npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-025-00708-8

Oral microbiome-derived biomarkers for non-invasive diagnosis of head and neck squamous cell carcinoma

Jingtai Zhi[1,2], Yibo Liang[1,2], Wang Zhao[1], Jie Qiao[1], Yongzhe Zheng[1], Xin Peng[1], Li Li[1] ✉, Xianfeng Wei[1] ✉ & Wei Wang[1] ✉

Mounting evidence suggests that sustained microbial dysbiosis is associated with the development of multiple cancers, while the species-level bacterial taxa and metabolic dysfunction of oral microbiome in patients with head and neck squamous cell carcinoma (HNSCC) remains unclear. In this cross-sectional study, comprehensive metagenomic and 16S rRNA amplicon sequencing analyses of oral swab samples from 172 patients were performed. Unsupervised clustering algorithms of relative microbial abundance profiles revealed three distinctive microbiome clusters. Based on the metagenomic and 16S rRNA amplicon sequencing data, machine learning-based methods were used to construct the HNSCC diagnostic classifier, which exhibited high area under the curve values of 0.78–0.89. Our study provided the first exhaustive metagenomic and 16S rRNA amplicon sequencing analyses to date, revealing that microbial-metabolic dysbiosis is a potential risk factor for HNSCC progression and therefore providing a robust theoretical basis for potential diagnostic and therapeutic strategies for HNSCC patients.

Head and neck squamous cell carcinoma (HNSCC) represents the majority of head and neck cancers and a heterogeneous group of tumors that arise from the squamous epithelium of the oral cavity, oropharynx, larynx and hypopharynx[1]. In the United States, HNSCC was predicted to have resulted in ~54,540 people and result in 11,580 deaths in 2023[2]. Even with multimodal treatment combining surgery, radiotherapy, and chemotherapy, locally advanced HNSCC has an aggressive clinical course with a high tendency for both local recurrence and distant metastasis, which is the main cause of death in HNSCC patients[3,4]. The current understanding of the carcinogenesis and management of HNSCC is insufficient to effectively control short-term and long-term morbidity.

Increasing evidence across multiple cancers indicates the microbial roles in cancer formation, diagnosis, prognosis, and treatment[5,6]. Currently, most proposed cancer-microbe relationships focus on gut microbiota alteration in patients with colorectal cancer (CRC). The close proximity facilitates the generation of abundant and direct interactions between the dysfunctional gut microbiome and CRC, for example, by supplying key metabolites for CRC proliferation, shaping the suppressive tumor microenvironment, and participating in chemoresistance[7].

Although the oral cavity similarly also has a close anatomical location to the pharynx and larynx, the relationship between the oral microbiome and HNSCC is unclear. Studies have recognized for decades that poor oral hygiene, periodontitis, and tooth loss are strongly associated with the development and/or progression of HNSCC[8], epidemiologically proving that oral microbial dysbiosis is a risk factor for HNSCC. Recent sporadic evidence has revealed an association between the microbiota and HNSCC[9,10]; however, a comprehensive understanding of microbiome dynamics across laryngeal and hypopharyngeal HNSCC progression is limited. In addition, most published studies have focused on nasopharyngeal and oral HNSCC[9,11,12], and the low taxonomical and functional resolution of 16S rDNA sequencing may limit the ability to interpret the results[6,13,14]. Given our lack of understanding regarding the functional and compositional changes in the microbiota during HNSCC onset and progression, the identification of certain microbes or a consortium of oral microbes with potential causative roles in hypopharyngeal and laryngeal HNSCC remains obscure.

To evaluate the oral microbial structural and functional alterations across stages of HNSCC carcinogenesis, we performed probabilistic partitioning of 16S rRNA amplicon sequencing data and revealed three

[1]Department of Otorhinolaryngology-Head and Neck Surgery, Tianjin First Central Hospital, Institute of Otolaryngology of Tianjin, Key Laboratory of Auditory Speech and Balance Medicine, Key Medical Discipline of Tianjin (Otolaryngology), Quality Control Centre of Otolaryngology, Tianjin First Central Hospital, Tianjin, PR China. [2]These authors contributed equally: Jingtai Zhi, Yibo Liang. ✉e-mail: daisylily20@sina.com; weixianfeng3@163.com; entwangwei@nankai.edu.cn

distinct microbiome clusters, thereby overcoming the considerable interpersonal variation in the microbial composition of the human oral cavity. We subsequently utilized metagenomic sequencing to thoroughly analyze shifts in microbiome composition and variations in the abundance of microbial metabolic genes across the three microbiome clusters. In addition, we developed the first metagenomic-based diagnostic classifier for HNSCC by utilizing the aforementioned data gained from these ground-breaking findings. Our findings may facilitate insights into the specific mechanisms underlying the pathophysiology of HNSCC, thereby providing a robust theoretical basis for diagnostic and therapeutic strategies.

Results

To provide a blueprint of the oral microbiome in HNSCC patients, we performed a comprehensive analysis of 16S rRNA amplicon and metagenomics sequencing (Fig. 1). Per the overall strategy, we performed 16S rRNA amplicon sequencing on oral swab samples collected from subjects with benign lesions ($n = 56$), subjects with precancerous lesions ($n = 29$), subjects with early-stage HNSCC ($n = 39$), and subjects with late-stage HNSCC ($n = 48$). After the rigorous and reliable sample testing and quality control process and the standardized data sequencing and analysis process (see Methods for details), species-level data were fitted to the DMM (Dirichlet multinomial mixture) model, and three microbiome clusters with distinct microbial proportions and clinical features were defined.

To provide a comprehensive understanding of microbial ecosystems and functional alterations, we performed additional metagenomic sequencing on samples from the 91 subjects, which was a sufficient number of samples for further study. The metagenome cohort included subjects with benign lesions ($n = 42$), subjects with precancerous lesions ($n = 17$), subjects with early-stage HNSCC ($n = 13$), and subjects with late-stage HNSCC ($n = 39$). After the sample testing process, quality control process, and standardized data sequencing and analysis processes (see "Methods" for details), rigorous and reliable metagenomic data were utilized to perform microbial composition analysis, functional alteration analysis, and contribution analysis sequentially. Finally, we constructed four classifiers based on our 16S rRNA amplicon sequencing and metagenomic data to evaluate the potential diagnostic value for oral microbial dysbiosis. In summary, our data may provide a reusable data resource for alterations in the oral microbiota in patients with HNSCC and provide new insights into HNSCC carcinogenesis.

Three microbiome clusters were identified that correlated with the HNSCC status using the 16S rRNA amplicon sequencing data

We started our data analysis with the comprehensive 16S rRNA amplicon sequencing cohort and aimed to explore the phylogenetic diversity of microbiome clusters. The species-level rarefaction curves approached plateaus with increasing sample size, indicating adequate 16S rDNA gene sequence coverage (Supplementary Fig. 1A). To gain insight into microbiome cluster structure and composition, the filtered 16S rDNA microbiome data were fitted to DMM models. The screening criterion was the presence of the genus in at least 50% of samples with a relative abundance of at least 0.1% on average. In this way, we could identify subclusters, estimate taxon abundances, and handle overdispersion associated with high-dimensional count data from microbiome clusters (see Supplementary Fig. 1B, C and Supplementary Table 1 for a comparison of the ordination results from DMM- and PAM-based clustering).

Notably, we identified three microbiome clusters designated in Fig. 2A as "C1–C3" and observed strong associations with pathological characteristics (Fisher exact test, $P < 0.001$). Among the 3 microbiome clusters, microbiome cluster 1 (C1) was strongly associated with advanced HNSCC, T3-T4 stage, TNM III-IV stage (Fisher exact test or chi-square test, $P < 0.001$, $P = 0.043$, and $P = 0.045$, respectively, Table 1). In total, the microbiome clusters in 44.8% (39/87) of the HNSCC samples and 57.4% (27/47) of the late-stage HNSCC samples were classified as microbiome cluster 1 (C1). The microbiome clusters in 50.0% (28/56) of samples of benign lesions were identified as microbiome cluster 2 (C2), and those in 72.4% (21/29) of precancerous lesion samples were identified as microbiome cluster 3 (C3). Microbiome clusters 2 and 3 together represented 80.4% (45/56) of the benign lesion samples. In summary, the overall microbiome composition aligns well with the clinical features according to the DMM clustering strategy, which highlights the value of our data for clinical-based diagnostic applications.

Next, we further explored the microbial composition alteration in three microbiome clusters. At the phylum and genus levels, the main members of each microbiome cluster were generally consistent (Supplementary Fig. 1D and Supplementary Fig. 1E). All microbiome clusters were represented by

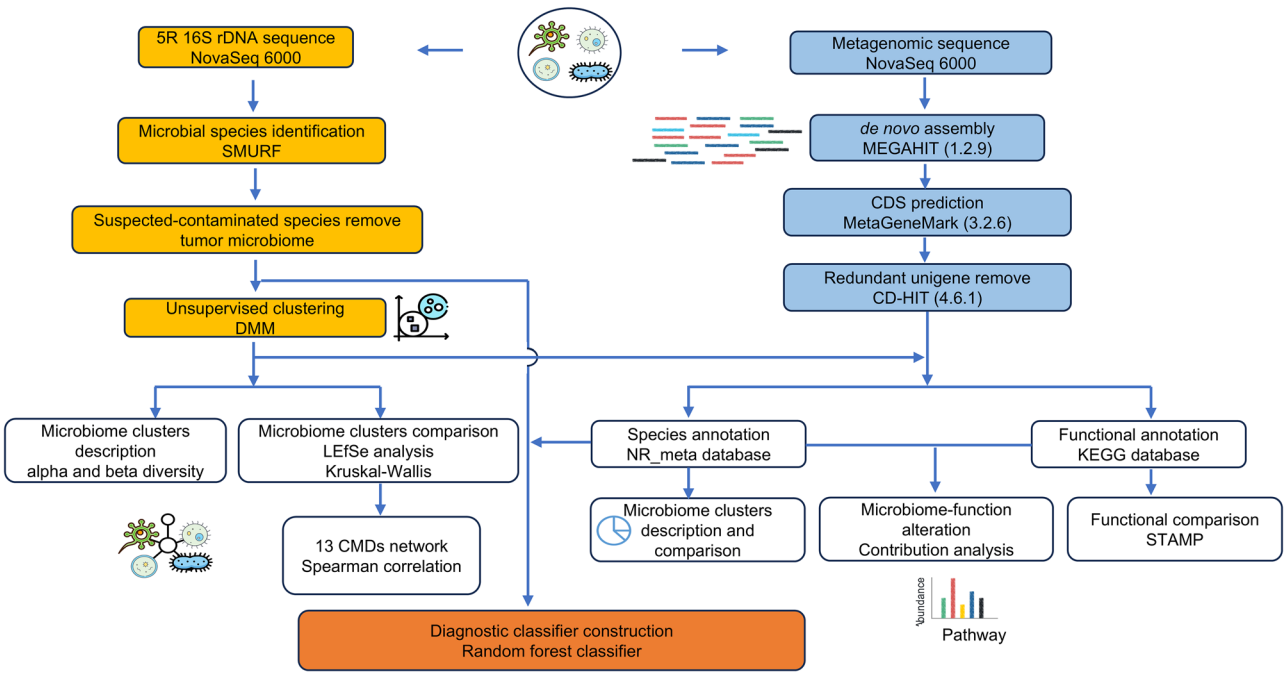


Fig. 1 | Flowchart of the comprehensive analysis of 16S rRNA amplicon and metagenomic sequencing data.

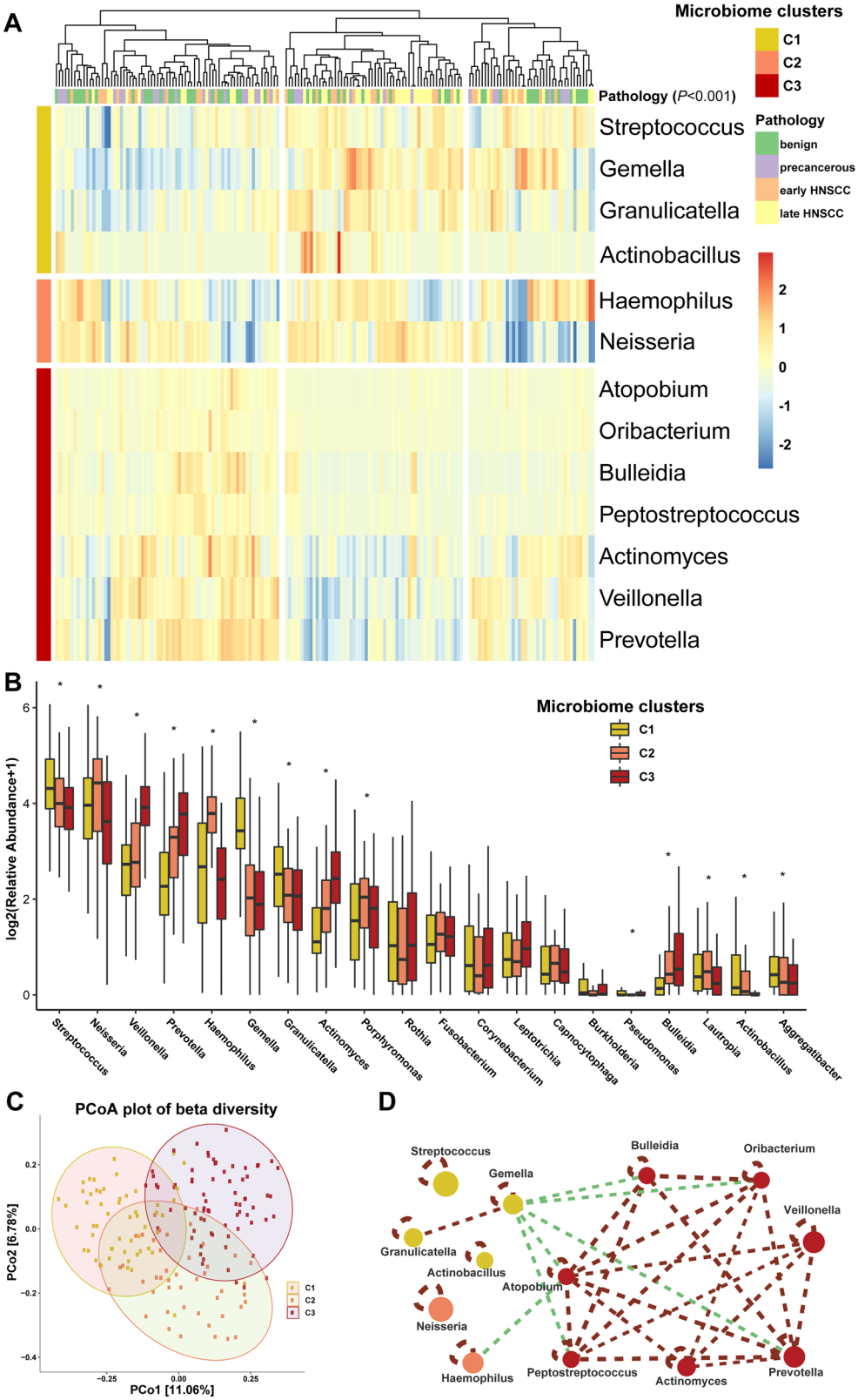


Fig. 2 | Overview of 16S rDNA sequencing based on the three microbiome clusters. A Heatmap depicting key microbiome profiles of 172 oral swab biopsies and corresponding pathological information from individuals with or without head and neck squamous cell carcinoma (HNSCC). The LEfSe algorithms were used to determine and evaluate the marker genera of the oral microbiotas of the three microbiome clusters. **B** Relative abundance of the top 20 genera was evaluated with box plots (Kruskal–Wallis test). $P < 0.05$ marked as *. **C** Principal coordinate analysis (PCoA) with the Bray–Curtis distance to assess the distributional differences in the oral microbiotas of the three microbiome clusters. The *P* value was calculated by the analysis of similarities (ANOSIM). The plots in yellow, pink, and red represent the three microbiome clusters. **D** Network visualization of the 13 marker genera of the oral microbiota. The three cluster-specific marker genera are labeled with yellow, pink, and red. The size of the nodes is proportional to the relative abundance. The width of the edge was the Spearman correlation between the two nodes (genera) across the study cohort. The edges are marked in red or green to represent positive or negative correlations, respectively.

Table 1 | Comparation of the pathological characteristics of the three microbiome clusters based on DMM models

	Microbiome cluster 1	Microbiome cluster 2	Microbiome cluster 3	χ^2 or F^*	*P* value
Age[a] (*N* = 172)	60.00 ± 13.38	51.86 ± 14.40	57.16 ± 13.11	4.541	0.012
Sex (*N* = 172)					
Female	9	15	21	4.268	0.118
Male	46	30	51		
Smoking status (*N* = 170)					
Nonsmokers	15	19	30	3.998	0.406
Smokers	32	20	35		
Former smoker	7	6	6		
Drinking status (*N* = 170)					
Nondrinkers	42	37	55	1.627	0.804
Drinkers	8	5	13		
Former drinkers	4	3	3		
Pathology (*N* = 172)					
Benign	11	28	17	42.365	<0.001
Precancerous	5	3	21		
Early-stage HNSCC	12	9	18		
Advanced HNSCC	27	5	16		
T stage (*N* = 87)					
Tis	0	0	4	14.762	0.043
T1	12	5	4		
T2	6	4	12		
T3	13	3	5		
T4	8	2	9		
N stage (*N* = 87)					
N0	25	12	26	7.137	0.255
N1	4	2	5		
N2	9	0	3		
N3	1	0	0		
M stage (*N* = 87)					
M0	39	14	33	1.795	0.552
M1	0	0	1		
TNM stage (*N* = 87)					
0	0	0	4	14.625	0.045
I	6	5	4		
II	6	4	10		
III	17	3	6		
IV	10	2	10		
p16 IHC staining (*N* = 77)					
No staining	18	8	27	2.600	0.646
Faint	4	2	2		
Intense	7	2	7		
Ki67 IHC staining[a] (*N* = 79)	54.84 ± 16.35	52.69 ± 18.21	59.519 ± 18.49	0.836	0.437

[a]Age and Ki67 IHC staining are compared with ANOVA and presented as mean ± SD.

Firmicutes, *Proteobacteria*, *Bacteroidetes*, *Actinobacteria*, and *Fusobacteria* at the phylum level. All microbiome clusters were represented by distinct proportions of *Streptococcus*, *Neisseria*, *Veillonella*, *Prevotella*, *Haemophilus*, *Gemella*, *Granulicatella*, *Actinomyces*, *Porphyromonas*, and *Rothia* at the genus level. However, the microbial proportions in different microbiome clusters varied remarkably, indicating that HNSCC-related microbiome cluster 1 has oral microbial dysbiosis.

Subsequently, we screened for taxa that distinguished the microbiome clusters using the LEfSe algorithm (Fig. 2A; see Supplementary Table 2 for the summary of LDA scores). In microbiome cluster 1, the representative members included *Streptococcus, Gemella, Granulicatella*, and *Actinobacillus* (Fig. 2A). Microbiome cluster 2 was dominated by *Haemophilus* and *Neisseria*. Microbiome cluster 3 was represented by members of *Atopobium, Oribacterium, Bulleidia, Peptostreptococcus, Actinomyces, Veillonella*, and *Prevotella*. The 13 aforementioned genera distinguished among the 3 microbiome clusters were termed the cluster-distinguished oral microbiota (CDM). Kruskal–Wallis analysis was further performed to evaluate the bacteria with significant differences among the three microbiome clusters, and the top 20 genera according to their average relative abundance are presented in Fig. 2B. These findings highlight distinct microbial compositions across the three microbiome clusters, suggesting potential functional or ecological differences in the oral microbiota.

To evaluate the diversity and richness of microorganisms among three microbiome clusters, the alpha and beta diversity was further calculated. As shown in Fig. 2C and Supplementary Fig. 1F, the three microbiome clusters derived from DMM modeling had partially significant alpha diversity (overall $P = 0.086$ while microbiome cluster 1 vs. microbiome cluster 2, $P < 0.05$) and distinct beta diversity ($P = 0.001$).

To gain insights into the interactions between microbiome clusters from an ecological perspective, we further investigated the correlations between the 13 CDM genera based on Spearman correlation. The correlation strength across the populations was also assessed (Fig. 2D and Supplementary Table 3). Figure 2D depicts the interaction network among marker genera of the oral microbiota, with edge widths representing the strength of associations. The width of each edge is proportional to the correlation coefficient between the connected marker genera, where thicker edges signify stronger correlations and thinner edges indicate weaker ones. Notably, cluster-specific bacteria formed distinct mutualistic networks that were negatively correlated with each other. For example, the seven genera in cluster 3 exhibit strong interconnections. These central species likely play a crucial role in the network, as indicated by analyses of global efficiency and weighting (Supplementary Table 3).

Variations in the abundance of microbial metabolic genes at the species level

Having determined the importance and clinical relevance of oral microbiome dysregulation, we sought to better understand species-level and functional variability among the three microbiome clusters using additional metagenomic analysis alongside 16S rDNA sequencing. Metagenomic sequencing was performed using a subset of 16S rRNA amplicon sequencing cohort ($N = 91$) and yielded an average of 7.09×10^7 raw reads (10.63 G) per sample. After removing host reads, an average of 8.62×10^6 microbial reads were retained (proportion of retained reads: 0.95–75.39%), generating 1,485,607 nonredundant microbial genes from assembled contigs (Supplementary Table 4). Among them, 865,194 genes (58%) were annotated to the KEGG database, and 452,923 (30%) were annotated to KOs (Supplementary Table 5). The Shannon index suggested that the three microbiome clusters had significantly distinct alpha diversity (Fig. 3A), while partial but significant separation was also observed among the microbiome clusters for taxonomic beta diversity (Fig. 3B), revealing a notable separation among three microbiome clusters for taxonomic profiles. Based on metagenomic sequencing data, we further analyzed the relatively abundant taxonomy among three microbiome clusters. In brief, metagenomic reads were aligned to a non-redundant reference database, and the abundance of each genus was determined by normalizing the mapped reads for each genus against the total reads per sample. These normalized values were subsequently utilized to cluster samples according to their microbial composition (Supplementary Table 6). To better visualize the wide range of

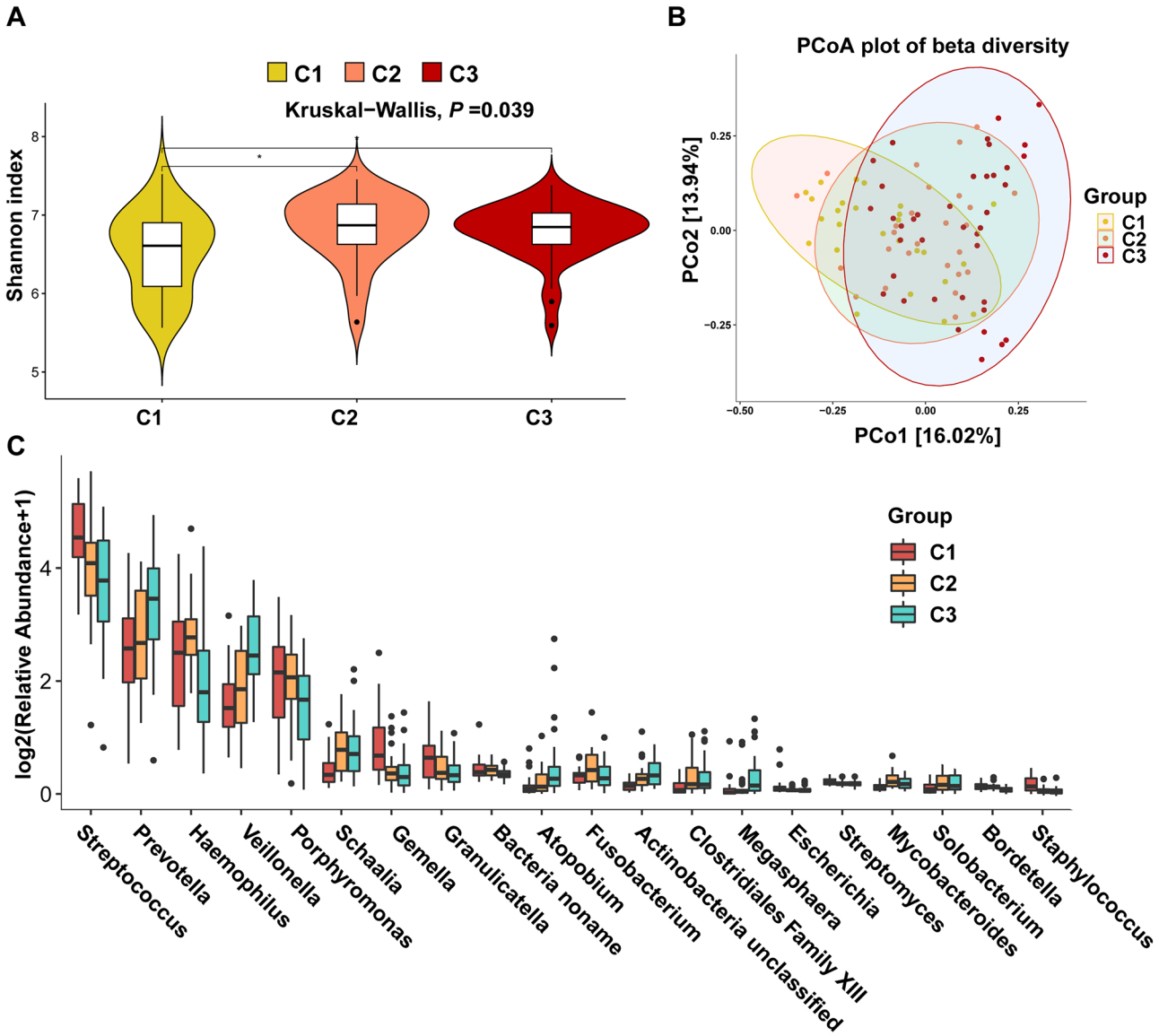


Fig. 3 | Overview of the metagenomic data for the three microbiome clusters. **A** Oral microbial alpha diversity among the three clusters was estimated with the Shannon index. The Kruskal–Wallis test and Wilcox test were performed to calculate statistical significance. **B** The beta diversity of the oral microbiota was assessed using principal coordinate analysis (PCoA) with the Bray–Curtis distance. The *P* value was calculated by the analysis of similarities (ANOSIM). Yellow, pink, and red plots represent the three microbiome clusters. **C** Relative abundance of the top 20 genera was evaluated with box plots (Kruskal–Wallis test). Red, orange, and green plots represent the three microbiome clusters.

data in bar charts, we applied a log2 transformation to the abundance data, as shown in Fig. 3C. Compared to 16S rDNA sequencing, metagenomic sequencing supported that most of the CDM genera were the dominant microbiota in corresponding microbiome clusters at the genus level (*Streptococcus* in microbiome cluster 1 and *Prevotella* in microbiome cluster 2, for example). By comparing the stacked bar plots of the 16S rRNA amplicon and metagenomic sequencing results, we found that the composition at the genus level of the two datasets was generally consistent (Supplementary Fig. 1F and Supplementary Fig. 2A). Thus, the correlation between the metagenomics and 16S rDNA gene sequencing data highlights the robustness of our results.

Based on the genus- and species-level metagenomic results, we then analyzed the microbial abundances in the various microbiome clusters (Supplementary Fig. 2A, B). Both genus- and species-level results reveal a notable prevalence of specific taxa, such as *Streptococcus* at the genus level and *Streptococcus mitis* at the species level, within certain clusters. This prevalence suggests that these taxa play key roles in shaping the community structure and may reflect important ecological interactions or functional significance within their respective clusters. The same methodology used in Fig. 3C was applied for species-level analysis, but with higher taxonomic resolution, as depicted in Fig. 4A. Reads were aligned to a comprehensive reference database to identify individual species, and species-level abundance was determined by normalizing the mapped reads for each species against the total reads per sample (Supplementary Table 7). Log2-transformed values were similarly employed for visualization purposes. As shown in Fig. 4A, the Kruskal–Wallis analysis at the species level revealed that the relative abundance of the top 20 bacteria differed significantly among the three microbiome clusters (C1–C3). Nineteen of the 20 species are members of one of the 13 CDM genera (Fig. 4A). *Streptococcus mitis* (Ranked No. 1), *Streptococcus pneumoniae* (Ranked No. 3), *Streptococcus pseudopneumoniae* (Ranked No. 9) and *Streptococcus infantis* (Ranked No. 14) were listed among the top 20 most abundant species in the metagenomic data. In summary, 16S

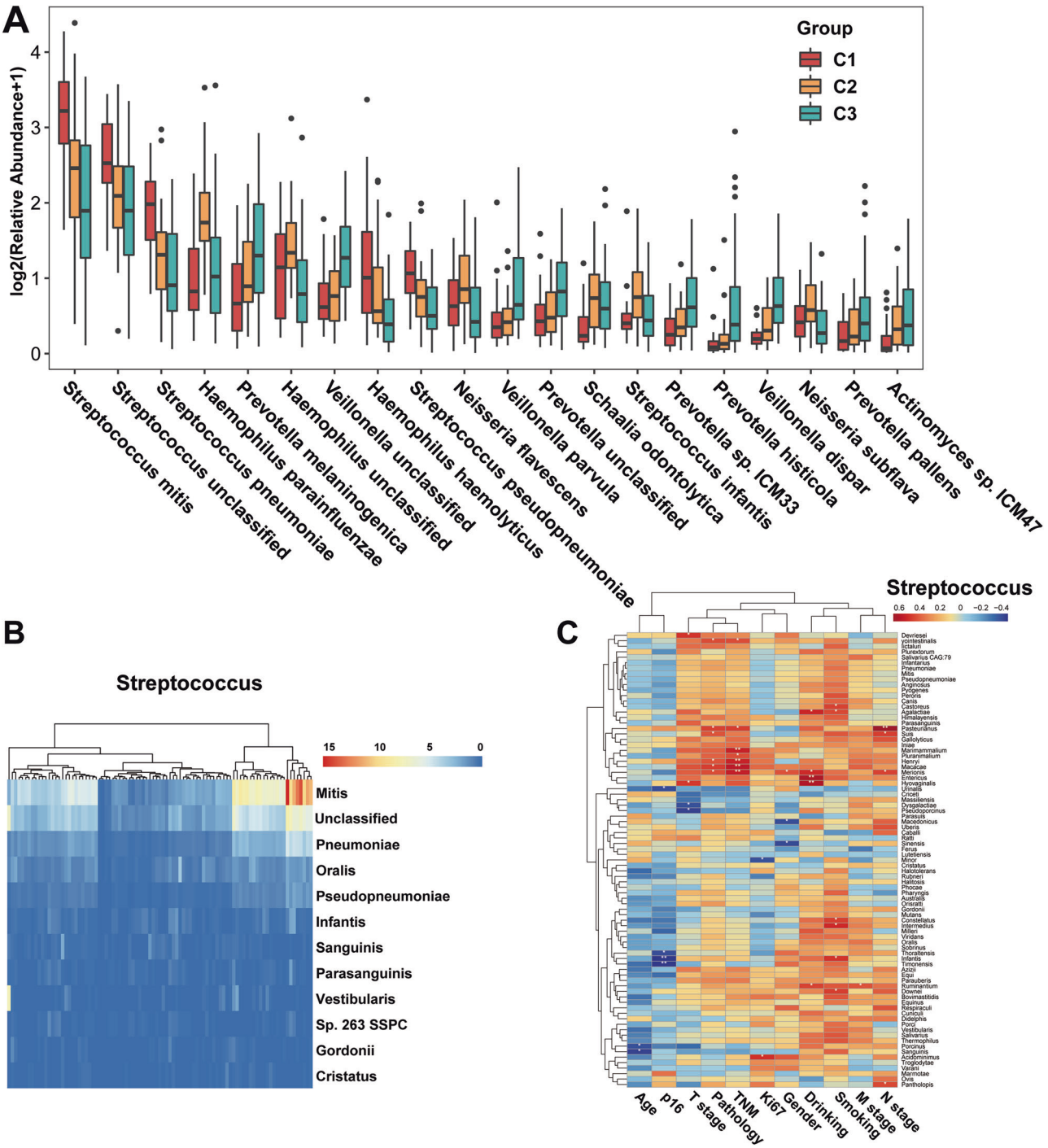


Fig. 4 | Metagenomic data were utilized to analyze the 3 microbiome clusters at the species level. A Relative abundance of the top 20 species was evaluated with box plots (Kruskal–Wallis test). Red, orange and green plots represent the three microbiome clusters. **B** Heatmap for the relative abundance of oral *Streptococcus* at the species level. **C** Heatmap for the correlation between the relative abundance of oral *Streptococcus* and clinical characteristics. Correlation was calculated by the Spearman correlation method. $P < 0.05$ marked as *, $P < 0.01$ marked as **. HNSCC subjects with complete clinical information were included in this analysis.

rDNA and metagenomic sequencing confirmed that the most prevalent genera in the oral samples from HNSCC patients are similar. These two distinct technologies provided reliable cross-validation for the most prevalent genus, namely, *Streptococcus*, in these samples. Since *Streptococcus* is abundant in oral samples, we further explored the species of this genus (Fig. 4B), revealing that *Streptococcus mitis* was the leading member of the *Streptococcus* genus. Heatmaps also illustrated that *Streptococcus unclassified* was the second most abundant species in multiple samples. This may imply some novel *Streptococcus* spp. with links to HNSCC (Fig. 4B).

To further explore the potential clinical applications, we performed the Spearman correlation test between the abundance of those *Streptococcus* species and the clinical features. Notably, we found differences in *Streptococcus* spp. that were significantly associated with multipole clinical features, including p16 protein and Ki67 staining (Fig. 4C).

In a recent study, p16 was strongly correlated with human papillomavirus (HPV) infection, and the positivity rate for HPV viral DNA testing of tumor tissue species with a positive p16 test was 90%[15]. HPV infection is one of the known causative factors of HNSCC. Further studies have shown that HPV-positive HNSCC patients have a better prognosis[15]. Specifically, *Streptococcus*

thoraltensis, *Streptococcus infantis*, and *Streptococcus timonensis* were significantly negatively correlated with p16 protein staining, suggesting that localized HPV infection may alter *Streptococcus* composition at the species level.

In general, Ki67 staining indicated whether the tumor cells were in an active state of proliferation. Since chemotherapy primarily inhibits the proliferation of tumor cells to achieve antitumor effects, chemotherapy efficacy is greater in patients with high Ki67 expression. *Streptococcus acidominimus* was significantly positively correlated with Ki67 expression, whereas *Streptococcus minor* was significantly negatively correlated with Ki67 expression. Therefore, it is possible that microbiome-host interactions have an effect on the proliferative activity of tumors based on our findings.

In addition, *Streptococcus henryi*, *Streptococcus macacae*, and *Streptococcus merionis* were significantly positively correlated with progressive HNSCC, whereas *Streptococcus dysgalactiae* and *Streptococcus pseudoporcinus* were significantly negatively correlated with HNSCC T stage. These enticing results indicate that the progression of HNSCC is accompanied by the evolution of microbiome clusters at both the genus and species levels. The causal relationship between different *Streptococcus* spp. and clinical pathological characteristics varied, indicating that different *Streptococcus* spp. may play distinct roles in the progression of HNSCC.

Functional analysis of metagenomic data revealed significant alterations in different microbiome clusters

To study the functional and metabolic changes in the oral microbiome clusters, STAMP was used to reveal the functional differences among the three microbiome clusters (Supplementary Table 8). As shown in Fig. 5A, 14 pathways were enriched in microbiome cluster 1, 2 pathways were enriched in microbiome cluster 2, and 4 pathways were enriched in microbiome cluster 3, suggesting that microbiome cluster 1 showed more activity than the other microbiome clusters according to the top 20 statistical KEGG items. To further analyze the functional differences, we conducted pairwise comparisons between different microbiome clusters. In the comparison between microbiome cluster 1 and microbiome cluster 2, both microbiome clusters occupied almost half of the enrichment pathways in the top 30 statistical KEGG items, and the top 5 significantly altered pathways were galactose metabolism, amino sugar and nucleotide sugar metabolism, glycolysis/gluconeogenesis, the phosphotransferase system (PTS), and oxidative phosphorylation (Fig. 5B). In the top 30 statistical KEGG items, microbiome cluster 1 showed more activity than microbiome cluster 3 according to the top 30 statistical KEGG items, and the top 5 microbiome cluster 1-upregulated KEGG items were ribosome, aminoacyl−tRNA biosynthesis, phosphotransferase system (PTS), ABC transporters, and amino sugar and nucleotide sugar metabolism Fig. 5C. In the comparison between microbiome cluster 2 and microbiome cluster 3, the pathways enriched in microbiome cluster 2 overwhelmingly dominated according to the top 30 statistical KEGG items, wherein 4 pathways, namely, aminoacyl−tRNA biosynthesis, carbon metabolism, ABC transporters, and ribosome, had the most significant upregulation. The above data indicated that the three microbiome clusters have unique bacterial metabolic features (Fig. 5D). By overlapping 4 STAMP analysis across the 3 microbiome clusters, we revealed that both the nitrogen metabolism pathway and PPP were detected in all comparisons and hence were further studied.

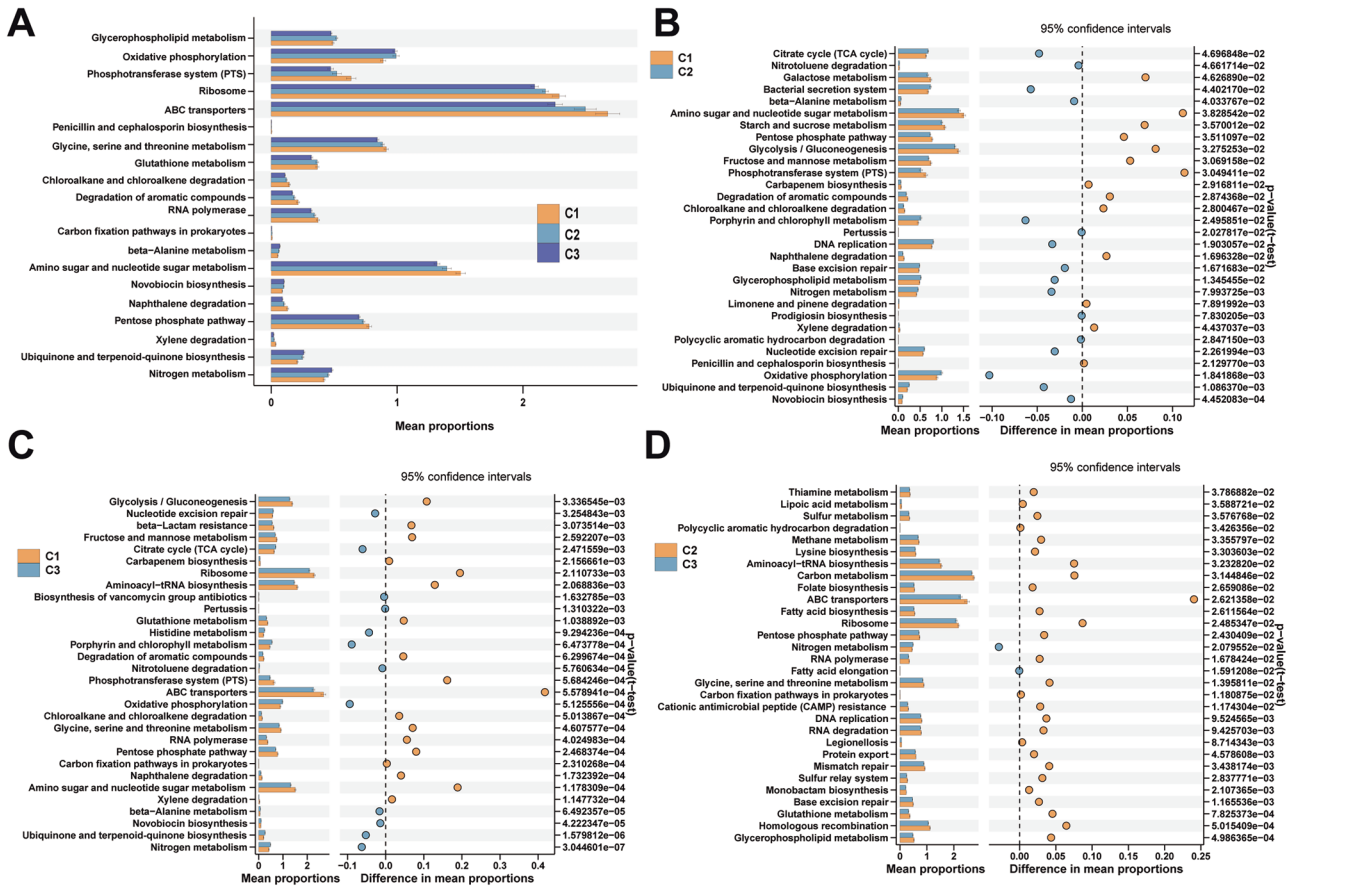


Fig. 5 | Differential KEGG pathways detected by STAMP among the 3 microbiome clusters. Significant differences between the 3 microbiome clusters (**A**), namely, microbiome cluster 1 and microbiome cluster 2 (**B**), microbiome cluster 1 and microbiome cluster 3 (**C**), and microbiome cluster 2 and microbiome cluster 3 (**D**), are presented. ANOVA and *t* tests were used to calculate the *P* values. The top 20 or 30 pathways with the lowest *P* values are shown in the figure. **A** and the left panels of (**B**–**D**) describe the relative abundance of enriched pathways, with different colors representing the relevant microbiome clusters. Bubbles in the right panels of (**B**–**D**) describe the difference in relative abundance between microbiome clusters. The color of the bubble is consistent with the microbiome cluster with higher abundance.

Contribution analysis revealed that the abundance changes in the oral microbiome mainly contributed to nitrogen metabolism and the PPP in different microbiome clusters

To further confirm the alterations in nitrogen metabolism and the PPP, we summed the relative abundances of relevant Unigenes and calculated the significance by the Kruskal−Wallis test. In microbiome cluster 1, nitrogen metabolism was significantly lower in relative abundance, while the PPP was significantly higher in relative abundance (Fig. 6A, B).

Next, we checked the contribution of each pathway to explore the correlation between changes in nitrogen metabolism and the PPP and changes in microbiome among the three clusters. As shown in Fig. 6C, D, stacked bar plots at the genus level showed that the top 3 genera associated with nitrogen metabolism were *Actinomyces*, *Veillonella*, and *Streptococcus*, while the top 3 genera associated with the PPP were *Streptococcus*, *Actinomyces*, and *Neisseria*. In microbiome cluster 1, *Streptococcus* contributed significantly more to nitrogen metabolism than *Actinomyces*, while in microbiome clusters 2 and 3, the opposite trend was observed. The contribution of *Streptococcus* to the PPP gradually decreased from microbiome cluster 1 to microbiome cluster 3. Taken together, the above analysis suggests that changes in the relative abundance of *Streptococcus* contribute significantly to changes in nitrogen metabolism and the PPP (Fig. 6C, D).

Moreover, we investigated the contribution of each species of *Streptococcus* to these two metabolic functions via an additional contribution analysis. *Streptococcus mitis*, *Streptococcus oralis*, and *Streptococcus pneumoniae* were the major contributors to nitrogen metabolism and the PPP, and the percentage of contribution of *Streptococcus mitis* to both pathways gradually decreased from microbiome cluster 1 to microbiome cluster 3. The above study suggests that changes in the species level of *Streptococcus* are one of the main reasons for the changes in nitrogen metabolism and the PPP among different taxa (Fig. 6E, F). In summary, changes in the relative abundance of oral bacteria at the genus or species level may affect the host by influencing nitrogen metabolism and the PPP, a mechanism closely related to carcinogenesis and progression.

Promising oral microbial classifiers based on metagenomic and 16S rRNA amplicon sequencing data were constructed to evaluate the predictive performance for HNSCC

Using microbiome data to construct predictive models for cancer diagnosis and prognosis has recently emerged as a promising strategy for enhancing cancer care. It promotes early detection, individualized treatment, and a better understanding of the microbiome's role in the development and progression of cancer. To develop prediction models for HNSCC, we employed a random forest classifier and trained it on the genus-level taxonomic profiles derived from our metagenomic data. Briefly, the genus-level metagenomic data were refined by including organisms with an abundance greater than 0.01 percent and present in at least 50 percent of samples. Among the models constructed in 5 replications with 10-fold cross-validation, the model with the highest area under the curve (AUC) values was chosen as the optimal model (0.8046) (Fig. 7A). The 8 genera that contributed to the prediction are listed in Fig. 7B and Supplementary Table 9. Most genera (5/8) significantly increased in oral samples from patients with benign lesions. In addition, the models based on the species profile of the oral microbiomes with an abundance greater than 0.01% and occurring in at least 80% of the samples were also evaluated. Interestingly, the classification prediction accuracy was 0.83 (Fig. 7C). The 12 species that significantly contributed to the prediction are listed in Fig. 7D and Supplementary Table 11. Among them, *Finegoldia magna* significantly increased in oral samples from HNSCC patients.

To further validate the accurate prediction of HNSCC using microbiome data, we analyzed the 16S rDNA gene in saliva samples. We established a module with core bacteria defined as those with an abundance of >0.01% in 20% of the samples, and we discovered that the models produced high AUCs, 0.8886 for the filtered-overall data at the genus level (Fig. 7E). There were 20 genera that contributed to the prediction (Fig. 7F), 16 genera significantly increased in oral samples from patients with benign lesions, while only 1 genus significantly increased in HNSCC ones. *Actinomyces*, *Bulleidia*, and *Peptostreptococcus*, which contributed significantly to the prediction (Supplementary Table 11), were consistent with those exhibiting significant changes in abundance. After selecting 13 CDM genera to build the prediction model, we discovered that the AUC value was only marginally diminished (0.7756, Supplementary Fig. 3 and Supplementary Table 12). This study demonstrates that it is possible to construct a prediction model for HNSCC using oral microbiome characteristics, whether based on metagenomic sequencing results or 16S rRNA amplicon sequencing results, and that the 13 oral CDM genera identified using 16S rRNA amplicon sequencing may have clinical significance.

Discussion

In this study, we conducted the most comprehensive microbial analysis in HNSCC to date. While sporadic 16S rDNA-based studies exist, to our knowledge, this marks the pioneering use of metagenomic-based microbial analysis in HNSCC research. Our study defined three distinct microbiome clusters and determine its associations with disease status, revealing the genus-level oral microbial alterations at different stages of HNSCC tumorigenesis. To comprehensively understand species-level characteristics and metabolic dysfunction, we performed additional metagenomic sequencing on 91 samples. Our study revealed that *Streptococcus spp*. was significantly correlated with clinical features, and had an impact on nitrogen metabolism pathways and the PPP pathway. Utilizing 16S rRNA amplicon sequencing and metagenomic data, four classifiers we constructed demonstrated to predict HNSCC with high accuracy, suggesting that the potential for screening HNSCC through the recognition of oral microbial dysbiosis. Our study is the first to apply both 16S rRNA amplicon and metagenomic sequencing strategies to a large number of samples, highlighting the significance of the oral microbial dysbiosis in HNSCC and investigating the HNSCC-related functional alterations in the oral microbiota.

The 5R-based 16S rRNA amplicon sequencing approach offers significant improvements over the commonly used V3–V4 amplicon sequencing method, particularly in coverage and resolution for bacterial species detection. By targeting five regions of the 16S rRNA gene (V2, V3, V5, V6, and V8) through multiplex PCR[6], the 5 R method captures ~68% of the full-length 16S sequence. This broader coverage provides a more comprehensive representation of microbial diversity and enables more accurate taxonomic classification, especially at the species level. Such enhanced resolution is crucial for identifying subtle dysbiotic shifts linked to disease states, including HNSCC. Furthermore, the method's ability to detect a wider range of bacterial taxa makes it particularly effective for profiling microbiome clusters in low-biomass samples, thereby strengthening microbiome research and biomarker discovery.

Accumulating evidence demonstrates that the oral microbiome is capable of ectopic colonization and produces an extraordinarily diverse array of microbial metabolites that have the potential to promote tumorigenesis by modulating pathways related to energy homeostasis, nutritional intake, and immunologic balance[16]. Due to the close anatomical location of the oral cavity to the pharynx and larynx and the continuous scouring effect of 1.5 L of saliva per day, the health and stability of the oral microbiome are of particular importance to the local microbial homeostasis in HNSCC patients. The association between dysbiosis of the oral microbiome and HNSCC has been confirmed epidemiologically[17].

Although pertinent research has been conducted since 2000[18], the relationship between *Streptococcus* and carcinogenesis in the oral cavity remains unclear. Using 16S rDNA sequencing data, our work revealed *Streptococcus* spp. as one of the primary groups of bacteria that differentiate mouth types. *Streptococcus* spp. were enriched in microbiome cluster 1 oral swab samples, which represent the majority of patients with advanced malignancies. *Streptococcus mitis*, *Streptococcus pneumoniae*, *Streptococcus pseudopneumoniae*, and *Streptococcus infantis* comprised four of the twenty most abundant species of microbiome, in accordance with the 16S sequencing results. *Streptococcal* spp. were significantly positively correlated

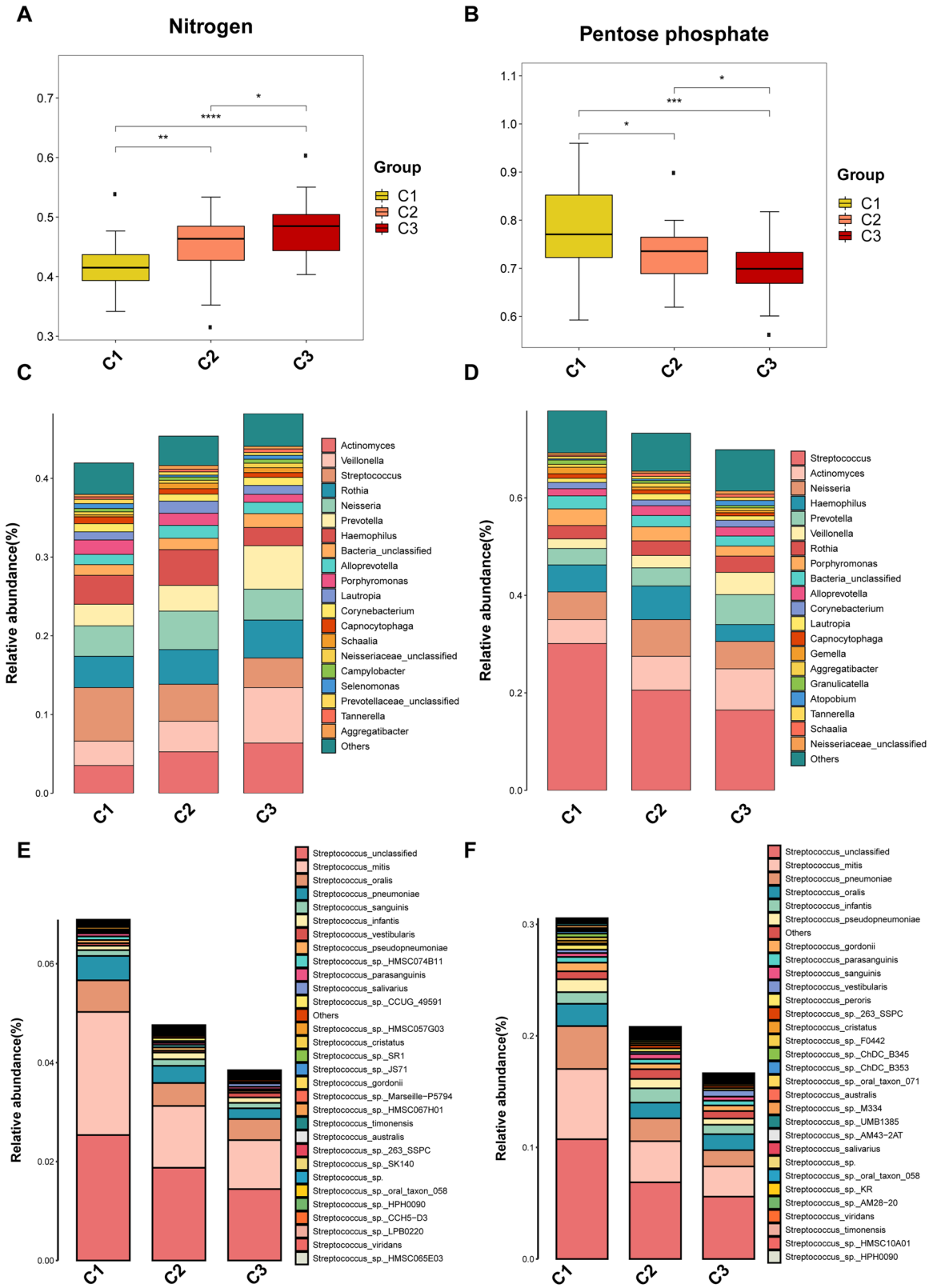
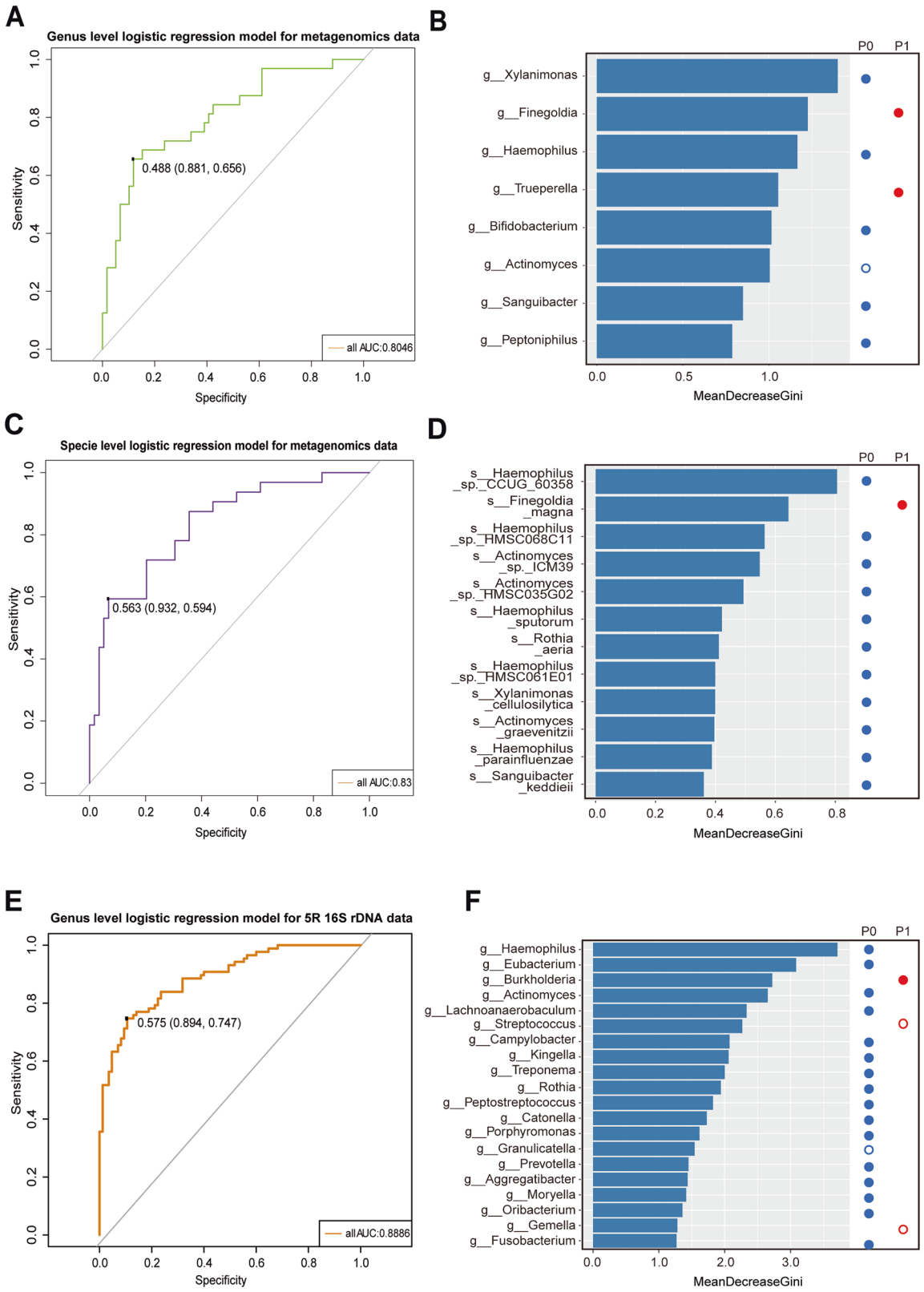


Fig. 6 | Metagenomic data revealed nitrogen metabolism and pentose phosphate pathway (PPP)-related metabolic differences among the 3 microbiome clusters. Kruskal–Wallis test-based analysis showing that the nitrogen metabolism pathway (**A**) and the PPP (**B**) were significantly distinguished among the 3 microbiome clusters. Contribution analysis of the top 20 abundant genera to nitrogen metabolism (**C**) and the PPP (**D**) among the three clusters. The stacked bar plots are presented in ascending order of the average abundance ranking of all samples. Stacked bar plots are used to display the relative abundance of *Streptococcal* spp. related to nitrogen metabolism (**E**) and the pentose phosphate pathway (**F**) among the 3 microbiome clusters. The legend (from top to bottom) and the graph (from bottom to top) are sorted based on the average abundance of all samples (from high to low). The sum of the abundance of the remaining Genera is labeled as "other", and do not participate in sorting. Contribution analysis was used to calculate the contribution of microbiota to both pathways.



with the majority of risk factors (tumor stage, Ki67, smoking, and alcohol consumption, etc.) and significantly negatively correlated with the prognostic protective factor, namely, the p16 protein, indicating that changes in oral streptococcal abundance play a driving role in HNSCC progression.

To determine whether functional differences exist between oral swabs, we performed metagenomic sequencing and functional analysis. Functional analysis of the oral microbiome revealed significant enrichment of the PPP and deletion of the nitrogen metabolism pathway in advanced HNSCC-related microbiome cluster 1. Intestinal nitrogen utilization has been proven to be necessary in maintaining carcinogenesis[19]. In multiple myeloma, an increased relative abundance of *Streptococcus* spp. in the gut changes systemic nitrogen recycling and accelerates cancer progression[20]. The PPP,

Fig. 7 | Promising oral microbial classifiers based on the oral microbiome data were constructed to evaluate the predictive performance for HNSCC. The predictive model was trained on the filtered genus (**A**, **B**) and species-level taxonomic profiles (**C**, **D**), respectively. **A** The ROC analysis of genus-level predictive model based on metagenomic data. **B** Gini index of the genera that contributed to the prediction. **C** The ROC analysis of species-level predictive model based on metagenomic data. **D** Gini index of the species that contributed to the prediction. **E**, **F** Prediction of HNSCC using random forest classifiers based on 16S rRNA amplicon sequencing data. **E** The ROC analysis of genus-level predictive model based on 16S rRNA amplicon sequencing data. **F** Gini index of the genera that contributed to the prediction. The cutoff value, true positive rate, and true negative rate are marked in the corresponding graphs and represented by the cutoff value (true positive rate, true negative rate). The right panel of the Gini coefficient bar plot showed significant differences of the markers that contributed to the corresponding prediction model. The red solid dot indicated a significant increase in oral samples from HNSCC patients, the blue solid circle indicated a significant increase in benign ones, and the hollow circle indicates no significant change between two groups. P0 = benign and precancerous groups, P1 = HNSCC. The Kruskal–Wallis test was used to calculate statistical significance.

which branches from glycolysis at the first committed step of glucose metabolism, is needed for the synthesis of ribose 5-phosphate (R5P) and NADPH. In cancer cells, the PPP plays a pivotal role in helping glycolytic cancer cells meet their anabolic demands and combat oxidative stress[21]. Abnormal regulation of the PPP in tumor cells is critical for managing the DNA damage response, metabolism, proliferation, and metastasis of cancer cells[21]. In the gut, microbes may participate in regulating PPP-related pathways in lung adenocarcinoma patients[22].

Our investigation showed substantial differences in nitrogen metabolism and PPP functional activity in the oral microbiome among three oral microbiome clusters. Further analysis revealed significant variations in nitrogen metabolism and PPP-related KOs among three microbiome clusters, and the resulting abnormalities in metabolite species and abundance may contribute to tumor growth. Nitrogen metabolism-related metabolites may have a more direct effect on HNSCC, given that oral microbiome metabolites can be carried directly to the major foci of HNSCC via saliva. *Streptococcus*, as the most abundant bacteria in the oral microbiome and the prototypical bacterium linked with nitrogen metabolism, has a considerable effect on metabolism, particularly nitrogen metabolism. Along with variations in nitrogen metabolism abundance, the percentage of *Streptococcus* (particularly *Streptococcus mitis*) was altered among three microbiome clusters, suggesting that *Streptococcus*-associated nitrogen metabolites and PPP metabolites may play a significant role in the evolution of HNSCC.

Due to the atypical early symptoms and the limitations of noninvasive detection techniques (for example, laryngoscopy has limitations such as high cost, poor patient experience, and floating accuracy depending on the operator experience), many HNSCC patients present with locally advanced disease, frequently with prominent involvement of lymph nodes. On the basis of previous findings, multiple oral microbiome signatures have been constructed with high accuracy for various cancers (such as pancreatic cancer, CRC, and oral cancer), and these signatures are considered a noninvasive detection method with potential clinical utility[5,23,24]. Based on 16S V3V4 sequencing data, numerous studies have developed highly accurate noninvasive diagnostic models[5,23–25]. Our study combined metagenomic sequencing and the 16S rRNA sequencing, to construct and characterize the different models at the genus or species level.

The AUC for constructing a prediction model using metagenomic genus-level data and species-level data was 0.8046 (95CI) and 0.83 (95CI), respectively, indicating that constructing a model using metagenomic data still requires sequencing results from larger sample sizes as well as a more appropriate prediction model to improve accuracy; thus, there is room for improvement. Using 16S rRNA sequencing data, we created a classifier of eight bacterial genera with a near-90% accuracy. In the future, 16S rRNA sequencing may be utilized as a cost-effective noninvasive diagnostic strategy for HNSCC screening due to its affordability and stability. In addition, genus-level models based on 13 CDM genera were similarly highly accurate, and their AUC was still close to 80%. This suggests that these 13 genera are clinically significant and that the causal relationship between changes in their oral relative abundance and the evolution of HNSCC needs to be further explored.

The inclusion or exclusion of the precancerous group in the analysis presents distinct considerations for model performance and clinical relevance. Combining benign and precancerous samples generally increases the sample size, enhancing the statistical power and robustness of the random forest model. This approach may capture a broader spectrum of microbial shifts during disease progression, potentially improving generalizability. However, excluding the precancerous group focuses the model solely on differentiating benign conditions from HNSCC, simplifying the classification task. In our analysis, while the removal of the precancerous group resulted in slightly lower ROC values overall, the performance based on key genera showed a marginal improvement, suggesting that certain microbial signatures may be more distinct in binary comparisons between benign and cancerous conditions. Despite these differences, the model's stability across both scenarios underscores the predictive potential of oral microbiomes for HNSCC detection and supports the utility of either approach depending on the clinical and research objectives (Supplementary Fig. 4A–D).

To clarify the potential applications of our predictive model and its relevance to non-invasive diagnosis, it is important to address the clinical context in which our findings are positioned. Although the data used in this study were collected from patients already diagnosed with HNSCC, this approach aligns with widely accepted strategies for biomarker discovery. The primary goal was to identify microbial and metabolic dysbiosis patterns associated with HNSCC that can serve as a foundation for diagnostic development.

Our predictive models, constructed using metagenomic and 16S rRNA sequencing data, demonstrated high performance with AUC values ranging from 0.78 to 0.89. These findings highlight the potential for using machine learning approaches to discern disease-associated microbiome profiles. While retrospective in nature, these models offer critical insights into microbiome clusters changes in HNSCC and form the groundwork for future studies that could focus on early disease detection.

The identification of distinctive microbiome clusters and species-level dysbiosis underscores the importance of microbial markers for advancing diagnostic capabilities. Furthermore, the metabolic dysfunction identified in the microbiome clusters suggests novel avenues for therapeutic intervention. Future prospective studies will be necessary to validate these biomarkers in pre-diagnostic settings and evaluate their utility for early detection, risk assessment, and monitoring disease progression.

By integrating microbiome and metabolic data, our study provides a robust theoretical basis for the continued development of non-invasive diagnostic tools and therapeutic strategies for HNSCC patients. These advancements hold promise for improving clinical outcomes through earlier diagnosis and targeted interventions.

The removal of contamination, particularly host-derived sequences, is a critical step in both 16S rRNA and metagenomic analyses to ensure the accuracy and reliability of microbial profiling. Contamination from the host genome or other exogenous sources can significantly skew results, leading to misinterpretation of microbial diversity and function. By rigorously aligning reads to the human reference genome and applying additional decontamination protocols, we ensure that the retained data reflects true microbial composition rather than artifacts of contamination. This is especially important in studies aiming to link microbial clusters to host health or disease, as even small amounts of contamination can obscure meaningful biological signals. Our approach, which removed an average of 91.38% of host reads in metagenomic data and identified 52 unique bacterial contaminants in 16S rRNA data, underscores the necessity of robust contamination removal strategies. These steps not only enhance the quality of

the dataset but also strengthen the validity of downstream analyses, such as taxonomic classification, functional profiling, and ecological inference. Therefore, contamination removal is not merely a technical formality but a foundational aspect of ensuring the integrity and interpretability of microbiome research.

Our study has a limitation lacking oral microbiota data from oral, oropharyngeal, and nasopharyngeal HNSCC, which was limited by the emergence of a large amount of clinical and mechanistic research. In our contaminant filtering process, we identified *Rothia mucilaginosa*, an oral commensal, as a potential contaminant due to its consistent detection in negative control samples. While this classification followed our predefined thresholding criteria, we recognize that the presence of *R. mucilaginosa* in controls may not solely indicate contamination, as it could also reflect true biological signals depending on the sample origin. Future studies could refine this step by integrating source-tracking models or adjusting abundance-based thresholds to better balance contaminant removal with the retention of legitimate community members.

In conclusion, our study combined 16S rRNA amplicon and metagenomic data for the first time to characterize three oral phenotypes significantly associated with HNSCC progression. We also analyzed the characteristic microbiome associated with these three oral phenotypes as well as the functional changes at the genus and species levels. In addition, we developed a highly accurate diagnostic classifier. This classifier could serve as a foundation for developing a reliable and accurate screening tool for HNSCC in clinical practice and for elucidating the roles of the microbiome in the etiology of HNSCC. In the future, further study is urgently needed to elucidate the contribution of oral microbial composition and function dysbiosis to the progression of HNSCC, and to explore whether identifying and correcting these dysbiosis can improve the diagnosis and management of HNSCC.

Methods

Patient enrollment

This study retrospectively analyzed 172 patients who underwent laryngeal and hypopharyngeal surgeries at Tianjin First Central Hospital from January 2022 to March 2023 (Table 2 and Supplementary Table 13). Clinical and pathological data were collected from medical records, with diagnoses confirmed by histopathological examination. Exclusion criteria included: (1) patients with immune/hereditary disorders, pregnancy, or coagulation dysfunction; (2) use of systemic antibiotics, immunosuppressive agents, or steroids within 12 weeks before screening; and (3) malignancies other than HNSCC (e.g., lymphoma or metastatic cancers). A total of 62 patients were excluded and the remaining 172 patients' samples were used in the following analysis.

Each patient underwent a physical examination by two physicians, with age, sex, and smoking history recorded. HNSCC patients received neck and chest imaging, and TNM staging was determined based on imaging, pathology, and biopsy findings. p16 and Ki67 staining results were obtained from pathology reports. All participants provided informed consent, and the study was approved by the local ethics committee (ID: 2023DZX46).

Sample collection

To assess oral microbiome changes during HNSCC tumorigenesis, we collected oral swabs for 16S rRNA amplicon and metagenomic sequencing. Patients fasted for at least 8 h before sampling, which occurred on the morning of surgery. Samples were analyzed only after pathological confirmation to ensure accurate disease classification.

A sterile, saline-moistened swab was used to collect microorganisms from the buccal mucosa, palate, and other oral sites. Six swabs per patient were immediately stored at −80°C. Blank collection tubes, left open for 30 s in the sampling area, served as negative controls. Preliminary testing confirmed sample quality and method viability. All procedures were conducted in a sterile operating room.

16S rRNA amplicon sequencing

To ensure sequencing accuracy, DNA was extracted from buccal swabs using the TIANGEN DP302-02 kit (CTAB method). For high-resolution microbial profiling, we performed 16S rRNA amplicon sequencing (5R 16S RNA method)[6], with sequencing conducted by Lc-Bio Technologies Co., Ltd (Hangzhou, China)[26]. Raw DNA reads were generated on a NovaSeq 6000 platform. Low-quality sequences (quality < 20, length < 100 bp) were removed using fqtrim (v0.94). Primer-aligned sequences from five amplified regions were used for analysis. The Short MUltiple Regions Framework (SMURF) method[13] enhanced microbiome profiling resolution by incorporating multiple 16S rDNA variable regions. As in prior studies[6], taxonomic classification relied on the Greengenes database (May 2013 version)[27]. To reduce noise, samples with <1000 reads and bacterial data with <0.1% relative abundance were excluded. Bacterial contamination was assessed based on frequency in negative controls. Bacteria found in >50% of control samples were deemed contaminants and filtered out. After decontamination, the remaining microbiome was analyzed (Supplementary Tables 14, 15).

Metagenomic sequencing

We performed 16S rRNA amplicon sequencing on 172 oral samples, with a subset ($N = 91$) subjected to metagenomic analysis. To ensure data reliability, we implemented rigorous quality control from DNA extraction to sequencing. DNA was extracted using the BIOTEKE AU46111-96 kit, quantified with a Qubit fluorescence quantifier (>200 ng), and verified by agarose gel electrophoresis. Library preparation was conducted using the TruSeq Nano DNA LT Library Kit (Illumina), followed by qPCR quantification. Libraries were qualified if the concentration exceeded 2 ng/μL, the main peak was ~350 bp, and no primer dimers were detected.

High-throughput sequencing was performed on a NovaSeq 6000 using paired-end PE150 mode. Raw sequencing data were preprocessed to remove low-quality sequences (quality value < 20, length <100 bp) using fqtrim (v0.94). Host sequences were filtered to improve microbial annotation accuracy. Clean reads were assembled using MEGAHIT (v1.2.9)[28], generating high-quality contigs (>500 bp), assessed using QUAST[29]. Quality metrics, including contig count, N50 values, and total assembly length, are detailed in Supplementary Table 16.

CDS prediction was performed on contigs (≥500 bp) using MetaGeneMark (v3.2.6). CD-HIT (v4.6.1) was used to remove redundancy (identity ≥ 95%, coverage ≥ 90%), selecting the longest sequences to construct a Unigene set (1,485,607 Unigenes)[30]. Clean reads were mapped to the Unigene reference using Bowtie2, filtering Unigenes with reads ≤2 across all samples.

Based on the Unigene reference sequence database, the clean read data of each sample were mapped to the reference using Bowtie2[31]. To provide a reliable Unigene list, we filtered out those Unigenes with reads of ≤2 in all samples. DIAMOND (0.9.14) was further used to map the Unigenes to the NR_meta database, which contains a total of 523,75954 microbial DNA sequences extracted from the NCBI non-redundant (NR) database[32]. The sequences were compared (BLASTP, *E* value ≤ 1e-5), and the best indicator for each Unigene comparison result was selected as the species classification. The final set of Unigenes was used for subsequent analysis, and their relative abundances were calculated based on the number of mapped reads and Unigene length using the following formula:

$$G_k = \frac{r_k}{L_k} * \frac{1}{\Sigma_{i=1}^{n} \frac{r_i}{L_i}} * 10^2$$

Where:

- G_k represents the relative abundance of Unigene k in a given sample.
- r_k is the number of reads mapped to Unigene k.
- L_k is the length of Unigene k.
- n is the total number of Unigenes in the sample.

This approach ensures an accurate estimation of Unigene abundance while minimizing redundancy and low-quality sequences.

Table 2 | Demographics and clinical characteristics of the study cohort

	Benign *N* = 56	Precancerous *N* = 29	Early-stage HNSCC *N* = 39	Advanced HNSCC *N* = 48
Age[a]	48 ± 12.93	54 ± 13.45	64 ± 8.70	63 ± 12.05
Sex				
Female	27	6	3	9
Male	29	23	36	39
Smoking status				
Nonsmokers	31	9	11	13
Smokers	22	15	22	28
Former smoker	2	5	6	6
Unknown or Not Applicable	1	0	0	1
Drinking status				
Nondrinkers	53	23	25	33
Drinkers	1	5	11	9
Former drinkers	1	1	3	5
Unknown or Not Applicable	1	0	0	1
T stage				
Tis	0	0	4	0
T1	0	0	15	6
T2	0	0	20	2
T3	0	0	0	21
T4	0	0	0	19
Unknown or Not Applicable	56	29	0	0
N stage				
N0	0	0	39	24
N1	0	0	0	11
N2	0	0	0	12
N3	0	0	0	1
Unknown or Not Applicable	56	29	0	0
M stage				
M0	0	0	39	47
M1	0	0	0	1
Unknown or Not Applicable	56	29	0	0
TNM stage				
0	0	0	4	0
I	0	0	15	0
II	0	0	20	0
III	0	0	0	26
IV	0	0	0	22
Unknown or Not Applicable	56	29	0	0
p16 IHC staining				
No staining	0	2	28	23
Faint	0	0	2	6
Moderate	0	1	6	9
Intense	0	0	1	2
Unknown or Not Applicable	56	26	1	8
Ki67 IHC staining				
Present	0	3	36	40
Unknown or Not Applicable	56	26	3	8

[a]Age is presented as mean ± SD.

Microbiome cluster identification based on the DMM and PAM algorithms

Core clusters of species-level data were screened, with a screening threshold for core clusters of species with an abundance greater than 0.1% in 50% of the samples. The clustering results were evaluated according to the DMM, and the optimal number of DMM classifications based on the combination of Laplace, AIC, and BIC indices was 3. The optimal number of PAM classifications based on the CHI index was 6 (Supplementary Table 1). We compared the classification results for the 6-classification scenario, and we found that the DMM model outperformed the PAM model in terms of discriminatory power; therefore, we concluded that the DMM model is more suitable for the data in this study. The classification results of the

DMM model were used for analysis. The DMM model was used to perform a cluster analysis to provide feedback on which cluster each sample belonged to, to obtain the percentage (contribution) of each sample in all clusters and to cluster the samples based on their contribution. Linear discriminant analysis (LDA) was performed to find the characteristic bacteria (microorganisms with an LDA of >3 and a P value of <0.05 were considered to be differentiated at the genus level, and those with uncertainty regarding species name were removed; the resulting samples with identified species names corresponded to 13 genera of bacteria).

For genus- and species- level data, box plots were used to compare the relative abundances of the top 20 genera across clusters. Statistical significance was assessed using the Kruskal–Wallis test. Due to the huge data deviation of the relative abundance in both genus- and species- level data (Supplementary Tables 6, 7), we do the log2 transformation for the visualization purpose.

Diversity analysis, statistical testing, and clinical correlation in microbiome and metagenomic data

For 16S rRNA amplicon sequencing, random sampling generated sparse curves to assess sequencing data validity and species abundance/diversity across samples. Species-stratified abundance was obtained for further analysis. Alpha diversity (Chao1 index) was analyzed at the species level using the Wilcoxon test (two-group comparisons) and Kruskal–Wallis test (three-group comparisons). Beta diversity, based on Bray–Curtis distances, was assessed by PCoA, with significance tested using ANOSIM and ANOVA. The Kruskal–Wallis test identified genera with significant abundance differences. The biological relevance of 13 characterized genera was determined using Spearman correlation coefficients, adjusting for covariates (age, sex, smoking, drinking) via a generalized linear model ($P < 0.05$, $r > 0.3$ or < -0.3).

Clinical differences among microbiome clusters were analyzed using chi-square, Fisher's exact, and ANOVA tests. For metagenomic data, alpha diversity was assessed using the Shannon index (QIIME 1.8.0), and significance was determined via the Kruskal–Wallis test. Beta diversity was also calculated using the Shannon index. Additionally, the Kruskal–Wallis test identified significant genus- and species-level abundance differences among clusters. Correlations between microbiome clusters, microbial species, and clinical variables were evaluated using Spearman correlation coefficients.

Functional annotation of Unigenes

We conducted functional analysis of the final Unigenes to explore metabolic features in our cohort. KEGG pathway annotation was performed using DIAMOND (v0.9.14)[32]. One-way ANOVA (STAMP) identified significant functional differences, with P values obtained via Student's t test and ANOVA[33].

For nitrogen and pentose phosphate metabolism, we assessed the relative abundance of associated Unigenes, comparing functional activity between groups using Student's t test. The top 20 most abundant KEGG orthologs (KOs) were visualized in R (3.4.4). Spearman correlation was used to analyze KO associations with clinicopathological features.

To infer microbial contributions to the tumor environment, we performed contribution analysis, assessing taxa's relative influence on microbiome structure and function via abundance and functional pathway analysis. Unigenes were annotated through the NR and KEGG databases, linking microbial taxa to metabolic functions. By summing relevant Unigene abundances, we determined each taxon's contribution to nitrogen and pentose phosphate metabolism. Streptococcal spp.'s role in these pathways was visualized using stacked bar plots.

Construction of a random forest-based diagnostic classifier

All microbiota data were used to establish classifiers with the RandomForest R package[26,34]. To optimize the best predictive model for diagnosing HNSCC, the metagenomic data and 16S rRNA amplicon sequencing data were used as input. For metagenomic data, species-level taxa were retained if their abundance exceeded 0.01% in ≥80% of samples, and genus-level taxa if in ≥50%. For 16S rRNA data, genera with abundances >0.01% in ≥20% of samples were included. Key taxa differentiating HNSCC from benign and precancerous groups were integrated into model construction. The dataset was randomly split (70% training, 30% testing) and trained using 5-times repeated 10-fold cross-validation. The classifier's performance was assessed via the area under the receiver operating curve (AUC) using the pROC package[35].

Data availability

All the data are free to use for academic purposes and available from the China National Center for Bioinformation (CRA013025 and CRA013024).

Code availability

All software used, with versions and non-default parameters, is described precisely and referenced in the method and technical validation section to ensure easy access and reproducibility.

Received: 4 September 2024; Accepted: 20 April 2025;
Published online: 07 May 2025

References

1. Johnson, D. E. et al. Head and neck squamous cell carcinoma. *Nat. Rev. Dis. Prim.* **6**, 92 (2020).
2. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **73**, 17–48 (2023).
3. Ferlay, J. et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. cancer* **144**, 1941–1953 (2019).
4. Mody, M., Rocco, J., Yom, S., Haddad, R. & Saba, N. J. L. Head and neck cancer. *PubMed* **398**, 2289–2299 (2021).
5. Sepich-Poore, G. D. et al. The microbiome and human cancer. *Science* **371**, https://doi.org/10.1126/science.abc4552 (2021).
6. Nejman, D. et al. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368**, 973–980 (2020).
7. Cai, J., Sun, L. & Gonzalez, F. J. Gut microbiota-derived bile acids in intestinal immunity, inflammation, and tumorigenesis. *Cell host microbe* **30**, 289–300 (2022).
8. Karmakar, S., Kar, A., Thakur, S. & Rao, V. U. S. Periodontitis and oral cancer-a striking link. *Oral. Oncol.* **106**, 104630 (2020).
9. Hayes, R. B. et al. Association of oral microbiome with risk for incident head and neck squamous cell cancer. *JAMA Oncol.* **4**, 358–365 (2018).
10. Frank, D. N. et al. A dysbiotic microbiome promotes head and neck squamous cell carcinoma. *Oncogene* **41**, 1269–1280 (2022).
11. Su, S. C. et al. Oral microbial dysbiosis and its performance in predicting oral cancer. *Carcinogenesis* **42**, 127–135 (2021).
12. Mäkinen, A. I. et al. Salivary microbiome profiles of oral cancer patients analyzed before and after treatment. *Microbiome* **11**, 171 (2023).
13. Fuks, G. et al. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* **6**, 17 (2018).
14. Shah, M. S. et al. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut* **67**, 882–891 (2018).
15. Mehanna, H. et al. Prognostic implications of p16 and HPV discordance in oropharyngeal cancer (HNCIG-EPIC-OPC): a multicentre, multinational, individual patient data analysis. *Lancet Oncol.* **24**, 239–251 (2023).
16. Zhang, S. et al. Human oral microbiome dysbiosis as a novel non-invasive biomarker in detection of colorectal cancer. *Theranostics* **10**, 11595–11606 (2020).
17. Xu, S., Zhang, G., Xia, C. & Tan, Y. H. Associations between poor oral health and risk of squamous cell carcinoma of the head and neck: a meta-analysis of observational studies. *J. Oral. Maxillofac. Surg.***77**, 2128–2142 (2019).
18. Tateda, M. et al. Streptococcus anginosus in head and neck squamous cell carcinoma: implication in carcinogenesis. *Int. J. Mol. Med.* **6**, 699–703 (2000).

19. Chen, H. et al. Urea cycle activation triggered by host-microbiota maladaptation driving colorectal tumorigenesis. *Cell Metab.* **35**, 651–66.e7 (2023).
20. Jian, X. et al. Alterations of gut microbiome accelerate multiple myeloma progression by increasing the relative abundances of nitrogen-recycling bacteria. *Microbiome* **8**, 74 (2020).
21. Patra, K. C. & Hay, N. The pentose phosphate pathway and cancer. *Trends Biochem. Sci.* **39**, 347–354 (2014).
22. Wang, S. et al. Gut microbiome was highly related to the regulation of metabolism in lung adenocarcinoma patients. *Front. Oncol.* **12**, 790467 (2022).
23. Nagata, N. et al. Metagenomic identification of microbial signatures predicting pancreatic cancer from a multinational study. *Gastroenterology* **163**, 222–238 (2022).
24. Flemer, B. et al. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**, 1454–1463 (2018).
25. Kartal, E. et al. A faecal microbiota signature with high specificity for pancreatic cancer. *Gut* **71**, 1359–1372 (2022).
26. Yang, J. et al. A distinct microbiota signature precedes the clinical diagnosis of hepatocellular carcinoma. *Gut Microbes* **15**, 2201159 (2023).
27. DeSantis, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
28. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics***31**, 1674–1676 (2015).
29. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
30. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics***28**, 3150–3152 (2012).
31. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. methods* **9**, 357–359 (2012).
32. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. methods* **12**, 59–60 (2015).
33. Parks, D. H., Tyson, G. W., Hugenholtz, P. & Beiko, R. G. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics***30**, 3123–3124 (2014).
34. Ai, D. et al. Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer. *Genes* **10**, https://doi.org/10.3390/genes10020112 (2019).
35. Kong, C. et al. Integrated metagenomic and metabolomic analysis reveals distinct gut-microbiome-derived phenotypes in early-onset colorectal cancer. *Gut* **72**, 1129–1142 (2023).

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant Nos. 82403470, 82401333), the Tianjin Key Medical Discipline Construction Project (TJYXZDXK-046A, from Tianjin Municipal Health Commission), Tianjin Health Research Project (TJSJMYXYC-D2-021, TJWJ2023XK013, from Tianjin Municipal Health Commission), and Tianjin Municipal Science and Technology Project (21JCYBJC01570, 24JCQNJC01170, from Tianjin Municipal Science and Technology Committee). The funders did not play any role in the research design, data collection, analysis, and interpretation, or manuscript writing. We thank Wenjuan Sun and Jianxin Li from Hangzhou LC-BIO Co., Ltd, for the technical assistance and help with the data analysis.

Author contributions

W.W. and W.X. designed the project. Z.J. collected data and performed the analysis. All authors wrote the manuscript draft, and Z.J. finalized the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41522-025-00708-8.

Correspondence and requests for materials should be addressed to Li Li, Xianfeng Wei or Wei Wang.