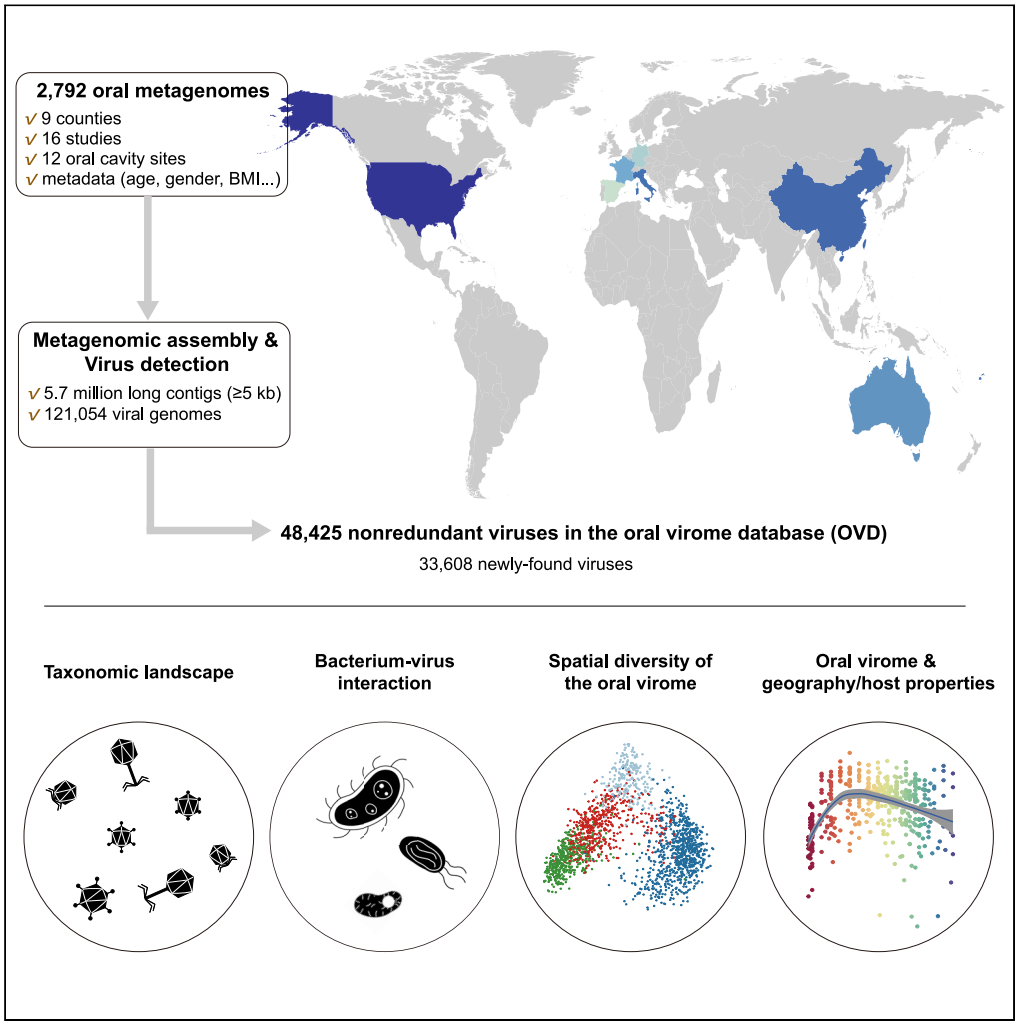
iScience

CellPress
OPEN ACCESS

Article

A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome



Shenghui Li, Ruochun Guo, Yue Zhang, ..., Hao Jin, Guangyang Wang, Qiulong Yan

lsh2@qq.com (S.L.)
qiulongy1988@163.com (Q.Y.)

Highlights

The Oral Virus Database comprises 48,425 viral genomes from 2,792 oral metagenomes

Novel Saccharibacteria phages and jumbo viruses are ubiquitously distributed

Oral virome shows a high degree of spatial variability

Salivary virome exhibits a characteristic age-dependent pattern

Li et al., iScience 25, 104418
June 17, 2022
https://doi.org/10.1016/j.isci.2022.104418

iScience

CellPress
OPEN ACCESS

Article

A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome

Shenghui Li,[1,2,3,8,9,*] Ruochun Guo,[2,8] Yue Zhang,[2,8] Peng Li,[2] Fang Chen,[1] Xifan Wang,[3,4] Jing Li,[5] Zhuye Jie,[6] Qingbo Lv,[2] Hao Jin,[2,7] Guangyang Wang,[1] and Qiulong Yan[1,*]

SUMMARY

The human oral cavity is a hotspot of numerous, mostly unexplored, viruses that are important for maintaining oral health and microbiome homeostasis. Here, we analyzed 2,792 publicly available oral metagenomes and proposed the Oral Virus Database (OVD) comprising 48,425 nonredundant viral genomes (≥5 kbp). The OVD catalog substantially expanded the known phylogenetic diversity and host specificity of oral viruses, allowing for enhanced delineation of some under-represented groups such as the predicted *Saccharibacteria* phages and jumbo viruses. Comparisons of the viral diversity and abundance of different oral cavity habitats suggested strong niche specialization of viromes within individuals. The virome variations in relation to host geography and properties were further uncovered, especially the age-dependent viral compositional signatures in saliva. Overall, the viral genome catalog describes the architecture and variability of the human oral virome, while offering new resources and insights for current and future studies.

INTRODUCTION

Viruses are perhaps the most abundant and diverse organisms on the Earth (Brüssow and Hendrix, 2002; Dion et al., 2020; Paez-Espino et al., 2016) and can contribute to human health and diseases (Jakobsen et al., 2020; Nakatsu et al., 2018; Norman et al., 2015). With the development of next-generation sequencing technologies and virus identification tools, previous studies suggested that the types of viruses inhabiting in humans were far more than the known viruses provided by the Reference Sequence (RefSeq) collection (Clooney et al., 2019; Gregory et al., 2020; Shkoporov et al., 2019), which had limited the exploration of the association between the virus and human health. Thus, expanding the genome resources of human-associated viruses is a key issue for virologists. For example, Soto-Perez et al. (2019) established the Human Virome Database (HuVirDB) via integrating virome samples from multiple body sites (including blood, stool, lung, and skin) and Tisza et al. (Tisza and Buck, 2021) developed the Cenote Human Virome Database (CHVD) from public metagenomic samples spanning gut, mouth, nose, skin, and vagina. Three recent studies had performed large-scale virus identification from public available fecal metagenomes and constructed the Gut Virome Database (GVD) (Gregory et al., 2020), Gut Phage Database (GPD) (Camarillo-Guerrero et al., 2021), and Metagenomic Gut Virus (MGV) catalog (Nayfach et al., 2021). These resources revealed the massive viral diversity in humans and would facilitate the exploration of the viral characteristics and host-virus interaction.

The importance of human oral virome has attracted increasing attention over the past few decades. Some ubiquitous viruses such as the Epstein-Barr virus (EBV) that infect in the oral cavity can probably contribute to oral cancers (Guidry et al., 2018), and the alterations of the oral viral community have been linked to a wide range of oral diseases such as periodontal disease (Ly et al., 2014), lichen planus (Carrozzo, 2008), and hand, foot, and mouth disease (Ho et al., 2021). Also, our previous study had shown that the diversity and structure of saliva and dental plaque viromes were distinctly changed in patients with rheumatoid arthritis and were correlated with their therapeutic plans (Guo et al., 2021b). In addition, longitudinal analysis had revealed that the oral virome of healthy subjects was highly diverse, individually specific, and temporally stable (Abeles et al., 2014; Pride et al., 2012). Because the understanding of oral virome is increasing rapidly, however, only a few studies explored the genome resources of the oral viruses. The

[1]Department of Microbiology, College of Basic Medical Sciences, Dalian Medical University, Dalian 116044, China

[2]Puensum Genetech Institute, Wuhan 430076, China

[3]Key Laboratory of Precision Nutrition and Food Quality, Department of Nutrition and Health, China Agricultural University, Beijing 100083, China

[4]Department of Obstetrics and Gynecology, Columbia University, New York, NY 10032, USA

[5]Department of Rheumatology and Immunology, Peking University People's Hospital, Beijing 100044, China

[6]Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark

[7]College of Food Science and Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China

[8]These authors contributed equally

[9]Lead contact

*Correspondence: lsh2@qq.com (S.L.), qiulongy1988@163.com (Q.Y.)

https://doi.org/10.1016/j.isci.2022.104418

CHVD database contained over 45,000 viral genomes from nearly 6,000 metagenomic samples from various body sites (Tisza and Buck, 2021), including 13,565 viruses that were assembled from 1,287 oral samples, which developed into the current largest viral genome reference of the human oral cavity. Meanwhile, a recent study analyzed four saliva samples using deep short-read and long-read metagenomic sequencing and generated hundreds of viral genomes, including high proportions of *Streptococcus* phages and jumbo phages (Yahara et al., 2021). Collectively, these efforts raised the requirement for expanding the genome resources and provided a systematic insight into the oral virome.

In this study, we analyzed 2,792 publicly available human oral metagenomic samples covering 12 oral cavity sites and nine countries to build a complementary genome catalog of oral virome. The catalog was named oral virome database (OVD) including 48,425 nonredundant viruses (de-replicated from 121,054 viruses with >95% nucleotide similarity) with a majority never found in existing databases. Using OVD, we gained preliminary insights into the taxonomic and host ranges, spatial diversity, and geographic and individual heterogeneity of the human oral virome.

RESULTS

Construction of a global human oral virus catalog

To address the material for oral virome research, we collected a total of 2,792 publicly metagenomic data originating from the human mouth and established the largest oral metagenomic dataset so far, spanning 16 studies from the USA, China, 5 European countries (France, Germany, Italy, Luxembourg, and Spain), and two Oceanian countries (Australia and Fiji) (Figure 1A; Table S1). This dataset contained data of samples from 12 oral cavity sites, including dental plaque (895 samples), saliva (696 samples), tongue (439 samples), buccal mucosa (356 samples), and others (406 samples). These samples represented 9.3 Tb high-quality non-human metagenomic data after being processed with a unified pipeline, and further generated a total of 5.7 million long contigs (≥5 kb; total length 73.3 Gb) via metagenomic assembly for each sample (Table S2). Using an integrated homology- and feature-based pipeline (see STAR Methods), approximately 2.1% (n = 121,054) of the contigs were detected as highly credible viral sequences. These viruses were finally clustered at >95% nucleotide similarity (Gregory et al., 2019) to generate a nonredundant oral virome database (OVD) with 48,425 viral operational taxonomic units (vOTUs). The length of vOTUs ranged from 5,004 bp to 433,840 bp, with an average length of 29,800 bp and an N50 length of 37,664 bp (Figure S1). We evaluated the completeness and contamination of the OVD catalog using CheckV (Nayfach et al., 2020), revealing 7.0% complete, 15.5% high-, 21.0% medium-, and 56.4% low-completeness viruses, and 99.5% of all these vOTUs were low contaminated (Figure 1B). 22.6% (n = 10,931) of the vOTUs with estimated high completeness and low contamination were classified as high-quality viruses (Table S3).

An accumulation curve for vOTUs was not yet reaching a plateau, indicating that the oral virome had not been fully captured by OVD (Figure 1C). We clustered the vOTUs into 3,572 approximately genus-level groups and 529 approximately family-level groups based on their protein similarity and gene sharing proportion following a previously reported method (Nayfach et al., 2021). Accumulation curve analysis showed that the OVD catalog appears to be approaching an asymptote at the approximate genus and family ranks.

We compared the vOTUs of OVD with several existing virus catalogs, including three human gut virus catalogs (i.e., GVD, GPD, and MGV), the oral viruses from CHVD, and the available viral genomes in the RefSeq database. Viral quality estimation based on CheckV as well as three other tools suggested that the confidence of vOTUs in OVD was comparable with the other catalogs (Figure S2). Remarkably, 74.6% of the CHVD oral viruses were found in OVD, which accounted for 20.9% of all OVD vOTUs (Figure 1D). Conversely, despite that the gut virus catalogs were constructed from a huge number of fecal metagenomes and viromes and contained more nonredundant viruses than the OVD, only less than 11% of OVD viruses were overlapped with them, suggesting a considerable habitat specificity of the human-associated virome. Moreover, the RefSeq viruses rarely overlapped with both gut and oral catalogs. These findings yielded that approximately 70% (33,608/48,425) of oral vOTUs were novel and thus highlighted the considerable novelty of human oral virome.

Taxonomic landscape and host range of oral viruses

Referring to a gene-sharing pipeline (Bin Jang et al., 2019) with updated reference, 45.8% (22,194/48,425) of vOTUs of the OVD catalog were taxonomically assigned into the viral families. These vOTUs were dominated by three viral families that belonged to the *Caudovirales* order, including *Siphoviridae*, *Myoviridae*,

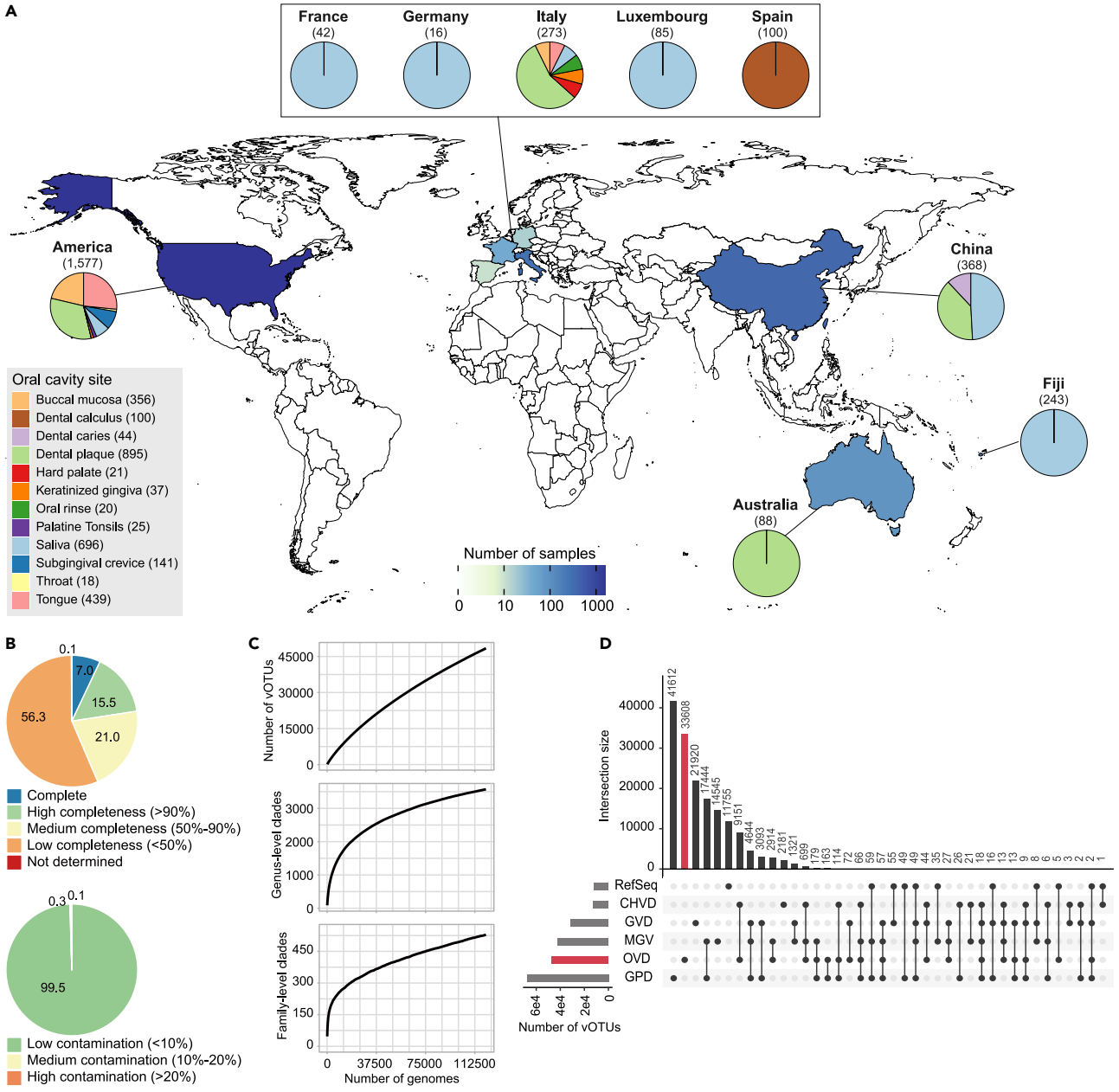


Figure 1. Overview of the oral virome database

(A) Map of the world showing the number of metagenomic samples per country and the distribution of oral cavity sites. Pie plots show the proportions of oral cavity sites for each country. Bracketed number indicates the number of metagenomic samples. Detailed information of all samples is provided in Table S1.

(B) CheckV estimation of completeness (upper panel) and contamination (bottom panel) of the OVD viruses.

(C) Accumulation curves for vOTUs (upper panel), approximately genus-level groups (middle panel), and approximately family-level groups (bottom panel) from the OVD catalog.

(D) UpSet plot shows the number of vOTUs shared by existing virome databases. CHVD, Cenote Human Virome Database (only oral viruses are included); GVD, Gut Virome Database; GPD, Gut Phage Database; MGV, Metagenomic Gut Virus.

and *Podoviridae* (Figure S3A), which were also the most dominant taxa in the human gut virome catalogs (Zuo et al., 2020). And members of the *crAss-like phages* (189 vOTUs), *Autographiviridae* (137 vOTUs), *Quimbyviridae* (120 vOTUs) (a putative viral family defined by Benler et al.'s study (Benler et al., 2021)), and *Microviridae* (23 vOTUs) frequently appeared in the remaining vOTUs. We performed host prediction of the vOTUs based on their homology of genome sequences (≥90% nucleotide identity and ≥30% viral

coverage) or CRISPR spacers (bit-score ≥45) to a large-scale oral prokaryotic genome collection (representing 3,569 species from over 50,000 metagenome-assembled genomes) (Zhu et al., 2021). This process predicted the prokaryotic hosts of 67.3% (32,570/48,425) of the vOTUs in OVD (Figure S3B). At the phylum level, the predicted hosts of oral viruses were dominated by *Firmicutes*, *Actinobacteriota*, *Proteobacteria*, and *Bacteroidota*, followed by *Fusobacteriota*, *Saccharibacteria*, and *Campylobacterota*, which were previously reported to be enriched in the human oral cavity (Human Microbiome Project, 2012; Zhu et al., 2021). Strikingly, 12.4% of the vOTUs were predicted to infect hosts belonged to more than one prokaryotic phylum, which was considerably larger than the proportion (6.7%) in gut viruses (Fisher's exact test, $p < 0.001$; Figure S3B), suggesting a relatively broad host range of most oral viruses.

According to genome-wide similarity at the protein level (Nishimura et al., 2017), we constructed a proteomic tree of 10,931 high-quality vOTUs, of which 56.3% could be assigned to a known viral family and 78.0% could be predicted to known prokaryotic hosts (Tables S3 and S4). The tree roughly showed that the viruses tend to cluster by both family-level taxonomies and potential host affiliations (Figure 2A), and a similar finding was also observed in multivariate analysis of the protein profiles of vOTUs (Figure S4). These findings suggested that the host adaptation had an important driving effect on genomic evolution of oral viruses, to a great extent in agreement with previous reports (Remold et al., 2008; Simmonds et al., 2019). The vOTUs of the largest family, *Siphoviridae*, were predicted to infect some of the most dominant bacteria in the human oral microbiota (Human Microbiome Project, 2012), including *Streptococcaceae* and *Actinomycetaceae* (Figures 2B and S5A). The *Myoviridae* viruses tended to infect *Proteobacteria* members, especially the *Pasteurellaceae* and *Neisseriaceae* species (Figure S5B), and the *Podoviridae* viruses were mostly grouped into a clade in the proteomic tree and part of them were predicted to infect the species of *Streptococcaceae* and *Actinomycetaceae*. For other viruses, the *Bacteroidaceae* species were dominant hosts of viruses belonging to *crAss-like viruses*, *Microviridae*, *Flandersviridae*, *Gratiaviridae*, and *Quimbyviridae*. Strikingly, we noticed that the vOTUs predicted to infect *Saccharibacteria* (formerly known as *TM7* which is ubiquitously distributed in the human oral cavity (Ferrari et al., 2014; He et al., 2015)) were mostly clustered in a single clade in the proteomic tree (Figure 2A). These predicted *Saccharibacteria* phages (n = 344; Table S5) were moderate in genome size (average 35.5 kbp) but highly prevalent in oral metagenomes across all samples (average prevalence rate, 25.9%) (Figures S6A and S6B). The taxa of these predicted *Saccharibacteria* phages were currently unknown, but the aforementioned clustering analysis could group them into two major family-level groups and seven major genus-level groups with remarkable stratification of bacterial hosts at the species level (Figure 3A). In particular, the seven genus-level groups showed large differences in their abundances (represented as the "abundance of reads", see STAR Methods) in samples from different countries and oral cavity sites (Figures S6C and S6D), highlighting a high geographic and spatial heterogeneity of them. Collectively, these results were important for understanding the host-virus interactions in the oral microbiome and will promote future studies in evolution and/or epidemiology scopes.

Recent studies have shown that the human oral virome is rich in jumbo viruses (Carr et al., 2020), a type of tailed bacteriophages with genome sizes more than 200 kbp. The OVD catalog contained 391 high-quality jumbo vOTUs with a high prevalence across all regions and oral sites (Figures S7A and S7B; Table S6). Similar to the *Saccharibacteria* phages, no jumbo vOTUs could be assigned into known viral families based on our current method/database. Instead, clustering analysis grouped the jumbo viruses into eight major approximately family-level groups that obviously differed in host predictions (Figures 3B and S8). Comparison of the annotatable functions revealed that, as expected, the jumbo viruses encoded a significantly higher proportion of auxiliary metabolic genes (AMGs) than the non-jumbo viruses (Fisher's exact test, $p = 0.04$; Figures 3C and S7C), suggesting a high metabolic potential of these viruses. Besides, 91.6% of jumbo vOTUs were lytic viruses, which is remarkably higher than that of non-jumbo viruses (Fisher's exact test, $p < 1e\text{-}10$; Figure 3D).

Spatial diversity, structure, and function of the oral virome

The oral cavity is a heterogeneous ecosystem containing distinct niches with remarkably different microbial communities (Xu et al., 2015). To explore the spatial specificity of oral virome, we undertook a compositional comparison of the vOTU profiles of four major oral cavity sites (i.e., buccal mucosa, dental plaque, saliva, and tongue), spanning a total of 1,909 deeply sequenced metagenomes (>5 million clean reads) for analysis. In terms of diversity, the salivary virome had the highest viral richness and diversity than other sites, followed by the tongue virome, buccal mucosa, and dental plaque (Wilcoxon rank-sum test, $p < 0.01$;

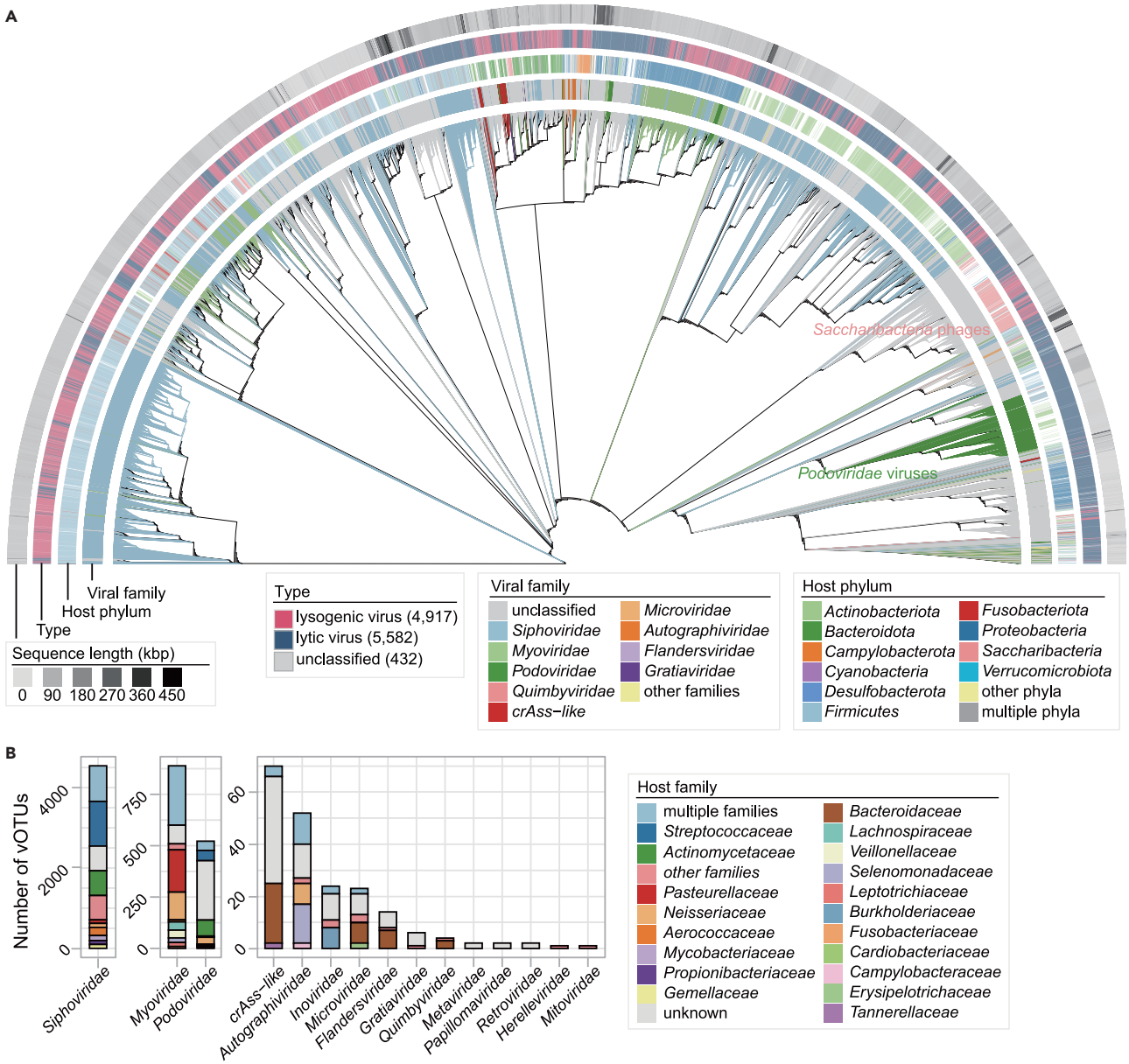


Figure 2. Phylogenomic analysis of high-quality vOTUs in OVD

(A) A proteomic tree of 10,931 high-quality vOTUs. The tree was generated using ViPTreeGen (Nishimura et al., 2017). Outer rings display metadata for each vOTU: innermost ring, viral family-level taxonomic assignments; ring 2, phylum-level host assignments; ring 3: lysogenic or lytic types; and outermost ring, the sequence length of the vOTUs. See Table S3 for details.

(B) Distribution of prokaryotic hosts of high-quality vOTUs. The vOTUs are grouped at the family level, and the host taxa are shown at the family level. The number of vOTUs that had more than one predicted host is labeled by light blue color.

Figure 4A). Analysis at the family level showed that several dominant viral families, such as *Siphoviridae* and *Myoviridae*, are universally present among all sites with similar distribution patterns (Figure S9), although almost all (20/21) families had statistically differences among four sites (Kruskal-Wallis rank-sum test, adjusted $p < 0.01$; Table S6). In addition, principal coordinates analysis (PCoA) based on the Bray-Curtis distance of vOTU profiles revealed a clear separation of viromes among different sites (Figure 4B).

Comparison at the vOTU level revealed that the majority of (57.7%) of 48,425 vOTUs were statistically enriched in the virome of one of the four oral cavity sites (Wilcoxon rank-sum test, adjusted $p < 0.01$; Table S8), suggesting

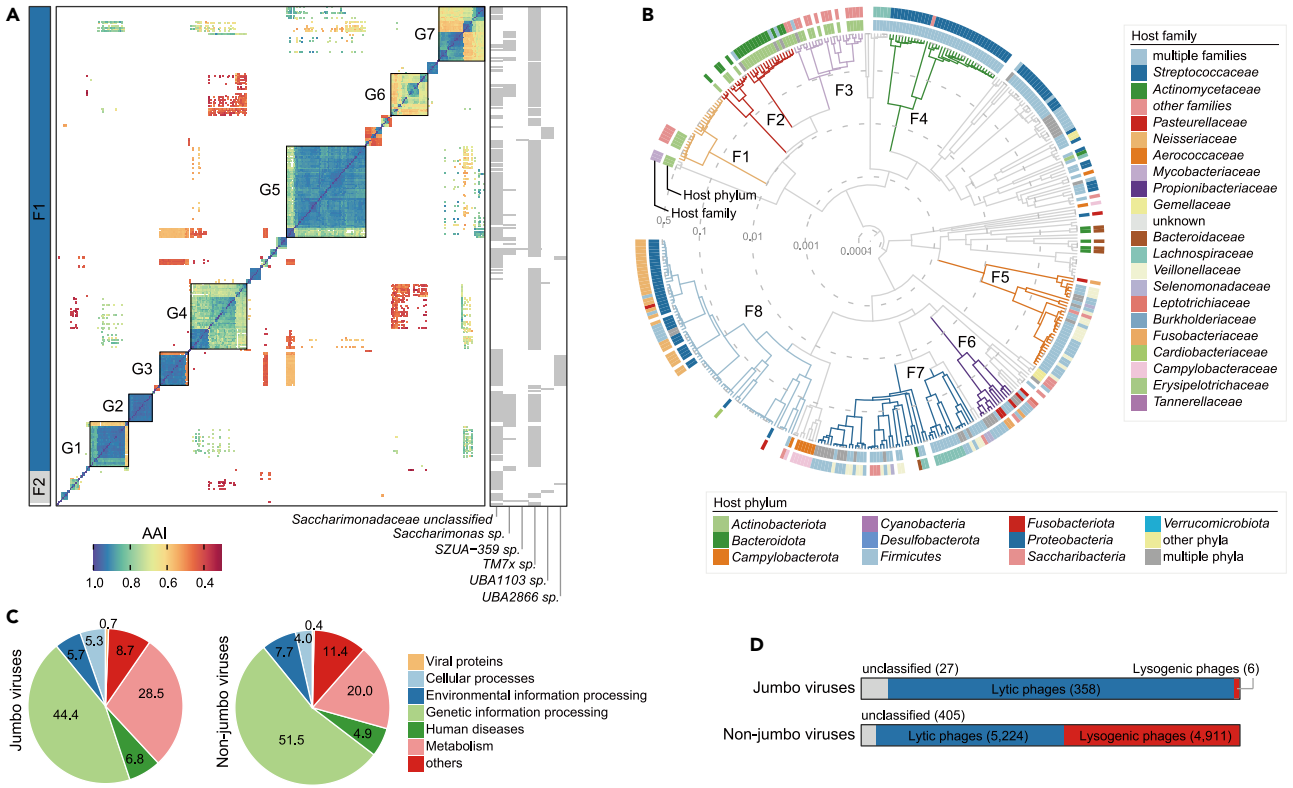


Figure 3. Phylogenomic analysis of the predicted *Saccharibacteria* phages and jumbo viruses

(A) Heatmap shows the pairwise proteomic similarly among 344 *Saccharibacteria* phages. F1–F2 and G1–G7 represent the approximately family-level and genus-level groups, respectively. Right panel represents the host predictions of vOTUs (gray tile: the phage was predicted to infect the corresponding species). AAI, average amino acid identity. See Table S5 for details.

(B) A proteomic tree of 391 jumbo viruses. The tree was generated using ViPTreeGen (Nishimura et al., 2017). F1–F8 represent the approximately family-level groups. Outer circles display the host assignments at the phylum and family levels. The branch lengths of the tree are labeled using logarithmic coordinate (showing by gray numbers). See also Figure S8.

(C) Pie plots show the functional distribution of Kyoto Encyclopedia of Genes and Genomes (KEGG)-annotated genes for jumbo (upper panel) and non-jumbo (bottom panel) viruses.

(D) The proportions of lysogenic and lytic types for jumbo and non-jumbo viruses. See Table S6 for details of subfigures (B–D).

strong niche specialization of viromes within individuals. The buccal mucosa-enriched vOTUs were more frequently concentrated in *Siphoviridae* compared with other sites, while the saliva-enriched vOTUs had more jumbo viruses, *crAss-like viruses*, and *Quimbyviridae* compared with other sites (Figure 4C). Of the enriched vOTUs with a predicted host, most of the buccal mucosa-enriched and tongue-enriched vOTUs were predicted to infect members of *Firmicutes*, while the dental plaque-enriched vOTUs had the highest proportions of predicted *Bacteroidota*, *Fusobacteriota*, and *Saccharibacteria* phages compared with other sites. Moreover, we compared the functional capacities of the site-associated vOTUs to elucidate the spatial specificity of oral virome functions. Although some basic viral functions (e.g., integrase, phage terminase, and ssDNA-binding protein) were ubiquitous across all oral cavity sites, there were numerous important functions that significantly differed in frequency among the enriched viruses of four sites (Figures 4D, S10; Table S9). For example, the saliva-enriched vOTUs had a higher frequency of two enzymes, dihydrofolate reductase (K00287) and thymidylate synthase (K00560), that were related to folate metabolism, the dental plaque-enriched vOTUs had a higher frequency for DNA (cytosine-5)-methyltransferase 1 (K00558, involving to DNA methylation) and phosphoadenosine phosphosulfate reductase (K00390, involving to sulfur metabolism), and two lysozymes (K01185 and K07273) were more frequent in the buccal mucosa-enriched vOTUs.

Geography and host properties describe the salivary and dental viromes

Finally, to investigate the variations of the oral virome in geography and host properties (i.e., gender, age, and body mass index [BMI]), we performed comparative analyses of vOTU profiles in two oral cavity sites

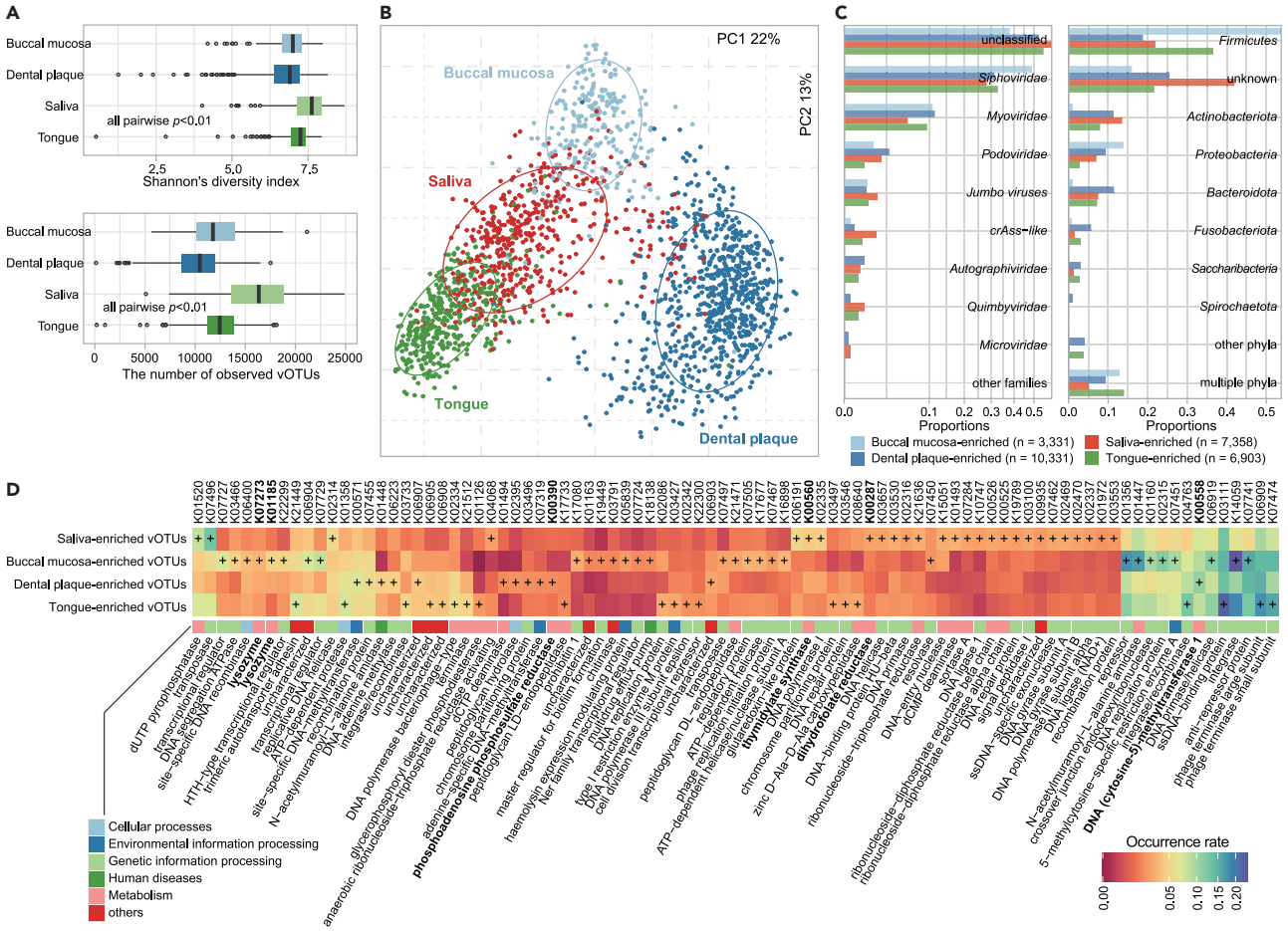


Figure 4. Comparison of viromes among four oral cavity sites

(A) Boxplots show the viral richness (the number of observed vOTUs, bottom panel) and diversity (Shannon's index, upper panel) of four sites. Wilcoxon rank-sum test was implemented between two groups.

(B) Principal coordinates analysis (PCoA) based on the Bray-Curtis distance of vOTU profiles. Samples are shown at the first and second principal coordinates (PC1 and PC2), and the ratio of variance contributed by these two PCs is shown. Ellipsoids represent an 80% confidence interval surrounding each group.

(C) Barplots show the proportions of site-enriched viruses grouping by their family-level taxa (left panel) and phylum-level host predictions (right panel). See Table S8 for details.

(D) Heatmap shows the occurrence rates of functions that exhibited as the top 50 functions in the vOTUs that significantly enriched in four oral sites. "+" represents the function with the highest occurrence rate in the corresponding vOTU groups. Occurrence rate represents the ratio of number of site-enriched viruses with corresponding KO to the total number of site-enriched viruses. Bold font shows several enzymes that are described as examples in Results. The comparison result of these functions is shown in Table S9.

with the largest sample size, saliva (containing 490 samples from six counties), and dental plaque (containing 772 samples from four counties). For the salivary virome, multivariate analyses revealed that three factors, including geography, age, and BMI, had considerably impacted the holistic viral profiles (permutational multivariate analysis of variance [PERMANOVA] $p < 0.001$ for all), whereas the effect size of gender was comparably slight (PERMANOVA $p = 0.023$; Figures 5A and 5C). The salivary viral compositions of American and European individuals were close in tendency and largely separated from that of Fiji individuals (Figures 5A and 5B). These geography-dependent variations may partly be explained by study bias but also potentially connected to the population differences in genetic background, lifestyle, or urbanization, as observed in the gut virome studies (Yan et al., 2021; Zuo et al., 2020). The viral richness and diversity significantly increase from children to adults and decrease from adults to elders, and the highest viral richness and diversity occurred at the ages between 20 and 30 years (Figure 5D). Several viral families such as *Myoviridae*, *Podoviridae*, and *Retroviridae* were enriched in the salivary virome of children compared with those of adults and elders, whereas the *crAss-like viruses* and *Quimbyviridae* were

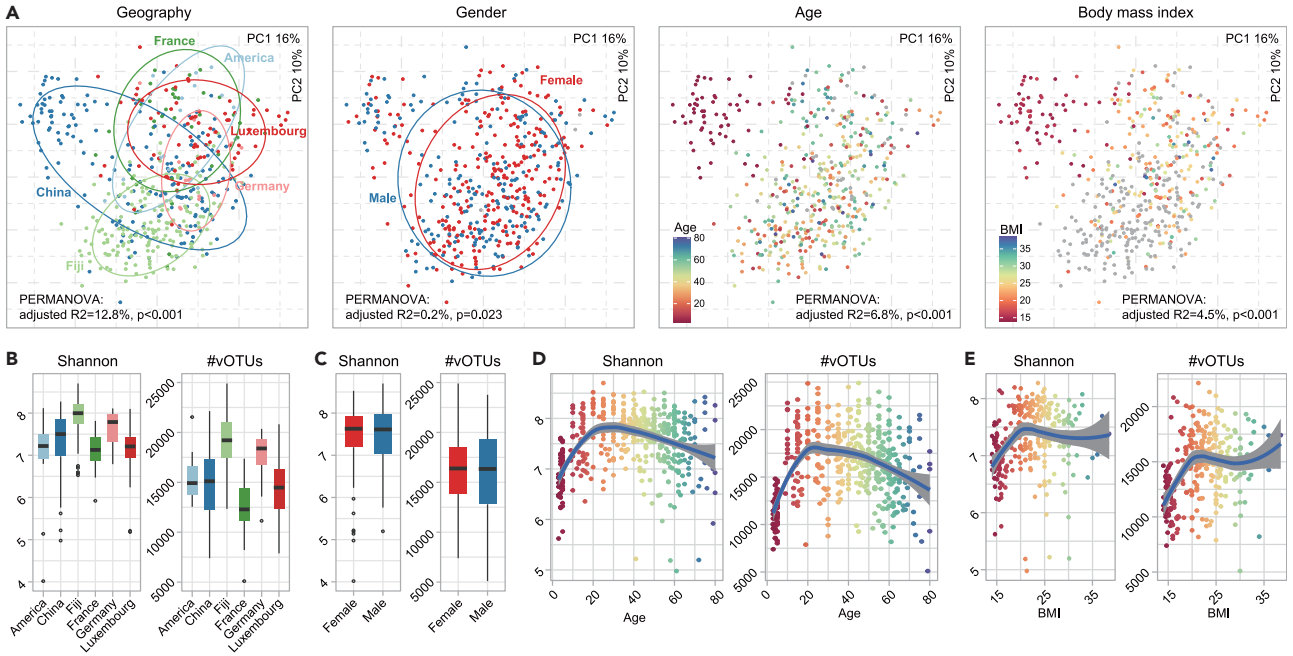


Figure 5. Alterations of the salivary virome across geography and host gender, age, and body mass index

(A) Principal coordinates analysis (PCoA) based on the Bray-Curtis distance of vOTU profiles of the salivary virome. Samples are shown at the first and second principal coordinates (PC1 and PC2), and the ratio of variance contributed by these two PCs is shown. Ellipsoids represent a 95% confidence interval surrounding each group. PERMANOVA, permutational multivariate ANOVA.

(B and C) Boxplots show the viral richness (the number of observed vOTUs) and diversity (Shannon's index) of the salivary samples grouped by their geography (B) and gender (C).

(D and E) Distribution of the viral richness and diversity of the salivary samples at different ages (D) and BMI (E). A smooth curve is formed based on the diversity index and the age/BMI of the samples using the *geom_smooth* function in the R platform. Points colored gray indicate the samples without available age/BMI data.

depleted in the children (Figure S11). Parallelly, we found that the viruses that infect *Fusobacteriota*, *Campylobacterota*, and *Euryarchaeota* decreased continuously with age, as well as the *Proteobacteria* phages in subjects under 40, whereas the *Firmicutes* phages increased continuously in subjects under 40 (Figure S12). Phages of *Actinobacteriota* and *Saccharibacteria* were enriched in adults aged 40 to 60. Besides, the individuals with BMI <18 appeared to have a markedly lower level of viral richness and diversity compared with the others (Figure 5E), but this phenomenon was not significant after adjusting for age and geography ($p = 0.09$).

For the dental plaque virome, multivariate analyses showed that the geography, gender, and age factors significantly impacted the variations of viral compositions (PERMANOVA $p < 0.001$ for geography and age and $p = 0.002$ for gender; Figure 6A). The viral richness and diversity of China and Australian individuals were remarkably higher than those of American and Italian individuals (Figure 6B), although the overall dental virome of American and Australian individuals was close. Females had a higher level of viral richness and diversity in their dental plaque than males, and middle-aged individuals had higher viral richness (not diversity) than the children and elders (Figures 6C and 6D). Besides, both viral composition and diversity seemed not to be correlated with the host's BMI (Figures 6A and 6E).

DISCUSSION

In this study, we reported a comprehensive catalog of 48,425 nonredundant viral sequences deriving from 2,792 public oral metagenomic samples across nine countries. The high coverage of the resulting viral clades at the approximate genus and family levels demonstrated the value of viral identification from the oral metagenomes. More importantly, nearly 70% of the viruses in the OVD catalog were newly found

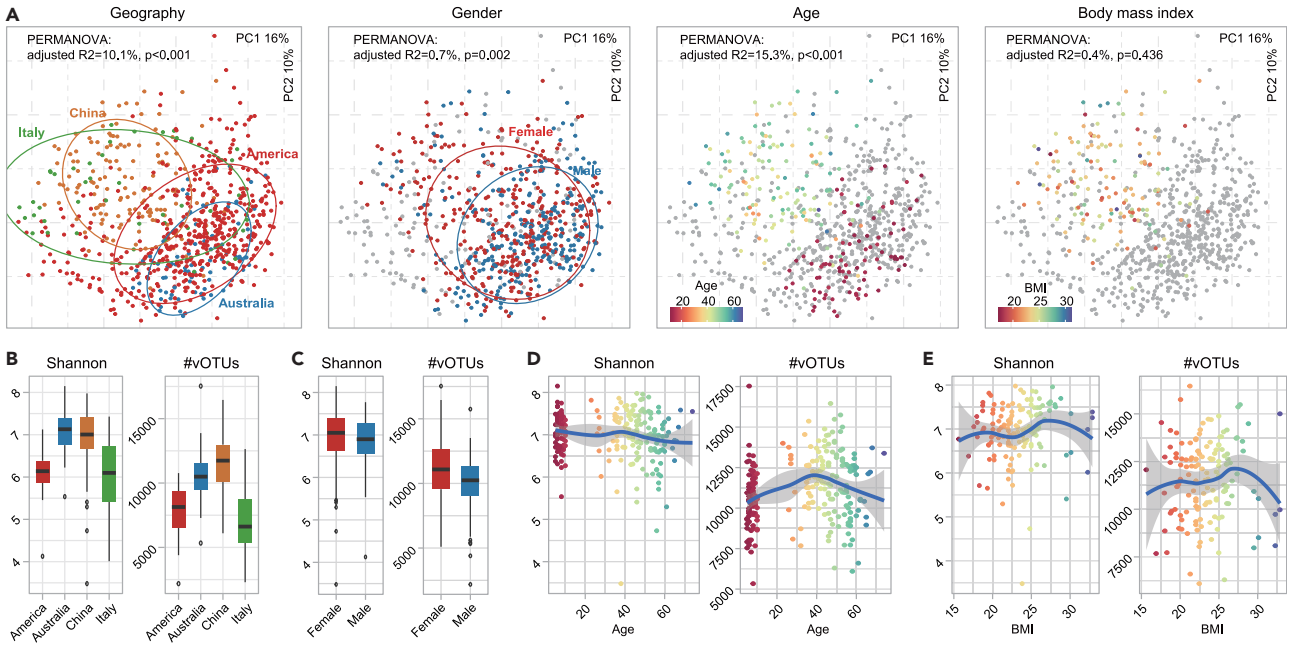


Figure 6. Alterations of the dental plaque virome across geography and host gender, age, and body mass index

(A) Principal coordinates analysis (PCoA) based on the Bray-Curtis distance of vOTU profiles of the dental plaque virome. Samples are shown at the first and second principal coordinates (PC1 and PC2), and the ratio of variance contributed by these two PCs is shown. Ellipsoids represent a 95% confidence interval surrounding each group. PERMANOVA, permutational multivariate ANOVA.

(B and C) Boxplots show the viral richness (the number of observed vOTUs) and diversity (Shannon's index) of the dental plaque samples grouped by their geography (B) and gender (C).

(D and E) Distribution of the viral richness and diversity of the dental plaque samples at different ages (D) and BMI (E). A smooth curve is formed based on the diversity index and the age/BMI of the samples using the *geom_smooth* function in the R platform. Points colored gray indicate the samples without available age/BMI data.

compared with the existing human virus catalogs or RefSeq viral genome reference, highlighting its value in exploring the previously uncharacterized "dark matter" of oral virus.

Taxonomic and host annotations of OVD described the phylogenetic view of the oral virome. At the viral family level, the known oral vOTUs were dominated by members of *Siphoviridae* and *Myoviridae*, and *Podoviridae*, highly similar with the composition of gut virome (Zuo et al., 2020). Several recently discovered viral taxa that are widespread in the human gut, such as *crAss-like viruses* (Edwards et al., 2019; Shkoporov et al., 2018) and *Quimbyviridae* (Benler et al., 2021), were also frequently present in OVD. In addition, OVD contained some important viral groups that are underrepresented in the gut microbiota, including the predicted phages of *Fusobacteriota* and *Saccharibacteria*. Fusobacteria are well established opportunistic pathogens of oral or colorectal cancers (Abed et al., 2020; Al-Hebshi et al., 2017; Harrandah et al., 2020; Kostic et al., 2012), and its phages might be also potential participators in the etiology of colorectal cancer (Nakatsu et al., 2018) (manuscript in preparation). *Saccharibacteria* are ubiquitous in oral microbiota and have been linked to multiple diseases (Bor et al., 2019; Kuehbacher et al., 2008), but their phages are uninvestigated. Here, we identified 176 high-quality vOTUs predicted to infect *Fusobacteriota* and 344 high-quality vOTUs predicted to infect *Saccharibacteria*, largely expanding the genome contents of these viruses. Besides, we identified 391 high-quality jumbo vOTUs from oral metagenomes and found that they were enriched in metabolism-associated genes and lytic viruses, probably linking to their nutrient availability conditions and pathogen elimination functions in the human body (Silveira and Rohwer, 2016).

We profiled the viral compositions of metagenomic samples and identified tremendous signatures in relation to spatial distribution, geographic, and individual heterogeneity of the oral virome; these results will be useful for experimental design (e.g., choosing of sampling sites) and mechanistic research in the future. Noticeably, we found that both salivary and dental plaque viromes were shaped by the individuals' age,

with rapidly increasing viral richness from child to adulthood and then slow decrease in old age. This age-dependent trajectory was similar to the observations in human gut virome (Gregory et al., 2020). Although the effect of virome dynamics needs further exploration, our results suggested that, in addition to the oral bacteria (Huang et al., 2020; Lira-Junior et al., 2018), the oral viruses may also correlate with maturation and aging of human beings.

In conclusion, the OVD catalog of over 48,000 viral genomes will largely improve further exploration of human oral virome both in depth and breadth.

Limitations of the study

A major limitation of the current OVD is the challenge of viral completeness (77.4% of viruses had <90% completeness) and the recovery of low-abundance viruses. State-of-the-art virome analysis strategies such as long-read metagenomic sequencing or virus-like particle (VLP) sequencing have been successfully used in uncovering the viral genome of fecal samples (Wang et al., 2021; Yahara et al., 2021) and are promising to obtain a more comprehensive landscape of the oral virome. More samples are still in need to address the representation of regions and oral cavity sites. Moreover, although we had provided preliminary information about the correlations between oral virome and host properties, the alteration of oral virome and its association with disease still need to be clarified.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human oral samples
- METHOD DETAILS
  - Preprocessing and assembly
  - Viral identification and decontamination
  - Viral clustering and gene calling
  - Viral taxonomy
  - Host prediction
  - Functional annotation
  - Phylogenetic analysis
  - The composition of human oral virome
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104418.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (81902037) and Dalian Science and Technology Innovation Fund (2020JJ27SN069).

AUTHOR CONTRIBUTIONS

S.L., R.G., and Q.Y. conceived and directed the study. Y.Z. performed data collection and investigation. S.L., R.G., Y.Z., F.C., Q.L., and G.W. carried out data processing and analyses. S.L. and R.G. drafted the manuscript. Z.J. and H.J. participated in design and coordination. P.L., X.W., and J.L. revised the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 7, 2021
Revised: March 26, 2022
Accepted: May 12, 2022
Published: June 17, 2022

REFERENCES

Abed, J., Maalouf, N., Manson, A.L., Earl, A.M., Parhi, L., Emgard, J.E.M., Klutstein, M., Tayeb, S., Almogy, G., Atlan, K.A., et al. (2020). Colon cancer-associated fusobacterium nucleatum may originate from the oral cavity and reach colon tumors via the circulatory system. Front. Cell. Infect. Microbiol. *10*, 400. https://doi.org/10.3389/fcimb.2020.00400.

Abeles, S.R., Robles-Sikisaka, R., Ly, M., Lum, A.G., Salzman, J., Boehm, T.K., and Pride, D.T. (2014). Human oral viruses are personal, persistent and gender-consistent. ISME J. *8*, 1753–1767. https://doi.org/10.1038/ismej.2014.31.

Al-Hebshi, N.N., Nasher, A.T., Maryoud, M.Y., Homeida, H.E., Chen, T., Idris, A.M., and Johnson, N.W. (2017). Inflammatory bacteriome featuring Fusobacterium nucleatum and Pseudomonas aeruginosa identified in association with oral squamous cell carcinoma. Sci. Rep. *7*, 1834. https://doi.org/10.1038/s41598-017-02079-3.

Altabtbaei, K., Maney, P., Ganesan, S.M., Dabdoub, S.M., Nagaraja, H.N., and Kumar, P.S. (2021). Anna Karenina and the subgingival microbiome associated with periodontitis. Microbiome *9*, 97. https://doi.org/10.1186/s40168-021-01056-3.

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A.B., Pevzner, P., and Koonin, E.V. (2021). Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome *9*, 78. https://doi.org/10.1186/s40168-021-01017-w.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. *37*, 632–639. https://doi.org/10.1038/s41587-019-0100-8.

Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats*8* (BMC Bioinformatics), pp. 1–8.

Bor, B., Bedree, J.K., Shi, W., McLean, J.S., and He, X. (2019). Saccharibacteria (TM7) in the human oral microbiome. J. Dent. Res. *98*, 500–509. https://doi.org/10.1177/0022034519831671.

Brito, I.L., Gurry, T., Zhao, S., Huang, K., Young, S.K., Shea, T.P., Naisilisili, W., Jenkins, A.P., Jupiter, S.D., Gevers, D., and Alm, E.J. (2019). Transmission of human-associated microbiota along family and social networks. Nat. Microbiol. *4*, 964–971. https://doi.org/10.1038/s41564-019-0409-6.

Brüssow, H., and Hendrix, R.W. (2002). Phage genomics: small is beautiful. Cell *108*, 13–16. https://doi.org/10.1016/s0092-8674(01)00637-7.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421–429. https://doi.org/10.1186/1471-2105-10-421.

Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D., and Lawley, T.D. (2021). Massive expansion of human gut bacteriophage diversity. Cell *184*, 1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.

Carr, V.R., Shkoporov, A.R., Gomez-Cabrero, D., Mullany, P., Hill, C., and Moyes, D.L. (2020). The human oral phageome is highly diverse and rich in jumbo phages. Preprint at bioRxiv. https://doi.org/10.1101/2020.07.06.186817.

Carrozzo, M. (2008). Oral diseases associated with hepatitis C virus infection. Part 2: lichen planus and other diseases. Oral Dis. *14*, 217–228. https://doi.org/10.1111/j.1601-0825.2007.01432.x.

Caselli, E., Fabbri, C., D'Accolti, M., Soffritti, I., Bassi, C., Mazzacane, S., and Franchi, M. (2020). Defining the oral microbiome by whole-genome sequencing and resistome analysis: the complexity of the healthy picture. BMC Microbiol. *20*, 120–219. https://doi.org/10.1186/s12866-020-01801-y.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Clooney, A.G., Sutton, T.D.S., Shkoporov, A.N., Holohan, R.K., Daly, K.M., O'Regan, O., Ryan, F.J., Draper, L.A., Plevy, S.E., Ross, R.P., and Hill, C. (2019). Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe *26*, 764–778.e5. https://doi.org/10.1016/j.chom.2019.10.009.

Dion, M.B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. Nat. Rev. Microbiol. *18*, 125–138. https://doi.org/10.1038/s41579-019-0311-5.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edwards, R.A., Vega, A.A., Norman, H.M., Ohaeri, M., Levi, K., Dinsdale, E.A., Cinek, O., Aziz, R.K., McNair, K., Barr, J.J., et al. (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat. Microbiol. *4*, 1727–1736. https://doi.org/10.1038/s41564-019-0494-6.

Ferrari, B., Winsley, T., Ji, M., and Neilan, B. (2014). Insights into the distribution and abundance of the ubiquitous candidatus Saccharibacteria phylum following tag pyrosequencing. Sci. Rep. *4*, 3957. https://doi.org/10.1038/srep03957.

Ganesan, S.M., Dabdoub, S.M., Nagaraja, H.N., Scott, M.L., Pamulapati, S., Berman, M.L., Shields, P.G., Wewers, M.E., and Kumar, P.S. (2020). Adverse effects of electronic cigarettes on the disease-naive oral microbiome. Sci. Adv. *6*, eaaz0108. https://doi.org/10.1126/sciadv.aaz0108.

Ghensi, P., Manghi, P., Zolfo, M., Armanini, F., Pasolli, E., Bolzan, M., Bertelle, A., Dell'Acqua, F., Dellasega, E., Waldner, R., et al. (2020). Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics. NPJ Biofilms Microbiomes *6*, 47. https://doi.org/10.1038/s41522-020-00155-7.

Goltsman, D.S.A., Sun, C.L., Proctor, D.M., DiGiulio, D.B., Robaczewska, A., Thomas, B.C., Shaw, G.M., Stevenson, D.K., Holmes, S.P., Banfield, J.F., and Relman, D.A. (2018). Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. Genome Res. *28*, 1467–1480. https://doi.org/10.1101/gr.236000.118.

Gomez, A., Espinoza, J.L., Harkins, D.M., Leong, P., Saffery, R., Bockmann, M., Torralba, M., Kuelbs, C., Kodukula, R., Inman, J., et al. (2017). Host genetic control of the oral microbiome in health and disease. Cell Host Microbe *22*, 269–278.e3. https://doi.org/10.1016/j.chom.2017.08.013.

Gregory, A.C., Zablocki, O., Zayed, A.A., Howell, A., Bolduc, B., and Sullivan, M.B. (2020). The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe *28*, 724–740.e8. https://doi.org/10.1016/j.chom.2020.08.003.

Gregory, A.C., Zayed, A.A., Conceicao-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., et al. (2019). Marine DNA viral macro- and microdiversity from Pole to Pole. Cell *177*, 1109–1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S., Draper, L.A., Gonzalez-Tortuero, E., Ross, R.P., and Hill, C. (2018). Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. Cell Host Microbe *24*, 653–664.e6. https://doi.org/10.1016/j.chom.2018.10.002.

Guidry, J.T., Birdwell, C.E., and Scott, R.S. (2018). Epstein–Barr virus in the pathogenesis of oral cancers. Oral Dis. *24*, 497–508. https://doi.org/10.1111/odi.12656.

Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitua, M.C., Vik, D., Sullivan, M.B., and Roux, S. (2021a). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome *9*, 37. https://doi.org/10.1186/s40168-020-00990-y.

Guo, R., Li, S., Zhang, Y., Zhang, Y., Wang, G., Ma, Y., and Yan, Q. (2021b). Dysbiotic oral and gut viromes in untreated and treated rheumatoid arthritis patients. Preprint at bioRxiv. https://doi.org/10.1101/2021.03.05.434018.

Harrandah, A.M., Chukkapalli, S.S., Bhattacharyya, I., Progulske-Fox, A., and Chan, E.K.L. (2020). Fusobacteria modulate oral carcinogenesis and promote cancer progression. J. Oral Microbiol. *13*, 1849493. https://doi.org/10.1080/20002297.2020.1849493.

He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.Y., Dorrestein, P.C., Esquenazi, E., Hunter, R.C., Cheng, G., et al. (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc. Natl. Acad. Sci. U S A *112*, 244–249. https://doi.org/10.1073/pnas.1419038112.

Heintz-Buschart, A., May, P., Laczny, C.C., Lebrun, L.A., Bellora, C., Krishna, A., Wampach, L., Schneider, J.G., Hogan, A., Beaufort, C.d., and Wilmes, P. (2016). Erratum: integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat. Microbiol. *2*, 16227–16313. https://doi.org/10.1038/nmicrobiol.2016.227.

Ho, S.X., Min, N., Wong, E.P.Y., Chong, C.Y., and Chu, J.J.H. (2021). Characterization of oral virome and microbiome revealed distinctive microbiome disruptions in paediatric patients with hand, foot and mouth disease. NPJ Biofilms Microbiomes *7*, 19. https://doi.org/10.1038/s41522-021-00190-y.

Huang, S., Haiminen, N., Carrieri, A.P., Hu, R., Jiang, L., Parida, L., Russell, B., Allaband, C., Zarrinpar, A., Vazquez-Baeza, Y., et al. (2020). Human skin, oral, and gut microbiomes predict chronological age. mSystems *5*, e00630-19. https://doi.org/10.1128/mSystems.00630-19.

Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. https://doi.org/10.1038/nature11234.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119–211. https://doi.org/10.1186/1471-2105-11-119.

Jakobsen, R.R., Haahr, T., Humaidan, P., Jensen, J.S., Kot, W.P., Castro-Mejia, J.L., Deng, L., Leser, T.D., and Nielsen, D.S. (2020). Characterization of the vaginal DNA virome in health and dysbiosis. Viruses *12*, 1143. https://doi.org/10.3390/v12101143.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457–D462. https://doi.org/10.1093/nar/gkv1070.

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 90. https://doi.org/10.1186/s40168-020-00867-0.

Kostic, A.D., Gevers, D., Pedamallu, C.S., Michaud, M., Duke, F., Earl, A.M., Ojesina, A.I., Jung, J., Bass, A.J., Tabernero, J., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. *22*, 292–298. https://doi.org/10.1101/gr.126573.111.

Kuehbacher, T., Rehman, A., Lepage, P., Hellmig, S., Folsch, U.R., Schreiber, S., and Ott, S.J. (2008). Intestinal TM7 bacterial phylogenies in active inflammatory bowel disease. J. Med. Microbiol. *57*, 1569–1576. https://doi.org/10.1099/jmm.0.47719-0.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Letunic, I., and Bork, P. (2021). Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296. https://doi.org/10.1093/nar/gkab301.

Lira-Junior, R., Akerman, S., Klinge, B., Bostrom, E.A., and Gustafsson, A. (2018). Salivary microbial profiles in relation to age, periodontal, and systemic diseases. PLoS One *13*, e0189374. https://doi.org/10.1371/journal.pone.0189374.

Low, S.J., Dzunkova, M., Chaumeil, P.A., Parks, D.H., and Hugenholtz, P. (2019). Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. Nature Microbiol. *4*, 1306–1315. https://doi.org/10.1038/s41564-019-0448-z.

Ly, M., Abeles, S.R., Boehm, T.K., Robles-Sikisaka, R., Naidu, M., Santiago-Rodriguez, T., and Pride, D.T. (2014). Altered oral viral ecology in association with periodontal disease. mBio *5*. e01133-01114. https://doi.org/10.1128/mBio.01133-14.

Manni, M., Berkeley, M.R., Seppey, M., Simao, F.A., and Zdobnov, E.M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Preprint at arXiv. 2106.11799.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H. (2016). Linking virus genomes with host taxonomy. Viruses *8*, 66. https://doi.org/10.3390/v8030066.

Nakatsu, G., Zhou, H., Wu, W.K.K., Wong, S.H., Coker, O.O., Dai, Z., Li, X., Szeto, C.H., Sugimura, N., Lam, T.Y.T., et al. (2018). Alterations in enteric virome are associated with colorectal cancer and survival outcomes. Gastroenterology *155*, 529–541.e5. https://doi.org/10.1053/j.gastro.2018.04.018.

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2020). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585. https://doi.org/10.1038/s41587-020-00774-7.

Nayfach, S., Paez-Espino, D., Call, L., Low, S.J., Sberro, H., Ivanova, N.N., Proal, A.D., Fischbach, M.A., Bhatt, A.S., Hugenholtz, P., and Kyrpides, N.C. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat. Microbiol. *6*, 960–970. https://doi.org/10.1038/s41564-021-00928-6.

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., and Goto, S. (2017). ViPTree: the viral proteomic tree server. Bioinformatics *33*, 2379–2380. https://doi.org/10.1093/bioinformatics/btx157.

Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C., Kambal, A., Monaco, C.L., Zhao, G., Fleshner, P., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell *160*, 447–460. https://doi.org/10.1016/j.cell.2015.01.002.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834. https://doi.org/10.1101/gr.213959.116.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature *536*, 425–430. https://doi.org/10.1038/nature19094.

Pride, D.T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R.A., 3rd, Loomer, P., Armitage, G.C., and Relman, D.A. (2012). Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. ISME J. *6*, 915–926. https://doi.org/10.1038/ismej.2011.169.

Remold, S.K., Rambaut, A., and Turner, P.E. (2008). Evolutionary genomics of host adaptation in vesicular stomatitis virus. Mol. Biol. Evol. *25*, 1138–1147. https://doi.org/10.1093/molbev/msn059.

Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., Poplin, R., and Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. Quant. Biol. *8*, 64–77. https://doi.org/10.1007/s40484-019-0187-4.

Shaiber, A., Willis, A.D., Delmont, T.O., Roux, S., Chen, L.-X., Schmid, A.C., Yousef, M., Watson, A.R., Lolans, K., Esen, Ö.C., et al. (2020). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol. *21*, 292–335. https://doi.org/10.1186/s13059-020-02195-w.

Shi, B., Lux, R., Klokkevold, P., Chang, M., Barnard, E., Haake, S., and Li, H. (2020). The subgingival microbiome associated with periodontitis in type 2 diabetes mellitus. ISME J. *14*, 519–530. https://doi.org/10.1038/s41396-019-0544-3.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A., McDonnell, S.A., Khokhlova, E.V., Draper, L.A., Forde, A., et al. (2019). The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe *26*, 527–541.e5. https://doi.org/10.1016/j.chom.2019.09.009.

Shkoporov, A.N., Khokhlova, E.V., Fitzgerald, C.B., Stockdale, S.R., Draper, L.A., Ross, R.P., and

Hill, C. (2018). ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat. Commun. *9*, 4781. https://doi.org/10.1038/s41467-018-07225-7.

Silveira, C.B., and Rohwer, F.L. (2016). Piggyback-the-Winner in host-associated microbial communities. NPJ Biofilms Microbiomes *2*, 16010. https://doi.org/10.1038/npjbiofilms.2016.10.

Simmonds, P., Aiewsakun, P., and Katzourakis, A. (2019). Prisoners of war - host adaptation and its constraints on virus evolution. Nat. Rev. Microbiol. *17*, 321–328. https://doi.org/10.1038/s41579-018-0120-2.

Soto-Perez, P., Bisanz, J.E., Berry, J.D., Lam, K.N., Bondy-Denomy, J., and Turnbaugh, P.J. (2019). CRISPR-Cas system of a prevalent human gut bacterium reveals hyper-targeting against phages in a human virome catalog. Cell Host Microbe *26*, 325–335.e5. https://doi.org/10.1016/j.chom.2019.08.008.

Tisza, M.J., and Buck, C.B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. Proc. Natl. Acad. Sci. U S A. *118*, e2023202118. https://doi.org/10.1073/pnas.2023202118.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804–810. https://doi.org/10.1038/nature06244.

Velsko, I.M., Fellows Yates, J.A., Aron, F., Hagan, R.W., Frantz, L.A.F., Loe, L., Martinez, J.B.R., Chaves, E., Gosden, C., Larson, G., and Warinner, C. (2019). Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. Microbiome *7*, 102–120. https://doi.org/10.1186/s40168-019-0717-3.

Wang, G., Li, S., Yan, Q., Guo, R., Zhang, Y., Chen, F., Tian, X., Lv, Q., Jin, H., and Ma, X. (2021). Optimization and Evaluation of Viral Metagenomic Amplification and Sequencing Methods toward a Genome-Level Resolution of the Human Fecal DNA Virome (Research Square). https://www.researchsquare.com/article/rs-1097721/latest.pdf.

Wang, Y., Wang, S., Wu, C., Chen, X., Duan, Z., Xu, Q., Jiang, W., Xu, L., Wang, T., Su, L., et al. (2019). Oral microbiome alterations associated with early childhood caries highlight the importance of carbohydrate metabolic activities. mSystems *4*, e00450-19. https://doi.org/10.1128/msystems.00450-19.

Wei, F., Sun, X., Gao, Y., Dou, H., Liu, Y., Su, L., Luo, H., Zhu, C., Zhang, Q., Tong, P., et al. (2021). Is oral microbiome of children able to maintain resistance and functional stability in response to short-term interference of ingesta? Protein Cell *12*, 502–510. https://doi.org/10.1007/s13238-020-00774-y.

Xu, X., He, J., Xue, J., Wang, Y., Li, K., Zhang, K., Guo, Q., Liu, X., Zhou, Y., Cheng, L., et al. (2015). Oral cavity contains distinct niches with dynamic microbial communities. Environ. Microbiol. *17*, 699–710. https://doi.org/10.1111/1462-2920.12502.

Yahara, K., Suzuki, M., Hirabayashi, A., Suda, W., Hattori, M., Suzuki, Y., and Okazaki, Y. (2021). Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. Nat. Commun. *12*, 27. https://doi.org/10.1038/s41467-020-20199-9.

Yan, Q., Wang, Y., Chen, X., Jin, H., Wang, G., Guan, K., Zhang, Y., Zhang, P., Ayaz, T., Liang, Y., et al. (2021). Characterization of the gut DNA and RNA viromes in a cohort of Chinese residents and visiting Pakistanis. Virus Evol. *7*, veab022. https://doi.org/10.1093/ve/veab022.

Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. *10*, 766. https://doi.org/10.15252/msb.20145645.

Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., Wu, X., Li, J., Tang, L., Li, Y., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat. Med. *21*, 895–905. https://doi.org/10.1038/nm.3914.

Zhu, J., Tian, L., Chen, P., Han, M., Song, L., Tong, X., Sun, X., Yang, F., Lin, Z., Liu, X., et al. (2021). Over 50,000 metagenomically assembled draft genomes for the human oral microbiome reveal new taxa. Dev. Reprod. Biol. S1672-S0229(21)00176-185. https://doi.org/10.1016/j.gpb.2021.05.001.

Zuo, T., Sun, Y., Wan, Y., Yeoh, Y.K., Zhang, F., Cheung, C.P., Chen, N., Luo, J., Wang, W., Sung, J.J.Y., et al. (2020). Human-gut-DNA virome variations across geography, ethnicity, and urbanization. Cell Host Microbe *28*, 741–751.e4. https://doi.org/10.1016/j.chom.2020.08.005.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
OVD database	This paper	https://github.com/RChGO/OVD
Scripts about virus identification	This paper	https://github.com/RChGO/virusDectect
CHVD database	(Tisza and Buck, 2021)	https://zenodo.org/record/4498884
GVD database	(Gregory et al., 2020)	https://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/Gregory_and_Zablocki_GVD_Jul2020/GVD_Viral_Populations
GPD database	(Camarillo-Guerrero et al., 2021)	http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/gut_phage_database
MGV database	(Nayfach et al., 2021)	https://portal.nersc.gov/MGV
NCBI RefSeq (virus)	https://ftp.ncbi.nlm.nih.gov/refseq/release/viral	https://ftp.ncbi.nlm.nih.gov/refseq/release/viral
Virus-Host DB	(Mihara et al., 2016)	ftp://ftp.genome.jp/pub/db/virushostdb/virushostdb.cds.faa.gz
crAss-phages database	(Guerin et al., 2018)	https://ars.els-cdn.com/content/image/1-s2.0-S1931312818305249-mmc6.zip
Viral proteins from Benler's study	(Benler et al., 2021)	ftp://ftp.ncbi.nih.gov/pub/yutinn/benler_2020/gut_phages
human genome GRCh38	NCBI	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38
BUSCO (bacteria)	(Manni et al., 2021)	https://busco-data.ezlab.org/v4/data/lineages/bacteria_odb10.2020-03-06.tar.gz
Oral SGBs database	(Zhu et al., 2021)	https://ftp.cngb.org/pub/SciRAID/Microbiome/human_oral_genomes/bowtie2_index
KEGG database	(Kanehisa et al., 2016)	https://www.kegg.jp/
Altabtbaei et al., 2021 sequencing reads	(Altabtbaei et al., 2021)	NCBI BioSample -see Table S2 for details
Brito et al., 2019 sequencing reads	(Brito et al., 2019)	NCBI BioSample -see Table S2 for details
Caselli et al., 2020 sequencing reads	(Caselli et al., 2020)	NCBI BioSample -see Table S2 for details
Ganesan et al., 2020 sequencing reads	(Ganesan et al., 2020)	NCBI BioSample -see Table S2 for details
Gomez et al., 2017 sequencing reads	(Gomez et al., 2017)	NCBI BioSample -see Table S2 for details
Ghensi et al., 2020 sequencing reads	(Ghensi et al., 2020)	NCBI BioSample -see Table S2 for details
Goltsman et al., 2018 sequencing reads	(Goltsman et al., 2018)	NCBI BioSample -see Table S2 for details
Heintz-Buschart et al., 2016 sequencing reads	(Heintz-Buschart et al., 2016)	NCBI BioSample -see Table S2 for details
HMP, sequencing reads	(Turnbaugh et al., 2007)	NCBI BioSample -see Table S2 for details
Shaiber et al., 2020 sequencing reads	(Shaiber et al., 2020)	NCBI BioSample -see Table S2 for details
Shi et al., 2020 sequencing reads	(Shi et al., 2020)	NCBI BioSample -see Table S2 for details
Velsko et al., 2019 sequencing reads	(Velsko et al., 2019)	NCBI BioSample -see Table S2 for details
Wang et al., 2019 sequencing reads	(Wang et al., 2019)	NCBI BioSample -see Table S2 for details
Wei et al., 2021 sequencing reads	(Wei et al., 2021)	NCBI BioSample -see Table S2 for details
Zeller et al., 2014 sequencing reads	(Zeller et al., 2014)	NCBI BioSample -see Table S2 for details
Zhang et al., 2015 sequencing reads	(Zhang et al., 2015)	NCBI BioSample -see Table S2 for details
Software and algorithms		
fastp v0.20.1	(Chen et al., 2018)	http://opengene.org/fastp/fastp.0.20.1
bowtie2 v2.4.1	(Langmead and Salzberg, 2012)	https://github.com/BenLangmead/bowtie2

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SPAdes v3.14.1	(Nurk et al., 2017)	https://github.com/ablab/spades
CheckV v0.7.0	(Nayfach et al., 2020)	https://bitbucket.org/berkeleylab/checkv
DeepVirFinder v1.0	(Ren et al., 2020)	https://github.com/jessieren/DeepVirFinder
VIBRANT v1.2.1	(Kieft et al., 2020)	https://github.com/AnantharamanLab/VIBRANT
VirSorter2 v2.2.2	(Guo et al., 2021a)	https://bitbucket.org/MAVERICLab/virsorter2/src/master/
hmmsearch v3.3.1	(Eddy, 2011)	http://hmmer.org/
BLAST v2.9.0	(Camacho et al., 2009)	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
Prodigal v2.6.3	(Hyatt et al., 2010)	https://github.com/hyattpd/Prodigal
DIAMOND v2.0.6.144	(Buchfink et al., 2015)	https://github.com/bbuchfink/diamond
MinCED v0.4.2	(Bland et al., 2007)	https://github.com/ctSkennerton/minced
ViPTreeGen v1.1.2	(Nishimura et al., 2017)	https://github.com/yosuken/ViPTreeGen
iTOL v6.3.2	(Letunic and Bork, 2021)	https://itol.embl.de/
R v4.0.3	https://www.r-project.org/	https://www.r-project.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Shenghui Li (lsh2@qq.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The viral sequences and annotation files of OVD have been deposited in the GitHub website with URL: https://github.com/RChGO/OVD/.
- The original codes used in the paper are provided in the following link: https://github.com/RChGO/virusDectect.
- The metadata of all 2,792 oral metagenomic samples, including NCBI BioSample ID, oral cavity site, host properties (geography, gender, age, and BMI), and data and assembly information, are available in Table S2. All other data reported in this paper will be shared by the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human oral samples

We reviewed multiple studies based on whole-metagenome sequencing of oral samples and collected a total of 2,792 oral metagenomic samples publicly available in the NCBI (https://www.ncbi.nlm.nih.gov/) database until May 2021. The metadata of all selected samples were summarized and listed in Tables S1 and S2.

METHOD DETAILS

Preprocessing and assembly

Raw reads were qualified via fastp v0.20.1 (Chen et al., 2018) with the options '-u 30 -q 20 -l 90 -y –trim_poly_g', and human reads were further removed by matching quality-filtered reads against the human genome GRCh38 with bowtie2 v2.4.1 (Langmead and Salzberg, 2012). The remaining clean reads of each sample were assembled into contigs using SPAdes v3.14.1 (Nurk et al., 2017) with the options '-meta -k 21,33,55' (for samples with read length <100 bp or less) or '-meta -k 21,33,55,77' (for samples with read length >100 bp). Total 5,670,086 contigs with ≥5 kbp sequence length (average length 12,934 bp, total length 73.3 Gbp) were used to further analyze.

Viral identification and decontamination

All assembled contigs (n = 5,670,086, length ≥5 kbp) were firstly assessed by CheckV v0.7.0 (Nayfach et al., 2020), and then contigs would be removed if their microbial genes are no less than 10 and five times the number of viral genes, resulting in a total of 310,627 contigs for further analyses. We identified potential viral sequences from the remaining contigs based on any of the following criteria: 1) contig whose viral genes was more than the number of microbial genes in CheckV (contigs with length <10 kbp and estimated as low-quality or undetermined contigs were removed); 2) contig with p-value <0.01 and score >0.90 in DeepVirFinder v1.0; 3) contig identified by VIBRANT v1.2.1 (Kieft et al., 2020) with default options. A total of 127,298 contigs were recognized as proposed viral sequences, including 107,424 identified by CheckV, 40,796 identified by DeepVirFinder, and 88,620 identified by VIBRANT. To decontaminate the viral sequences, according to the previous study (Gregory et al., 2020), we searched the bacterial universal single-copy orthologs (BUSCO) (Manni et al., 2021) within viral sequence using hmmsearch (Eddy, 2011) with default options and calculated the ratio of the number of BUSCO to the total number of genes in each viral sequence (BUSCO ratio). Then we removed high-contaminated viral sequences with ≥5% BUSCO ratio, and the remaining contigs (n = 121,054) were considered as the final viral sequences for each sample.

Viral clustering and gene calling

The viral sequences were de-replicated based on the following steps: 1) all viral sequences (n = 121,054) were aligned in pairs using BLASTn v2.9.0 (Camacho et al., 2009) with the options '-evalue 1e-10 -word_size 20 -num_alignments 99999'. 2) viral sequences which shared 95% nucleotide identity across 75% of the sequence were clustered into a viral operational taxonomic unit (vOTU) using the custom scripts (https://github.com/RChGO/virusDectect). 3) For each vOTU, the longest viral sequence was considered as the representative sequence and used for further analyses. Total 48,425 vOTU sequences clustered from all oral samples were integrated into the oral virome database (OVD). In addition, the shared vOTUs between different virome databases were identified based on the same steps as above.

We further clustered 48,425 vOTUs of OVD into approximately genus-level and family-level groups respectively. Firstly, totaling 1,846,359 putative protein sequences in vOTUs were called using Prodigal v2.6.3 (Hyatt et al., 2010) with options '-p meta', and pairwise protein sequence alignments were performed using DIAMOND v2.0.6.144 (Buchfink et al., 2015) with the options '-e 1e-5 –max-target-seqs 99999'. Then we calculated the percentage of shared gene and average amino acid identity (AAI) between each pair of vOTUs. According to the previous study (Nayfach et al., 2021), at the family level, we kept the connections between vOTUs with >20% AAI and >10% genes shared. At the genus level, we kept the connections between vOTUs with >50% AAI and >20% genes shared. Finally, clustering was performed based on connections between vOTUs using MCL with the option '-I 1.2' for the family level or '-I 2' for the genus level. All vOTUs in OVD represented approximately 529 family- and 3,572 genus-level groups.

Viral taxonomy

Taxonomic annotation of viral sequences was performed based on protein sequence alignment to the combined database derived from Virus-Host DB (Mihara et al., 2016) downloaded in May 2021, crAss-like protein sequences from Guerin's study (Guerin et al., 2018) and viral protein sequences from Benler's study (Benler et al., 2021). To implement accurate family-level taxonomic classification, we firstly aligned proteins of viral sequences from NCBI RefSeq against the combined database using DIAMOND with the options '–query-cover 50 –subject-cover 50 –id 30 –min-score 50 –max-target-seqs 10'. A viral sequence was annotated to the viral family level when over a quarter of its proteins were matched to the same family.

Host prediction

The oral microbial genome catalogue was derived from 3,569 prokaryotic species (containing over 50,000 metagenome-assembled genomes) from a previous study (Zhu et al., 2021). Based on the catalogue, the virus-host prediction was performed using two bioinformatic methods that included CRISPR-spacer matches and prophage blasts. For CRISPR-spacer matches, we firstly predicted CRISPR spacer sequences from the oral prokaryotic genome catalogue using MinCED v0.4.2 (Bland et al., 2007) with the option '-minNR 2', and then assigned a host to the virus if host CRISPR spacer sequence was

matched to the viral genome (bit-score ≥45) using BLASTn with options '-evalue 1e-5 -word_size 8 -num_alignments 99999' (Gregory et al., 2020). For prophage blasts, the viral sequence was blasted against host genome sequences, and assigned a host if the viral sequence was exactly matched to the host genome at ≥90% nucleotide identity and ≥30% viral coverage (Gregory et al., 2020). For 48,425 vOTUs in OVD, 57.0% (27,622/48,425) could be predicted to hosts by CRISPR-spacer matches, and 32.5% (15,737/48,425) could be predicted by prophage blasts, while 22.3% (10,789/48,425) were predicted by both two approaches.

Functional annotation

Functional annotation of viral proteins was performed based on the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (downloaded in December 2020) (Kanehisa et al., 2016) using DIAMOND with the options '–query-cover 50 –subject-cover 50 -e 1e-5 –min-score 50 –max-target-seqs 50'. Each protein was assigned a KEGG orthologue (KO) on the basis of the best-hit gene in the database.

Prediction of viral lifestyle (i.e., lytic or lysogenic phage) was implemented based on VIBRANT v1.2.1 (Kieft et al., 2020).

Phylogenetic analysis

Phylogenetic analysis was implemented based on high-quality vOTUs with ≥90% completeness and <10% contamination assessed by the CheckV algorithm. Firstly, by referring to the phylogenetic approach described by Low's study (Low et al., 2019), we used single-copy protein markers to try to construct a phylogenetic tree based on 4,019 *Caudovirales* genomes from NCBI RefSeq downloaded in January 2021. Single-copy protein markers were defined using the script provided by Nayfach's study (https://github.com/snayfach/MGV/blob/master/marker_gene_tree/marker_tree.py) (Nayfach et al., 2021). However, there were only 4 single markers with ≥10% prevalence among *Caudovirales* viruses, which displayed the very limited phylogenetic signal. A similar situation occurred in phylogenetic analysis of high-quality vOTUs in OVD (3 single markers), which made it difficult to construct a phylogenetic tree using single-copy protein markers of vOTUs. To perform phylogenetic analysis, we used another phylogenetic approach that was based on genome-wide similarities. We generated a viral proteomic tree of high-quality vOTUs in OVD using ViPTreeGen v1.1.2 (Nishimura et al., 2017) that provided the appropriate choice for understanding previously unknown viral genomes. The proteomic tree was then visualized using iTOL v6.3.2 (Letunic and Bork, 2021).

The composition of human oral virome

To explore the spatial specificity of oral virome, all vOTUs in OVD were profiled in oral metagenomes from 4 major oral cavity sites (i.e., buccal mucosa, dental plaque, saliva, and tongue), spanning a total of 1,909 deeply sequenced metagenomes (>5 million clean reads; Table S2). Clean reads in each metagenomic sample were mapped to vOTUs in OVD using bowtie2 with the options '–end-to-end –fast –no-head –no-unal -u 5000000 –no-sq'. The abundance profile of vOTUs in each sample was generated by aggregating the number of reads mapped to each vOTU, resulting an "abundance of reads" for each vOTU. The relative abundance of vOTUs was its abundance divided by the number of total mapped reads in each sample. The relative abundance profile at the family level was generated by aggregating the relative abundance of vOTUs assigned to the same family.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed in R v4.0.3. Alpha diversities were estimated based on the relative abundance of vOTUs: 1) Shannon diversity index was calculated using the function *diversity* with the argument index = shannon; 2) The number of observed vOTUs was counts of unique vOTUs in each sample. Principal coordinates analysis (PCoA) was performed based on the Bray-Curtis distance of vOTU profiles using the function *pcoa*. Permutational multivariate analysis of variance (PERMANOVA) was performed using the function *adonis*, and ADONIS R^2 was adjusted using the function *RsquareAdj*. When using PERMANOVA, each host property was analyzed after adjusting the other properties. Statistical significance was verified using the function *wilcox.test* and *kruskal.test*. P-values were adjusted using the function *p.adjust* with the argument method = BH, and an adjusted P-value < 0.01 was considered statistical significance.

Data visualizations except for proteomic tree were also carried out using the function *ggplot2* in R. The map of the world was created using the function *geom_polygon* that could load map data provided by the function *map_data*. For the analyses about the effect of individual age and BMI on viral diversities and abundance, regression line in scatter plots was added using the function *geom_smooth* with the argument method = loess.