*Research Reports: Clinical*

Journal of Dental Research
1–10

DOI: 10.1177/00220345231152802
journals.sagepub.com/home/jdr

# Prediction of Future Caries in 1-Year-Old Children via the Salivary Microbiome

**R. Raksakmanut[1], P. Thanyasrisung[2], S. Sritangsirikul[3,4], K. Kitsahawong[5], A.L. Seminario[6], W. Pitiphat[7], and O. Matangkasombut[2,8]**

**Abstract**

Dental caries is the most common chronic disease in children that causes negative effects on their health, development, and well-being. Early preventive interventions are key to reduce early childhood caries prevalence. An efficient strategy is to provide risk-based targeted prevention; however, this requires an accurate caries risk predictor, which is still lacking for infants before caries onset. We aimed to develop a caries prediction model based on the salivary microbiome of caries-free 1-y-old children. Using a nested case-control design within a prospective cohort study, we selected 30 children based on their caries status at 1-y follow-up (at 2 y old): 10 children who remained caries-free, 10 who developed noncavitated caries, and 10 who developed cavitated caries. Saliva samples collected at baseline before caries onset were analyzed through 16S rRNA gene sequencing. The results of β diversity analysis showed a significant difference in salivary microbiome composition between children who remained caries-free and those who developed cavitated caries at 2 y old (analysis of similarities, Benjamini-Hochberg corrected, $P = 0.042$). The relative abundance of *Prevotella nanceiensis*, *Leptotrichia* sp. HMT 215, *Prevotella melaninogenica*, and *Campylobacter concisus* in children who remained caries-free was significantly higher than in children who developed cavitated caries (Wilcoxon rank sum test, $P = 0.024$, 0.040, 0.049, and 0.049, respectively). These taxa were also identified as biomarkers for children who remained caries-free (linear discriminant analysis effect size, linear discriminant analysis score = 3.69, 3.74, 3.53, and 3.46). A machine learning model based on these 4 species distinguished between 1-y-old children who did and did not develop cavitated caries at 2 y old, with an accuracy of 80%, sensitivity of 80%, and specificity of 80% (area under the curve, 0.8; 95% CI, 44.4 to 97.5). Our findings suggest that these salivary microbial biomarkers could assist in predicting future caries in caries-free 1-y-old children and, upon validation, are promising for development into an adjunctive tool for caries risk prediction for prevention and monitoring.

**Keywords:** dental caries, biomarkers, microbiota, machine learning, infant, saliva

[1]Graduate Program in Oral Biology and Center of Excellence on Oral Microbiology and Immunology, Faculty of Dentistry, Chulalongkorn University, Wang-Mai, Pathumwan, Bangkok, Thailand
[2]Department of Microbiology and Center of Excellence on Oral Microbiology and Immunology, Faculty of Dentistry, Chulalongkorn University, Wang-Mai, Pathumwan, Bangkok, Thailand
[3]Department of Pediatric Dentistry, Faculty of Dentistry, Chulalongkorn University, Wang-Mai, Pathumwan, Bangkok, Thailand
[4]PhD Program in Oral Sciences, Faculty of Dentistry, Khon Kaen University, Muang District, Khon Kaen, Thailand
[5]Division of Pediatric Dentistry, Department of Preventive Dentistry, Faculty of Dentistry, Khon Kaen University, Muang District, Khon Kaen, Thailand
[6]Department of Pediatric Dentistry, School of Dentistry, University of Washington, WA, USA
[7]Division of Dental Public Health, Department of Preventive Dentistry, Faculty of Dentistry, Khon Kaen University, Muang District, Khon Kaen, Thailand
[8]Research Laboratory of Biotechnology, Chulabhorn Research Institute, Laksi, Bangkok, Thailand

A supplemental appendix to this article is available online.

**Corresponding Author:**
O. Matangkasombut, Department of Microbiology and Center of Excellence on Oral Microbiology and Immunology, Faculty of Dentistry, Chulalongkorn University, 34 Henri-Dunant Road, Wang-Mai, Pathumwan, Bangkok, 10330, Thailand.
Email: oranart.m@chula.ac.th, oranart.chula@gmail.com

## Introduction

Early childhood caries (ECC) is among the most common chronic diseases in children globally, with prevalence increasing drastically during the first 5 y of life (Tinanoff et al. 2019). ECC is considered a critical public health concern due to its profound impacts on children's growth, development, and quality of life and its economic burden (Casamassimo et al. 2009). Providing preventive interventions as early as possible, especially within the first year of life, could be a key approach to reduce overall caries prevalence in children (Pitts et al. 2019). Importantly, targeting preventions to children with high risk of developing caries would be the most efficient strategy, particularly in limited resource settings. Yet, an accurate caries prediction tool, which is essential for target group identification, is still lacking (Fontana et al. 2020).

Existing caries management systems, such as CAMBRA and CariesCare International, are useful for identifying individual risk factors for personalized prevention and treatment but have limitations for predicting future disease (Mejàre et al. 2014; Christian et al. 2020). Among the risk factors evaluated in these systems, past caries experience is considered the most powerful predictor (Hänsel Petersson et al. 2013; Mejàre et al. 2014). Thus, there is a significant limitation for their application in infants before any caries initiates.

According to the ecologic plaque hypothesis, dental caries is caused by oral microbial dysbiosis within the biofilm resulting from changes in the oral environment (Marsh 1994; Simon-Soro and Mira 2015). The long process of cariogenesis consists of multiple stages (Takahashi and Nyvad 2011), each with different microbial contributors (Li et al. 2004; Takahashi and Nyvad 2008). There is evidence that the oral microbiome reflects conditions in the oral cavity deriving from other cariogenic factors (Grassl et al. 2016; Hemadi et al. 2017) and could be an independent predictor of dental caries.

A few recent studies proposed novel caries prediction models derived from supervised machine learning based on oral microbiome data, with an accuracy of 81% to 93.1% and with use of a variety of microbial biomarkers (Teng et al. 2015; Xu et al. 2018; Grier et al. 2021). However, most of these studies were performed in children at the age of 2 to 6 y. Since the oral microbiome changes rapidly as children grow and their dentition and diet also change drastically in the first 3 y of life (Dashper et al. 2019), data from these studies might not be applicable to children of different ages, especially those in the first year of life. Therefore, there is a critical gap in our knowledge on the oral microbiome of young infants in relation to their risk of caries development in subsequent years.

The aims of this study were to investigate the salivary microbial composition of caries-free 1-y-old children and to identify microbial predictors for dental caries onset in the following year. We used a nested case-control design within a prospective cohort and machine learning to develop caries prediction models to distinguish 1-y-old caries-free children who remained caries-free from those who developed caries 1 y later.

## Materials and Methods

### Study Samples

This report conforms to the STROBE guideline for observational study. All saliva samples used were stored samples collected in our previous prospective study conducted among 1-y-old children in Khon Kaen, Thailand (Sritangsirikul et al. 2021), described briefly in the Appendix. This study protocol was approved by the Human Research Ethics Committee, Faculty of Dentistry, Chulalongkorn University (HREC-DCU 2021-032). Saliva (~1 mL) was collected with a dropper, and after an aliquot was taken for previous studies, remaining samples were stored in 25% glycerol at −80 °C. Among the cohort's 568 participants, 113 were caries-free with >0.1 mL of the saliva sample remaining. Oral examination for caries was performed via modified World Health Organization (1997) diagnostic criteria, and cavitated and noncavitated caries were recorded separately.

In brief, using a nested case-control design, we selected 30 saliva samples collected at baseline from 113 caries-free participants based on their caries status at 12-mo follow-up and saliva sample quantity (at least 0.4 mL): 10 children who remained caries-free (caries-free to caries-free; F2F), 10 children who developed at least 1 noncavitated caries lesion (white spot lesion) without any cavitated caries (caries-free to white spot; F2W), and 10 children who developed at least 1 cavitated caries lesion (caries-free to decay; F2D). A previous study on salivary microbiome in children used species accumulation curves to show that sequence data reached saturation at 10 samples per set (Zhu et al. 2018), suggesting that this sample size (10 per group) is sufficient for microbiome analysis. An additional 10 samples were selected from cavitated caries lesion–free 1-y-old children as a testing group for model validation.

### Sample Processing and Analysis

The overview of study design and analysis pipeline is shown in Figure 1. Detailed methods, including sample preparation, 16S rRNA gene sequencing, bioinformatics, and statistical methods, are described in the Appendix. All 16S rRNA sequence reads and generated data sets are publicly available through NCBI accession PRJNA866487.

## Results

### Characteristics of Participants and Samples

All 3 groups of children were similar in age, gender, status of human milk consumption, age weaned from breastfeeding, number of erupted teeth at 1 and 2 y old, and salivary pH (Table). Mean ± SD age at the time of enrolment was 12.7 ± 1.3, 12.6 ± 1.1, and 12.4 ± 0.5 mo for children who remained caries-free (F2F), those who developed noncavitated caries (F2W), and those with cavitated caries (F2D) at 12-mo-follow-up, respectively.

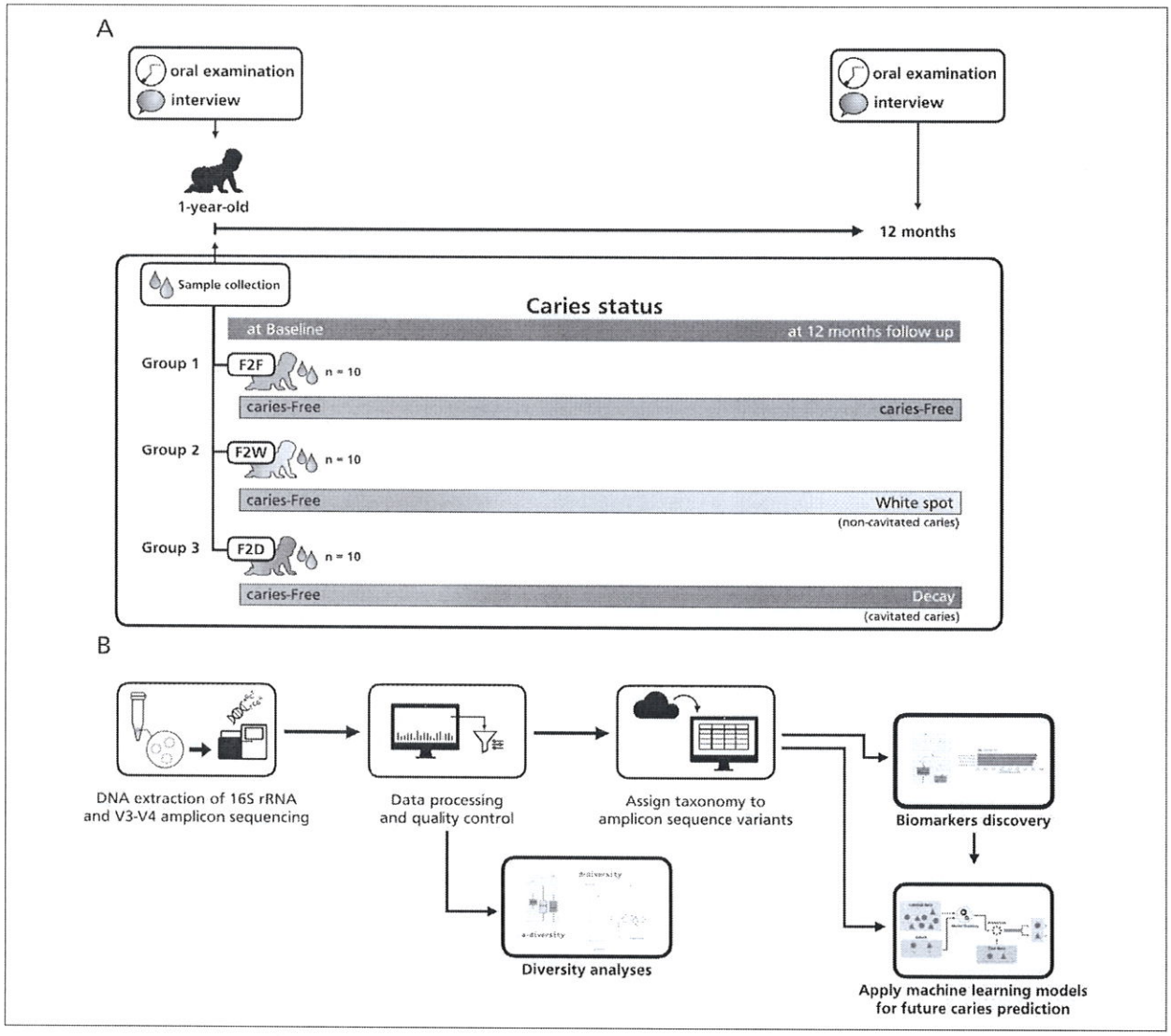


**Figure 1.** Study scheme overview. (**A**) Saliva samples were collected from caries-free 1-y-old children at baseline, while caries status was examined at baseline and 12-mo-follow-up. Samples were grouped into caries-free to caries-free (F2F), caries-free to white spot (F2W), and caries-free to decay (F2D), based on caries status at follow-up. (**B**) Analysis pipeline consisted of DNA extraction, 16S ribosomal RNA gene amplification (V3 to V4 hypervariable regions), DNA sequencing, quality control, diversity analyses, taxonomic assignments, biomarkers discovery, and use of machine learning to generate models for future caries prediction.

The salivary microbiota determination by 16S rRNA gene sequencing produced an average of >57,000 reads per sample after quality control and amplicon sequence variant identification. The minimum frequency among all samples was 18,207 with an average of 28,489.7 per sample, representing 945 unique features, 160 distinct species-level taxonomic assignments, and 74 genera. The top 15 most abundant genera accounted for 94.7% of the overall composition across all samples, with *Streptococcus* being the most abundant (29.2%), followed by *Alloprevotella* (12.5%), *Veillonella* (9.99%), *Haemophilus* (8.8%), *Leptotrichia* (6.2%), and *Neisseria* (6.1%; Fig. 2A). The abundance was not significantly different when compared among groups ($P > 0.05$, Kruskal-Wallis). The top 25 most abundant species accounted for 87.9% of the overall composition. Among these, 6 were unclassified (Fig. 2B).

## *Salivary Microbiome Diversity among Caries-free Children with Different Future Caries Status*

Diversity, within sample ($\alpha$ diversity) and between samples ($\beta$ diversity), was analyzed among the 3 groups with different future caries outcomes at the amplicon sequence variant level. For $\alpha$ diversity, no significant difference was found in terms of Shannon and Chao1 $\alpha$ diversity among the 3 groups ($P = 0.11$

**Table.** Characteristics of Participants by Comparison Group.

| Variable | F2F ($n$ = 10) | F2W ($n$ = 10) | F2D ($n$ = 10) | $P$ Value |
|---|---|---|---|---|
| Age, mo | 12.7 ± 1.3 | 12.6 ± 1.1 | 12.4 ± 0.5 | 0.99[a] |
| Sex | | | | 0.87[b] |
| Male | 50 | 60 | 50 | |
| Female | 50 | 40 | 50 | |
| No. of teeth | | | | |
| At 1 y old | 5.7 ± 2.2 | 5.4 ± 2.2 | 4 ± 1.9 | 0.19[a] |
| At 2 y old | 17.4 ± 2.3 | 16.6 ± 1.3 | 15.8 ± 1.3 | 0.28[a] |
| Weaned (human milk) | 90 | 90 | 60 | 0.15[b] |
| Age weaned from breastfeeding, mo | 6.6 ± 6.7 | 5.8 ± 6.0 | 8.6 ± 8.6 | 0.78[a] |
| pH of saliva | 7.2 ± 0.3 | 7.2 ± 0.6 | 7.4 ± 0.6 | 0.57[c] |
| At 2 y old[d] | | | | |
| dmfs | 0 | 0 | 5.4 ± 4.9 | <0.001[a] |
| $d_1$mfs | 0 | 6.1 ± 2.1 | 8.6 ± 4.7 | <0.001[a] |

Values are presented as mean ± SD or %.
F2D, developed cavitated caries; F2F, caries-free children who remained caries-free; F2W, developed noncavitated caries.
[a]Kruskal-Wallis test.
[b]Chi-square test.
[c]Analysis of variance.
[d]dmfs is the sum of the number of decayed, missing due to caries, and filled tooth surfaces in the deciduous teeth. $d_1$mfs includes $d_1$ (detectable enamel lesion with a sound surface) or worse as decayed. There was no m (missing due to caries) and f (filled tooth/surface) found in any children.

and 0.30, respectively, Kruskal-Wallis; Fig. 2C, D). However, there was a significant difference between F2F and F2W groups in terms of the relative evenness of species richness (Pielou evenness, $P = 0.024$, Kruskal-Wallis with Benjamini-Hochberg correction; Fig. 2E). For β diversity, principal coordinates analysis based on unweighted UniFrac distance showed that the salivary microbiota of children who stayed caries-free (F2F) was significantly different from those who developed cavitated caries at follow-up (F2D; $P = 0.042$, analysis of similarities with Benjamini-Hochberg correction; Fig. 2F). Thus, we focused on the comparison between these groups in subsequent analyses.

### *Microbial Biomarkers of Future Caries among Caries-free Children*

To identify potential biomarkers for caries prediction, univariate tests on each species-level taxonomic unit showed that the relative abundance of *Prevotella nanceiensis*, *Leptotrichia* sp. HMT 215, *Prevotella melaninogenica*, and *Campylobacter concisus* was significantly higher in F2F when compared with F2D ($P = 0.024$, $P = 0.04$, $P = 0.049$, and $P = 0.049$, respectively, Wilcoxon rank sum test; Fig. 3A–D). By using linear discriminant analysis effect size (LEfSe), a biomarker discovery method (Segata et al. 2011), these 4 species were also identified as biomarkers of the F2F group (Fig. 3E).

### *Caries Prediction Models Building from Salivary Microbiome Data*

Our findings showed that the differences in salivary microbiome structure between F2F and F2D could be observed prior to the onset of dental caries. Thus, these data were used as a training data set to generate a random forest model based on the relative abundance of the overall salivary microbiota composition. When this model was tested on a testing data set of an additional 10 samples from 1-y-old cavitated caries lesion–free children (Appendix Table 2), it gave an accuracy of 70%, sensitivity of 80%, and specificity of 60% (area under the curve [AUC], 0.7; 95% CI, 34.8 to 93.3; Fig. 4A).

To improve the predictive accuracy and simplify the models for ease of future applications, potential biomarkers were selected through 2 approaches. First, in the important features model, Boruta and recursive feature elimination algorithms selected future caries–discriminatory taxa based on their importance to model accuracy: *Leptotrichia* sp. HMT 215, *C. concisus*, *P. nanceiensis*, *Leptotrichia goodfellowii*, and *P. melaninogenica* (Fig. 3F, G). The predictive performance of this model was comparable to that of the all-taxa model (AUC, 0.7; 95% CI, 34.8 to 93.3; Fig. 4B). Second, by using LEfSe, a differential abundance model was constructed from the 4 species that showed significant differences in abundance between F2F and F2D (*P. nanceiensis*, *Leptotrichia* sp. HMT 215, *P. melaninogenica*, and *C. concisus*). This approach improved the model performance, giving an accuracy of 80%, sensitivity of 80%, and specificity of 80% (AUC, 0.8; 95% CI, 44.4 to 97.5; Fig. 4C).

Furthermore, single-species models, generated from each of the 5 biomarkers from the important features model, showed that only *C. concisus* abundance exhibited comparable predictive performance to the differential abundance model (AUC, 0.8; 95% CI, 44.4 to 97.5; Fig. 4D). The other single-species models did not perform as well (Appendix Fig. 2).

## Discussion

Dental caries is a complex and multifactorial biofilm-mediated disease (Takahashi and Nyvad 2011). *Streptococcus mutans*

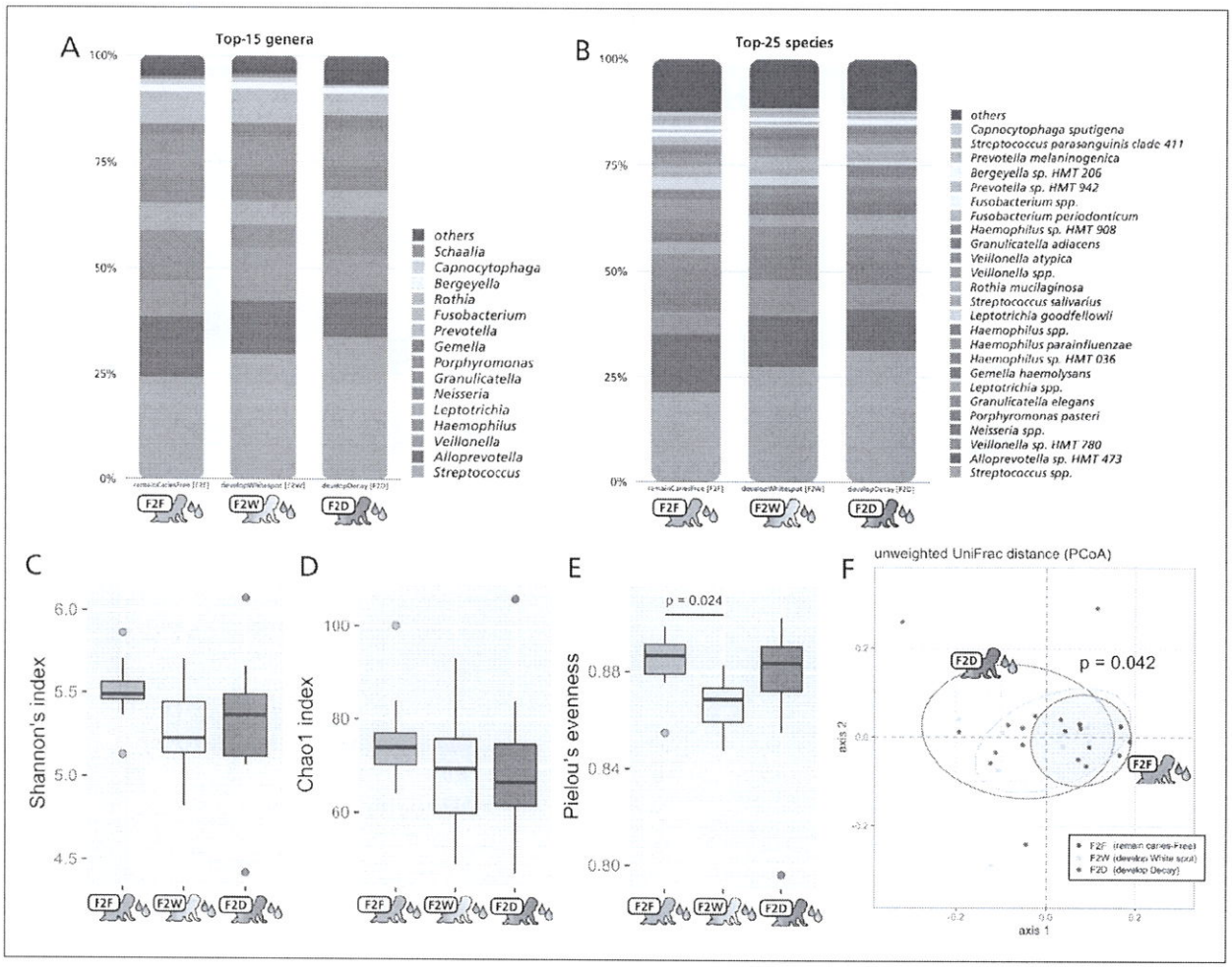


**Figure 2.** Relative abundance of bacterial genera (**A**) and species (**B**) in the microbiome of caries-free children by caries status at 12-mo follow-up in saliva samples based on eHOMD database. F2F children remained caries-free; F2W children developed white spot lesions (noncavitated caries) within 12 mo; and F2D children developed decay (cavitated caries) within 12 mo. Diversity analyses: box plots of alpha diversity, measured by Shannon diversity index (**C**), Chao1 index (**D**), and Pielou evenness (**E**). Green box plots represent F2F group; yellow, F2W; and red, F2D. Horizontal line, median; box, IQR; vertical lines, 95% CI; circle, outlier. (**F**) Principal coordinate analysis plot based on the unweighted UniFrac distance matrix, with samples colored by groups: green dots represent F2F samples; yellow, F2W; and red, F2D.

and *Lactobacillus* spp. are well-known cariogenic bacteria associated with ECC, and their levels have been included as risk factors in certain caries risk assessment tools (Liu et al. 2019). However, the levels of these bacteria could reflect current caries status rather than predict future caries (Hong and Hu 2010). In contrast, changes in oral microbiota resulting from the combined effects of various risk and protective factors precede the establishment of cariogenic bacteria and should therefore be useful as a tool for caries prediction (Dashper et al. 2019). An accurate prediction tool is essential for planning targeted preventive interventions for high-risk infants before caries initiate, but this is still lacking. In this work, we built caries prediction models based on bacterial biomarkers in the salivary microbiome of 1-y-old caries-free children that can distinguish children who remained caries-free from those who developed cavitated caries 1 y later. To our knowledge, this is the first study that developed caries prediction models specifically for this important age group.

Our results showed that, up to 1 y before caries onset, the salivary microbiome composition of 1-y-old children who developed clinically detectable cavitated caries at 2-y-old follow-up was significantly different from that of children who remained caries-free (Fig. 2F). We used supervised machine learning to identify bacterial biomarkers and constructed models for caries prediction. The all-taxa model performed reasonably well in predicting caries in a test data set (Fig. 4A). However, models that use information of all taxa may suffer from overfitting when they learn too many details, including noise, during training. Therefore, we generated simplified models based on selected taxa to increase model performance

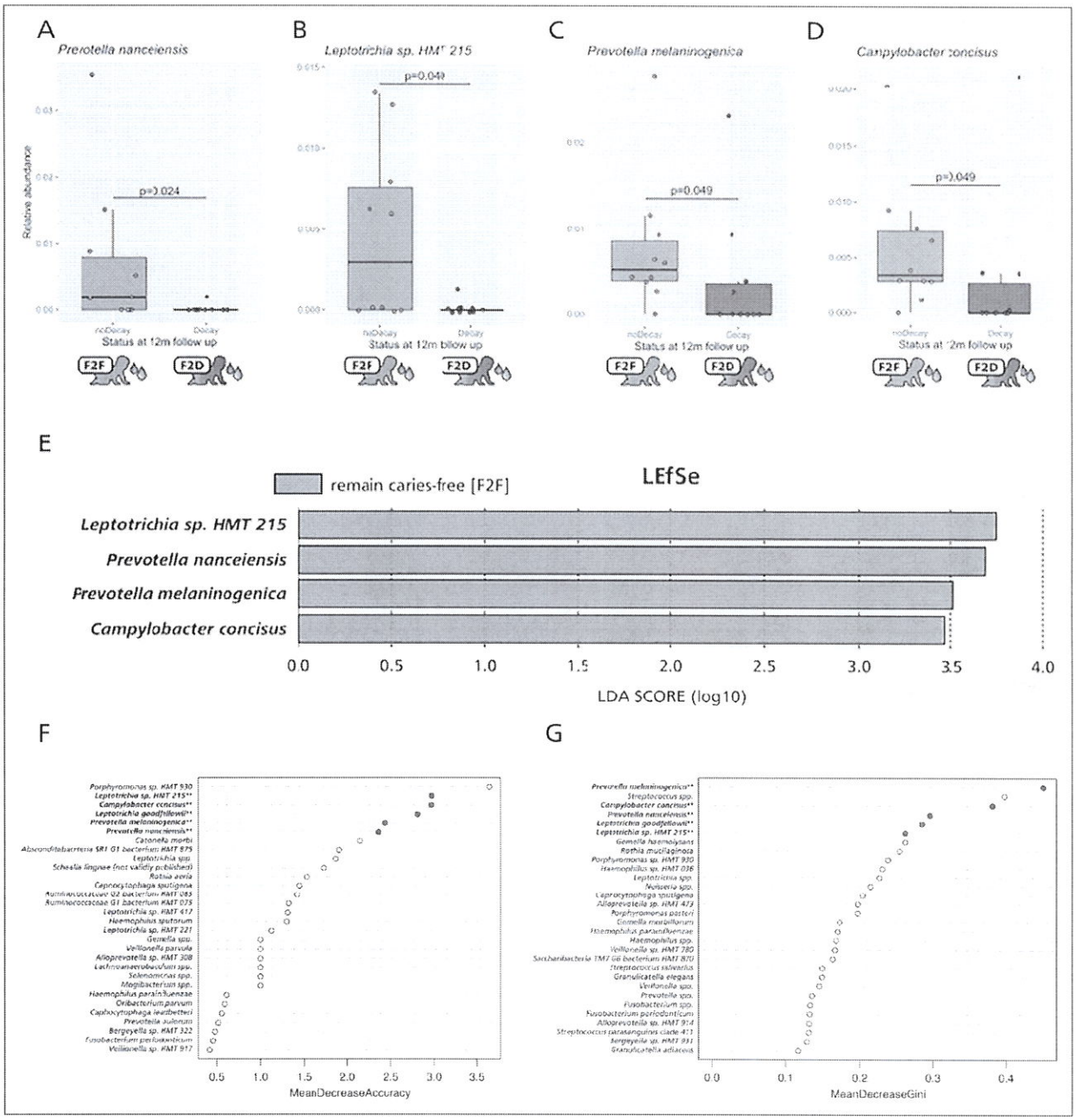


**Figure 3.** Differentially abundant microbial taxa based on caries status at 12-mo follow-up among children who remained caries-free (F2F; light gray box plots [green in the online version]) and those who developed cavitated caries (F2D; dark gray boxplots [red online]) according to the Wilcoxon rank sum tests. A significant difference was observed for (**A**) *Leptotrichia* sp. HMT 215, (**B**) *Prevotella nanceiensis*, (**C**) *Prevotella melaninogenica*, and (**D**) *Campylobacter concisus*. Horizontal line, median; box, IQR; vertical lines, 95% CI. (**E**) Histogram of the linear discriminant analysis scores for differentially abundant features among groups (threshold score set to 2.0). Variable importance plots show most discriminatory bacterial taxa ranked in descending order of importance according to the accuracy of the model: mean decrease in (**F**) accuracy and (**G**) Gini coefficient. **Confirmed important features by Boruta algorithm and recursive feature elimination, shown in black dots (red dots in the online version). LDA, linear discriminant analysis; LEfSe, linear discriminant analysis effect size.

and ease future implementation in clinical applications. The important features model was constructed from 5 species—*Leptotrichia* sp. HMT 215, *C. concisus*, *P. nanceiensis*, *L. goodfellowii*, and *P. melaninogenica*—selected by Boruta algorithm, considered the most powerful selection method of random forests (Kursa and Rudnicki 2010; Degenhardt et al.

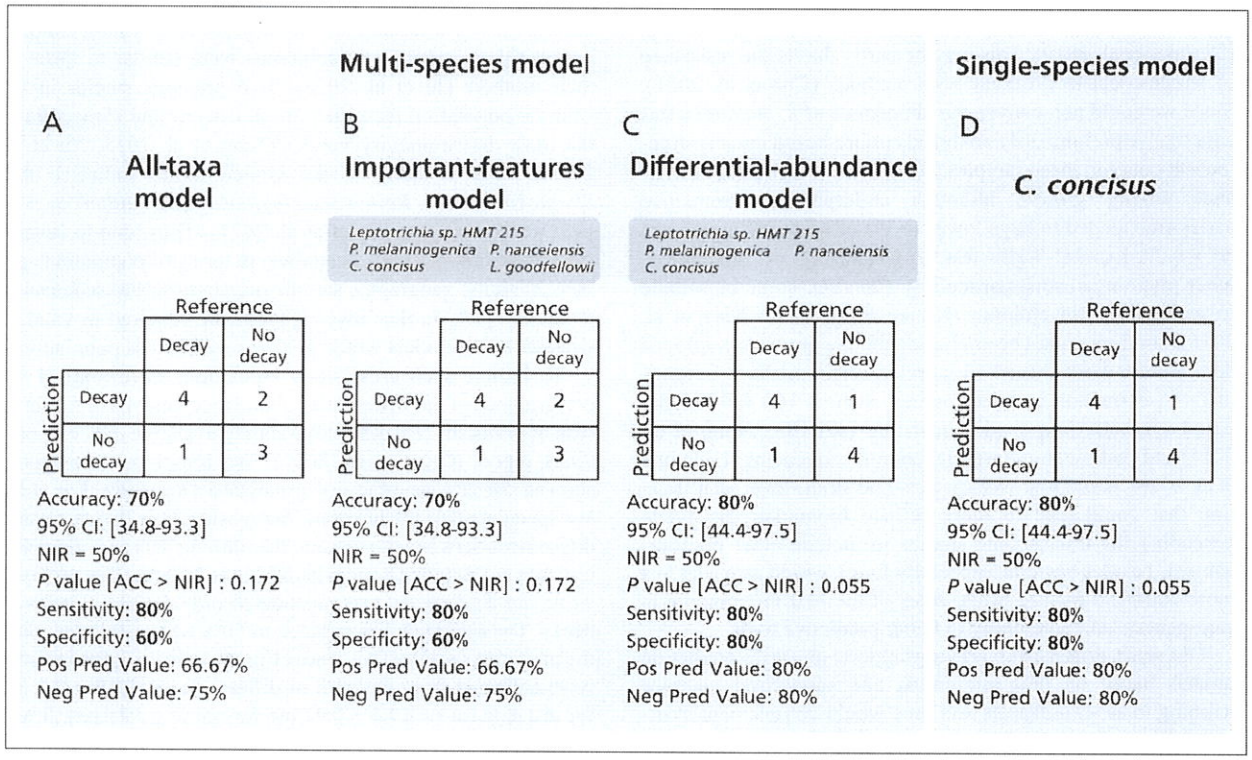


**Figure 4.** Confusion matrix evaluations of random forest models based on their relative abundance: (**A**) all taxa, (**B**) important features (*Leptotrichia* sp. HMT 215, *Prevotella nanceiensis*, *Prevotella melaninogenica*, *Campylobacter concisus*, *Leptotrichia goodfellowii*), (**C**) differential abundance (*Leptotrichia* sp. HMT 215, *P. nanceiensis*, *P. melaninogenica*, *C. concisus*), and (**D**) *C. concisus*. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and *P* value (accuracy [ACC] > no information rate [NIR]) of models are shown.

2017). The selection was confirmed by recursive feature elimination, a popular method for discovering a minimal set of features that delivers good prediction (Diaz-Uriarte and Alvarez de Andres 2006; Degenhardt et al. 2017). We also constructed a differential abundance model based on LEfSe, which identified the same set of biomarkers except *L. goodfellowii* (Fig. 4B, C). Our results suggest that the differential abundance model gave the highest level of accuracy for caries prediction up to 1 y before caries onset. Although the *C. concisus* single-species model gave similar accuracy in this test data set, there is high possibility that this species may be absent in some samples in a larger population. Thus, an effective model should include a balanced combination of appropriate biomarkers.

*C. concisus* was one of the most important features (Fig. 3F, G; Appendix Fig. 3) and was the only single-species model that performed well in caries prediction (Fig. 4D). *C. concisus* abundance was higher in children who stayed caries-free (F2F), which is consistent with a previous study (Al-Hebshi et al. 2019), suggesting that it might play a protective role against cariogenesis. Because of its ability to form biofilm (Lavrencic et al. 2012) and to survive in low pH (Kaakoush et al. 2016), *C. concisus* may be able to compete with cariogenic biofilm producers, such as *S. mutans*. Previous studies reported that *C. concisus* was associated with gingival inflammation and periodontal disease onset (Macuch and Tanner 2000). However, a high prevalence of *C. concisus* in the saliva of healthy persons also suggests that it could be a part of the normal human oral microbiota (Zhang et al. 2010). Our results suggest that a low abundance of *C. concisus* may be a predictor of future caries, but its mechanistic roles require further studies.

Our best model gave 80% accuracy in this test data set, so we examined the cases with incorrect prediction in detail (Appendix Table 3). In the case incorrectly predicted to develop caries (T06), although no cavitated caries developed, the $d_1s$ index increased from 4 at baseline to 12 at follow-up. In contrast, the case incorrectly predicted to remain caries-free (T03) developed only 1 cavitated caries lesion at follow-up (Appendix Table 2). Thus, the model may not perform well in borderline scenarios. A way to improve the accuracy of the models is by increasing training data. We recognize the limitation posed by the small sample size used in this preliminary analysis and expect that the accuracy would be improved in future analyses of a larger number of children.

Another limitation of this study is the use of 16S rRNA gene sequencing, which provided considerably limited taxonomic resolution below the genus level due to the high sequence similarity of the 16S rRNA gene, especially with the use of partial amplicons (Caudill and Brayton 2022). For instance, the genus *Streptococcus* was not classified well at the species level (Fig. 2B). In addition, the genus *Streptococcus*

was less abundant than previous reports in 1-y-old children (Dashper et al. 2019). This may be partly due to the resistance of streptococci to different lysis methods (Cho et al. 2021). Since we could not analyze the abundance of *S. mutans* in this data set, we preliminarily analyzed culture-based mutans streptococci count in these samples. We found no significant difference among groups, although children who remained caries-free tended to have lower levels of mutans streptococci ($P = 0.09$, Kruskal-Wallis test; Appendix Fig. 6). In addition, other microorganisms, particularly *Candida*, could contribute to cariogenesis in children (Klinke et al. 2014; Xiao et al. 2018), but they cannot be analyzed with the methods employed here. Future works based on more powerful methods in terms of DNA extraction and sequencing, such as 16S full length–based synthetic long-read sequencing (sFL16S; Jeong et al. 2021) and shallow shotgun metagenomic sequencing (Hillmann et al. 2018), could improve species- and strain-level identification that could lead to more refined biomarker discovery. Extending the microbiome studies to include other microorganisms besides bacteria, especially fungi, would also lead to a more complete picture of the roles of the oral microbiome in cariogenesis and potentially to better predictive tools.

Our study differs from previous reports on caries prediction models based on oral microbiota and supervised machine learning in several aspects and thus would complement existing evidence. First, we focused on 1-y-old children, while others studied older children (2 to 6 y; Teng et al. 2015; Xu et al. 2018; Grier et al. 2021). Since the salivary microbiome of young children changes drastically during early life, it was reported that the microbiome of 1-y-olds differs in terms of $\alpha$ and $\beta$ diversity from other ages from 2 to 48 mo (Dashper et al. 2019). It is likely that the biomarkers of different age groups could differ; thus, we chose to target the population that is key to prevention of early disease. Second, the prediction models by Teng et al. (2015) and Xu et al. (2018) were developed by learning from the differences in the oral microbiota between caries-free and caries-affected children. Therefore, these models made predictions for high risk of ECC based on similarity of the microbiota to that of the caries-active state. However, since the oral microbial community gradually transforms through various stages of caries development, the oral microbiota of precaries and caries-active states are not necessarily similar (Takahashi and Nyvad 2008). In contrast, our models learned exclusively from caries-free children with different future caries status, which is a more direct approach for caries prediction among healthy subjects before caries onset. Third, our model achieved 80% accuracy for caries prediction up to 1 y before caries detection, while previous models by Grier et al. (2021) showed similar levels of accuracy at 6 mo before caries diagnosis but not as accurate when examining enrollment visit samples. Furthermore, for the first time, our study developed prediction models based on a Southeast Asian population. Since the diversity of the oral microbiome varies by geography and race/ethnicity (Gupta et al. 2017), our study provides new information in a population that had not been explored. Although these differences were reported among adults, the oral microbiome may also differ in young children because their salivary microbiomes were similar to those of their mothers (Jo et al. 2021). Two previous studies in the Chinese population identified *Streptococcus* and *Prevotella* as the most discriminatory genera (Teng et al. 2015; Xu et al. 2018), while a study in the United States indicated that *Streptococcus* sp., *Rothia mucilaginosa*, and *Veillonella parvula* were important (Grier et al. 2021). Thus, due to several differences among various studies, in terms of population age, race/ethnicity, geography, sample preparation, and sequencing methodologies, further investigations are required to validate whether these models would perform well across populations.

To initiate such cross-study validation, we examined the performance of our differential-abundance model in a subset of data of children <2 y old from publicly available data of a previous report (Grier et al. 2021). We found that the model showed 58.82% accuracy of prediction (Appendix Fig. 4A). We speculate that the lower accuracy was likely due to several differences between the test and the training data sets. First, the test data set included noncavitated caries for precaries samples, while our training data set considered only cavitated caries as decay. The accuracy was reduced to 50% when analyzing only the precaries samples (Appendix Fig. 4B). Second, the samples were collected from children of different ages (20.5 ± 2.6 mo for test data set vs. 12.6 ± 0.97 mo for training data set). It was shown that the salivary microbiome of 1- and 2-y-old children differs significantly (Dashper et al. 2019). In addition, the differences in the race/ethnicity of study populations and other methodologies may affect microbiome data and model performance. Nevertheless, we found that the relative abundance of *C. concisus* and *P. nanceiensis* appeared to be higher in children who remained caries-free as compared with those with ECC within 2 y, although these were not statistically significant (Appendix Fig. 5A, B).

Altogether, our analysis of saliva samples exclusively from caries-free 1-y-old children results in a promising microbiota-based caries prediction model for this important age group. The results showed that a low abundance of 4 bacterial biomarkers in 1-y-old infants was associated with caries development at 12-mo follow-up. These findings could be further developed to integrate with risk-based caries management systems for targeted prevention to reduce overall incidence of dental caries in young children. In conclusion, the caries prediction model based on salivary microbial biomarkers offers a promising approach to accurate and bias-free caries prediction for 1-y-old children prior to caries onset.

## Author Contributions

R. Raksakmanut, contributed to design, data acquisition, analysis, and interpretation, drafted and critically revised the manuscript; P. Thanyasrisung, contributed to design, data analysis and interpretation, drafted and critically revised the manuscript; S. Sritangsirikul, K. Kitsahawong, contributed to design and data acquisition, critically revised the manuscript; A.L. Seminario, contributed to design and data interpretation, critically revised the manuscript; W. Pitiphat, contributed to conception and design, data interpretation, critically

revised the manuscript; O. Matangkasombut, contributed to conception and design, data analysis and interpretation, drafted and critically revised the manuscript. All authors gave their final approval and agree to be accountable for all aspects of the work.

## Acknowledgments

The authors thank Professor Timothy A. DeRouen, Department of Oral Health Sciences, School of Dentistry, University of Washington, for his contributions and guidance on the cohort study. They also appreciate the suggestions on machine learning from Dr. Ploy N. Pratanwanich, Department of Mathematics and Computer Science, Faculty of Science, and Proadpran Punyabukkana, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Ratchadaphisecksomphot Endowment Fund (Center of Excellence on Oral Microbiology and Immunology) and a Faculty Research Grant (DRF64023), Faculty of Dentistry, Chulalongkorn University Bangkok, Thailand. The original cohort study was funded by the US National Institutes of Health (grant D43TW009071) and the Chronic Inflammatory and Systemic Diseases Associated with Oral Health Research Group, Faculty of Dentistry, Khon Kaen University, Khon Kaen, Thailand. R.R. was partly supported by a scholarship from the College of Dental Medicine, Rangsit University, Pathum Thani, Thailand.

## ORCID iDs

R. Raksakmanut https://orcid.org/0000-0001-5103-3310
P. Thanyasrisung https://orcid.org/0000-0002-1378-695X
O. Matangkasombut https://orcid.org/0000-0003-4029-6830

## References

Al-Hebshi NN, Baraniya D, Chen T, Hill J, Puri S, Tellez M, Hasan NA, Colwell RR, Ismail A. 2019. Metagenome sequencing-based strain-level and functional characterization of supragingival microbiome associated with dental caries in children. J Oral Microbiol. 11(1):1557986.

Casamassimo PS, Thikkurissy S, Edelstein BL, Maiorini E. 2009. Beyond the dmft: the human and economic cost of early childhood caries. J Am Dent Assoc. 140(6):650–657.

Caudill MT, Brayton KA. 2022. The use and limitations of the 16S rRNA sequence for species classification of anaplasma samples. Microorganisms. 10(3):605.

Cho HS, Choi M, Lee Y, Jeon H, Ahn B, Soundrarajan N, Hong K, Kim JH, Park C. 2021. High-quality nucleic acid isolation from hard-to-lyse bacterial strains using PMAP-36, a broad-spectrum antimicrobial peptide. Int J Mol Sci. 22(8):4149.

Christian B, Calache H, Adams G, Hall M, Dashper S, Gibbs L, Gussy M. 2020. A methodological study to assess the measurement properties (reliability and validity) of a caries risk assessment tool for young children. J Dent. 95:103324.

Dashper SG, Mitchell HL, Le Cao KA, Carpenter L, Gussy MG, Calache H, Gladman SL, Bulach DM, Hoffmann B, Catmull DV, et al. 2019. Temporal development of the oral microbiome and prediction of early childhood caries. Sci Rep. 9(1):19732.

Degenhardt F, Seifert S, Szymczak S. 2017. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform. 20(2):492–503.

Diaz-Uriarte R, Alvarez de Andres S. 2006. Gene selection and classification of microarray data using random forest. BMC Bioinform. 7:3.

Fontana M, Carrasco-Labra A, Spallek H, Eckert G, Katz B. 2020. Improving caries risk prediction modeling: a call for action. J Dent Res. 99(11):1215–1220.

Grassl N, Kulak NA, Pichler G, Geyer PE, Jung J, Schubert S, Sinitcyn P, Cox J, Mann M. 2016. Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. Genome Med. 8(1):44.

Grier A, Myers JA, O'Connor TG, Quivey RG, Gill SR, Kopycka-Kedzierawski DT. 2021. Oral microbiota composition predicts early childhood caries onset. J Dent Res. 100(6):599–607.

Gupta VK, Paul S, Dutta C. 2017. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. Front Microbiol. 8:1162.

Hänsel Petersson G, Ericson E, Isberg PE, Twetman S. 2013. Caries risk assessment in young adults: a 3-year validation of clinical guidelines used in public dental service. Acta Odontol Scand. 71(6):1645–1650.

Hemadi AS, Huang R, Zhou Y, Zou J. 2017. Salivary proteins and microbiota as biomarkers for early childhood caries risk assessment. Int J Oral Sci. 9(11):e1.

Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, Knight R, Knights D. 2018. Evaluating the information content of shallow shotgun metagenomics. mSystems. 3(6):e00069-00018.

Hong X, Hu DY. 2010. Salivary *Streptococcus mutans* level: value in caries prediction for 11-12-year-old children. Community Dent Health. 27(4):248–252.

Jeong J, Yun K, Mun S, Chung W-H, Choi S-Y, Nam Y-D, Lim MY, Hong CP, Park C, Ahn YJ, et al. 2021. The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. Sci Rep. 11(1):1727.

Jo R, Yama K, Aita Y, Tsutsumi K, Ishihara C, Maruyama M, Takeda K, Nishinaga E, Shibasaki K, Morishima S. 2021. Comparison of oral microbiome profiles in 18-month-old infants and their parents. Sci Rep. 11(1):861.

Kaakoush NO, Thomas DS, Ruzayqat MM, Lynch D, Leach ST, Lemberg DA, Day AS, Mitchell HM. 2016. *Campylobacter concisus* utilizes blood but not short chain fatty acids despite showing associations with Firmicutes taxa. Microbiology. 162(8):1388–1397.

Klinke T, Urban M, Lück C, Hannig C, Kuhn M, Krämer N. 2014. Changes in *Candida* spp., mutans streptococci and lactobacilli following treatment of early childhood caries: a 1-year follow-up. Caries Res. 48(1):24–31.

Kursa MB, Rudnicki WR. 2010. Feature selection with the Boruta package. J Stat Softw. 36(11):1–13.

Lavrencic P, Kaakoush NO, Huinao KD, Kain N, Mitchell HM. 2012. Investigation of motility and biofilm formation by intestinal *Campylobacter concisus* strains. Gut Pathog. 4(1):22.

Li J, Helmerhorst EJ, Leone CW, Troxler RF, Yaskell T, Haffajee AD, Socransky SS, Oppenheim FG. 2004. Identification of early microbial colonizers in human dental biofilm. J Appl Microbiol. 97(6):1311–1318.

Liu J-F, Hsu C-L, Chen L-R. 2019. Correlation between salivary mutans streptococci, lactobacilli and the severity of early childhood caries. J Dent Sci. 14(4):389–394.

Macuch PJ, Tanner AC. 2000. *Campylobacter* species in health, gingivitis, and periodontitis. J Dent Res. 79(2):785–792.

Marsh PD. 1994. Microbial ecology of dental plaque and its significance in health and disease. Adv Dent Res. 8(2):263–271.

Mejàre I, Axelsson S, Dahlén G, Espelid I, Norlund A, Tranæus S, Twetman S. 2014. Caries risk assessment: a systematic review. Acta Odontol Scand. 72(2):81–91.

Pitts NB, Baez RJ, Diaz-Guillory C, Donly KJ, Alberto Feldens C, McGrath C, Phantumvanit P, Seow WK, Sharkov N, Songpaisan Y, et al. 2019. Early childhood caries: IAPD Bangkok declaration. Int J Paediatr Dent. 86(2):384–386.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12(6):R60.

Simon-Soro A, Mira A. 2015. Solving the etiology of dental caries. Trends Microbiol. 23(2):76–82.

Sritangsirikul S, Kitsahawong K, Matangkasombut O, Seminario AL, DeRouen TA, Pitiphat W. 2021. Breastfeeding duration and the risk of dental caries: a prospective cohort study. Int J Clin Prev Dent. 17(3):189.

Takahashi N, Nyvad B. 2008. Caries ecology revisited: microbial dynamics and the caries process. Caries Res. 42(6):409–418.
Takahashi N, Nyvad B. 2011. The role of bacteria in the caries process: ecological perspectives. J Dent Res. 90(3):294–303.
Teng F, Yang F, Huang S, Bo C, Xu ZZ, Amir A, Knight R, Ling J, Xu J. 2015. Prediction of early childhood caries via spatial-temporal variations of oral microbiota. Cell Host Microbe. 18(3):296–306.
Tinanoff N, Baez RJ, Diaz Guillory C, Donly KJ, Feldens CA, McGrath C, Phantumvanit P, Pitts NB, Seow WK, Sharkov N, et al. 2019. Early childhood caries epidemiology, aetiology, risk assessment, societal burden, management, education, and policy: global perspective. Int J Paediatr Dent. 29(3):238–248.
World Health Organization. 1997. Oral health surveys: basic methods. 4th ed. Geneva (Switzerland): World Health Organization.
Xiao J, Grier A, Faustoferri RC, Alzoubi S, Gill AL, Feng C, Liu Y, Quivey RG, Kopycka-Kedzierawski DT, Koo H, et al. 2018. Association between oral candida and bacteriome in children with severe ECC. J Dent Res. 97(13):1468–1476.
Xu H, Tian J, Hao W, Zhang Q, Zhou Q, Shi W, Qin M, He X, Chen F. 2018. Oral microbiome shifts from caries-free to caries-affected status in 3-year-old Chinese children: a longitudinal study. Front Microbiol. 9:2009.
Zhang L, Budiman V, Day AS, Mitchell H, Lemberg DA, Riordan SM, Grimm M, Leach ST, Ismail Y. 2010. Isolation and detection of *Campylobacter concisus* from saliva of healthy individuals and patients with inflammatory bowel disease. J Clin Microbiol. 48(8):2965–2967.
Zhu C, Yuan C, Ao S, Shi X, Chen F, Sun X, Zheng S. 2018. The predictive potentiality of salivary microbiome for the recurrence of early childhood caries. Front Cell Infect Microbiol. 8:423.