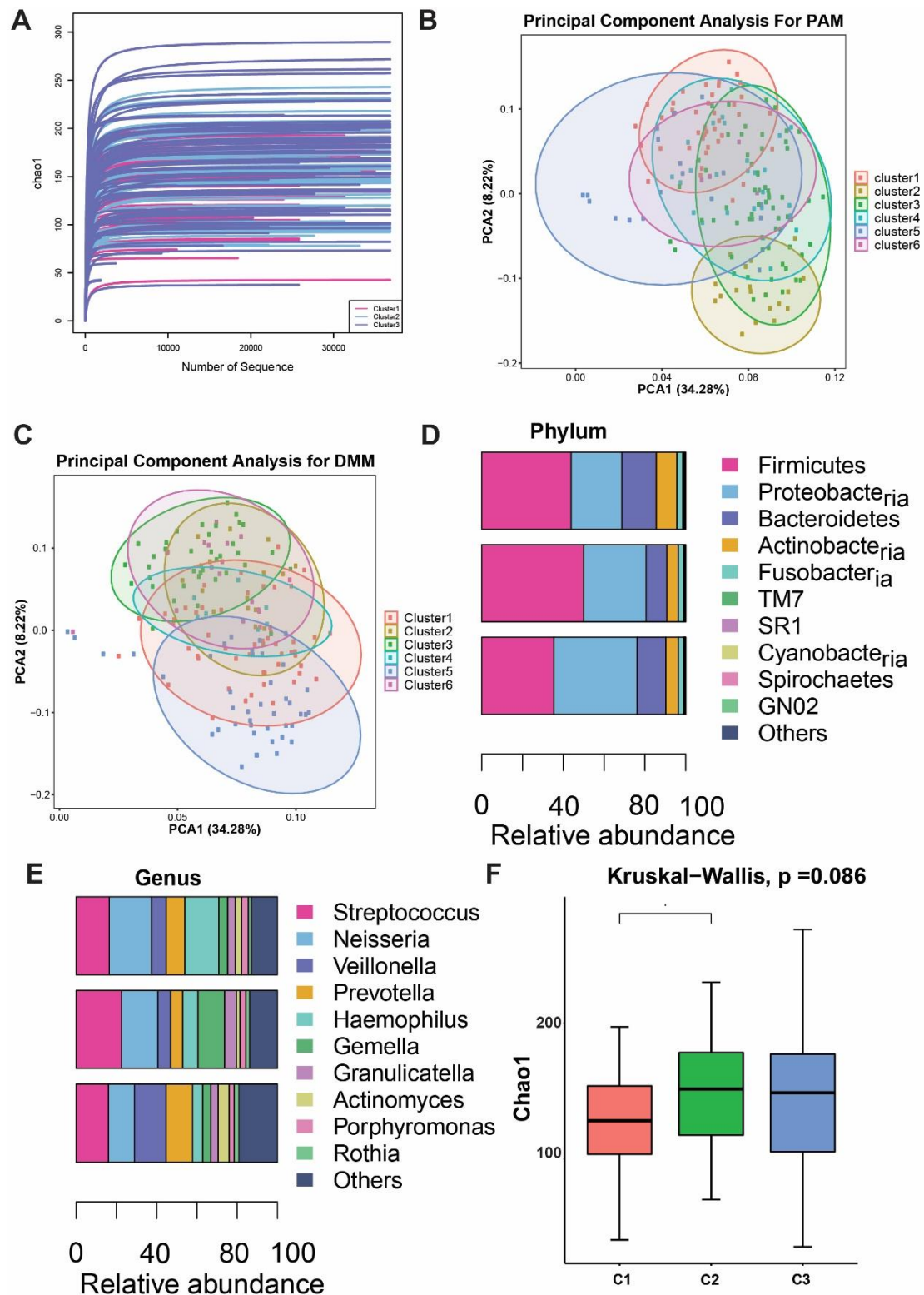
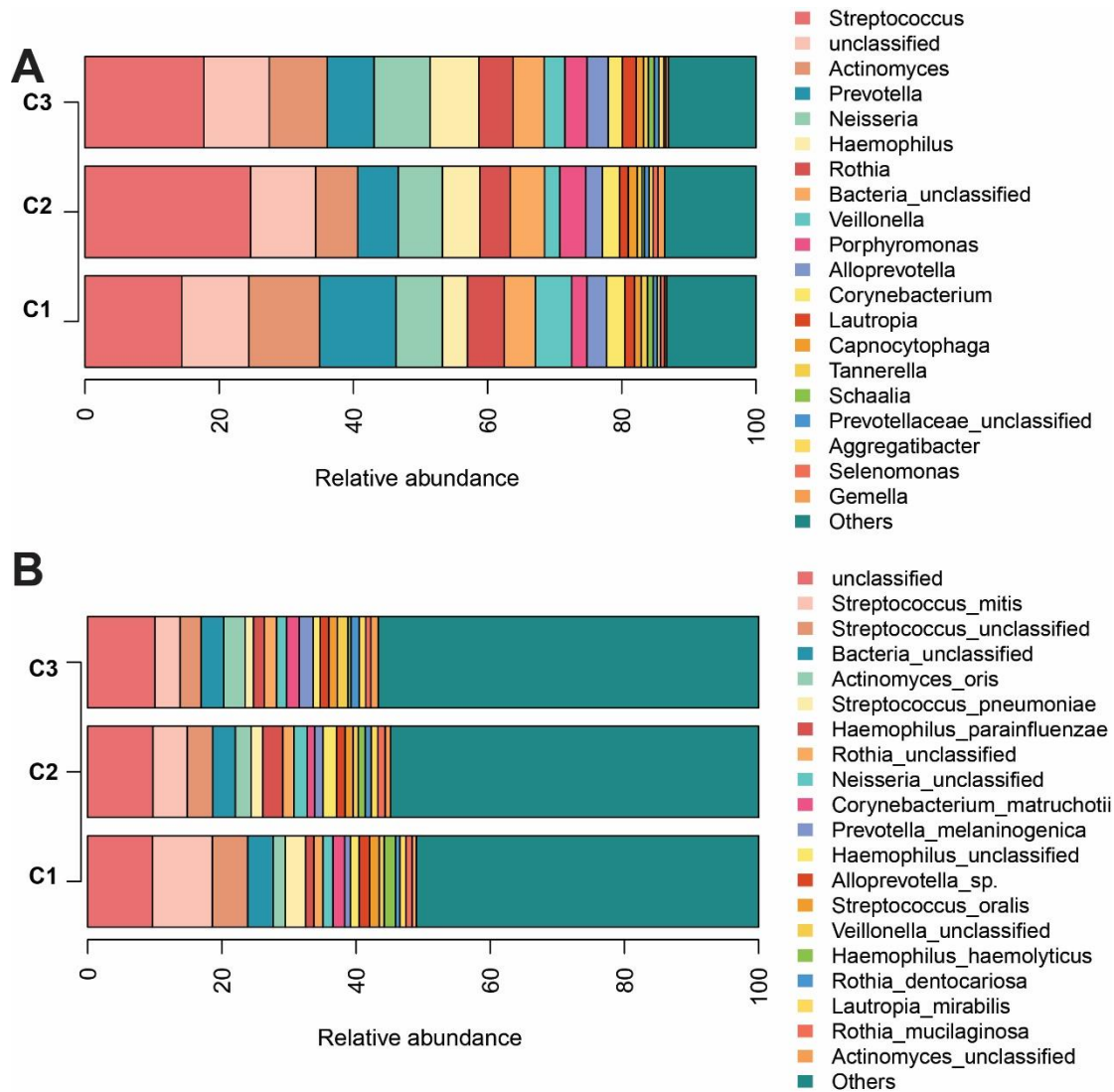


Supplementary Figures

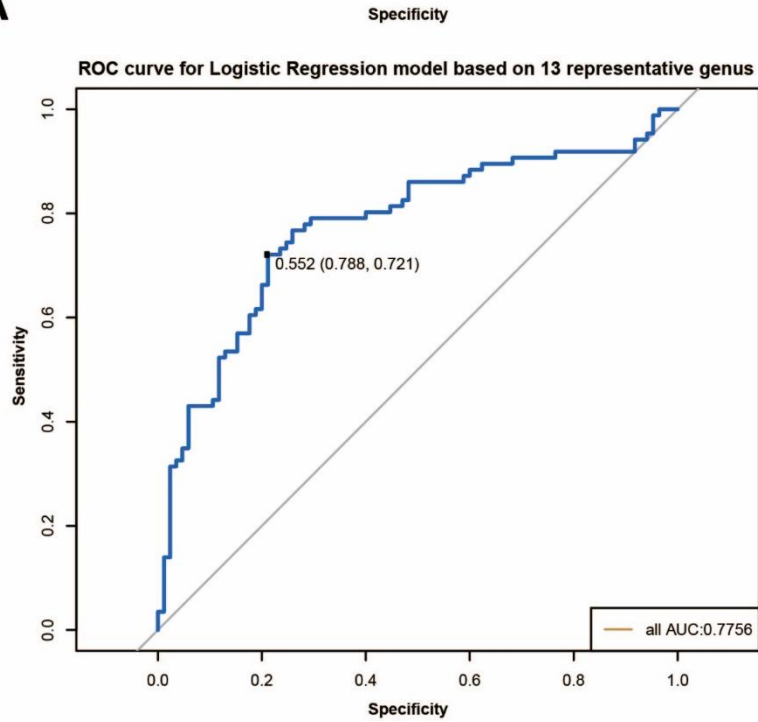


Supplementary Figure 1. Characterization of 16S rDNA gene sequencing for oral microbiome clusters in HNSCC carcinogenesis. (A) Rarefaction curves directly reflect the rationality of sequencing data. (B-C) Comparison of DMM- (B) and PAM-based (C) detection of microbiome clusters using PCA. (D-E) Stacked bar plot of the

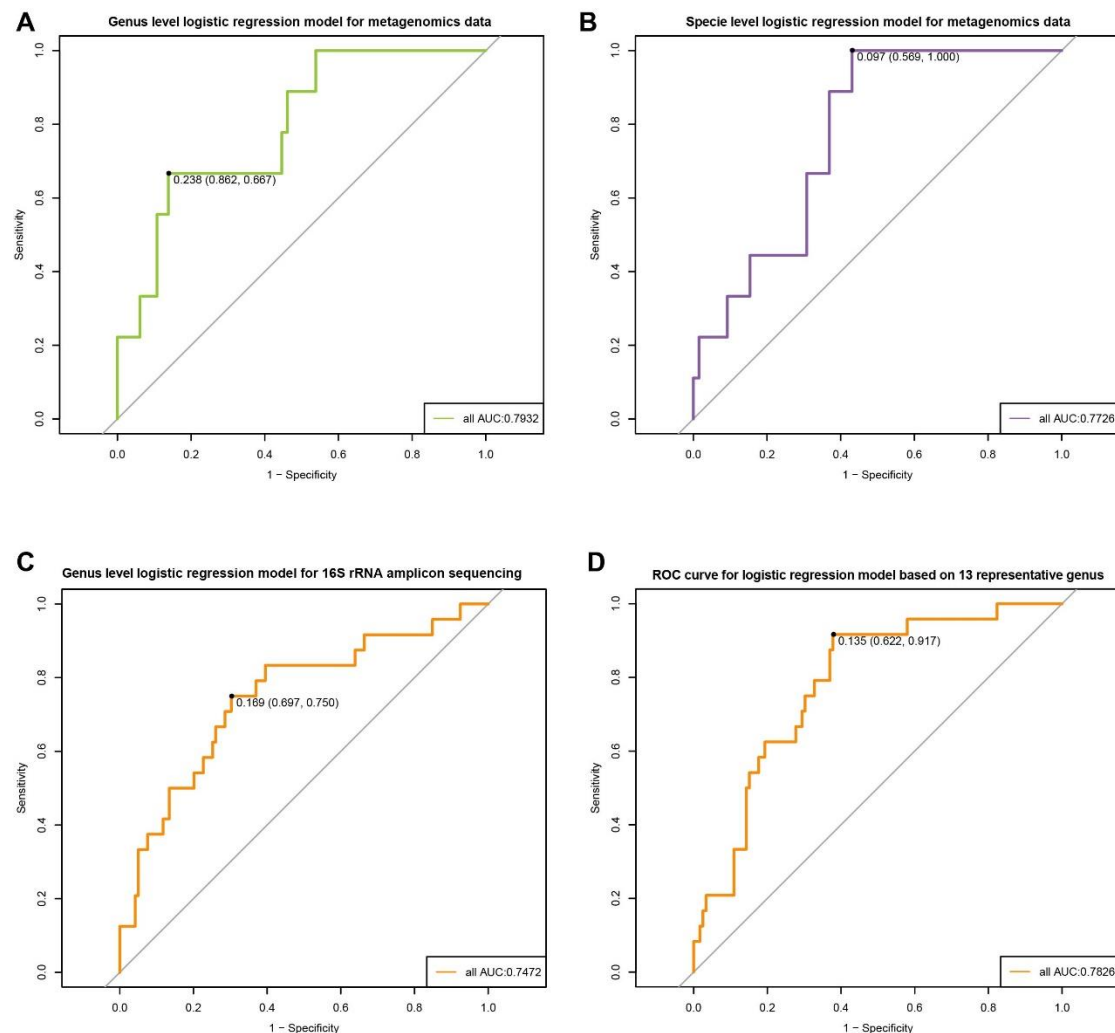
mean relative abundance of the top 10 oral phyla (E) and genera (F) among the three clusters. (F) Oral microbial alpha diversity among the three clusters was estimated by the Chao 1 index. The Kruskal–Wallis test was used to calculate statistical significance.



Supplementary Figure 2. Relative Abundance of Taxa Across Microbiome Clusters. (A) The genus-level relative abundance of the oral microbiota across three microbiome clusters (C1, C2, and C3). The x-axis represents the identified genera, and the y-axis indicates their relative abundance as a percentage of the total microbiome clusters composition. The data show that certain genera, such as *Streptococcus* and *Actinomyces*, exhibit substantial dominance in specific clusters. (B) The species-level relative abundance of the oral microbiota across the same microbiome clusters (C1, C2, and C3). The x-axis displays the identified species, while the y-axis represents their relative abundance. Similar to the genus-level trends, species such as *Streptococcus mitis* and *Haemophilus parainfluenzae* demonstrate significant dominance within their respective microbiome clusters.

A

Supplementary Figure 3. Prediction of HNSCC based on 13 CDM genera. 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).



Supplementary Figure 4: Oral microbial classifiers constructed based on oral microbiome data were evaluated for predictive performance for HNSCC. (A) ROC analysis of the genus-level predictive model based on metagenomic data. (B) ROC analysis of the species-level predictive model based on metagenomic data. (C) Prediction of HNSCC using random forest classifiers based on 16S rRNA amplicon sequencing data. (D) Prediction of HNSCC using random forest classifiers based on 13 CDM genera. All models were trained using benign and HNSCC samples exclusively.

Supplementary tables:

Supplementary Table 1. Robustness of the PAM- and DMM-based approaches.
 Supplementary Table 2. The summary of linear discriminant analysis scores.
 Supplementary Table 3. The Spearman correlation coefficient of 13 clusterdistinguished oral microbiota (CDM).

Supplementary Table 4. Detailed information on the host genome contamination rates for each metagenomics sequenced.

Supplementary Table 5. KO abundance data of metagenomic sequencing.

Supplementary Table 6. Taxonomy abundance data of metagenomic sequencing based on metagenomic genus data.

Supplementary Table 7. Taxonomy abundance data of metagenomic sequencing based on metagenomic species data.

Supplementary Table 8. The differential KEGG pathways detected by STAMP among 3 microbiome clusters.

Supplementary Table 9. Eight markers involved in HNSCC classifier construction based on metagenomic genus data.

Supplementary Table 10. Twelve markers involved in HNSCC classifier construction based on metagenomic species data.

Supplementary Table 11. Twenty markers involved in HNSCC classifier construction based on 16S rRNA amplicon

Supplementary Table 12. Thirteen CDM involved in HNSCC classifier construction based on 16S rRNA amplicon sequencing.

Supplementary Table 13. Demographics and clinical feature of the study.

Supplementary Table 14. Microbial abundance data of 16S rRNA amplicon sequencing.

Supplementary Table 15. A comprehensive list of 52 unique bacterial contaminants.

Supplementary Table 16. Detailed assembly quality metrics for the 91 samples that underwent metagenomic sequencing.