Original Article

Oral microbiome and risk of malignant esophageal lesions in a high-risk area of China: A nested case-control study

Fangfang Liu[1*], Mengfei Liu[1*], Ying Liu[1*], Chuanhai Guo[1], Yunlai Zhou[2], Fenglei Li[3], Ruiping Xu[4], Zhen Liu[1], Qiuju Deng[1], Xiang Li[1], Chaoting Zhang[1], Yaqi Pan[1], Tao Ning[1], Xiao Dong[2], Zhe Hu[1], Huanyu Bao[1], Hong Cai[1], Isabel Dos Santos Silva[5], Zhonghu He[1], Yang Ke[1]

[1]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Laboratory of Genetics, Peking University Cancer Hospital & Institute, Beijing 100142, China; [2]Novogene Co., Ltd, Beijing 100080, China; [3]Hua County People's Hospital, Anyang 456400, China; [4]Anyang Cancer Hospital, Anyang 455000, China; [5]Department of Non-communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK

*These authors contributed equally to this work.

Correspondence to: Prof. Yang Ke. Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Laboratory of Genetics, Peking University Cancer Hospital & Institute, Beijing 100142, China. Email: keyang@bjmu.edu.cn; Prof. Zhonghu He. Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Laboratory of Genetics, Peking University Cancer Hospital & Institute, Beijing 100142, China. Email: zhonghuhe@foxmail.com; Prof. Isabel Dos Santos Silva. Department of Non-communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK. Email: Isabel.silva@lshtm.ac.uk; Prof. Hong Cai. Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Laboratory of Genetics, Peking University Cancer Hospital & Institute, Beijing 100142, China. Email: drhcai@gmail.com.

Abstract

Objective: We aimed to prospectively evaluate the association of oral microbiome with malignant esophageal lesions and its predictive potential as a biomarker of risk.

Methods: We conducted a case-control study nested within a population-based cohort with up to 8 visits of oral swab collection for each subject over an 11-year period in a high-risk area for esophageal cancer in China. The oral microbiome was evaluated with 16S ribosomal RNA (rRNA) gene sequencing in 428 pre-diagnostic oral specimens from 84 cases with esophageal lesions of severe squamous dysplasia and above (SDA) and 168 matched healthy controls. DESeq analysis was performed to identify taxa of differential abundance. Differential oral species together with subject characteristics were evaluated for their potential in predicting SDA risk by constructing conditional logistic regression models.

Results: A total of 125 taxa including 37 named species showed significantly different abundance between SDA cases and controls (all $P<0.05$ & false discovery rate-adjusted $Q<0.10$). A multivariate logistic model including 11 SDA lesion-related species and family history of esophageal cancer provided an area under the receiver operating characteristic curve (AUC) of 0.89 (95% CI, 0.84–0.93). Cross-validation and sensitivity analysis, excluding cases diagnosed within 1 year of collection of the baseline specimen and their matched controls, or restriction to screen-endoscopic-detected or clinically diagnosed case-control triads, or using only bacterial data measured at the baseline, yielded AUCs>0.84.

Conclusions: The oral microbiome may play an etiological and predictive role in esophageal cancer, and it holds promise as a non-invasive early warning biomarker for risk stratification for esophageal cancer screening programs.

Keywords: Early warning biomarker; esophageal squamous cell carcinoma; oral microbiome; risk prediction

Submitted Nov 25, 2020. Accepted for publication Dec 16, 2020.
doi: 10.21147/j.issn.1000-9604.2020.06.07
View this article at: https://doi.org/10.21147/j.issn.1000-9604.2020.06.07

Introduction

Esophageal cancer is the seventh most common cancer worldwide (1). Fifty-five percent of new cases occur annually in China, and 90% of these are esophageal squamous cell carcinoma (ESCC) (2). Tobacco and alcohol consumption are well-established risk factors for esophageal cancer in western countries, but contribute little to ESCC incidence in high-risk areas such as Anyang of China (2). The etiology of ESCC needs to be further investigated (2). Most ESCC cases are diagnosed at an advanced stage which confers an unfavorable prognosis. Early detection has been shown to improve survival and reduce mortality from the disease (2), with upper gastrointestinal endoscopic screening being widely accepted as an optimal secondary prevention strategy for esophageal lesions of severe dysplasia and above (SDA), including severe squamous dysplasia, carcinoma *in situ* (CIS), and ESCC, in high-risk populations. However, this approach has disadvantages such as its potential for complications (3). Identification of high-risk individuals in the general population through the use of minimally-invasive biomarkers could help to maximize the benefits of endoscopic screening by targeting those most likely to benefit.

Emerging evidence has linked the human microbiome with diseases such as cancer. The development of microbiome-based risk prediction models for some types of cancer, such as colorectal cancer, has opened new research avenues, by demonstrating that the microbiome may be a valid non-invasive biomarker of risk (4,5). Oral bacteria, in particular periodontal pathogens, together with indicators of oral health (e.g., tooth loss) have been reported to be associated with ESCC and its precursor lesions (2,6,7). The anatomic proximity of the esophagus to the oral cavity likely renders the esophagus vulnerable to the effects of oral dysbiosis (8). We thus hypothesized that the oral microbiome is associated with the risk of developing SDA lesions and thus may be useful as a non-invasive biomarker of risk for SDA lesions. Only a few studies have so far characterized the oral microbiome in SDA lesions and interpretation of their findings has been hampered by the fact they relied on a single measurement taken from a one-off oral specimen collection, did not use optimal methods of statistical analysis for differential comparison of taxa, and did not evaluate the predictive value of oral bacteria (7,9). Given the dynamics of the human microbiome (5), prospective follow-up studies with repeat specimen sampling are warranted to better understand the role of oral microbiome in malignant esophageal lesions.

The present case-control study, nested within a population-based cohort in a high-risk area for esophageal cancer in rural China with collection of multiple (up to 8) oral swabs over an 11-year follow-up period (10,11), aims to assess the association between oral microbiome and the risk of esophageal SDA and to investigate the potential value of this biomarker in predicting risk.

Materials and methods

Study design and participants

The subjects for this nested case-control study were selected from the prospective population-based endoscopic Anyang Esophageal Cancer Cohort Study (AECCS; 9,035) and its oral sub-cohort (4,073) in rural Anyang, China, as previously described (10,11). Eligible participants (i.e. permanent residents in cluster-sampled villages, aged 25–65 years, with no prior history of cancer, cardiovascular illness, or infection with Hepatitis B, Hepatitis C, or Human Immunodeficiency Viruses) were visited in their villages a maximum of 8 times for collection of oral swabs including 3 visits at 2.5-year intervals from 2006 to 2013 (endoscopic inspection of the esophagus was also performed at each visit), and 5 bi-annual visits from 2013 to 2015 (*Figure 1*).

Cases and controls were selected from AECCS who had provided a baseline oral swab at enrollment into the cohort (*Figure 1*). Cases included both screen-endoscopic-detected SDA cases and clinically diagnosed SDA cases diagnosed after collection of the baseline oral swab, but prior to July 2017, when follow-up of the cohort for the present analysis ended. The clinically diagnosed SDA cases were identified through annual active door-to-door interviews and through passive linkage with claims data from the New Rural Cooperative Medical Scheme. For each case, two controls were randomly selected among cohort members who did not have SDA at the time of diagnosis of the case (incidence density sampling) matching on gender, village of residence, age at cohort entry (5-year intervals), and number and timing (±1-year) of oral swab collection and endoscopic examination.

The follow-up time for the included cases and controls was estimated from time of enrolment into the cohort to the time of a SDA diagnosis for cases, and corresponding time for their two matched controls. The median follow-up time was calculated by using the reversed Kaplan-Meier method (12).

The study was performed in accordance with the Declaration of Helsinki. Research protocols were approved

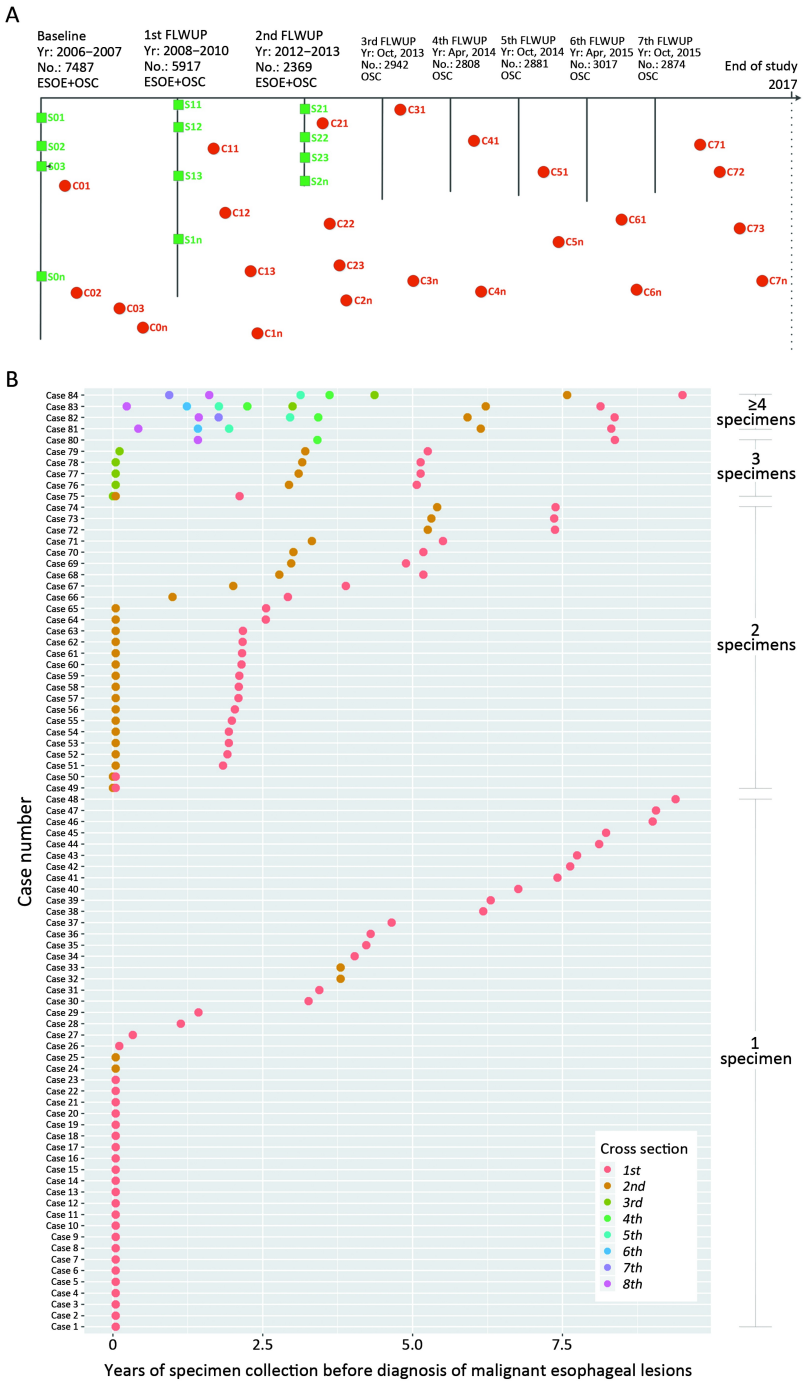


Figure 1 SDA cases and their oral specimens used in this nested case-control study from Anyang, China, 2006–2017. (A) Visual overview of SDA cases taken from the prospective AECCS cohort and its oral sub-cohort (Primary outcome: SDA lesions). C denotes clinically diagnosed SDA cases (32 cases) which were identified by active follow-up. Interval SDA cases "C71, C72, C73...C7n" occurred sequentially at interval 8 from the 7th follow-up to July 1st, 2017. S denotes screen-endoscopic-detected SDA cases (52 cases) diagnosed by endoscopy. Screened SDA cases "S21, S22, S23...S2n" were diagnosed sequentially at the 2nd follow-up; (B) Number and time frame of oral specimen collection for 84 SDA cases. Points are plotted according to years of oral specimen collection prior to diagnosis of SDA lesions. Color indicates cross-sections at which specimens were collected. AECCS, Anyang Esophageal Cancer Cohort Study; ESOE, endoscopic screening of esophagus; FLWUP, follow-up; No., number; OSC, oral swab collection; SDA, severe dysplasia and above; Yr., year.

by the Institutional Review Board of the Peking University Cancer Hospital & Institute. All participants provided written informed consent.

Oral specimen and questionnaire data collection

Using saline-moistened cotton swabs, exfoliated oral cells were collected from the upper and lower lips, left and right sides of the hard palate, the buccal mucosa, the top and the bottom of the tongue, and the surface of the gingiva. Cells were rinsed with 0.9% saline solution and frozen at −80 ℃ pending testing after centrifugation (10,11). A total of 143 pre-diagnosis oral specimens were provided by the 84 cases (48 cases provided only one specimen; 26 provided two; 6 provided three; and 4 provided five or more specimens). Of these 143 specimens, 49 were collected within 15 d before the diagnosis of the SDA lesion, which were all provided by screen-endoscopic-detected SDA cases (*Figure 1*).

A one-on-one computer-aided interview on demographic characteristics and potential risk factors for esophageal cancer (~50 items) was administered by a trained interviewer at the baseline visit conducted at enrolment into the cohort.

Laboratory handling and bioinformatics

DNA was extracted using the E.Z.N.A. Mag-Bind Tissue DNA Kit (Omega Bio-Tek, Inc., Norcross, USA). The 16S ribosomal (rRNA) gene V3–V4 regions were amplified using universal primers (341F 5'-CCTAYGGGRBGCA SCAG-3' and 806R 5'-GGACTACNNGGGTATCTA AT-3') and sequenced on the Ion S5 XL sequencing platform.

Multiplexed and barcoded sequences were deconvoluted. High-quality sequences were obtained according to the Cutadapt (V1.9.1) quality-controlled process. Chimera sequences were detected using the UCHIME algorithm and then removed. Filtered sequence reads were clustered into operational taxonomic units (OTUs). OTUs with a mean relative abundance ≥0.001% were assigned to taxa using the expanded Human Oral Microbiome Database (eHOMD) with ≥97% sequence similarity. From 428 oral specimens (143 from cases; 285 from controls), we obtained 32,917,908 ($\bar{x}\pm s$, 76,911±10,444) high-quality sequence reads, with similar numbers of reads per specimen for both case and control groups (*Supplementary Table S1*). A total of 15 phyla, 44 classes, 79 orders, 147 families, 324 genera, and 720 species were identified and included in our analysis.

Quality control

Specimens from any given case-control triad were included in the same batch and tested blindly. Ten replicate aliquots of oral cell DNA from eight volunteers were mixed and included in the 5 sequencing batches (2 replicates per batch) as quality control samples. The intra-plate and inter-plate coefficients of variation (CV) for the Shannon diversity index and observed-species of the quality control samples were all <7.0% (*Supplementary Table S2*). Rarefaction curves and the species-accumulation boxplot indicate sufficient sequence depth and adequate sample size, respectively (*Supplementary Figure S1*,*S2*).

Statistical analysis

Dataset description

To obtain stable measurements of bacterial populations and to account for heterogeneity in the number and timing of specimens from different subjects, a full averaged dataset was produced for bacterial abundance comparison and prediction model establishment. This dataset contained a total of 84 SDA cases providing 143 oral specimens and 168 matched controls providing 285 oral specimens (*Figure 1*). The bacterial population values for each specimen provided by an individual were averaged to produce single values at each taxonomic level (e.g., class, species) for that individual.

Overall diversity comparison

Trends of α diversity (Shannon index) with years of specimen collection prior to diagnosis of malignant esophageal lesions were evaluated using linear mixed-effects (LME) regression (LME function in R) by treating the subject as a random effect. Differences in α diversity between cases and controls were also analyzed by LME regression. Differences in overall bacterial community composition (β diversity) according to case and control status were assessed with permutational multivariate analysis of variance (PERMANOVA; adonis function in R) by treating matched case-control triads as strata.

Association analysis

To compare relative abundance of taxa in SDA cases and controls at each level (phylum to species), DESeq (DESeq2 package, R) with variance and mean linked by local multivariable regression, which is an optimal method for microbiome data analysis, was performed based on the full averaged dataset (13,14). Taxa were considered

significantly differentially abundant between groups if P<0.05 & the false discovery rate (FDR)-adjusted Q<0.10.

Prediction analysis

To establish a final prediction model for risk of SDA lesions and determine which species should be retained in the final prediction model, analysis was carried out based on the full averaged dataset as follows (*Supplementary Figure S3*). For each of the named and cultured differential species selected by DESeq analysis, multiple species-specific classifiers (low carriage *vs.* high carriage), derived from a series of cut-off points ranging from quantiles 5% to 95% (5% per step) of the relative abundance in the control group, were evaluated in separate univariate conditional logistic regression models (dependent variable: SDA status). Taking both error probability and effect size into consideration, the optimal classifiers for each species with the lowest sum of odds ratio rank and reverse P value rank, together with subject characteristics were included in the multivariate conditional logistic model. Their retention in the final prediction model was determined using the Akaike information criterion (step AIC function, MASS package, R). The area under the receiver operating characteristic curve (AUC) and the DeLong test were adopted to evaluate the performance of the prediction model. Leave-one triad-out cross-validation was used to estimate the generalization error on the basis of predicted probabilities for each case-control triad from models built on all the remaining triads.

Temporal stability assessment

To assess the temporal stability of the relative abundance of oral species within and between individuals, we used the metrics of mean, standard deviation, and CV as employed by Utter *et al* (15). A total of 128 specimens provided from 10 cases and 18 controls (each with three or more serial specimens) were included in this analysis. For each species, means and CVs for each individual were calculated based on the relative abundance of three or more specimens from this individual. The mean CV (intra-individual CV) was the mean of all the CVs calculated from all included individuals; the overall CV was calculated based on the relative abundances from all specimens provided by all included individuals.

Sensitivity analysis

To reduce the likelihood of reverse causation, the following sensitivity analyses were performed: 1) including only cases diagnosed more than 1 year after collection of the baseline specimen and their matched controls, but using the average microbiome data from all their collected oral specimens (strictly averaged dataset; 55 cases with 87 specimens and 110 controls with 176 specimens); 2) including all enrolled study subjects, but using only microbiome data from their oral specimens collected at baseline (full baseline dataset; 84 cases and 168 controls, each with a single baseline specimen); and 3) including only cases diagnosed more than 1 year after collection of the baseline specimen and their matched controls, and using only microbiome data from their oral specimens collected at baseline (strict baseline dataset; 55 cases and 110 controls, each with a single baseline specimen). Also, stratified analysis was carried out by separating screen-endoscopic-detected and clinically diagnosed case-control triads. Model performance was also recalculated using 75th quantile cut-off points instead of optimal thresholds for classification of low *vs.* high carriage of the predictive bacteria.

All multivariate models included level of education, type of employment, cigarette smoking, alcohol consumption, and family history of esophageal cancer unless otherwise specified. All analysis was carried out using R statistical software (Version 3.4.3; R Foundation for Statistical Computing, Vienna, Austria). P values less than 0.05 (two-sided) were considered to be statistically significant.

Results

Participant characteristics

Median follow-up time for study participants was of 8.7 (interquartile range: 5.2–9.7) years. Cases and controls were of a similar age and gender (matching variables) and had a similar educational level, type of employment, and cigarette smoking and alcohol intake habits. Cases were, however, more likely to have a family history of esophageal cancer than controls (15.5% *vs.* 7.1%, P=0.037) (*Table 1*).

Overall microbiome diversity in relation to malignant esophageal lesions

No significant trend over years of specimen collection prior to diagnosis of malignant esophageal lesions in the Shannon diversity index was found for SDA cases (P=0.124) or controls (P=0.425) (*Supplementary Figure S4*). Between groups, cases showed a slightly higher Shannon diversity index than controls (P=0.044). Cases differed significantly from controls in overall oral microbiome composition (β diversity) neither when measured by unweighted (P=0.248) nor when measured by weighted UniFrac distances (P=0.590) (*Supplementary Figure S4*).

Table 1 Selected demographic and baseline behavioral characteristics of cases of malignant esophageal lesions and matched controls from Anyang, China, 2006–2017

Variables*	Cases (N=84)**	Controls (N=168)##	P***
Age (IQR) (year)	57 (51–61)	56 (51–61)	0.140
Gender			NA
Female	40 (47.6)	80 (47.6)	
Male	44 (52.4)	88 (52.4)	
Education level			0.520
Primary school (1–6 years) or below	53 (63.1)	100 (59.5)	
Junior high school (7–9 years) or above	31 (36.9)	68 (40.5)	
Type of employment			0.680
Farming	66 (78.6)	135 (80.4)	
Non-farming	18 (21.4)	33 (19.6)	
Cigarette smoking*			0.890
No	53 (63.1)	105 (62.5)	
Yes	31 (36.9)	63 (37.5)	
Alcohol consumption#			0.300
No	72 (85.7)	136 (81.0)	
Yes	12 (14.3)	32 (19.0)	
Family history of esophageal cancer			0.037
No	71 (84.5)	156 (92.9)	
Yes	13 (15.5)	12 (7.1)	

IQR, interquartile range; NA, not applicable; SDA, severe dysplasia and above; *, Cigarette smoking is defined as consumption of one cigarette or more per day for at least 12 months; #, Alcohol consumption is defined as consumption of Chinese liquor twice per week or more for at least 12 months; **, Cases were subjects with esophageal lesions of severe dysplasia and above (SDA) including severe squamous dysplasia, carcinoma *in situ*, and esophageal squamous cell carcinoma; ##, Controls were subjects without SDA lesions. ***, P values were calculated by univariate conditional logistic regression analyses.

Taxa associated with malignant esophageal lesions

Based on DESeq analysis, a higher abundance of 15 taxa including 6 species was found to be associated with decreased risk of SDA lesions, and a higher abundance of 110 taxa including 66 species was associated with increased risk of SDA lesions (all P<0.05 & Q<0.10). Of the 72 species with differential abundance between cases and controls, 37 were named and cultured according to eHOMD (*Figure 2*; *Supplementary Table S3*). The species Fusobacterium nucleatum which is known to be associated with periodontal diseases (16,17), and all of its higher taxonomic levels were among the above taxa with positive associations.

Species-level prediction model for malignant esophageal lesions

A total of 11 species of 37 named and cultured differential species selected by DESeq analysis, together with family history of esophageal cancer were retained in the final model predicting risk of SDA lesions. These species and their corresponding optimal cut-off points for relative abundance are shown in *Table 2*. Higher carriage of the predictive species was associated with increased risk of SDA lesions, with adjusted ORs ranging from 1.98 (Prevotella baroniae) to 10.93 (Lachnoanaerobaculum umeaense). For Fusobacterium nucleatum, the adjusted OR was 3.85 (95% CI, 1.12–13.24).

The AUC was 0.89 (95% CI, 0.84–0.93) for the final model, which was constructed based on the full averaged dataset (*Figure 3*). Leave-one triad-out cross-validation provided similar AUC statistics (AUC, 0.89; 95% CI, 0.88–0.89). After exclusion of cases which were diagnosed within 1 year of collection of the baseline specimen and matched controls for these cases, the $AUC_{\text{strictly averaged dataset}}$ was 0.85 (95% CI, 0.80–0.91). When stratifying by case type, the AUC for screen-endoscopic-detected cases and matched controls was 0.90 (95% CI, 0.86–0.95) and the AUC for clinically diagnosed SDA cases and matched controls was 0.88 (95% CI, 0.81–0.94) (*Supplementary*

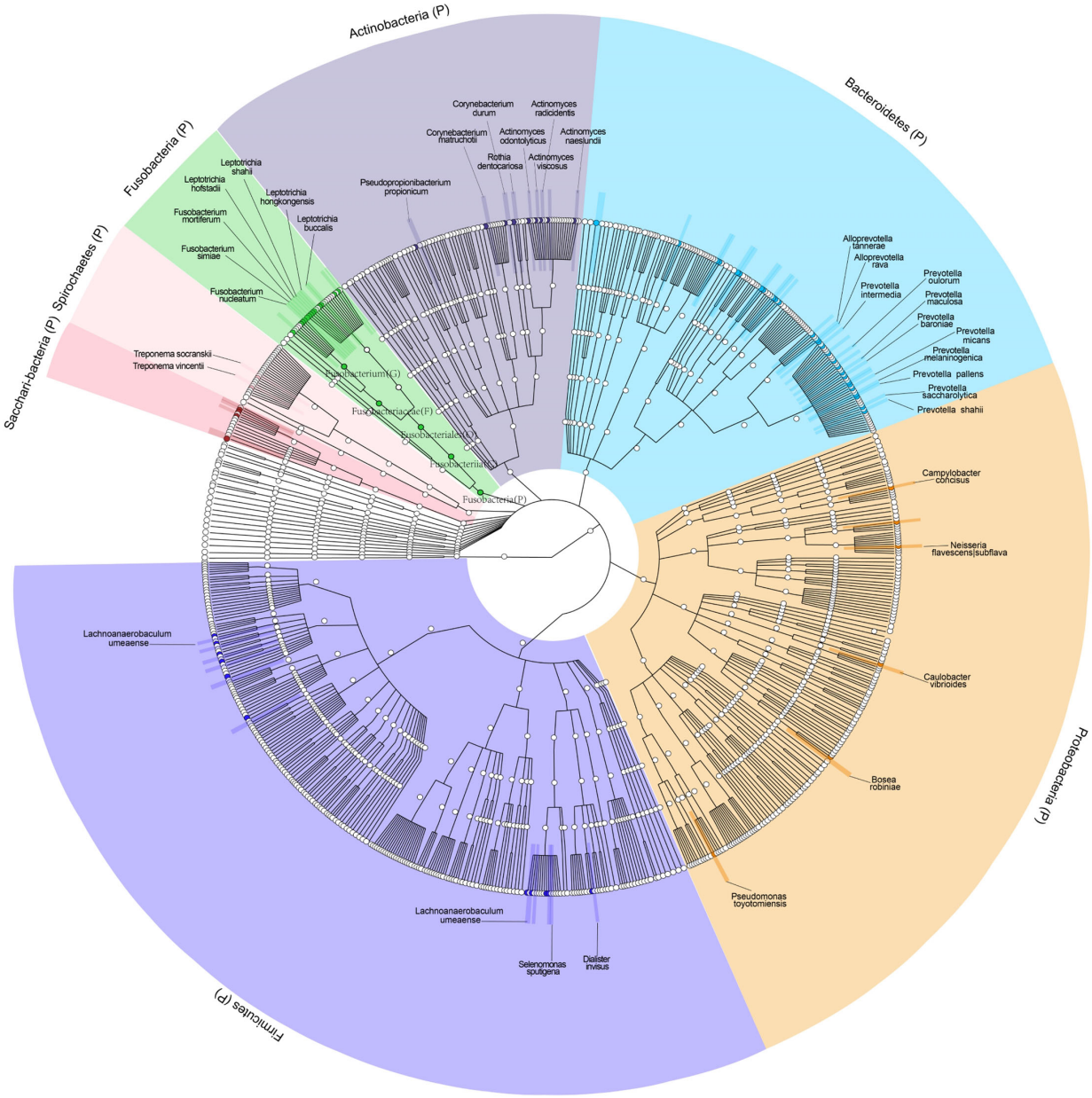


Figure 2 Phylogenetic tree of taxa associated with malignant esophageal lesions. A total of 125 taxa including 72 species (marked by colorful bars around the phylogenetic tree) was found to be associated with risk of SDA lesions (all P<0.05 & FDR-adjusted Q<0.10). Among these 72 species, 37 of them were named and cultured according to eHOMD (marked with species names in the Figure). The species *Fusobacterium nucleatum* which is known to be associated with periodontal diseases, and all of its higher taxonomic levels (marked in the green area of the Figure) were among the above taxa with positive associations. (P), (C), (O), (F), and (G) indicate bacterial taxa at the level of Phylum, Class, Order, Family, and Genus.

Figure S5). Additionally, when analysis was limited to baseline specimens, the AUCs were also similar [$AUC_{full\ baseline\ dataset}$: 0.84 (95% CI, 0.79–0.89); $AUC_{strict\ baseline\ dataset}$: 0.85 (95% CI, 0.79–0.91)]. When the 75th quantile was used as the cut-off point, the AUCs remained above 0.78 (*Supplementary Figure S6*).

Temporal stability of predictive species

For the 11 predictive species, shifts in the relative abundance over time (~8 years) within a single individual

Table 2 Structure and OR of oral microbiome-based prediction model for risk of malignant esophageal lesions in Anyang, China, 2006–2017

Predictive species*	n (%)**		OR (95% CI)***	
	Cases (N=84)	Controls (N=168)	Univariate	Multivariate
Actinomyces odontolyticus				
Low carriage [<90% quantile (0.3677%)]	65 (77.4)	151 (89.9)	Ref.	Ref.
High carriage [≥90% quantile (0.3677%)]	19 (22.6)	17 (10.1)	2.58 (1.25–5.29)	2.16 (0.80–5.88)
Actinomyces viscosus				
Low carriage [<95% quantile (1.5110%)]	72 (85.7)	159 (94.6)	Ref.	Ref.
High carriage [≥95% quantile (1.5110%)]	12 (14.3)	9 (5.4)	3.74 (1.29–10.89)	7.70 (1.75–33.87)
Dialister invisus				
Low carriage [<75% quantile (0.0655%)]	49 (58.3)	126 (75.0)	Ref.	Ref.
High carriage [≥75% quantile (0.0655%)]	35 (41.7)	42 (25.0)	2.43 (1.31–4.51)	2.32 (1.02–5.30)
Fusobacterium mortiferum				
Low carriage [<80% quantile (0.0016%)]	58 (69.0)	134 (79.8)	Ref.	Ref.
High carriage [≥80% quantile (0.0016%)]	26 (31.0)	34 (20.2)	1.94 (1.01–3.71)	4.64 (1.73–12.50)
Fusobacterium nucleatum				
Low carriage [<30% quantile (0.5956%)]	9 (10.7)	51 (30,4)	Ref.	Ref.
High carriage [≥30% quantile (0.5956%)]	75 (89.3)	117 (69.6)	5.64 (2.13–14.88)	3.85 (1.12–13.24)
Lachnoanaerobaculum umeaense				
Low carriage [< 95% quantile (0.4933%)]	67 (79.8)	159 (94.6)	Ref.	Ref.
High carriage [≥95% quantile (0.4933%)]	17 (20.2)	9 (5.4)	6.75 (2.24–20.41)	10.93 (2.24–53.38)
Leptotrichia hofstadii				
Low carriage [<35% quantile (0.0782%)]	11 (13,1)	59 (35.1)	Ref.	Ref.
High carriage [≥35% quantile (0.0782%)]	73 (86.9)	109 (64.9)	3.65 (1.77–7.50)	2.36 (0.89–6.25)
Prevotella baroniae				
Low carriage [<60% quantile (0.0051%)]	37 (44.0)	101 (60.1)	Ref.	Ref.
High carriage [≥60% quantile (0.0051%)]	47 (56.0)	67 (39.9)	2.15 (1.19–3.89)	1.98 (0.86–4.58)
Prevotella melaninogenica				
Low carriage [<30% quantile (0.1074%)]	10 (11.9)	51 (30.4)	Ref.	Ref.
High carriage [≥30% quantile (0.1074%)]	74 (88.1)	117 (69.6)	4.79 (1.94–11.87)	3.26 (1.05–10.19)
Prevotella shahii				
Low carriage [<90% quantile (0.0446%)]	62 (73.8)	151 (89.9)	Ref.	Ref.
High carriage [≥90% quantile (0.0446%)]	22 (26.2)	17 (10.1)	3.54 (1.65–7.63)	2.37 (0.81–6.90)
Rothia dentocariosa				
Low carriage [<70% quantile (0.1806%)]	47 (56.0)	117 (69.6)	Ref.	Ref.
High carriage [≥70% quantile (0.1806%)]	37 (44.0)	51 (30.4)	2.04 (1.11–3.75)	2.66 (1.12–6.31)

OR, odds ratio; 95% CI, 95% confidence interval; Ref., reference category; SDA, severe dysplasia and above. *, The optimal cut-off point for relative abundance of each predictive species was listed in the brackets; **, The analyzed dataset (full averaged dataset) contained a total of 84 SDA cases providing 143 oral specimens and 168 matched controls providing 285 oral specimens. For all specimens produced by each subject, bacterial abundance at the species level was averaged to produce a single value for that subject; ***, OR and 95% CI were derived by univariate conditional logistic regression analysis and multivariate conditional logistic regression analysis including all species listed in the Table as well as family history of esophageal cancer.

were generally fluctuations around an individual mean, which did not exhibit any increasing or decreasing trend (*Supplementary Figure S7*). These species had lower intra-individual CVs within each subject (average of intra-individual CVs=107.9%) than overall CVs across all specimens provided by all included subjects with multiple

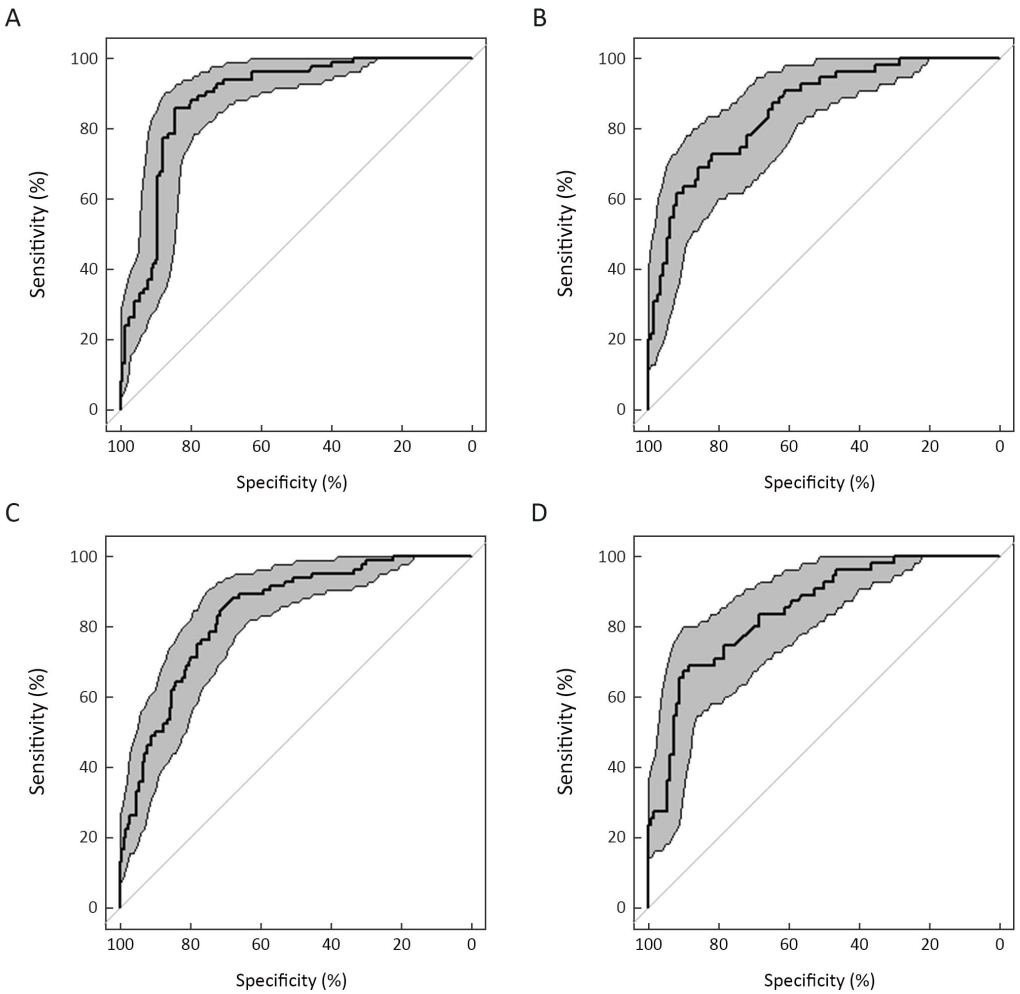
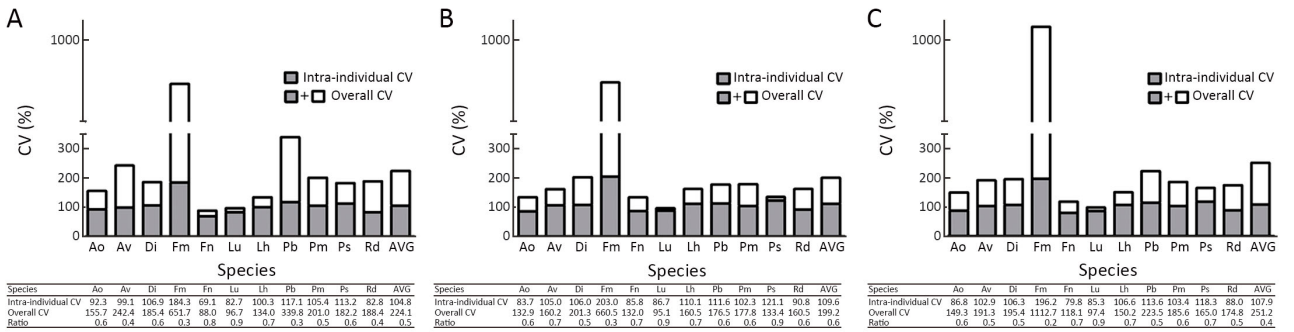


Figure 3 Performance of oral microbiome-based prediction model using the optimal cut-off point for risk of malignant esophageal lesions in Anyang, China, 2006–2017. The predictive model included 11 oral species and family history of esophageal cancer. ROC for this prediction model constructed based on (A) full averaged dataset [AUC: 0.89 (95% CI, 0.84–0.93)]; (B) strictly averaged dataset [AUC: 0.85 (95% CI, 0.80–0.91)]; (C) full baseline dataset [AUC: 0.84 (95% CI, 0.79–0.89)]; (D) strict baseline dataset [AUC: 0.85 (95% CI, 0.79–0.91)]. For A and B datasets, and for all specimens produced by each subject, bacterial abundance at the species level was averaged to produce a single value for that subject. AUC, area under the receiver operating characteristic curve; 95% CI, 95% confidence interval; ROC, receiver operating characteristic; SDA, severe dysplasia and above.

sampling (average of overall CVs=251.2%), resulting in an average ratio of intra-individual CV and overall CV of 0.4 (*Figure 4*; *Supplementary Table S4*), and no appreciable discrepancy was found in cases and controls.

Discussion

One of the key problems in current microbiome-oncology research is the lack of prospective longitudinal studies, and the execution of such studies within the microbiome field is challenging but is urgently needed to provide direct evidence of causation (18). In this first dynamic longitudinal investigation of the causative and predictive role of oral microbiome in malignant esophageal lesions, we show that specific oral species are differentially abundant with respect to disease status, and a panel of 11 bacteria can accurately distinguish SDA cases from healthy controls. It seems likely that the oral microbiome has an etiological role in esophageal cancer, and it holds promise as a non-invasive early warning biomarker for risk stratification for esophageal cancer screening programs. The oral microbiome presents an opportunity to better understand esophageal cancer and how it might be prevented.



Species	Ao	Av	Di	Fm	Fn	Lu	Lh	Pb	Pm	Ps	Rd	AVG
Intra-individual CV	92.3	99.1	106.9	184.3	69.1	82.7	100.3	117.1	105.4	113.2	82.8	104.8
Overall CV	155.7	242.4	185.4	651.7	88.0	96.7	134.0	339.8	201.0	182.2	188.4	224.1
Ratio	0.6	0.4	0.6	0.3	0.8	0.9	0.7	0.3	0.5	0.6	0.4	0.5

Species	Ao	Av	Di	Fm	Fn	Lu	Lh	Pb	Pm	Ps	Rd	AVG
Intra-individual CV	83.7	105.0	106.0	203.0	85.8	86.7	110.1	111.6	102.3	121.1	90.8	109.6
Overall CV	132.9	160.2	201.3	660.5	132.0	95.1	160.5	176.5	177.8	133.4	160.5	199.2
Ratio	0.6	0.7	0.5	0.3	0.7	0.9	0.7	0.6	0.6	0.9	0.6	0.6

Species	Ao	Av	Di	Fm	Fn	Lu	Lh	Pb	Pm	Ps	Rd	AVG
Intra-individual CV	86.8	102.9	106.3	196.2	79.8	85.3	106.6	113.6	103.4	118.3	88.0	107.9
Overall CV	149.3	191.3	195.4	1112.7	118.1	97.4	150.2	223.5	185.6	165.0	174.8	251.2
Ratio	0.6	0.5	0.5	0.2	0.7	0.9	0.7	0.5	0.6	0.7	0.5	0.4

Figure 4 Intra-individual CVs and overall CVs for predictive oral species in subjects with multiple sampling from Anyang, China, 2006–2017. A total of 28 subjects (C) including (A) 10 SDA cases and (B) 18 controls (each with three or more serial specimens) were included in this analysis. For each predictive species, CVs for each individual were calculated based on the relative abundance of all specimens from this individual. Mean CV (intra-individual CV) was the mean of all CVs calculated from all included individuals. Overall CV was calculated based on the relative abundances from all specimens provided by all included individuals. AVG, average; CV, coefficient of variation; SDA, severe dysplasia and above; Ao, *Actinomyces odontolyticus*; Av, *Actinomyces viscosus*; Di, *Dialister invisus*; Fm, *Fusobacterium mortiferum*; Fn, *Fusobacterium nucleatum*; Lu, *Lachnoanaerobaculum umeaense*; Lh, *Leptotrichia hofstadii*; Pb, *Prevotella baroniae*; Pm, *Prevotella melaninogenica*;Ps, *Prevotella shahii*; Rd, *Rothia dentocariosa*.

Cross-sectional studies and case-control studies have reported distinct differences in upper digestive tract microbiome between gastroesophageal reflux disease (19-21), Barrett's esophagus (19-22), esophageal adenocarcinoma (EAC) (19,23), esophageal squamous dysplasia (24), or ESCC (6,7,9) cases and controls. Additionally, poor oral health, including poor periodontal health, tooth loss, and irregular teeth brushing, has repeatedly reported to be linked with the risk of malignant esophageal lesions (2,6,7,25,26), supporting the hypothesis that oral health-related microbial environment (e.g. oral dysbiosis) may play a role in the carcinogenesis of esophageal epithelium. However, only one study to date has prospectively examined whether upper digestive tract microbiome influences risk for subsequent esophageal cancer. In a nested case-control study conducted in USA, Peters *et al.* evaluated oral bacteria using 16S rRNA gene sequencing in prediagnostic mouthwash specimens from n=81/160 EAC and n=25/50 ESCC cases/matched controls (7). They found that several specific species were associated with cancer risk (For EAC, *Tannerella forsythia* and *Streptococcus pneumoniae* with P<0.05; For ESCC, *Prophyromonas gingivalis* with a P value of 0.09). In our study, at the species level, we found that dozens of oral bacteria were associated with malignant esophageal lesions. Using a larger sample size and a more appropriate statistical method for abundance comparison (Deseq *vs.* Conditional logistic regression) may partially explain the larger number of cancer related-species we found. Our results are in keeping with the current concept that mixed communities of pathogens collectively drive disease progression, rather than individual species working in isolation (13,27). The molecular mechanisms by which the microbiome may be involved in the aetiopathogenesis of cancer have been extensively discussed. All the proposed mechanisms, including genomic integration, genotoxicity, inflammation, immunity and metabolism, seem to ultimately converge on final common pathways of enhanced capacity of replication and dedifferentiation, and prolonged host cell survival (18). Further study about the oncogenic mechanisms by which the oral microbiome, alone or alongside with environmental and host factors, may initiate and/or drive the carcinogenesis of esophageal cancer is warranted.

All 11 SDA lesion-associated oral species included in this prediction model were anaerobic bacteria. Four of these (*Actinomyces odontolyticus*, *Actinomyces viscosus*, *Lachnoanaerobaculum umeaense*, and *Rothia dentocariosa*) were Gram-positive, and all the others were Gram-negative. For the most part, these bacteria live in harmony with the host, generally in a commensal state. However, under certain circumstances, this commensal relationship may break down, and these bacteria may be involved in human disease. While most of these 11 bacteria have been reported to have associations with dental cavities and periodontal diseases, some have a linkage with autoimmune diseases (e.g., *Lachnoanaerobaculum umeaense* induces animal models of celiac disease), and some are correlated with cancer (e.g., Prevotella melaninogenica shows increased abundance in oral cancer patients) (28,29).

Fusobacterium nucleatum is a well-known periodontal pathogen identified as one bacterium among these 11

predictive bacteria (17). This bacterium has frequently been found to be enriched in colorectal cancer tissues, and it has been suggested that it influences colorectal carcinogenesis through activation of cellular proliferation pathways, and by suppression of the antitumor immune response (30). Given the proximity of the esophagus to the oral cavity, *Fusobacterium nucleatum* may also play a role in esophageal cancer. Using real-time polymerase chain reaction (PCR), Yamamura *et al.* reported that 23% of ESCC tumor tissues contain *Fusobacterium nucleatum* DNA, which is greater than that in normal adjacent esophageal tissue (P=0.021). Moreover the presence of *Fusobacterium nucleatum* is associated with significantly shorter survival time in patients with ESCC (17). *Prophyromonas gingivalis* has also been found to be associated with ESCC (6,31) but we did not observe a significant difference in its relative abundance in cases of malignant esophageal lesions and their matched controls in our study population. One possible explanation is that these studies assessed for the presence of *Prophyromonas gingivalis* rather than its relative abundance. A recent report from a study also conducted in Henan, China supported our findings (32). This study showed that tumor tissues had a greater abundance of *Fusobacterium* than paired non-tumor tissues (67 pairs), but no significant difference in the abundance of *Porphyromonas* was observed. Altogether, certain oral bacterial species, including *Fusobacterium nucleatum*, might contribute to and predict the carcinogenesis of malignant esophageal lesions. Identification and manipulation of carcinogenic oral bacteria may offer actionable strategies for prevention of this highly fatal disease.

Sensitivity analysis of the performance of this predictive model showed that after excluding cases diagnosed within 1 year of collection of the baseline specimen and their matched controls, the AUC remained high. More importantly, this model performed well for both endoscopically screened and clinically diagnosed case-control triads. Identification of early esophageal lesions is a primary concern of endoscopic screening, as early lesions are of greater clinical and public-health importance than clinically diagnosed SDA cases. Most of the latter are advanced lesions which are less likely to benefit from treatment (33). When analysis was limited to the baseline data of the AECCS cohort, the final prediction model stably yielded good discriminatory results. Additionally, we found that the overall CV across all specimens from all individuals with multiple sampling was about 2.3 times higher than the intra-individual CV (251.2% *vs.* 107.9%), indicating the time-stability of these species. Consistent with our findings, previous studies have also suggested that the abundance of core members of the oral microbiome is fairly stable over time, although more precise microbiome estimates can be obtained by measurement at multiple time points (34,35). These findings indicate our model may also be generalized in settings without intensive sampling, where single-time only specimen collection is employed. Altogether, the oral microbiome holds promise as a non-invasive early warning biomarker for risk stratification for esophageal cancer screening programs.

Due to the high incidence of ESCC in China, endoscopic surveillance of esophageal cancer has come to be viewed as an important national undertaking. Since 2006, more than 1 million endoscopies sponsored by the Chinese government have been carried out in several regions of high ESCC incidence (36,37). The cost of endoscopic examination is high, and endoscopy is an invasive procedure. Therefore, identification of high-risk subjects in the general population is a strategy which is cost-effective. We previously established an easy-to-use risk prediction model for ESCC using demographic and lifestyle factors (37). Use of the model in screening could have allowed 27% of subjects 60 years or younger and 9% of subjects older than 60 years to avoid endoscopy without missing SDAs, which means that approximately 16.6% of endoscopies in total could have been avoided. Oral microbiome markers could be combined in the future with demographic/lifestyle factors to construct a more comprehensive and accurate prediction model for malignant esophageal lesions. Incorporation of model-based risk assessment into large-scale screening programs for esophageal cancer with endoscopic examination of only high-risk individuals identified by the model may render screening programs safer and more cost-effective.

Although this is the first population-based nested case-control study with multiple sampling, its limitations should be noted. First, due to the prospective matched case-control design, the added predictive value of matching factors such as age could not be evaluated. Second, despite access to a repository of >40,000 oral specimens from the AECCS cohort, the absolute number of case-control triads included in this study with multiple specimens was still relatively small due to the low incidence of SDA lesions which rendered some temporal analysis inaccurate. Third, an independent external cohort is needed to validate the results of this study.

Conclusions

This prospective study shows that specific members of the

oral microbiome are associated with the subsequent risk of malignant esophageal lesions, and a model based upon a panel of 11 lesion-associated oral species achieves excellent classification performance. This lends support to the hypothesis that the oral microbiome may play a causative and predictive role in the aetiopathogenesis of esophageal cancer, and raises the possibility that the non-invasive microbiome biomarkers, alone or in combination with other factors, may enable risk-stratification of esophageal cancer screening programs in the future. Our findings have implications for a personalized approach to primary and secondary prevention of esophageal cancer. Further studies are needed to validate our findings and to elucidate mechanisms of the causal relationship.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 30930102, 82073626, 81502855, 81773501), the National Key R&D program of China (No. 2016YFC0901404), the National Special Programme of Scientific and Technological Resources Investigation (No. 2019FY101102), the Digestive Medical Coordinated Development Center of Beijing Hospitals Authority (No. XXZ0204), the Beijing Natural Science Foundation (No. 7182033), the Beijing Municipal Administration of Hospital's Youth Programme (No. QML20171101), and the Science Foundation of Peking University Cancer Hospital (No. 2020-7).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Abnet CC, Arnold M, Wei WQ. Epidemiology of esophageal squamous cell carcinoma. Gastroenterology 2018;154:360-73.
3. Codipilly DC, Qin Y, Dawsey SM, et al. Screening for esophageal squamous cell carcinoma: recent advances. Gastrointest Endosc 2018;88:413-26.
4. Flemer B, Warren RD, Barrett MP, et al. The oral microbiota in colorectal cancer is distinctive and predictive. Gut 2018;67:1454-63.
5. Gilbert JA, Blaser MJ, Caporaso JG, et al. Current understanding of the human microbiome. Nat Med 2018;24:392-400.
6. Yuan X, Liu Y, Kong J, et al. Different frequencies of *Porphyromonas gingivalis* infection in cancers of the upper digestive tract. Cancer Lett 2017;404:1-7.
7. Peters BA, Wu J, Pei Z, et al. Oral microbiome composition reflects prospective risk for esophageal cancers. Cancer Res 2017;77:6777-87.
8. Ajayi TA, Cantrell S, Spann A, et al. Barrett's esophagus and esophageal cancer: Links to microbes and the microbiome. PLoS Pathog 2018;14:e1007384.
9. Chen X, Winckler B, Lu M, et al. Oral microbiota and risk for esophageal squamous cell carcinoma in a high-risk area of China. PLoS One 2015;10:e0143603.
10. Liu F, Guo F, Zhou Y, et al. The Anyang esophageal cancer cohort study: study design, implementation of fieldwork, and use of computer-aided survey system. PLoS One 2012;7:e31602.
11. Zhang C, Liu F, Pan Y, et al. Incidence and clearance of oral human papillomavirus infection: A population-based cohort study in rural China. Oncotarget 2017; 8:59831-44.
12. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials 1996;17:343-6.
13. Hayes RB, Ahn J, Fan X, et al. Association of oral microbiome with risk for incident head and neck squamous cell cancer. JAMA Oncol 2018;4:358-65.
14. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
15. Winer RL, Kiviat NB, Hughes JP, et al. Development and duration of human papillomavirus lesions, after initial infection. J Infect Dis 2005;191:731-8.
16. Griffen AL, Beall CJ, Campbell JH, et al. Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. ISME J 2012; 6:1176-85.
17. Yamamura K, Baba Y, Nakagawa S, et al. Human microbiome *Fusobacterium nucleatum* in esophageal cancer tissue is associated with prognosis. Clin Cancer Res 2016;22:5574-81.
18. Scott AJ, Alexander JL, Merrifield CA, et al. International Cancer Microbiome Consortium consensus statement on the role of the human microbiome in carcinogenesis. Gut 2019;68:1624-32.
19. Blackett KL, Siddhi SS, Cleary S, et al. Oesophageal bacterial biofilm changes in gastro-oesophageal reflux

disease, Barrett's and oesophageal carcinoma: association or causality? Aliment Pharmacol Ther 2013;37:1084-92.
20. Yang L, Lu X, Nossa CW, et al. Inflammation and intestinal metaplasia of the distal esophagus are associated with alterations in the microbiome. Gastroenterology 2009;137:588-97.
21. Liu N, Ando T, Ishiguro K, et al. Characterization of bacterial biota in the distal esophagus of Japanese patients with reflux esophagitis and Barrett's esophagus. BMC Infect Dis 2013;13:130.
22. Macfarlane S, Furrie E, Macfarlane GT, et al. Microbial colonization of the upper gastrointestinal tract in patients with Barrett's esophagus. Clin Infect Dis 2007;45:29-38.
23. Zaidi AH, Kelly LA, Kreft RE, et al. Associations of microbiota and toll-like receptor signaling pathway in esophageal adenocarcinoma. BMC Cancer 2016; 16:52.
24. Yu G, Gail MH, Shi J, et al. Association between upper digestive tract microbiota and cancer-predisposing states in the esophagus and stomach. Cancer Epidemiol Biomarkers Prev 2014;23:735-41.
25. Di Pilato V, Freschi G, Ringressi MN, et al. The esophageal microbiota in health and disease. Ann N Y Acad Sci 2016;1381:21-33.
26. Nwizu NN, Marshall JR, Moysich K, et al. Periodontal disease and incident cancer risk among postmenopausal women: results from the women's health initiative observational cohort. Cancer Epidemiol Biomarkers Prev 2017;26:1255-65.
27. Jakubovics NS, Yassin SA, Rickard AH. Community interactions of oral streptococci. Adv Appl Microbiol 2014;87:43-110.
28. Mager DL, Haffajee AD, Devlin PM, et al. 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;3:27.
29. Lerner A, Aminov R, Matthias T. Dysbiosis may trigger autoimmune diseases via inappropriate post-translational modification of host proteins. Front Microbiol 2016;7:84.
30. Sato Y, Yamagishi J, Yamashita R, et al. Inter-individual differences in the oral bacteriome are greater than intra-day fluctuations in individuals. PLoS One 2015;10:e0131607.
31. Gao S, Li S, Ma Z, et al. Presence of *Porphyromonas gingivalis* in esophagus and its association with the clinicopathological characteristics and survival in patients with esophageal cancer. Infect Agent Cancer 2016;11:3.
32. Shao D, Vogtmann E, Liu A, et al. Microbial characterization of esophageal squamous cell carcinoma and gastric cardia adenocarcinoma from a high-risk region of China. Cancer 2019;125:3993-4002.
33. Pennathur A, Gibson MK, Jobe BA, et al. Oesophageal carcinoma. Lancet 2013;381:400-12.
34. Costello EK, Lauber CL, Hamady M, et al. Bacterial community variation in human body habitats across space and time. Science 2009;326:1694-7.
35. Wang J, Jia Z, Zhang B, et al. Tracing the accumulation of *in vivo* human oral microbiota elucidates microbial community dynamics at the gateway to the GI tract. Gut 2020;69:1355-6.
36. Zhao P, Dai M, Chen W, et al. Cancer trends in China. Jpn J Clin Oncol 2010;40:281-5.
37. Liu M, Liu Z, Cai H, et al. A model to identify individuals at high risk for esophageal squamous cell carcinoma and precancerous lesions in regions of high prevalence in China. Clin Gastroenterol Hepatol 2017;15:1538-46.e7.

Cite this article as: Liu F, Liu M, Liu Y, Guo C, Zhou Y, Li F, Xu R, Liu Z, Deng Q, Li X, Zhang C, Pan Y, Ning T, Dong X, Hu Z, Bao H, Cai H, Dos Santos Silva I, He Z, Ke Y. Oral microbiome and risk of malignant esophageal lesions in a high-risk area of China: A nested case-control study. Chin J Cancer Res 2020;32(6):742-754. doi:10.21147/j.issn.1000-9604.2020.06.07

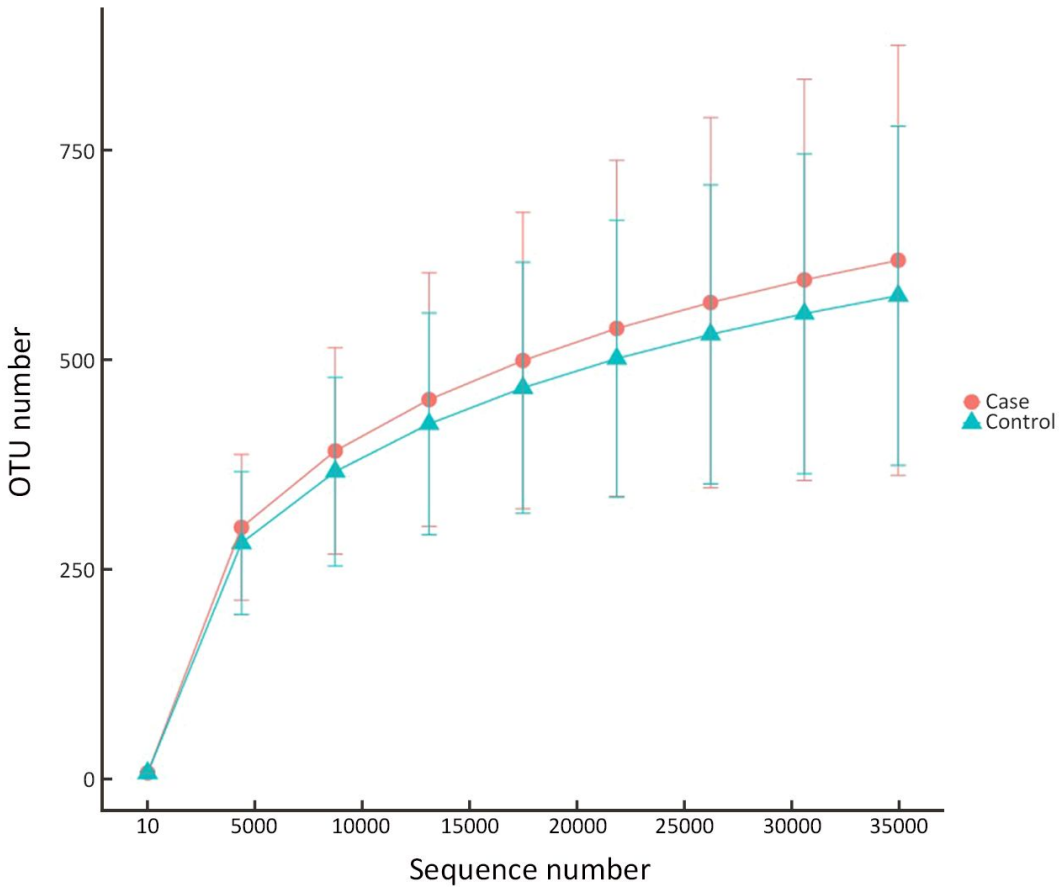


Figure S1 Rarefaction curves of OTU number. The analysis was based on 84 SDA cases (providing 143 oral specimens) and 168 matched controls (providing 285 oral specimens). OTU, operational taxonomic unit; SDA, severe dysplasia and above.

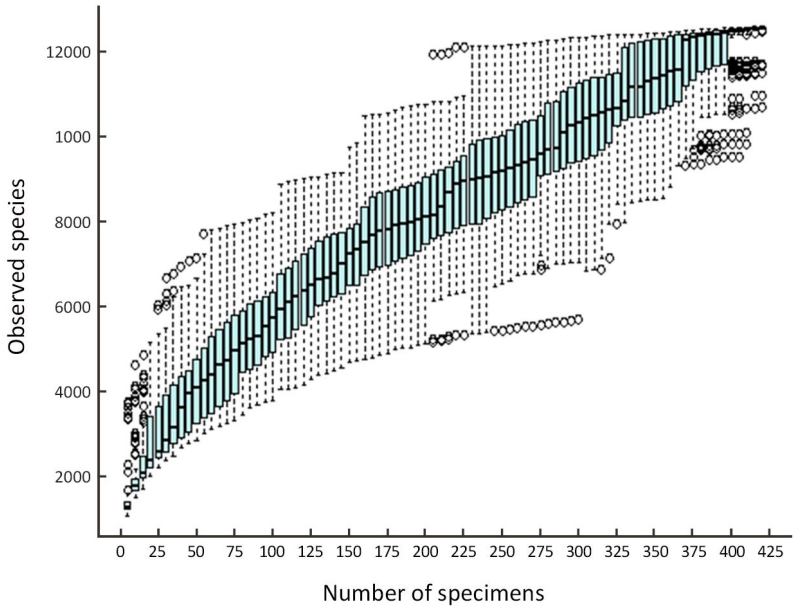


Figure S2 Species accumulation boxplot of observed species. The analysis was based on 428 oral specimens provided from 84 SDA cases and 168 matched controls. SDA, severe dysplasia and above.

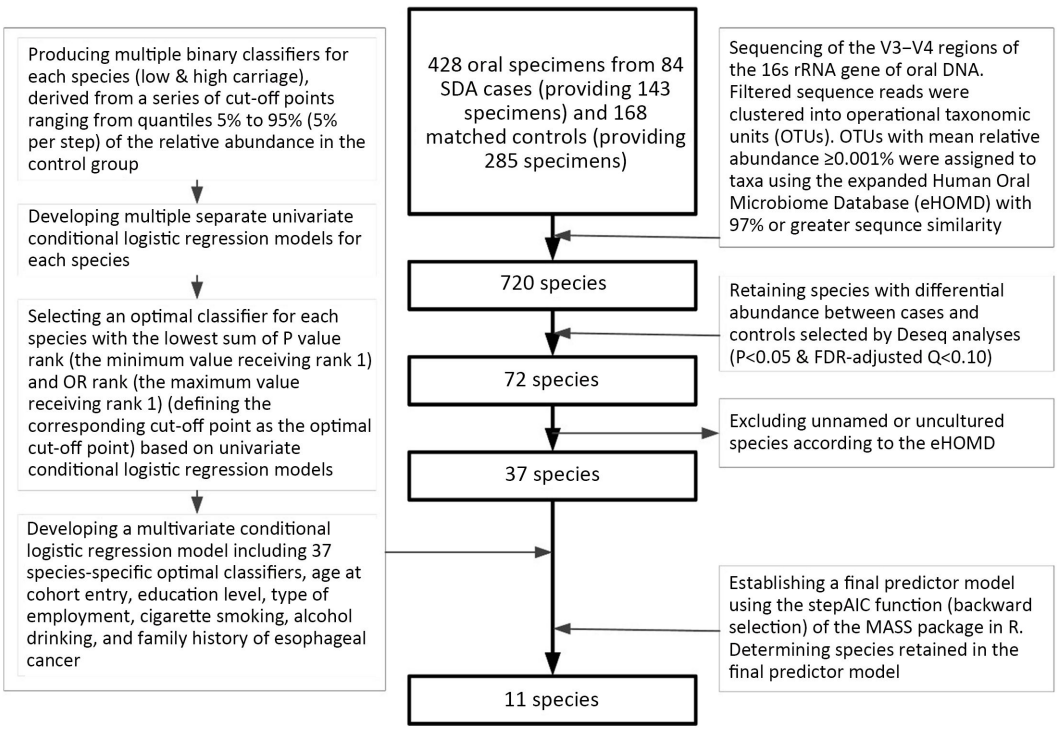


Figure S3 Flowchart for selecting oral species included in the final oral microbiome-based prediction model for risk of malignant esophageal lesions in Anyang, China, 2006–2017. AIC, Akaike information criterion; DNA, deoxyribonucleic acid; eHOMD, expanded Human Oral Microbiome Database; FDR, false discovery rate; OR, odds ratio; OTU, operational taxonomic unit; rRNA, ribosomal RNA; SDA, severe dysplasia and above.

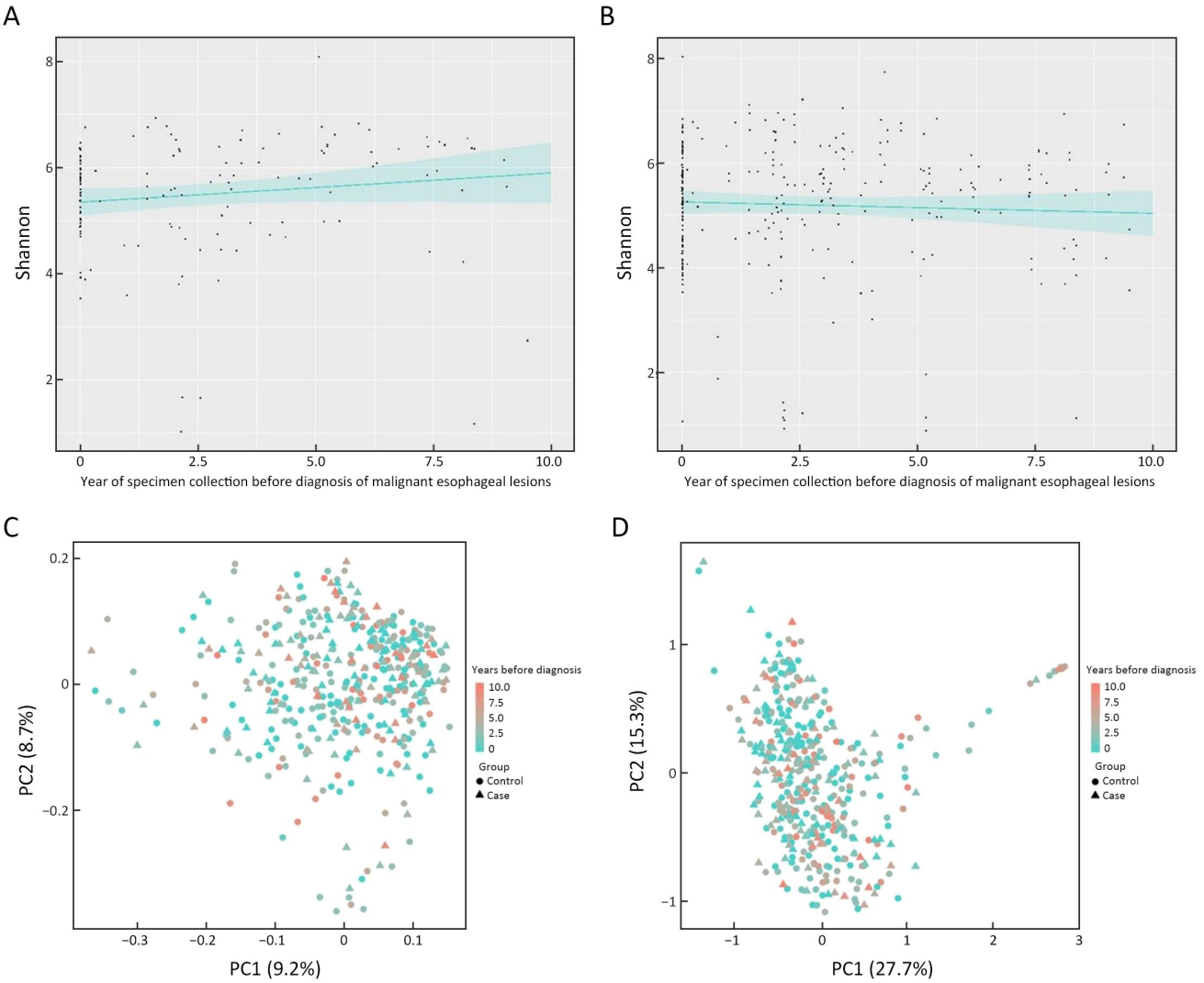


Figure S4 α diversity and β diversity according to years of oral specimen collection before diagnosis of malignant esophageal lesions by case-control status. Trends of α diversity (Shannon index) with years of specimen collection prior to diagnosis of SDA lesions were evaluated using linear mixed-effects (LME) regression (blue lines, shading indicates 95% CI). No significant trend was found for SDA cases (P=0.124) (A) or controls (P=0.425) (B). Between groups, cases showed a higher Shannon diversity index than controls (P=0.044); PERMANOVA models showed that SDA cases differed significantly from controls in overall oral microbiome composition (β diversity) neither when measured by unweighted (P=0.248) (C) nor when measured by weighted UniFrac distances (P=0.590) (D). 95% CI, 95% confidence interval; PCoA, principal coordinate analysis; PERMANOVA, permutational multivariate analysis of variance; SDA, severe dysplasia and above.

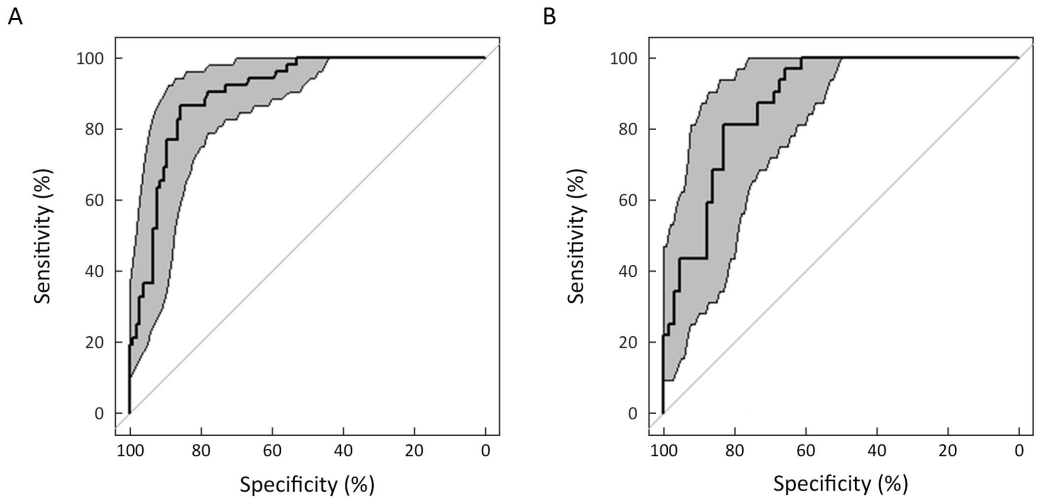


Figure S5 Performance of oral microbiome-based model for prediction risk of malignant esophageal lesions stratified by case type in Anyang, China, 2006–2017. (A) ROC based on 52 screen-endoscopic-detected SDA cases (providing 84 oral specimens) and 104 matched controls (providing 162 oral specimens) [AUC: 0.90 (95% CI, 0.86–0.95)]; (B) ROC based on 32 clinically diagnosed SDA cases (providing 59 oral specimens) and 64 matched controls (providing 123 oral specimens) [AUC: 0.88 (95% CI, 0.81–0.94)]. ROC, receiver operating characteristic; SDA, severe dysplasia and above; AUC, area under the receiver operating characteristic curve; 95% CI, 95% confidence interval; ESCC, esophageal squamous cell carcinoma.

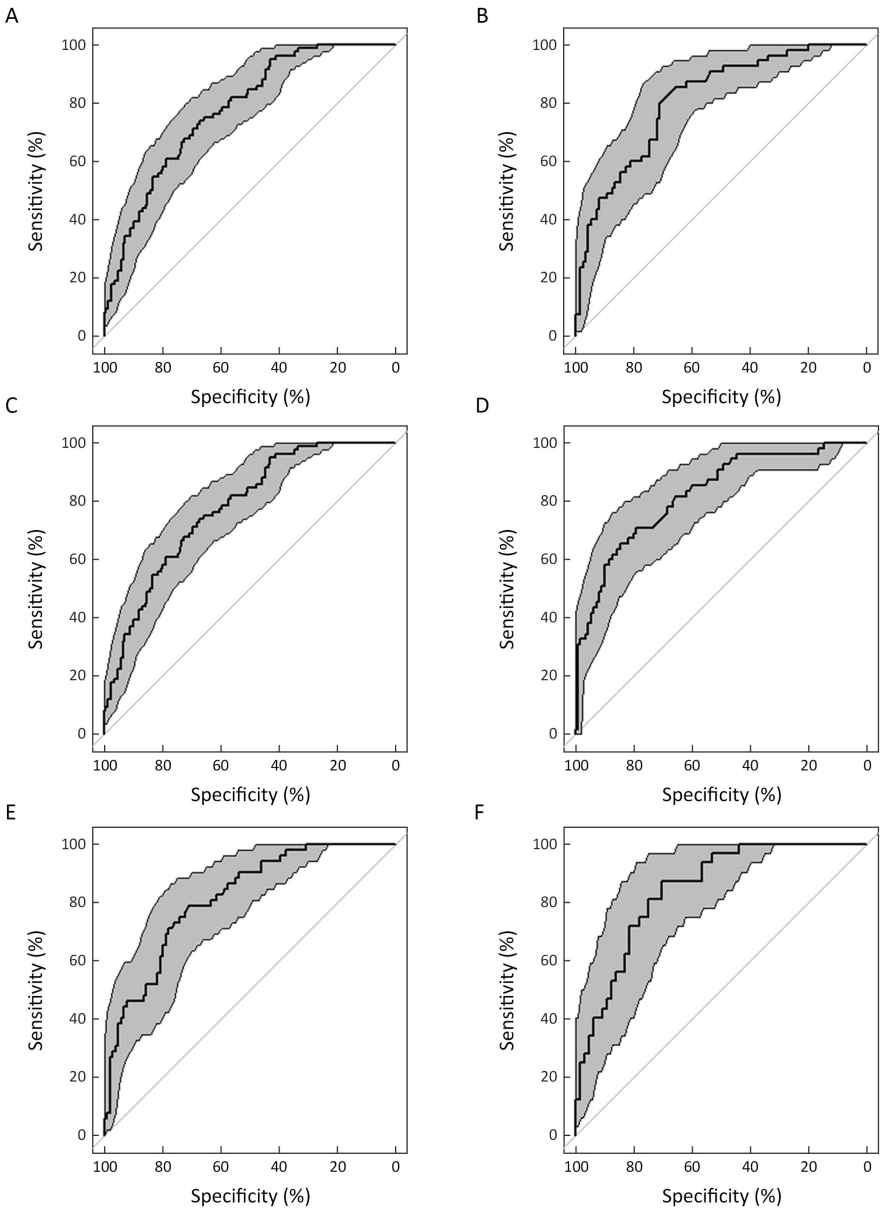


Figure S6 Performance of oral microbiome-based prediction model using the 75th quantile cut-off point for risk of malignant esophageal lesions in Anyang, China, 2006–2017. ROC for the prediction model based on (A) full averaged dataset (84 SDA cases + 168 matched controls) [AUC: 0.78 (95% CI, 0.72–0.83)]; (B) strictly averaged dataset (55 cases + 110 matched controls) [AUC: 0.81 (95% CI, 0.74–0.88)]; (C) full baseline dataset (84 cases + 168 matched controls) [AUC: 0.78 (95% CI, 0.72–0.84)]; (D) strict baseline dataset (55 cases + 110 matched controls) [AUC: 0.82 (95% CI, 0.76–0.89)]; (E) 52 screen-endoscopic-detected SDA cases and 104 matched controls [AUC: 0.81 (95% CI, 0.75–0.88)]; (F) 32 clinically diagnosed SDA cases and 64 matched controls [AUC: 0.84 (95% CI, 0.77–0.92)]. ROC, receiver operating characteristic; SDA, severe dysplasia and above; AUC, area under the receiver operating characteristic curve; 95% CI, 95% confidence interval.

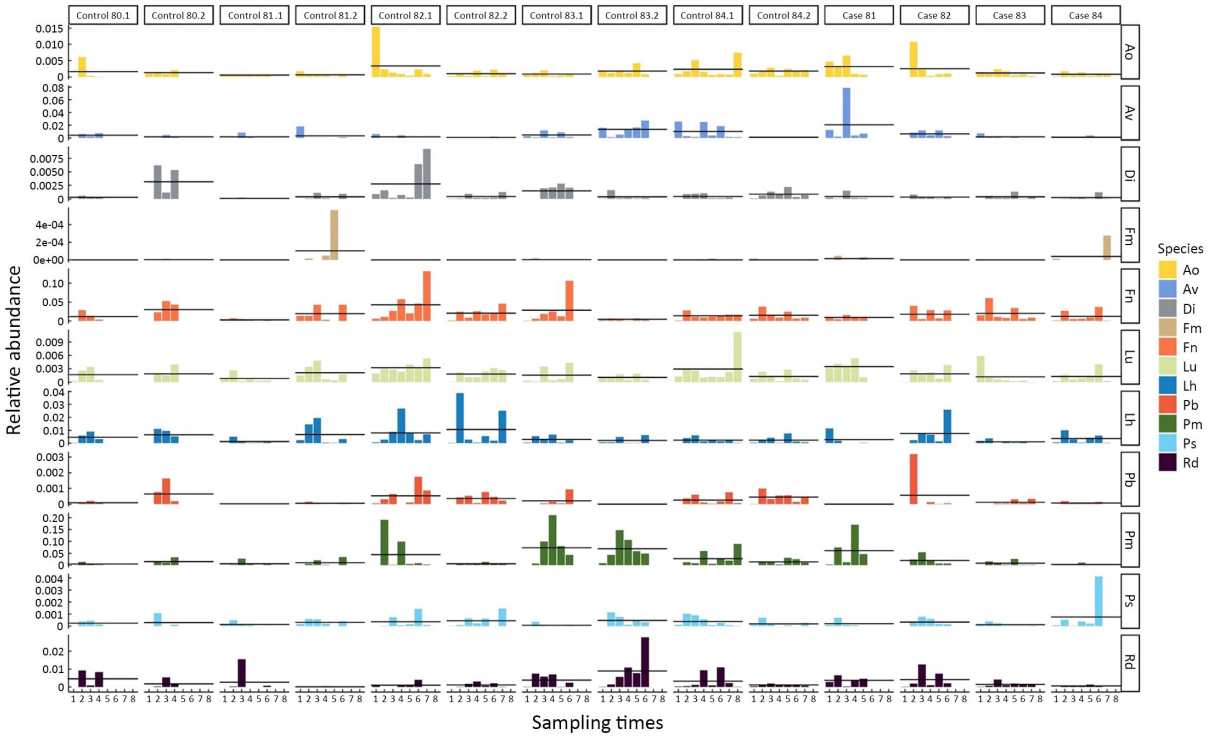


Figure S7 Distribution of relative abundance of predictive species among individuals with four or more specimens in Anyang, China, 2006–2017. A total of 86 specimens provided from 4 SDA cases and 10 controls (with four or more serial specimens) were included in this analysis. For each predictive species, the mean relative abundance, calculated across four or more specimens for each individual, is marked by the dark line. SDA, severe dysplasia and above; Ao, *Actinomyces odontolyticus*; Av, *Actinomyces viscosus*; Di, *Dialister invisus*; Fm, *Fusobacterium mortiferum*; Fn, *Fusobacterium nucleatum*; Lu, *Lachnoanaerobaculum umeaense*; Lh, *Leptotrichia hofstadii*; Pb, *Prevotella baroniae*; Pm, *Prevotella melaninogenica*; Ps, *Prevotella shahii*; Rd, *Rothia dentocariosa*.

Table S1 Number of filtered sequence reads per specimen*

Cross-sections	Total (N=428 specimens)		Cases (N=143 specimens)**		Controls (N=285 specimens)**	
	n	Reads ($\bar{x}\pm s$)	n	Reads ($\bar{x}\pm s$)	n	Reads ($\bar{x}\pm s$)
1	243	76,104±10,786	81	76,285±10,730	162	76,013±10,846
2	109	78,327±9,729	38	78,198±8,970	71	78,396±10,174
3	20	77,439±14,943	7	84,419±21,085	13	73,680±9,356
4	12	79,233±7,502	4	83,016±5,076	8	77,342±8,071
5	12	77,158±8,048	4	82,153±3,490	8	74,660±8,670
6	10	77,583±9,607	2	71,930±11,659	8	78,996±9,372
7	7	74,729±5,573	2	76,054±5,862	5	74,200±6,064
8	15	77,513±8,925	5	73,569±9,946	10	79,486±8,184
Total	428	76,911±10,444	143	77,385±10,733	285	76,673±10,306

*, Quality filtering on the raw reads was performed under specific filtering conditions to obtain the high-quality clean reads according to the Cutadapt (Version 1.9.1, http://cutadapt.readthedocs.io/en/stable/) quality control process. The UCHIME algorithm (UCHIME Algorithm, http://www.drive5.com/usearch/manual/uchime_algo.html) was used to detect chimera sequences, and then the chimera sequences were removed. Finally, clean reads were obtained. **, Cases were subjects with esophageal lesions of severe dysplasia and above including severe squamous dysplasia, carcinoma *in situ*, and esophageal squamous cell carcinoma. Controls were subjects without SDA lesions.

Table S2 Coefficients of variation for Shannon diversity index and number of species observed among quality control samples

Indices*	Intra-plate CV (%)			Inter-plate CV (%)
	1st quantile	Median	3rd quantile	
Shannon	0.80	1.82	1.94	3.70
Observed-species	2.21	2.94	5.00	6.57

CV, coefficients of variation; DNA, deoxyribonucleic acid; *, Indices were calculated with QIIME (Version1.7.0).

Table S3 Species with differential abundance in cases of malignant esophageal lesions and controls based on DESeq analysis in Anyang, China, 2006–2017

Species*	baseMean**	$\log_2$ (fold change)±SE	P***	Q****
Actinomyces naeslundii	278.57	0.75±0.21	<0.001	0.006
Actinomyces odontolyticus	149.37	0.40±0.17	0.022	0.100
Actinomyces radicidentis	31.02	0.58±0.21	0.007	0.049
Actinomyces viscosus	336.96	0.59±0.19	0.002	0.022
Alloprevotella rava	20.51	0.57±0.21	0.006	0.048
Alloprevotella tannerae	32.72	0.50±0.20	0.011	0.061
Bosea robiniae	111.12	−0.66±0.23	0.004	0.037
Campylobacter concisus	101.78	0.70±0.19	<0.001	0.005
Caulobacter vibrioides	697.47	−0.55±0.24	0.022	0.100
Corynebacterium durum	42.36	0.57±0.21	0.007	0.049
Corynebacterium matruchotii	303.41	0.91±0.22	<0.001	0.001
Dialister invisus	73.40	0.79±0.22	<0.001	0.006
Fusobacterium mortiferum	24.39	0.64±0.24	0.009	0.057
Fusobacterium nucleatum	1,358.57	0.53±0.16	0.001	0.015
Fusobacterium simiae	278.13	0.91±0.26	0.001	0.009
Lachnoanaerobaculum umeaense	151.14	0.54±0.17	0.001	0.015
Leptotrichia buccalis	693.34	0.82±0.23	<0.001	0.005
Leptotrichia hofstadii	297.77	0.93±0.22	<0.001	0.001
Leptotrichia hongkongensis	381.05	0.71±0.22	0.001	0.013
Leptotrichia shahii	50.77	0.53±0.22	0.015	0.082
Neisseria flavescens|subflava	722.49	0.49±0.21	0.020	0.099
Prevotella baroniae	13.11	0.74±0.22	0.001	0.008
Prevotella intermedia	227.10	0.54±0.23	0.022	0.100
Prevotella maculosa	27.34	0.45±0.19	0.020	0.099
Prevotella melaninogenica	699.20	0.49±0.21	0.021	0.099
Prevotella micans	12.19	0.87±0.22	<0.001	0.003
Prevotella oulorum	58.03	0.50±0.22	0.023	0.100
Prevotella pallens	133.88	0.54±0.22	0.014	0.076
Prevotella saccharolytica	24.58	0.78±0.21	<0.001	0.005
Prevotella shahii	16.42	0.77±0.19	<0.001	0.001
Pseudomonas toyotomiensis	34.37	−0.59±0.23	0.010	0.061
Pseudopropionibacterium propionicum	136.13	0.59±0.23	0.011	0.061
Rothia dentocariosa	164.18	0.59±0.21	0.005	0.040
Selenomonas dianae	21.72	0.60±0.21	0.005	0.040
Selenomonas sputigena	62.57	0.46±0.20	0.022	0.100
Treponema socranskii	58.19	0.73±0.22	0.001	0.013
Treponema vincentii	44.34	0.61±0.23	0.008	0.053

FDR, false discovery rate; SDA, severe dysplasia and above; SE, standard error. *, A total of 84 SDA cases providing 143 oral specimens and 168 matched controls providing 285 oral specimens were included in this analysis. For all specimens produced by each subject, bacterial abundance at each taxonomic level were averaged to produce a single value for that subject. All named and cultured species with a FDR-adjusted Q<0.10 are included in the Table; **, “baseMean” is the average of normalized count values, dividing by DESeq size factors, taken over all samples; ***, P values were obtained by using differential gene expression analysis based on the negative binomial distribution (DESeq) in the DESeq2 package, adjusting for education level, type of employment, cigarette smoking, alcohol consumption, and family history of esophageal cancer; ****, FDR-adjusted P value, the FDR adjustment was conducted at the species level.

Table S4 Temporal stability of relative abundance of predictive oral species within an individual with three or more specimens from Anyang China, 2006–2017*

Subjects	Statistics	Species										
		Ao	Av	Di	Fm	Fn	Lu	Lh	Pb	Pm	Ps	Rd
Control 75.1	M	362.2	1,336.7	6.0	0	3,328.9	81.0	54.1	0	4,882.2	21.2	1,343.0
	CV	68.9	107.2	57.7	0	154.1	33.7	129.5	0	66.4	166.4	84.6
Control 75.2	M	164.8	1,256.2	3.6	0	155.2	100.8	336.6	0	1,286.2	12.5	386.2
	CV	89.6	68.3	129.1	0	63.4	81.1	85.3	0	68.5	101.1	82.7
Control 76.2	M	220.6	150.6	50.9	2.6	1,480.4	195.7	705.4	0	132.5	32.2	172.2
	CV	123.5	55.8	74.8	40.8	50.3	96.6	93.8	0	38.9	144.6	32.0
Control 77.1	M	301.9	1,527.1	7.9	0	1,055.6	189.5	185.0	12.5	480.3	12.8	121.1
	CV	84.9	142.3	105.5	0	76.1	53.8	72.8	81.2	128.1	20.7	47.8
Control 77.2	M	38.9	545.5	7.0	0.6	986.4	124.9	1,329.1	2.3	174.1	6.0	300.8
	CV	54.9	116.5	97.0	173.2	52.6	113.6	126.0	173.2	74.9	77.0	77.2
Control 78.1	M	54.2	13.6	6.1	0	469.8	117.8	195.1	2.3	129.3	22.3	20.3
	CV	32.7	13.0	45.2	0	62.2	109.0	134.6	86.6	72.7	106.7	73.4
Control 78.2	M	370.4	226.3	41.9	0	2,453.9	229.1	105.5	3.9	354.9	3.2	1,017.6
	CV	123.1	31.0	144.1	0	90.6	76.6	43.5	88.1	91.4	173.2	48.0
Control 79.2	M	103.6	82.9	1.0	0	237.1	191.5	45.1	1.6	258.0	4.1	6.9
	CV	60.3	159.4	173.2	0	136.3	150.4	170.0	97.3	142.8	139.7	45.5
Control 83.1	M	92.3	447.4	150.0	0.3	2,822.6	153.9	294.0	19.8	7,335.3	7.2	387.5
	CV	68.1	102.6	79.6	244.9	138.6	118.8	87.2	185.3	105.7	192.3	85.9
Control 83.2	M	180.6	1,335.7	40.2	0	449.3	106.9	213.0	0	6,843.3	46.3	896.2
	CV	66.7	69.6	153.7	0	58.2	55.9	125.8	0	73.2	90.4	111.5
Control 81.1	M	56.8	167.4	8.6	0	335.0	79.1	113.6	1.6	547.6	13.6	278.5
	CV	32.1	200.5	111.9	0	78.0	116.6	167.8	127.0	189.5	133.5	223.1
Control 81.2	M	65.8	337.0	39.1	10.5	1,930.3	209.1	662.4	5.0	1,024.6	32.6	26.8
	CV	81.8	213.4	131.9	215.0	97.9	81.6	124.5	122.6	132.6	72.0	70.0
Control 80.1	M	161.4	417.2	29.2	0	1,176.9	165.4	442.4	8.2	488.6	24.6	465.7
	CV	183.7	83.6	86.8	0	108.0	93.6	85.0	109.3	127.9	86.2	103.2
Control 80.2	M	127.8	157.8	318.5	0.3	2,975.8	184.3	637.6	64.2	1,492.4	29.8	179.4
	CV	40.2	132.9	95.9	200.0	78.2	88.5	76.5	114.5	93.6	172.5	137.9
Control 82.1	M	335.0	177.5	276.6	0	4,274.8	319.8	794.1	52.8	4,371.6	35.7	105.4
	CV	159.4	140.3	129.1	0	101.4	34.7	111.2	118.1	169.0	149.1	134.5
Control 82.2	M	97.5	77.4	43.2	0	2,076.1	179.6	1,057.6	35.3	561.0	44.3	119.6
	CV	81.3	98.6	110.1	0	66.8	61.2	143.5	76.7	74.0	120.2	102.1
Control 84.1	M	234.5	1,003.5	44.3	0.2	1,343.0	294.0	219.2	26.0	2,756.1	38.3	326.3
	CV	108.0	109.6	102.6	282.8	56.1	116.8	90.4	108.0	112.3	102.3	133.7
Control 84.2	M	179.0	131.7	89.0	0.2	1,544.2	125.4	228.4	43.6	1,378.8	18.5	120.9
	CV	47.3	45.5	79.3	264.6	75.9	77.5	115.2	74.0	79.5	131.5	42.1
Case 75	M	175.1	64.0	112.8	0	2,263.2	417.4	1,302.4	15.3	390.4	51.6	198.7
	CV	88.4	63.0	104.2	0	58.0	104.2	96.8	46.3	67.2	110.2	108.9
Case 76	M	34.4	180.3	3.9	2,480.6	383.7	23.0	85.7	2.3	354.4	1.4	2,446.7
	CV	114.0	72.9	86.7	173.0	17.3	81.9	73.4	94.3	80.4	173.2	85.0

Table S4 (*continued*)

Table S4 (*continued*)

Subjects	Statistics	Species										
		Ao	Av	Di	Fm	Fn	Lu	Lh	Pb	Pm	Ps	Rd
Case 77	M	852.8	541.3	180.0	0	1,586.6	212.1	114.6	13.7	487.5	16.4	613.3
	CV	136.4	54.0	66.4	0	113.9	90.5	95.6	162.5	81.7	62.4	94.7
Case 78	M	141.3	686.2	74.3	1.1	3,618.8	235.2	810.7	10.8	300.2	81.9	537.8
	CV	15.3	150.8	129.6	173.2	69.5	84.0	67.0	88.6	89.3	115.3	39.5
Case 79	M	440.2	230.2	485.5	0	3,054.4	440.0	866.8	0	2,318.6	24.2	93.2
	CV	103.7	149.6	108.0	0	18.0	69.5	46.9	0	115.7	61.0	86.8
Case 83	M	122.8	176.3	39.6	0	2,034.8	118.4	107.5	11.7	7,76.3	11.4	157.3
	CV	63.7	146.5	116.5	0	99.1	173.6	120.8	124.7	126.3	121.1	83.7
Case 81	M	321.7	2,070.1	45.3	1.6	937.0	344.9	269.9	0	6,127.9	19.7	372.4
	CV	78.3	157.1	136.9	141.6	46.3	45.2	183.8	0	110.1	141.8	61.7
Case 80	M	547.7	69.5	284.1	0	1,790.1	569.6	1,316.7	7.1	864.5	65.0	62.6
	CV	95.4	26.0	70.3	0	69.6	0.4	83.9	86.8	129.5	65.1	83.3
Case 82	M	254.2	647.4	33.1	0	1,817.2	186.1	735.1	56.4	1,887.3	34.7	425.2
	CV	161.3	74.1	86.3	0	87.6	72.1	130.6	227.9	104.3	85.9	114.3
Case 84	M	84.8	121.4	26.7	4.1	1,253.7	122.3	340.0	6.6	237.2	76.4	70.2
	CV	66.6	97.4	163.7	249.6	111.6	106.0	104.4	105.5	149.6	196.2	70.0
Controls	Overall M	171.0	496.2	75.4	0.9	1,721.7	177.8	437.7	19.4	2,255.9	25.3	320.6
	Overall CV	132.9	160.2	201.3	660.5	132.0	95.1	160.5	176.5	177.8	133.4	160.5
	Intra-individual CV	83.7	105.0	106.0	203.0	85.8	86.7	110.1	111.6	102.3	121.1	90.8
Cases	Overall M	259.6	503.1	100.3	174.0	1,783.7	237.6	520.5	14.3	1,469.9	38.2	415.4
	Overall CV	155.7	242.4	185.4	651.7	88.0	96.7	134.0	339.8	201.0	182.2	188.4
	Intra-individual CV	92.3	99.1	106.9	184.3	69.1	82.7	100.3	117.1	105.4	113.2	82.8
Total	Overall M	200.8	498.5	83.7	59.1	1,742.5	197.9	465.5	17.7	1,991.8	29.7	352.4
	Overall CV	149.3	191.3	195.4	1,112.7	118.1	97.4	150.2	223.5	185.6	165.0	174.8
	Intra-individual CV	86.8	102.9	106.3	196.2	79.8	85.3	106.6	113.6	103.4	118.3	88.0

M, mean relative abundance; CV, coefficient of variation; SDA, severe dysplasia and above; Ao, *Actinomyces odontolyticus*; Av, *Actinomyces viscosus*; Di, *Dialister invisus*; Fm, *Fusobacterium mortiferum*; Fn, *Fusobacterium nucleatum*; Lu, *Lachnoanaerobaculum umeaense*; Lh, *Leptotrichia hofstadii*; Pb, *Prevotella baroniae*; Pm, *Prevotella melaninogenica*; Ps, *Prevotella shahii*; Rd, *Rothia dentocariosa*. *, Cases were subjects with esophageal lesions of severe dysplasia and above (SDA) including severe squamous dysplasia, carcinoma *in situ*, and esophageal squamous cell carcinoma. Controls were subjects without SDA lesions. For each predictive species (column), M and CV are listed. CV is presented as a percentage. In rows (i.e. Control 75.1), M and CV are calculated based on the relative abundance of three or more specimens from the respective individual. M and CV rows with the heading "Overall" were calculated from all specimens provided by included subjects (each with three or more serial specimens). The intra-individual CV row is mean of all CVs calculated from all included individuals.