# 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

23 [j] State Key Laboratory for Infectious Diseases Prevention and Control, Institute for Viral Disease
24 Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, P. R. China
25
26 Running Title: Oral Microbiome Niche Specialization in NPC
27
28 #Address Correspondence to Weimin Ye (weimin.ye@ki.se); Zhe Zhang
29 (zhangzhe@gxmu.edu.cn); and Yi Zeng (zengyicdc@sina.com)
30
31 *Present Address: Center for Translational Microbiome Research, Department of Microbiology,
32 Tumor and Cancer Biology, Karolinska Institutet, Stockholm, Sweden.
33
34 Justine Debelius, Tingting Huang, and Yonglin Cai contributed equally to this work. Authorship
35 order was determined on seniority.
36
37 Weimin Ye, Zhe Zhang, and Yi Zeng contributed equally to this work. Authorship order was
38 established by consortium policy.
39
40

**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

75 **Introduction**

76

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

88

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

97

98 Several of these NPC risk factors may affect oral health and the oral microbiome (16–19). Poor
99 oral hygiene is a risk factor for NPC, and a recent small scale study suggested differences in the
100 oral microbiome between NPC patients and controls prior to radiation treatment (20, 21).
101 However, it failed to fully address the way in which a NPC-related risk factor might confound
102 this relationship; especially with regard to smoking. Therefore, we carried out a population-
103 based case-control study in an endemic area of southern China (22). We analyzed microbial
104 communities from 499 untreated incident NPC cases and 495 age and sex frequency-matched
105 controls, and addressed the relationship between NPC status and the oral microbiome adjusted
106 for potential confounders.
107
108 **Results**
109
110 Study participants were recruited from the Wuzhou region in Southern China between 2010 and
111 2014 as part of a large population-based case-control study (22). Saliva was collected during
112 interview. After sequencing and denoising to ASVs, there were 1066 subject's samples which had
113 sufficiently high-quality sequences and clinical information to be retained for analysis (Figure S1).
114 Preliminary investigation suggested the microbiota of a small number of former smokers were
115 highly heterogenous (n=72, 33 cases, 39 controls; Figure S2). We excluded former smokers from
116 the final analysis, retaining 994 individuals (Table S1; Figure S1).
117
118 We aimed to address the relationship between NPC and the oral microbiome, adjusted for
119 potential confounders. As a result, we looked for factors which might affect the oral microbiome
120 at a community level. Our primary confounders included oral hygiene and health, tobacco use,

121 family history of NPC, and tea consumption (14, 16, 18, 19, 23–27). We also considered a
122 history of oropharyngeal inflammation, and the region where an individual lived and alcohol use,
123 as covariates primarily expected to affect the microbiome, as well as salted fish consumption,
124 which is primarily seen as a risk factor for NPC (15, 28, 29).
125
126 When comparing alpha diversity between cases and controls, we found that NPC cases showed
127 significantly fewer overall ASVs, reduced phylogenetic diversity, and reduced Shannon diversity
128 compared to controls (rank sum $p < 0.001$; Figure 1a; Table S2).; these findings did not change
129 after adjustment for covariates which were significantly associated with alpha diversity (Figure
130 1b; Tables S3-S5). Hence, this suggests that patients newly diagnosed with NPC have lower
131 overall microbial diversity than healthy controls. Our results agree with a small study of the oral
132 microbiome in NPC patients (n=90), which also found reduced alpha diversity (21). Unlike other
133 body sites, there is no clear relationship between salivary microbiome richness and the health of
134 the microbial community.
135
136 Similarly, when comparing global community patterns (beta-diversity) via Adonis models
137 minimally adjusted for sex, age and sequencing run, we found significant differences between
138 NPC cases and controls, both based on unweighted UniFrac distance as well as for weighted
139 UniFrac and Bray-Curtis distances (FDR $p< 0.001$, 9999 permutations; Figures 1c,d and S3a,b)
140 (30–32). Compared to the potential confounders in the same setting, NPC status was the
141 strongest explanatory factor for unweighted UniFrac distance, more than five times the effect
142 size of the next strongest variable, as well as the second-strongest factor for weighted UniFrac-
143 and Bray-Curtis distances, just after tobacco use. There was no statistically significant difference

144 in dispersion between cases and controls in any metric, supporting the idea that the differences
145 are due to consistent differences between cases and controls ($p > 0.55$, 999 permutations; Figure
146 1d). Significance persisted in more fully adjusted Adonis models including potential confounders
147 with robust differences in community patterns.
148
149 These findings establish that NPC status and smoking are strongly associated with differences in
150 the oral microbiome in our population; the association with NPC is especially strong with regard
151 to presence and absence of organisms (as emphasized by unweighted UniFrac), but second only
152 to smoking with regard to abundances (as captured by weighted UniFrac and Bray-Curtis). We
153 found no evidence that these associations are driven by community heterogeneity; they are,
154 however, robust under adjustment for observed confounders, and in the case of the unweighted
155 UniFrac distances, unlikely to be the result of confounding by unobserved factors due to the
156 crushing dominance of the signal for NPC status. Since we recruited incident, treatment-naive
157 patients, it is also implausible that the observed differences in microbiome composition are
158 treatment-related (21, 22). Taken together, our findings provide strong evidence for a clear
159 difference in the oral microbiome between patients with NPC and healthy controls.
160
161 Since the relationship between the microbiome and NPC status was strongest in unweighted
162 UniFrac distance, which focuses on presence and absence, we evaluated the relationship between
163 ASV prevalence and disease in a fully adjusted log binomial model. To limit spurious
164 correlations, we defined presence as a relative abundance greater than 0.02% and focused on
165 ASVs present in at least 10% of samples (n=245, Figure S4). We identified 53 ASVs which were
166 significantly different between cases and controls (FDR $p < 0.05$; Figure 1e; Table S6). The large

167 majority of these ASVs were more prevalent in controls and came from a wide variety of
168 taxonomic clades, which may suggest a somewhat stochastic loss of ASVs in NPC patients,
169 rather than a systematic loss of specific organisms (Table S6). This finding is in line with our
170 alpha diversity findings, and may indicate overall community instability. In contrast, two ASVs
171 were more prevalent in cases: a member of genus *Lactobacillus* (Lact-eca9) and a *Granulicatella*
172 ASV (Gran-7770).
173
174 To evaluate whether NPC status affected abundance-based partitioning of the microbial
175 community, we applied Phylofactor (33). Our model looked for phylogenetic clades which
176 differentiated NPC cases and controls, adjusting for potential confounders (Figure 2, Table S7).
177 Of the twelve factors examined, nine were associated with disease status. The primary partition
178 in the data suggested a *Granulicatella* ASV (Gran-7770) was 3.4 (95% CI 2.4, 4.9) fold more
179 abundant in NPC cases compared to controls. The third factor identified was second
180 *Granulicatella* ASV (Gran-5a37) as less abundant in cases. Both ASVs were also associated with
181 smoking status. We identified three large-scale shifts in microbial abundance associated with
182 NPC status. The remaining factors associated with NPC status were all single ASVs which
183 differentiated cases and controls, none of which differed in prevalence (Table S6, S7).
184
185 Based on the significant difference in abundance and prevalence of ASVs from genus
186 *Granulicatella* between cases and controls, we further explored this genus. We identified a total
187 of 14 ASVs in the dataset; three were prevalent enough to be included in our feature-based
188 analyses (Gran-5a37, Gran-7770, and Gran-6959). In 972 (97.8%) individuals, the abundant
189 ASVs were the only *Granulicatella* present. When blasted against the Human Oral Microbiome

190 Database (HOMD), the ASV sequences mapped to two cultured species with more than 99.5%
191 accuracy: *Granulicatella elegans* (*G. elegans*) which included Gran-6959 and *Granulicatella*
192 *adiacens* (*G. adiacens*; Gran-7770 and Gran-5a37) (34). Strikingly, we found our two abundant
193 *G. adiacens* ASVs differ by a single nucleotide: Gran-7770 carries a G at nucleotide 119 of our
194 sequence (corresponding approximately to 458 in the full length 16s rRNA sequence) while
195 Gran-5a37 carries an A.
196
197 Gran-7770 was found to be 26% more prevalent among cases, while Gran-5a37 was among the
198 51 ASVs less prevalent in cases (PR 0.81 [95% CI 0.74, 0.88]; Table S6]). Both ASVs were also
199 significantly associated with smoking status: Gran-7770 was more prevalent in smokers (PR
200 1.48, [95% CI 1.29, 1.70]) and Gran-5a37 less prevalent (PR 0.74, [95% CI 0.67, 0.81]). There
201 was not a significant relationship between Gran-6959 (*G. elegans*) and either disease status (PR
202 0.94 [95% CI 0.88, 1.00]) or tobacco use (PR 0.97 [95% CI 0.90, 1.06]).
203
204 We found that 993 out of 994 individuals carried at least one *G. adiacens* with a relative
205 abundance of at least 0.02%: 330 (33.2%) carried only Gran-5a37, 316 (31.8%) carried Gran-
206 7770 alone, and 347 (34.9%) carried both. Among individuals who were classified as carrying
207 only one ASV (Gran-7770 alone or Gran 5a37 alone), the "present" ASV was at least 50-fold
208 more abundant than the other variant. We used a multinomial logistic regression to confirm that
209 disease status was significantly associated with variants an individual carried: compared to the
210 odds of carrying Gran-5a37 alone, cases had significantly higher odds of carrying both ASVs
211 and, again, significantly higher odds of carrying Gran-7770 alone (Figure 3a). Although smokers

212 were more likely to have both ASVs or Gran-7770 alone, there was no significant interaction
213 between smoking and disease status.
214
215 We also investigated how the presence of a *G. adiacens* variant structured the overall microbial
216 community. We filtered the full ASV table to remove any *Granulicatella* ASVs and used the
217 reduced table to re-calculate beta diversity metrics. The *Granulicatella*-free community
218 recapitulated the patterns seen in the full community well (Mantel $R^2 > 0.91$; p=0.001, 999
219 permutations). We found significant differences between individuals who carried Gran-7770,
220 both, or Gran-5a37 in weighted and unweighted UniFrac distances and Bray Curtis; all three
221 metrics show clear separation in PCoA space (p=0.001, 999 permutations; Figure 3b; Figure S5).
222 In unweighted UniFrac space (Figure 3b), the separation was primarily along PC2, likely
223 corresponding to the separation along PC2 seen between cases and controls (Figure 1d).
224 Furthermore, we found that the *G. adiacens* variant explained 16% of the variation attributed to
225 case-control status in unweighted UniFrac distance and 15% of the variation in weighted
226 UniFrac distance. Our results suggest that the *G. adiacens* variant carried by an individual is
227 significantly associated with community structure, and may be a route by which NPC status
228 shapes the oral microbiome.
229
230 We used a SparCC-based network analysis to identify other community members *Granulicatella*
231 might interact with to exert an effect on the microbiome (35). We were able to identify five
232 networks: one pair of co-occurring ASVs, two pairs of co-excluding ASVs, one three-member
233 network of co-occurring ASVs and a large 29-member network of co-occurring and co-excluding
234 ASVs (Figures 4a). This main network consisted of two clusters of a total of 20 organisms which

235 were positively correlated with a *Granulicatella* variant; the main members of the networks
236 belonged to *Veillonella*, *Streptococcus*, and *Prevotella*. Blasting against HOMD, we identified
237 two additional pairs of ASVs that co-excluded between the two nodes but mapped to the same
238 clones: Stre-900d and Stre-0531 (*Streptococcus parasanguinis* clade 411) and *Prevotella*
239 *melaninogenica* (Prev-b7f2 and Prev-71e7; Figure 4b; Table S8) (34).
240
241 We hypothesize the co-excluding networks of ASVs, centered around *Granulicatella*, may
242 reflect partial niche specialization. Previous work suggests quorum sensing networks can form
243 between the core species, and that metabolic changes occur in these networks (36, 37). These
244 closely related organisms occupy the same niches within these metabolic networks, however,
245 strain-specific variation may either respond to or promote disease-associated transformation.
246 Without culture-based experimentation, it is difficult to determine how these organisms may
247 function in concert. One major challenge for in-silico validation is the limited resolution of
248 existing databases; our results exceed the OTU-based resolution used in construction and span a
249 less frequently characterized hypervariable region.
250
251 Within the context of NPC in an endemic region, we hypothesize the oral microbiome may act
252 through several potential mechanisms. The oral microbiome has been suggested to contribute to
253 local tumorigenesis through immune regulation or oncogenic metabolites such as acetaldehyde
254 or nitrosamines (38). An *in silico* study suggested that commercially available strains of *G.*
255 *adiacens* and co-abundant organisms encode genes involved in nitrate and nitrite reduction (39).
256

257 In summary, we have demonstrated a difference in the oral microbial community between NPC
258 patients and healthy controls in an endemic area of southern China, which cannot be explained
259 by other measured factors. The difference is associated with both a loss of community richness
260 and differences among specific organisms, including closely related ASVs from genus
261 *Granulicatella*. In addition, we identified a network of co-occurring and co-excluding ASVs
262 which included these *Granulicatella* variants. These results strongly suggest a relationship
263 between the oral microbiome and nasopharyngeal carcinoma status in untreated patients.
264

265 **Acknowledgements**

266

267 The authors which to thank the study participants, the field work team for the NPCGEE project,
268 and the Wuzhou Health System Key Laboratory for Nasopharyngeal Carcinoma Etiology and
269 Molecular Mechanism and the Key Laboratory of High-Incidence-Tumor Prevention &
270 Treatment (Guangxi Medical University), especially Suhua Zhong, Xiling Xiao, for the
271 processing of salivary samples. The data was stored in the Department of Medical Epidemiology
272 and Biostatistics; the authors wish to thank the IT group for their assistance.
273

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

## Data Availability

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

## Author contributions

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

## Methods

*Survey metadata and sample collection*

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

302 approved by the Institutional Review Board or Ethical Review Board at all participating centers.
303 All study participants provided written or oral informed consent.
304
305 A questionnaire covering demographics, diet, residential, occupational, medical and family
306 history was administered in a structured interview. Sample collection occurred at the interview.
307 Participants were asked not to eat nor chew gum for 30 minutes prior to sample collection. Saliva
308 samples (2ml-4ml) were collected into 50ml falcon tubes with a Tris-EDTA buffer.
309
310 Demographic characteristics of the study population were compared using a two-sided t-test for
311 continuous covariates (age) and a chi-squared test for categorical covariates. Tests were
312 conducted using scipy 0.19.1 in python 3.5.5 (40).
313
314 *DNA extraction, PCR, and sequencing*
315 Saliva DNA was extracted using a two-step protocol including the sample pre-processing with
316 lysozyme lysis and bead beating, and the TIANamp blood DNA kit (Beijing, China). The 16s
317 rRNA amplicon library was amplified with 341F/805R primers (CCTACGGGNGGCWGCAG,
318 GACTACHVGGGTATCTAATCC) (41, 42). Samples were amplified with 20 cycles of a
319 program with 30 seconds at 98°C for melting, 30 second at 60°C, and 30 seconds at 72°C.
320 Samples were barcoded in a second PCR step (41). DNA clean-up was performed using
321 Agentcourt AMPure XP purification kit. DNA volume and purity were measured on an Agilent
322 2100 Bioanalyzer system and Real-time polymerase chain reaction. Sequencing was performed
323 at Beijing Genome Institute (BGI) on an Illumina MiSeq using a 2x300bp paired end strategy.
324

325 *Denoising, Annotation and Filtering*

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

337

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

342

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

347

348 ASV-based analyses were performed on a representative subset: those with at least 0.02%
349 relative abundance in at least 10% of samples (n=245). A Mantel test was applied to Bray Curtis
350 distance showed a correlation of 0.96 between the filtered matrix rarefied to 5000
351 sequences/sample and the full table distance matrix (p=0.001, 999 permutations); the mantel
352 corresponding correlation for UniFrac distance was 0.76 (p=0.001, 999 permutations; Figure S3)
353 (30, 32, 50).
354
355 The sequences and identifiers for the abundant ASVs are listed in supplemental file 2. ASVs are
356 identified by the first 4 letters of their lowest taxonomic assignment and the first 4 characters of a
357 MD5 hash of the sequence. The full taxonomic assignment and MD5 hashes can be found in
358 Table S6.
359
360 *Diversity Analyses*
361 Diversity analyses were performed using samples rarefied to 6,500 sequences.
362
363 Alpha diversity was calculated as observed ASVs, Shannon diversity, and Faith's phylogenetic
364 diversity using q2-diversity in QIIME 2 (51, 52). Potentially significant alpha diversity
365 categories predictors were identified using a rank-sum test in scipy 0.19.1 (40). A p-value of 0.05
366 was considered the threshold for borderline significance for inclusion in a subsequent regression
367 model. Alpha diversity was then evaluated in a multivariate ordinary least squares (OLS)
368 regression model adjusted for age, sex and sequencing run number. A final model for each
369 metric was selected by forward selection using models which resulted in decreasing AIC. We
370 checked for the normality of residuals by plotting. The relative contribution of each covariate to

371 that metric was estimated by a "leave one out" approach. Regressions were performed in
372 Statsmodels (v. 0.9.0) (53). For visualization, we calculated z-normalized alpha diversity using
373 the mean and standard deviation in diversity for the controls. Alpha diversity was plotted using
374 boxenplots in Seaborn 0.9.0 (54, 55).
375
376 Beta diversity was measured using the unweighted UniFrac, weighted UniFrac, and Bray-Curtis
377 metrics on rarefied data (q2-diversity) (30–32). Beta diversity was compared using Adonis in the
378 R *vegan* library (v 2.5-2) adjusted for host age, sex, and sequencing run, with 9999 permutations
379 (56–58). We used a permdisp test with 999 permutations and the centroid estimate to test for the
380 presence of differences in within-group variation implemented is scikit-bio 0.5.4 (www.scikit-
381 bio.org).(59) Uncorrected p-values of less than 0.05 were considered to have significant
382 dispersion, since we were more concerned about false positives than false negatives. Principal
383 coordinate analyses (PCoA)s were visualized using Emperor (v. 1.0.0b18) and seaborn v. 0.9.0
384 in matplotlib v. 2.2.3 (54, 60).
385
386 *ASV regression model*
387 To look at the relationship between ASV prevalence and disease and smoking status, we used a
388 log binomial regression which was approximated via a Poisson regression with robust standard
389 errors, implemented via base function *glm* in R and the robust error mechanism implemented via
390 packages *lmtest* (v 0.9) and *sandwich* (v. 2.5) in R 3.5 (58, 61–63). The model was adjusted for
391 age, sex, sequencing run, residential community, and the number of missing or repaired teeth.
392 "Presence" was defined as a relative abundance of 1 / 5000, which corresponded to the
393 shallowest sequencing depth for the abundant counts. ASVs which were present in more than

394 1000 samples were excluded from prevalence analysis. A Benjamini-Hochberg FDR corrected p-
395 value of 0.05 was considered significant.
396
397 *Phylofactor*
398 Phylofactor (v. 0.01) was used to look at the relationship between disease status and
399 phylogenetic partitioning between clades (33). Phylofactor is a compositionally aware technique
400 which uses isometric log transforms over an unrooted phylogenetic tree to model differences in
401 the data. This allows the partitioning of data into polyphyletic clades. The Phylofator
402 multivariate model for each partition was modeled with an OLS regression considering
403 diagnosis, adjusted for residential community, age, sex, number of missing or repaired teeth,
404 tobacco use, and sequencing run. We looked at the first 12 factors using the default parameters,
405 which optimized for explaining maximal variance. The cladogram, and regression coefficient
406 plots were generated in seaborn (54).
407
408 *Granulicatella*
409 Total *Granulicatella* was identified by filtering the full ASV table for any ASV assigned to the
410 genus. Species-level assignments were made by blasting each ASV against the Human Oral
411 Microbiome Database using the online tool; species-level assignments were taken for the
412 cultured species with the best match (34). We treated the abundance of Gran-6959 as the *G.*
413 *elegens* abundance and the combined abundance of Gran-5a37 and Gran-7770 as the *G. adiacens*
414 abundance throughout. One sample which did not contain Gran-7770 or Gran-5a37 was excluded
415 from *G. adicans* associated analyses.
416

417 We used a multinomial logistic regression model, implemented in the *nnet* library (v. 0.8) in R to
418 look at whether the carriage of Gran-5a377 alone, Gran-7770 alone, or both ASVs was
419 associated with smoking and disease status (64). The regression was adjusted for age, sex,
420 sequencing run, number of missing or repaired teeth, residential community, the relative
421 abundance of *G. adiacens*, and the relative abundance of *G. elegens*. Having Gran-5a37 was
422 considered the reference group for the multinomial regression.
423
424 The effect of *Granulicatella* on alpha and beta diversity was calculated by first filtering out all
425 Granulicatella ASVs from the table, and then rarifying to 6250 sequences/sample before
426 diversity calculations. Adonis coefficients were calculated in a model adjusted for *G. adiacens*
427 abundance, sequencing run, age, sex, residential community, number of missing or repaired
428 teeth, tobacco use, and disease status. The proportion of disease status explained by comparing a
429 model excluding the Granulicatella variant minus the model including the variant over the model
430 excluding the variant.
431
432 *Network Analysis*
433 We used the Sparse Cooccurrence Network Investigation for Compositional data (SCNIC;
434 https://github.com/shafferm/SCNIC) in QIIME 2 (q2-SCNIC) to perform network analysis on
435 the abundant ASVs. The correlation network was built using SparCC, and the network was built
436 using edges with a correlation co-efficient of at least 0.3, allowing both co-occurrence and co-
437 exclusion (35). Network clusters were identified by finding the most connected node and
438 following all positively correlated nodes in the trimmed SparCC network. Networks were
439 visualized in Cytoscape (v. 3.7.1) using a perfuse-weighted network layout (65). Nodes which

440 were anti-correlated with a single node in the main cluster were trimmed for the sake of
441 visualization; these are labeled with the correlation coefficient.
442
443 The phylogenetic tree of core network members was visualized using ete3 (v. 3.1.1) in python
444 3.6 (66).
445
446
447 **References**
448

449 1. Abdul-Rasoul M, Habib H, Al-Khouly M. 2006. "The honeymoon phase" in children with
450 type 1 diabetes mellitus: frequency, duration, and influential factors. Pediatr Diabetes
451 7:101–7.
452 2. Vogtmann E, Goedert JJ. 2016. Epidemiologic studies of the human microbiome and
453 cancer. Br J Cancer 114:237–42.
454 3. Wroblewski LE, Peek RM, Wilson KT. 2010. Helicobacter pylori and gastric cancer:
455 Factors that modulate disease risk. Clin Microbiol Rev.
456 4. Bhatt AP, Redinbo MR, Bultman SJ. 2017. The role of the microbiome in cancer
457 development and therapy. CA Cancer J Clin 67:326–344.
458 5. Lu H, Wu Z, Xu W, Yang J, Chen Y, Li L. 2011. Intestinal microbiota of HBV cirrhotic
459 patients. Microb Ecol 61:693–703.
460 6. Ling Z, Liu X, Cheng Y, Jiang X, Jiang H, Wang Y, Li L. 2015. Decreased Diversity of
461 the Oral Microbiota of Patients with Hepatitis B Virus-Induced Chronic Liver Disease: A
462 Pilot Project. Sci Rep 5.

463 7. Guerrero-Preston R, Godoy-Vitorino F, Jedlicka A, Rodríguez-Hilario A, González H,
464 Bondy J, Lawson F, Folawiyo O, Michailidi C, Dziedzic A, Thangavel R, Hadar T,
465 Noordhuis MG, Westra W, Koch W, Sidransky D. 2016. 16S rRNA amplicon sequencing
466 identifies microbiota associated with oral cancer, human papilloma virus infection and
467 surgical treatment. Oncotarget 7:51320–51334.
468 8. Torcia MG. 2019. Interplay among Vaginal Microbiome, Immune Response and Sexually
469 Transmitted Viral Infections. Int J Mol Sci 20.
470 9. Li N, Ma WT, Pang M, Fan QL, Hua JL. 2019. The commensal microbiota and viral
471 infection: A comprehensive review. Front Immunol 10.
472 10. Wei K-R, Zheng R-S, Zhang S-W, Liang Z-H, Li Z-M, Chen W-Q. 2017. Nasopharyngeal
473 carcinoma incidence and mortality in China, 2013. Chin J Cancer 36:90.
474 11. Chua MLK, Wee JTS, Hui EP, Chan ATC. 2016. Nasopharyngeal carcinoma. Lancet
475 (London, England) 387:1012–1024.
476 12. Lieberman PM. 2014. Virology. Epstein-Barr virus turns 50. Science 343:1323–5.
477 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,
478 Liao C-T, Chang T-CJ. 2017. Familial aggregation of nasopharyngeal carcinoma in
479 Taiwan. Oral Oncol 73:10–15.
480 14. Long M, Fu Z, Li P, Nie Z. 2017. Cigarette smoking and the risk of nasopharyngeal
481 carcinoma: a meta-analysis of epidemiological studies. BMJ Open 7:e016582.
482 15. Barrett D, Ploner A, Chang ET, Liu Z, Zhang C-X, Liu Q, Cai Y, Zhang Z, Chen G,
483 Huang Q-H, Xie S-H, Cao S-M, Shao J-Y, Jia W-H, Zheng Y, Liao J, Chen Y, Lin L,
484 Ernberg I, Adami H-O, Huang G, Zeng Y, Zeng Y-X, Ye W. 2019. Past and Recent Salted
485 Fish and Preserved Food Intakes Are Weakly Associated with Nasopharyngeal Carcinoma

486 Risk in Adults in Southern China. J Nutr.

487 16. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan
488 A, Ley RE, Gevers D, Clark AG. 2015. Host genetic variation impacts microbiome
489 composition across human body sites. Genome Biol 16:191.

490 17. Wu J, Peters BA, Dominianni C, Zhang Y, Pei Z, Yang L, Ma Y, Purdue MP, Jacobs EJ,
491 Gapstur SM, Li H, Alekseyenko A V, Hayes RB, Ahn J. 2016. Cigarette smoking and the
492 oral microbiome in a large study of American adults. ISME J 10:2435–46.

493 18. Belstrøm D, Sembler-Møller ML, Grande MA, Kirkby N, Cotton SL, Paster BJ, Twetman
494 S, Holmstrup P. 2018. Impact of Oral Hygiene Discontinuation on Supragingival and
495 Salivary Microbiomes. JDR Clin Transl Res 3:57–64.

496 19. Kilian M, Chapple ILC, Hannig M, Marsh PD, Meuric V, Pedersen AML, Tonetti MS,
497 Wade WG, Zaura E. 2016. The oral microbiome – an update for oral healthcare
498 professionals. Br Dent J 221:657–666.

499 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,
500 Zheng Y, Liao J, Chen Y, Ernberg I, Vaughan TL, Adami H-O, Huang G, Zeng Y, Zeng
501 Y-X, Ye W. 2016. Oral Hygiene and Risk of Nasopharyngeal Carcinoma-A Population-
502 Based Case-Control Study in China. Cancer Epidemiol Biomarkers Prev 25:1201–7.

503 21. Zhu X-X, Yang X-J, Chao Y-L, Zheng H-M, Sheng H-F, Liu H-Y, He Y, Zhou H-W.
504 2017. The Potential Effect of Oral Microbiota in the Prediction of Mucositis During
505 Radiotherapy for Nasopharyngeal Carcinoma. EBioMedicine 18:23–31.

506 22. Ye W, Chang ET, Liu Z, Liu Q, Cai Y, Zhang Z, Chen G, Huang Q-H, Xie S-H, Cao S-M,
507 Shao J-Y, Jia W-H, Zheng Y, Liao J, Chen Y, Lin L, Liang L, Ernberg I, Vaughan TL,
508 Huang G, Zeng Y, Zeng Y-X, Adami H-O. 2017. Development of a population-based

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

510 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,
511 Zheng Y, Liao J, Chen Y, Ernberg I, Vaughan TL, Adami H-O, Huang G, Zeng Y, Zeng
512 Y-X, Ye W. 2016. Oral Hygiene and Risk of Nasopharyngeal Carcinoma-A Population-
513 Based Case-Control Study in China. Cancer Epidemiol Biomarkers Prev 25:1201–7.

514 24. Wu J, Peters BA, Dominianni C, Zhang Y, Pei Z, Yang L, Ma Y, Purdue MP, Jacobs EJ,
515 Gapstur SM, Li H, Alekseyenko A V, Hayes RB, Ahn J. 2016. Cigarette smoking and the
516 oral microbiome in a large study of American adults. ISME J 10:2435–46.

517 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,
518 Liao C-T, Chang T-CJ. 2017. Familial aggregation of nasopharyngeal carcinoma in
519 Taiwan. Oral Oncol 73:10–15.

520 26. Yuan X, Long Y, Ji Z, Gao J, Fu T, Yan M, Zhang L, Su H, Zhang W, Wen X, Pu Z,
521 Chen H, Wang Y, Gu X, Yan B, Kaliannan K, Shao Z. 2018. Green Tea Liquid
522 Consumption Alters the Human Intestinal and Oral Microbiome. Mol Nutr Food Res
523 62:e1800178.

524 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
525 P-J, Chen I-H, Yang C-S, Hildesheim A, Chen C-J. 2012. Lowered risk of nasopharyngeal
526 carcinoma and intake of plant vitamin, fresh fish, green tea and coffee: a case-control
527 study in Taiwan. PLoS One 7:e41779.

528 28. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, Chen M-X, Chen Z-H, Ji G-Y,
529 Zheng Z-D-X, Mujagond P, Chen X-J, Rong Z-H, Chen P, Lyu L-Y, Wang X, Wu C-B,
530 Yu N, Xu Y-J, Yin J, Raes J, Knight R, Ma W-J, Zhou H-W. 2018. Regional variation
531 limits applications of healthy gut microbiome reference ranges and disease models. Nat

532 Med 24:1532–1535.

533 29. Fan X, Peters BA, Jacobs EJ, Gapstur SM, Purdue MP, Freedman ND, Alekseyenko A V,
534 Wu J, Yang L, Pei Z, Hayes RB, Ahn J. 2018. Drinking alcohol is associated with
535 variation in the human oral microbiome in a large study of American adults. Microbiome
536 6:59.

537 30. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing
538 microbial communities. Appl Env Microbiol 71:8228–8235.

539 31. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and Qualitative
540 Diversity Measures Lead to Different Insights into Factors That Structure Microbial
541 Communities. Appl Environ Microbiol 73:1576–1585.

542 32. Sørensen T. 1948. A method of establishing groups of equal amplitude in plant sociology
543 based on similarity of species content and its application to analyses of the vegetation on
544 Danish commons. I kommission hos E. Munksgaard, København.

545 33. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N,
546 David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level
547 associations in microbiome datasets. PeerJ 5:e2969.

548 34. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. 2018. New Insights into
549 Human Nostril Microbiome from the Expanded Human Oral Microbiome Database
550 (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. mSystems
551 3:e00187-18.

552 35. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data.
553 PLoS Comput Biol 8:e1002687.

554 36. Chalmers NI, Palmer RJ, Cisar JO, Kolenbrander PE. 2008. Characterization of a

555 Streptococcus sp.-Veillonella sp. Community Micromanipulated from Dental Plaque. J
556 Bacteriol 190:8145–8154.
557 37. Palmer RJ, Diaz PI, Kolenbrander PE. 2006. Rapid succession within the Veillonella
558 population of a developing human oral biofilm in situ. J Bacteriol 188:4117–24.
559 38. Gholizadeh P, Eslami H, Yousefi M, Asgharzadeh M, Aghazadeh M, Kafil HS. 2016.
560 Role of oral microbiome on oral cancers, a review. Biomed Pharmacother 84:552–558.
561 39. Hyde ER, Andrade F, Vaksman Z, Parthasarathy K, Jiang H, Parthasarathy DK,
562 Torregrossa AC, Tribble G, Kaplan HB, Petrosino JF, Bryan NS. 2014. Metagenomic
563 analysis of nitrate-reducing bacteria in the oral cavity: implications for nitric oxide
564 homeostasis. PLoS One 9:e88645.
565 40. Jones E, Oliphant T, Peterson P, others. {SciPy}: Open source scientific tools for
566 {Python}.
567 41. Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. 2011.
568 Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea.
569 ISME J 5:1571–9.
570 42. Hugerth LW, Wefer HA, Lundin S, Jakobsson HE, Lindberg M, Rodin S, Engstrand L,
571 Andersson AF. 2014. DegePrime, a Program for Degenerate Primer Design for Broad-
572 Taxonomic-Range PCR in Microbial Ecology Studies. Appl Environ Microbiol 80:5116–
573 5123.
574 43. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open
575 source tool for metagenomics. PeerJ 4:e2584.
576 44. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander
577 H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A,

578 Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da
579 Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF,
580 Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A,
581 Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen
582 S, Jarmusch AK, Jiang L, Kaehler BD, Kang K Bin, Keefe CR, Keim P, Kelley ST,
583 Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X,
584 Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ,
585 Melnik A V., Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias
586 LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen
587 LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song
588 SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ,
589 Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M,
590 Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ,
591 Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive,
592 scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–
593 857.

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

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

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

601 DM, Orwoll E, Manary M, Mirarab S, Knight R. 2018. Phylogenetic Placement of Exact
602 Amplicon Sequences Improves Associations with Clinical Information. mSystems
603 3:e00021-18.
604 48. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL,
605 Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for
606 ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–8.
607 49. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid
608 assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol
609 73:5261–7.
610 50. Mantel N. 1967. The detection of disease clustering and a generalized regression
611 approach. Cancer Res 27:209–220.
612 51. Shannon CE, E. C. 2001. A mathematical theory of communication. ACM SIGMOBILE
613 Mob Comput Commun Rev 5:3.
614 52. Faith DP, Baker AM. 1992. Phylogenetic diversity (PD) and biodiversity conservation:
615 some bioinformatics challenges. Evol Bioinform Online 2:121–128.
616 53. JS Seabold JP. 2010. Statsmodels: Econometric and Statistical Modeling with Python.
617 Proc 9th Python Sci Conf.
618 54. Waskom M, Botvinnik O, O'Kane D, Hobson P, Ostblom J, Lukauskas S, Gemperline
619 DC, Augspurger T, Halchenko Y, Cole JB, Warmenhoven J, Ruiter J de, Pye C, Hoyer S,
620 Vanderplas J, Villalba S, Kunter G, Quintero E, Bachant P, Martin M, Meyer K, Miles A,
621 Ram Y, Brunner T, Yarkoni T, Williams ML, Evans C, Fitzgerald C, Brian, Qalieh A.
622 2018. mwaskom/seaborn: v0.9.0 (July 2018).
623 55. Hofmann H, Kafadar K, Wickham H. 2011. Letter-value plots: Boxplots for large dataThe

American Statistican.

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.

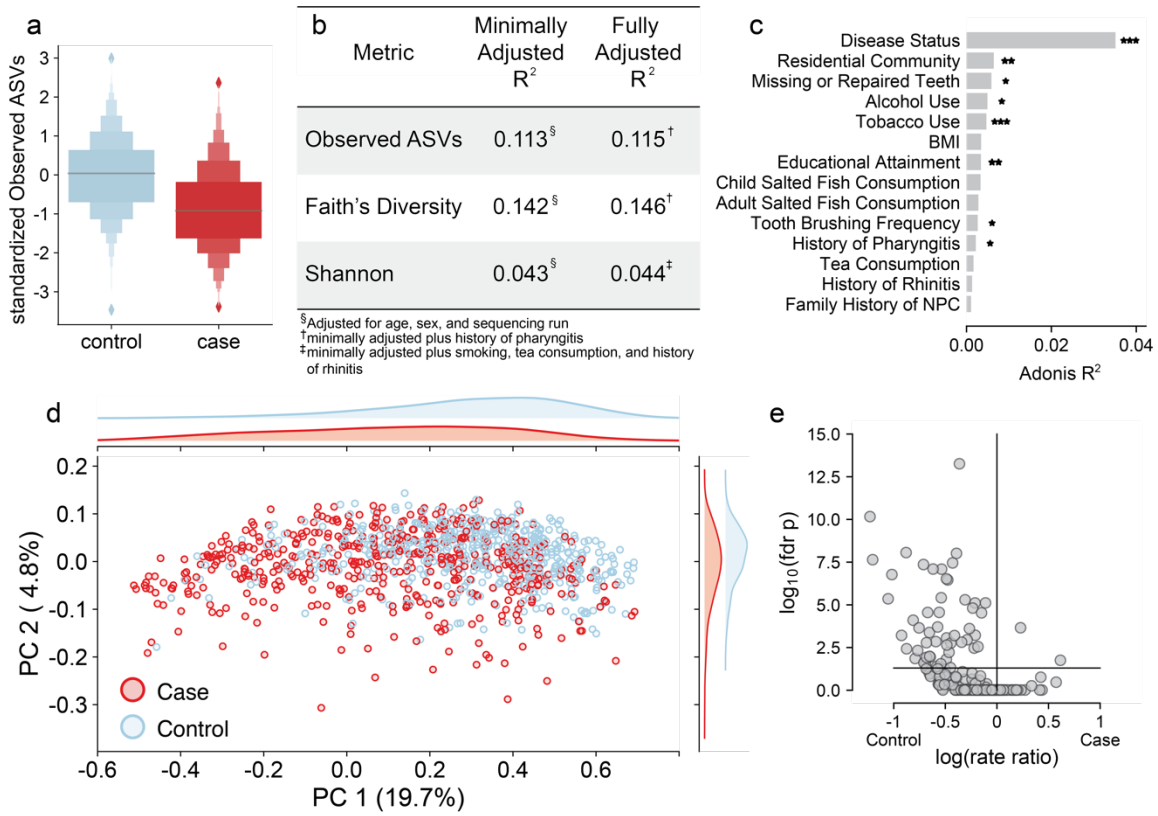
61. Barros AJ, Hirakata VN. 2003. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol 3:21.

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 SFourth. 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




| Metric | Minimally Adjusted $R^2$ | Fully Adjusted $R^2$ |
|---|---|---|
| Observed ASVs | 0.113[§] | 0.115[†] |
| Faith's Diversity | 0.142[§] | 0.146[†] |
| Shannon | 0.043[§] | 0.044[‡] |

[§]Adjusted for age, sex, and sequencing run
[†]minimally adjusted plus history of pharyngitis
[‡]minimally adjusted plus smoking, tea consumption, and history of rhinitis

**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; $p < 1x10^{-12}$). 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

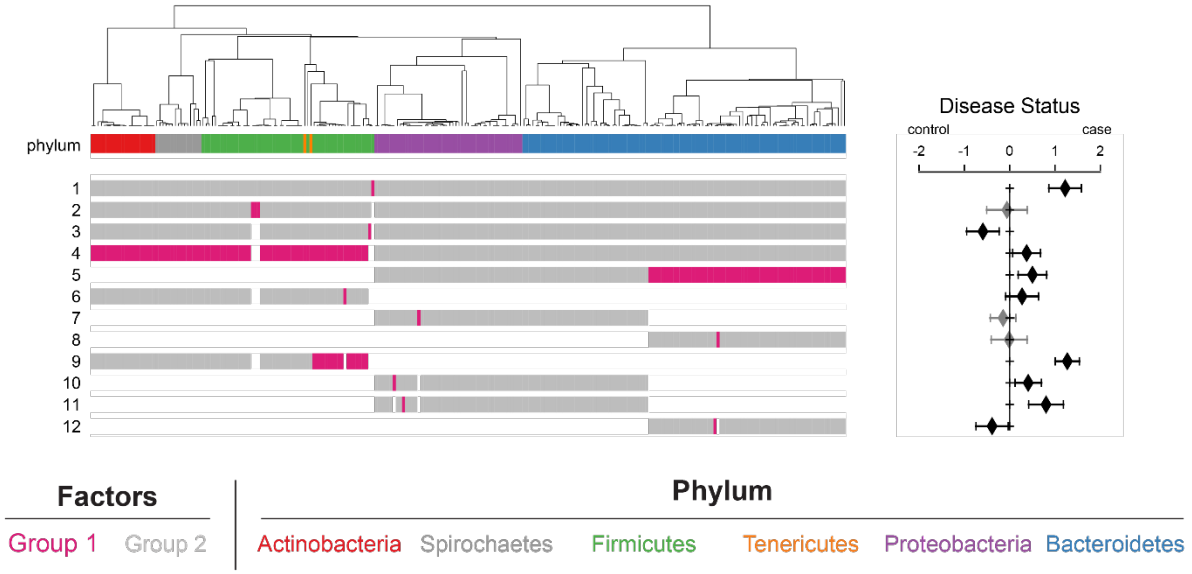



**Figure 2. There are significant associations between phylogenetic partitioning of the taxa and NPC status.**

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 $a < 0.05$, gray indicates a non-significant association.

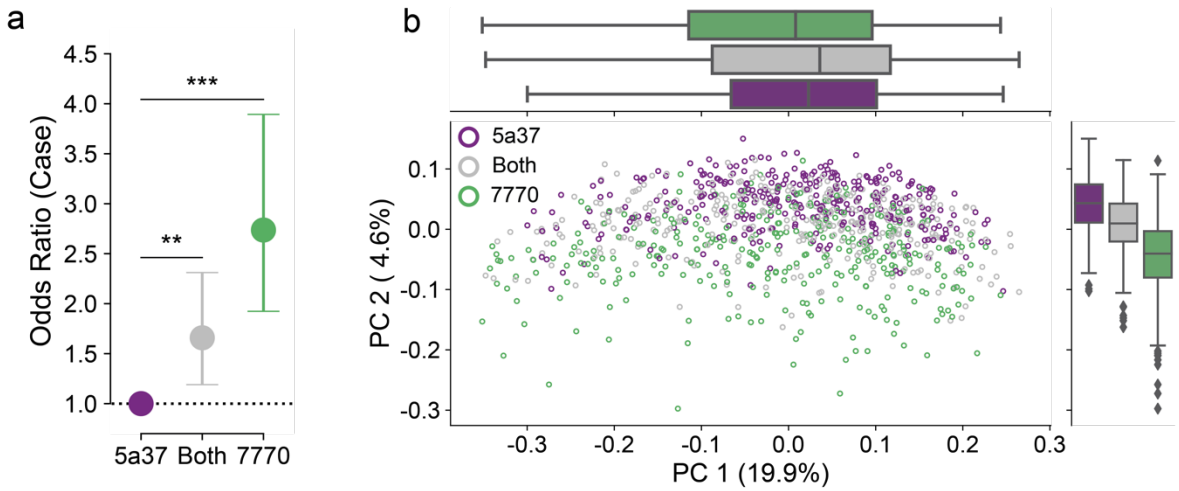



**Figure 3. The *Granulicatella adiacens* variant predicts community structure.** (**a**) NPC cases have significantly higher odds of carrying both Gran-5a37 and Gran-7770 than Gran-5a37 alone, and again, significantly higher odds than carrying either Gran-5a37 and Gran-7770 or Gran-7770. (**b**) In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

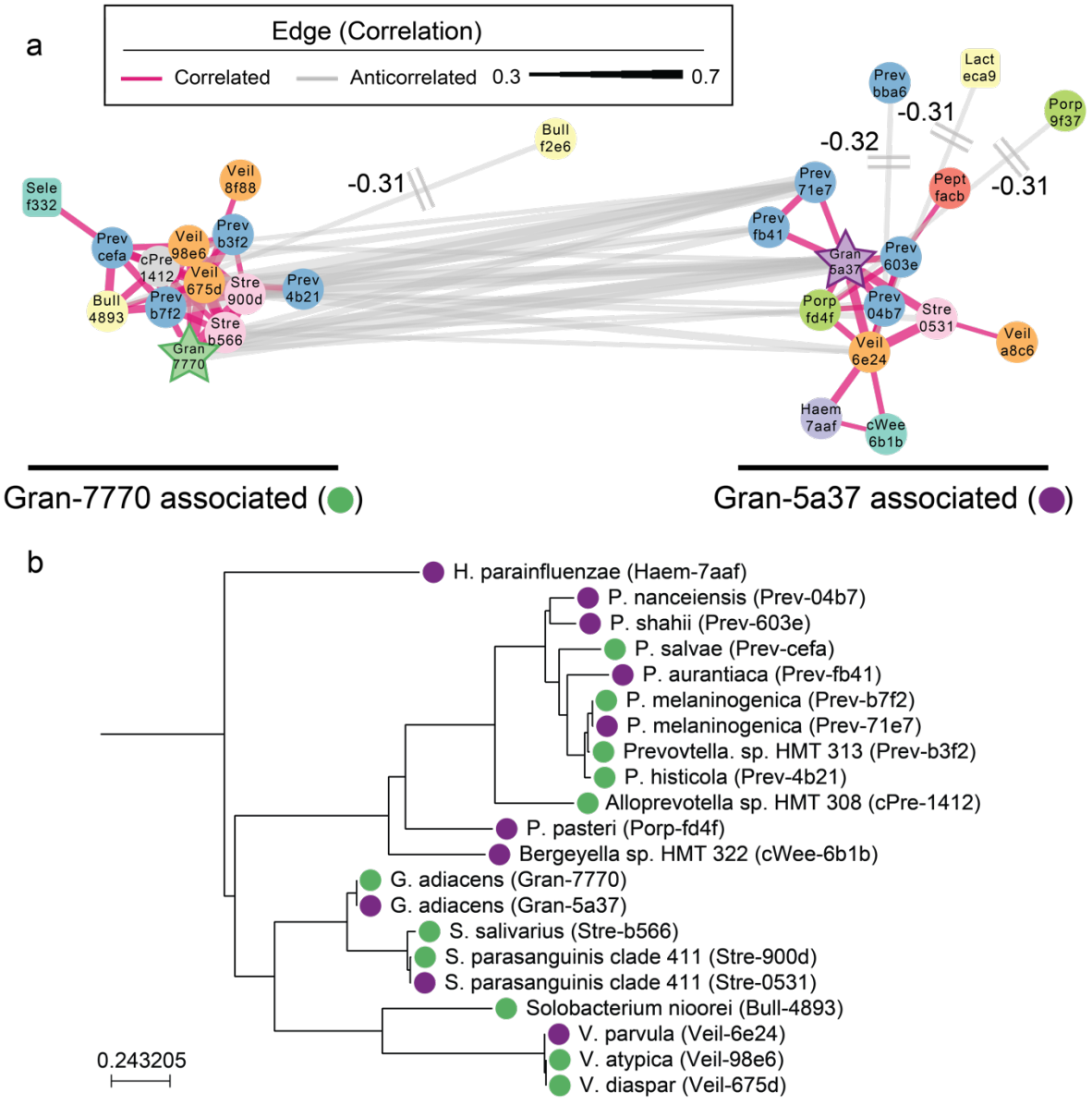



Figure 4. ***Granulicatella adiacens*** **variants set at the center of a network of closely related co-occurring organisms. (a)** SparCC-based network analysis for co-occurring and co-excluding ASVs for all subjects showed a large network with two clusters with common core structures. The color and 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