A core salivary microbiome shows the high prevalence of bacterial members yet variability across human populations

Xinwei Ruan [1], Jiaqiang Luo[1], Pangzhen Zhang[1] and Kate Howell[1,] *

[1]School of Agriculture and Food, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville 3010, Australia

Xinwei Ruan: ORCID: 0000-0002-1963-3807

Jiaqiang Luo: ORCID: 0000-0001-6459-3309

Pangzhen Zhang: ORCID: 0000-0002-9794-2269

* Corresponding author. khowell@unimelb.edu.au; ORCID 0000-0001-6498-0472

Abstract

Background

Human saliva contains diverse bacterial communities, reflecting human health status, dietary patterns and contributing to variability in the sensory perception of food. Many descriptions of salivary microbiome diversity compare commonalities and differences with reference to a diseased state, but the composition of healthy saliva has not been described.

Results

Here, we use a meta-analysis approach to define and explore the core membership of the human salivary microbial community by collecting and re-analysing raw 16S rRNA amplicon sequencing data from 47 studies with 2206 saliva samples. We found 68 core bacterial taxa that were consistently detected. Differences induced by various host intrinsic and behaviour factors, including gender, age, geographic location, tobacco usage, and alcohol consumption, were evident. The core of the salivary microbiome was verified by collecting and analysing saliva in an independent study.

Conclusion

These results suggest that the methods used can effectively define a core microbial community in human saliva with high prevalence. The core salivary microbiome demonstrated both stability and variability among populations. Geographic location was identified as the host factor with the largest effect in shaping salivary microbiota. The independent analysis underlined that the impact of geographic variation is likely due to diet.

Background

Human saliva plays an essential role in influencing the sensory perception of foods and beverages and driving the purchase decisions of consumers. When food is taken into the mouth, mixing and mastication allow a semi-solid bolus to be formed, and at the same time, aroma and flavour to be released from the food [1]. The variation in perceived responses can be attributed to the inter-individual variability in salivary composition, especially salivary microorganisms. As a complex

ecosystem, the human oral cavity hosts thousands of bacterial taxa, accompanied by interactions with other microorganisms [2]. It is an ecological system that contains many distinct sub-niches, including saliva, dental plaques, gingival sulcus, epithelial cells on the cheek, tongue, and teeth [3]. High heterogeneity has been reported between the composition of microbial communities that colonise on different sites [4]. The saliva is recognised as a reservoir with microorganisms from all ecological niches in the human mouth with long-term stability [5]. The ensemble of microorganisms and the expressed genetic material in human saliva is known as the "salivary microbiome".

The contribution of salivary microbiome to sensory perception of foods has been described by various studies [6-8]. However, the diverse conclusions suggest that the role of the salivary microbiome may be confounded by inter-individual variance. Meanwhile, comparing results from different studies introduces significant technical and bioinformatic biases [9] especially when studies have targeted different 16S rRNA hypervariable regions for amplification [10]. Characterising the microbial communities commonly found in most human saliva regardless of the study-specific variation could help establish the connection between salivary composition and food preference. On this basis, the shifts from a common salivary microbiome by diseases or host lifestyle factors will also be more prominent. A meta-analysis can summarise the existing knowledge and identify the commonalities and differences in salivary microbiota between people from various backgrounds.

The core oral microbiome of a healthy human has been tried to be defined for more than ten years [11]. The core microbiome is described as the common group of microbes that are important for host biological function [12]. Defining the core only depends on occupancy frequency does not necessarily reflect the underlying host-microbes functional relationship. However, it provides a foundation for prioritising members adapted to the host environment [13]. A variety of studies have been devoted to discovering the changes in human salivary microbiota based on different conditions. The variability in the microbial profile of human saliva has not only been associated with oral diseases [14-16] but also various chronic diseases that do not occur in the oral cavity, such as diabetes [17] pancreatic cancer

[18] and Polycystic ovary syndrome [19]. Although the shift in human salivary composition caused by diseases has been studied for decades, our understanding of the impact of host intrinsic and behaviour factors is still limited.

Many host characteristics have been shown to have an impact on the composition of the salivary microorganisms, including age [20], diet [21, 22], ethnicity [23], gender [24], smoking [25], alcohol use [26], circadian rhythm [27], body mass index [28], and the type of stimulation [29]. Some studies have correlated the diverse microbiome with the distinct sensory responses between consumer groups [30, 31] It has also been reported that people from different countries are colonised with distinct salivary bacterial communities [32]. Li et al. analysed the human oral microbiome from Africa, Alaska, and Germany and reported differences between the human groups living in various climate conditions [33]. However, no study to date has demonstrated a clear global pattern in salivary microbial composition.

In this paper, we collected raw 16S rRNA sequences of human salivary microbiota from 47 publicly available datasets spanning 15 different countries. These raw data were systematically re-analysed and pooled together to define a core salivary microbiome. We classified all sequences into operational taxonomic units (OTUs) at 97% identity against the Human Oral Microbiome Database (HOMD) to minimise the technical variation induced by comparing data from different hypervariable regions. It allowed us to make a comparison between studies and reduced the redundancy in the dataset for defining the "core". Using the metadata acquired with raw data, we also investigated the influences of several host factors and technical factors on human salivary microbiota. Factors that showed a potentially strong impact on shaping microbial communities in saliva were selected, and the taxa as potential biomarkers were identified, and linked with functional predictions. Finally, saliva samples were collected from independent, healthy individuals and analysed for microbial composition to confirm the results found from the global dataset. These data confirmed the composition of the core microbiome members, but the verification of geographic origin was not possible. Our study contributes to fundamental understandings of the stable and differential salivary microbiome across

healthy adult populations. We have identified bacteria linked to particular identities of participants and points to salivary microbiome composition being linked to diet, rather than ethnic origin.

Methods

1. Literature search and data collection

To acquire sufficient data from healthy human saliva, available public studies related to human salivary microbiota were systematically reviewed. A literature search was performed using the combination of relative terms in EMBASE, MEDLINE and Web of Science for the studies published before November 2020 using the terms described in supplementary data (Table S1). A supplementary dataset search in NCBI's Sequence Read Archive (SRA) was also performed using the search term 'salivary microbiome'. The included studies met the following criteria: 1) Having samples from participants without any diagnosed disease state. For studies investigating the influence of certain kind of disease on salivary microbiota, only samples collected from healthy controls were included in further analysis; 2) Using whole human saliva collected by spitting, swab, mouth washing or oral rinsing, samples exclusively extracted from any specific oral spot, like tongue surface, parotid gland, supragingival plaque, were excluded; 3) Using 16 rRNA gene high-throughput sequencing and sequenced with the Illumina MiSeq platform; 4) Having and sequence file with quality score and associated metadata, information about geographic locations are required; 5) Having freely available sequencing data; 6) Sequencing data correctly separated according to the metadata. Raw sequence data acquired from the healthy individuals of selected studies were downloaded from SRA, European Nucleotide Archive (ENA), using SRA Toolkit. The files were converted to the FASTQ formats if necessary.

2. 16S rRNA gene sequence processing

Sequence data from each selected study were processed separately using QIIME2 (version.2020.2) [34]. Sequences with primers were trimmed with "q2-cutadapt" (https://github.com/qiime2/q2-cutadapt) to retain the targeted hypervariable regions. The demultiplexed paired-end sequences were

130 firstly joined by "q2-vsearch" (https://github.com/qiime2/q2-vsearch), then subjected to a quality
131 filter with a minimum quality of Q30. The remaining reads were then clustered into operational
132 taxonomic units (OTUs) at 97% similarity against the expanded Human Oral Microbiome Database
133 (eHOMD) version 15.1 [35] by the closed-reference OTU picking command. Reads that failed to
134 match a reference sequence in the HOMD database were discarded. The chimeras and features with a
135 frequency ≤10 or detected in a single sample were also removed. The resulting tables and sequences
136 from all studies were merged by QIIME2's merge and merge-seqs commands. Samples with <2000
137 reads were also removed. Taxonomic annotations were assigned to the representative sequences of
138 each OTU using the HOMD database. For the factor groups containing samples with unknown
139 metadata, the unknown sample was removed from the group before the downstream analyses.

140 **3. Diversity measures in R**

141 The merged OTU table was exported into BIOM format. Further analyses were carried out in R
142 (version 4.1.0) with custom scripts as detailed below. Samples with >2,000 reads were retained and
143 processed with four normalisations: 1) Rarefying samples to 5,000 (Rarefaction, RAR); 2) Samples
144 were rarefied to 5,000 and converted to relative abundance (Rarefied Total-sum Scaling, RRA); 3)
145 Samples were converted to relative abundance directly (Total-sum scaling, TSS); 4) As described by
146 Romano et al. [36], zeros were added to data through the count zero multiplicative approach using
147 the *cmultRepl* function of the *zCompostions* package [37] in R (Centred Log-ratio, CLR).
148
149 The alpha-diversity of all samples grouped by studies were calculated in the form of Chao1, Shannon,
150 and Simpson's diversity indices. The beta-diversity was assessed at different taxonomic levels,
151 including OTU, species, genus, family, order, class, and phylum level. For the OTU level, the beta-
152 diversity of data processed with the first three normalisations were determined using the weighted
153 UniFrac distances and Bray-Curtis dissimilarities. Euclidean distances were calculated for all
154 normalisations. At the other levels, Bray-Curtis dissimilarity was used for the first three
155 normalisations, while Euclidean distances were combined with all the normalisations. Permutational
156 multivariate analysis of variance (PERMANOVA) using the *adonis2* function of *vegan* package [38]

157 with 999 permutations was conducted to investigate the statistical differences caused by different
158 factors, adjusting for study.

159 **4. Defining the core microbiome**

160 The core microbiome was determined based on the abundance-occupancy pattern. Two methods
161 adapted from Shade and Stopnisek [39] and Wu et al. [13] were used. For the methods adapted from
162 the study of Shade and Stopnisek [39], samples were rarefied to 5,000. Both Bray-Curtis similarity
163 and weighted Unifrac distance were used to determine the contribution in the percentage of the
164 prospective core set to the overall beta diversity. For the method adapted from Wu et al. [13], OTUs
165 were filtered out with the mean relative abundance (MRA) bigger than 0.1% and the presence in more
166 than 75% of samples or 100% occupancy in more than 10 studies. To investigate the bacteria-bacteria
167 interactions in salivary microbial communities, the co-occurrence network was constructed using
168 pairwise Spearman's correlation based on relative abundance. The Spearman's correlation was
169 calculated using the *rcorr* function in the *Hmisc* R package [40] and visualised by Cytoscape v3.8.2
170 [41]. A correlation with Spearman's correlation coefficient > 0.5 or < -0.5 and p-value < 0.01 is
171 considered as statistically robust and shown in the network.

172 **5. Differential abundance analyses**

173 **5.1 Random Forest**

174 The normalised abundance of taxa in the phylum, class, order, family, genus, species and OTU level
175 were classified against each provided metadata categories to determine which factor has the largest
176 effect on the salivary microbiota. A random forest classifier was created in R using the *randomForest*
177 package [42] with default parameters. We used the *randomForest* (importance = TRUE, proximity =
178 TRUE) function to generate the classification model for seven categories. For four categories that
179 have the random forest classifier with average error rate lower that 20%, including hypervariable
180 region (5.0%), geographic location (10.4%), tobacco usage (12.7%), sample type (13.8%),
181 differentialt taxa were defined using cross-validation. Cross-validation was performed by the *rfcv()*
182 function for selecting appropriate features. The *varImpPlot* function was used to show the importance

183 of features in the classification. The importance of features and the cross-validation curve were
184 visualized by using the *ggplot2* package [43] in R.

185 **5.2 Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC)**

186 ANCOM-BC [44] were performed to identify the taxa with different relative abundance between
187 Chinese and Western samples. Function *ANCOMBC* were used with holm-bonferroni false discovery
188 rate correction and other default parameters. The hypervariable regions used by different studies were
189 used as the covariate.

190 **6. Functional prediction**

191 Microbial metagenomes were inferred from 16S rRNA gene-based bacterial profiles, and the
192 functional prediction were conducted based on Kyoto Encyclopedia of Gene and Genomes (KEGG)
193 database [45] using the default pipeline in Phylogenetic Investigation of Communities by
194 Reconstruction of Unobserved States 2 (PICRUSt2) (Douglas et al., 2020). The ANCOM-BC analysis
195 was used to identify the differential abundant KEGG pathways by geographic location, adjusting for
196 hypervariable regions. At the same time, a random forest model was established for distinguishing
197 Chinese and Western samples, and the importance of pathways was measured using mean decreased
198 accuracy. Spearman's correlation was performed to assess the relationship between the relative
199 abundance of differential pathways and genera. The significant correlations were visualised using the
200 *corrplot* package [46] in R.

201 **7. Comparison between Chinese and Western people on an independent cohort**

202 Saliva samples were collected from 26 participants (aged 20-60 years) recruited for a wine assessment
203 experiment and consisted of 13 Chinese and 13 Western wine experts (Table 1). The study was
204 approved by the Office for Research Ethics and Integrity of the University of Melbourne (Ethics ID:
205 1852616). Each group had six female panellists and seven male panellists. The Western panellists
206 were defined as people who have lived in Australia for more than ten years. Chinese panellists were
207 defined as people who were born in China and had lived in Australia for no more than 18 months.
208 Bacteria genomic DNA was extracted from human saliva using QIAGEN® MagAttract® PowerSoil®
209 DNA KF Kit [47] and subjected to 16S rRNA amplicon sequencing on the Illumina platform

following the Earth Microbiome Project protocols (https://earthmicrobiome.org/protocols-and-standards/16s/). The raw data are available in NCBI Sequence Read Archive, with accession number PRJNA786805.

The raw sequences were processed using the same pipeline in the meta-analysis as described in section 2. Additionally, raw sequencing reads were denoised into zero-radios OTUs (zOTUs) by UNOISE3 pipeline [48], and taxonomically classified by classifiers trained on the full-length 16S rRNA gene SILVA v138 [49] database and eHOMD v15.22 [35], respectively. The affiliation between each ZOTUs and the originating OTU was determined using a customised code adapted from Stopnisek and Shade [50] and available at https://github.com/XINWEIR/SalivaryMicrobiome_MetaAnalysis. The relative abundance of taxa at the genus level in this cohort was used as the test set for the random forest model trained using the genus-level OTU assignment information in the meta-analysis.

Results

1. Inclusion of studies and sequences

In this study, we extracted the 16S rRNA gene amplicon sequencing data of healthy human saliva from 47 studies (Figure 1A). We abstracted data from subjects who had no diagnosed disease state, hereafter named “healthy”. Of course, subjects could have had subclinical diseases or may have altered health status for undisclosed reasons, but we considered that this would be true of the wider human population and therefore able to be included in our study. A total of 107,005,868 high quality 16S rRNA sequences were obtained. After removing all samples below 2,000 reads, 2206 samples with 909 features were retained. The retained samples included studies from 15 countries (Table S2). Most studies were conducted in three geographic regions: North America, Europe, and China (Figure 1B). Most sequences included in this meta-analysis were generated from the hypervariable region “V3-V4” and “V4”. Samples from studies targeted at “V1-V2” or the “V4-V5" region were classified as “others” (Figure 1C; Table S3). Similarly, saliva samples collected by unconventional methods like

237 “swab” were also classified as “others”. People who recorded a smoking habit, whether they smoke e-
238 cigarettes or tobacco, were categorised as smokers. Similarly, people who drink alcohol, regardless of
239 the frequency or the type of alcohol consumed, were classified as drinkers. The given age was treated
240 as a categorical factor, classifying into “18-30”, “31-55”, and “56+”. The samples without associated
241 information to a particular category or classified as “others” were excluded for downstream analyses
242 related to the impact of this category. For example, the following analyses measuring the effect of
243 hypervariable region on microbial profiles included only comparisons between the V3-V4 and V4
244 region (Figure 1C).

245 **2. Intrinsic and lifestyle factors have a significant effect on the host salivary microbiome**

246 Large variability between studies was observed in the number of reads, taxonomic profile, and alpha
247 diversity (Figure 2; Figure S1). Phylum *Bacteroidetes*, *Proteobacteria*, *Firmicutes*, *Fusobacteria* were
248 dominated among all studies, while their proportion varies (Figure 2A). When studies were grouped
249 by the geographic locations they originated from (coloured in Figure 2B, C, D), there is generally no
250 difference between their intra-community diversity, represented by Shannon, Chao1, and Simpson
251 indices. Only one study conducted in Qatar showed relatively lower Chao 1 index and higher
252 Simpson’s diversity indices than studies from other locations. However, it is hard to decide whether
253 such variation is caused by the geographic location or other technical variations.
254
255 Because of the large disparity of methodologies amongst the studies used in our global analysis, we
256 applied several different strategies for normalisation as described in Methods. When investigating the
257 influences of different categories using permutational multivariate analysis of variance
258 (PERMANOVA) tests, these normalising methods were combined as appropriate with different
259 distance metrics, including Bray-Curtis, weighted UniFrac, and Euclidean distance. Overall, the effect
260 of rarefaction (RAR), total-sum scaling (TSS), and the rarefied relative abundance transformation
261 (RRA) were very similar in the result of PERMANOVA (Figure 3A-D; Table S4). Meanwhile, the
262 centred log-ratio transformation (CLR) enlarged the variance induced by an unwanted technical
263 factor, namely the amplified hypervariable region, at all taxonomic levels.

264

265 The beta-diversity analyses showed that all metadata categories measured have a significant ($p<0.001$)
266 effect on the bacterial profile of human saliva at all taxonomic levels, adjusted for the study effect
267 (Table S4). However, only limited variation among samples has been explained by these factors ($R^2 <$
268 10%). In contrast, "study" accounts for around 35% of the variability between samples. At the OTU
269 level, the combination of weighted UniFrac distance and the total-sum scaling (TSS) transformed data
270 best minimised the variability raised by different hypervariable regions (Figure 3D). The results of
271 the unconstrained principal coordinate analysis (PCoA) are in agreement with the results of
272 PERMANOVA. When using Bray-Curtis dissimilarity and Euclidean distance, the samples separated
273 distinctly according to the hypervariable regions in PCoA plots, whereas the plot constructed using
274 the weighted UniFrac revealed the clusters formed by the samples from different geographic locations
275 (Figure S2). A distinct separation of samples from three main geographic locations (Figure 3F), with
276 more than half (58.0%) of the variance explained by the first two dimensions, using weighted UniFrac
277 distance. In contrast, the differences between locations were confounded by which hypervariable
278 regions were sequenced in the PCoA plot for Bray-Curtis dissimilarity (Figure 3E). The results
279 suggest that host intrinsic and lifestyle factors significantly influence the microbial profile in human
280 saliva, regardless of the variation induced by technical factors.

281 **3. A core microbiome is defined from saliva from healthy humans**

282 Despite the large intra- and inter-study variability, many OTUs still showed a consistently high
283 presence and relative abundance across studies (Figure S3). These persistent OTUs detected across
284 studies with different protocols could be functionally important for the salivary microbiome of healthy
285 adults. We wanted to identify the most widespread microbial taxa within a specific population that
286 allows us to better understand the broad structure of microbiomes and their potential functional
287 consequences [12]. The abundance and occurrence frequency of taxa are two important criteria used
288 to define the "core salivary microbiome". Conventionally, thresholds on these two parameters filter all
289 taxa detected, and taxa that meet both criteria can be classified as the core salivary microbiome [13].
290 Recently, a more standardised procedure based on abundance-occupancy distribution was proposed

291 [39]. We employed both strategies to define the core salivary microbiome to identify the microbial
292 features with high persistence and robustness in human saliva. Considering the sequences involved in
293 this meta-analysis were collected from studies targeted at different hypervariable regions, close-
294 referenced clustering at 97% identity were used to cluster sequences into OTUs. In addition, taxa
295 defined at 100% sequence identity may increase the redundancy in the dataset [39]. Therefore, the
296 core salivary microbiome was defined using the clustered OTUs at 97% identity. To begin, the core
297 OTUs were determined by filtering all OTUs based on mean relative abundance and occurrence
298 frequency using the criteria described in the Methods [13]. In total, 11.6% of all OTUs (105 OTUs)
299 were included as Core 1 (Figure 4A: MRA + OCC; Table S5). Meanwhile, according to the method
300 proposed by Shade and Stopnisek [39], OTUs were ranked depending on their occupancy across
301 studies, and the contribution of top-ranked OTUs to beta-diversity was expressed by Bray-Curtis
302 similarity and weighted UniFrac distance. Two groups of the core microbiome were prioritised by
303 these two indices, to give different inclusions in the core, consisting of the top 69 OTUs (using
304 weighted UniFrac; Figure 4A: BC) and 94 OTUs (using Bray-Curtis; Figure 4A: wUF) OTUs (Figure
305 S4).
306
307 Overall, sixty-eight OTUs were shared across all three methods (Figure 4A), accounting for 7.5% of
308 all OTUs detected and 72.5% of all 16S rRNA gene sequences after clustering and filtering (Figure
309 4B, C). *Firmicutes* account for nearly half (46.4 %) of all core OTUs, while only one OTU belongs
310 to *Saccharibacteria*. The mean relative abundance (MRA) of each OTU in sub-groups classified by
311 different factors was also measured (Figure 4D). On average, the core OTUs were highly prevalent
312 (73.2% ± 3.4% of cumulative relative abundance) in saliva samples across different levels in
313 subgroups classified by age, gender, geographic locations, hypervariable regions, sample type,
314 smoking, and drinking habits. The core OTUs were clustered into four main groups based on their
315 distribution pattern in sub-groups (Figure 4E). The eight OTUs affiliated to Cluster 1 showed overall
316 high abundance in all sub-groups. Cluster 2 consists of core OTUs with a slightly lower mean relative
317 abundance than Cluster 1 and higher intra-group variability. Notably, although having a higher MRA
318 than some members of Cluster 1, "476_9291" was still classified as Cluster 2. The reason could be its

319 biased presence in sub-groups. For example, the relative abundance of "476_9291" is higher in
320 samples from China than other two locations. The other two clusters contain OTUs with lower MRA
321 than Cluster 1 and 2, while variations can still be observed within sub-groups.
322
323 Bacteria-bacteria interactions play an important role in shaping microbiota. Therefore, co-occurrence
324 network analysis could be a useful approach for finding the most important part of the microbial
325 community. We applied a network analysis built by Spearman's correlations to investigate whether
326 the core OTUs defined were also important to the structure of the co-occurrence pattern. A correlation
327 with Spearman's correlation coefficient >0.5 or < -0.5 and *p*-value < 0.01 is considered statistically
328 robust [51, 52]. The resulting co-occurrence network contains 293 nodes and 1,424 significant
329 correlations (edges) (Figure S5). Small modules with less than seven nodes were not displayed.
330 Although most core OTUs have relatively low connectivity, they tend to associate with each other
331 rather than with rare OTUs. Nine OTUs were identified as potential "hub" OTUs based on their
332 centrality and the number of links in the network (Figure S5, 6). Because of the central position in the
333 network, the hub taxa are regarded as the key contributor to community stability. Two core OTUs
334 were also identified as hub taxa in the network, which are "122BU057" (*Megasphaera*
335 *micronuciformis*) and "524_3631" (*Veillonella atypica*). Compared to other "hub" taxa, they showed
336 lower connectivity and relatively high betweenness centrality (Figure S6).

337 **4. Geographic location is the host factor with the largest impact on bacterial composition**

338 To investigate which metadata category has the largest impact on the salivary microbiome, we
339 established random forest models to link the seven categories described above and a new category,
340 study, with the salivary microbiota data at seven taxonomic levels (OTU, phylum, class, order, family,
341 genus, and species). The effect of four normalisation methods was compared using the error rate
342 generated by random forest classification. In total, 224 random forest models were constructed
343 (Figure 5A). Among four normalisation methods, total-sum scaling produced the models that were, on
344 average, the most accurate. Generally, the random forest models built with microbial communities at
345 OTU levels have the lowest error rate (mean = 13.0%), while the models constructed at phylum levels

346 have the highest (mean =26.8%). The model built with the hypervariable region used for sequencing
347 was also the category that showed the lowest error rate (Figure 3G). Geographic locations
348 demonstrated the second important impact on the bacterial communities, with the lowest error rate
349 among biological factors. The random forest model constructed by the other two categories, sample
350 type and tobacco usage, also showed a relatively lower error rate than other categories. Study
351 constructed the models with high error rates at phylum (44.3% ± 2.5%) and class level (29.3% ±
352 2.1%). However, the error rate of models built with study rapidly dropped with the increase of
353 taxonomic levels, reaching 10.5% ± 5.1% at OTU level. Gender and age range led to poorly
354 performing models.
355
356 We wanted to determine whether the defined core microbiome could be used as biomarkers to
357 differentiate people categorised by intrinsic and lifestyle factors. The random forest models showed
358 high accuracy at OTU level were used (i.e., geographic location and smoking factors). The differential
359 OTUs induced by hypervariable regions and sample types were also analysed to exclude the influence
360 of technical factors. We further performed ten-fold cross-validation five times to measure the
361 importance of OTUs used to train the model. All OTUs before the point that the cross-validation error
362 curve starts to stabilise were defined as important OTUs. In total, we defined 59, 57, 34 and 70
363 important OTUs as biomarkers to differentiate samples according to geographic location, smoking
364 habit, hypervariable region, and sample type, respectively (Figure S7). Of these, 28, 10, 13 and 22
365 biomarkers were also classified as "core" (Figure 5B; Figure S8). Although 31 core OTUs showed the
366 importance in discriminating samples according to geographic location and smoking, nearly half of
367 them (15 OTUs) had the possibility of being confounded by technical factors (Figure 5B). After
368 excluding the OTUs that could be influenced by other factors, core OTU "322AK152" (*Bergeyella*
369 *sp.HMT_322*) was the OTU with the highest contribution to the classification of samples from three
370 geographic locations. Meanwhile, "122BU057" (*Megasphaera micronuciformis*) showed the highest
371 importance among the core differential OTUs specific to smoking, followed by "524_3631"
372 (*Veillonella atypica*). We were surprised to find that these two OTUs were the only two OTUs that
373 were defined as hub taxa in the co-occurrence network analysis (Figure S6).

374 **5. The salivary microbiota as biomarkers to differentiate Chinese and Western**

375 We further analysed the changes caused by geographic locations in higher taxonomic hierarchies,
376 where many differences have been revealed. Of particular interest were taxa under
377 phylum *Synergistetes* and *Spirochaetes*, Class *Mollicutes* and *Betaproteobacteria*,
378 Family *Clostridiales*, and genus *Prevotella*. Interestingly, many taxa showed a higher relative
379 abundance in the Chinese samples, both compared to North American samples and compared to
380 European samples (Figure 6A). It suggested that the variance induced by geographic locations may be
381 dominated by the differences between samples from Chinese and Western people. Therefore, we
382 combined samples from North America and Europe into a single group, "Western." Compared to the
383 Chinese grouping, the Western group has significantly lower within-sample diversity (alpha-diversity)
384 (Wilcoxon rank-sum test, $p < 0.001$; Figure 6B, C). Next, we examined the differences between
385 Chinese and Western in the salivary microbiota at the genus and species level (Table S6, S7). Besides
386 establishing a random forest model, we also identified differential taxa using ANCOM-BC, adjusting
387 for the hypervariable region. We found 48 genera identified as significantly different by both methods
388 (Figure 6D, Table S6). Among them, *Arachnia*, *Filifactor*, *Ottowia*, *Neisseria*, *Aggregatibacter*, one
389 genus from *Gracilibacteria*, one genus from *Clostridiales*, and three other genera belonging to the
390 family *Peptostreptococcaceae* were strongly (standardized effect size>10, mean decreased
391 accuracy >10) enriched in Chinese samples.
392 Meanwhile, *Prevotella*, *Scardovia*, *Bergeyell*, *Veillonella*, *Oribacterium*, and one genus belonging to
393 the family *Erysipelotrichaceae* were strongly (standardized effect size>7 mean decreased
394 accuracy >10) enriched in Western samples.
395
396 Finally, we performed the functional prediction-based 16S rRNA gene profiles to investigate whether
397 differences in the salivary microbiota between Chinese and Western affect its function. Two methods,
398 ANCOM-BC, and random forest model were used to identify which pathways were differential
399 between Chinese and Western. The result of ANCOM-BC indicated that 69 pathways related to
400 metabolism were differentially abundant between the two groups. The random forest classification

401 model established using KEGG pathways demonstrated an error rate of 10.01% and revealed 46
402 differential pathways. Among them, thirty pathways belonging to nine upper pathways (level 2) were
403 simultaneously defined by two methods as differing in abundance between Chinese and Western
404 (Figure 6E, Table S8). A variety of pathways was in higher abundance in Chinese samples. The
405 enrichment of these pathways in Chinese samples was mainly associated with the increased
406 abundance of *Neisseria* and *Lautropia* and the depleted abundance of *Prevotella*, *Veillonella*,
407 and *Atopobium*. Notably, three lipid metabolism pathways enriched in Chinese samples, including
408 "Ether lipid metabolism" (ko00565), "alpha-Linolenic acid metabolism" (ko00592), and "Linoleic
409 acid metabolism" (ko00591), have the highest standardised effect size (W statistics, Table S8). The
410 enrichment of these pathways related to lipid metabolism has been positively associated with the
411 higher abundance of *Neisseri*a in Chinese. *Neisseria* may have also contributed to the pathway
412 "Carotenoid biosynthesis" (ko00906). Another metabolic pathway related to the metabolism of
413 terpenoids and polyketides, "Sesquiterpenoid and triterpenoid biosynthesis" (ko00909), showed a
414 positive correlation with a genus belong to *Peptostreptococcaceae*. In contrast, only one pathway
415 named "Flavone and flavonol biosynthesis" (ko00944) was enriched in the saliva samples from
416 Western. A strong positive correlation has been demonstrated between this pathway and the increased
417 abundance of *Veillonella* in the samples from the Western grouping.

418 **6. Validation of the core in an independent Australian cohort**

419 To validate the prevalence of the core OTUs in human saliva, we collected saliva samples from 13
420 Chinese and 13 Western participants in Melbourne and sequenced the extracted DNA with 515F-
421 806R primers. In total, 841,188 high-quality 16S rRNA sequences were obtained, which clustered into
422 397 OTUs with 97% identity to the HOMD database. Among them, the core OTUs we defined in the
423 meta-analysis showed high relative abundance (78.3 ± 6.9%) in all collected samples. To increase the
424 accuracy of the OTU assignment, we denoised the sequencing reads using the UNOISE3 pipeline and
425 generated ZOTUs with 100% sequence identity. After re-clustering the core OTUs defined in the
426 meta-analysis to ZOTUs, we observed that 59 of the identified OTUs in this independent dataset
427 consisted of 87 ZOTUs, and thus made up close to 80% of the relative abundance (Figure 7A).

428 Although some sequences belonging to the same ZOTU are clustered to different OTUs, all the core
429 OTUs contain at least one highly abundant ZOTU. The taxonomic profiles of the global core
430 annotated by the HOMD database and the ZOTUs annotated by the SILVA database were very
431 similar at the genus level (Figure 8A).
432
433 We wanted to verify the observed differences between the OTUs in saliva samples from Chinese and
434 Western people in this independent dataset. Although two groups did not differ significantly when
435 considering the Shannon diversity index (Wilcoxon rank-sum test, $p = 0.073$; Figure 8B), the Chinese
436 group showed a higher Chao 1 Index than the Western group (Wilcoxon rank-sum test, $p < 0.001$;
437 Figure 8C), which is in agreeance with the result of the meta-analysis. Meanwhile, no significant
438 differences were observed between the two groups by beta-diversity analyses (Bray-Curtis and
439 weighted uniFrac distance, Table S9). We used the random forest classification model constructed
440 using the genus level profile of the large-scale dataset to predict the Chinese and Western samples in
441 this independent study. This dataset's genus-level relative abundance table was prepared from both the
442 OTU table with 97% identity to HOMD v15.1 database (Figure 8D) and the ZOTU table annotated by
443 the HOMD v15.22 database (Figure 8E). The accuracy of both predictions was relatively low, with
444 57.7% for the OTU table and 50% for the ZOTU table. Interestingly, most Western samples were
445 correctly classified, while most samples from Chinese participants were classified as being 'Western'
446 in this analysis.
447
448 **Discussion**

449 There is ample knowledge on the disease-affected salivary microbiota, yet our perspectives to the
450 bacteria present in healthy humans remains limited. Our systematic selection of studies, together with
451 the re-analysis of the 16S rRNA amplicon sequencing data from 47 studies offers a comprehensive
452 description of the salivary microbiome presented in adults without diagnosed disease. Our study has
453 defined the core members of salivary bacterial communities across 2211 samples from 47 studies and
454 has used metadata captured in these studies to investigate the role of different intrinsic and extrinsic

455 factors on the occurrence of these core. It is clear that core members differ between geographic
456 locations of collected saliva, and our analysis shows that Chinese participants are different from
457 Western participants (encompassing European and North American studies). A prediction of the
458 pathways enriched in each collective indicates that bacterial metabolic pathways are likely to
459 influence the aroma and flavour perception of foods. These results show that despite the core
460 microbial members of saliva being common across humans, there are differences, likely due to diet.
461 We suggest that the aroma and flavour of foods and beverages are likely to be differently affected in
462 healthy humans across the globe, meaning that preference and consumption of different foods is likely
463 to be prioritised. These results have important consequences for food and beverage design,
464 composition, and dietary advice across the globe.
465
466 Based on the abundance-occupancy pattern, the definition of core microbiome highlighted the
467 persistent and conserved microbial communities in human saliva across the globe. Here, we compared
468 two approaches adapted from two studies (references) to defining the core. The method adapted from
469 the study of Wu et al. [13] is a relatively conventional strategy that has been chosen by many other
470 studies investigating different ecosystems, such as soil [53], compost [54], wastewater treatment
471 plants [55], and human's intestinal system [56]. The thresholds were simply setting on each taxon's
472 mean relative abundance and occupancy across all samples. For studies aimed to determine the spatial
473 or temporal core microbiome, additional thresholds will be added on the prevalence of taxa within
474 sub-groups. However, the thresholds used by different studies are usually arbitrary. Some studies have
475 even adopted only abundance or occupancy alone as criteria for defining the core. Therefore, a
476 generalised approach for defining the core microbial members from diverse datasets based on
477 abundance-occupancy was also applied in this study [39]. Rather than over space or time, we
478 determined the occupancy of OTUs according to their detection over study. When evaluating the
479 contribution of top-ranked OTUs in occupancy to the beta-diversity of the community, we further
480 used weighted Unifrac due to its effectiveness in minimising the biases induced by the selection of
481 hypervariable regions.

482 The resulting core members identified by these methods have a lot in common. A majority of the core
483 OTUs defined by Shade's method is also included in the cores defined by Wu's method, suggesting
484 the recently developed multi-step approach is effective in determining taxa with high prevalence. The
485 general high relative abundance of the core across different sub-groups emphasised the utility of this
486 pipeline in identifying the persistent members across diverse datasets. Most of the core salivary
487 microbiota we defined had been proposed in previous studies as prevalent bacteria in the human oral
488 cavity that persistently span across different individuals [11, 57-59]. The dominant genus of the core
489 we defined, *Streptococcus*, *Neisseria*, and *Prevotella*, were concluded as core human salivary
490 microbiome by a recent study based on the MG-RAST data [60]. Ten OTUs belonging to genus
491 *Streptococcus* were included in the "core", two of which were classified to the cluster with overall the
492 highest relative abundance across all dimensions. The prevalence of *Streptococcus* we observed is
493 consistent with a previous study defining the healthy core from the 454 pyrosequencing results of
494 three individuals [11]. The most abundant core OTU we found, *Streptococcus*
495 *oralis* subspecies *dentisani*, has been documented in previous studies as potential oral health-
496 promoting organisms and being highly abundant at various oral niches of healthy humans [61, 62].
497
498 We further conducted a co-occurrence network analysis to investigate the role of these core
499 microbiota in shaping the microbial community and found the co-existence between many members
500 of the core (Figure S5). The presence of the rare OTU that became the hub suggests that although
501 some taxa are not persistently detected across the community, they may still be important for the
502 overall structure of the salivary microbiome. It has been proposed that the oral microbiota of healthy
503 individuals is both homeostatic and dynamic [57]. The core microbial members consistently present in
504 human saliva identified here may explain the stability of oral microbiota to some extent.
505 Moreover, we conducted an independent study to verify the prevalence of the core defined from the
506 published studies. The high relative abundance and occurrence of the original core OTUs in this
507 independent cohort suggest that the core human salivary microbiome we defined can be applied to
508 different datasets. After re-clustering the core OTUs to ZOTUs with 100% sequence identity, we may
509 conclude that the same members constitute the core, even if different taxonomic resolution is applied.

510

511 Besides the core microbiome, there are “variable” microbiota in human microbial communities, which
512 vary among individuals because of unique lifestyle and genetic factors [63]. We performed analyses
513 for beta-diversity of samples (Figure 3) and random forest classifications (Figure 5A), demonstrating
514 several factors-both technical and physiological-significantly discriminated between sub-populations.
515 Because of the high heterogeneity between studies in their methodology, large inter-study variability
516 was the main factor that affected the observed salivary microbiota [64, 65]. As one of the main
517 technical factors that may induce the variation, the impact of chosen hypervariable regions for
518 sequencing on driving microbial community structures has been confirmed by our study. In addition,
519 the selection of different primers for the same region and DNA extraction methods may also lead to
520 technical variations [66]. However, studies’ choices of primers and DNA extraction protocols are
521 more diverse than hypervariable regions (Table S2), making it difficult to group them into categories
522 as simple as for the variable regions. By adjusting the analyses by study, the variability caused by
523 study-specific technical factors other than hypervariable regions could also be covered. For future
524 studies, it is important to establish a standardised DNA extraction and sequencing protocol on a global
525 scale. During the literature search, we found that the inter-study variation may also be attributed to the
526 criteria of recruiting participants. Although all samples included in this meta-analysis were collected
527 from the control groups and population without specific diagnosed disease, the definition of 'healthy'
528 varies. For example, the use of antibiotics was not always considered as an exclusion criterion, and
529 when included, different time intervals were adopted. However, for this study, we aimed to construct
530 a microbial community that reflects the salivary microbiota of real-life consumers. Therefore, samples
531 collected from individuals without known severe systemic diseases were included.

532

533 Besides the influence of variables related to study design, this study has also revealed the importance
534 of various host intrinsic and lifestyle factors. We observed the change in salivary microbial
535 composition induced by smoking habits. As a recognised risk factor to oral health, the role of smoking
536 in shaping the human oral microbiome has received increasing attention [67]. Previous studies have
537 reported the change of *Megasphaera micronuciformis* caused by smoking in the microbial profile of

538 the human tongue surface [68] and upper gastrointestinal tract [69]. In agreement with these findings,
539 we identified a core OTU belonging to *Megasphaera micronuciformis* as a biomarker for reported
540 smoking. In addition, we found the differential abundance of an OTU belonging to *Veillonella atypica*
541 between smokers and non-smokers, which agrees with a study where this species already found to be
542 increased in the saliva [70]. Another strong determinant of the salivary microbiome we defined was
543 "sample type". As we found in this study, the difference between the mouthwash sample and the other
544 two collection methods is greater than the difference between the stimulated and unstimulated saliva.
545 Contrary to the result of Jo et al. [29], the OTU belonging to *Neisseria flava* has not been identified as
546 differential taxa for the type of saliva. The microbial composition of alcohol drinkers and non-
547 drinkers was also found to be different, while the small sample size acquired hindered our ability to
548 draw large-scale downstream conclusions. Although the significant differences we observed in
549 salivary microbiota were also attributed to the gender and age of participants, the variations they
550 explained are relatively low compared to other factors.
551
552 Geographic location has been identified as the host physiological factor with the largest impact on
553 salivary microbiota (Figure 5A). Although it only explained limited variability between samples'
554 microbial profiles, the observed variations were robust to the heterogeneity induced by different
555 hypervariable regions used (Figure 3E, F). To date, little is known regarding the influence of
556 geographic locations on the human salivary microbiome. A comparative study reported the
557 differences in saliva microbial composition in Alaskans, Africans, and Germans [33]. To our
558 knowledge, the geographically structured microbial communities have not been observed in healthy
559 human saliva based on the large-scale dataset used in this study. Our result also demonstrated several
560 core OTUs that may differentiate saliva samples from North America, Europe, and China. It suggests
561 that the global prevalent core microbiota is not necessary to be stable across populations. Given the
562 high abundance and occupancy frequency of the core microbiome, we would expect these taxa to be
563 effective indicators to predict the geographic background of saliva donors.
564

565 Due to the sometimes large differences in culture and lifestyle between Western and non-Western
566 populations, we further grouped our data into Western and Chinese samples. The comparison between
567 Western and non-Western populations has already been applied to the human gut microbiota, whereas
568 less is known about the saliva microbiota [71]. Our study found a difference in the abundance of
569 *Veillonella* spp between saliva from Western and Chinese people, where *Veillonella* was generally
570 higher in Western samples. Such differences may influence the flavone and flavonol biosynthesis
571 pathways in the oral cavity. Our previous study revealed the Western-born and Chinese-born wine
572 experts had different responses to the astringency of wine. Since flavonol is a well-known constituent
573 of wine-related to the bitterness and astringency perception [72], we would hypothesis that the
574 enrichment of *Veillonella* in Western may affect their sensitivity to the phenolic compounds in wine.
575 It has also been suggested that the regular consumption of flavonoid-rich foods, such as oolong tea,
576 may increase the abundance of *Veillonella* spp. in human saliva [73]. These influences can potentially
577 be the bridges that link the differences between Chinese and Western groups in sensory evaluation
578 and their salivary microbiota together. Because the amplicon sequencing data cannot speak directly to
579 the functional sequences of the observed difference, shotgun metagenome sequencing will be
580 necessary to verify the exact association. *Prevotella* abundance has previously been reported to be
581 enriched in the gut microbiota of non-Western populations [74], which is opposed to our observation
582 in saliva. However, there is substantial species-level diversity in *Prevotella* (Table S7), making it
583 plausible that different species belonging to *Prevotella* may respond differently to the geographic
584 background [75].
585
586 Although the independent dataset we collected did not show differences between samples taken from
587 Chinese and Western participants, the results of random forest classification may lead to some
588 interesting hypotheses. The prediction of models revealed that most of the Chinese samples in this
589 cohort were classified as Western. The donors of these samples were wine experts, of Chinese
590 ethnicity, born in China and living in Australia for no more than 18 months. Recent studies reported
591 that immigrants from Asia experience a "Westernization" of gut microbiota induced by dietary
592 acculturation [74, 75]. We hypothesise that such a phenomenon may also happen in salivary

593 microbiota. It may suggest that dietary pattern is a more important determinant than ethnicity in
594 shaping the salivary microbiota of the participants, leading to variation among different geographic
595 locations that we observed in the meta-analysis.
596

597 **Conclusions**

598 In summary, we have defined a core bacterial community in saliva from healthy humans, and this core
599 demonstrated both stability and variability among populations. The prevalence of the core members of
600 the saliva microbiome has been confirmed in an independent cohort. We have revealed the influence
601 of various host factors, such as geographic locations, incidence of smoking and drinking, on the
602 salivary microbiome. We also identified microbial and functional biomarkers to differentiate the
603 Chinese and Western people, underlying the potential relationship between salivary microbiota and
604 sensory perception. Results in this work will provide foundational information to inform future
605 studies to understand the similarities and differences in saliva microbial composition, potentially
606 associating oral to aroma and flavour perception of foods.
607

608 **Availability of data and materials**

609 The sequencing data supporting the conclusion of the meta-analysis in this article are available in
610 publicly accessible databases (full details can be found in Table S2). The sequencing data generated
611 and/or analysed during the current study are available in the NCBI Bioproject repository,
612 PRJNA786805 (https://www.ncbi.nlm.nih.gov/bioproject/786805). Original scripts generated during
613 the current study are available in Github
614 (https://github.com/XINWEIR/SalivaryMicrobiome_MetaAnalysis).
615

Acknowledgements

This study was funded by the Faculty of Veterinary and Agricultural Sciences at the University of Melbourne. JL and XR gratefully acknowledge a Melbourne Research Scholarship administered by the University of Melbourne.

Legends-Results

Figure 1. Overview of literature search procedure and metadata of included studies. a) Large-scale literature searching and data filtering process, followed by the number of samples submitted to the bioinformatic analyses; b) The locations of studies, the scale of symbols that reflect the number of samples of each study; c) Distribution of metadata categories.

Figure 2. Summary of taxonomic composition and alpha diversity of included studies. A) The mean community composition of each study at the phylum level; The alpha-diversity measured by B) Shannon index; C) Chao 1 index; D) Simpson's index, the colour of boxes stands for the geographic location of the studies. The horizontal bars within boxes represent medians. The tops and bottoms of boxes represent the 75th and 25th percentiles, respectively.

Figure 3. The variability in human salivary microbiota have been explained by different factors. Among them, hypervariable regions and geographic locations have the largest impact. The effect of the categories on the clustering of the sample was measured using PERMANOVA at four taxonomic levels: family **(A),** Genus **(B),** species **(C)** and OTU level **(D).** The colour indicates the different combinations of normalisation (TSS, Total-sum scaling; RRA, Rarefied relative abundance; CLR, Centred log ratio) and indices (BC, Bray-Curtis; EUC, Euclidean; wUF, weighted uniFrac). Because the results of rarefication (RAR) were very close to TSS and RRA, they were not displayed in the figures. Principal coordinate analysis (PCoA) with Bray-Curtis **(E)** and weighted uniFrac **(F)** showing the differences between samples from North America, Europe, and China.

Figure 4. The core OTUs defined by abundance-occupancy pattern. A) Venn diagram showing the interaction between three methods used to define the core. Sixty-eight OTUs were defined as the

642 core for all methods. (MRA+OCC: The thresholds were setting on mean relative abundance and
643 occupancy to define the core; BC: The method adapted from Shade and Stopnisek using Bray-Curtis
644 similarity; wUF: The method adapted from Shade and Stopnisek using weighted uniFrac distance). **B)**
645 Pie chart showing the number of the core (pink) versus other OTUs (blue) identified in percentage. **C)**
646 Pie chart showing the relative abundance of the core and other OTUs across all samples. **D)** Relative
647 abundance of 68 core OTUs across subgroups classified by seven categories. **E)** Heatmap showing the
648 log-transformed mean relative abundance of each core OTU at each level of different categories.
649

650 **Figure 5. Salivary microbiome members which significantly contribute to categorisation of**
651 **metadata. Random Forest models showed the impact of categories on salivary microbiome and**
652 **the core OTUs contributing to accuracy of these models. A)** Error rate (%) for the random forest
653 classifications conducted with samples grouped by eight different categories. **B)** Phylogenetic tree
654 indicates the taxonomic information of 68 core OTUs. The coloured squares between the tree and the
655 annotation of phylum indicate the OTUs that were defined by the Random Forest model as
656 "important" for distinguishing between different levels in each category. The bars on the outmost ring
657 showing the mean relative abundance of each OTU.

658 **Figure 6. Distinct microbial profiles are evident in the saliva samples from Chinese and Western**
659 **adults. A)** Taxonomic hierarchies show the relative enrichment of taxa in three geographic locations
660 at phylum through species level. Coloured nodes indicate log2-fold increase in median abundance of
661 the group in x-axis (pink) or y-axis (blue). Only taxa showed significant changes (false discovery rate-
662 adjusted Wilcoxon rank sum $q < 0.05$) are displayed. **B)** and **C)** Comparison of salivary microbial
663 alpha diversity between the Chinese and Western samples, calculated by Chao1 (B: $p < 0.001$,
664 Wilcoxon rank-sum test) and Shannon index (C: $p < 0.001$, Wilcoxon rank-sum test). **D)** Differential
665 abundant genera identified between saliva from Chinese and Western samples. The panel on the left
666 indicates the standardised effect sizes (W statistic) estimated via the difference on relative abundance
667 using ANCOM-BC (taxa enriched in Western samples have a value shifted to right, whereas taxa
668 enriched in Chinese samples have a value shifted to left); The panel in the middle shows the relative

669 abundance of selected genera; the panel on the right indicates the Mean Decrease Accuracy of the
670 random forest model established. **E)** Spearman's correlation coefficients were calculated between
671 each pairwise comparison of differential genus and KEGG pathway. Only significantly correlated
672 comparisons (p<0.01, FDR adjusted Spearman's rank correlation) are displayed. The only Western-
673 enriched pathway is marked in pink.

674 **Figure 7. An independent cohort verifies the definition of core microbiome membership**
675 **but cannot classify based on cultural background. A)** Alluvial plot showing the affiliation of
676 ZOTUs to their originating core OTUs defined in the meta-analysis. **B)** and **C)** Comparison of salivary
677 microbial alpha diversity between the Chinese and Western samples, calculated by Shannon (B: p =
678 0.073, Wilcoxon rank-sum test) and Chao1 index (C: p< 0.001, Wilcoxon rank-sum test). **D)** and **E)**
679 The prediction of the cultural backgrounds of the samples according to the random forest
680 classification model constructed using the genus profiles of samples in the meta-analysis. The genus
681 level profiles of samples processed by **D)** closed-reference clustering with 97% sequence identity and
682 **E)** UNOISE3 denoising with 100% sequence identity were used as the test set.

683 **Legends-Supplementary Figures**

684 **Figure S1 Average sequencing depth and rarefaction curve for the complete 16S rRNA dataset.**
685 **A)** Mean read number of samples from each study. The dash line indicates that all samples below this
686 depth (depth = 2,000) have been removed. **B)** The rarefaction curve reflects the increase of sample's
687 Shannon index with sequencing depth. The curve was basically stabilised at sequencing depth = 2000.
688

689 **Figure S2 The effect of different combinations between normalisations and distance matrices on**
690 **reducing the impact of hypervariable regions in the PCoA plot.** PCoA plots showing Bray-Curtis
691 dissimilarity **(A-C)**, weighted uniFrac distance **(D-F)**, Euclidean distance **(G-J)** under rarefaction
692 (RAR) **(A, D, G)**, total-sum scaling (TSS) **(B, E, H)**, rarefied relative abundance transformation
693 (RRA) **(C, F, I)** and centred log-ratio transformation **(J)**. Percentage of variances explained by the
694 first two principal coordinates are shown on the axes.

Figure S3 Phylogenetic tree showing the presence and absence of top 500 OTUs with highest mean relative abundance. The colour strips on the innermost ring indicates which phylum the OTUs belong to. The presence of coloured circles on the 17 rings in the middle indicate that an OTU was found in a specific level of a sub-group. The grey bars on the outermost layer represent how many studies that an OTU presented. The scale lines are used to highlight the number of 10, 20 and 40.

Figure S4 The percentage contribution of the top-ranked OTUs to the beta-diversity of the dataset. The beta-diversity is calculated for the whole dataset and for only the top-ranked OTUs using both Bray-Curtis similarity (A) and weighted uniFrac distance (B). The percentage contribution of top-ranked OTUs is calculated by dividing the beta-diversity among top-ranked OTUs using the beta-diversity of the whole dataset. The dash lines indicate the last points at which the increase on the contribution is 2%. All OTUs before this point (the point itself was also included) were defined as "core".

Figure S5 Bacterial co-occurrence network verifies the role of identified core salivary microbiome members. Small modules with less than seven nodes were not displayed. The size of nodes is proportioned to the connectivity of nodes (node degree). Core OTUs from Figure 4 are indicated as yellow, while rare OTUs are in green. The edges between nodes represent the strength of the correlation (Spearman's correlation coefficient, $\rho \geq 0.5$, $p < 0.01$). The shape of "hub" OTUs are indicated as squares with the OTU name displayed.

Figure S6 The Betweenness and Closeness centrality of OTUs involved in the network analysis. The "hub" OTUs were identified as OTUs with either high connectivity (node degree) or centrality (betweenness **(A)** and closeness **(B)** centrality). The accession number of "hub" OTUs are indicated. The core OTUs are marked in yellow, while the rare OTUs were coloured in green.

722 **Figure S7 The optimal number for defining the biomarker OTUs of four categories.** The
723 contribution of the OTUs used to differentiate the levels in each category on ten-fold cross-validation.
724 The OTUs were ranked in the order of importance. The dash lines represent the point at which the
725 curve starts to become overall stable.
726
727 **Figure S8 The top important OTUs identified by Random Forest classification model**
728 **established by the relative abundance of human salivary microbiota.** The differential OTUs
729 defined were ranked in descending order of their importance (Mean Decrease Accuracy). The colour
730 of bars reflects the phylum level information of OTUs. The Mean Decrease Accuracy bar of core
731 OTUs were marked with bolded black borders.
732

733 ## Legends-Supplementary Tables

734 **Table S1** The terms used for searching in databases, including Medline, EMBASE, and Web of
735 Science.
736
737 **Table S2** Studies included after large-scale literature searches that met all the inclusion criteria.
738
739 **Table S3** The full metadata used in this study, including 2206 samples with unique accession
740 numbers.
741
742 **Table S4** The influence of seven factors at seven taxonomic levels on human salivary microbial
743 communities, measured by PERMANOVA with adonis2 function (permutation = 999).
744 PERMANOVA models were adjusted for study.
745
746 **Table S5** Core OTUs of adults' saliva microbiome defined by **a)** method adapted from Wu et al.
747 (2019) ("1"= yes, "0"=no), **b)** method adapted from Shade and Stopnisek (2019).
748

749 **Table S6** Genus with differential abundance between samples from Western and Chinese people
750 identified by both ANCOM-BC and Random Forest model, adjusted for hypervariable regions
751 sequenced.
752
753 **Table S7** Species with differential abundance between samples from Western and Chinese people
754 identified by both ANCOM-BC and Random Forest model, adjusted for hypervariable regions
755 sequenced.
756
757 **Table S8** KEGG pathways with differential abundance between samples from Western and Chinese
758 people identified by both ANCOM-BC and Random Forest model, adjusted for hypervariable regions
759 sequenced.
760
761 **Table S9** The differences between Chinese and Western participants in the independent cohort,
762 measured by PERMANOVA with adonis2 function (permutation = 999). The PERMANOVA model
763 was adjusted for the gender and age range of participants.

764 **Table 1** Studies included after large-scale literature searches that met all the inclusion criteria.

Study	Database	Accession Number	Location	Hypervariable Region	Number of Samples	Sample Type
[76]	SRA	PRJNA323410	India	V3-V4	51	Unstimulated
[77]	SRA	PRJNA577839	Europe	V3-V4	37	Stimulated
[78]	SRA	SRP125370	India	V3-V4	12	Other
[79]	SRA	PRJNA380250	United States	V1-V2/V4-V5	4	Unstimulated
[80]	ENA	PRJEB9010	Europe	V3-V4	70	Stimulated
[81]	SRA	PRJNA438728	India	V3-V4	30	Mouthwash
[82]	SRA	PRJNA321534	United States	V4	18	Unstimulated
[83]	SRA	PRJNA361501	Europe	V3-V4	99	Other
[84]	ENA	PRJEB37445	China	V3-V4	436	Stimulated
[57]	ENA	PRJEB11529	Canada	V3-V4	96	Stimulated
[85]	SRA	PRJNA609244	Europe	V1-V2	13	Unstimulated
[86]	SRA	PRJNA503603	China	V3-V4	24	Unstimulated
[87]	SRA	PRJNA495719	China	V3-V4	22	Unstimulated
[88]	ENA	PRJEB37299	Europe	V4	11	Stimulated
[89]	SRA	PRJNA386665	China	V4	127	Unstimulated
[90]	SRA	PRJNA578492	China	V3-V4	14	Unstimulated
[91]	SRA	PRJNA326866	Europe	V1-V2	20	Unstimulated
[92]	SRA	PRJNA484857	China	V3-V4	15	Unstimulated
[93]	ENA	PRJEB21767	Europe	V3-V4	47	Stimulated

[94]	SRA	PRJNA586897	China	V3-V4	64	Unstimulated
[95]	SRA	SRP113577	China	V4-V5	71	Unstimulated
[96]	SRA	PRJNA292800	United States	V4	28	Unstimulated
[97]	SRA	PRJNA602902	United States	V3-V4	119	Stimulated
[98]	SRA	PRJNA623352	Europe	V3-V4	17	Unstimulated
[99]	SRA	PRJNA356414	Europe	V3-V4	10	Unstimulated
[100]	SRA	PRJNA602902	United States	V3-V4	40	Stimulated
[101]	SRA	PRJNA587625	Qatar	V3-V4	73	Unstimulated
[102]	SRA	PRJNA612815	Europe	V1-V3/V3-V4/V4-V5/V6-V8	44	Other
[103]	SRA	PRJNA413706	United States	V4	30	Stimulated
[104]	SRA	PRJNA601054	China	V3-V4	27	Unstimulated
[105]	SRA	PRJNA321349	United States	V3-V4	20	Stimulated
[106]	SRA	PRJNA578951	China	V3-V4	22	Unstimulated
[107]	SRA	PRJNA539937	United States	V3-V4	25	Other
[108]	SRA	PRJNA421234	New Zealand	V3-V4	7	Unstimulated
[109]	Qiita	10823	United States	V4	150	Mouthwash
[110]	SRA	PRJNA306560	China	V4	18	Stimulated
[110]	SRA	PRJNA414355	China	V3-V4	37	Unstimulated
[111]	SRA	PRJNA587078/	China	V3-V4	21	Unstimulated
[112]	SRA	PRJNA556311	China	V3-V4	20	Unstimulated

[113]	ENA	PRJEB18476	Europe	V4	11	Stimulated
[114]	SRA	PRJNA414682	China	V3-V4	20	Unstimulated
[115]	SRA	PRJNA634162	United States	V4	75	Other
[116]	SRA	PRJNA515166	Malaysia	V3-V4	72	Unstimulated
[117]	SRA	PRJNA542018	China	V4	10	Unstimulated
[118]	SRA	PRJNA586723	China	V3-V4	60	Stimulated and unstimualted
[119]	SRA	PRJNA534340	China	V3-V4	120	Stimulated and unstimualted
[120]	SRA	PRJNA598080	Europe	V3-V4	12	Unstimulated

References

1. Mosca AC, Chen J. Food-saliva interactions: Mechanisms and implications. Trends Food Sci Technol. 2017;66:125-34. https://doi.org/https://doi.org/10.1016/j.tifs.2017.06.005.
2. Liu D, Jiang X, Zheng HJ, Xie B, Wang H, He T, et al., editors. The modularity of microbial interaction network in healthy human saliva: Stability and specificity. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017 13-16 Nov. 2017.
3. Acharya A, Chan Y, Kheur S, Jin LJ, Watt RM, Mattheos N. Salivary microbiome in non-oral disease: A summary of evidence and commentary. Arch Oral Biol. 2017;83:169-73. https://doi.org/https://doi.org/10.1016/j.archoralbio.2017.07.019.
4. Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. Proceedings of the National Academy of Sciences. 2016;113(6):E791-E800. https://doi.org/10.1073/pnas.1522149113.
5. Shaw L, Ribeiro ALR, Levine AP, Pontikos N, Balloux F, Segal AW, et al. The Human Salivary Microbiome Is Shaped by Shared Environment Rather than Genetics: Evidence from a Large Family of Closely Related Individuals. mBio. 2017;8(5):e01237-17. https://doi.org/doi:10.1128/mBio.01237-17.
6. Muñoz-González C, Cueva C, Ángeles Pozo-Bayón M, Victoria Moreno-Arribas M. Ability of human oral microbiota to produce wine odorant aglycones from odourless grape glycosidic aroma precursors. Food Chem. 2015;187:112-9. https://doi.org/https://doi.org/10.1016/j.foodchem.2015.04.068.
7. Parker M, Onetto C, Hixson J, Bilogrevic E, Schueth L, Pisaniello L, et al. Factors Contributing to Interindividual Variation in Retronasal Odor Perception from Aroma Glycosides: The Role of Odorant Sensory Detection Threshold, Oral Microbiota, and Hydrolysis in Saliva. Journal of Agricultural and Food Chemistry. 2020;68(38):10299-309. https://doi.org/10.1021/acs.jafc.9b05450.
8. Piombino P, Genovese A, Esposito S, Moio L, Cutolo PP, Chambery A, et al. Saliva from Obese Individuals Suppresses the Release of Aroma Compounds from Wine. PLoS One. 2014;9(1):e85611. https://doi.org/10.1371/journal.pone.0085611.
9. De Filippis F, Parente E, Zotta T, Ercolini D. A comparison of bioinformatic approaches for 16S rRNA gene profiling of food bacterial microbiota. Int J Food Microbiol. 2018;265:9-17. https://doi.org/https://doi.org/10.1016/j.ijfoodmicro.2017.10.028.
10. Soriano-Lerma A, Pérez-Carrasco V, Sánchez-Marañón M, Ortiz-González M, Sánchez-Martín V, Gijón J, et al. Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. Sci Rep. 2020;10(1):13637. https://doi.org/10.1038/s41598-020-70141-8.
11. Zaura E, Keijser BJF, Huse SM, Crielaard W. Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol. 2009;9(1):259. https://doi.org/10.1186/1471-2180-9-259.
12. Risely A. Applying the core microbiome to understand host–microbe systems. J Anim Ecol. 2020;89(7):1549-58. https://doi.org/https://doi.org/10.1111/1365-2656.13229.
13. Wu L, Ning D, Zhang B, Li Y, Zhang P, Shan X, et al. Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nat Microbiol. 2019;4(7):1183-95. https://doi.org/10.1038/s41564-019-0426-5.
14. Guo R, Zheng Y, Zhang L, Shi J, Li W. Salivary microbiome and periodontal status of patients with periodontitis during the initial stage of orthodontic treatment. Am J Orthod Dentofacial Orthop. 2021;159(5):644-52. https://doi.org/https://doi.org/10.1016/j.ajodo.2019.11.026.

814 15. Relvas M, Regueira-Iglesias A, Balsa-Castro C, Salazar F, Pacheco JJ, Cabral C, et al.
815 Relationship between dental and periodontal health status and the salivary microbiome:
816 bacterial diversity, co-occurrence networks and predictive models. Sci Rep. 2021;11(1):929.
817 https://doi.org/10.1038/s41598-020-79875-x.
818 16. Wang K, Lu W, Tu Q, Ge Y, He J, Zhou Y, et al. Preliminary analysis of salivary
819 microbiome and their potential roles in oral lichen planus. Sci Rep. 2016;6(1):22943.
820 https://doi.org/10.1038/srep22943.
821 17. Sabharwal A, Ganley K, Miecznikowski JC, Haase EM, Barnes V, Scannapieco FA.
822 The salivary microbiome of diabetic and non-diabetic adults with periodontal disease. J
823 Periodontol. 2019;90(1):26-34. https://doi.org/https://doi.org/10.1002/JPER.18-0167.
824 18. Torres PJ, Fletcher EM, Gibbons SM, Bouvet M, Doran KS, Kelley ST.
825 Characterization of the salivary microbiome in patients with pancreatic cancer. PeerJ.
826 2015;3:e1373.
827 19. Lindheim L, Bashir M, Münzker J, Trummer C, Zachhuber V, Pieber TR, et al. The
828 Salivary Microbiome in Polycystic Ovary Syndrome (PCOS) and Its Association with
829 Disease-Related Parameters: A Pilot Study. Front Microbiol. 2016;7(1270).
830 https://doi.org/10.3389/fmicb.2016.01270.
831 20. Nomura Y, Kakuta E, Okada A, Otsuka R, Shimada M, Tomizawa Y, et al. Oral
832 Microbiome in Four Female Centenarians. Applied Sciences. 2020;10(15):5312.
833 21. Hansen TH, Kern T, Bak EG, Kashani A, Allin KH, Nielsen T, et al. Impact of a
834 vegan diet on the human salivary microbiota. Sci Rep. 2018;8(1):5847.
835 https://doi.org/10.1038/s41598-018-24207-3.
836 22. Kato I, Vasquez A, Moyerbrailean G, Land S, Djuric Z, Sun J, et al. Nutritional
837 Correlates of Human Oral Microbiome. J Am Coll Nutr. 2017;36(2):88-98.
838 https://doi.org/10.1080/07315724.2016.1185386.
839 23. Liu K, Chen S, Huang J, Ren F, Yang T, Long D, et al. Oral Microbiota of Children Is
840 Conserved across Han, Tibetan and Hui Groups and Is Correlated with Diet and Gut
841 Microbiota. Microorganisms. 2021;9(5):1030.
842 24. Renson A, Jones HE, Beghini F, Segata N, Zolnik CP, Usyk M, et al.
843 Sociodemographic variation in the oral microbiome. Ann Epidemiol. 2019;35:73-80.e2.
844 https://doi.org/https://doi.org/10.1016/j.annepidem.2019.03.006.
845 25. Wu J, Peters BA, Dominianni C, Zhang Y, Pei Z, Yang L, et al. Cigarette smoking
846 and the oral microbiome in a large study of American adults. The ISME Journal.
847 2016;10(10):2435-46. https://doi.org/10.1038/ismej.2016.37.
848 26. Fan X, Peters BA, Jacobs EJ, Gapstur SM, Purdue MP, Freedman ND, et al. Drinking
849 alcohol is associated with variation in the human oral microbiome in a large study of
850 American adults. Microbiome. 2018;6(1):59. https://doi.org/10.1186/s40168-018-0448-x.
851 27. Takayasu L, Suda W, Takanashi K, Iioka E, Kurokawa R, Shindo C, et al. Circadian
852 oscillations of microbial and functional composition in the human salivary microbiome. DNA
853 Res. 2017;24(3):261-70. https://doi.org/10.1093/dnares/dsx001.
854 28. Wu Y, Chi X, Zhang Q, Chen F, Deng X. Characterization of the salivary microbiome
855 in people with obesity. PeerJ. 2018;6:e4458.
856 29. Jo R, Nishimoto Y, Umezawa K, Yama K, Aita Y, Ichiba Y, et al. Comparison of oral
857 microbiome profiles in stimulated and unstimulated saliva, tongue, and mouth-rinsed water.
858 Sci Rep. 2019;9(1):16124. https://doi.org/10.1038/s41598-019-52445-6.
859 30. Cattaneo C, Riso P, Laureati M, Gargari G, Pagliarini E. Exploring Associations
860 between Interindividual Differences in Taste Perception, Oral Microbiota Composition, and
861 Reported Food Intake. Nutrients. 2019;11(5):1167.

31. Gardner A, So PW, Carpenter GH. Intraoral Microbial Metabolism and Association with Host Taste Perception. J Dent Res. 2020;99(6):739-45. https://doi.org/10.1177/0022034520917142.
32. Henne K, Li J, Stoneking M, Kessler O, Schilling H, Sonanini A, et al. Global analysis of saliva as a source of bacterial genes for insights into human population structure and migration studies. BMC Evol Biol. 2014;14(1):190. https://doi.org/10.1186/s12862-014-0190-3.
33. Li J, Quinque D, Horz H-P, Li M, Rzhetskaya M, Raff JA, et al. Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. BMC Microbiol. 2014;14(1):316. https://doi.org/10.1186/s12866-014-0316-1.
34. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852-7. https://doi.org/10.1038/s41587-019-0209-9.
35. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database. 2010;2010. https://doi.org/10.1093/database/baq013.
36. Romano S, Savva GM, Bedarf JR, Charles IG, Hildebrand F, Narbad A. Meta-analysis of the Parkinson's disease gut microbiome suggests alterations linked to intestinal inflammation. npj Parkinson's Disease. 2021;7(1):27. https://doi.org/10.1038/s41531-021-00156-z.
37. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics Intellig Lab Syst. 2015;143:85-96. https://doi.org/https://doi.org/10.1016/j.chemolab.2015.02.019.
38. Oksanen J, Guillaume; BF, Michael; F, Roeland; K, Pierre; L, Dan; M, et al. vegan: Community Ecology Package. R package version 2.5-7.; 2020.
39. Shade A, Stopnisek N. Abundance-occupancy distributions to prioritize plant core microbiome membership. Curr Opin Microbiol. 2019;49:50-8. https://doi.org/https://doi.org/10.1016/j.mib.2019.09.008.
40. Frank E Harrell Jr wcfCDamo. Hmisc: Harrell Miscellaneous. R package version 4.5-0. 2021. https://CRAN.R-project.org/package=Hmisc.
41. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504.
42. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.
43. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016. https://ggplot2.tidyverse.org.
44. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nature Communications. 2020;11(1):3514. https://doi.org/10.1038/s41467-020-17041-7.
45. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2013;42(D1):D199-D205. https://doi.org/10.1093/nar/gkt1076.
46. Wei T, Simko V. R package 'corrplot': Visualization of a Correlation Matrix (Version 0.90). 2021. https://github.com/taiyun/corrplot.
47. Marotz C, Amir A, Humphrey G, Gaffney J, Gogul G, Knight R. DNA extraction for streamlined metagenomics of diverse environmental samples. BioTechniques. 2017;62(6):290-3. https://doi.org/10.2144/000114559.

48. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. 2016:081257. https://doi.org/10.1101/081257.
49. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-D6. https://doi.org/10.1093/nar/gks1219.
50. Stopnisek N, Shade A. Persistent microbiome members in the common bean rhizosphere: an integrated analysis of space, time, and plant genotype. The ISME Journal. 2021;15(9):2708-22. https://doi.org/10.1038/s41396-021-00955-5.
51. Domin H, Zurita-Gutiérrez YH, Scotti M, Buttlar J, Hentschel Humeida U, Fraune S. Predicted Bacterial Interactions Affect in Vivo Microbial Colonization Dynamics in Nematostella. Front Microbiol. 2018;9(728). https://doi.org/10.3389/fmicb.2018.00728.
52. Fan P, Nelson CD, Driver JD, Elzo MA, Peñagaricano F, Jeong KC. Host genetics exerts lifelong effects upon hindgut microbiota and its association with bovine growth and immunity. The ISME Journal. 2021;15(8):2306-21. https://doi.org/10.1038/s41396-021-00925-x.
53. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, et al. A global atlas of the dominant bacteria found in soil. Science. 2018;359(6373):320-5. https://doi.org/doi:10.1126/science.aap9516.
54. Wang Y, Gong J, Li J, Xin Y, Hao Z, Chen C, et al. Insights into bacterial diversity in compost: Core microbiome and prevalence of potential pathogenic bacteria. Sci Total Environ. 2020;718:137304. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.137304.
55. Giordano C, Boscaro V, Munz G, Mori G, Vannini C. Summer holidays as break-point in shaping a tannery sludge microbial community around a stable core microbiota. Sci Rep. 2016;6(1):30376. https://doi.org/10.1038/srep30376.
56. Salonen A, Salojärvi J, Lahti L, de Vos WM. The adult intestinal core microbiota is determined by analysis depth and health status. Clinical Microbiology and Infection. 2012;18(s4):16-20. https://doi.org/https://doi.org/10.1111/j.1469-0691.2012.03855.x.
57. Hall MW, Singh N, Ng KF, Lam DK, Goldberg MB, Tenenbaum HC, et al. Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. npj Biofilms and Microbiomes. 2017;3(1):2. https://doi.org/10.1038/s41522-016-0011-0.
58. Simón-Soro Á, Tomás I, Cabrera-Rubio R, Catalan MD, Nyvad B, Mira A. Microbial Geography of the Oral Cavity. J Dent Res. 2013;92(7):616-21. https://doi.org/10.1177/0022034513488119.
59. Yao T, Han X, Guan T, Zhai C, Liu C, Liu C, et al. Exploration of the microbiome community for saliva, skin, and a mixture of both from a population living in Guangdong. Int J Legal Med. 2021;135(1):53-62. https://doi.org/10.1007/s00414-020-02329-6.
60. Oliveira SG, Nishiyama RR, Trigo CAC, Mattos-Guaraldi AL, Dávila AMR, Jardim R, et al. Core of the saliva microbiome: an analysis of the MG-RAST data. BMC Oral Health. 2021;21(1):351. https://doi.org/10.1186/s12903-021-01719-5.
61. Conrads G, Westenberger J, Lürkens M, Abdelbary MMH. Isolation and Bacteriocin-Related Typing of Streptococcus dentisani. Frontiers in Cellular and Infection Microbiology. 2019;9(110). https://doi.org/10.3389/fcimb.2019.00110.
62. López-López A, Camelo-Castillo A, Ferrer MD, Simon-Soro Á, Mira A. Health-Associated Niche Inhabitants as Oral Probiotics: The Case of Streptococcus dentisani. Front Microbiol. 2017;8(379). https://doi.org/10.3389/fmicb.2017.00379.
63. Zarco M, Vess T, Ginsburg G. The oral microbiome in health and disease and the potential impact on personalized dental medicine. Oral Dis. 2012;18(2):109-20. https://doi.org/https://doi.org/10.1111/j.1601-0825.2011.01851.x.

960 64. Cornejo-Granados F, Gallardo-Becerra L, Leonardo-Reza M, Ochoa-Romo JP,
961 Ochoa-Leyva A. A meta-analysis reveals the environmental and host factors shaping the
962 structure and function of the shrimp microbiota. PeerJ. 2018;6:e5382.
963 65. Holman DB, Brunelle BW, Trachsel J, Allen HK, Bik H. Meta-analysis To Define a
964 Core Microbiota in the Swine Gut. mSystems. 2017;2(3):e00004-17.
965 https://doi.org/doi:10.1128/mSystems.00004-17.
966 66. Wright RJ, Langille MGI, Walker TR. Food or just a free ride? A meta-analysis
967 reveals the global diversity of the Plastisphere. ISME J. 2021;15(3):789-806.
968 https://doi.org/10.1038/s41396-020-00814-9.
969 67. Roberts FA, Darveau RP. Microbial protection and virulence in periodontal tissue as a
970 function of polymicrobial communities: symbiosis and dysbiosis. Periodontol 2000.
971 2015;69(1):18-27. https://doi.org/https://doi.org/10.1111/prd.12087.
972 68. Sato N, Kakuta M, Hasegawa T, Yamaguchi R, Uchino E, Kobayashi W, et al.
973 Metagenomic analysis of bacterial species in tongue microbiome of current and never
974 smokers. npj Biofilms and Microbiomes. 2020;6(1):11. https://doi.org/10.1038/s41522-020-
975 0121-6.
976 69. Vogtmann E, Flores R, Yu G, Freedman ND, Shi J, Gail MH, et al. Association
977 between tobacco use and the upper gastrointestinal microbiome among Chinese men. Cancer
978 Causes Control. 2015;26(4):581-8. https://doi.org/10.1007/s10552-015-0535-2.
979 70. Karabudak S, Ari O, Durmaz B, Dal T, Basyigit T, Kalcioglu MT, et al. Analysis of
980 the effect of smoking on the buccal microbiome using next-generation sequencing
981 technology. J Med Microbiol. 2019;68(8):1148-58.
982 https://doi.org/https://doi.org/10.1099/jmm.0.001003.
983 71. Prasoodanan P. K V, Sharma AK, Mahajan S, Dhakan DB, Maji A, Scaria J, et al.
984 Western and non-western gut microbiomes reveal new roles of Prevotella in carbohydrate
985 metabolism and mouth–gut axis. npj Biofilms and Microbiomes. 2021;7(1):77.
986 https://doi.org/10.1038/s41522-021-00248-x.
987 72. García-Estévez I, Ramos-Pineda AM, Escribano-Bailón MT. Interactions between
988 wine phenolic compounds and human saliva in astringency perception. Food Funct.
989 2018;9(3):1294-309.
990 73. Liu Z, Guo H, Zhang W, Ni L. Salivary Microbiota Shifts under Sustained
991 Consumption of Oolong Tea in Healthy Adults. Nutrients. 2020;12(4):966.
992 74. Vangay P, Johnson AJ, Ward TL, Al-Ghalith GA, Shields-Cutler RR, Hillmann BM,
993 et al. US Immigration Westernizes the Human Gut Microbiome. Cell. 2018;175(4):962-
994 72.e10. https://doi.org/https://doi.org/10.1016/j.cell.2018.10.029.
995 75. Peters BA, Yi SS, Beasley JM, Cobbs EN, Choi HS, Beggs DB, et al. US nativity and
996 dietary acculturation impact the gut microbiome in a diverse US population. The ISME
997 Journal. 2020;14(7):1639-50. https://doi.org/10.1038/s41396-020-0630-6.
998 76. Acharya A, Chan Y, Kheur S, Kheur M, Gopalakrishnan D, Watt R, et al. Salivary
999 microbiome of an urban Indian cohort and patterns linked to subclinical inflammation. Oral
1000 diseases. 2017;23(7):926-40.
1001 77. Anderson A, Al-Ahmad A, Schlueter N, Frese C, Hellwig E, Binder N. Influence of
1002 the long-term use of oral hygiene products containing stannous ions on the salivary
1003 microbiome–a randomized controlled trial. Scientific reports. 2020;10(1):1-8.
1004 78. Bhushan B, Yadav A, Singh S, Ganju L. Diversity and functional analysis of salivary
1005 microflora of Indian Antarctic expeditionaries. Journal of oral microbiology.
1006 2019;11(1):1581513.
1007 79. Cabral DJ, Wurster JI, Flokas ME, Alevizakos M, Zabat M, Korry BJ, et al. The
1008 salivary microbiome is consistent between subjects and resistant to impacts of short-term

1009 hospitalization. Scientific Reports. 2017;7(1):11040. https://doi.org/10.1038/s41598-017-
1010 11427-2.
1011 80. Cameron SJS, Huws SA, Hegarty MJ, Smith DPM, Mur LAJ. The human salivary
1012 microbiome exhibits temporal stability in bacterial diversity. FEMS Microbiology Ecology.
1013 2015;91(9). https://doi.org/10.1093/femsec/fiv091.
1014 81. Chaudhari DS, Dhotre DP, Agarwal DM, Gaike AH, Bhalerao D, Jadhav P, et al. Gut,
1015 oral and skin microbiome of Indian patrilineal families reveal perceptible association with
1016 age. Scientific Reports. 2020;10(1):5685. https://doi.org/10.1038/s41598-020-62195-5.
1017 82. Chen C, Hemme C, Beleno J, Shi ZJ, Ning D, Qin Y, et al. Oral microbiota of
1018 periodontal health and disease and their changes after nonsurgical periodontal therapy. The
1019 ISME Journal. 2018;12(5):1210-24. https://doi.org/10.1038/s41396-017-0037-1.
1020 83. Collado MC, Engen PA, Bandín C, Cabrera-Rubio R, Voigt RM, Green SJ, et al.
1021 Timing of food intake impacts daily rhythms of human salivary microbiota: a randomized,
1022 crossover study. The FASEB Journal. 2018;32(4):2060-72.
1023 84. Debelius JW, Huang T, Cai Y, Ploner A, Barrett D, Zhou X, et al. Subspecies niche
1024 specialization in the oral microbiome is associated with nasopharyngeal carcinoma risk.
1025 Msystems. 2020;5(4):e00065-20.
1026 85. Hijazi K, Morrison RW, Mukhopadhya I, Martin B, Gemmell M, Shaw S, et al. Oral
1027 bacterial diversity is inversely correlated with mucosal inflammation. Oral Diseases.
1028 2020;26(7):1566-75. https://doi.org/https://doi.org/10.1111/odi.13420.
1029 86. Ji Y, Liang X, Lu H. Analysis of by high-throughput sequencing: Helicobacter pylori
1030 infection and salivary microbiome. BMC Oral Health. 2020;20(1):84.
1031 https://doi.org/10.1186/s12903-020-01070-1.
1032 87. Jiang Q, Liu J, Chen L, Gan N, Yang D. The Oral Microbiome in the Elderly With
1033 Dental Caries and Health. Frontiers in cellular and infection microbiology. 2019;8(442).
1034 https://doi.org/10.3389/fcimb.2018.00442.
1035 88. Kumpitsch C, Moissl-Eichinger C, Pock J, Thurnher D, Wolf A. Preliminary insights
1036 into the impact of primary radiochemotherapy on the salivary microbiome in head and neck
1037 squamous cell carcinoma. Scientific reports. 2020;10(1):1-12.
1038 89. Lee W-H, Chen H-M, Yang S-F, Liang C, Peng C-Y, Lin F-M, et al. Bacterial
1039 alterations in salivary microbiota and their association in oral cancer. Scientific reports.
1040 2017;7(1):1-11.
1041 90. Lin M, Li X, Wang J, Cheng C, Zhang T, Han X, et al. Saliva microbiome changes in
1042 patients with periodontitis with and without chronic obstructive pulmonary disease. Frontiers
1043 in cellular infection microbiology. 2020;10:124.
1044 91. Lindheim L, Bashir M, Münzker J, Trummer C, Zachhuber V, Pieber TR, et al. The
1045 salivary microbiome in polycystic ovary syndrome (pcos) and its association with disease-
1046 related parameters: a pilot study. Frontiers in microbiology. 2016;7:1270.
1047 92. Liu Y, Zhang Q, Hu X, Chen F, Hua H. Characteristics of the salivary microbiota in
1048 cheilitis granulomatosa. Med Oral Patol Oral Cir Bucal. 2019;24(6):e719-e25.
1049 https://doi.org/10.4317/medoral.23041.
1050 93. Lundmark A, Hu YO, Huss M, Johannsen G, Andersson AF, Yucel-Lindberg T.
1051 Identification of salivary microbiota and its association with host inflammatory mediators in
1052 periodontitis. Frontiers in cellular infection microbiology. 2019;9:216.
1053 94. Menon R, Gomez A, Brandt B, Leung Y, Gopinath D, Watt R, et al. Long-term
1054 impact of oral surgery with or without amoxicillin on the oral microbiome-A prospective
1055 cohort study. Scientific reports. 2019;9(1):1-10.
1056 95. Niu C, Dong T, Jiang W, Gao L, Yuan K, Hu X, et al. Pregnancy-Related Ecological
1057 Shifts of Salivary Microbiota and its Association with Salivary Sex Hormones. 2020.

1058 96. Ozga AT, Sankaranarayanan K, Tito RY, Obregon-Tito AJ, Foster MW, Tallbull G, et
1059 al. Oral microbiome diversity among Cheyenne and Arapaho individuals from Oklahoma.
1060 American journal of physical anthropology. 2016;161(2):321-7.
1061 97. Pushalkar S, Paul B, Li Q, Yang J, Vasconcelos R, Makwana S, et al. Electronic
1062 Cigarette Aerosol Modulates the Oral Microbiome and Increases Risk of Infection. iScience.
1063 2020;23(3):100884. https://doi.org/https://doi.org/10.1016/j.isci.2020.100884.
1064 98. Relvas M, Regueira-Iglesias A, Balsa-Castro C, Salazar F, Pacheco JJ, Cabral C, et al.
1065 Assessing the impact of dental and periodontal statuses on the salivary microbiome: a global
1066 oral health scale. medRxiv. 2020.
1067 99. Russo E, Bacci G, Chiellini C, Fagorzi C, Niccolai E, Taddei A, et al. Preliminary
1068 comparison of oral and intestinal human microbiota in patients with colorectal cancer: a pilot
1069 study. Frontiers in microbiology. 2018;8:2699.
1070 100. Sierra MA, Li Q, Pushalkar S, Paul B, Sandoval TA, Kamer AR, et al. The influences
1071 of bioinformatics tools and reference databases in analyzing the human oral microbial
1072 community. Genes. 2020;11(8):878.
1073 101. Sohail MU, Elrayess MA, Al Thani AA, Al-Asmakh M, Yassine HM. Profiling the
1074 oral microbiome and plasma biochemistry of obese hyperglycemic subjects in Qatar.
1075 Microorganisms. 2019;7(12):645.
1076 102. Soriano-Lerma A, Pérez-Carrasco V, Sánchez-Marañón M, Ortiz-González M,
1077 Sánchez-Martín V, Gijón J, et al. Influence of 16S rRNA target region on the outcome of
1078 microbiome studies in soil and saliva samples. Scientific reports. 2020;10(1):1-13.
1079 103. Stewart CJ, Auchtung TA, Ajami NJ, Velasquez K, Smith DP, De La Garza II R, et
1080 al. Effects of tobacco smoke and electronic cigarette vapor exposure on the oral and gut
1081 microbiota in humans: a pilot study. PeerJ. 2018;6:e4693.
1082 104. Sun X, Li M, Xia L, Fang Z, Yu S, Gao J, et al. Alteration of salivary microbiome in
1083 periodontitis with or without type-2 diabetes mellitus and metformin treatment. Scientific
1084 reports. 2020;10(1):1-14.
1085 105. Tian N, Faller L, Leffler DA, Kelly CP, Hansen J, Bosch JA, et al. Salivary gluten
1086 degradation and oral microbial profiles in healthy individuals and celiac disease patients.
1087 Applied and environmental microbiology. 2017;83(6):e03330-16.
1088 106. Tong Y, Zheng L, Qing P, Zhao H, Li Y, Su L, et al. Oral microbiota perturbations
1089 are linked to high risk for rheumatoid arthritis. Frontiers in cellular infection microbiology.
1090 2020;9:475.
1091 107. Urbaniak C, Lorenzi H, Thissen J, Jaing C, Crucian B, Sams C, et al. The influence of
1092 spaceflight on the astronaut salivary microbiome and the search for a microbiome biomarker
1093 for viral reactivation. Microbiome. 2020;8(1):1-14.
1094 108. Vesty A, Gear K, Biswas K, Radcliff FJ, Taylor MW, Douglas RG. Microbial and
1095 inflammatory-based salivary biomarkers of head and neck squamous cell carcinoma. Clinical
1096 and experimental dental research. 2018;4(6):255-62.
1097 https://doi.org/https://doi.org/10.1002/cre2.139.
1098 109. Vogtmann E, Chen J, Kibriya MG, Amir A, Shi J, Chen Y, et al. Comparison of oral
1099 collection methods for studies of microbiota. Cancer Epidemiology and Prevention
1100 Biomarkers. 2019;28(1):137-43.
1101 110. Wang T, Yu L, Xu C, Pan K, Mo M, Duan M, et al. Chronic fatigue syndrome
1102 patients have alterations in their oral microbiome composition and function. PloS one.
1103 2018;13(9):e0203503.
1104 111. Wang Q, Rao Y, Guo X, Liu N, Liu S, Wen P, et al. Oral microbiome in patients with
1105 oesophageal squamous cell carcinoma. Scientific reports. 2019;9(1):1-9.

1106 112. Wang X, Zhao Z, Tang N, Zhao Y, Xu J, Li L, et al. Microbial community analysis of
1107 saliva and biopsies in patients with oral lichen planus. Frontiers in microbiology.
1108 2020;11:629.
1109 113. Wolf A, Moissl-Eichinger C, Perras A, Koskinen K, Tomazic PV, Thurnher D. The
1110 salivary microbiome as an indicator of carcinogenesis in patients with oropharyngeal
1111 squamous cell carcinoma: A pilot study. Scientific reports. 2017;7(1):1-10.
1112 114. Xun Z, Zhang Q, Xu T, Chen N, Chen F. Dysbiosis and Ecotypes of the Salivary
1113 Microbiome Associated With Inflammatory Bowel Diseases and the Assistance in Diagnosis
1114 of Diseases Using Oral Bacterial Profiles. Frontiers in microbiology. 2018;9(1136).
1115 https://doi.org/10.3389/fmicb.2018.01136.
1116 115. Yano Y, Hua X, Wan Y, Suman S, Zhu B, Dagnall CL, et al. Comparison of Oral
1117 Microbiota Collected Using Multiple Methods and Recommendations for New
1118 Epidemiologic Studies. Msystems. 2020;5(4):e00156-20.
1119 116. Yeo L-F, Aghakhanian FF, Tan JS, Gan HM, Phipps ME. Health and saliva
1120 microbiomes of a semi-urbanized indigenous tribe in Peninsular Malaysia. F1000Research.
1121 2019;8.
1122 117. Yu FY, Wang QQ, Li M, Cheng Y-H, Cheng Y-SL, Zhou Y, et al. Dysbiosis of saliva
1123 microbiome in patients with oral lichen planus. BMC microbiology. 2020;20(1):1-12.
1124 118. Zhu C, Yuan C, Wei F-Q, Sun X-Y, Zheng S-G. Comparative evaluation of
1125 peptidome and microbiota in different types of saliva samples. Ann Transl Med.
1126 2020;8(11):686-. https://doi.org/10.21037/atm-20-393.
1127 119. Zhu C, Yuan C, Wei FQ, Sun XY, Zheng SG. Intraindividual Variation and Personal
1128 Specificity of Salivary Microbiota. J Dent Res. 2020;99(9):1062-71.
1129 https://doi.org/10.1177/0022034520917155.
1130 120. Ziganshina EE, Sagitov II, Akhmetova RF, Saleeva GT, Kiassov AP, Gogoleva NE,
1131 et al. Comparison of the Microbiota and Inorganic Anion Content in the Saliva of Patients
1132 with Gastroesophageal Reflux Disease and Gastroesophageal Reflux Disease-Free
1133 Individuals. BioMed Research International. 2020;2020:2681791.
1134 https://doi.org/10.1155/2020/2681791.
1135

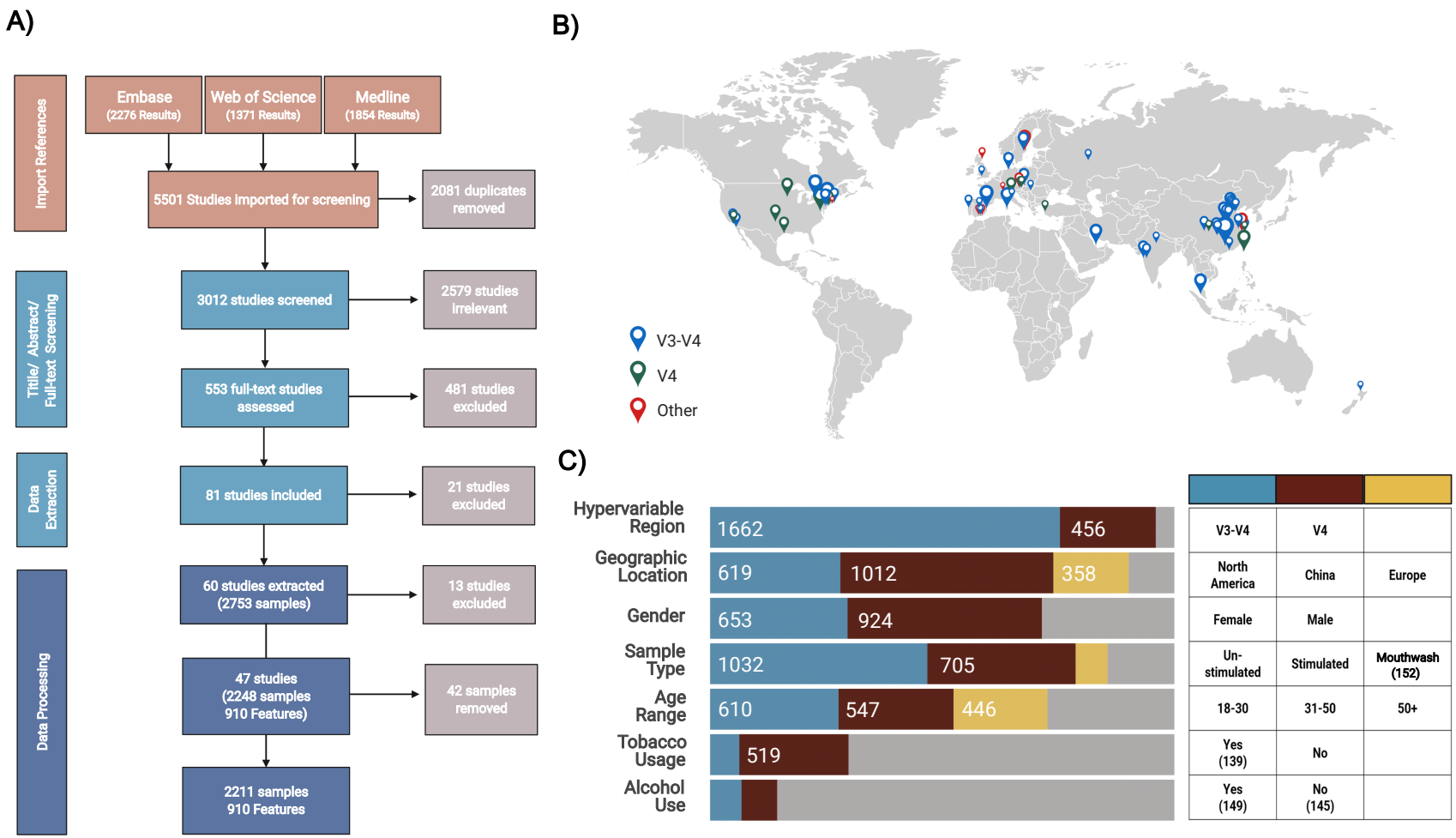

Figure 1. Overview of literature search procedure and metadata of included studies. a) Large-scale literature searching and data filtering process, followed by the number of samples submitted to the bioinformatic analyses; b) The locations of studies, the scale of symbols that reflect the number of samples of each study; c) Distribution of metadata categories.

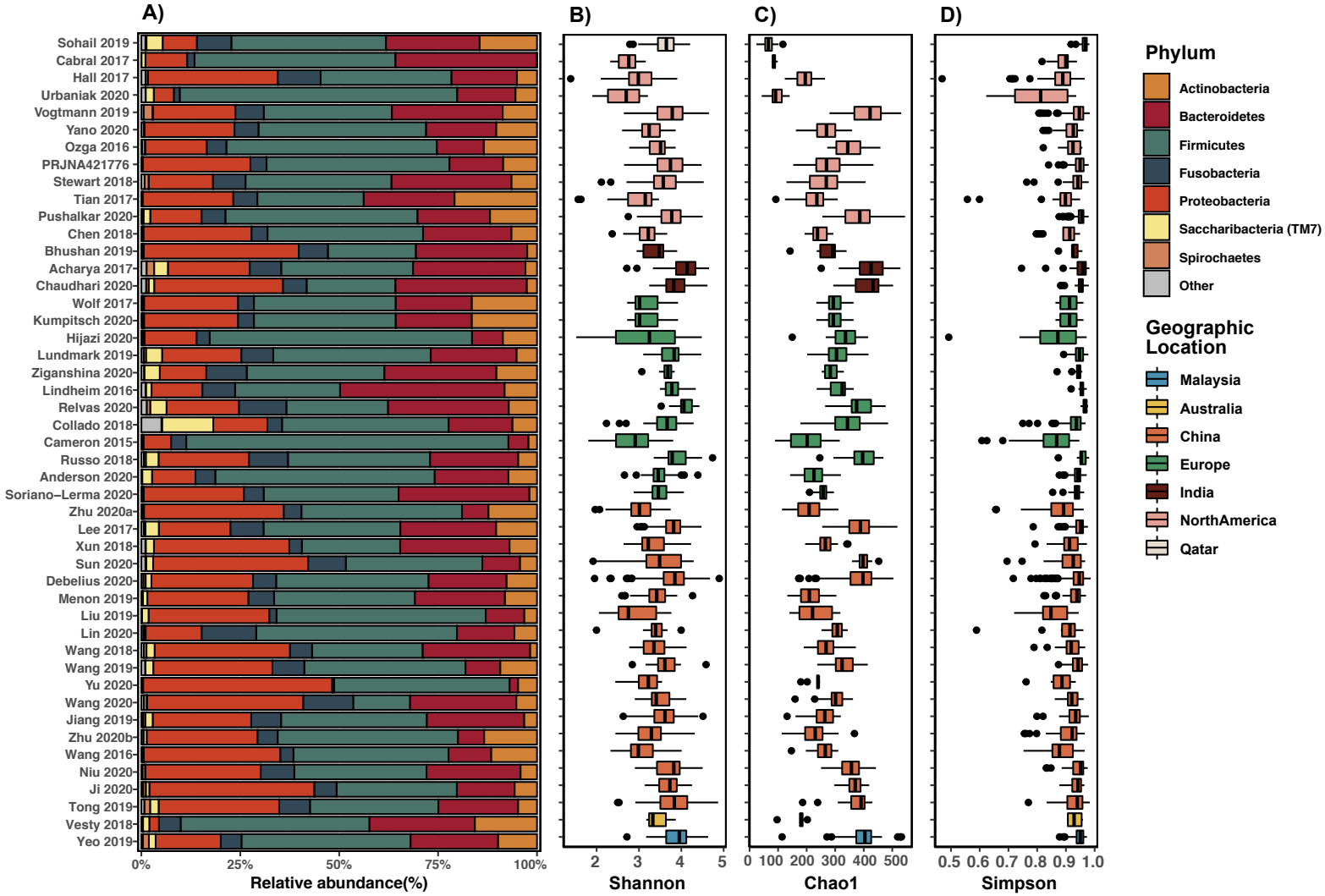

Figure 2. Summary of taxonomic composition and alpha diversity of included studies. A) The mean community composition of each study at the phylum level; The alpha-diversity measured by **B)** Shannon index; **C)** Chao 1 index; **D)** Simpson's index, the colour of boxes stands for the geographic location of the studies. The horizontal bars within boxes represent medians. The tops and bottoms of boxes represent the 75th and 25th percentiles, respectively.

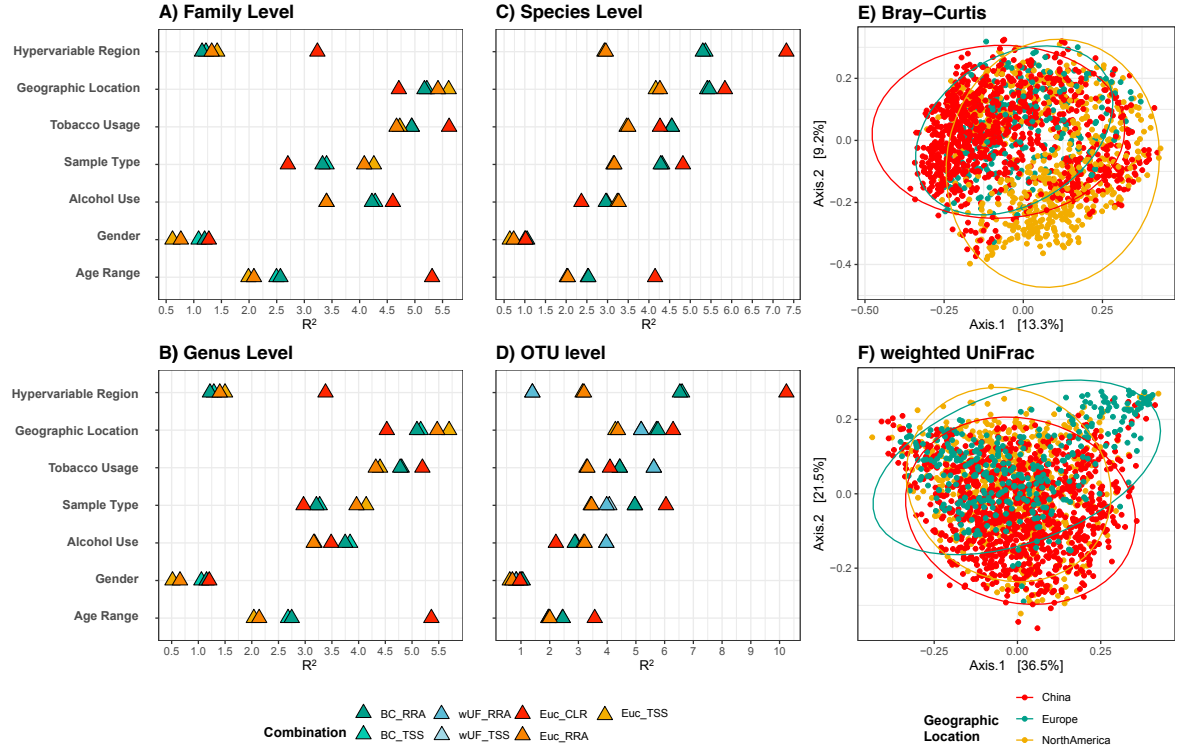

Figure 3. The variability in human salivary microbiota have been explained by different factors. Among them, hypervariable regions and geographic locations have the largest impact. The effect of the categories on the clustering of the sample was measured using PERMANOVA at four taxonomic levels: family **(A),** Genus **(B),** species **(C)** and OTU level **(D).** The colour indicates the different combinations of normalisation (TSS, Total-sum scaling; RRA, Rarefied relative abundance; CLR, Centred log ratio) and indices (BC, Bray-Curtis; EUC, Euclidean; wUF, weighted uniFrac). Because the results of rarefication (RAR) were very close to TSS and RRA, they were not displayed in the figures. Principal coordinate analysis (PCoA) with Bray-Curtis **(E)** and weighted uniFrac **(F)** showing the differences between samples from North America, Europe, and China.

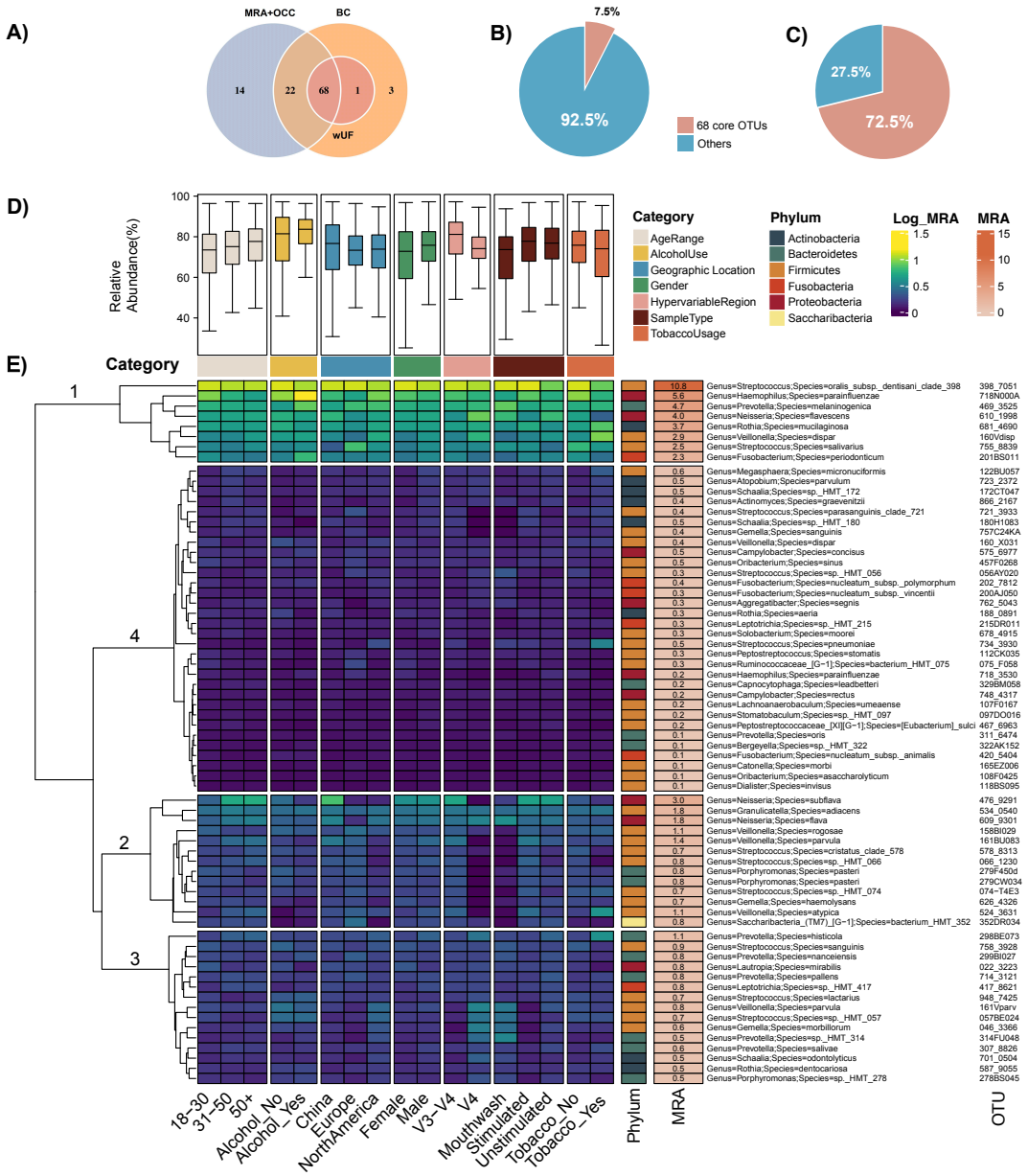

Figure 4. The core OTUs defined by abundance-occupancy pattern. A) Venn diagram showing the interaction between three methods used to define the core. Sixty-eight OTUs were defined as the core for all methods. (MRA+OCC: The thresholds were setting on mean relative abundance and occupancy to define the core; BC: The method adapted from Shade and Stopnisek using Bray-Curtis similarity; wUF: The method adapted from Shade and Stopnisek using weighted uniFrac distance). **B)** Pie chart showing the number of the core (pink) versus other OTUs (blue) identified in percentage. **C)** Pie chart showing the relative abundance of the core and other OTUs across all samples. **D)** Relative abundance of 68 core OTUs across subgroups classified by seven categories. **E)** Heatmap showing the log-transformed mean relative abundance of each core OTU at each level of different categories.

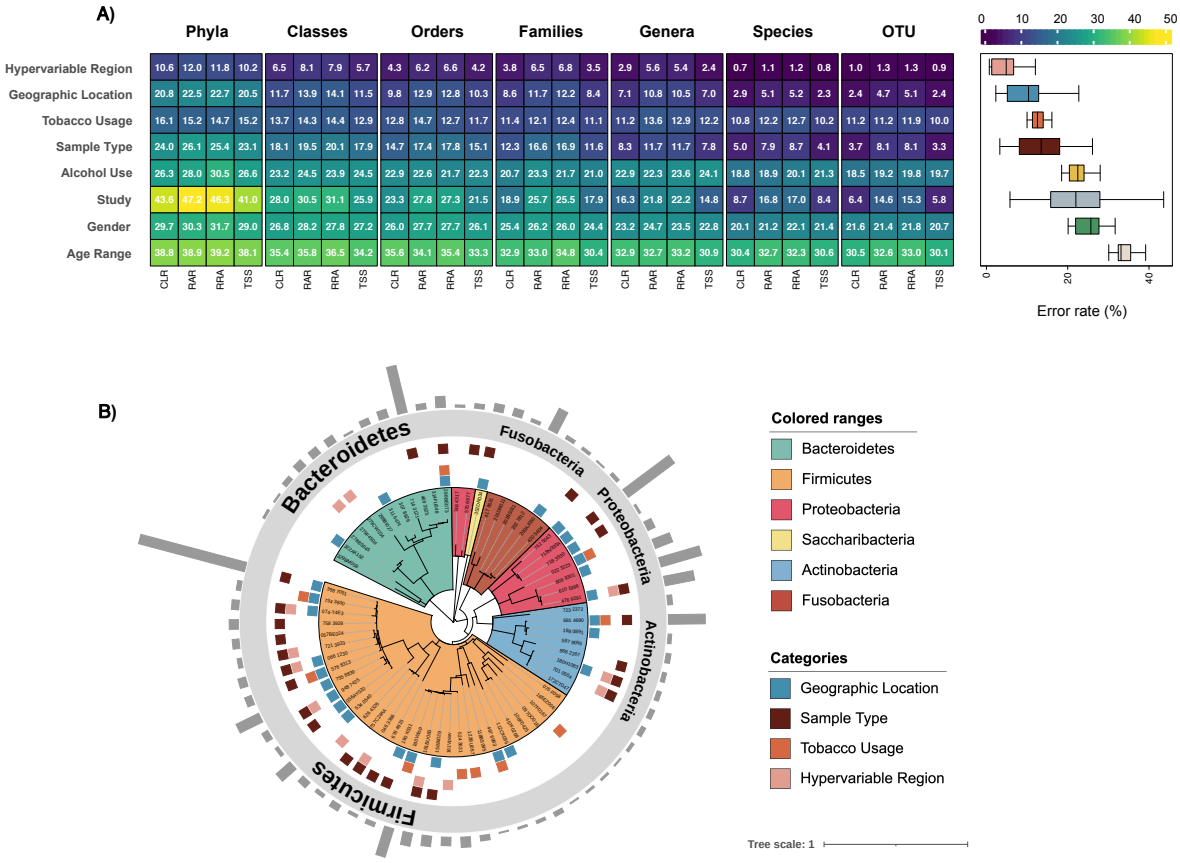

Figure 5. Salivary microbiome members which significantly contribute to categorisation of meta data. Random Forest models showed the impact of categories on salivary microbiome and the core OTUs contributing to accuracy of these models. A) Error rate (%) for the random forest classifications conducted with samples grouped by eight different categories. **B)** Phylogenetic tree indicates the taxonomic information of 68 core OTUs. The coloured squares between the tree and the annotation of phylum indicate the OTUs that were defined by the Random Forest model as "important" for distinguishing between different levels in each category. The bars on the outmost ring showing the mean relative abundance of each OTU.

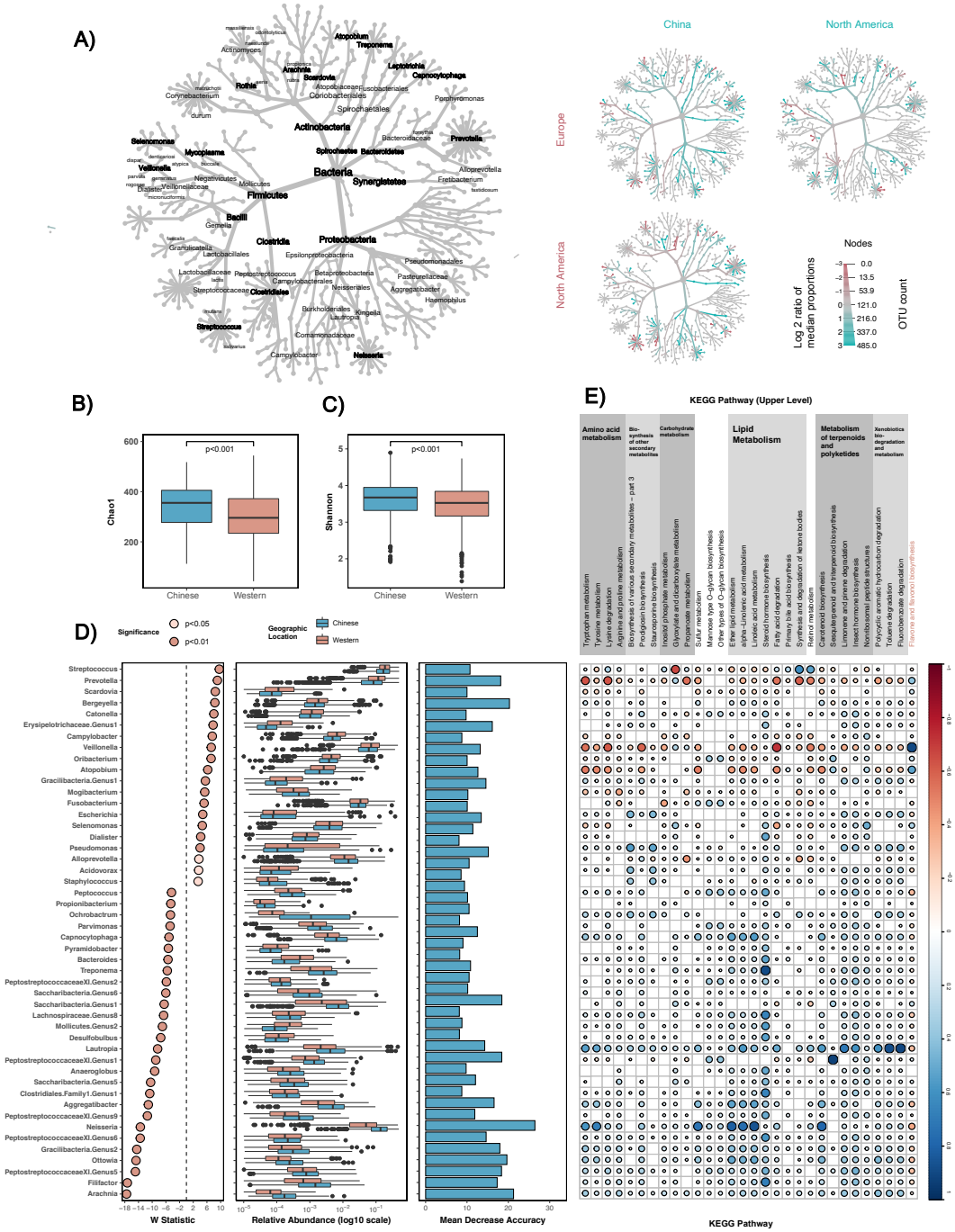

Figure 6. Distinct microbial profiles are evident in the saliva samples from Chinese and Western adults. A) Taxonomic hierarchies show the relative enrichment of taxa in three geographic locations at phylum through species level. Coloured nodes indicate log2-fold increase in median abundance of the group in x-axis (pink) or y-axis (blue). Only taxa showed significant changes (false discovery rate-adjusted Wilcoxon rank sum q < 0.05) are displayed. **B)** and **C)** Comparison of salivary microbial alpha diversity between the Chinese and Western samples, calculated by Chao1 (B: $p < 0.001$, Wilcoxon rank-sum test) and Shannon index (C: $p < 0.001$, Wilcoxon rank-sum test). **D)** Differential abundant genera identified between saliva from Chinese and Western samples. The panel on the left indicates the standardised effect sizes (W statistic) estimated via the difference on relative abundance using ANCOM-BC (taxa enriched in Western samples have a value shifted to right, whereas taxa enriched in Chinese samples have a value shifted to left); The panel in the middle shows the relative abundance of selected genera; the panel on the right indicates the Mean Decrease Accuracy of the random forest model established. **E)** Spearman's correlation coefficients were calculated between each pairwise comparison of differential genus and KEGG pathway. Only significantly correlated comparisons (p<0.01, FDR adjusted Spearman's rank correlation) are displayed. The only Western-enriched pathway is marked in pink.

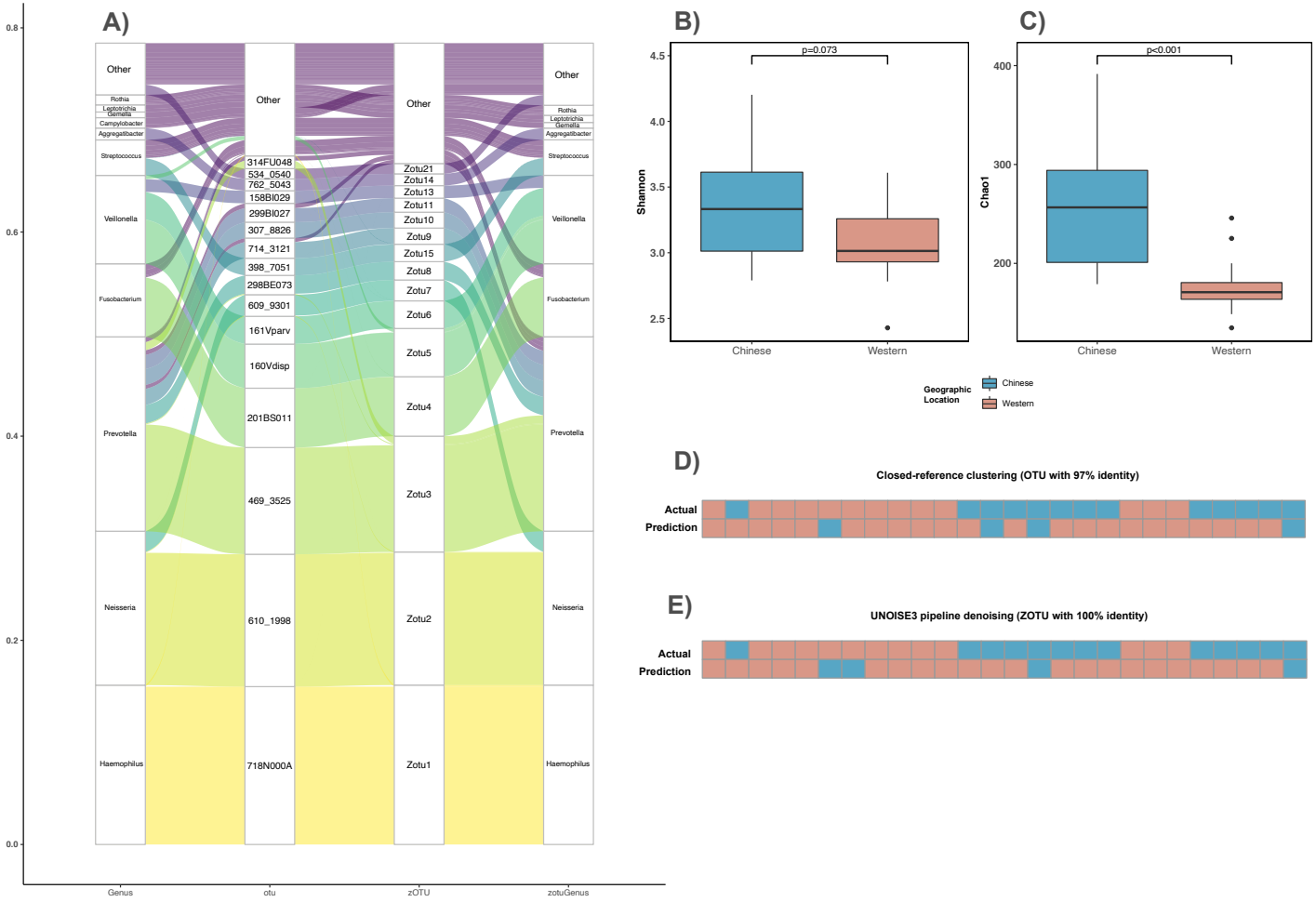

Figure 7. Validation of the findings in the meta-analysis in an independent cohort. A) Alluvial plot showing the affiliation of ZOTUs to their originating core OTUs defined in the meta-analysis. **B)** and **C)** Comparison of salivary microbial alpha diversity between the Chinese and Western samples, calculated by Shannon (B: $p = 0.073$, Wilcoxon rank-sum test) and Chao1 index (C: $p < 0.001$, Wilcoxon rank-sum test). **D)** and **E)** The prediction of the cultural backgrounds of the samples according to the random forest classification model constructed using the genus profiles of samples in the meta-analysis. The genus level profiles of samples processed by **D)** closed-reference clustering with 97% sequence identity and **E)** UNOISE3 denoising with 100% sequence identity were used as the test set.