

Optimizing methods for virome analysis based on studies of a synthetic viral community

Jiayi Duan,¹ Andrew D. Marques,¹ Matthew Hogenauer,¹ Young Hwang,¹ Yanjia Zhang,² Aaron Timperman,^{2,3} Stephanie Higgins,^{4,5} Naomi G. Wilson,^{4,5} Elizabeth Aine Fitts,¹ Haeun Karissa Lim,¹ Kyle Bittinger,^{4,5} Ahmed M. Moustafa,^{4,5} Ronald G. Collman,^{1,6} Frederic D. Bushman¹

AUTHOR AFFILIATIONS See affiliation list on p. 15.

ABSTRACT Studies of whole viral populations—the “virome”—are yielding exciting new insights into biological systems, but methods are still being optimized. Here, we describe generation and use of a synthetic viral community and its use to evaluate technical challenges arising in virome analysis. We spiked the mock community into different human sample types, then passed the samples through different virus enrichment protocols and analyzed by Illumina sequencing. Compared with direct metagenomic sequencing, VLP enrichment protocols greatly increased viral read yields from stool and saliva. Four methods for DNA amplification were compared, with three showing over-amplification of small circular ssDNA viruses, most notably GenomiPhi. Studies of viral particle stability in the presence of nuclease showed that most viral genomes were stable when protected in viral particles, but phage MS2 RNA was unexpectedly labile under some of the conditions tested. Comparison of Illumina 1,000-cycle sequencing versus 300-cycle sequencing showed that longer reads supported generation of longer viral genome assemblies. We tested bacteriophage T4 DNA modified with glucosyl-hydroxymethylcytosine (ghmC) and hydroxymethylcytosine (hmC) and found that both were readily detected, though the recovery of ghmC-modified DNA was reduced compared with T4 genomes with unmodified cytosine. These studies together with published data help provide guidance for virome researchers optimizing analytical protocols.

IMPORTANCE A challenge in characterizing the human virome in health and disease is identifying optimal methods for enriching the viral content of samples. Due to the tremendous abundance and diversity of viruses, capturing as broad of a range of viruses as possible for analysis is difficult and potentially complicated by unrecognized biases. This report presents the use of a synthetic viral community for methods optimization in virome studies and illustrates the feasibility and challenges of current virus enrichment strategies for high-throughput virome analysis of different human sample types.

KEYWORDS virus, virome, bacteriophage, DNA sequencing, virus enrichment, method optimization

The planetary virome is vast. Virus-like particles (VLPs) on Earth are believed to number $\sim 10^{31}$ (1–4). Rich sea water can harbor $\sim 10^7$ VLPs per mL (5, 6). Human stool can contain $\sim 10^9$ VLPs per gram (7–10). Viruses can affect their hosts in numerous ways, both positively and negatively (9, 11–24). Thus it is often of interest to characterize whole viral populations in a biome to assess the influence of these populations on function.

Due to the vast diversity of viruses, only a small fraction have been characterized, thus many reads from virome sequencing lack close matches in existing databases. The rapid evolution of viral genomes and emergence of variants also make it difficult

Editor Sepideh Pakpour, The University of British Columbia Okanagan, Kelowna, BC, Canada

Address correspondence to Ronald G. Collman, collmanr@penmedicine.upenn.edu, or Frederic D. Bushman, bushman@penmedicine.upenn.edu.

The authors declare no conflict of interest.

See the funding table on p. 15.

Received 23 February 2026

Accepted 27 April 2026

Published 2 June 2026

Copyright © 2026 Duan et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

to know whether sequences in “metagenomic dark matter” are viral in origin (25–27). Moreover, viral sequences often represent only a tiny fraction of nucleic acids in a biological sample, so the overwhelming majority of non-viral nucleic acids complicates the capture and identification of viral sequences. Over 90% of human gut bacterial genomes carry integrated prophages, making it challenging to distinguish free viral particles from integrated sequences (28). Thus, it is often of interest to investigate the virome content in samples after enriching for VLPs before analysis by sequencing. In favorable cases, this will allow tentative identification of previously unstudied sequences as potentially virus-like. Many useful protocols for VLP enrichment have been reported (8, 9, 11, 12, 18, 29–40), but no single method is ideal for all applications.

Greatly complicating virome analysis is the heterogeneity of viral particles. Viral genomes can be comprised of RNA or DNA. Genomes can be single-stranded or double-stranded, linear or circular, and segmented or continuous. Viral particles can be of many morphologies, and enclosed in one or more lipid membranes. Viral particles can range in size from 1.5 μm in length (pandoravirus) to 25 nm (Adeno-associated virus). Genome sizes range from 2.8 Mb for Pandoravirus to 2–3.9 Kb for Anelloviruses (4, 41, 42). Genomes of satellite viruses and virus-like RNAs can be even smaller, comprised of structured circular molecules as little as 200–400 nt in length (27, 43, 44).

Here, we describe steps toward optimizing methods for VLP enrichment based on assays with a synthetic viral community comprised of known viruses. Several previous publications have also investigated virome methods with spike-ins of known viruses (18, 31, 33–36, 39, 45–48), which complement the work described here. As new virome research programs are initiated, interest turns to capturing viruses that might have previously been lost during purification, motivating further rounds of method optimization. We included eight viruses in our mock community, named VirMock1, spanning a wide range of viral types (Table 1). Several questions pertinent to optimizing methods were addressed. Multiple methods for VLP enrichment have been reported (e.g., 8, 18, 29, 32, 34, 37, 38, 40, 45–47, 49)—how do these perform versus simple metagenomic DNA sequencing or RNA-seq on the same samples? When only small amounts of enriched viral genomes are available, amplification methods are often used—how much do these distort the relative abundances of viral genomes in a sample? How does the choice of sequencing method affect the data recovered? VLP fractions are often treated with nucleases (50) to remove host cell nucleic acids—how frequently do these treatments also destroy viral genomes? Bacteriophage DNA has been reported to be modified by at least 10 forms of covalent modification (51–53)—how often do these disrupt recovery in typical virome protocols? A complication is that many factors can influence the answers to these questions. Here, we start with several published protocols and report surveys using our synthetic viral community to address some of the parameters influencing the output virome data produced. Examples of optimized protocols are presented in Document S1.

RESULTS

Comparison of methods for recovering VLP contigs

We initially compared viral sequence recovery after VLP enrichment versus simple metagenomic DNA or RNA sequencing of the unfractionated samples. For this, we first monitored just the viruses naturally occurring in the stool community analyzed. Three methods were compared for VLP enrichment from stool samples (Table 2), using methods derived from published protocols and used in multiple studies, here termed VP1 (46), VP2 (47), and VP3 (45). All methods involved homogenization of stool in SM buffer, centrifugation to remove debris, filtration, nuclease treatment, and nucleic acid extraction. The methods differ in the details, including amount of starting material used, extent of initial dilution, filtration method, whether samples were concentrated after filtration, nature of the nuclease treatment, whether proteinase K treatment and phenol extraction were performed, and the nucleic acid purification kit chosen. Extracted RNA was subjected to cDNA synthesis and DNA and cDNA pooled and sequenced

TABLE 1 Viruses included in the synthetic viral community VirMock1

Virus	Enveloped?	RNA or DNA?	Single-stranded (ss) or double-stranded (ds)?	Morphology	Genome size (bp)	Host
Enterobacteria phage T4	No	DNA	ds	Tailed phage	168,903	<i>Escherichia coli</i> DH10B
Enterobacteria phage lambda	No	DNA	ds	Tailed phage	48,502	<i>Escherichia coli</i> DH10B
Adeno-associated virus vector (AAV)	No	DNA	ss	Icosahedral	3,300	Synthetic vector grown in human cell line
Vaccinia virus WR (VV)	Yes	DNA	ds	Brick-shaped	194,711	BSC-1, African green monkey kidney cells
Pseudomonas phage phi6	Yes	RNA	ds	Spherical	13,385 (all segments total)	<i>Pseudomonas syringae</i> LM2691
Phage MS2	No	RNA	ss	Icosahedral	3,569	<i>Escherichia coli</i> C3000
Murine hepatitis virus strain A59 (MHV)	Yes	RNA	ss	Spherical	31,335	17Cl-1, murine fibroblast cell line
Enterobacteria phage M13	No	DNA	ss	Filamentous	6,407	<i>Escherichia coli</i> C3000

together. All three methods were tested on a common sample of homogenized stool (Fig. 1A through D). For comparison, we also analyzed viral sequence recovery from the same stool samples in pipelines based on conventional metagenomic DNA sequencing and total RNA-seq. Output sequencing reads were analyzed using Sunbeam (54), which includes Cenote-Taker2 for viral content annotation (55), then compared over the percentage of reads mapping to the viral contigs generated from the data.

Results with VP1–VP3 showed recovery of viral contig-aligning reads (DNA and cDNA combined) reaching an average of 62.9 (± 2.98)% to 73.0 (± 3.60)% of total reads, while metagenomic DNA- and RNA-seq returned only an average of 2.37 (± 0.04)% and 0.128 (± 0.011)%, respectively (Fig. 1A through D). The majority of recovered contigs annotated as Caudoviricetes, tailed dsDNA bacteriophages, in both virome sequencing and metagenomic sequencing (Fig. S1B and G through I). Metagenomic sequencing was less efficient at recovering viral genomes, due in part to dilution with DNA from other sources. Thus, we conclude that VP1–VP3 are all effective at enriching for virus-like particles and increasing recovery of virus-like contigs from stool.

We next carried out a similar comparison for saliva and BAL (Fig. 1E through H). The two represent sample types with lower biomasses than stool, though saliva is still rich in viruses. BAL of healthy individuals, in contrast, is quite sparse for viruses, cells and biological materials generally, thus yielding a low amount of DNA for sequencing (49, 56–59). For analysis of these liquid samples, an enrichment protocol named VP4 was devised, paralleling VP1 but with reduced sample dilution and use of filters with larger pore sizes, which was added in an attempt to accommodate the higher viscosity of some airway samples.

Saliva undergoing VP4 had on average 34.8 (± 2.86)% of reads annotated as viral (Fig. 1H). Saliva analyzed using metagenomic DNAseq and RNAseq without enrichment yielded only 2.10 (± 0.26)% and 0.179 (± 0.026)% viral reads, respectively (Fig. 1F). For BAL, VP4 resulted in no recovery of viral contigs (Fig. 1G); repeating the bioinformatic analysis using geNomad (60) for viral annotation returned similar result (Fig. S2). Processing of samples with low biomass through the VLP enrichment protocols, including BAL, OP wash and nasopharyngeal swab eluates, commonly resulted in sample loss. This indicates that other analytical approaches may be needed for low-biomass samples (discussed below) (49, 56–59).

To ask how these methods perform with viruses of widely differing particle compositions, we next turned to tests of known viruses spiked into human-derived samples.

Assembling the VirMock1 synthetic viral community

Viral particles vary extensively in genome sizes, structures, and nucleic acid types. As a tool for methods optimization, we thus assembled a community with a wide range

TABLE 2 Steps in each of the three virus enrichment protocols focused on analysis of stool samples

Steps	VP1	VP2	VP3
Homogenization	200 mg stool resuspended in 10 mL SM buffer	50 mg stool resuspended in 500 μ L SM buffer	500 mg stool resuspended in 10 mL SM buffer
Centrifugation	4,000 $\times g$ for 10 min	17,000 $\times g$ for 3 min	5,000 $\times g$ for 10 min (twice)
Filtration	0.22- μ m Steriflip vacuum filter	40- μ m pore-size strainer; then 0.8- μ m PES centrifugal filter, 17,000 $\times g$ for 1 min	0.45- μ m PES syringe filter (twice); 0.5M NaCl and 10% PEG-8000. Incubate overnight (16 h) at 4°C; 5,000 $\times g$ for 20 min to collect pellet; Pellet suspended in 400 μ L SM buffer + 400 μ L chloroform to extract; 2,500 $\times g$ for 5 min
Reconcentration	Amicon Ultra-15 centrifugal filter, 4,000 $\times g$ for 10min intervals till ~ 250 μ L left; Add 10 mL SM buffer and repeat the previous centrifugation step		
Nuclease treatment	10 $\times$ DNase buffer + 2 μ L DNase I + 1 μ L RNase I (Roche) for every 100 μ L of post-viomeprep sample, 37°C incubation for 30 min Add 1 μ L of 0.5 M EDTA and incubate at 75°C for 10 min		Aspirate the aqueous phase and mix it with 10 mM CaCl ₂ , 50 mM MgCl ₂ , 8 U TURBO DNase I and 20 U RNase I, 37°C incubation for 1 h; Incubate at 70°C for 10 min
Proteinase K treatment			Proteinase K and 10% SDS for 20 min at 56°C
Nucleic acid extraction	QIAamp Viral RNA mini kit		100 μ L phage lysis buffer. Incubate at 65°C for 10 min; Phenol/chloroform/isoamyl alcohol extraction (twice); Centrifuge at 8,000 $\times g$ for 5 min; DNeasy blood and tissue purification kit

of virus types (VirMock1), where each virus was well characterized and straightforward to grow (Table 1). We included viruses of multiple morphologies with RNA genomes (ϕ i6, MS2, and MHV) and DNA genomes (AAV, lambda, M13, T4, and VV), including both single-stranded and double-stranded representatives.

To analyze an example of unusual covalent DNA modification, we included phage T4 and two mutant derivatives (Table S1) and assessed the effects of the DNA modification of each strain on virome recovery (see below). In our standard VirMock1 community, we used wild-type T4 only.

We acknowledge that different viruses behave differently, and the applicability of the study result is limited to the viruses in VirMock1. In future studies, it will be beneficial to incorporate more varieties of viruses, such as giant viruses, to provide more comprehensive insights into virome analysis.

Use of VirMock1 to test protocols for virus recovery from stool

To test virome recovery using VP1–VP3, we spiked the VirMock1 community into a common stool sample or into SM buffer. Purified virome nucleic acids were analyzed using Illumina short-read sequencing. Sequence reads were (i) directly mapped to genomes of VirMock1 community members to quantify VirMock1 recovery and (ii) used to generate contigs, followed by contig annotation, and then mapping of reads back onto viral contigs to quantify all viruses in the samples, not limited to VirMock1. Each sample type was analyzed in triplicate (Fig. 2). Fig. 2A through G quantifies the relative proportions of reads of VirMock1 and total viruses annotated; Fig. 2H through N shows the relative proportions of the VirMock1 community members recovered. As a check, unspiked biological samples were analyzed for representation of the VirMock1 viruses, and no significant matches were detected (Table S2).

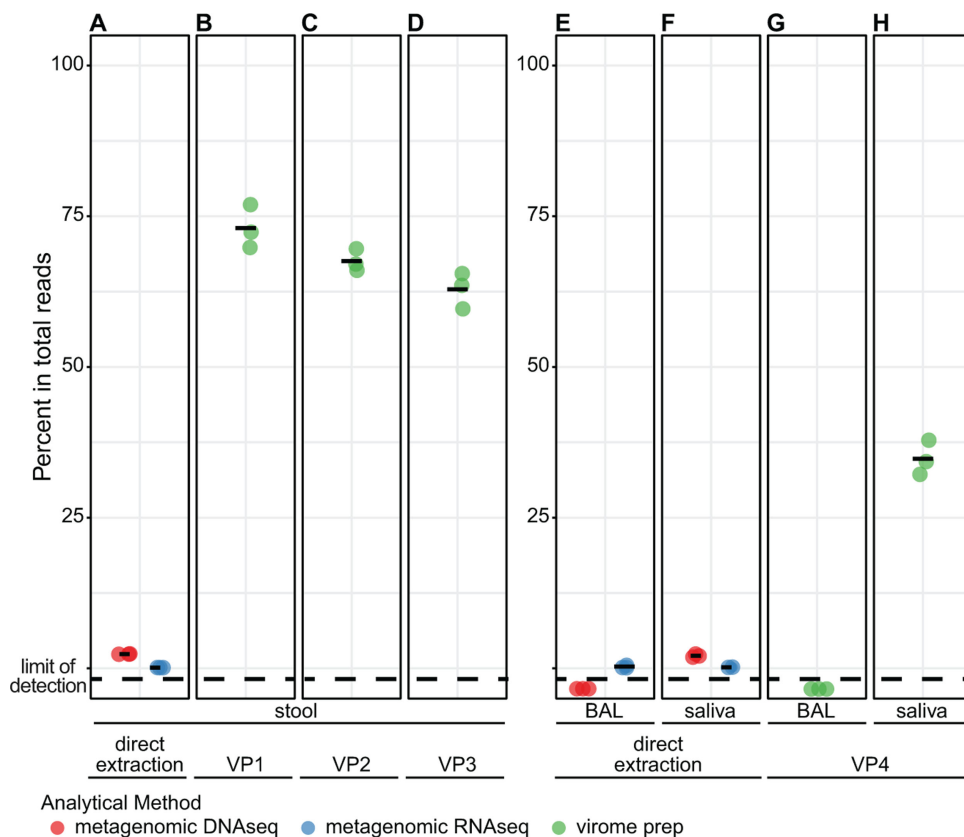


FIG 1 Comparison of viral sequence recovery after virus particle enrichment followed by sequencing, versus direct metagenomic sequencing of total DNA or RNA. (A–D) Percentages of reads mapped to viruses annotated by Cenote-Taker2 in stool that was analyzed by (A) direct extraction of DNA (red) or RNA (blue), or after viral particle enrichment using the (B) VP1, (C) VP2, or (D) VP3 protocols (green), with the mean of triplicate measurements represented by a black bar. (E–H) Percentages of reads mapped to viruses annotated by Cenote-Taker2 in (E and G) BAL or (F and H) saliva that was analyzed by (E and F) direct extraction and metagenomic DNAseq (red) and RNAseq (blue), or after (G and H) viral particle enrichment using the VP4 protocol (green), with the mean of triplicate measurements represented by a black bar. Dots below the limit of detection represent samples, with no viral contigs detected.

As a reference, the VirMock1 community was mixed with SM buffer, treated with nuclease (DNase I and bovine pancreatic RNase, Methods S1), and directly extracted (Fig. 2A and H). After direct extraction, an average of 13.1 (± 0.976)% of reads aligned to the VirMock1 genome sequences (Fig. 2A, green dots). Assembly of all viral contigs and mapping of reads back to contigs yielded a similar result (Fig. 2A, tan dots).

A similar analysis was performed after carrying out the VP1-VP3 purification protocols on VirMock1 spiked into a common stool sample (Fig. 2E through G), or spiked into SM buffer (Fig. 2B through D). After spiking into the stool sample, for VP1, viral contigs aligned to an average of 72.8 (± 4.41)% of reads (Fig. 2E, tan dots); numbers were slightly lower for VP2 (64.5% ± 8.00 %) (Fig. 2F, tan dots) and VP3 (63.1% ± 3.06 %) (Fig. 2G, tan dots). The proportions of the VirMock1 sequences in these samples were much lower, in the range of 0.1%–0.8% of all reads, but this reflects dilution by the large numbers of viral sequences authentically present in the stool samples and recovered after viral enrichment.

In contrast, the recovery of VirMock1 sequences was more variable after spiking into SM buffer only (Fig. 2B through D). VP2, the method involving the least dilution of the sample, showed the best recovery, while heavy losses of VirMock1 were seen with VP1 and VP3. This comparison is consistent with the idea that dilution can impair recovery,

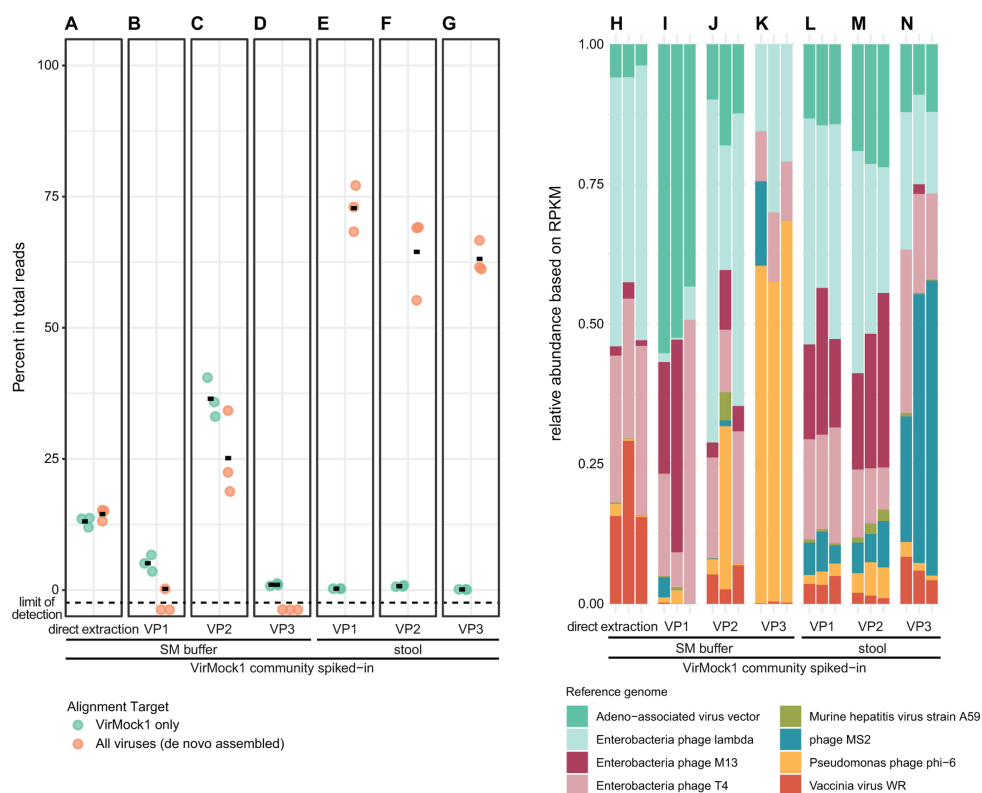


FIG 2 Comparison of methods for purifying viral particles from stool. (A–G) Percentage of reads mapping to viruses from VirMock1 or any *de novo* assembled viral contigs after VirMock1 was spiked into (A–D) SM buffer or (E–G) stool. Samples were (A) nuclease treated and then directly extracted for nucleic acid, or first enriched for viral particles using (B and E) VP1, (C and F) VP2, or (D and G) VP3. Reads aligning to any virus as annotated by Cenote-Taker2 are indicated in tan; reads annotating as aligning to a virus in VirMock1 are shown in green; the mean of triplicate measurements is represented by the black bar. Dots below the limit of detection represent samples with no viral contigs detected. (H–N) Relative abundance of each VirMock1 reference virus in samples where VirMock1 was spiked into (H–K) SM buffer or (L–N) stool. Samples were (H) nuclease treated and then directly extracted for nucleic acid, or first enriched for viral particles using (I and L) VP1, (J and M) VP2, or (K and N) VP3.

and that biological materials such as stool may improve recovery, potentially via “carrier” effects.

Fig. 2H through N shows the VirMock1 lineages detected in each sample as read out from the Illumina sequencing data. Direct extraction of VirMock1 treated only with nuclease before purification showed robust representation of six of the VirMock1 viruses, but two RNA viruses, phage MS2 and murine hepatitis virus, were present in only trace amounts. All eight VirMock1 viruses were detected after spiking into stool, again suggesting a possible “carrier” effect of the biological sample, in which the presence of stool components may stabilize labile viruses and facilitate their recovery. Potential carrier effects were also seen with saliva, where recovery was better than for BAL, oropharyngeal (OP) wash or SM buffer (Fig. S3A through E). Principal coordinate analysis (PCoA) indicated that experimental treatments affected the viruses recovered: sequencing results clustered more closely when spiked into stool than spiked into SM buffer, again consistent with the potential carrier effect of stool components (Fig. S4).

Fig. S5 shows the annotation of all viral contigs—mostly stool derived—in each specimen. The samples of stool with VirMock1 spiked in were dominated by Caudoviricetes (Fig. S5A through C), an abundant lineage of tailed bacteriophages with double-stranded DNA genomes, and Malgrandaviricetes, a group containing common single-stranded DNA bacteriophages, both widely reported in human stool (13, 24, 37, 61–67).

The nature of background contamination in VirMock1 could also be assessed in the sequence data. Analyzing VirMock1-spiked SM buffer with Kraken2 showed that the majority of the reads not assigned to VirMock1 viruses were from bacterial and mammalian host cells used to propagate the viruses (Fig. S6A and H). Identification and removal of background sequences are important steps in virome analysis (68–71).

Distortions of recovered communities associated with different DNA amplification and library preparation methods

We next tested how different DNA amplification methods affected virome recovery. Since stool is rich in viruses and may not require amplification, we focused instead on liquid samples undergoing VP4. We assessed VirMock1 (i) spiked into saliva and purified via VP4, (ii) spiked into SM buffer and purified via VP4, or (iii) spiked into SM buffer, nuclease-treated, and directly extracted. Samples were then reverse-transcribed and amplified with one of four commercially available amplification kits: GenomiPhiV3 (72), primary template-directed amplification (PTA) (73), whole transcriptome amplification (WTA2) (74), or multiple annealing and looping-based amplification cycles (MALBACs) (75) prior to Illumina DNA sequencing. These include multiple displacement amplification and quasi-linear amplification, with detailed features of each method specified in Table S3. We also included analysis of VirMock1 with no amplification. Samples were analyzed in triplicate.

In the VirMock1-spiked SM buffer samples treated with nuclease and directly extracted without amplification (Fig. 3C), the sequencing result generally reflected the input VirMock1 composition (Table S4). Samples showed good recovery of six of the VirMock1 viruses, and trace amount of phi6 and MS2. Roughly similar patterns were seen for the VP4-purified unamplified samples (Fig. 3A through C), except in these samples MHV and AAV showed increased representation after purification.

Different amplification methods resulted in different biases for the VirMock1 community members. GenomiPhi greatly favored M13 (Fig. 3D through F) probably reflecting rolling circle replication of the small circular ssDNA genome, as has been reported in several previous studies (33, 76, 77). The other three methods showed less extreme distortions of the viral populations (Fig. 3G through O). PTA did not show such overamplification of M13 but resulted in considerable amplification of vaccinia virus (Fig. 3G through I). A repeat experiment with PTA showed less positive bias for vaccinia, while M13 was over-amplified even more dramatically (Fig. S7E through H). WTA2 and MALBAC also showed lower and variable levels of overamplification of M13 in some conditions (Fig. 3J through O). Amplification using MALBAC showed overall similar relative abundance of VirMock1 viruses compared with unamplified, except for a moderate bias favoring MHV and M13 (Fig. 3M through O).

Fig. 3P summarizes these data by PCoA based on Bray-Curtis distances between the measured VirMock1 compositions of samples. Samples clustered by amplification method (Fig. 3P; PERMANOVA, Bray-Curtis, $R^2 = 0.719$, $P = 0.001$). Genomiphi was the most extreme outlier, with the other methods closer to unamplified but clustering separately. Thus we conclude that the GenomiPhi method is too biased for routine use, and the other methods recovered most viruses effectively, though each has its own modest biases; it will be useful to characterize further amplification methods over additional sample types in future work.

We also compared methods for second-strand cDNA synthesis after reverse transcription (RT; Fig. S8). Viral genome recovery was actually efficient without implementing any specific step for second strand synthesis after RT using our custom protocol (Fig. S8A and H), likely due to the spontaneous second-strand synthesis initiated by RT under the conditions used. However, addition of Klenow polymerase in a subsequent step appeared to diminish the variability in viral proportions (Fig. S8O; linear regression, $P < 2 \times 10^{-16}$). Thus, in our favored protocol (Methods S1), we recommend a second-strand synthesis step.

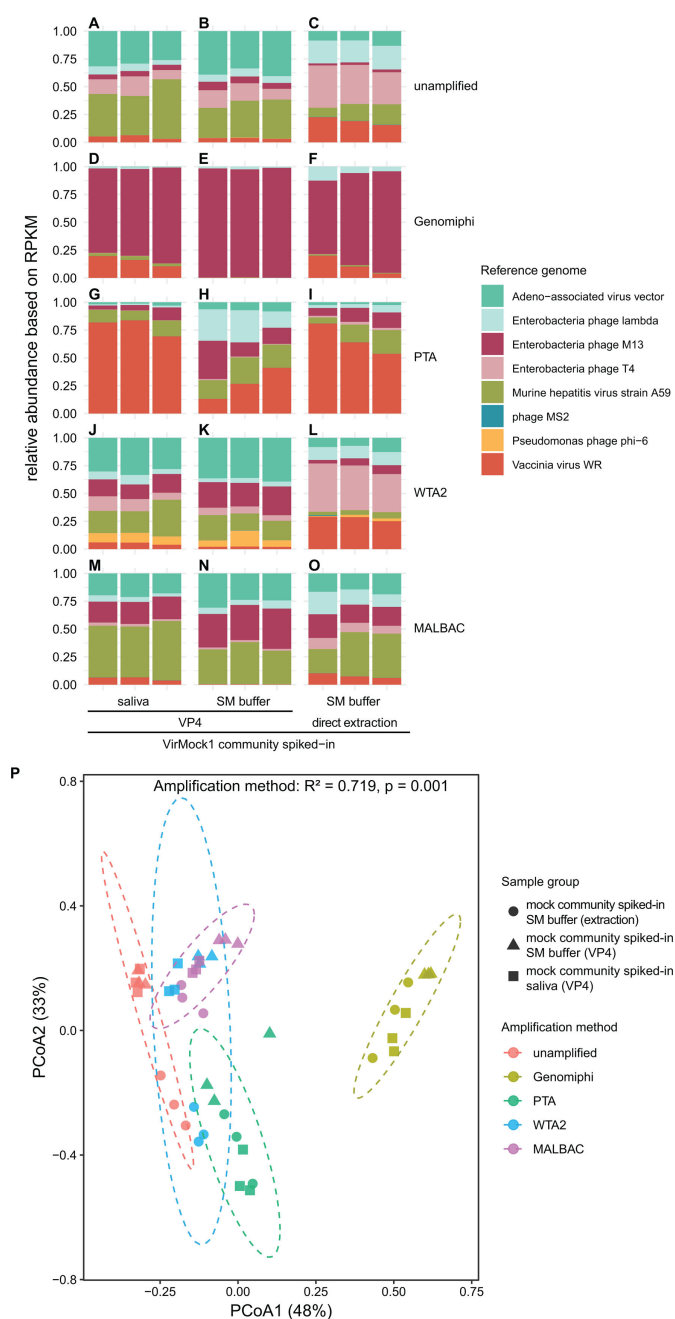


FIG 3 Comparison of the effects of different DNA amplification methods on viral genome recovery. (A–O) Relative abundance of each VirMock1 reference virus. Samples were compared after addition of VirMock1 to (A, D, G, J, and M) saliva or (B, E, H, K, and N) SM buffer and purified using a modified protocol with elements of VP1 and VP2 optimized for saliva (VP4), or (C, F, I, L, O) after addition of VirMock1 into SM buffer, nuclease treated and extracted immediately afterwards. Samples were analyzed by Illumina sequencing without amplification, or after amplification using (D–F) Genomiphi, (G–I) PTA, (J–L) WTA2 with 17 PCR cycles, or (M–O) MALBAC with 17 PCR cycles. (P) PCoA was performed using Bray-Curtis distances computed based on relative abundance of VirMock1 reference virus in each sample. Variation in community composition across different amplification methods were assessed using permutational multivariate analysis of variance (PERMANOVA) on the Bray-Curtis distances ($R^2 = 0.719$, $P = 0.001$). Ellipses showing the 95% CI around the group mean were drawn for sample groups with more than three samples, assuming a multivariate t-distribution. Amplification methods (represented by different colors) and samples (represented by different shapes) were as in panels A–O.

Virus types differ in their sensitivity to nuclease treatments

VLP enrichment methods typically involve a nuclease step following VLP enrichment and before nucleic acid purification to remove free nucleic acids (8, 16, 18, 19, 29, 31–34, 36–39, 48, 61, 62, 76, 78–80). However, such a step has the danger of degrading nucleic acid within particles if there is any degree of accessibility of encapsulated viral genomes (8, 78, 79, 81). We thus incubated VirMock1 in SM buffer with a titrated mixture of DNase and RNase over a range of concentrations, and quantified viral genome recovery. Quantification was carried out using Illumina DNA sequencing (Fig. 4A) or qPCR (Fig. 4B). We also compared results after spiking VirMock1 into stool (Fig. 4C). Recovery of each virus in VirMock1 was quantified relative to the measured input as represented by the untreated condition.

Recovery of viruses varied among types. Most viruses were insensitive to nuclease up to 10 times (10×) the concentration used in typical VLP preparations (Fig. 4A and B). An exception was bacteriophage MS2, which was sharply reduced in abundance by both measures at even 0.1× nuclease concentration (Fig. 4A and B). The decreased MS2 was associated with a proportional increase of other viruses in the sequencing data (Fig. 4A). MS2 was also lost when spiked into stool in two out of three replicates even without nuclease treatment, perhaps due to handling or pre-existing nuclease activity in stool (Fig. 4C). In other cases where MS2 was spiked into stool (Fig. 2 and 5 below), MS2 genomes were detected, but in reduced proportion relative to other VirMock1 viruses. Other viruses showed less susceptibility to nuclease; phi6 showed some decrease in relative abundance when spiked into stool (Fig. 4C). Control experiments showed that free nucleic acids were reduced in abundance efficiently by 1× nuclease treatment when passed through VP4 (Table S5). Thus, we find that most viruses withstood the nuclease treatment robustly, but one non-enveloped linear RNA virus (MS2) was quite sensitive.

Comparing Illumina short-read versus longer read sequencing

To begin to examine how the choice of sequencing method affects the output contig length, we compared the new Illumina 1,000-cycle kit and the standard 300-cycle kit. For this, we used the same sequencing libraries generated from a common stool sample spiked with VirMock1 and purified using VP1 (Fig. 2L and 5). Samples were sequenced and analyzed using our contig assembly and annotation pipeline (Methods S1).

The proportions of virus types in the stool samples were similar between the 1,000-cycle kit and 300-cycle kits for both the VirMock1-spiked (Fig. 5A, B, C, and E) and unspiked samples (Fig. 5D and F). Mock community members were detected after spike-in of VirMock1 (Fig. 5A and B), and an unbiased virus annotation emphasized Caudoviricetes and Malgrandaviricetes, as expected for stool samples (Fig. 5C through F).

We next compared the distribution of contig lengths. For comparison, we normalized the length of each contig by the total number of bases sequenced in that sample. The 1,000-cycle kit resulted in more total contigs and viral contigs of longer length than the 300-cycle kit for both the VirMock1-spiked and unspiked stool samples (Fig. 5G and H). The median length of contigs assembled from spiked and unspiked stool was 576 and 587 bp, respectively, using the 1,000-cycle kit, and 449 and 450bp, respectively, using the 300-cycle kit (Table S6). Thus, use of the longer length Illumina sequencing platform resulted in longer length contigs (Wilcoxon rank-sum, $P < 0.001$). As expected, different sequencing depths also resulted in different length contigs (linear regression, $R^2 = 0.338$, $P = 1.8\text{e-}16$; Table S7; Fig. S9).

Beginning to assess the possible influence of covalent DNA modification

Bacteriophage DNA has been reported to be subject to at least ten forms of DNA modification (51–53). RNA broadly has been reported to be subject to >100 forms of covalent modification (82). Both have the potential to disrupt recovery of viral genomes.

As one step toward assessing the effects these modifications, we compared three genotypes of bacteriophage T4: wild-type T4 (termed T4ghmC), T4hmC, and T4C. For

wild-type T4, >50% of C residues are modified to glucosyl-hydroxymethylcytosine, which can block attack by host cell nucleases (53, 83, 84). T4hmC is a mutant strain that

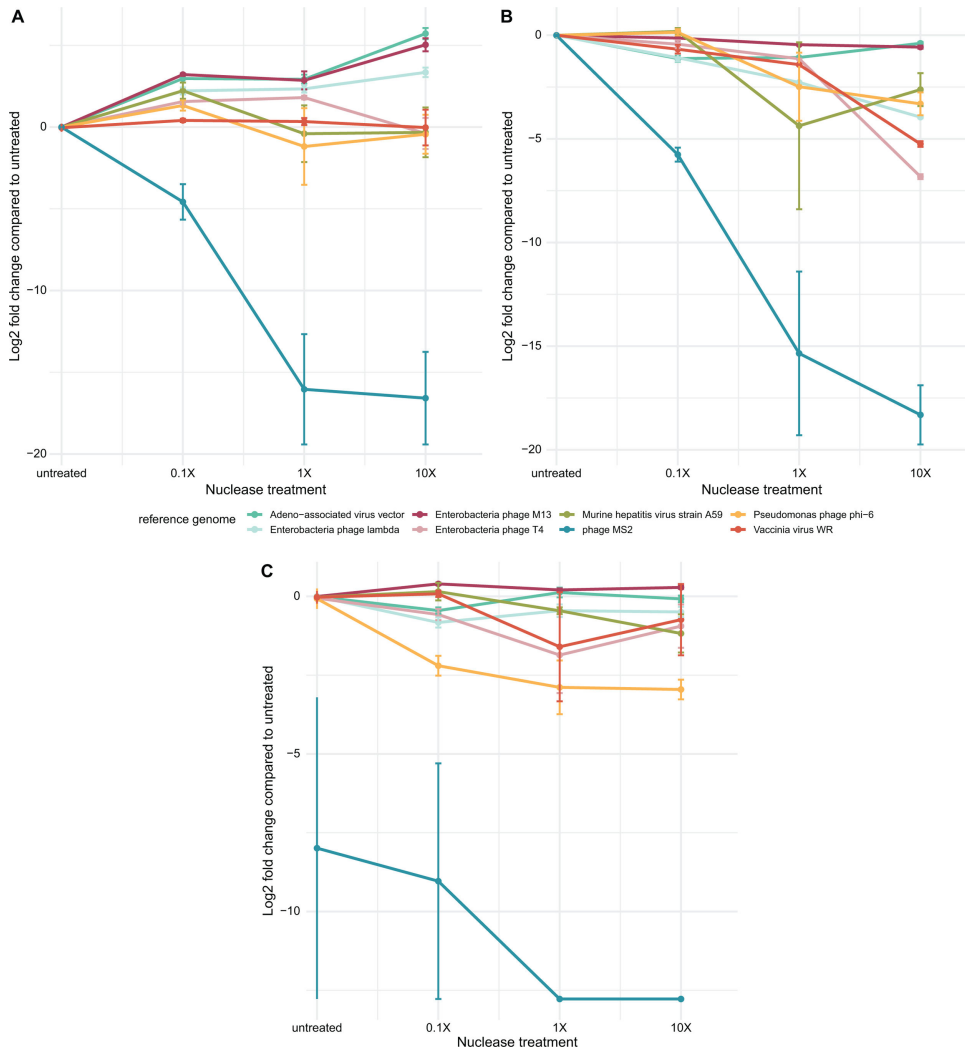


FIG 4 Assessing the stability of viral genomes within viral particles in the presence of nucleases. Nuclease treatment titrated from no added nuclease to 10 times of 1× nuclease was performed in the presence of VirMock1. 1× nuclease included 10× DNase buffer, 2 μL of Roche DNase I, and 1 μL of bovine pancreatic RNase for every 100 μL of post-virome prep sample (Methods S1). Dots represent the mean log₂ fold changes of relative abundance or genome copy number of a reference virus when compared with the mean relative abundance or genome copy number of that virus in the untreated control triplicates, with lines connecting the mean log₂ fold changes across nuclease concentrations and error bars representing the standard errors of the mean log₂ fold change. Spearman's rank correlations were computed between the log₂ fold changes of relative abundance and nuclease concentration to summarize the overall trend with increasing nuclease treatment. (A) Log₂ fold change of relative abundance of each virus (calculated based on RPKM) after VirMock1 underwent different nuclease treatments, followed by Illumina sequencing. Spearman's rank correlation, AAV: $\rho = 0.907$, $P = 4.7 \times 10^{-5}$; M13: $\rho = 0.842$, $P = 5.9 \times 10^{-4}$; lambda: $\rho = 0.928$, $P = 1.3 \times 10^{-5}$; VV: $\rho = 0.367$, $P = 0.24$; MHV: $\rho = -0.108$, $P = 0.74$; phi6: $\rho = -0.0864$, $P = 0.79$; T4: $\rho = 0$, $P = 1$; MS2: $\rho = -0.901$, $P = 6.3 \times 10^{-5}$. (B) Log₂ fold change of genome copy number of each virus after VirMock1 underwent different nuclease treatments, followed by qPCR. Spearman's rank correlation, AAV: $\rho = -0.324$, $P = 0.3$; M13: $\rho = -0.885$, $P = 1.3 \times 10^{-4}$; lambda: $\rho = -0.972$, $P = 1.4 \times 10^{-7}$; VV: $\rho = -0.972$, $P = 1.4 \times 10^{-7}$; MHV: $\rho = -0.777$, $P = 2.9 \times 10^{-3}$; phi6: $\rho = -0.756$, $P = 4.5 \times 10^{-3}$; T4: $\rho = -0.95$, $P = 2.3 \times 10^{-6}$; MS2: $\rho = -0.907$, $P = 4.7 \times 10^{-5}$. (C) Log₂ fold change of relative abundance of each virus (calculated based on RPKM) after VirMock1 was spiked into stool and underwent VP1 with different nuclease treatments, followed by Illumina sequencing. Spearman's rank correlation, AAV: $\rho = 0.043$, $P = 0.89$; M13: $\rho = 0.389$, $P = 0.21$; lambda: $\rho = -0.345$, $P = 0.27$; VV: $\rho = 0.173$, $P = 0.59$; MHV: $\rho = -0.734$, $P = 6.6 \times 10^{-3}$; phi6: $\rho = -0.734$, $P = 6.6 \times 10^{-3}$; T4: $\rho = -0.367$, $P = 0.24$; MS2: $\rho = -0.415$, $P = 0.18$.

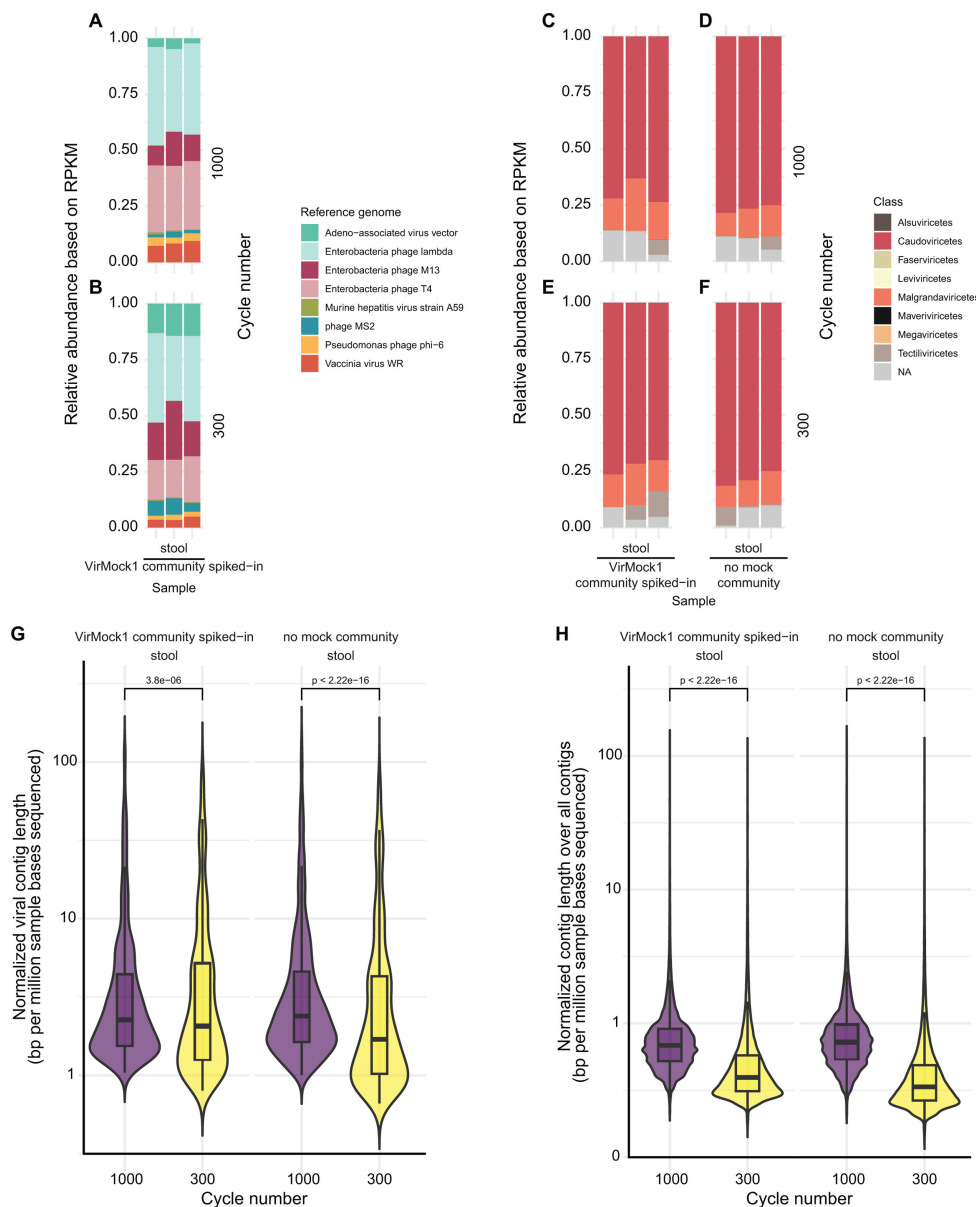


FIG 5 Comparison of Illumina sequencing with the Illumina 1,000-cycle kit and 300-cycle kit platforms. Stool was unspiked or spiked with VirMock1 community, underwent VP1 and Nextera XT library preparation, and sequenced using either the 300-cycle kit or 1,000-cycle kit. (A and B) Relative abundance of each reference virus in VirMock1-spiked stool when sequenced using (A) the 1,000-cycle kit or (B) the 300-cycle kit. (C–F) Relative abundance of classes of viruses as annotated by Cenote-Taker2 in (C) VirMock1-spiked stool when sequenced using the 1000-cycle kit or (E) the 300-cycle kit, and (D) unspiked stool when sequenced using the 1,000-cycle kit or (F) the 300-cycle kit. (G) Distribution of the normalized length of the viral contigs output by the 1,000-cycle kit and 300-cycle kit in VirMock1 community-spiked and unspiked stool. (H) Distribution of the normalized length of all contigs (not only viral) output by the 1,000-cycle kit and 300-cycle kit in VirMock1 community-spiked and unspiked stool. Contig lengths were normalized to total sample sequencing depth (as expressed by base pairs per million sample bases sequenced) and plotted on a log₁₀ scale. Violin plots show the full distribution of normalized contig lengths, with embedded boxplots indicating the median and interquartile range. Pairwise differences between the kits were evaluated using Wilcoxon rank-sum tests, with formatted *P*-values shown above each comparison. Purple: 1,000-cycle kit. Yellow: 300-cycle kit.

produces only hydroxymethylcytosine. T4C harbors mutations in genes required for substituting HMC for dCTP and thus only contains unmodified cytosines in genomes. The

effects of these modifications were previously documented in DNA sequencing studies as effects on inter-pulse distances in Pacific Biosciences DNA sequencing (53).

To test the effects of these modifications in our Illumina sequencing protocol, we first verified the presence of each DNA modification in the three T4 strains using LC-MS/MS (Fig. S10). We also purified each DNA and showed that sensitivity to restriction enzyme digestion was as expected—that is, DNA containing glucosyl-hydroxymethylcytosine and hydroxymethylcytosine were protected from digestion by Alu I, while DNA containing cytosine was sensitive (Fig. S11). These studies documented the predominance of the expected forms of cytosine in each T4 strain.

We then asked whether these modifications disrupted genome sequencing. We mixed each T4 strain with bacteriophage lambda, which lacks the genome modifications mentioned above, and compared recovery after extraction, library preparation and DNA sequencing (Fig. 6). When mixed with lambda, the relative abundance of T4ghmC assayed by sequencing was on average 2.6-fold lower than that expected from the qPCR values for total genomes (Fig. 6; Table S8). For T4C and T4hmC, there was no major reduction in relative abundance by sequencing compared with quantification by qPCR (Fig. 6; Table S8). Thus, hydroxymethylation does not detectably inhibit detection by sequencing, but glucosyl-hydroxymethylation did appear to diminish detection modestly (Kruskal-Wallis, $P = 0.027$). Similar conclusions were reached by Zhai et al. (39). Given that the Nextera method was used to prepare the sequencing library, we conjecture that the transposon-based step used to attach amplification primers may be partially inhibited by glucosyl-hydroxymethylation, because after subsequent amplification steps, the DNA will be unmodified. Thus, the modified genomes were detectable in our samples, though with different efficiencies. It will be useful to document effects of further forms of DNA and RNA modification on virome assays.

DISCUSSION

Here, we describe steps toward optimizing virome purification methods, in part by taking advantage of tests with a mock community comprised of eight diverse viruses (VirMock1). Previously, multiple groups have published studies of optimizing methods for virome analysis (8, 18, 29–34, 45–47). All of these are valuable, but in any biochemical manipulation, there are typically multiple alternatives at each step, so that the parameter space is large. Here, we start with well-established protocols and work to fill in more of this space with control experiments. In addition, goals in virome studies have been changing over time. Early workers were generally content with recovering large numbers of viruses reproducibly and were less concerned with capturing every possible virus in a community. As the field has matured, interest has grown in the “unknown unknowns,” and method optimization has broadened to include difficult-to-recover viral types that might once have been passed over.

Some of our main conclusions include the following. Multiple published protocols can enrich virus particles from stool and saliva; the usefulness of enrichment steps for low-biomass samples, such as BAL and OP wash, is uncertain. Conventional metagenomic DNA sequencing recovered dsDNA viral genomes, and RNA-seq recovered RNA viral genomes, as expected, but the viral-enrichment protocols greatly increased the fraction of reads attributable to contigs annotated as viral. Amplification methods introduce distortions into the data and should be used with caution. Most viral particles can withstand nuclease treatment, but MS2 was sensitive in some settings; one conjecture is that the structure of the MS2 particle, which includes a portal protein embedded in the icosahedral protein shell, somehow exposes the viral genome to degradation, though other explanations are possible. Longer read Illumina sequencing allowed recovery of longer viral contigs. DNA modifications found in phage T4 genomes did not disrupt capture in the pipelines used, though the presence of glucosyl-hydroxymethylcytosine did reduce recovery modestly.

Based on our assays, we prefer VP1 for analysis of stool samples, and VP4 for analysis of saliva samples based on their overall effectiveness in virus recovery, availability for

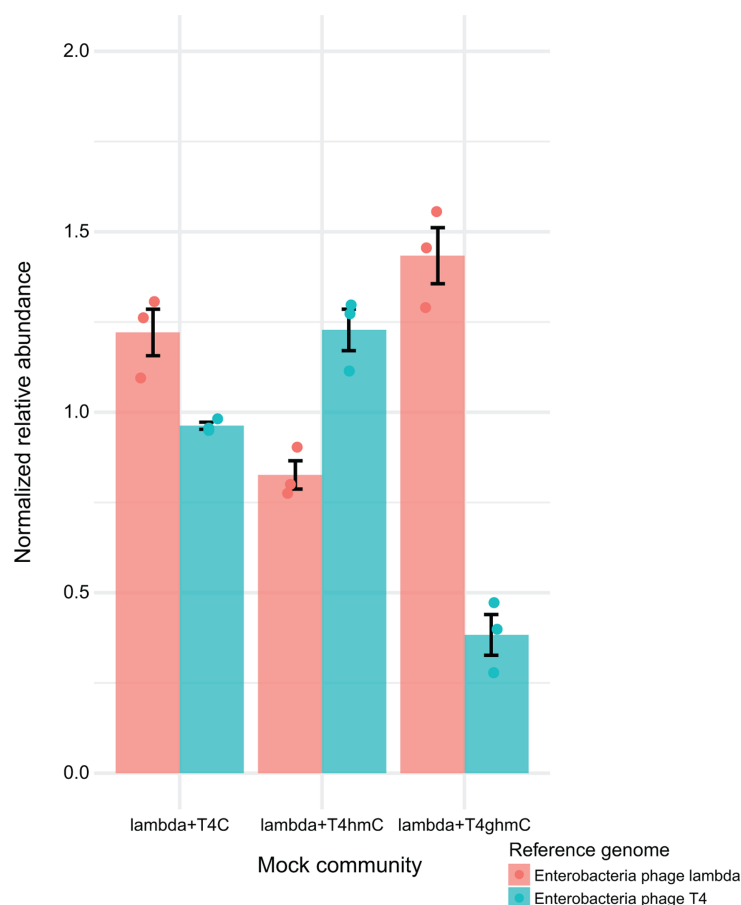


FIG 6 Assessing the potential inhibitory effect of ghmC and hmC on recovery of modified T4 viral genomes. T4 strain with (hmC, ghmC) or without (C) genome base modifications was mixed with lambda, directly extracted for nucleic acid, and subjected to Illumina sequencing. The graph shows the relative abundance of T4 and the lambda control by RPKM from sequencing normalized to that of the input genome copy number from qPCR. Statistical analyses were performed separately for each species. Differences in normalized relative abundance of T4 across groups were assessed using a Kruskal–Wallis test ($\chi^2 = 7.2$, $df = 2$, $P = 0.027$). ghmC: glucosyl-hydroxymethylcytosine. hmC: hydroxymethylcytosine.

implementation, and higher throughput. Protocols for each are presented in Document S1. Ongoing work aims to address limitations in current methods, such as potential sample loss during dilutions and handling, and challenges with low-biomass samples.

Short-read and long-read sequencing each have advantages and limitations. Short-reads yield a large amount of data at relatively low cost, but the shorter sequencing length (around 300 bp) presents challenges to microbial genome assembly, especially for sequences with repetitive regions or high-GC content (85). This limitation could be overcome using single-molecule long-read sequencing, as in Oxford Nanopore or PacBio technologies. However, these can be lower-throughput, less accurate, and more expensive (86). In addition, many viral genomes are below 5 kb in length, and so might be removed in steps to purify longer molecules prior to sequence acquisition. As an initial approach to finding a sequencing platform that balances these factors, we compared the Illumina 1,000-cycle kit and 300-cycle kit. The 1,000-cycle kit generated more viral contigs of longer lengths and is suited to capturing short viral genomes. It will be useful to assess the utility of the different methods more fully.

Choices of virome enrichment methods are dependent on the sample type and goals of the experiment. For stool and saliva samples, VP1 and VP2 are reasonable as standard methods; VP3 may be useful in more specialized settings. In studies where small circular

single-stranded DNA viruses are the target of interest, it may actually be advantageous to use GenomiPhi to bias recovery in their favor, though GenomiPhi is not optimal under other circumstances. Viral particle enrichment is likely useful for analysis of stool and saliva, where rich viral communities are present along with a complex background. For dilute low-biomass samples, such as BAL, OP wash, and nasopharyngeal swab eluate, losses during handling become major factors. Recovery was better for VirMock1 in stool or saliva, indicating a potential carrier effect of the complex sample background. For low-biomass samples, direct nucleic acid sequencing without enrichment may be best; optimization tests are ongoing. We note that a limitation of this paper is that there are multiple different steps in each protocol, each of which could be carried out in multiple ways, so that not all methods were fully queried; it will be valuable in future studies to explore more of the experimental parameter space.

Multiple additional methods can be used for virome analysis, each of which has strengths and weaknesses. Hybridization capture is one useful method (87), in which viral sequences are enriched from a complex mixture by binding to a library of tagged oligonucleotides. This can be useful when scanning for known viruses, for example, in identifying a virus of human cells responsible for a disease outbreak. This method is not useful for studying viruses that are not complementary to the capture primers, commonly including bacteriophages.

Some of the most extreme variations in viral genomes and related genetic elements were not examined here. Some of the giant viruses (Phylum Nucleocytoviricota) have huge particle sizes and genomes (41), and may be lost during handling. We included vaccinia as a representative of this phylum, but other Nucleocytoviricota are much larger and would be useful to study. At the other extreme, small circular RNAs of only a few hundred bases have been found in diverse sample types (43), some of which may be viruses. Further work to quantify recovery of these elements would be useful.

Sample composition, including the moisture and softness of stool, viscosity of saliva, etc., and effects of storage conditions will also likely influence virome data. Preliminary examination of different storage conditions reveals that processing the stool using VP1 immediately after stool solubilization, after overnight storage at 4°C, or undergoing one round of freeze-thaws from –80°C did not affect the virome enrichment result (data not shown). Further investigation of more parameters would be helpful.

In summary, where possible, it is useful to enrich virome specimens for particles before sequencing to optimize capture of previously unknown viral lineages. This work presents checks on a variety of steps, taking advantage of samples spiked with a community of known composition. However, there are multiple reasonable alternatives at each step; here we review outcomes over a variety of protocol choices, but many more alternative methods could reasonably be considered. It would be useful if multiple groups carrying out such methods optimization studies would publish their work to accelerate further technology refinement.

MATERIALS AND METHODS

Human samples

OP wash involved subjects swishing and gargling 10 mL of 1% saline. Saliva was collected (approximately 5 mL) by subjects spitting into a sterile container. Lower respiratory tract samples, specifically BAL, were obtained via bronchoscopy (Protocol #810851). To minimize contamination, upper respiratory tract samples were collected prior to anesthesia, and for bronchoscopy, the first lavage return was discarded.

Materials and reagents used

Key reagents are listed in Table S9.

Virome enrichment, sequencing, and analysis methods

Viral particle enrichment was carried out essentially as described. Detailed protocols are in the supplemental material (Methods S1; Tables S10 to S14).

ACKNOWLEDGMENTS

We are grateful to members of the Bushman and Collman laboratories for help and suggestions and to participants who provided samples. Thanks go to Laurie Zimmerman for help with artwork. We thank Dr. Mark Goulian, Dr. Stuart Isaacs, and Dr. Susan Weiss for providing viruses. We thank Illumina for making the 1,000-cycle kit available.

J.D., A.D.M., R.G.C., and F.D.B. conceived the study; Y.Z. and A.T. carried out mass spec analysis; J.D., A.D.M., M.H., Y.H., A.F., and K.L. carried out biochemical analysis; S.H., A.M.M., N.W., and K.B. carried out metagenomic DNaseq and RNAseq analysis; J.D., A.D.M., and N.W. carried out bioinformatic analysis; J.D., A.D.M., R.G.C., and F.D.B. wrote the paper.

This work was supported in part by The PennCHOP Microbiome Program and NIH grants U54AG089323, P30AI045008, and U19AI174998.

AUTHOR AFFILIATIONS

¹Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

²Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA

³Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

⁴Division of Gastroenterology, Hepatology & Nutrition, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

⁵Center for Microbial Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

⁶Pulmonary and Critical Care Division, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

PRESENT ADDRESS

Matthew Hogenauer, UT Health MD Anderson Cancer Center, Houston, Texas, USA

AUTHOR ORCIDs

Kyle Bittinger  <http://orcid.org/0000-0003-3472-5934>

Ahmed M. Moustafa  <http://orcid.org/0000-0002-9949-6936>

Ronald G. Collman  <http://orcid.org/0000-0002-0508-3701>

Frederic D. Bushman  <http://orcid.org/0000-0003-4740-4056>

FUNDING

Funder	Grant(s)	Author(s)
National Institutes of Health	U54AG089323	Ronald G. Collman Frederic D. Bushman
National Institutes of Health	P30AI045008	Ronald G. Collman
National Institutes of Health	U19AI174998	Frederic D. Bushman

AUTHOR CONTRIBUTIONS

Jiayi Duan, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing | Andrew D. Marques, Conceptualization,

Data curation, Formal analysis, Investigation, Project administration, Supervision, Writing – original draft, Writing – review and editing | Matthew Hogenauer, Investigation | Young Hwang, Investigation, Methodology, Validation | Yanjia Zhang, Formal analysis, Investigation, Validation, Writing – review and editing | Aaron Timperman, Resources, Supervision | Stephanie Higgins, Investigation, Writing – review and editing | Naomi G. Wilson, Data curation, Formal analysis, Software | Elizabeth Aine Fitts, Investigation | Haeun Karissa Lim, Investigation | Kyle Bittinger, Resources, Supervision | Ahmed M. Moustafa, Resources, Supervision, Writing – review and editing | Ronald G. Collman, Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review and editing | Frederic D. Bushman, Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

Sample information and raw sequences are available in the National Center for Biotechnology Information Sequence Read Archive under BioProject ID [PRJNA1348896](https://ncbi.nlm.nih.gov/bioproject/PRJNA1348896) (Table S15). Intermediate data files are deposited at Zenodo: <https://zenodo.org/records/17436149>. Raw LC-MS/MS data for DNA nucleoside composition analysis (enzymatic hydrolysis of genomic DNA into individual nucleosides) are deposited in the MassIVE repository under accession number [MSV000099555](https://massive.ucsf.edu/MSV000099555). All bioinformatics scripts used are deposited at GitHub: <https://github.com/jduan7/VLP-enrichment-methods-optimization-2025>.

ETHICS APPROVAL

All human subject research was conducted under protocols approved by the University of Pennsylvania Institutional Review Board (IRB). The samples were collected under protocols #849982, #842613, #810851, #851540, #828422, #834299, and #814859. All samples were collected after informed consent.

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

Document S1 (mSystems00188-26-s0001.docx). Additional protocol details.

Methods S1 (mSystems00188-26-s0002.docx). Additional experimental details.

Supplemental Figure Legends (mSystems00188-26-s0003.docx). Legends for Fig. S1 to S11.

Supplemental Figures (mSystems00188-26-s0004.pdf). Fig. S1 to S11.

Supplemental Tables (mSystems00188-26-s0005.xlsx). Tables S1 to S15.

REFERENCES

- Mushegian AR. 2020. Are there 10^{31} virus particles on earth, or more, or fewer? *J Bacteriol* 202:e00052-20. <https://doi.org/10.1128/jb.00052-20>
- Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci USA* 96:2192–2197. <https://doi.org/10.1073/pnas.96.5.2192>
- Crockford PW, Bar On YM, Ward LM, Milo R, Halevy I. 2023. The geologic history of primary productivity. *Curr Biol* 33:4741–4750. <https://doi.org/10.1016/j.cub.2023.09.040>
- Koonin EV, Kuhn JH, Dolja VV, Krupovic M. 2024. Megataxonomy and global ecology of the virosphere. *ISME J* 18:18. <https://doi.org/10.1093/ismej/wrad042>
- Bergh Ø, Børshheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found in aquatic environments. *Nature* 340:467–468. <https://doi.org/10.1038/340467a0>
- Suttle CA. 2007. Marine viruses—major players in the global ecosystem. *Nat Rev Microbiol* 5:801–812. <https://doi.org/10.1038/nrmicro1750>
- Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F. 2003. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 185:6220–6223. <https://doi.org/10.1128/JB.185.20.6220-6223.2003>
- Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. 2009. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4:470–483. <https://doi.org/10.1038/nprot.2009.10>
- Liang G, Bushman FD. 2021. The human virome: assembly, composition and host interactions. *Nat Rev Microbiol* 19:514–527. <https://doi.org/10.1038/s41579-021-00536-5>
- Liang G, Gao H, Bushman FD. 2022. The pediatric virome in health and disease. *Cell Host Microbe* 30:639–649. <https://doi.org/10.1016/j.chom.2022.04.006>

11. Liang G, Cobián-Güemes AG, Albenberg L, Bushman F. 2021. The gut virome in inflammatory bowel diseases. *Curr Opin Virol* 51:190–198. <https://doi.org/10.1016/j.coviro.2021.10.005>
12. Garmaeva S, Gulyaeva A, Sinha T, Shkoporov AN, Clooney AG, Stockdale SR, Spreckels JE, Sutton TDS, Draper LA, Dutilh BE, Wijmenga C, Kurilshikov A, Fu J, Hill C, Zhernakova A. 2021. Stability of the human gut virome and effect of gluten-free diet. *Cell Rep* 35:109132. <https://doi.org/10.1016/j.celrep.2021.109132>
13. Carding SR, Davis N, Hoyles L. 2017. Review article: the human intestinal virome in health and disease. *Aliment Pharmacol Ther* 46:800–815. <https://doi.org/10.1111/apt.14280>
14. Jiang L, Lang S, Duan Y, Zhang X, Gao B, Chopyk J, Schwanemann LK, Ventura-Cots M, Bataller R, Bosques-Padilla F, et al. 2020. Intestinal virome in patients with alcoholic Hepatitis. *Hepatology* 72:2182–2196. <https://doi.org/10.1002/hep.31459>
15. Nakatsu G, Zhou H, Wu WKK, Wong SH, Coker OO, Dai Z, Li X, Szeto C-H, Sugimura N, Lam TY-T, Yu AC-S, Wang X, Chen Z, Wong MC-S, Ng SC, Chan MTV, Chan PKS, Chan FKL, Sung JJ-Y, Yu J. 2018. Alterations in enteric virome are associated with colorectal cancer and survival outcomes. *Gastroenterology* 155:529–541. <https://doi.org/10.1053/j.gastro.2018.04.018>
16. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, Stappenbeck TS, McGovern DPB, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ, Wang D, Parkes M, Virgin HW. 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160:447–460. <https://doi.org/10.1016/j.cell.2015.01.002>
17. Tetz G, Brown SM, Hao Y, Tetz V. 2018. Parkinson's disease and bacteriophages as its overlooked contributors. *Sci Rep* 8:10812. <https://doi.org/10.1038/s41598-018-29173-4>
18. Castro-Mejía JL, Muhammed MK, Kot W, Neve H, Franz CMAP, Hansen LH, Vogensen FK, Nielsen DS. 2015. Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* 3:64. <https://doi.org/10.1186/s40168-015-0131-4>
19. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, Ryan FJ, Draper LA, Plevy SE, Ross RP, Hill C. 2019. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host Microbe* 26:764–778. <https://doi.org/10.1016/j.chom.2019.10.009>
20. Gogokhia L, Buhrke K, Bell R, Hoffman B, Brown DG, Hanke-Gogokhia C, Ajami NJ, Wong MC, Ghazaryan A, Valentine JF, Porter N, Martens E, O'Connell R, Jacob V, Scherl E, Crawford C, Stephens WZ, Casjens SR, Longman RS, Round JL. 2019. Expansion of bacteriophages is linked to aggravated intestinal inflammation and colitis. *Cell Host Microbe* 25:285–299. <https://doi.org/10.1016/j.chom.2019.01.008>
21. Nanda AM, Thormann K, Frunzke J. 2015. Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. *J Bacteriol* 197:410–419. <https://doi.org/10.1128/JB.02230-14>
22. Nishiyama H, Endo H, Blanc-Mathieu R, Ogata H. 2020. Ecological structuring of temperate bacteriophages in the inflammatory bowel disease-affected gut. *Microorganisms* 8:8. <https://doi.org/10.3390/microorganisms8111663>
23. Qv L, Mao S, Li Y, Zhang J, Li L. 2021. Roles of gut bacteriophages in the pathogenesis and treatment of inflammatory bowel disease. *Front Cell Infect Microbiol* 11:755650. <https://doi.org/10.3389/fcimb.2021.755650>
24. Sausset R, Petit MA, Gaboriau-Routhiau V, De Paepe M. 2020. New insights into intestinal phages. *Mucosal Immunol* 13:205–215. <https://doi.org/10.1038/s41385-019-0250-5>
25. Aggarwala V, Liang G, Bushman FD. 2017. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mob DNA* 8:12. <https://doi.org/10.1186/s13100-017-0095-y>
26. Gago S, Elena SF, Flores R, Sanjuán R. 2009. Extremely high mutation rate of a hammerhead viroid. *Science* 323:1308. <https://doi.org/10.1126/science.1169202>
27. Koonin EV, Lee BD. 2025. Diversity and evolution of viroids and viroid-like agents with circular RNA genomes revealed by metatranscriptome mining. *Nucleic Acids Res* 53:53. <https://doi.org/10.1093/nar/gkae1278>
28. Pei Z, Liu Y, Chen Y, Pan T, Sun X, Wang H, Ross RP, Lu W, Chen W. 2024. A universe of human gut-derived bacterial prophages: unveiling the hidden viral players in intestinal microecology. *Gut Microbes* 16:2309684. <https://doi.org/10.1080/19490976.2024.2309684>
29. d'Humières C, Touchon M, Dion S, Cury J, Ghazlane A, Garcia-Garcera M, Bouchier C, Ma L, Denamur E, P.C.Rocha E. 2019. A simple, reproducible and cost-effective procedure to analyse gut phageome: from phage isolation to bioinformatic approach. *Sci Rep* 9:11331. <https://doi.org/10.1038/s41598-019-47656-w>
30. Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL. 2015. Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLoS One* 10:e0134802. <https://doi.org/10.1371/journal.pone.0134802>
31. Kleiner M, Hooper LV, Duerkop BA. 2015. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* 16:7. <https://doi.org/10.1186/s12864-014-1207-4>
32. Kohl C, Brinkmann A, Dabrowski PW, Radonić A, Nitsche A, Kurth A. 2015. Protocol for metagenomic virus detection in clinical specimens. *Emerg Infect Dis* 21:48–57. <https://doi.org/10.3201/eid2101.140766>
33. Parras-Moltó M, Rodríguez-Galet A, Suárez-Rodríguez P, López-Bueno A. 2018. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* 6:119. <https://doi.org/10.1186/s40168-018-0507-3>
34. Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, Coleman ML, Breitbart M, Sullivan MB. 2016. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* 4:e2777. <https://doi.org/10.7717/peerj.2777>
35. Iker BC, Bright KR, Pepper IL, Gerba CP, Kitajima M. 2013. Evaluation of commercial kits for the extraction and purification of viral nucleic acids from environmental and fecal samples. *J Virol Methods* 191:24–30. <https://doi.org/10.1016/j.jviromet.2013.03.011>
36. Hall RJ, Wang J, Todd AK, Bissielo AB, Yen S, Strydom H, Moore NE, Ren X, Huang QS, Carter PE, Peacey M. 2014. Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J Virol Methods* 195:194–204. <https://doi.org/10.1016/j.jviromet.2013.08.035>
37. Milani C, Casey E, Lugli GA, Moore R, Kaczorowska J, Feehily C, Mangifesta M, Mancabelli L, Duranti S, Turroni F, Bottacini F, Mahony J, Cotter PD, McAuliffe FM, van Sinderen D, Ventura M. 2018. Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. *Microbiome* 6:145. <https://doi.org/10.1186/s40168-018-0527-z>
38. Daly GM, Bexfield N, Heaney J, Stubbs S, Mayer AP, Palser A, Kellam P, Drou N, Caccamo M, Tiley L, Alexander GJM, Bernal W, Heeney JL. 2011. A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. *PLoS One* 6:e28879. <https://doi.org/10.1371/journal.pone.0028879>
39. Zhai X, Gobbi A, Kot W, Krych L, Nielsen DS, Deng L. 2024. A single-stranded based library preparation method for virome characterization. *Microbiome* 12:219. <https://doi.org/10.1186/s40168-024-01935-5>
40. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JL. 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466:334–338. <https://doi.org/10.1038/nature09199>
41. Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. 2021. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol* 19:e3001430. <https://doi.org/10.1371/journal.pbio.3001430>
42. Kaczorowska J, van der Hoek L. 2020. Human anelloviruses: diverse, omnipresent and commensal members of the virome. *FEMS Microbiol Rev* 44:305–313. <https://doi.org/10.1093/femsre/fuaa007>
43. Zheludev IN, Edgar RC, Lopez-Galiano MJ, de la Peña M, Babaian A, Bhatt AS, Fire AZ. 2024. Viroid-like colonists of human microbiomes. *Cell* 187:6521–6536. <https://doi.org/10.1016/j.cell.2024.09.033>
44. Lee BD, Neri U, Roux S, Wolf YI, Camargo AP, Krupovic M, RNA Virus Discovery Consortium, Simmonds P, Kyripides N, Gophna U, Dolja VV, Koonin EV. 2023. Mining metatranscriptomes reveals a vast world of viroid-like circular RNAs. *Cell* 186:646–661. <https://doi.org/10.1016/j.cell.2022.12.039>
45. Shkoporov AN, Ryan FJ, Draper LA, Forde A, Stockdale SR, Daly KM, McDonnell SA, Nolan JA, Sutton TDS, Dalmaso M, McCann A, Ross RP, Hill C. 2018. Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6:68. <https://doi.org/10.1186/s40168-018-0446-z>
46. Liang G, Zhao C, Zhang H, Mattei L, Sherrill-Mix S, Bittinger K, Kessler LR, Wu GD, Baldassano RN, DeRusso P, Ford E, Elovitz MA, Kelly MS, Patel MZ, Mazhani T, Gerber JS, Kelly A, Zemel BS, Bushman FD. 2020. The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* 581:470–474. <https://doi.org/10.1038/s41586-020-2192-1>

47. Conceição-Neto N, Zeller M, Lefrère H, De Bruyn P, Beller L, Deboutte W, Yinda CK, Lavigne R, Maes P, Van Ranst M, Heylen E, Matthijssens J. 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep* 5:16532. <https://doi.org/10.1038/srep16532>
48. Li L, Deng X, Mee ET, Collot-Teixeira S, Anderson R, Schepelmann S, Minor PD, Delwart E. 2015. Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent. *J Virol Methods* 213:139–146. <https://doi.org/10.1016/j.jviromet.2014.12.002>
49. Young JC, Chehoud C, Bittinger K, Bailey A, Diamond JM, Cantu E, Haas AR, Abbas A, Frye L, Christie JD, Bushman FD, Collman RG. 2015. Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. *Am J Transplant* 15:200–209. <https://doi.org/10.1111/ajt.13031>
50. Kramná L, Kolářová K, Oikarinen S, Pursiheimo J-P, Ilonen J, Simell O, Knip M, Veijola R, Hyöty H, Cinek O. 2015. Gut virome sequencing in children with early islet autoimmunity. *Diabetes Care* 38:930–933. <https://doi.org/10.2337/dc14-2490>
51. Warren RA. 1980. Modified bases in bacteriophage DNAs. *Annu Rev Microbiol* 34:137–158. <https://doi.org/10.1146/annurev.mi.34.100180.001033>
52. Loenen WAM, Raleigh EA. 2014. The other face of restriction: modification-dependent enzymes. *Nucleic Acids Res* 42:56–69. <https://doi.org/10.1093/nar/gkt747>
53. Bryson AL, Hwang Y, Sherrill-Mix S, Wu GD, Lewis JD, Black L, Clark TA, Bushman FD. 2015. Covalent modification of bacteriophage T4 DNA inhibits CRISPR-Cas9. *mBio* 6:e00648-15. <https://doi.org/10.1128/mBio.0648-15>
54. Clarke EL, Taylor LJ, Zhao C, Connell A, Lee JJ, Fett B, Bushman FD, Bittinger K. 2019. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome* 7:46. <https://doi.org/10.1186/s40168-019-0658-x>
55. Tisza MJ, Belford AK, Domínguez-Huerta G, Bolduc B, Buck CB. 2021. Cenote-taker 2 democratizes virus discovery and sequence annotation. *Virus Evol* 7:veaa100. <https://doi.org/10.1093/ve/veaa100>
56. Abbas A, Diamond JM, Chehoud C, Chang B, Kotzin JJ, Young JC, Imai I, Haas AR, Cantu E, Lederer DJ, Meyer KC, Milewski RK, Olthoff KM, Shaked A, Christie JD, Bushman FD, Collman RG. 2017. The perioperative lung transplant virome: torque teno viruses are elevated in donor lungs and show divergent dynamics in primary graft dysfunction. *Am J Transplant* 17:1313–1324. <https://doi.org/10.1111/ajt.14076>
57. Abbas AA, Young JC, Clarke EL, Diamond JM, Imai I, Haas AR, Cantu E, Lederer DJ, Meyer K, Milewski RK, Olthoff KM, Shaked A, Christie JD, Bushman FD, Collman RG. 2019. Bidirectional transfer of anelloviridae lineages between graft and host during lung transplantation. *Am J Transplant* 19:1086–1097. <https://doi.org/10.1111/ajt.15116>
58. Merenstein C, Liang G, Whiteside SA, Cobián-Güemes AG, Merlino MS, Taylor LJ, Glascock A, Bittinger K, Tanes C, Graham-Wooten J, Khatib LA, Fitzgerald AS, Reddy S, Baxter AE, Giles JR, Oldridge DA, Meyer NJ, Wherry EJ, McGinniss JE, Bushman FD, Collman RG. 2021. Signatures of COVID-19 severity and immune response in the respiratory tract microbiome. *medRxiv*:2021.04.02.21254514. <https://doi.org/10.1101/2021.04.02.21254514>
59. Taylor LJ, Keeler EL, Bushman FD, Collman RG. 2022. The enigmatic roles of anelloviridae and redondoviridae in humans. *Curr Opin Virol* 55:101248. <https://doi.org/10.1016/j.coviro.2022.101248>
60. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC. 2024. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 42:1303–1312. <https://doi.org/10.1038/s41587-023-01953-y>
61. Kim MS, Park EJ, Roh SW, Bae JW. 2011. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl Environ Microbiol* 77:8062–8070. <https://doi.org/10.1128/AEM.06331-11>
62. Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, Ndao IM, Warner BB, Tarr PI, Wang D, Holtz LR. 2015. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat Med* 21:1228–1234. <https://doi.org/10.1038/nm.3950>
63. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P, Kyrpides NC. 2021. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* 6:960–970. <https://doi.org/10.1038/s41564-021-00928-6>
64. Ogilvie LA, Jones BV. 2015. The human gut virome: a multifaceted majority. *Front Microbiol* 6:918. <https://doi.org/10.3389/fmicb.2015.00918>
65. Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* 4:e08490. <https://doi.org/10.7554/eLife.08490>
66. Santiago-Rodríguez TM, Hollister EB. 2022. Unraveling the viral dark matter through viral metagenomics. *Front Immunol* 13:1005107. <https://doi.org/10.3389/fimmu.2022.1005107>
67. Tisza MJ, Buck CB. 2021. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc Natl Acad Sci USA* 118:118. <https://doi.org/10.1073/pnas.2023202118>
68. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, Lauder A, Sherrill-Mix S, Chehoud C, Kelsen J, Conrad M, Collman RG, Baldassano R, Bushman FD, Bittinger K. 2017. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* 5:52. <https://doi.org/10.1186/s40168-017-0267-5>
69. Jurasz H, Pawłowski T, Perlejewski K. 2021. Contamination issue in viral metagenomics: problems, solutions, and clinical perspectives. *Front Microbiol* 12:745076. <https://doi.org/10.3389/fmicb.2021.745076>
70. Shaffer JP, Marotz C, Belda-Ferre P, Martino C, Wandro S, Estaki M, Salido RA, Carpenter CS, Zaramela LS, Minich JJ, Bryant M, Sanders K, Fraraccio S, Ackermann G, Humphrey G, Swafford AD, Miller-Montgomery S, Knight R. 2021. A comparison of DNA/RNA extraction protocols for high-throughput sequencing of microbial communities. *Biotechniques* 70:149–159. <https://doi.org/10.1016/btn-2020-0153>
71. Brennan C, Shaffer JP, Belda-Ferre P, Mohanty I, Weng Y, Cantrell K, Ackermann G, Allaband C, Bryant M, Farmer S, González A, McDonald D, Martino C, Meehan MJ, Rahman G, Salido RA, Schwartz T, Song SJ, Tribelhorn C, Tubb HM, Dorrestein PC, Knight R. 2025. Streamlined extraction of nucleic acids and metabolites from low- and high-biomass samples using isopropanol and matrix tubes. *Microbiol Spectr* 13:e01912-25. <https://doi.org/10.1128/spectrum.01912-25>
72. Dean FB, Nelson JR, Giesler TL, Lasken RS. 2001. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* 11:1095–1099. <https://doi.org/10.1101/gr.180501>
73. Gonzalez-Pena V, Natarajan S, Xia Y, Klein D, Carter R, Pang Y, Shaner B, Annu K, Putnam D, Chen W, Connolly J, Pruett-Miller S, Chen X, Easton J, Gawad C. 2021. Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc Natl Acad Sci USA* 118:118. <https://doi.org/10.1073/pnas.2024176118>
74. Ward BHK. 2021. Transplex whole transcriptome amplification of RNA from low cell-number samples. *J Biomol Tech* 21
75. Zong C, Lu S, Chapman AR, Xie XS. 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338:1622–1626. <https://doi.org/10.1126/science.1229164>
76. Haagmans R, Charity OJ, Baker D, Telatin A, Savva GM, Adriaenssens EM, Powell PP, Carding SR. 2025. Assessing bias and reproducibility of viral metagenomics methods for the combined detection of faecal RNA and DNA viruses. *Viruses* 17:155. <https://doi.org/10.3390/v17020155>
77. Yilmaz S, Allgaier M, Hugenholtz P. 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods* 7:943–944. <https://doi.org/10.1038/nmeth1210-943>
78. Adriaenssens EM, Farkas K, Harrison C, Jones DL, Allison HE, McCarthy AJ. 2018. Viromic analysis of wastewater input to a river catchment reveals a diverse assemblage of RNA viruses. *mSystems* 3:e00025-18. <https://doi.org/10.1128/mSystems.00025-18>
79. Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C. 2021. Biases in viral metagenomics-based detection, cataloguing and quantification of bacteriophage genomes in human faeces, a review. *Microorganisms* 9:524. <https://doi.org/10.3390/microorganisms9030524>
80. Shah SA, Deng L, Thorsen J, Pedersen AG, Dion MB, Castro-Mejía JL, Silins R, Romme FO, Sausset R, Jessen LE, et al. 2023. Expanding known viral diversity in the healthy infant gut. *Nat Microbiol* 8:986–998. <https://doi.org/10.1038/s41564-023-01345-7>
81. Acheson NH, Tamm I. 1970. Ribonuclease sensitivity of Semliki forest virus nucleocapsids. *J Virol* 5:714–717. <https://doi.org/10.1128/JVI.5.6.714-717.1970>
82. Zhang Y, Lu L, Li X. 2022. Detection technologies for RNA modifications. *Exp Mol Med* 54:1601–1616. <https://doi.org/10.1038/s12276-022-00821-0>
83. Rifat D, Wright NT, Varney KM, Weber DJ, Black LW. 2008. Restriction endonuclease inhibitor IPI* of bacteriophage T4: a novel structure for a dedicated target. *J Mol Biol* 375:720–734. <https://doi.org/10.1016/j.jmb.2007.10.064>

84. Bair CL, Black LW. 2007. A type IV modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified DNAs. *J Mol Biol* 366:768–778. <https://doi.org/10.1016/j.jmb.2006.11.051>
85. Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nat Methods* 8:61–65. <https://doi.org/10.1038/nmeth.1527>
86. Warburton PE, Sebra RP. 2023. Long-Read DNA sequencing: recent advances and remaining challenges. *Annu Rev Genomics Hum Genet* 24:109–132. <https://doi.org/10.1146/annurev-genom-101722-103045>
87. Ceballos-Garzon A, Comtet-Marre S, Peyret P. 2024. Applying targeted gene hybridization capture to viruses with a focus to SARS-CoV-2. *Virus Res* 340:199293. <https://doi.org/10.1016/j.virusres.2023.199293>