Long-read shotgun metagenome sequencing using PromethION uncovers novel bacteriophages, their abundance, and interaction with host bacterial immunity in the oral microbiota

Koji Yahara^{1*}, Masato Suzuki¹, Aki Hirabayashi¹, Yutaka Suzuki², Yusuke Okazaki³

¹ Antimicrobial Resistance Research Center, National Institute of Infectious Diseases, Tokyo, Japan; ² Laboratory of Systems Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo.; ³ Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

*Correspondence to: Koji Yahara, Antimicrobial Resistance Research Center, National Institute of Infectious Diseases, Aoba-cho 4-2-1, Higashimurayama-shi, Tokyo, 189-0002 Japan. Tel: +81-42-202-6080. Email: k-yahara@nih.go.jp

16

17 **Abstract**

18

19 Bacteriophages (phages), or bacterial viruses, are very diverse and highly abundant worldwide,
20 including human microbiomes. Although a few metagenomic studies have focused on oral
21 phages, they relied on short-read sequencing. Here, we conducted a long-read metagenomic
22 study of human saliva for the first time using PromethION that requires a smaller amount of
23 DNA than PacBio. Our analyses, which integrated both PromethION and HiSeq data of >30 Gb
24 per sample, revealed N50 ranging from 187-345 kb and thousands of contigs with >1 kb
25 accounting for > 99% of all contigs on which 94-96% of HiSeq reads were mapped. We
26 identified hundreds of viral contigs (95 phages and 333 prophages on an average per sample);
27 0-43.8% and 12.5-56.3% of the “most confident” phages and prophages, respectively, didn’t
28 cluster with those reported previously and were identified as novel. Our integrated analyses
29 identified highly abundant oral phages/prophages, including a novel *Streptococcus* phage
30 cluster and nine jumbo phages/prophages. Interestingly, 86% of the phage cluster and 67% of
31 the jumbo phages/prophages contained remote homologs of antimicrobial resistance genes,
32 suggesting their potential role as a source of recombination to generate new resistance genes.
33 Pan-genome analysis of the phages/prophages revealed remarkable diversity, identifying 0.3%
34 and 86.4% of the genes as core and singletons, respectively. Functional annotation revealed
35 that the highest fraction of the core genes was enriched in phage morphogenesis, followed by
36 the fraction enriched in host cellular processes. Furthermore, our study suggested that oral
37 phages present in human saliva are under selective pressure for escaping CRISPR immunity.

38 (250/250 words)

39

40 **Importance**

41

42 Despite the abundance and grave implications oral bacterial viruses in health and disease, little
43 is known regarding the different groups of oral bacterial viruses, their relative abundances
44 under various conditions, and their activities. We provided answers to these questions for the

first time utilizing a recently developed sequencer that can capture and sequence long DNA fragments, including viruses, and requires only a small amount of DNA input, making it suitable for analyzing human oral samples. We identified hundreds of viral sequences, including “jumbo” viruses and a distinctive group of highly abundant oral viruses, which often contained parts of antimicrobial resistance genes; the entire repertoire of these viral genes showed remarkable diversity and supported a recently proposed hypothesis that phages modulate oral microbiota through multiple mechanisms. We also revealed genomic signs of coevolution of viruses and host bacteria that have been missed in large viromic studies in humans.

(147/150 words)

Keywords: metagenome, microbiome, virome, oral, phage, coevolution, PromethION, long-read, CRISPR

58 Introduction

59 Human microbiomes are of enormous interest to researchers (1, 2) and have been model
60 systems for studying polymicrobial communities (3, 4). Interspecies networks within the
61 microbiome can modulate energy metabolism pathways and affect human health. The two
62 main human microbiomes are intestinal and oral microbiota, which harbor hundreds of
63 coexisting species, including bacteria and viruses. Among them, bacteriophages (phages), or
64 bacterial viruses, which are the most abundant and diverse biological entities in the planet, are
65 present in both intestinal and oral microbiota with implications in health and disease; however,
66 there have been few recent studies on their role in the human microbiome(3, 5, 6).

67 With the rise of next-generation metagenome sequencing technologies, human gut virome
68 studies have increased rapidly (7). For example, there have been more attempts to characterize
69 “healthy gut phageome” since a study in 2016 conducted deep-sequencing of DNA from
70 virus-like particles (VLP) and revealed the presence of completely assembled phage genomes
71 in 64 healthy individuals around the world (8). Attempts have also been made to explore the
72 associations between human gut virome alterations and diseases (7, 9, 10). However, only a
73 small number of metagenomic studies focused on oral phage communities (5), including a
74 study in 2014 that reported an alteration of virome composition in subjects with periodontal
75 disease (11). The largest oral virome study was conducted in 2015, which generated and
76 analyzed more than 100 Gb shotgun sequencing data from 25 samples (20 dental plaque
77 specimens and 5 salivary) to primarily explore the phage-bacteria interaction network (12).
78 More recently, shotgun metagenome sequencing of 3,042 samples from various environments,
79 including the human oral cavity, was conducted in a project aimed at uncovering the Earth’s
80 virome (13), which achieved an almost 3-fold increase in the metagenome samples (14).
81 Re-analysis of the Earth’s virome data revealed signatures of genetic conflict invoked by the
82 coevolution of phages and host oral bacteria enriched in the human oral cavity (15), suggesting
83 it as an attractive system to study coevolution using metagenomic data. Such metagenomic

signs of coevolution have been missed in previous large viromic studies in humans (3).

These previous studies, however, were based on short-read sequencing data generated using Illumina sequencer. However, using contigs assembled from short-reads as short as 2 kb, a popular phage search tool could only identify up to 15% of phage regions, while its sensitivity increases to ~90% in contigs assembled from reads > 20 kb (16). Therefore, in principle, it is much more desirable to use long-read sequencing data to explore phages and their relationships with host bacteria in metagenomic analyses. Recently, a pioneering study used PacBio sequencing and conducted a long-read metagenomic exploration of extrachromosomal mobile genetic elements, such as phages in the human gut, and generated contigs of 71 plasmids and 11 phages including complete genomes of 5 diverse crAssphages from 12 human fecal samples (17). PacBio sequencing, however, requires relatively high DNA input (~5µg for standard library protocol) (18), which is usually difficult to obtain from standard amounts of human oral specimens; our preliminary experiment showed that the average amount of DNA extracted from 1 mL saliva of healthy individuals is 2.4 µg, which is smaller than the requirement of PacBio.

In this study, we conducted a long-read shotgun metagenomic study using PromethION, a recently developed high-throughput nanopore sequencer that requires a smaller amount of DNA than PacBio and is thus, revolutionary for oral microbiota research. We also used Illumina HiSeq for sequencing the same samples, enabling error correction and abundance quantification of contigs assembled from the long-reads. Our analyses integrated the deep-sequencing data of PromethION and HiSeq and uncovered hundreds of metagenome-assembled viral genomes, including a highly abundant phage cluster and jumbo phages/prophages (with > 200 kb); the characteristics of these genes and the metagenomic signs of coevolution indicated that the oral phages can evade CRISPR immunity.

RESULTS

DNA amount and metagenome sequencing.

The concentrations of DNA extracted by the method of Kim, Suda, Hattori et al. (19) from two 1 mL saliva samples taken from four healthy volunteers are shown in Fig. 1A. The range of concentrations varied from 24.5-94.1 ng/μl (average was 63.8 ng/μl), corresponding to at least 1.2 μg total DNA per sample, which meets the input requirement (1.0 μg) of the ligation sequencing kit for PromethION. The amount of metagenome sequencing data obtained for the different samples of each individual using HiSeq and PromethION is shown in Fig. 1B. The range of sequencing data obtained through HiSeq varied from 37.2 Gb to 55.5 Gb for HiSeq (purple in Figure 1B), while it was 38.8 Gb to 90.1 Gb for PromethION (red in Fig. 1B).

Preprocessing and viral dark matter in human saliva.

Preprocessing of the HiSeq data using the EDGE pipeline (20) discarded 0.15-0.22% reads and trimmed 0.31-0.41% of bases by the initial quality control. It then removed 0.20-0.93% of the filtered reads that were mapped to the human genome, which was unexpectedly small; this was likely because of our protocol using a saliva DNA collection kit for microbes and viruses followed by the enzymatic DNA extraction. Preprocessing of the PromethION data discarded 17.9-28.2% of bases (9.1-18.3 Gb) by excluding reads with average quality score < 7 (Fig. 1B), and then removed 0.2-9.7% of the filtered reads that were mapped to the human genome.

Taxonomic profiling using Kraken (21) and its full database for assigning taxonomic labels to the short-reads assigned 0.02-0.2% to “Viruses,” whereas 33.0-55.8% remained unmapped to the database, which is expected to correspond to the viral dark matter manifesting an inability to obtain functional or taxonomic annotations for most of the surveyed sequence space (22). The proportion of unmapped reads was smaller than those (63–93%) reported in marine viral metagenomics (23) but was found to be still dominant in human saliva.

Length, N50 of contigs assembled from the long-reads.

Using a large amount of long-reads (13 kb length on average, Fig. S1), we assembled the long-reads using the recently developed assembler Flye (24) with the “--meta” option followed by error correction based on

mapping the HiSeq short-reads to assembled contigs. 94-96% of the HiSeq reads were mapped to the assembled contigs, confirming they well represented viral diversity in the environment. The number of contigs (≥ 1 kb, accounting for $> 99\%$ of all contigs in each sample) after the assembly and error correction ranges 2865-5574 per sample with an average of 3802 (Fig. 2A). The N50 ranges from 187 kb to 345kb, with an average of 249 kb (Fig. 2B). Nucleotide sequences of all the contigs after the assembly and error correction are downloadable at <https://figshare.com/s/e211dd1ab1a77ab94e6f>. On the contrary, if we conducted the hybrid assembly implemented in SPAdes (25), the N50 was found to be much smaller (11.8 kb on average, dashed line in Fig. 2B).

Novel phages/prophages and their taxonomic assignments. For the contigs after the long-read assembly and error correction, we conducted computational classification using VirSorter and identified phage or prophage genome sequences based on the presence of viral ‘hallmark’ genes encoding for components found in many virus particles, together with a reference database of genomes from many viruses (26, 27). The phage or prophage regions identified by VirSorter are classified into category 1 (“most confident”), 2 (“likely”), 3 (“possible”) phages and category 4 (“most confident”), 5 (“likely”), 6 (“possible”) prophages (Fig. S2). The “phages” here include various types of viral sequences outside of the main host chromosome, for example, extrachromosomal prophages (27). The number of each viral category identified in this study is shown in Fig. S2, and the list of all identified viral sequences is shown in Table S1. Because of difficulty to automatically and reliably exclude false positives from the “possible” candidates, we focused on the “most confident” and “likely” phages and prophages (the nucleotide sequences are available at <https://figshare.com/s/e211dd1ab1a77ab94e6f>). We also applied Contig Annotation Tool (CAT) (28) for taxonomic classification of the contigs based on amino acid sequence searching of each ORF against NCBI nr database followed by a voting approach by summing all scores from ORFs supporting a certain taxonomic classification (superkingdom, phylum, class, order, family, genus, and species, separately) and checking if the summation exceeds a cutoff value

(by default $0.5 \times$ summed scores supporting a superkingdom across ORFs). The number of contigs classified as viral by CAT was, however, only ten across the four samples, all of which were included in those by VirSorter.

In order to examine the novelty of the contigs, for each phage or prophage, we conducted genome clustering (29) with oral viral sequences stored in the largest database (IMG/VR v2.0) of cultured and uncultured DNA viruses (30) developed based on the Earth's virome project (13). Those remaining not clustered with a viral sequence in the database (indicated as empty in “clustered with IMG/VR v2.0” column in Table S1), were considered as novel, and its number and proportion stratified by the “most confident” and “likely” categories are shown in Fig. 3. Among the “most confident” phages and prophages, 0.0-43.8% and 12.5-46.4% were novel, while among the “likely” phages and prophages, 41.7-59.3% and 55.7-73.7% were novel, respectively (colored in yellow in Fig. 3).

Furthermore, for each phage or prophage, we conducted taxonomic assignments based on the extent (network) of gene sharing between a query and the reference viral genomes using vConTACT v2.0 (31). The genus-level assignment was possible only for 3, 1, 1, 3 phages in each of the four samples, respectively. Family-level assignment (indicated in “assigned family” column in Table S1) was possible for 35.9-50.7% of the known phages clustered with the viral sequences in the IMG/VR v2.0 database, and for 38.9-53.8% of the novel phages (Fig. S3). However, the family-level assignment could not be achieved for any of the prophages.

Highly abundant phages/prophages. The estimated abundance of each viral sequence is shown in Table S1. Rank abundance in each sample (Fig. S4) seems to follow the power distribution as reported in ocean and lake virome studies (32, 33). For top 10% abundant viral sequences in each sample (“top 10% among category 1, 2, 4, 5” in Table S1), we conducted a recently developed gene calling specifically designed for phage genomes (34) that are very compact and often have overlapping adjacent genes (10.5-39.4% in the top 10% abundant viral sequences). The iterative protein searches using the HMM-HMM-based lightning-fast iterative sequence search (HHblits) tool that represents both query and UniProt database

sequences by profile hidden Markov models (HMM) for the detection of remote homology (35), annotated 53.5% (11840 out of 22113) genes. After excluding 32 questionable viral sequences (“top 10% among category 1,2,4,5 (questionable)” in Table S1) without carrying any annotated gene related to phage morphogenesis or transposase, we identified 129 predominant phages/prophages across the 4 samples, 61.7% of which were novel because they did not cluster with the viral sequences in the IMG/VR v2.0 database. Self-alignment of each viral contig and manual examination of the dot plot confirmed the integrity and absence of redundancy in their assembly. Nucleotide sequences of the predominant 129 phages/prophages are downloadable at <https://figshare.com/s/87a80593aa656ac5567b>.

A proteomic tree based on genomic similarity (normalized tBLASTx scores) (Fig. S5) (36) revealed a notable cluster (Fig. S5) that contains 21.7% (28 across the 4 samples, Table S2) of the predominant phages/prophages, which was largest among clusters that can be seen in the tree. Contigs carrying the predominant phages/prophages accounted for 19.0%, 28.8%, 3.4%, and 8.8% abundance (Table S1) among the “most confident” and “likely” viral contigs of each sample. Analysis of the cluster using a larger proteomic tree, including other reference viral sequences, revealed its location distinctively in a *Siphoviridae* and among *Streptococcus* phages (Fig. 4). Examination of their genes revealed that 86% (24 out of the 28) of them encoded remote homologs of antimicrobial resistance genes with > 99% estimated probability to be (at least partly) homologous to the gene sequences (“CDS related to resistance” Table S2). Genomic context and coverage of HiSeq read of one of the predominant prophages is shown in Fig. 5, in which three remote homologs of antimicrobial resistance genes are colored in red. The remote homologs of genes for phage resistance proteins and acid resistance proteins were also found in 7% and 11%, respectively.

Jumbo phages/prophages. The distribution of the size of the phages and prophages is shown in Fig. S6. The average, median, and interquartile range (IQR) were 57.9 kb, 44.4 kb, 33.9-76.9 kb for the phages, respectively. For the first time in oral microbiota publications, we discovered six phages with genomes larger than 200 kb (Fig. S6, Table 1), that were classified

as jumbo, and had been rarely found but recently began to be identified across Earth's ecosystems (37, 38). We also discovered five jumbo (>200 kb) prophages. Nucleotide sequences of the jumbo phages and prophages are downloadable at <https://figshare.com/s/95f9c3cbb074b2782ccf>. Manual examination based on self-alignment and dot plots revealed one of them (contig_659 of the 3rd sample) contained a redundant sequence region spanning approximately 160 kb that might be resulted from an assembly error. After excluding it, 90% of them (nine out of the ten) had at least one phage hallmark gene. One of the remaining contigs with 223995 bp without a phage hallmark gene (contig_811 in the 2nd sample) was classified as a phage by VirSorter, but 8.9 kb region in the middle (94241-104074) encodes three remote homologs of Type IV secretory system protein for conjugative DNA transfer as well as that of plasmid segregation protein. These results suggested it is rather a plasmid-like element as reported previously (39).

For the jumbo phages and prophages, as explained above, we conducted the gene calling specifically designed for phage genomes (34) and found that 13.2-32.6% of the genes were annotated. Regarding the annotated genes, similar to the highly abundant oral phages/prophages examined above, remote homologs of antimicrobial resistance genes ("Note" in Table 1) were found in 67% (six out of the nine) of the phages/prophages carrying the phage hallmark genes. An example of jumbo prophage embedded in a bacterial chromosome is shown in Fig. 6, in which seven remote homologs of antimicrobial resistance genes with > 99% estimated probability are located (reds in Fig. 6). Six out of the seven were remote homologs of beta-lactamases and formed a cluster in a genomic region (approximately 15.5 kb-17.7 kb; Fig. 6). 31-75 amino acid residues were in the pairwise HMM alignment between the query and beta-lactamases in the UniProt database. The seventh is a remote homolog of *drmA* (77 amino acid residues were in the pairwise HMM alignment) that carries out the export of the antibiotics (40), which was also found in the other three jumbo phages/prophages (Table 1). The taxonomy of the contig was predicted to be the *Treponema* genus, consisting of dozens of species in the human oral microbiota (41). Predicted taxonomy

of another contig containing a jumbo prophage was *Patescibacteria*, the recently proposed Candidate Phyla Radiation (CPR) lineage that encompasses mostly unculturable bacterial taxa with relatively small genome sizes (42), including ubiquitous members of the human oral microbiota (43). Overall, there was no overlap in the predicted taxonomies among the jumbo phages/prophages carrying the phage hallmark genes, suggesting they are not confined to specific phylogenetic groups in the human oral microbiota.

Characteristics of genes encoded in oral phages/prophages. The pan-genome analysis of all the identified viral sequences to create a gene presence or absence matrix revealed remarkable diversity. Among 115621 different genes (i.e., the number of rows in the matrix), only 0.3% (309) genes were “core” and present (based on 70% amino acid sequence identity by BLASTp) in all the four samples, while 86.4% (998641) genes were singletons (Fig. 7). The amino acid sequence alignments of the 309 core genes are downloadable at <https://figshare.com/s/9f76b265f23e23d1e63f>. As much as 94.8% (293 out of 309) of the core genes were annotated as hypothetical by a standard annotation program Prokka compatible with the pan-genome analysis (44). We then conducted the iterative protein searches using the HHblits tool that detected remote homologs with > 99% estimated probability for 301 out of the 309 (Table S3). A breakdown of their functional categories (Fig. 7) shows 38.9% were homologs of uncharacterized proteins (gray), while the two most dominant annotated functional categories are phage morphogenesis (25.9%, red) and host cellular processes (14.6%, green) such as transcriptional regulators. Of the remaining 20.6%, DNA/RNA metabolism (6.3%, yellow) and phage lysis (3.3%, light purple) accounted for the half, while DNA recombination (light yellow), virulence (purple), nuclease (blue), methyltransferase (orange), prophage antirepressor (brown), transposase (dark gray), and host’s phage susceptibility (khaki) accounted for the remaining half.

Distribution of CRISPR spacers and its implication to coevolution between oral phages and bacteria. Finally, for all contigs, we detected CRISPR arrays and spacer sequences. In total, we detected 16187 unique spacers across the four samples. The number of

spacer sequences counted separately in bacterial contigs and phages/prophages (excluding the “possible” sequences) is shown in Fig. 8A. The average number of spacers in bacterial contigs and prophage sequences was 2503 and 668, respectively (Fig. 8A). Whereas, the spacers in CRISPR arrays were also found in the phage sequences, as reported in other phages (45): 16 spacers on an average in three samples except for the 4th sample (Fig. 8A). . The BLAST search of the spacers in CRISPR arrays located in the phage sequences against two datasets of oral phage/prophage sequences (either in IMG/VR v2.0 database or all the assembled viral contigs in its own sample including the category 3 and 6 (“possible”) candidates), revealed almost no homologous sequences (“protospacers”) (Fig. 8B). The number was only one or five in the 2nd sample and zero in the other three samples. Whereas, the BLAST search of the spacers in CRISPR arrays in the bacterial contigs and prophage sequences against the IMG/VR v2.0 database revealed the on an average 22.2% and 20.3% of the spacers had homologous sequences (“protospacers”) in the database, respectively. In contrast, when the BLAST search was conducted against all the assembled viral contigs in each sample, the average proportion decreased to 1.8% for spacers both in the bacterial contigs and in the prophage sequences. The difference between the two conditions was statistically highly significant ($p < 10^{-15}$, chi-square test). Further examination of the result of BLAST search of spacers against IMG/VR v2.0 database and that of the genome clustering conducted above (Fig. 3) revealed that among the oral viral contigs in IMG/VR v2.0 database carrying the protospacers, only 5.2% showed nucleotide similarity enough to be clustered with those identified in the present study, which was significantly lower ($p < 10^{-15}$, chi-square test) than the overall proportion of clustering between the two datasets of viral contigs (36.0%, Fig. 3). In other words, most of the oral viral contigs in the IMG/VR v2.0 database carrying the protospacers were outside those identified in the present study.

Discussion

The present study, for the first time, analyzed as much as > 30 Gb metagenomic data of both PromethION and HiSeq per sample. The analysis revealed 40.0-56.3% of the phages and 49.0-72.5% of the prophages were novel (Fig. 3), and identified as much as ten jumbo oral phages/prophages (including the plasmid-like element) in an oral environment, significantly increasing our knowledge about “who is there” in the human oral cavity. Such jumbo phages were previously found among approximately 20 host bacterial genus (38, 46, 47) and recently began to be identified across Earth’s ecosystems (37). However, so far, there is no publication of those in the human oral cavity. Long-read sequencing using PromethION will boost the discovery of such novel jumbo phages/prophage more in various environments.

The analysis also enabled quantification of the abundance of each assembled viral sequence (Table S1), and for the first time, identified the highly abundant oral phages/prophages, some of which formed the novel distinctive *Streptococcus* phage cluster (Fig. 4). There has been no study that demonstrated the abundance of oral phages and quantitatively answered what kind of phages are not only present but also abundant in the human oral cavity. In the *Streptococcus* phage cluster, there was a “most confident” phage (contig_2233 in the 2nd sample) that was similar to a known *Streptococcus* phage SM1 (NC_004996 in Fig. S7) and clustered with 7 viral sequences in the IMG/VR 2.0 database (Table S1). Whereas, 85.7% (24 out of the 28) seemed to be novel as they did not cluster. The genome clustering used cutoffs for obtaining pairwise hits, which are more stringent than the normalized tBLASTx scores used for constructing the proteomic tree. 87.5% of them (21 out of the 24) were “likely” prophages (Fig. S5), of which 2 had genome synteny to the “most confident” phage in the cluster (Fig. S7). Regions outside the synteny of the “likely” prophages did not encode a gene for phage morphogenesis, and might not be viral but host bacterial sequences at both ends mistakenly extracted by VirSorter, making the “likely” prophages not clustered with viral sequences in the IMG/VR 2.0 database. We should keep in mind if that is the case, the proportion of novel “likely” prophages, as well as of jumbo prophages, could be overestimated.

A notable feature in oral microbiota revealed from the largest deep-sequencing data of both PromethION and HiSeq was high frequency of the remote homologs of antimicrobial resistance genes; 67% of the jumbo phages/prophages and 86% of the predominant *Streptococcus* phages cluster. There has been controversy as to how prevalent the antimicrobial resistance genes are really in phages, given that the genes are only rarely directly encoded in publicly available phage genome sequences (48). The remote homologs formed the pairwise HMM alignment not entirely but partly with the antimicrobial resistance genes, and are perhaps unfunctional. However, they might play a role as a source of recombination or horizontal gene transfer to generate a new antimicrobial resistance gene sequence. In the human oral cavity, it is known that the “mosaic” penicillin-resistant genes are generated by recombination among oral streptococci (49) and that the massive diversity of organisms and a large amount of extracellular DNA in oral biofilm matrices are expected to create opportunities of recombination or horizontal gene transfer (4).

Regarding the analysis of the distribution of CRISPR spacers and protospacers, a previous study reported that protospacers were detectable as a small fraction (from 1% to 19%, on average ~7%) of the spacers (50). The proportion in our study obtained by the BLAST search of the spacers against the oral viral sequences in the IMG/VR v2.0 database was slightly higher than the upper limit (19%). A more recent study of the human gut microbiome (51) reported that the proportion was 23.5% when IMG/VR v2.0 database was used, which is comparable to the findings of our study. When the BLAST search was instead conducted against all the assembled viral contigs in each sample, the proportion was significantly decreased. The possible explanation for this reduction that the assembled viral contigs account for only a small part of the overall oral virome is nullified because 94-96% of the HiSeq reads were mapped to them. However, it suggests that the oral phages currently in human saliva could be under selective pressure escaping from CRISPR immunity, given hundreds of thousands of CRISPR spacer groups are transcribed in the human oral cavity (52). Previously, it was reported that oral streptococcal CRISPR spacers and viruses carrying protospacers coexisted in human saliva of

a subject, and CRISPRs in some subjects were just as likely to match viral sequences from other subjects as they were to match viruses from the same subject (53). Compared to the previous study that examined 3473 unique spacers confined to streptococci of 4 subjects (53), we expanded the number to > 4-fold across various bacterial species. Therefore, the result presented in our study is probably more general and reliable than the previous study in understanding the overall relationship between CRISPR spacers and protospacers in the oral microbiota. A recent review also pointed out the possibility of genetic conflict invoked by the coevolution of oral phages and host bacteria using classic Red Queen dynamics (4). The possibility was recently suggested in another study that revealed signatures of highly elevated recombination among oral phage genomes, potentially reflecting the evolutionary conflict in which phages continuously change their genomic sequences by recombination to escape cleavage by the host bacterial immune system in the human oral cavity (15).

Recently, a framework of viral, long-read metagenomics via MinION was proposed, and one of its main advantages was to capture more and longer genomic islands (54). The genomic island was defined in a previous study by coverage of less than 20% of the median coverage of the entire contig and >500 bp in length (55). However, such a decrease in coverage can result from a long read-assembled contig that contains too erroneous regions for correction by short-read mapping. In particular, long-read assembly of low coverage contig is challenging, and detection of a genomic island in such a contig was not practical. This issue is worth to be addressed in future works by exploiting the improvement of per-base accuracy and per-cost throughput of the long-read sequencing platforms.

A remarkable feature we found in the pan-genome analysis was the high diversity of genes encoded by oral phages/prophages: as much as 86.4% of the genes were singletons (i.e., specific to each sample). It is consistent with previous studies of 16S rRNA-based metagenomics reporting individuals' oral microbiota are highly specific at the species level. Our study further demonstrated the specificity at the gene level based on the very deep shotgun

metagenome sequencing.

Furthermore, we conducted identification and HMM-based computational characterization of the core genes of oral phages (most of which were initially annotated as hypothetical) to deepen our understanding of what they are doing. The high fraction of the core genes for host cellular processes after phage morphogenesis supports a recently proposed notion that phages modulate the oral microbiome through multiple mechanisms and represent an additional level of balance required for eubiosis (4). Experimental characterization of the genes in the future will further deepen our understanding of their functions and roles.

In summary, our study demonstrates, for the first time, the power of long-read metagenomics utilizing PromethION in uncovering novel and abundant bacteriophages, characteristics of their genes, and their interaction with host bacterial immunity. Our study will provide a solid basis for utilizing PromethION to study bacteriophages and host bacteria simultaneously, and further explore the viral dark matter in various environments.

Materials and Methods

Saliva collection, DNA extraction, library preparation, and metagenome sequencing.

Two samples of 1 mL saliva were successively collected and stored using a kit specialized for microbial and viral DNA/RNA (OMNIgene ORAL OM-501) from each of four healthy volunteers (2 men and 2 women aged 35 to 65 years old). From the four individuals, two saliva samples for each were taken and subjected to DNA extraction using an enzymatic method (19). The extracted DNA samples were stored in 50 µL pure water. From the paired samples per person, we used one with a smaller amount of extracted DNA for Nextera XT library construction and genome sequencing with the Illumina HiSeq 2 x 150 bp paired-end run protocol, and another for library construction using the ligation sequencing kit (SQK-LSK109) and PromethION sequencing.

Preprocessing and initial taxonomic profiling.

We used EDGE pipeline version 1.5 (20) for preprocessing (trimming or filtering out reads, and removal of reads mapped to the human genome) of the HiSeq data. For the PromethION data, we used MinIONQC (56) to check diagnostic plots and NanoFilt (57) (“-q 6 --headcrop 75” option) to filter out reads with average quality less than 7 and trim 75 nucleotides from the beginning of a read. We then used Minimap2 (58) to find and remove PromethION reads mapped to the human genome. Initial taxonomic profiling using HiSeq reads was conducted using Kraken (21) and its full database created in March 2017 following https://github.com/mw55309/Kraken_db_install_scripts as a part of the EDGE pipeline. We executed it implemented in metaWRAP (59) as a module that can directly subsample the same number (approximately 120 million) of reads in each sample.

Assembly, error correction, and quality assessment.

We used Flye (24) with the “--meta and --genome-size 200 m (a value much higher than known bacterial genome size)” option to assemble the preprocessed PromethION long-reads, followed by mapping the preprocessed HiSeq short-reads to assembled contigs using bowtie2 (60) with “—very-sensitive” option. We then conducted an error correction based on the mapping using a single run of Pilon for each assembled contig (61). It took several days from the mapping to error correction. We also conducted a hybrid assembly implemented in SPAdes with “—meta --nanopore” option (25). Quality assessment for the assembled contigs was conducted using MetaQUAST (62).

Viral sequence identification, taxonomic classification, abundance quantification, and clustering of assembled contigs.

We used VirSorter (26) with the “-db 2” option using the “Viromes” database. We also used CAT (28) with the default database and taxonomy information created in January 2019 and included in the package. Distribution of length of viral sequences identified by VirSorter was examined using JMP Pro version 13 (SAS Institute, Cary, NC, USA). The abundance of each viral sequence was estimated by FastViromeExplorer (63) using randomly selected 123519885 read-pairs (equal to the number of the least sequenced sample) from each of the four samples. A proteomic tree of highly abundant viral sequences (top 10% in each sample) was constructed using ViPTreeGen, and a larger proteomic tree,

including the highly abundant viral sequence cluster and other reference viral sequences, was constructed using ViPTree (36) in which their genome alignment was also visually examined. The viral sequences were clustered with oral viral sequences registered in IMG/VR v2.0 database (30) according to a “Viral genome clustering” procedure included in a nontargeted virus sequence discovery pipeline (from step 8 to 11 in (29), using cutoffs of nucleotide sequence similarity $\geq 90\%$, covered length $\geq 75\%$, and covered length requiring at least one contig of $> 1,000$ -bp length. Taxonomic assignment of the contigs was conducted using vConTACT v.2.0 (31) with “--db 'ProkaryoticViralRefSeq85-Merged' --pcs-mode MCL --vcs-mode ClusterONE” option.

Functional annotation and pan-genome analysis of viral sequences. The prediction of protein-coding genes in the phages/prophages was conducted using PHANOTATE (34) implemented in multiPhATE (64). For each predicted gene, we conducted iterative protein searches using HHblits (35), which represents both query and database sequences by profile hidden Markov models (i.e., condensed representation of multiple sequence alignments specifying, for each sequence position, the probability of observing each of the 20 amino acids) instead of single sequences for the detection of remote homology. We used the clustered uniprot20_2016_02 database (http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/), which covers essentially all of the sequence universe by clustering the UniProt database (65) from EBI/SIB/PIR and the non-redundant (nr) database from the NCBI. For all hits with $> 99\%$ probability of being true positives, we visualized their genomic locations using Geneious software (Biomatters Ltd., Auckland, New Zealand) and individually examined each annotation. We did not include uncharacterized genes in the visualization because the number was too many for the software. Furthermore, we conducted pan-genome analyses using the Roary pipeline with the “-i 70 -group_limit 500000” option (66) after the prediction of protein-coding genes for every contig using Prokka software with the “-kingdom Viruses” option (44). For the core genes, we also conducted the iterative protein searches using HHblits,

individually examined each annotation, and manually made the functional categorization.

Analyses of distribution of CRISPR spacers. CRISPR arrays were predicted on all contigs using a command-line version of the program CRISPRDetect (67). For each sample, spacers were extracted from the output files and searched using the BLASTn-short function from the BLAST+ package (68) against either oral viral contigs in IMG/VR v2.0 database or those assembled in this study in each sample. The cutoffs were set with at least 95% identity over the whole spacer length and allowing only 1–2 SNPs at the 5' end of the sequence, according to a procedure in the Earth's virome project (13).

Ethics. This study was approved by the ethics committees of National Institute of Infectious Diseases (approval number 931).

Data availability. The data of PromethION and HiSeq after the quality control and removal of human reads were deposited at DDBJ (with JGA accession number JGAS00000000186) and will be mirrored at NCBI under BioProject accession PRJDB9452.

Acknowledgments

The computational calculations were done at the Human Genome Center at the Institute of Medical Science (the University of Tokyo). This work was supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (19H04846 to K.Y.) and JSPS KAKENHI Grant Number 16H06279 (PAGS). We thank Yutaka Suzuki laboratory members for PromethION sequencing. We thank Ryan Wick and Nick Loman for the instruction of the long-read assembly and thank Matthew B Sullivan, Masahira Hattori, Yosuke Nishimura, Tetsuya Hayashi, Takuro Nunoura, Yoshihiro Takagi, Masaki Shintani and So Nakagawa for discussions.

Figure Legends

Fig. 1. (A) Concentration of DNA extracted from saliva samples. The prefix numbers “2_” and “1_” indicate individual numbers; e.g., “1_1” and “1_2” indicate the 1st and 2nd saliva samples collected from the 1st individual **(B) The amount of generated sequence data.** “Q ≥ 7” (Orange) represents a subset of PromethION reads with average quality score ≥ 7 .

Fig. 2. Assembly statistics. (A) The number of contigs with ≥ 1 kb. (B) N50. The dashed line indicates the average of contigs generated by the hybrid assembly rather than the long-read assembly.

Fig. 3. The number and proportion of novel viral sequences identified in each sample and stratified by the “most confident” and “likely” phages and prophages. Yellow: not clustered with viral sequences in the IMG/VR v2.0 database, and regarded as the novel.

Fig. 4. Proteomic tree showing a cluster of highly abundant oral phages. The red lines at the top right and bottom right indicates the cluster consisting of the 28 abundant phages found across the 4 samples and listed in Table S2.

Fig. 5. A highly abundant oral prophage. Top: coverage plot based on mapping short-reads against it. Bottom: Genes characterized by the HMM-based iterative protein searches are shown with colors according to their functional categories.

507

508 **Fig. 6. A jumbo prophage.** This Fig. is similar to Fig. 5. The prophage region is indicated as a
509 rectangle in the coverage plot spanning the entire contig.

510

511 **Fig. 7. (A) The number of the gene present in 4, 3, 2, 1 out of the four samples.** The core
512 (4/4) genes are colored in orange, while singletons are colored in grey. **(B) Breakdown of the**
513 **core genes.** The number of genes in each category is shown in the pie chart. More detailed
514 information on each gene is shown in Table S3.

515

516 **Fig. 8. (A) The number of spacers in CRISPR arrays found in the bacterial contig, phage,**
517 **and prophage regions. (B) The number of protospacers located in oral viral contigs in**
518 **either IMG/VR v2.0 database (left) or those assembled in each sample (right).**

519

520

Table

Table 1. Detected jumbo phages and prophages.

Sample	Length (bp)	Type	Num. of phage hallmark genes	Taxonomy of the contig*	Note
1	263225	prophage	1	Patescibacteria group: 0.62 <i>Candidatus</i> <i>Gracilibacteria</i> : 0.62	Encoding a remote homolog of <i>drrA</i> that carries out export of the antibiotics
1	215658	prophage	6	<i>Treponema</i> (genus): 0.93	Fig 6
2	294299	phage	4	<i>Selenomonas</i> (genus): 0.59	Encoding a remote homolog of a beta-lactamase superfamily domain
2	270161	prophage	10	<i>Lachnospiraceae</i> (family): 0.60	Encoding remote homologs of three metallo-beta-lactamase domain proteins and <i>vanZ</i>
2	223995	phage (plasmid-like element)	0	<i>Streptococcus</i> sp. (species):	Highly abundant (top 10%)

				0.64	
2	201060	phage	1	<i>Fusobacterium</i>	Encoding a remote
				(genus): 0.86	homolog of <i>drrA</i> that
					carries out export of
					the antibiotics
3	217059	prophage	2	unclassified	
				Bacteria: 0.50	
3	201353	phage	5	<i>Leptotrichia</i>	Encoding remote
				(genus): 0.83	homologs of two
					multidrug-resistant
					proteins and <i>drrA</i> that
					carries out export of
					the antibiotics
4	246016	phage	1	<i>Campylobacter</i>	
				(genus): 0.96	
4	200581	prophage	1	<i>Leptotrichia</i>	
				(genus): 0.84	

525 *predicted by Contig Annotation Tool (CAT) (28): the support values for prediction of the
526 taxonomy of the contig are indicated in parentheses.

527

528

529

530 **References**

531

- 532 1. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. 2018. Current
533 understanding of the human microbiome. *Nat Med* 24:392-400.
- 534 2. Kolodziejczyk AA, Zheng D, Elinav E. 2019. Diet-microbiota interactions and
535 personalized nutrition. *Nat Rev Microbiol* 17:742-753.
- 536 3. De Sordi L, Lourenco M, Debarbieux L. 2019. The Battle Within: Interactions of
537 Bacteriophages and Bacteria in the Gastrointestinal Tract. *Cell Host Microbe*
538 25:210-218.
- 539 4. Baker JL, Bor B, Agnello M, Shi W, He X. 2017. Ecology of the Oral Microbiome:
540 Beyond Bacteria. *Trends Microbiol* 25:362-374.
- 541 5. Edlund A, Santiago-Rodriguez TM, Boehm TK, Pride DT. 2015. Bacteriophage and
542 their potential roles in the human oral cavity. *J Oral Microbiol* 7:27423.
- 543 6. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI. 2012. Going viral:
544 next-generation sequencing applied to phage populations in the human gut. *Nat Rev*
545 *Microbiol* 10:607-17.
- 546 7. Garmaeva S, Sinha T, Kurilshikov A, Fu J, Wijmenga C, Zhernakova A. 2019.
547 Studying the gut virome in the metagenomic era: challenges and perspectives. *BMC*
548 *Biol* 17:84.
- 549 8. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. 2016. Healthy
550 human gut phageome. *Proc Natl Acad Sci U S A* 113:10400-5.
- 551 9. Zuo T, Lu XJ, Zhang Y, Cheung CP, Lam S, Zhang F, Tang W, Ching JYL, Zhao R,
552 Chan PKS, Sung JJY, Yu J, Chan FKL, Cao Q, Sheng JQ, Ng SC. 2019. Gut mucosal
553 virome alterations in ulcerative colitis. *Gut* 68:1169-1179.
- 554 10. Han M, Yang P, Zhong C, Ning K. 2018. The Human Gut Virome in Hypertension.
555 *Front Microbiol* 9:3150.

- 556 11. Ly M, Abeles SR, Boehm TK, Robles-Sikisaka R, Naidu M, Santiago-Rodriguez T,
557 Pride DT. 2014. Altered oral viral ecology in association with periodontal disease.
558 MBio 5:e01133-14.
- 559 12. Wang J, Gao Y, Zhao F. 2016. Phage-bacteria interaction network in human oral
560 microbiome. Environ Microbiol 18:2143-58.
- 561 13. Paez-Espino D, Eloie-Fadrosch EA, Pavlopoulos GA, Thomas AD, Huntemann M,
562 Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome.
563 Nature 536:425-30.
- 564 14. Paez-Espino D, Roux S, Chen IA, Palaniappan K, Ratner A, Chu K, Huntemann M,
565 Reddy TBK, Pons JC, Llabres M, Eloie-Fadrosch EA, Ivanova NN, Kyrpides NC. 2019.
566 IMG/VR v.2.0: an integrated data management and analysis system for cultivated and
567 environmental viral genomes. Nucleic Acids Res 47:D678-D686.
- 568 15. Meier-Kolthoff JP, Uchiyama J, Yahara H, Paez-Espino D, Yahara K. 2018.
569 Investigation of recombination-intense viral groups and their genes in the Earth's
570 virome. Sci Rep 8:11496.
- 571 16. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a
572 better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16-21.
- 573 17. Suzuki Y, Nishijima S, Furuta Y, Yoshimura J, Suda W, Oshima K, Hattori M,
574 Morishita S. 2019. Long-read metagenomic exploration of extrachromosomal mobile
575 genetic elements in the human gut. Microbiome 7:119.
- 576 18. Kingan SB, Heaton H, Cudini J, Lambert CC, Baybayan P, Galvin BD, Durbin R,
577 Korlach J, Lawniczak MKN. 2019. A High-Quality De novo Genome Assembly from a
578 Single Mosquito Using PacBio Sequencing. Genes (Basel) 10.
- 579 19. Kim SW, Suda W, Kim S, Oshima K, Fukuda S, Ohno H, Morita H, Hattori M. 2013.
580 Robustness of gut microbiota of healthy adults in response to probiotic intervention
581 revealed by high-throughput pyrosequencing. DNA Res 20:241-53.
- 582 20. Li PE, Lo CC, Anderson JJ, Davenport KW, Bishop-Lilly KA, Xu Y, Ahmed S, Feng S,

- Mokashi VP, Chain PS. 2017. Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res* 45:67-80.
21. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15:R46.
22. Brum JR, Ignacio-Espinoza JC, Kim EH, Trubl G, Jones RM, Roux S, VerBerkmoes NC, Rich VI, Sullivan MB. 2016. Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc Natl Acad Sci U S A* 113:2436-41.
23. Hurwitz BL, Sullivan MB. 2013. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* 8:e57355.
24. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37:540-546.
25. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32:1009-15.
26. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3:e985.
27. Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 4.
28. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. 2019. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* 20:217.
29. Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. 2017. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat Protoc* 12:1673-1682.
30. Paez-Espino D, Chen IA, Palaniappan K, Ratner A, Chu K, Szeto E, Pillay M, Huang J, Markowitz VM, Nielsen T, Huntemann M, TB KR, Pavlopoulos GA, Sullivan MB,

- Campbell BJ, Chen F, McMahon K, Hallam SJ, Denev V, Cavicchioli R, Caffrey SM, Streit WR, Webster J, Handley KM, Salekdeh GH, Tsesmetzis N, Setubal JC, Pope PB, Liu WT, Rivers AR, Ivanova NN, Kyrpides NC. 2017. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res* 45:D457-D465.
31. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* 37:632-639.
32. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano SI. 2019. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. *Environ Microbiol* 21:4740-4754.
33. Yoshida T, Nishimura Y, Watai H, Haruki N, Morimoto D, Kaneko H, Honda T, Yamamoto K, Hingamp P, Sako Y, Goto S, Ogata H. 2018. Locality and diel cycling of viral production revealed by a 24 h time course cross-omics analysis in a coastal region of Japan. *ISME J* 12:1287-1295.
34. McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA. 2019. PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* 35:4537-4542.
35. Remmert M, Biegert A, Hauser A, Soding J. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9:173-5.
36. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. 2017. ViPTree: the viral proteomic tree server. *Bioinformatics* 33:2379-2380.
37. Al-Shayeb B, Sachdeva R, Chen LX, Ward F, Munk P, Devoto A, Castelle CJ, Olm MR, Bouma-Gregson K, Amano Y, He C, Meheust R, Brooks B, Thomas A, Lavy A, Matheus-Carnevali P, Sun C, Goltsman DSA, Borton MA, Sharrar A, Jaffe AL, Nelson TC, Kantor R, Keren R, Lane KR, Farag IF, Lei S, Finstad K, Amundson R, Anantharaman K, Zhou J, Probst AJ, Power ME, Tringe SG, Li WJ, Wrighton K, Harrison S, Morowitz M, Relman DA, Doudna JA, Lehours AC, Warren L, Cate JHD,

637 Santini JM, Banfield JF. 2020. Clades of huge phages from across Earth's ecosystems.
638 Nature doi:10.1038/s41586-020-2007-4.

639 38. Yuan Y, Gao M. 2017. Jumbo Bacteriophages: An Overview. Front Microbiol 8:403.
640 39. Octavia S, Sara J, Lan R. 2015. Characterization of a large novel phage-like plasmid in
641 Salmonella enterica serovar Typhimurium. FEMS Microbiol Lett 362:fnv044.

642 40. Kaur P. 1997. Expression and characterization of DrrA and DrrB proteins of
643 Streptomyces peucetius in Escherichia coli: DrrA is an ATP binding protein. J Bacteriol
644 179:569-75.

645 41. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A, Wade
646 WG. 2010. The human oral microbiome. J Bacteriol 192:5002-17.

647 42. Lemos LN, Medeiros JD, Dini-Andreote F, Fernandes GR, Varani AM, Oliveira G,
648 Pylro VS. 2019. Genomic signatures and co-occurrence patterns of the ultra-small
649 Saccharimonadia (phylum CPR/Patescibacteria) suggest a symbiotic lifestyle. Mol
650 Ecol 28:4259-4271.

651 43. Bor B, Bedree JK, Shi W, McLean JS, He X. 2019. Saccharibacteria (TM7) in the
652 Human Oral Microbiome. J Dent Res 98:500-509.

653 44. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics
654 30:2068-9.

655 45. Chenard C, Wirth JF, Suttle CA. 2016. Viruses Infecting a Freshwater Filamentous
656 Cyanobacterium (Nostoc sp.) Encode a Functional CRISPR Array and a
657 Proteobacterial DNA Polymerase B. mBio 7.

658 46. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, Archie EA,
659 Turnbaugh PJ, Seed KD, Blekhman R, Aarestrup FM, Thomas BC, Banfield JF. 2019.
660 Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat
661 Microbiol 4:693-700.

662 47. Thanki AM, Brown N, Millard AD, Clokie MRJ. 2019. Genomic Characterization of
663 Jumbo Salmonella Phages That Effectively Target United Kingdom Pig-Associated

664 Salmonella Serotypes. *Front Microbiol* 10:1491.

665 48. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit MA. 2017. Phages rarely
666 encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J*
667 11:237-247.

668 49. Dowson CG, Coffey TJ, Kell C, Whiley RA. 1993. Evolution of penicillin resistance in
669 *Streptococcus pneumoniae*; the role of *Streptococcus mitis* in the formation of a low
670 affinity PBP2B in *S. pneumoniae*. *Mol Microbiol* 9:635-43.

671 50. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. 2017. The
672 CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes.
673 *mBio* 8.

674 51. Soto-Perez P, Bisanz JE, Berry JD, Lam KN, Bondy-Denomy J, Turnbaugh PJ. 2019.
675 CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting
676 against Phages in a Human Virome Catalog. *Cell Host Microbe* 26:325-335 e5.

677 52. Lum AG, Ly M, Santiago-Rodriguez TM, Naidu M, Boehm TK, Pride DT. 2015.
678 Global transcription of CRISPR loci in the human oral cavity. *BMC Genomics* 16:401.

679 53. Pride DT, Salzman J, Relman DA. 2012. Comparisons of clustered regularly
680 interspaced short palindromic repeats and viromes in human saliva reveal bacterial
681 adaptations to salivary viruses. *Environ Microbiol* 14:2564-76.

682 54. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ,
683 Sullivan MB, Temperton B. 2019. Long-read viral metagenomics captures abundant
684 and microdiverse viral populations and their niche-defining genomic islands. *PeerJ*
685 7:e6800.

686 55. Mizuno CM, Ghai R, Rodriguez-Valera F. 2014. Evidence for metaviromic islands in
687 marine phages. *Front Microbiol* 5:27.

688 56. Lanfear R, Schalamun M, Kainer D, Wang W, Schwessinger B. 2019. MinIONQC: fast
689 and simple quality control for MinION sequencing data. *Bioinformatics* 35:523-525.

690 57. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack:

visualizing and processing long-read sequencing data. *Bioinformatics* 34:2666-2669.

58. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094-3100.

59. Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6:158.

60. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357-9.

61. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963.

62. Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32:1088-90.

63. Tithi SS, Aylward FO, Jensen RV, Zhang L. 2018. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* 6:e4227.

64. Ecale Zhou CL, Malfatti S, Kimbrel J, Philipson C, McNair K, Hamilton T, Edwards R, Souza B. 2019. multiPhATE: bioinformatics pipeline for functional annotation of phage isolates. *Bioinformatics* 35:4402-4404.

65. UniProt C. 2010. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 38:D142-8.

66. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691-3.

67. Biswas A, Staals RH, Morales SE, Fineran PC, Brown CM. 2016. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* 17:356.

68. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

718

719

A. **B.**

■ clustered or ■ not with IMG/VR v.2.0 database

Figure 4

Figure 5

A. **B.**