Sub-species niche specialization in the oral microbiome is associated with nasopharyngeal carcinoma risk

Justine W. Debelius^{a*}, Tingting Huang^{a,b}, Yonglin Cai^{c,d}, Alexander Ploner^a, Donal Barrett^a, Xiaoying Zhou^{e,f}, Xue Xiao^g, Yancheng Li^{c,d}, Jian Liao^h, Yuming Zheng^{c,d}, Guangwu Huang^g, Hans-Olov Adami^{a,i}, Yi Zeng^{j,#}, Zhe Zhang^{g,#}, Weimin Ye^{a,#}

^a Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

^b Department of Radiation Oncology, The First Affiliated Hospital of Guangxi Medical University, Nanning, P. R. China

^c Department of Cancer Prevention Center, Wuzhou Red Cross Hospital, Wuzhou, P. R. China;

^d Wuzhou Health System Key Laboratory for Nasopharyngeal Carcinoma Etiology and Molecular Mechanism, Wuzhou, P. R. China

^e Life Science Institute, Guangxi Medical University, Nanning, P. R. China;

^f Key Laboratory of High-Incidence-Tumor Prevention & Treatment (Guangxi Medical University), Ministry of Education, Nanning, P. R. China

^g Department of Otolaryngology-Head & Neck Surgery, First Affiliated Hospital of Guangxi Medical University, Nanning, P. R. China

^h Cangwu Institute for Nasopharyngeal Carcinoma Control and Prevention, Wuzhou, P. R. China

ⁱ Clinical Effectiveness Research Group, Institute of Health, University of Oslo, Oslo, Norway

^j State Key Laboratory for Infectious Diseases Prevention and Control, Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, P. R. China

Running Title: Oral Microbiome Niche Specialization in NPC

#Address Correspondence to Weimin Ye (weimin.ye@ki.se); Zhe Zhang (zhangzhe@gxmu.edu.cn); and Yi Zeng (zengyicdc@sina.com)

*Present Address: Center for Translational Microbiome Research, Department of Microbiology, Tumor and Cancer Biology, Karolinska Institutet, Stockholm, Sweden.

Justine Debelius, Tingting Huang, and Yonglin Cai contributed equally to this work. Authorship order was determined on seniority.

Weimin Ye, Zhe Zhang, and Yi Zeng contributed equally to this work. Authorship order was established by consortium policy.

Abstract

Oral health and changes in the oral microbiome have been associated with both local and systemic cancer. Poor oral hygiene is a known risk factor for Nasopharyngeal Carcinoma (NPC), a virally-associated head and neck cancer endemic to southern China. We explored the relationship between NPC and the oral microbiome using 16s rRNA sequencing in a study of 499 NPC patients and 495 population-based age and sex frequency-matched controls from an endemic area of Southern China. We found a significant reduction in community richness in cases compared to controls. Differences in the overall microbial community structure between cases and controls could not be explained by other potential confounders; disease status explained 5 times more variation in the Unweighted UniFrac distance than the next most explanatory variable. In feature-based analyses, we identified a pair of co-excluding *Granulicatella adiacens* (*G. adicans*) amplicon sequence variants (ASVs) which were strongly associated with NPC status and differed by a single nucleotide. The *G. adicans* variant an individual carried was also associated with the overall microbial community based on beta diversity. Co-occurrence analysis suggested the two *G. adicans* ASVs sit at the center of two co-excluding clusters of closely related organisms. Our results suggest there are differences in the oral microbiome between NPC patients and healthy controls and these may be associated with both a loss of microbial diversity and niche specialization among closely related commensals.

64 **Importance**

65

66 The relationship between oral health and the risk of Nasopharyngeal Carcinoma has been

67 previously established. However, the role of oral microbiome has not been evaluated in the

68 disease in a large epidemiological study. This paper clearly establishes a difference in the oral

69 microbiome between NPC patients and healthy controls which cannot be explained by other

70 confounding factors. It furthermore identifies a pair of closely related co-excluding organisms

71 associated with the disease, highlighting the importance of modern methods for single nucleotide

72 resolution in 16s rRNA sequence characterization. To the best of our knowledge, this is one of

73 the first examples of cancer-associated niche-specialization of the oral microbiome.

74

Introduction

The microbiome, including the oral microbiome, is emerging as an important factor in cancer and carcinogenic processes. *Helicobacter pylori* is perhaps the best known example of an oncogenic bacteria; sequencing studies have implicated a new set of organisms including *Fusobacterium nucleatum* in colorectal cancer, although cancer-associated differences in the microbiome have been seen across a wide variety of tumor types (1–3). Proposed mechanisms have included genotoxicity, immune regulation, and the production of oncogenic metabolites (4). Furthermore, differences in the local microbiome have been shown in cancer-associated viral infections including chronic Hepatitis infection and Human Papilloma Virus (5–8). The microbiome is known to contribute to and modulate infection, persistence, re-activation, and transmission of oncogenic viruses, while viral infection may contribute to immune-regulation of the microbiome (9).

Nasopharyngeal carcinoma (NPC) is a virally-associated cancer endemic to areas including southern China, Southeast Asia, the Arctic, and the Middle East/North Africa. In these regions, the incidence rate of NPC is more than 20 times higher than the rest of the world (10). Infection with the Epstein-Barr virus (EBV) is the most widely accepted and extensively studied etiological factor, although its prevalence in adult population worldwide approaches 90% (11, 12). The geographic and cultural differences associated with NPC incidence suggest that both genetic susceptibility, and environmental/life-style factors such as cigarette smoking, salted fish consumption, and alcohol drinking, contribute to carcinogenesis (11, 13–15).

Several of these NPC risk factors may affect oral health and the oral microbiome (16–19). Poor oral hygiene is a risk factor for NPC, and a recent small scale study suggested differences in the oral microbiome between NPC patients and controls prior to radiation treatment (20, 21). However, it failed to fully address the way in which a NPC-related risk factor might confound this relationship; especially with regard to smoking. Therefore, we carried out a population-based case-control study in an endemic area of southern China (22). We analyzed microbial communities from 499 untreated incident NPC cases and 495 age and sex frequency-matched controls, and addressed the relationship between NPC status and the oral microbiome adjusted for potential confounders.

Results

Study participants were recruited from the Wuzhou region in Southern China between 2010 and 2014 as part of a large population-based case-control study (22). Saliva was collected during interview. After sequencing and denoising to ASVs, there were 1066 subject's samples which had sufficiently high-quality sequences and clinical information to be retained for analysis (Figure S1). Preliminary investigation suggested the microbiota of a small number of former smokers were highly heterogenous (n=72, 33 cases, 39 controls; Figure S2). We excluded former smokers from the final analysis, retaining 994 individuals (Table S1; Figure S1).

We aimed to address the relationship between NPC and the oral microbiome, adjusted for potential confounders. As a result, we looked for factors which might affect the oral microbiome at a community level. Our primary confounders included oral hygiene and health, tobacco use,

family history of NPC, and tea consumption (14, 16, 18, 19, 23–27). We also considered a history of oropharyngeal inflammation, and the region where an individual lived and alcohol use, as covariates primarily expected to affect the microbiome, as well as salted fish consumption, which is primarily seen as a risk factor for NPC (15, 28, 29).

When comparing alpha diversity between cases and controls, we found that NPC cases showed significantly fewer overall ASVs, reduced phylogenetic diversity, and reduced Shannon diversity compared to controls (rank sum $p < 0.001$; Figure 1a; Table S2); these findings did not change after adjustment for covariates which were significantly associated with alpha diversity (Figure 1b; Tables S3-S5). Hence, this suggests that patients newly diagnosed with NPC have lower overall microbial diversity than healthy controls. Our results agree with a small study of the oral microbiome in NPC patients ($n=90$), which also found reduced alpha diversity (21). Unlike other body sites, there is no clear relationship between salivary microbiome richness and the health of the microbial community.

Similarly, when comparing global community patterns (beta-diversity) via Adonis models minimally adjusted for sex, age and sequencing run, we found significant differences between NPC cases and controls, both based on unweighted UniFrac distance as well as for weighted UniFrac and Bray-Curtis distances (FDR $p < 0.001$, 9999 permutations; Figures 1c,d and S3a,b) (30–32). Compared to the potential confounders in the same setting, NPC status was the strongest explanatory factor for unweighted UniFrac distance, more than five times the effect size of the next strongest variable, as well as the second-strongest factor for weighted UniFrac- and Bray-Curtis distances, just after tobacco use. There was no statistically significant difference

in dispersion between cases and controls in any metric, supporting the idea that the differences are due to consistent differences between cases and controls ($p > 0.55$, 999 permutations; Figure 1d). Significance persisted in more fully adjusted Adonis models including potential confounders with robust differences in community patterns.

These findings establish that NPC status and smoking are strongly associated with differences in the oral microbiome in our population; the association with NPC is especially strong with regard to presence and absence of organisms (as emphasized by unweighted UniFrac), but second only to smoking with regard to abundances (as captured by weighted UniFrac and Bray-Curtis). We found no evidence that these associations are driven by community heterogeneity; they are, however, robust under adjustment for observed confounders, and in the case of the unweighted UniFrac distances, unlikely to be the result of confounding by unobserved factors due to the crushing dominance of the signal for NPC status. Since we recruited incident, treatment-naïve patients, it is also implausible that the observed differences in microbiome composition are treatment-related (21, 22). Taken together, our findings provide strong evidence for a clear difference in the oral microbiome between patients with NPC and healthy controls.

Since the relationship between the microbiome and NPC status was strongest in unweighted UniFrac distance, which focuses on presence and absence, we evaluated the relationship between ASV prevalence and disease in a fully adjusted log binomial model. To limit spurious correlations, we defined presence as a relative abundance greater than 0.02% and focused on ASVs present in at least 10% of samples ($n=245$, Figure S4). We identified 53 ASVs which were significantly different between cases and controls ($\text{FDR } p < 0.05$; Figure 1e; Table S6). The large

majority of these ASVs were more prevalent in controls and came from a wide variety of taxonomic clades, which may suggest a somewhat stochastic loss of ASVs in NPC patients, rather than a systematic loss of specific organisms (Table S6). This finding is in line with our alpha diversity findings, and may indicate overall community instability. In contrast, two ASVs were more prevalent in cases: a member of genus *Lactobacillus* (Lact-eca9) and a *Granulicatella* ASV (Gran-7770).

To evaluate whether NPC status affected abundance-based partitioning of the microbial community, we applied Phylofactor (33). Our model looked for phylogenetic clades which differentiated NPC cases and controls, adjusting for potential confounders (Figure 2, Table S7). Of the twelve factors examined, nine were associated with disease status. The primary partition in the data suggested a *Granulicatella* ASV (Gran-7770) was 3.4 (95% CI 2.4, 4.9) fold more abundant in NPC cases compared to controls. The third factor identified was second *Granulicatella* ASV (Gran-5a37) as less abundant in cases. Both ASVs were also associated with smoking status. We identified three large-scale shifts in microbial abundance associated with NPC status. The remaining factors associated with NPC status were all single ASVs which differentiated cases and controls, none of which differed in prevalence (Table S6, S7).

Based on the significant difference in abundance and prevalence of ASVs from genus *Granulicatella* between cases and controls, we further explored this genus. We identified a total of 14 ASVs in the dataset; three were prevalent enough to be included in our feature-based analyses (Gran-5a37, Gran-7770, and Gran-6959). In 972 (97.8%) individuals, the abundant ASVs were the only *Granulicatella* present. When blasted against the Human Oral Microbiome

Database (HOMD), the ASV sequences mapped to two cultured species with more than 99.5% accuracy: *Granulicatella elegans* (*G. elegans*) which included Gran-6959 and *Granulicatella adiacens* (*G. adiacens*; Gran-7770 and Gran-5a37) (34). Strikingly, we found our two abundant *G. adiacens* ASVs differ by a single nucleotide: Gran-7770 carries a G at nucleotide 119 of our sequence (corresponding approximately to 458 in the full length 16s rRNA sequence) while Gran-5a37 carries an A.

Gran-7770 was found to be 26% more prevalent among cases, while Gran-5a37 was among the 51 ASVs less prevalent in cases (PR 0.81 [95% CI 0.74, 0.88]; Table S6)]. Both ASVs were also significantly associated with smoking status: Gran-7770 was more prevalent in smokers (PR 1.48, [95% CI 1.29, 1.70]) and Gran-5a37 less prevalent (PR 0.74, [95% CI 0.67, 0.81]). There was not a significant relationship between Gran-6959 (*G. elegans*) and either disease status (PR 0.94 [95% CI 0.88, 1.00]) or tobacco use (PR 0.97 [95% CI 0.90, 1.06]).

We found that 993 out of 994 individuals carried at least one *G. adiacens* with a relative abundance of at least 0.02%: 330 (33.2%) carried only Gran-5a37, 316 (31.8%) carried Gran-7770 alone, and 347 (34.9%) carried both. Among individuals who were classified as carrying only one ASV (Gran-7770 alone or Gran 5a37 alone), the “present” ASV was at least 50-fold more abundant than the other variant. We used a multinomial logistic regression to confirm that disease status was significantly associated with variants an individual carried: compared to the odds of carrying Gran-5a37 alone, cases had significantly higher odds of carrying both ASVs and, again, significantly higher odds of carrying Gran-7770 alone (Figure 3a). Although smokers

were more likely to have both ASVs or Gran-7770 alone, there was no significant interaction between smoking and disease status.

We also investigated how the presence of a *G. adiacens* variant structured the overall microbial community. We filtered the full ASV table to remove any *Granulicatella* ASVs and used the reduced table to re-calculate beta diversity metrics. The *Granulicatella*-free community recapitulated the patterns seen in the full community well (Mantel $R^2 > 0.91$; $p=0.001$, 999 permutations). We found significant differences between individuals who carried Gran-7770, both, or Gran-5a37 in weighted and unweighted UniFrac distances and Bray Curtis; all three metrics show clear separation in PCoA space ($p=0.001$, 999 permutations; Figure 3b; Figure S5). In unweighted UniFrac space (Figure 3b), the separation was primarily along PC2, likely corresponding to the separation along PC2 seen between cases and controls (Figure 1d). Furthermore, we found that the *G. adiacens* variant explained 16% of the variation attributed to case-control status in unweighted UniFrac distance and 15% of the variation in weighted UniFrac distance. Our results suggest that the *G. adiacens* variant carried by an individual is significantly associated with community structure, and may be a route by which NPC status shapes the oral microbiome.

We used a SparCC-based network analysis to identify other community members *Granulicatella* might interact with to exert an effect on the microbiome (35). We were able to identify five networks: one pair of co-occurring ASVs, two pairs of co-excluding ASVs, one three-member network of co-occurring ASVs and a large 29-member network of co-occurring and co-excluding ASVs (Figures 4a). This main network consisted of two clusters of a total of 20 organisms which

were positively correlated with a *Granulicatella* variant; the main members of the networks belonged to *Veillonella*, *Streptococcus*, and *Prevotella*. Blasting against HOMD, we identified two additional pairs of ASVs that co-excluded between the two nodes but mapped to the same clones: Stre-900d and Stre-0531 (*Streptococcus parasanguinis* clade 411) and *Prevotella melaninogenica* (Prev-b7f2 and Prev-71e7; Figure 4b; Table S8) (34).

We hypothesize the co-excluding networks of ASVs, centered around *Granulicatella*, may reflect partial niche specialization. Previous work suggests quorum sensing networks can form between the core species, and that metabolic changes occur in these networks (36, 37). These closely related organisms occupy the same niches within these metabolic networks, however, strain-specific variation may either respond to or promote disease-associated transformation. Without culture-based experimentation, it is difficult to determine how these organisms may function in concert. One major challenge for in-silico validation is the limited resolution of existing databases; our results exceed the OTU-based resolution used in construction and span a less frequently characterized hypervariable region.

Within the context of NPC in an endemic region, we hypothesize the oral microbiome may act through several potential mechanisms. The oral microbiome has been suggested to contribute to local tumorigenesis through immune regulation or oncogenic metabolites such as acetaldehyde or nitrosamines (38). An *in silico* study suggested that commercially available strains of *G. adiacens* and co-abundant organisms encode genes involved in nitrate and nitrite reduction (39).

In summary, we have demonstrated a difference in the oral microbial community between NPC patients and healthy controls in an endemic area of southern China, which cannot be explained by other measured factors. The difference is associated with both a loss of community richness and differences among specific organisms, including closely related ASVs from genus *Granulicatella*. In addition, we identified a network of co-occurring and co-excluding ASVs which included these *Granulicatella* variants. These results strongly suggest a relationship between the oral microbiome and nasopharyngeal carcinoma status in untreated patients.

Acknowledgements

The authors wish to thank the study participants, the field work team for the NPCGEE project, and the Wuzhou Health System Key Laboratory for Nasopharyngeal Carcinoma Etiology and Molecular Mechanism and the Key Laboratory of High-Incidence-Tumor Prevention & Treatment (Guangxi Medical University), especially Suhua Zhong, Xiling Xiao, for the processing of salivary samples. The data was stored in the Department of Medical Epidemiology and Biostatistics; the authors wish to thank the IT group for their assistance.

We acknowledge funding by the National Cancer Institute of the NIH grant under Award Number R01CA115873 [Principal Investigator (PI): H.-O. Adami; co-PIs: Y. Zeng, Y.X. Zeng]; from the Swedish Research Council (2015-02625, 2015-06268, 2017-05814 to Ye); the National Natural Science Foundation of China (81272983 to Z. Zhang); and the Guangxi Natural Science Foundation (2013GXNSFGA019002 to Z. Zhang). T. Huang is partly supported by a grant from China Scholarship Council (201408450018).

280

281 **Data Availability**

282

283 Raw sequencing data, feature table, and metadata are available from the corresponding author
284 upon request.

285

286 **Author contributions**

287

288 The study approach was conceived by GH, HA, WY, YZ, and ZZ. AP, DB, JWD , TH, WY, and
289 YC, refined the study design for this project. JL, YC, YL, and YZ were responsible for sample
290 collection and management. DB, SZ, and XX performed the lab work, supervised by TH, XX,
291 XZ and ZZ. AP, JWD, and TH and performed statistical modeling and refinement; bioinformatic
292 and biostatistical analyses were performed by JWD. WY contributed to the supervision and
293 coordination of the project. JWD and TH wrote the manuscript; AP provided critical edits. All
294 authors reviewed and approved the final submission.

295

296 **Methods**

297

298 *Survey metadata and sample collection*

299 Participant recruitment has been previously described (22). Briefly, incident cases of NPC in the
300 Guangxi Autonomous Region between 2010 and 2013 were invited to participate in the full
301 study. Age and sex matched controls were selected from the total population. The study was

approved by the Institutional Review Board or Ethical Review Board at all participating centers.

All study participants provided written or oral informed consent.

A questionnaire covering demographics, diet, residential, occupational, medical and family history was administered in a structured interview. Sample collection occurred at the interview. Participants were asked not to eat nor chew gum for 30 minutes prior to sample collection. Saliva samples (2ml-4ml) were collected into 50ml falcon tubes with a Tris-EDTA buffer.

Demographic characteristics of the study population were compared using a two-sided t-test for continuous covariates (age) and a chi-squared test for categorical covariates. Tests were conducted using scipy 0.19.1 in python 3.5.5 (40).

DNA extraction, PCR, and sequencing

Saliva DNA was extracted using a two-step protocol including the sample pre-processing with lysozyme lysis and bead beating, and the TIANamp blood DNA kit (Beijing, China). The 16s rRNA amplicon library was amplified with 341F/805R primers (CCTACGGGNGGCWGCAG, GACTACHVGGGTATCTAATCC) (41, 42). Samples were amplified with 20 cycles of a program with 30 seconds at 98°C for melting, 30 second at 60°C, and 30 seconds at 72°C.

Samples were barcoded in a second PCR step (41). DNA clean-up was performed using Agentcourt AMPure XP purification kit. DNA volume and purity were measured on an Agilent 2100 Bioanalyzer system and Real-time polymerase chain reaction. Sequencing was performed at Beijing Genome Institute (BGI) on an Illumina MiSeq using a 2x300bp paired end strategy.

Denoising, Annotation and Filtering

Samples were demultiplexed using an in-house script. Adaptors were trimmed and paired end sequences were joined using VSEARCH (v. 2.7) (43). Paired sequences were loaded into the November 2018 release of QIIME 2 (44). Sequences were quality filtered (q2-quality-filter) and denoised using deblur (v. 1.0.4; q2-deblur) with the default parameters on 420 bp amplicons to generate amplicon sequence variants (ASVs) (45, 46). A phylogenetic tree was built using fragment insertion into the August 2013 Greengenes 99% identity tree backbone with q2-fragment-insertion; taxonomic assignments were made with a naïve Bayesian classifier trained against the same reference (q2-feature-classifier) (47–49). In cases where the classifier or reference database was unable to describe a taxonomic level (for instance, a missing genus), the taxonomy was described by inheriting the lowest defined level using a custom python script. Following sequencing and denoising, 24,763,933 high quality reads were retained.

Any sample with fewer than 1000 reads after denoising was excluded, leaving 1074 saliva samples and 9 negative or single organism controls. Additionally, samples missing information on tobacco use, defined information about tooth brushing frequency, or an undefined residential region ($n=8$) were excluded (Figure S1).

Preliminary investigation suggested that the microbial communities for former smokers ($n=72$) were highly heterogenous (Figure S2). Sensitivity analyses suggest their exclusion does not alter the major community-level differences. Therefore, they were excluded, leaving a total of 994 individuals in the analysis.

ASV-based analyses were performed on a representative subset: those with at least 0.02% relative abundance in at least 10% of samples (n=245). A Mantel test was applied to Bray Curtis distance showed a correlation of 0.96 between the filtered matrix rarefied to 5000 sequences/sample and the full table distance matrix (p=0.001, 999 permutations); the mantel corresponding correlation for UniFrac distance was 0.76 (p=0.001, 999 permutations; Figure S3) (30, 32, 50).

The sequences and identifiers for the abundant ASVs are listed in supplemental file 2. ASVs are identified by the first 4 letters of their lowest taxonomic assignment and the first 4 characters of a MD5 hash of the sequence. The full taxonomic assignment and MD5 hashes can be found in Table S6.

Diversity Analyses

Diversity analyses were performed using samples rarefied to 6,500 sequences.

Alpha diversity was calculated as observed ASVs, Shannon diversity, and Faith's phylogenetic diversity using q2-diversity in QIIME 2 (51, 52). Potentially significant alpha diversity categories predictors were identified using a rank-sum test in scipy 0.19.1 (40). A p-value of 0.05 was considered the threshold for borderline significance for inclusion in a subsequent regression model. Alpha diversity was then evaluated in a multivariate ordinary least squares (OLS) regression model adjusted for age, sex and sequencing run number. A final model for each metric was selected by forward selection using models which resulted in decreasing AIC. We checked for the normality of residuals by plotting. The relative contribution of each covariate to

that metric was estimated by a “leave one out” approach. Regressions were performed in Statsmodels (v. 0.9.0) (53). For visualization, we calculated z-normalized alpha diversity using the mean and standard deviation in diversity for the controls. Alpha diversity was plotted using boxenplots in Seaborn 0.9.0 (54, 55).

Beta diversity was measured using the unweighted UniFrac, weighted UniFrac, and Bray-Curtis metrics on rarefied data (q2-diversity) (30–32). Beta diversity was compared using Adonis in the R *vegan* library (v 2.5-2) adjusted for host age, sex, and sequencing run, with 9999 permutations (56–58). We used a permdisp test with 999 permutations and the centroid estimate to test for the presence of differences in within-group variation implemented in scikit-bio 0.5.4 (www.scikit-bio.org). (59) Uncorrected p-values of less than 0.05 were considered to have significant dispersion, since we were more concerned about false positives than false negatives. Principal coordinate analyses (PCoA)s were visualized using Emperor (v. 1.0.0b18) and seaborn v. 0.9.0 in matplotlib v. 2.2.3 (54, 60).

ASV regression model

To look at the relationship between ASV prevalence and disease and smoking status, we used a log binomial regression which was approximated via a Poisson regression with robust standard errors, implemented via base function *glm* in R and the robust error mechanism implemented via packages *lmtest* (v 0.9) and *sandwich* (v. 2.5) in R 3.5 (58, 61–63). The model was adjusted for age, sex, sequencing run, residential community, and the number of missing or repaired teeth. “Presence” was defined as a relative abundance of 1 / 5000, which corresponded to the shallowest sequencing depth for the abundant counts. ASVs which were present in more than

1000 samples were excluded from prevalence analysis. A Benjamini-Hochberg FDR corrected p-value of 0.05 was considered significant.

Phylofactor

Phylofactor (v. 0.01) was used to look at the relationship between disease status and phylogenetic partitioning between clades (33). Phylofactor is a compositionally aware technique which uses isometric log transforms over an unrooted phylogenetic tree to model differences in the data. This allows the partitioning of data into polyphyletic clades. The Phylofactor multivariate model for each partition was modeled with an OLS regression considering diagnosis, adjusted for residential community, age, sex, number of missing or repaired teeth, tobacco use, and sequencing run. We looked at the first 12 factors using the default parameters, which optimized for explaining maximal variance. The cladogram, and regression coefficient plots were generated in seaborn (54).

Granulicatella

Total *Granulicatella* was identified by filtering the full ASV table for any ASV assigned to the genus. Species-level assignments were made by blasting each ASV against the Human Oral Microbiome Database using the online tool; species-level assignments were taken for the cultured species with the best match (34). We treated the abundance of Gran-6959 as the *G. elegans* abundance and the combined abundance of Gran-5a37 and Gran-7770 as the *G. adiacens* abundance throughout. One sample which did not contain Gran-7770 or Gran-5a37 was excluded from *G. adiacens* associated analyses.

We used a multinomial logistic regression model, implemented in the *nnet* library (v. 0.8) in R to look at whether the carriage of Gran-5a377 alone, Gran-7770 alone, or both ASVs was associated with smoking and disease status (64). The regression was adjusted for age, sex, sequencing run, number of missing or repaired teeth, residential community, the relative abundance of *G. adiacens*, and the relative abundance of *G. elegans*. Having Gran-5a37 was considered the reference group for the multinomial regression.

The effect of *Granulicatella* on alpha and beta diversity was calculated by first filtering out all *Granulicatella* ASVs from the table, and then rarifying to 6250 sequences/sample before diversity calculations. Adonis coefficients were calculated in a model adjusted for *G. adiacens* abundance, sequencing run, age, sex, residential community, number of missing or repaired teeth, tobacco use, and disease status. The proportion of disease status explained by comparing a model excluding the *Granulicatella* variant minus the model including the variant over the model excluding the variant.

Network Analysis

We used the Sparse Cooccurrence Network Investigation for Compositional data (SCNIC; <https://github.com/shafferm/SCNIC>) in QIIME 2 (q2-SCNIC) to perform network analysis on the abundant ASVs. The correlation network was built using SparCC, and the network was built using edges with a correlation co-efficient of at least 0.3, allowing both co-occurrence and co-exclusion (35). Network clusters were identified by finding the most connected node and following all positively correlated nodes in the trimmed SparCC network. Networks were visualized in Cytoscape (v. 3.7.1) using a perfuse-weighted network layout (65). Nodes which

were anti-correlated with a single node in the main cluster were trimmed for the sake of visualization; these are labeled with the correlation coefficient.

The phylogenetic tree of core network members was visualized using ete3 (v. 3.1.1) in python 3.6 (66).

References

1. Abdul-Rasoul M, Habib H, Al-Khouly M. 2006. “The honeymoon phase” in children with type 1 diabetes mellitus: frequency, duration, and influential factors. *Pediatr Diabetes* 7:101–7.
2. Vogtmann E, Goedert JJ. 2016. Epidemiologic studies of the human microbiome and cancer. *Br J Cancer* 114:237–42.
3. Wroblewski LE, Peek RM, Wilson KT. 2010. *Helicobacter pylori* and gastric cancer: Factors that modulate disease risk. *Clin Microbiol Rev.*
4. Bhatt AP, Redinbo MR, Bultman SJ. 2017. The role of the microbiome in cancer development and therapy. *CA Cancer J Clin* 67:326–344.
5. Lu H, Wu Z, Xu W, Yang J, Chen Y, Li L. 2011. Intestinal microbiota of HBV cirrhotic patients. *Microb Ecol* 61:693–703.
6. Ling Z, Liu X, Cheng Y, Jiang X, Jiang H, Wang Y, Li L. 2015. Decreased Diversity of the Oral Microbiota of Patients with Hepatitis B Virus-Induced Chronic Liver Disease: A Pilot Project. *Sci Rep* 5.

7. Guerrero-Preston R, Godoy-Vitorino F, Jedlicka A, Rodríguez-Hilario A, González H, Bondy J, Lawson F, Folawiyo O, Michailidi C, Dziedzic A, Thangavel R, Hadar T, Noordhuis MG, Westra W, Koch W, Sidransky D. 2016. 16S rRNA amplicon sequencing identifies microbiota associated with oral cancer, human papilloma virus infection and surgical treatment. *Oncotarget* 7:51320–51334.
8. Torcia MG. 2019. Interplay among Vaginal Microbiome, Immune Response and Sexually Transmitted Viral Infections. *Int J Mol Sci* 20.
9. Li N, Ma WT, Pang M, Fan QL, Hua JL. 2019. The commensal microbiota and viral infection: A comprehensive review. *Front Immunol* 10.
10. Wei K-R, Zheng R-S, Zhang S-W, Liang Z-H, Li Z-M, Chen W-Q. 2017. Nasopharyngeal carcinoma incidence and mortality in China, 2013. *Chin J Cancer* 36:90.
11. Chua MLK, Wee JTS, Hui EP, Chan ATC. 2016. Nasopharyngeal carcinoma. *Lancet* (London, England) 387:1012–1024.
12. Lieberman PM. 2014. Virology. Epstein-Barr virus turns 50. *Science* 343:1323–5.
13. Huang S-F, Hsiao J-H, Young C-K, Chien H-T, Kuo C-F, See L-C, Luo S-F, Huang L-H, Liao C-T, Chang T-CJ. 2017. Familial aggregation of nasopharyngeal carcinoma in Taiwan. *Oral Oncol* 73:10–15.
14. Long M, Fu Z, Li P, Nie Z. 2017. Cigarette smoking and the risk of nasopharyngeal carcinoma: a meta-analysis of epidemiological studies. *BMJ Open* 7:e016582.
15. Barrett D, Ploner A, Chang ET, Liu Z, Zhang C-X, Liu Q, Cai Y, Zhang Z, Chen G, Huang Q-H, Xie S-H, Cao S-M, Shao J-Y, Jia W-H, Zheng Y, Liao J, Chen Y, Lin L, Ernberg I, Adami H-O, Huang G, Zeng Y, Zeng Y-X, Ye W. 2019. Past and Recent Salted Fish and Preserved Food Intakes Are Weakly Associated with Nasopharyngeal Carcinoma

Risk in Adults in Southern China. *J Nutr*.

16. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, Clark AG. 2015. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* 16:191.
17. Wu J, Peters BA, Dominianni C, Zhang Y, Pei Z, Yang L, Ma Y, Purdue MP, Jacobs EJ, Gapstur SM, Li H, Alekseyenko A V, Hayes RB, Ahn J. 2016. Cigarette smoking and the oral microbiome in a large study of American adults. *ISME J* 10:2435–46.
18. Belstrøm D, Sembler-Møller ML, Grande MA, Kirkby N, Cotton SL, Paster BJ, Twetman S, Holmstrup P. 2018. Impact of Oral Hygiene Discontinuation on Supragingival and Salivary Microbiomes. *JDR Clin Transl Res* 3:57–64.
19. Kilian M, Chapple ILC, Hannig M, Marsh PD, Meuric V, Pedersen AML, Tonetti MS, Wade WG, Zaura E. 2016. The oral microbiome – an update for oral healthcare professionals. *Br Dent J* 221:657–666.
20. Liu Z, Chang ET, Liu Q, Cai Y, Zhang Z, Chen G, Xie S-H, Cao S-M, Shao J-Y, Jia W-H, Zheng Y, Liao J, Chen Y, Ernberg I, Vaughan TL, Adami H-O, Huang G, Zeng Y, Zeng Y-X, Ye W. 2016. Oral Hygiene and Risk of Nasopharyngeal Carcinoma-A Population-Based Case-Control Study in China. *Cancer Epidemiol Biomarkers Prev* 25:1201–7.
21. Zhu X-X, Yang X-J, Chao Y-L, Zheng H-M, Sheng H-F, Liu H-Y, He Y, Zhou H-W. 2017. The Potential Effect of Oral Microbiota in the Prediction of Mucositis During Radiotherapy for Nasopharyngeal Carcinoma. *EBioMedicine* 18:23–31.
22. Ye W, Chang ET, Liu Z, Liu Q, Cai Y, Zhang Z, Chen G, Huang Q-H, Xie S-H, Cao S-M, Shao J-Y, Jia W-H, Zheng Y, Liao J, Chen Y, Lin L, Liang L, Ernberg I, Vaughan TL, Huang G, Zeng Y, Zeng Y-X, Adami H-O. 2017. Development of a population-based

cancer case-control study in southern china. *Oncotarget* 8:87073–87085.

23. Liu Z, Chang ET, Liu Q, Cai Y, Zhang Z, Chen G, Xie S-H, Cao S-M, Shao J-Y, Jia W-H, Zheng Y, Liao J, Chen Y, Ernberg I, Vaughan TL, Adami H-O, Huang G, Zeng Y, Zeng Y-X, Ye W. 2016. Oral Hygiene and Risk of Nasopharyngeal Carcinoma-A Population-Based Case-Control Study in China. *Cancer Epidemiol Biomarkers Prev* 25:1201–7.
24. Wu J, Peters BA, Dominianni C, Zhang Y, Pei Z, Yang L, Ma Y, Purdue MP, Jacobs EJ, Gapstur SM, Li H, Alekseyenko A V, Hayes RB, Ahn J. 2016. Cigarette smoking and the oral microbiome in a large study of American adults. *ISME J* 10:2435–46.
25. Huang S-F, Hsiao J-H, Young C-K, Chien H-T, Kuo C-F, See L-C, Luo S-F, Huang L-H, Liao C-T, Chang T-CJ. 2017. Familial aggregation of nasopharyngeal carcinoma in Taiwan. *Oral Oncol* 73:10–15.
26. Yuan X, Long Y, Ji Z, Gao J, Fu T, Yan M, Zhang L, Su H, Zhang W, Wen X, Pu Z, Chen H, Wang Y, Gu X, Yan B, Kaliannan K, Shao Z. 2018. Green Tea Liquid Consumption Alters the Human Intestinal and Oral Microbiome. *Mol Nutr Food Res* 62:e1800178.
27. Hsu W-L, Pan W-H, Chien Y-C, Yu KJ, Cheng Y-J, Chen J-Y, Liu M-Y, Hsu M-M, Lou P-J, Chen I-H, Yang C-S, Hildesheim A, Chen C-J. 2012. Lowered risk of nasopharyngeal carcinoma and intake of plant vitamin, fresh fish, green tea and coffee: a case-control study in Taiwan. *PLoS One* 7:e41779.
28. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, Chen M-X, Chen Z-H, Ji G-Y, Zheng Z-D-X, Mujagond P, Chen X-J, Rong Z-H, Chen P, Lyu L-Y, Wang X, Wu C-B, Yu N, Xu Y-J, Yin J, Raes J, Knight R, Ma W-J, Zhou H-W. 2018. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat*

Med 24:1532–1535.

29. Fan X, Peters BA, Jacobs EJ, Gapstur SM, Purdue MP, Freedman ND, Alekseyenko A V, Wu J, Yang L, Pei Z, Hayes RB, Ahn J. 2018. Drinking alcohol is associated with variation in the human oral microbiome in a large study of American adults. *Microbiome* 6:59.
30. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Env Microbiol* 71:8228–8235.
31. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and Qualitative Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Appl Environ Microbiol* 73:1576–1585.
32. Sørensen T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. I kommission hos E. Munksgaard, København.
33. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* 5:e2969.
34. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. 2018. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. *mSystems* 3:e00187-18.
35. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 8:e1002687.
36. Chalmers NI, Palmer RJ, Cisar JO, Kolenbrander PE. 2008. Characterization of a

Streptococcus sp.-Veillonella sp. Community Micromanipulated from Dental Plaque. J Bacteriol 190:8145–8154.

37. Palmer RJ, Diaz PI, Kolenbrander PE. 2006. Rapid succession within the Veillonella population of a developing human oral biofilm in situ. J Bacteriol 188:4117–24.

38. Gholizadeh P, Eslami H, Yousefi M, Asgharzadeh M, Aghazadeh M, Kafil HS. 2016. Role of oral microbiome on oral cancers, a review. Biomed Pharmacother 84:552–558.

39. Hyde ER, Andrade F, Vaksman Z, Parthasarathy K, Jiang H, Parthasarathy DK, Torregrossa AC, Tribble G, Kaplan HB, Petrosino JF, Bryan NS. 2014. Metagenomic analysis of nitrate-reducing bacteria in the oral cavity: implications for nitric oxide homeostasis. PLoS One 9:e88645.

40. Jones E, Oliphant T, Peterson P, others. {SciPy}: Open source scientific tools for {Python}.

41. Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. 2011. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. ISME J 5:1571–9.

42. Hugerth LW, Wefer HA, Lundin S, Jakobsson HE, Lindberg M, Rodin S, Engstrand L, Andersson AF. 2014. DegePrime, a Program for Degenerate Primer Design for Broad-Taxonomic-Range PCR in Microbial Ecology Studies. Appl Environ Microbiol 80:5116–5123.

43. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

44. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A,

Brislaw CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang K Bin, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciulek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik A V., Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Priesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37:852–857.

45. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* 10:57–9.

46. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* 2:e00191-16.

47. Janssen S, McDonald D, Gonzalez A, Navas-Molina JA, Jiang L, Xu ZZ, Winker K, Kado

- DM, Orwoll E, Manary M, Mirarab S, Knight R. 2018. Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *mSystems* 3:e00021-18.
48. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6:610–8.
49. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–7.
50. Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res* 27:209–220.
51. Shannon CE, E. C. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mob Comput Commun Rev* 5:3.
52. Faith DP, Baker AM. 1992. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evol Bioinform Online* 2:121–128.
53. JS Seabold JP. 2010. Statsmodels: Econometric and Statistical Modeling with Python. *Proc 9th Python Sci Conf*.
54. Waskom M, Botvinnik O, O’Kane D, Hobson P, Ostblom J, Lukauskas S, Gemperline DC, Augspurger T, Halchenko Y, Cole JB, Warmenhoven J, Ruiter J de, Pye C, Hoyer S, Vanderplas J, Villalba S, Kunter G, Quintero E, Bachant P, Martin M, Meyer K, Miles A, Ram Y, Brunner T, Yarkoni T, Williams ML, Evans C, Fitzgerald C, Brian, Qalieh A. 2018. *mwaskom/seaborn: v0.9.0* (July 2018).
55. Hofmann H, Kafadar K, Wickham H. 2011. Letter-value plots: Boxplots for large dataThe

American Statistician.

56. McArdle BH, Anderson MJ. 2001. FITTING MULTIVARIATE MODELS TO COMMUNITY DATA: A COMMENT ON DISTANCE-BASED REDUNDANCY ANALYSIS. *Ecology* 82:290–297.
57. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O’Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2018. *vegan: Community Ecology Package*.
58. R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria.
59. Anderson MJ. 2006. Distance-Based Tests for Homogeneity of Multivariate Dispersions. *Biometrics* 62:245–253.
60. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R, Vazquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* 2:16.
61. Barros AJ, Hirakata VN. 2003. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med Res Methodol* 3:21.
62. Zeileis A. 2006. Object-Oriented Computation of Sandwich Estimators. *J Stat Softw* 16:1–16.
63. Zeileis A. 2004. Econometric Computing with {HC} and {HAC} Covariance Matrix Estimators. *J Stat Softw* 11:1–17.
64. Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S* Fourth. Springer, New York.

- 647 65. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski
648 B, Ideker T. 2003. Cytoscape: a software environment for integrated models of
649 biomolecular interaction networks. *Genome Res* 13:2498–504.
- 650 66. Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and
651 Visualization of Phylogenomic Data. *Mol Biol Evol* 33:1635–1638.
- 652
- 653

Figure 1. The oral microbiome differs between patients with nasopharyngeal carcinoma

and healthy controls. (a) NPC cases (red) have significantly lower microbial richness compared to cases (blue; $p < 1 \times 10^{-12}$). The horizontal line in the boxlin represents the median, the large box the interquartile region, increasingly smaller boxes are the upper and lower eighths, sixteens, etc. in the data, reflecting the distribution. This difference is reflected in **(b)** the correlation coefficients from a multivariate regression model. **(c)** Adonis testing with a model adjusted for age, sex, and sequencing run shows that for unweighted UniFrac distance, NPC diagnosis has more than five times the explanatory power of the next most important variable, residential community. For 9999 permutations, FDR-adjusted $p < 0.001$ ***; $p < 0.01$ **; $p < 0.05$ *. **(d)** Principal coordinates analysis (PCoA) of unweighted UniFrac shows separation between cases (red) and controls (blue) along PC1 and PC2. Upper and right panels reflect the density distribution along each axis. The axes are labeled with the variation they explain. In unweighted

667 UniFrac, PC1 explains 19.7% and PC2 explains 4.8% of the variation. A volcano plot of (e) the
 668 Poisson regression coefficient for disease status vs the log p-value reflects reduced diversity. The
 669 horizontal line indicates significant at a Benjamini-Hochberg corrected p-value of less than 0.05.
 670

Figure 2. There are significant associations between phylogenetic partitioning of the taxa and NPC status.

The phylogenetic tree with the first 12 phylofactor-based clade partitions is shown on the left. The top row is colored by phylum, the associated color is shown below. The isometric log transformation is taken as the ratio of the tips highlighted in pink over those highlighted in gray and passed into the regression model to predict the coefficient shown in the forest plot. Clades which are excluded from that factor appear white in the row. The forest plot to the right shows the estimated increase in the factor associated with case control status based on fitting the ratio in a linear regression adjusted for age, sex, sequencing run, number of missing or repaired teeth, tobacco use, and residential community. Error bars are 95% confidence interval for the regression coefficient. Black bars indicate significance at $\alpha < 0.05$, gray indicates a non-significant association.

Figure 3. The *Granulicatella adiacens* variant predicts community structure. (a) NPC cases have significantly higher odds of carrying both Gran-5a37 and Gran-7770 than Gran-5a37 alone, and again, significantly higher odds than carrying either Gran-5a37 and Gran-7770 or Gran-7770. (b) In unweighted UniFrac space, we see separation based on the *G. adiacens* variant along PC2.

Figure 4. *Granulicatella adiacens* variants set at the center of a network of closely related co-occurring organisms. (a) SparCC-based network analysis for co-occurring and co-excluding ASVs for all subjects showed a large network with two clusters with common core structures. The color and shape of the nodes are genus-specific. The two *G. adiacens* variants are highlighted as stars: Gran-5a37 in purple and Gran-7770 in green. Correlated edges are shown in pink, anti-correlated edges are grey. The sides of each network are labeled with their associated *G. adiacens* variant. (b) Phylogenetic tree of the core ASVs from the network (positively

700 correlated with either Gran-7770 or Gran-5a37). Tips are labeled by their association with Gran-
701 7770 (Green) or Gran-5a37 (Purple).

702