The ISME Journal (2013) 7, 1710–1724
© 2013 International Society for Microbial Ecology All rights reserved 1751-7362/13
www.nature.com/ismej

ORIGINAL ARTICLE

Association between living environment and human oral viral ecology

Refugio Robles-Sikisaka[1], Melissa Ly[1], Tobias Boehm[2], Mayuri Naidu[1], Julia Salzman[3] and David T Pride[1,4]

[1]*Department of Pathology, University of California, San Diego, CA, USA;* [2]*College of Dental Medicine, Western University of Health Sciences, Pomona, CA, USA;* [3]*Departments of Biochemistry and Statistics, Stanford University School of Medicine, Stanford, CA, USA and* [4]*Department of Medicine, University of California, San Diego, CA, USA*

The human oral cavity has an indigenous microbiota known to include a robust community of viruses. Very little is known about how oral viruses are spread throughout the environment or to which viruses individuals are exposed. We sought to determine whether shared living environment is associated with the composition of human oral viral communities by examining the saliva of 21 human subjects; 11 subjects from different households and 10 unrelated subjects comprising 4 separate households. Although there were many viral homologues shared among all subjects studied, there were significant patterns of shared homologues in three of the four households that suggest shared living environment affects viral community composition. We also examined CRISPR (clustered regularly interspaced short palindromic repeat) loci, which are involved in acquired bacterial and archaeal resistance against invading viruses by acquiring short viral sequences. We analyzed 2 065 246 CRISPR spacers from 5 separate repeat motifs found in oral bacterial species of *Gemella*, *Veillonella*, *Leptotrichia* and *Streptococcus* to determine whether individuals from shared living environments may have been exposed to similar viruses. A significant proportion of CRISPR spacers were shared within subjects from the same households, suggesting either shared ancestry of their oral microbiota or similar viral exposures. Many CRISPR spacers matched virome sequences from different subjects, but no pattern specific to any household was found. Our data on viromes and CRISPR content indicate that shared living environment may have a significant role in determining the ecology of human oral viruses.

The ISME Journal (2013) **7,** 1710–1724; doi:10.1038/ismej.2013.63; published online 18 April 2013
Subject Category: microbial population and community ecology
Keywords: household; saliva; virus; virome; CRISPR; microbiome

Introduction

The study of the human microbiome focuses predominantly on bacterial flora (Bik *et al.*, 2006; Gao *et al.*, 2007; Costello *et al.*, 2009); however, there are numerous reports of viral communities inhabiting different body sites (Bachrach *et al.*, 2003; Breitbart *et al.*, 2003; Bik *et al.,* 2006; Gao *et al.*, 2007; Breitbart *et al.*, 2008; Costello *et al.*, 2009; Reyes *et al.*, 2010). Although there are viruses of Eukarya inhabiting these sites (Victoria *et al.*, 2009; Foulongne *et al.*, 2012; Lysholm *et al.*, 2012), the majority of the viruses that have been identified in these habitats are bacteriophage (Breitbart *et al.*, 2008; Willner *et al.*, 2009). As has been demonstrated in other environments (Breitbart *et al.*, 2002; Angly *et al.*, 2006), viral communities in the human body are diverse and generally composed of numerous different genotypes. These communities have also been hypothesized to serve as reservoirs of gene function for human microbial communities (Canchaya *et al.*, 2003; Rohwer and Thurber, 2009; Pride *et al.*, 2012a) and may serve particular importance in the mobilization of antibiotic resistance (Colomer-Lluch *et al.*, 2011a,b). Although the constituents of human viral communities may vary substantially within individuals over time, much of the gene function present in viruses of the human oral cavity remains conserved (Pride *et al.*, 2012a).

Similar to viral communities, the relative abundance of bacterial species is variable; however, their presence in the oral cavity of different human subjects is generally conserved (Pride *et al.*, 2012a). Given this relative conservation, bacteriophage in the environment might find suitable bacterial hosts in a variety of human subjects. There is some evidence that suggests that the environment has a role in the shaping of bacterial and viral

Correspondence: DT Pride, Department of Pathology, University of California, San Diego, 9500 Gilman Drive, MC 0612, La Jolla, CA 92093-0612, USA.
E-mail: dpride@ucsd.edu

Received 26 September 2012; revised 16 January 2013; accepted 18 March 2013; published online 18 April 2013

communities in humans (Willner *et al*., 2009). We had previously performed a study of a small number of human subjects and found two subjects from the same household (SHH) who shared many homologous sequences between their viral communities (Pride *et al*., 2012a). Each of these subjects shared similar bacterial biota but their relative abundances were dissimilar, suggesting that shared environment and not the relative abundances of bacteria in the oral cavity may be a significant determinant of viral community constituents in humans.

There is much interest in understanding how our microbiota are established and deciphering how exposure to microbes might ultimately affect community ecology. For human viral communities, little is known about to which viruses human subjects are exposed. CRISPRs (clustered regularly interspaced short palindromic repeats) are part of the CRISPR/Cas system in bacteria and archaea and are considered to be part of an adaptive immune response against invading viruses (Barrangou *et al*., 2007). They function by acquiring short sequences from invading viruses and utilizing these sequences to resist subsequent exposures to those viruses through nucleic acid interference (Bhaya *et al*., 2011; Brouns *et al*., 2008; Hale *et al*., 2009; Marraffini and Sontheimer, 2009; Sorek *et al*., 2008; Young *et al*., 2012). Because CRISPR loci acquire and accumulate short viral sequences (Pourcel *et al*., 2005; Marraffini and Sontheimer, 2009), they can be used to track viral exposures (Vergnaud *et al*., 2007; Andersson and Banfield, 2008; Zhang *et al*., 2010; Pride *et al*., 2012b; Rho *et al*., 2012). In this study, we examined the viromes and CRISPRs from 21 human subjects, 11 independent and 10 from among 4 separate households, in an attempt to gain a better understanding of the role that shared environment has in viral community membership and viral exposures in the human oral cavity.

Experimental Procedures

Enrollment of human subjects

Recruitment of each subject and enrollment in the current study was approved by the University of California, San Diego, CA, USA and the Western University Administrative Panels on Human Subjects in Medical Research. Each subject received a baseline periodontal examination before or at the time of saliva collection, which included measurements of probing depths, clinical attachment loss, Gingival Index, Plaque Index and gingival irritation (Loe, 1967). A minimum of 3 ml of saliva was collected from each subject and immediately frozen at −20 °C until further processing. All subjects enrolled were free from non-restored carious lesions and were in good overall periodontal health, with a diagnosis no greater than mild gingivitis. Exclusion criteria for the study included antibiotic administration during the 3 months before sample collection and preexisting medical conditions that could result in substantial immunosuppression. During their visit, subjects who enrolled were asked if they had members of their household willing to participate in the study.

Isolation of viruses, virome sequencing and screening for contamination

Saliva from human subjects was filtered sequentially through 0.45- and 0.2-μm filters to remove cellular debris, and the remaining fraction purified on a cesium chloride gradient as previously described (Pride *et al*., 2012a). Only the fraction at the density of most known viruses (Murphy *et al*., 1995) was retained. Viruses were then further purified on an Amicon YM-100 column (Millipore, Inc., Bellerica, MA, USA), treated with DNASE I, followed by lysis and DNA purification using Qiagen UltraSens virus kit (Qiagen, Valencia, CA, USA). Resulting DNA was amplified using GenomiPhi V2 MDA amplification (GE Healthcare, Pittsburgh, PA, USA), fragmented to roughly 100–200 bp using a Bioruptor (Diagenode, Denville, NJ, USA), created into libraries using the Ion Plus Fragment Library Kit according to the manufacturer's instructions and sequenced using 314 chips on an Ion Torrent Personal Genome Machine (Life Technologies, Grand Island, NY, USA) (Rothberg *et al*., 2011) producing an average read length of approximately 100 bp for each sample. Because of the sequencing error rate and the possibility for low-complexity reads, each read was trimmed according to modified Phred scores of 0.5 using CLC Genomics Workbench 4.65 (CLC Bio USA, Cambridge, MA, USA), and low-complexity reads (where >25% of the length were due to homopolymer tracts) were removed before further analysis. After trimming and removal of low-complexity reads, any remaining reads with substantial length variation (<50 nucleotides or >200 nucleotides) or reads with ambiguous characters were removed from the analysis. Remaining reads were analyzed using CLC Genomics Workbench 4.65 to construct assemblies based on 98% identity with a minimum of 50% read overlap, consistent with criteria developed to discriminate between highly related viruses (Breitbart *et al*., 2002). Because the shortest reads were 50 nucleotides, the minimum tolerable overlap was 25 nucleotides, and the average overlap was no <50 nucleotides depending on the characteristics of each virome. Contigs <200 bp were removed from further study, and the remaining contigs were assigned to their corresponding phylum based on their BLASTX best hits from the National Center for Biotechnology Information (NCBI) non-redundant database with an E-score cutoff value of 10^{-5}. Specific viral homologues were determined by parsing BLASTX results for known viral genes, including replication, structural, transposition, restriction/modification and hypothetical, and other genes previously found in viruses for which the E-score was at least 10^{-5}. Analysis of shared homologues present in each virome was performed by creating custom BLAST databases for each virome, comparing each database with all the other viromes using BLASTN analysis (E-score $<10^{-5}$), and normalization to the size of the smaller virome. Principal coordinates analysis was performed on homologous virome reads using binary Sorensen distances using QIIME (Caporaso *et al.*, 2010). Read mapping of viromes to a combined database of viruses (www.phantome.org; ftp://ftp.ncbi.nih.gov/genomes/Viruses/) was performed using CLC Genomics Workbench 4.65 and were mapped using 98% identity over a minimum of 50% of the read length. The read mapping results were reported as the proportion of reads mapping to any virus compared with the total number of virome reads. Virome sequences are available for download in the MG-RAST database (metagenomics.anl.gov/) under the project 'Household Study,' under individual accession numbers 451142.3, 451143.3, 451144.3, 451145.3, 451146.3, 451147.3, 451148.3, 451149.3, 451150.3, 451151.3, 451152.3, 451153.3, 451154.3, 451155.3, 451156.3, 451157.3, 451158.3, 451159.3, 451160.3, 451161.3 and 451162.3.

CRISPR sequencing and analysis

From each subject, genomic DNA was prepared from saliva using the Qiagen QIAamp DNA MINI Kit (Qiagen), with the addition of a bead beating step using Lysing Matrix B (MPBio, Solon, OH, USA) before DNA extraction. CRISPR sequences were amplified based on primers designed from the palindromic repeat sequences of various CRISPRs (Supplementary Table S2). Primer pairs for Gemella haemolysans group I (GHI), Veillonella species group I (VSI), Leptotrichia buccalis group I (LBI), Streptococcus group II (SGII) and SGI CRISPR spacers contain 10-nucleotide barcode sequences (represented by the 'X') and were used to amplify CRISPR sequences from each subject (Supplementary Tables S2 and 10). Reaction conditions included 44 μl Platinum High-Fidelity PCR Mastermix (Invitrogen, Carlsbad, CA, USA), 1 μl of each of the forward and reverse primer (10 mmol each) and 4 μl DNA template. The following were used as cycling parameters: 2 min initial denaturation at 94 °C, followed by 30 cycles of denaturation (15 s at 95 °C), annealing (15 s), and extension (2 min at 72 °C), followed by a final extension (10 min at 72 °C). CRISPR amplicons were gel extracted using the Qiagen MinElute Kit (Qiagen). Molar equivalents were determined for each product using an Agilent Bioanalyzer HS DNA Kit (Agilent, Santa Clara, CA, USA), and each were pooled into equimolar equivalents. Resulting pools were sequenced on 314 chips using an Ion Torrent Personal Genome Machine according to the manufacturer's instructions (Life Technologies) (Rothberg *et al.*, 2011). Barcoded sequences were then binned according to 100% matching barcodes. Each read was trimmed according to modified Phred scores of 0.5, and low-complexity reads and reads with ambiguous characters were removed before further analysis. Only those reads that had 100% matching sequences to both the 5′ and the 3′ end of the CRISPR repeat motifs were used for further evaluation. Spacers were defined as any nucleotide sequences (length $\geqslant$20) in between repeat motifs. To group spacers according to their trinucleotide content, we first compiled the trinucleotide content for all spacers and added them to a database. For each sequence, the difference in trinucleotide content was compared between all possible pairs of sequences regardless of overall spacer length, as length differences between identical spacers over the length of the shorter spacer would only account for small differences in total trinucleotide content. The sum of the differences for all sequence pairs was then determined for all sequences, and then spacers were binned together if their differences were less than the s.d. from the mean overall difference. Charts were created that included the total number of spacers with a specified number of trinucleotide differences. To validate the technique, random data sets of spacers were created with known numbers of mutations (including indels and polymorphisms). For each specified number of indels or polymorphisms, 1000 data sets with 5000 spacers each were created. For each data set, we took a single spacer and created 10, 100 or 500 different mutations in that spacer. We either created a single mutation in each spacer, or we created 2, 3 or 4 mutations. We then introduced these spacers back into the data set and binned the spacers according to their trinucleotide content differences. When a single mutation was introduced, regardless of whether the source of that mutation was an indel or a polymorphism, the mutated spacers were always binned together into a single spacer group.

For each subject evaluated, a database of spacer groups was generated, and databases were compared to determine shared spacer groups and to create heatmaps using Java Treeview (Saldanha, 2004). Beta diversity was determined using binary Sorensen distances and was used as input for principal coordinates analysis using Qiime (Caporaso *et al.*, 2010). Spacers from each subject were subjected to BLASTN analysis based on the NCBI non-redundant database. Hits were considered significant based on bit scores $\geqslant$45, which roughly correlated to 2 nucleotide differences over the length of a 30 nucleotide spacer. CRISPR spacers for each subject were used to search a database of the virome reads for matches from all viromes combined, and the number of spacer matches per virome read was used to create heatmaps. The heatmaps were normalized by the total number of spacer matches per virome read and were generated using Java Treeview (Saldanha, 2004). Rarefaction analysis was performed based on spacer group richness estimates of 10 000 iterations using EcoSim (Lee *et al.*, 2005). Good's coverage was determined as the estimation of the number of singletons in the population (n), compared with the total number of sequences (N), using the equation $(1-(n/N))\times 100$ (Good, 1953).

CRISPR loci from *Streptococcus oralis* were amplified using primers 5′-CGCCAAAAATCCGTATGAAA-3′ and 5′-TCGTAAAGTGTGGGCTCTCC-3′ and *Streptococcus thermophilus* CRISPR loci were amplified using primers 5′-CTGAGATTAATAGTGCGATTACG-3′ and 5′-GCTGGATATTCGTATAACATGTC-3′. CRISPR amplicons were gel extracted using the Qiagen MinElute Kit (Qiagen) and sequenced in both the directions using conventional Sanger sequencing.

Statistics

To assess whether virome reads or spacer groups had significant overlap between the set of individuals within the SHH, we performed a permutation test. We simulated the distribution of the fraction of overlapping reads within a group of individuals from mutually exclusive households that were randomly chosen across all the different households (DHHs) for numbers of individuals equal to the sizes of household nos. 1, 2, 3 and 4, respectively. For each set, we computed the summed fraction of randomly chosen spacer groups or virome reads and from these computed an empirical null distribution of statistics. The fraction computed resulted from 10 000 iterations for both the spacer groups and virome reads. For the CRISPR spacer groups, 1000 spacer groups were sampled in each iteration, and 10 000 reads were sampled in each iteration for the virome reads. The s.d. was computed from the percentage of homologous virome reads or spacer groups over the 10 000 iterations. For each household, an empirical null distribution of statistics was determined from an equal number of subjects from DHHs. The observed statistic per household was referred to this distribution, and the *P*-value was computed as the fraction of times the simulated statistic for the SHH exceeded the simulated statistic for the DHHs.

Results

Isolation and sequencing of salivary viruses

We recruited and sampled saliva from 21 human subjects in good overall periodontal health (Table 1). Eleven of the subjects were from DHHs, and 10 of the subjects were from 4 separate households (3 subjects in household no.1, 3 subjects in household no. 2, 2 subjects in household no.3 and 2 subjects in household no.4). All of the subjects residing in the SHHs were unrelated; however, 3 of the subjects (DHH5, DHH11 and SHH4-1) were related but had resided in DHHs for at least 9 years (Table 1). DNA viruses were isolated from the saliva of each subject through sequential filtration followed by cesium chloride density gradient centrifugation (Pride *et al.*, 2012a). We then sequenced 10 685 110 virome reads (average of 508 815 reads per subject) using semiconductor sequencing (Rothberg *et al.*, 2011). To ensure that the viral communities were properly separated from potential contaminating cellular elements, we screened each virome against 16S ribosomal RNA and reference human databases. We found the vast majority of the viromes were free of substantial numbers of homologues to human DNA or bacterial 16S ribosomal RNA (Table 2). The largest number of human homologues from a virome (DHH2) measured at 0.17% (592 of 346 558) of the sequenced reads. These data indicate that the viromes studied have a relatively insignificant amount of measurable contaminating DNA from cellular elements.

Table 1 Study subjects

Subject	*Age*	*Ethnicity*	*Sex*	*Time in same household (Years)*	*Nature of relationship*
Household 1					
SHH1-1	24	Caucasian	Male	1	Housemate
SHH1-2	24	Asian	Male	6	Housemate
SHH1-3	54	Asian	Female	6	Housemate
Household 2					
SHH2-1	52	Asian	Male	3	Housemate
SHH2-2	23	Asian	Female	3	Spouse
SHH2-3	28	Filipino	Male	3	Spouse
Household 3					
SHH3-1	32	Asian	Male	5	Spouse
SHH3-2	23	Filpino	Female	5	Spouse
Household 4					
SHH4-1[a]	34	Asian	Female	5	Spouse
SHH4-2	36	African-American	Male	5	Spouse
Different households					
DHH1	38	Latino	Female	NA	NA
DHH2	32	Indian	Female	NA	NA
DHH3	23	Asian	Female	NA	NA
DHH4	28	Asian	Male	NA	NA
DHH5[b]	58	Asian	Female	NA	NA
DHH6	50	Caucasian	Male	NA	NA
DHH7	28	African-American	Male	NA	NA
DHH8	34	Caucasian	Male	NA	NA
DHH9	25	Caucasian	Female	NA	NA
DHH10	28	Caucasian	Male	NA	NA
DHH11[c]	59	Asian	Female	NA	NA

[a] Daughter of DHH5 (not residing together for 9 years).
[b] Mother of SHH4-1, sibling of DHH11 (not residing together for 15 years).
[c] Sibling of DHH5.

Virome composition

We assembled the reads from each virome into larger contigs (Table 2) to help identify the most abundant viruses present in the saliva of each subject. Each contig then was subjected to BLASTX analysis using the NCBI non-redundant database to identify homologues present in the viromes. Similar to other studies (Breitbart *et al.*, 2008; Pride *et al.*, 2012a), the

Table 2 Virome reads and contigs

	Reads							*Contigs*		
Sample name	*Number of reads*	*Length*	*No. trimmed*	*Homopolymers*	*Final reads*[a]	*No. 16S*[b]	*No. human*[c]	*Number*	*Average length*	*G+C content*
DHH1	976 195	76	716 966	407	258 822	0	9	296	421	63%
DHH2	454 834	105	107 909	367	346 558	5	592	725	648	53%
DHH3	587 699	109	113 707	217	473 775	0	0	850	799	50%
DHH4	620 084	123	36 166	303	583 615	0	91	985	946	46%
DHH5	631 944	122	59 203	170	572 571	1	0	471	749	46%
DHH6	465 530	120	30 234	68	435 228	0	0	520	849	45%
DHH7	314 147	96	185 370	46	128 731	0	5	473	595	62%
DHH8	493 089	106	304 940	56	188 093	0	1	377	450	60%
DHH9	584 909	115	64 087	37	520 785	0	0	500	658	53%
DHH10	564 033	113	38 441	65	525 527	0	2	313	788	55%
DHH11	477 488	122	36 459	52	440 977	0	2	150	1728	45%
SHH1-1	895 260	88	244 456	311	650 493	0	11	1214	796	48%
SHH1-2	1 267 857	87	418 037	1579	848 241	0	12	476	1042	45%
SHH1-3	594 679	87	180 273	710	413 696	0	3	553	1071	52%
SHH2-1	1 401 276	54	982 471	170	418 635	0	2	257	971	48%
SHH2-2	1 226 398	92	250 489	3063	972 846	0	28	437	1381	40%
SHH2-3	1 150 652	88	311 904	6636	832 112	0	13	110	1853	44%
SHH3-1	798 431	74	302 571	205	495 655	0	5	399	1007	40%
SHH3-2	436 059	95	59 497	70	376 492	0	111	89	1177	51%
SHH4-1	463 034	115	32 585	56	430 393	0	58	712	716	49%
SHH4-2	919 430	83	147 361	204	771 865	0	1	383	647	52%

[a] Final number of reads after trimming and removal of homopolymer reads.
[b] Based on BLASTN analysis (E-score $<10^{-5}$) of a composite 16S rRNA database, including the full Ribosomal Database Project (RDP), Greengenes, National Center for Biotechnology Information (NCBI) and Silva databases.
[c] Based on BLASTN analysis (E-score $<10^{-5}$) of NCBI human reference genome assemblies.

vast majority of the homologues found were to bacteriophage, while only a few homologues to viruses of Eukarya (herpesviruses and circoviruses) were found. By using the host taxonomy of each homologue as the putative taxonomy for each contig, we determined the host taxonomy of our viral communities at the Phlyum level. We found that most of the subjects had a predominance of viruses from the phyla Proteobacteria or Firmicutes (Figure 1), which includes the genera *Streptococcus*, *Gemella* and *Veillonella*. We also found many Fusobacteria, which includes the genus *Leptotrichia*. With the exception of household no.3, there were similar proportions of phyla for the subjects within the SHHs; however, similar patterns were also observed in subjects from DHHs.

We mapped the virome reads from each subject to a database of known bacteriophage (www.phantome.org) to determine the proportion of virome reads from each subject that mapped to known bacteriophage. In some subjects, a substantial proportion of the reads mapped to a single bacteriophage in the database, suggesting that these bacteriophage are highly abundant (Supplementary Table S1). This is contrary to results we previously reported for human salivary viromes in which the oral viral populations were predicted to be evenly distributed in all the subjects studied (Pride *et al.*, 2012a). For subject DHH5, 13% (75 581 reads) of the virome reads mapped to Streptococcus phage 5093, while 10% (53 880) of the reads from subject DHH11 map to the same phage. These subjects were related (Table 1) but had not resided in the SHH for the past 15 years. Subject SHH4-1, the daughter of subject DHH5, had a much lower abundance of virome reads (<1%) that mapped to Streptococcus phage 5093. Numerous other bacteriophage were identified by mapping virome reads against the database of bacteriophage, including Actinomyces phage AV-1

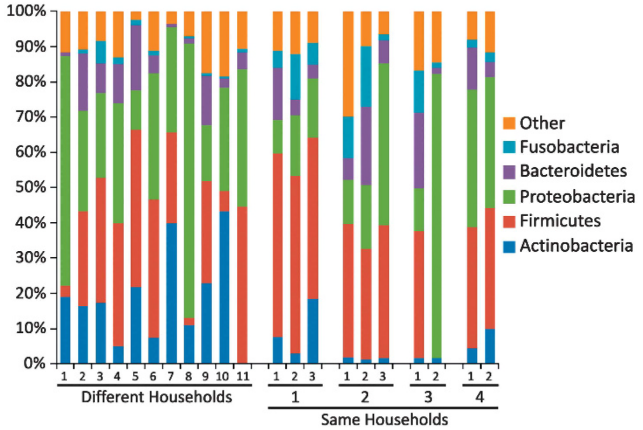

Figure 1 Putative host biological assignments for viral contigs from human saliva from all the subjects. The percentage of contigs assigned to each biological group based on BLASTX best-hits is shown for each subject.

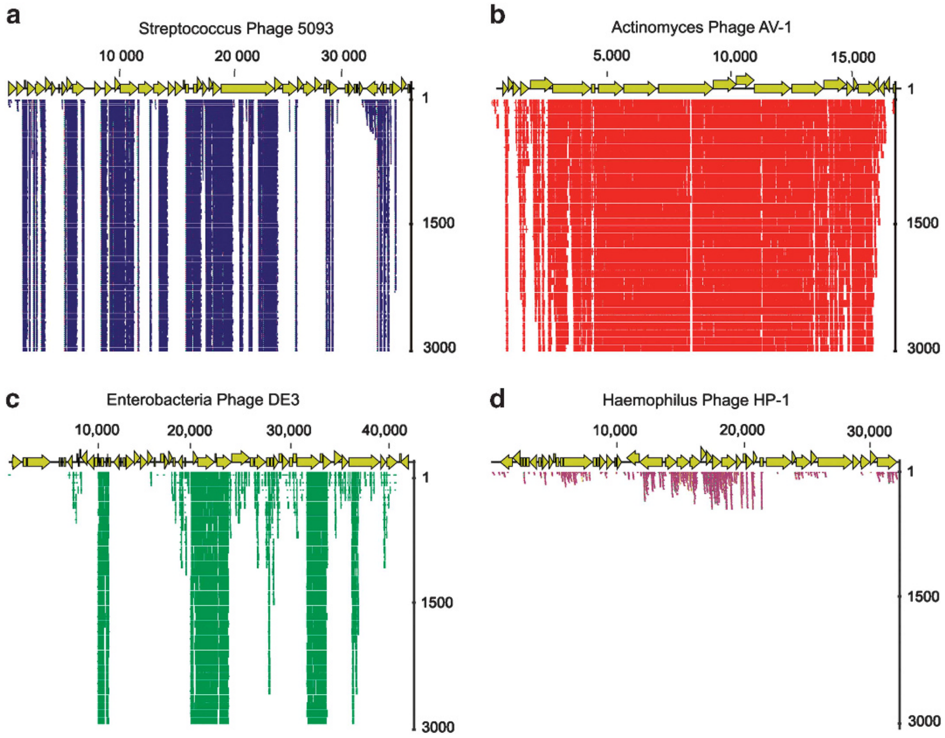

Figure 2 Read mappings of viromes to select viruses. Panel (**a**) demonstrates the mapping of reads from subject DHH5 to Streptococcus phage 5093 (75 581 of the 572 571 reads). Panel (**b**) demonstrates the mapping of reads from subject DHH3 to Actinomyces phage AV-1 (21 076 of the 473 775 reads). Panel (**c**) demonstrates the mapping of reads from subject DHH8 to Enterobacteria phage DE3 (18 776 of the 188 093 reads). Panel (**d**) demonstrates the mapping of reads from subject SHH3-2 to Haemophilus phage HP-1 (1406 of the 376 492 reads). The y axis demonstrates the total number of reads mapping to an individual portion of each virus.

(4% in DHH3) and Enterobacteria phage DE3 (10% in DHH8) (Table 2). Nearly the entire bacteriophage genomes of Streptococcus phage 5093, Actinomyces Phage AV-1 and Enterobacteriaphage DE3 were present in these viral communities (Figure 2). We found other bacteriophage present in these communities whose full genome likely is present, including Haemophilus phage HP-1, despite representing <1% of the virome reads in subject SHH3-2 (Figure 2d).

Shared viral homologues within households

To determine whether there was an association between living environment and the composition of the oral viral community, we compared the virome reads among all the subjects studied. Using BLASTN analysis to find homologues between all subjects, we found a higher percentage of shared homologues between some household members (Figure 3a and Supplementary Figure S1). For household no.1, SHH1-1 and SHH1-2 shared the most homologous sequences, while SHH1-3 shared the most homologous sequences with DHH9. For household no.2, SHH2-2 shared the most homologous sequences with SHH2-3, and SHH2-1 shared the most homologous sequences with DHH11. For household no.3, SHH3-1 and SHH3-2 shared the most homologous sequences. Subjects SHH4-1 and SHH4-2 shared the most homologous sequences with subjects outside of their household. We next tested whether the fraction of homologous virome reads within each household would be greater than that between DHHs by randomly sampling the virome reads >10 000 iterations. We found that there were significant relationships found in households nos.1, 3, and 4 ($P<0.0001$ for each) (Table 3). Thus, three of the four households studied harbor subjects with a higher fraction of homologous sequences among their viromes than would be expected if they resided in DHHs, suggesting that shared living environment has a role in shaping salivary viral ecology. We also used principal coordinates analysis of the viral homologues to determine whether there was an association between shared living environment and virome sequences (Figure 3b). The analysis supported that there was some association among the virome sequences within households nos.1 and 3. Subjects DHH5 and SHH4-1 (mother and daughter) did not have a significant association; however, DHH5 and DHH11 (siblings) shared many virome sequences (Figure 3a).

Grouping CRISPR spacer

CRISPR loci expand at the 5′ end as new virus-specific spacers are added after each viral encounter.

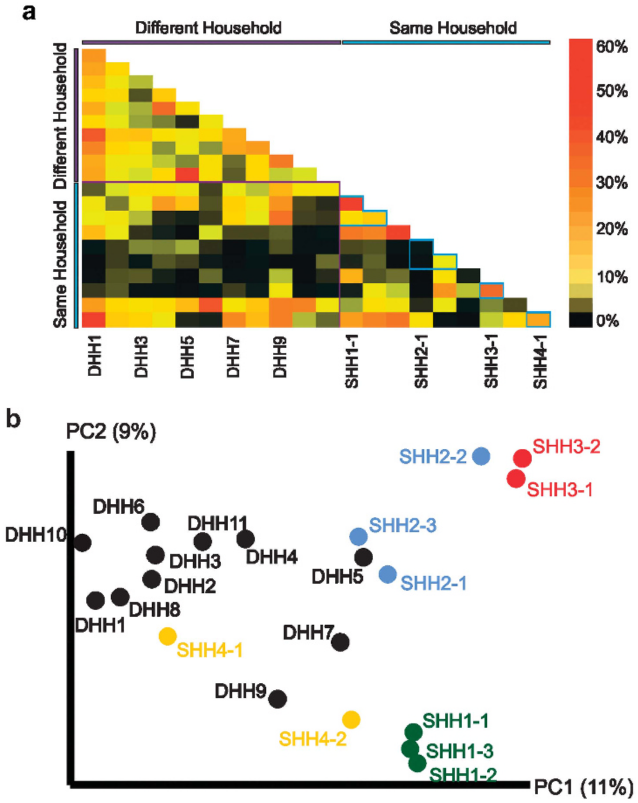

Figure 3 Homologous reads matrix (**a**) and principal coordinates analysis (**b**) of viromes from all the subjects. The matrix demonstrates the percentage of homologous reads shared between all subjects. The top half of the matrix represents comparisons between subjects residing in DHHs, the bottom half of the matrix represents comparisons between subjects from the four separate shared households and the middle portion of the matrix outlined in purple represents comparisons between DHHs and the four separate shared households. Comparisons between subjects who reside in the SHH are outlined in blue. The principal coordinates analysis is based on beta diversity metrics of all shared virome reads from all the subjects. Household no.1 is represented in green, household no.2 is represented in cyan, household no. 3 is represented in red and household no.4 is represented in yellow.

By analyzing the CRISPR loci, we could trace past viral exposures. Because our analysis of salivary viruses was based on a single time point, we chose to explore CRISPRs in each subject to track past viral exposures. Rather than examine individual CRISPR loci, we chose a metagenomic approach. The benefits of a metagenomic approach were that we could examine multiple loci simultaneously, explore loci from bacteria not known to be present or to harbor CRISPR loci and examine CRISPR loci from abundant and relatively rare microbes. Although this approach produced substantial data from many likely streptococcal species, the major limitation was that CRISPR spacers could not be ordered into individual loci or attributed to individual loci or species.

We developed a technique to bin CRISPR spacers based on trinucleotide content so that spacers targeting similar viruses would be grouped together.

Table 3 Viral homologues between households

Subject	*Percentage of homologous reads within households*[a]	*Percentage of homologous reads for different households*[a]	P-*value, same vs different households*[b]
Household 1	23.10 ± 0.43	3.99 ± 1.14	< **0.0001**
Household 2	2.59 ± 0.17	4.01 ± 1.14	0.9031
Household 3	67.24 ± 0.46	1.91 ± 1.01	< **0.0001**
Household 4	9.76 ± 0.30	1.92 ± 1.00	≤ **0.0001**

[a] Based on the mean of 10 000 iterations. A total of 10 000 random reads were sampled per iteration.
[b] Empirical *P*-value based on the fraction of times the estimated percentage of homologous reads for each household exceeds that for different households. Bold values represent *P*-values $\leqslant$0.01.

We chose such a technique because depending on the characteristics of the data set, similar spacers that harbor multiple polymorphisms could be binned together without setting an arbitrary similarity cutoff. When binning spacers based on trinucleotide content, those spacers that differ by a single polymorphism or indel (insertion or deletion) will have similar trinucleotide content differences. We validated the technique by generating sets of 5000 random spacer sequences. Then using these spacer data sets, we took a single spacer and created random mutations along its length. We introduced 10, 100 or 500 mutated spacers to the random spacer data set, and then tested whether the technique could group these mutated spacers into a single spacer group. When a single polymorphism or indel was introduced, the mutated spacers were always binned together in a single spacer group regardless of the number of mutated spacers that were introduced (Table 4). Similar results were also produced when we introduced two polymorphisms or indels. The technique was less robust when three or four polymorphisms or indels were introduced, with the mutated spacers forming more than a single spacer group. Although the spacers formed more groups when more than two polymorphisms were introduced, these groups were always populated only by other mutated spacers.

CRISPR spacer metagenomes

We sequenced CRISPR loci using primers based on five separate repeat motifs found in oral bacteria *Gemella haemolysans* (GHI), *Veillonella* sp. (VSI), *Leptotrichia buccalis*, *Streptococcus gordonii* (SGII) and *Streptococcus mutans* (SGI) (Supplementary Table S2). Because the streptococcal repeat motifs have been shown to be present in many different streptococcal species (Pride *et al.*, 2011), we presumed the same principle might apply to GHI, VSI and LBI CRISPR repeat motifs. We identified 2 065 246 CRISPR spacers from the 21 subjects studied. For GHI, we identified 360 429 CRISPR spacers in 1706 spacer groups based on trinucleotide content (Supplementary Table S3). The distribution of the trinucleotide content differences between the GHI spacers demonstrated that the majority of the spacers were either identical or highly similar, with only 0.003% of the GHI spacers identified as having polymorphisms or indels that necessitated grouping according to trinucleotide content (Supplementary Figure S2A). Similar results were produced for VSI (0.003% had polymorphisms), LBI (0.009% had polymorphisms), SGII (0.002% had polymorphisms) and SGI spacers (0.001% had polymorphisms) (Supplementary Tables S4–7 and Supplementary Figures S2B–E). These data indicate that there were relatively few polymorphisms in each population of CRISPR spacers, potentially the result of sequencing error.

Table 4 Grouping of mutated spacers

	Number of mutated spacers evaluated		
	10 Spacers[a]	*100 Spacers*[a]	*500 Spacers*[a]
Number of polymorphisms			
1	1.00[b]	1.01[b]	1.00[b]
2	1.10	1.01	1.69
3	1.81	4.72	13.00
4	2.35	8.59	24.87
Number of indels			
1	1.00	1.02	1.00
2	1.01	1.01	1.00
3	1.29	1.65	4.78
4	2.19	16.95	15.67
Number of polymorphisms and indels			
1[c]	1.24	2.00	2.28
2[d]	2.76	9.88	17.35

[a]Results indicate the number of groups to which the mutated spacers are assigned based on 1000 iterations.
[b]Number of spacer groups the mutated spacers comprise.
[c]Includes one indel and one polymorphism.
[d]Includes two indels and two polymorphisms.

We subjected the CRISPR spacer population in each subject to rarefaction analysis to estimate how thoroughly each subject had been sampled for each CRISPR type. The shapes of many of the curves indicated that further sampling would not have identified many more CRISPR spacer groups (Figure 4). For example, LBI spacers for each subject reached asymptote (Figure 4c), with the exception of one subject. LBI CRISPRs had relatively low spacer richness compared with other CRISPR types, while SGI CRISPRs (Figure 4e) generally had the highest spacer richness. Shared living environment generally did not predict CRISPR spacer richness. Although a few curves represented continued expanding diversity (Figure 4), the Good's Coverage estimates were relatively high (range from 90 to 100) and indicated that the majority of unique spacers in each subject and CRISPR spacer type had been adequately sampled (Supplementary Tables S3–7).

Shared CRISPR spacers

To improve our understanding of the conservation of CRISPR spacer groups in all the subjects, we

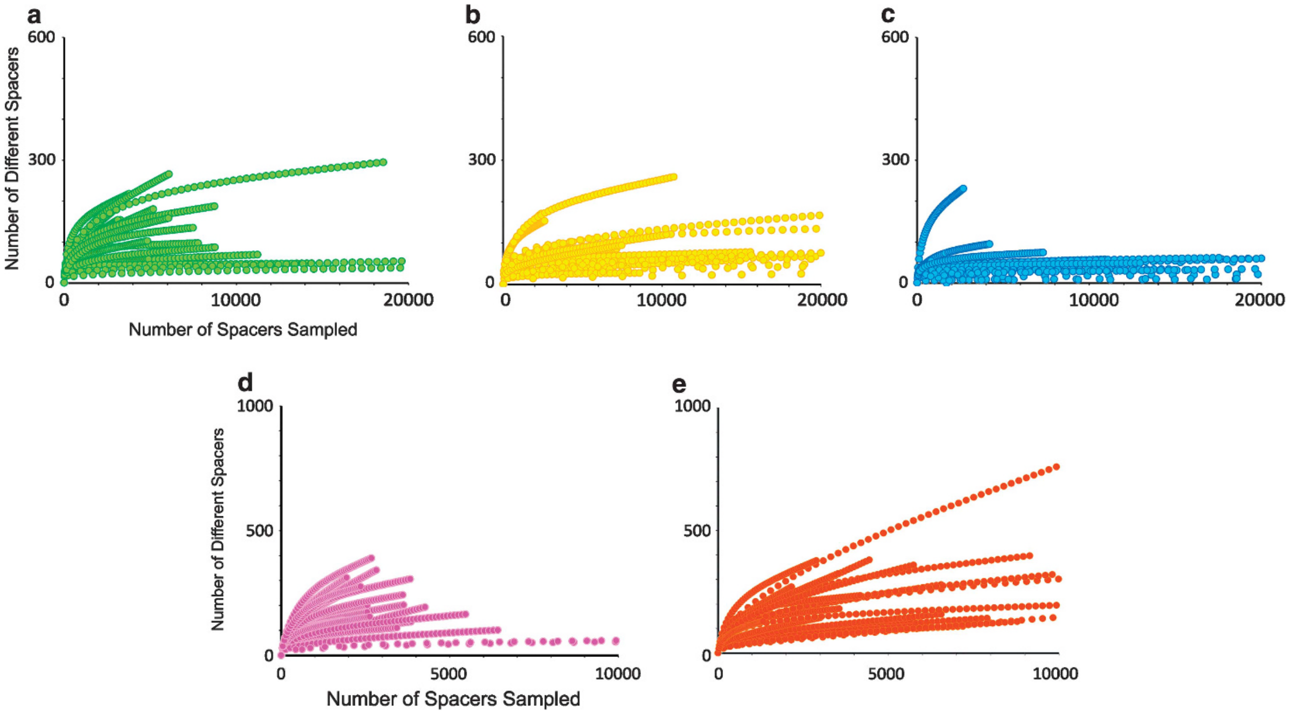

Figure 4 Rarefaction analysis of CRISPR spacers in the saliva of all the subjects. Rarefaction curves were created using 10 000 random iterations based on spacer group richness. (**a**) GHI CRISPR spacers; (**b**) VSI CRISPR spacers; (**c**) LBI CRISPR spacers; (**d**) SGII CRISPR spacers; (**e**) SGI CRISP spacers.

generated heatmaps (Figure 5). Each subject had a large proportion of CRISPR spacer groups that were subject specific for all CRISPR spacer types, while a minority of the spacers was shared between different subjects. For most subjects, no specific trend of shared spacer groups was seen; however, there were some subjects who shared a proportion of spacer groups depending on the CRISPR type. For example, subjects DHH4 and DHH5 had similar trends in GHI spacer group conservation (Figure 5a) and subjects DHH1, DHH2, DHH8 and DHH11 shared a substantial portion of SGII spacer groups (Figure 5d).

We determined the proportion of CRISPR spacer groups shared among all the subjects and found a trend similar to that found for viromes. In household no.2, subject SHH2-2 shares the most CRISPR spacer groups with subject SHH2-3 for GHI, VSI, LBI, SGII and SGI type CRISPR spacers (Figures 6a–e). Subjects SHH3-1 and SHH3-2 in household no.3 also share the most spacer groups for all types of CRISPR spacers (Figure 6) and viromes (Figure 3). As was also seen for viromes, there were generally fewer spacer groups shared between the subjects in household no.4 than for other households. We also measured the fraction of shared spacer groups that were present in each household and compared with that from an equal number of randomly selected DHHs. In each instance, the fraction of shared spacer groups within a household was greater than would be expected when comparing individuals from DHHs (Table 5), and most results were highly significant ($P<0.0001$). Only VSI and LBI CRISPR

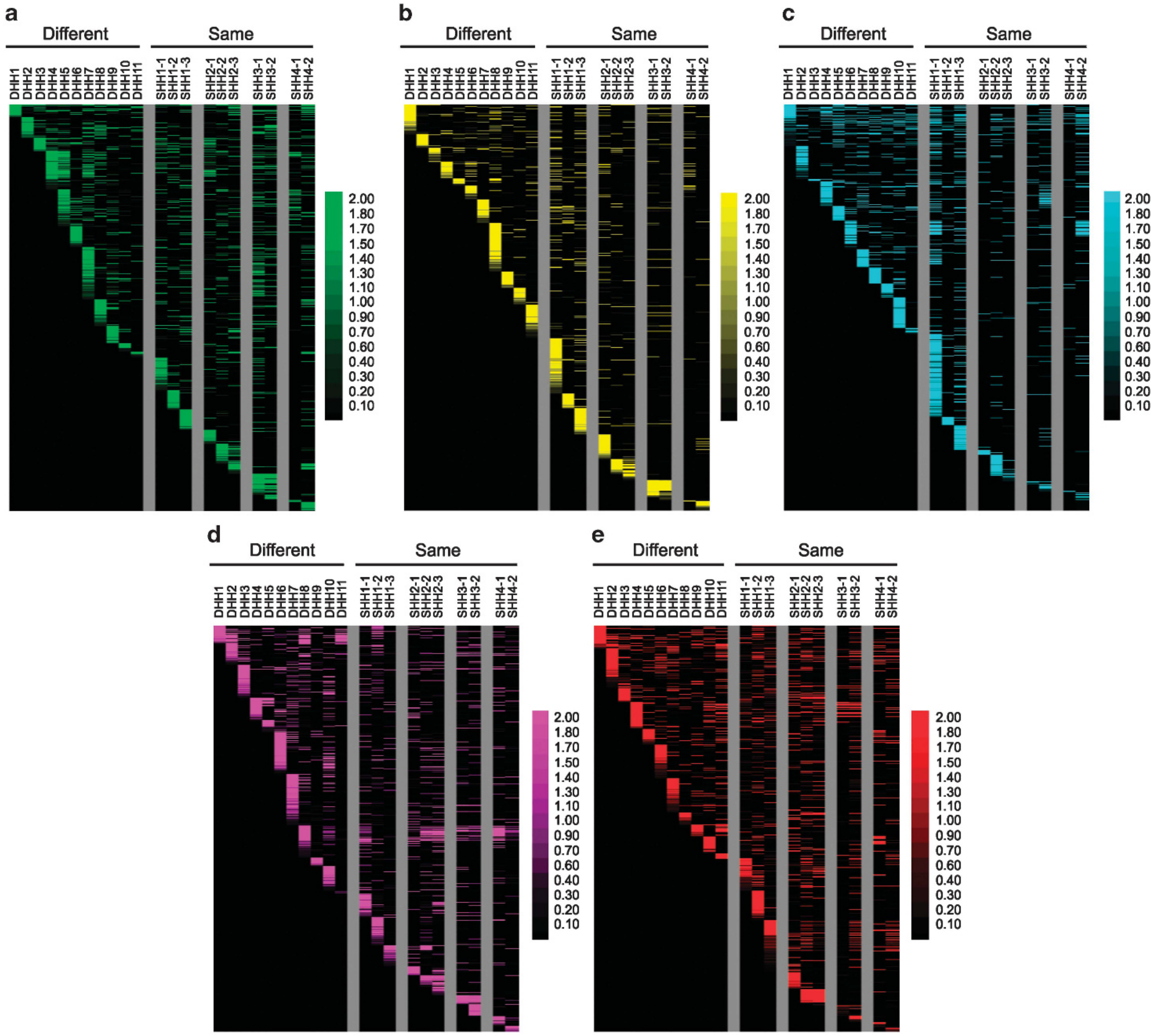

Figure 5 Heatmaps of the CRISPR spacer groups in all the subjects. Each row represents a unique spacer group and the columns represent subjects from DHHs or subjects from the four separate shared (same) households. The intensity scale bar is located to the right. (**a**) GHI CRISPR spacers; (**b**) VSI CRISPR spacers; (**c**) LBI CRISPR spacers; (**d**) SGII CRISPR spacers and (**e**) SGI CRISPR spacers.

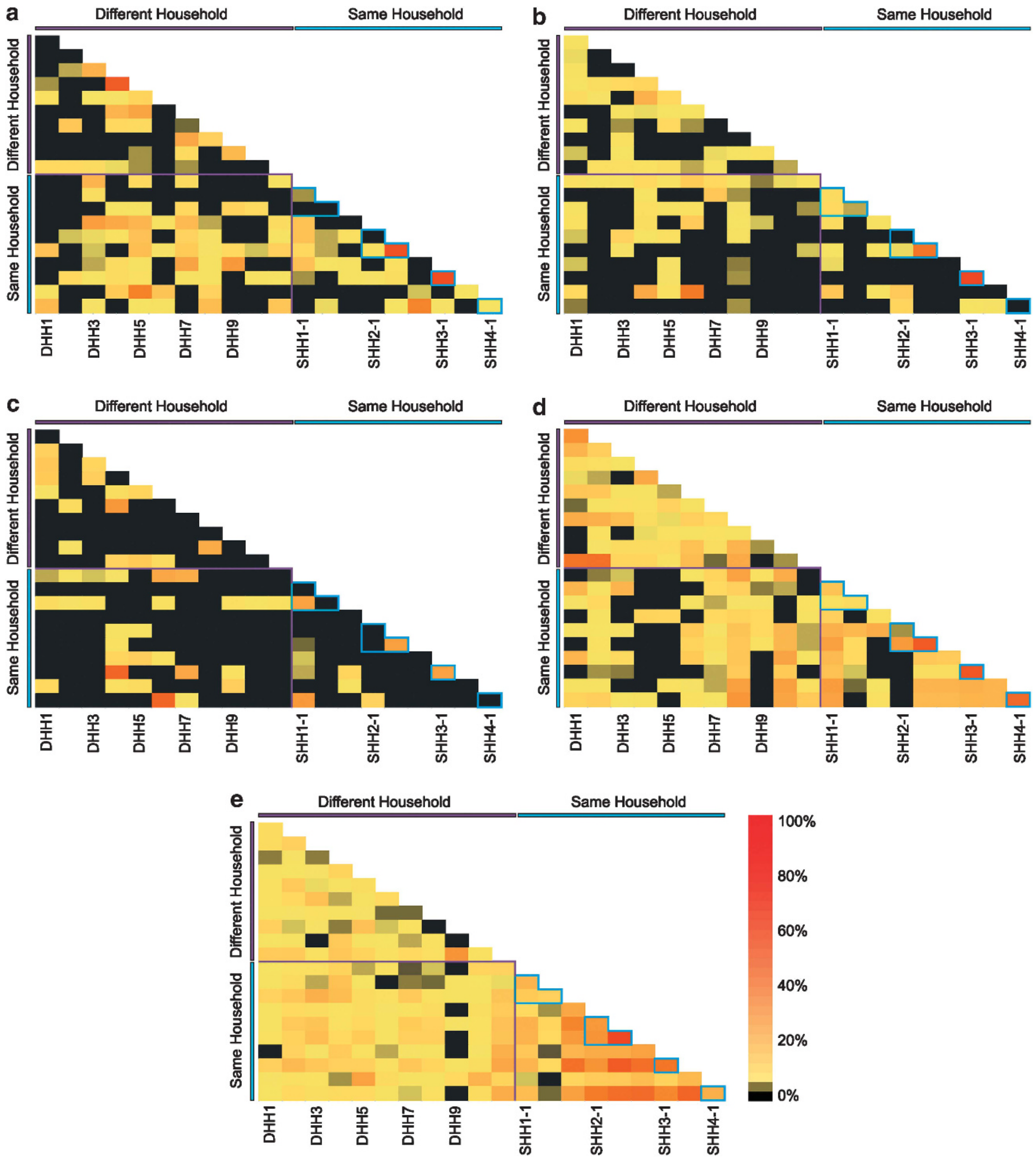

Figure 6 CRISPR spacer group matrix from all the subjects. The matrix demonstrates the percentage of shared CRISPR spacer groups between all the subjects. The top half of the matrix represents comparisons between subjects residing in DHHs, the bottom half of the matrix represents comparisons between subjects from the four separate shared households and the middle portion of the matrix outlined in purple represents comparisons between DHHs and the four separate shared households. Comparisons between subjects who reside in the SHH are outlined in blue. (**a**) GHI CRISPR spacers; (**b**) VSI CRISPR spacers; (**c**) LBI CRISPR spacers; (**d**) SGII CRISPR spacers and (**e**) SGI CRISPR spacers.

types in household no.4 did not show a significant association between household and CRISPR spacer groups. Of the three related subjects (DHH5, DHH11 and SHH4-1), there were few spacer groups shared. We also sequenced several individual CRISPR loci from *S. oralis* and *S. thermophilus* and found a high diversity of CRISPR loci among the different subjects (Supplementary Figure S3). Interestingly, we found a few shared spacers among the different subjects, but the CRISPR spacer order was not conserved (Supplementary Figure S3B). These data indicate that the CRISPR spacers in these loci may have been acquired independently and suggest that some of the shared CRISPR spacers among our subject population are the result of independent CRISPR locus evolution rather than vertical or horizontal transmission.

We examined each of the CRISPR types by principal coordinates analysis to determine whether close relationships existed in CRISPR spacer

Table 5 Shared CRISPR spacers within households

Subject	*Percentage of homologous reads within household*[a]	*Percentage of homologous reads for different households*[a]	P-*value, same vs different households*[b]
Gemella (GHI)			
Household 1	32.63 ± 2.02	7.38 ± 2.05	**<0.0001**
Household 2	47.29 ± 1.97	7.37 ± 2.02	**<0.0001**
Household 3	53.04 ± 1.56	3.22 ± 1.61	**<0.0001**
Household 4	8.83 ± 0.89	3.22 ± 1.61	**0.0058**
Veillonella (VSI)			
Household 1	18.31 ± 1.57	2.63 ± 1.39	**<0.0001**
Household 2	25.15 ± 1.57	2.62 ± 1.39	**<0.0001**
Household 3	64.50 ± 1.51	1.16 ± 0.93	**<0.0001**
Household 4	5.77 ± 0.73	1.17 ± 0.95	0.0993
Leptotrichia (LBI)			
Household 1	36.89 ± 1.52	1.94 ± 1.26	**<0.0001**
Household 2	30.00 ± 1.64	9.74 ± 6.72	**0.0046**
Household 3	20.62 ± 1.28	4.33 ± 4.65	**0.0149**
Household 4	6.30 ± 0.77	4.28 ± 4.62	0.2163
Gordonii (SGII)			
Household 1	33.26 ± 2.18	4.52 ± 1.53	**<0.0001**
Household 2	54.02 ± 1.79	4.51 ± 1.55	**<0.0001**
Household 3	43.23 ± 1.57	1.93 ± 1.27	**<0.0001**
Household 4	36.89 ± 1.52	1.94 ± 1.26	**<0.0001**
Mutans (SGI)			
Household 1	46.31 ± 2.24	8.60 ± 3.36	**<0.0001**
Household 2	85.15 ± 2.52	8.62 ± 3.39	**<0.0001**
Household 3	31.95 ± 1.46	4.03 ± 2.48	**<0.0001**
Household 4	16.36 ± 1.78	4.05 ± 2.50	**<0.0001**

Abbreviation: CRISPR, clustered regularly interspaced short palindromic repeat.

[a] Based on the mean of 10 000 iterations. A total of 1000 random spacers were sampled per iteration.

[b] Empirical *P*-value based on the fraction of times the estimated percentage of homologous reads for each household exceeds that for different households. Bold values indicate *P*-values ≤0.01.

repertoires from subjects in shared living environments (Supplementary Figure S4). For household no.3, a close relationship in their CRISPR spacers is demonstrated for all CRISPR spacer types, except LSI (Supplementary Figure S4C). Subjects SHH2-2 and SHH2-3 in household no.2 also share close relationships in their CRISPR spacer content. There was little relationship discovered between the subjects in household no.4, with the exception of SGII CRISPR spacers (Supplementary Figure S4D). In general, there were closer relationships within households for the streptococcal CRISPRs (Supplementary Figures S4D–E) than for the other CRISPR types (Supplementary Figures S4A–C). These data suggest that shared living environment may influence viral exposures in human saliva; however, these results could also be explained by shared ancestry of CRISPR loci.

CRISPR spacers match virome reads

We tested whether the CRISPR spacers had any homologues in the NCBI non-redundant database to better define the group of viruses to which our subjects may have had previous exposures. Only a single homologue was found for GHI spacers among the 1706 spacer groups from all the subjects (Supplementary Table S8). There were few homologues in the database to the 1990 VSI spacer groups, with all of them found in the *Veillonella parvula* DSM 2008 genome. We found homologues to LBI spacers matching sequences outside of the CRISPR loci in the genomes of *L. buccalis* and *Sebaldella termitidis*, both belonging to the phylum Fusobacteria. There were abundant homologues found to SGII and SGI spacers (Supplementary Tables S8 and 9), likely as a result of the large database of previously sequenced streptococci and streptococcal viruses.

Although there were relatively few CRISPR spacers that had homologues in the NCBI non-redundant database, we tested whether there were homologues present in the virome reads from all the subjects. In nearly every circumstance, there were more virome reads than NR database homologues matching CRISPR spacers (Supplementary Table S8), indicating that oral spacer repertoires were specifically adapted to oral viromes. Similar to a previous study (Pride *et al.*, 2012b), there was no specific pattern of CRISPR spacers repertoires matching virome reads within that same subject, but rather many subjects have CRISPR spacers that match viruses from other subjects (Supplementary Figure S5). Similar results for SGI and SGII spacers were found for virome contigs to those that were found for virome reads (Supplementary Figure S6). Indeed, the vast majority of the virome reads that matched CRISPR spacers were found in the viromes of subjects DHH4, DHH5, DHH6, DHH11 and SHH1-2. Because the majority of the matches were to streptococcal viruses (Supplementary Table S9), these data suggest that those subjects were highly enriched for streptococcal viruses.

There was a much lower percentage of each type of CRISPR spacer that had homologues in the NR database (Figure 7a) than in the virome reads (Figure 7b). We found a small percentage of CRISPR spacers that matched virome reads for GHI, VSI and LBI type CRISPRs, indicating that these viromes likely have been targeted by these CRISPR types (Figures 7a–c). Importantly, *Veillonella* and *Leptotrichia* are genera of bacteria that usually live in the anaerobic environment of the subgingival crevice. Our data indicate that viruses of anaerobic bacteria were also present in the salivary environment (Figures 1 and 7). There were many more CRISPR spacers that matched virome reads for the streptococcal CRISPR types SGI and SGII. Most interestingly, 75% of the SGII CRISPR spacer groups in subject DHH1 were homologous to virome reads from the subjects in this study. Similar results were found for SGII and SGI CRISPR types in other subjects, including DHH2, DHH11 and SHH4-1. We believe the high percentage of CRISPR spacers

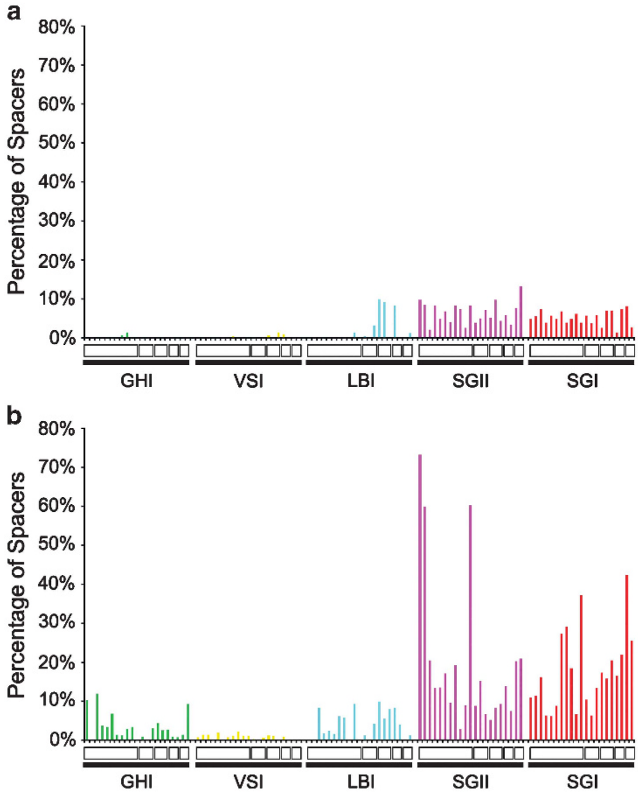

Figure 7 Percentage of the CRISPR spacer groups with homologues in the NCBI non-redundant database (**a**) and in the viromes of all the subjects (**b**). The y axis represents the percentage of the CRISPR spacer groups, and the x axis represents the different CRISPR spacer types. The boxes on the x axis represent the DHHs, where the 11 subjects from DHHs are represented by the large box (from DHH1 to DHH11), and the subjects in the four separate households are represented by the smaller boxes (from household no.1 subjects to household no.4 subjects, left to right).

matching virome reads represents specific adaptation of bacteria in the human oral cavity to viruses commonly encountered in the salivary environment.

Discussion

Our data suggest that shared living environments influence the viral microbiota of the human oral cavity. Although not every subject in the SHH demonstrated this trend, the viruses in households nos.1, 3 and 4 shared a significantly greater than expected fraction of homologous reads (Table 3). Although it seems plausible that virome constituents and viral exposures would directly reflect the bacterial communities present, we chose not to evaluate the bacterial communities for the following reasons: (1) we previously have demonstrated that inverse relationships exist between certain viruses and their host bacteria in the human oral cavity (Pride *et al*., 2012a), (2) much of the variation in CRISPR communities likely exists at the strain level and thus cannot be evaluated through 16S rRNA sequencing, and (3) there have already been relatively thorough evaluations of the bacterial communities present in saliva (Bachrach *et al*., 2003; Lazarevic *et al*., 2009; Nasidze *et al*., 2009; Bik *et al*., 2010; Pride *et al*., 2011a; Pride *et al*., 2011). Thus, we do not believe that sequencing of 16S rRNA would clarify the effects of bacteria on the viral community constituents within households.

Contamination of viromes with nucleic acids from human and bacterial cells could potentially influence the results in any study of human viral communities. We have developed highly stringent criteria to ensure that the viromes studied are relatively free of cellular contamination to prevent apparent similarity between viromes because of shared contaminating nucleic acids. Generally, most, if not all, of the cellular material was removed through sequential filtering through 0.2-μm filters, though free contaminating nucleic acids could remain. Centrifugation of the remaining materials on a cesium chloride density gradient separated the viral fraction from free nucleic acids and cellular materials, and the addition of a DNase digestion step was added to remove any naked nucleic acids before further processing. We then screened all sequenced virome reads for human DNA by BLASTN analysis against the NCBI human genome assemblies. None of the viromes we sequenced had a significant percentage of reads that were homologous to the human genome (Table 2), and those virome reads were removed before further analysis to ensure that these homologues could not affect the analysis of homologous sequences between viromes. Because the putative phyla distribution in these viromes is similar to that found for the bacterial communities in saliva (Pride *et al*., 2012a, Pride *et al.,* 2011), we also screened for evidence of contaminating bacterial DNA by BLASTN analysis against a composite 16S rRNA database. Of the 21 viromes we sequenced, only DHH2 had a significant number of 16S rRNA homologues (Table 2). All five 16S rRNA homologues were to staphylococcal genomes, which are generally not normal flora of the human oral cavity. No significant numbers of DHH2 virome reads were homologous to staphylococcal genomes, suggesting that these genomes were not contaminated with staphylococcal nucleic acids. In total, our analysis indicates that these viromes are relatively free of contaminating cellular nucleic acids. We also replicated the virome read data for household no.2 to demonstrate that the data were reproducible. However, because of the lack of homologous sequences among the subjects within household no.2 (Table 3), the replicates were not informative. The replicate data were combined with the original virome reads for the subjects in household no.2, and each treated as a single virome throughout the study.

The extraordinary diversity of viruses in human ecosystems can often serve as a limiting factor in assembly and analysis of human viral communities.

We previously showed that viral communities in the human oral cavity are inhabited by many different viral genotypes with a relatively even distribution (Pride *et al.*, 2012a). The results of those data were highly dependent on the ability of assembly tools to create contig spectra that were representative of the viral population. We used several different assemblers on our virome reads; however, most produced similar contig spectra consistent with an evenly distributed population using PHACCS (Angly *et al.*, 2005). We noticed in our analysis of the virome communities from the subjects in this study that in many of the viromes a substantial proportion of the reads mapped to a single virus (Figure 2 and Supplementary Table S1). Virome read distributions from subjects DHH5, DHH8 and DHH11 suggested that Streptococcus phage 5093, Actinomyces phage AV-1 and Enterobacteria phage DE3 or highly similar viruses were abundant in these subjects and that the populations were not evenly distributed. There are other tools such as MaxiΦ (Angly *et al.*, 2006) available for analysis of viromes, but they also are based on contig spectra, which we do not believe provide an accurate description of the viromes produced in this study.

Because we did not evaluate CRISPR spacers derived from Enterobacteria or *Actinomyces*, we could not discern a bacterial response to their viruses being present in high abundance in this study (Figure 2 and Supplementary Table S1). Surprisingly, we did not find CRISPR spacers matching Streptococcus phage 5093 in the subjects harboring this virus, but we did find spacers matching this virus in other subjects (data not shown). There was a substantial repertoire of CRISPR spacers found for SGI and SGII CRISPR spacers that matched many different species of Streptococcus (Supplementary Table S9). Interestingly, the spacers generally were not identical to those of the streptococcal CRISPR loci in these genomes but rather were identical to putative lysogenic phage in the streptococcal genomes, which limited our ability to use existing CRISPR loci as a template for assembly of the CRISPR spacers sequenced in this study. We found many spacers for SGI and SGII type spacers that matched bacteria other than streptococci, which we had not found in our previous analysis of these types of CRISPR spacers. We believe that the discovery of these spacers is the result of the greater sequence depth provided in this study and suggests that these CRISPR spacer types may be present in other organisms beyond the Genus *Streptococcus*.

Our data supports that shared living environment affects the composition of the oral viral community (Figure 3b), but is only one of many possible factors that might affect oral viral ecology. Other factors that might affect shared viral ecology include diet, shared bacterial biota or shared ancestry of the study subjects, which might explain the similar CRISPR and virome profiles from siblings DHH5 and DHH11. By the very nature of saliva, its viruses may be transitory and thus not the optimal means of assessing the effects of shared living environments on viral populations. The oral biofilm might represent a more stable population of viruses; however, its relatively low biomass complicates isolation and identification of virome constituents. Based on our findings, the length of time in the shared household may not be a critical factor affecting the viral flora, as subject SHH1-1 has viruses similar to subjects SHH1-2 and SHH1-3 after only a year in the SHH, while subjects SHH2-1 and SHH2-2 share few viral constituents after 3 years in the SHH. We believe that viral exposures are a critical determinant of oral viral ecology, and the CRISPR spacer repertoires from the subjects in all the households strongly suggest that each household member has been exposed to similar viral populations (Table 5). With the exception of VSI and LBI spacers for household no.4, the data strongly support that viral exposures are associated with shared living environment for all households and CRISPR type. Although it is possible that some CRISPR loci are similar merely as a result of shared ancestry of bacterial strains, we believe that this would not explain the full extent of the shared spacers in some subjects. Using data from the CRISPRs web server (Grissa *et al.*, 2007), we estimated that the average streptococcal CRISPR locus has 16 spacers. Thus for household no.3, each subject would need to share 6–7 average-sized SGII CRISPR loci and 10–11 identical SGI CRISPR loci to reproduce the results presented here. Because these CRISPR loci belong to many unidentified species and strains of *Streptococcus* and the short-read metagenomics approach used to sequence the CRISPR spacers, we were not able to assemble and determine the spacer order or the strain from which each spacer was derived. We identified *S. oralis* and *S. thermophilus* CRISPR loci by more conventional techniques and demonstrated that individuals who share some CRISPR spacers often have relatively disparate CRISPR loci (Supplementary Figure S3).

As we continue to explore the human microbiome, we now have a greater understanding of those features that affect viral community ecology and those viruses to which human subjects are exposed. We previously demonstrated that for many host/virus relationships in human saliva there is an inverse relationship, indicating that the relative abundance of bacteria in the oral cavity does not necessarily reflect the relative abundances of certain viruses (Pride *et al.*, 2012a). Although these viruses must be inexorably linked to their host bacteria, as long as a suitable host is present, viruses may be capable of colonizing the oral cavity. Thus, we believe that exposure to viruses is a critical determinant of viral community membership. Our findings that many of the subjects in shared living environments also share similar CRISPR repertoires suggests that they have been exposed to similar viruses; however, additionally this association could be secondary to shared bacterial strains between subjects. Regardless of their method of acquisition, these CRISPR spacers likely endow these human subjects with similar capacities to resist oral viruses. Because there were similar viruses in subjects in some of the shared living environments, developing similar CRISPR repertoires would provide a selective advantage to those CRISPR-bearing microbes in those shared environments. Interestingly, almost all subjects share some streptococcal CRISPR spacers (Figures 6d and e), suggesting that they all have encountered similar streptococcal viruses. We believe the CRISPR spacers that are shared between subjects may represent a practical benefit of CRISPR-mediated resistance, where they provide innate immunity against viruses they have yet to encounter but to which they carry matching spacer sequences.

Acknowledgements

This study was supported by the Robert Wood Johnson Foundation, the Burroughs Wellcome Fund and NIH 1K08AI085028 to DTP.

Author Contributions

DTP conceived and designed experiments. RR-S, ML and MN performed the experiments. DTP, RR-S and JS analysed the data. TB collected specimens. DTP wrote the manuscript.

References

Andersson AF, Banfield JF. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**: 1047–1050.

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P *et al.* (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**: 41.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

Bachrach G, Leizerovici-Zigmond M, Zlotkin A, Naor R, Steinberg D. (2003). Bacteriophage isolation from human saliva. *Lett Appl Microbiol* **36**: 50–53.

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.

Bhaya D, Davison M, Barrangou R. (2011). CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. In: Bassler BL, Lichten M, Schupbach G (eds) *Annual Review Genetics* Vol 45. Annual Reviews: Palo Alto, pp 273–297.

Bik EM, Eckburg PB, Gill SR, Nelson KE, Purdom EA, Francois F *et al.* (2006). Molecular analysis of the bacterial microbiota in the human stomach. *Proc Natl Acad Sci USA* **103**: 732–737.

Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF *et al.* (2010). Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J* **4**: 962–974.

Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, Felts B *et al.* (2008). Viral diversity and dynamics in an infant gut. *Res Microbiol* **159**: 367–373.

Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P *et al.* (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**: 6220–6223.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D *et al.* (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99**: 14250–14255.

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP *et al.* (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**: 960–964.

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brussow H. (2003). Phage as agents of lateral gene transfer. *Curr Opin Microbiol* **6**: 417–424.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Colomer-Lluch M, Imamovic L, Jofre J, Muniesa M. (2011a). Bacteriophages carrying antibiotic resistance genes in fecal waste from cattle, pigs, and poultry. *Antimicrob Agents Chemother* **55**: 4908–4911.

Colomer-Lluch M, Jofre J, Muniesa M. (2011b). Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PLoS One* **6**: e17549.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.

Foulongne V, Sauvage V, Hebert C, Dereure O, Cheval J, Gouilh MA *et al.* (2012). Human Skin Microbiota: High Diversity of DNA Viruses Identified on the Human Skin by High Throughput Sequencing. *PLoS One* **7**: e38499.

Gao Z, Tseng CH, Pei Z, Blaser MJ. (2007). Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci USA* **104**: 2927–2932.

Grissa I, Vergnaud G, Pourcel C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**: 172.

Good IJ. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**: 237–264.

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L *et al.* (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**: 945–956.

Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osteras M *et al.* (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* **79**: 266–271.

Lee SG, Kim CM, Hwang KS. (2005). Development of a software tool for in silico simulation of Escherichia coli using a visual programming environment. *J Biotechnol* **119**: 87–92.

Loe H. (1967). The Gingival Index, the Plaque Index and the Retention Index Systems. *J Periodontol* **38**: Suppl:610–616.

Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K *et al.* (2012). Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS One* **7**: e30875.

Marraffini LA, Sontheimer EJ. (2009). Invasive DNA, chopped and in the CRISPR. *Structure* **17**: 786–788.

Murphy FA, Fauquet CM, Bishop. DHL, Ghabrial SA, Jarvis AW, Martelli GP *et al.* (1995). *Virus Taxonomy: Sixth Report of the International Committee on Taxonomy of Viruses*. Springer-Verlag: New York, NY, USA, Vol. Supplement 10.

Nasidze I, Li J, Quinque D, Tang K, Stoneking M. (2009). Global diversity in the human salivary microbiome. *Genome Res* **19**: 636–643.

Pourcel C, Salvignol G, Vergnaud G. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**: 653–663.

Pride DT, Salzman J, Haynes M, Rohwer F, Davis-Long C, White RA 3rd *et al.* (2012a). Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J* **6**: 915–926.

Pride DT, Salzman J, Relman DA. (2012b). 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 DT, Sun CL, Salzman J, Rao N, Loomer P, Armitage GC *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.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F *et al.* (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**: 334–U381.

Rho M, Wu Y-W, Tang H, Doak TG, Ye Y. (2012). Diverse CRISPRs Evolving in Human Microbiomes. *PLoS Genet* **8**: e1002441.

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

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M *et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348–352.

Saldanha AJ. (2004). Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**: 3246–3248.

Sorek R, Kunin V, Hugenholtz P. (2008). CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Micro* **6**: 181–186.

Vergnaud G, Li Y, Gorge O, Cui Y, Song Y, Zhou D *et al.* (2007). Analysis of the three *Yersinia pestis* CRISPR loci provides new tools for phylogenetic studies and possibly for the investigation of ancient DNA. *Adv Exp Med Biol* **603**: 327–338.

Victoria JG, Kapoor A, Li LL, Blinkova O, Slikas B, Wang CL *et al.* (2009). Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol* **83**: 4642–4651.

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, 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.

Young JC, Dill BD, Pan CL, Hettich RL, Banfield JF, Shah M *et al.* (2012). Phage-Induced Expression of CRISPR-Associated Proteins Is Revealed by Shotgun Proteomics in Streptococcus thermophilus. *PLoS One* **7.**

Zhang J, Abadia E, Refregier G, Tafaj S, Boschiroli ML, Guillard B *et al.* (2010). Mycobacterium tuberculosis complex CRISPR genotyping: improving efficiency, throughput and discriminative power of ‘spoligotyping’ with new spacers and a microbead-based hybridization assay. *J Med Microbiol* **59** Pt 3: 285–294.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)