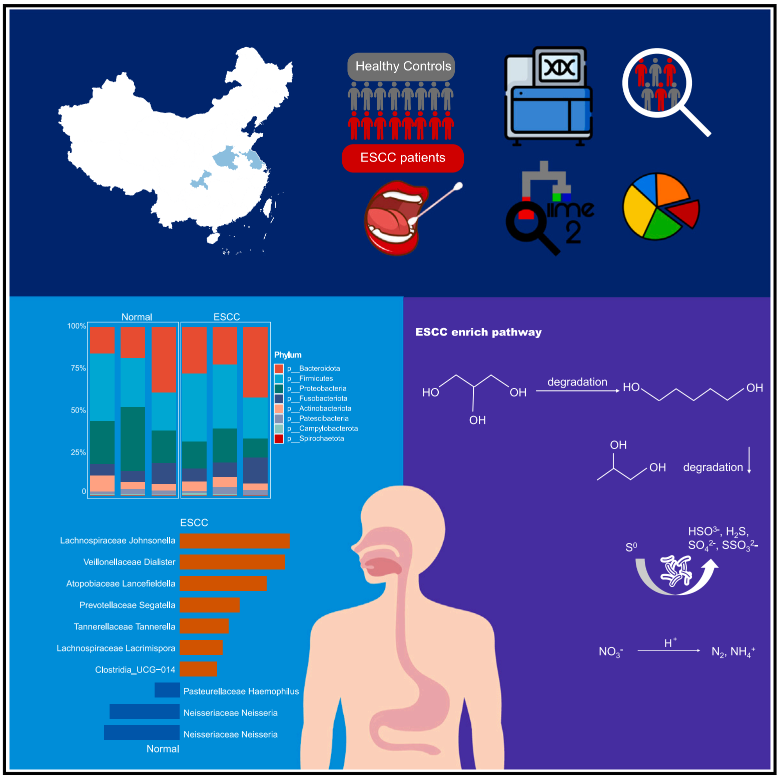
iScience

Article

A cross-cohort study identifies potential oral microbial markers for esophageal squamous cell carcinoma

Graphical abstract



Highlights

- A non-invasive, cost-effective predictive model for screening ESCC has been developed
- *Neisseria perflava* and *Haemophilus parainfluenzae* were markers to identify ESCC
- Sulfur oxidation, and nitrate reduction were enriched in ESCC

Authors

Yanxiang Yu, Lei Xia, Zhouxuan Wang, Tong Zhu, Lujun Zhao, Saijun Fan

Correspondence

zhutong@irm-cams.ac.cn (T.Z.),
zhaolujun@tjmuch.com (L.Z.),
fansaijun@irm-cams.ac.cn (S.F.)

In brief

Microbiome; Oral microbiology; Cancer

Yu et al., 2024, iScience *27*, 111453
December 20, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.111453

CellPress

iScience

CellPress
OPEN ACCESS

Article

A cross-cohort study identifies potential oral microbial markers for esophageal squamous cell carcinoma

Yanxiang Yu,[2,4] Lei Xia,[1,2,3,4] Zhouxuan Wang,[2] Tong Zhu,[2,*] Lujun Zhao,[1,*] and Saijun Fan[2,5,*]
[1]Department of Radiation Oncology, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin's Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin 300060, China
[2]Tianjin Key Laboratory of Radiation Medicine and Molecular Nuclear Medicine, Institute of Radiation Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Tianjin 300192, China
[3]Department of Cancer Center, The Second Affiliated Hospital of Chongqing Medical University, Chongqing 401336, China
[4]These authors contributed equally
[5]Lead contact
*Correspondence: zhutong@irm-cams.ac.cn (T.Z.), zhaolujun@tjmuch.com (L.Z.), fansaijun@irm-cams.ac.cn (S.F.)
https://doi.org/10.1016/j.isci.2024.111453

SUMMARY

Current screening methods for esophageal squamous cell carcinoma (ESCC) face challenges such as low patient compliance and high costs. This study aimed to develop a model based on oral microbiome data for identifying ESCC. By analyzing 249 oral flora samples, we identified microbial markers associated with ESCC and constructed random forest classifiers that distinguished patients with ESCC from controls, achieving an area under the ROC curve (AUC) of 0.87. Key ESCC-associated microbial markers included *Neisseria perflava* and *Haemophilus parainfluenzae*. The classifier was validated within the cohort, attaining an AUC of 0.93. For comparison, traditional tumor markers carcinoembryonic antigen (CEA) and squamous cell carcinoma antigen (SCC-Ag) yielded AUCs of 0.84. Functional analysis identified pathways linked to ESCC, such as glycerol degradation and nitrate reduction. This study suggests a potential noninvasive method for detecting ESCC, offering a more accessible and accurate alternative to current screening methods.

INTRODUCTION

Esophageal cancer is the seventh most prevalent cancer, with 604,100 new cases and an estimated 544,000 deaths in 2020.[1] Esophageal squamous cell carcinoma (ESCC) accounts for approximately 84% of these cases.[2,3] Over the past five decades, the five-year overall survival rate for patients with esophageal cancer in the United States has improved from 3.6% to 21.1%. However, this improvement is observed mainly in patients diagnosed with localized and locoregional disease.[4] Patients with advanced-stage ESCC have a poor prognosis, with five-year survival rates of 18.5% in the United States and 36.9% in China.[5] In contrast, early-stage ESCC can often be cured with endoscopic resection.[6] The most common method for identifying aberrant lesions is endoscopy with iodine staining, but its widespread use is limited by low patient compliance, high costs, and the need for specialized expertise, particularly in high-risk, economically disadvantaged regions.[7]

Oral bacteria can translocate to distal sites and influence distant areas through metabolic processes and immune responses.[8] Differences in the oral microbiota have been observed between patients with esophageal cancer and healthy individuals.[9,10] Profiles based on 16S ribosomal RNA more accurately reflect the authentic microbial consortia.[11]

Endoscopy with iodine staining is the most common strategy for the identification of aberrant lesions in clinical practice. Therefore, we analyzed three cohorts of 249 samples collected from three provinces in China.

To address the microbiome batch effect, we utilized cross-cohort analysis to mitigate the impact of biological and technical variables across different studies, facilitating the identification of significant changes.[11–13] In this study, we constructed a random forest (RF) algorithm using a set of 43 important amplicon sequence variants (ASVs), achieving an area under the curve (AUC) of 0.88. This model effectively distinguished between patients with esophageal cancer and noncancer controls. Further functional enrichment analysis revealed that microbiome alterations in ESCC were marked by the upregulation of nitrate reduction and chondroitin sulfate degradation, and the downregulation of the superpathway converting glycerol to 1,3-propanediol, indicating a specific metabolic profile characteristic of ESCC.

Using oral microorganisms as biomarkers offers a noninvasive and cost-effective alternative to endoscopic screening. This study presents a model for detecting ESCC on the basis of

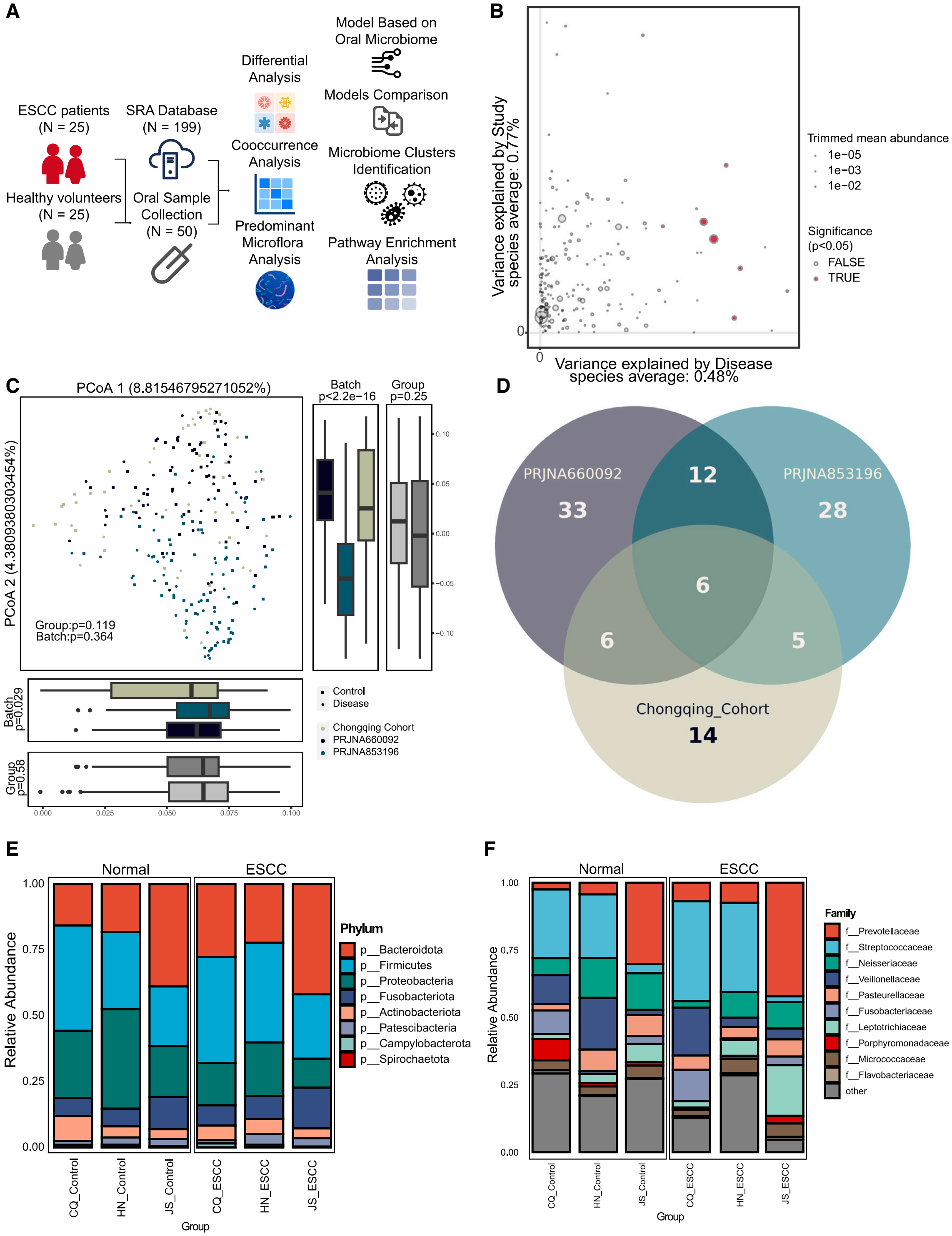


Figure 1. Alterations in the oral microbial composition of the different groups

(A) The flow diagram of this research.

(B) The dispersion of variance due to group distinctions was plotted against the variance resulting from study-specific influences for each individual ASV. The ASVs demonstrating notable differential abundance were depicted in red, with the size of the markers proportional to the abundance of the ASVs. The *p* values obtained from a (ANOVA)-type approach were used to assess the potential confounding variables.

(legend continued on next page)

oral microbiome data, providing a promising approach for diagnosis and improved patient outcomes.

RESULTS

Differences in the oral microbiome

To investigate the differences in the oral microbiome between patients with ESCC and healthy individuals, we conducted 16S rRNA sequencing on 25 patients with ESCC and 25 matched healthy volunteers in Chongqing, China. To identify broader and more common differences, we included two additional datasets, encompassing three distinct cohorts from three provinces in China (Figure 1A; Figure S1). The baseline demographic and clinical characteristics of the participants are detailed in Table S1. Previous studies have shown that batch effects can influence the results of clinical microbiome analyses. Therefore, we initially examined disease-associated and cohort-specific variations in each ASV to assess potential confounding factors (Figure 1B). The "batch" variable accounted for a significant proportion of the variance in ASVs, more so than the illness status itself (Figures 1B and 1C). This batch effect also substantially influenced microbial communities and individual taxa, as evidenced by considerable fluctuations in β-diversity across the studies. To minimize potential biases, we standardized the analysis of 16S rRNA data using a consistent analytical pipeline and confined further investigations to subjects within a single cohort. For integrated analyses, where necessary, we used two-sided blocked Wilcoxon rank-sum tests, with "batch" serving as a covariate, to adjust for batch effects.

Alterations in the oral microbial composition in ESCC

The comparative analysis of the oral microbiome between patients with esophageal cancer and the healthy cohort did not reveal a statistically significant difference ($p = 0.119$, Figure 1C). Similarly, microbial diversity metrics, such as the Shannon index and Simpson's index, did not significantly differ between the two groups (Figure S2). These findings remained consistent even after applying a two-sided blocked Wilcoxon rank-sum test to control for batch effects, with "batch" serving as the blocking variable. Thus, adjusting for batch effects did not alter the significance of the diversity indices, confirming the initial observations.

At the ASV level, differential abundance analysis across three studies revealed that 57 ASVs were significantly distinct in the PRJNA660092 cohort, 51 in the PRJNA853196 cohort, and 31 in the Chongqing cohort, with 6 ASVs consistently differing across all cohorts (Figure 1D). In the oral cavities of patients with esophageal cancer, taxa from the *Neisseria* family and *Rothia* genus were reduced, whereas the *Atopobiaceae* and *Selenomonadaceae* families, along with genera such as *Megasphaera* and *Dialister*, were more prevalent in healthy individuals (Figure S3).

The composition of the oral microbiota included members of the phyla *Actinobacteria*, *Fusobacteria*, *Patescibacteria*, *Campylobacterota*, *Firmicutes*, *Proteobacteria*, and *Spirochaetota* (Figure 1E). At the family level, the microbiota was predominantly composed of *Prevotellaceae* and *Streptococcaceae*, followed by *Neisseriaceae*, *Veillonellaceae*, *Pasteurellaceae*, *Fusobacteriaceae*, *Leptotrichiaceae*, *Porphyromonadaceae*, *Micrococcaceae*, and *Flavobacteriaceae* (Figure 1F).

Identification of bacterial biomarkers at the species level for diagnosing ESCC through cross-cohorts

Cross-cohort analyses are crucial for identifying unique biomarkers that enable consistent and accurate disease diagnosis. As shown in Figure 1D, 74 ASVs, annotated to 62 genera, exhibited statistically significant differences in at least one study (Figure 2). Among these genera, 35 were found to be less abundant in patients with cancer (Table S2). Consistent findings across all three studies indicated that patients with ESCC had a reduced presence of genera such as *Neisseria*, *Megasphaera*, and *Rothia*. In contrast, the families *Atopobiaceae* and *Selenomonadaceae*, along with the genus *Dialister*, were more prevalent (Figure S3). These observed shifts in microbial composition align with previous research on oral microbiome changes in smokers.[14],[15]

Co-occurrence and clustering analysis of the microbiota

The co-occurrence network for differential ASVs was constructed using the FastSpar method, which efficiently infers robust correlation networks within microbial communities.[16] Despite occasional negative correlations, a predominant pattern of positive correlations was observed among the differentially abundant ASVs, indicating a prevalent symbiotic association within the microbial network. Using the MCODE plugin, three modules were identified from this network, as shown in Figure 3B. Module one consisted of 9 nodes and 37 edges, with the biomarkers *Leptotrichia wadei* and *Segatella salivae* being notably more abundant and positively correlated with other biomarkers, including *Clostridia_UCG-014*, *Absconditabacteriales_(SR1)*, *Alloprevotella rava*, *Megasphaera micronuciformis*, *Veillonella atypica*, *Lancefieldella parvula*, and *Lachnoanaerobaculum gingivalis*. The second module, containing 4 nodes and 6 edges, was marked by the presence of

(C) Principal coordinate analysis (PCoA) was applied to samples from four studies (control, $n = 120$; disease, $n = 129$), utilizing Bray-Curtis distances. The analysis revealed no significant variation in the overall composition of the oral microbiota across batches ($p = 0.364$) or between groups ($p = 0.119$). However, differences were observed within the PCoA 1 and PCoA 2 axes. Beta diversity p values, calculated via PERMANOVA based on Bray-Curtis distances, supported these findings. Studies were differentiated by color coding, and groups were denoted by distinct shapes. Boxplots positioned in the upper-right and bottom-left quadrants display the distribution of samples along the first two principal coordinates, categorized by study and group, respectively. The p values for the first and second principal components were ascertained using a two-sided Kruskal-Wallis test, considering both study and group variables. The boxplots depict the interquartile range (25th–75th percentile), with the median represented by a central bold line. Whiskers extended to values within 1.5 times the interquartile range (IQR), and outliers are indicated as individual points.

(D) Venn diagram delineated the intersection of distinct ASVs between healthy controls and patients with ESCC.

(E) The relative distribution of bacterial phyla in healthy individuals and those with ESCC was assessed across three studies.

(F) The relative distribution of bacterial families in healthy individuals and those with ESCC was assessed across three studies.

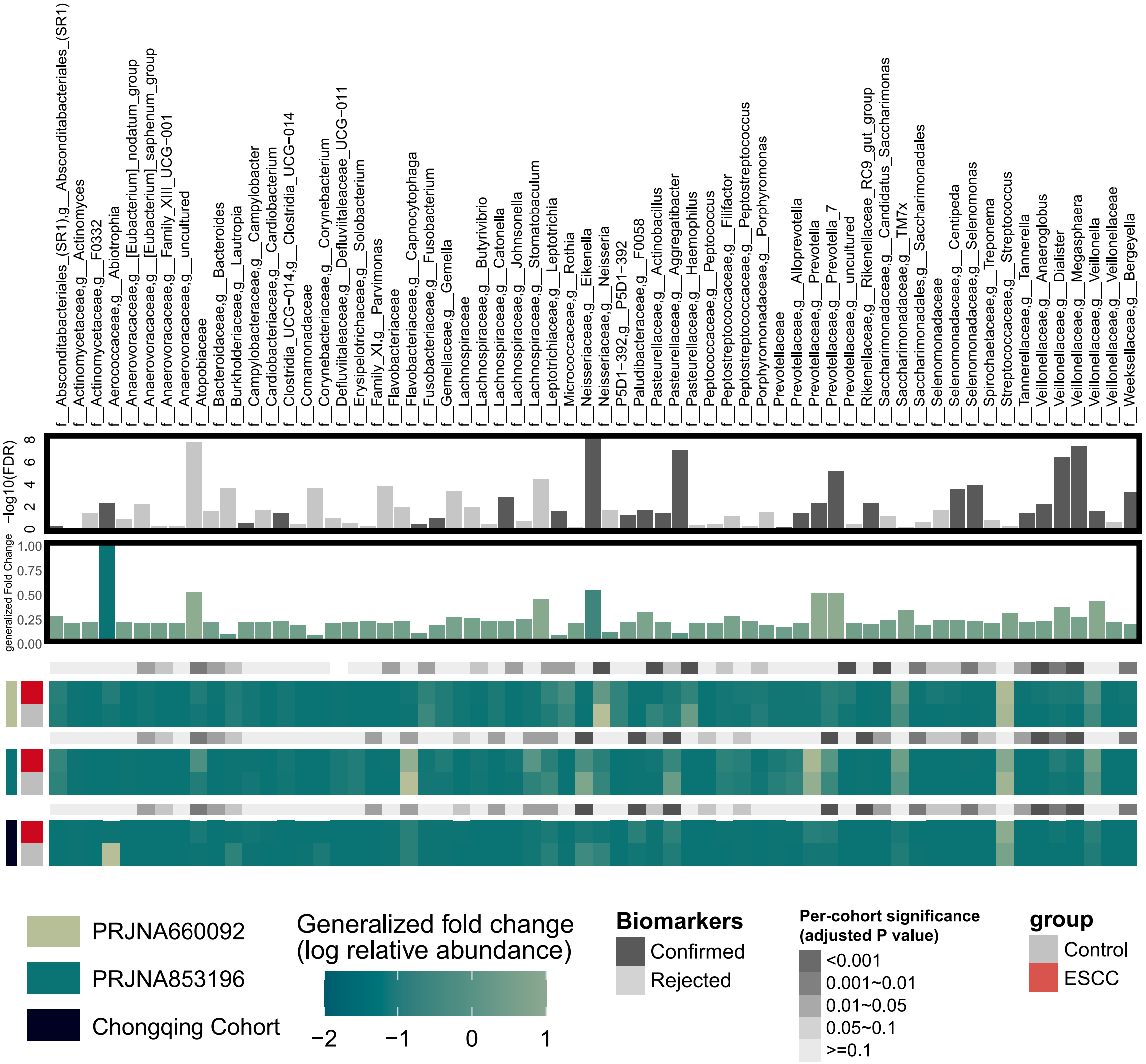


Figure 2. Co-occurrence analysis of ESCC-associated oral microbial ASVs

The top bar graph shows 74 ASVs, annotated to 62 genera, with significant differences in at least one cohort ($p < 0.05$, two-sided blocked Wilcoxon rank-sum test), 43 of which are designated in dark gray as important ASVs for future random forest modeling. The middle bar graph illustrates the generalized fold change (gFC) of these ASVs, with pink representing the 39 species more prevalent in ESCC and blue representing the 45 less prevalent species. Below, heatmaps in grayscale and color represent the significance and gFC of these species within individual cohorts, respectively.

Absconditabacteriales_(SR1), which was prevalent in patients with colorectal cancer and type 2 diabetes mellitus.[17,18] The third module consisted of 3 nodes and 3 edges, with *Selenomonas flueggei*, *Centipeda periodontii*, and *Anaeroglobus geminatus* exhibiting a symbiotic relationship, as illustrated in Figure 3A.

To gain a deeper understanding of these results, the co-occurrence patterns of these important ASVs were organized into three distinct clusters. Clusters 1 and 3, which were more common in patients with ESCC, were composed primarily of taxa from the orders *Veillonellales* and *Selenomonadales*, including genera such as *Megasphaera*, *Selenomonas*, *Anaeroglobus*, *Centipeda*, and *Dialister*, as shown in Figure 4. Previous studies have revealed an inverse relationship between the efficacy of immunotherapy for esophageal cancer and the abundance of taxa within the *Veillonellales* or *Selenomonadales* orders.[19] In contrast, cluster 2, which was more common in healthy individuals, exhibited a diverse range of taxonomic classifications.

Microbial classifier for ESCC

Using the RF algorithm, we constructed stratified 10-fold cross-validation RF models by aggregating all the samples. These models achieved an AUC of 0.87 in differentiating esophageal cancer patients from controls, with an accuracy of 0.78, a sensitivity of 0.78, a specificity of 0.70, a precision of 0.79, and an F1 score of 0.78 (Figure 5A). The top-ranking ASVs identified as features included *Neisseria perflava*, *Haemophilus parainfluenzae*, *Lancefieldella parvula*, *Neisseria flavescens*, *Johnsonella*, *Clostridia_UCG-014*, *Dialister invisus*, *Lacrimispora sphenoides*, *Segatella salivae*, and *Tannerella serpentiformis* (Figure 5B).

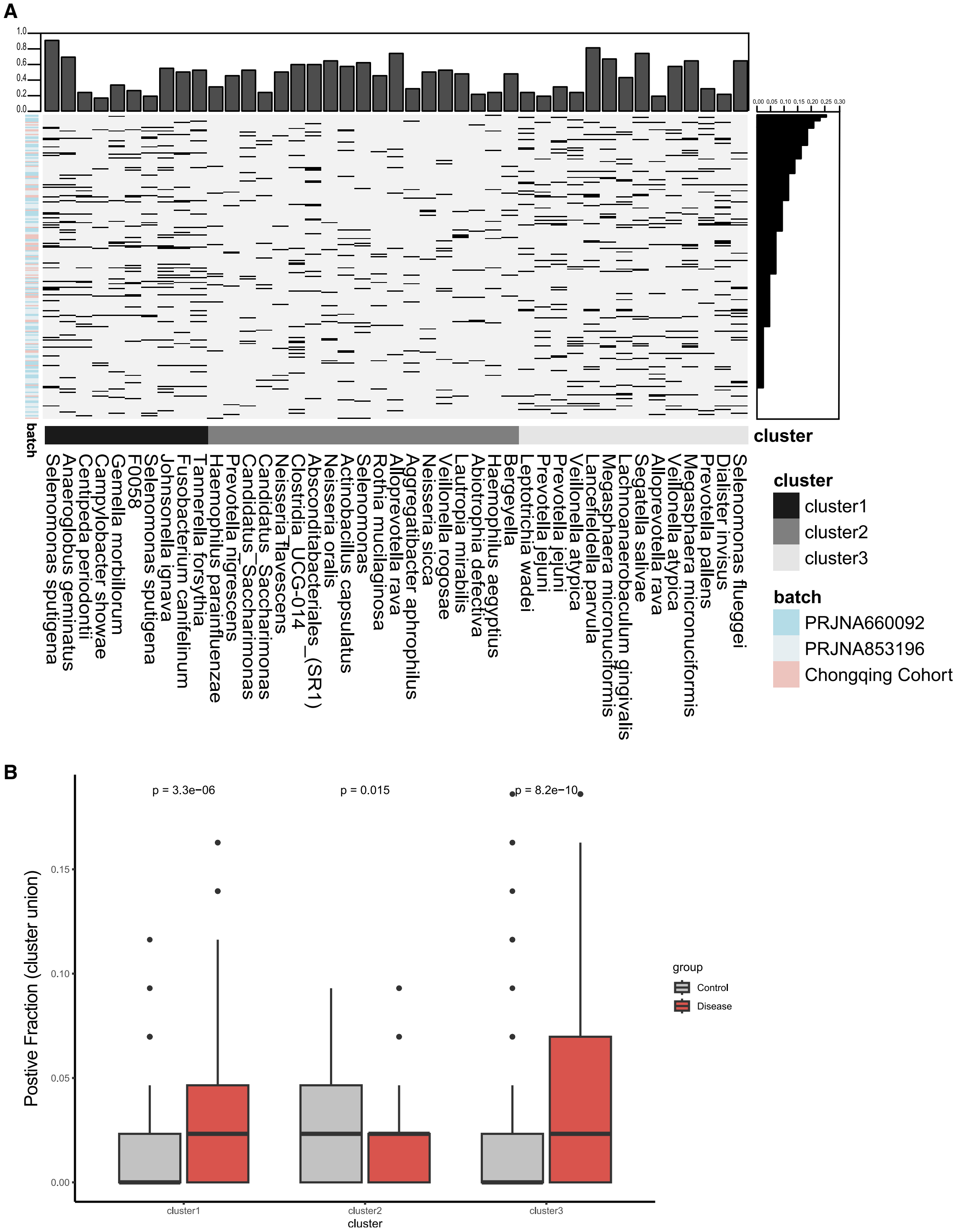


Figure 3. Co-occurrence analysis of oral microbial species associated with ESCC identified three distinct clusters

(A) In the study encompassing 249 independent ESCC patient samples, the heatmap illustrates the presence of core microbial marker species within each sample. The ordering of samples is based on the cumulative count of positive markers, and the clustering of marker species utilizes the Jaccard index of positive sample occurrences, yielding 3 distinct clusters. The X axis represents the genus level of the bacteria; however, if the genus level is uncultured or unannotated, only the family level is displayed.

(B) The bar plot shows the fraction of samples that are positive for marker species clusters (defined as the union of positive marker species) broken down by disease condition. The error bars indicate the standard deviation of the positive fraction calculated. Statistically significant associations between groups and marker species clusters were identified via a two-sided blocked Wilcoxon rank-sum test.

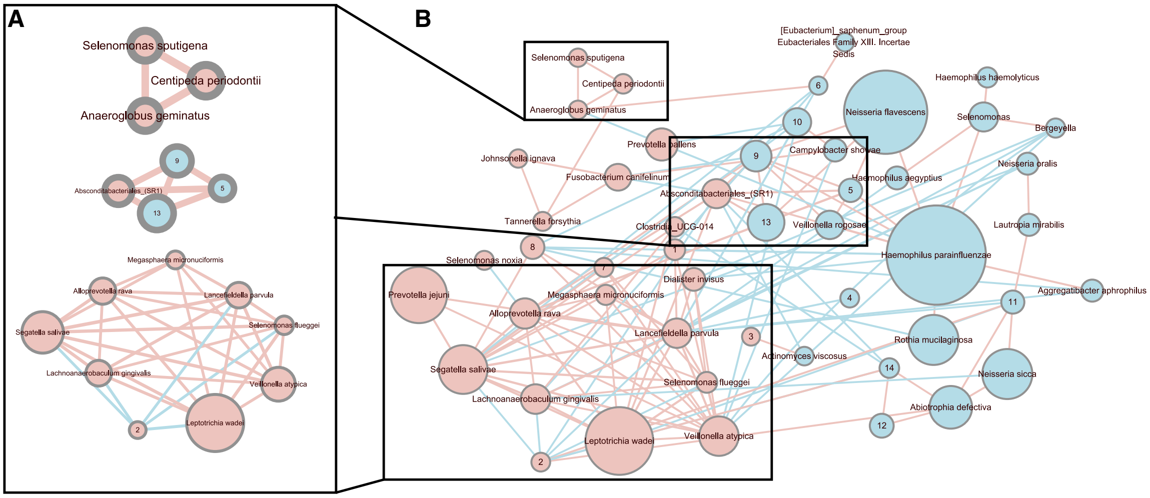


Figure 4. Microbial correlation networks for biomarkers

(A) Modules were generated via the MCODE application from (B).

(B) The correlation network depicted differential ASVs between the disease and control groups, showcasing 50 ASVs. In this network, the node size corresponds to the average abundance of ASVs; biomarker ASVs are labeled with species names, while other ASVs are identified by node numbers. The connections between nodes, or edges, reflect correlations: thickness indicates the strength, and color denotes the type (pink for positive, blue for negative).

Validation of ESCC

To ensure the reliability and consistency of the identified features across multiple studies, we conducted study-to-study transfer validation and leave-one-dataset-out (LODO) validation on the complete sample set. For all the ASV models, the AUC values for study-to-study transfer validation ranged from 0.83 to 0.94, with an average of 0.86 (Figure 6A). Specifically, when these studies served as training datasets, the mean AUCs for PRJNA660092, PRJNA853196, and the Chongqing cohort were 0.90, 0.87, and 0.81, respectively. Conversely, when these studies functioned as testing datasets, the mean AUC values were 0.88, 0.80, and 0.84, respectively. Notably, PRJNA660092 emerged as the optimal training set because of its comparatively higher training AUC, which is likely attributed to its geographical intermediary position between the other two research areas. Additionally, the AUC values for LODO analysis ranged from 0.80 to 0.88, with an average of 0.84 (Figure S4). The PRJNA66092 dataset demonstrated superior predictive accuracy, suggesting that geographic location and eating habits may play crucial roles in variations in the oral microbiome. Therefore, augmenting the dataset with patient data from diverse regions may refine diagnostic precision.

We evaluated the diagnostic potential of various feature sets, including differential ASVs and all salient features. After applying recursive feature elimination, we examined the diagnostic efficacy of critical ASVs. The AUC values for study-to-study transfer validation ranged from 0.85 to 0.97, with a mean of 0.92 (Figure 6B). Moreover, the AUC values for LODO analysis ranged from 0.85 to 0.91, with an average of 0.88 (Figure S5). These findings confirm that the accuracy of the method based on important ASVs is superior to that of all ASVs in the diagnosis of esophageal cancer.

Carcinoembryonic antigen (CEA) and squamous cell carcinoma antigen (SCC-Ag) are known diagnostic and prognostic markers for esophageal cancer.[20,21] We built classifiers based on these blood-based markers, achieving AUCs of 0.86 and 0.84, respectively (Figures 7A–7D). In the Chongqing cohort, the AUC of the ESCC-associated microbial markers was 0.92 (Figures 7E and 7F). Subsequently, the costs associated with endoscopic screening for esophageal cancer, chest helical computed tomography, chest X-ray examination, CEA, and SCC-Ag tests were assessed.[22–28] Our analysis revealed that the cost of 16S rRNA sequencing is comparatively lower (Table 1). Considering the cost and practicality of sample collection, a model based on the oral microbiome is more accessible and accurate. This model not only provides a greater AUC value but also offers a noninvasive and cost-effective alternative to traditional blood-based markers.

Microbial functional changes in ESCC

Oral bacteria engage in various pathways that can either hinder or promote disease progression. To understand these functional shifts within the microbiome under different disease conditions, we analyzed 83 pathways that were differentially expressed between healthy individuals and patients with cancer ($p < 0.05$, with p values adjusted via the Benjamini-Hochberg method). The top five differential pathways are shown in Figure 8. Notably, superpathways of sulfur oxidation, along with nitrate reduction and chondroitin sulfate degradation, were significantly enriched in patients with cancer (Figure 8; Table S3). The nitrate reduction pathway, in particular, was notably influenced by the genera *Schaalia*, *Veillonella*, *Rothia*, and *Neisseria*. Reduced nitrate concentrations are linked to increased susceptibility to esophageal cancer.[29] In contrast, the control samples exhibited enrichment of the superpathway of glycerol degradation to 1,3-propanediol. The concentration of ethanol affects the rate of the superpathway of glycerol degradation to 1,3-propanediol. As the ethanol concentration increased, the degradation rate decreased. Given

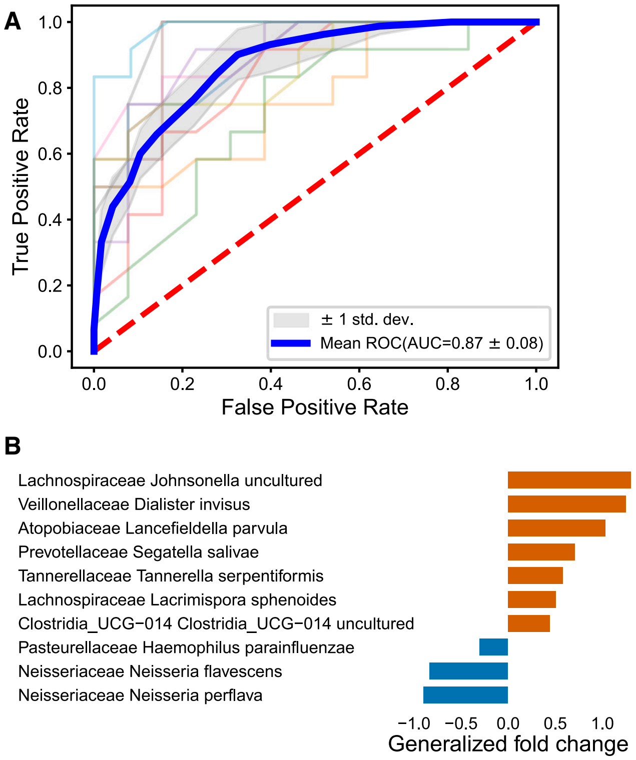


Figure 5. Performance of the oral microbiome in discriminating between cancerous conditions and control samples

(A) The AUC for the optimized models, which were developed using all ASVs, delineates the distinction between control and cancer samples. The different colored lines represent the ROC curves for each of the 10-fold cross-validations. Data are represented as mean ± SD.

(B) The top 10 biomarkers, each corresponding to a unique ASV, were selected to develop random forest (RF) models that differentiate cancerous conditions from control groups. The taxonomic details of these ASVs at the family, genus, and species levels are shown. The biomarkers were ranked by their importance in the RF model, with *p* values calculated via a two-sided blocked Wilcoxon rank-sum test. The generalized fold change was visually represented through color gradients.

that many patients with ESCC have a habit of alcohol consumption, we hypothesize that this habit may lead to a downregulation of the superpathway of glycerol degradation to 1,3-propanediol in the oral cavity of these patients.

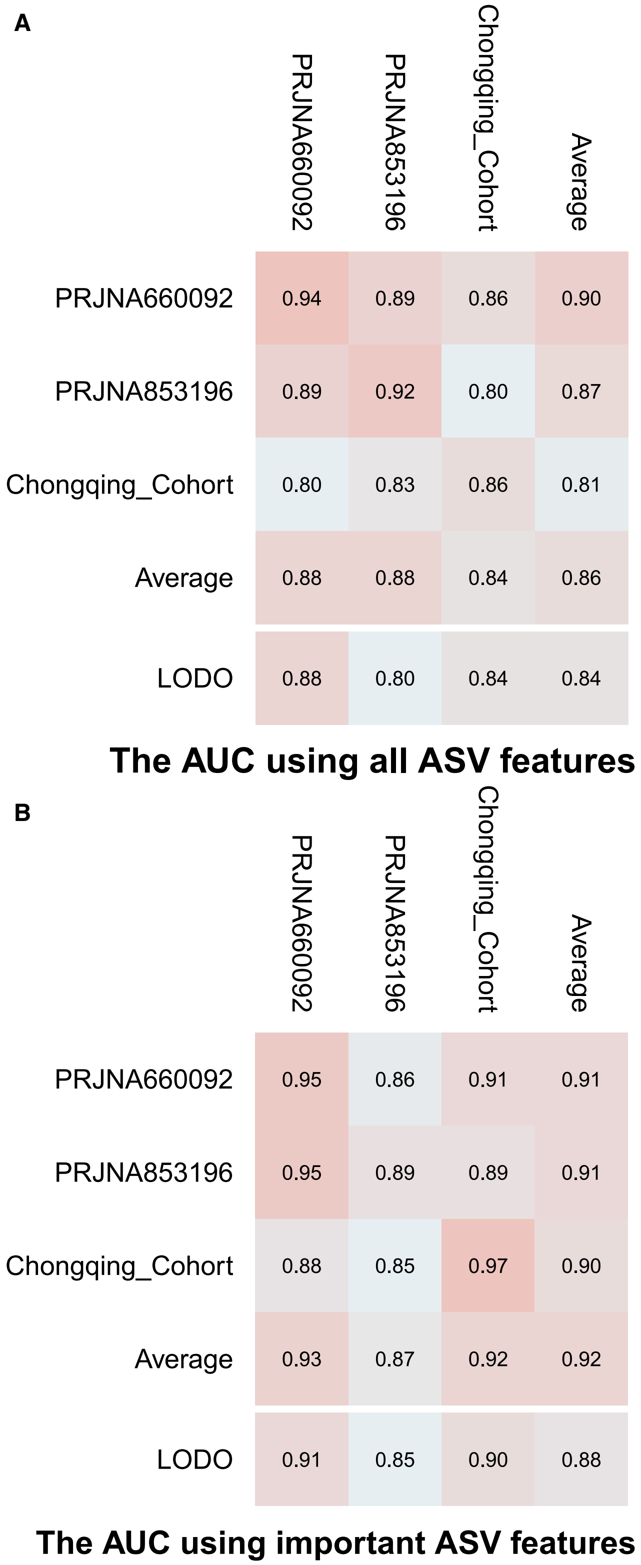
DISCUSSION

In contrast to the pronounced disparities in alpha and beta diversity observed in the gut microbiota of individuals with colon cancer or inflammatory bowel disease compared with healthy volunteers, no such differences were detected in the oral microbiota diversity indices between patients with esophageal cancer and healthy individuals.[11,12,30,31] The use of feature biomarkers can promote the accuracy of model from 0.84 to 0.87 and increased its explanatory ability. Among all the important ASVs, *Veillonella* may be a characteristic that indicates the presence of esophageal squamous cells, which is in line with previous research.[32] Our study revealed that *Neisseria perflava* was also a predictive feature of esophageal squamous carcinoma and inhibited the progression of oral squamous cell carcinoma. These findings may be explained by the similar pathology of squamous cell carcinomas.[33] Moreover, *Atopobium* secretes lactic acid into the oral cavity, where it can travel to tumor primary or metastatic tumor sites through physiological processes such as eating and swallowing. This can change the pH of the surrounding area, which can aid the tumor immune escape and metastasis. Numerous studies have demonstrated that *Fusobacteriaceae* promotes the development of ESCC. However, in this study, *Fusobacteriaceae* was upregulated in patients with esophageal cancer in a single cohort. Further studies are needed to determine the potential utility of *Fusobacteriaceae* as a predictor of esophageal squamous carcinoma within the oral microbiome. This investigation highlights the complex bacterial interactions that may either contribute to or inhibit disease manifestation. The functional pathways enriched in ESCC included those involved in glycerol degradation, sulfur oxidation, and nitrate reduction, whereas pathways related to ubiquinol and menaquinol biosynthesis were more prominent in healthy individuals.

Cancer liquid biopsies are also promising biomarkers for detecting ESCC. Previous studies have reported good performance in terms of blood parameters and extracellular vesicle (EV) protein content.[34,35] The classification of the EV protein is more precise but relatively costly. After collecting and comparing data on the levels of CEA and SCC-Ag in our cohort, we found that the ESCC-associated microbial biomarkers in this cohort exhibited better accuracy and specificity.

During the sample collection phase of our research, oral bacterial sample collectors reported that patients diagnosed with esophageal cancer exhibited suboptimal oral environments in comparison with their healthy counterparts. A substantial proportion of patients with cancer demonstrated a distinct lack of adherence to oral hygiene practices, including a notable absence of regular dental visits for professional teeth cleaning. As previous studies reported, various factors, including dietary practices, oral hygiene conditions, and alcohol and tobacco consumption, can influence the composition of the oral microbiota. Moreover, the resemblance of the oral microbiome in individuals with esophageal cancer to that found in populations who smoke suggests a potential mechanistic link through which smoking may contribute to the onset of ESCC.[14,36] Patients with esophageal cancer usually have a poor prognosis due to osteoporosis, which is exacerbated by the activation of the chondroitin sulfate breakdown pathway by oral bacteria.[37] Consequently, we hypothesize that an impaired oral microbiome may indeed constitute a significant risk factor and poor prognostic indicator for patients with esophageal cancer.

Our functional prediction aforementioned revealed that the superpathway of glycerol degradation to 1,3-propanediol was activated in patients with esophageal cancer. The 1,3-propanediol level, which increases in patients as the tumor progresses, has been reported as a noninvasive diagnostic feature for stomach cancer.[38] Esophageal cancer and gastric cancer exhibit certain pathological similarities; therefore, 1,3-propanediol may serve as a noninvasive biomarker for esophageal cancer.



	PRJNA660092	PRJNA853196	Chongqing_Cohort	Average
PRJNA660092	0.94	0.89	0.86	0.90
PRJNA853196	0.89	0.92	0.80	0.87
Chongqing_Cohort	0.80	0.83	0.86	0.81
Average	0.88	0.88	0.84	0.86
LODO	0.88	0.80	0.84	0.84

The AUC using all ASV features

	PRJNA660092	PRJNA853196	Chongqing_Cohort	Average
PRJNA660092	0.95	0.86	0.91	0.91
PRJNA853196	0.95	0.89	0.89	0.91
Chongqing_Cohort	0.88	0.85	0.97	0.90
Average	0.93	0.87	0.92	0.92
LODO	0.91	0.85	0.90	0.88

The AUC using important ASV features

Figure 6. Predictive accuracy of all ASVs and important ASVs across studies and determination of the minimal set of features required for the detection of ESCC

The cross-prediction matrix provides the predictive value for distinguishing between cancerous and noncancerous samples using random forest (RF) models.

(A and B) The AUC was calculated utilizing all ASVs (A) or important ASVs (B).

Our research presents a low-cost, noninvasive screening technique for detecting esophageal cancer, demonstrating better accuracy and specificity than traditional tumor biomarkers such as CEA and SCC-Ag. However, this study has limitations in terms of experimental validation. To address this, we strengthened the evidence from other aspects of the study design and provided various types of validations for the identified microbial markers. Another limitation of this study stems from the absence of certain information in public databases, which required us to compare the predictive efficacy of CEA and SCC models solely within our dataset against that of ESCC-associated microbial markers. This also makes it challenging to analyze the independent association between the presence of ESCC and the oral microbiome with proper adjustments for these lifestyle factors. More clinical information in public databases could improve the predictive accuracy of the model.

Limitations of the study

The limitations of this study include the inability to analyze certain confounders due to the absence of specific information in public databases. Consequently, we were unable to thoroughly examine the associations between sex or gender and other independent confounders with oral flora. Our analysis was restricted to comparing the capability of oral flora to distinguish between patients with squamous esophageal carcinoma and healthy controls, relative to CEA and SCC-Ag, within the collected cohort. Additionally, the study lacks experimental validation, highlighting the need for further research.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Saijun Fan (fansaijun@irm-cams.ac.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The raw 16S rRNA gene sequencing data reported in this paper have been deposited in the National Genomics Data Center: (https://ngdc.cncb.ac.cn/gsa/) (ProjectID: PRJCA025484). The raw 16S rRNA gene sequencing data from PRJNA660092 and PRJNA853196 cohort were downloaded from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/).
- The code and analysis scripts of this study are available on Zenodo (https://doi.org/10.5281/zenodo.14180648).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We acknowledge funding from the National Natural Science Foundation of China (NSFC, 82273577 to S.F. and 82203978 to T.Z.). We sincerely thank all our laboratory members for their kind support, especially Shuya Zhao, Yichi

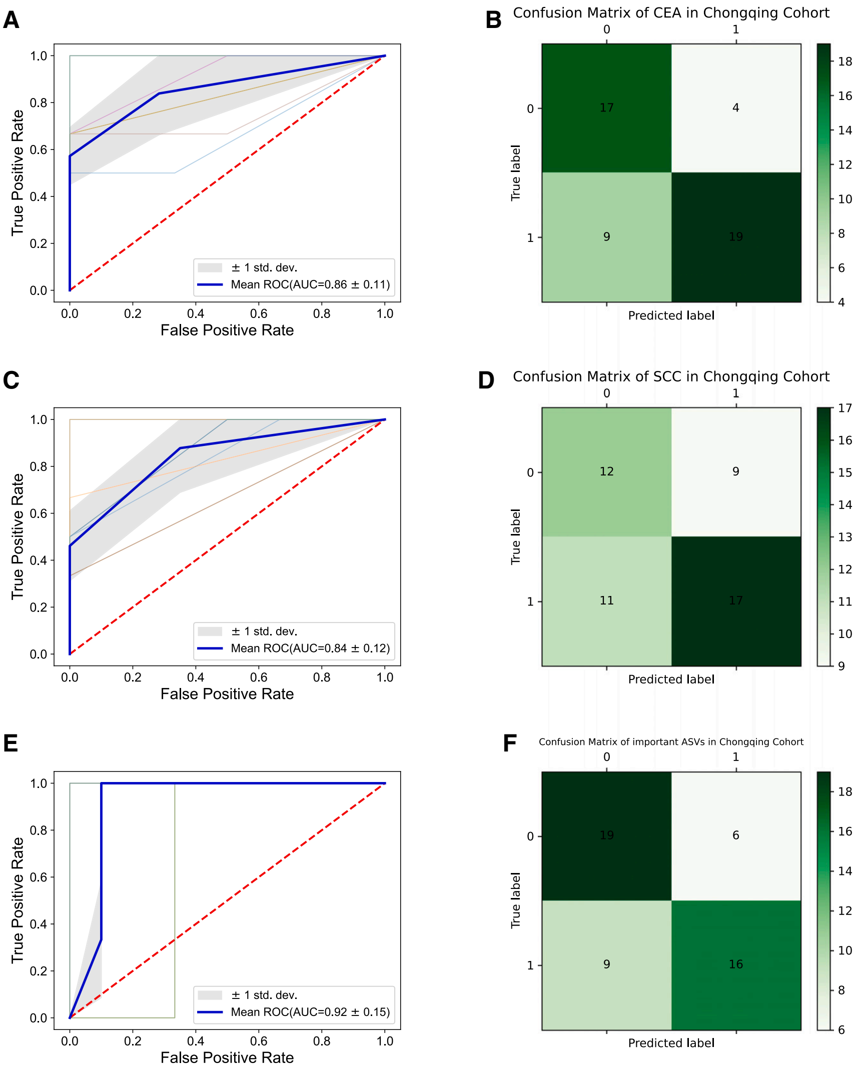


Figure 7. Predictive accuracy of CEA, SCC-Ag, and important ASVs across studies and determination of the minimal set of features required for the detection of ESCC

The cross-prediction matrix provides the predictive value for distinguishing between cancerous and noncancerous samples using random forest (RF) models.

(A, C, and E) The AUC was calculated utilizing CEA (A), SCC-Ag (C), or important ASVs (E). The different colored lines represent the ROC curves for each of the 10-fold cross-validations.

(B, D, and F) The confusion matrix based on CEA (B), SCC-Ag (D), or important ASVs (F). Data are represented as mean ± SD.

Huang, Jiazhe Li, Tong Yuan, Wenzhe Kong, Songling Zhang, Jin Jiang, and Ruirui Wang for their help with our research.

AUTHOR CONTRIBUTIONS

Conceptualization, S.F., L.Z., and T.Z.; methodology, S.F. and T.Z.; investigation, L.X. and Y.Y.; writing – original draft, Y.Y.; writing – review and editing, S.F., L.Z., T.Z., L.X., and Z.W.; funding acquisition, S.F. and T.Z.; resources, T.Z. and Z.W.; supervision, T.Z., L.X., and Z.W.

Table 1. The cost of screening methods for esophageal cancer

Project	Cost per person
Endoscopic screening for esophageal cancer	$194–$400
Helical computed tomography	$130–$560
Chest X-ray	$28–$200
SCC-Ag (squamous cell carcinoma antigen)	$28–$200
CEA (carcinoembryonic antigen)	$12–$28
16S rDNA sequencing	$2–$22

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, the authors used ChatGPT in order to improve the language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

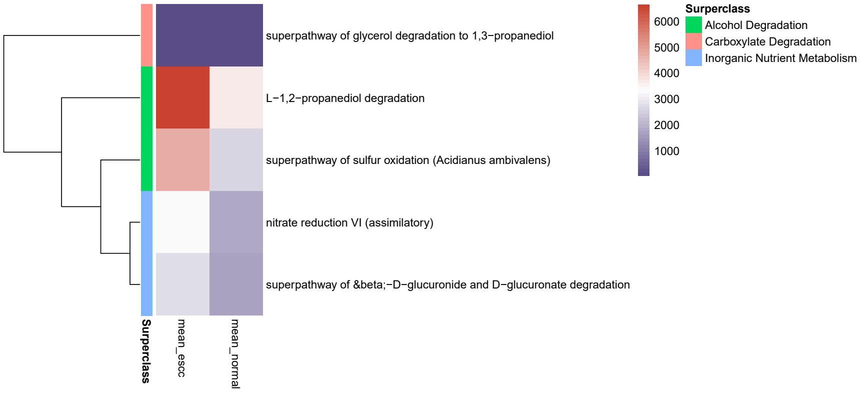


Figure 8. Alterations of function in control and cancer

The variance in functional pathway levels between patients with ESCC and healthy controls was compared. Statistical significance was determined using a two-sided blocked Wilcoxon rank-sum test, with pathways demonstrating p values less than 0.05. The heatmap also displayed abundance through varying color intensities.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111453.

Received: June 4, 2024
Revised: August 13, 2024
Accepted: November 19, 2024
Published: November 22, 2024

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. *71*, 209–249.

2. Middleton, D.R.S., Mmbaga, B.T., Menya, D., Dzamalala, C., Nyakunga-Maro, G., Finch, P., Mlombe, Y., Schüz, J., McCormack, V., Kigen, N., et al. (2022). Alcohol consumption and oesophageal squamous cell cancer risk in east Africa: findings from the large multicentre ESCCAPE case-control study in Kenya, Tanzania, and Malawi. Lancet. Glob. Health *10*, e236–e245.

3. Arnold, M., Ferlay, J., van Berge Henegouwen, M.I., and Soerjomataram, I. (2020). Global burden of oesophageal and gastric cancer by histology and subsite in 2018. Gut *69*, 1564–1571.

4. Shah, M.A., Altorki, N., Patel, P., Harrison, S., Bass, A., and Abrams, J.A. (2023). Improving outcomes in patients with oesophageal cancer. Nat. Rev. Clin. Oncol. *20*, 390–407.

5. An, L., Zheng, R., Zeng, H., Zhang, S., Chen, R., Wang, S., Sun, K., Li, L., Wei, W., and He, J. (2023). The survival of esophageal cancer by subtype in China with comparison to the United States. Int. J. Cancer *152*, 151–161.

6. Obermannová, R., Alsina, M., Cervantes, A., Leong, T., Lordick, F., Nilsson, M., van Grieken, N., Vogel, A., and Smyth, E.C. (2022). Oesophageal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann. Oncol. *33*, 992–1004.

7. Wu, I.-C., Syu, H.-Y., Jen, C.-P., Lu, M.-Y., Chen, Y.-T., Wu, M.-T., Kuo, C.-T., Tsai, Y.-Y., and Wang, H.-C. (2018). Early identification of esophageal squamous neoplasm by hyperspectral endoscopic imaging. Sci. Rep. *8*, 13797.

8. Baker, J.L., Mark Welch, J.L., Kauffman, K.M., McLean, J.S., and He, X. (2024). The oral microbiome: diversity, biogeography and human health. Nat. Rev. Microbiol. *22*, 89–104.

9. Zhao, Q., Yang, T., Yan, Y., Zhang, Y., Li, Z., Wang, Y., Yang, J., Xia, Y., Xiao, H., Han, H., et al. (2020). Alterations of Oral Microbiota in Chinese Patients With Esophageal Cancer. Front. Cell. Infect. Microbiol. *10*, 541144.

10. Jiang, Z., Wang, J., Qian, X., Zhang, Z., and Wang, S. (2023). Oral microbiota may predict the presence of esophageal squamous cell carcinoma. J. Cancer Res. Clin. Oncol. *149*, 4731–4739.

11. Wu, Y., Jiao, N., Zhu, R., Zhang, Y., Wu, D., Wang, A.-J., Fang, S., Tao, L., Li, Y., Cheng, S., et al. (2021). Identification of microbial markers across populations in early detection of colorectal cancer. Nat. Commun. *12*, 3063.

12. Ning, L., Zhou, Y.-L., Sun, H., Zhang, Y., Shen, C., Wang, Z., Xuan, B., Zhao, Y., Ma, Y., Yan, Y., et al. (2023). Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts. Nat. Commun. *14*, 7135.

13. Xiao, L., Zhang, F., and Zhao, F. (2022). Large-scale microbiome data integration enables robust biomarker identification. Nat. Comput. Sci. *2*, 307–316.

14. Suzuki, N., Nakano, Y., Yoneda, M., Hirofuji, T., and Hanioka, T. (2022). The effects of cigarette smoking on the salivary and tongue microbiome. Clin. Exp. Dent. Res. *8*, 449–456.

15. Huang, Q., Wu, X., Zhou, X., Sun, Z., Shen, J., Kong, M., Chen, N., Qiu, J.-G., Jiang, B.-H., Yuan, C., and Zheng, Y. (2023). Association of cigarette smoking with oral bacterial microbiota and cardiometabolic health in Chinese adults. BMC Microbiol. *23*, 346.

16. Watts, S.C., Ritchie, S.C., Inouye, M., and Holt, K.E. (2019). FastSpar: rapid and scalable correlation estimation for compositional data. Bioinformatics *35*, 1064–1066.

17. Russo, E., Gloria, L.D., Nannini, G., Meoni, G., Niccolai, E., Ringressi, M.N., Baldi, S., Fani, R., Tenori, L., Taddei, A., et al. (2023). From adenoma

to CRC stages: the oral-gut microbiome axis as a source of potential microbial and metabolic biomarkers of malignancy. Neoplasia *40*, 100901.

18. Rungrueang, K., Yuma, S., Tantipoj, C., Khovidhunkit, S.O.P., Fuangtharnthip, P., Thuramonwong, T., Suwattipong, M., and Supa-amornkul, S. (2021). Oral Bacterial Microbiomes in Association with Potential Prediabetes Using Different Criteria of Diagnosis. Int. J. Environ. Res. Public Health *18*, 7436.
19. Wu, H., Leng, X., Liu, Q., Mao, T., Jiang, T., Liu, Y., Li, F., Cao, C., Fan, J., Chen, L., et al. (2023). Intratumoral Microbiota Composition Regulates Chemoimmunotherapy Response in Esophageal Squamous Cell Carcinoma. Cancer Res. *83*, 3131–3144.
20. Yang, Y., Huang, X., Zhou, L., Deng, T., Ning, T., Liu, R., Zhang, L., Bai, M., Zhang, H., Li, H., and Ba, Y. (2019). Clinical use of tumor biomarkers in prediction for prognosis and chemotherapeutic effect in esophageal squamous cell carcinoma. BMC Cancer *19*, 526.
21. Verma, R., Sattar, R.S.A., Nimisha, A., Kumar, A., Kumar, A., Sharma, A.K., Sumi, M.P., Ahmad, E., Ali, A., Mahajan, B., and Saluja, S.S. (2021). Cross-talk between next generation sequencing methodologies to identify genomic signatures of esophageal cancer. Crit. Rev. Oncol. Hematol. *162*, 103348.
22. Verberne, C.J., Wiggers, T., Grossmann, I., de Bock, G.H., and Vermeulen, K.M. (2016). Cost-effectiveness of a carcinoembryonic antigen (CEA) based follow-up programme for colorectal cancer (the CEA Watch trial). Colorectal Dis. *18*, O91–O96.
23. Lee, H.-Y., Park, E.-C., Jun, J.K., Choi, K.S., and Hahm, M.-I. (2010). Comparing upper gastrointestinal X-ray and endoscopy for gastric cancer diagnosis in Korea. World J. Gastroenterol. *16*, 245–250.
24. Yu, L., Xu, X., and Niu, S. (2020). Should computed tomography and bronchoscopy be routine examinations for chronic cough? J. Thorac. Dis. *12*, 5238–5242.
25. Wei, W.-Q., Yang, C.-X., Lu, S.-H., Yang, J., Li, B.-Y., Lian, S.-Y., and Qiao, Y.-L. (2011). Cost-benefit analysis of screening for esophageal and gastric cardiac cancer. Chin. J. Cancer *30*, 213–218.
26. Graham, R.A., Wang, S., Catalano, P.J., and Haller, D.G. (1998). Postsurgical Surveillance of Colon Cancer: Preliminary Cost Analysis of Physician Examination, Carcinoembryonic Antigen Testing, Chest X-Ray, and Colonoscopy. Ann. Surg. *228*, 59–63.
27. Chan, Y.M., Ng, T.Y., Ngan, H.Y.S., and Wong, L.C. (2002). Monitoring of Serum Squamous Cell Carcinoma Antigen Levels in Invasive Cervical Cancer: Is It Cost-Effective? Gynecol. Oncol. *84*, 7–11.
28. Johnson, J.S., Spakowicz, D.J., Hong, B.-Y., Petersen, L.M., Demkowicz, P., Chen, L., Leopold, S.R., Hanson, B.M., Agresta, H.O., Gerstein, M., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat. Commun. *10*, 5029.
29. Rogers, M.A., Vaughan, T.L., Davis, S., and Thomas, D.B. (1995). Consumption of nitrate, nitrite, and nitrosodimethylamine and the risk of upper aerodigestive tract cancer. Cancer Epidemiol. Biomarkers Prev. *4*, 29–36.
30. Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med. *25*, 679–689.
31. Halfvarson, J., Brislawn, C.J., Lamendella, R., Vázquez-Baeza, Y., Walters, W.A., Bramer, L.M., D'Amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. Nat. Microbiol. *2*, 17004.
32. Sato, N., Kakuta, M., Hasegawa, T., Yamaguchi, R., Uchino, E., Kobayashi, W., Sawada, K., Tamura, Y., Tokuda, I., Murashita, K., et al. (2020). Metagenomic analysis of bacterial species in tongue microbiome of current and never smokers. npj Biofilms Microbiomes *6*, 11.
33. Shen, X., Zhang, B., Hu, X., Li, J., Wu, M., Yan, C., Yang, Y., and Li, Y. (2022). Neisseria sicca and Corynebacterium matruchotii inhibited oral squamous cell carcinomas by regulating genome stability. Bioengineered *13*, 14094–14106.
34. Min, Y., Deng, W., Yuan, H., Zhu, D., Zhao, R., Zhang, P., Xue, J., Yuan, Z., Zhang, T., Jiang, Y., et al. (2024). Single extracellular vesicle surface protein-based blood assay identifies potential biomarkers for detection and screening of five cancers. Mol. Oncol. *18*, 743–761.
35. Lu, F., Yang, L., Luo, Z., He, Q., Shangguan, L., Cao, M., and Wu, L. (2024). Laboratory blood parameters and machine learning for the prognosis of esophageal squamous cell carcinoma. Front. Oncol. *14*, 1367008.
36. Huang, F.-L., and Yu, S.-J. (2018). Esophageal cancer: Risk factors, genetic association, and treatment. Asian J. Surg. *41*, 210–215.
37. Sugase, T., Sugimura, K., Kanemura, T., Takeoka, T., Yamamoto, M., Shinno, N., Hara, H., Omori, T., Yasui, M., and Miyata, H. (2023). Long-term changes in bone mineral density in postoperative patients with esophageal cancer. Ann. Gastroenterol. Surg. *7*, 419–429.
38. Jung, Y.J., Seo, H.S., Kim, J.H., Song, K.Y., Park, C.H., and Lee, H.H. (2021). Advanced Diagnostic Technology of Volatile Organic Compounds Real Time analysis Analysis From Exhaled Breath of Gastric Cancer Patients Using Proton-Transfer-Reaction Time-of-Flight Mass Spectrometry. Front. Oncol. *11*, 560591.
39. Wang, F.-H., Zhang, X.-T., Li, Y.-F., Tang, L., Qu, X.-J., Ying, J.-E., Zhang, J., Sun, L.-Y., Lin, R.-B., Qiu, H., et al. (2021). The Chinese Society of Clinical Oncology (CSCO): Clinical guidelines for the diagnosis and treatment of gastric cancer, 2021. Cancer Commun. *41*, 747–795.
40. Jia, Y., Jin, S., Hu, K., Geng, L., Han, C., Kang, R., Pang, Y., Ling, E., Tan, E.K., Pan, Y., and Liu, W. (2021). Gut microbiome modulates Drosophila aggression through octopamine signaling. Nat. Commun. *12*, 2698.
41. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857.
42. Bader, G.D., and Hogue, C.W.V. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinf. *4*, 2.
43. Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C., and Langille, M.G.I. (2020). PICRUSt2 for prediction of metagenome functions. Nat. Biotechnol. *38*, 685–688.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human oral flora samples	Chongqing in China	N/A
Deposited data		
16S rRNA gene sequencing data	This study	GSA: PRJCA025484
16S rRNA gene sequencing data	Zhao et al.[9]	GEO: PRJNA660092
16S rRNA gene sequencing data	Jiang et al.[10]	GEO: PRJNA853196
Oligonucleotides		
Forward primer 341F (5′-CCTAYG GGRBGCASCAG-3′)	This paper	N/A
Reverse primer 806R (5′-GGACTA CNNGGGTATCTAAT-3′)	This paper	N/A
Software and algorithms		
FastQC v0.12.1 Babraham Bioinformatics	FastQC v0.12.1 Babraham Bioinformatics	RRID: SCR_014583
QIIME2 Version 2023.7	https://qiime2.org	RRID: SCR_021258
R version 4.4.0	http://www.r-project.org	RRID: SCR_001905
Python version 3.11.5	https://www.python.org/	SCR_008394
FastSpar	https://github.com/scwatts/fastspar	Version 1.0.0
Cytoscape Version 3.10.1	https://cytoscape.org	RRID: SCR_003032
Custom code and script used in this study	This paper	https://doi.org/10.5281/zenodo.14180648

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Study design and participants

Human participants

Twenty-five patients with esophageal squamous cell carcinoma, who were not excluded by the criteria, along with twenty-five age- and gender-matched healthy volunteers, were recruited at the Second Affiliated Hospital of Chongqing Medical University between April 23, 2023, and October 5, 2023. Pathological confirmation and diagnosis of esophageal cancer were conducted according to the guidelines of the Chinese Society of Clinical Oncology.[39] The exclusion criteria included individuals with systemic diseases, those who had taken antibiotics, proton pump inhibitors, prebiotics, or other preparations within the previous month, and those with a history of oral ulcers in the preceding six months. Additionally, patients who had undergone any form of tumor treatment, including surgery, radiotherapy, chemotherapy, or immunotherapy, were excluded. Matched healthy controls, confirmed to be cancer-free by helical CT, were also enrolled in a 1:1 ratio in this study. Age, sex, gender, BMI, smoking and drinking habits, family history of tumors, race, ethnicity, ancestry, and province of the individuals are reported in Table S1.

Ethics approval

Written informed consent was obtained from each participant, ensuring adherence to the ethical principles outlined in the World Medical Association's Declaration of Helsinki (2008) and the Belmont Report. The experimental protocol was approved by Human Ethics Committee of The Second Affiliated Hospital of Chongqing Medical University with the approval number 117/2023.

METHOD DETAILS

Public data collection

In the context of esophageal squamous cell carcinoma (ESCC), we collected 16S rRNA sequencing data for both ESCC patients and healthy individuals from the National Center for Biotechnology Information (NCBI) database. This investigation focused on two studies that provided comprehensive sample metadata and utilized high-throughput sequencing techniques targeting the V3-V4 hypervariable regions of the 16S rRNA gene. The raw sequencing datasets associated with these studies were constructed using the SRA toolkit, version 3.0.6, from the Sequence Read Archive (SRA), corresponding to the accession numbers PRJNA660092 and PRJNA853196.[9,10]

Clinical sample collection

The samples was collected via sterile brushes. The participants were required to abstain from food, smoking, and alcohol for 2 h prior to the collection process. Samples were obtained from the posterior pharyngeal wall and gingiva to ensure comprehensive coverage of the oral cavity surfaces. After collection, the samples were immediately stored at −80°C to preserve their integrity until DNA extraction.

In our clinical laboratory, the serum CEA and SCC-Ag levels were detected using radioimmunoassay (RIA). A 3-mL venous blood sample was collected and centrifuged at 3,500 rpm for 10 min at room temperature. After centrifugation, the processed samples were subjected to immunoassay analysis using the fully automated DiaSorin LIAISON XL and ARCHITECT i2000SR chemiluminescence systems.

Sequencing processing

Each sample was frozen in dry ice and sent to Novogene Bioinformatics Technology Co., Ltd (Beijing, China, https://cn.novogene.com/). Total bacterial DNA extraction and sequencing were performed according to standard protocols. Specifically, genomic DNA was extracted from the collected samples using the cetyltrimethylammonium bromide (CTAB) method and diluted to a standard concentration of 1 ng/μL with sterile water. Polymerase chain reaction (PCR) was performed with universal primers targeting the V3-V4 region of the 16S rRNA genes, specifically 341F (5′-CCTAYGGGRBGCASCAG-3′) and 806R (5′-GGACTACNNGGGTATCTAAT-3′). Sequencing libraries were prepared using the NEB Next Ultra II FS DNA PCR-free Library Prep Kit (New England Biolabs, USA, Catalog #: E7430L) following the manufacturer's recommendations, with indexes added subsequently. The library quality was assessed using Qubit and real-time PCR for quantification, and a bioanalyzer for size distribution detection. Upon approval of library quality control, libraries were pooled according to their effective concentration and target data output requirements. The pooled libraries were then sequenced on the Illumina NovaSeq 6000 platform with P250 sequencing, generating 250 bp paired-end reads through synthesis by sequencing.[40] High-throughput sequencing produced raw image data files, which were converted into raw sequencing reads via base calling analysis. The results were stored in FASTQ (fq) file format, containing both the sequence information of the reads and their corresponding sequencing quality information. The quality of the raw data was assessed using FastQC (version 0.12.1), excluding reads with a quality score below 30 or shorter than 150 bp.

Microorganism analysis pipeline

All the raw sequencing data were processed consistently on the Quantitative Insights Into Microbial Ecology 2 (QIIME2-2023.7) platform.[41] The primers were removed using Cutadapt (version 4.5), and the "join-pairs" function of the VSEARCH plugin in QIIME2 was used to merge the reads. The joined reads were then input into the Deblur plugin to construct the ASV feature table, with singletons filtered out. After quality filtering, 18,960,674 reads remained, with an average of 76,147 reads per sample. A total of 4,133 ASVs were detected after clustering sequences at 99% similarity with the SILVA database (version 138) using the classify-sklearn algorithm in the feature classifier plugin. ASVs that could not be precisely annotated to species were reassigned to the most similar sequences in the same genus or family using NCBI BLAST. The mean taxon abundance was assessed at different taxonomic levels, including the species, genus, family, class, and phylum levels, for both the esophageal cancer and control groups. Low-abundance ASVs, with a relative abundance less than 0.1% in at least 20% of each study, were excluded from subsequent analysis.

Confounder analysis

To assess the potential confounding variables in relation to ESCC for a specific amplicon sequence variant (ASV), an analysis of variance (ANOVA)-type approach was employed.[12] To determine the total variance explained by the abundance of a particular ASV, we utilized a linear regression model that incorporated both ESCC status and confounding factors as independent variables. Variance estimations were performed using ranked values to address the non-Gaussian distribution of microbiome abundance data.

Univariate meta-analysis for the identification of ESCC-associated oral microbial species

The significance of differential abundance for individual amplicon sequence variants (ASVs) was assessed using a two-sided blocked Wilcoxon rank-sum test, facilitated by the "coin" package (version 1.4.3) in R (version 4.4.0).[30] To mitigate potential confounding influences from batch effects, each ASV was analyzed individually, with data stratified by cohort. Permutation tests were conducted within each stratum to generate a conditional null distribution, accounting for variances in block composition and magnitude. Next, *p*-values were corrected using the false discovery rate (FDR) approach to address multiple hypothesis testing. Additionally, the magnitude of disparity between control and esophageal squamous cell carcinoma (ESCC) samples was quantified using the generalized fold change (gFC) methodology.

Model construction and feature selection

The random forest (RF) algorithm, a hierarchical ensemble of decision trees, is particularly adept at handling microbiome data, and is often represented as a sparse matrix with intricate variable interdependencies. We utilized the RF algorithm's robust feature selection capability to construct models using the scikit-learn package (version 1.3.0) in Python (version 3.11.5), incorporating all ASVs. Stratified 10-fold cross-validation was used to fine-tune the training and testing datasets, ensuring accurate classification of cancer versus control samples. The most predictive features, identified as "important features," were further characterized as "biomarkers" on the basis of their performance in the top RF model. Model efficacy was assessed using metrics such as the area under the curve (AUC), accuracy, sensitivity, specificity, precision, and F1 score. To enhance model reliability and reduce complexity, we utilized the recursive feature elimination (RFE) method. This process begins with a differential feature analysis to pinpoint potential discriminative features. An RF model was then trained using the scikit-learn package in Python (version 3.11.5) and subjected to stratified 10-fold cross-validation to differentiate between ESCC patients and controls. The RFE step was instrumental in refining the model by systematically eliminating less informative features, culminating in a set of discriminative features. The top-performing features, identified by the highest AUC value, were deemed "important ASVs," forming the final feature set for the model. These ASVs are pivotal for understanding disease progression and may serve as biomarkers for ESCC.

Model evaluation

Leave-one-dataset-out (LODO) validation and study-to-study transfer validation methodologies were used to assess the generalizability of microbial-based adenoma classifiers. These approaches evaluate classifier performance across diverse contexts, considering geographic variability and technical disparities in microbial data acquisition and processing across different patient cohorts. The study-to-study transfer validation involved developing classifiers within a single study and subsequently assessing them on all remaining cohorts, as illustrated by the off-diagonal elements in Figure 6.

To determine the optimal performance of pivotal features within study-to-study transfer validation and LODO validation, models were constructed using three distinct input feature sets: (1) the entirety of amplicon sequence variants (ASVs), (2) differential ASVs, and (3) all essential features. The objective was to ascertain whether a minimal subset of critical features could enhance accuracy. A selection of the highest-ranking essential features was consistently incorporated into the minimal subset *a priori*. Using methodologies identical to those used in study-to-study transfer validation and LODO validation, the average area under the curve (AUC) for each testing study was determined and represented as individual points in Figures 6A and 6B. A comparative analysis of the predictive values within the testing set was then conducted across models with varying input feature sets.

Microbial ecological analysis

Species diversity and evenness within a community were assessed using alpha diversity measures, specifically the Shannon and Simpson indices. Nonmetric multidimensional scaling (NMDS) analysis was performed using the vegan package (version 2.6.8) within the R software framework, which is based on a normalized table of ASV abundances. Beta diversity was measured by the Bray−Curtis dissimilarity metric, which compares microbial community structures across samples. Variations in bacterial community composition between different disease groups or cohorts were analyzed using 999 permutations within a permutational multivariate analysis of variance (PERMANOVA).

Species co-occurrence and cluster analysis in ESCC

Microbial communities often collaborate synergistically to increase their pathogenicity, resilience, or colonization capacity. To facilitate a comprehensive understanding, co-occurrence networks were formulated and depicted. FastSpar (version 1.0.0), an accelerated and more efficient variant of the SparCC algorithm, was used to swiftly construct correlation networks and determine *p* values.[16] SparCC ensures the robustness and transferability of the component data by accounting for the diversity and sparsity of community members. Correlation coefficients were assessed using an average of 50 inference iterations, adhering to the default threshold for strength. *p* values were derived from 1,000 bootstrap correlations. For subsequent visualization in Cytoscape (V3.10.1), only correlation coefficients with *p* values less than 0.05 and magnitudes exceeding 0.3 were selected.[11] The MCODE algorithm was utilized to evaluate modular structures and clusters of tightly interconnected nodes, using the default parameter settings.[42]

To investigate the co-occurrence patterns of biomarkers, we converted the relative abundances of biomarker ASVs into binary values, labeling them as either "positive" or "negative". A sample was considered "positive" if the relative abundance of a biomarker ASV exceeded a threshold of 0.[11] We then created a binarized marker-by-sample matrix and analyzed it using the Jaccard index.

Functional profile analysis

PICRUSt2, an amalgamation of existing open-source methodologies, facilitates the prediction of genomes from environmentally sampled 16S rRNA gene sequences, enabling the estimation of microbial abundance and diversity. Consequently, the functional attributes of the oral microbiome were deduced from 16S rRNA sequences using PICRUSt2 (V2.5.2) following established protocols.[43] Functional profiles with a relative abundance of less than 1x10ˆ-5 in more than 80% of the samples and presence in fewer than three studies were excluded. Differential analysis and generalized fold change (gFC) calculations were conducted on pathway profiles, similar to the ASV profiles described in the data preprocessing section. The individual contributions of each ASV to the aggregate differential pathways were then assessed. This contribution was quantified as the proportion of the functional abundance of a single ASV relative to the cumulative functional abundance of all ASVs within a specific pathway. Finally, all differential pathways were categorized on basis of their gFC scores.

QUANTIFICATION AND STATISTICAL ANALYSIS

Results are given as mean (± standard deviation) if not otherwise indicated. Multiple group comparisons of data that adhered to a normal distribution were performed using analysis of variance (ANOVA)-type approach. The *p* values obtained from a two-sided blocked Wilcoxon rank-sum test were used to determine statistical significance. Two-sided Kruskal–Wallis test was used in PCoA analysis. P-values less than 0.05 were considered statistically significant, and *p*-values less than 0.01 were considered highly significant.