Sub-species niche specialization in the oral microbiome is associated with nasopharyngeal carcinoma risk in an endemic area of southern China

Justine W. Debelius[1] *, Tingting Huang[1, 2] *, Yonglin Cai[3, 4] *, Alexander Ploner[1], Donal Barrett[1], Xiaoying Zhou[5, 6], Xue Xiao[7], Yancheng Li[3, 4], Jian Liao[8], Yuming Zheng[3, 4], Guangwu Huang[7], Hans-Olov Adami[1,9], Yi Zeng[10 §], Zhe Zhang[7 §], Weimin Ye[1 §]

[1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[2] Department of Radiation Oncology, The First Affiliated Hospital of Guangxi Medical University, Nanning, P. R. China

[3] Department of Cancer Prevention Center, Wuzhou Red Cross Hospital, Wuzhou, P. R. China;

[4] Wuzhou Health System Key Laboratory for Nasopharyngeal Carcinoma Etiology and Molecular Mechanism, Wuzhou, P. R. China

[5] Life Science Institute, Guangxi Medical University, Nanning, P. R. China;

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

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

[8] Cangwu Institute for Nasopharyngeal Carcinoma Control and Prevention, Wuzhou, P. R. China

[9]Clinical Effectiveness Research Group, Institute of Health, University of Oslo, Oslo, Norway

26 [10] State Key Laboratory for Infectious Diseases Prevention and Control, Institute for Viral
27 Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing,
28 P. R. China
29
30
31 Weimin Ye - Department of Medical Epidemiology and Biostatistics, Karolinska Institutet,
32 Nobels väg 12A, PO Box 281, Stockholm, SE-171 77, Sweden. Tel: +46-8-5248 6184; E-
33 mail: weimin.ye@ki.se.
34 Zhe Zhang - Department of Otolaryngology-Head & Neck Surgery, First Affiliated Hospital
35 of Guangxi Medical University, Nanning, P. R. China (zhangzhe@gxmu.edu.cn)
36 * *First authors Contributed equally;*
37 § *Last authors who contributed equally.*

Summary

Nasopharyngeal carcinoma (NPC) is a globally rare cancer, with a unique geographic distribution. In endemic areas including Southern China, the incidence is more than 20 times higher than the rest of the world.[1] Although recent evidence suggests poor oral hygiene is a risk factor for NPC,[2] it remains unknown whether the disease status is associated with changes in the oral microbiome. Therefore, we carried out a population-based case-control study in an endemic area of southern China.[3] We analyzed microbial communities from 499 untreated incident NPC cases and 495 age and sex frequency-matched controls. Here, we show the oral microbiome is altered in patients with NPC: patients have lower microbial diversity and significant changes in the overall structure of their microbial communities which cannot be attributed to other factors. Furthermore, the combination of two closely related amplicon sequence variants (ASVs) from *Granulicatella adiacens* an individual carried were predicted by disease status. These ASVs sat at the center of a network of closely-related co-excluding organisms, suggesting that NPC may be associated with subtle changes in the oral microbiome.

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.[3] Saliva was collected during interview. After sequencing and denoising to ASVs, samples from 1066 subjects 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).

62 We aimed to address the relationship between NPC and the oral microbiome, adjusted for
63 potential confounders. As a result, we looked for factors which might affect the oral
64 microbiome at a community level. Our primary confounders included oral hygiene and
65 health,[2,4,5] tobacco use,[6,7] family history of NPC,[8,9] alcohol use,[10,11] and tea consumption.[12,13]
66 We also considered a history of oropharyngeal inflammation, and the region where an
67 individual lived[14] as covariates primarily expected to affect the microbiome, as well as salted
68 fish consumption, which is primarily seen as a risk factor for NPC.[15]
69
70 When comparing alpha diversity between cases and controls, we found that NPC cases
71 showed significantly fewer overall ASVs, reduced phylogenetic diversity, and reduced
72 Shannon diversity compared to controls (rank sum $p < 0.001$; Figure 1a; Table S2); these
73 findings did not change after adjustment for covariates which were significantly associated
74 with alpha diversity (Figure 1b; Tables S3-S5). Hence, this suggests that patients newly
75 diagnosed with NPC have lower overall microbial diversity than healthy controls. Our results
76 agree with a small study of the oral microbiome in NPC patients (n=90), which also found
77 reduced alpha diversity.[16] Unlike other body sites, there is no clear relationship between
78 salivary microbiome richness and the health of the microbial community.
79

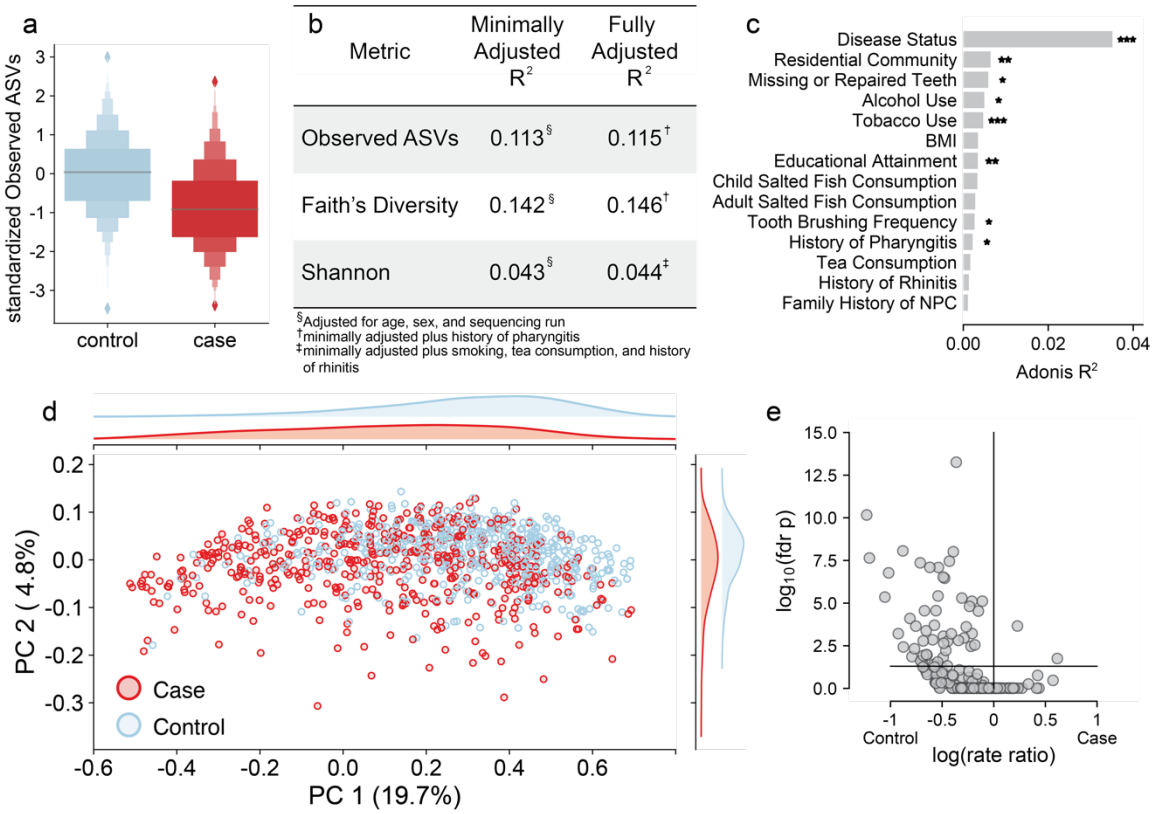

Figure 1. The oral microbiome differs between patients with nasopharyngeal carcinoma and healthy controls. (**a**) NPC cases (red) have significantly lower microbial richness compared to cases (blue; $p < 1x10^{-12}$). The horizontal line in the boxlin represents the median, the large box the interquartile region, increasingly smaller boxes are the upper and lower eighths, sixteenths, 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 UniFrac, PC1 explains 19.7% and PC2 explains 4.8% of the variation. A volcano plot of (e) the Poisson regression coefficient for disease status vs the log p-value reflects reduced diversity. The horizontal line indicates significant at a Benjamini-Hochberg corrected p-value of less than 0.05.

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[17] as well as for weighted UniFrac[18] and Bray-Curtis distances (FDR $p< 0.001$, 9999 permutations; Figures 1c,d and S3a,b). 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

104 UniFrac- and Bray-Curtis distances, just after tobacco use. There was no statistically
105 significant difference in dispersion between cases and controls in any metric, supporting the
106 idea that the differences are due to consistent differences between cases and controls ($p >$
107 0.55, 999 permutations; Figure 1d). Significance persisted in more fully adjusted Adonis
108 models including potential confounders with robust differences in community patterns.
109
110 These findings establish that NPC status and smoking are strongly associated with
111 differences in the oral microbiome in our population; the association with NPC is especially
112 strong with regard to presence and absence of organisms (as emphasized by unweighted
113 UniFrac), but second only to smoking with regard to abundances (as captured by weighted
114 UniFrac and Bray-Curtis). We found no evidence that these associations are driven by
115 community heterogeneity; they are, however, robust under adjustment for observed
116 confounders, and in the case of the unweighted UniFrac distances, unlikely to be the result of
117 confounding by unobserved factors due to the crushing dominance of the signal for NPC
118 status. Since we recruited incident, treatment-naive patients,[3,16] it is also implausible that the
119 observed differences in microbiome composition are treatment-related. Taken together, our
120 findings provide strong evidence for a clear difference in the oral microbiome between
121 patients with NPC and healthy controls.
122
123 Since the relationship between the microbiome and NPC status was strongest in unweighted
124 UniFrac distance, which focuses on presence and absence, we evaluated the relationship
125 between ASV prevalence and disease in a fully adjusted log binomial model. To limit
126 spurious correlations, we defined presence as a relative abundance greater than 0.02% and
127 focused on ASVs present in at least 10% of samples (n=245, Figure S4). We identified 53
128 ASVs which were significantly different between cases and controls (FDR $p < 0.05$; Figure

129 1e; Table S6). The large majority of these ASVs were more prevalent in controls and came
130 from a wide variety of taxonomic clades, which may suggest a somewhat stochastic loss of
131 ASVs in NPC patients, rather than a systematic loss of specific organisms (Table S6). This
132 finding is in line with our alpha diversity findings, and may indicate overall community
133 instability. In contrast, two ASVs were more prevalent in NPC cases: a member of genus
134 *Lactobacillus* (Lact-eca9) and a *Granulicatella* ASV (Gran-7770).
135
136 To evaluate whether NPC status affected abundance-based partitioning of the microbial
137 community, we applied Phylofactor.[19] Our model looked for phylogenetic clades which
138 differentiated NPC cases and controls, adjusting for potential confounders (Figure 2, Table
139 S7). Of the twelve factors examined, nine were associated with disease status. The primary
140 partition in the data suggested a *Granulicatella* ASV (Gran-7770) was 3.4 (95% CI 2.4, 4.9)
141 fold more abundant in NPC cases compared to controls. The third factor identified was
142 second *Granulicatella* ASV (Gran-5a37) as less abundant in cases. Both ASVs were also
143 associated with smoking status. We identified three large-scale shifts in microbial abundance
144 associated with NPC status. The remaining factors associated with NPC status were all single
145 ASVs which differentiated cases and controls, none of which differed in prevalence (Table
146 S6, S7).
147

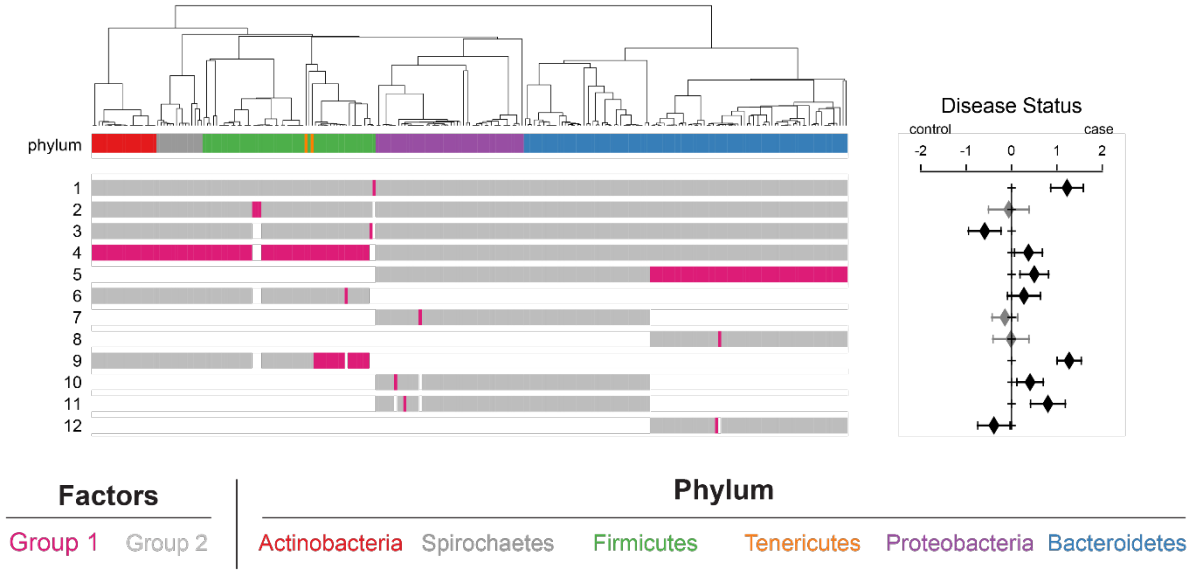

Figure 2. There are significant associations between phylogenetic partitioning of the taxa and NPC status. The phylogenetic tree with the first 12 phylofactor-based clade partitions is shown on the left. The top row is colored by phylum, the associated color is shown below. The isometric log transformation 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 intervals for the regression coefficient. Black bars indicate significance at $a < 0.05$, gray indicates a non-significant association.

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 to their corresponding assignment: *Granulicatella elegans* (*G. elegans*) which included Gran-6959 and *Granulicatella adiacens* (*G. adiacens*; Gran-7770 and Gran-5a37).[20] 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 16s rRNA sequence) while Gran-5a37 carries an A.

171 Gran-7770 was found to be 26% more prevalent among cases, while Gran-5a37 was among
172 the 51 ASVs less prevalent in cases (Prevalence Ratio [PR] 0.81 [95% CI 0.74, 0.88]; Table
173 S6]). Both ASVs were also significantly associated with smoking status: Gran-7770 was
174 more prevalent in smokers (PR 1.48, [95% CI 1.29, 1.70]) and Gran-5a37 less prevalent (PR
175 0.74, [95% CI 0.67, 0.81]). There was not a significant relationship between Gran-6959 (*G.*
176 *elegans*) and either disease status (PR 0.94 [95% CI 0.88, 1.00]) or tobacco use (PR 0.97
177 [95% CI 0.90, 1.06]).
178
179 We found that 993 out of 994 individuals carried at least one *G. adiacens* with a relative
180 abundance of at least 0.02%: 330 (33.2%) carried only Gran-5a37, 316 (31.8%) carried Gran-
181 7770 alone, and 347 (34.9%) carried both. Among individuals who were classified as
182 carrying only one ASV (Gran-7770 alone or Gran 5a37 alone), the "present" ASV was at
183 least 50-fold more abundant than the other variant. We used a multinomial logistic regression
184 to confirm that disease status was significantly associated with variants an individual carried:
185 compared to the odds of carrying Gran-5a37 alone, cases had significantly higher odds of
186 carrying both ASVs and, again, significantly higher odds of carrying Gran-7770 alone
187 (Figure 3a). Although smokers were more likely to have both ASVs or Gran-7770 alone,
188 there was no significant interaction between smoking and disease status.
189
190

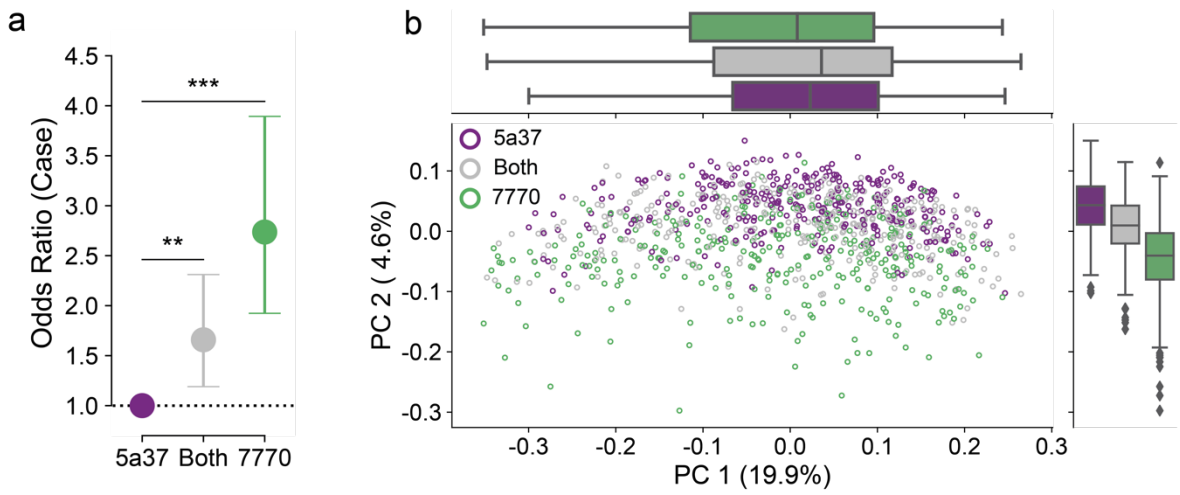

Figure 3. The *Granulicatella adiacens* variant predicts community structure. (**a**) NPC cases have significantly higher odds of carrying both Gran-5a37 and Gran-7770 than Gran-5a37 alone, and again, significantly higher odds than carrying either Gran-5a37 and Gran-7770 or Gran-7770. (**b**) In unweighted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

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.

212 We used a SparCC-based network analysis to identify other community members
213 *Granulicatella* might interact with to exert an effect on the microbiome.[21] We were able to
214 identify five networks: one pair of co-occurring ASVs, two pairs of co-excluding ASVs, one
215 three-member network of co-occurring ASVs and a large 29-member network of co-
216 occurring and co-excluding ASVs (Figures 4a). This main network consisted of two clusters
217 of a total of 20 organisms which were positively correlated with a *Granulicatella* variant; the
218 main members of the networks belonged to *Veillonella*, *Streptococcus*, and *Prevotella*.
219 Blasting against HOMD, we identified two additional pairs of ASVs that co-excluded
220 between the two nodes but mapped to the same clones: Stre-900d and Stre-0531
221 (*Streptococcus parasanguinis* clade 411) and *Prevotella melaninogenica* (Prev-b7f2 and
222 Prev-71e7; Figure 4b; Table S8).[20]
223
224 We hypothesize the co-excluding networks of ASVs, centered around *Granulicatella*, may
225 reflect partial niche specialization. Previous work suggests quorum sensing networks can
226 form between the core species,[22,23] and that metabolic changes occur in these networks. We
227 hypothesize these closely correlated organisms occupy the same niches within these
228 metabolic networks, however, strain-specific variation may either respond to or promote
229 disease-associated transformation. Without culture-based experimentation, it is difficult to
230 determine how these organisms may function in concert. One major challenge for in-silico
231 validation is the limited resolution of existing databases; our results exceed the OTU-based
232 resolution and span a less frequently characterized hypervariable region.
233

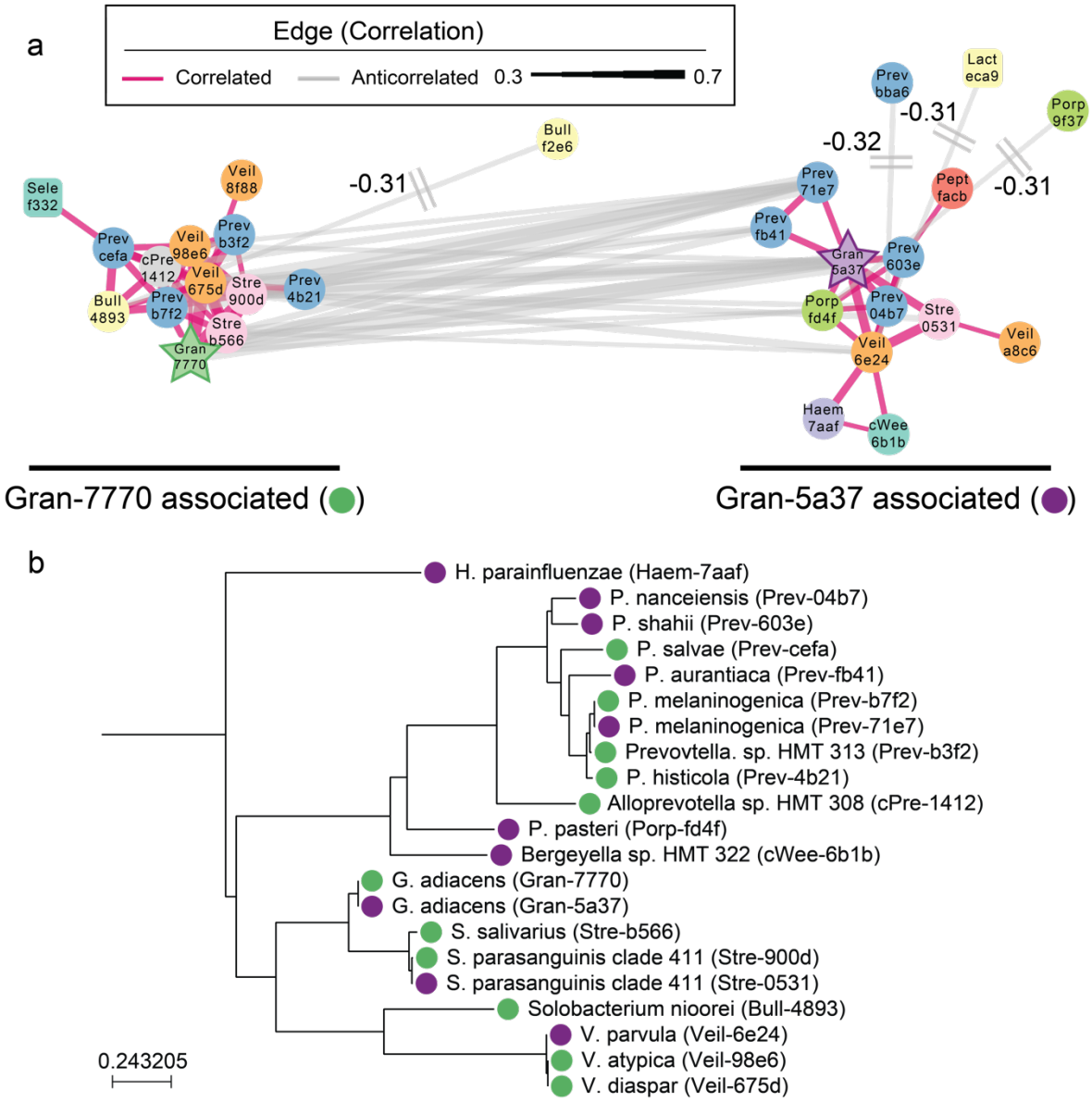

Figure 4. ***Granulicatella adiacens* variants set at the center of a network of closely related co-occurring organisms. (a)** SparCC-based network analysis for co-occurring and co-excluding ASVs for all subjects showed a large network with two clusters with common core structures. The color and shape 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 correlated with either Gran-7770 or Gran-5a37). Tips are labeled by their association with Gran-7770 (Green) or Gran-5a37 (Purple).

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.[24] An *in silico* study suggested that commercially available

248 strains of *G. adiacens* and co-abundant organisms encode genes involved in nitrate and nitrite
249 reduction.[25]
250
251 Alternatively, we propose the possibility of an NPC-specific mechanism, in which the
252 microbiome interacts with the Epstein-Barr Virus (EBV). Infection with EBV is the most
253 widely accepted etiological factor for NPC, and butyrate, a well-known product of microbial
254 fermentation, has been linked to EBV reactivation,[26] a necessary step in NPC oncogenesis.[27]
255 The local microbiota has also been suggested to be involved in the acquisition and
256 persistence of oncogenic viral infections at other sites, for example, the interaction between
257 the vaginal microbiome and the human papillomavirus.[28] We therefore hypothesize the oral
258 microbiome and potentially the nasopharyngeal microbiome, may work in concert to lead to
259 high risk EBV infection in the nasopharyngeal epithelium, leading to NPC. However,
260 prospective studies are needed to determine whether the microbiome contributes to EBV
261 infection, or if differences in the oral microbiota only reflect EBV infection and NPC-related
262 stress.
263
264 In summary, we have demonstrated a difference in the oral microbial community between
265 NPC patients and healthy controls in an endemic area of southern China, which cannot be
266 explained by other measured factors. The difference is associated with both a loss of
267 community richness and differences among specific organisms, including closely related
268 ASVs from genus *Granulicatella*. In addition, we identified a network of co-occurring and
269 co-excluding ASVs which included these *Granulicatella* variants. These results strongly
270 suggest a relationship between the oral microbiome and nasopharyngeal carcinoma status in
271 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, Karolinska Institutet; the authors wish to thank them for their assistance.

We acknowledge funding from the Swedish Research Council (2015-02625, 2015-06268, 2017-05814, PI Dr. W. Ye); the National Natural Science Foundation of China (81272983, PI Dr. Z. Zhang); and the Guangxi Natural Science Foundation (2013GXNSFGA019002, PI Dr. Z. Zhang). The field work of the NPCGEE study was funded by the National Cancer Institute of the NIH (Award Number R01CA115873, PI H.-O. Adami). T. Huang is partly supported by a grant from China Scholarship Council.

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 HA, YZ, GH, ZZ and WY. YC, DB, WY, TH, JWD, and AP refined the study design for this project. YC, YL, JL and YZ were responsible for sample collection and management. DB performed the lab work, supervised by TH, XZ, XX, ZZ, and WY. Bioinformatics and biostatistical analyses were performed by JWD; TH and AP

297 contributed to statistical modeling and refinement. WY contributed to the supervision and
298 coordination of the project. JWD and TH wrote the manuscript; AP provided critical edits.
299 All authors reviewed and approved the final submission.
300
301 **Methods**
302
303 *Survey metadata and sample collection*
304 Participant recruitment has been previously described.[3] Briefly, incident cases of NPC in
305 Guangdong Province and Guangxi Autonomous Region between 2010 and 2013 were invited
306 to participate in the study. Age and sex matched controls were selected from the total
307 population. The current study was approved by the Institutional Review Board or Ethical
308 Review Board at all participating centers. All study participants provided written or oral
309 informed consent.
310
311 A questionnaire covering demographics, diet, residential, occupational, medical and family
312 history was administered in a structured interview. Sample collection occurred at the
313 interview. Participants were asked not to eat nor chew gum for 30 minutes prior to sample
314 collection. Saliva samples with volumes (2ml-4ml) were collected into 50ml falcon tubes
315 with a Tris-EDTA buffer.
316
317 Demographic characteristics of the study population were compared using a two-sided t-test
318 for continuous covariates (age) and a chi-squared test for categorical covariates. Tests were
319 conducted using scipy 0.19.1[29] in python 3.5.5.
320
321 *DNA extraction, PCR, and sequencing*

322 Saliva DNA was extracted using a two-step protocol including the sample pre-processing
323 with lysozyme lysis and bead beating, and the TIANamp blood DNA kit (Beijing, China).
324 The 16s rRNA amplicon library was amplified with 341F/805R primers
325 (CCTACGGGNGGCWGCAG, GACTACHVGGGTATCTAATCC).[30,31] Samples were
326 amplified with 20 cycles of a program with 30 seconds at 98°C for melting, 30 second at
327 60°C, and 30 seconds at 72°C. Samples were barcoded in a second PCR step.[30] DNA clean-
328 up was performed using Agentcourt AMPure XP purification kit. DNA volume and purity
329 were measured on an Agilent 2100 Bioanalyzer system and Real-time polymerase chain
330 reaction. Sequencing was performed at Beijing Genome Institute (BGI) on an Illumina MiSeq
331 using a 2x300bp paired end strategy.
332
333 *Denoising, Annotation and Filtering*

334 Samples were demultiplexed using an in-house script. Adaptors were trimmed and paired end
335 sequences were joined using VSEARCH (v. 2.7).[32] Paired sequences were loaded into the
336 November 2018 release of QIIME 2.[33] Sequences were quality filtered (q2-quality-filter)[34]
337 and denoised using deblur (v. 1.0.4; q2-deblur)[35] with the default parameters on 420 bp
338 amplicons to generate amplicon sequence variants (ASVs). A phylogenetic tree was built
339 using fragment insertion into the August 2013 Greengenes 99% identity tree backbone with
340 q2-fragment-insertion;[36,37] taxonomic assignments were made with a naïve Bayesian
341 classifier trained against the same reference (q2-feature-classifier).[38] In cases where the
342 classifier or reference database was unable to describe a taxonomic level (for instance, a
343 missing genus), the taxonomy was described by inheriting the lowest defined level using a
344 custom python script. Following sequencing and denoising, 24,763,933 high quality reads
345 were retained.
346

347 Any sample with fewer than 1000 reads after denoising was excluded, leaving 1074 saliva
348 samples and 9 negative or single organism controls. Additionally, samples missing
349 information on tobacco use, defined information about tooth brushing frequency, or an
350 undefined residential region (n=8) were excluded (Figure S1).
351
352 Preliminary investigation suggested that the microbial communities for former smokers
353 (*n*=72) were highly heterogenous (Figure S2). Sensitivity analyses suggest their exclusion
354 does not alter the major community-level differences. Therefore, they were excluded, leaving
355 a total of 994 individuals in the analysis.
356
357 ASV-based analyses were performed on a representative subset: those with at least 0.02%
358 relative abundance in at least 10% of samples (n=245). A Mantel test[39] was applied to Bray
359 Curtis distance[40] and showed a correlation of 0.96 between the filtered matrix rarefied to
360 5000 sequences/sample and the full table distance matrix (p=0.001, 999 permutations); the
361 mantel corresponding correlation for UniFrac distance[41] was 0.76 (p=0.001, 999
362 permutations; Figure S3).
363
364 The sequences and identifiers for the abundant ASVs are listed in supplemental file 2. ASVs
365 are identified by the first 4 letters of their lowest taxonomic assignment and the first 4
366 characters of a MD5 hash of the sequence. The full taxonomic assignment and MD5 hashes
367 can be found in Table S6.
368
369 *Diversity Analyses*
370 Diversity analyses were performed using samples rarefied to 6,500 sequences.
371

372 Alpha diversity was calculated as observed ASVs, Shannon diversity,[42] and Faith's
373 phylogenetic diversity[43] using q2-diversity in QIIME 2. Potentially significant alpha diversity
374 predictors were identified using a rank-sum test in scipy 0.19.1.[29] A p-value of 0.05 was
375 considered the threshold for borderline significance for inclusion in a subsequent regression
376 model. Alpha diversity was then evaluated in a multivariate ordinary least squares (OLS)
377 regression model adjusted for age, sex and sequencing run number. A final model for each
378 metric was selected by forward selection using models which resulted in decreasing Akaike
379 information criterion (AIC). We checked for the normality of residuals by plotting. The
380 relative contribution of each covariate to that metric was estimated by a "leave one out"
381 approach. Regressions were performed in Statsmodels (v. 0.9.0).[44] For visualization, we
382 calculated z-normalized alpha diversity using the mean and standard deviation in diversity for
383 the controls. Alpha diversity was plotted using boxenplots in Seaborn 0.9.0.[45,46]
384
385 Beta diversity was measured using the unweighted UniFrac,[17] weighted UniFrac,[18] and Bray-
386 Curtis[40] metrics on rarefied data (q2-diversity). Beta diversity was compared using Adonis in
387 the R vegan library (v 2.5-2) adjusted for host age, sex, and sequencing run, with 9999
388 permutations.[47–49] We used a permdisp test with 999 permutations and the centroid estimate
389 to test for the presence of differences in within-group variation implemented in scikit-bio
390 0.5.4 (www.scikit-bio.org).[50] Uncorrected p-values of less than 0.05 were considered to have
391 significant dispersion, since we were more concerned about false positives than false
392 negatives. Principal coordinate analyses (PCoA)s were visualized using Emperor[51] (v.
393 1.0.0b18) and with seaborn[45] v. 0.9.0 in matplotlib v. 2.2.3.
394
395
396 *ASV regression model*

397 To look at the relationship between ASV prevalence and disease and smoking status, we used
398 a log binomial regression which was approximated via a Poisson regression with robust
399 standard errors,[52] implemented via base function *glm* in R and the robust error mechanism
400 implemented via packages *lmtest* (v 0.9) and *sandwich* (v. 2.5) in R 3.5.[49,53,54] The model was
401 adjusted for age, sex, sequencing run, residential community, and the number of missing or
402 repaired teeth. "Presence" was defined as a relative abundance of 1 / 5000, which
403 corresponded to the shallowest sequencing depth for the abundant counts. ASVs which were
404 present in more than 1000 samples were excluded from prevalence analysis. A Benjamini-
405 Hochberg FDR corrected p-value of 0.05 was considered significant.
406
407 *Phylofactor*
408 Phylofactor (v. 0.01) was used to look at the relationship between disease status and
409 phylogenetic partitioning between clades.[19] Phylofactor is a compositionally aware technique
410 which uses isometric log transforms over an unrooted phylogenetic tree to model differences
411 in the data. This allows the partitioning of data into polyphyletic clades. The Phylofactor
412 multivariate model for each partition was modeled with an OLS regression considering
413 diagnosis, adjusted for residential community, age, sex, number of missing or repaired teeth,
414 tobacco use, and sequencing run. We looked at the first 12 factors using the default
415 parameters, which optimized for explaining maximal variance. The cladogram, and
416 regression coefficient plots were generated in seaborn.[45]
417
418 *Granulicatella*
419 Total *Granulicatella* was identified by filtering the full ASV table for any ASV assigned to
420 the genus. Species-level assignments were made by blasting each ASV against the Human
421 Oral Microbiome Database using the online tool;[20] species-level assignments were taken for

422 the cultured species with the best match. We treated the abundance of Gran-6959 as the *G.*
423 *elegens* abundance and the combined abundance of Gran-5a37 and Gran-7770 as the *G.*
424 *adiacens* abundance throughout.
425
426 We used a multinomial logistic regression model, implemented in the *nnet* library (v. 0.8) in
427 R to look at whether the carriage of Gran-5a377 alone, Gran-7770 alone, or both ASVs was
428 associated with smoking and disease status.[55] The regression was adjusted for age, sex,
429 sequencing run, number of missing or repaired teeth, residential community, the relative
430 abundance of *G. adiacens*, and the relative abundance of *G. elegens*. Having Gran-5a37 was
431 considered the reference group for the multinomial logistic regression.
432
433 The effect of *Granulicatella* on alpha and beta diversity was calculated by first, filtering out
434 all *Granulicatella* ASVs from the table, and then rarifying to 6250 sequences/sample before
435 diversity calculations. Adonis coefficients were calculated in a model adjusted for *G.*
436 *adiacens* abundance, sequencing run, age, sex, residential community, number of missing or
437 repaired teeth, tobacco use, and disease status. The proportion of disease status explained by
438 comparing a model excluding the *Granulicatella* variant minus the model including the
439 variant over the model excluding the variant.
440
441 *Network Analysis*
442 We used the Sparse Cooccurrence Network Investigation for Compositional data (SCNIC;
443 https://github.com/shafferm/SCNIC) in QIIME 2 (q2-SCNIC) to perform network analysis on
444 the abundant ASVs in current and never smokers. The correlation network was built using
445 SparCC, and the network was built using edges with a correlation co-efficient of at least 0.3,
446 allowing both co-occurrence and co-exclusion.[21] Network clusters were identified by finding

447 the most connected node and following all positively correlated nodes in the trimmed SparCC
448 network. Networks were visualized in Cytoscape (v. 3.7.1) using a perfuse-weighted network
449 layout.[56] Nodes which were anti-correlated with a single node in the main cluster were
450 trimmed for the sake of visualization; these are labeled with the correlation coefficient.
451
452 The phylogenetic tree of core network members was visualized using ete3 (v. 3.1.1) in
453 python 3.6.[57]
454
455 **References**

456 1. Wei, K.-R. *et al.* Nasopharyngeal carcinoma incidence and mortality in China, 2013.
457 *Chin. J. Cancer* **36**, 90 (2017).
458 2. Liu, Z. *et al.* Oral Hygiene and Risk of Nasopharyngeal Carcinoma-A Population-
459 Based Case-Control Study in China. *Cancer Epidemiol. Biomarkers Prev.* **25**, 1201–7
460 (2016).
461 3. Ye, W. *et al.* Development of a population-based cancer case-control study in southern
462 china. *Oncotarget* **8**, 87073–87085 (2017).
463 4. Kilian, M. *et al.* The oral microbiome – an update for oral healthcare professionals. *Br.*
464 *Dent. J.* **221**, 657–666 (2016).
465 5. Belstrøm, D. *et al.* Impact of Oral Hygiene Discontinuation on Supragingival and
466 Salivary Microbiomes. *JDR Clin. Transl. Res.* **3**, 57–64 (2018).
467 6. Long, M., Fu, Z., Li, P. & Nie, Z. Cigarette smoking and the risk of nasopharyngeal
468 carcinoma: a meta-analysis of epidemiological studies. *BMJ Open* **7**, e016582 (2017).
469 7. Wu, J. *et al.* Cigarette smoking and the oral microbiome in a large study of American
470 adults. *ISME J.* **10**, 2435–46 (2016).
471 8. Huang, S.-F. *et al.* Familial aggregation of nasopharyngeal carcinoma in Taiwan. *Oral*

472 *Oncol.* **73**, 10–15 (2017).

473 9. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across
474 human body sites. *Genome Biol.* **16**, 191 (2015).

475 10. Chen, L. *et al.* Alcohol Consumption and the Risk of Nasopharyngeal Carcinoma: A
476 Systematic Review. *Nutr. Cancer* **61**, 1–15 (2009).

477 11. Fan, X. *et al.* Drinking alcohol is associated with variation in the human oral
478 microbiome in a large study of American adults. *Microbiome* **6**, 59 (2018).

479 12. Yuan, X. *et al.* Green Tea Liquid Consumption Alters the Human Intestinal and Oral
480 Microbiome. *Mol. Nutr. Food Res.* **62**, e1800178 (2018).

481 13. Hsu, W.-L. *et al.* Lowered risk of nasopharyngeal carcinoma and intake of plant
482 vitamin, fresh fish, green tea and coffee: a case-control study in Taiwan. *PLoS One* **7**,
483 e41779 (2012).

484 14. He, Y. *et al.* Regional variation limits applications of healthy gut microbiome
485 reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).

486 15. Barrett, D. *et al.* Past and Recent Salted Fish and Preserved Food Intakes Are Weakly
487 Associated with Nasopharyngeal Carcinoma Risk in Adults in Southern China. *J. Nutr.*
488 (2019). doi:10.1093/jn/nxz095

489 16. Zhu, X.-X. *et al.* The Potential Effect of Oral Microbiota in the Prediction of Mucositis
490 During Radiotherapy for Nasopharyngeal Carcinoma. *EBioMedicine* **18**, 23–31 (2017).

491 17. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing
492 microbial communities. *Appl Env. Microbiol* **71**, 8228–8235 (2005).

493 18. Lozupone, C. A., Hamady, M., Kelley, S. T. & Knight, R. Quantitative and Qualitative
494 Diversity Measures Lead to Different Insights into Factors That Structure Microbial
495 Communities. *Appl. Environ. Microbiol.* **73**, 1576–1585 (2007).

496 19. Washburne, A. D. *et al.* Phylogenetic factorization of compositional data yields

497 lineage-level associations in microbiome datasets. *PeerJ* **5**, e2969 (2017).

498 20. Escapa, I. F. *et al.* New Insights into Human Nostril Microbiome from the Expanded
499 Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the
500 Human Aerodigestive Tract. *mSystems* **3**, e00187-18 (2018).

501 21. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data.
502 *PLoS Comput. Biol.* **8**, e1002687 (2012).

503 22. Chalmers, N. I., Palmer, R. J., Cisar, J. O. & Kolenbrander, P. E. Characterization of a
504 Streptococcus sp.-Veillonella sp. Community Micromanipulated from Dental Plaque.
505 *J. Bacteriol.* **190**, 8145–8154 (2008).

506 23. Palmer, R. J., Diaz, P. I. & Kolenbrander, P. E. Rapid succession within the
507 Veillonella population of a developing human oral biofilm in situ. *J. Bacteriol.* **188**,
508 4117–24 (2006).

509 24. Gholizadeh, P. *et al.* Role of oral microbiome on oral cancers, a review. *Biomed.*
510 *Pharmacother.* **84**, 552–558 (2016).

511 25. Hyde, E. R. *et al.* Metagenomic analysis of nitrate-reducing bacteria in the oral cavity:
512 implications for nitric oxide homeostasis. *PLoS One* **9**, e88645 (2014).

513 26. Luka, J., Kallin, B. & Klein, G. Induction of the Epstein-Barr virus (EBV) cycle in
514 latently infected cells by n-butyrate. *Virology* **94**, 228–231 (1979).

515 27. Hirayama, T. & Ito, Y. A new view of the etiology of nasopharyngeal carcinoma.
516 *Prev. Med. (Baltim).* **10**, 614–22 (1981).

517 28. Mitra, A. *et al.* The vaginal microbiota, human papillomavirus infection and cervical
518 intraepithelial neoplasia: what do we know and where are we going next? *Microbiome*
519 **4**, 58 (2016).

520 29. Jones, E., Oliphant, T., Peterson, P. & others. SciPy: Open source scientific tools for
521 Python.

522 30. Herlemann, D. P. *et al.* Transitions in bacterial communities along the 2000 km
523 salinity gradient of the Baltic Sea. *ISME J.* **5**, 1571–9 (2011).
524 31. Hugerth, L. W. *et al.* DegePrime, a Program for Degenerate Primer Design for Broad-
525 Taxonomic-Range PCR in Microbial Ecology Studies. *Appl. Environ. Microbiol.* **80**,
526 5116–5123 (2014).
527 32. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile
528 open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
529 33. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data
530 science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
531 34. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from
532 Illumina amplicon sequencing. *Nat. Methods* **10**, 57–9 (2013).
533 35. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence
534 Patterns. *mSystems* **2**, e00191-16 (2017).
535 36. Janssen, S. *et al.* Phylogenetic Placement of Exact Amplicon Sequences Improves
536 Associations with Clinical Information. *mSystems* **3**, e00021-18 (2018).
537 37. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for
538 ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**, 610–8 (2012).
539 38. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for
540 rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ.*
541 *Microbiol.* **73**, 5261–7 (2007).
542 39. Mantel, N. The detection of disease clustering and a generalized regression approach.
543 *Cancer Res.* **27**, 209–220 (1967).
544 40. Sørensen, T. *A method of establishing groups of equal amplitude in plant sociology*
545 *based on similarity of species content and its application to analyses of the vegetation*
546 *on Danish commons.* (I kommission hos E. Munksgaard, 1948).

547 41. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing
548 microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–35 (2005).
549 42. Shannon, C. E. & E., C. A mathematical theory of communication. *ACM SIGMOBILE*
550 *Mob. Comput. Commun. Rev.* **5**, 3 (2001).
551 43. Faith, D. P. & Baker, A. M. Phylogenetic diversity (PD) and biodiversity conservation:
552 some bioinformatics challenges. *Evol Bioinform Online* **2**, 121–128 (1992).
553 44. JS Seabold, J. P. Statsmodels: Econometric and Statistical Modeling with Python.
554 *Proc. 9th Python Sci. Conf.* (2010).
555 45. Waskom, M. *et al.* mwaskom/seaborn: v0.9.0 (July 2018). (2018).
556 doi:10.5281/ZENODO.1313201
557 46. Hofmann, H., Kafadar, K. & Wickham, H. *Letter-value plots: Boxplots for large data*.
558 *The American Statistican* (2011).
559 47. McArdle, B. H. & Anderson, M. J. FITTING MULTIVARIATE MODELS TO
560 COMMUNITY DATA: A COMMENT ON DISTANCE-BASED REDUNDANCY
561 ANALYSIS. *Ecology* **82**, 290–297 (2001).
562 48. Oksanen, J. *et al.* vegan: Community Ecology Package. (2018).
563 49. R Core Team. R: A Language and Environment for Statistical Computing. (2018).
564 50. Anderson, M. J. Distance-Based Tests for Homogeneity of Multivariate Dispersions.
565 *Biometrics* **62**, 245–253 (2006).
566 51. Vázquez-Baeza, Y. *et al.* EMPeror: a tool for visualizing high-throughput microbial
567 community data. *Gigascience* **2**, 16 (2013).
568 52. Barros, A. J. & Hirakata, V. N. Alternatives for logistic regression in cross-sectional
569 studies: an empirical comparison of models that directly estimate the prevalence ratio.
570 *BMC Med. Res. Methodol.* **3**, 21 (2003).
571 53. Zeileis, A. Object-Oriented Computation of Sandwich Estimators. *J. Stat. Softw.* **16**,

572 1–16 (2006).

573 54. Zeileis, A. Econometric Computing with {HC} and {HAC} Covariance Matrix

574 Estimators. *J. Stat. Softw.* **11**, 1–17 (2004).

575 55. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, 2002).

576 56. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of

577 biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).

578 57. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and

579 Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

580