medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

The mouth of America: the oral microbiome profile of the US population

Anil K. Chaturvedi, PhD [1*], Emily Vogtmann, PhD, MPH [1*], Jianxin Shi, PhD [1], Yukiko Yano, PhD, MPH [1], Martin J. Blaser, MD [2], Nicholas A. Bokulich, PhD [3], J. Gregory Caporaso, PhD [4], Maura L. Gillison, MD, PhD [5], Barry I. Graubard, PhD [1], Xing Hua, PhD [1], Autumn G. Hullings, PhD, MPH [6], Lisa Kahle, BA [7], Rob Knight, PhD [8], Shilan Li, PhD [9], Jody McLean, MPH [10], Vaishnavi Purandare, MS [1], Yunhu Wan, PhD [1], Neal D. Freedman, PhD, MPH [11*], and Christian C. Abnet, PhD, MPH [1*]

[*] These authors contributed equally to the work

[1] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD., USA.
[2] Center for Advanced Biotechnology and Medicine, Rutgers, Piscataway, NJ, USA.
[3] Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland.
[4] Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA.
[5] Department of Thoracic and Head and Neck Medical Oncology, MD Anderson Cancer Center, Houston, TX, USA.
[6] Department of Nutrition, University of North Carolina, Chapel Hill, NC, USA.
[7] Information Management Services, Calverton, MD, USA
[8] Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA.
[9] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA.
[10] National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD. USA.
[11] Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD., USA.

Manuscript details

Abstract: 350
Text: 3,046
Tables and Figures: 5 figures
References: 35

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Funding: This study was funded in part by the Intramural Program of the US National Cancer Institute, National Institutes of Health, by a grant from the US Food and Drug Administration's Center for Tobacco Products (to CCA), and by an NIH National Cancer Institute Informatics Technology for Cancer Research Award (to JGC, 1U24CA248454-01). The NHANES oral HPV protocol that collected oral rinse samples was supported by a grant from the National Institute for Dental and Craniofacial Research (NIDCR) to MLG.

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Correspondence:

Christian C. Abnet, PhD
Director, Metabolic Epidemiology Branch
Division of Cancer Epidemiology and Genetics, National Cancer Institute
9609 Medical Center Drive, Rm. 6E410
Rockville, MD. 20852
Email: abnetc@mail.nih.gov

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

ABSTRACT

Importance: The oral microbiome is increasingly recognized to play key roles in human health and disease; yet, population-representative characterizations are lacking.

Objective: Characterize the composition, diversity, and correlates of the oral microbiome among US adults.

Design: Cross-sectional population-representative survey.

Setting: The National Health and Nutrition Examination Survey (NHANES, 2009-2012), a stratified multistage probability sample of the US population.

Participants: NHANES participants aged 18-69 years (n=8,237, representing 202,314,000 individuals).

Exposures: Demographic, socioeconomic, behavioral, anthropometric, metabolic, and clinical characteristics.

Main outcomes: Oral microbiome, characterized through 16S rRNA sequencing. Microbiome metrics were alpha diversity (number of observed Amplicon Sequence Variants [ASV], Faith's Phylogenetic diversity, Shannon-Weiner Index, and Simpson Index); beta diversity (unweighted UniFrac, weighted UniFrac, and Bray-Curtis dissimilarity); and prevalence and relative abundance at taxonomic levels (phylum through genus). Analyses accounted for the NHANES complex sample design.

Results: Among US adults aged 18-69 years, the oral microbiome encompassed 37 bacterial phyla, 99 classes, 212 orders, 446 families, and 1,219 genera. Five phyla—*Firmicutes, Actinobacteria, Bacteroidetes, Proteobacteria,* and *Fusobacteria* and six genera—*Veillonella, Streptococcus, Prevotella7, Rothia, Actinomyces, and Gemella*, were present in nearly all US adults (weighted-prevalence >99%). These genera also

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

were the most abundant, accounting for 65.7% of abundance. Observed ASVs showed a quadratic pattern with age (peak at 30 years), was similar by sex, significantly lower among non-Hispanic White individuals, and increased with higher body mass index (BMI) categories, alcohol use, and periodontal disease severity. All covariates together accounted for a modest proportion of oral microbiome variability, as measured by beta diversity (unweighted UniFrac=8.7%, weighted UniFrac=7.2%, and Bray-Curtis=6.3%). By contrast, relative abundance of a few genera explained a high percentage of variability in beta diversity (weighted UniFrac: *Aggregatibacter*=22.4%, *Lactococcus*=21.6%, *Haemophilus*=18.4%). Prevalence and relative abundance of numerous genera were significantly associated (Bonferroni-corrected Wald-p<0.0002) with age, race and ethnicity, smoking, BMI categories, alcohol use, and periodontal disease severity.

Conclusions: We provide a contemporary reference standard for the oral microbiome of the US adult population. Our results indicate that a few genera were universally present in US adults and a different set of genera explained a high percentage of oral microbiome diversity across the population.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

BACKGROUND

The community of bacteria in the human oral cavity, the oral microbiome, is increasingly recognized to play a role in health and disease.[1-3] The oral microbiome is involved in digestion and metabolism of nutrients (e.g., dietary nitrates, carbohydrates) and carcinogens (e.g., conversion of ethanol to acetaldehyde), blood pressure and metabolic homeostasis,[1,4,5] and regulation of oral and systemic immunity and inflammation.[1,4] Importantly, certain oral bacteria contribute to periodontitis and poor oral health,[6] and alterations in the oral microbiome have been associated with several chronic diseases, including cancer (upper-aerodigestive, colorectal), diabetes, cardiovascular disease, and neurological disorders.[1,3,4,7-11]

Understanding the relevance of the oral microbiome in health and disease requires a population reference standard of the composition of the oral bacterial ecosystem and prevalence, relative abundance, and interrelationships of taxa, established through characterization of the oral microbiome in a representative sample/population. Such a reference standard could serve as a critical comparator to understand the disease relevance of alterations in the oral microbiome. However, few population-representative characterizations exist in the literature.[12] Most prior studies of the oral microbiome have been conducted in non-representative groups or highly selective subgroups. [2,13-15]

Here, we address the lack of population-representative studies through characterization of the composition and diversity of the oral microbiome as well as demographic, socioeconomic, behavioral, metabolic, and clinical correlates of the oral microbiome in a representative sample of the US adult population aged 18-69 years.[16]

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

METHODS

Study population and biospecimens

We conducted this study using data from the National Health and Nutrition Examination Survey (NHANES), a complex stratified multistage cluster probability sample of the civilian, non-institutionalized US population.[17] The survey includes a household interview component for collection of data on demographics, diet, tobacco use, and medical history and a mobile examination center (MEC) component for health, dental, anthropometric, and biochemical examinations and biospecimen collection.[17]

We conducted this study using oral rinse samples collected for oral human papillomavirus (HPV) evaluations in two consecutive NHANES cycles (2009-2010 and 2011-2012).[18] Briefly, NHANES incorporated an oral HPV protocol in four cycles (2009-2016) for participants aged 14-69 years.[19,20] Response rates for the MEC component were 68.5% and 69.5% in 2009-2010 and 2011-2012, respectively. The NHANES protocol was approved by the National Center for Health Statistics (NCHS) Ethics Review Board (ERB)[21] and all participants provided written informed consent.

Biospecimens and laboratory methods

Participants were provided a 10 mL oral rinse (Scope or saline) sample and completed 3 alternating swish and gargle cycles of 5 seconds each. Oral rinses were shipped weekly to the Gillison laboratory (Ohio State University) and stored at 4°C until further processing.[18] Oral rinses were pelleted, lysed, and digested using DNase-free RNase A and proteinase K. DNA was extracted on the QIAsymphony SP instrument using the QIAGEN Virus/Bacteria Midi kit and Pathogen Complex 800 program and stored at -20°C (**Supplemental methods**).[18]

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

16S rRNA gene sequencing

Extracted DNA from n=9,997 participants was shipped to the Knight laboratory (University of California, San Diego) on dry ice and the V4 region of the 16S rRNA gene was PCR-amplified using the 515F-806R primers and sequenced using Illumina HiSeq 2500 with paired-end sequencing of 125 bp;[22] however, only data from forward sequencing runs were used for analyses due to a lack of overlapping reads. A total of 19 individual HiSeq sequencing runs were conducted, with repeated sequencing of runs, plates, or plate-lanes, as necessary (**Supplemental methods**).

Of note, NCHS/NHANES policy at the time (circa 2015) did not allow targeting human DNA for sequencing due to participant confidentiality concerns, which precluded metagenomic sequencing. Consequently, our data lack species-level resolution and imputed bacterial functionality. Also, there are geographic bacterial niches within the oral cavity;[1,2,23,24] thus, our results from oral rinses need to be interpreted as the generalized/sentinel microbiome profile of the biofilms within the oral cavity.

Details regarding quality control (QC) samples are described in the **Supplemental methods**. QC analyses showed acceptable test-retest reproducibility; importantly, given no evidence for plate/batch effects we did not employ batch normalization procedures (**Figures S1-S2**).

Bioinformatics methods and microbiome metrics

The FASTQ files were demultiplexed using QIIME 1.[25] The forward reads from each sequencing run were processed using DADA2 and Amplicon Sequence Variants (ASV) were predicted.[26,27] The phylogenetic tree was generated using QIIME 2 and taxonomy was assigned using the SILVA v123 database.[28,29] For primary analyses, we

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

created an ASV table that filtered out a non-bacterial ASV (SV1032, mapping to a mitochondrial pseudogene) that was identified in data quality checks (**Supplemental methods)**.

The resulting ASV tables were rarefied 10 times at 10,000 reads per sample and averaged for calculation of alpha and beta diversity metrics. We calculated four alpha diversity metrics using ASV tables: number of observed ASVs, a measure of richness of samples; Faith's Phylogenetic Diversity [Faith's PD], a phylogeny-weighted richness; Shannon-Weiner Index, a measure of evenness of samples; and Simpson Index, a measure of both richness and evenness of samples (**Supplemental Methods**).

We also calculated three beta diversity distance matrices for n x n comparisons across participants as measures of differences in the bacterial community: unweighted UniFrac, which considers presence/absence of bacterial ASVs to estimate phylogeny-weighted distance between participants; weighted UniFrac, which considers relative abundance of bacterial ASVs to estimate phylogeny-weighted distance between participants; and Bray-Curtis dissimilarity, which considers relative abundance of bacterial ASVs without phylogenetic weighting (**Supplemental Methods**).

Taxonomic analyses were based on non-rarefied data but were restricted to samples with at least 10,000 reads. Bacterial taxonomy was characterized across phylum (L2), class (L3), order (L4), family (L5), and genus (L6) levels, with data expressed as prevalence (presence/absence) and relative abundance (ratio of the number of each taxon-specific sequences and sum of all taxon-specific sequences in an individual). We primarily present results at the phylum (L2) and genus (L6) levels; characterizations at all levels are presented in **Tables S1-S6**.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Statistical methods

All analyses accounted for the NHANES complex sampling design through the use of the cycle-specific NHANES sample weights (examination weights adjusted for non-response) and the strata and multistage clustered sample design. We created post-stratification adjustments to the cycle-specific examination weights to account for non-availability of oral microbiome results (18.4%) from MEC participants aged 18-69 years (**Table S7**). Specifically, NHANES MEC weights were post-stratified to the age by sex by race/ethnicity cycle-specific distribution of the US population. Post-stratified weights were divided by two given the use of two NHANES cycles.

We describe the structure and diversity of the oral microbiome through descriptive statistics. Results are presented as percentages, means, medians, and quantiles, as appropriate. Taxa by taxa relative abundance correlations were calculated using sample weighted Pearson's correlation coefficients (**Tables S8-S12**). Population totals were estimated using post-stratified weights.

Analyses considered demographic and socioeconomic factors (age, modeled using restricted 5-knot cubic regression splines, sex, race and ethnicity, education, marital status, income-to-poverty ratio), anthropometrics (measured body mass index [BMI], categorized), risk behaviors (smoking, alcohol use), medical conditions (diabetes and hypertension), oral health (periodontal disease, tooth count, and edentulism), and use of prescription medications within the past 30 days (antibiotics, antilipidemics, respiratory inhalants, and for gastroesophageal reflux). Race and ethnicity was self-reported as Mexican American, Other Hispanic, non-Hispanic White, non-Hispanic Black, and non-Hispanic other race and ethnicity (including Native American, Pacific

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Islander, Asian and multi-racial). These variables were chosen a priori for adjustment in statistical models (variable details and code in **Table S13**).

Correlates of the four alpha diversity metrics were evaluated using multiple linear regression with each metric as the dependent variable. Results from these models are expressed as predictive margins, analogous to adjusted means.[30] For the three beta diversity matrices, we estimated sample weighted principal coordinate analysis (PCoA) vectors and utilized the first 100 PCoAs (scree plots in **Figures S3**) as outcomes in linear regression models.

We utilized *FastAdonis*[31] to estimate partial and full-model explained variability in the beta diversity matrices from each variable and combinations of variables using the ANOVA principle, while accounting for the complex design. Similarly, we estimated the contribution of each genus to variability in the beta diversity matrices, unadjusted for covariates or correlations across genera.[31]

Our analyses of covariate associations examined genus-level associations with prevalence of genera using binary logistic regression, results presented as odds ratio [OR]. Associations with relative abundance used Poisson regression, with genus-level sequence counts as the outcomes and the total sequence counts as the offset; results are presented as risk ratios [RR]. We note that microbiome relative abundance data seldom follow a Poisson distribution and show zero inflation, which results in variance inflation. It is, however, difficult to implement zero-inflated regression (e.g., zero-inflated negative binomial or Poisson regression) while accounting for the NHANES design because of recognized issues with model convergence. Thus, we utilized complex survey design-based Poisson regression, which provides robust variances that account

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

for extra-Poisson variability. Logistic and Poisson regression models were adjusted for covariates noted herein, except respiratory inhalant drugs (excluded due to sparse sample sizes).

Associations with alpha and beta diversity utilized two-sided Wald-p<0.05 for statistical significance. Analyses of the genus-level prevalence and relative abundance were restricted to genera with at least 1% prevalence (n=229 genera) and utilized a Bonferroni-corrected Wald-p-value threshold of 0.0002 (=0.05/229) for statistical significance. Analyses were conducted in SAS-callable SUDAAN version 10.4.1 (Raleigh, North Carolina) and R v4.2.0.

RESULTS

Among 8,237 US adults aged 18-69 years (representing 202,314,000 individuals), the oral microbiome encompassed 37 bacterial phyla, 99 classes, 212 orders, 446 families, and 1,219 genera. Five phyla—*Firmicutes, Actinobacteria, Bacteroidetes, Proteobacteria,* and *Fusobacteria* **(Figure 1A**) and six genera—*Veillonella, Streptococcus, Prevotella7, Rothia, Actinomyces, and Gemella*, (**Figure 1B**) were present in nearly all US adults (weighted prevalence >99%). These genera also had the highest relative abundances (**Figure 1B inset**), collectively accounting for 65.7% of the genus-level total abundance—*Streptococcus* (32.6%), *Rothia* (11.5%), *Prevotella7* (8.0%), *Veillonella* (6.3%), *Gemella* (4.8%), and *Actinomyces* (2.7%). At the population-level, however, genus-level prevalence was only moderately correlated with its relative abundance (rho=0.35), as most prevalent genera had very low relative abundances.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Alpha diversity associations

The estimated mean oral microbiome alpha diversity in US adults aged 18-69 years was 128 (95% CI=126-130) for observed ASVs, 14.5 (95% CI=14.3-14.7) for Faith's PD, 4.58 (95% CI=4.55-6.61) for the Shannon-Weiner Index, and 0.90 (95% CI=0.89-0.90) for the Simpson Index.

Across years of age, the number of observed ASVs showed a non-linear quadratic pattern, with an increase from ages 18 to 30 years, peak at age 30 years, and subsequent decline after 30 (**Figure 2**). The number of observed ASVs was similar for men and women. Compared to non-Hispanic White individuals (124 ASVs), observed ASVs were significantly higher for individuals of all other race and ethnicity groups (**Figure 3**, Mexican American (137), other Hispanic (132), non-Hispanic Black (134), and other non-Hispanic (131), including non-Hispanic Asian (127), a group first oversampled in NHANES 2011-2012). Observed ASVs decreased with increasing education and increased with increasing BMI, alcohol use (among ever drinkers), and increasing severity of periodontal disease (**Figure 3**). Notably, the number of observed ASVs was substantially lower in edentulous individuals (75 vs. population average=128). Use of prescription medications over the past 30 days, including antibiotics, antilipidemic medications, and medications for hypertension and gastroesophageal reflux were associated with lower observed ASVs, while inhaled respiratory drug use did not alter this metric (**Figure 3**). Similar associations were observed for the other alpha diversity metrics (**Figures S4-S6**).

Beta diversity associations

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

A modest percentage (<9%) of the overall variability in beta diversity was explained by the demographic, socioeconomic, behavioral, and health covariates—8.7%, 7.2%, and 6.3%, for unweighted UniFrac, weighted UniFrac, and Bray-Curtis matrices, respectively (**Figure 4A-B**). Dominant factors in weighted UniFrac analyses for beta diversity variability were periodontal disease (3.3%), age (1.1%), and smoking (2.4%). These covariates and antibiotic use were also significantly associated with the first PCoA axis across the three beta diversity matrices (**Tables S14-S16**).

By contrast, a few genera explained a surprisingly high percentage of variability in weighted UniFrac in particular—relative abundance of 3 genera: *Aggregatibacter* (22.4%), *Lactococcus* (21.6%), and *Haemophilus* (18.4%) (**Figure C-F)**. Importantly, these genera had marked variation in average prevalence and relative abundance compared to each other—*Aggregatibacter* prevalence=67.3%, relative abundance=0.5%; *Lactococcus* prevalence=5.5%, relative abundance=0.007%; and *Haemophilus* prevalence=98.3%, relative abundance=5.6%—underscoring their influence on beta diversity, particularly weighted UniFrac, regardless of prevalence or relative abundance.

Genus-level associations

The prevalence and/or relative abundance of genera with major contributions to variability in weighted UniFrac (*Aggregatibacter, Lactococcus, Haemophilus)* also had associations with several covariates. For example, prevalence of *Aggregatibacter* was significantly associated with race and ethnicity, smoking, periodontal disease, and antibiotic use; relative abundance of *Haemophilus* was 5-times higher among antibiotic

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

users than non-users and 3-times higher among obese individuals compared to normal weight individuals (**Tables S17-S18).**

Numerous bacterial genera were significantly associated with particular covariates, some with multiple covariates, at a Bonferroni-corrected p-value threshold ($p<0.0002$) (**Figure 5A**). For example, prevalence of *Alloscardovia* and of *Lachnospiraceae* was significantly associated with age, race and ethnicity, smoking, alcohol use, and periodontal disease. Likewise, relative abundance of *Castellaniella* and *Phascolarctobacterium* was significantly associated with age, race and ethnicity, periodontal disease, alcohol use, and BMI. The genera associated with each covariate generally differed for prevalence and relative abundance, except for periodontal disease where both measures were associated with mostly the same genera (**Figure S7**).

Similar to alpha diversity, prevalence and relative abundance of several genera were non-linearly associated with age, with varying age peaks (**Figures 5B and 5C**). Prevalence of *Desulfomicrobium* was >5-times higher in each of the four race and ethnicity groups compared to non-Hispanic White individuals (**Table S17**). Prevalence of *Metascardovia* was 13-times higher in current smokers vs. never smokers. Current smoking was also associated with increased prevalence of several genera in the *Bifidobacteriaceae* family (*Metascardovia, Scardovia, Parascardovia, Aeriscardovia, Alloscardovia*). In contrast, prevalence of *Neisseria* was 3-times higher in never smokers than current smokers. Prevalence of *Castellaniella* increased with increasing alcohol use and was ~5-times higher among heavy vs. never drinkers.

Strong associations were also observed for relative abundance (**Table S18**) by race and ethnicity (*Murdochiella* 5-18-times higher abundance in non-White races and

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

ethnicities vs. non-Hispanic White individuals), smoking (*Metascardovia* 14-times higher abundance in current vs. never smokers), and BMI categories (*Coprobacter* 10-times higher abundance among adults with obesity vs. normal weight individuals).

Increasing severity of periodontal disease was generally associated with increased prevalence and increased relative abundance of several genera. Despite the lack of resolution in our results at the species level, both prevalence and relative abundance of red complex bacterial genera (*Porphyromonas, Tannerella,* and *Treponema*) were significantly higher among participants with severe periodontitis when compared to those with no periodontal disease (**Tables S17-S18**).

DISCUSSION

Our study provides a contemporary reference standard for the oral microbiome of the US adult population aged 18-69 years during 2009-2012 (representing 202,314,000 individuals). We show that the oral microbiome is a complex ecosystem encompassing 37 phyla and over 1,000 genera. Yet, a limited set of genera (*Veillonella, Streptococcus, Prevotella7, Rothia, Actinomyces, and Gemella)* were observed in all adults (>99%) and accounted for a majority of the genus-level relative abundance (66%), indicating the existence of a limited universal oral microbiome in US adults.[2,12,13] A different, limited set of genera (*Aggregatibacter, Lactococcus, Haemophilus*) were associated with a high proportion of variability in one measure of oral microbiome beta diversity (>60% in weighted UniFrac), indicating the existence of key genera that potentially influence oral microbiome diversity across individuals. A wide range of demographic, behavioral, and metabolic/clinical covariates (age, race and ethnicity, smoking, alcohol consumption,

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

and periodontal disease) were consistently associated with multiple oral microbiome metrics—alpha diversity, beta diversity, and genus-level prevalence and relative abundance—underscoring potential modifiability of the oral microbiome.

Some of the alpha diversity, beta diversity, and genus-level associations we report have previously been noted in the literature and may have plausible biological explanations. [32,33] Such observations include enrichment of *Actinobacteria* among current smokers (particularly several genera in the *Bifidobacteriaceae* family)[32] and red complex bacteria in those with severe periodontal disease.[33] Nonetheless, several of our key observations remain unexplained. The non-linear associations of age with oral microbiome alpha diversity (peaking at age 30) and prevalence and relative abundance of several genera (varying age peaks), has not been previously reported. The numerous associations of race and ethnicity with alpha and beta diversity metrics and genus-level prevalence and relative abundance, although previously reported,[11,2] were not explained by the socioeconomic, behavioral, anthropometric, metabolic, and clinical factors we included for statistical adjustment.[14,34] Additional hypotheses and research are needed to identify potential biological or alternative explanations. For example, the non-linear association with age could arise from biological ageing or reflect birth cohort effects. However, we could not evaluate these birth cohort effects because the narrow calendar years in our study makes age equivalent to birth cohort. Likewise, associations with race and ethnicity could reflect multifactorial social determinants of health, such as early life exposures or access to health services,[34] which were not considered herein.

Our observations suggest that the interpretation of alpha diversity as an indicator of health/disease may be context-specific. Although high alpha diversity (particularly of

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

the gut microbiome) is generally believed to be an indicator of good health/outcomes,[35] we show that oral microbiome alpha diversity (as measured by observed number of ASVs and after adjustment for a wide range of covariates) was higher among those with severe periodontal disease and overweight and obesity but substantially lower among edentulous individuals. The lower richness in edentulous individuals suggests the important role of teeth in creating diverse niches in the oral environment.

We note key study limitations and strengths. Our evaluation was restricted to 16S rRNA gene sequencing because metagenomic sequencing was precluded by NCHS policy. Thus, the data resolution is through the genus-level and we lacked species differentiation and imputed bacterial functionality, which metagenomic sequencing would have permitted. The strengths of our study are the population representative nature of the data, large sample size, and the comprehensive array of demographic, socioeconomic, and behavioral data and uniform, high-quality data, and unindicated (i.e., assessment independent of health status) health examinations.

In summary, we present the oral microbiome profile of US adults—colloquially, the mouth of America. Importantly, NCHS, CDC has made available all the results to the research community (publicly available data can be accessed at the NHANES website: https://wwwn.cdc.gov/Nchs/Nhanes/omp/ and data considered restricted can be accessed at NCHS Research Data Centers: https://www.cdc.gov/rdc/index.htm). Our comprehensive characterization herein, and the anticipated use of the data by the scientific community, portends deeper understanding of the role of the oral microbiome in health and disease.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

REFERENCES

1. Tuganbaev T, Yoshida K, Honda K. The effects of oral microbiota on health. *Science*. May 27 2022;376(6596):934-936. doi:10.1126/science.abn1890

2. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. *Nature*. Jun 13 2012;486(7402):207-14. doi:10.1038/nature11234

3. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 3/13/2012 2012;13(4):260-270. Not in File. doi:nrg3182 [pii];10.1038/nrg3182 [doi]

4. Takahashi N. Oral Microbiome Metabolism: From "Who Are They?" to "What Are They Doing?". *J Dent Res*. Dec 2015;94(12):1628-37. doi:10.1177/0022034515606045

5. Raizada MK, Joe B, Bryan NS, et al. Report of the National Heart, Lung, and Blood Institute Working Group on the Role of Microbiota in Blood Pressure Regulation: Current Status and Future Directions. *Hypertension*. Jul 31 2017;doi:10.1161/HYPERTENSIONAHA.117.09699

6. Pihlstrom BL, Michalowicz BS, Johnson NW. Periodontal diseases. *Lancet*. Nov 19 2005;366(9499):1809-20. doi:10.1016/S0140-6736(05)67728-8

7. Hayes RB, Ahn J, Fan X, et al. Association of Oral Microbiome With Risk for Incident Head and Neck Squamous Cell Cancer. *JAMA Oncol*. Mar 1 2018;4(3):358-365. doi:10.1001/jamaoncol.2017.4777

8. Wu Z, Han Y, Wan Y, et al. Oral microbiome and risk of incident head and neck cancer: A nested case-control study. *Oral Oncol*. Feb 2023;137:106305. doi:10.1016/j.oraloncology.2022.106305

9. Vogtmann E, Hua X, Yu G, et al. The Oral Microbiome and Lung Cancer Risk: An Analysis of 3 Prospective Cohort Studies. *J Natl Cancer Inst*. Nov 14 2022;114(11):1501-1510. doi:10.1093/jnci/djac149

10. Fan X, Alekseyenko AV, Wu J, et al. Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut*. Jan 2018;67(1):120-127. doi:10.1136/gutjnl-2016-312580

11. Huybrechts I, Zouiouich S, Loobuyck A, et al. The Human Microbiome in Relation to Cancer Risk: A Systematic Review of Epidemiologic Studies. *Cancer Epidemiol Biomarkers Prev*. Oct 2020;29(10):1856-1868. doi:10.1158/1055-9965.EPI-20-0288

12. Renson A, Jones HE, Beghini F, et al. Sociodemographic variation in the oral microbiome. *Ann Epidemiol*. Jul 2019;35:73-80 e2. doi:10.1016/j.annepidem.2019.03.006

13. Nearing JT, DeClercq V, Van Limbergen J, Langille MGI. Assessing the Variation within the Oral Microbiome of Healthy Adults. *mSphere*. Sep 30 2020;5(5)doi:10.1128/mSphere.00451-20

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

14. Yang Y, Zheng W, Cai Q, et al. Racial Differences in the Oral Microbiome: Data from Low-Income Populations of African Ancestry and European Ancestry. *mSystems*. Nov 26 2019;4(6)doi:10.1128/mSystems.00639-19

15. Ruan X, Luo J, Zhang P, Howell K. The salivary microbiome shows a high prevalence of core bacterial members yet variability across human populations. *NPJ Biofilms Microbiomes*. Oct 20 2022;8(1):85. doi:10.1038/s41522-022-00343-7

16. Vogtmann E, Chaturvedi AK, Blaser MJ, et al. Representative oral microbiome data for the US population: the National Health and Nutrition Examination Survey. *Lancet Microbe*. Feb 2023;4(2):e60-e61. doi:10.1016/S2666-5247(22)00333-0

17. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD:. *U S Department of Health and Human Services, Centers for Disease Control and Prevention, in , 2010*. 2014. 2014.

18. Chaturvedi AK, Graubard BI, Broutian T, et al. NHANES 2009-2012 Findings: Association of Sexual Behaviors with Higher Prevalence of Oral Oncogenic Human Papillomavirus Infections in U.S. Men. *Cancer Res*. 6/15/2015 2015;75(12):2468-2477. Not in File. doi:0008-5472.CAN-14-2843 [pii];10.1158/0008-5472.CAN-14-2843 [doi]

19. Chaturvedi AK, Graubard BI, Broutian T, Xiao W, Pickard RKL, Kahle L, Gillison ML. Prevalence of Oral HPV Infection in Unvaccinated Men and Women in the United States, 2009-2016. *JAMA*. Sep 10 2019;322(10):977-979. doi:10.1001/jama.2019.10508

20. Gillison ML, Broutian T, Pickard RK, et al. Prevalence of oral HPV infection in the United States, 2009-2010. *JAMA*. 2/15/2012 2012;307(7):693-703. Not in File. doi:jama.2012.101 [pii];10.1001/jama.2012.101 [doi]

21. National Center for Health Statistics C. https://www.cdc.gov/nchs/nhanes/irba98.htm

22. Walters W, Hyde ER, Berg-Lyons D, et al. Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems*. Jan-Feb 2016;1(1)doi:10.1128/mSystems.00009-15

23. Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. *Proc Natl Acad Sci U S A*. Feb 9 2016;113(6):E791-800. doi:10.1073/pnas.1522149113

24. Wilbert SA, Mark Welch JL, Borisy GG. Spatial Ecology of the Human Tongue Dorsum Microbiome. *Cell Rep*. Mar 24 2020;30(12):4003-4015 e3. doi:10.1016/j.celrep.2020.02.097

25. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. May 2010;7(5):335-6. doi:10.1038/nmeth.f.303

26. Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. *F1000Res*. 2016;5:1492. doi:10.12688/f1000research.8986.2

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

27. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*. Jul 2016;13(7):581-3. doi:10.1038/nmeth.3869

28. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*. Aug 2019;37(8):852-857. doi:10.1038/s41587-019-0209-9

29. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. Jan 2013;41(Database issue):D590-6. doi:10.1093/nar/gks1219

30. Graubard BI, Korn EL. Predictive margins with survey data. *Biometrics*. Jun 1999;55(2):652-9. doi:10.1111/j.0006-341x.1999.00652.x

31. Li S, Vogtmann E, Graubard BI, Gail MH, Abnet CC, Shi J. fast.adonis: a computationally efficient non-parametric multivariate analysis of microbiome data for large-scale studies. *Bioinform Adv*. 2022;2(1):vbac044. doi:10.1093/bioadv/vbac044

32. Yang Y, Zheng W, Cai QY, et al. Cigarette smoking and oral microbiota in low-income and African-American populations. *J Epidemiol Community Health*. Dec 2019;73(12):1108-1115. doi:10.1136/jech-2019-212474

33. Rocas IN, Siqueira JF, Jr., Santos KR, Coelho AM. "Red complex" (Bacteroides forsythus, Porphyromonas gingivalis, and Treponema denticola) in endodontic infections: a molecular approach. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. Apr 2001;91(4):468-71. doi:10.1067/moe.2001.114379

34. Byrd DA, Carson TL, Williams F, Vogtmann E. Elucidating the role of the gastrointestinal microbiota in racial and ethnic health disparities. *Genome Biol*. Aug 3 2020;21(1):192. doi:10.1186/s13059-020-02117-w

35. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nat Med*. Apr 10 2018;24(4):392-400. doi:10.1038/nm.4517

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure legends

Figure 1: Shows the prevalence of phyla (**panel A**) and prevalence of genera (**panel B**) in the oral microbiome among US adults aged 18-69 years, NHANES 2009-2012; bars represent weighted percent prevalence and circles represent weighted mean relative abundance. The inset in panel B shows the proportionate weighted mean relative abundance for genera with >1% relative abundance and for the remaining genera.

Figure 2: Shows the adjusted predictive margins (solid line) and 95% confidence intervals (dashed lines) for the observed number of amplicon sequence variants (ASVs) by age (in years) in the oral microbiome of US adults aged 18-69 years, NHANES 2009-2012. Age was modeled through 5-knot restricted cubic splines.

Predictive margins were estimated in survey-design adjusted linear regression models, with concomitant adjustment for sex, race and ethnicity, education, marital status, income-to-poverty ratio, measured body mass index [BMI] categories, risk behaviors (smoking, alcohol use), medical conditions (diabetes and hypertension), oral health (periodontal disease, tooth count, and edentulism), and use of prescription medications within the past 30 days (antibiotics, antilipidemics, respiratory inhalants, and for gastroesophageal reflux). Race and ethnicity was self-reported and used as categorized by NHANES. Please see Table S13 for variable definitions/code.

Figure 3: Shows the adjusted predictive margins (circles) and 95% confidence intervals (error bars) for the observed number of amplicon sequence variants (ASVs) in the oral microbiome of US adults aged 18-69 years (NHANES 2009-2012) by each covariate.

Predictive margins were estimated in survey-design adjusted linear regression models, with concomitant adjustment for age (modeled as 5-knot restricted cubic splines), sex, race and ethnicity, education, marital status, income-to-poverty ratio, measured body mass index [BMI] categories, risk behaviors (smoking, alcohol use), medical conditions (diabetes and hypertension), oral health (periodontal disease, tooth count, and edentulism), and use of prescription medications within the past 30 days (antibiotics, antilipidemics, respiratory inhalants, and for gastroesophageal reflux). Race and ethnicity was self-reported and used as categorized by NHANES. Please see Table S13 for variable definitions/code. An asterisk denotes $p<0.05$ for the covariate in the adjusted model.

Figure 4: **Panel A** shows the covariate-specific (horizontal bars) and overall (horizontal bars) proportion of variability in beta diversity matrices (unweighted UniFrac, weighted UniFrac, and Bray-Curtis) in the oral microbiome of US adults aged 18-69 years, NHANES 2009-2012. The fastAdonis algorithm was used for estimation. The overall estimate was based on models that incorporated concomitant adjustment for age (modeled as 5-knot restricted cubic splines), sex, race and ethnicity, education, marital status, income-to-poverty ratio, measured body mass index [BMI] categories, risk behaviors (smoking, alcohol use), medical conditions (diabetes and hypertension), oral health (periodontal disease, tooth count, and edentulism), and use of prescription medications within the past 30 days (antibiotics, antilipidemics, respiratory inhalants, and for gastroesophageal reflux). Race and ethnicity was self-reported and used as collected by NHANES. Covariate-specific estimates are unadjusted. Please see Table S13 for variable definitions/code.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

The proportion of variability in the unweighted UniFrac beta diversity matrix from prevalence of each genus is shown in **panel C** and from relative abundance of key genera is shown in **panel D**. **Panel E** shows the proportion of variability in the weighted UniFrac beta diversity matrix from relative abundance of key genera. **Panel F** shows the proportion of variability in the Bray-Curtis beta diversity matrix from relative abundance of key genera. The fastAdonis algorithm was used for estimation and genus-specific estimates are unadjusted for covariates or correlations across prevalence or relative abundance of genera.

Figure 5: **Panel A** shows for key covariates the number of genera with statistically significant (at a Bonferroni-corrected $p<0.0002$) associations with genus-specific relative abundance (light grey bars) and genus-specific prevalence (dark grey bars) for ages 18-69 years, NHANES 2009-2012. Associations with relative abundance were estimated in Poisson regression models and with prevalence were estimated in binary logistic regression models. **Panel B** shows the adjusted odds ratio (OR) by age (18-69 years) for prevalence of genera with statistically significant associations with age at a Bonferroni-corrected $p<0.0002$. **Panel C** shows the model-predicted genus-specific sequence counts by age (18-69 years) for genera with statistically significant association of genus-specific relative abundance with age at Bonferroni-corrected $p<0.002$.

All models included concomitant adjustment for age (modeled as 5-knot restricted cubic splines), sex, race and ethnicity, education, marital status, income-to-poverty ratio, measured body mass index [BMI] categories, risk behaviors (smoking, alcohol use), medical conditions (diabetes and hypertension), oral health (periodontal disease, tooth count, and edentulism), and use of prescription medications within the past 30 days (antibiotics, antilipidemics, and for gastroesophageal reflux). Race and ethnicity was self-reported and used as categorized by NHANES. Please see Table S13 for variable definitions/code.

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure 1A

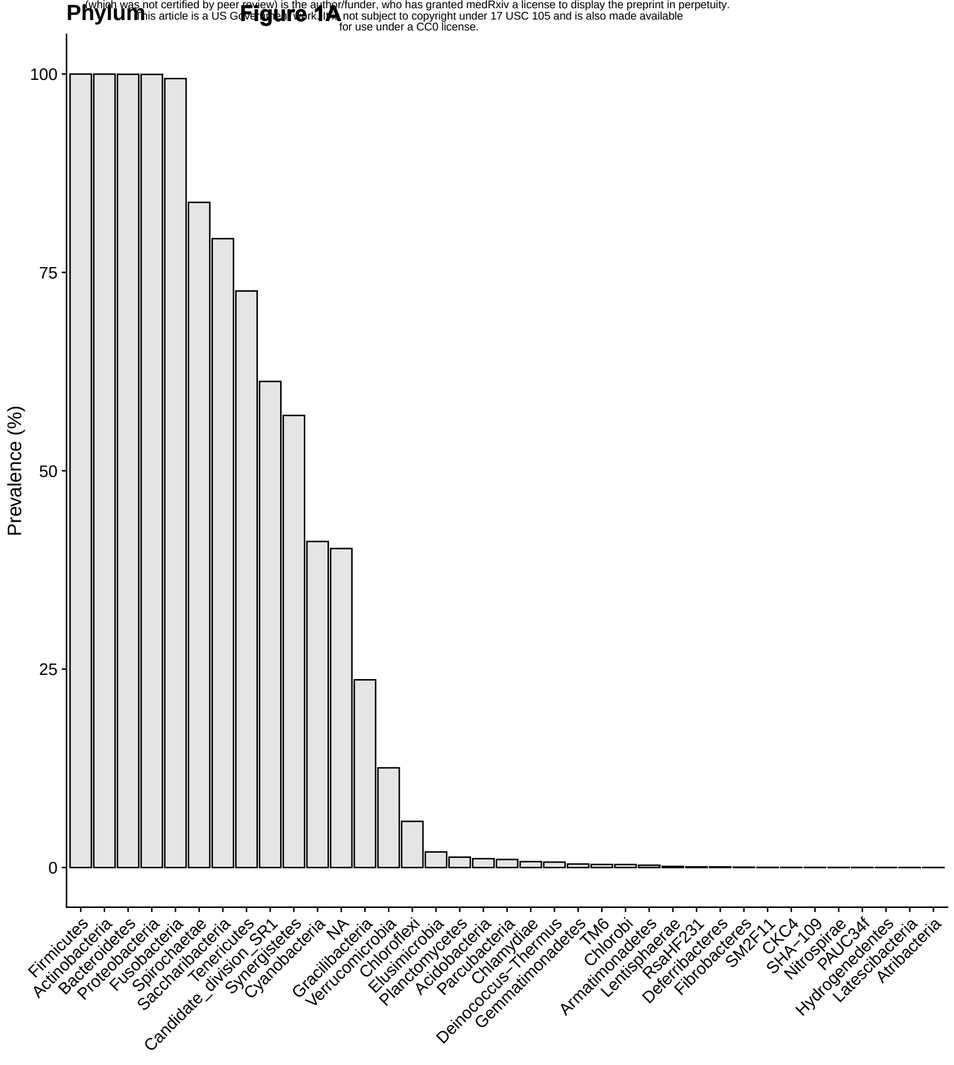
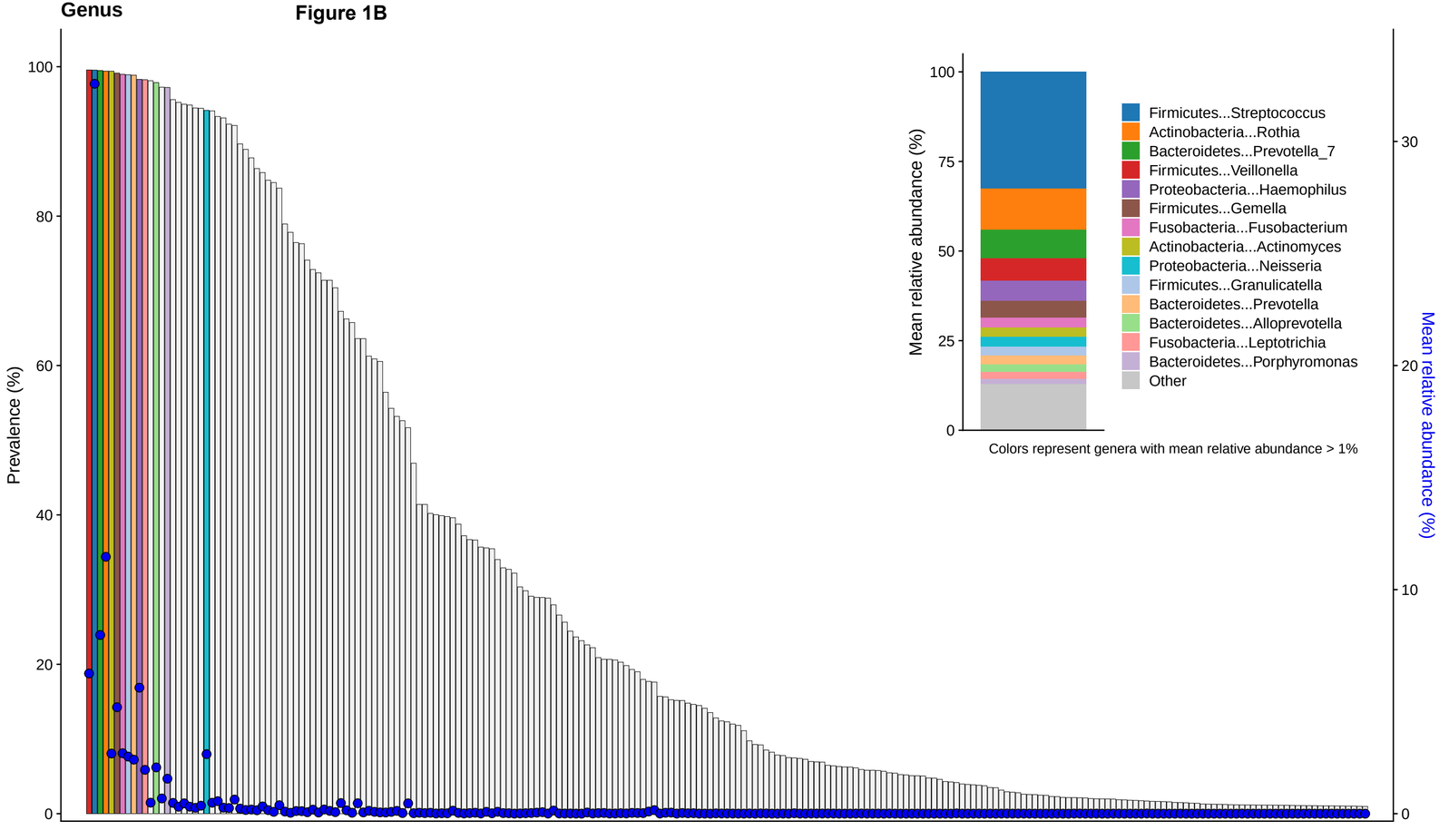


Figure 1B



medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure 2: Association of oral microbiome alpha diversity with age, NHANES 2009-2012

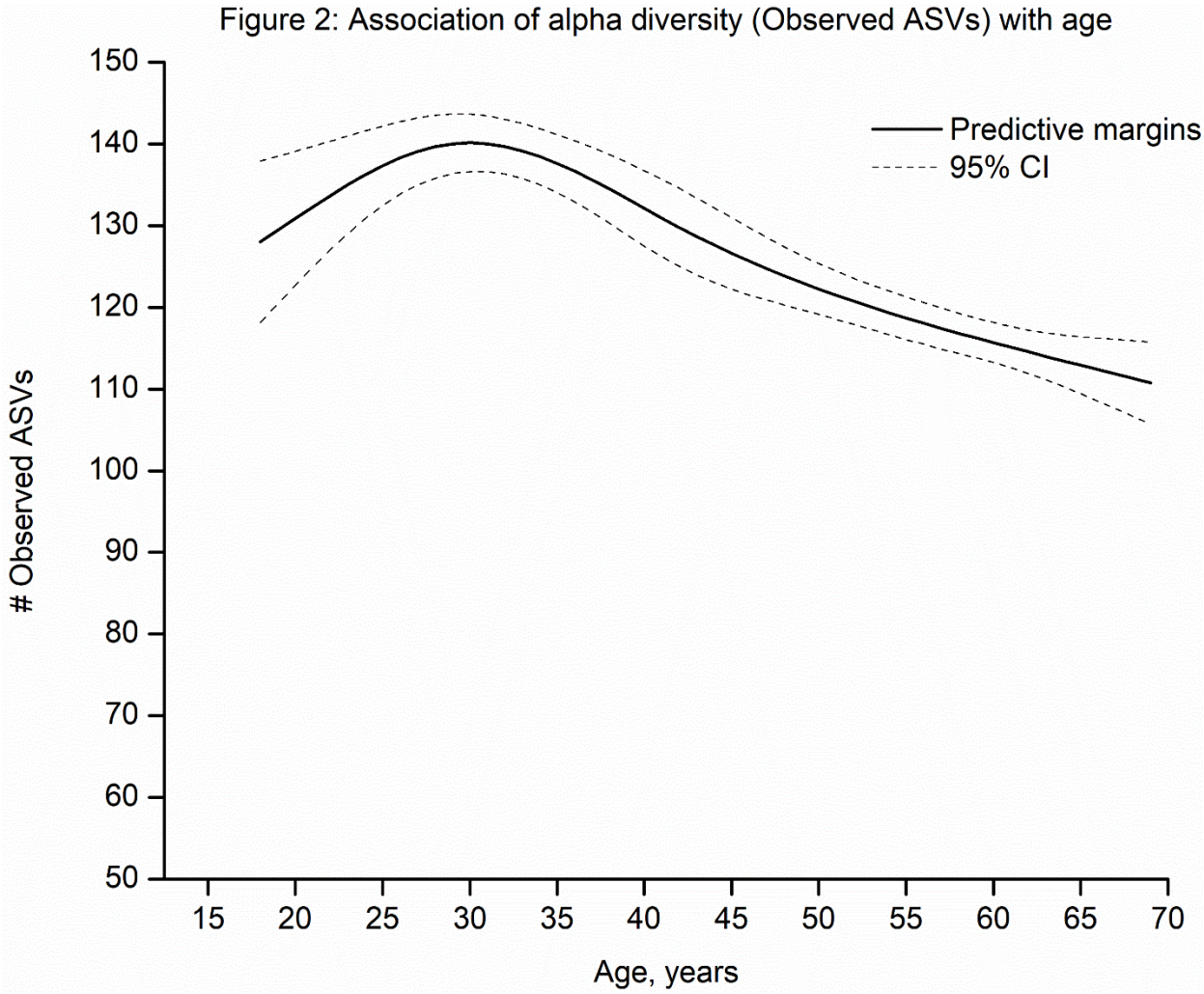
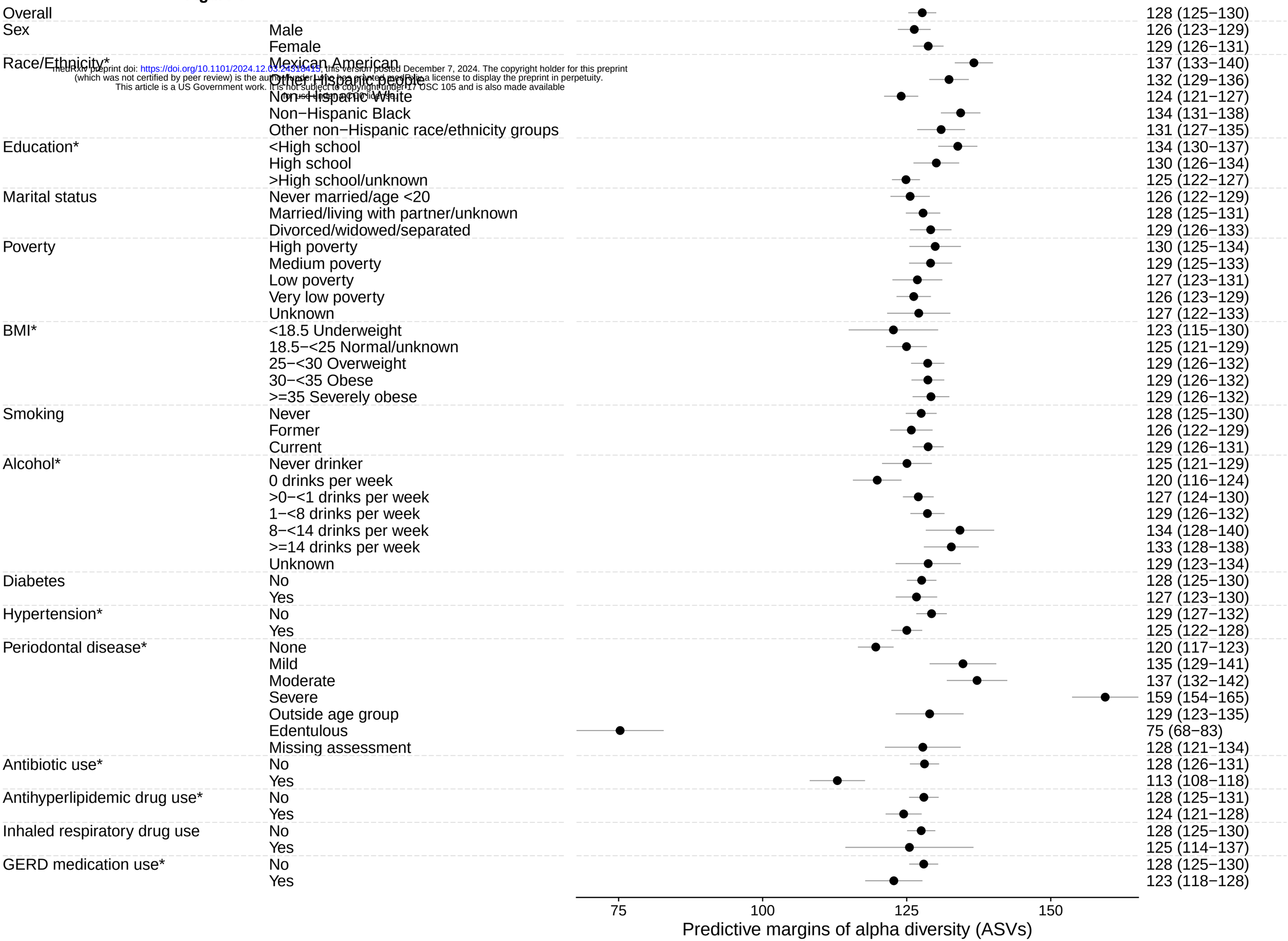


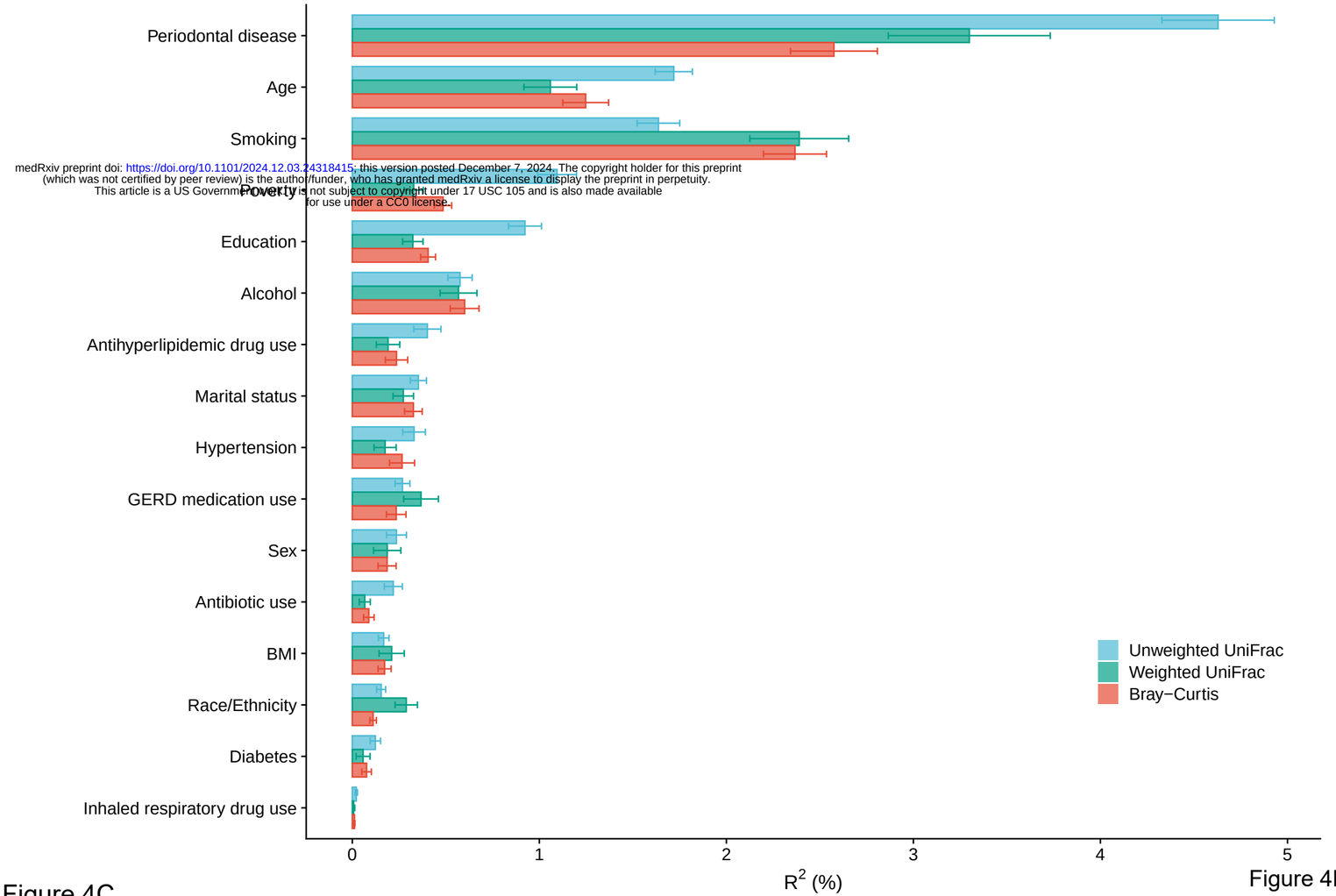
Figure 3

medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

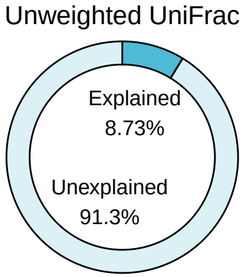


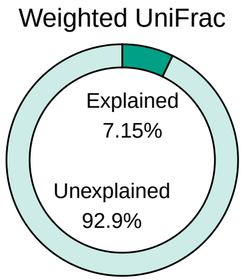
medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available
for use under a CC0 license.

A Figure 4A



B Figure 4B





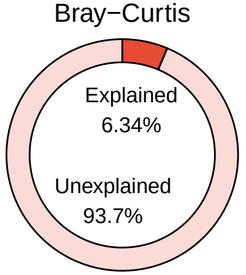


Figure 4C

C Unweighted UniFrac (presence)

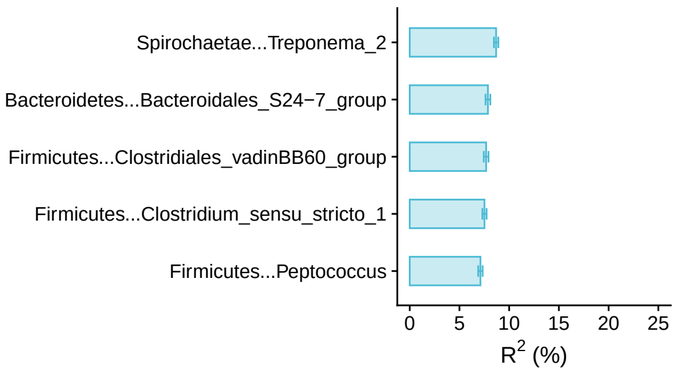


Figure 4D

D Unweighted UniFrac (relative abundance)

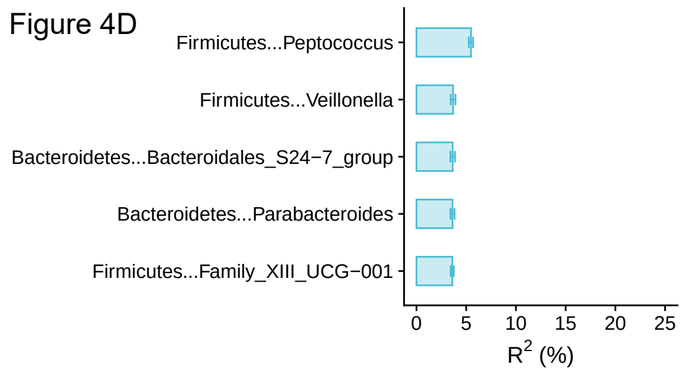


Figure 4E

E Weighted UniFrac

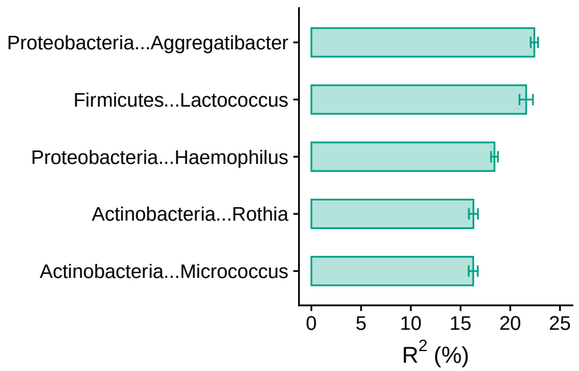
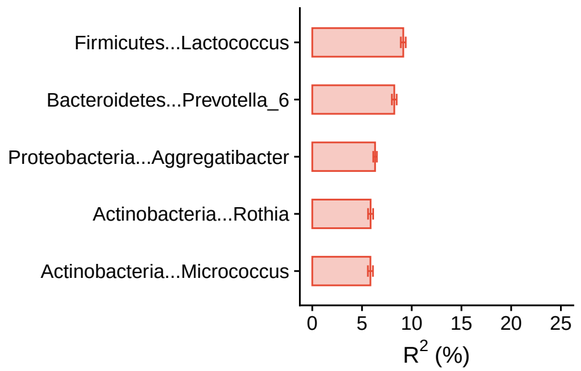
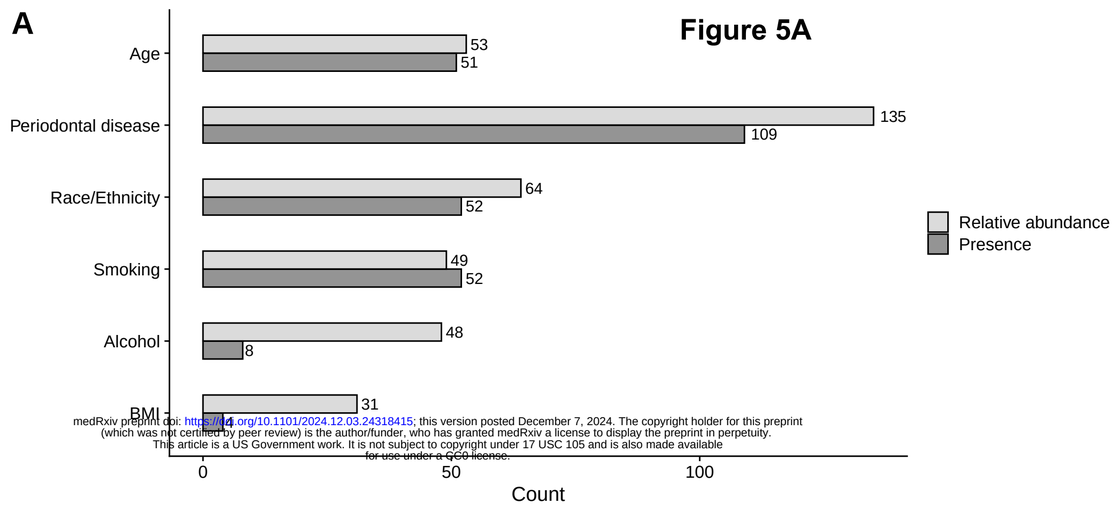
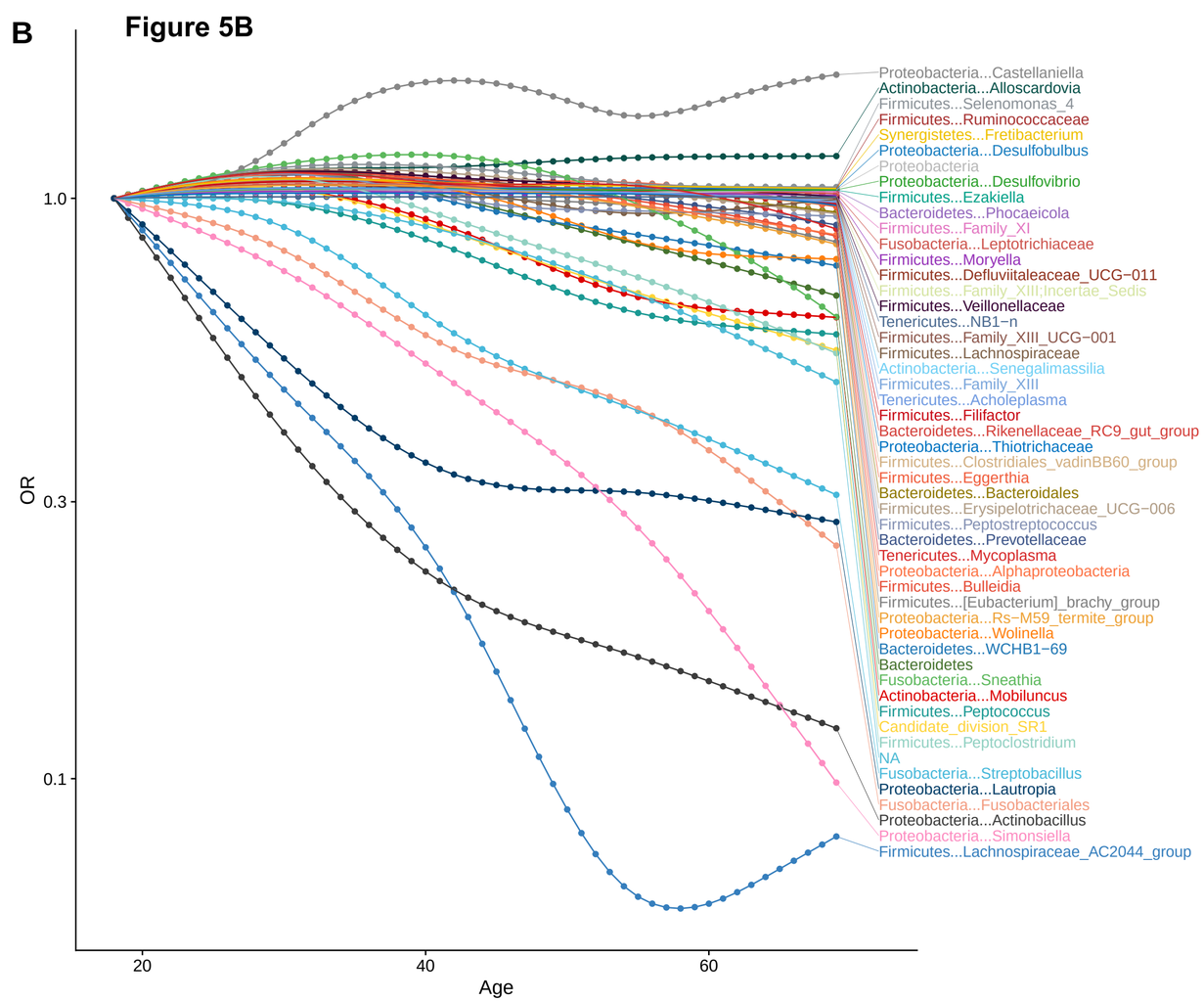
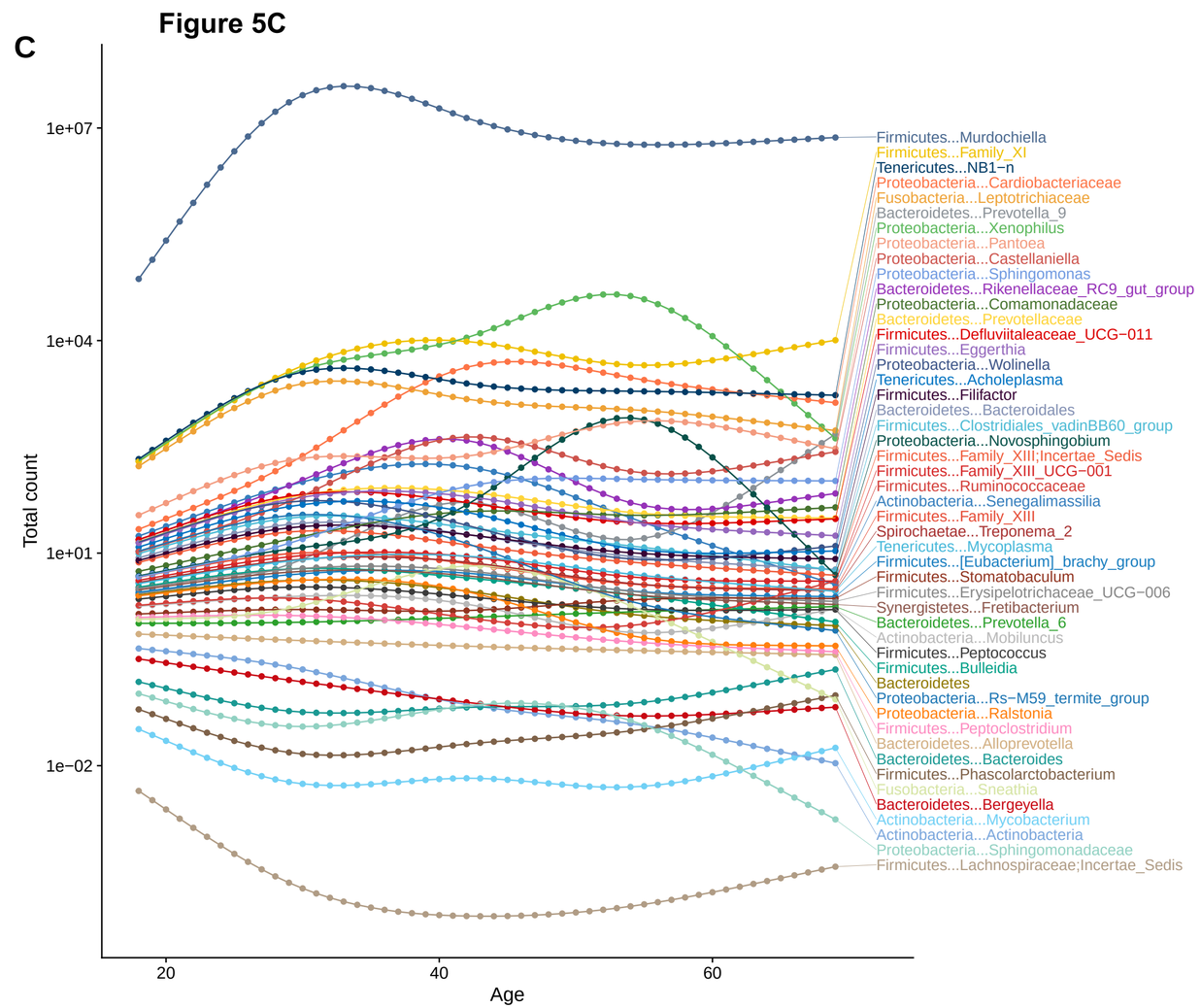


Figure 4F

F Bray−Curtis





medRxiv preprint doi: https://doi.org/10.1101/2024.12.03.24318415; this version posted December 7, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available
for use under a CC0 license.