environmental microbiology

Environmental Microbiology (2015) doi:10.1111/1462-2920.12923

Phage–bacteria interaction network in human oral microbiome

Jinfeng Wang,[†] Yuan Gao[†] and Fangqing Zhao*
Computational Genomics Lab, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China.

Summary

Although increasing knowledge suggests that bacteriophages play important roles in regulating microbial ecosystems, phage–bacteria interaction in human oral cavities remains less understood. Here we performed a metagenomic analysis to explore the composition and variation of oral dsDNA phage populations and potential phage–bacteria interaction. A total of 1,711 contigs assembled with more than 100 Gb shotgun sequencing data were annotated to 104 phages based on their best BLAST matches against the NR database. Bray–Curtis dissimilarities demonstrated that both phage and bacterial composition are highly diverse between periodontally healthy samples but show a trend towards homogenization in diseased gingivae samples. Significantly, according to the CRISPR arrays that record infection relationship between bacteria and phage, we found certain oral phages were able to invade other bacteria besides their putative bacterial hosts. These cross-infective phages were positively correlated with commensal bacteria while were negatively correlated with major periodontal pathogens, suggesting possible connection between these phages and microbial community structure in oral cavities. By characterizing phage–bacteria interaction as networks rather than exclusively pairwise predator–prey relationships, our study provides the first insight into the participation of cross-infective phages in forming human oral microbiota.

Introduction

There is a growing interest in exploring the microbiomes for their close connection with human health (Methe *et al*., 2012). These studies primarily focused on bacterial communities at multiple body site niches but neglected viruses (Koren *et al*., 2011; Ravel *et al*., 2011; Qin *et al*., 2012; Aagaard *et al*., 2014), though viruses have been clearly demonstrated to impact microbial communities in many environmental ecosystems (Sullivan *et al*., 2005; Kunin *et al*., 2008; Rohwer and Thurber, 2009). It is also widely believed that bacteriophages, which account for a portion of viruses, have a major and direct impact on the structure of bacterial communities by implementing selection pressures (Donlan, 2009; Koskella, 2013). Given the potential effect of phages on bacterial communities and human health, several studies focused on the phage–bacterial host dynamics (Levin and Bull, 2004; Perez-Brocal *et al*., 2013; Reyes *et al*., 2013). However, studies on human virome and phage–bacteria interaction are still in their early days (Reyes *et al*., 2012).

Among various human body niches, oral cavity is a special ecosystem because of its intermittent oxygen exposure and isolation. It is also a reservoir of abundant bacteria and viruses. These microbes are reported to relate with oral and many other diseases, such as periodontitis, preterm birth and cardiovascular disease (Darveau, 2010; Genco and Van Dyke, 2010; Aagaard *et al*., 2014). Recently, Pride *et al*. (2012b) employed high throughput sequencing to study oral virome and demonstrated viruses prevalently exist in human oral cavities and the majority of them are bacteriophages. They also indicated that phage community is individual specific and remains stable over a long time without disturbance (Abeles *et al*., 2014). Nonetheless, association between phage composition and their host environment has also been reported. For example, unrelated subjects from the same family have similar phage communities compared with those from different households (Robles-Sikisaka *et al*., 2013). Most recently, they further discovered that the oral virome in periodontitis patients, rather than healthy subjects, tend to have homogeneous community structure, which suggests an altered oral phage community is associated with periodontal disease (Ly *et al*., 2014). However, corresponding alteration tendency were not observed on bacterial communities, which obscures our understanding on pathogenic mechanism of periodontal disease.

Actually, discordance of bacteria and phages was not rare at community scale in human microbiome. For instance, clear distinction between Crohn's disease and

Received 9 January, 2015; revised 22 May, 2015; accepted 25 May, 2015. *For correspondence. E-mail zhfq@mail.biols.ac.cn; Tel. (+8610) 64869325; Fax (+8610) 64874346. [†]Equal contributors.

control groups was observed based on bacteria diversity but not on phage (Perez-Brocal *et al.*, 2013). In another example from oral cavity, phage communities were obviously gender-consistent, whereas similar patterns could not be detected for the bacterial communities (Abeles *et al.*, 2014). A possible explanation for the discordance of diversities is the existence of cross-infective phages (CIPs). Such cross-infections involve multiple hosts and phages, which display one-to-many or many-to-many relationships and complicate correspondence between the two communities (Weitz *et al.*, 2013). It was reported that CIPs have great impact on diversities of bacterial community (Weitz *et al.*, 2013). Thus, recognition of CIPs in oral cavity by reconstructing phage–bacterial infection network will be of great necessity for understanding interactions between phages and bacteria.

Existing recognition of cross-infections mainly comes from isolation-based approaches, which needs data accumulation from a mass of laboratory work (Holmfeldt *et al.*, 2007; Flores *et al.*, 2011; 2013). Moreover, these approaches were low throughput and only provided partial information on the association between phages and their bacterial hosts. Metagenomic analysis based on clustered regularly interspaced short palindromic repeat (CRISPR) in turn was suggested to be a worthwhile approach to detect phage–bacteria interaction in complex communities (Weitz *et al.*, 2013). As an important microbial defense system against foreign genetic elements, CRISPR arrays universally exist in genomes of bacteria and archaea (Barrangou *et al.*, 2007). Such loci have alternately arranged direct repeat (DR) and spacer sequences, which are usually dozens of nucleotides in length. When a bacterium encounters a bacteriophage that has never encountered before, bacterial CRISPR-associated complex can cleave a short segment of exogenous DNA from the invader and integrate the segment into its CRISPR loci as a novel spacer (Horvath and Barrangou, 2010). As the CRISPR arrays record the infection relationship between bacteria and phage by direct repeats and spacers, respectively, they were frequently used to identify phages from metagenomic assemblies as well as to investigate the interactions between certain phages and their bacterial hosts (Pride *et al.*, 2011; 2012a; Roux *et al.*, 2014).

Our previous study demonstrated the association between certain phages and periodontal status and sample types, and their abundance was consistent with corresponding bacterial hosts (Wang *et al.*, 2013). To provide expanding insights into the oral viruses, in the present study, we performed a metagenomic survey on the composition and variation of the phage populations from WGS data of deep sequencing. As purification of virus-like particles (VLPs) might lose some viral components and considerable bacteria in nature are lysogens (Thurber *et al.*, 2009), we conducted metagenomic sequencing of both 25 oral specimens without VLPs purification and 15 viruses-enriched specimens. A phage–bacteria infection network was reconstructed based on CRISPR elements identified from NGS reads and assemblies. By viewing oral phage–bacteria interaction as networks rather than as exclusively pairwise predator–prey relationships in isolation, the effect of specific phages on the oral microbiome and its possible association with periodontally healthy status were studied.

Results

Characteristics of the sequencing data and assemblies

We collected a total of 40 salivary or dental plaque specimens from 40 human subjects with different periodontal status. Metagenomic DNA was first separately isolated from 25 of these 40 specimens and sequenced directly. By this way, a total of 832 887 202 reads were generated, with an average of 33.3 million reads per sample (Table S1). Sequencing reads from each sample were then merged into three separate datasets according to the sample types, including saliva of periodontally healthy individuals, plaque of periodontally healthy individuals and plaque of periodontally diseased individuals. We next *de novo* assembled the merged datasets individually, yielding about 1.4, 0.05 and 0.07 million contigs respectively.

After gene prediction and annotation (BLASTP against the NCBI nr database, *E*-value $< 10^{-3}$) for the contigs no shorter than 500 bp of the three assembled datasets, approximately 0.04–1.0 million contigs could be assigned to microbial sequences, and 107–239 of them were phage sequences (Supplementary Fig. S1). The remaining 15 out of the 40 collected specimens were pooled into three samples, precipitated with polyethylene glycol and subjected to deep sequencing, which generated 230 331 556 more reads. These sequencing data were utilized to generate more phage contigs from human oral cavities and facilitate a more complete genome assembly. After quality filtering and manual editing, a total of 1,711 phage contigs were assembled from all of the sequencing data. Most of these contigs were 0.5–2.0 kb in length and 5–10 × in *k*-mer coverage (Fig. 1A). According to reference and overlapping contigs, 17 nearly complete genomes were manually assembled, eight of which with coverages greater than 15 × and without gaps were shown in Supplementary Fig. S2.

A catalog of oral phages based on annotated assemblies

To characterize the content of the oral phage communities, we first assigned the metagenomic assemblies to known phage species (BLASTP against the NCBI nr database, *E*-value < 10^{-3}). Considering the rapid evolution of virus and insufficiency of viral genomes in currently available databases and specificity of oral phages, we chose a lower identity (20%) as criteria for phage identification than that for bacteria. As a result, 104 unique phage species belonging to three families (*Siphoviridae*, *Myoviridae* and *Podoviridae*) were identified from the salivary and dental plaque samples, with a threshold of 20% sequence identities (Fig. 1B). Most of these identifiable phages were classified as species of *Actinomyces* phage, *Corynebacterium* phage, *Lactococcus* phage, *Mycobacterium* phage, *Pseudomonas* phage, *Rhodococcus* phage, *Streptococcus* phage and *Yersinia* phage, which accounted for nearly 70% (30.1 million) of ~43.6 million normalized reads assigned to phage contigs. Among them, *Streptococcus* phage and *Pseudomonas* phage were also the most diverse phages, including 10 and 13 annotated phage species respectively. However, when observing their sequence identities (Fig. 1B, left panel), we found that relatively few contigs (~24%) had more than 80% identities, and only one *Actinomyces* phage species, four *Streptococcus* phage species and six *Pseudomonas* phage species showed values exceeding 90%. The low identity of the vast majority of the contigs confirms the reasonability of our classification criteria.

Sequence variation of the oral phage communities can be further illustrated via two cases present in our shotgun metagenomic assemblies. In one case, all contigs with *Actinomyces* phage homologues were manually assembled to five distinct genomes more than 10 kb, which covered most contiguous regions of the reference genome (Supplementary Fig. S2). As none of the *Actinomyces* phage genome was sequenced so far except AV-1, we assigned all of the five genomes to this species with approximately 30–150 × coverages. Only one of these assembled genomes was closely related to the AV-1 at a 91% sequence identity, whereas the remaining four had relatively low sequence identities in the range of from 36% to 77%, suggesting that *Actinomyces* phages in the collected oral samples should be from at least five distinct species. In the other case, phylogenetic trees were built using the predicted gene sequences of the head protein and tail protein, both of which were two common structural genes in phages (Supplementary Fig. S3). By clustering with known phage species, both of the genes showed great diversity and variation and several predicted homologous sequences form major branches with no known phage species clustering together, which suggests oral cavity is still an unexplored niche for viruses.

From the annotated contigs, we found several 'hybrid contigs' consisting of both phage and bacteria parts (Supplementary Fig. S4). These 'hybrid contigs' were confirmed to be authentic and not assembly artefacts by both experimental and computational approaches. One of these contigs (C25356877_3.0_0.434966, ~1.2 kb) contained four predicted genes, three of which (ZP_13037070-ZP_13037072) were assigned to the bacteria *Aggregatibacter actinomycetemcomitans*, whereas the fourth (NP_852764) was annotated as gene of *Haemophilus* phage Aaphi23 with a 100% sequence identity. The same phenomenon was also observed in another contig (Supragingival_plaque_LANL_766674, ~ 1.2 kb) from the HMP assemblies. Among the three predicted genes, *Haemophilus* phage Aaphi23 was identified as source of the first gene (NP_852773), but the other two (WP_005578502 and WP_005578504) were from *A. actinomycetemcomitans*. The lengths of the three sequences are 447 bp, 321 bp and 285 bp, respectively, and all sequence identities of them exceeded 97%. These assemblies present the possibility that the bacteria *A. actinomycetemcomitans* could be invaded by some *Haemophilus* phage. Intriguingly, we also identified a reversed form of such multiple infections in contig Supragingival_plaque_LANL_767792 (~1.2 kb), which linked bacteria *Haemophilus influenzae* KR494 with phage *Aggregatibacter* phage S1249. Considering the relatively high identities in each above case and the fact that *Aggregatibacter* and *Haemophilus* are two different genera, these results strongly suggest a cross-infection relationship between the two microorganisms.

Natural structure and variation of the oral phage communities

We investigated the structure and variation of oral phage communities based on randomly sequencing of total metagenomic DNA. As shown in Supplementary Fig. S5, although family of *Siphoviridae* was described as the dominant taxon in a previous study (Ly *et al.*, 2014), we did not observe such overwhelming proportion in our data. By contrast, percentages of *Podoviridae* were consistently low according to previous observation, but it displayed high abundance in our untreated samples. To further describe community structures, the number of species and their abundance were also calculated in each sample type (Fig. 1B, right panel). For the saliva of healthy individuals, *Yersinia* phage phiR1-37 was the most abundant species, representing about 13.9% of the salivary phage communities, followed by *Lactococcus* phage KSY1 (8.4%) and *Pseudomonas* phage phiKZ (7.8%), whereas the other 53 phages have relatively low abundances (<5%). For the plaque of healthy individuals, *Lactococcus* phage 1,706 (7.5%), *Actinomyces* phage AV-1 (7.0%), *Corynebacterium* phage P1201 (5.4%) and *Streptococcus* phage (22.0%) were detected at higher proportions comparing with the salivary phage

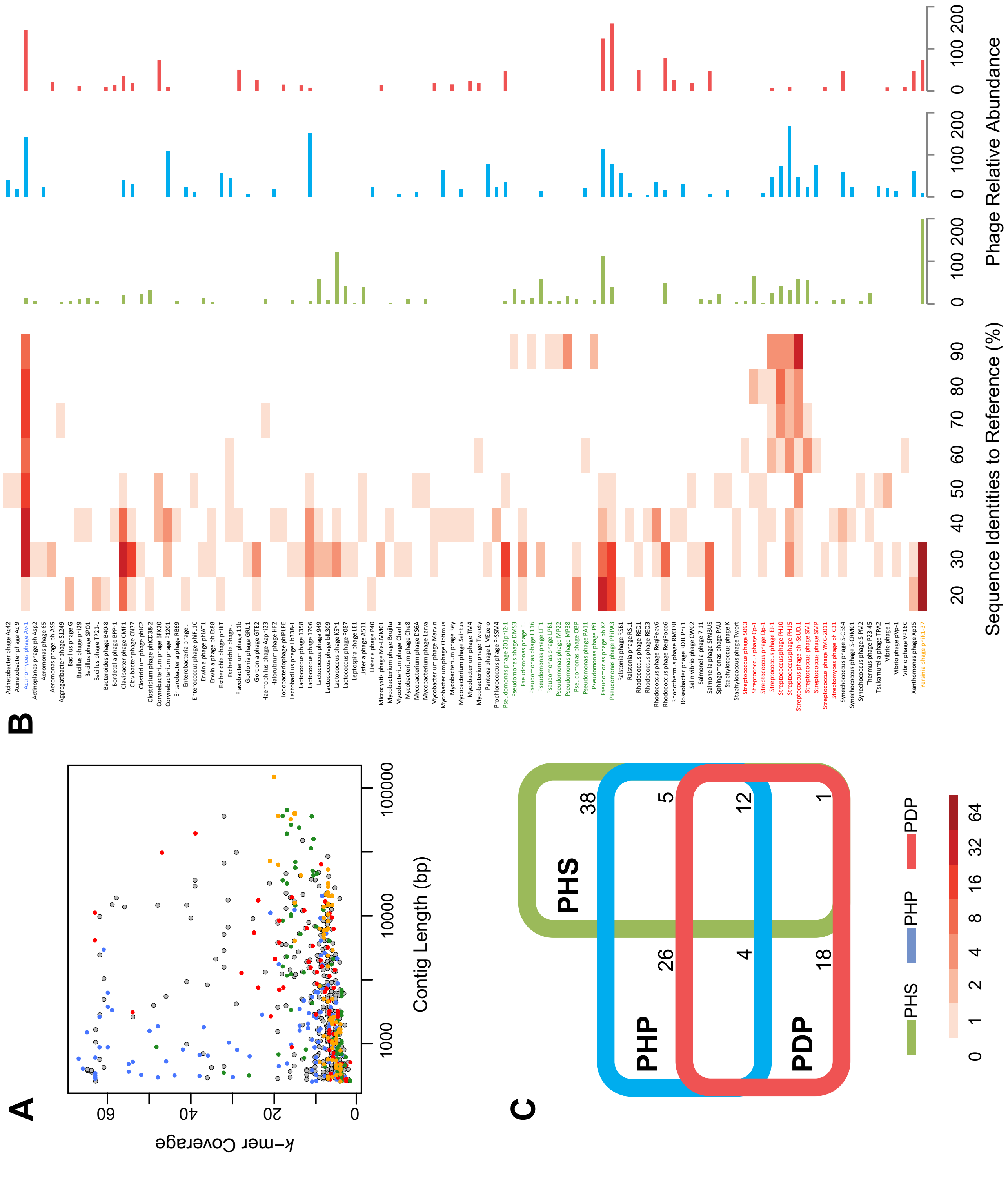
A
k-mer Coverage
0
20
40
60
Contig Length (bp)
1000
10000
100000
B
Acinetobacter phage Ac42
Acinetobacter phage Acj9
Actinomyces phage Av-1
Actinoplanes phage phiAsp2
Aeromonas phage 65
Aeromonas phage phiAS5
Aggregatibacter phage S1249
Bacillus phage G
Bacillus phage phi29
Bacillus phage SPO1
Bacillus phage TP21-L
Bacteroides phage B40-8
Bordetella phage BPP-1
Clavibacter phage CMP1
Clavibacter phage CN77
Clostridium phage phiC2
Clostridium phage phiCD38-2
Corynebacterium phage BFK20
Corynebacterium phage P1201
Enterobacteria phage RB69
Enterobacteria phage...
Enterococcus phage phiFL1C
Erwinia phage phiAT1
Erwinia phage phiEt88
Escherichia phage phiKT
Escherichia phage...
Flavobacterium phage 11b
Gordonia phage GRU1
Gordonia phage GTE2
Haemophilus phage Aaphi23
Halorubrum phage HF2
Iodobacteriophage phiPLPE
Lactobacillus phage Lb338-1
Lactococcus phage 1358
Lactococcus phage 1706
Lactococcus phage 949
Lactococcus phage bIL309
Lactococcus phage KSY1
Lactococcus phage P087
Leptospira phage LE1
Listeria phage A511
Listeria phage P40
Microcystis phage Ma-LMM01
Mycobacterium phage Brujita
Mycobacterium phage Charlie
Mycobacterium phage Che8
Mycobacterium phage DS6A
Mycobacterium phage Larva
Mycobacterium phage Marvin
Mycobacterium phage Optimus
Mycobacterium phage Rey
Mycobacterium phage Saintus
Mycobacterium phage TM4
Mycobacterium phage Tweety
Pantoea phage LIMEzero
Prochlorococcus phage P-SSM4
Pseudomonas phage 201phi2-1
Pseudomonas phage DMS3
Pseudomonas phage EL
Pseudomonas phage F116
Pseudomonas phage LIT1
Pseudomonas phage LPB1
Pseudomonas phage MP22
Pseudomonas phage MP38
Pseudomonas phage OBP
Pseudomonas phage PA11
Pseudomonas phage Pf1
Pseudomonas phage phiKZ
Pseudomonas phage PhiPA3
Ralstonia phage RSB1
Ralstonia phage RSL1
Rhodococcus phage REQ1
Rhodococcus phage REQ3
Rhodococcus phage ReqiPepy6
Rhodococcus phage ReqiPoco6
Rhodothermus phage RM378
Roseobacter phage RDJL Phi 1
Salinivibrio phage CW02
Salmonella phage 7-11
Salmonella phage SPN3US
Sphingomonas phage PAU
Staphylococcus phage K
Staphylococcus phage Twort
Streptococcus phage 5093
Streptococcus phage Cp-1
Streptococcus phage Dp-1
Streptococcus phage EJ-1
Streptococcus phage PH10
Streptococcus phage PH15
Streptococcus phage phi-SsUD.1
Streptococcus phage SM1
Streptococcus phage SMP
Streptococcus phage YMC-2011
Streptomyces phage phiC31
Synechococcus phage S-CBS4
Synechococcus phage S-CRM01
Synechococcus phage S-PM2
Thermus phage P23-45
Tsukamurella phage TPA2
Vibrio phage 1
Vibrio phage pVp-1
Vibrio phage VP16C
Xanthomonas phage Xp15
Yersinia phage phiR1-37
20
30
40
50
60
70
80
90
Sequence Identities to Reference (%)
0
100
200
0
100
200
0
100
200
Phage Relative Abundance
C
PHS
PHP
PDP
38
5
12
1
26
4
18
PHS
PHP
PDP
0
1
2
4
8
16
32
64

Fig. 1. Putative oral phages and their distribution in different sample types.
A. Characteristics of phage contigs assembled from our WGS sequencing reads. Each contig is represented by a dot. The length of the contig is shown on the *x*-axis and the *k*-mer coverage on the *y*-axis. Coloured dots show four high abundant phage contigs, which are *Actinomyces* phages (blue dots), *Pseudomonas* phages (green dots), *Streptococcus* phages (red dots) and *Yersinia* phages (orange dots).
B. Heatmap shows relative abundance (colour key) of each phage, and the sequence identity (*x*-axis) to its close relative in known viral genome databases (left panel). The bar chart shows the relative abundance (*x*-axis) of each phage taxon in saliva of periodontal health (PHS, green bar), dental plaque of periodontal health (PHP) and disease (PDP) (right panel). Relative abundance of the putative phages was evaluated by dividing the hit reads of each taxon by the total number of reads and normalizing to 10^7. Square roots were calculated for the resulting values to facilitate data visualization.
C. Shared and unique phages in comparison between PHS, PHP and PDP.

communities. As for the plaque of periodontal disease, slightly increased *Actinomyces* phage AV-1 (11.1%), and substantially decreased *Streptococcus* phage (1.9%) and *Lactococcus* phage (1.6%) were counted. In general, salivary samples contained the most diverse phage species (Fig. 1C), while plaques of periodontal disease exhibited the least diversity in spite of its much larger data size than plaques of periodontally healthy samples (130 million versus 80 million reads). Most of the phage species were unique for each sample type, and only 12 species were shared and constituted the core phage populations in the oral samples.

To inspect the variation of the phage communities between periodontally healthy and diseased samples, we measured the Shannon–Wiener diversities and Bray–Curtis dissimilarities using taxonomic matrices of the dental plaque samples (Fig. 2A and B). The median Shannon indices were 1.20 (ranging from 0.17 to 1.65) and 0.74 (ranging from 0 to 1.93) in periodontally healthy and diseased samples, respectively, but the difference is not significant ($P = 0.39$, Mann–Whitney *U* test, Fig. 2A). As to the Bray–Curtis dissimilarities (Fig. 2B), the distances between periodontally healthy samples (median = 0.93, ranging from 0.67 to 1.00) were significantly greater ($P < 0.0001$, *U* test) than distances between periodontal disease samples (median = 0.72, ranging from 0.29 to 0.94), which signifies that phage communities may differ from one another for healthy individuals, whereas showed a trend towards homogenization for samples from diseased gingivae.

To study the association between phages and bacteria in human oral cavities, we also compared the variation of bacterial communities between periodontally healthy and diseased samples (Fig. 2C and D). Comparison of alpha diversities showed that the median Shannon indices of the bacterial communities were 2.60 in periodontally healthy and 2.73 in diseased samples respectively ($P = 0.14$, *U* test, Fig. 2C). The Bray–Curtis distances between periodontally healthy samples (median = 0.48, ranging from 0.21 to 0.79) were significantly greater ($P = 0.008$, *U* test) than they were between periodontal disease samples (median = 0.40, ranging from 0.17 to 0.73, Fig. 2D). This pattern resembled what we found in the phage communities, which is suggestive of the potential contribution or response of oral phages to microbial composition and community variation.

The phage–bacteria interaction network in human oral cavities

To further investigate the interplay between oral phages and bacteria, we attempted to infer the phage–bacteria interaction according to DR and spacer sequences recorded in CRISPR arrays. A total of 124 534 bacterial genomes or scaffolds were downloaded from the HMP and used to identify CRISPR arrays. Accordingly, 1,424 DRs were recognized from the inferred CRISPR arrays. These DRs were queried to detect CRISPR

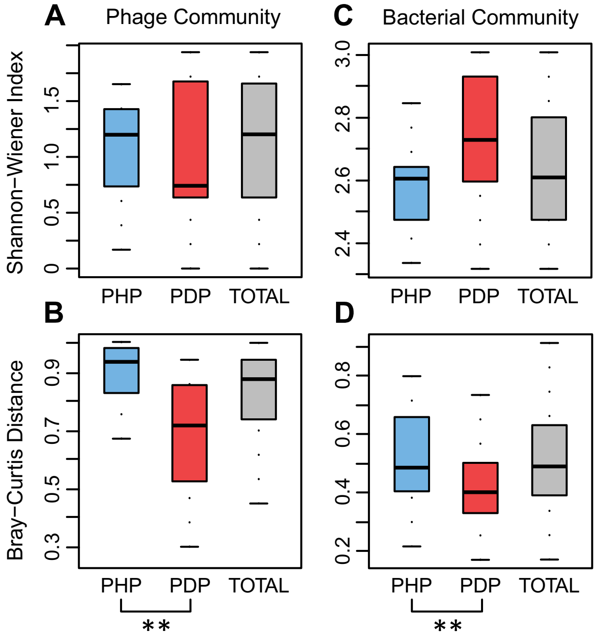

Fig. 2. Species richness and variation of oral phage communities associated with periodontal status.
A and B. Comparison of Shannon–Wiener diversities and Bray–Curtis dissimilarities of oral phage communities between PHP (blue box), PDP (red box) and total samples (grey box).
C and D. Shannon–Wiener diversities and Bray–Curtis dissimilarities of bacterial communities in PHP and PDP. The star represents the significant difference (Two asterisks denote $P < 0.01$, Mann–Whitney *U* test).

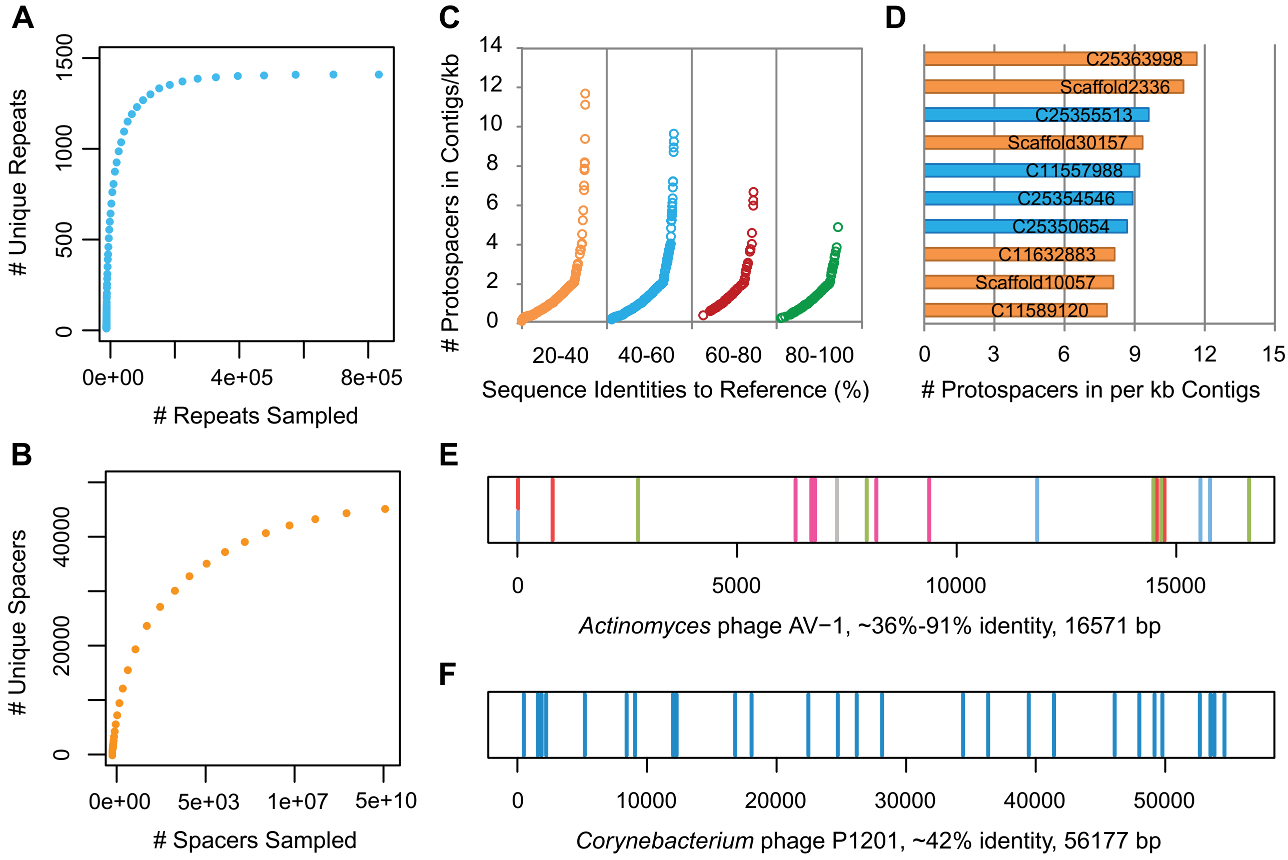

Fig. 3. CRISPR elements and protospacers identified from oral samples.
A and B. Discovery rate of new elements as a function of number of direct repeats (DRs) or spacers identified from sequencing reads. The rarefaction analysis was repeated 1,000 times with random sample order.
C. Number of protospacers located in per kb putative phage contigs. Dots in different colours represent contigs with different sequence identities to reference phage genomes.
D. Top 10 phage contigs with the most abundant protospacers. Yellow bars denote phage contigs having 20–40% sequence identities to reference phage genomes, whereas blue bars denote phage contigs having 40–60% sequence identities.
E. Protospacers located in five contigs of putative *Actinomyces* phages. To visualize the distribution of all protospacers present in *Actinomyces* phages, one of these contigs with the highest sequence identity (91%) to the genome of *Actinomyces* phages AV-1 was chosen as a template. The distinct line colours represent protospacers of different contigs.
F. Protospacers located in one phage contig having approximately 56 kb length, with 42% identity to the genome of *Corynebacterium* phage P1201.

elements using BLASTN against the sequencing reads (*E*-value < 10^{-5}). Our detection identified 966 582 DRs and 15 893 240 spacers in the sequenced samples. A total of 1,411 DRs and 46 279 spacers were consequently acquired after removing identical sequences (Fig. 3A and B). Of these detected spacers, 13.5% (6,270) could be aligned to the phage contigs that contained no corresponding DRs, and these spacers were denoted as protospacers. A significant amount of protospacers were found in putative phage contigs with sequence identities of 20–100% to reference phage genomes (Fig. 3C). The prevalence of protospacers confirmed that these contigs were likely to be derived from phages. Interestingly, we found that the most divergent contigs (20–40% sequence identity) contained relatively more protospacers than the contigs having close relatives in known phage databases, indicating that these novel phages may have higher infection activities in human oral cavities. For example, several contigs annotated as *Actinomyces* phages contained the largest number of protospacers (Fig. 3D), which is consistent with their high abundance in dental plaque samples. In addition, there was no location preference along the genomes for the detected protospacers (Fig. 3E and F).

We then established the connections between oral phages and bacteria based on the CRISPR arrays detected in the sequencing reads and the protospacers located in phage contigs and genomes. Briefly, a phage was linked to a bacterial species only if the spacers aligned to this phage were sandwiched between two DRs of that bacterial species in at least one read. In most cases, one phage contig was exclusively linked to a single bacterial species, which implies that the majority of oral phages followed a one-to-one infection model, i.e., one phage most possibly invaded only one bacterial host. For example, each putative *Actinomyces* phage or *Corynebacterium* phage contig only contained protospacers corresponding to the CRISPR elements of one species of *Actinomyces* (Fig. 3E) or *Corynebacterium* (Fig. 3F) respectively. Besides these host-specific phages, a small fraction of phages were found to connect with multiple bacterial species, which indicated that they should follow a one-to-many model. The hybrid contigs as shown in Supplementary Fig. S4 provide good evidence of a CIP and its two hosts. Based on the CRISPR arrays identified in this study, we constructed a phage–bacteria cross-infection network in the human oral cavities (Fig. 4), where nodes and edges represented phages/bacteria and their invasion relationships respectively (Fig. 5). Within this network, 27 bacterial species were connected by 36 phage genomes and 18 assembled contigs. These phages and bacteria totally formed three disjoint clusters and the largest one contained a vast majority of CIPs. Although a CIP usually linked two bacteria in the network, we observed that some phages linked three or more. For example, the phage that is represented by Node 1 (Fig. 5) jointed three bacteria, *Streptococcus anginosus*, *Actinomyces graevenitzii* and *Haemophilus parainfluenzae*. Several bacteria even formed closed loops by connections of CIPs. As the centre of the network, *Streptococcus* was the bacterial species with most phages linked (30 distinct phages), followed by *Actinomyces* (12 phages), *Fusobacterium* (six phages) and *Aggregatibacter* (six phages). Most cross-infections happened in different species of the same bacterial genus (primarily *Streptococcus* or *Actinomyces*). However, there were also many cross-infections that took place across genera or higher taxonomic levels. As expected, phages (Nodes 2 and 3 in Fig. 5) capable of invading both *Aggregatibacter* and *Haemophilus* species, which we predicted in the annotated contigs (Supplementary Fig. S4), were also observed in the network.

When classifying the bacteria in the network into types of periodontal pathogens and commensals based on their pathogenicity, we found that cross-infection mainly existed in bacterial species of the same type (i.e., within periodontal pathogens or within commensal bacteria), whereas cross-infection between the bacterial species of different types was the minority. Although the well-known pathogen of periodontitis, *Porphyromonas gingivalis*, was not included in the network, we identified *Campylobacter*, *Fusobacterium* and *Prevotella*, which were reported to form biofilm for attachment of *P. gingivalis*. Interestingly, all these three genera were linked by CIPs to commensal bacteria, which suggested that a subset of CIPs may infect both periodontal pathogens and commensals and thus probably play a regulatory role for both populations.

Statistical relation between CIPs and bacterial community

A following statistical analysis showed that periodontitis and subgingival samples had less CIPs compared with periodontally healthy and supragingival samples respectively (Supplementary Fig. S6). To detect if there is any correlation between the CIPs and bacterial communities, we compared the relative abundance of the CIPs (Fig. 6A and B, *x*-axis) with commensal bacteria (Fig. 6A, *y*-axis) or with periodontal pathogens (Fig. 6B, *y*-axis) within each of the 50 oral samples. These 50 samples included 25 HMP samples and 25 our collected samples without VLP enrichment. Reads mapping to phages or bacteria were counted to calculate relative abundances of certain group of phages (CIPs or total phages) and bacteria (commensal bacteria or periodontal pathogens) in each sample. Relative abundance of phages in each dataset was normalized to 10^7 reads, and that of bacteria was normalized to 10^5 reads. Because the abundance of phages varied greatly among specimens, we logarithmically transformed their abundance for better visualization. Strikingly, abundance of CIPs was significantly positive correlated with commensal bacteria (Fig. 6A, Pearson correlation coefficient, $R = 0.5235$, $P = 0.000113$), whereas showed a negative correlation with major periodontal pathogens (Fig. 6B, $R = -0.5179$, $P = 0.000143$), indicating the strong association between CIPs and both bacterial populations. To verify if the same tendencies also work for the total phage populations (including CIPs and other identified phages), we further evaluated abundance variations of total phages (Fig. 6C and D, *x*-axis) and commensal bacteria (Fig. 6C, *y*-axis) or periodontal pathogens (Fig. 6D, *y*-axis) for the same samples. However, none of the comparisons exhibited significant correlations ($R = 0.1197$–0.2245), which suggested that these CIPs are not randomly selected subsets of the total phage community and their potential roles in regulating microbial community structure are indispensable.

Discussion

We previously demonstrated that the abundance of certain oral phages varied with different periodontal status (i.e., periodontal health or disease) and sample types (i.e., supragingival or subgingival plaque), and their abundance was consistent with corresponding bacterial hosts (Wang *et al*., 2013), which suggested possible involvement of phages in shaping oral microbial ecosystems. Accordingly, a metagenomic survey on the composition and variation of oral phage populations was performed in this study in order to provide expanding insights into the phage–bacteria interaction and its association with the variation of microbial community in the human oral cavities. Our extremely deep WGS sequencing generated

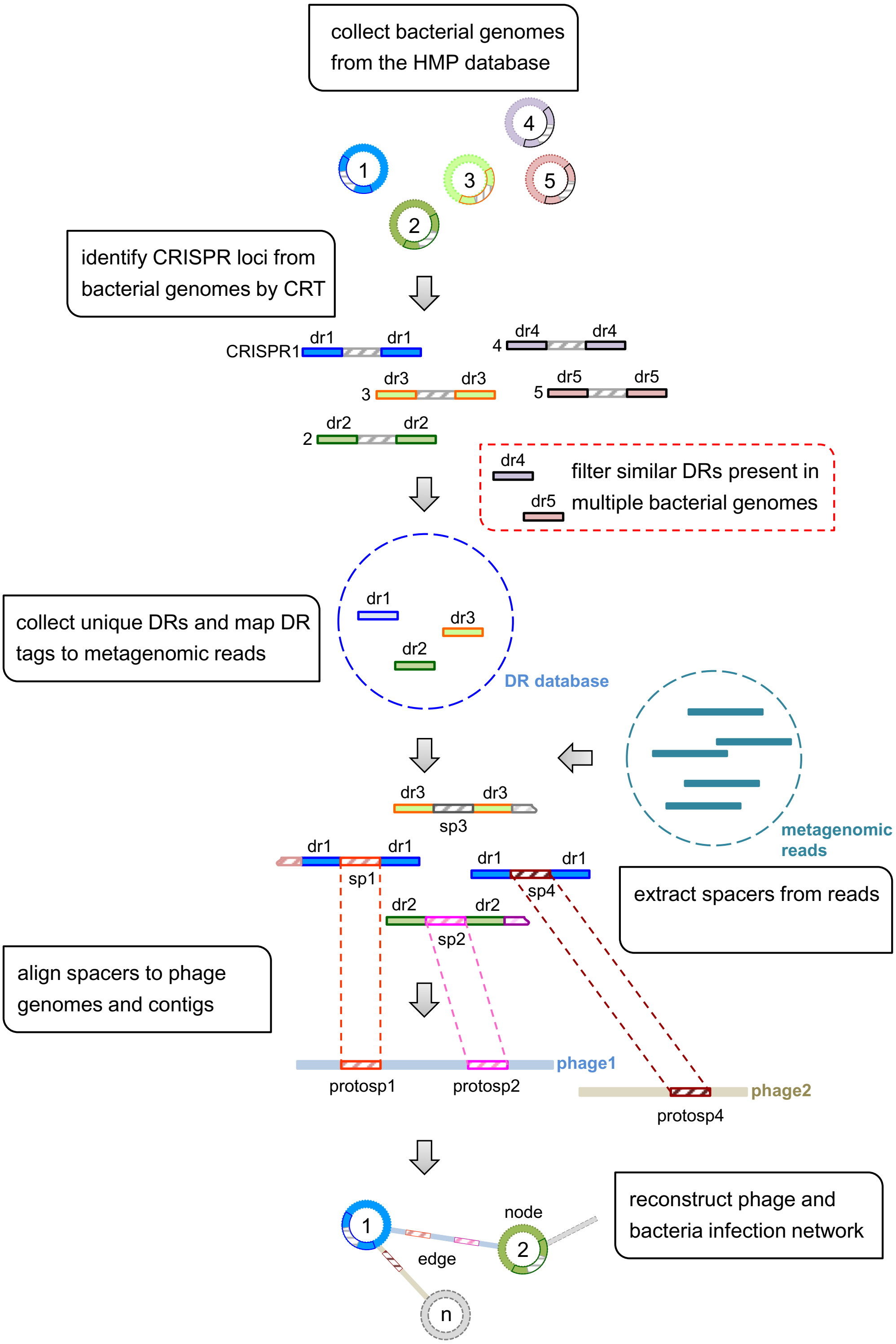
collect bacterial genomes from the HMP database
4
1
3
5
2
identify CRISPR loci from bacterial genomes by CRT
CRISPR1
dr1
dr1
4
dr4
dr4
3
dr3
dr3
5
dr5
dr5
2
dr2
dr2
dr4
dr5
filter similar DRs present in multiple bacterial genomes
collect unique DRs and map DR tags to metagenomic reads
dr1
dr3
dr2
DR database
metagenomic reads
dr3
dr3
sp3
dr1
dr1
sp1
dr1
dr1
sp4
dr2
dr2
sp2
extract spacers from reads
align spacers to phage genomes and contigs
phage1
protosp1
protosp2
phage2
protosp4
reconstruct phage and bacteria infection network
node
1
edge
2
n

Fig. 4. An analysis workflow of phage–bacteria interaction network. Briefly, bacterial genomes were first downloaded from the HMP database and identify CRISPR loci in them by CRT. Then similar DRs shared by different bacterial species were filtered, and only unique DRs were used to build a DR database. Based on this DR database, CRISPR loci in NGS reads were identified. CRISPR-contained reads were required to be aligned at least twice by the same DR in the same direction and the interspace [spacers (Sp)] between each two alignments should be no shorter than 20 bp. The spacers were anchored and extracted from NGS reads. The resulting spacers were aligned to phage contigs and genomes retrieved from NCBI and phantome to detect protospacers using BLASTN. According to the connection between DR-spacer-phage [protospacer, (Protosp)], phage–bacteria infection network was reconstructed. In this network, bacteria with unique DR and phages carrying two or more protospacers were taken as nodes while the CRISPR arrays detected in NGS reads were edges.

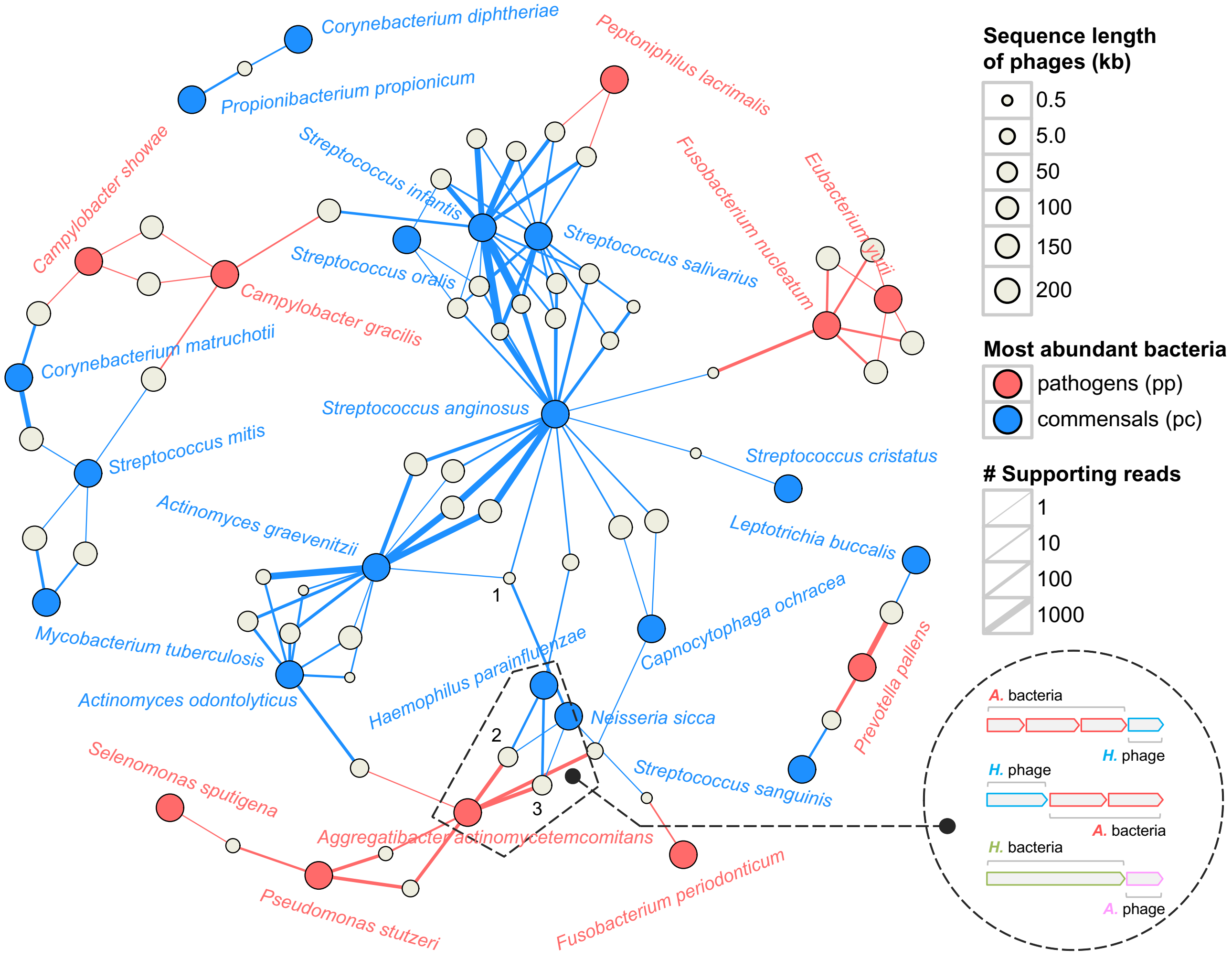

Fig. 5. A phage–bacteria interaction network between cips and the most abundant bacteria in the human oral cavities. Within the network, phages contigs or genomes are represented as white nodes. The size of these nodes indicates their sequence length. The most abundant bacteria are shown as coloured nodes, which are classified into periodontal pathogens (red nodes) and commensal bacteria (blue nodes). The edge between a phage node and a bacteria node indicates their infection history which is recorded in the CRISPR arrays. The edge width represents the number of reads supporting this connection. The black dotted circle in the lower right shows the schematic diagram of three 'hybrid contigs' consisting of both phage and bacteria parts. Red, blue, green and pink pentagons represent predicted genes of *A. actinomycetemcomitans* (*A.* Bacteria), *Haemophilus* phage (*H.* Phage), *Haemophilus influenza* (*H.* Bacteria) and *Aggregatibacter* phage (*A.* Phage) respectively.

more than 1 billion NGS reads, which is the largest metagenomic data for human oral cavities reported in a single study so far.

Previous studies revealed that the success rate for assigning short reads to known viruses was extremely low (Bench *et al.*, 2007; Xu *et al.*, 2011), especially for highly divergent or novel viruses. Unfortunately, very few phage species in oral virome have been described except for the highly abundant phages such as *Streptococcus* phage, *Actinomyces* phage, *Veillonella* phage and *Leptotrichia* phage (Willner *et al.*, 2011; Pride *et al.*, 2012b; Robles-Sikisaka *et al.*, 2013). A similar situation was found in our previous work (Wang *et al.*, 2013), in which only several dominant oral phages were detected despite we merged 2 × 101 bp PE-reads into ~180 bp sequences. In this study, we first performed *de novo* assembly from all sequencing reads, and then used long contig sequences to determine their taxonomic classification. In addition, the taxonomic classification for many of these phage contigs was further evaluated by protospacer locations, particularly for those with low sequence identities (36–44%) to reference phage genomes (Fig. 3C–F). Although more than 100 phage species are identified in this study, this number is still likely to be underestimated considering high phylogenetic diversities and low similarities of related phage contigs to each other with the same reference (Fig. 1B, Supplementary Figs S2, S3).

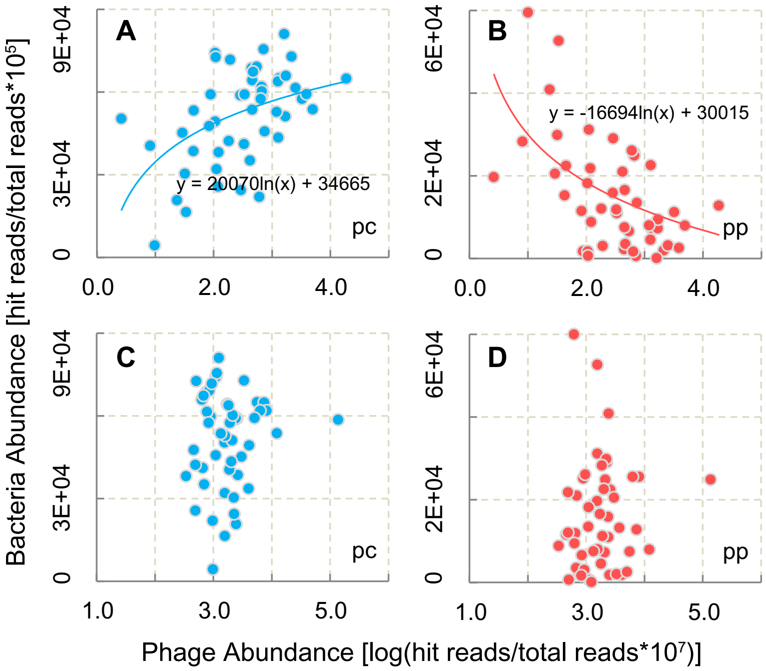

Fig. 6. Quantitative relation between oral phages and bacterial communities.
A and B. The correlation between the abundance of cips (*x*-axis) and the abundance of periodontal pathogens (red dots) and commensal bacteria (blue dots) (*y*-axis). Red and blue curves are the logarithmic trend lines.
C and D. The correlation between the abundance of all oral phages (*x*-axis) and bacterial populations (*y*-axis). Each dot represents one sample ($n = 50$). Reads mapped to phages or bacteria were counted to calculate the abundance of certain group of phages (cips or total phages) and bacteria (commensal bacteria or periodontal pathogens) in each sample. Relative abundance of bacteria was normalized to 10^5, and that of phages was normalized to 10^7. Because the abundance of phages varied greatly among specimens, it was logarithmically transformed to facilitate visualization.

Our findings on the oral phage communities provided a new understanding of their richness and variations, and further confirmed their associations with different periodontal status and sample types (Fig. 2). In the comparison of the phage communities under different periodontal status, the species richness of oral phages in people with healthy gingiva was slightly higher than in periodontitis individuals (Fig. 2A), similar to the gut viral communities associated with Crohn's disease (Perez-Brocal *et al.*, 2013). In addition, community structure was highly similar to each other for periodontitis, whereas it was more distinct between periodontally healthy individuals (Fig. 2B). This clustering pattern was also observed in a recent study, in which the VLPs were enriched from periodontal samples and amplified by multiple displacement amplification (MDA) (Ly *et al.*, 2014). These findings in several studies together suggested that the variation of oral phage communities should be involved in various human diseases. However, the variation trends are not the same for every disease. Similar with periodontitis, smaller variations were observed within the cystic fibrosis group than the healthy group (Willner *et al.*, 2009). By contrast, a larger variability were found between Crohn's disease samples rather than healthy samples (Perez-Brocal *et al.*, 2013). As oral cavities and respiratory tracts represent relatively open environments, whereas intestinal tracts are closed spaces, the divergence of the variation trends may be largely due to the environmental difference. Besides that, distinct pathogenesis of these diseases may also be one of many possible factors resulting in the divergence.

Now it is generally believed that periodontitis is closely related to oral microflora disorders (Darveau, 2010; Curtis *et al.*, 2011). Significant difference has been uncovered in bacterial communities between periodontally healthy people and patients suffered from periodontitis either by 16S pyrosequencing or by WGS sequencing (Griffen *et al.*, 2012; Liu *et al.*, 2012; Wang *et al.*, 2013; Li *et al.*, 2014). In our present work, the variation tendency of bacterial community structure is consistent with that of phage communities (Fig. 2B and D), implying that phages may play some roles in morphogenesis of oral microbial community and consequently affect the periodontally healthy status. However, the correlated tendency cannot provide details of the links between the phages and their bacterial hosts, let alone recognition of specific phages that may participate in regulating the oral microbial ecology. Therefore, we built the phage–bacteria interaction relationship according to DRs and spacers recorded in the CRISPR arrays and statistically evaluated the correlations between certain phages and bacterial community. According to CRISPR arrays, most phages only connect with their putative bacterial hosts, indicating that one phage normally invaded one bacterial species only. Notably, there are some phages connected with more than one bacterial species, which suggests the likelihood of cross-infection (Fig. 5). Actually, cross-infections have long been verified in many studied ecosystems (Weitz *et al.*, 2013). Although none of the phages was recognized to be cross-infective in human ecosystem so far, a previous study discovered that spacers of streptococcal CRISPRs in human saliva were not always aligned to streptococcal genomes, and they speculated the presence of phages with broad host range (Pride *et al.*, 2011). Interestingly, in some of our contigs, fragment of a phage species is connected with another bacterial species rather than its putative bacterial host (Supplementary Fig. S4), further confirming the existence of cross-infection in oral microbes.

Here we summarize several evidence that supports the connections in the cross-infection network (Supplementary Fig. S7), including that (i) dozens of distinct reads support one connection; (ii) spacers supporting one connection can be aligned to different loci of the same phage genome or contig; (iii) multiple phage genomes or contigs link the same two bacteria species; and (iv) certain connections such as cross-infection between bacteria *Aggregatibacter* and *Haemophilus* can also be found in metagenomic assemblies of both the HMPs and ours (Supplementary Fig. S4). Moreover, the connection in the network is independent of the number of reference phage genomes, assembled phage contigs, and the abundance of bacteria in oral cavities. For example, although more than 10 genomes of *Streptococcus* phages have been sequenced while only one genome of *Actinomyces* phages could be taken as reference, both *Streptococcus* phages and *Actinomyces* phages are dominant taxon in the network. In another case, plenty of *Pseudomonas* phage contigs were assembled in this study, but few of them are recognized as CIPs. Additionally, only one species of *Capnocytophaga* and *Prevotella* appear in the network, respectively, though both of them are highly abundant oral bacteria with multiple species (Costello *et al.*, 2009; Huttenhower *et al.*, 2012). As for the cross-infection between *Campylobacter* and *Streptococcus*, CRISPR array-based analyses indicated that six distinct reads support the connection between *Campylobacter gracilis* and *S. infantis* (Fig. 5 and Supplementary Fig. S7C), although no solid experimental evidence supports such a connection. Moreover, *C. gracilis* is connected with *S. mitis* by another phage genome, which further reveals the possibility of cross-infection between these two bacterial genera. These findings indicate the robustness of this network and reveal that the CIPs we found are not likely to be false positives generated from noises in WGS sequencing or microbial community background. However, several potential limitations of our method might be silent on some phage–host interactions and thus simplify the network diagram. For example, (i) not all interactions are mediated by CRISPR/Cas immune defense (especially for bacteria without CRISPR arrays) and may never have the chance to record such interactions; (ii) some CRISPR-mediated interactions may not be recorded in CRISPR arrays; and (iii) the relatively small size of CRISPR arrays in bacteria genomes may need a great amount of sequencing reads to cover them. In addition, we filtered out the DRs that were 'similar' across bacterial species because these 'similar' DRs would confuse us in determining host source. To ensure robustness of the network, only the DRs and those belonging exclusively to unique bacterial species were retained to build a custom database, although we know some real biological signals might be lost.

Quantitative comparisons between oral phages and bacterial communities in our study provide statistical evidences that CIPs are not a randomly selected subset of total community (Fig. 6). Significant correlation suggests there might be a close relationship between these phages and bacterial population dynamics. In previous studies, local adaptation of phages to their bacterial hosts was widely confirmed (Vos *et al.*, 2009; Koskella *et al.*, 2011). By analyzing the data from laboratory culture, Brown *et al.* demonstrated that high abundant susceptible strains of *Escherichia coli* spurred phage proliferation, which would feed back to control bacterial richness (Brown *et al.*, 2006). Higher richness of *Flectobacillus* was found after adding phages into ecosystem in another experiment using natural populations and *in situ* incubations (Weinbauer *et al.*, 2007). In our study, we found that CIPs were more abundant in samples rich in periodontal commensal bacteria (Fig. 6A). This positive correlation raises the possibility that abundance variances of these CIPs may affect populations of multiple hosts and consequently maintain diversity and dynamic balance of oral bacterial community. This was also supported by our observation of the relative stability of alpha-diversity in bacterial communities from different oral samples (Fig. 2B). Quantitative comparisons were only conducted based on the samples originated from directly metagenomic sequencing in this study, because our virus-enriched (15 samples pooled into 3, filtered, precipitated and then amplified) samples were not quantitative data. Such enrichment may bring in biases. For example, only most known viruses distributed in certain CsCl density layers could be selectively enriched and further required amplification (Thurber *et al.*, 2009). It is also well known that whole-genome amplification, such as the linker amplified shotgun library and MDA, usually produce artefacts and tend to amplify certain types of viruses (Kim and Bae, 2011), which hampers quantitative analysis of viromes (Yilmaz *et al.*, 2010). The sample size involved in statistical analysis would be increased if quantitative linker amplification approaches for virome were used (Duhaime and Sullivan, 2012; Duhaime *et al.*, 2012; Hurwitz *et al.*, 2013; Solonenko *et al.*, 2013; Sullivan, 2015).

It is well known that CIPs can prey on other bacteria besides their common prey species. For instance, cyanophages infecting cyanobacterium *Prochlorococcus* can also attack *Synechococcus* (Sullivan *et al.*, 2003). As shown in our phage–bacteria interaction network, although most of CIP connections occurred within periodontal pathogens or commensal bacteria, a subset of them linked commensal bacteria to periodontal pathogens (Fig. 5). Accordingly, the high number of CIPs in oral environment may also directly or indirectly restrain growth of pathogenic bacteria when regulating commensal flora, which also accounts for a significant negative correlation between CIPs and periodontal pathogens (Fig. 6B). In other words, the richness of periodontal pathogens would be increased if the biomass of CIPs decreases. A similar process was shown in one experiment of artificially removing phages from environmental samples, which resulted in increases of several previously rare bacterial species (Bouvier and del Giorgio, 2007). According to these findings, we speculate that CIPs rather than the whole phage population may play some roles in forming microbial populations and communities. These phages might be important factors that engage in preventing or aggravating deterioration of the oral ecosystem. The potential role also suggests a possible utilization of CIPs on oral microbe biocontrol and phage therapy of oral diseases.

Experimental procedures

Ethics statement

Subject recruitment and sample collection were approved by the Ethical Committee of Peking University and performed at the School and Hospital of Stomatology, Peking University. All subjects provided written informed consent before participation.

Subject recruitment and sample collection

Inclusion criteria for subjects were mature non-smoking people, 30–65 years of age, free of systemic diseases, without prosthetic dental appliances, had never received professional cleaning or other periodontal therapy, had not taken antibiotics in the past 3 months, and at least 6 h after tooth brushing and 2 h after eating. After clinically monitoring the periodontal status, subjects were divided into chronic periodontitis patients or healthy controls. Periodontal health was characterized as no probing depth or attachment loss exceeding 2 mm at any site. Periodontitis was required to exhibit no less than 4 mm probing depth and 6 mm attachment loss in at least four non-adjacent interproximal sites.

We collected a total of 40 oral specimens from 40 human subjects. Twenty dental plaque samples were collected from 10 periodontitis patients and 10 periodontally healthy subjects. Additionally, to gather more oral phages, approximately 5 ml saliva was collected from each of 20 healthy subjects. For each individual, dental plaques of 3–4 sites were sampled and pooled together as one sample to get adequate yields of microbial DNA for WGS sequencing. Each plaque sample was placed in a sterile 1.5 ml centrifuge tube containing 50 μl phosphate buffer solutions (PBS; pH = 8.0). Each salivary sample was kept in a sterile 50 ml centrifuge tube. All samples were collected by dentists or periodontists in the hospital of stomatology under strict uniform protocol. After collection, specimens were immediately transported to the laboratory and stored at −80 °C until DNA extraction.

Specimen processing and WGS sequencing

Metagenomic DNA was isolated from 25 (20 dental plaque specimens and five salivary) specimens individually using a QIAamp DNA Mini Kit (Qiagen). The other 15 salivary specimens were pooled into three samples and precipitated VLPs using PEG 8000 (Amresco, Solon, OH, USA). Briefly, approximately 25 ml pooled saliva was homogenized in equal volumes of PBS solution and then centrifuged at 4,500 g for 30 min to remove cellular debris. The supernatant was serially filtered through 0.45 and 0.22 μm filters. The filtrate was stored overnight at 4°C after adding PEG 8000 at a final concentration of 10%, and subsequently centrifuged at 13 000 *g* for 30 min to gather pellets. DNA was isolated from the resultant pellets using the same reagent and protocol as described above and amplified with REPLI-g Midi polymerase (Qiagen).

For the preparation of sequencing libraries, either metagenomic DNA (~0.5 μg DNA per sample, 25 samples) or whole-genome amplified DNA (~5 μg, three pooled samples) were quantified using a Qubit Fluorometer (Invitrogen) before sheared into about 180 bp fragments with a Covaris Focused-ultrasonicator (Covaris). Libraries were prepared sequentially on the sheared fragments using a TruSeq Sample Preparation Kit v2 (Illumina) according to the manufacturer's instructions. Fragment size and integrity of the libraries were assessed by a Fragment Analyzer (Advanced Analytical Technologies). Libraries with different indexes were pooled into three lanes and 2 × 101 bp paired-end (PE) sequenced on an Illumina HiSeq 2000 platform in the Research Facility Center at Beijing Institutes of Life Science (BIOLS), Chinese Academy of Sciences. The sequencing data were deposited under accession number SRP033553 in NCBI SRA.

Additional specimen retrieval and data acquirement

Additional 25 WGS datasets of dental plaque samples were obtained from the HMP database (http://www.hmpdacc.org/). Basic information on these datasets was provided in the Supplementary Materials Table S1. Three assembly files generated from salivary, subgingival and supragingival samples as well as bacterial genomes in the HMP database were also downloaded for CRISPR identification and network reconstruction. Viral and phage genomes were retrieved from NCBI (ftp://ftp.ncbi.nih.gov/genomes) and the PhAnToMe website (http://www.phantome.org/Downloads/DNA/all_sequences/), and these genomes were then integrated to construct a custom database after removing duplicates.

De novo *assembly and sequence annotation*

Low-quality PE reads were filtered from raw sequencing data of each sample before assembly using Illumina's CASAVA pipeline v1.8 with default parameters. The high-quality reads then were aligned to a human genome assembly hg19 using the backtrack algorithm of BWA v0.5.9 (Li and Durbin, 2009), allowing up to five mismatches per 100 bp, to remove contaminating human sequences. The remaining non-human reads were classified and merged into three datasets according to the sample types. The merged datasets were independently using SOAPdenovo2 with the multi-*k*-mer option (-K 41 -m 63) (Li *et al.*, 2010). Only contigs no shorter than 500 bp were retained for further analysis. These contigs and those retrieved from the HMP were submitted to predict open reading frames (ORFs) using MetaGeneMark v2.8 (Zhu *et al.*, 2010), and the predicted proteins were subsequently compared against the NCBI non-redundant (nr) amino acid sequence database using BLASTP. For each query protein, S_{best} was set as the bitscore of the best BLAST hit, and all the BLAST hits that have bitscore higher than S_{best}*0.95 were collected. According to the functional annotation of these collected BLAST hits, a majority-rule consensus approach was applied to determine the function of the query protein. Gene sequences of the predicted head protein and tail protein were extracted from the annotated contigs, translated to amino acid sequences and built a phylogenetic tree with 1,000 bootstraps by MEGA 5.02 (Tamura *et al.*, 2011).

Taxonomic classification and abundance estimation

For taxonomic classification of oral phages in the annotated datasets, the lowest common ancestor (LCA) of all taxa in the collected BLAST hits was computed and used to determine taxonomic origin. A contig was classified as a putative phage if all annotated ORFs in this contig were assigned to the phage or phage ORFs in this contig were more than annotated ORFs of any other taxa. Relative abundance of recognizable phage species was evaluated for the samples originated from metagenomic DNA of saliva and dental plaque. As for the samples originated from whole-genome amplified DNA, taxonomic classification other than abundance estimation was performed to facilitate phage contig construction and genome assembly but avoid introducing unintended biases. The reads mapping to the given contigs or genomes were counted to calculate abundances of phages and bacteria in each sample.

CRISPR identification and network reconstruction

We adopted a modified pipeline used in a previous study for searching of DRs and spacers (Stern *et al.*, 2012). DR identification was first performed by CRT using default settings for each bacterial genome downloaded from the HMP database (Bland *et al.*, 2007). The identified DRs were then integrated for searching of spacers. Briefly, the database constructed based on the DRs was queried by 100 bp metagenomic reads of each sample using BLASTN. Two hits of DR in the same direction on a read with a minimum of 90% identity along 90% of its length were accepted when the interspace is longer than 20 bp. When only one such hit was found, an additional search at both ends of the read will retrieve potentially incomplete DR longer than 8 bp. We then extracted the interspaces from reads as CRISPR spacers and performed a following BLASTN analysis of the spacers against identified phage contigs and genomes mentioned above, in which a significant hit (identity ≥ 90%, length ≥ 20 bp and e-value ≤ 0.005) inferred phage source of the spacer (protospacer). Thus, the above processes associated the phage contigs and genomes with the initial bacterial genomes through the CRISPR arrays detected in the metagenomic reads and were used for reconstruction of the bacteria and phage infection network. It should be noted that several necessary filtration steps were also adopted within the procedure to avoid possible false positives. For example, we filtered out similar DR sequences shared by different bacterial species that would confuse following analysis of cross-infection as their host source cannot be determined. Another filtration was aimed at potentially wrongly identified phage contigs that were matched by spacers, in which all contigs that could also be matched by corresponding DR of the spacers at a lower threshold (e-value ≤ 0.1) were removed. To make this method easier to follow, a flowchart has been provided (Fig. 4). We then focused on the most abundant oral bacteria for two groups (periodontal pathogens and commensals) reported previously (Wang *et al.*, 2013). All aligning contigs of spacers that have a direct or indirect connection with these bacterial hosts were used for network reconstruction.

Statistical analysis and data visualization

Shannon–Wiener indexes and Bray–Curtis distances were calculated by the R package vegan (http://CRAN.R-project.org/package=vegan). Briefly, the taxonomic matrices at the genus level were used as inputs, where each element was the number of phages or bacteria from a given sample. The variation of dissimilarities was compared between plaques of periodontally healthy and diseased samples, and the significance of the variation was assessed by Mann–Whitney rank sum test. The threshold of statistical significance was set at $P < 0.05$. To visualize the phage–bacteria interaction, a custom R script and the igraph package (Csardi and Nepusz, 2006) were used to plot the network that taking phage and bacteria as nodes and their predation relationships recorded in CRISPR arrays as edges. Pearson correlation coefficients were computed to evaluate the relationships between phages and their bacterial hosts. A Bonferroni correction was used to test statistical significance of the Pearson correlations.

Authors' contributions

J. Wang performed the experiment and data analysis, interpreted the results and wrote the manuscript. Y. Gao contributed to data analysis and result interpretation, and edited the manuscript. F. Zhao designed the study, assembled and annotated contigs, and revised the manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (91131013, 31300110). We thank S. Wei and Y. Zhang (School and Hospital of Stomatology, Peking University) for their gracious help in sampling. We also thank L. Hou and L. Zhang (Research Facility Center at Beijing Institutes of Life Science, Chinese Academy of Science) for their sequencing assistance.

Competing interests

The authors declare no competing interests.

References

Aagaard, K., Ma, J., Antony, K.M., Ganu, R., Petrosino, J., and Versalovic, J. (2014) The placenta harbors a unique microbiome. *Sci Transl Med* **6:** 237ra65.

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.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., *et al*. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315:** 1709–1712.

Bench, S.R., Hanson, T.E., Williamson, K.E., Ghosh, D., Radosovich, M., Wang, K., and Wommack, K.E. (2007) Metagenomic characterization of Chesapeake bay virioplankton. *Appl Environ Microbiol* **73:** 7629–7641.

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. *BMC Bioinform* **8:** 209.

Bouvier, T., and del Giorgio, P.A. (2007) Key role of selective viral-induced mortality in determining marine bacterial community composition. *Environ Microbiol* **9:** 287–297.

Brown, S.P., Le Chat, L., De Paepe, M., and Taddei, F. (2006) Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. *Curr Biol* **16:** 2048–2052.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009) Bacterial community variation in human body habitats across space and time. *Science* **326:** 1694–1697.

Csardi, G., and Nepusz, T. (2006) The igraph software package for complex network research. *Inter J, Complex Syst* **1695:** 1–9.

Curtis, M.A., Zenobia, C., and Darveau, R.P. (2011) The relationship of the oral microbiotia to periodontal health and disease. *Cell Host Microbe* **10:** 302–306.

Darveau, R.P. (2010) Periodontitis: a polymicrobial disruption of host homeostasis. *Nat Rev Microbiol* **8:** 481–490.

Donlan, R.M. (2009) Preventing biofilms of clinically relevant organisms using bacteriophage. *Trends Microbiol* **17:** 66–72.

Duhaime, M.B., and Sullivan, M.B. (2012) Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **434:** 181–186.

Duhaime, M.B., Deng, L., Poulos, B.T., and Sullivan, M.B. (2012) Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ Microbiol* **14:** 2526–2537.

Flores, C.O., Meyer, J.R., Valverde, S., Farr, L., and Weitz, J.S. (2011) Statistical structure of host-phage interactions. *Proc Natl Acad Sci USA* **108:** E288–E297.

Flores, C.O., Valverde, S., and Weitz, J.S. (2013) Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *Isme J* **7:** 520–532.

Genco, R.J., and Van Dyke, T.E. (2010) Reducing the risk of CVD in patients with periodontitis. *Nat Rev Cardiol* **7:** 479–480.

Griffen, A.L., Beall, C.J., Campbell, J.H., Firestone, N.D., Kumar, P.S., Yang, Z.K., *et al*. (2012) Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *Isme J* **6:** 1176–1185.

Holmfeldt, K., Middelboe, M., Nybroe, O., and Riemann, L. (2007) Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their *Flavobactetium* hosts. *Appl Environ Microbiol* **73:** 6730–6739.

Horvath, P., and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327:** 167–170.

Hurwitz, B.L., Deng, L., Poulos, B.T., and Sullivan, M.B. (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol* **15:** 1428–1440.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., *et al*. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207–214.

Kim, K.H., and Bae, J.W. (2011) Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol* **77:** 7663–7668.

Koren, O., Spor, A., Felin, J., Fak, F., Stombaugh, J., Tremaroli, V., *et al*. (2011) Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc Natl Acad Sci USA* **108:** 4592–4598.

Koskella, B. (2013) Phage-mediated selection on microbiota of a long-lived host. *Curr Biol* **23:** 1256–1260.

Koskella, B., Thompson, J.N., Preston, G.M., and Buckling, A. (2011) Local biotic environment shapes the spatial scale of bacteriophage adaptation to bacteria. *Am Nat* **177:** 440–451.

Kunin, V., He, S., Warnecke, F., Peterson, S.B., Martin, H.G., Haynes, M., *et al*. (2008) A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res* **18:** 293–297.

Levin, B.R., and Bull, J.J. (2004) Population and evolutionary dynamics of phage therapy. *Nat Rev Microbiol* **2:** 166–173.

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li, R.Q., Zhu, H.M., Ruan, J., Qian, W.B., Fang, X.D., Shi, Z.B., *et al*. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20:** 265–272.

Li, Y., He, J., He, Z., Zhou, Y., Yuan, M., Xu, X., *et al*. (2014) Phylogenetic and functional gene structure shifts of the oral microbiomes in periodontitis patients. *Isme J* **8:** 1879–1891.

Liu, W., Schat, H., Bliek, M., Chen, Y., McGrath, S.P., George, G., *et al*. (2012) Knocking out ACR2 does not affect arsenic redox status in *Arabidopsis thaliana*: implications for as detoxification and accumulation in plants. *PLoS ONE* **7:** e42408.

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.

Methe, B.A., Nelson, K.E., Pop, M., Creasy, H.H., Giglio, M.G., Huttenhower, C., *et al.* (2012) A framework for human microbiome research. *Nature* **486:** 215–221.

Perez-Brocal, V., Garcia-Lopez, R., Vazquez-Castellanos, J.F., Nos, P., Beltran, B., Latorre, A., and Moya, A. (2013) Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. *Clin Transl Gastroenterol* **4:** e36.

Pride, D.T., Sun, C.L., Salzman, J., Rao, N., Loomer, P., Armitage, G.C., *et al.* (2011) Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Res* **21:** 126–136.

Pride, D.T., Salzman, J., and Relman, D.A. (2012a) Comparisons of clustered regularly interspaced short palindromic repeats and viromes in human saliva reveal bacterial adaptations to salivary viruses. *Environ Microbiol* **14:** 2564–2576.

Pride, D.T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R.A., *et al.* (2012b) Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *Isme J* **6:** 915–926.

Qin, J.J., Li, Y.R., Cai, Z.M., Li, S.H., Zhu, J.F., Zhang, F., *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490:** 55–60.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S.K., McCulle, S.L., *et al.* (2011) Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci USA* **108:** 4680–4687.

Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., and Gordon, J.I. (2012) Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol* **10:** 607–617.

Reyes, A., Wu, M., McNulty, N.P., Rohwer, F.L., and Gordon, J.I. (2013) Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proc Natl Acad Sci USA* **110:** 20236–20241.

Robles-Sikisaka, R., Ly, M., Boehm, T., Naidu, M., Salzman, J., and Pride, D.T. (2013) Association between living environment and human oral viral ecology. *Isme J* **7:** 1710–1724.

Rohwer, F., and Thurber, R.V. (2009) Viruses manipulate the marine environment. *Nature* **459:** 207–212.

Roux, S., Hawley, A.K., Beltran, M.T., Scofield, M., Schwientek, P., Stepanauskas, R., *et al.* (2014) Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta- genomics. *Elife* **3**.

Solonenko, S.A., Ignacio-Espinoza, J.C., Alberti, A., Cruaud, C., Hallam, S., Konstantinidis, K., *et al.* (2013) Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14:** 320.

Stern, A., Mick, E., Tirosh, I., Sagy, O., and Sorek, R. (2012) CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res* **22:** 1985–1994.

Sullivan, M.B. (2015) Viromes, not gene markers, for studying double-stranded DNA virus communities. *J Virol* **89:** 2459–2461.

Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424:** 1047–1051.

Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005) Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3:** 790–806.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28:** 2731–2739.

Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4:** 470–483.

Vos, M., Birkett, P.J., Birch, E., Griffiths, R.I., and Buckling, A. (2009) Local adaptation of bacteriophages to their bacterial hosts in soil. *Science* **325:** 833.

Wang, J.F., Qi, J., Zhao, H., He, S., Zhang, Y.F., Wei, S.C., and Zhao, F.Q. (2013) Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Sci Rep* **3:** 1843.

Weinbauer, M.G., Hornak, K., Jezbera, J., Nedoma, J., Dolan, J.R., and Simek, K. (2007) Synergistic and antagonistic effects of viral lysis and protistan grazing on bacterial biomass, production and diversity. *Environ Microbiol* **9:** 777–788.

Weitz, J.S., Poisot, T., Meyer, J.R., Flores, C.O., Valverde, S., Sullivan, M.B., and Hochberg, M.E. (2013) Phage-bacteria infection networks. *Trends Microbiol* **21:** 82–91.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.E., Silva, J., *et al.* (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* **4:** e7370.

Willner, D., Furlan, M., Schmieder, R., Grasis, J.A., Pride, D.T., Relman, D.A., *et al.* (2011) Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc Natl Acad Sci USA* **108:** 4547–4553.

Xu, B.L., Liu, L.C., Huang, X.Y., Ma, H., Zhang, Y., Du, Y.H., *et al.* (2011) Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. *PLoS Pathog* **7:** e1002369.

Yilmaz, S., Allgaier, M., and Hugenholtz, P. (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods* **7:** 943–944.

Zhu, W.H., Lomsadze, A., and Borodovsky, M. (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38:** e132.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Numbers of ORFs in each putative phage contig. A contig was classified as a putative phage if all annotated ORFs in this contig were assigned to the phage or phage

ORFs in this contig were more than annotated ORFs of any other phages. The number of phage ORFs in each contig is shown on *x*-axis and bacterial ORFs on *y*-axis. Each blue circle shows a putative phage contig. Circular area represents the number of unclassified ORFs within a contig.

Fig. S2. Eight phage contigs presented nearly complete genomes. Colourized pentagons represent complete genes with known or unknown functions, and shaded pentagons represent partial genes. Peak charts show the length (*x*-axis) and coverage (*y*-axis) of contigs. The best-hit and sequence identity to reference phage genome (BLASTP) are given below each contig.

Fig. S3. Phylogenetic analysis based on head (A) and tail proteins (B). Predicted gene sequences of head (blue nodes) and tail proteins (red nodes) were extracted from our contigs, translated to amino acids and built phylogenetic trees with homologous sequences of the NCBI database respectively. The numbers on the branching represent bootstrap values.

Fig. S4. Fragments of a phage species connected with another bacterial species rather than its putative bacterial host. (A) A contig of our assemblies. (B and C) Contigs from the HMP. Pentagons with different color represent putative phage or bacterial genes. GenBank access number for each gene is given at the middle of each pentagon. Sequence identity (BLASTP) and the start and end position of each gene are shown above and below each pentagon respectively.

Fig. S5. Percentages of phage families in saliva of periodontal health (PHS), dental plaque of periodontal health (PHP) and disease (PDP). (A) Composition of phage families measured in our study. (B) Composition of phage families shown in a previous study (Ly *et al.*, 2014).

Fig. S6. Distributions of CIPs in different types of oral samples. (A) Relative abundance of CIPs (*y*-axis) in dental plaque of periodontal health (PHP) and disease (PDP). (B) Relative abundance of CIPs in subgingival (SUB) and supragingival (SUP) samples. Reads mapping to CIPs were counted to calculate their abundance in each sample ($n = 10$). The star represents the significant difference (Asterisk denotes $P < 0.05$, Mann–Whitney *U*-test).

Fig. S7. Several lines of evidence that support the connections in the cross-infection network. (A) Spacers from 11 distinct reads locate on a phage contig (Supragingival_plaque_LANL_312552, 9476bp), which supports the connection between bacteria *Aggregatibacter actinomycetemcomitans* (*Aa*) and *Pseudomonas stutzeri* (*Ps*). *Aa* dr represents direct repeats identified from *A. actinomycetemcomitans* genome. *Ps* dr represents direct repeats identified from *P. stutzeri* genome. (B) Spacers supporting the connections between *Streptococcus salivarius* (*Ss*), *S. infantis* (*Si*) and *S. anginosus* (*Sa*) can be aligned to different loci of the same phage contig (C35702196_63.0_0.370, 10475bp). (C) Six distinct reads support the connection between *Campylobacter gracilis* and *Streptococcus infantis*, and *Campylobacter gracilis* is connected with *Streptococcus mitis* by another phage genome.

Table S1. Samples used in this study.