824 repository. Phenotype data for the COHRA2 study are available at dbGaP phs001591.v1.p1 upon application. All of the code to reproduce the analyses in this paper is available at *https://github.com/blostein/ECCPaper1.*

Competing interests The authors declare that they have no competing interests

Funding This work was funded by the National Institutes for Health, National Institute for Dental and Craniofacial research grant R01 DE014899. FB was funded by the National Institutes for Health, National Institute for Dental and Craniofacial research grant F31 DE029992.

Authors contribution FB analyzed sequencing data and wrote the initial draft of the paper. DB provided code review. MM, BF, and DM were responsible for the design and collection of the cohort data. ES performed the laboratory preparation for sequencing. FB, BF, KB, ED, KS, and MD contributed to the conceptualization of the analysis and analytic plan. All authors read and approved the final manuscript.

Acknowledgments Not applicable

References

1. Fleming E, Afful J. Prevalence of Total and Untreated Dental Caries Among Youth: United States, 2015–2016. NCHS Data Brief. 2018 Apr;(307):1–8.
2. Statement on Early Childhood Caries [Internet]. American Dental Association. 2000 [cited 2021 Jun 8]. Available from: https://www.ada.org/en/about-the-ada/ada-positions-policies-and-statements/statement-on-early-childhood-caries.

3. Heilmann A, Tsakos G, Watt RG. Oral Health Over the Life Course BT - A Life Course Perspective on Health Trajectories and Transitions. In: Burton-Jeangros C, Cullati S, Sacker A, Blane D, editors. Cham: Springer International Publishing; 2015. p. 39–59. Available from: https://doi.org/10.1007/978-3-319-20484-0_3.
4. Martins-Júnior PA, Vieira-Andrade RG, Corrêa-Faria P, et al. Impact of Early Childhood Caries on the Oral Health-Related Quality of Life of Preschool Children and Their Parents. Caries Res [Internet]. 2013;47(3):211–8. Available from: https://www.karger.com/DOI/10.1159/000345534.
5. Pitts NB, Zero DT, Marsh PD, et al. Dental caries. Nat Rev Dis Prim. 2017;3:17030.
6. Gomez A, Nelson KE. The Oral Microbiome of Children: Development, Disease, and Implications Beyond Oral Health. Microb Ecol [Internet]. 2016/09/14. 2017 Feb;73(2):492–503. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27628595.
7. Mark Welch JL, Dewhirst FE, Borisy GG. Biogeography of the Oral Microbiome: The Site-Specialist Hypothesis. Annu Rev Microbiol [Internet]. 2019 Sep 8;73(1):335–58. Available from: https://doi.org/10.1146/annurev-micro-090817-062503.
8. Bhaumik D, Manikandan D, Foxman B. Cariogenic and oral health taxa in the oral cavity among children and adults: A scoping review. Arch Oral Biol. 2021 Jun;129:105204.
9. Fakhruddin KS, Ngo HC, Samaranayake LP. Cariogenic microbiome and microbiota of the early primary dentition: A contemporary overview. Oral Dis [Internet]. 2018 Jul 3;0(0). Available from: https://doi.org/10.1111/odi.12932.
10. Marsh PD, Zaura E. Dental biofilm: ecological interactions in health and disease. 2017;44:12–22.
11. Lif Holgerson P, Esberg A, Sjödin A, West CE, Johansson I. A longitudinal study of the development of the saliva microbiome in infants 2 days to 5 years compared to the microbiome in adolescents. Sci Rep [Internet]. 2020;10(1):9629. Available from: https://doi.org/10.1038/s41598-020-66658-7.
12. Dzidic M, Collado MC, Abrahamsson T, et al. Oral microbiome development during childhood: an ecological succession influenced by postnatal factors and associated with tooth decay. ISME J [Internet]. 2018;12(9):2292–306. Available from: https://doi.org/10.1038/s41396-018-0204-z.
13. Gussy M, Mnatzaganian G, Dashper S, et al. Identifying predictors of early childhood caries among Australian children using sequential modelling: Findings from the VicGen birth cohort study. J Dent [Internet]. 2020;93:103276. Available from: https://www.sciencedirect.com/science/article/pii/S0300571220300105.
14. Dashper SG, Mitchell HL, Lê Cao K-A, et al. Temporal development of the oral microbiome and prediction of early childhood caries. Sci Rep. 2019 Dec;9(1):19732.
15. Grier A, Myers JA, O'Connor TG, et al. Oral Microbiota Composition Predicts Early Childhood Caries Onset. J Dent Res. 2021 Jun;100(6):599–607.
16. Nyvad B, Takahashi N. Integrated hypothesis of dental caries and periodontal diseases. J Oral Microbiol [Internet]. 2020 Jan 7;12(1):1710953. Available from: https://pubmed.ncbi.nlm.nih.gov/32002131.
17. Hurley E, Barrett MPJ, Kinirons M, et al. Comparison of the salivary and dentinal microbiome of children with severe-early childhood caries to the salivary microbiome of caries-free children. BMC Oral Health. 2019 Jan;19(1):13.
18. Manzoor M, Lommi S, Furuholm J, et al. High abundance of sugar metabolisers in saliva of children with caries. Sci Rep [Internet]. 2021;11(1):4424. Available from: https://doi.org/10.1038/s41598-021-83846-1.
19. Jiang S, Gao X, Jin L, Lo ECM. Salivary microbiome diversity in caries-free and caries-affected children. Int J Mol Sci. 2016;17(12).
20. Kim B-S, Han D-H, Lee H, Oh B. Association of Salivary Microbiota with Dental Caries Incidence with Dentine Involvement after 4 Years. J Microbiol Biotechnol. 2018 Mar;28(3):454–64.
21. Belstrøm D, Fiehn N-E, Nielsen CH, et al. Altered bacterial profiles in saliva from adults with caries lesions: a case-cohort study. Caries Res. 2014;48(5):368–75.
22. Shade A. Diversity is the question, not the answer. ISME J [Internet]. 2017;11(1):1–6. Available from: https://doi.org/10.1038/ismej.2016.118.
23. Mukherjee C, Moyer CO, Steinkamp HM, et al. Acquisition of oral microbiota is driven by environment, not host genetics. Microbiome [Internet]. 2021;9(1):54. Available from: https://doi.org/10.1186/s40168-020-00986-8.
24. Sulyanto RM, Thompson ZA, Beall CJ, Leys EJ, Griffen AL. The Predominant Oral Microbiota Is Acquired Early in an Organized Pattern. Sci Rep [Internet]. 2019;9(1):10550. Available from: https://doi.org/10.1038/s41598-019-46923-0.
25. Ramadugu K, Bhaumik D, Luo T, et al. Maternal Oral Health Influences Infant Salivary Microbiome. J Dent Res. 2021 Jan;100(1):58–65.
26. Mark Welch JL, Utter DR, Rossetti BJ, et al. Dynamics of tongue microbial communities with single-nucleotide resolution using oligotyping. Front Microbiol [Internet]. 2014;5:568. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2014.00568.
27. Zaura E, Brandt BW, Prodan A, et al. On the ecosystemic network of saliva in healthy young adults. ISME J [Internet]. 2017;11(5):1218–31. Available from: https://doi.org/10.1038/ismej.2016.199.
28. Relvas M, Regueira-Iglesias A, Balsa-Castro C, et al. Relationship between dental and periodontal health status and the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models. Sci Rep [Internet]. 2021;11(1):929. Available from: https://doi.org/10.1038/s41598-020-79875-x.
29. Custodio R, Johnson E, Liu G, Tang CM, Exley RM. Commensal Neisseria cinerea impairs Neisseria meningitidis microcolony development and reduces pathogen colonisation of epithelial cells. PLOS Pathog [Internet]. 2020;16(3):1–21. Available from: https://doi.org/10.1371/journal.ppat.1008372.
30. Kim WJ, Higashi D, Goytia M, et al. Commensal Neisseria Kill Neisseria gonorrhoeae through a DNA-Dependent Mechanism. Cell Host Microbe. 2019 Aug;26(2):228–39.e8.
31. Mashima I, Nakazawa F. The influence of oral Veillonella species on biofilms formed by Streptococcus species. Anaerobe. 2014 Aug;28:54–61.
32. Liu J, Wu C, Huang I-H, Merritt J, Qi F. Differential response of Streptococcus mutans towards friend and foe in mixed-species cultures. Microbiology [Internet]. 2011/05/12. 2011 Sep;157(Pt 9):2433–44. Available from: https://pubmed.ncbi.nlm.nih.gov/21565931.

33. Guo L, Shokeen B, He X, Shi W, Lux R. Streptococcus mutans SpaP binds to RadD of Fusobacterium nucleatum ssp. polymorphum. Mol Oral Microbiol [Internet]. 2017;32(5):355–64. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/omi.12177.
34. Thurnheer T, Karygianni L, Flury M, Belibasakis GN. Fusobacterium Species and Subspecies Differentially Affect the Composition and Architecture of Supra- and Subgingival Biofilms Models. Front Microbiol [Internet]. 2019 Jul 30;10:1716. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31417514.
35. Biyikoğlu B, Ricker A, Diaz PI. Strain-specific colonization patterns and serum modulation of multi-species oral biofilm development. Anaerobe. 2012 Aug;18(4):459–70.
36. Rosier BT, Buetas E, Moya-Gonzalvez EM, Artacho A, Mira A. Nitrate as a potential prebiotic for the oral microbiome. Sci Rep [Internet]. 2020;10(1):12895. Available from: https://doi.org/10.1038/s41598-020-69931-x.
37. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, et al. The oral metagenome in health and disease. ISME J. 2012 Jan;6(1):46–56.
38. Edlund A, Yang Y, Yooseph S, et al. Uncovering complex microbiome activities via metatranscriptomics during 24 hours of oral biofilm assembly and maturation. Microbiome [Internet]. 2018;6(1):217. Available from: https://doi.org/10.1186/s40168-018-0591-4.
39. Son MR, Shchepetov M, Adrian PV, et al. Conserved mutations in the pneumococcal bacteriocin transporter gene, blpA, result in a complex population consisting of producers and cheaters. MBio. 2011;2(5).
40. Qi F, Chen P, Caufield PW. The group I strain of Streptococcus mutans, UA140, produces both the lantibiotic mutacin I and a nonlantibiotic bacteriocin, mutacin IV. Appl Environ Microbiol. 2001 Jan;67(1):15–21.
41. Eren AM, Borisy GG, Huse SM, Mark JL. Oligotyping analysis of the human oral microbiome. 2014.
42. Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci [Internet]. 2016 Feb 9;113(6):E791 LP-E800. Available from: http://www.pnas.org/content/113/6/E791.abstract.
43. Mager DL, Ximenez-Fyvie LA, Haffajee AD, Socransky SS. Distribution of selected bacterial species on intraoral surfaces. J Clin Periodontol. 2003 Jul;30(7):644–54.
44. Neiswanger K, McNeil DW, Foxman B, et al. Oral Health in a Sample of Pregnant Women from Northern Appalachia (2011–2015). Int J Dent [Internet]. 2015;2015:469312–76. Available from: http://umich.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwrV1LSwMxEA5qD3oRX2i1Ihw86GFtm0d391i1SxHB4gO8hWSTSKFWWdu7_8F_6C9x-lmhYfuG9TJ5nfH-UuEv9BArkwWXZGwZzUQuU8-5tCK32sTh5YqDlHbecLiLa4yDSF7CVuFK5.
45. Robins JM, Gail MH, Lubin JH. More on “Biased selection of controls for case-control analyses of cohort studies”. Biometrics. 1986 Jun;42(2):293–9.
46. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol [Internet]. 2013/06/21. 2013 Sep;79(17):5112–20. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23793624.
47. Callahan BJ, McMurdie PJ, Rosen MJ, et al. DADA2: High resolution sample inference from Illumina amplicon data. Nat Methods [Internet]. 2016;13(7):581–3. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4927377/.
48. Escapa IF, Huang Y, Chen T, et al. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome [Internet]. 2020;8(1):65. Available from: https://doi.org/10.1186/s40168-020-00841-w.
49. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome [Internet]. 2018;6(1):226. Available from: https://doi.org/10.1186/s40168-018-0605-2.
50. Kuhn M. caret: Classification and Regression Training [Internet]. 2021. Available from: https://cran.r-project.org/package=caret.
51. John CR. MLeval: Machine Learning Model Evaluation [Internet]. 2020. Available from: https://cran.r-project.org/package=MLeval.
52. Holmes I, Harris K, Quince C. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. PLoS One [Internet]. 2012 Feb 3;7(2):e30126. Available from: https://doi.org/10.1371/journal.pone.0030126.
53. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics [Internet]. 2008;9(1):559. Available from: https://doi.org/10.1186/1471-2105-9-559.
54. Bodenhofer U, Bonatesta E, Horejs-Kainrath C, Hochreiter S. msa: an R package for multiple sequence alignment. Bioinformatics. 2015;31(24):3997–9.
55. Schliep K, Potts, et al. Intertwining phylogenetic trees and networks. Methods Ecol Evol. 2017;8(10):1212–20.
56. Yu G, Smith D, Zhu H, Guan Y, Lam TT-Y ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol [Internet]. 2017;8(1):28–36. Available from: http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12628/abstract.
57. Beghini F, McIver LJ, Blanco-Miguez A, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. bioRxiv [Internet]. 2020; Available from: https://www.biorxiv.org/content/early/2020/11/21/2020.11.19.388223.
58. Tamames J, Puente-Sánchez F, SqueezeMeta AH, Portable, Fully Automatic Metagenomic Analysis Pipeline. Front Microbiol [Internet]. 2019 Jan 24;9:3349. Available from: https://pubmed.ncbi.nlm.nih.gov/30733714.

Figures

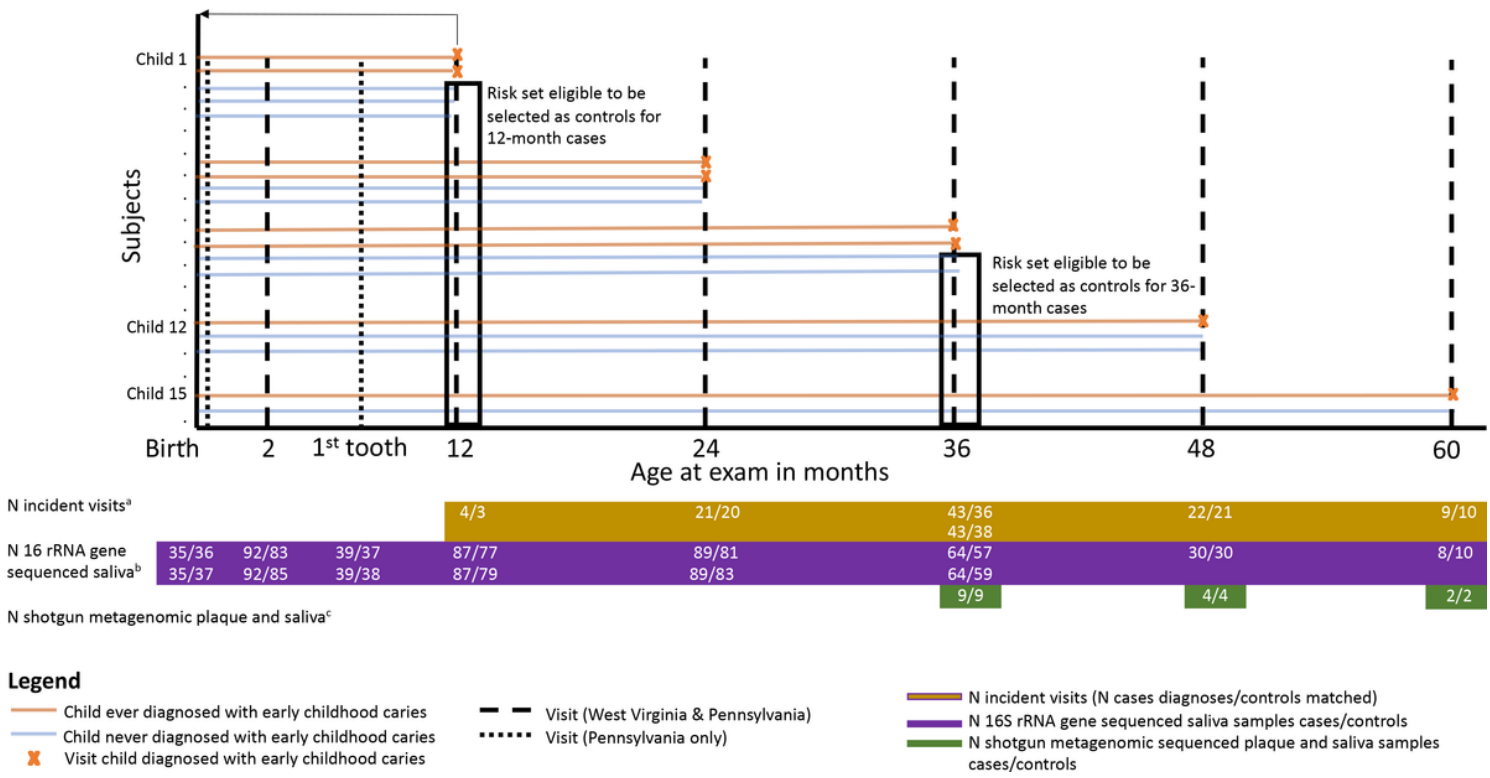

Figure 1

Cohort visit and incidence-density case-control sampling schematic for 189 children from Appalachia. Children were followed from birth until 60-months of age, attending regularly scheduled study visits (black vertical lines). Children who were diagnosed with white spots or enamel lesions were selected as cases (orange bars). For each visit at which cases were diagnosed (incident-visit, orange Xs) a similar number of controls were selected from the group of children free of white spots and enamel lesions at that time (risk set, black vertical rectangles). Children could be selected as a control more than once or as a case and a control. Child 12 (never is diagnosed with white spots nor enamel lesions during follow-up) is in both the 12-month and 36-month risk sets. Child 15 does not have enamel lesions or white spots at 36-months but does at 60-months and could therefore be selected as a control at 36-months and as a case at 60-months. In our sample, 1 child was selected as a control at both the 36- and 60-month visit, and 1 child was selected as a control at the 36-month visit and a case at the 60-month visit. The number of diagnosed cases/matched control children at each visit is shown in the gold box both excluding (top row) and including (bottom row) the second sampling of these twice-sampled children. All the available saliva samples from the incident-visit and all preceding visits for cases and selected controls were sequenced for the V4 region of the 16S rRNA gene (purple box, top row including, bottom row excluding second record for twice-sampled children). For a subsample of 15 children diagnosed with enamel lesions at or after the 36-month visit and their 15 matched-controls both saliva and plaque samples from the visit of diagnosis/matching were shotgun metagenomic sequenced (green box).

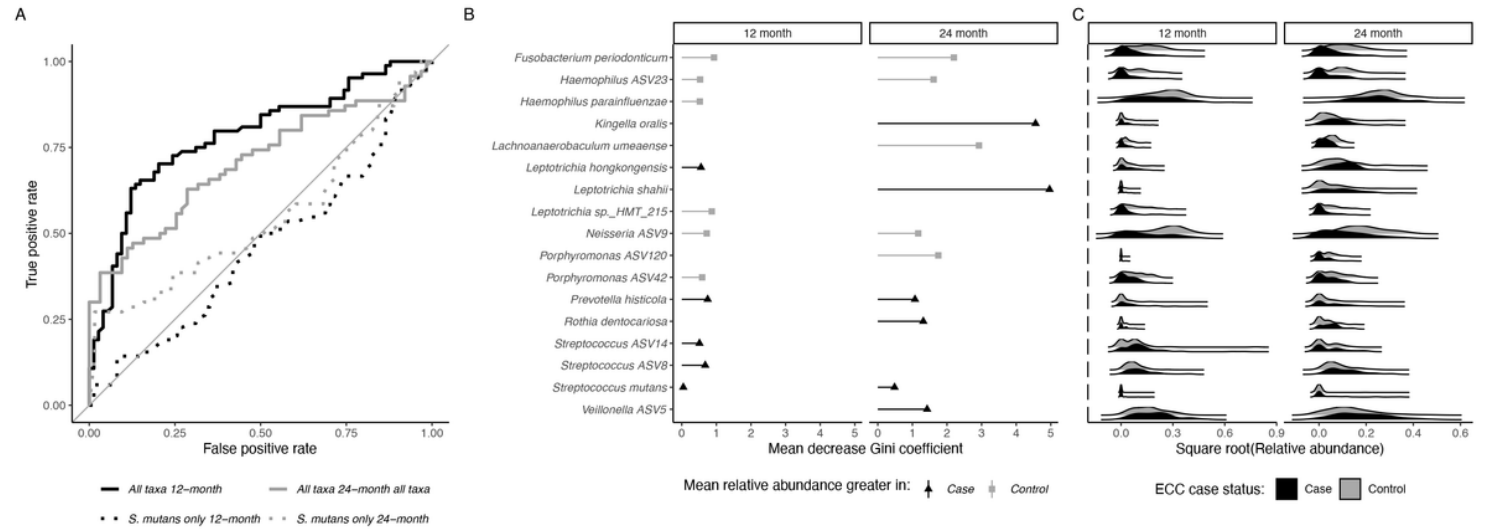

Figure 2

Taxa-wide supervised 5-repeat, 10-fold random forest classification models outperform Streptococcus mutans only model at predicting future early childhood caries status when using 12- (n=158) and 24-month (n=133) 16S rRNA gene amplicon sequenced saliva samples of children from Appalachia in an incidence-density sampled case control study (Center for Oral Health Research in Appalachia 2 cohort): A) Area under the curve receiver operating curves from supervised random forests predicting future early childhood caries using the 273 most prevalent and abundant amplicon sequence variants (solid lines) vs using the amplicon identified as *S. mutans* only (dotted lines) at 12-months (black lines) and 24-months (grey lines) B) Importance plots showing the top ten most important amplicon sequence variants from the 12- and 24-month supervised random forest classifiers performed on 273 amplicon sequence variants, as determined by mean decrease in the Gini coefficient, with the importance of the *S. mutans* amplicon included for comparison. C) Joy plots showing the relative abundance distribution of the top ten most important amplicon sequence variants and *S. mutans* among cases (black) and controls (grey) at the 12- and 24-month visits

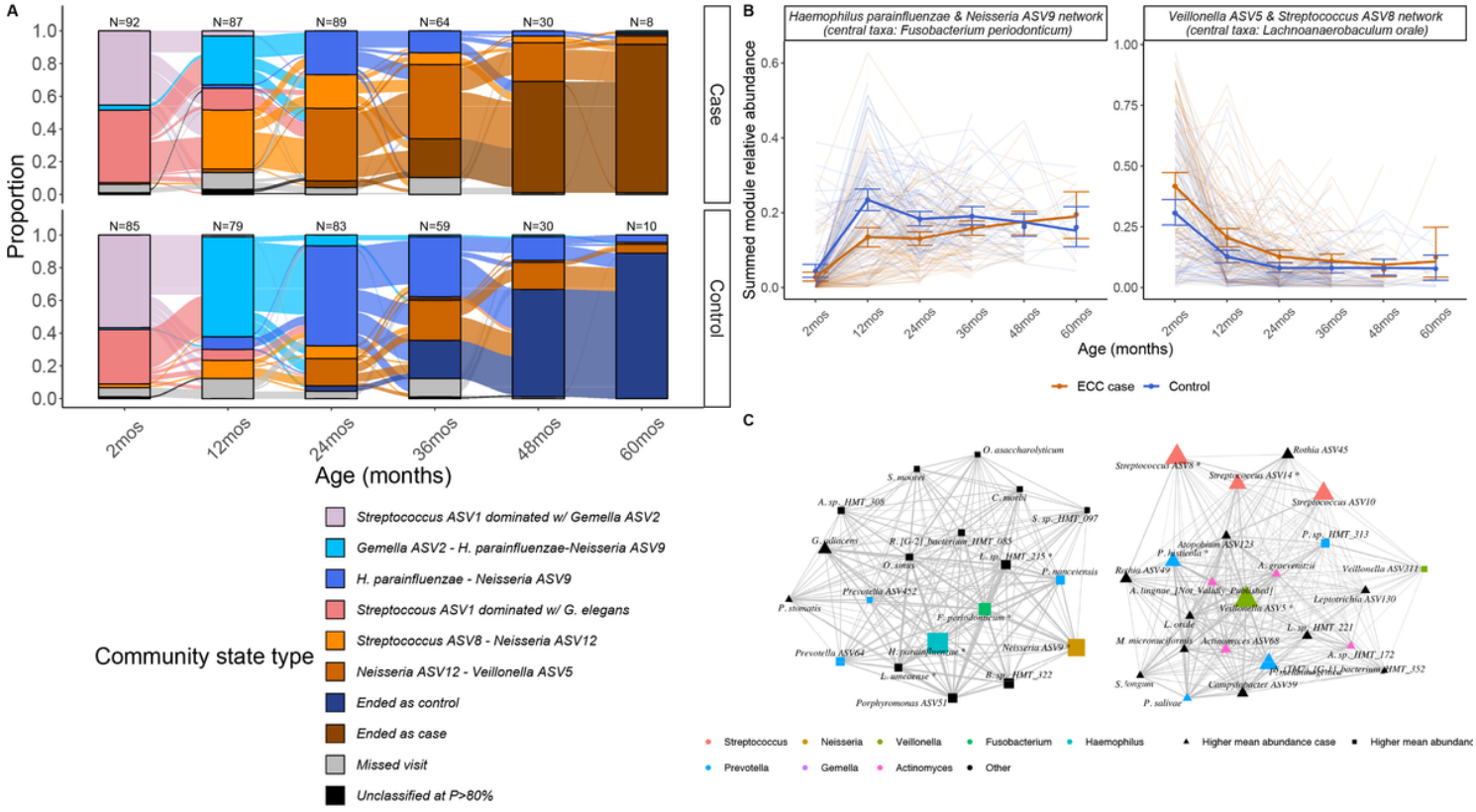

Figure 3

Two different unsupervised clustering techniques, community state typing (A) and weighted cooccurrence networks (B & C) identify overlapping and early childhood caries-associated groups of cooccurring taxa when performed on 855 longitudinal, 16S rRNA gene sequenced saliva samples from 189 children from Appalachia (191 records) in an incidence-sampled case-control study (Center for Oral Health Research in Appalachia 2 study). A) Alluvial plot showing the proportion of the sample in each community state type at each visit and the transitions between visits, faceted by early childhood caries case status. Bars are annotated at the top with the sample size N of cases and controls (excluding children who missed visit or did not have cleaned 16S rRNA gene amplicon saliva data for that visit) B) Spaghetti plots showing the summed module relative abundance of two of the five identified network modules from weighted co-occurrence networks. Networks were named using the two most abundant amplicon sequence variants in the network and the most central amplicon sequence variant. Summed module relative abundance calculated by summing the relative abundance of all amplicon sequence variants assigned to the same cluster. Thin, transparent lines are individuals over time, thick lines represent smoothed means, dots and bars are mean and bootstrapped 95% confidence intervals at each visit. C) Network graphs of the two network modules shown in (B). Amplicon sequence variants (nodes) that were more abundant in cases are shown as triangles, those more abundant in controls are shown as squares. Larger nodes represent more abundant amplicon sequence variants, and nodes are colored by genus. Amplicon sequence variants which were among the top ten most important features in the supervised random forests are annotated with an asterisk, *. Thicker edges represent stronger correlations between amplicon sequence variants.

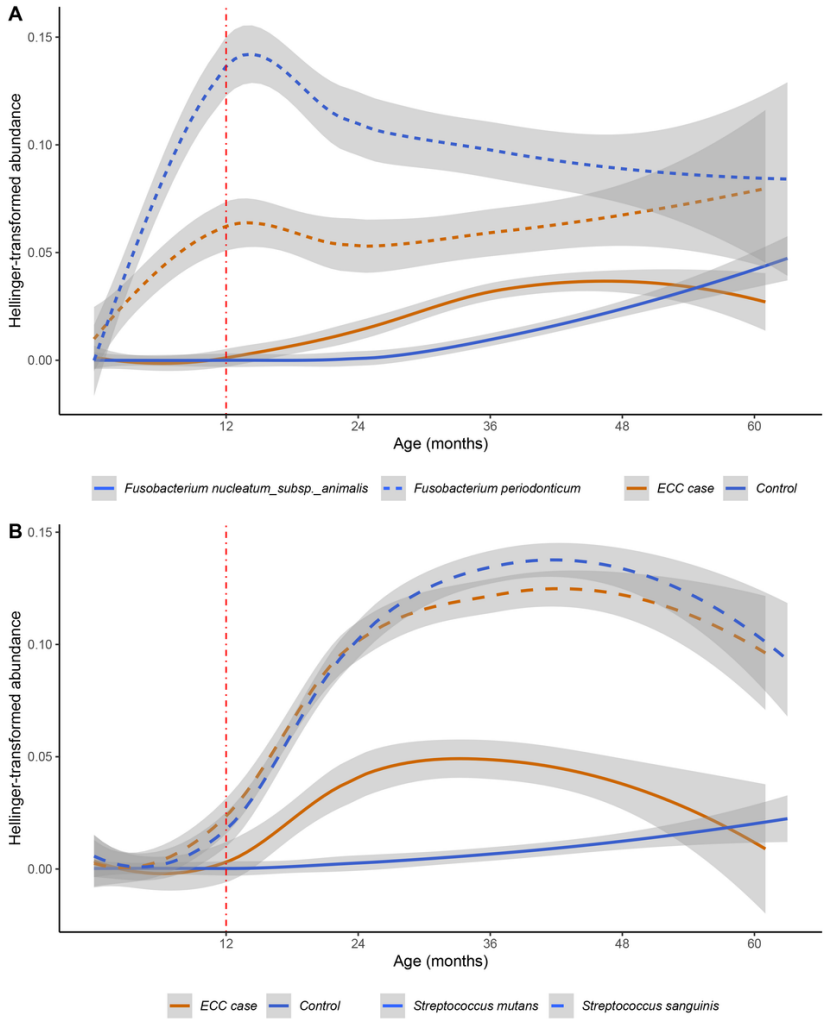

Figure 4

Intragenera and intergenera longitudinal correlations between Fusobacterium and Streptococcus species of interest in 16S rRNA gene amplicon saliva samples from 189 children from Appalachia (191 records) in an incidence-sampled case-control study (Center for Oral Health Research in Appalachia 2 study. A) *Fusobacterium periodonticum* (dashed line) increased in relative abundance more among controls (blue line) than cases (orange line), peaking shortly after 12-months of age (red dashed and dotted line). *Fusobacterium nucleatum subspecies animalis* (solid line) increased in relative abundance only after 12-months of age, with a larger increase among cases (orange) than controls (blue). B) Both *Streptococcus sanguinis* (dashed line) and *Streptococcus mutans* (solid line) increased in relative abundance starting at around 12-months (red dashed and dotted line). Cases (orange) had a larger increase in *S. mutans* than controls (blue), and this increase was time-correlated with the increase in *Fusobacterium nucleatum subspecies animalis* (3B – solid line)

Figure 5

Incident-visit plaque and saliva samples exhibited significantly different abundances of KEGG ortholog groups among 15 early childhood caries cases and 15 incidence-density matched controls selected from the Center for Oral Health Research in Appalachia 2 study. A) Volcano plots showing the -log10 pvalue and log2 fold change between cases and controls of KEGG orthologs in plaque and saliva samples. Points are colored black if Benjamini-Hochberg *P* value>0.05 and by the first top-level KEGG annotation from the KEGG hierarchy of the KEGG ortholog if the adjusted *P* value<0.05. The top 6 most significant KEGG orthologs are annotated with the name of the KEGG ortholog and the taxa in which that KEGG ortholog was found in our sample B) Count of KEGG orthologs with Benjamini-Hochberg adjusted *P* value<0.05 by 3rd level KEGG annotation (y-axis), faceted by top level KEGG annotation (colors same as in 3A).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.xlsx
- AdditionalFile2.pdf
- AdditionalFile3.xlsx
- AdditionalFile4.xlsx
- AdditionalFile5.xlsx
- AdditionalFile6.xlsx
- AdditionalFile7.xlsx
- AdditionalFile8.xlsx
- AdditionalFile9.xlsx
- AdditionalFile10.xlsx
- AdditionalFile11.xlsx