Functional Profiling of Saliva Microbiome is Essential for Oral Cancer Prediction

Jiung-Wen Chen  
National Cheng Kung University

Wei-Fan Chiang  
Chi Mei Medical Center

Jer-Horng Wu (✉ enewujh@ncku.edu.tw)  
National Cheng Kung University  https://orcid.org/0000-0003-4926-2995

Yuh-Ling Chen  
National Cheng Kung University College of Medicine

Wei-Sheng Wu  
National Cheng Kung University

Li-Wha Wu  
National Cheng Kung University College of Medicine

Research

Keywords: oral cancer, microbiome dysbiosis, machine learning, microbial ecology, function

DOI: https://doi.org/10.21203/rs.3.rs-45330/v1

License: ☺️  nf This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Background: The association between microbiome and host disease has been documented in oral cancer, one of the leading cancers worldwide. Huge efforts are made to use the profile of oral microbiome and distinct signature species as markers to distinguish oral cancer patients from healthy individuals. The previous results, however, remain inconclusive. The assembly mechanisms of oral microbiome and their response to changes in oral carcinogenesis also have not been characterized. Here, using 16S rRNA gene amplicon sequencing and in-silico function prediction approaches, we analyzed the saliva microbiome in cohorts of orally healthy, oral verrucous hyperplasia, and oral cancer at taxon and function levels, and compared their corresponding predictive performance of oral cancer using machine learning algorithms.

Results: Analyses of diversity and phylogenetic profiles of bacterial communities in saliva showed that microbiome dysbiosis was significantly linked to oral health status. As oral health deteriorated, the number of core species (>75% prevalence) as a percentage of overall species richness declined. In line with the null model-based analysis, taxonomic and functional assemblies of saliva microbiomes were primarily governed by the stochastic processes. Correspondingly, the quantitative assessment of partitioned beta-diversity suggested extremely high species turnover but low function turnover, revealing a functional redundancy of the oral ecosystem. Functional beta-diversity in salvia microbiome shifted from turnover to nestedness during carcinogenesis of oral verrucous hyperplasia, but this pattern was not observed at the taxon level. Moreover, using both taxon and function data as training features for machine learning-aided prediction on host health status supports a superior predictive performance when using functional profiling. Similar results were also obtained and validated using publicly accessible data.

Conclusions: Our results suggest that altered oral bacterial communities are highly associated with carcinogenesis of oral verrucous hyperplasia. Partly owing to high taxonomic turnover and stochastic assembly processes of the oral ecosystem, discovering oral microbial consortia as universal biomarkers for oral cancer may prove difficult and arduous. Functional profiles are relatively stable and evolve a nestedness pattern during oral carcinogenesis, serving as a new benchmark to study the interplay of the oral microbiome and host health in the future.

Background

Oral cancer (cancer of the lip or oral cavity) causes approximately 180,000 deaths a year worldwide [1]. Oral squamous cell carcinoma (OSCC) accounts for more than 90% of oral cancers [2]. OSCC is one of the most common cancers in males, and the overall 5-year relative mortality rate for OSCC at all stages of diagnosis is approximately 65% in the United States [3]. Since the 5-year survival rate for OSCC is highly stage-dependent: 82% at stage I and 36% at stage IV [4], early-stage OSCC prognosis is favorable, and these patients have a better quality of life after treatment [5]. To better understand the progress of malignancy and to identify OSCC earlier, oral potentially malignant disorders (OPMDs) are studied. OPMD describes a diverse group of conditions, including leukoplakia, erythroplakia, oral submucous fibrosis,
and oral verrucous hyperplasia (OVH), that may precede OSCC development [6]. The presence of OPMD signifies a risk of malignant transformation, though the transformation rate varies, ranging from less than 1% to over 30% [6, 7]. Smoking, alcohol, and betel quid chewing are major risk factors [8]. OVH is one of the most common precancers observed in the oral cavity of betel quid chewers and has the highest transformation rate (up to 21.3%) [6, 7]. Therefore, it is crucial to identify prognostic factors for OSCC development in OPMDs, and saliva can serve as a non-invasive source of biological signatures for OSCC [9].

The microorganisms present in the human oral cavity are referred to as the human oral microbiome, and they play an essential role in the etiology of oral and systemic diseases [10]. Several previous studies have reported the association between oral microbiome and oral cancers, but the results have not always been consistent. For example, the genus *Actinomyces* was reported to be associated with tumor development in one study [11], but the opposite microbial pattern was identified in another study [12]. These inconsistent observations can be attributed to study design, confounding variables, sequencing artifacts, and/or taxonomic resolution [6, 13]. Alternatively, the inconsistency can be explained by microbial “functional redundancy”, defined as the coexistence of different taxa that can perform the same function [14]. A recent systemic review of studies on the microbiome of OSCC patients reported that the tumor-associated microbiome presented similar functional potentials regardless of variations in taxonomic profiles [15]. Thus, taxonomic information in conjunction with functional profiles may shed some light on the variation of the oral microbiome. Continuous efforts to update analysis tools of the microbiome in terms of taxon and function (e.g., finer resolution of taxa [16, 17] and up-to-date metagenome prediction software [18, 19]) should provide opportunities to facilitate the understanding of the interplay between the microbiome and host health.

Microbial ecology concepts have been applied to microbiome studies to evaluate the differences in microbial communities separated by host phenotypes or health status. In exploratory studies of case-control microbiome analysis, researchers have hypothesized that microbiomes differ between host phenotypes, and distinct clusters have been generated when visualized in ordination spaces. However, the stressors produce substantial perturbations to microbes and drive the microbial community assembly into distinct patterns [20]. To decipher how microbiomes respond to stressors, we must first learn the patterns of microbial community assembly and the underlying mechanisms shaping the microbiome. The between-sample diversity (beta diversity) is often used to measure the differences between samples and can be disentangled into nestedness and turnover components: the former is a non-random process of species loss or gain, while the latter is the replacement of some species by others [21]. These patterns are microbial responses to the deterministic (e.g., competition, predation, and environmental filtering) or stochastic processes (e.g., birth, death, migration, and colonization) [22, 23]. It is crucial to quantify the stochasticity of the microbial community assembling in an ecosystem, especially when the primary goal is to identify specific taxa or consortium between phenotypes. However, for the oral microbiome, the aforementioned ecological characteristics and their responses to disease development, such as oral carcinogenesis, have not yet been studied.
In the present study, to understand the changes in microbial communities associated with oral health status, we assessed oral microbiomes of orally healthy (normal), OPMD (specifically OVH), and OSCC cohorts through 16S rRNA amplicon sequencing at a single nucleotide resolution. To investigate the role of ecological patterns in healthy and diseased oral microbiomes, both taxonomic profiles and functional potentials were studied in terms of the dichotomy of beta diversity (nestedness and turnover), along with the stochasticity ratio. We validated our results with publicly available data using the same pipeline and then used machine learning techniques to predict oral health status.

**Methods**

**Study participants and sample collection**

Healthy participants were recruited from the oral screening activity, and saliva samples of OVH and OSCC patients were withdrawn from the biobank in Chi Mei Medical Center (CMMC) in Liouying, Taiwan, with the approval of the Institutional Review Board (No.10612-L02) of CMMC. The saliva collection complied with the following protocol. Participants were interviewed and instructed to refrain from eating, drinking, or using oral hygiene products for at least 1 hour prior to saliva collection, and to rinse their mouth with drinking water. Five minutes after oral rinsing, participants were instructed to spit into a 50 mL centrifuge tube, which was kept on ice, and were cautioned not to cough up sputum. A total of 5 mL of saliva was collected from each participant within a 30-minute time frame. Saliva samples were then centrifuged at 2,600 × g at 4 °C for 15 min. One milliliter of the supernatant was transferred to a new centrifuge tube for other research, and the rest of the saliva supernatant was treated with RNase Inhibitor (Ambion, Austin, TX, USA) and stored at −80 °C for further analysis. The samples were processed and frozen within 30 minutes after collection.

**DNA extraction, PCR, and 16S rRNA gene sequencing**

Bacterial genomic DNA was extracted from saliva samples using a QIAamp DNA Mini Kit (Qiagen, Germany) according to the manufacturer’s spin column protocol. The extracted DNA was amplified using a barcoded *Bacteria*-specific primer set (341F/806R) that targets the V3–V4 hypervariable region of the 16S rRNA gene. The PCR amplicons were sequenced on a MiSeq platform (Illumina, USA) using v3 Chemistry Kits (2 × 300 bp). The detailed sequencing protocol has been described previously [24].

**Bioinformatics analyses**

**16S rRNA sequence processing**

High-throughput amplicon sequencing data were analyzed on the QIIME 2 platform (v2019.4) [25]. After the primers at both ends were trimmed, raw sequences were quality filtered, denoised, merged, and chimera filtered using DADA2 to produce amplicon sequence variants (ASVs) [17]; the maximum number
of expected errors was set at 3. The denoised ASVs with lengths outside the interval between 400 and 450 nt were excluded from the subsequent analysis. Taxonomic annotation of ASVs was performed by a customized classifier trained on the expanded Human Oral Microbiome Database (version 15.1) [26] using the q2-feature-classifier plugin with default settings.

**Diversity analysis**

Alpha and beta diversity indices were calculated using the q2-diversity plugin at a rarefaction depth of 43,313 reads per sample. A phylogenetic tree was constructed using the q2fragmentinsertion plugin for phylogenetic alpha (Faith’s phylogenetic diversity) and beta diversity (UniFrac) measurements. Kruskal–Wallis rank-sum test was used to compare the differences between the alpha diversity indices among cohorts. The contribution of participant age, oral health status (healthy, OVH, and OSCC), and lifestyle factors (alcohol, betel nut, or cigarette consumption) were analyzed using Adonis with 9999 permutations. Distance-based permutational multivariate analysis of variance (PERMANOVA) was used to test the significant difference levels of the centroid of beta diversity metrics among cohorts in the ordination space of principal co-ordinate analysis (PCoA). For the observed significant PERMANOVA results, a permutational test for homogeneity of multivariate dispersion (PERMDISP) was then performed with 9999 permutations to determine the within-group homogeneity of dispersion. The Benjamini-Hochberg procedure was applied to control the false discovery rate (FDR) for multiple testing by statsmodels (0.10.2) [27]. To evaluate the respective contribution of turnover and nestedness components to beta-diversity as a whole, we calculated the multiple-site dissimilarity (Sørensen-based, $\beta_{SØR}$), and the partitioning dissimilarities that accounted only for turnover (Simpson-based, $\beta_{SIM}$) and for nestedness ($\beta_{NES}$) components, respectively [21].

**Core microbiome analysis**

Core microbiome analysis was performed using a customized Python script and visualized using matplotlib-venn (0.11.5). The feature table was first converted to incidence data (presence/absence), and the prevalence of each taxon in each cohort was calculated. If the prevalence of a given taxon was greater than 75%, it was considered a core species in a cohort [28, 29]. A Venn diagram was used to illustrate the distinct and shared core species between cohorts.

**In silico metagenome prediction**

The metagenomic content was developed in silico from the denoised 16S rRNA genes using PICRUSt2 [18]. HMMER (www.hmmer.org), EPA-NG, and GAPPA were performed to place ASVs into reference phylogeny [30, 31]. The castor R package was subsequently used for hidden state prediction to infer gene family copy numbers [32]. Finally, the EC number abundances were predicted based on the adjusted gene family abundances. To infer pathway abundances, MinPath was applied to identify a set of minimum pathways based on the predicted gene families [33]. Default settings were used to regroup EC numbers to MetaCyc reactions and further inferred to MetaCyc pathway abundances [34].

**Statistical testing of differential abundance**
Linear discriminant analysis (LDA) effect size (LEfSe) was applied to identify differentially abundant species and metabolic pathways among cohorts [35]. The input of the frequency matrix was rarefied to the same depth and then transformed into a relative abundance matrix. The significance level was 0.05 for the Kruskal–Wallis test, and the cutoff of the logarithmic LDA scores was 3.

**Stochasticity ratio estimation**

To evaluate the drivers of the microbiome assembly, the stochasticity ratio was estimated using the null-model-based index based on the Bray-Curtis dissimilarity metric, as described previously [36]. For the null model algorithm, proportional taxa/pathway occurrence frequency and richness were applied to generate random microbial/functional communities [37]. Samples in each cohort shared the sample regional taxa/pathway pool in the null model algorithm.

**Public data acquisition, processing, and re-analysis**

Academic search systems, including Google Scholar and PubMed, were used to find studies published in the last five years (2015–2020) with the search terms “oral microbiome”, “saliva microbiome”, and “OSCC”. We included 16S rRNA amplicon-based studies (Table S1) with publicly available sequences and metadata indicating OSCC or control for each sample. To compare the results, we only included studies with samples collected by non-invasive collection methods (oral swab, oral rinse, or saliva samples) and excluded studies that used tissue biopsies and those without sample metadata. Studies with the OSCC cohort consisting of both the oral cavity and oropharynx types were also included. The raw sequence processing, diversity analysis, and core species/metabolic pathway analysis were performed as described in previous sections.

**Prediction using machine learning analysis**

The QIIME2 q2-sample-classifier plugin [38] was used to predict sample health statuses based on taxonomic profiles and predicted functional profiles generated from this study, and the publicly available dataset was re-analyzed. Input data were randomly split into 80% for training and 20% for testing. The Random Forest classifier was applied for supervised machine learning. Cross-validated recursive feature elimination was applied for feature selection, with 5% of features eliminated at each iteration. Hyperparameters were automatically tuned using a random grid search with 5-fold cross-validation. Based on taxonomic and functional profiles, respectively, we performed the analysis procedure 100 times with different random seeds and recorded the testing accuracy ratio for each iteration. The resulting accuracy ratio data was tested using an independent $t$-test to determine the statistical significance of the machine learning prediction results. Area under the receiver-operating characteristic curve (AUROC) metrics were calculated using the scikit-learn package [39]. For multi-class classification, micro-average was used. The data in each study was trained and validated separately to minimize the experimental batch effects.
Results

Cohort descriptions and sequencing quality

Saliva samples collected from 75 male participants, including healthy controls (normal, n = 27), non-recurrent OVH patients with > 8-year follow-up (2011 December–2019 November) (NRVH, n = 21), and patients having primary OVH followed by OSCC development within 8 years follow-up (OSCC, n = 27), were included in this study (Table S2). Statistical analysis of the participants’ metadata (age and lifestyle factors) showed that the differences in the studied cohorts were significant for age between normal and OSCC cohorts, and for betel nut chewing habits between OVH and OSCC cohorts (Table S3). Illumina high-throughput sequencing generated a total of 14,261,633 raw sequences targeting the V3–V4 region of the 16S rRNA gene. After sequence denoising, 8,522,211 denoised reads were retained from 75 samples, with an average of 113,629 ± 33,379 high-quality sequences per library. The plateau rarefaction curves indicated that the sequencing depth was sufficient for downstream analysis (Figure S1).

Diversity analysis of salivary microbiomes

When multiple alpha diversity indices (observed ASVs, Pielou’s evenness, Shannon index, and Faith’s phylogenetic diversity) of the salivary microbial communities were compared among the cohorts, the observed ASVs were significantly higher ($p = 0.044$ before FDR adjustment) only between the healthy (192 ± 44) and the NRVH (164 ± 57) cohorts (Fig. 1A and Table S4). The mean observed ASVs for the OSCC cohort (174 ± 53) were lower than that of the control cohort without statistical significance ($p = 0.119$ before FDR adjustment). However, no significant difference was found for all indices of alpha diversity after FDR adjustment (Table S4). To evaluate the effects of risk factors, including participant age, oral health status, and lifestyle on the changes of the microbial community, UniFrac distance-based Adonis analysis was performed. As shown in Figs. 1B and 1C, oral health status was detected as the strongest explanatory power ($\text{Adonis } R^2 = 0.037$ for unweighted UniFrac and 0.062 for weighted UniFrac), and was the only factor to significantly differentiate the cohorts (FDR-adjusted $p < 0.05$). Age and betel nut chewing habits, although exhibiting significant distinction between some cohorts (Table S3), were not significantly associated with changes in oral microbial communities. The UniFrac-based beta diversity distribution of salivary microbiomes from the cohorts was visualized using a PCoA plot (Fig. 1D-E), and showed random distribution on the ordination space. Pairwise PERMDISP analysis further confirmed that the dispersion effect was not found among cohorts based on the unweighted UniFrac metric ($p_{\text{PERMDISP}} = 0.1669$); however, this effect was observed between normal and diseased (NRVH/OSCC) cohorts. In particular, the dispersion effect reached a significant level between the normal and NRVH cohorts (FDR-adjusted $p_{\text{PERMDISP}} = 0.0362$) in the weighted UniFrac distance measurement, suggesting heterogeneous dispersion of abundant taxa in salivary microbiota in correspondence with the oral health status.

Core salivary microbiome
To further compare the differences in salivary microbiomes among cohorts, we investigated the “core” species, defined as the taxa commonly present in the saliva of each cohort with prevalence > 75% [28, 29]. The number of core species was 55 (67.14 ± 11.06% of total abundance), 39 (47.24 ± 14.96% of total abundance), and 30 (44.52 ± 12.81% of total abundance) in normal, NRVH, and OSCC cohorts, respectively, and 24 species taxa were universal in the saliva samples from all cohorts, even when the oral health status altered (Fig. 2A). Five species (Anaeroglobus geminatus, Porphyromonas gingivalis, Prevotella oulorum, Saccharibacteria (TM7) [G5] bacterium HMT-356, and Tannerella forsythia) were specific to the NRVH cohort, while two species (Capnocytophaga sputigena and Catonella morbi) were specific to the OSCC cohort. One species (Dialister invisus) was specific to the NRVH and OSCC cohort. Interestingly, a decreased trend in overall core species richness (gamma diversity) was clearly observed with deteriorating oral health status (Fig. 2B): from 12.53% in the normal cohort to 6.34% in the OSCC cohort.

The LEfSe analysis revealed the core species that were significantly abundant in each group (p < 0.05, LDA score > 10^3) (Fig. 2C). A total of 15 species taxa were enriched in the normal cohort compared to the two diseased cohorts. These included two unclassified taxa related to Prevotella and Selenomonas genera, three uncultured species (Saccharibacteria (TM7) [G-1] bacterium HMT-352, Leptotrichia sp. HMT-417, and Actinomyces sp. HMT-180), and 10 known species: Prevotella melaninogenica, Veillonella atypica, Streptococcus salivarius, Haemophilus parainuenzae, Veillonella dispar, Prevotella pallens, Megasphaera micronuciformis, Campylobacter concisus, Prevotella salivae, and Rothia mucilaginosa. For the NRVH cohort, only Veillonella parvula and Rothia dentocariosa were significantly abundant (p < 0.05, LDA score > 10^3), while four species, Capnocytophaga sputigena, Prevotella oris, Peptostreptococcus stomatis, and Parvimonas micra, were abundant specifically in the OSCC cohort. The results as a whole show the alteration in the salivary microbiome with the progress of oral carcinogenesis.

**Differences in predicted metagenome among cohorts**

We applied PICRUSt2, an updated version of a widely used metagenomic prediction tool [40], to infer the functional profiles of the microbial communities using the denoised ASVs. The nearest-sequenced taxon index (NSTI) of 81.68% of reads was less than 0.15 (Figure S2A), suggesting the high-quality metagenome predicted [40]. LEfSe analysis identified 26, 7, and 24 inferred pathways that were significantly abundant specific to normal, NRVH, and OSCC cohorts, respectively (Figure S2B). By categorizing these pathways to higher classes, we found that most of them belong to amino acid biosynthesis (10 pathways) and cofactor, prosthetic group, electron carrier, and vitamin biosynthesis (8 pathways) for the normal cohort; while cell structure biosynthesis (5 pathways), fatty acid and lipid biosynthesis (4 pathways), and nucleoside and nucleotide metabolism (3 pathways for biosynthesis; 2 for degradation) were abundant in the OSCC cohort (Fig. 3). Only three pathways belonging to TCA cycles and nucleic acid processing were found to be significantly higher in abundance in specific relation to the NRVH cohort.
Partitioning turnover and nestedness

To determine the ecological drivers shaping the salivary microbiomes among cohorts, we quantitatively compared the assemblage dissimilarities of the salivary microbiomes based on taxonomic and functional profiles. For assemblage assessment of taxonomic data, the species nestedness was 0.042, 0.058, and 0.043, (sustainably lower than the turnover: 0.822, 0.802, and 0.841) for the normal, NRVH, and OSCC cohorts, respectively (Fig. 4), showing that the differentiation of salivary microbiomes was influenced by the species turnover. By contrast, the functional profiles of the salivary microbiomes were relatively stable: the mean multi-site Sørensen dissimilarity related to pathways (0.456 ± 0.044) was lower than that related to species taxa (0.869 ± 0.013) by approximately 50%. Notably, the numerical distributions of function nestedness (0.180, 0.187, and 0.295) and function turnover (0.291, 0.209, and 0.204) were relatively similar in each of the three cohorts (normal, NRVH, and OSCC, respectively), as compared to species-based Sørensen dissimilarity (Fig. 4). Conspicuously, the ratio of the function nestedness to turnover increased from 0.620 for normal, 0.896 for NRVH, to 1.449 for OSCC, suggesting that the effect of nestedness gradually dominates the functional differentiation of salivary microbiomes during the process of oral carcinogenesis.

To validate these findings, a null model-based analysis of stochasticity with taxonomic and functional profiles was performed. Figure S3 shows that the stochastic process dominated the salivary microbiome assembly, with a higher stochastic ratio of inferred pathways (85.97 ± 16.73% to 93.04 ± 4.02%) than that of species taxa (61.88 ± 13.75 to 64.44 ± 13.11%). The distribution patterns of taxa-based stochastic ratio among cohorts were relatively similar. By contrast, a far wider distribution spectra for the function-based stochastic ratios were observed for the OSCC cohort than for the normal cohort. This finding supports a shift of the underlying mechanism of the functional alternation towards the deterministic process. Regardless of the observed high species turnover of saliva microbial communities, the dominance of functional nestedness in the OSCC cohort suggests that a set of distinct microbial functions, which may be associated with oral carcinogenesis, evolved accordingly in the oral cavity.

Comparisons with other studies

To generalize the idea of using the functional potential as a valid signature for oral cancer detection, we compared the taxonomic and functional diversity data of this study with those of previous studies. We included two studies for the comparison (see justifications in Table S1). PCoA plots (Fig. 5A-B) based on Bray-Curtis distance metric revealed that the percentage of variance explained by the first two axes was higher (47.06%) using functional profiles than that using taxonomic profiles (20.61%). The PERMANOVA results showed that at least one cohort was significantly different from the rest of the cohorts (namely, normal vs. NRVH and OSCC) in both taxonomic ($p_{\text{PERMANOVA}}=0.0001$, Fig. 5A) and functional profiles ($p_{\text{PERMANOVA}}=0.0001$, Fig. 5B). However, the PERMDISP results suggest heterogeneous dispersion in the taxonomic profile ($p_{\text{PERMDISP}}=0.0061$, Fig. 5A), but not in the functional profile ($p_{\text{PERMDISP}}=0.1267$, Fig. 5B). Although limited data regarding the pre-cancer cohorts are available from previous studies, the
results shown here further support the concept that the functional profile of the oral microbiota is more effectively related to the oral health status.

Because all the previous studies focused on the analysis of 16S rRNA sequences with a 97% similarity threshold using a clustering approach, many potential signature taxa were reported at the genus level. We re-analyzed the sequences with the same analytical pipeline to achieve phylogenetic resolution down to the species level. Although the core species (> 75% prevalence) were found in these studies, no taxon could be identified across all three studies for the normal and OSCC cohorts (Fig. 5C-D). Instead, we found a comparatively high number of pathways between studies when comparing functional profiles, although there was still no common pathway across studies obtained for the normal cohort (Fig. 5E). For the OSCC cohort, four pathways involved in nucleotide biosynthesis (UMP biosynthesis I) and cell structure biosynthesis (UDP-N-acetylmuramoyl-pentapeptide biosynthesis I and II, and peptidoglycan biosynthesis I) were prevalent across the studies (Fig. 5F).

**Machine learning to predict oral health status**

To further study the discriminability of taxonomic and functional profiles for predicting the occurrence of oral cancer, we trained random forest models with both profiles separately. The datasets from each study were processed independently in accordance with the analytical process illustrated in Fig. 6A. The mean accuracy ratio (the mean of the ratios of predicted accuracy to the accuracy of random guess, 100 iterations) was higher when using the functional profile compared to when using the taxonomic profile (Fig. 6B and Figure S4), although statistical significance was detected in only two of the three datasets (Wolf et al. and Zhao et al., with $p = 0.043$ and 0.001, respectively). In addition, we bootstrapped the receiver-operating characteristic (ROC) analysis 100 times and demonstrated that, for the three datasets, AUROC was higher when functional profiles were used to distinguish OSCC samples from healthy controls, compared when using taxonomic profiles (Fig. 6C-D).

**Discussion**

In this study, we found that altered oral microbiomes were significantly associated with oral health status. A number of previous studies have reported the association between the change of oral microbiome and oral health status, as well as risk factors including betel nut chewing, cigarette smoking, and alcohol consumption [41–44]. A recent investigation of multiple factors that potentially affect the oral microbiome assemblage between nasopharyngeal carcinoma patients and healthy controls showed that differences in oral microbial communities were most associated with oral health status and smoking [45]. In agreement with previous studies, our findings also showed that oral microbiome may be linked to the progression of OVH carcinogenesis, and the statistical significance was especially strong when more weights were put on rare ASVs (unweighted UniFrac measure). Cigarette smoking and other risk factors like betel nut usage and alcohol drinking, however, were not the main contributors to differences in oral microbiomes in the present study. Additionally, although the within-group diversity was not significantly
different among cohorts, a decreased trend of core microbiome from healthy or OVH to OSCC was observed when the total diversity of each cohort (gamma diversity) was considered. Consistent with our results, previous studies have reported that 9.6–13.1% of detected bacterial species are common among healthy individuals [28, 29]. Gamma diversity can be an index related to oral health status. Together, our findings showed dysbiosis of the core salivary microbiome as the patients developed OSCC, implying that the salivary microbiome is linked to oral health status.

More specifically, several enriched species were associated with the health status of the oral cavity. Previous studies [12, 46–48] have reported that the bacterial genera *Lautropia, Rothia, Streptococcus*, and *Veillonella* were enriched in the healthy control or pre-cancer lesion groups, whereas the genera *Fusobacterium, Capnocytophaga, Dialister, Selenomonas, Parvimonas, Peptostreptococcus*, and *Treponema* exhibited higher abundance in the OSCC cohort. In line with these studies, we also identified signature species belonging to the genera *Rothia, Streptococcus*, and *Veillonella* within the non-cancer cohorts, while other species belonging to *Capnocytophaga* and *Parvimonas* were more abundant specifically in the OSCC cohort. However, several species (*Prevotella melaninogenica, Streptococcus salivarius*, and *Rothia mucilaginosa*) that were highly abundant in patients with OPMD or OSCC [49–51] are identified as signatures for the healthy cohort in our study. Besides, many genera/species found in the present study have not been previously reported, and vice versa. This inconsistency between studies could be attributed to the experimental design (e.g., sample types or hypervariable regions of the 16S rRNA gene [52]), the bioinformatics analysis pipeline (e.g., sequence denoising approaches [53] and reference databases [54]), the genetics of studied cohorts (e.g., racial factors [55]), and the complexity of oral carcinogenesis [56]. Alternatively, our results suggest that the inconsistencies may be due to the extremely high species turnover of oral ecosystems, and the dominance of the stochastic process in shaping microbial communities in saliva (Fig. 4 and Figure S3). As revealed by the component partitioning of dissimilarity in the present study, microbes in saliva are subject to a high population dynamic [57], representing an extraordinarily dynamic ecosystem; consequently, detecting bacterial species as stable or universal signatures to reflect the oral health status can be challenging.

We found that the stable function profiles remained distinguishable, despite the high variation and fluctuation of the microbial populations. Previous studies have illustrated that microbial communities of the oral cavity are highly dynamic and individual-specific [58–60]. This is not surprising; the oral cavity represents an open environment that is exposed to exogenous bacteria and rapid changes in environmental factors in addition to host effects [61]. The stable function profiles but highly varied species composition in the salivary ecosystem may be accounted for by functional redundancy that has been proposed for other microbial ecosystems, such as soil and the human gut [62, 63].

To our knowledge, this is the first report to disentangle the contribution of the turnover and nestedness of both taxonomic and functional compositions in three different states of oral cavity ecosystem (orally healthy, OPMD, and OSCC). Notably, the primary mechanism contributing to the differences in functional profiles shifted from turnover to nestedness in accordance with OVH carcinogenesis (Fig. 4). Since nestedness reflects the effect of environmental filtering [64, 65], a possible explanation is that the
diseased status of the oral cavity (OVH and OSCC) was a more influential environmental filtering factor than the healthy one, resulting in the loss of specific metabolic niches and an increase in nestedness components during the shift from healthy to diseased status. This can be further supported by the null-model-based stochasticity analysis (Figure S3) that revealed a more deterministic influence, which has been linked to the nestedness effect [66], in the metabolic pathway profiles in the NRVH and OSCC cohorts compared to healthy controls.

The re-analysis of OSCC-associated oral microbiomes from publicly accessible studies reinforced the use of functional profiles for more consistent results. The relationship between oral microbiome and OSCC remains uncertain and has not been intensively studied. To circumvent various technical challenges that potentially yielded different results and inconclusive interpretations, we re-analyzed the raw sequences with the same bioinformatic analysis pipeline, which generated more comparable results across different studies, as shown in the function-based PCoA ordination (Fig. 5). These more comparable results may be partly attributable to the bias (due to different hypervariable regions of 16S rRNA) being alleviated by inserting high-resolution amplicons into a reference phylogenetic tree [67], an improved inference process included in PICRUSt2. However, the accuracy of the taxonomic classifier based on phylogenetic inference was varied to some extent; meanwhile, precisely comparing the ASVs of different sequence regions in taxonomy remains challenging [68]. Thus, it is not surprising to observe a within-group dispersion effect based on taxonomic profiles and to identify no common core taxon among three studies in healthy control and OSCC cohorts, respectively. No common core metabolic pathway was identified in healthy controls either. This is counterintuitive; oral microbiota should possess a number of essential common metabolic pathways for survival [69]. Likely, the dietary patterns and lifestyles of the control groups from different geographical regions can serve as a stochastic factor that drives the functional profiles to be dissimilar to one another [70]. As for the functional profiles in the OSCC cohorts, the common core pathways (nucleotide biosynthesis and bacterial cell structure) are related to generate building blocks for DNA and RNA synthesis as well as cell wall formation. Unlike the pathways for metabolites such as fatty acids and amino acids, which can be utilized exogenously by the host cell to generate complex host-microbe interactions [71, 72], the enrichment of these common pathways could be the consequence of functional alterations of microbial communities in response to changing oral health status. Further study is required to investigate the relationship between enriched metabolic pathways and the host. Raw sequence sharing, along with completely publicly available metadata, will also enable to reproduce, compare, and validate results across studies through meta-analysis. This is especially crucial for hard-to-obtain disease-associated samples (e.g., OPMD and OSCC in the present study) in order to achieve clinical application of disease progression prediction.

In several previous studies, machine learning-aided models were trained for disease prediction using taxonomic profiles derived from 16S rRNA genes [73–75]. Since the saliva microbiome was characterized as a highly dissimilar, high species turnover, and stochasticity-dominated entity (Fig. 4 and Figure S3), using taxonomic data as input features for classification and prediction tasks can be suboptimal. However, the functional potential of the saliva microbiome was much more similar and displayed a nestedness pattern, where deterministic processes gradually dominated as the oral health status
deteriorated. In this study, we demonstrated that predictive accuracy could be improved by “transforming” the taxonomic data into functional profiles with a widely used metagenome prediction tool. The consistently higher mean accuracy using functional profiles as the input feature could be primarily attributed to the lower variation of the pathway data. Although it has long been discussed whether functional profiles outperform taxonomic profiles for microbiome classification tasks [76], the predicted functional profiles from OSCC-associated individuals were reported similar despite the variation of the taxonomic profiles [15]. Besides, a recent shotgun metagenomic study [77] also supports our results by discriminating healthy controls from dental caries cohorts more effectively using pathways than taxa. In contrast to our findings, another study with gut microbiome datasets [78] showed a worse performance from function-based searches compared to those using the taxa-based strategy. This conflict may be due to the lower stability of oral microbiomes relative to gut microbiomes [79]. Also, the updated metagenome prediction tool improved its accuracy by allowing genome prediction via ASVs, expanding the database of reference genomes, and employing a more stringent method of generating pathway-level functional profiles [18]. These technical improvements make it successful for predicting phenotypes via functional profiles in our study. One limitation of using predicted functional profiles is the loss of function data attributed to a fraction of microbes without genomic information. This can be improved in the future by incorporating both metagenomics and culturomics to expand the microbiome database [80].

Conclusions

Overall, this study has revealed that altered oral microbial communities are highly associated with oral health status (normal, OVH, and OVH-associated OSCC). From the perspective of microbial ecology, because the oral ecosystems are high taxonomic turnover (i.e., high variance and fluctuation), any attempt of discovering oral microbial consortia as biomarkers for oral cancer would be a daunting task. Compared to taxonomic profiles, functional profiles are relatively stable and display a nestedness pattern in the process of oral carcinogenesis, thus serving as a suitable alternative for OSCC signatures. Machine learning models yielded a better mean predictive performance by using functional profiles as training data when compared to using taxonomic profiles. The re-analysis of datasets from the published microbiomes further confirmed the feasibility of using functional profiles for OSCC prediction. This function-based method serves as a new benchmark to study the interplay of the oral microbiome and host health in the future. Further studies are warranted to elucidate the association and causation between the ecological pattern of microbiota and the host, as well as how to improve the predictive performance by understanding the ecological assemblage of the microbiome.

Abbreviations

ASV: Amplicon sequence variant
AUROC: Area under the receiver-operating characteristic curve
FDR: False discovery rate
LDA: Linear discriminant analysis
LEfSe: Linear discriminant analysis effect size
NSTI: Nearest-sequenced taxon index
NRVH: Non-recurrent verrucous hyperplasia
OPMD: Oral potentially malignant disorder
OSCC: Oral squamous cell carcinoma
OVH: Oral verrucous hyperplasia
PCoA: Principal co-ordinate analysis
PERMANOVA: Permutational multivariate analysis of variance
PERMDISP: Permutation analyses of multivariate dispersions
ROC: Receiver-operating characteristic

Declarations

Ethics approval and consent to participate

All participants in the present study were recruited from the Chi Mei Medical Center (CMMC) in Liouying, Taiwan, with the approval of the Institutional Review Boards of CMMC.

Consent for publication

Not applicable.

Availability of data and material

Raw sequence data obtained in this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB39064. QIIME 2 artifacts and analysis scripts used in this study are available under open-source licenses from https://github.com/mft-lab/OVH-study.

Competing interests

The authors declare that they have no competing interests.
Funding

This study was supported by the Ministry of Science and Technology, Taiwan (grant numbers 107-2321-B-006-007, 108-2321-B-006-010, and 108-2221-E-006 -160 -MY3).

Authors' contributions

JHW, YLC, and LWW conceived the study. JHW designed the analysis. WFC collected and processed the saliva samples. JWC performed experiments and data analysis. JWC and JHW drafted the manuscript. JHW and LWW revised the manuscript and edited the final version. All authors helped edit the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Gina Lee's assistance in sample collection, and Dr. Wei-Yu Chen and Mr. Han-De Chen for technical assistance in the microbiome analysis. This manuscript was edited by Enago Editing.

References

1. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Pineros M, Znaor A, Bray F: Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019, 144(8):1941-1953.
2. Markopoulos AK: Current aspects on oral squamous cell carcinoma. Open Dent J 2012, 6:126-130.
3. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2020. Ca-Cancer J Clin 2020, 70(1):7-30.
4. Liu SY, Lu CL, Chiu CT, Yen CY, Liaw GA, Chen YC, Liu YC, Chiang WF: Surgical outcomes and prognostic factors of oral cancer associated with betel quid chewing and tobacco smoking in Taiwan. Oral Oncol 2010, 46(4):276-282.
5. Bilal S, Doss JG, Cella D, Rogers SN: Quality of life associated factors in head and neck cancer patients in a developing country using the FACT-H&N. J Cranio Maxill Surg 2015, 43(2):274-280.
6. Warnakulasuriya S: Oral potentially malignant disorders: A comprehensive review on clinical aspects and management. Oral Oncol 2020, 102:104550.
7. Chiang WF, Liu SY, Lin JF, Chiu SF, Gou SB, Chiu CT, Chang CH: Malignant development in patients with oral potentially malignant disorders detected through nationwide screening: Outcomes of 5-year follow-up at a single hospital. Head & neck 2020, 42(1):67-76.
8. Lin WJ, Jiang RS, Wu SH, Chen FJ, Liu SA: Smoking, alcohol, and betel quid and oral cancer: a prospective cohort study. J Oncol 2011, 2011:525976.
9. Tang HK, Wu ZY, Zhang JP, Su B: Salivary IncRNA as a potential marker for oral squamous cell carcinoma diagnosis. Mol Med Rep 2013, 7(3):761-766.
10. Gao L, Xu TS, Huang G, Jiang S, Gu Y, Chen F: Oral microbiomes: more and more importance in oral cavity and whole body. *Protein Cell* 2018, 9(5):488-500.

11. Mukherjee PK, Wang H, Retuerto M, Zhang H, Burkey B, Ghannoum MA, Eng C: Bacteriome and mycobiome associations in oral tongue cancer. *Oncotarget* 2017, 8(57):97273-97289.

12. Zhao HS, Chu M, Huang ZW, Yang X, Ran SJ, Hu B, Zhang CP, Liang JP: Variations in oral microbiota associated with oral cancer. *Sci Rep-Uk* 2017, 7.

13. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ: Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* 2017, 8.

14. Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O’Connor MI, Ackermann M, Hahn AS, Srivastava DS, Crowe SA et al: Function and functional redundancy in microbial systems. *Nat Ecol Evol* 2018, 2(6):936-943.

15. Al-Hebshi NN, Borgnakke WS, Johnson NW: The microbiome of oral squamous cell carcinomas: a functional perspective. *Current Oral Health Reports* 2019, 6(2):145-160.

16. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, Kightley EP, Thompson LR, Hyde ER, Gonzalez A et al: Deblur rapidly resolves single-nucleotide community sequence patterns. *Msystems* 2017, 2(2).

17. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP: DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016, 13(7):581-583.

18. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI: PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* 2020, 38(6):685-688.

19. Wemheuer F, Taylor JA, Daniel R, Johnston E, Meinicke P, Thomas T, Wemheuer B: Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ Microbiome* 2020, 15(1).

20. Zaneveld JR, McMinds R, Thurber RV: Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nat Microbiol* 2017, 2(9).

21. Baselga A: Partitioning the turnover and nestedness components of beta diversity. *Global Ecol Biogeogr* 2010, 19(1):134-143.

22. Stegen JC, Freestone AL, Crist TO, Anderson MJ, Chase JM, Comita LS, Cornell HV, Davies KF, Harrison SP, Hurlbert AH et al: Stochastic and deterministic drivers of spatial and temporal turnover in breeding bird communities. *Global Ecol Biogeogr* 2013, 22(2):202-212.

23. Zhou JZ, Ning DL: Stochastic community assembly: Does it matter in microbial ecology? *Microbiol Mol Biol R* 2017, 81(4).

24. Chen W-Y, Ng TH, Wu J-H, Chen J-W, Wang H-C: Microbiome dynamics in a shrimp grow-out pond with possible outbreak of acute hepatopancreatic necrosis disease. *Sci Rep-Uk* 2017, 7(1):9395.

25. Bolyen E, Rideout JR, Dillon MR, Bokulich N, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F et al: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019, 37(8):852-857.
26. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP: New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *Msystems* 2018, 3(6).

27. Seabold S, Perktold J: Statsmodels: Econometric and statistical modeling with python. In: *Proceedings of the 9th Python in Science Conference: 2010*. Scipy: 61.

28. Willis JR, Gonzalez-Torres P, Pittis AA, Bejarano LA, Cozzuto L, Andreu-Somavilla N, Alloza-Trabado M, Valentin A, Ksiezopolska E, Company C et al: Citizen science charts two major "stomatotypes" in the oral microbiome of adolescents and reveals links with habits and drinking water composition. *Microbiome* 2018, 6.

29. Takeshita T, Kageyama S, Furuta M, Tsuboi H, Takeuchi K, Shibata Y, Shimazaki Y, Akifusa S, Ninomiya T, Kiyohara Y et al: Bacterial diversity in saliva and oral health-related conditions: the Hisayama Study. *Sci Rep-Uk* 2016, 6.

30. Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Fouri T, Stamatakis A: EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic biology* 2018, 68(2):365-369.

31. Czech L, Stamatakis A: Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. *Plos One* 2019, 14(5):e0217050.

32. Louca S, Doebeli M: Efficient comparative phylogenetics on large trees. *Bioinformatics* 2017, 34(6):1053-1055.

33. Ye Y, Doak TG: A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *Plos Comput Biol* 2009, 5(8):e1000465.

34. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD: The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Research* 2020, 48(D1):D445-D453.

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

36. Ning DL, Deng Y, Tiedje JM, Zhou JZ: A general framework for quantitatively assessing ecological stochasticity. *P Natl Acad Sci USA* 2019, 116(34):16892-16898.

37. Gotelli NJ: Null model analysis of species co-occurrence patterns. *Ecology* 2000, 81(9):2606-2621.

38. Bokulich NA, Dillon MR, Bolyen E, Kaehler BD, Huttley GA, Caporaso JG: q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J Open Res Softw* 2018, 3(30).

39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V: Scikit-learn: Machine learning in Python. *Journal of machine learning research* 2011, 12(Oct):2825-2830.

40. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Thurber RLV, Knight R et al: Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013, 31(9):814-821.
41. Wu J, Peters BA, Dominianni C, Zhang YL, Pei ZH, Yang LY, Ma YF, Purdue MP, Jacobs EJ, Gapstur SM et al: Cigarette smoking and the oral microbiome in a large study of American adults. *Isme J* 2016, 10(10):2435-2446.

42. Hernandez BY, Zhu XM, Goodman MT, Gatewood R, Mendiola P, Quinata K, Paulino YC: Betel nut chewing, oral premalignant lesions, and the oral microbiome. *Plos One* 2017, 12(2).

43. Fan XZ, Peters BA, Jacobs EJ, Gapstur SM, Purdue MP, Freedman ND, Alekseyenko AV, Wu J, Yang LY, Pei ZH et al: Drinking alcohol is associated with variation in the human oral microbiome in a large study of American adults. *Microbiome* 2018, 6.

44. Sampaio-Maia B, Caldas IM, Pereira ML, Perez-Mongiovi D, Araujo R: The oral microbiome in health and its implication in oral and systemic diseases. *Advances in Applied Microbiology, Vol 97* 2016, 97:171-210.

45. Debelius JW, Huang T, Cai Y, Ploner A, Barrett D, Zhou X, Xiao X, Li Y, Liao J, Zheng Y: Sub-species niche specialization in the oral microbiome is associated with nasopharyngeal carcinoma risk in an endemic area of southern China. *bioRxiv* 2019:782417.

46. Schmidt BL, Kuczynski J, Bhattacharya A, Huey B, Corby PM, Queiroz ELS, Nightingale K, Kerr AR, DeLacure MD, Veeramachaneni R et al: Changes in abundance of oral microbiota associated with oral cancer. *Plos One* 2014, 9(6).

47. Lee WH, Chen HM, Yang SF, Liang C, Peng CY, Lin FM, Tsai LL, Wu BC, Hsin CH, Chuang CY et al: Bacterial alterations in salivary microbiota and their association in oral cancer. *Sci Rep-Uk* 2017, 7.

48. Wolf A, Moissl-Eichinger C, Perras A, Koskinen K, Tomazic PV, Thurnher D: The salivary microbiome as an indicator of carcinogenesis in patients with oropharyngeal squamous cell carcinoma: A pilot study. *Sci Rep-Uk* 2017, 7.

49. Amer A, Whelan A, Al-Hebshi NN, Healy CM, Moran GP: Acetaldehyde production by Rothia mucilaginosa isolates from patients with oral leukoplakia. *J Oral Microbiol* 2020, 12(1):1743066.

50. Pushalkar S, Ji X, Li Y, Estilo C, Yegnanarayana R, Singh B, Li X, Saxena D: Comparison of oral microbiota in tumor and non-tumor tissues of patients with oral squamous cell carcinoma. *Bmc Microbiol* 2012, 12:144.

51. Mager DL, Haffajee AD, Devlin PM, Norris CM, Posner MR, Goodson JM: The salivary microbiota as a diagnostic indicator of oral cancer: A descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. *J Transl Med* 2005, 3.

52. Tremblay J, Singh K, Fern A, Kirton ES, He SM, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG: Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* 2015, 6.

53. Nearing JT, Douglas GM, Comeau AM, Langille MGI: Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *Peerj* 2018, 6.

54. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D et al: Best practices for analysing microbiomes. *Nat Rev Microbiol* 2018, 16(7):410-422.
55. Yang YH, Zheng W, Cai QY, Shrubsole MJ, Pei ZH, Brucker R, Steinwandel M, Bordenstein SR, Li ZG, Blot WJ et al: **Racial differences in the oral microbiome: Data from low-income populations of African ancestry and European ancestry.** *Msystems* 2019, 4(6).

56. Tanaka T, Ishigamori R: **Understanding carcinogenesis for fighting oral cancer.** *J Oncol* 2011, 2011:603740.

57. Xu X, He JZ, Xue J, Wang Y, Li K, Zhang KK, Guo Q, Liu XH, Zhou Y, Cheng L et al: **Oral cavity contains distinct niches with dynamic microbial communities.** *Environ Microbiol* 2015, 17(3):699-710.

58. Hall MW, Singh N, Ng KF, Lam DK, Goldberg MB, Tenenbaum HC, Neufeld JD, Beiko RG, Senadheera DB: **Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity.** *Npj Biofilms and Microbiomes* 2017, 3.

59. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N et al: **Moving pictures of the human microbiome.** *Genome Biol* 2011, 12(5).

60. Mukherjee C, Beall CJ, Griffen AL, Leys EJ: **High-resolution ISR amplicon sequencing reveals personalized oral microbiome.** *Microbiome* 2018, 6.

61. Wade WG: **The oral microbiome in health and disease.** *Pharmacol Res* 2013, 69(1):137-143.

62. Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Yunta RG, Okuda S, Vandeputte D, Valles-Colomer M, Hildebrand F, Chaffron S et al: **Species-function relationships shape ecological properties of the human gut microbiome.** *Nat Microbiol* 2016, 1(8).

63. Mendes LW, Tsai SM, Navarrete AA, de Hollander M, van Veen JA, Kuramae EE: **Soil-borne microbiome: Linking diversity to function.** *Microbial Ecol* 2015, 70(1):255-265.

64. Chase JM, Myers JA: **Disentangling the importance of ecological niches from stochastic processes across scales.** *Philos T R Soc B* 2011, 366(1576):2351-2363.

65. Daniel J, Gleason JE, Cottenie K, Rooney RC: **Stochastic and deterministic processes drive wetland community assembly across a gradient of environmental filtering.** *Oikos* 2019, 128(8):1158-1169.

66. Amend AS, Cobian GM, Laruson AJ, Remple K, Tucker SJ, Poff KE, Antaky C, Boraks A, Jones CA, Kuehu D et al: **Phytobiomes are compositionally nested from the ground up.** *PeerJ* 2019, 7.

67. Janssen S, McDonald D, Gonzalez A, Navas-Molina JA, Jiang LJ, Xu ZJZ, Winker K, Kado DM, Orwell E, Manary M et al: **Phylogenetic placement of exact amplicon sequences improves associations with clinical information.** *Msystems* 2018, 3(3).

68. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttenhower C, Caporaso JG: **Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin.** *Microbiome* 2018, 6.

69. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS et al: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, 486(7402):207-214.

70. Gupta VK, Paul S, Dutta C: **Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity.** *Front Microbiol* 2017, 8.
71. Koh A, De Vadder F, Kovatcheva-Datchary P, Backhed F: From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. Cell 2016, 165(6):1332-1345.

72. Rohrig F, Schulze A: The multifaceted roles of fatty acid synthesis in cancer. Nat Rev Cancer 2016, 16(11):732-749.

73. Xu H, Tian J, Hao WJ, Zhang Q, Zhou Q, Shi WH, Qin M, He XS, Chen F: Oral microbiome shifts from caries-free to caries-affected status in 3-year-old Chinese children: a longitudinal study. Front Microbiol 2018, 9.

74. Teng F, Yang F, Huang S, Bo CP, Xu ZZ, Amir A, Knight R, Ling JQ, Xu J: Prediction of early childhood caries via spatial-temporal variations of oral microbiota. Cell Host & Microbe 2015, 18(3):296-306.

75. Ai LY, Tian HY, Chen ZF, Chen HM, Xu J, Fang JY: Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. Oncotarget 2017, 8(6):9546-9556.

76. Xu ZJ, Malmer D, Langille MGI, Way SF, Knight R: Which is more important for classifying microbial communities: who's there or what they can do? Isme J 2014, 8(12):2357-2359.

77. Baker JL, Morton JT, Dinis M, Alvarez R, Tran NC, Knight R, Edlund A: Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. bioRxiv 2019:804443.

78. Su XQ, Jing GC, Sun Z, Liu L, Xu ZJ, McDonald D, Wang ZB, Wang HL, Gonzalez A, Zhang YF et al: Multiple-disease detection and classification across cohorts via microbiome search. Msystems 2020, 5(2).

79. Ding T, Schloss PD: Dynamics and associations of microbial community types across the human body. Nature 2014, 509(7500):357-360.

80. Bilen M, Dufour JC, Lagier JC, Cadoret F, Daoud Z, Dubourg G, Raoult D: The contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. Microbiome 2018, 6.

Figures
Figure 1

A

Observed ASVs

Normal (n=27) NRVH (n=21) OSCC (n=27)

B

Host status
Cigarette
Betel nut
Age
Alcohol

Adonis R²

C

Host status
Cigarette
Age
Betel nut
Alcohol

Adonis R²

D

PERMANOVA: p=0.0245
PERMDISP: p=0.1669

E

PERMANOVA: p=0.0052
PERMDISP: p=0.0295

| Group 1 | Group 2 | FDR-adjusted p_{PERMDISP} |
|---------|---------|----------------------------|
| Normal  | NRVH    | 0.2381                     |
| Normal  | OSCC    | 0.2312                     |
| NRVH    | OSCC    | 0.5654                     |

| Group 1 | Group 2 | FDR-adjusted p_{PERMDISP} |
|---------|---------|----------------------------|
| Normal  | NRVH    | 0.0362                     |
| Normal  | OSCC    | 0.0699                     |
| NRVH    | OSCC    | 0.4975                     |
Figure 1

Differences in oral microbiomes among normal, NRVH, and OSCC cohorts. (A) Boxplots illustrating the distribution of observed ASVs. Each dot represents a subject. The normal cohort has a higher mean observed ASVs compared to the other cohorts. However, no statistical significance is shown after FDR adjustment. (B-C) Adonis analysis based on (B) unweighted and (C) weighted UniFrac distance metrics shows the effect (R2) of factors with the oral microbiome. * indicates FDR-adjusted p < 0.05 and *** indicates FDR-adjusted p < 0.001. (D-E) Principal coordinate analysis (PCoA) plots of taxonomic profiles based on (D) unweighted and (E) weighted UniFrac distance metrics. Marginal kernel densities visualize the distribution of microbial diversity along both axes. The pairwise PERMDISP reveals the dispersion effect (FDR-adjusted p < 0.05) between normal and NRVH cohorts.
Figure 1
Figure 1

Differences in oral microbiomes among normal, NRVH, and OSCC cohorts. (A) Boxplots illustrating the distribution of observed ASVs. Each dot represents a subject. The normal cohort has a higher mean observed ASVs compared to the other cohorts. However, no statistical significance is shown after FDR adjustment. (B-C) Adonis analysis based on (B) unweighted and (C) weighted UniFrac distance metrics shows the effect (R2) of factors with the oral microbiome. * indicates FDR-adjusted p < 0.05 and *** indicates FDR-adjusted p < 0.001. (D-E) Principal coordinate analysis (PCoA) plots of taxonomic profiles based on (D) unweighted and (E) weighted UniFrac distance metrics. Marginal kernel densities visualize the distribution of microbial diversity along both axes. The pairwise PERMDISP reveals the dispersion effect (FDR-adjusted p < 0.05) between normal and NRVH cohorts.
Figure 2

A

B

C

| Species/Gene | Normal | NR VH | OSCC |
|--------------|--------|-------|------|
| Prevotella melaninogena |  | | |
| Veillonella atypica |  | | |
| Prevotella unclassified |  | | |
| Streptococcus salivarius |  | | |
| Haemophilus parainfluenzae |  | | |
| Saccharibacteria (TM7) [G-1] bacterium HMT-352 |  | | |
| Veillonella dispar |  | | |
| Prevotella pallens |  | | |
| Selenomonas unclassified |  | | |
| Megasphaera micronucleiformis |  | | |
| Campylobacter concisus |  | | |
| Leptotrichia sp. HMT-417 |  | | |
| Prevotella salivae |  | | |
| Rothia mucilaginosa |  | | |
| Actinomyces sp. HMT-180 |  | | |
| Veillonella parvula |  | | |
| Rothia dentocariosa |  | | |
| Capnocytophaga sputigena |  | | |
| Prevotella oris |  | | |
| Peptostreptococcus stomatis |  | | |
| Parvimonas micra |  | | |

LDA SCORE (log 10)
Figure 2

Core microbiome analysis. (A) Venn diagram of core microbiomes among cohorts. The core is defined as the species taxa present in saliva with $\geq 75\%$ of prevalence. (B) The fraction of core species number to overall species richness in each cohort. (C) LEfSe reveals the distribution of core species displaying the abundance significantly higher (LDA $>$ log103) among cohorts. The asterisk (*) indicates a taxon that was annotated only to the genus level.
Figure 2

A. Venn diagram showing the overlap of samples in different groups.

B. Bar chart showing the percentage of core microbiome in Normal, NRVH, and OSCC groups.

C. Bar chart showing the LDA scores of different bacterial species in Normal, NRVH, and OSCC groups.
Figure 2

Core microbiome analysis. (A) Venn diagram of core microbiomes among cohorts. The core is defined as the species taxa present in saliva with \( \geq 75\% \) of prevalence. (B) The fraction of core species number to overall species richness in each cohort. (C) LEfSe reveals the distribution of core species displaying the abundance significantly higher (LDA > log103) among cohorts. The asterisk (*) indicates a taxon that was annotated only to the genus level.
Figure 3
Figure 3

Distribution of signature pathways. The signature pathways, which abundances are significantly higher concerning each studied cohort, are detected using LEfSe. The inferred pathways are collapsed to each category based on Metacyc’s pathway ontology. Colored boxes indicate a higher rank of the categories.
Figure 3
Figure 3

Distribution of signature pathways. The signature pathways, which abundances are significantly higher concerning each studied cohort, are detected using LEfSe. The inferred pathways are collapsed to each category based on Metacyc’s pathway ontology. Colored boxes indicate a higher rank of the categories.
Figure 4

![Species and Pathway Comparison](image)

- **Species**:
  - Normal: Sørensen dissimilarity (β_{SØR})
  - NRVH: Sørensen dissimilarity (β_{SØR})
  - OSCC: Sørensen dissimilarity (β_{SØR})

- **Pathway**:
  - Normal: Nestedness (β_{NES}) and Turnover (β_{SIM})
  - NRVH: Nestedness (β_{NES}) and Turnover (β_{SIM})
  - OSCC: Nestedness (β_{NES}) and Turnover (β_{SIM})
Figure 4

Multiple-site beta diversity (Sørensen dissimilarity) and corresponding nestedness and turnover components. The dissimilarities were analyzed in terms of species and metabolic pathway profiles in each cohort.
Figure 4
Figure 4

Multiple-site beta diversity (Sørensen dissimilarity) and corresponding nestedness and turnover components. The dissimilarities were analyzed in terms of species and metabolic pathway profiles in each cohort.
Figure 5

A

Bray-Curtis (Taxon)

PERMANOVA: p=0.0001
PERMDISP: p=0.0061

B

Bray-Curtis (Function)

PERMANOVA: p=0.0001
PERMDISP: p=0.1267

C

Taxonomic profile – normal

D

Taxonomic profile – OSCC

E

Metabolic pathway – normal

F

Metabolic pathway – OSCC

| Group 1   | Group 2   | FDR-adjusted p<sub>PERMDISP</sub> |
|-----------|-----------|-----------------------------------|
| Normal    | NRVH      | 0.3067                            |
| Normal    | OSCC      | 0.0018                            |
| NRVH      | OSCC      | 0.3067                            |
|           |           |                                   |
| Normal    | NRVH      | 0.8007                            |
| Normal    | OSCC      | 0.1914                            |
| NRVH      | OSCC      | 0.3423                            |
Figure 5

Comparison of diversity and core analysis between taxonomic and functional profiles from different studies. (A-B) Principal coordinate analysis (PCoA) plots of (A) taxonomic and (B) metabolic pathway profiles based on Bray-Curtis distance metric from previous studies and this study. The dispersion effect was found between normal and OSCC cohorts (FDR-adjusted pPERMDISP = 0.0018) using taxonomic profiles. The percentages of explained variance are displayed along with each principal coordinate axis. (C-F) Venn diagrams reveal common core species/pathways (prevalence > 75%) in normal and OSCC cohorts, respectively. (Taxonomic profiles of normal (C) and OSCC (D) cohorts; metabolic pathway profiles of normal (E) and OSCC (F) cohorts).
Figure 5

A. Bray-Curtis (Taxon)

PERMANOVA: p=0.0001
PERMDISP: p=0.0061

| Group 1  | Group 2 | FDR-adjusted p_FDRDISP |
|---------|---------|------------------------|
| Normal  | NRVH    | 0.3067                 |
| Normal  | OSCC    | 0.0018                 |
| NRVH    | OSCC    | 0.3067                 |

B. Bray-Curtis (Function)

PERMANOVA: p=0.0001
PERMDISP: p=0.1267

| Group 1  | Group 2 | FDR-adjusted p_FDRDISP |
|---------|---------|------------------------|
| Normal  | NRVH    | 0.8007                 |
| Normal  | OSCC    | 0.1914                 |
| NRVH    | OSCC    | 0.3423                 |

C. Taxonomic profile – normal

D. Taxonomic profile – OSCC

E. Metabolic pathway – normal

F. Metabolic pathway – OSCC
Figure 5

Comparison of diversity and core analysis between taxonomic and functional profiles from different studies. (A-B) Principal coordinate analysis (PCoA) plots of (A) taxonomic and (B) metabolic pathway profiles based on Bray-Curtis distance metric from previous studies and this study. The dispersion effect was found between normal and OSCC cohorts (FDR-adjusted pPERMDISP = 0.0018) using taxonomic profiles. The percentages of explained variance are displayed along with each principal coordinate axis. (C-F) Venn diagrams reveal common core species/pathways (prevalence > 75%) in normal and OSCC cohorts, respectively. (Taxonomic profiles of normal (C) and OSCC (D) cohorts; metabolic pathway profiles of normal (E) and OSCC (F) cohorts).
Figure 6

A

Raw sequences
→ DADA2 pipeline
→ Taxa annotation
→ PICRUSSt2 pipeline
→ Pathway abundances
→ Taxa (species) abundances
→ rarefaction
→ Supervised classification
→ Mean accuracy ratio & AUROC

100 iterations of random training-testing split (80%-20%)

Key: Input/output
Method/pipeline

B

Accuracy ratio

Function Taxon

p = 0.136
p = 0.001
p = 0.043

C

Wolf et al. Zhao et al. This study
Mean AUROC=0.723 Mean AUROC=0.702 Mean AUROC=0.783
Std Dev=0.200 Std Dev=0.135 Std Dev=0.091

D

Wolf et al. Zhao et al. This study
Mean AUROC=0.525 Mean AUROC=0.702 Mean AUROC=0.776
Std Dev=0.150 Std Dev=0.135 Std Dev=0.076
Figure 6

Evaluating functional profile as an alternative signature for OSCC detection using machine learning. (A) Flow chart illustrating the process of obtaining predicted results from published sequences. Blue boxes indicate input or output files, and red boxes indicate bioinformatics methods or pipelines. (B) The mean accuracy ratios of 100 iterations of the randomly split dataset (80% training and 20% testing) based on taxonomic and functional profiles. The accuracy ratio is defined as the predicted accuracy to the accuracy of a random guess. Vertical bars indicate 95% confidence intervals. (C-D) The 2D-density plots of ROC curves from 100 iterations demonstrate a higher mean AUROC using (C) functional profiles to distinguish OSCC from normal cohorts compared to that using (D) taxonomic profiles.
Figure 6

A

Raw sequences

DADA2 pipeline

Taxa annotation

PICRUST2 pipeline

Pathway abundances

Taxa (species) abundances

rarefaction

Supervised classification

100 iterations of random training-testing split (80%-20%)

Mean accuracy ratio & AUROC

Key:

Input/output

Method/pipeline

B

Accuracy ratio

Function

Taxon

C

Wolf et al.

Zhao et al.

This study

Mean AUROC = 0.723

Mean AUROC = 0.798

Mean AUROC = 0.783

Std Dev = 0.030

Std Dev = 0.098

Std Dev = 0.091

D

Wolf et al.

Zhao et al.

This study

Mean AUROC = 0.652

Mean AUROC = 0.702

Mean AUROC = 0.776

Std Dev = 0.350

Std Dev = 0.135

Std Dev = 0.076
Figure 6

Evaluating functional profile as an alternative signature for OSCC detection using machine learning. (A) Flow chart illustrating the process of obtaining predicted results from published sequences. Blue boxes indicate input or output files, and red boxes indicate bioinformatics methods or pipelines. (B) The mean accuracy ratios of 100 iterations of the randomly split dataset (80% training and 20% testing) based on taxonomic and functional profiles. The accuracy ratio is defined as the predicted accuracy to the accuracy of a random guess. Vertical bars indicate 95% confidence intervals. (C-D) The 2D-density plots of ROC curves from 100 iterations demonstrate a higher mean AUROC using (C) functional profiles to distinguish OSCC from normal cohorts compared to that using (D) taxonomic profiles.

Supplementary Files

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

- OVHSupplFig.pptx
- OVHSupplFig.pptx
- SupplementaryTablev6.docx
- SupplementaryTablev6.docx