Supporting Information

Willner et al. 10.1073/pnas.1000089107

SI Materials and Methods

Oropharyngeal Sampling. Oropharyngeal swab samples were collected in July 2007 from individuals with no symptoms of respiratory infection as assessed by a prescreening questionnaire. A signed consent form was obtained from each subject before sample collection, as required by the San Diego State University Institutional Review Board. Study subjects ranged in age from 23 to 56 y old and consisted of 8 females and 11 males. Samples were obtained by swabbing the area posterior and superior to the palatopharyngeal arch (lower border of the nasopharynx) on both the right and left sides with a sterile swab. Swabs were replaced into their original self-sealing specimen containers and transported to the laboratory.

Oropharyngeal Sample Processing and Viral Isolation. Upon arrival at the laboratory, 2 mL of SM buffer (50 mM Tris·HCl, pH 7.5, 100 mM NaCl, 8 mM $MgSO_4$, 0.01% gelatin) was added to the oropharyngeal swabs. Samples were then treated with DTT to break up mucus as described in ref. 1 and subsequently filtered through a 0.8-μm polycarbonate filter (Sterlitech) and 0.45-μm Millex-HV filters. The filtrate was brought to a density of 1.15 $g{\cdot}mL^{-1}$ by addition of solid cesium chloride (CsCl). This sample was overlaid onto a CsCl step gradient to concentrate viral particles as detailed in ref. 2). Ten microliters of each viral concentrate was vacuum filtered onto a 0.02-μm Anodisc (Millipore) and stained with SYBR Gold (Invitrogen), and virus-like particles were visualized using epifluorescence microscopy.

Viral concentrates from 19 individual samples were comb xined to form a pooled sample. Half of the pooled sample was filtered with a 0.22-μm Millex-HV filter to remove cellular material. DNase I was added to the sample to a final concentration of 10 $\mu g{\cdot}mL^{-1}$, followed by incubation at 37 °C for 1 h to degrade free DNA. The other half of the pooled sample was treated with chloroform but not filtered. Twenty microliters of chlorofrom was added per milliliter of viral concentrate, and the sample was then incubated at 4 °C for 2 h and centrifuged at 2155 × *g* for 15 min. The supernatant containing the viral concentrate was collected and treated with DNase I as described above. DNA was extracted from the filtered and chloroformed samples using a cetyltrimethylammonium bromide (CTAB)/formamide protocol (3). Viral DNA was sequenced at the Joint Genome Institute using the 454 GS-FLX platform. The chloroformed metagenome contained 215,281 sequences with an average length of 206 bp, and the filtered metagenome contained 245,025 sequences with an average length of 219 bp.

Saliva Sampling for Metagenomic Study. For the metagenomic study, three subjects donated saliva samples at three time points over a 3-mo period from February 2008 to April 2008. All subjects had no preexisting medical conditions and were determined to be periodontally healthy on the basis of full baseline periodontal examinations performed before the study. A minimum of 3 mL of saliva was collected at each time point, and saliva was stored at −20 °C until processing for metagenomic sequencing.

Saliva Sample Processing and Viral Isolation. Saliva samples were combined with an equal volume of SM buffer. Viruses were concentrated as described above. Saliva viral concentrates were filtered at 0.22 μm and treated with DNase I. Viral DNA was extracted using a CTAB/formamide protocol and sent for sequencing at the W. M. Keck Center at the University of Illinois, which uses the 454 titanium platform.

Initial Processing of Metagenomic Sequences. The two oropharyngeal and nine saliva metagenomic libraries were de-replicated and then compared with the nonredundant database at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) using BLASTn and tBLASTx (4). Sequences were assigned taxonomy on the basis of the most significant BLAST similarity with an *e*-value $<10^{-5}$ and a minimum alignment length of 50 bp. All sequences classified as microbial were compared with the ACLAME prophage database using BLASTn and tBLASTx to distinguish prophage sequences from microbial genomic sequences (5). Sequences with the best BLAST similarities to eukaryotic and microbial genomes were removed from the metagenomes before subsequent analyses.

Bioinformatics for Oropharyngeal Metagenomes. Sequences with best BLASTn similarities to viral genomes were mapped to the genomes of Epstein-Barr virus, *Escherichia coli* phage T3, *Propionibacterium acnes* phage PA6, and *Streptococcus mitis* phage SM1 by using BLAT and visualized with the Integrated Genome Browser (6, 7). Reference sequences for viral genomes were obtained from NCBI (http://www.ncbi.nlm.nih.gov). Continuous coverage of the SM1 genome at the amino acid level and pblA and pblB genes was calculated using a Perl script. Contigs were assembled using the 454 gsAssembler with a minimum overlap length of 35 bp and 98% identity and compared with the nonredundant database for taxonomic assignment. Phage genome annotations were obtained from Pajunen et al. (8) for phage T3 , Farrar et al. (9) for phage PA6, and Siboo et al. (10) for phage SM1. Counts of significant similarities to T3 in other environmental metagenomes were obtained by comparing the T3 genome to the environmental samples database (env_nt) at NCBI by using BLASTn.

Viral community composition was determined using GAAS, based on tBLASTx comparisons of all noneukaryotic and nonmicrobial sequences to a database containing all complete viral genomes currently available at NCBI (11). GAAS parameters were at least 40% identity with a minimum relative alignment length of 80% and an *e*-value cutoff of 10^{-5}. Viral diversity was estimated using the PHACCs program, with input contig spectra generated by Circonspect (http://biome.sdsu.edu/circonspect) and average genome size estimated by GAAS (11, 12). Cross-contig spectra were generated using Circonspect and used for a Monte Carlo simulation (as described in ref. 13) to determine the percentage of species shared and permuted between the two oropharyngeal metagenomes.

Bioinformatics and Statistics for Salivary Metagenomes. Sequences with significant similarity (*e*-value $<10^{-5}$, >30% identity over at least 80% of the query length) to the pblA and pblB gene regions of phage SM1 were extracted from the salivary metagenomes. Continuous coverage of the pblA and pblB genes was calculated with a Perl script. Sequences extracted from each metagenome were compared in a bidirectional pairwise fashion using cd-hit-2d-est with a 90% identity cutoff to determine the percentage of sequences shared (14). The similarity index between metagenomes was calculated as the number of pblA or pblB sequences in metagenome 1 shared with metagenome 2 plus the number of sequences in metagenome 2 shared with metagenome 1 divided by the total number of sequences. The dissimilarity matrix was constructed by subtracting all similarity values from 1 and then used as an input to multidimensional scaling in R (15). Coverage of pblA and pblB genes in each metagenome was assessed by dividing the gene into bins of 20, 50, or 100 bp and counting the number of

sequences covering each bin. Coverage over all bins was compared using the XIPE program, which uses nonparametric statistical methods to compare two empirical distributions (16). Coverage dissimilarity was calculated as the proportion of bins identified as having different coverage by XIPE. Results were the same whether 20-, 50-, or 100-bp bins were used. The correlation between coverage and sequence similarity was calculated using the correlation test procedure in R with the "Spearman" option (15).

To search for flanking viral sequences that would be indicative of the horizontal gene transfer of pblA and pblB, saliva metagenomic sequences were assembled using the 454 gsAssembler to form contigs. Contigs were aligned to the genome of phage SM1 using both BLASTn and tBLASTx (4). Contigs with similarities to the pblA and/or pblB genes as well as potential flanking sequences (i.e., sequences at the end that were not similar to the SM1 genome) were selected for further analysis. The contigs in total and the excised potential flanking sequences were compared with the nonredundant database using both BLASTn and tBLASTx. In the majority of cases, these sequences were unknown. For some contigs, flanking sequences could be identified using tBLASTx. For example, two contigs from subject 1 on day 90 were found to contain the pblA gene with flanking sequence from the tail protein of a phage of *Streptococcus pyogenes* NZ131. However, there were few contigs containing such sequences, and the use of contigs is somewhat problematic because flanking sequences may be the result of poor assemblies.

Bacterial Strains. The *S. mitis* SF100 and PS344 strains were provided courtesy of Paul Sullam and Ho Seong Seo (University of California, San Francisco). *S. mitis* SF100 contains the complete SM1 prophage including pblA and pblB genes (10). *S. mitis* PS344 contins the SM− prophage with the pblA and pblB genes deleted (10). *S. mitis* strains were grown on blood agar (tryptic soy agar supplemented with 5% sheep's blood) or Todd Hewitt broth (THB) at 5% CO_2 at 37 °C.

PCR Screening of Saliva Samples. Total DNA was isolated from saliva samples using the method of Quinque et al. (17). Positive control DNA was prepared from an overnight culture of *S. mitis* SF100 grown in THB. DNA was extracted using the Nucleospin tissue kit (Macherey-Nagel) with the addition of a Gram-positive lysis step as described in the manufacturer's instructions. Negative control DNA was similarly prepared from *S. mitis* PS344.

The PCR mixture (50 μL total) contained target DNA, 1× Taq Buffer, 0.2 mM dNTPs, 1 μM of each primer, and 1 unit Taq DNA polymerase. The forward primer (pblA1456F) was 5′-ACCGCAGAGGCAGCGAATGC-3′, and the reverse primer (pblA2222F) was 5′-CCAGGCCATAGACGCAGCCG-3′. Primers were designed using primer BLAST at NCBI to generate primer sequences and amplicon unique to SM1 pblA (18). The thermocycler conditions were the following: 5 min at 94°; 30 cycles of 1 min at 94°; 1 min at 58° with a −0.5° touchdown; 1 min at 72°; and 10 min at 72°. PCR products were checked for size on a 1% agarose gel and prepared for sequencing using the Accu-Prep PCR purification kit (Bioneer). PCR products were sequenced at the San Diego State University Microchemical Core Facility using an ABI Prism 3100 Genetic Analyzer.

Sequences from PCR products were trimmed and aligned to each other and the reference sequence using ClustalW (19). Nucleotide sequences were translated in all six frames using TranSeq and were aligned to the translated reference sequence to determine the correct translation (20). Translated PCR product sequences were then aligned to the SM1 reference sequence (GeneID: 1461276) both with and without the presence of pblA homolog sequences (21). Homolog sequences were as follows: *S. pyogenes* MGAS315 SpyM3_1104 (GeneID: 1009419), *Streptococcus pneumoniae* 70585 SP70585_0072 (GeneID: 7683049), *Streptococcus agalactiae* λ Sa1 pblA (GeneID: 1013400), *S. agalactiae* λ Sa03 pblA (GeneID: 3686919), *S. pyogenes* M1GAS Spy_1448 (GeneID: 901501), *Enterococcus faecalis* V583 tape measure (GeneID: 1200878), *S. pyogenes* M1GAS SpyM3_1313 (GeneID: 1009628), *E. faecalis* V583 tail protein (GeneID: 1199267), and *S. pyogenes* phage 315.5 tail protein (GeneID: 1257924) (22–25). Alignments were visualized using JalView (26). A phylogenetic tree was created on the basis of the aligned amino acid sequences using MrBayes 3.1 (27). Four independent Monte Carlo Markov chains were run for 500,000 generations using the mixed amino acid model option.

Phage Induction Assays. Phage inductions were performed using an adaptation of the protocol described in ref. 21. Overnight cultures of *S. mitis* SF100 were grown in THB for 16 h. Cultures were diluted 1:10 in fresh THB and incubated for 30 min. Cultures were treated with one of six treatments: 0.25 μg/mL mitomycin C, red wine diluted 1:100, white wine 1:10, cola diluted 1:10, nicotine 1:10 (2.4 mg), or soy sauce diluted 1:10. The nictoine treatment consisted of Johnson Creek Original Smoke Juice (Vapure), which contains 24 mg/mL of nicotine. An untreated culture was used as a control. These treatments were selected on the basis of overnight growth curves of *S. mitis* SF100 with a variety of treatments added. All cultures were incubated for 3 h and then filtered using a 0.45-μm Millex-HV filter to remove remaining bacterial cells.

Phage particles were enumerated using a flow cytometry procedure (28). Samples were fixed with 0.5% glutaraldehyde for 30 min at 4 °C. Samples were then flash frozen in liquid nitrogen and stored at −80 °C until analysis. For analysis, samples were thawed at room temperature and diluted 1:100 in Tris-ethylenediaminetetraacetic acid buffer. One unit of DNase was added to each diluted sample and allowed to incubate at room temperature for 15 min to eliminate free DNA in the sample. Fresh SYBR Green I (0.5×) was then added to each sample to stain for DNA and incubated at 80 °C for 10 min in the dark. After incubation, samples were allowed to cool in the dark at room temperature for 5 min. Internal standard beads (0.75-μm-diameter YG fluorescent latex microspheres; Polysciences) were added to each sample at a concentration of 1×10^6 beads per sample. Samples were analyzed using a FACSAria flow cytometer (Becton Dickinson) using FACSDiva software. The cytometer threshold was set on green fluorescence while the machine flow rate analyzed ~1,000 events per second. Contour plots were generated on a side scatter *x* axis and a green fluorescence *y* axis on biexponential scales. Two separate electronic gates were generated for enumeration of virus and beads. Samples were collected until 1×10^5 bead events were detected. Viral positive events between different induction conditions had a minimum of ~1600 viral counts and a maximum of ~420,000 viral counts per 1×10^5 bead events. Three replicate experiments for each treatment were conducted, and statistical significance was assessed using randomization tests as implemented in the R function permtest (15).

1. Willner D, et al. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 4: e7370.
2. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4:470–483.
3. Sambrook J (2001) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), 3rd Ed.
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
5. Leplae R, Hebrant A, Wodak SJ, Toussaint A (2004) ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res* 32 (Database issue):D45–D49.
6. Kent WJ (2002) BLAT: The BLAST-like alignment tool. *Genome Res* 12:656–664.
7. Nicol JW, Helt GA, Blanchard SG Jr, Raja A, Loraine AE (2009) The Integrated Genome Browser: Free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 25:2730–2731.
8. Pajunen MI, Elizondo MR, Skurnik M, Kieleczawa J, Molineux IJ (2002) Complete nucleotide sequence and likely recombinatorial origin of bacteriophage T3. *J Mol Biol* 319:1115–1132.

9. Farrar MD, et al. (2007) Genome sequence and analysis of a Propionibacterium acnes bacteriophage. *J Bacteriol* 189:4161–4167.
10. Siboo IR, Bensing BA, Sullam PM (2003) Genomic organization and molecular characterization of SM1, a temperate bacteriophage of Streptococcus mitis. *J Bacteriol* 185:6968–6975.
11. Angly FE, et al. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLOS Comput Biol* 5: e1000593.
12. Angly F, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* 6:41.
13. Angly FE, et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4:e368.
14. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
15. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna) Available at http://www.R-project.org.
16. Rodriguez-Brito B, Rohwer F, Edwards RA (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* 7:162.
17. Quinque D, Kittler R, Kayser M, Stoneking M, Nasidze I (2006) Evaluation of saliva as a source of human DNA for population and association studies. *Anal Biochem* 353: 272–277.
18. National Center for Biotechnology Information (2008) *Primer-BLAST* Available at http://www.ncbi.nlm.nih.gov/tools/primer-blast.
19. Larkin MA, et al. (2007) ClustalW and ClustalX version 2.0. *Bioinformatics* 23:2947–294.
20. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277.
21. Bensing BA, Siboo IR, Sullam PM (2001) Proteins PblA and PblB of Streptococcus mitis, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. *Infect Immun* 69:6186–6192.
22. Beres SB, et al. (2002) Genome sequence of a serotype M3 strain of group A Streptococcus: Phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci USA* 99:10078–10083.
23. Ferretti JJ, et al. (2001) Complete genome sequence of an M1 strain of Streptococcus pyogenes. *Proc Natl Acad Sci USA* 98:4658–4663.
24. Paulsen IT, et al. (2003) Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis. *Science* 299:2071–2074.
25. Tettelin H, et al. (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. *Proc Natl Acad Sci USA* 99:12391–12396.
26. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2: A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189–1191.
27. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
28. Brussaard CPD (2009) Enumeration of bacteriophages using flow cytometry. *Methods Mol Biol* 501:97–111.

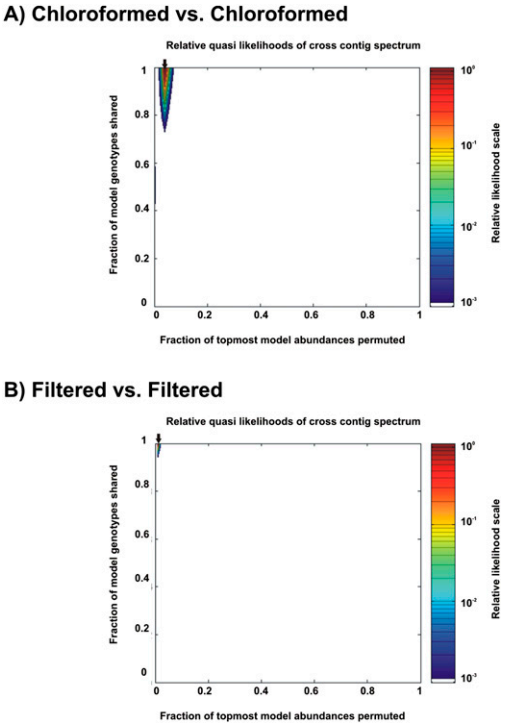

Fig. S1. Monte Carlo analysis of control cross-contig spectra for oropharyngeal metagenomes. Cross-contigs were generated for each metagenome compared to itself (chloroformed in *A*, and filtered in *B*) using Circonspect. Cross-contig spectra for each metagenome versus itself indicated that nearly all species were shared, although very few were permuted. It is expected that a metagenome versus itself would share 100% of sequences with 0% permuted; however, the small deviations seen here are the result of the sampling mechanism implemented in Circonspect.

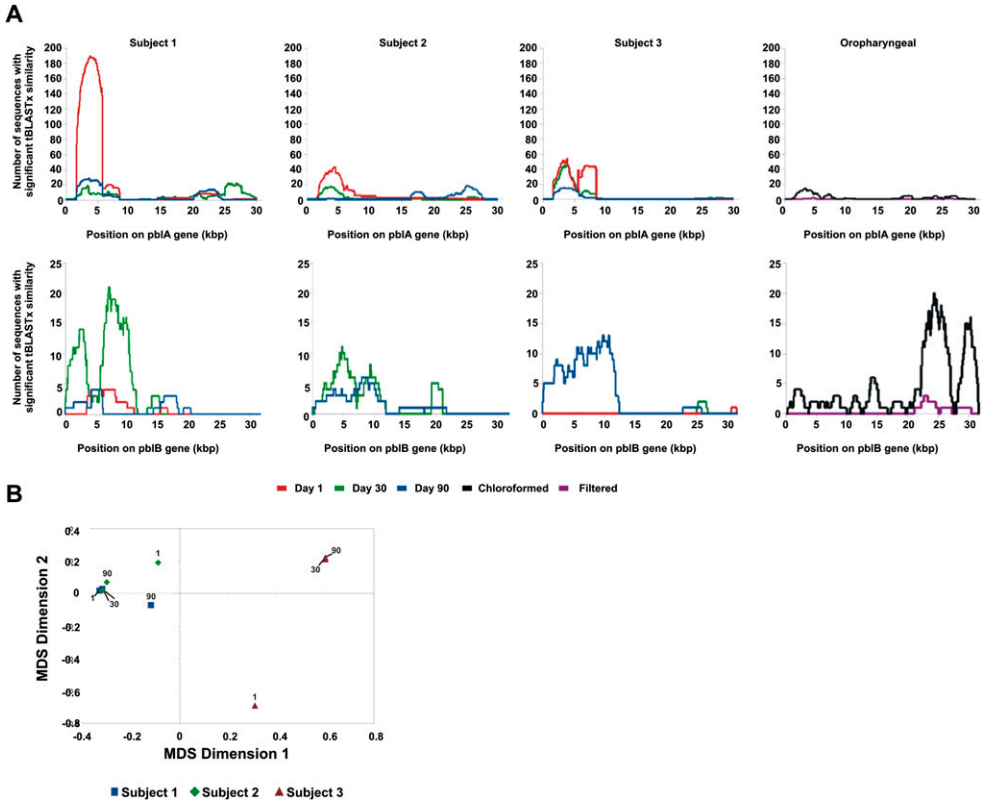

Fig. S2. (*A*) Coverage of the pblA and pblB genes in salivary and oropharyngeal metagenomes. The *x* axis depicts the position on the pblA/pblB gene, and the *y* axis indicates the number of metagenomic sequences covering each position on the gene. A sequence was considered to cover a position if the index was included in the best BLAST alignment of the metagenomic sequence to the gene. (*B*) Scatterplot of multidimensional scaling (MDS) coordinates for sequence dissimilarities between pblA and pblB sequences in salivary metagenomes. Points that are farther apart represent sequence sets that are less similar. The time point for each salivary sample (1, 30, or 90 d) is indicated adjacent to each point on the graph.

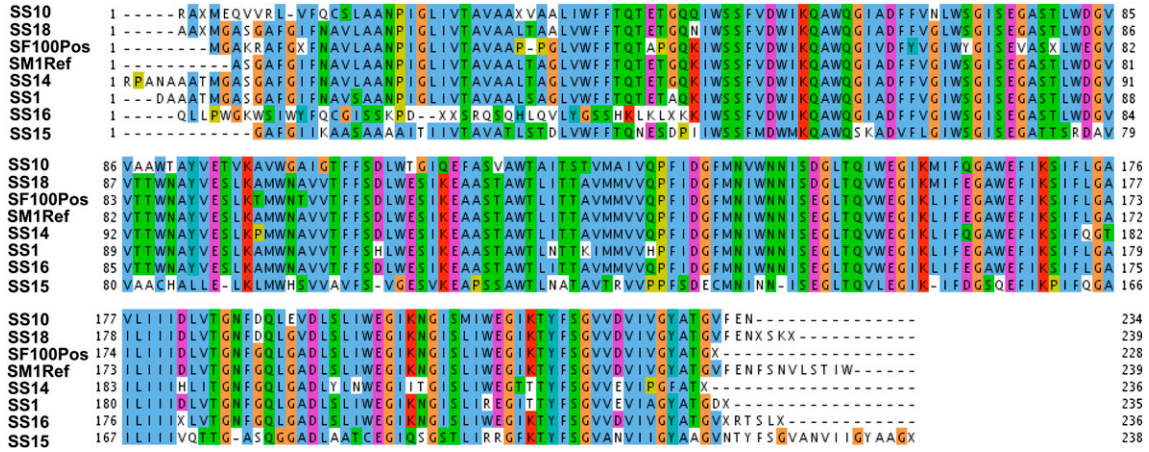

Fig. S3. Nucleotide alignment of saliva pblA sequences with the *S. mitis* SF100 positive control sequence and the SM1 reference sequence.

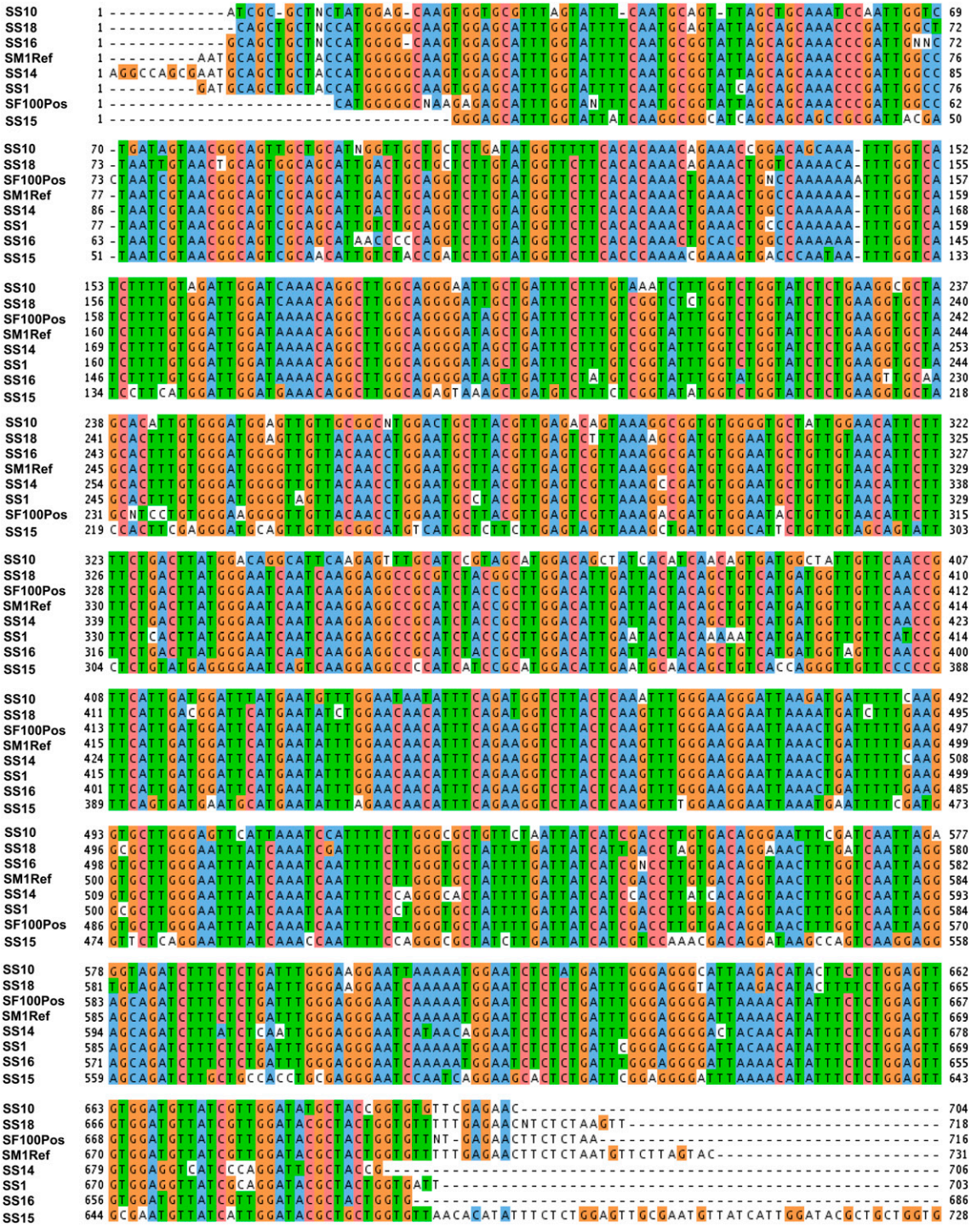

Fig. S4. Amino acid alignment of saliva pblA sequences with the *S. mitis* SF100 positive control sequence and the SM1 reference sequence. The saliva and positive control sequences were translated in all six frames and aligned to determine the correct reading frame before alignment.

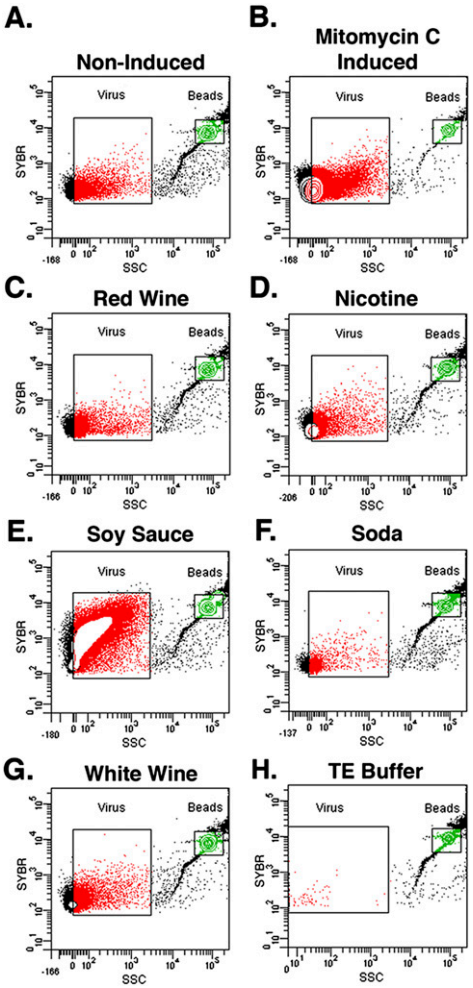

Fig. S5. Representative flow cytometry data of viral induction. Samples were not induced (*A*) or induced under different conditions (*B*–*G*). Samples were processed and analyzed as described in *Materials and Methods*. Contour plots are displayed on biexponential scales with side scatter (SSC) on the *x* axis and SYBR Green I (SYBR) on the *y* axis. Electronic gates were created to collect bead events and virus events as indicated. (*H*) Beads alone in Tris-ethylenediaminetetraacetic acid buffer to determine background signals and generate electronic gate for virus counts.

Table S1. Count of BLASTn similarities to *E. coli* phage T3 in environmental metagenomes

Sample name	BLASTn identities (*e*-value $<10^{-5}$)	Source
Chloroformed oropharyngeal viruses	523	This study
Marine viruses	63	(1)
Human gut microbiome	1	(2)
Coral viruses	1	(3)
Freshwater viruses	35	(4)
Mosquito viruses	5	(4)

Similarities were determined by BLASTn (*e*-value $<10^{-5}$).

1. Angly FE, et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4:e368.
2. Kurokowa, et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 14:169–181.
3. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4:470–483.
4. Dinsdale, et al. (2008) Functional metagenomic profiling of nine biomes. *Nature* 452:629–632.

Table S2. Dissimilarity matrix for pblA sequences between salivary metagenomes

	S1D1	S1D30	S1D90	S2D1	S2D30	S2D90	S3D1	S3D30	S3D90
S1D1	0	0.14	0.82	0.62	0.51	0.44	1	1	1
S1D30	0.14	0	0.72	0.46	0.55	0.58	1	1	1
S1D90	0.82	0.72	0	0.81	0.61	0.58	0.99	0.95	0.96
S2D1	0.62	0.46	0.81	0	0.50	0.29	0.98	0.80	0.77
S2D30	0.51	0.55	0.61	0.50	0	0.36	1	1	1
S2D90	0.44	0.58	0.58	0.29	0.36	0	1	0.96	0.98
S3D1	1	1	0.99	0.98	1	1	0	0.99	0.98
S3D30	1	1	0.95	0.80	1	0.96	0.99	0	0.34
S3D90	1	1	0.96	0.77	1	0.98	0.98	0.34	0

Dissimilarity was calculated as the proportion of unshared sequences between metagenomes as determined by cd-hit-2d-est.

Table S3. Dissimilarity matrix for coverage of pblA genes between salivary metagenomes

	S1D1	S1D30	S1D90	S2D1	S2D30	S2D90	S3D1	S3D30	S3D90
S1D1	0	0.07	0.13	0.10	0.23	0.10	0.30	0.10	0.13
S1D30	0.07	0	0.23	0	0.13	0	0.37	0	0.10
S1D90	0.13	0.23	0	0.27	0.20	0.20	0.10	0.20	0.13
S2D1	0.10	0	0.27	0	0.13	0	0.43	0	0
S2D30	0.23	0.13	0.20	0.13	0	0.13	0.50	0.03	0.10
S3D1	0.30	0.10	0.13	0	0.10	0	0.13	0	0

Dissimilarity was calculated by dividing the pblA gene into 20-bp bins and using the XIPE program to determine which bins were over-represented in each metagenome upon pairwise comparison with every other metagenome. The dissimilarity index is the proportion of nonidentically distributed bins.