

Additional file 1

Unraveling the Diagnostic and Prognostic Signatures of Oral Microbiota in Head and Neck Cancer

Hojun Sung^{1,2#}, Dong-Wook Hyun^{1,2#}, Tae Woong Whon³, Pil Soo Kim^{1,2}, Hyun Sik Kim^{1,2}, Jae-Yun Lee¹, So-Yeon Lee², Jee-Won Choi², Ji-Ho Yoo^{2,4}, Mi-Ja Jung³, Ji-Hyun Yun¹, June-Young Lee^{1,2}, Euon Jung Tak¹, Yun-Seok Jeong², Seung Woo Kim⁵, Moonseung Baeg⁵, Young-Gyu Eun⁶, Young Chan Lee^{6*} and Jin-Woo Bae^{1,2*}

¹Department of Life and Nanopharmaceutical Sciences, Kyung Hee University, Seoul, Republic of Korea.

²Department of Biology, Kyung Hee University, Seoul, Republic of Korea.

³Kimchi Functionality Research Group, World Institute of Kimchi, Gwangju, Republic of Korea.

⁴Department of Biomedical and Pharmaceutical Sciences, Kyung Hee University, Seoul, Republic of Korea.

⁵Department of Otolaryngology-Head and Neck Surgery, Veterans Health Service Medical Center, Seoul, Republic of Korea

⁶Department of Otolaryngology-Head and Neck Surgery, School of Medicine, Kyung Hee University, #892 Dongnamro, Gangdong-gu, Seoul 05278, Republic of Korea.

#These authors contributed equally to this study: Hojun Sung, Dong-Wook Hyun.

*Correspondence: Jin-Woo Bae (baejw@khu.ac.kr), Young Chan Lee (medchan@khu.ac.kr)

Supplementary Information includes:

- Detailed information of methods and materials
- Fig. S1 to S5

Supplementary Information

- Detailed information of methods and materials

Differential analysis of microbial abundance

To identify taxa that show significant differences between groups, we used ANCOM-BC [1] and performed it using the “ancombc2” function of the R package ANCOMBC. The processed microbial abundance table was input in the raw count state without any normalization and was transformed through the clr-transformation provided by the “ancombc2” function for further analyses. Structural zero, which refers to taxa found only in specific groups but not in their counterpart groups, was ignored. For taxa at the phylum level, only those with a minimum prevalence of 25% across all samples were targeted, while at the genus and species level, only those with a prevalence of at least 20% were targeted to avoid false positives caused by minor taxa. The Benjamini-Hochberg's (BH) method was used for the false discovery rate (FDR) correction, and taxa showing statistical significance with an adjusted p value (q value) of less than 0.05 were presented. To ensure consistent interpretation, differential analyses were conducted at both the genus and species levels. However, taxa presented in the main figures (e.g., Fig. 2B) were aggregated at the genus level for consistency across visualizations, whereas species-level differential results are summarized in Supplementary Table 1.

Correlation and network analysis

Potential microbial interactions were investigated using SECOM [2]. This tool, proficient in determining both linear and nonlinear correlations between microbial taxa in microbiome data, enabled the identification of significant correlations. The genera present in at least 50% of total samples from both HNSCC patients and controls were selected for analysis. Both distance correlation [3] and Spearman correlation [4] were employed, ensuring the inclusion of only statistically validated relationships. Relationships characterized by a zero distance correlation coefficient and a non-zero Spearman correlation coefficient were excluded. This approach aids in establishing a robust network, mitigating the incorporation of false positive correlations. Significant correlations among the selected taxa were visualized as a heatmap plot using the R package pheatmap (version 1.0.12) [5]. Hierarchical clustering was conducted using the “complete” method in the R package pheatmap. Network analyses based on the significant relationships were conducted and visualized using Cytoscape software [6], where each node denotes a distinct taxon and edges denote significant relationships between them. Edge colors, red for positive and blue for negative correlations, convey the directional tendencies of these relationships. The network structures were then analyzed, focusing on the number of edges, betweenness centrality [7], and average clustering coefficient associated with each node [8].

1. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. *Nat Commun.* 2020;11(1):3514.
2. Lin H, Eggesbo M, Peddada SD. Linear and nonlinear correlation estimators unveil undescribed taxa interactions in microbiome data. *Nat Commun.* 2022;13(1):4946.
3. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Stat.* 2007;35(6):2769-94.
4. Spearman C. The proof and measurement of association between two things. *Am J Psychol.* 1904;15(1):72–101.
5. Kolde R. pheatmap: Pretty Heatmaps. 2019. <https://CRAN.R-project.org/package=pheatmap>.
6. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-504.
7. Freeman LC. A Set of Measures of Centrality Based on Betweenness. *Sociometry.* 1977;40(1):35–41.1. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature.* 1998;393(6684):440-2.
8. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature.* 1998;393(6684):440-2.

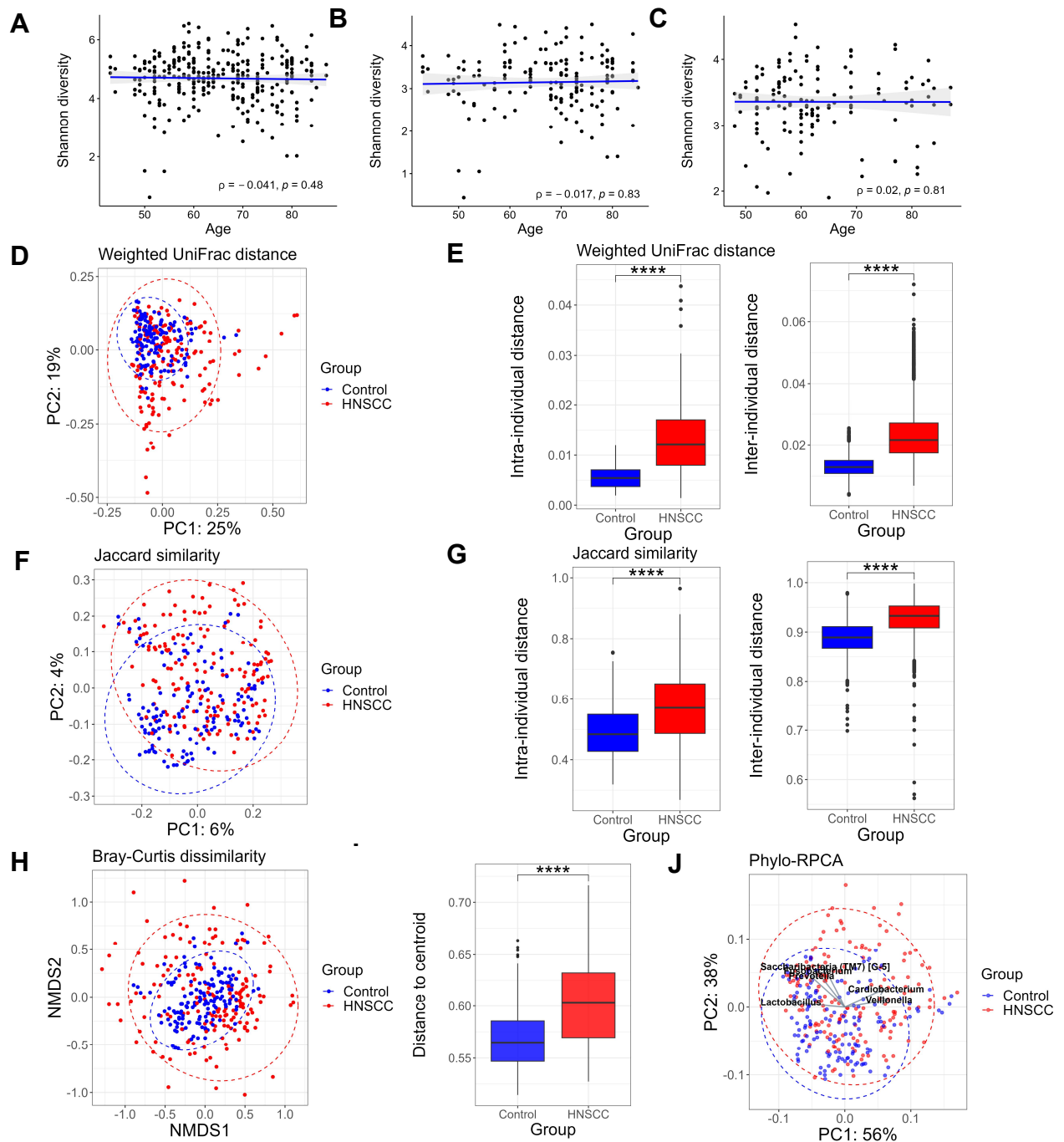


Fig. S1 Age exerts no influence on alpha diversity, and oral microbiota differ across metrics between HNSCC and controls, marked by high inter-individuality in HNSCC. **A-C** Correlation between demographic factor of age and Shannon diversity values among all samples (**A**, $n = 305$), within the HNSCC patients (**B**, $n = 164$) and the cancer-free control group (**C**, $n = 141$). Spearman's rank correlation coefficient and statistical p value were presented. **D, F** Comparison of beta diversity presented as PCoA using weighted UniFrac (**D**, $R^2=0.033$, $p < 0.001$) and Jaccard similarity distance (**F**, $R^2=0.014$, $p < 0.001$) between HNSCC patients and control groups. A significant difference between the two groups was confirmed by PERMANOVA. Dotted ellipses were drawn based on a 95% confidence level for a multivariate t-distribution. The percentage of the variation explained by each principal coordinate is presented on the axes. **E, G** Comparison of the distribution of inter- and intra-individual distances within each group using weighted UniFrac (**e**) and Jaccard similarity distance (**G**). **H** NMDS plot of saliva samples based on Bray-Curtis dissimilarity (PERMANOVA, $R^2=0.024$, $p < 0.001$). **I** Comparison of values of distance to the centroid within each group. **J** Phylo-RPCA plot of HNSCC and control groups (PERMANOVA, $R^2=0.054$, $p < 0.001$). Dotted ellipses were drawn based on a 95% confidence level for a multivariate t-distribution. Arrows represent taxonomic features with a higher driving force. Selected features are presented on the PCA plot. The median with upper and lower quartiles and whiskers extended to 1.5 interquartile range (IQR) are shown. Outliers are presented as black dots. A significant difference is determined using the Wilcoxon rank sum test and presented as an asterisk. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

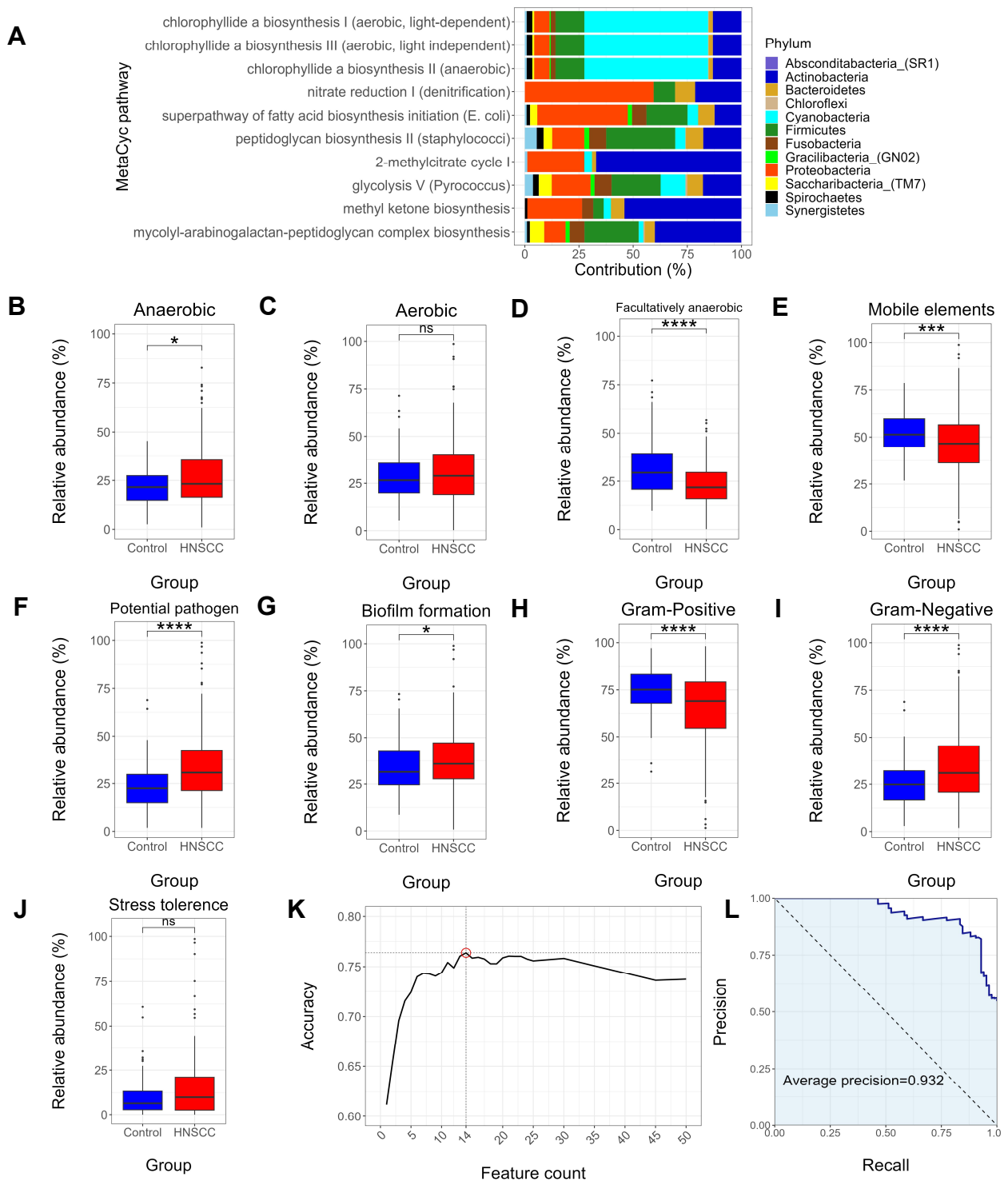


Fig. S2 Functional inference of oral microbiota reveals distinct phenotypes and contributing taxa, and the optimal RF model is discovered at 14 features with strong performance. **A** Relative abundance of the main contributors to the top five enriched functions in HNSCC patients ($n = 164$) and cancer-free control ($n = 141$) groups, respectively. The five functions at the top of the plot were enriched in the HNSCC group, and the five functions at the bottom of the plot were enriched in the control group. **B-J** Comparison of relative abundance of phenotypic traits based on Bugbase between HNSCC patients and cancer-free control groups. A significant difference is tested using the Wilcoxon rank sum test and presented as an asterisk. ns = not significant, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. **K** Recursive feature elimination plot. The RF model with the highest accuracy in predicting samples was selected automatically while sequentially removing genera from a total of 153 genera. The vertical dotted line shows feature counts for the finest model, and the horizontal dotted line shows the highest accuracy. The empty red circle indicates the peak of accuracy, representing the selected model. **L** Precision-Recall (PR) curve with average precision of the optimal RF model in the validation cohort (HNSCC, $n = 84$; controls, $n = 70$).

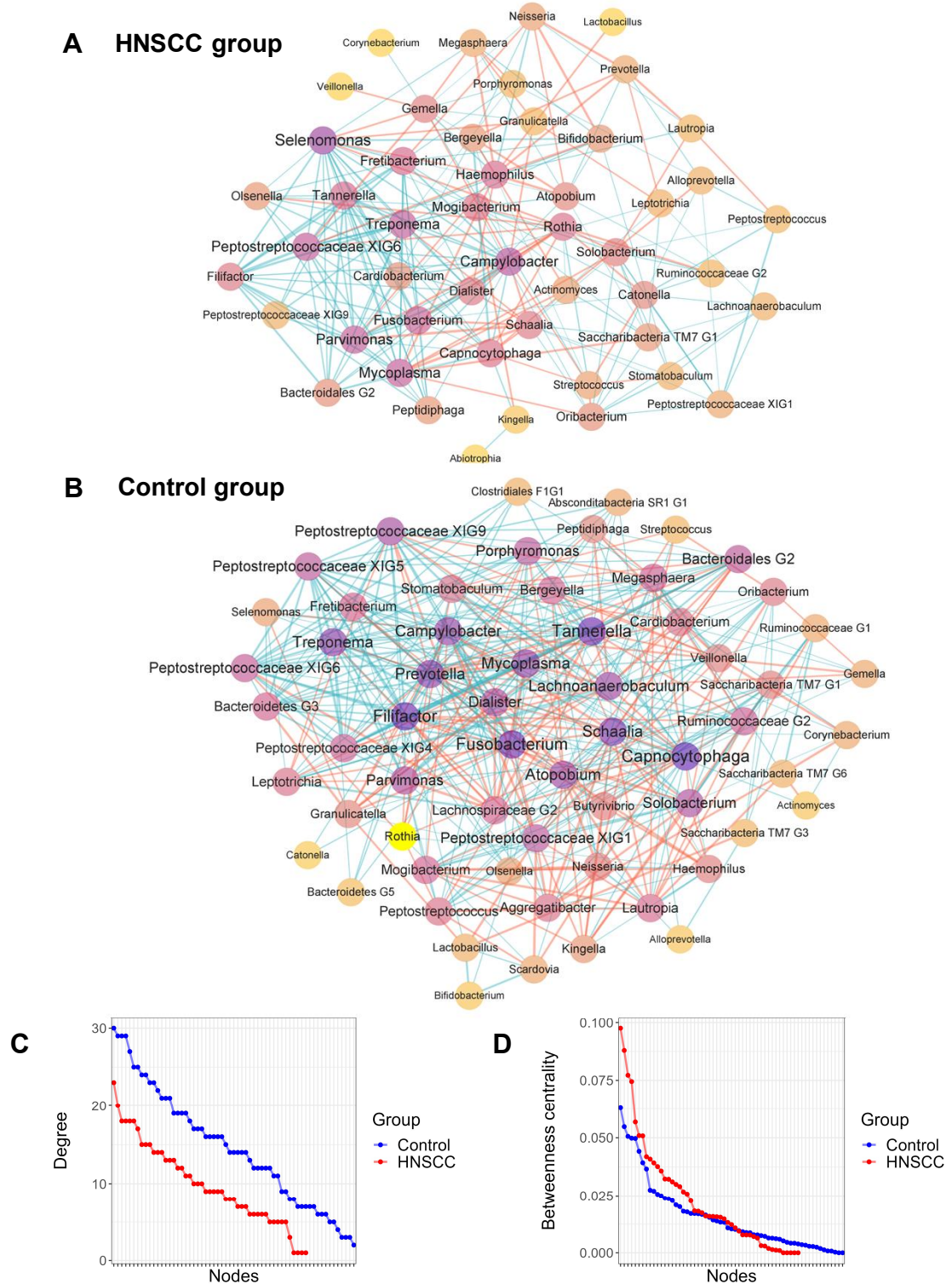


Fig. S3 Comparison of the network structure of the oral microbiota in HNSCC patients and cancer-free controls. A, B Networks of oral microbiota in HNSCC patients (**A**, $n = 168$) and cancer-free controls (**B**, $n = 141$). Nodes represent individual taxa, and edges indicate significant relationships. The color of a node represents the node's degree, and a darker color signifies a higher degree. Red and blue edges denote positive and negative correlations between taxa, respectively. **C** A distribution plot showing the degrees of individual taxa in both groups. The X axis represents each node within the networks, and the Y axis indicates the node's degree, defined as the number of edges and arranged in descending order. **D** A distribution plot showing the betweenness centrality values of individual taxa in both groups. The X axis represents each node within the networks, and the Y axis indicates each node's betweenness centrality value, arranged in descending order. 49 nodes in HNSCC patients' network and 61 nodes in cancer-free controls' network.

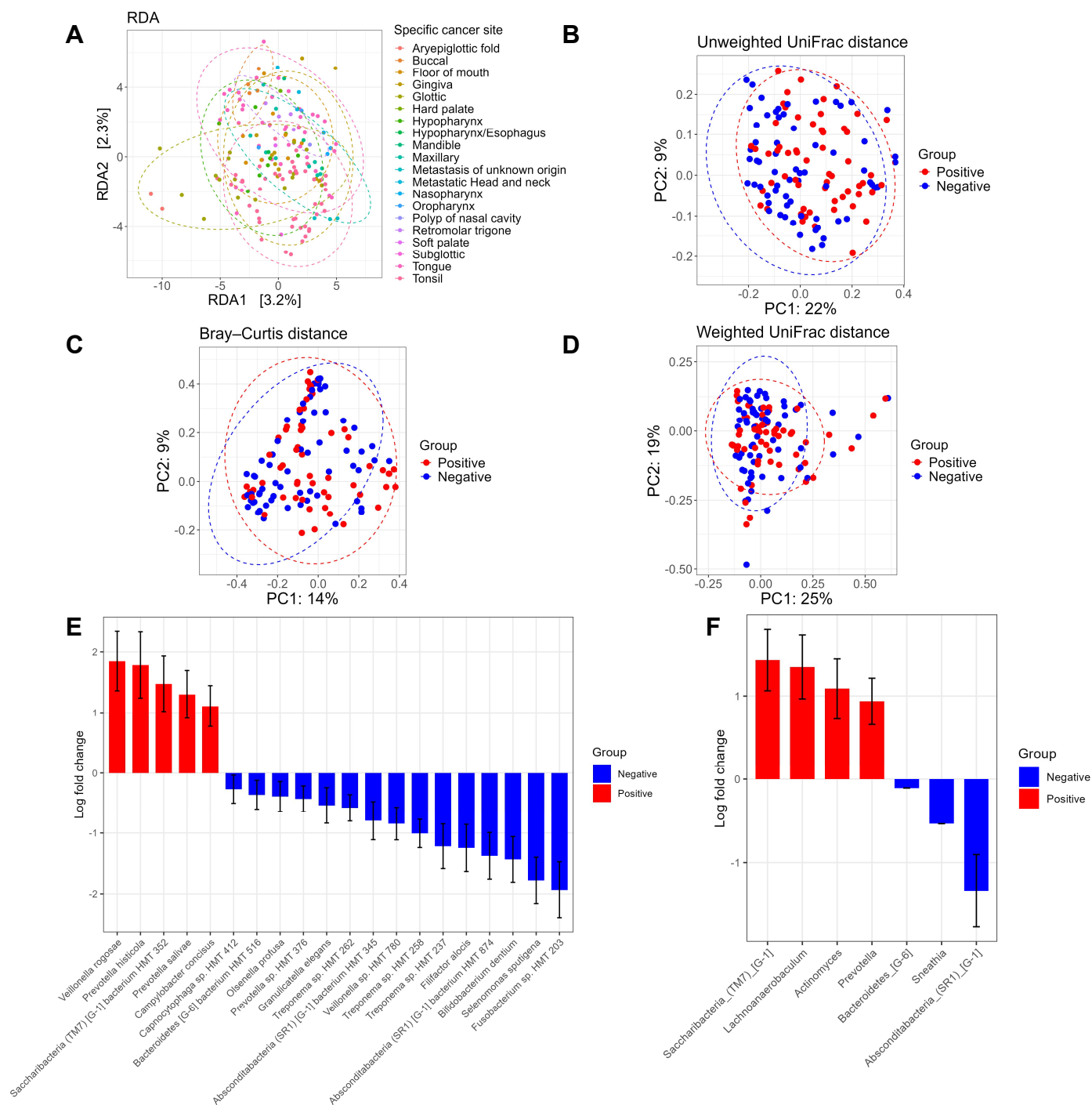


Fig. S4 HPV16-associated microbial variation in HNSCC. **A** RDA ordination plot of HNSCC samples ($n=168$) based on the specific site of cancer. The percentage of variation explained by the first two axes is presented on the X and Y axes. **B-D** Beta diversity comparison using unweighted UniFrac distance (**B**, $R^2=0.023$, $p=0.006$), Bray-Curtis dissimilarity (**C**, $R^2=0.022$, $p<0.001$) and weighted UniFrac distance (**D**, $R^2=0.012$, $p=0.175$) between HPV16 positive ($n=65$) and HPV16 negative ($n=59$) patients. Significant differences were confirmed by PERMANOVA. Dotted ellipses represent 95% confidence level for a multivariate t-distribution. Variance explained by each PCoA axis is shown. **E** Waterfall plots showing the significantly enriched taxa between two groups at the species level (**E**) and the genus level (**F**). Natural log fold changes with BH-adjusted $p<0.05$ are shown; error bars denote standard errors.

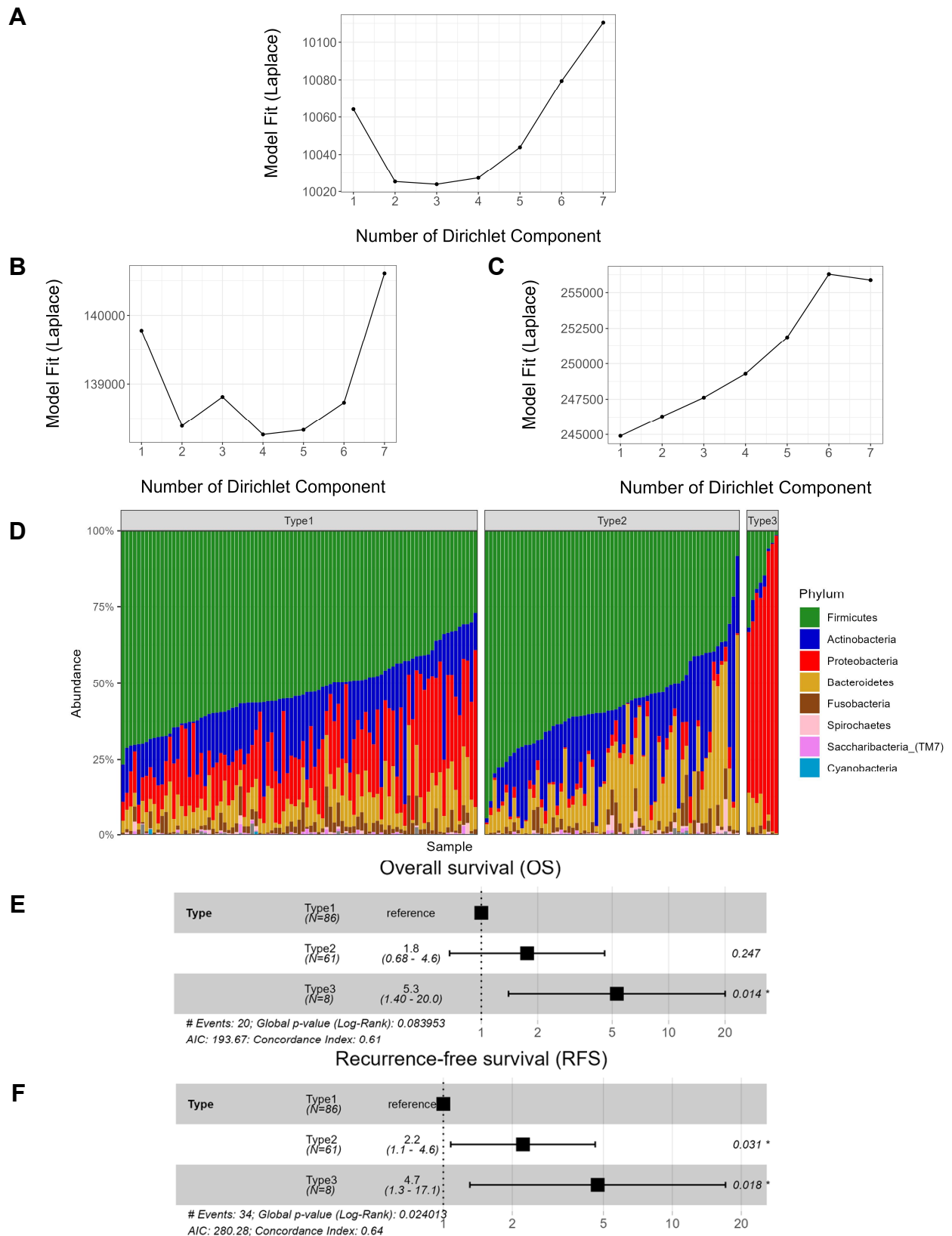


Fig. S5 The DMM identifies three distinct oral microbiota community types in HNSCC, with the Type 3 community associated with a poor prognosis. **A-C** Fitness of model using Laplace approximation. The Dirichlet multinomial mixture (DMM) model was used to identify the optimal number of Dirichlet components k at the phylum (**A**), species (**B**), and ASV (**C**) levels. While three components were optimal at the phylum level, the model couldn't determine an optimal k component at the species and ASV levels. **D** Relative abundance at the phylum level of all samples from HNSCC patients. Samples are grouped according to the annotated community types from the DMM model. **E, F** Hazard ratios with 95% confidence intervals from the Cox proportional hazards model showing probabilities of overall (**E**) and recurrence-free survival (**F**) across distinct DMM community types (Type 1, $n = 86$; Type 2, $n = 61$; Type 3, $n = 8$). The Type 1 community served as the reference. Significance was determined using the Wald test and log-rank test.