

# Metagenomic Analysis of the Viral Communities in Fermented Foods<sup>∇†</sup>

Eun-Jin Park,<sup>1,2‡</sup> Kyoung-Ho Kim,<sup>1,3‡</sup> Guy C. J. Abell,<sup>4</sup> Min-Soo Kim,<sup>1</sup>  
Seong Woon Roh,<sup>1</sup> and Jin-Woo Bae<sup>1\*</sup>

Department of Life and Nanopharmaceutical Sciences and Department of Biology, Kyung Hee University, Seoul 130-701, Republic of Korea<sup>1</sup>; Department of Food Bioengineering, Jeju National University, Jeju 690-756, Republic of Korea<sup>2</sup>; Department of Microbiology, Pukyong National University, Pusan 608-737, Republic of Korea<sup>3</sup>; and CSIRO, Marine and Atmospheric Research and Wealth from Oceans, National Research Flagship, GPO Box 1538, Hobart, Tasmania, Australia<sup>4</sup>

Received 4 August 2010/Accepted 15 December 2010

Viruses are recognized as the most abundant biological components on Earth, and they regulate the structure of microbial communities in many environments. In soil and marine environments, microorganism-infecting phages are the most common type of virus. Although several types of bacteriophage have been isolated from fermented foods, little is known about the overall viral assemblages (viromes) of these environments. In this study, metagenomic analyses were performed on the uncultivated viral communities from three fermented foods, fermented shrimp, kimchi, and sauerkraut. Using a high-throughput pyrosequencing technique, a total of 81,831, 70,591 and 69,464 viral sequences were obtained from fermented shrimp, kimchi and sauerkraut, respectively. Moreover, 37 to 50% of these sequences showed no significant hit against sequences in public databases. There were some discrepancies between the prediction of bacteriophages hosts via homology comparison and bacterial distribution, as determined from 16S rRNA gene sequencing. These discrepancies likely reflect the fact that the viral genomes of fermented foods are poorly represented in public databases. Double-stranded DNA viral communities were amplified from fermented foods by using a linker-amplified shotgun library. These communities were dominated by bacteriophages belonging to the viral order *Caudovirales* (i.e., *Myoviridae*, *Podoviridae*, and *Siphoviridae*). This study indicates that fermented foods contain less complex viral communities than many other environmental habitats, such as seawater, human feces, marine sediment, and soil.

Viruses are the most abundant biological components, with global numbers estimated at  $10^{31}$  (9). Prior to the 1990s, our knowledge of environmental viral communities was limited by the available methodology (3, 25). Viruses cannot be cultivated in the absence of host cells, and the inability to culture most microorganisms under standard laboratory conditions has hindered progress in the field of viral ecology (9). Recently, it has become possible to direct whole metagenome sequencing of uncultured viral assemblages in a process termed “viral metagenomics,” and this advance has dramatically expanded our insight into understanding of viral diversity (39). Researchers are now using this approach to explore viral communities in a wide range of environments, including soil (16), sea (48), potable water (40), activated sludge (33), coral (30), and hot springs (43), as well as in human samples from feces (7, 8, 55), blood (10), and the respiratory tract (51).

In many cultures, fermented foods form a large part of the human diet, and these food products contain complex communities of microorganisms and viruses. Human contact with fermented foods not only results in a direct link between food microbial ecosystems and the ecology of gastrointestinal mi-

crobiota, but it also affects human health (44, 55). Food fermentation is driven by a rich and dense microbial consortia that include bacteria, archaea, and yeast, which are organisms that coexist and interact within a given environment. Fermented foods such as yogurt and cheese are consumed worldwide, and these foodstuffs are commonly prepared by using a selected starter culture. In contrast, kimchi, sauerkraut, and fermented seafood products are fermented by microorganisms present in the raw ingredients. Although kimchi is a fermented vegetable product (Chinese cabbage mixed with various vegetable ingredients and condiments) with a long tradition in Korea, it is now consumed worldwide (more than 1,500,000 metric tons are consumed annually) and is considered to be one of the five healthiest foods (<http://eating.health.com/2008/02/01/worlds-healthiest-foods/>). Sauerkraut is a fermented cabbage side dish consumed commonly in Eastern Europe and North America, and it is prepared with shredded cabbage and 2 to 3% salt (49). Fermented and salted seafood products such as shrimp, shellfish, and anchovies are widely produced and consumed in Asia and Europe (37).

Microorganisms are known to influence the organoleptic qualities and human health benefits of fermented foods. Recent deep-sequencing studies have comprehensively investigated the population structures and dynamics of bacteria and archaea in various types of fermented food ecosystems (20, 37). Despite these new insights, very little is known about the ecological roles that viruses play in fermented food ecosystems. Although a few studies have used plaque formation assays to isolate bacteriophages from kimchi and sauerkraut (27, 28, 54),

\* Corresponding author. Mailing address: Department of Life and Nanopharmaceutical Sciences and Department of Biology, Kyung Hee University, Seoul 130-701, Republic of Korea. Phone: 82 2 961 2312. Fax: 82 2 961 0244. E-mail: baejw@khu.ac.kr.

‡ E.-J.P. and K.-H.K. contributed equally to this study.

† Supplemental material for this article may be found at <http://aem.asm.org/>.

∇ Published ahead of print on 23 December 2010.

a more traditional culture-dependent methodology has only been used for one restricted bacteriophage ecology study on sauerkraut (28). To date, there has been no comprehensive investigation of the overall viral communities in fermented foods.

Human gastrointestinal microbiota can be affected by the viral community composition of fermented foods, and thus it is important to understand these complex communities and their potential impacts on human health. However, there has been very limited investigation into the viral and corresponding microbial diversity of fermented food communities. This study reports metagenomic analyses of three different fermented food samples (fermented shrimp, kimchi, and sauerkraut) and describe the viral and bacterial communities they contain.

## MATERIALS AND METHODS

**Sample preparation.** Two fermented vegetable products (kimchi and sauerkraut) and a fermented seafood product (fermented shrimp) were selected for this experiment. Kimchi (Chinese cabbage [*baechukimchi* in Korean]) was obtained from Chongga Food (Seoul, Korea) and fermented in the laboratory at 12°C for 3 days to reach optimal ripeness (pH 4.2 to 4.7). Sauerkraut was prepared in the laboratory using a typical recipe in which fresh cabbage was shredded and mixed with 2.0% (wt/wt) salt. The sauerkraut mixture was then placed in a sterile glass jar and fermented for 7 days at 15°C. The fermented shrimp sample was obtained from a fish market (Incheon, Korea), transported immediately to the laboratory, and analyzed within 12 h. The fermented shrimp, kimchi, and sauerkraut had pH values of 6.8, 4.4, and 4.0, respectively.

For analysis of viral communities, food particles were removed by centrifugation at  $2,500 \times g$  for 20 min at room temperature. The supernatant was serially filtered through sterile gauze and 8- $\mu\text{m}$ -pore-size nitrocellulose filter paper (Whatman, Ltd., United Kingdom), followed by 0.45- and 0.20- $\mu\text{m}$ -pore-size syringe filters (Sartorius, Germany). To collect virus samples, the fermented shrimp, kimchi and sauerkraut filtrates (approximately 700, 500, and 300 ml, respectively) were ultracentrifuged at  $100,000 \times g$  for 4 h at 4°C in a swinging-bucket rotor (Optima L-100 K, Beckman Coulter, Fullerton, CA). Pelleted viruses were resuspended in 1 ml of deionized water prior to viral DNA extraction.

**TEM.** Virus-containing samples were passed through a 0.20- $\mu\text{m}$ -pore-size syringe filter and then loaded onto a cesium chloride (CsCl) density gradient and ultracentrifuged at  $88,250 \times g$  for 2 h. A fraction (1.2 to 1.5 g ml<sup>-1</sup>) was collected and dialyzed against 10 volumes of Tris-EDTA (TE) buffer. The samples were then reloaded onto a second CsCl density gradient, and the process repeated to increase viral purity. Fractions were pooled and resuspended in 10 volumes of TE buffer, followed concentration with centrifugal filters (Amicon Ultra-15, 100 kDa). Purified viruses were examined by transmission electron microscopy (TEM). Copper grids with 200-mesh carbon-coated Formvar were floated on a droplet of virus sample, negatively stained with 2% uranyl acetate for 30 s, washed twice with deionized water, and then air dried. Images ( $\times 50,000$  and  $\times 100,000$  magnifications) were collected by using an energy-filtering transmission electron microscope (Libra 120; Carl Zeiss, Germany).

**Viral DNA extraction and LASL-PCR amplification for pyrosequencing.** Viral DNA was extracted from concentrated viral suspensions, as described previously (41). The resulting DNA was purified by using an UltraClean microbial DNA isolation kit (Mo Bio Laboratories, California) and then quantified with a spectrophotometer (Nanodrop Technologies, Rockland, DE) prior to PCR amplification. Extracted viral DNA was amplified by using a linker-amplified shotgun library (LASL) technique (11). Amplification of PCR products was confirmed by examining 1- $\mu\text{l}$  reaction mixtures by 1.0% agarose gel electrophoresis (0.5 $\times$  TAE buffer). PCRs were purified by using a PCR product purification kit (SolGent).

**Viral metagenome pyrosequencing.** LASL-amplified viral DNA (11) was sequenced with a genome sequencer (GS) FLX titanium apparatus (Roche 454 Life Science) (29). An eighth of a PicoTiter plate was used for sequencing each of the viral metagenomes from the three food samples. The sequences obtained in the present study were deposited in GenBank under accession number SRP002583.

**Viral metagenome analysis.** LASL linker sequences were removed from raw pyrosequencing reads by using the Wordfinder program in the Emboss package (35). Sequence reads longer than 100 bp that contained fewer than two ambig-

uous bases (N's) were selected for further analysis using dedicated Python scripts; 12 to 17% of the sequences were removed at this stage. The sequences were compared to the GenBank database using BLASTN, and those showing greater than 98% identity and 90% coverage with *Escherichia coli* and *E. coli* phage DE3 strains were considered likely contaminants; 11 to 16% of the sequences were excluded from analysis at this step. These artifacts appeared to arise from residual *E. coli* DNA in the enzyme preparations used for the ligation and amplifications steps.

BLASTX analyses were performed against the GenBank nonredundant (nr) protein database (1) and the ACLAME protein database (23, 24). The two results were then merged. After the ratio of E-value and score for best hits from both analyses were compared, the ACLAME results were prioritized in the process of classification when the log (E-value) ratio of ACLAME's best hit to GenBank's best hit was lower than 0.5 and the score ratio was higher than 0.75. Since the ACLAME database represents sequences from phages, plasmids, prophages, and transposons, the classification of cellular organisms and nonphage viruses was based on the GenBank results. Sequences with no hits or hits with E-values  $>0.001$  were regarded as unknown. Classification of phage hosts was based on host information in the ACLAME database. Classification of viruses was performed according to the National Center for Biotechnology Information (NCBI) viral classification system, as proposed originally by the International Committee on Taxonomy of Viruses. MEGAN (MEtaGenome ANalyzer) software was used to compare the taxonomic groups of viruses present in the three samples (13).

Viral metagenome structure and diversity estimations were determined by alpha diversity analysis using the CAMERA 2 server (<http://camera.calit2.net/>). Briefly, the Circonspect tool (<http://sourceforge.net/projects/circonspect/>) was used to generate average contig spectra from over 10 assemblies of 10,000 randomly sampled sequences. Minimal assembly parameters were set at 98% identity and a 35-bp overlap. To obtain an identical read length, all sequences were trimmed to 100 bp prior to sequence selection and assembly. Average genome length was calculated on the basis of BLAST results (E-value  $< 10^{-5}$ ) against 510 complete phage genomes (Rohwer lab's phage genome database). Two parameters, average contig spectra and average genome length, were examined with the Phage Communities using the Contig Spectra (PHACCS) tool (2). Community structure and diversity estimates were determined by using the rank-abundance model that showed the lowest error. To identify similar proteins, the metabolic profile of each virome was assessed by using BLASTX (E-value  $< 10^{-5}$ ) against the SEED-subsystems database, which contains a collection of genes with related metabolic functions (<http://www.theseed.org>). These assessments were performed by using the MG-RAST service (31).

**Viral contig assembly and analysis.** Contigs were assembled by the Genome Sequencer De Novo Assembler (Roche) using all of the reads of each sample. The assembler used a minimum length and percent identity of overlaps of 20 and 98%, respectively. Default values were used for all other parameters during assembly. Coverage of the largest contigs was calculated by multiplying the read number and average read length and then dividing this value by the contig length. The Glimmer program (14) was used to predict metagenomic open reading frames (ORFs) from contigs. BLASTX was then used to compare predicted ORFs and intergenic sequences to the CAMERA databases (45). When these sequences were compared to CAMERA's nonidentical protein database, it was evident that almost all contigs contained ORFs and intergenic regions with similarity to phage or prophage gene sequences. The Artemis and ACT programs (12) were used to compare contigs to their cognate phage genomes.

**Bacterial 16S rRNA gene amplification and pyrosequencing.** Bacterial DNA was extracted from each sample (1 g of product) by using a bead-beating method (53). Extracted DNA was purified and quantified as previously described (5). Hypervariable regions (V1 to V3) of the bacterial 16S rRNA gene were amplified from bacterial DNA extracts using the universal primer pair 8F (5'-ACGAGT TGTGATCMTGGCTCAG-3') and 518R (5'-ACWTTACCGCGGCTGCTGG-3') (4). The PCR conditions were as follows: 1 min at 94°C; 18 cycles of denaturation (1 min at 94°C), annealing (30 s at 60°C), and extension (30 s at 72°C); followed by a final elongation (10 min at 72°C). Barcodes comprising sequences of 4 to 6 nucleotide (nt) sequences (ACGT, GTACT, and TCATCG) were included at the 5' ends of primers to analyze all three samples in a single 454 pyrosequencing run. The resulting sequences were assigned to the appropriate sample based on these barcodes. Agarose gel electrophoresis (1.0% gel; 0.5 $\times$  TAE buffer) of each PCR mixture (1  $\mu\text{l}$ ) was used to confirm that the amplified products were the correct size (530 nt). PCR products were subsequently purified by using a QIAquick PCR purification kit (Qiagen, Valencia, CA), and equal amounts of each barcoded PCR product (1  $\mu\text{g}$ ) were pooled (total, 3  $\mu\text{g}$ ). PCR products that contained purified bacterial 16S rRNA genes were amplified by

emulsion PCR prior to sequencing in an eighth of a PicoTiter plate, as described above.

**Analysis of bacterial 16S rRNA gene sequences.** After removal of barcode sequences, reads shorter than 300 bp and containing more than one ambiguous base (N) were excluded. The Ribosomal Database Project Release 10 (RDP) classifier was used for archaeal and bacterial taxonomic classification at the genus level, according to Bergey's taxonomy (50). Bacterial community structures were assigned to ratios according to the numbers of RDP-classified sequences in each sample.

## RESULTS AND DISCUSSION

**Morphologies of viruses in fermented foods.** TEM indicated that the majority of purified particles resembled well-known viral morphologies (tailed, filamentous, and polyhedral) (see Fig. S1 in the supplemental material). Most morphotypes were 50 to 100 nm in diameter, although some viruses were <50 nm and >100 nm in diameter. TEM detected myovirus-like (see Fig. S1A, S1C, and S1E in the supplemental material) and siphoviruslike (see Fig. S1B and S1D in the supplemental material) viruses, as well as three virions that formed unidentified clusters (see Fig. S1F in the supplemental material).

**Taxonomic composition of viral metagenomes.** A total of 309,445 viral metagenome sequences were obtained. After pre-processing, 81,831, 70,591, and 69,464 viral sequences were identified for fermented shrimp, kimchi, and sauerkraut, with average read lengths of 373, 379, and 345 bp, respectively. No significant identity (E-value of >0.001 or significant matches) was found between 46% (fermented shrimp), 37% (kimchi), and 50% (sauerkraut) of the selected sequences and protein sequences in the extant GenBank and ACLAME databases (see Fig. S2 in the supplemental material). "Known" sequences with E-values of <0.001 were classified as virus, prophage, plasmid, bacteria, eukarya, or archaea, according to the best hit. Specifically, 40, 32, and 34% (fermented shrimp, kimchi, and sauerkraut, respectively) of the known sequences were most similar to viral protein sequences (see Fig. S2 in the supplemental material); 29, 18, and 24% of the identified sequences showed similarities to prophage sequences; and virus-related sequences, including both virus and prophage, accounted for 50 to 69% of known sequences of the three fermented foods. Unlike culture-dependent investigations of phage diversity, which require specific microbial hosts, metagenomic analyses such as that performed in the present study produce large numbers of viral sequences that have no or very low identity to sequences in public databases. This phenomenon has been reported previously for viromes in other environments, including an Antarctic lake (25), seawater (3), activated sludge (33), and reclaimed and potable water (40), as well as from human fecal and respiratory tract samples (8, 51). The higher ratio of unknown genes might also be related to the short read length (100 to 400 bp) of metagenome sequences causing fewer homolog than long read sequences (>750 bp) (15, 52). Despite a virus-targeting extraction methodology, a number of the "known" viral metagenome sequences showed a very strong similarity to bacterial sequences. This finding may be due to the incorporation of host genes into viral genomes, uncharacterized prophage sequences and the bias of databases toward microbial rather than viral genome sequences (11, 25, 40).

In reads identified as viral in origin, the majority (99.3 to

99.9%) of sequence showed the greatest similarity to phages (see Fig. S2 in the supplemental material). Double-stranded phages accounted for more than 99.9% of the phages amplified using the LASL technique, and they belonged to the order *Caudovirales* (*Myoviridae*, *Podoviridae*, and *Siphoviridae*). The LASL method amplifies only double-stranded DNA (dsDNA), so diversity of RNA viruses and/or single-stranded DNA (ssDNA) viruses would not be investigated in the present study. While the *Siphoviridae* family was abundant in the fermented shrimp (53.55%) and sauerkraut (60.07%), the most abundant viral family in kimchi was the *Podoviridae* (52.82%).

This study used the LASL amplification technique to examine dsDNA viruses in fermented food viromes. However, these samples only contained a few dsDNA viral families that all belonged to the order *Caudovirales*. Other environmental habitats such as an Antarctic lake or paddy soil have been shown to contain ssDNA viruses related to other viral families (e.g., *Circoviridae*, *Geminiviridae*, *Inoviridae*, *Nanoviridae*, and *Microviridae*) (Fig. 3). The abundance of ssDNA viruses can be assessed by using the multiple displacement amplification method (21, 42) and the diversity of ssDNA viruses in fermented food requires further study. Of the dsDNA viruses, the *Siphoviridae* and *Myoviridae* families are known to infect bacteria. Their prevalence in the fermented shrimp and sauerkraut samples is consistent with a previous study that used conventional plaque assays to examine bacteriophage diversity in commercial sauerkraut and which isolated a total 171 phages (28).

Figure 1 shows the relative abundance of viral genera in the three viromes, according to their taxonomic classifications. T4- and T7-like viruses were distributed evenly in all three samples, but various genera, including phiKMV, SP6, SPO1, N4, lambda, phi29, T1, and 44AHJD-like viruses, were unevenly distributed or sometimes detected in only one sample. A small number of sequences related to the *Phyconaviridae* family were identified in the fermented shrimp (0.4%) and kimchi (0.1%) samples. Members of the *Phyconaviridae* family are the largest viruses known to infect harmful eukaryotic algae, including phytoplankton species responsible for algal blooms (22). Given that the fermented shrimp is made from raw shrimp and associated seawater, these viruses may originate from the marine environment, and their existence in the fermented shrimp sample may simply reflect the source of the raw material.

The predicted hosts were examined for all of the top hits showing similarities to phage sequences and the relative abundance of phage host was noted for the three phage families (*Myoviridae*, *Podoviridae*, and *Siphoviridae*) (see Fig. S3 in the supplemental material). The three fermented products showed differences in host distribution, which suggests the presence of distinct phage communities. A more even distribution of phage host was detected in the fermented shrimp sample than in the kimchi or sauerkraut.

**Comparison of host diversity of bacteriophages to bacterial diversity in each fermented food.** A total of 101,752 16S rRNA gene sequences were PCR amplified from bacterial metagenomic DNA obtained from three fermented foods. Bar-coded pyrosequencing was used to determine that fermented shrimp, kimchi, and sauerkraut generated 15,549, 40,000, and 21,419 bacterially derived sequences, respectively. At the genus

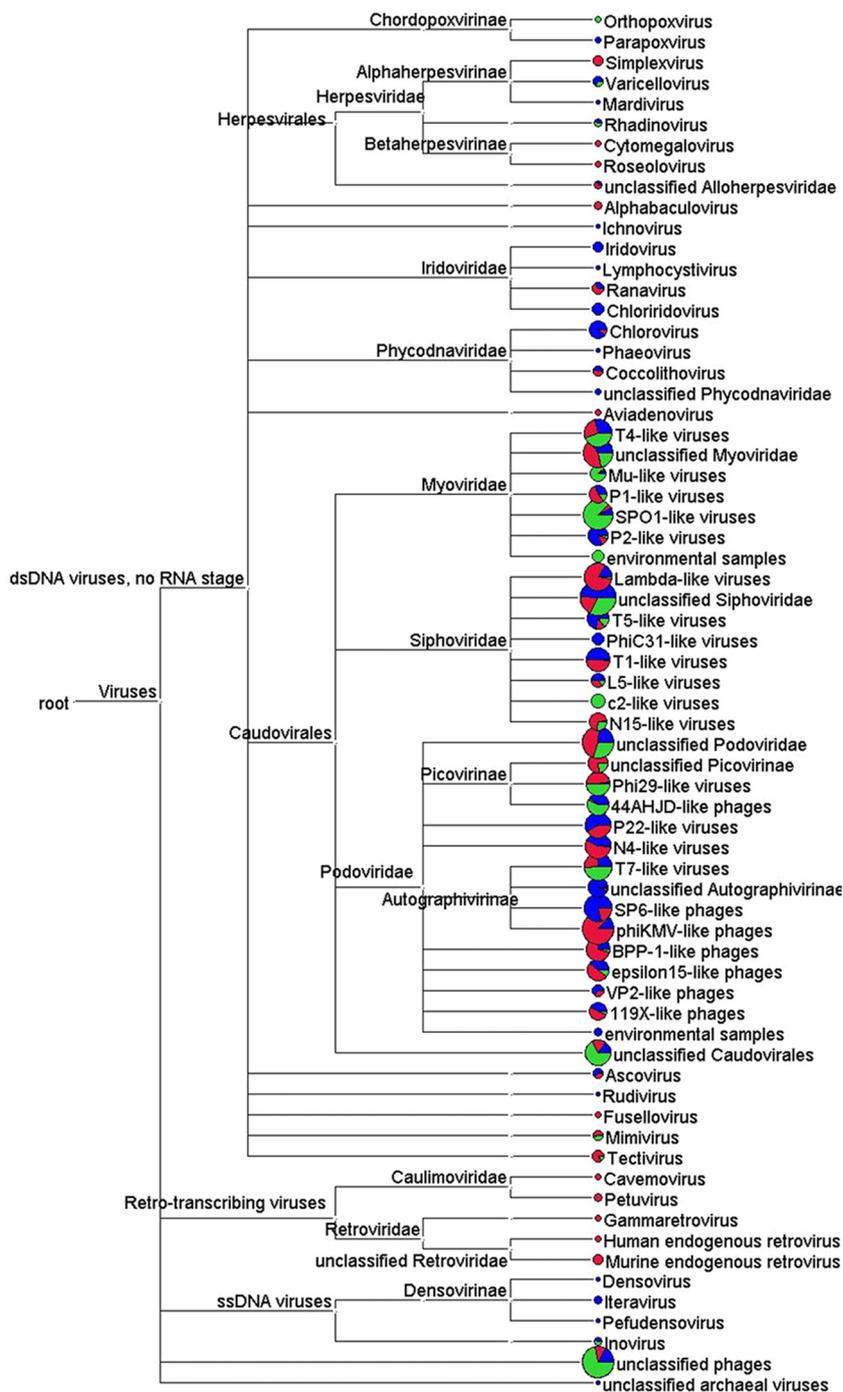


FIG. 1. Overview of relative abundance and distribution of viral taxonomic groups based on MEGAN analysis of viral genotypes from the three fermented food viral metagenomes. Colors: blue, fermented shrimp; red, kimchi; green, sauerkraut.

level, 45 to 99% of the reads could be classified into three to five bacterial genera, while the three viral metagenomes indicated the presence of nine different types of phage and prophage host (Fig. 2). The relative abundance of genera in the bacterial community was then compared to viral community phage and prophage host abundance in the three fermented foods (Fig. 2). Although a similar bacterial community was observed in both fermented vegetable foods, i.e., kimchi and

sauerkraut, the fermented shrimp sample exhibited a distinctly different bacterial community. At the genus level, *Pseudomonas* (31.6%), *Cupriavidus* (7.6%), and *Propionibacterium* (5.6%) were assigned to core bacterial groups in the fermented shrimp, whereas 98.5 and 88.6% of the total reads were assigned to only three genera in kimchi and sauerkraut. Previously, culture-dependent methodologies have identified *Leuconostoc*, *Lactobacillus*, and *Pediococcus* as dominant

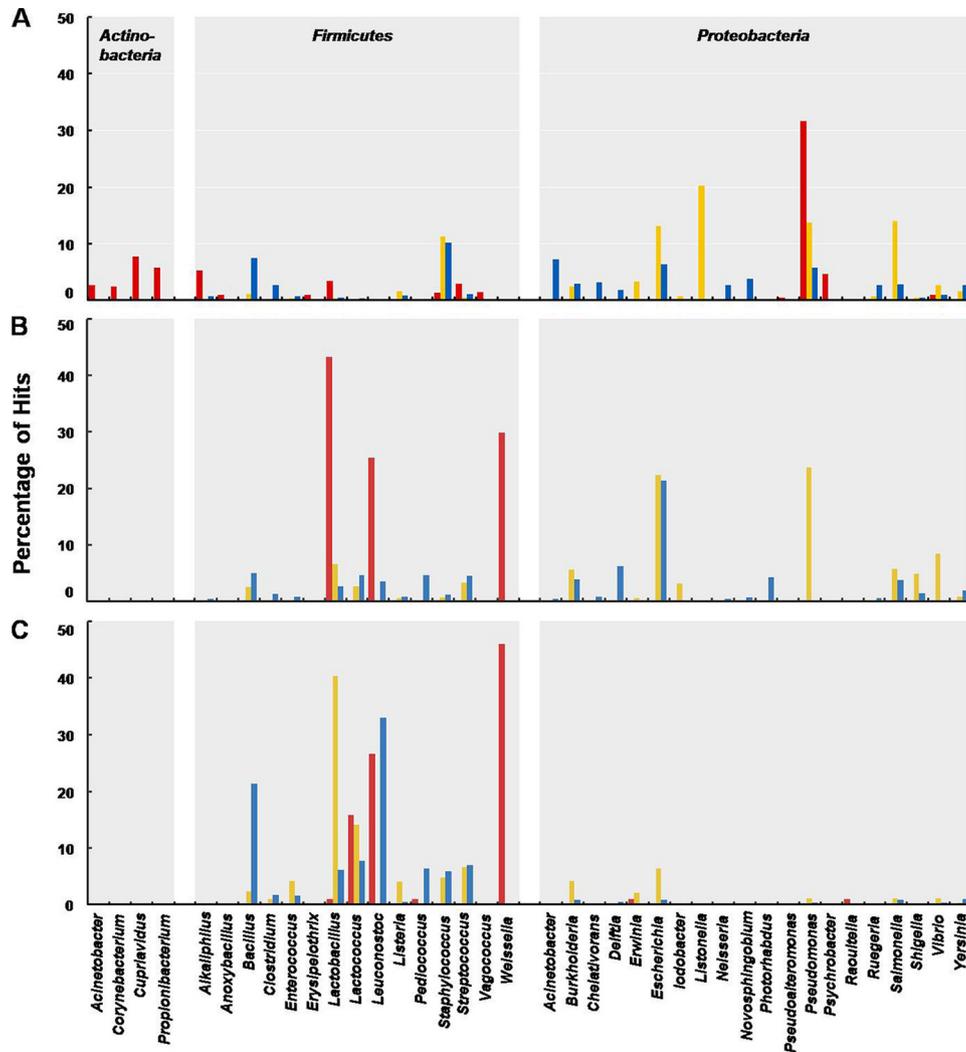


FIG. 2. Comparison of host diversity of phage and prophages to bacterial diversity in each of the three fermented foods. Phage host and prophage host diversity were identified comparing to the GenBank and ACLAME databases, and bacterial community was determined based on 16S rRNA gene sequences. (A) Fermented shrimp; (B) kimchi; (C) sauerkraut. Colors: red, 16S rRNA gene-based bacterial community; yellow, phage host distribution; blue, prophage host distribution.

bacterial genera in sauerkraut fermentation (34). Moreover, previous culture-dependent and -independent studies of bacterial species in kimchi have demonstrated the predominance of several genera of lactic acid bacteria, including *Weissella* and *Leuconostoc* (5, 32). Similarly, the present study revealed that *Lactobacillus* (43.2%), *Weissella* (29.9%) and *Leuconostoc* (25.4%) were the dominant genera in kimchi and that *Weissella* (46.1%), *Leuconostoc* (26.6%), and *Lactococcus* (15.9%) were the most common in sauerkraut.

In any environment, the abundance of host bacterial strains is thought to relate to phage abundance. A long-term study provides total bacterial and viral numbers for a given environment, allowing comparison of phage-host interactions (46). Although these methods indicate apparent population sizes, they cannot provide data for the abundance of a specific phage or its co-occurrence with a corresponding host strain. In the present study, metagenomic sequences were used to examine the composition of viral and host communities (Fig. 2).

For the fermented shrimp, the major genera of phage hosts were predicted to be *Listonella* (20.2%), *Escherichia* (13.0%), and *Staphylococcus* (11.2%). In contrast, there was a more even distribution of the prophage hosts *Staphylococcus* (10.1%), *Bacillus* (7.4%), and *Acinetobacter* (7.2%). Analysis of kimchi reads indicated that the phage hosts were associated predominantly with the genera *Pseudomonas* (23.6%), *Escherichia* (22.3%), and *Vibrio* (8.3%), while the genera *Escherichia* (21.2%), *Delftia* (6.2%), and *Bacillus* (4.9%) represented the core group of prophage hosts. There was little similarity between the microbiomes and viromes in either the fermented shrimp or the kimchi. In sauerkraut, however, the dominant bacterial strains and predicted viral hosts were remarkably similar. The most abundant phage host genera were *Lactobacillus* (40.3%), *Lactococcus* (14.1%), and *Streptococcus* (6.6%). The dominant prophage host genera were predicted to be *Leuconostoc* (32.9%), *Bacillus* (21.3%), and *Lactococcus* (7.6%). This agreement between microbiome and virome may be due to the

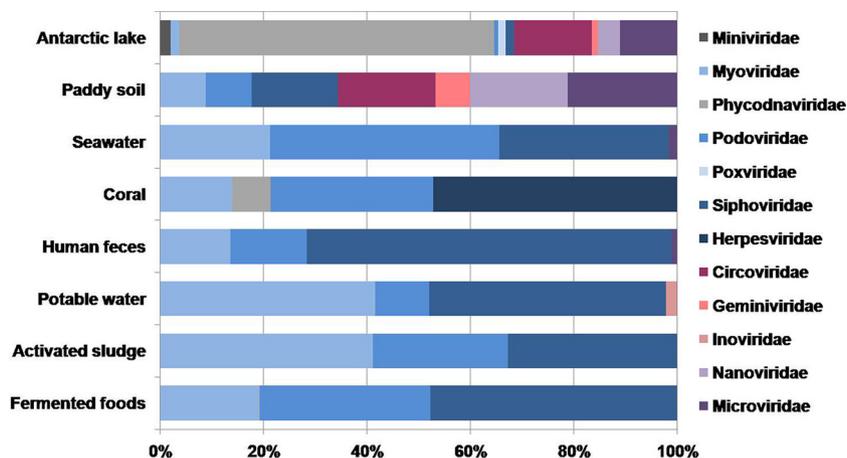


FIG. 3. Comparison of viral communities represented as viral families in fermented foods, potable water, human feces, seawater, coral, paddy soil, and an Antarctic lake. The percent representations of divisions in each virome are shown. The activated sludge viral metagenome sequences ( $n = 109$ ) (33), the potable water viral metagenome sequences ( $n = 231,715$ ) (40), the human feces data obtained and combined from two independent studies ( $n = 95$ ) (7, 8), the coral metagenome sequences ( $n = 1,580$ ) (30), the seawater data set ( $n = 131$ ) (11), the paddy soil data ( $n = 785$ ) (21), and the Antarctic lake virome ( $n = 3,360$ ) (25) were obtained from each of the studies, and the fermented-food virome was obtained from the present study.

fact that the databases contain more phages that infect lactic acid bacteria than any other group. It is important to consider, however, that these analyses are based on currently available viral databases and that some dominant but previously unreported viruses may have been omitted from the viral abundance results. Bacteriophage host prediction represents another area of uncertainty, since viral genomes are highly diverse and using homology comparison to discern the host of unknown viruses could lead to the wrong conclusions.

**Discrepancy between the diversity of bacteriophage hosts and bacterial diversity.** In any environment, the richness and quantity of viruses will largely be determined by the corresponding microbial community. Numerous studies have demonstrated high similarities between the microbiomes and host distribution of viromes within the same environment and low similarities between those from different environments (9, 36). In fermented food preparations, the microbial ecosystem is effectively enclosed within a container, and therefore these microbial communities may develop specifically enriched community profiles. Thus, the low species richness of viral communities in fermented foods may reflect the relatively low diversity of the dominant microbial hosts. However, the present study shows that both fermented shrimp and kimchi exhibit a discrepancy between the dominant bacterial genera and the predicted host distributions of phage and prophage (Fig. 2). Only in sauerkraut did the major bacterial groups (*Leuconostoc* and *Lactococcus*) reflect the major predicted hosts of phages and prophages. This discrepancy between host and dominant microbial profile may be explained by the current shortage of databases containing food viromes. The genomic characterization of phages infecting the genera *Weissella* and *Leuconostoc*, the predominant genera in sauerkraut and kimchi, has not been reported thus far, whereas the complete genome sequence of phage isolated from commercial sauerkraut has been reported (26). Thus, it is not surprising that a better agreement was observed between the distribution patterns of bacteria and phage host in the analysis of sauerkraut than in the fermented

shrimp and kimchi. At the time of writing, the ACLAME and NCBI databases only contained 465 and 579 phage genomes, respectively. In contrast, the prophage sequences in the ACLAME database numbered 754. Compared to the total number of completed archaeal or bacterial genome sequences ( $>1,000$ ), the number of phage genomes sequenced is small, especially given the size difference between the cumulative phage and bacterial genomes (a total length of only 3.8 Mbp for 579 phage genomes compared to 3.6 Gbp for 1,089 bacterial genomes). Furthermore, there was a significant difference between the phage and prophage host distributions in the ACLAME database. We identified here a large number of unknown viral sequences and found inconsistencies between viral host and bacterial diversity. These findings are likely due to the lack of fermented food phage sequences in the sequence databases and to the intrinsic limitations of using homology to known viral hosts for predicting the hosts of unknown viruses.

**Diversity estimation and metabolic profiles of fermented food viromes.** The PHACCS online tool was used to perform diversity estimates for the three fermented food viromes. The fermented shrimp exhibited a much greater species richness than kimchi or sauerkraut (see Table S1 in the supplemental material), which showed relatively low richness compared to environments such as surface seawater (11), human feces (8), marine sediment (6), and soil (16). However, these diversity comparisons should be considered carefully, since different research conditions and PHACCS analysis parameters were

TABLE 1. Overview of contig assemblies of viral metagenomes from three fermented foods

Sample	Total % assembled sequences	Largest contig size (bp)	No. of contigs >100 bp	No. of singletons	No. of contigs >5,000 bp
Fermented shrimp	72	13,325	11,096	23,581	6
Kimchi	70	43,456	3,266	20,635	27
Sauerkraut	82	57,863	3,019	8,458	26

TABLE 2. Structure and diversity estimation by using PHACCS analysis of viral metagenomes from three fermented foods

Sample	Rank-abundance model	Avg genome size (bp)	Error	Richness	Evenness	Most abundant genotype (%)	Shannon index
Fermented shrimp	Lognormal	39,839	168	7,310	0.663	10.40	5.90
Kimchi	Power	48,220	72.8	279	0.867	9.57	4.88
Sauerkraut	Lognormal	36,494	444	108	0.887	8.98	4.15

used in the various studies. Moreover, a high degree of uncertainty in estimating richness was reported (16). The Shannon index and evenness also indicated that fermented shrimp contained a more diverse viral assembly than the other fermented foods. However, the most abundant genotype occupied 9 to 10% of the total assembly in all fermented food.

The SEED database was used to examine the metabolic potential encoded by viral genes. The three fermented food viromes shared a common viral metabolic profile, which comprised 27 classifications (see Fig. S4 in the supplemental material). The analysis of metabolic potential showed that the major groups of functionally related genes encoded proteins involved in virulence and DNA or carbohydrate metabolism. Of all genes classified from fermented shrimp, kimchi, and sauerkraut, 26.0, 45.4, and 33.3%, respectively, were predicted to be involved in virulence. Most of these (69.5 to 91.5%) were subcategorized as prophage- and transposon-related genes.

**Assembly of metagenomes.** The contig assemblies of metagenome sequences from each sample are presented in Tables 1 and 2. Interestingly, a large number (70 to 82%) of these metagenome sequences assembled into several thousand contigs. Despite a similar percentage of assembled sequences, the fermented shrimp metagenome contained a different assembly pattern than the other two fermented foods. The largest contig from the fermented shrimp metagenome (13,325 bp) was much smaller than that of kimchi (43,456 bp) or sauerkraut (57,863 bp). Moreover, 6, 27, and 26 contigs of >5,000 bp were assembled from fermented shrimp, kimchi, and sauerkraut, respectively (Table 1; see also Table S1 in the supplemental material), and most of these contigs showed at least one hit with a phage-related gene. Since the assembled contigs of kimchi and sauerkraut were longer and more numerous than those of fermented shrimp, it is likely that these fermented vegetables have lower viral diversity, a finding that is in agreement with the PHACCS analysis. As shown in Table S1, a large number of reads could be assembled into single contigs and many contigs showed high coverage (between 10- and 70-fold). In the sauerkraut metagenome, contigs showing relatively high coverage (30 to 70X) were observed in greater numbers than in the kimchi viral metagenome. This finding implies that the major viral genotypes identified in sauerkraut occupy a larger proportion of the total viral community than those in kimchi. Contigs >5,000 bp are shown in Table S1 in the supplemental material.

Similarities to previously described viral genomes can be observed by comparing the structures of several viral metagenome contigs. In contig 03074 (43,456 bp in length, 71-fold coverage), which was assembled from the kimchi metagenome, a large proportion of the predicted ORFs show similarities to sequences from *Pseudomonas* phage LKD16. A TBLASTX comparison between contig 03074 and the *Pseudomonas* phage LKD16 genome demonstrated similar structures in the two

sequences (see Fig. S5A in the supplemental material). However, there is low average amino acid identity (39%) with known proteins, which suggests that contig03074 represents the putative genome of a previously undescribed phage (see Table S2A in the supplemental material).

**Significance of the investigation of viral communities in fermented foods.** Metagenomic approaches present a fascinating opportunity to identify previously uncultured microorganisms and to understand their biodiversity, function, interactions, and evolution in different environments (3, 8, 9, 11, 38). In this discipline, the recent introduction of large-scale high-throughput DNA sequencing techniques has allowed access to a large body of metagenomic sequencing information, circumventing the need for cultivation or gene cloning. To date, large metagenomic sequencing projects have mainly focused on bacterial ecology (13, 16, 37), although more recently the focus of studies has broadened to include the investigation of eukarya (17), archaea (37), and viruses (25, 40, 43, 47). Over the past decade, the emphasis of viral research in foods has been on bacteriophages and, in particular, on the roles that they play in the inhibition of dairy starter culture growth and control of pathogenic microorganisms (18).

Metagenomic studies of environmental viral communities have revealed that viruses are more highly abundant and widespread in marine environments than previously thought (3, 6, 11, 60). Our viral metagenomic data, which were obtained from three samples that represent different types of fermented food, have also shown large numbers of previously undiscovered viral sequences. Viral metagenomic studies of various environments, such as human feces, marine water, and fermented foods, represent a valuable avenue for investigating the genetic and biochemical diversity of viruses. Such studies help further our understanding of the ecological roles played by viruses in the environment.

#### ACKNOWLEDGMENTS

We thank Forest Rohwer (San Diego State University) for helpful discussion and critical reading of the manuscript.

E.-J.P. was supported by a National Research Foundation grant funded by the Korean Government (NRF-2009-2-C00179). This study was supported by the Technology Development Program for Agriculture and Forestry.

#### REFERENCES

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.
- Angly, F., et al. 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinform.* **6**:41.
- Angly, F. E., et al. 2006. The marine viromes of four oceanic regions. *PLoS Biol.* **4**:e368.
- Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* **71**:7724–7736.

5. **Bae, J. W., et al.** 2005. Development and evaluation of genome-probing microarrays for monitoring lactic acid bacteria. *Appl. Environ. Microbiol.* **71**:8825–8835.
6. **Breitbart, M., et al.** 2004. Diversity and population structure of a near-shore marine-sediment viral community. *Proc. Biol. Sci.* **271**:565–574.
7. **Breitbart, M., et al.** 2008. Viral diversity and dynamics in an infant gut. *Res. Microbiol.* **159**:367–373.
8. **Breitbart, M., et al.** 2003. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**:6220–6223.
9. **Breitbart, M., and F. Rohwer.** 2005. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**:278–284.
10. **Breitbart, M., and F. Rohwer.** 2005. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* **39**:729–736.
11. **Breitbart, M., et al.** 2002. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* **99**:14250–14255.
12. **Carver, T., et al.** 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**:2672–2676.
13. **Claesson, M. J., et al.** 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One* **4**:e6669.
14. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**:4636–4641.
15. **Edwards, R. A., and F. Rohwer.** 2005. Viral metagenomics. *Nat. Rev. Microbiol.* **3**:504–510.
16. **Fierer, N., et al.** 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* **73**:7059–7066.
17. **Ghannoum, M. A., et al.** 2010. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog.* **6**:e1000713.
18. **Gorski, A., and B. Weber-Dabrowska.** 2005. The potential role of endogenous bacteriophages in controlling invading pathogens. *Cell. Mol. Life Sci.* **62**:511–519.
19. Reference deleted.
20. **Humblot, C., and J. P. Guyot.** 2009. Pyrosequencing of tagged 16S rRNA gene amplicons for rapid deciphering of the microbiomes of fermented foods such as pearl millet slurries. *Appl. Environ. Microbiol.* **75**:4354–4361.
21. **Kim, K. H., et al.** 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl. Environ. Microbiol.* **74**:5975–5985.
22. **Larsen, J. B., A. Larsen, G. Bratbak, and R. A. Sandaa.** 2008. Phylogenetic analysis of members of the *Phycodnaviridae* virus family, using amplified fragments of the major capsid protein gene. *Appl. Environ. Microbiol.* **74**:3048–3057.
23. **Leplae, R., A. Hebrant, S. J. Wodak, and A. Toussaint.** 2004. ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res.* **32**:D45–D49.
24. **Leplae, R., G. Lima-Mendez, and A. Toussaint.** 2006. A first global analysis of plasmid encoded proteins in the ACLAME database. *FEMS Microbiol. Rev.* **30**:980–994.
25. **Lopez-Bueno, A., et al.** 2009. High diversity of the viral community from an Antarctic lake. *Science* **326**:858–861.
26. **Lu, Z., E. Altermann, F. Breidt, and S. Kozyavkin.** 2010. Sequence analysis of *Leuconostoc mesenteroides* bacteriophage Phi1-A4 isolated from an industrial vegetable fermentation. *Appl. Environ. Microbiol.* **76**:1955–1966.
27. **Lu, Z., F. Breidt, Jr., H. P. Fleming, E. Altermann, and T. R. Klaenhammer.** 2003. Isolation and characterization of a *Lactobacillus plantarum* bacteriophage, phiJL-1, from a cucumber fermentation. *Int. J. Food Microbiol.* **84**:225–235.
28. **Lu, Z., F. Breidt, V. Plengvidhya, and H. P. Fleming.** 2003. Bacteriophage ecology in commercial sauerkraut fermentations. *Appl. Environ. Microbiol.* **69**:3192–3202.
29. **Margulies, M., et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376–380.
30. **Marhaver, K. L., R. A. Edwards, and F. Rohwer.** 2008. Viral communities associated with healthy and bleaching corals. *Environ. Microbiol.* **10**:2277–2286.
31. **Meyer, F., et al.** 2008. The metagenomics RAST server: a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* **9**:386.
32. **Nam, Y. D., H. W. Chang, K. H. Kim, S. W. Roh, and J. W. Bae.** 2009. Metatranscriptome analysis of lactic acid bacteria during kimchi fermentation with genome-probing microarrays. *Int. J. Food Microbiol.* **130**:140–146.
33. **Parsley, L. C., et al.** 2010. Census of the viral metagenome within an activated sludge microbial assemblage. *Appl. Environ. Microbiol.* **76**:2673–2677.
34. **Plengvidhya, V., F. Breidt, Jr., Z. Lu, and H. P. Fleming.** 2007. DNA fingerprinting of lactic acid bacteria in sauerkraut fermentations. *Appl. Environ. Microbiol.* **73**:7697–7702.
35. **Rice, P., I. Longden, and A. Bleasby.** 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**:276–277.
36. **Rodriguez-Brito, B., et al.** 2010. Viral and microbial community dynamics in four aquatic environments. *ISME J.* **4**:739–751.
37. **Roh, S. W., et al.** 2010. Investigation of archaeal and bacterial diversity in fermented seafood using barcoded pyrosequencing. *ISME J.* **4**:1–16.
38. **Rohwer, F., D. Prangishvili, and D. Lindell.** 2009. Roles of viruses in the environment. *Environ. Microbiol.* **11**:2771–2774.
39. **Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren.** 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**:84–89.
40. **Rosario, K., C. Nilsson, Y. W. Lim, Y. Ruan, and M. Breitbart.** 2009. Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* **11**:2806–2820.
41. **Sambrook, J., and D. W. Russell.** 2001. *Molecular cloning: a laboratory manual.* Cold Spring Harbor Laboratory Press, New York, NY.
42. **Schoenfeld, T., et al.** 2009. Functional viral metagenomics and the next generation of molecular tools. *Trends Microbiol.* **18**:20–29.
43. **Schoenfeld, T., et al.** 2008. Assembly of viral metagenomes from Yellowstone hot springs. *Appl. Environ. Microbiol.* **74**:4164–4174.
44. **Scott, R., and W. C. Sullivan.** 2008. Ecology of fermented foods. *Hum. Ecol. Rev.* **15**:25–31.
45. **Seshadri, R., S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier.** 2007. CAMERA: a community resource for metagenomics. *PLoS Biol.* **5**:e75.
46. **Shapiro, O. H., A. Kushmaro, and A. Brenner.** 2010. Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *ISME J.* **4**:327–336.
47. **Suttle, C. A.** 2005. Viruses in the sea. *Nature* **437**:356–361.
48. **Venter, J. C., et al.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**:66–74.
49. **Viander, B., M. Maki, and A. Palva.** 2003. Impact of low salt concentration, salt quality on natural large-scale sauerkraut fermentation. *Food Microbiol.* **20**:391–395.
50. **Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole.** 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**:5261–5267.
51. **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.
52. **Wommack, K. E., J. Bhavsar, and J. Ravel.** 2008. Metagenomics: read length matters. *Appl. Environ. Microbiol.* **74**:1453–1463.
53. **Yeates, C., M. R. Gillings, A. D. Davison, N. Altavilla, and D. A. Veal.** 1998. Methods for microbial DNA extraction from soil for PCR amplification. *Biol. Proc. Online* **1**:40–47.
54. **Yoon, S. S., R. Barrangou-Pouey, F. Breidt, Jr., T. R. Klaenhammer, and H. P. Fleming.** 2002. Isolation and characterization of bacteriophages from fermenting sauerkraut. *Appl. Environ. Microbiol.* **68**:973–976.
55. **Zhang, T., et al.** 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* **4**:e3.