Genomics Proteomics Bioinformatics 20 (2022) 246–259

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa

Jie Zhu[1,2,#], Liu Tian[1,2,#], Peishan Chen[1,3], Mo Han[1,4], Liju Song[1,2], Xin Tong[1,2], Xiaohuan Sun[1], Fangming Yang[1], Zhipeng Lin[1], Xing Liu[1], Chuan Liu[1], Xiaohan Wang[1], Yuxiang Lin[1], Kaiye Cai[1], Yong Hou[1], Xun Xu[1,2], Huanming Yang[1,5], Jian Wang[1,5], Karsten Kristiansen[1,4], Liang Xiao[1,3,6], Tao Zhang[1,4], Huijue Jia[1,2,*], Zhuye Jie[1,2,4,*]

[1] *BGI-Shenzhen, Shenzhen 518083, China*
[2] *Shenzhen Key Laboratory of Human Commensal Microorganisms and Health Research, BGI-Shenzhen, Shenzhen 518083, China*
[3] *Shenzhen Engineering Laboratory of Detection and Intervention of Human Intestinal Microbiome, BGI-Shenzhen, Shenzhen 518083, China*
[4] *Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Copenhagen DK-2100, Denmark*
[5] *James D. Watson Institute of Genome Sciences, Hangzhou 310058, China*
[6] *BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China*

Received 26 November 2020; revised 13 May 2021; accepted 23 August 2021
Available online 4 September 2021

Handled by Fangqing Zhao

KEYWORDS

Metagenomics;
Human oral microbiome;
Metagenome-assembly genome;
Genome catalog;
Gender

Abstract The oral cavity of each person is home to hundreds of bacterial species. While taxa for oral diseases have been studied using culture-based characterization as well as amplicon sequencing, metagenomic and genomic information remains scarce compared to the fecal microbiome. Here, using metagenomic shotgun data for 3346 oral **metagenomic** samples together with 808 published samples, we obtain 56,213 metagenome-assembled genomes (MAGs), and more than 64% of the 3589 species-level genome bins (SGBs) contain no publicly available genomes. The resulting genome collection is representative of samples around the world and contains many genomes from candidate phyla radiation (CPR) that lack monoculture. Also, it enables the discovery of new taxa such as a genus *Candidatus* Bgiplasma within the family Acholeplasmataceae. Large-scale metagenomic

[*] Corresponding authors.
E-mail: jiezhuye@genomics.cn (Jie Z), jiahuijue@genomics.cn (Jia H).

[#] Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2021.05.001

data from massive samples also allow the assembly of strains from important oral taxa such as *Porphyromonas* and *Neisseria*. The oral microbes encode genes that could potentially metabolize drugs. Apart from these findings, a strongly male-enriched *Campylobacter* species was identified. Oral samples would be more user-friendly collected than fecal samples and have the potential for disease diagnosis. Thus, these data lay down a genomic framework for future inquiries of the **human oral microbiome**.

Introduction

The human microbiome has been implicated in a growing number of diseases. The majority of microbial cells are believed to reside in the large intestine [1] and cohorts with fecal metagenomic data contain over 1000 individuals [2,3].

For the oral microbiome, culture-based characterization as well as marker gene sequencing techniques has been applied in many oral bacteria-associated disease studies, such as cystic fibrosis with *Streptococcus oralis* [4], colorectal cancer (CRC) with *Fusobacterium nucleatum* [5], and even Alzheimer's disease with *Porphyromonas gingivalis* [6]. With the ultra-fast growing culture-independent next generation sequencing technologies, hundreds of metagenomic shotgun-sequenced samples have been available from the Human Microbiome Project (HMP) and for the research of rheumatoid arthritis [1,7,8]. Several other diseases studied by accurate and robust metagenome-wide association studies (MWAS) using gut microbiome data indicated potential contributions from the oral microbiome in disease etiology [9–14]. Although the MWAS on rheumatoid arthritis was based on a *de novo* assembled reference gene catalog for the oral microbiome [8], analyses on genomes would be more desirable. For diagnosis of some diseases, gathering oral samples shows advantages like higher sensitivity and accuracy, better prognosis, and better patient satisfaction than fecal samples. Moreover, collecting oral samples is more convenient as it is easy to operate and could be taken at a fully controlled setting witnessed by trained professionals. Unlike the anaerobic environment for the gut microbiome, the oral microbiome is believed to be well covered by culturing [15], and the analyses by 16S rRNA gene sequencing or PCR are common.

Recent published large-scale metagenomic assembly efforts mostly included fecal metagenomic data [16–18]. It is not clear how much is missing for the oral microbiome. The most well-characterized oral microbial database, the expanded Human Oral Microbiome Database (eHOMD), has about 750 oral taxa; however, only 57% of the oral bacterial species have been officially named, 13% have been cultivated yet remain unnamed, and 30% are uncultivated [19]. A representative genome catalog for genome-resolved analysis is strongly required.

In this study, we present 3346 new oral metagenomic samples. A total of 56,213 medium- and high-quality metagenome-assembled genomes (MAGs) are constructed based on our collection together with previously published 808 samples. The 56,213 MAGs together with 190,309 public genomes were clustered into 3589 oral species-level clades. New taxa as well as the substantially complemented genomic content of known species are revealed. We provide a genome reference that is highly representative of metagenomic samples not used in assembly and could facilitate culturing and functional screens, as well as disease diagnosis and modulation based on the oral microbiome.

Results

Draft genomes assembled from oral metagenomic data

We performed shotgun sequencing on 2284 saliva and 391 tongue dorsum samples from the 4D-SZ cohort (trans-omics, with more time points in future studies, based in China) [2,11,20], and 671 saliva samples from five ethnic groups from Yunnan Province, China. Over 43.19 terabytes of sequence data were generated. Together with 808 published oral samples from 5 studies [8,21–24] that have not been used in a recent large-scale assembly study [18], a total of 4154 oral samples with metagenomic data were obtained (Table S1). A single sample assembly, single sample binning strategy was used. In brief, the metagenomic shotgun reads were assembled into contigs using metaSPAdes [25] independently. Contigs longer than 1500 bp (average 14,094 contigs per sample) were binned by MetaBAT2 [26] in each sample, leading to 163,718 MAGs. After quality control by CheckM [27], 56,213 MAGs which agreed the medium-quality standards [28] (> 50% completeness, < 10% contamination) were retained for further analysis (**Figure 1**A; Table S2). Of these, 15,013 MAGs reached the standards for high-quality (> 90% completeness, < 5% contamination). Moreover, 73.25% of high-quality MAGs and 26.57% of medium-quality MAGs had at least 18 types of standard 20 amino acids decoded by tRNAs (Figure S1A). All three (23S, 16S, and 5S) rRNA genes were present in 27.28% of high-quality MAGs and 13.13% of medium-quality MAGs (Figure S1B).

New genomes from the new samples

To assess the novelty of our assembled genomes, we comprehensively incorporated 190,309 existing isolate and metagenome-assembled genomes from NCBI RefSeq, eHOMD [29], and recent publications [16–18,30,31] (Figure 1A; Table S1) with reconstructed MAGs in this study. A total of 246,522 genomes were obtained.

Species-level genome bins (SGBs) were computed for the 246,522 genomes following multiple steps (Figure 1A, see Methods and methods for details), defined as at least 95% average nucleotide identity (ANI) and at least 30% overlap of the aligned genomes. The clustering analysis resulted in 31,525 SGBs. Among them, 3589 SGBs with oral reference genomes from eHOMD or oral MAGs were defined as oral SGBs (Figure 1B). In the oral SGBs, 2313 clusters (64% of the total oral species) only contained MAGs from this study (denoted uSGBs for unknown SGBs), some of which were

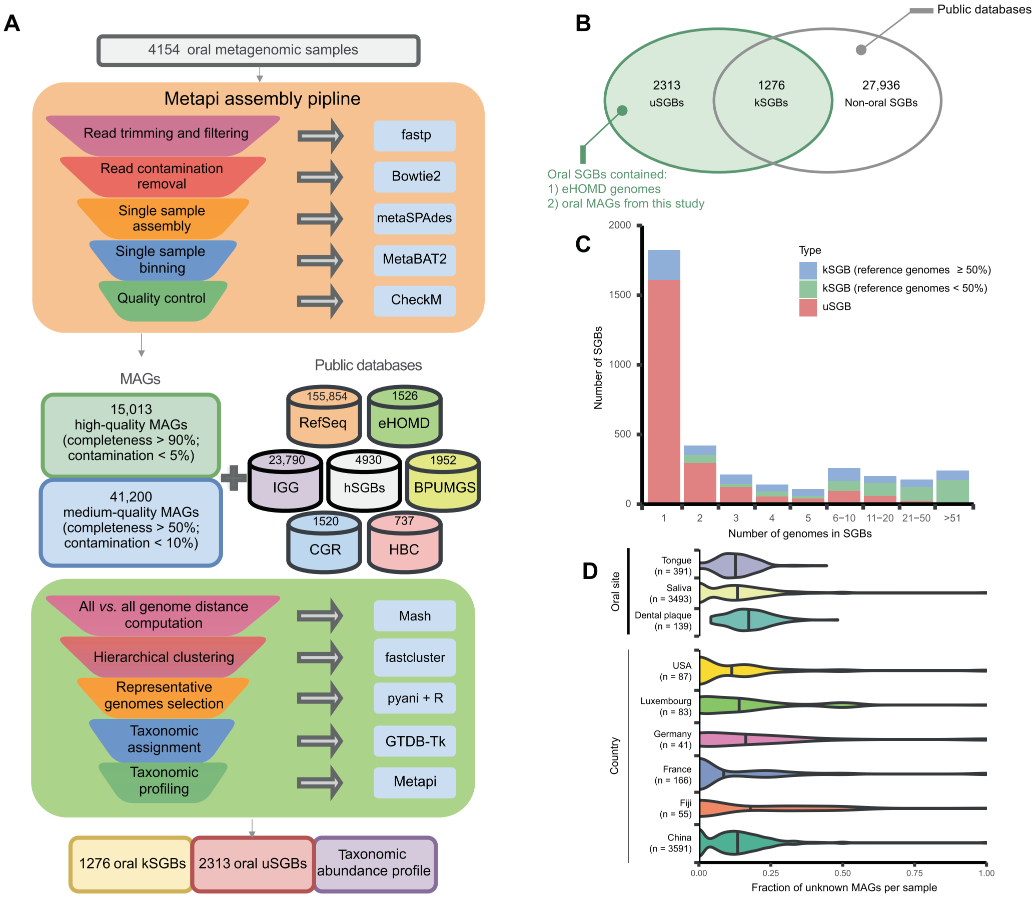

Figure 1 A total of 3589 oral SGBs assembled from 4154 (3346 new samples) meta-analyzed oral metagenomes
A. The workflow for SGB construction. **B.** Reconstructed MAGs and genomes from public databases were clustered into 31,525 SGBs based on ANI. A total of 3589 oral SGBs (green background) which contained eHOMD genomes or oral MAGs from this study were divided into 1276 kSGBs and 2313 uSGBs. The kSGBs contained pubilc genomes, and the uSGBs only consisted of MAGs from this study. **C.** Genome number distribution of uSGBs and kSGBs. **D.** Distribution of fraction of unknown MAGs in each sample from different oral sites and geographic origins. Numbers in brackets indicate the sample size. SGB, species-level genome bin; kSGB, known SGB; uSGB, unknown SGB; MAG, metagenome-assembled genome; ANI, average nucleotide identity; eHOMD, the expanded Human Oral Microbiome Database.

repeatedly captured in our data, with more than 50 genomes each (Figure 1C); the 1276 known oral SGBs (kSGBs) could be further divided according to the percentage of reference genomes in the cluster. Interestingly, kSGBs with < 50% reference genomes outnumbered kSGBs with ≥ 50% reference genomes for clusters containing more than 10 genomes (Figure 1C), underscoring the discovery power of large metagenomic cohorts. MAGs assembled from the 4D-SZ cohort made the greatest contribution to the number of uSGBs, followed by the rheumatoid arthritis cohort and the Yunnan cohort (Figure S2A). The number of medium- and high-quality MAGs per sample increased with the microbial sequence bases, reaching saturation at 15 gigabases (Figure S2B). Regarding the ratio of new MAGs in the samples, we retrieved a greater fraction of previously unknown genomes in dental samples than in saliva or tongue samples, even though we had many more saliva samples than tongue samples and published dental samples (Figure 1D). This ratio also appeared to differ between cohorts, with less than 10% unknown MAGs for samples from France or USA, and more newly matched uSGBs for samples from Fiji, Germany, and Luxembourg (Figure 1D). The large cohort available from this study is crucial for the retrieval of novel oral species, contributing over 2000 uSGBs, which greatly expands our knowledge of oral microbiome diversity.

Close to 90% representation of oral metagenomic data by the genomes

We next examined the ability of this species-level genome set to represent the metagenomic shotgun data. We assessed the percentage of reads that could align to cultured genomes (eHOMD) only and cultured complemented by metagenomically assembled genomes (Table S3). The 1526 genomes from eHOMD led to a median mapping rate of 67.86%; the 4930 representative human SGBs from a recent large-scale assembly study [18] led to a median mapping rate of 80.12%; the 3589 representative oral SGBs from the current study led to a median mapping rate of 88.29%, especially for metagenomes from USA and Germany; and the 81 saliva and subgingival

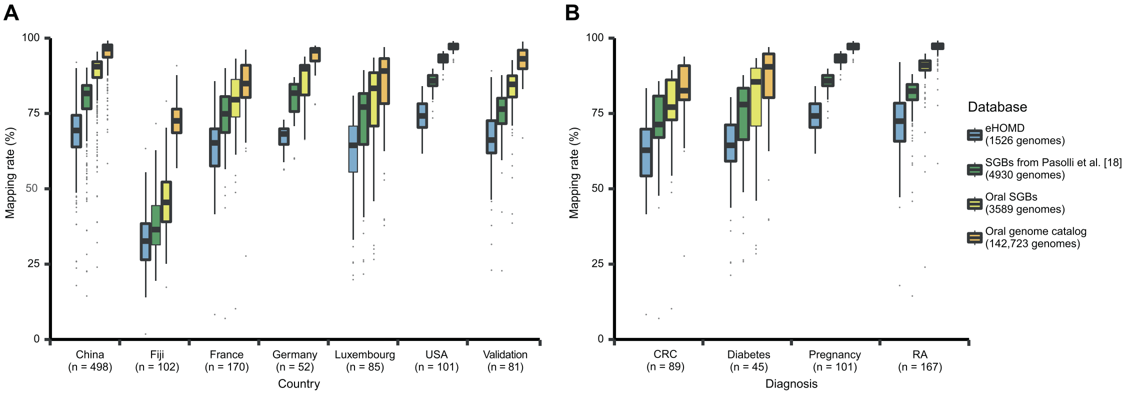

Figure 2 The expanded genome set substantially increases the mappability of oral metagenomes
Reads from all the 808 public samples, 100 randomly selected samples from Shenzhen and Yunnan in China, and 81 additional verified samples which haven't been assembled were mapped against databases including eHOMD, representative genomes of human SGBs from Pasolli et al. [18], representative genomes of oral SGBs, and all genomes from oral genome catalog. **A.** Mappability of samples from different country and a verified data set. **B.** Mappability of the public samples with diseases as well as pregnancy. The numbers in brackets under the boxes indicate the sample size. CRC, colorectal cancer; RA, rheumatoid arthritis.

metagenomes from three verified cohorts [32–34] that were not used in the assembly process even led to a median mapping rate of 84.67% (**Figure 2**A; Table S1). Across physiological states, our SGBs well represented pregnant samples from USA (reaching a median mapping rate of 93.22%), rheumatic arthritis (RA) samples (reaching a median mapping rate of 91.00%), and diabetes samples (reaching a median mapping rate of 85.49%) (Figure 2B). Such a high degree of representation of metagenomic data across geography, ethnicity, age, and physiological states suggests that the expanded genomic content of oral SGBs could serve as a starting point for quantitative taxonomic and functional analyses of the human oral microbiome.

While with all genomes from the oral genome set, the median mapping rate rose to 94.36%. The remaining unmapped reads were mainly classified as homo using Kraken2 [35], especially for metagenomes from France and Luxembourg (Figure S3A). For Fiji, besides homo contamination, 12.07% of unmapped reads were classified as bacteria, and 57.40% of unmapped reads were still unclassified (Figure S3B).

Taxonomic landscape of the oral microbial genomes

The taxonomic classification of 3589 representative genomes of oral SGBs was assigned using GTDB-Tk [36–40] with external Genome Taxonomy Database (GTDB) release 89. Similar to the gut microbiome, Firmicutes took up the largest number of branches (1248 SGBs, 44,217 genomes). The other SGBs distributed into 15 phyla, including major human oral phyla such as Actinobacteria (490 SGBs, 14,574 genomes), Bacteroidetes (368 SGBs, 23,976 genomes), Proteobacteria (364 SGBs, 49,270 genomes), Campylobacterota (280 SGBs, 3295 genomes), and Fusobacteriota (145 SGBs, 2124 genomes) (**Figure 3**A; Table S3). uSGBs accounted for 64.45% in the total reconstructed phylogenetic branch length, with 78.57% of the diversity in the Campylobacterota phylum contributed by the new uSGBs, followed by 73.66% for Patescibacteria and 72.41% for Fusobacteriota (Figure 3B), which seemed overlooked by culturing studies. We estimated that there was the median of 210 SGBs with a relative abundance higher than 0.1% per sample (Figure S2C). Besides, uSGB also had very high abundance, explained for 68.10% of richness and 65.23% of relative abundance per sample (Figure S2D and E). Our MAGs greatly expanded the species' or strains' diversity within each phylum. As many as 596 SGBs from 4028 genomes belonged to the candidate superphylum of Patescibacteria (Parcubacteria, also known as OD1), which only had 157 kSGBs with 178 reference genomes. We noted a few not so well studied phyla that were interesting in analogy to the gut microbiome. *Akkermansia* is the only genus from the Verrucomicrobiota phylum in the human gut and intensively pursued for its role in health and diseases, and the Verrucomicrobiota and Spirochaetota phyla take up a greater fraction in Hadza hunter gatherers compared to developed countries [41]. Here, we identified 67 SGBs from 958 genomes belonging to the Spirochaetota phylum, while only 32 kSGBs with 121 reference genomes were available for Spirochaetota. Moreover, we identifed 142 kSGBs with 1526 genomes belonging to the Patescibacteria phylum (candidate division Saccharibacteria, TM7) (Table S3).

In addition, 265 of the 2313 uSGBs had taxonomic information until order or family level, but cannot be annotated to a known genus. The top 3 uSGB classified families were Saccharimonadaceae (17.99%), Streptococcaceae (12.88%), and Campylobacteraceae (9.51%), whereas the top 3 assigned genera were *Streptococcus* (12.88%), *Campylobacter* (7.65%), and *TM7x* (5.92%) (Figure 3C).

A new genus with small genomes

In the Acholeplasmataceae family of the Bacilli class, a number of our uSGBs with high-quality MAGs formed a clade with shallow branches, which was distinct from the *Acholeplasma* and *Candidatus* Phytoplasma genera (**Figure 4**A). The genome size of this genome-defined genus (temporarily denoted as *Candidatus* Bgiplasma) was 0.69 ± 0.05 Mb, which was similar to that of *Candidatus* Phytoplasma (0.64 ± 0.14 Mb), but much smaller than that of *Acholeplasma*

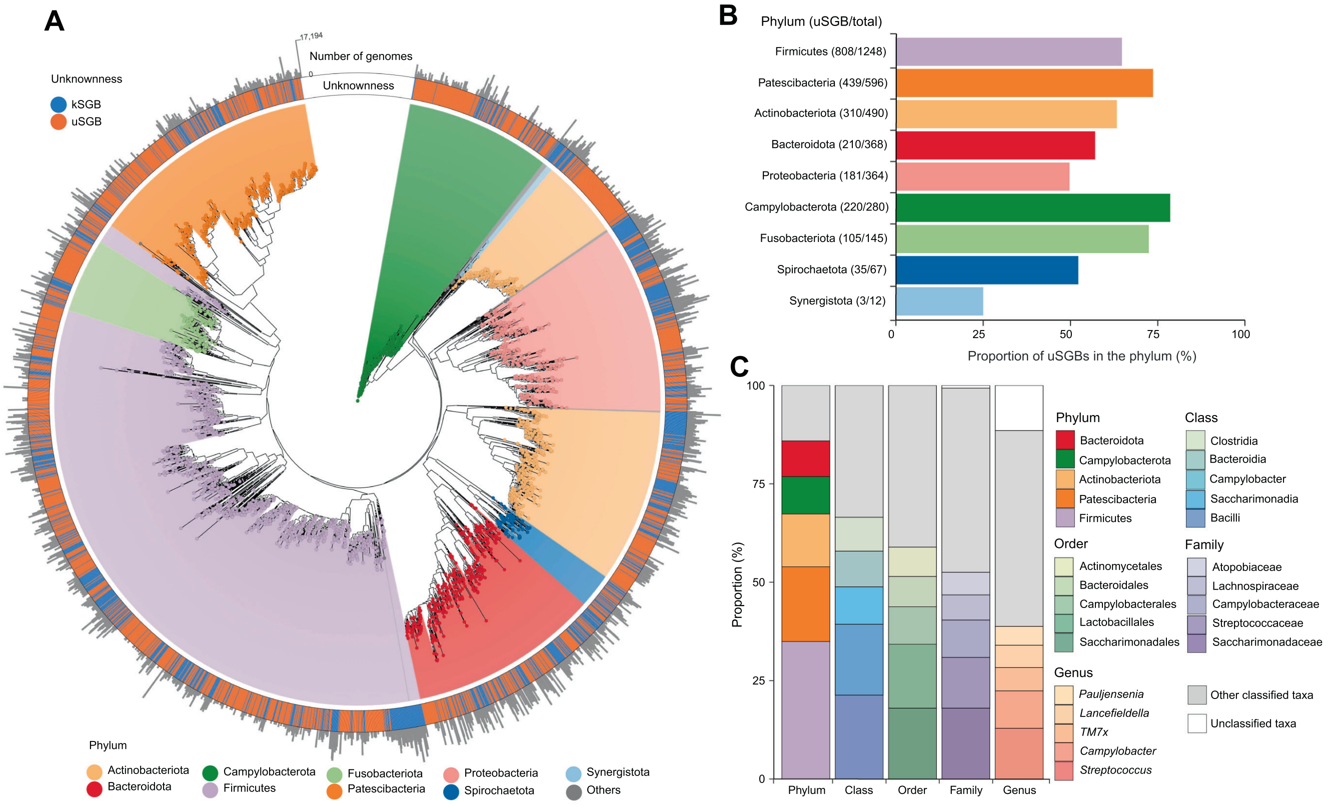

Figure 3 Phylogeny of representative oral SGBs

A. Oral-associated microbial phylogenetic tree of representative genomes from SGBs. Clades and their backgrounds are colored based on the phylum level. Unknownness of SGBs is shown in the ring. Outer histogram with $\log_{10}$ axis shows the genome numbers of SGBs. **B.** Proportion of uSGBs at the phylum level. Numbers of uSGBs and total SGBs are shown in brackets. **C.** Taxonomic composition of the 2313 uSGBs compared to GTDB. Five most frequently observed taxa at the phylum, class, order, famlily, and genus levels are shown in different colors; unclassified taxa are filled in white; and the remaining lineages are grouped as "other classified taxa" and shown in gray. GTDB, Genome Taxonomy Database.

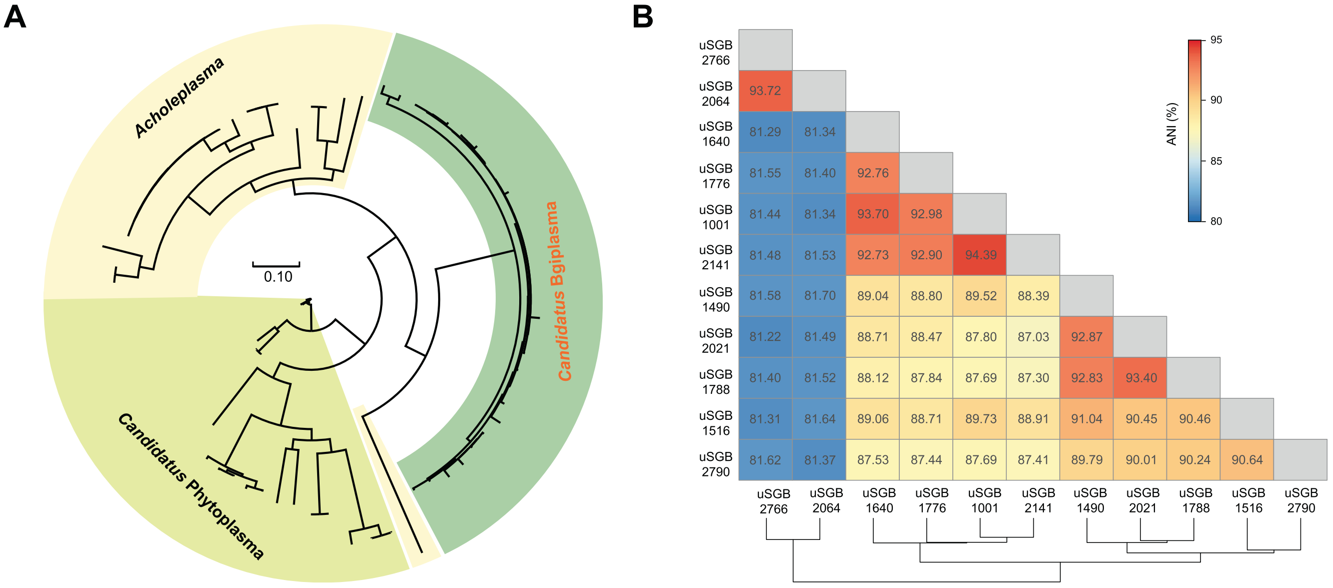

Figure 4 A new candidatus genus is found within the Acholeplasmataceae family

A. Phylogenetic tree drawn from all MAGs in the new candidatus genus *Candidatus* Bgiplasma and public genomes in the Acholeplasmataceae family. **B.** ANI among all uSGBs in the *Candidatus* Bgiplasma genus indicates two species clades (ANI < 85%).

(1.50 ± 0.20 Mb). However, the genome of *Candidatus* Bgiplasma was complete according to single-copy marker genes in CheckM (Table S2). The GC contents of the three clades were also different. The *Candidatus* Bgiplasma genus was more toward normal GC content (34.57% ± 0.21%), but not as low as those of *Acholeplasma* (30.99% ± 1.75%) and *Candidatus* Phytoplasma (25.98% ± 2.68%) (Table S5). Despite the lack of deep branches, the ANI distribution of uSGBs within the *Candidatus* Bgiplasma genus showed two separate groups at species-level divergence (ANI < 85%) (Figure 4B), illustrating diversity within this new genus. These 11 uSGBs comprising 29 MAGs contributed more than 0.1% relative abundance in 209 of 4154 (5.03%) samples, indicating that this genus is a potentially important but so far uncharacterized clade in the oral microbiome.

All genes of the *Candidatus* Bgiplasma genus were annotated by eggNOG mapper [42,43] and the rate of annotation was 81.83% (Table S5). Within the compact genome of *Candidatus* Bgiplasma, Clusters of Orthologous Groups (COG) categories of replication, recombination and repair, posttranslational modification, protein turnover, chaperones, and inorganic ion transport and metabolism are found (Figure S4).

Distribution of species and strains

The new samples from this study differed in oral microbiome composition compared to published samples across geography (**Figure 5**A). Three groups of Chinese samples (Shenzhen: southern China, Yunnan: southwest China, Beijing: northern China) appeared to be as distinct from each other as those from any of the other geographic groups. Both the 4D-SZ and Yunnan samples abundantly contained many uSGBs, such as *Neisseria flavescens*, *Neisseria* spp., and *Porphyromonas* spp., and kSGBs, such as *Porphyromonas gingivalis* (an opportunistic oral pathogen), which was rare in the other cohorts (Figure 5B). Pregnant samples from USA contained *Fannyhessea vaginae* (the vaginal pathogen previously known as *Atopobium vaginae* [44]), *Urinacoccus*, *etc*., which were of much lower abundance in other cohorts (Figure 5B). Samples from

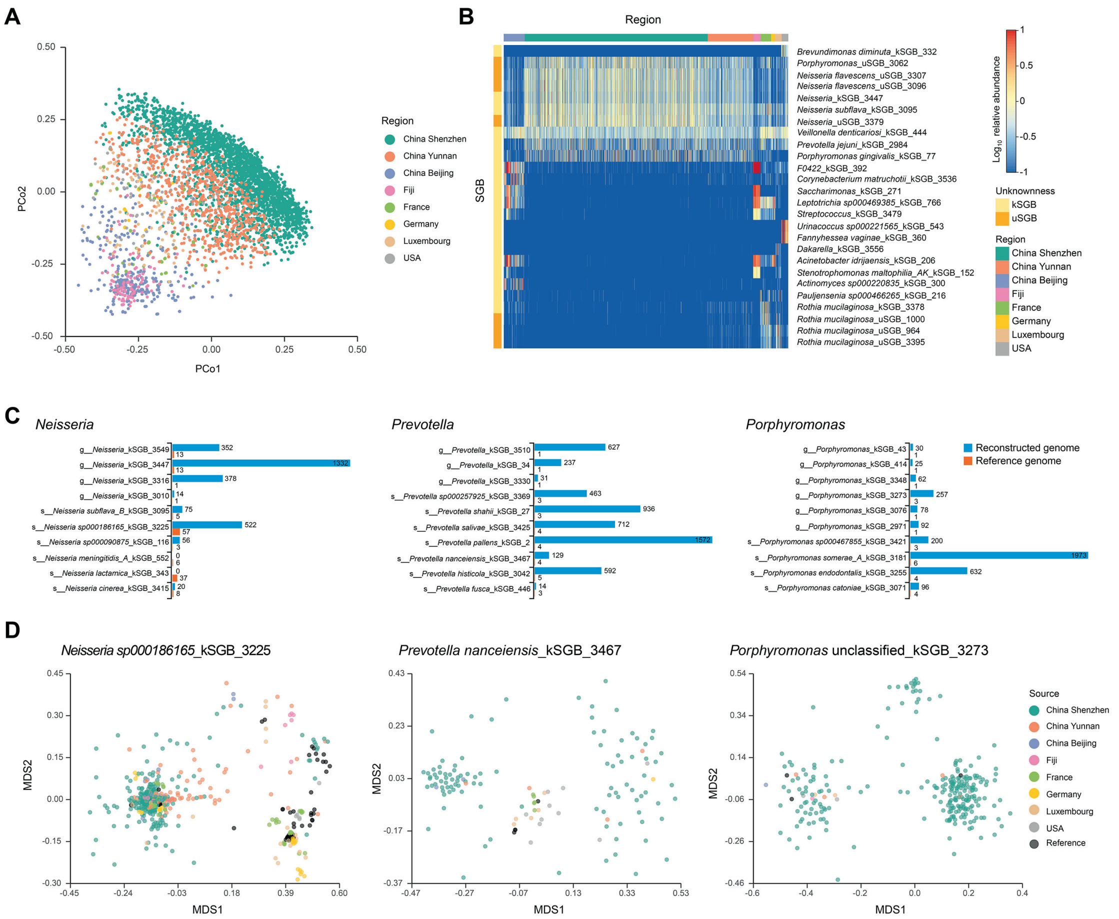

Figure 5 Geographic distribution of oral SGBs and strains

A. PCoA plot based on Bray-Curtis distances of oral SGB relative abundance profile highlights distinct microbial communities among different original populations. **B.** Heatmap of top 5 abundant SGBs in each geographic region. Columns are samples from different regions separated by colors. Unknownness and classification of top SGBs are shown on the left and right of the heatmap, respectively. **C.** Bar plot showing genome numbers of top 10 most frequent species from common oral genus. Reconstructed MAGs largely extend the reference set. **D.** MDS showing the distribution of assembled MAGs and reference genomes based on ANI. Only high-quality MAGs and reference genomes are presented here. Points are genomes colored by country. PCoA, principal coordinate analysis; MDS, multidimensional scaling.

Fiji, although not well mapped (Figure 2A), showed high levels of a few SGBs that were also seen in the RA study from Beijing, China, including an SGB from *Saccharimonas* (Figure 5B).

At the strain level, the new samples from the current study greatly expanded the genome collection for common taxa such as *Neisseria* spp., *Porphyromonas* spp., and *Prevotella* spp. (Figure 5C and D). The numbers of publicly available reference genomes for the top 10 most abundant species in the genera *Porphyromonas* and *Prevotella* were less than 10, and less than 100 for the genus *Neisseria*. Here we obtained more than 1000 genomes for a few of the species and increased the diversity in all the species in these genera (Figure 5C). And rarefaction analyses revealed that the number of genes in both the public reference genome and our reconstructed MAGs was significantly higher than the public reference genome only (Figure S5). The phylogenetic structure of select species with a large number of genomes also suggests strain-level variations (subspecies). The *Prevotella nanceiensis*_kSGB_3467, for example, included 3 reference genomes that were similar to a few genomes from developed countries, while our samples contributed two large clusters that were more distantly related (Figure 5D).

Potential functions in drug metabolism

Many human-targeted drugs are reported to be metabolized to their inactive form by the human gut microbiome or to be affected by the gut bacteria [45–47]. We conducted literature survey and curated experiments which confirmed that microbiome genes could metabolize drugs. To address the potential of drug metabolite of oral microbiome, we mapped genes that could metabolize drugs to our oral SGB genomic contents. Interestingly, we found that many oral communities contained homologs to those gut bacteria encoding enzymes relating up to 41 human-targeted drugs (Table S6). With the same method, we also identified 20 genes that encode key enzymes involved in 12 human diseases, and 6 genes that produce non-traditional antibacterial therapeutic compounds in our oral SGB genomic contents (Table S6). These results suggest that the oral microbiome may play an important role in medical therapy and disease development. Moreover, there were a total of 2696 SGBs containing a β-glucuronidase enzyme that can metabolite anti-cancer drug Gemcitabine (2′,2′-difluorodeoxycytidine) into its inactive form [47]. Besides, 456 SGBs had an agmatine gene for for anti-type II diabetes drug Metformin; 225 SGBs had a tyrosine decarboxylase encoding gene for anti-parkinson drug L-dopa; and 1733 oral SGBs contained genes that can produce small molecules taurine and 5-aminovalerate which are potential drugs for autism spectrum disorder. Only a few SGBs contained *CutC*/*CutD* genes which encode key enzymes for trimethylamine, a metabolite with a high cardiovascular event risk.

A strongly male-enriched *Campylobacter* species

Gender-associated differences were observed in the saliva biochemical parameters, oral disease, and microbiota [48–50]. It is necessary to know the differences in oral microbe distribution between females and males in the natural population. We thus built a random forest classifier for gender with all 3589 oral SGBs in saliva samples of our 4D-SZ cohort (Shenzhen, China, 1025 females and 959 males). The Area Under Curve (AUC) of all SGBs for gender was 0.792 [95% confidence interval (CI): 0.772–0.811]. Four of the top 5 most important oral SGBs were from uSGBs. The random forest importance of the first ranking uSGB (g__*Campylobacter_A*_uSGB_1674) was five times more than the second ranking uSGB (g__*Stomatobaculum*_uSGB_1040) (**Figure 6**A). The AUC of g__*Campylobacter_A*_uSGB_1674 was 0.722 (95% CI: 0.700–0.744), showing the similar predict power for gender as all 3589 oral SGBs (Figure 6B).

We next examined whether g__*Campylobacter_A*_uSGB_1674 is a conservative gender-related bacterium in saliva by integrating other existing cohort data across different populations (see Materials and methods). As shown in Figure 6C, g__*Campylobacter_A*_uSGB_1674 was detected in oral samples from multiple regions and displayed a strongly male-enriched pattern. Moreover, the difference between genders of this uSGB was only observed in saliva, but not in dental or tongue (Figure 6D). An odds ratio (OR) is a measure of association between an exposure and an outcome, and OR can be adjusted by available confounders [such as age, body mass index (BMI), and healthy status] in all cohort data by logistic regression. In our 4D-SZ cohort, the OR of g__*Campylobacter_A*_uSGB_1674 was 2.796 (OR > 1 means male-enriched; 95% CI: 2.504–3.137); the ORs of g__*Campylobacter_A*_uSGB_1674 for ZellerG_2014 (Germany and France, 93 females and 128 males) was 1.322 (95% CI: 1.175–1.531), for Heintz_2016 (Luxembourg, 47 females and 36 males) was 1.407 (95% CI: 1.131–1.865), for IlanaB_2019 (Fiji, 50 females and 52 males) was 5.913 (95% CI: 2.972–13.610), for ZhangX_2015 (Beijing, China, 121 females and 32 males) was 2.115 (95% CI: 1.524–3.099), and for Yunnan (Yunnan, China, 441 females and 229 males) was 6.533 (95% CI: 4.962–8.884). GoltsmanDSA_2018 was excluded because it only had female samples. Virulence factor (VF) analysis showed that 15 genes of 22 identified VFs in this uSGB were flagella associated, suggesting its migration ability adapted to saliva (Table S7).

We suspected that g__*Stomatobaculum*_uSGB_1674 could adapt to male-enriched metabolites. Therefore, the gene function prediction was performed based on genome annotation. As a result, 47 previously reported male-enriched metabolites [50] were involved in metabolic steps that could be catalyzed by g__*Stomatobaculum*_uSGB_1674 (Table S7).

We wondered whether g__*Campylobacter_A*_uSGB_1674 is linked to human diseases. The OR for dental calculus (1849 healthy and 429 dental calculus samples) was 1.294 (OR > 1 means disease-enriched; 95% CI: 1.163–1.443), which suggests that g__*Campylobacter_A*_uSGB_1674 in saliva is a risk of dental calculus. In contrast, the OR for type I diabetes (40 healthy and 40 type I diabetes samples) was 0.795 (95% CI: 0.649–0.938), for RA dental (dental samples collected from 51 controls and 91 RA patients) was 0.566 (95% CI: 0.408–0.756), and for RA saliva (saliva samples collected from 80 controls and 73 RA patients) was 0.598 (95% CI: 0.419–0.815) (Figure S6).

New disease markers according to the oral genomes

To illustrate the utility of our genome collection in metagenomic studies including MWAS, we reanalyzed the dental and saliva microbiome data from RA patients and controls

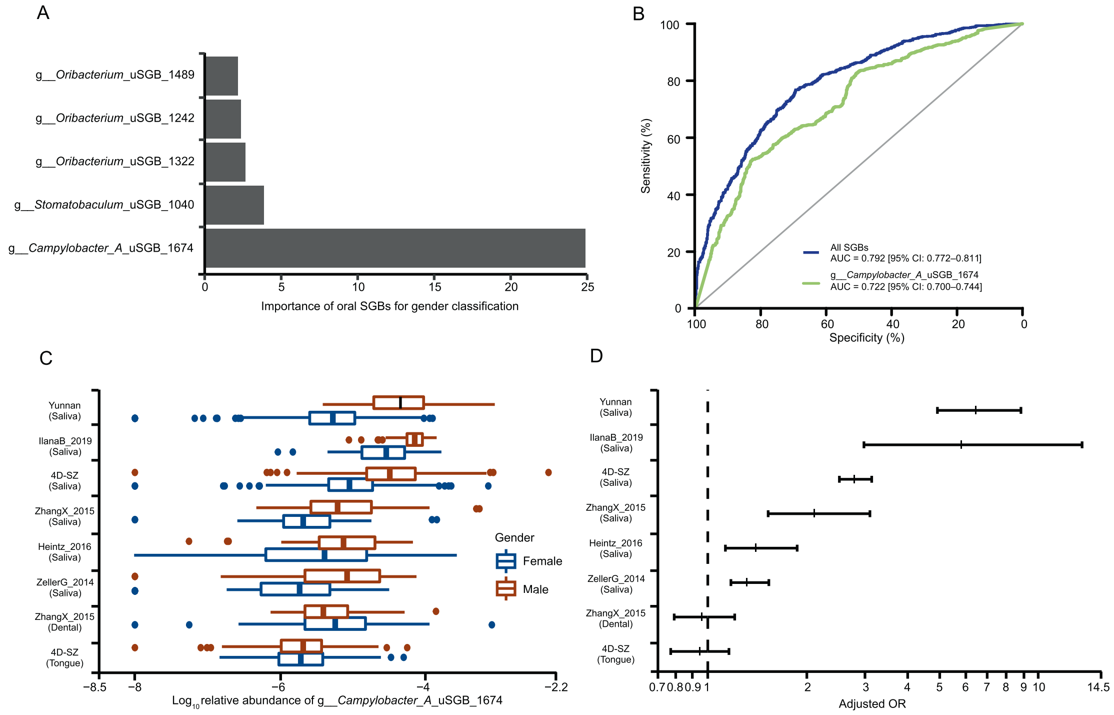

Figure 6 The strongly male-enriched g__*Campylobacter_A*_uSGB_1674

A. The random forest importance of gender classifier. Here shows the top 5 important oral SGBs for gender classifier. **B.** AUC plot of random forest classifiers with all oral SGBs and a single uSGB g__*Campylobacter_A*_uSGB_1674. The out of bag prediction from random forest and true measure is used. **C.** Relative abundance ($\log_{10}$ transformation) of g__*Campylobacter_A*_uSGB_1674 between females and males across different studies. **D.** OR (OR $>$ 1 means male-enriched) of g__*Campylobacter_A*_uSGB_1674 across different studies. g__*Campylobacter_A*_uSGB_1674 ($\log_{10}$ transformation) was regressed against gender and adjusted potential confounders such as age, BMI, and healthy status if available using generalized linear model. The middle of the bar is OR. The upper and lower limits are the 95% CI of the OR. AUC, area under curve; OR, odds ratio; CI, confidence interval; BMI, body mass index.

[8]. For better confidence in the markers regardless of the cohorts, we only analyzed the SGBs containing $>$ 10 genomes. Similar to the original study, oral markers selected by a 5× 10-fold cross-validated gradient boosting model (GBM) included a number of Gram-negative bacteria, *e.g.*, *Haemophilus* spp. and *Aggregatibacter* spp., enriched in dental samples from healthy volunteers, while only a *Pseudomonas* SGB and an *Enterococcus* SGB were selected for RA samples (Figure S7A). Interestingly, the two new RA dental markers appeared more abundant in control saliva samples. The strongest marker from healthy saliva remained *Lactococcus lactis* [7]. *Lactobacillus paracasei* and *Streptococcus infantarius* were identified, reminiscent of the beneficial effects of *Lactobacillus casei* gavage in a rat model of RA [51,52]. The assembled genomes allowed matching of different species in the *Veillonella* genus as RA saliva markers. Moreover, *Pauljensenia*, a genus recently renamed from *Actinomyces* [53], was identified as a highly predictive marker of RA. As *Actinomyces* spp. are the basis for dental attachment of oral bacteria [54], the potential contribution of *Pauljensenia* spp. to periodontitis in RA patients remains to be explored. In addition, the dental microbiome was obviously deranged, consistent with epidemiology [7].

A set of saliva samples from CRC patients and controls from France are also available [55]. Here, we found *Pauljensenia* spp. to be control-enriched, along with *Acinetobacter radioresistens*, *Lachnoanaerobaculum* sp., and *Catonella* sp., *etc*. (Figure S7B). *Streptococcus thermophilus*, a species previously found to be enriched in fecal samples from control or adenoma compared to CRC patients [56], was also identified in control saliva. The markers enriched in CRC oral samples were more unexpected than CRC fecal samples. Besides *Porphyromonas* spp. and *Prevotella maculosa*, we found a *Lachnospiraceae* SGB (potentially TMA-producing and consistent with gut results [10,57–59]), *Capnocytophaga leadbetteri*, and *Cardiobacterium hominis*, *etc*. (Figure S7B). Thus, the substantially expanded collection of oral microbial genomes enabled the discovery of new disease markers and the genomic representation of previously reported markers, facilitating the shift from fecal to oral microbiome-based diagnosis and therapeutics.

Discussion

In summary, we provided the largest set of oral metagenomic shotgun data, assembled 56,213 draft genomes for the human oral microbiome, including 2313 new species as well as many new strains of known species. More than 94.36% of metagenomic reads can now be mapped to the expanded genome catalog for oral microbiomes, enabling much more comprehensive profiling of these communities. Many uSGBs are not currently represented by cultured isolates. The results illustrate that culturomics has not even exhausted the microbial complexity in the more accessible body sites and that metagenomic data for large cohorts of non-fecal samples have great potential. A number of taxa with compact genomes were identified in this study, such as CPR and Mollicutes. Much remains to be elucidated for the metabolic requirement of small bacteria in the oral microbiome. Oral bacteria also contribute to the discovery of new CRISPR-Cas systems [60]. Species with thousands of metagenomic and isolated genomes would be amenable to microbial genome-wide association studies [61] to discover virulence factors, drug resistance, and more commensal functions, which has so far only be possible for pathogens.

In this study, we identified a conservative saliva- and male-specific uSGB g__*Campylobacter_A*_uSGB_1674 across cities and countries. It was assembled from our 4D-SZ cohort. What is more interesting is that g__*Campylobacter_A*_uSGB_1674 is only significantly different in saliva, not in tongue or dental. Genome analysis showed that this saliva-specific difference of gender may be due to its flagella enrichment function which powers its swimming ability in saliva. Function annotation by eggNOG showed that many male-enriched metabolites can be catalyzed by g__*Campylobacter_A*_uSGB_1674, suggeting that it can be more adaptive to male oral metabolome. Further experimental validation of this microbe should be carried out to verify its potential functions.

Although we have reconstructed tremendous MAGs, the oral metagenomic assembly still faces many challenges. For example, the host rate is about 80% in saliva, and high-quality assembly requires further expanded sample size and sequencing depth (Figure S2B). Many homos read contamination were still found after host removal in public data (Figure S3A). A more sensitive host removal method is needed to reduce the influence of host sequences. The oral cavity contains diverse forms of microbes such as protozoa and fungi (like *Candida* as a common oral infection in the immune deficiency population and infants). However, no fungus was observed from our data, probably due to a lack of such samples. Our new oral genome set enables us to identify new biomarkers for RA and CRC, but further validation would depend on more patient samples. By collecting more samples with a variety of diseases, the diversity and quality of the oral genome set will be further improved and would benefit human society by providing health management and disease prevention.

Materials and methods

Collection of the oral microbiomic samples

The 2675 oral metagenomic samples (2284 saliva and 391 tongue samples) from the Chinese 4D-SZ cohort and 671 saliva samples from the Yunnan cohort were newly collected in this study from 2017 to 2018 (Table S1, sheets 4 and 5). The 4D-SZ cohort is a young cohort composed of 2675 people from Shenzhen, China. In those people, 2162 people have age and gender information, with an average age of 30.27 (± 5.58) years old, including 1051 males and 1111 females. No complex diseases were observed in this cohort. And the population of the Yunnan cohort comes from six regions in Yunnan Province of China with an average age of 31.75 (± 10.78) years old, including 260 males and 487 females.

The 4D-SZ samples were self-collected by the volunteers during a physical examination using the MGIEasy Fecal Sample Collection Kit (Catalog No. 10000035265, BGI, Shenzhen, China) containing a room temperature stabilizing reagent to preserve the metagenome [62]. For the saliva sample, a double concentration of stabilizing reagent kit was used and 2 ml saliva was collected. The samples from Yunnan Province were self-collected using a commercial kit (Catalog No. 401103, Zeesan, Xiamen, China). Yunnan samples were approved by The First Affiliated Hospital of Kunming Medical University, China. The collected samples were temporarily stored in freezers at −80 °C and then transported to China National GeneBank (CNGB), Shenzhen on dry ice via commercial logistics.

DNA extraction, sequencing, and quality control

DNA extraction of the stored samples within the next few months was performed using the MagPure Stool DNA KF Kit B (Catalog No. MD5115, Magen, Guangzhou, China) from 1 ml of each sample [63]. Metagenomic sequencing was done on the BGISEQ-500 platform (BGI, Shenzhen, China) [64] (100 bp of paired-end reads for all samples and four libraries were constructed for each lane) and generated 101.4 billion pairs of raw reads (mean: 75.8 million paired-end reads per sample; standard deviation: 14.1 million paired-end reads). DNA was extracted in the same way as above. Sequencing was performed on the BGISEQ-500 machines and generated 26.5 billion single-end 100 bp length reads (mean: 39.5 million single-end reads per sample; standard deviation: 8.2 million single-end reads). After filtering and trimming with strict filtration standards (not less than mean quality phred score 20 and not shorter than 51 bp read length) using fastp (v0.19.4) [65], host read contamination removing using Bowtie2 (v2.3.5) [66] (human genome GRCh38) and seqtk (v1.3) [67], and quality control, we totally got 54.9 billion high-quality paired-end reads and 7.1 billion high-quality single-end reads.

Collection of public metagenomic samples

A total of 808 public oral metagenomic datasets were downloaded from NCBI Sequence Read Archive database (SRA: SRP029441, ERP006678, SRP133047, ERP110622, and SRP07256), which came from five different studies [8,21–24] (Table S1, sheet 3) that have been reported previously. These published data were used to do assembly and profile.

To illustrate the representativeness of the assembled genome set for new data, additional 81 oral metagenomes from three validation cohorts were downloaded from NCBI SRA database (SRA: ERP016024, SRP018108, and SRP052958) (Table S1).

Metagenomic *de novo* assembly, binning, and quality assessment

The high-quality paried-end and single-end reads were individually assembled using the assembly module of metapi pipeline with different max kmer cutoff by different max read length of each sample applying SPAdes (v3.13.0) [68] (paired-end reads with option --meta [25]). After we obtained draft genomes on the contig level of each sample, the reads were mapped back to each assemblies using BWA-MEM (v0.7.17) [69] with default parameters and the contig depth was calculated by jgi_summarize_bam_contig_depths. Then using MetaBAT2 (v2.12.1) [26], we performed metagenomic binning individually for each sample. Finally, we obtained a total of 163,718 bins. After MAG quality assignment by CheckM (v1.0.12) [28] lineage workflow, 15,013 high-quality (completeness > 90% and contamination < 5%) bins and 41,200 medium-quality (completeness > 50% and contamination < 10%) bins (Table S2) were generated based on MIMAG standard [27]. The 16S rRNA sequences in the MAGs were searched by Barrnap (v0.9; https://github.com/tseemann/barrnap) with parameters "--reject 0.01 --evalue 1e-3", and the tRNA sequences in the MAGs were searched by tRNAscan-SE (v2.0.3) [70] with the default parameters.

Clustering metagenomic genomes into SGBs

The 56,213 reconstructed genomes and 190,309 reference genomes were grouped into SGBs by a two-step clustering strategy as reported previously [18] with a slight modification. In the first step, an all-versus-all genetic distance matrix between the 246,522 genomes was carried out using Mash (v2.0) [71] ("-k 21 -s 1e4" for sketching). Then, hierarchical clustering with average linkage and genetic distance cutoff of 0.05 on the distance matrix by fastcluster [72] generated 33,008 clusters. Because the Mash will underestimate the distance between the incomplete genomes [73] and split same-species genomes into multiple SGBs, we performed clustering base on ANI in the second step. First, we divided the SGBs into kSGBs and uSGBs according to with or without reference genomes. Then, a representative genome was selected for each SGB. For the kSGB, the genome with the largest genome size was selected. For the uSGB, all MAGs were ranked by completeness (in descending order), contamination (in ascending order), coverage (in descending order), strain heterogeneity (in ascending order), and N50 (in descending order), and a representative genome was selected as the one minimizing the sum of the five ranks. We recalculated the more precise genetic distance using pyani (v0.2.9) [74] (option '-m ANIb') for the pairs of representative genomes with mash distances less than 0.95 and only left ANI with genome coverage above 0.3. Following hierarchical clustering with complete linkage based on ANI > 95%, 12,911 representative genomes with mash distances less than 0.95 were merged to 11,427 new clusters. Finally, we obtained 31,525 SGBs by a two-step clustering strategy. In this dataset, only 3589 SGBs included eHOMD genomes or oral MAGs, and they were named as oral SGBs and further divided into 2313 uSGBs and 1276 kSGBs.

Reconstruction of the human oral microbiome phylogenetic structure

The phylogenetic trees of 3589 representative genomes of SGBs (Figure 3A) and 76 genomes of the Acholeplasmataceae family were both built using the 400 PhyloPhlAn markers with the parameters "--diversity high --fast --min_num_markers 80" by the PhyloPhlAn2 [75]. As input data for PhyloPhlAn2, proteomes were predicted using Prodigal (v2.6.3) [37] with default parameters. Following tools with their set of parameters were used in the configuration files: Diamond (v0.9.22.123) [76] with the parameters "blastp --quiet --threads 1 --outfmt 6 --more-sensitive --id 50 --max-hsps 35 -k 0"; Mafft (v7.407) [77] with the "--anysymbol" option; Trimal (v1.4.rev15) [78] with the "-gappyout" option; Iqtree (v1.6.12) [79] with the parameters "-quiet -nt AUTO -m LG". After that, 3433 representative genomes of SGBs contained more than 80 markers, and the phylogenetic tree in Figure 3A was generated using GraPhlAn (v1.1.3) [80]. All genomes of the Acholeplasmataceae family were remained and the phylogenetic tree in Figure 4A was generated using FigTree (v1.4.4).

Taxonomic and function analyses of SGBs

The taxonomic classification of 3589 representative genomes of SGBs was assigned using the GTDB-Tk (v0.3.2) [36–40] classification workflow with external GTDB release 89. Although some kSGBs already have taxonomy labels, we still used GTDB-Tk to classify them, because GTDB-Tk has its own taxonomy classification system that is different from the NCBI taxonomy database. Then above the genus level, we manually removed the classification tag with a single letter suffix (Table S3). Those suffixes were used to indicate the taxa needed to be subdivided based on the current GTDB reference tree. We used eggNOG mapper (v1.0.3) [42] to do genome-wide functional annotation through orthology assignment.

Mapping rate comparison between different oral-related genome databases

The mapping rates of oral metagenomic reads align to four different oral-related genome databases (eHOMD, SGBs from Pasolli et al. [18], oral SGBs, and oral genome catalog) were compared based on the statistics summary of Bowtie2's results (Table S4). First, we randomly selected 100 oral metagenomic samples from each of the 4D_SZ and Yunnan cohorts. With all 808 public samples and 81 additional verified samples which haven't been assembled (Table S1, sheet 2), a total of 1089 oral metagenomic samples were mapped to these databases respectively using Bowtie2 (v2.3.5) with SE model (-U) and default parameters. For acceleration calculation, all 142,723 oral genomes were divided into 19 subsets and indexed by Bowtie2. Reads were aligned to these databases in turns, and only unmapped reads (samtools fastq -f 4) were retained for the next alignment.

Taxonomic assignment for unmapped reads by Kraken2

Kraken2 is a taxonomic classification system using exact *k*-mer matches and has a higher error tolerance in alignment than Bowtie2. The reads which were unmapped to oral genome set were aligned to the database (maxikraken2_1903_140GB, March 2019) which includes archaea, bacteria, fungi, protozoa, viral, and human from Loman Lab (https://lomanlab.github.io/mockcommunity/mc_databases.html) by Kraken 2.0.8-beta with default parameters.

Metapi for oral SGB metagenomic profiling

The quantification of species relative abundance of oral metagenomic samples was performed with the taxonomic profiling module of metapi pipeline: 1) build the oral genome index of oral representative SGBs by Bowite2; 2) align the high-quality reads of each sample to the oral genome index using Bowtie2 with the parameters "--end-to-end --very-sensitive --seed 0 --time -k 2 --no-unal --no-discordant -X 1200"; 3) obtain the normalized contig depths by using jgi_summarize_bam_contig_depths; and 4) convert the normalized contig depth to the relative abundance of each SGB for each sample base on the correspondence of contigs and genome. Finally, we merged the relative abundances of all representative SGBs to generate a taxonomic profile.

Principal coordinate analysis and heatmap for human oral profile

Principal coordinate analysis (PCoA) of human oral profile was done using the dudi.pco function in ade4 R package based on bray distance from vegan 2.5.2 R package. The mean top 10 most abundant SGBs from every study were merged (total 27 SGBs) to visual in the pheatmap R package.

Pangenome and phylogenetic analyses of kSGBs and uSGBs

From the taxonomic profiling results of 4154 oral metagenomic samples, the most prevalent eight genera were selected based on the ranking of average relative abundance (in descending order), occurrence frequency (in descending order), and oral genome number/SGB size (in descending order), including *Prevotella*, *Neisseria*, *Streptococcus*, *Veillonella*, *Porphyromonas*, *Fusobacterium*, *Pauljensenia*, and *Haemophilus*. Then, we chose the top 10 prevalent species for each genus to do pangenome analysis. All genomes of each SGB were annotated by prokka (v1.13.7) [81] and constructed to the pangenome database via PanPhlAn (v1.2) [82]. Finally, the gene-family presence/absence profile matrix was transformed to a zero/one matrix for reference genomes and the genomes of each SGB were reconstructed to perform rarefaction analysis. Accumulation curves (Figure S5) based on the number of core genes of each SGB were bootstrapped ten times at each sampling interval. The observation of the intra-SGB phylogenetic structure of *Neisseria sp000186165*_kSGB_3225, *Prevotella nanceiensis*_kSGB _3467, and *Porphyromonas* unclassified_kSGB_3273 was performed by the nonmetric multidimensional scaling analysis using the metaMDS function of R package vegan (v2.5.2).

Identification of a novel saliva- and male-specific oral uSGB g__*Campylobacter_A*_uSGB_1674

We built a RandomForest classifier for gender with 3589 oral SGBs in saliva samples of our 4D-SZ cohort (1025 females and 959 males) by randomForest 4.6-14 R package. Receiver operating characteristic (ROC) curve was plotted with pROC R package. The most importance SGB for the gender classifier was identified to be g__*Campylobacter_A*_uSGB_1674 (Figure 6A). Generalized linear model (GLM) analysis was performed to further confirm the association between g__*Campylobacter_A*_uSGB_1674 and gender after adjusting potential confounders. The ORs extracted from the GLM were plotted in Figure 6D with R package ggplot2. The association between g__*Campylobacter_A*_uSGB_1674 and diseases was also tested with GLM after adjusting confounders and the ORs were plotted in Figure S6. For 4D-SZ, we adjusted age, antibiotics usage during last six months, dietary_structure, disease_oral_ulcer, disease_caries, BMI, dental_calculus, inflammatory_gingivitis, and decayed_tooth. For ZellerG_2014 and Heintz_2016, we adjusted age and healthy status. For IlanaB_2019 and Yunnan, we adjusted age only. For ZhangX_2015, we adjusted age, BMI, and healthy status.

Disease markers according to the oral genomes

The metagenome-wide association between 3589 SGB profiles and diseases for previously published CRC and RA studies was done using a GLM with adjustment for potential confounders such as gender, age, and BMI (Table S1). BMI is only available for RA. Species relative abundance was asin-sqrt transformed as previously described [83]. Non-oral SGBs were excluded. Correction for multiple hypothesis tests was done using false discovery rate (FDR). We predicted disease status using GBM in the caret R package, such that 80% of the samples were randomly sampled for each estimator. The depth of the tree at each estimator was not limited, but leaves were restricted to have at least 30 instances. We used 4000 estimators with a learning rate of 0.002. All the oral marker SGBs with FDR < 1% were included in the model as predictors. To avoid overfitting, 5× 10-fold cross-validated ROC was used to measure the model performance. VarImp function was used to extract the GBM importance.

Ethical statement

The study was approved by the Institutional Review Board (IRB) of BGI-Shenzhen (Nos. BGI-IRB19121 and BGI-IRB17162) and the ethics committee of No.1 Affiliated People's Hospital of Kunming Medical University [(2017) Ethics review L No.14], China. Informed consent was obtained from each participant.

Code availability

The pipeline used in this study is available at https://github.com/ohmeta/metapi. The scripts of figures are available at https://github.com/ohmeta/oral-assembly.

Data availability

All sequence data are available at CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) (CNSA: CNP0000687 for the 4D-SZ cohort and CNP0001221 for the Yunnan cohort). Oral SGB genomes and genome annotations are available at Microbiome Database of National Genomics Data Center (https://cngb.org/microbiome/genomecatalog/human_oral/). Oral SGB genomes and genome annotations have also been deposited in the Genome Sequence Archive [84] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (BioProject: PRJCA003731), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa.

CRediT author statement

Jie Zhu: Conceptualization, Methodology, Software, Visualization, Writing - original draft. **Liu Tian:** Conceptualization, Methodology, Visualization, Writing - original draft. **Peishan Chen:** Investigation. **Mo Han:** Investigation. **Liju Song:** Investigation. **Xin Tong:** Investigation. **Xiaohuan Sun:** Investigation, Writing - review & editing. **Fangming Yang:** Investigation. **Zhipeng Lin:** Investigation. **Xing Liu:** Investigation. **Chuan Liu:** Investigation. **Xiaohan Wang:** Investigation. **Yuxiang Lin:** Investigation. **Kaiye Cai:** Investigation. **Yong Hou:** Supervision. **Xun Xu:** Supervision. **Huanming Yang:** Supervision. **Jian Wang:** Supervision. **Karsten Kristiansen:** Writing - review & editing. **Liang Xiao:** Supervision. **Tao Zhang:** Supervision. **Huijue Jia:** Conceptualization, Writing - review & editing, Supervision, Project administration. **Zhuye Jie:** Conceptualization, Methodology, Visualization, Writing - original draft, Project administration. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work is supported by the National Key Program for Infectious Diseases of China (Grant No. 2018ZX10714-002). We gratefully acknowledge Dr. Junkun Niu, Dr. Yang Sun, and Dr. Yinglei Miao from the First Affiliated Hospital of Kunming Medical University for Yunnan sample collection. And we also thank colleagues at BGI-Shenzhen and CNGB, Shenzhen for sample collection, DNA extraction, library construction, sequencing, and discussions.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.05.001.

ORCID

ORCID 0000-0002-5789-6296 (Jie Zhu)
ORCID 0000-0002-4421-5105 (Liu Tian)
ORCID 0000-0002-6968-4303 (Peishan Chen)
ORCID 0000-0002-4404-297X (Mo Han)
ORCID 0000-0001-9935-181X (Liju Song)
ORCID 0000-0002-4849-6522 (Xin Tong)
ORCID 0000-0002-3411-5995 (Xiaohuan Sun)
ORCID 0000-0001-6979-7423 (Fangming Yang)
ORCID 0000-0003-2948-9578 (Zhipeng Lin)
ORCID 0000-0003-1741-3437 (Xing Liu)
ORCID 0000-0003-3716-7050 (Chuan Liu)
ORCID 0000-0003-4658-4171 (Xiaohan Wang)
ORCID 0000-0002-9634-3993 (Yuxiang Lin)
ORCID 0000-0003-0593-1878 (Kaiye Cai)
ORCID 0000-0002-0420-0726 (Yong Hou)
ORCID 0000-0002-5338-5173 (Xun Xu)
ORCID 0000-0002-0858-3410 (Huanming Yang)
ORCID 0000-0003-4949-4753 (Jian Wang)
ORCID 0000-0002-6024-0917 (Karsten Kristiansen)
ORCID 0000-0003-0836-4397 (Liang Xiao)
ORCID 0000-0003-2765-2802 (Tao Zhang)
ORCID 0000-0002-3592-126X (Huijue Jia)
ORCID 0000-0003-0833-6134 (Zhuye Jie)

References

[1] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012;486:207–14.

[2] Jie Z, Liang S, Ding Q, Li F, Tang S, Wang D, et al. A transomic cohort as a reference point for promoting a healthy human gut microbiome. Med Microecol 2021;100039.

[3] Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 2016;352:565–9.

[4] Whiley RA, Fleming EV, Makhija R, Waite RD. Environment and colonisation sequence are key parameters driving cooperation and competition between *Pseudomonas aeruginosa* cystic fibrosis strains and oral commensal streptococci. PLoS One 2015;10: e0115513.

[5] Drewes JL, White JR, Dejea CM, Fathi P, Iyadorai T, Vadivelu J, et al. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. NPJ Biofilms Microbiomes 2017;3:34.

[6] Dominy SS, Lynch C, Ermini F, Benedyk M, Marczyk A, Konradi A, et al. Porphyromonas gingivalis in Alzheimer's disease brains: evidence for disease causation and treatment with small-molecule inhibitors. Sci Adv 2019;5:eaau3333.

[7] Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 2017;550:61–6.

[8] Zhang X, Zhang D, Jia H, Feng Q, Wang D, Liang D, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med 2015;21:895–905.

[9] Atarashi K, Suda W, Luo C, Kawaguchi T, Motoo I, Narushima S, et al. Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. Science 2017;358:359–65.

[10] Jie Z, Xia H, Zhong SL, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun 2017;8:845.

[11] Liu X, Tang S, Zhong H, Tong X, Jie Z, Ding Q, et al. A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases. Cell Discov 2021;7:9.

[12] Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature 2014;513:59–64.

[13] Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev Microbiol 2016;14:508–22.

[14] Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 2017;66:70–8.

[15] Chen T, Yu WH, 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 (Oxford) 2010;2010:baq013.

[16] Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature 2019;568:499–504.

[17] Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature 2019;568:505–10.

[18] Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 2019;176:649–62.

[19] Willis JR, Gabaldón T. The human oral microbiome in health and disease: from sequences to ecosystems. Microorganisms 2020;8:308.

[20] Jie Z, Chen C, Hao L, Li F, Song L, Zhang X, et al. Life history recorded in the vagino-cervical microbiome. Genomics Proteomics Bioinformatics 2022;20:304–21.

[21] Brito IL, Gurry T, Zhao S, Huang K, Young SK, Shea TP, et al. Transmission of human-associated microbiota along family and social networks. Nat Microbiol 2019;4:964–71.

[22] Goltsman DSA, Sun CL, Proctor DM, DiGiulio DB, Robaczewska A, Thomas BC, et al. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. Genome Res 2018;28:1467–80.

[23] Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol 2017;2:16180.

[24] Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 2014;10:766.

[25] Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res 2017;27:824–34.

[26] Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 2019;7:e7359.

[27] Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;25:1043–55.

[28] Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 2017;35:725–31.

[29] Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP, et al. New insights into human nostril microbiome from the expanded Human Oral Microbiome Database (eHOMD): a resource for the microbiome of the human aerodigestive tract. mSystems 2018;3:e00187-18.

[30] Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol 2019;37:186–92.

[31] Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol 2019;37:179–85.

[32] Lassalle F, Spagnoletti M, Fumagalli M, Shaw L, Dyble M, Walker C, et al. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. Mol Ecol 2018;27:182–95.

[33] Liu B, Faller LL, Klitgord N, Mazumdar V, Ghodsi M, Sommer DD, et al. Deep sequencing of the oral microbiome reveals signatures of periodontal disease. PLoS One 2012;7:e37919.

[34] Shi B, Chang M, Martin J, Mitreva M, Lux R, Klokkevold P, et al. Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis. mBio 2015;6:e01926-14.

[35] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20:257.

[36] Eddy SR, Pearson WR. Accelerated profile HMM searches. PLoS Comput Biol 2011;7:e1002195.

[37] Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 2012;28:2223–30.

[38] Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018;9:5114.

[39] Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 2010;11:538.

[40] Price MN, Dehal PS, Arkin AP. FastTree 2 — approximately maximum-likelihood trees for large alignments. PLoS One 2010;5: e9490.

[41] Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut microbiome of the Hadza hunter-gatherers. Nat Commun 2014;5:3654.

[42] Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 2017;34:2115–22.

[43] Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47:D309–14.

[44] Fredricks DN, Fiedler TL, Marrazzo JM. Molecular identification of bacteria associated with bacterial vaginosis. N Engl J Med 2005;353:1899–911.

[45] Cully M. Microbiome therapeutics go small molecule. Nat Rev Drug Discov 2019;18:569–72.

[46] Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature 2018;555:623–8.

[47] Taylor MR, Flannigan KL, Rahim H, Mohamud A, Lewis IA, Hirota SA, et al. Vancomycin relieves mycophenolate mofetil-induced gastrointestinal toxicity by eliminating gut bacterial beta-glucuronidase activity. Sci Adv 2019;5:eaax2358.

[48] Ioannidou E. The sex and gender intersection in chronic periodontitis. Front Public Health 2017;5:189.

[49] Minty M, Loubières P, Canceill T, Azalbert V, Burcelin R, Tercé F, et al. Gender-associated differences in oral microbiota and salivary biochemical parameters in response to feeding. J Physiol Biochem 2021;77:155–66.

[50] Zaura E, Brandt BW, Prodan A, Teixeira de Mattos MJ, Imangaliyev S, Kool J, et al. On the ecosystemic network of saliva in healthy young adults. ISME J 2017;11:1218–31.

[51] Pan H, Guo R, Ju Y, Wang Q, Zhu J, Xie Y, et al. A single bacterium restores the microbiome dysbiosis to protect bones from destruction in a rat model of rheumatoid arthritis. Microbiome 2019;7:107.

[52] Pan H, Guo R, Zhu J, Wang Q, Ju Y, Xie Y, et al. A gene catalogue of the Sprague-Dawley rat gut metagenome. Gigascience 2018;7:giy055.

[53] Nouioui I, Carro L, García-López M, Meier-Kolthoff JP, Woyke T, Kyrpides NC, et al. Genome-based taxonomic classification of the phylum Actinobacteria. Front Microbiol 2018;9:2007.

[54] Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci U S A 2016;113:E791–800.

[55] Schmidt TS, Hayward MR, Coelho LP, Li SS, Costea PI, Voigt AY, et al. Extensive transmission of microbes along the gastrointestinal tract. Elife 2019;8:e42693.

[56] Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun 2015;6:6528.

[57] Guertin KA, Li XS, Graubard BI, Albanes D, Weinstein SJ, Goedert JJ, et al. Serum trimethylamine N-oxide, carnitine, choline, and betaine in relation to colorectal cancer risk in the alpha tocopherol, beta carotene cancer prevention study. Cancer Epidemiol Biomarkers Prev 2017;26:945–52.

[58] Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 2019;25:679–89.

[59] Xu R, Wang Q, Li L. A genome-wide systems analysis reveals strong link between colorectal cancer and trimethylamine N-oxide (TMAO), a gut microbial metabolite of dietary meat and fat. BMC Genomics 2015;16:S4.

[60] Shmakov S, Abudayyeh O, Makarova K, Wolf Y, Gootenberg J, Semenova E, et al. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. Mol Cell 2015;60:385–97.

[61] Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet 2017;18:41–50.

[62] Han M, Hao L, Lin Y, Li F, Wang J, Yang H, et al. A novel affordable reagent for room temperature storage and transport of fecal samples for metagenomic analyses. Microbiome 2018;6:43.

[63] Yang F, Sun J, Luo H, Ren H, Zhou H, Lin Y, et al. Assessment of fecal DNA extraction protocols for metagenomic studies. Gigascience 2020;9:giaa071.

[64] Fang C, Zhong H, Lin Y, Chen B, Han M, Ren H, et al. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. Gigascience 2018;7:1–8.

[65] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90.

[66] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.

[67] Shen W, Le S, Li Y, Hu F, Zou Q. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 2016;11:e0163962.

[68] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77.

[69] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.

[70] Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol 2019;1962:1–14.

[71] Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132.

[72] Mullner D. fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. J Stat Softw 2013;53:1–18.

[73] Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 2017;11:2864–8.

[74] Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal Methods 2016;8:12–24.

[75] Segata N, Bornigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun 2013;4:2304.

[76] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12:59–60.

[77] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772–80.

[78] Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25:1972–3.

[79] Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015;32:268–74.

[80] Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ 2015;3:e1029.

[81] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9.

[82] Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods 2016;13:435–8.

[83] Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 2012;13:R79.

[84] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.