RESEARCH

Open Access

Exploring the microbiome of oral epithelial dysplasia as a predictor of malignant progression

Robyn J. Wright[1*], Michelle E. Pewarchuk[2], Erin A. Marshall[2], Benjamin Murray[1], Miriam P. Rosin[3,4], Denise M. Laronde[3,5], Lewei Zhang[5,6], Wan L. Lam[2], Morgan G. I. Langille[1,7] and Leigha D. Rock[1,7,8,9,10]

Abstract

A growing body of research associates the oral microbiome and oral cancer. Well-characterized clinical samples with outcome data are required to establish relevant associations between the microbiota and disease. The objective of this study was to characterize the community variations and the functional implications of the microbiome in low-grade oral epithelial dysplasia (OED) using 16S rRNA gene sequencing from annotated archival swabs in progressing (P) and non-progressing (NP) OED. We characterised the microbial community in 90 OED samples — 30 swabs from low-grade OED that progressed to cancer (cases) and 60 swabs from low-grade OED that did not progress after a minimum of 5 years of follow up (matched control subjects). There were small but significant differences between P and NP samples in terms of alpha diversity as well as beta diversity in conjunction with other clinical factors such as age and smoking status for both taxa and functional predictions. Across all samples, the most abundant genus was *Streptococcus*, followed by *Haemophilus*, *Rothia*, and *Neisseria*. Taxa and predicted functions were identified that were significantly differentially abundant with progression status (all Ps and NPs), when samples were grouped broadly by the number of years between sampling and progression or in specific time to progression for Ps only. However, these differentially abundant features were typically present only at low abundances. For example, *Campylobacter* was present in slightly higher abundance in Ps (1.72%) than NPs (1.41%) and this difference was significant when Ps were grouped by time to progression. Furthermore, several of the significantly differentially abundant functions were linked to the *Campylobacteraceae* family in Ps and may justify further investigation. Larger cohort studies to further explore the microbiome as a potential biomarker of risk in OED are warranted.

Keywords Oral epithelial dysplasia, Oral cancer, 16S rRNA sequencing, Oral microbiota, Microbiome

*Correspondence:
Robyn J. Wright
Robyn.Wright@dal.ca
[1] Department of Pharmacology, Dalhousie University, Halifax, Canada
[2] Department of Integrative Oncology, British Columbia Cancer Research Centre, Vancouver, Canada
[3] Department of Cancer Control Research, British Columbia Cancer Research Centre, Vancouver, Canada
[4] Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, Canada
[5] Faculty of Dentistry, University of British Columbia, Vancouver, Canada
[6] Oral Biopsy Service, Vancouver General Hospital, Vancouver, Canada
[7] Beatrice Hunter Cancer Research Institute, Halifax, Canada
[8] Faculty of Dentistry, Dalhousie University, Halifax, Canada
[9] Department of Pathology, Faculty of Medicine, Dalhousie University, Halifax, Canada
[10] Department of Anatomical Pathology, QEII Hospital, Nova Scotia Health, Halifax, Canada

Background

Oralcancers pose a major public health challenge. In 2020, there were > 350,000 new cases of oral cancer worldwide, with > 175,00 deaths [1]. In North America, there were 35,310 cases and 7,110 deaths, respectively [2]. In Canada, oral cancer represents 3.3% of overall cancer burden in males and 1.5% in females, ranking above stomach, liver, brain, testicular and cervical cancer [3]. The 5-year survival rate for oral cancers remains < 50% in much of the world [1], mainly due to late stage diagnosis [4, 5]. Early detection is vital to the improvement of this prognosis [6]. While the costs to health care systems for oral cancer treatments are large [7], these are decreased with earlier intervention [8].

Oral cancer is frequently preceded by an oral potentially malignant lesion (OPML) [9]. However, even when an OPML is discovered early, clinical management can be challenging, as it is difficult to discriminate between indolent and aggressive disease. A histological diagnosis of oral epithelial dysplasia (OED) provides an indication of risk for high-grade (severe) dysplasia, however, it is a relatively poor prognosticator for low-grade (mild or moderate) dysplasia, which represents the majority of cases [10]. A major barrier to the clinical management of low-grade lesions is the inability to accurately discriminate between indolent and transformative disease [11–15]. Treatment can bear significant morbidity, and given that most low-grade lesions will not progress, it is correspondingly important to avoid overtreatment [16]. Biomarkers are needed to supplement epithelial dysplasia grading in order to triage low-grade OED according to their risk of progression [9, 17].

A growing body of research draws attention to the connection between the microbiome and cancer in the oral cavity. Since *Helicobacter pylori* was first demonstrated to be a causative agent in gastric cancer [18], many efforts have been made to explore the relationship between bacteria and cancer in other body sites. These efforts uncovered additional associations, including *Salmonella typhi* with gallbladder cancer [19], *Bacteroides fragilis* with colon cancer [20], and *Porphyromonas gingivalis* and *Fusobacterium nucleatum* with colorectal cancers [21–23]. In the context of oral cancer, there is growing evidence of the carcinogenicity of bacteria that have been found to inhibit apoptosis, activate cellular proliferation, promote cellular invasion and induce chronic inflammation, although these findings are primarily from in vitro and animal studies [24]. However, given the oral cavity contains more than 700 microbial species that form complex and diverse communities (microbiomes), the microbiome-host interaction is believed to extend far beyond the involvement of a select few species [25]. There has been an increase in the number of studies exploring the microbiome associated with oral squamous cell carcinoma (OSCC) [24, 26–29]. Evidence is growing that polymorphic microbiomes can have an impact on cancer, and have recently been added as an enabling characteristic to the Hallmarks of Cancer [30].

Definitive evidence that the microbiome plays a causal role in the development of oral cancer remains lacking. Shifts in the microbial populations that colonize human tissues have been shown to affect host biological pathways through the output of small molecules and metabolites [31]. Although some research has investigated the diversity of the microbiome in the context of OPML, the studies are cross-sectional or have focused on differences in the oral microbiota associated with health and disease [32, 33]. Longitudinal studies that focus on the relationship between the oral microbiome and the malignant transformation of OED are needed**.**

Well-characterized clinical samples with long follow-up data are required to establish relevant associations between the microbiota and disease. Unlike some anatomical sites, the oral cavity is easily accessible for sampling using non-invasive techniques, for example brush swabs from OPMLs.

The objectives of this study were to characterize the community variations and the functional implications of the microbiome in low-grade OED with known outcome using 16S rRNA gene sequencing from annotated archival brush swabs.

Methods

Patient population and study design

This project used data and samples already collected from subjects enrolled in the Oral Cancer Prediction Longitudinal (OCPL) study, an ongoing prospective cohort study being conducted in British Columbia that has recruited over 600 patients with biopsy confirmed OED. Subjects were identified through a centralized population-based biopsy service, the BC Oral Biopsy Service (OBS), where community dentists and surgeons across British Columbia (estimated population 5.2 million, in 2021 [34]) send biopsies for histopathological diagnosis. Patients with a diagnosis of low-grade OED were referred for follow up to OCPL oral dysplasia clinics where they were invited to participate in the OCPL study and were subsequently followed at 6-month intervals, creating an extensive biobank with associated demographic, clinical, histological, and outcome data. Participants were recruited by written informed consent. Details about the OCPL cohort recruitment, sample collection, and participant follow-up procedures have been published previously [13, 35]. As part of the OCPL Study, a cytology brush swab (Innovatek Inc.) was used to collect exfoliated cells from the primary oral lesion by stroking with pressure 10 times in one direction, turning the swab 180° and stroking 10 times in the opposite direction. The brush tip was broken off and placed into a 2 mL vial containing PreservCyt solution preservative (Hologic, Canada). All samples were barcoded, labeled and stored in a secure biobank. Comparative biopsies were performed on the index lesion every 24 months or upon significant clinical change by a certified oral medicine specialist. Histopathological grading was done at the Oral Biopsy Service and was reviewed and confirmed by the study pathologist (LZ), using diagnostic criteria established by the WHO [36].

The present study used a nested, retrospective matched case–control design. Inclusion criteria included being an OCPL study participant with a biopsy confirmed diagnosis of low-grade (mild or moderate) OED [36], and a baseline cytological specimen available. Participants with a previous history of oral cancer were excluded. Demographic, risk-habit, clinical and outcome information of eligible participants were obtained from the study database. Cases consisted of lesions that progressed to severe dysplasia, carcinoma *in-situ* or OSCC (progressors (P)). Controls consisted of lesions that did not progress after a minimum of 5 years of follow-up (non-progressors (NP)). Cases and controls were matched by anatomical site, age (±6 years), sex, ethnicity, smoking status, and alcohol consumption. A sample size of 83 (28 cases, 55 controls) was required to detect a significant difference with a ratio of two controls to one case, a hypothetical proportion of controls with 20% exposure, and a hypothetical proportion of 50% of cases with exposure, with a significance level of 5% and 80% power on 2-tailed tests in an unmatched case–control study design (OpenEpi® Version 3.01 software) [37]. Thirty cases and 60 controls were pulled to allow for a reasonable margin of error. A simple random sampling method of all eligible participants was employed. Ethics approval for the present study was obtained from the UBC BC Cancer (H20-00,809) and the Dalhousie University Research Ethics Boards (2020–5102). To ensure that the clinical characteristics of the samples would not bias the observations of microbial taxa and associations with progression, a Cox proportional hazards model within the R package survminer [38] was utilized to determine whether time to progression was associated with clinical characteristics: alcohol consumption (drinks per week), smoking history (pack-years), age (years) and grade of dysplasia.

Sample preparation and 16S rRNA sequencing

Lesion cytology brush swabs from baseline visits were pulled from the OCPL biobank. DNA isolation was performed using the DNeasy Blood & Tissue Kit (QIAGEN, Hilden Germany). An enzymatic lysis buffer preparation and incubation period was followed by the addition of proteinase K, followed by vortex and incubation. Extraction steps were conducted as per the manufacturer's instructions. DNA was eluted using 50 µl nuclease free distilled water and stored at − 20 °C. Quantification and quality assessment were performed using a spectrophotometer (NanoDrop ND-100; PEQLAB Biotechnologie, Erlangen, Germany).

A 16S rRNA gene sequencing approach was employed to uncover associations between microbial taxa and malignant progression. A single round of PCR (25 cycles) was done using Platinum SuperFi II DNA Polymerase–High-Fidelity PCR Enzyme for preamplification and primers targeting the V1-V3 region (27Fmod forward primer = AGRGTTTGATCMTGGCTCAG; 519R reverse primer = GWATTACCGCGGCKGCTG) [39] of the 16S rRNA gene. PCR products were verified visually by gel electrophoresis. Amplicon fragments were sent to the Integrated Microbiome Resource at Dalhousie University (https://imr.bio) and were PCR-amplified using high-fidelity Phusion Plus® polymerase (New England Biolabs Inc.). Amplified DNA concentrations were then normalized, pooled, and sequenced on an Illumina MiSeq using 300 bp paired end read chemistry.

Bioinformatic analyses

Read processing

Raw forward reads were imported into QIIME2 v2022.2 [40] for processing. Due to low reverse read quality, these were not used. Briefly, primers were trimmed using Cutadapt [41], reads were quality filtered using the default parameters within the quality-filter plugin and reads were denoised using the DADA2 denoising algorithm [42] with 5 errors allowed. Taxonomy was assigned to the resulting Amplicon Sequence Variants (ASVs) using the scikit-learn [43] naïve bayes classifier trained on the full-length 16S rRNA gene SILVA reference database (version 138) [44] downloaded from the QIIME2 website on 4th July 2022. ASVs were removed from further analysis that were: unclassified at the phylum level, classified as mitochondria or chloroplasts, had a maximum abundance of < 10 reads per sample or present in < 3 samples. The resulting ASVs were subsequently classified using a local BLAST [45]search against the SILVA reference database (version 138) [44] and with the scikit-learn [43] naïve bayes classifier trained on the full-length 16S rRNA gene Human Oral Microbiome Database (HOMD; version 15.22) [46]. Unless otherwise stated, the taxonomic classifications used are those obtained from the naïve bayes classifier trained on the full-length 16S rRNA gene HOMD (version 15.22). A phylogenetic tree was built using SEPP [47] with a reference phylogeny created using the SILVA reference database (version 128) [44].

Rarefaction curves were visualised, and all samples were found to have sufficient sampling depth (mean 19,082 reads per sample; range 7,907–48,793). To assess whether shifts in the oral microbiome community were associated with changes in community metabolism, the PICRUSt2 tool was used to predict metabolic pathways as well as Enzyme Commission (EC) numbers [48]. This tool uses 16S rRNA gene sequences and previously published genomic information to estimate the metabolic capabilities of microbial communities. Downstream bioinformatic analyses were performed using R (v4.1.2), Python (v3.9.12) and RStudio (v2022.02.1).

Statistical analysis

Raw count tables were normalised by rarefying to the lowest read depth, conversion to relative abundances or conversion to centered log ratios (CLR). The phylogenetic tree was collapsed at different taxonomic ranks using the R package Phyloseq [49]. Alpha diversity was assessed using Chao1 richness, Shannon diversity, Simpson's diversity, Simpson's evenness, and Faith's phylogenetic diversity on rarefied data. Beta diversity between Ps and NPs was assessed using Weighted UniFrac distance [50] on rarefied data as well as the compositionally-aware Phylogenetic Isometric Log Ratio (PhILR) distance [51] on raw count tables and visualized using a principal coordinate analysis (PCoA). Alpha and beta diversity metrics as well as ordinations were calculated using the Python package scikit-bio [52]. Mann–Whitney U tests were run using the Python package scipy [53] to determine whether there were differences in alpha diversity between Ps and NPs. To assess the association between microbial composition and progressor status, PERMANOVA tests were run using the Adonis function within the R package vegan [54] with the following metadata variables: Age, Sex, Ethnicity, Alcohol intake, Smoking status, Grade of dysplasia and Anatomical site. The matched P/NP groupings were given to the model using the strata option to constrain permutations and the tests were run separately with: (i) Progression grouped to Ps/NPs; (ii) Progression grouped to NPs or groupings of the time to progression (< 1, 1–2, 2–3, 3–4, 4–6 or 6 + years); or (iii) the number of months to progression for Ps only. PERMANOVA tests were run separately for Weighted UniFrac and PhILR distances (as calculated above).

To identify taxonomic features, predicted EC numbers or MetaCyc metabolic pathways associated with progression status, we ran MaAsLin2 [55] using the Maaslin2 R package. These tests were run separately with: (i) Progression grouped to Ps/NPs; (ii) Progression groups to NPs or groupings of the time to progression (< 1, 1–2, 2–3, 3–4, 4–6 or 6 + years); or (iii) the number of months to progression for Ps only as fixed effects. The matched P/NP groupings were given to the model as random effects and the NP group was used as the reference. These tests were run with and without the inclusion of the other clinical variables (age, sex, ethnicity [with white used as the reference], alcohol intake, smoking status, grade of dysplasia and lesion site) as fixed effects and with both relative abundance and CLR-transformed data. Taxa, EC numbers or pathways were considered to be significantly differentially abundant between groups if they had a *q*-value of 0.25 (the default in MaAsLin2). JarrVis [56, 57] was used to visualise the links between taxonomy and the top 10 differentially abundant EC numbers. The alpha diversity of ASVs contributing to the top 10 differentially abundant EC numbers were calculated as in the FuncDiv R package (https://github.com/gavinmdouglas/FuncDiv) [58].

Alpha and beta diversity analyses as well as PERMANOVA and differential abundance tests were run in the same way for taxonomic data at the ASV, species or genus level as well as on the PICRUSt2 output at the predicted pathway or enzyme level. All code used for analysis can be found at https://github.com/R-Wright-1/OED_microbiome and https://doi.org/10.5281/zenodo.7093667. Sequences were deposited at Gene Expression Omnibus (GEO) with accession GSE198811.

Results

Demographic analysis of the cohort

Participants were followed to an average of 83.8 months (Table 1; Supp. Table S1A). The average age at diagnosis was 60.2 years and the ratio of males ($n = 42$) to females ($n = 53$) was almost equal. Participants identified as primarily white (81%), followed by Asian (11%), and South or East Asian (6%). The majority of participants reported having never smoked (60%) and were non- or light alcohol drinkers (90%). Ninety samples were included in the study: 30 cases (progressors; OED that progressed to severe dysplasia, *CIS* or SCC) and 60 controls (non-progressors; OED that did not progress after a minimum of 5 years of follow-up). Samples were primarily from the lateral or ventral tongue or from the floor of mouth. Progressors (Ps) and non-progressors (NPs) were matched by clinical and demographic variables, and there were therefore no significant differences in lesion site, age, sex, smoking history or alcohol consumption between Ps and NPs ($p > 0.05$). Cohorts were followed for a comparable amount of time. Given that the oral cavity may be exposed to a variety of environmental carcinogens, and lifestyle is a large factor contributing to this, we assessed the effects of various clinical characteristics on the progression status of our cohort. The age, smoking status, and alcohol intake did not differ significantly between P and NP sample sets

Table 1 Clinicopathological information of patient cohort

	All (%)[£]	**No progression[*] (%)[£]**	**Progression[*] (%)[£]**
Total	*n* = 90	*n* = 60	*n* = 30
Length of follow-up Median months (range)	83.8 (12.9 to 181.4)	85.3 (18.0 to 172.5)	73.7 (12.9 ± 181.4)
Age at diagnosis Mean (years ± SD)	60.2 ± 10.4	60.6 ± 10.2	59.3 ± 10.9
Sex			
Male	42 (47)	28 (47)	14 (47)
Female	48 (53)	32 (53)	16 (53)
Ethnicity			
White	73 (81)	49 (82)	24 (80)
Asian	10 (11)	6 (10)	4 (13)
South or East Asian	5 (6)	5 (8)	0 (0)
Other	2 (2)	0 (0)	2 (7)
Smoking history[a]			
Never	54 (60)	35 (58)	19 (63)
Ever	36 (40)	25 (42)	11 (37)
Alcohol consumption[b]			
Non/light	81 (90)	54 (90)	27 (90)
Heavy	9 (10)	6 (10)	3 (10)
Lesion site[c]			
Low Risk	12 (13)	8 (13)	4 (13)
High Risk	78 (87)	52 (87)	26 (87)
Grade of dysplasia			
Mild dysplasia	42 (47)	33 (55)	9 (30)
Moderate dysplasia	48 (53)	27 (45)	21 (70)
Time to progression category			
< 1 year	-	-	5 (17)
1 – 2 years			5 (17)
2 – 3 years			5 (17)
3 – 4 years			6 (20)
4 – 6 years			5 (17)
> 6 years			4 (13)

[*] Progression = progression to severe dysplasia, carcinoma in situ, or squamous cell carcinoma; No progression = no progression to severe dysplasia, carcinoma in situ, or squamous cell carcinoma after a minimum of five years of follow-up

[£] Column percentage reported

[a] Never smoker < 100 cigarettes in lifetime; Ever smoker > 100 cigarettes in lifetime

[b] Heavy alcohol consumption is defined as consumption of more than 14 drinks per week for females and 21 drinks per week for men. One alcoholic drink was defined as 8 oz of beer, 5 oz of wine or 1 oz of spirits

[c] High Risk = floor of mouth, soft palate, and tongue; Low Risk = all other sites

(Mann–Whitney U tests $p > 0.384$) and were not associated with disease onset time (time to progression; Cox proportional hazards test $p > 0.3$; Fig. 1A), and the cohorts were followed for a comparable amount of time (Table 1). 53% ($n = 48$) of samples exhibited moderate OED and 47% ($n = 42$) demonstrated mild OED. As expected, grade of dysplasia was significant for risk of progression (coefficient 0.96, $p = 0.0194$), with a greater proportion of Ps exhibiting moderate dysplasia (70%). Therefore, to control for potential confounding, multivariable analyses were performed.

Establishing a profile of microbial taxa in progressing and non-progressing OED

Diversity of oral microbial communities

Since external clinical factors did not have a significant impact on disease onset, we investigated whether changes in the microenvironment may be allowing for the colonization of additional organisms not previously known to be associated with the oral cavity which may impact the progression of OED. ASVs were classified taxonomically using the HOMD version 15.22 [46] (see Supp. Results and Supp. Figs. S1 and S2 for the results obtained using different taxonomic classifiers/databases).

To assess the diversity at the species level within the microbial communities of P and NP OED samples (Supp. Table S1B), we used five different alpha-diversity metrics (Fig. 1B). Although Ps and NPs showed similar levels of species diversity according to Shannon and Simpson's diversity metrics, there was a relative decrease in diversity according to Chao1 richness and Faith's PD (Mann–Whitney U test, $p=0.006$ and $p=0.014$, respectively) and an increase in Simpson's evenness (Mann–Whitney U test $p=0.029$) for the Ps relative to NPs. This may indicate that while there is a reasonably equal abundance of species within each sample set, the number of unique species may be lower in Ps, and they may be more closely related in phylogenetic space. Therefore, we observe lower relative diversity of microbes in the oral cavity of Ps.

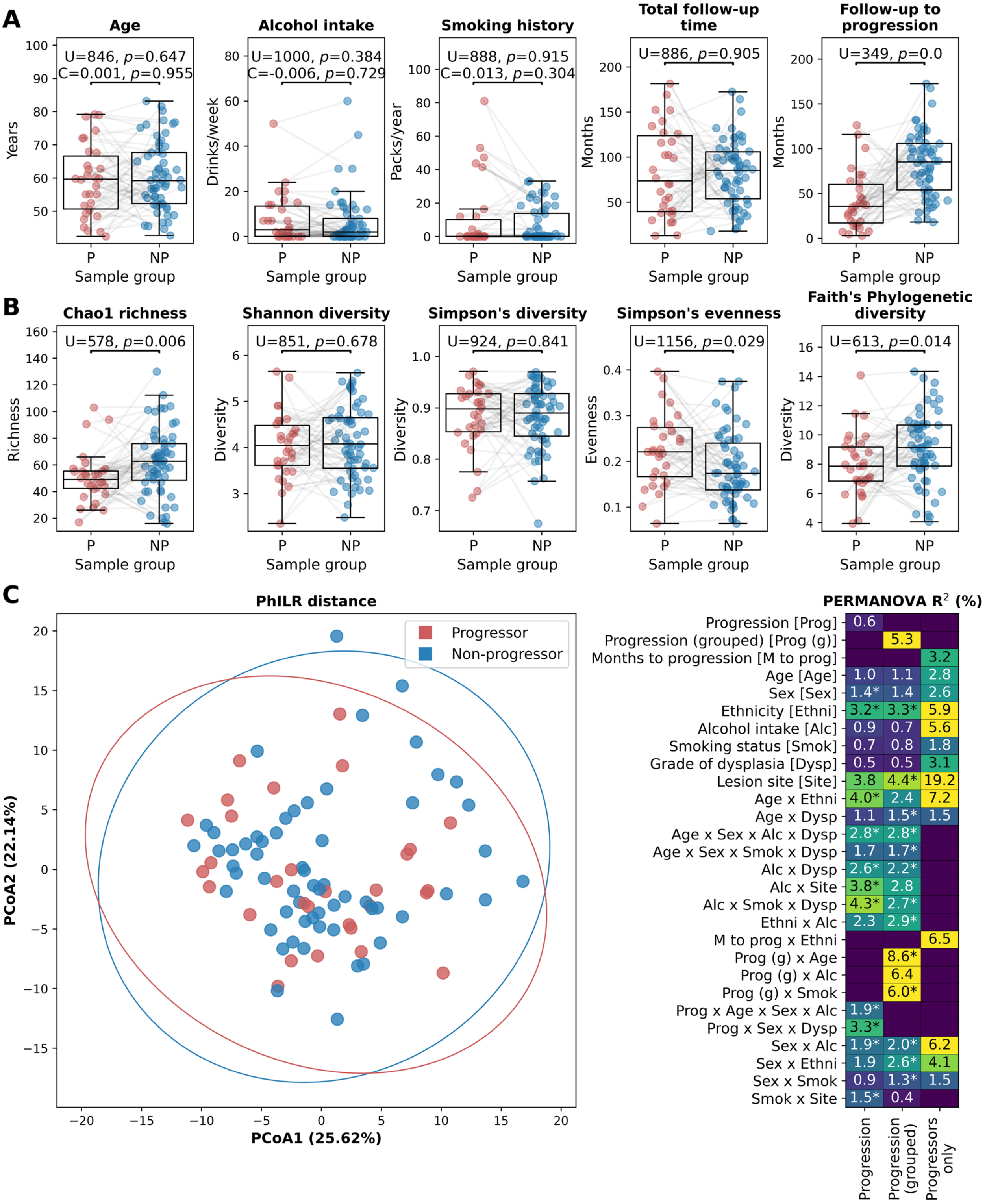
We compared the microbial diversity of Ps and NPs using both PhILR (Fig. 1C) and Weighted UniFrac distance (Supp. Fig. S3) with PERMANOVA tests with age, sex, ethnicity, alcohol intake, smoking status, grade of dysplasia, lesion site and progression grouped in one of three ways: (i) Ps and NPs (Progression; residual $R^2=6.1\%$ and 0.056, respectively); (ii) Ps grouped by the number of years to progression (<1, 1–2, 2–3, 3–4, 4–6 and 6+; Progression (Grouped); Residual $R^2=5.2\%$ and 0.050, respectively) and NPs; (iii) Ps only with the number of months to progression (Progressors only; Residual $R^2=0$ for both). Information on the matching of Ps and NPs was also given to the first two models. While there was no statistical difference observed in any of these groupings (PERMANOVA $p>0.05$; Fig. 1C and Supp. Fig. S3) and the Principal Coordinates Analysis (PCoA) of PhILR distance showed no clustering of P and NP samples – indicating that diversity was not affected by progression status alone – there were some significant ($p\leq 0.05$) differences with some other clinical variables (sex, ethnicity and smoking status) and interactions between progression and the other clinical variables (Fig. 1C and Supp. Fig. S3). Of note, for PhILR distance there were significant interactions between (i) progression and age, sex and alcohol intake ($R^2=1.9\%$, $p=0.027$) or sex and grade of dysplasia ($R^2=3.3\%$, $p=0.017$) (ii) progression (grouped) and age ($R^2=8.6\%$, $p=0.025$) or smoking status ($R^2=6\%$, $p=0.004$). The results were similar for Weighted UniFrac distance (Supp. Fig. S3), with diversity not being significantly affected by progression status alone, with significant interactions between (i) progression and sex ($R^2=2.4\%$, $p=0.011$), age, sex and alcohol intake ($R^2=1.5\%$, $p=0.043$) or sex and ethnicity ($R^2=0.7\%$, $p=0.049$) and (ii) progression (grouped) and either age ($R^2=8.2\%$, $p=0.036$) or smoking status ($R^2=6.3\%$, $p=0.001$). While there were no significant differences found with any of the variables for (iii) Ps only, the lesion site was contributing to large (non-significant) differences in beta diversity ($R^2=19.2\%$ or 20.5% for PhILR or Weighted UniFrac distance, respectively).

Phylum- and genus-level shifts in microbial communities

When taxa in Ps and NPs were compared, small differences in abundance were observable up to the phylum level of classification (Supp. Fig. S4). We have examined the abundance of taxa using both relative abundance, as this is what the majority of studies to date have used, and CLR abundance, as this accounts for the compositionality of microbiome data [59]. For the relative abundances, higher values indicate that more sequences belonging to a particular taxon are present. For the CLR abundance, a zero value indicates that the abundance of a taxon is equal to the mean $\log_2$ abundance of all taxa, with positive or negative values indicating higher or lower abundances than the mean $\log_2$ relative abundance,

(See figure on next page.)

Fig. 1 Clinical summary and sample diversity at the species level. **A** Age, alcohol intake, smoking status, and follow-up time were similar between groups. **B** Species diversity within the microbial communities of 30 progressing (P) and 60 non-progressing (NP) oral epithelial dysplasia (OED) samples using five different alpha-diversity metrics: Chao1 richness, Shannon diversity, Simpson's diversity, Simpson's evenness, and Faith's phylogenetic diversity. U and *p*-values in **A** and **B** were determined by Mann–Whitney U tests and lines indicate matched Ps and NPs. For age, alcohol intake and smoking history, Cox proportional hazard test coefficients (C) and *p*-values are shown in addition to the Mann–Whitney U test statistics. Lines between points indicate matched Ps/NPs and boxes show the median, upper and lower quartiles while whiskers show the range of the data (1.5 times the interquartile range). **C** Principal Coordinates Analysis (PCoA) and PERMANOVA tests using Phylogenetic Isometric Log Ratio (PhILR) distance at the species level (Supp. Table S1C). Ellipses show the confidence interval (3 standard deviations) for each group and the values shown on each axis label indicate the proportion of sample variation accounted for by that axis. The heatmap in **C** shows PERMANOVA R^2 values for all separate variables (shown with abbreviations in the first instance) that were added to the models as well as all interactions between variables with an $R^2\geq 5\%$ and/or with $p\leq 0.05$ (denoted with an asterisk). All PERMANOVA R^2 and p values are shown in Supp. Table S1C. The columns show PERMANOVA tests for Progression (P/NP), Progression (grouped; NP and P grouped to <1, 1–2, 2–3, 3–4, 4–6 or 6+years for time to progression) and Progressors only with the specific follow-up time. The equivalent of **C** for Weighted UniFrac distance is shown in Supp. Fig. S3



	Progression	Progression (grouped)	Progressors only
Progression [Prog]	0.6		
Progression (grouped) [Prog (g)]		5.3	
Months to progression [M to prog]			3.2
Age [Age]	1.0	1.1	2.8
Sex [Sex]	1.4*	1.4	2.6
Ethnicity [Ethni]	3.2*	3.3*	5.9
Alcohol intake [Alc]	0.9	0.7	5.6
Smoking status [Smok]	0.7	0.8	1.8
Grade of dysplasia [Dysp]	0.5	0.5	3.1
Lesion site [Site]	3.8	4.4*	19.2
Age x Ethni	4.0*	2.4	7.2
Age x Dysp	1.1	1.5*	1.5
Age x Sex x Alc x Dysp	2.8*	2.8*	
Age x Sex x Smok x Dysp	1.7	1.7*	
Alc x Dysp	2.6*	2.2*	
Alc x Site	3.8*	2.8	
Alc x Smok x Dysp	4.3*	2.7*	
Ethni x Alc	2.3	2.9*	
M to prog x Ethni			6.5
Prog (g) x Age		8.6*	
Prog (g) x Alc		6.4	
Prog (g) x Smok		6.0*	
Prog x Age x Sex x Alc	1.9*		
Prog x Sex x Dysp	3.3*		
Sex x Alc	1.9*	2.0*	6.2
Sex x Ethni	1.9	2.6*	4.1
Sex x Smok	0.9	1.3*	1.5
Smok x Site	1.5*	0.4	

Fig. 1 (See legend on previous page.)

respectively. In both sample groups, the dominant phyla were *Firmicutes*, *Proteobacteria*, *Actinobacteria*, *Bacteroidetes*, and *Fusobacteria*; a slight decrease in *Firmicutes* and a slight increase in *Proteobacteria* in Ps was observed for both relative and CLR abundance (Supp. Fig. S4A). At the genus-level, *Streptococcus* was the most abundant, with *Haemophilus*, *Neisseria*, and *Rothia* also making up high abundances for both relative and CLR abudance (Fig. 2; Supp. Fig. S4E). The relative abundances of these genera between NPs and Ps appear quite similar, other than an apparent decrease in *Streptococcus* in Ps.

Known oral cavity colonizers did not correlate with progression status

Most of the more abundant taxa identified in this study are known to be associated with human oral health [60, 61]. However, *Streptococcus* spp., *Haemophilus*, and *Fusobacterium*have been correlated with oral cancer and epithelial precursor lesions [26, 27, 33, 60, 62, 63]. Furthermore, some of the less abundant taxa identified, including *Campylobacter, Prevotella, Pseudomonas,* and *Rothia*, have been shown to be differentially enriched in studies that have investigated OSCC or OPML by swab or biopsy [26, 27, 33, 64]. We therefore carried out differential abundance testing between Ps and NPs using MaAsLin2 [55] – which allows the inclusion of the matched P/NP grouping as well as other metadata – at the genus and species levels in order to see whether there was any indication that these taxa are significantly associated ($q \leq 0.25$, the default cut-off in MaAsLin2) with progression status. We ran these tests both with and without the other clinical variables (age, sex, ethnicity, alcohol intake, smoking status, grade of dysplasia and lesion site) as well as with relative abundance or CLR-transformed count tables. There were no genera that were significantly differentially abundant with progression status or time to progression with the relative abundance data, however, there were some with the CLR abundance data. For the CLR abundance, no genera were significantly differentially abundant between Ps and NPs, however, when Ps were grouped to the time to progression (< 1, 1–2, 2–3, 3–4, 4–6 or 6 + years), one genus was significant in the CLR-transformed data, without or with the inclusion of the other clinical variables: *Campylobacter*, which was typically more abundant in Ps than NPs, although the magnitude and direction of this difference varied depending on the time to progression (Fig. 2, Supp. Fig. S5 and Supp. Table S2). There were 124 genera that were significantly differentially abundant with follow-up time in the Ps only (without the inclusion of the other clinical variables), nine of which were among the top 25 most abundant genera; *Actinomyces*, *Aggregatibacter*, *Alloprevotella*, *Campylobacter*, *Lautropia*, *Prevotella*, *Ralstonia*, *Schaalia* and *Selenomonas*. However, none of these were significantly differentially abundant when the other clinical variables were also included (Fig. 2 and Supp. Table S2).

At the species level there were – as for the genus level – no taxa that were significantly differentially abundant with either progression status or time to progression with the relative abundance data, but there were with the CLR abundance. There were 271 species that were significantly differentially abundant between Ps and NPs (five when the other clinical variables were also included) the majority of which were present in only very low abundances (Supp. Table S3). Of these 271 species, only five were present within the top 40 most abundant species (Fig. 3 and Supp. Table S3): two of these were higher in abundance in Ps than NPs (*Bergeyella* sp. HMT 322 and *Lautropia mirabilis*), two were lower in abundance in Ps than NPs (unclassified *Veillonella* and *Ralstonia pickettii*) and one was very similar in abundance between Ps and NPs, with slightly higher CLR abundance in NPs and higher relative abundance in Ps (unclassified *Prevotella*). None of these were also significantly differentially abundant with the inclusion of the clinical variables. Three species were significantly differentially abundant when Ps were grouped to the time to progression (< 1, 1–2, 2–3, 3–4, 4–6 or 6 + years), two of which were detected without (*Gemella morbillorum* and *Neisseria elongata*) and one with the inclusion of the other clinical variables (*Prevotella pallens*), although none of these were within the 40 most abundant species (Fig. 3 and Supp. Table S3).

(See figure on next page.)

Fig. 2 Prevalence, abundance and differential abundance of the top 25 most abundant genera in progressor and non-progressor samples. Phylogenetic tree showing the class of each genus and a heatmap showing mean prevalence (blue scale; left). Heatmaps showing mean abundance of genera in progressor (P) or non-progressor (NP) samples (left) or differential abundance (centre) are shown alongside boxplots showing abundance in all samples (right) for each of relative abundance and CLR abundance. In the boxplots, each sample is shown as an individual point and boxes show the median, upper and lower quartiles while whiskers show the range of the data (1.5 times the interquartile range). MaAsLin2 tests for differential abundance were run with (+ M) and without the other clinical variables (metadata, M; age, sex, ethnicity, alcohol intake, smoking status, grade of dysplasia and lesion site). As above for PERMANOVA tests, progression was grouped in one of three ways: (i) Ps and NPs (P *vs* NP); (ii) Ps grouped by the number of years to progression (< 1, 1–2, 2–3, 3–4, 4–6 and 6 +; P *vs* NP); and (iii) Ps only with the number of months to progression (P only). For (i) and (ii) the matched P/NP grouping was given to the model so only matched controls were used. Genera were determined to be differentially abundant and are shown in black in the heatmap if they had $q \leq 0.25$ (the default in MaAsLin2). White denotes that they were not significantly differentially abundant

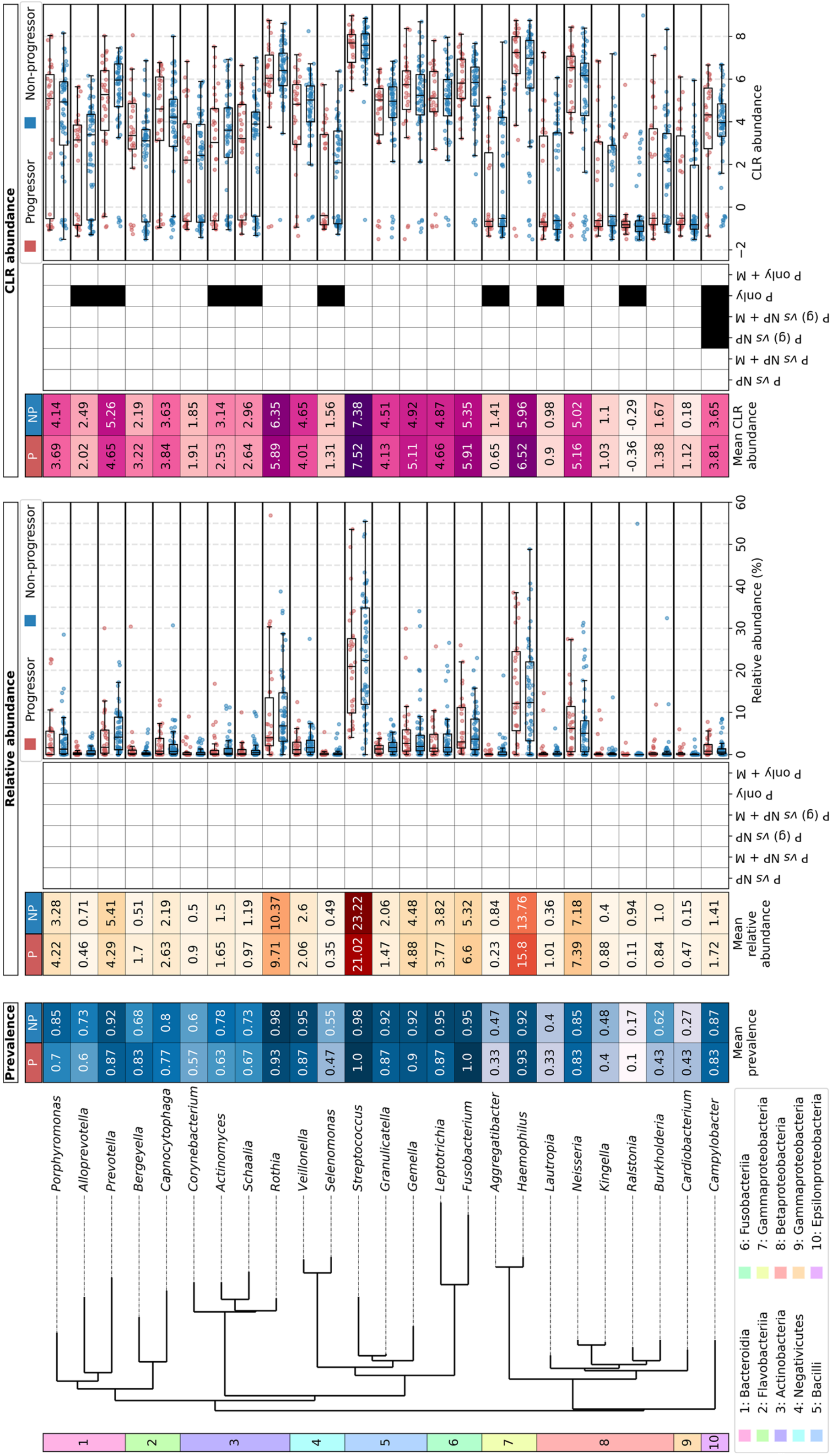


Fig. 2 (See legend on previous page.)

No species were significantly differentially abundant with time to progression for Ps only.

Metabolic impact of oral microbiome changes associated with progression

Colonizing bacteria are known to interact with their hosts through both direct and indirect mechanisms, including receptor interactions and the release of nutrients or metabolic products into the microenvironment. Therefore, we next examined the involvement of specific EC numbers or metabolic pathways in NPs and Ps using PICRUSt2 to generate predicted metagenomes for all samples based on the taxa present (Supp. Tables S1D and S1E). The alpha and beta diversity of profiles generated by PICRUSt2 for functions at the level of both genes for EC numbers and MetaCyc metabolic pathways were similar to that of the taxa (Fig. 1 and Supp. Figs. S6 and S7). Ps have significantly lower Chao1 richness and significantly higher Simpson's evenness than NPs (Mann–Whitney U test $p < 0.05$) and while progression status did not have a significant effect on Bray–Curtis beta diversity on its own, it did have a significant interaction with several of the other clinical variables (*e.g.*, sex, smoking status, lesion site and grade of dysplasia; Supp. Figs. S6 and S7 and Table S1C).

We carried out differential abundance testing using MaAsLin2 [55] at both the enzyme and the pathway level, with both relative and CLR abundance and both with and without the other clinical variables (age, sex, ethnicity, alcohol intake, smoking status, grade of dysplasia and lesion site). We initially examined the pathway level, and, as with the taxa, there were no significant differences between Ps and NPs for the relative abundance data, but there were five pathways that were significantly differentially abundant between Ps and NPs for the CLR abundance data without the other clinical variables, with one of these pathways (PWY0-1533; generally higher in abundance in NPs than Ps) still being differentially abundant with the inclusion of the other clinical variables (Supp. Table S4 and Supp. Fig. S8). When Ps were grouped to the time to progression (<1, 1–2, 2–3, 3–4, 4–6 or 6+years), there were 21 pathways that were significantly differentially abundant between one or more groups and the NPs without the other clinical variables, six of which were also significant with the inclusion of the other clinical variables (P164-PWY, P562-PWY, PWY-5265, PWY-6608 and RHAMCAT-PWY; along with one additional pathway; Supp. Fig. S8). For progressors only, there was only one pathway that was significantly differentially abundant (for the CLR abundance with the other clinical variables); P221-PWY (Supp. Table S4 and Supp. Fig. S8).

When we examined the enzymes, there were again very few that were significantly differentially abundant with progression status for the relative abundance, but there were a large amount that were for the CLR abundance (Supp. Table S5). We therefore focus only on the enzymes that were identified both with and without the inclusion of the other clinical variables: six for Ps *vs* NPs and nine for Ps grouped to the time to progression *vs* NPs (Supp. Fig. S9; no enzymes were significantly differentially abundant with time to progression for progressors only). In order to explore the links between taxonomy and function, we examined the ASVs that were contributing to these 15 EC numbers and found that the alpha diversity tended to be higher as the abundance of the EC number increased (Supp. Fig. S10). We used JarrVis to collapse the ASVs contributing to these EC numbers at the family level and visualise those families/EC numbers with >100 gene copies on average within a sample grouping (after rarefying; Fig. 4). Two of the EC numbers were removed by this filtering due to low abundance. This revealed that some families were the only abundant contributors to some enzymes and that these contributions could come from only a single sample group, *i.e.*, the contribution of Campylobacteraceae to enzymes EC:2.3.1.203, EC:2.4.1.290, EC:2.4.1.291, EC:2.6.1.34, EC:2.7.8.36, EC:3.2.2.30 and EC:4.2.1.135 from the 2–3 year time to progression group. There were also a few families for which there were only abundant contributions from one sample grouping (*e.g.*, Propionibacteriaceae, NP; Corynebacteriaceae, Bacillaceae, unclassified Proteobacteria and unclassified Rhizobiales, 6+years; Acetobacteraceae, 3–4 years; Actinomycetaceae, 1–2 years; unclassified Bacteria, 2–3 years; Neisseriaceae, 4–6 years

(See figure on next page.)

Fig. 3 Prevalence, abundance and differential abundance of the top 40 most abundant species in progressor and non-progressor samples. Phylogenetic tree showing the class of each species and a heatmap showing mean prevalence (blue scale; left). Heatmaps showing mean abundance of species in progressor (P) or non-progressor (NP) samples (left) or differential abundance (centre) are shown alongside boxplots showing abundance in all samples (right) for each of relative abundance and CLR abundance. In the boxplots, each sample is shown as an individual point and boxes show the median, upper and lower quartiles while whiskers show the range of the data (1.5 times the interquartile range). MaAsLin2 tests for differential abundance were run with (+ M) and without the other clinical variables (metadata, M; age, sex, ethnicity, alcohol intake, smoking status, grade of dysplasia and lesion site). As above for PERMANOVA tests, progression was grouped in one of three ways: (i) Ps and NPs (P *vs* NP); (ii) Ps grouped by the number of years to progression (<1, 1–2, 2–3, 3–4, 4–6 and 6+; P *vs* NP); and (iii) Ps only with the number of months to progression (P only). For (i) and (ii) the matched P/NP grouping was given to the model so only matched controls were used. Species were determined to be differentially abundant and are shown in black in the heatmap if they had $q \leq 0.25$ (the default in MaAsLin2). White denotes that they were not significantly differentially abundant

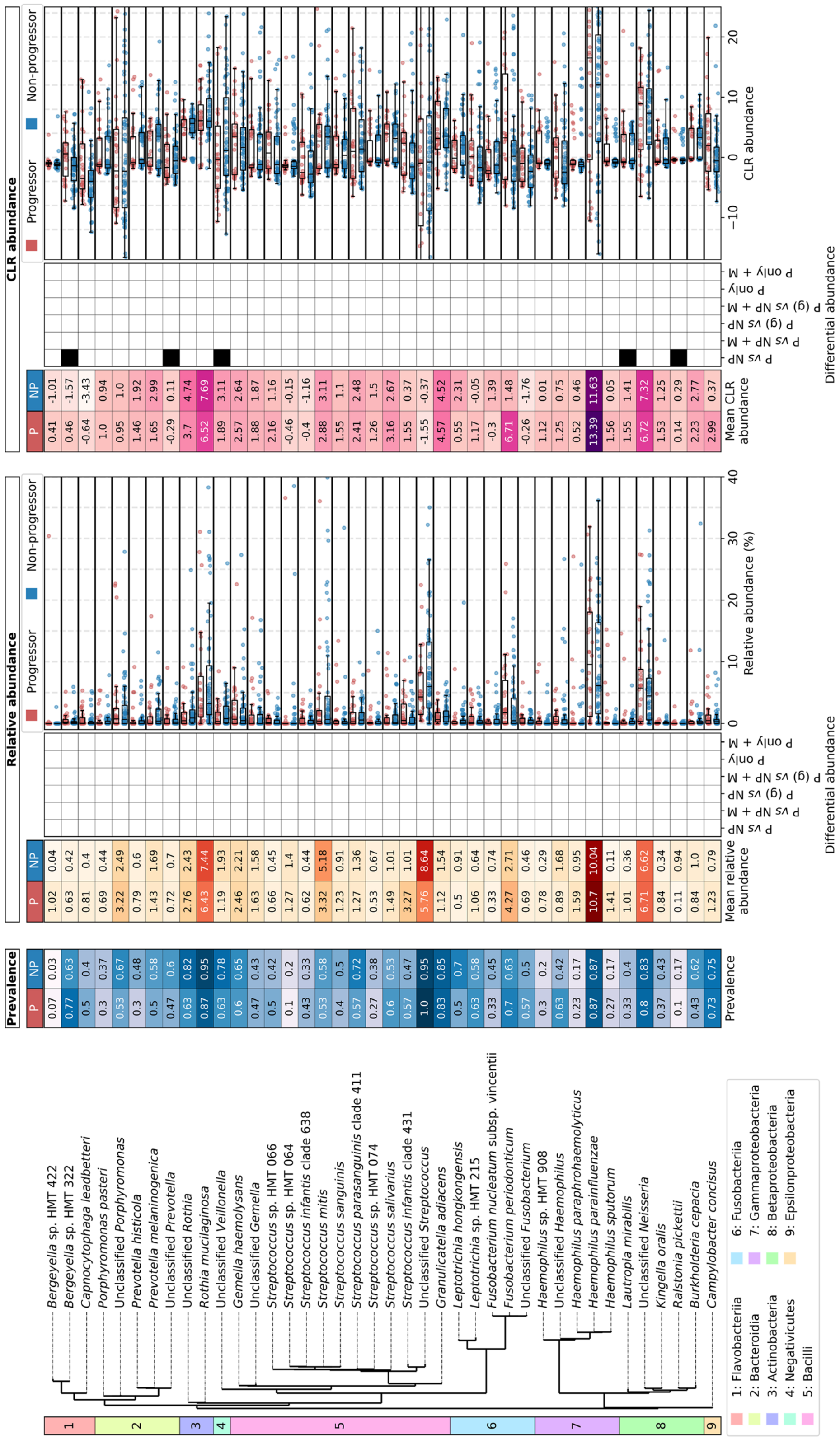


Fig. 3 (See legend on previous page.)

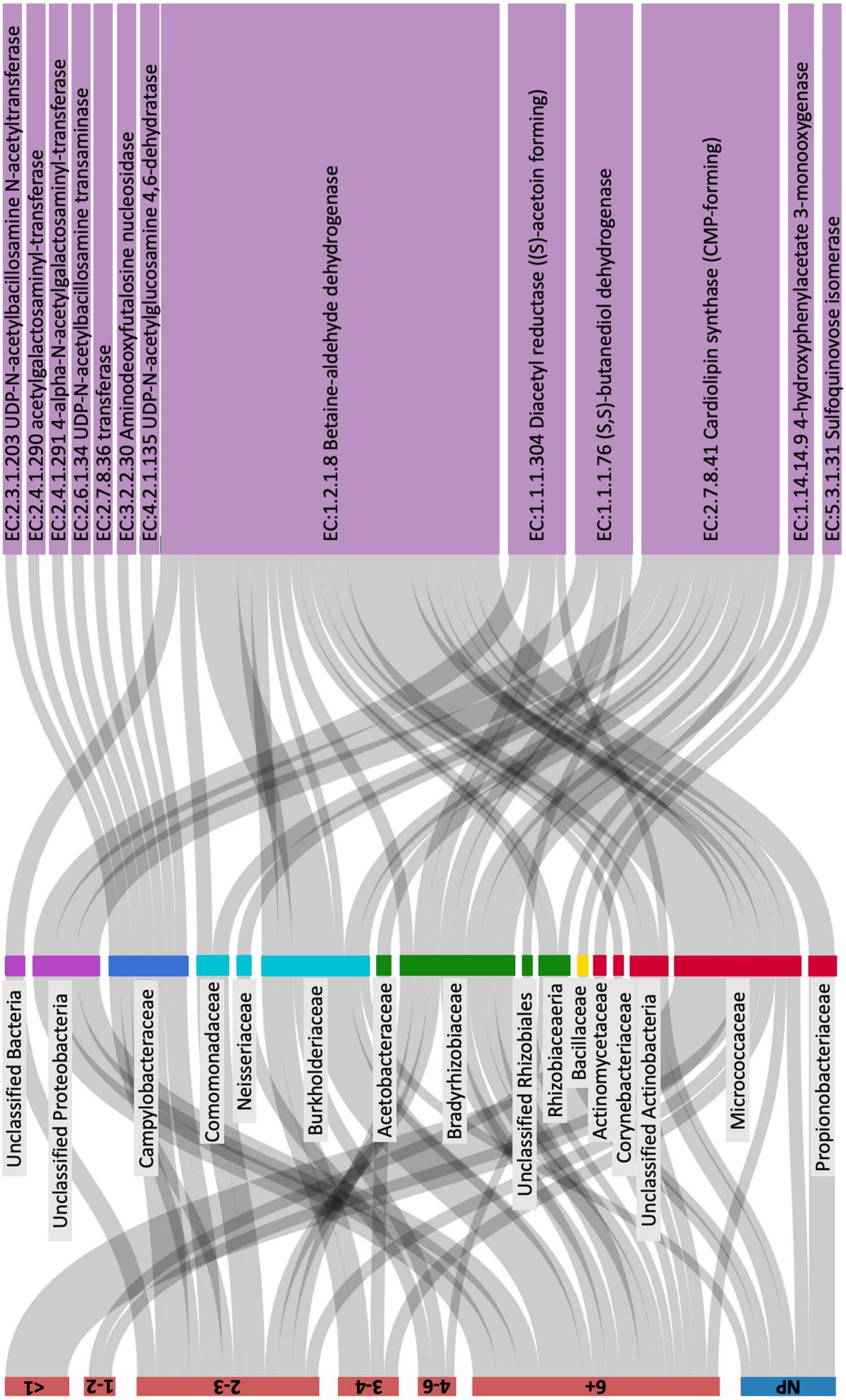


Fig. 4 JarrVis sankey plot showing the links between samples grouped by years to progression (left), microbial families (middle) and predicted EC numbers (right). Microbial families from different classes have differently coloured nodes. The sizes of nodes and connecting lines correspond to mean abundance values (rarefied gene copy numbers) for the contribution to microbial families and predicted EC numbers within each sample grouping. Only families/EC numbers with a mean abundance of > 100 copies are shown

and Comomonadaceae, 2–3 years), and some that only had abundant contributions from either NPs or those groups with a longer time to progression (*e.g.*, Burkholderiaceae, unclassified Actinobacteria and Rhizobiaceae). This also shows that in those that progressed soon after sampling (*i.e.,* the < 1 year group), there was only one family (Micrococcaceae) abundant enough to be visualised, and this family only contributed to one EC number (EC:1.2.1.8 Betaine-aldehyde dehydrogenase).

Discussion

This is the first study to use archival lesion swabs for the characterization of bacteria within OED. It is also the first to compare the diversity between progressing and non-progressing OED in a longitudinal design and to report on the functional potential of the bacteriome associated with risk of progression in OED. Our study has revealed that the bacterial make-up of the OED niche is similar to what has previously been described in the normal oral cavity, OPML and OSCC at the phylum and genus levels [29, 60, 65–67]. We also found that although participant factors such as age or smoking status have an overall larger impact on the oral microbiome than progression status, once these are controlled for there are also small differences in the microbiome that can be attributed to progression status (Fig. 1). In cross-sectional studies, differential relative abundance of certain genera, such as *Streptococcus*, *Prevotella, Campylobacter, Pseudomonas*, and *Fusobacterium,*is often seen between normal and diseased states [26, 27, 33, 64]. While there were slight differences in relative abundances of these genera, only *Campylobacter* showed significant differences in CLR-transformed abundance while controlling for clinical variables. *Campylobacter*has indeed been shown to be associated with OSCC tissue and is often higher in abundance in tumour tissue than in normal tissue [27, 68–72]. Furthermore, *Campylobacter spp.* has shown increased abundance in oral leukoplakia compared to contra-lateral controls [33]. However, taxonomic features at the species level that were determined to be significantly associated with either Ps or NPs were low abundance, and therefore are not likely to be large contributors to the niche. To determine whether these slight changes in abundances are significant at either genera or species level resolutions, it will likely take substantially larger sample sizes.

More recently, the literature has pointed to the role of the microbial metabolome and how the sum of the community as a whole may play a larger role in influencing the tissue microenvironment than any species alone [73, 74]. The concept of functional redundancy may explain how compositional variations of the microbiome associated with OPML and OSCC may collectively be contributing to a dysbiotic community. Our functional prediction analysis identified 15 enzymes (EC numbers) that were significantly differentially abundant with the time to progression in Ps (Supp. Figs. S9 and S10). In particular, seven of the identified enzymes were linked to the abundance of the *Campylobacteraceae* family in 2–3 year Ps. Several of these (EC:2.4.1.290, EC:2.4.1.291, EC:2.6.1.34) are related to the N-linked glycosylation, which was first described for the bacterial species *Campylobacter jejuni,* which belongs to this family [75], as well as protein glycosylation in general (EC:2.3.1.203). In particular, *N,N*–bacillosamine, a substrate and intermediate of these processes, may contribute to the pathogenicity of the bacteria and play a role as a virulence factor [76]. Amino sugar and nucleotide sugar metabolism were also associated with multiple enzymes (EC:2.3.1.203, EC:2.6.1.34, EC:4.2.1.135) and nucleotide sugar metabolism may provide glycosyl donors for glycosylation [77, 78]. While the specifc mechanisms by which glycoproteins affect pathogencity and virulence are not well-known, they may either be expressed on the surface of bacteria, where they are important for adhesion to host cells and thus the initiation of infection, or secreted by the bacteria, which may allow for the evasion of the host's immune system [79]. The repeated appearance of enzymes involved in glycosylation may indicate that those bacteria which are inhabiting the oral cavity in pre-malignant lesions that progress behave more pathogenically. However, it is uncertain as to which *Campylobacter* species are particularly abundant and whether they are pathogenic or commensal in more aggressive disease.

Given that our study design compared samples taken at an early stage of disease (mild /moderate OED), the lack of significant differences between P and NPs may indicate that changes in overall diversity as well as taxonomic shifts occur at later stages in progression, or perhaps detectable only after OSCC has been established. Further to this, a study looking at OSCC, normal, and OPML found that while OSCC samples clustered based on beta-diversity, pre-cancer and normal samples were mixed, indicating that there was not a great difference in the diversity between these groups [65]. This may support the notion that microbial changes in diversity change at a later stage. Larger studies that investigate the full spectrum of expertly graded OED are necessary.

A potential limitation of this work is that the samples were stored at ultra-low temperatures for a considerable period of time from date of collection to DNA extraction (mean 15.7 years, range 8.9 – 24.3 years). It is unknown how long-term storage may have affected such small samples with a potentially small biomass. However, this should also be viewed as strength, as older samples have longer follow-up and as a result, more robust outcome data. Alpha-diversity differed between Ps and NPs; however, beta diversity did not differ significantly among these groups alone (although it did in conjunction with other participant metadata). One of the reasons for this lack of significance in beta diversity may be due to a relatively small sample size. It is possible that the study suffered from a type II error due to lack of statistical power. However, prior to moving forward with a full-scale study, a pilot study of these small and invaluable samples was necessary. A significant strength is that this study examined patients with known outcomes, who had samples taken prior to developing disease. However, a limitation is that no longitudinal sampling points have been evaluated. A comparison between longitudinal samples can provide insightful results on the temporal changes. Future studies that employ repeated sampling are warranted. A caveat to this work is that 16S rRNA sequencing is not always capable of providing a high enough resolution to differentiate between closely related genera. In addition, functional profiles were established via prediction analysis (PICRUSt2) [48]. A metagenome sequencing approach may yield more comprehensive data for taxonomic assignment to the species level and provide more direct information on metabolic pathways for functional profiling based on pathway component genes.

Conclusions

In conclusion, for the first time, we have characterized the microbiome of low-grade OED with known outcome using 16S rRNA gene sequencing from annotated archival swabs. At the genus level, known oral cavity colonizers did not correlate with progression. The collective metabolic impact of the bacteriome trends toward a depletion of several enzymes that have been previously linked to cancer in progressing oral lesions but requires a larger sample size to show this more clearly. Having shown that quality NGS data can be obtained from archival oral swabs, larger prospective cohort studies to further explore the taxa and the function of the microbiome as a potential biomarker of risk in OED are warranted.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-023-02911-5.

Additional file 1. Supplementary results text and figures.

Additional file 2. Supplementary tables 1A-E showing (**A**) Participant and sample metadata, (**B**) Output from QIIME2 showing amplicon sequence variant (ASV), ASV name, assigned taxonomy, DNA sequence and abundance in samples, (**C**) PERMANOVA results, (**D**) Output from PICRUSt2 showing EC number, description of the enzyme produced and abundance in samples, and (**E**) Output from PICRUSt2 showing pathways, a description of the pathway and abundance in samples.

Additional file 3. Supplementary tables 2A-L showing MaAsLin2 differential abundance test results at the genus level.

Additional file 4. Supplementary tables 3A-L showing MaAsLin2 differential abundance test results at the species level.

Additional file 5. Supplementary tables 4A-L showing MaAsLin2 differential abundance test results on PICRUSt2 predicted metagenomes at the pathway level.

Additional file 6. Supplementary tables 5A-L showing MaAsLin2 differential abundance test results on PICRUSt2 predicted metagenomes at the enzyme commission number level.

Acknowledgements

Not applicable.

Authors' contributions

LDR and EAM were responsible for the project design. EAM, MEP, DML, MPR, LZ and LDR, contributed to data acquisition. RJW, EAM, MEP, BM, MGIL, WLL, and LDR contributed to data analysis, interpretation of results, and manuscript preparation. LDR was the principal investigator of this project. All authors have read, edited, and approved the final manuscript and agree to be accountable for the content of the work.

Funding

This work was supported by grants from the National Institute of Health (R01DE13124), the National Institute of Dental and Craniofacial Research (R01DE17013) and the Canadian Institutes of Health Research—Network for Canadian Oral Health Research New Frontiers Seed Grant (NGH – 154223; NFSG2019-17). RJW was funded by an I3V-DMRF Dr. David H. Hubel postdoctoral fellowship and EAM was funded by the Vanier Canada Scholars Program.

Availability of data and materials

Data is publicly available at the Gene Expression Omnibus (GEO) with accession GSE198811.

Declarations

Ethics approval and consent to participate

Study protocol and ethical approval for the OCPL study were obtained from the University of British Columbia and BC Cancer Agency Research Ethics Boards (UBC BC Cancer REB) (H98-61224). Participants were recruited by written informed consent. Ethics approval for the present study was obtained from the UBC BC Cancer (H20-00809) and the Dalhousie University Research Ethics Boards (2020–5102). All methods were carried out in accordance with the relevant regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 3 October 2022 Accepted: 25 March 2023
Published online: 06 April 2023

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7–30.
3. Canadian Cancer Statistics. Canadain Cancer Society, Statistics Canada, the Public Helath Agency of Canada, in collaboration with the provinical and territorial cancer registries. 2019. https://www.cancer.ca/statistics.

4. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. Oral Oncol. 2009;45(4–5):309–16.
5. Brocklehurst P, Kujan O, Glenny AM, Oliver R, Sloan P, Ogden G, et al. Screening programmes for the early detection and prevention of oral cancer. Cochrane Database Syst Rev. 2010;11:CD004150.
6. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014;64(1):9–29.
7. Amarasinghe H, Jayasinghe RD, Dharmagunawardene D, Attygalla M, Scuffham PA, Johnson N, et al. Economic burden of managing oral cancer patients in Sri Lanka: a cross-sectional hospital -based costing study. BMJ Open. 2019;9(7):e027661.
8. Pollaers K, Massingham I, Friedland PL, Farah CS. The economic burden of oral squamous cell carcinoma in Australia. J Oral Pathol Med Off Publ Int Assoc Oral Pathol Am Acad Oral Pathol. 2019;48(7):588–94.
9. Warnakulasuriya S, Johnson NW, van der Waal I. Nomenclature and classification of potentially malignant disorders of the oral mucosa. J Oral Pathol Med Off Publ Int Assoc Oral Pathol Am Acad Oral Pathol. 2007;36(10):575–80.
10. Napier SS, Speight PM. Natural history of potentially malignant oral lesions and conditions: an overview of the literature. J Oral Pathol Med Off Publ Int Assoc Oral Pathol Am Acad Oral Pathol. 2008;37(1):1–10.
11. Speight PM, Khurram SA, Kujan O. Oral potentially malignant disorders: risk of progression to malignancy. Oral Surg Oral Med Oral Pathol Oral Radiol. 2018;125(6):612–27.
12. Rosin MP, Cheng X, Poh C, Lam WL, Huang Y, Lovas J, et al. Use of allelic loss to predict malignant risk for low-grade oral epithelial dysplasia. Clin Cancer Res Off J Am Assoc Cancer Res. 2000;6(2):357–62.
13. Zhang L, Poh CF, Williams M, Laronde DM, Berean K, Gardner PJ, et al. Loss of heterozygosity (LOH) profiles–validated risk predictors for progression to oral cancer. Cancer Prev Res Phila Pa. 2012;5(9):1081–9.
14. Rock LD, Rosin MP, Zhang L, Chan B, Shariati B, Laronde DM. Characterization of epithelial oral dysplasia in non-smokers: first steps towards precision medicine. Oral Oncol. 2018;78:119–25.
15. Rock LD, Laronde DM, Lin I, Rosin MP, Chan B, Shariati B, et al. Dysplasia should not be ignored in lichenoid mucositis. J Dent Res. 2018;97(7):767–72.
16. Nankivell P, Mehanna H. Oral dysplasia: biomarkers, treatment, and follow-up. Curr Oncol Rep. 2011;13(2):145–52.
17. Dionne KR, Warnakulasuriya S, Zain RB, Cheong SC. Potentially malignant disorders of the oral cavity: current practice and future directions in the clinic and laboratory. Int J Cancer. 2015;136(3):503–15.
18. Kim SS, Ruiz VE, Carroll JD, Moss SF. Helicobacter pylori in the pathogenesis of gastric cancer and gastric lymphoma. Cancer Lett. 2011;305(2):228–38.
19. Nagaraja V, Eslick GD. Systematic review with meta-analysis: the relationship between chronic Salmonella typhi carrier status and gall-bladder cancer. Aliment Pharmacol Ther. 2014;39(8):745–50.
20. Toprak NU, Yagci A, Gulluoglu BM, Akin ML, Demirkalem P, Celenk T, et al. A possible role of Bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis. 2006;12(8):782–6.
21. Ahn J, Segers S, Hayes RB. Periodontal disease, Porphyromonas gingivalis serum antibody levels and orodigestive cancer mortality. Carcinogenesis. 2012;33(5):1055–8.
22. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res. 2012;22(2):299–306.
23. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012;22(2):292–8.
24. Perera M, Al-Hebshi NN, Speicher DJ, Perera I, Johnson NW. Emerging role of bacteria in oral carcinogenesis: a review with special reference to perio-pathogenic bacteria. J Oral Microbiol. 2016;8:32762.
25. Zhao H, Chu M, Huang Z, Yang X, Ran S, Hu B, et al. Variations in oral microbiota associated with oral cancer. Sci Rep. 2017;7(1):11773.
26. Al-Hebshi NN, Nasher AT, Maryoud MY, Homeida HE, Chen T, Idris AM, et al. Inflammatory bacteriome featuring Fusobacterium nucleatum and Pseudomonas aeruginosa identified in association with oral squamous cell carcinoma. Sci Rep. 2017;7(1):1834.
27. Perera M, Al-Hebshi NN, Perera I, Ipe D, Ulett GC, Speicher DJ, et al. Inflammatory bacteriome and oral squamous cell carcinoma. J Dent Res. 2018;97(6):725–32.
28. Yang CY, Yeh YM, Yu HY, Chin CY, Hsu CW, Liu H, et al. Oral microbiota community dynamics associated with oral squamous cell carcinoma staging. Front Microbiol. 2018;9:862.
29. Zhang L, Liu Y, Zheng HJ, Zhang CP. The oral microbiota may have influence on oral cancer. Front Cell Infect Microbiol. 2019;9:476.
30. Hanahan D. Hallmarks of cancer: new dimensions. Cancer Discov. 2022;12(1):31–46.
31. Elinav E, Garrett WS, Trinchieri G, Wargo J. The cancer microbiome. Nat Rev Cancer. 2019;19(7):371–6.
32. Mok SF, Karuthan C, Cheah YK, Ngeow WC, Rosnah Z, Yap SF, et al. The oral microbiome community variations associated with normal, potentially malignant disorders and malignant lesions of the oral cavity. Malays J Pathol. 2017;39(1):1–15.
33. Amer A, Galvin S, Healy CM, Moran GP. The Microbiome of Potentially Malignant Oral Leukoplakia Exhibits Enrichment for Fusobacterium, Leptotrichia, Campylobacter, and Rothia Species. Front Microbiol. 2017;8:2391.
34. Government of British Columia. Population estimates. Available from: https://www2.gov.bc.ca/gov/content/data/statistics/people-population-community/population/population-estimates
35. Zhang L, Williams M, Poh CF, Laronde D, Epstein JB, Durham S, et al. Toluidine blue staining identifies high-risk primary oral premalignant lesions with poor outcome. Cancer Res. 2005;65(17):8017–21.
36. Barnes L, Eveson HW, Reichart P, Sidransky D. World Health Organization Classification of Tumours. Pathology and genetics of head and neck tuours. Lyon: IARC Presss; 2005.
37. Dean, AG, Sillivan, KM, Soe, MM. OpenEpi: open source epidemiologic statistics for public health. Version 3.01. 2013. Available from: www.OpenEpi.com
38. Kassambara, A. Kosinski, M. Biecek, P. survminer: Drawing Survival Curve using "ggplot2." R package; 2021. 10.1186/s12903-023-02911-5. Available from: https://rpkgs.datanovia.com/survminer/index.html.
39. Allen HK, Bayles DO, Looft T, Trachsel J, Bass BE, Alt DP, et al. Pipeline for amplifying and analyzing amplicons of the V1–V3 region of the 16S rRNA gene. BMC Res Notes. 2016;2(9):380.
40. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852–7.
41. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17(1):10.
42. 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):581–3.
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(85):2825–30.
44. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-596.
45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
46. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database. 2010;2010(0):baq013–baq013.
47. Janssen S, McDonald D, Gonzalez A, Navas-Molina JA, Jiang L, Xu ZZ, et al. Phylogenetic placement of exact amplicon sequences improves associations with clinical information. Chia N, editor. mSystems. 2018;3(3):e00021-18.
48. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020;38(6):685–8.
49. McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. Watson M, editor. PLoS ONE. 2013;8(4):e61217.
50. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011;5(2):169–72.

51. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. eLife. 2017;6:e21887.
52. Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, et al. Scikit-bio: a bioinformatics library for data scientists, students, and developers. 2020. Available from: http://scikit-bio.org.
53. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72.
54. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community ecology package. 2020.
55. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, et al. Multivariable association discovery in population-scale meta-omics studies. PLoS Comput Biol. 2021;17(11):e1009442.
56. Comeau AM, Douglas GM, Langille MGI. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. Eisen J, editor. mSystems. 2017;2(1). Available from: https://journals.asm.org/doi/10.1128/mSystems.00127-16. [cited 2022 Feb 5].
57. Dhwani, D. JarrVis. https://github.com/dhwanidesai/JarrVis.
58. Douglas GM, Kim S, Langille MGI, Shapiro BJ. Efficient computation of contributional diversity metrics from microbiome data with FuncDiv. Bioinformatics. 2023;39(1).
59. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;15(8):2224.
60. Pushalkar S, Ji X, Li Y, Estilo C, Yegnanarayana R, Singh B, et al. Comparison of oral microbiota in tumor and non-tumor tissues of patients with oral squamous cell carcinoma. BMC Microbiol. 2012;20(12):144.
61. Nearing JT, DeClercq V, Van Limbergen J, Langille MGI. Assessing the variation within the oral microbiome of healthy adults. mSphere. 2020;5(5):e00451-20.
62. Karpiński TM. Role of Oral Microbiota in Cancer Development. Microorganisms. 2019;7(1):E20.
63. Guerrero-Preston R, Godoy-Vitorino F, Jedlicka A, Rodríguez-Hilario A, González H, Bondy J, et al. 16S rRNA amplicon sequencing identifies microbiota associated with oral cancer, human papilloma virus infection and surgical treatment. Oncotarget. 2016;7(32):51320–34.
64. Schmidt BL, Kuczynski J, Bhattacharya A, Huey B, Corby PM, Queiroz ELS, et al. Changes in abundance of oral microbiota associated with oral cancer. PLoS ONE. 2014;9(6):e98741.
65. Li Z, Chen G, Wang P, Sun M, Zhao J, Li A, et al. Alterations of the oral microbiota profiles in chinese patient with oral cancer. Front Cell Infect Microbiol. 2021;11:780067.
66. Pushalkar S, Mane SP, Ji X, Li Y, Evans C, Crasta OR, et al. Microbial diversity in saliva of oral squamous cell carcinoma. FEMS Immunol Med Microbiol. 2011;61(3):269–77.
67. Chattopadhyay I, Verma M, Panda M. Role of oral microbiome signatures in diagnosis and prognosis of oral cancer. Technol Cancer Res Treat. 2019;1(18):1533033819867354.
68. Nie F, Wang L, Huang Y, Yang P, Gong P, Feng Q, et al. Characteristics of microbial distribution in different oral niches of oral squamous cell carcinoma. Front Cell Infect Microbiol. 2022;12:905653.
69. Arthur RA, Dos Santos BR, Ximenez JPB, Merlin BL, de Andrade MR, Neto JV, et al. Microbiome and oral squamous cell carcinoma: a possible interplay on iron metabolism and its impact on tumor microenvironment. Braz J Microbiol Publ Braz Soc Microbiol. 2021;52(3):1287–302.
70. Downes JE, Swann PG, Holmes RS. Ultraviolet light-induced pathology in the eye: associated changes in ocular aldehyde dehydrogenase and alcohol dehydrogenase activities. Cornea. 1993;12(3):241–8.
71. Yang K, Wang Y, Zhang S, Zhang D, Hu L, Zhao T, et al. Oral microbiota analysis of tissue pairs and saliva samples from patients with oral squamous cell carcinoma - a pilot study. Front Microbiol. 2021;12:719601.
72. Zhang Z, Yang J, Feng Q, Chen B, Li M, Liang C, et al. Compositional and functional analysis of the microbiome in tissue and saliva of oral squamous cell carcinoma. Front Microbiol. 2019;10:1439.
73. Al-Hebshi NN, Borganakke WS, Johnson NW. The microbiome of oral squamous cell carcinomas: a functional perspectivethe oral microbiota may have influence on oral cancer. Curr Oral Health Rep. 2019;6:145–60.
74. Tian L, Wang XW, Wu AK, Fan Y, Friedman J, Dahlin A, et al. Deciphering functional redundancy in the human microbiome. Nat Commun. 2020;11(1):6217.
75. Wacker M, Linton D, Hitchen PG, Nita-Lazar M, Haslam SM, North SJ, et al. N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli. Science. 2002;298(5599):1790–3.
76. Morrison MJ, Imperiali B. The renaissance of bacillosamine and its derivatives: pathway characterization and implications in pathogenicity. Biochemistry. 2014;53(4):624–38.
77. Samuel G, Reeves P. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydr Res. 2003;338(23):2503–19.
78. Mikkola S. Nucleotide sugars in chemistry and biology. Mol Basel Switz. 2020;25(23):5755.
79. Lin B, Qing X, Liao J, Zhuo K. Role of protein glycosylation in host-pathogen interaction. Cells. 2020;9(4):1022.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.