1 **Title:** DNase treatment improves viral enrichment in agricultural soil viromes

2

3 **Running Title:** DNase treatment of soil viromes

4

5 **Authors**: Jackson W. Sorensen[1], Laura A. Zinke[1], Anneliek M. ter Horst[1], Christian Santos-
6 Medellin[1], Alena Schroeder[1], Joanne B. Emerson[1,2*]

7 1. Department of Plant Pathology, University of California, Davis, CA, USA

8 2. Genome Center, University of California, Davis, CA, USA

9

10 ***Corresponding Author Contact**: Joanne B Emerson jbemerson@ucdavis.edu

11

12

Abstract

The small genomes of most viruses make it difficult to fully capture viral diversity in metagenomes dominated by DNA from cellular organisms. Viral size-fraction metagenomics (viromics) protocols facilitate enrichment of viral DNA from environmental samples, and these protocols typically include a DNase treatment of the post-0.2 µm viromic fraction to remove contaminating free DNA prior to virion lysis. However, DNase may also remove desirable viral genomic DNA (*e.g.*, contained in virions compromised due to frozen storage or laboratory processing), suggesting that DNase-untreated viromes might be useful in some cases. In order to understand how virome preparation with and without DNase treatment influences the resultant data, here we compared 15 soil viromes (7 DNase-treated, 8 untreated) from 8 samples collected from agricultural fields prior to tomato planting. DNase-treated viromes yielded significantly more assembled viral contigs, contained significantly less non-viral microbial DNA, and recovered more viral populations (vOTUs) through read mapping. However, DNase-treated and untreated viromes were statistically indistinguishable, in terms of ecological patterns across viral communities. Although results suggest that DNase treatment is preferable where possible, in comparison to previously reported total metagenomes from the same samples, both DNase-treated and untreated viromes were significantly enriched in viral signatures by all metrics compared, including a ~225 times greater proportion of viral reads in untreated viromes compared to total metagenomes. Thus, even without DNase treatment, viromics was preferable to total metagenomics for capturing viral diversity in these soils, suggesting that preparation of DNase-untreated viromes can be worthwhile when DNase treatment is not possible.

Importance

Viromics is becoming an increasingly popular method for characterizing soil viral communities. DNase treatment of the viral size fraction prior to DNA extraction is meant to reduce

39 contaminating free DNA and is a common step within viromics protocols to ensure sequences
40 are of viral origin. However, some samples may not be amendable to DNase treatment due to
41 viral particles being compromised either in storage (i.e. frozen) or during other sample
42 processing. To date, the effect of DNase treatment on the recovery of viruses and downstream
43 ecological interpretations of soil viral communities is not thoroughly understood. This work
44 sheds light on these questions and indicates that while DNase treatment of soil viromes
45 improves recovery of viral populations, this improvement is modest in comparison to the gains
46 made by viromics over total soil metagenomics. Further, DNase treatment may not be
47 necessary to observe the ecological patterns structuring soil viral communities.
48
49 **Introduction**
50 Viruses infect all three domains of life and play key roles not only in human health but
51 also in agriculture and global nutrient cycling (1–5). They are important in oceanic food webs,
52 and our understanding of their role in soils is growing rapidly (2, 6–14). Viral abundances are
53 estimated to range from 10^7 to 10^{10} virions per gram in soil (6, 11), and measurements from
54 transmission electron microscopy suggest that up to 28% of microbial cells in soil are actively
55 infected by viruses (15–17). Through metagenomic approaches, soil viral populations have
56 been implicated in soil carbon cycling and microbial community dynamics in changing
57 environments, including in thawing permafrost and other peatlands (18–20).
58 The study of soil viral communities has lagged behind analogous efforts in marine
59 systems, in part because the complex and heterogeneous nature of soil presents unique
60 challenges for recovering viral DNA (2, 10, 11, 14, 21). Although marine viral ecology has
61 benefitted from a viromics approach, in which purified, concentrated viral particles are subjected
62 to DNA extraction and metagenomic sequencing (12, 22, 23), most recent soil viral ecological
63 studies have focused on recovering viral signatures from total soil metagenomes (10, 19, 20,
64 24). Bioinformatic advances in viral contig identification (*e.g.*, through recognition of viral

hallmark genes and other viral sequence signatures) (25–28) and efforts to compile viral reference databases that include partial and putative viral genomes (1, 19, 29) have improved our ability to recognize viral genomic sequences in soil metagenomes. However, despite these advances, our ability to catalog soil viral diversity is still largely gated by the low prevalence of viral DNA in total soil metagenomes, which tend to be dominated by bacterial and archaeal sequences (10, 19, 30).

Fortunately, viral size-fractionation protocols (*e.g.*, passage of a sample through a 0.2 µm filter to remove most cells), initially used in marine and other aquatic systems (12, 22, 23, 31–34), have also been applied to soil (11, 35), and recent data suggest that these protocols can enrich the viral signal in sequencing data (10, 19, 21, 30). Through iterative steps of mechanical and/or chemical desorption and centrifugation, virus-sized particles are separated from the soil matrix and from microbial cells, and DNA can then be directly extracted and sequenced from this viral size fraction to generate a shotgun metagenome, known as a virome (11, 18, 19, 30, 36, 37). Our group has shown that this approach can greatly increase both the number of viral populations and the proportion of viral DNA in the produced sequencing data from soil viromes, compared to total metagenomes (19, 30). For example, in agricultural soils, on average 30 times more contigs were identified as viral and 585 times more reads were recruited to viral genomes in viromes, compared to total metagenomes from the same samples (30).

A common step in laboratory viromics protocols is treatment with DNase after the post-0.2 µm (viral) fraction has been purified and enriched but before DNA extraction (11, 21, 30, 37). Under the assumption that most viral particles (virions) remain intact with their genomic contents protected at this stage, a DNase treatment is meant to reduce the amount of extracellular and/or free, “relic” DNA (38) that may have been co-enriched with the virions. The amount of “relic” DNA in a given soil sample presumably varies widely, depending on the soil, and the amount recovered in a given metagenomic or viromic library will also depend on the laboratory

procedure(s) used to prepare the DNA (38–40). Estimates of relic DNA in soil vary (38, 39), but one study suggested that, on average, 40.7% of soil 16S rRNA gene amplicon sequences are relic (38). One meta-analysis of viromes (predominantly from freshwater, saline, and human gut environments, with none from soils) determined that a range of 0.2 % to 40.3 % of viromic reads were mapped to non-viral microbial genomes, suggesting the potential for substantial non-viral DNA contamination in some cases (41). However, the amount of free DNA contamination in soil viromes and the potential impact of this DNA on downstream analyses has yet to be thoroughly considered.

Although these prior results would suggest that DNase treatment is an important step in the process of preparing a virome, virions themselves can be compromised prior to DNA extraction, such that DNase treatment of these compromised virions may remove the very viral genomic DNA that was meant to be enriched. Virions can be compromised both naturally through degradation in the environment and potentially during sample collection, transportation, storage, and/or laboratory processing (42–44). In some cases, particularly if the virions were compromised after removal from the field, it may be desirable to recover DNA from these compromised virions. The successful enrichment of viral DNA via viromics without DNase treatment has been previously observed, for example, from a hypersaline lake and from soil (peat) samples stored frozen (19, 34). This suggests that, in cases in which DNase treatment of a virome is not possible due to loss of viral DNA, preparation of a virome that has not undergone DNase treatment may still be worthwhile. However, direct comparisons of DNase-treated and untreated viromes from the same samples have not been made in soil (or any other environment, to our knowledge), nor have these two types of viromes been placed in the context of recoverable viral sequences from total metagenomes.

Here we sought to better understand the differences between soil viromes prepared with and without DNase treatment, in order to more thoroughly evaluate the utility of non-DNase-treated soil viromes (hereafter, untreated viromes). Considering 15 viromes (7 DNase-treated

(previously reported, with one having failed at the library construction step (30)), and 8 untreated (new in this study)) from 8 agricultural soil samples, this study compares overall sequence complexity, assembly success, proportions of recoverable viral contigs, the percentage of viral reads, viral taxonomic diversity, and the downstream ecological interpretations that would be derived from these two treatments. We hypothesized that treatment with DNase would increase the recovery of viral contigs by decreasing the overall sequence complexity and improving assembly and, therefore, that DNase treatment would be preferable, where possible. We also suspected that the overall patterns of viral community beta-diversity across samples would not be significantly influenced by DNase treatment and that untreated viromes would yield substantially more recognizable viral sequences than the total metagenomes that were previously sequenced from these same samples (30).

Materials and Methods

Sample collection and soil processing:

Our sampling design and soil collection have been described previously (30). Briefly, eight agricultural tomato plots near the UC Davis campus (38°32'08"N, 121°46'22" W) were sampled on April 23rd of 2018. Each of the plots had been treated with one of four biochar amendments (650 °C pyrolyzed pine feedstock, 650 °C pyrolyzed coconut shell, 800 °C pyrolyzed almond shell, or no biochar control) on November 8th, 2017 as part of an ongoing study to investigate the impact of biochar treatment on agricultural production (Table S1). Tomato seedlings had not yet been planted at the time of sampling (the field was fallow). The top 30 cm of soil was collected using a 2.5 cm diameter probe, and a total of 8 probe cores per plot were combined into a single sterile bag per sample and transported on ice to the laboratory, where each sample was sieved through an 8 mm mesh.

Viral purification and DNA extraction for viromics

The eight DNase-treated viromes were prepared as previously described (30), and in the current study, the same soils were also prepared without DNase treatment, for a total of 16 samples. Laboratory processing for all samples was the same up to the DNase treatment step. Briefly, viromes were generated for each sample from 50 g of fresh soil separated into two 50 mL conical tubes, following a previous protocol (11) with slight modifications (30). To each of the two tubes per sample, 37.5 mL of 0.02 μm filtered AKC' extraction butter (10% PBS, 10 g/L potassium citrate, 1.44 g/L Na_2HPO_4, 0.24 g/L KH_2PO_4, 36.97 g/L $MgSO_4$) (37) was added. Tubes were briefly vortexed to homogenize the soil slurry and then shaken at 400 RPM for 15 minutes on an orbital shaker. Subsequently, each tube was vortexed for an additional 3 minutes before undergoing centrifugation at 4,700 x g for 15 minutes to pellet the soil. The two supernatants from the same sample were then filtered through a 0.22 μm polyethersulfone filter to remove most cells and combined into a 70 mL polycarbonate ultracentrifuge tube, which was centrifuged for three hours at 4 °C and 32,000 x g to pellet viral particles. Taking care not to disturb the pellet, the supernatant was discarded and the viral pellet resuspended in 200 μL of ultrapure water. The eight untreated samples (no DNase treatment) proceeded directly to DNA extraction at this point. To the eight samples designated for DNase treatment, as previously described (30), 30 units of RQ1 RNase-free DNase and 30 μL of 10X DNase buffer (Promega Corp., Madison, WI, USA) were added, and samples were incubated at room temperature for two hours before stopping the reaction with 30 μL of the DNase stop solution (Promega Corp. Madison, WI, USA). The eight DNase-treated samples underwent DNA extraction at this point. For both DNase-treated and untreated viromes, DNA was extracted using the DNeasy PowerSoil Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions with slight modifications, as previously described (30).

Library construction and sequencing

Libraries were constructed and sequenced by the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center. The DNA Hyper Prep library kit (Kapa Biosystems-Roche, Basel, Switzerland) was used for all libraries. A single lane of Illumina HiSeq 4000 paired-end 150 bp sequencing was used to generate all of the sequencing data with a targeted sequencing depth of 4 Gbp per sample. Raw sequences can be found under the BioProject accession PRJNA646773.

Sequence processing, assembly, and identifying viral contigs

All viromes were bioiniformatically processed from raw sequencing data (*i.e.*, those that were previously reported were re-processed here). Sequencing reads were quality-trimmed and primers removed using Trimmomatic (45). MEGAHIT was used with the “meta” preset to individually assemble each virome, using the paired quality-trimmed reads and a minimum contig length of 10 kbp (46). Analyses of overall assembly statistics (Figure 1) were performed on these data. Putative viral contigs were then identified using VirSorter in decontamination mode, retaining any contigs that were assigned as higher confidence categories (1, 2, 4, or 5), in accordance with established recommendations (10, 18, 25, 28). Analyses considering viral contigs not yet dereplicated into populations (*i.e.*, Figures 2A-C) were performed on these data.

Kmer analyses of virome sequence complexity

Kmer counting was performed using the khmer software package version 2.1.1 (47). Reads were first k-mer error trimmed using the command “trim-low-abund.py” before k-mers of size 31 were counted in each virome using the script “load-into-counting.py”.

Taxonomic identification of bacterial and archaeal 16S rRNA gene content in viromes

SortMeRNA was used with its internal SILVA bacterial and archaeal 16S rRNA gene databases (version V119) to identify partial 16S rRNA gene sequences present in the reads

194 from each virome (48, 49). Reads found to contain partial 16S rRNA gene sequences were then
195 classified using the Ribosomal Database Project classifier trained with the RDP training set and
196 a confidence cutoff of 0.8 (50). Classifications were collapsed at the phylum level to create a
197 phylum-by-sample table in order to investigate changes in the relative abundances of phyla
198 across DNase treatments.
199
200 *Viral population (vOTU) identification, read mapping, and vOTU detection criteria for ecological*
201 *analyses*
202 VirSorter-identified viral contigs (described above) were dereplicated through clustering,
203 using the 'psi-cd-hit.pl' command of CDHIT (51) with a minimum alignment length equal to 85%
204 of the smaller contig and minimum percent identity equal to 95%, in accordance with best
205 practices for identifying viral populations (vOTUs) (37). The resulting representative seed contig
206 sequences from each cluster were then used as our set ("database") of vOTUs for further
207 analysis. vOTU representative seed sequences were annotated using prodigal (52) and then
208 grouped into viral clusters (VCs) and taxonomically identified using vConTACT2 with its
209 "ProkaryoticViralRefSeq85-Merged" database (53).
210 In order to perform community ecological analyses, the relative abundances of each
211 vOTU in each sample were assessed by read mapping to the reference database of vOTUs.
212 Specifically, quality-trimmed reads were mapped to the database of vOTUs at a minimum
213 identity of 90% using BBMap (54). The resulting SAM files were then converted into sorted and
214 indexed BAM files using samtools (55). The trimmed pileup coverage and read count
215 abundance of each vOTU were calculated using BamM parse to generate tables of vOTU
216 abundances (average coverage depth) in each sample (56). We used bedtools to calculate the
217 per-base coverage for each vOTU in each sample, requiring that >75% of the vOTU contig
218 length be covered by at least one read for detection in a given virome (also known as "breadth")
219 (57, 58). The vOTU coverage tables generated to this point were considered in analyses with

220 "relaxed" detection criteria, meaning that reads mapped to vOTUs assembled from any sample
221 were included. For analyses using "stringent" detection criteria, we also required that, for a
222 given vOTU to be considered detected in a virome, an assembled contig from that same virome
223 and/or another virome within the same DNase treatment group had to be in the same >=95%
224 nucleotide identity vOTU cluster. In other words, that same vOTU (viral "species") must have
225 been assembled from a virome in the same treatment group, mimicking a condition in which
226 only that treatment had been performed and thus only vOTUs from that treatment would be in
227 the reference database for read mapping. The resulting vOTU coverage tables were used for
228 downstream ecological and statistical analyses.
229
230 *Ecological and statistical analyses*
231 After generating the vOTU tables, all ecological and statistical analyses were performed
232 in R (59). The vegan package was used to calculate Bray-Curtis dissimilarities (function vegdist)
233 using vOTU relative abundances, perform PERMANOVA (function adonis), and correlate
234 matrices (Mantel tests, function mantel) (60). Kruskal.test from the stats package was used to
235 perform the Kruskal-Wallis rank sum test. Boxplots were constructed using ggplot2 and
236 tanglegrams using the dendextend package (61, 62). For both Mantel tests and tanglegram
237 analyses comparing DNase treated and untreated viromes, the untreated virome from plot PN-L
238 was dropped from the analysis because its paired DNase-treated virome failed at the library
239 construction step.
240
241 *Data Availability*
242 All viromes analyzed and presented in the current study have been deposited in the
243 NCBI SRA under BioProject PRJNA646773.
244
245 **Results**

Comparison of metagenomic assembly success from DNase-treated and untreated viromes

We sampled eight agricultural plots that had been treated with four different biochar amendments (30) and generated two viromes (one treated with DNase and one untreated) from each sample. The DNase-treated viromes were part of a prior study (30) and the untreated viromes are new here. These 16 viromes were sequenced to a depth of 4 Gbp (range 3.65-4.53 Gbp), apart from a single DNase-treated virome from which library construction failed, as previously described (Table S2) (30). Despite equimolar DNA contributions to the sequenced pool of libraries, untreated samples recovered a greater number of sequencing reads compared to their DNase-treated counterparts (Kruskal-Wallis p= 0.02, untreated median = 28,008,452, DNase-treated median = 26,847,586, Supplementary Figure 1). However, after quality filtering, there was no significant difference in the number of reads between treatment types (Kruskal-Wallis p = 0.08). Overall, DNase-treated viromes assembled into significantly more contigs (average 917 DNase-treated contigs, 513 untreated, Kruskal-Wallis p = 0.002) and had a longer total assembly length than their paired untreated viromes (Figure 1). However, the average contig lengths and N50s (*i.e.* the shortest contig length where half the assembly length is represented in longer contigs and half in shorter contigs) were statistically indistinguishable between the two treatments (Figure 1, Table S3).

Sequence complexity and proportion of cellular organism-derived reads in DNase-treated compared to untreated viromes

We suspected that decreased sequence complexity in DNase-treated viromes contributed to the observed significant improvement in assembly, presumably due to the degradation of “free” DNA (*e.g.*, from bacteria and archaea, as opposed to viruses). We tested this in two ways: first by comparing the k-mer complexity between the two approaches, and second by comparing the 16S rRNA gene recovery. DNase-treated viromes tended to have more abundant k-mers and fewer singleton k-mers than their untreated counterparts (Figure

2D), and DNase-treated viromes had significantly fewer total k-mers per sample (Kruskal-Wallis p=0.002). We next asked whether the reduced complexity of the DNase-treated viromes could be attributable to a depletion of non-viral (e.g., bacterial and archaeal) DNA. Indeed, DNase-treated viromes had significantly fewer reads identifiable as 16S rRNA gene fragments by approximately two-fold (on average, 0.013% for DNase-treated, compared to 0.028% for untreated samples, Kruskal-Wallis p = 0.03, Figure 2E, Table S4). Based on taxonomic classification of these 16S rRNA gene fragments, nine of the 12 most abundant phyla across the dataset had a significantly lower abundance in the DNase-treated viromes (Figure 2F, Table S5), with Acidobacteria, Actinobacteria, and Candidatus Saccharibacteria showing the most significant differences between treatments. Bacteroidetes was the only phylum to increase in abundance in the DNase-treated viromes, and DNase treatment had no significant effect on Planctomycetes or Verrucomicrobia relative abundances.

Viral contig and viral population (vOTU) recovery from DNase-treated compared to untreated viromes

We next wanted to assess whether treating viromes with DNase prior to DNA extraction had an influence on our ability to recover viral contigs and, subsequently, viral populations (vOTUs). We identified putative viral contigs from each single-sample assembly using VirSorter, retaining only viral contigs from the higher confidence categories (categories 1, 2, 4, and 5) (10, 18, 25). Overall, DNase-treated viromes assembled significantly more putative viral contigs (424 median, 303-475 range) in comparison to untreated viromes (226 median, 131-332 range, Figure 2A, Kruskal-Wallis P value <0.01, Table S4).

Thus far, contigs from the same viral population could have been counted multiple times (*e.g.*, in assemblies from different viromes). In order to evaluate recovery of unique viral populations (vOTUs), we clustered all of the putative viral contigs from both DNase-treated and untreated viromes at 95% nucleotide identity into vOTUs (36). We then categorized these

vOTUs into three groups, according to their occurrence within and/or across treatments, as follows: vOTUs containing contigs assembled from solely DNase-treated viromes (DNase vOTUs), solely from DNase untreated viromes (NoDNase vOTUs), or assembled in viromes from both treatments (shared vOTUs). In total, we identified 2,176 vOTUs, of which 1,121 were classified as DNase vOTUs, 421 as NoDNase, and 634 as shared (Figure 3, Data Set S1). Thus, DNase treatment resulted in approximately 1.7 times greater assembly of viral populations. However, of the 1,121 vOTUs that were assembled solely in DNase-treated viromes, 1,016 (90.6%) were detected in untreated viromes through read mapping, meaning that DNA from the vast majority of these vOTUs was present in the untreated viromes but did not sufficiently assemble.

Comparing the proportions of virus-derived reads in DNase-treated and untreated viromes

With the set of 2,176 vOTUs as references for read mapping, we next sought to determine whether the smaller number of 16S rRNA gene reads was accompanied by an increase in viral reads in DNase-treated compared to untreated viromes. We mapped the quality-filtered reads from each sample to the dereplicated reference set of all vOTUs. A significantly higher number and fraction of reads from DNase-treated viromes mapped to vOTUs (on average, ~3.2 million, or 17% of reads per sample), compared to untreated viromes (on average, ~1.9 million, or 9% of reads per sample, Figure 2C, Table S4). DNase treatment improved viral enrichment more than two-fold, compared to untreated viromes.

Patterns in the taxonomy and types of vOTUs assembled from DNase-treated and untreated viromes

We next wanted to determine whether there were differences in the types of vOTUs recovered in DNase-treated compared to untreated viromes. We performed whole-genome, network-based clustering of predicted proteins, using vConTACT2 (53), to cluster groups of

vOTUs at approximately the genus level into viral clusters, or VCs (63). vConTACT2's collection of viral genomes from the NCBI RefSeq database (ViralRefSeq-prokaryotes-v85) was included in this analysis for assigning taxonomy, as previously described (53). Of the 2,176 total vOTUs, 1,457 (67.0%) clustered into 744 VCs. 599 VCs (80.5%) contained vOTUs assembled in both treatments, while 131 VCs (17.7%) exclusively contained vOTUs assembled from DNase-treated viromes and 14 VCs (1.8%) exclusively contained vOTUs assembled from untreated viromes. Only 43 vOTUs were assigned taxonomy, based on clustering in the same VC as a reference sequence, and these vOTUs accounted for 0.3%-1.6% of the total viral community abundance (based on read mapping) in each virome. Considering these limited taxonomic assignments, DNase-treated and untreated viromes generally did recover the same taxonomic groups, namely the Caudovirales families *Siphoviridae* (10 VCs), *Myoviridae* (5 VCs), and *Podoviridae* (4 VCs). We note that these results are based on the current taxonomies for the relevant reference sequences, but phage taxonomy is actively undergoing revision by the International Committee on Taxonomy of Viruses (ICTV), and these groups of Caudovirales have been recommended for removal as taxonomic groups (64).

Recovery of relatively rare compared to abundant vOTUs by treatment

We addressed relative abundance patterns (for example, whether recovered vOTUs tended to be relatively abundant or rare) by comparing the proportions of reads recruited to vOTUs in the three different vOTU source categories (*i.e.*, occurrence in assemblies within and/or across treatments, described earlier, Figure 4). The shared vOTUs (those assembled in at least one virome from both treatments) recruited on average 57.7% and 59.5% of mapped reads from DNase-treated and untreated viromes, respectively, despite these shared vOTUs only accounting for approximately 29% of the total vOTUs (634/2,176, Figure S2). While vOTUs uniquely assembled from DNase-treated viromes accounted for 52% of all vOTUs (1,121/2,176), they only recruited an average of 32.5% of mapped reads from DNase-treated viromes and a similar but slightly lower percentage of mapped reads from untreated viromes

(28.9%) (Figure S2). These results led us to suspect that the vOTUs uniquely assembled in DNase-treated viromes tended to be relatively rare (low abundance), compared to shared vOTUs in the dataset. To address this, we constructed frequency plots of the mean relative abundances of vOTUs by category (treatment-specific or shared) (Figure 3). In both DNase-treated and untreated viromes, the distribution of treatment-specific vOTU abundances was shifted to the left (indicating lower abundances), compared to the abundances of shared vOTUs (Kruskal-Wallis $p<0.001$). In untreated viromes, vOTUs assembled only from untreated viromes had similar mean relative abundances to vOTUs assembled only from DNase-treated viromes. In contrast, in DNase-treated viromes, vOTUs assembled only from DNase-treated viromes had significantly higher relative abundances compared to vOTUs assembled only from untreated viromes (Kruskal-Wallis $p<0.001$). In short, while there were some vOTUs that were both uniquely assembled in one treatment and in high abundance in one or both treatments, the vast majority of the treatment-specific vOTUs were in low relative abundance, compared to those that were assembled in both DNase-treated and untreated viromes.

Ecological inferences from DNase-treated compared to untreated viromes

In order to better understand how DNase treatment, or a lack thereof, might influence downstream ecological interpretations from soil viromic data, we applied and compared two different sets of vOTU detection criteria. For both analyses, we followed the same established best practices for considering a vOTU to be “detected” in a given sample (36). The first set of detection criteria, which we refer to as “relaxed”, considers data from reads mapped to all 2,176 reference vOTUs (i.e., vOTUs assembled from any virome in this study). The second set of criteria, referred to as “stringent”, removed from consideration reads that mapped to vOTUs that were only assembled from the other treatment group. This stringent set of criteria was meant to mimic a dataset in which only one treatment had been performed (DNase treated or untreated), as would be expected for most viromic studies. DNase-treated viromes had significantly higher

perceived richness (alpha diversity) than their untreated counterparts for both the relaxed (on average 1,128 vs. 985 vOTUs, Table S6, Figure 4AB, Kruskal-Wallis p = 0.003) and stringent (on average 980 vs. 494 vOTUs per untreated, Table S6, Kruskal-Wallis p = 0.001) criteria. While both DNase-treated and untreated viromes had lower observed richness using the stringent criteria, the untreated samples showed a greater decrease (approximately two-fold) in richness between the relaxed and stringent criteria.

We also wanted to determine whether DNase treatment affected analyses of viral community structure. Using both the relaxed and stringent criteria for vOTU detection, we calculated pairwise Bray-Curtis dissimilarities between viromes. With the relaxed criteria, there was no significant difference in viral community composition attributable to DNase treatment (Figure 4C, PerMANOVA p = 0.952), but the application of the stringent criteria did result in a significant effect of DNase treatment (Figure 4D, p = 0.002). We also wanted to assess whether one set of viromes exhibited a greater amount of variation than the other. When using the relaxed vOTU criteria, the beta dispersion (i.e., the breadth of beta diversity within a group) of the DNase-treated and untreated viromes was statistically indistinguishable (Homogeneity of multivariate dispersions p = 0.430). When applying the stringent vOTU detection criteria, the DNase-treated viromes trended towards showing greater beta dispersion, but the difference was not statistically significant (Homogeneity of multivariate dispersions p = 0.075).

Finally, we previously observed a strong East-to-West gradient effect on viral community composition in these agricultural fields, using DNase-treated viromes only (30). Under the assumption that this gradient effect was real, we assessed our ability to detect this effect in both the DNase-treated and untreated viromes. We observed a significant East-West structuring of viral community composition in both sets of viromes, using both relaxed and stringent criteria (Figure 4 CD, Table S7). We further confirmed the robustness of viral community compositional patterns to DNase treatment by testing for correlations between the Bray-Curtis community dissimilarity matrices derived from DNase-treated compared to untreated viromes, using Mantel

tests. The observed beta-diversity patterns (i.e., how samples were grouped according to viral community composition) were highly correlated between DNase-treated and untreated viromes, according to both stringent and relaxed vOTU detection criteria (Mantel R = 0.87 for relaxed criteria, R=0.83 for stringent criteria, both p = 0.002). This result was further reinforced in a tanglegram, which showed highly similar hierarchical clustering of samples according to viral community composition between the two virome treatments, independent of vOTU detection criteria (Figure 4 EF). A single sample (AS-L) clustered differently in tanglegrams derived from DNase-treated compared to untreated data when using the stringent vOTU detection criteria only. This was the only sample that lacked a paired sample from the same East-West position in the field, owing to the necessary removal of the single successful virome from that plot (plot PN-L) in this analysis, because its matched DNase-treated virome failed at the library construction step. Otherwise, each pair of samples from the same field column grouped together in all four hierarchical clusters (DNase-treated vs. untreated and relaxed vs. stringent vOTU detection criteria).

Comparing viral recovery from untreated viromes and total soil metagenomes

In most metrics that we have compared to this point, DNase-treated viromes have outperformed untreated viromes, but we wanted to know the extent to which untreated viromes could still improve viral sequence recovery and reduce bacterial and archaeal DNA content in viromes, compared to total soil metagenomes. We previously analyzed total soil metagenomes from these same samples (30), which showed an average of 2.2% viral contig content (compared to 45% for untreated viromes in this study, a ~20-fold improvement), 0.04% 16S rRNA gene reads (compared to 0.02% for untreated viromes here), and an average of 0.04% of reads mapping to vOTUs (compared to 9.2% for untreated viromes here, a ~225-fold improvement). Further, the ecological patterns observed in this study were robust to different DNase treatments (Figure 4), and we wanted to know the extent to which mining total

metagenomes for viral signatures would yield the same patterns. For example, a highly significant effect of spatial structuring (E-W gradient effect) on viral community composition was observed for untreated viromes here, even with the stringent detection criteria (PerMANOVA p=0.003), and we wanted to know the extent to which this pattern could also have been recovered from the total soil metagenomes. While this result was reproduced with viral communities recovered from the total soil metagenomes, the significance was borderline (PerMANOVA p=0.045).

Discussion

DNase treatment of viromes reduced contamination and sequence complexity, consistent with removal of free DNA

We have shown that DNase treatment of viromes significantly reduced sequence complexity and lowered the amount of contaminating cell-derived DNA (measured as 16S rRNA gene fragments) by about two-fold. Sequence complexity has long been a challenge for assembling environmental metagenomes and can result in high fragmentation of genomes from low-abundance species (65, 66). Thus, we suspect that the observed decrease in sequence complexity in DNase-treated viromes was responsible for the larger, more contiguous assemblies from DNase-treated viromes, and it is reasonable to assume that this reduction in sequence complexity resulted from free ("relic") DNA depletion as a result of successful DNase treatment.

Relic DNA (sometimes called environmental DNA, eDNA, or free DNA) is not contained within a viable cell or virion and has been shown to artificially increase the observed richness of microbial communities in some soils (38, 39), presumably by allowing for the detection of locally dead or extinct microbial taxa (38). Studies have also suggested that the presence of relic DNA can obscure or minimize patterns in beta diversity (38, 67, 68), but here we observed that both

454 DNase-treated and untreated viromes produced viral communities with highly correlated beta-
455 diversity patterns (Figure 4). Although there was a single sample that clustered differently in
456 DNase-treated compared to untreated viromes when using the stringent vOTU detection criteria,
457 we attribute this difference predominantly to the lack of a successful replicate matching sample
458 in the same column of the field, rather than differences in relic DNA composition between
459 treatments.
460
461 *Viromics without DNase treatment might be particularly useful for samples stored frozen*
462 The laboratory protocol for generating viromes requires equipment that is unlikely to be
463 available or practical to run in the field, precluding immediate processing of samples collected
464 from distant field sites (19, 34). Even samples from nearby sites may need to be stored
465 temporarily, as a relatively small number of samples can be processed for viromics at a time (6-
466 12 per ~2 days in our lab, but this will depend on available equipment and personnel). Frozen
467 storage can preserve *in situ* community composition (69–72), and ideally, virions would be
468 frozen in cryoprotectant or similar to preserve their integrity, but the compatibility of
469 cryoprotectants with various viromics protocols is not well known. Thus, in some cases, direct
470 freezing of samples may be necessary (19). We have previously shown that freezing can
471 prohibit the use of DNase on aquatic viromes, resulting in viral DNA yields below detection limits
472 (34), and anecdotally, we see have seen similar results from soils stored frozen (data not
473 shown).
474 Encouragingly, work from our group has shown that viromes prepared without DNase
475 treatment (untreated viromes) from frozen peat soils can still substantially improve vOTU
476 recovery, compared to total metagenomes (19). Similarly, hypersaline lake water stored frozen
477 yielded predominantly viral sequences in viromes that did not undergo DNase treatment (34). In
478 combination with the complete depletion of DNA after DNase treatment in these hypersaline
479 lake samples, it is reasonable to suspect that some virions became compromised by freezing,

such that DNase treatment removed valuable viral genomic DNA contained in degraded virions that may have been intact in the field. Studies in pure culture support virion degradation through freezing; for example, coliphages from wasterwater showed decreased viability after prolonged storage in frozen wastewater, and *Bacillus subtilis* bacteriophage viability decreased multiple logs after only two hours of frozen storage with no cryoprotectant (43, 44).

Here, to ensure that we could get sufficient DNA for sequencing from both treatments for a direct comparison, we compared fresh samples with and without DNase treatment. While results from this and prior studies converge to suggest that skipping DNase treatment is likely to be a good option for viromics from samples stored frozen, future comparisons would benefit from including a combination of samples processed fresh and after frozen storage.

Recommendations for future viral ecology studies

We have shown here that DNase treatment produced better assemblies, more viral contigs, fewer 16S rRNA gene reads (indicative of bacterial and archaeal DNA), and more viral reads in comparison to not treating with DNase. However, both kinds of viromes substantially outperformed total soil metagenomes in these metrics. Together, these results suggest that soil viromics with DNase treatment is the best approach for interrogating soil viral ecology, where possible, but soil virome preparation without DNase treatment can be better than total soil metagenomics when DNase treatment is not an option. Prior work suggests that these results may be generalizable to viromics in other ecosystems as well (34), but to our knowledge, direct comparisons of these approaches have not been made in other ecosystems. The decision of what approach to take is inherently dependent on the questions being asked, along with the logistics of sample collection, storage (and possibly shipment), and performing the laboratory sample processing. Where possible, we recommend processing soil viromes soon after sampling and including a DNase treatment after virion purification and prior to virion lysis.

505 However, even without DNase, soil viromes substantially enrich for viral sequences in
506 comparison to total soil metagenomes. Across multiple studies, including fresh, frozen,
507 agricultural, and peat soils in various combinations (this study and (19, 30)), viromics (with or
508 without DNase) seems to outperform total metagenomics for soil viral community investigations.
509 Still, only a tiny fraction of soil types and a few combinations of laboratory procedures have
510 been attempted, so assessing the broad generalizability of the observed trends will require
511 expanding our investigations across diverse terrestrial and other ecosystems. Thus far, the
512 extra effort required to purify virions from soil prior to DNA extraction seems to be worthwhile,
513 even without DNase treatment.
514

515 **Acknowledgements**

516 We thank Sanjai Parikh and Danielle Gelardi for designing and maintaining the field
517 experiment from which all samples for this study were derived. We thank Sara Geonczy,
518 Winston Bess, and Rose Bolle for contributing to discussions of this work.
519 This work was supported by the UC Davis College of Agricultural and Environmental
520 Sciences and Department of Plant Pathology (new lab start-up to JBE). This work was also
521 supported in part by the USDA National Institute of Food and Agriculture, Hatch project number
522 CA-D-PPA-2464-H/accession number 1016718. CSM was supported by an award from the U.S.
523 Department of Energy, Office of Science, Office of Biological and Environmental Research,
524 Genomic Science Program, #DE-SC0020163 (grant to JBE). The sequencing was performed by
525 the UC Davis Genome Center's DNA Technologies and Expression Analysis Core, supported
526 by NIH Shared Instrumentation Grant 1S10OD010786-01. Any opinions, findings, conclusions,
527 or recommendations expressed in this manuscript are those of the authors and do not
528 necessarily reflect the view(s) of any funding agencies or institutions.
529
530

References

1. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. 2020. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. Cell Host Microbe 28:724-740.e8.
2. Emerson JB. 2019. Soil Viruses: A New Hope. mSystems 4:1–4.
3. Brum JR, Ignacio-espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S. 2015. Patterns and ecological drivers of ocean viral communities. Science (80-) 348:1261498-1–11.
4. Suttle CA. 2007. Marine viruses - Major players in the global ecosystem. Nat Rev Microbiol 5:801–812.
5. Wang X, Wei Z, Yang K, Wang J, Jousset A, Xu Y, Shen Q, Friman VP. 2019. Phage combination therapies for bacterial wilt disease in tomato. Nat Biotechnol 37:1513–1520.
6. Williamson KE, Fuhrmann JJ, Wommack KE, Radosevich M. 2017. Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. Annu Rev Virol 4:201–219.
7. Pratama AA, Elsas JD Van. 2018. The ‘ Neglected ’ Soil Virome – Potential Role and Impact. Trends Microbiol 26:649–662.
8. Bonetti G, Trevathan-Tackett SM, Carnell PE, Macreadie PI. 2019. Implication of viral infections for greenhouse gas dynamics in freshwater wetlands: Challenges and perspectives. Front Microbiol 10.
9. Kuzyakov Y, Mason-Jones K. 2018. Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. Soil Biol Biochem 127:305–317.
10. Trubl G, Hyman P, Roux S, Abedon ST. 2020. Coming-of-age characterization of soil viruses: A user’s guide to virus isolation, detection within metagenomes, and viromics.

Soil Syst 4:1–34.

11. Trubl G, Solonenko N, Chittick L, Solonenko SA, Rich VI, Sullivan MB. 2016. Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. PeerJ 4:1–24.

12. Brum JR, Sullivan MB. 2015. Rising to the challenge: Accelerated pace of discovery transforms marine virology. Nat Rev Microbiol 13:147–159.

13. Breitbart M. 2012. Marine Viruses: Truth or Dare. Ann Rev Mar Sci 4:425–448.

14. Kimura M, Jia ZJ, Nakayama N, Asakawa S. 2008. Ecology of viruses in soils: Past, present and future perspectives. Soil Sci Plant Nutr 54:1–32.

15. Bowatte S, Newton PCD, Takahashi R, Kimura M. 2010. High frequency of virus-infected bacterial cells in a sheep grazed pasture soil in New Zealand. Soil Biol Biochem 42:708–712.

16. Takahashi R, Bowatte S, Taki K, Ohashi Y, Asakawa S, Kimura M. 2011. High frequency of phage-infected bacterial cells in a rice field soil in Japan. Soil Sci Plant Nutr 57:35–39.

17. Takahashi R, Saka N, Honjo H, Asakawa S, Kimura M. 2013. Comparison of the frequency of visibly infected bacterial cells between the soil and the floodwater in two Japanese rice fields. Soil Sci Plant Nutr 59:331–336.

18. Trubl G, Jang H Bin, Roux S, Emerson JB, Solonenko N, Vik DR, Solden L, Ellenbogen J, Runyon AT, Bolduc B, Woodcroft BJ, Saleska SR, Tyson GW, Wrighton KC, Sullivan MB, Rich VI. 2018. Soil viruses are underexplored players in ecosystem carbon processing. mSystems 3:1–21.

19. ter Horst AM, Santos-Medellin C, Sorensen JW, Zinke LA, Wilson RM, Johnston ER, Trubl G, Pett-Ridge J, Blazewicz SJ, Hanson PJ, Chanton JP, Schadt CW, Kostka JE, Emerson JB. 2020. Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. bioRxiv 1-.

20. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang H Bin, Singleton CM,

583 Solden LM, Naas AE, Boyd JA, Hodgkins SB, Wilson RM, Trubl G, Li C, Frolking S, Pope
584 PB, Wrighton KC, Crill PM, Chanton JP, Saleska SR, Tyson GW, Rich VI, Sullivan MB.
585 2018. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol
586 3:870–880.
587 21. Göller PC, Haro-Moreno JM, Rodriguez-Valera F, Loessner MJ, Gómez-Sanz E. 2020.
588 Uncovering a hidden diversity: Optimized protocols for the extraction of dsDNA
589 bacteriophages from soil. Microbiome 8:1–16.
590 22. Thurber R V., Haynes M, Breitbart M, Wegley L, Rohwer F. 2009. Laboratory procedures
591 to generate viral metagenomes. Nat Protoc 4:470–483.
592 23. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, Brum J, Polz MF,
593 Boyle EA, Sullivan MB. 2011. A simple and efficient method for concentration of ocean
594 viruses by chemical flocculation. Environ Microbiol Rep 3:195–202.
595 24. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M,
596 Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome.
597 Nature 536:425–430.
598 25. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: Mining viral signal from
599 microbial genomic data. PeerJ 2015:1–20.
600 26. Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: Automated recovery, annotation and
601 curation of microbial viruses, and evaluation of viral community function from genomic
602 sequences. Microbiome 8:1–23.
603 27. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F. 2020.
604 Identifying viruses from metagenomic data using deep learning. Quant Biol 8:64–77.
605 28. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-huerta G, Delmont TO, Pratama AA,
606 Gazitúa MC, Vik D, Sullivan MB, Roux S. 2021. VirSorter2: a multi-classifier, expert-
607 guided approach to detect diverse DNA and RNA viruses. Microbiome 9:1–13.
608 29. Roux S, Páez-Espino D, Chen IMA, Palaniappan K, Ratner A, Chu K, Reddy TBK,

609 Nayfach S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA,
610 Kyrpides NC. 2021. IMG/VR v3: an integrated ecological and evolutionary framework for
611 interrogating genomes of uncultivated viruses. Nucleic Acids Res 49:D764–D775.
612 30. Santos-Medellin C, Zinke LA, ter Horst AM, Gelardi DL, Parikh SJ, Emerson JB. 2021.
613 Viromes outperform total metagenomes in revealing the spatiotemporal patterns of
614 agricultural soil viral communities. ISME J https://doi.org/10.1101/2020.08.06.237214.
615 31. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer
616 F. 2002. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U
617 S A 99:14250–14255.
618 32. Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D. 2008.
619 Assembly of viral metagenomes from Yellowstone hot springs. Appl Environ Microbiol
620 74:4164–4174.
621 33. Rodriguez-Brito B, Li LL, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues
622 C, Dinsdale E, Edwards R, Felts B, Haynes M, Liu H, Lipson D, Mahaffy J, Martin-
623 Cuadrado AB, Mira A, Nulton J, Pašić L, Rayhawk S, Rodriguez-Mueller J, Rodriguez-
624 Valera F, Salamon P, Srinagesh S, Thingstad TF, Tran T, Thurber RV, Willner D, Youle
625 M, Rohwer F. 2010. Viral and microbial community dynamics in four aquatic
626 environments. ISME J 4:739–751.
627 34. Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfielda JF. 2012.
628 Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly.
629 Appl Environ Microbiol 78:6309–6320.
630 35. Williamson KE, Radosevich M, Wommack KE. 2005. Abundance and Diversity of Viruses
631 in Six Delaware Soils 71:3119–3125.
632 36. Roux S, Adriaenssens EM, Dutilh BE, Koonin E V., Kropinski AM, Krupovic M, Kuhn JH,
633 Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M,
634 Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA,

635 Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee KB, Malmstrom
636 RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit MA, Putonti C, Rattei
637 T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan
638 MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS,
639 Whiteson KL, Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T,
640 Young MJ, Yutin N, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. 2019. Minimum information
641 about an uncultivated virus genome (MIUVIG). Nat Biotechnol 37:29–37.
642 37. Trubl G, Roux S, Solonenko N, Li YF, Bolduc B, Rodríguez-Ramos J, Eloe-Fadrosh EA,
643 Rich VI, Sullivan MB. 2019. Towards optimized viral metagenomes for double-stranded
644 and single-stranded DNA viruses from challenging soils. PeerJ 2019:1–24.
645 38. Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. 2016. Relic DNA is
646 abundant in soil and obscures estimates of soil microbial diversity. Nat Microbiol 2:1–6.
647 39. Lennon JT, Muscarella ME, Placella SA, Lehmkuhl BK. 2018. How , When , and Where
648 Relic DNA Affects Microbial Diversity. MBio 9:1–14.
649 40. Emerson JB, Adams RI, Román CMB, Brooks B, Coil DA, Dahlhausen K, Ganz HH,
650 Hartmann EM, Hsu T, Justice NB, Paulino-Lima IG, Luongo JC, Lymperopoulou DS,
651 Gomez-Silvan C, Rothschild-Mancinelli B, Balk M, Huttenhower C, Nocker A,
652 Vaishampayan P, Rothschild LJ. 2017. Schrödinger's microbes: Tools for distinguishing
653 the living from the dead in microbial ecosystems. Microbiome 5:86.
654 41. Roux S, Krupovic M, Debroas D, Forterre P, Enault F. 2013. Assessment of viral
655 community functional potential from viral metagenomes may be hampered by
656 contamination with cellular sequences. Open Biol 3.
657 42. Jończyk E, Kłak M, Międzybrodzki R, Górski A. 2011. The influence of external factors on
658 bacteriophages-review. Folia Microbiol (Praha) 56:191–200.
659 43. Nagai T. 2019. Viability of Bacillus subtilis (natto) bacteriophages after freezing and
660 thawing. Japan Agric Res Q 53:7–12.

661 44. Torrella F, López JP, Banks CJ. 2003. Survival of indicators of bacterial and viral
662 contamination in wastewater subjected to low temperatures and freezing: Application to
663 cold climate waste stabilisation ponds. Water Sci Technol 48:105–112.
664 45. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina
665 sequence data. Bioinformatics 30:2114–2120.
666 46. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node
667 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
668 Bioinformatics 31:1674–1676.
669 47. Brown CT, Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R,
670 Charbonneau A, Constantinides B, Edvenson G, Fay S, Fenton J, Fenzl T, Fish J,
671 Garcia-Gutierrez L, Garland P, Gluck J, González I, Guermond S, Guo J, Gupta A, Herr
672 JR, Howe A, Hyer A, Härpfer A, Irber L, Kidd R, Lin D, Lippi J, Mansour T, McA'Nulty P,
673 McDonald E, Mizzi J, Murray KD, Nahum JR, Nanlohy K, Nederbragt AJ, Ortiz-Zuazaga
674 H, Ory J, Pell J, Pepe-Ranney C, Russ ZN, Schwarz E, Scott C, Seaman J, Sievert S,
675 Simpson J, Skennerton CT, Spencer J, Srinivasan R, Standage D, Stapleton JA,
676 Steinman SR, Stein J, Taylor B, Trimble W, Wiencko HL, Wright M, Wyss B, Zhang Q,
677 Zyme E. 2015. The khmer software package: Enabling efficient nucleotide sequence
678 analysis. F1000Research 4:1–12.
679 48. Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: Fast and accurate filtering of ribosomal
680 RNAs in metatranscriptomic data. Bioinformatics 28:3211–3217.
681 49. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO.
682 2013. The SILVA ribosomal RNA gene database project: Improved data processing and
683 web-based tools. Nucleic Acids Res 41:590–596.
684 50. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid
685 assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol
686 73:5261–5267.

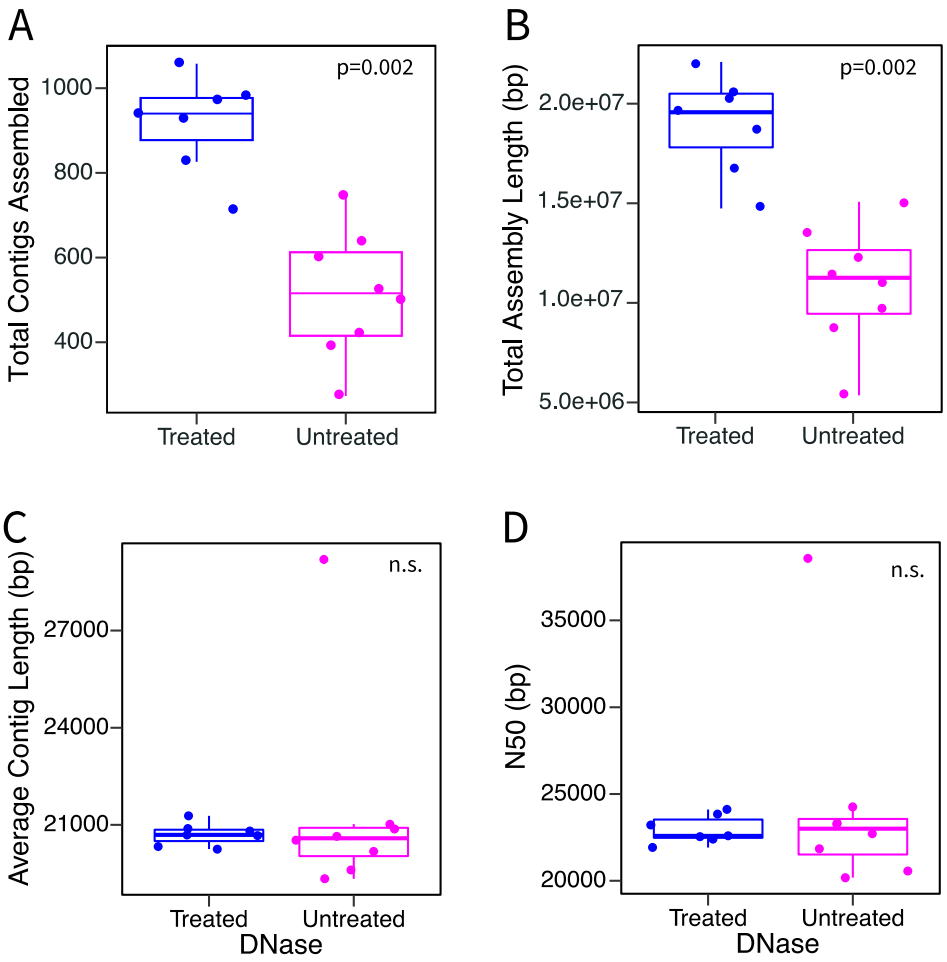
687 51. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-
688 generation sequencing data. Bioinformatics 28:3150–3152.
689 52. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal:
690 Prokaryotic gene recognition and translation initiation site identification. BMC
691 Bioinformatics 11.
692 53. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR,
693 Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic
694 assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing
695 networks. Nat Biotechnol 37:632–639.
696 54. Bushnell B. BBTools. https://sourceforge.net/projects/bbmap/.
697 55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
698 Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics
699 25:2078–2079.
700 56. Lamberton T, Imelfort M. BamM. https://github.com/Ecogenomics/BamM.
701 57. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
702 features. Bioinformatics 26:841–842.
703 58. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek B, Morowitz MJ, Banfield JF. 2020.
704 InStrain enables population genomic analysis from metagenomic data and rigorous
705 detection of identical microbial strains.
706 59. R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna,
707 Austria.
708 60. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR,
709 O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2017. vegan:
710 Community Ecology Package.
711 61. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New
712 York.

713 62. Galili T. 2015. dendextend: An R package for visualizing, adjusting and comparing trees
714 of hierarchical clustering. Bioinformatics 31:3718–3720.
715 63. Bolduc B, Jang H Bin, Doulcier G, You ZQ, Roux S, Sullivan MB. 2017. vConTACT: An
716 iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria.
717 PeerJ 2017:1–26.
718 64. Turner D, Kropinski AM, Adriaenssens EM. 2021. A Roadmap for Genome-Based Phage
719 Taxonomy. Viruses 13:1–10.
720 65. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics,
721 from sampling to analysis. Nat Biotechnol 35:833–844.
722 66. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. 2014. Tackling soil
723 diversity with the assembly of large, complex metagenomes Adina. Proc Natl Acad Sci
724 111:4904–4909.
725 67. Carini P, Delgado-Baquerizo M, Hinckley ELS, Holland-Moritz H, Brewer TE, Rue G,
726 Vanderburgh C, McKnight D, Fierer N. 2020. Unraveling the effects of spatial variability
727 and relic DNA on the temporal dynamics of soil microbial communities. MBio 11:1–16.
728 68. Fierer N. 2017. Embracing the unknown: Disentangling the complexities of the soil
729 microbiome. Nat Rev Microbiol 15:579–590.
730 69. Pavlovska M, Prekrasna I, Parnikoza I, Dykyi E. 2021. Soil sample preservation strategy
731 affects the microbial community structure. Microbes Environ 36:1–7.
732 70. Carroll IM, Ringel-Kulka T, Siddle JP, Klaenhammer TR, Ringel Y. 2012. Characterization
733 of the Fecal Microbiota Using High-Throughput Sequencing Reveals a Stable Microbial
734 Community during Storage. PLoS One 7:1–7.
735 71. Kia E, MacKenzie BW, Middleton D, Lau A, Waite DW, Lewis G, Chan YK, Silvestre M,
736 Cooper GJS, Poppitt SD, Taylor MW. 2016. Integrity of the human faecal microbiota
737 following long-term sample storage. PLoS One 11:1–8.
738 72. Lee KM, Adams M, Klassen JL. 2019. Evaluation of DESS as a storage medium for

739 microbial community analysis. PeerJ 2019.

740

741 **Figures**



742

743 **Figure 1** Assembly comparisons of DNase-treated and untreated viromes. Each point is one
744 virome, with comparisons according to: (A) total contigs assembled, (B) total assembly length,
745 (C) average contig length, and (D) N50 (contig length where half the assembly length is
746 represented in longer contigs and half in shorter contigs). Boxes show the interquartile range
747 and median value. Whiskers extend to the furthest non-outlying datapoint. P-values show the
748 significance of Kruskal Wallis tests between DNase-treated (n=7) and untreated (n=8) samples.
749 Insignificant results (p-values >0.05) are listed as n.s.

750

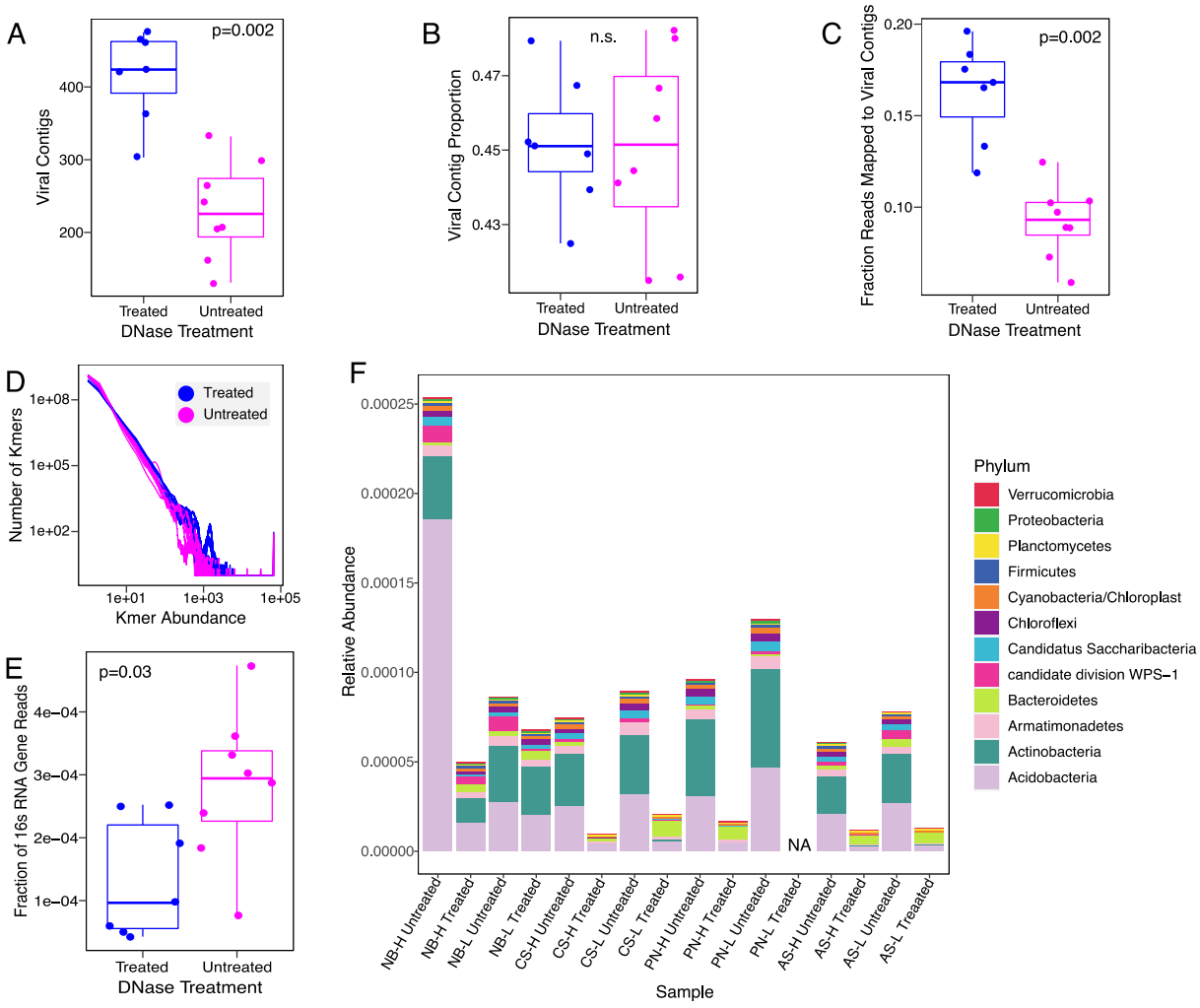
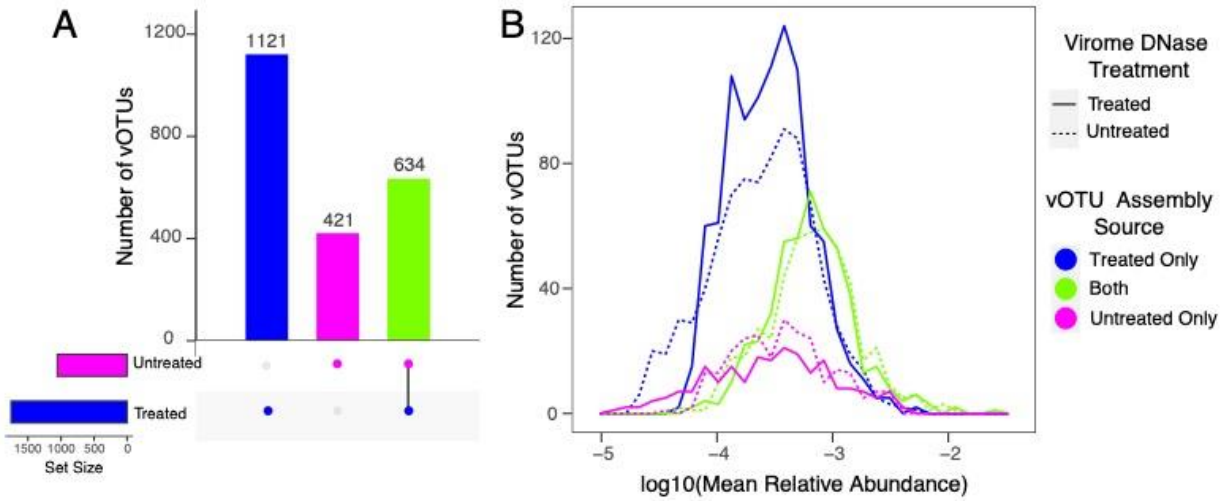


Figure 2 Differences in sequencing content between DNase-treated and untreated viromes. (A) Number of VirSorter identified viral contigs assembled per virome and (B) their proportion of the total number of contigs per virome. (C) The proportion of reads from each sample that mapped to VirSorter identified viral contigs. (D) Frequency plot of kmers showing kmer abundance on the x axis, and the number of kmers with that abundance on the y-axis; each line is one virome. (E) Proportion of reads that contain partial 16S rRNA gene sequences as identified via SortMeRNA. (F) Relative abundances of the top 12 most abundant phyla according to partial 16S rRNA gene sequences. The y-axis displays the number of reads containing 16S rRNA gene fragments from each of the top 12 phyla as a proportion of the total number of quality-trimmed

761 reads in each virome. DNase treated and untreated viromes from the same plot are placed next
762 to each other for ease of comparison. NA indicates no data. For all boxplots (panels A, B, C,
763 and E), boxes show the interquartile range and median value with whiskers extending to the
764 furthest non-outlying datapoint, and P-values show the significance of Kruskal Wallis tests
765 between DNase-treated (n=7) and untreated (n=8) viromes. Insignificant results (p-values
766 >0.05) are displayed as n.s. (not significant).
767
768



769

770 **Figure 3** Numbers and relative abundances of vOTUs, according to the origin(s) of the
771 assembled contigs contained within each vOTU. (A) Number of vOTUs that contained contigs
772 clustered at 95% identity assembled from: DNase treated only (blue), untreated only (pink), or
773 both (green) types of viromes. (B) Distribution of vOTU mean relative abundances across
774 viromes within each DNase treatment group, colored according to the assembly source(s) of
775 viral contigs within each vOTU. Relative abundances are derived from read mapping, such that
776 vOTUs with contigs solely assembled from one treatment could have been detected in viromes
777 from the other treatment via read recruitment.

778

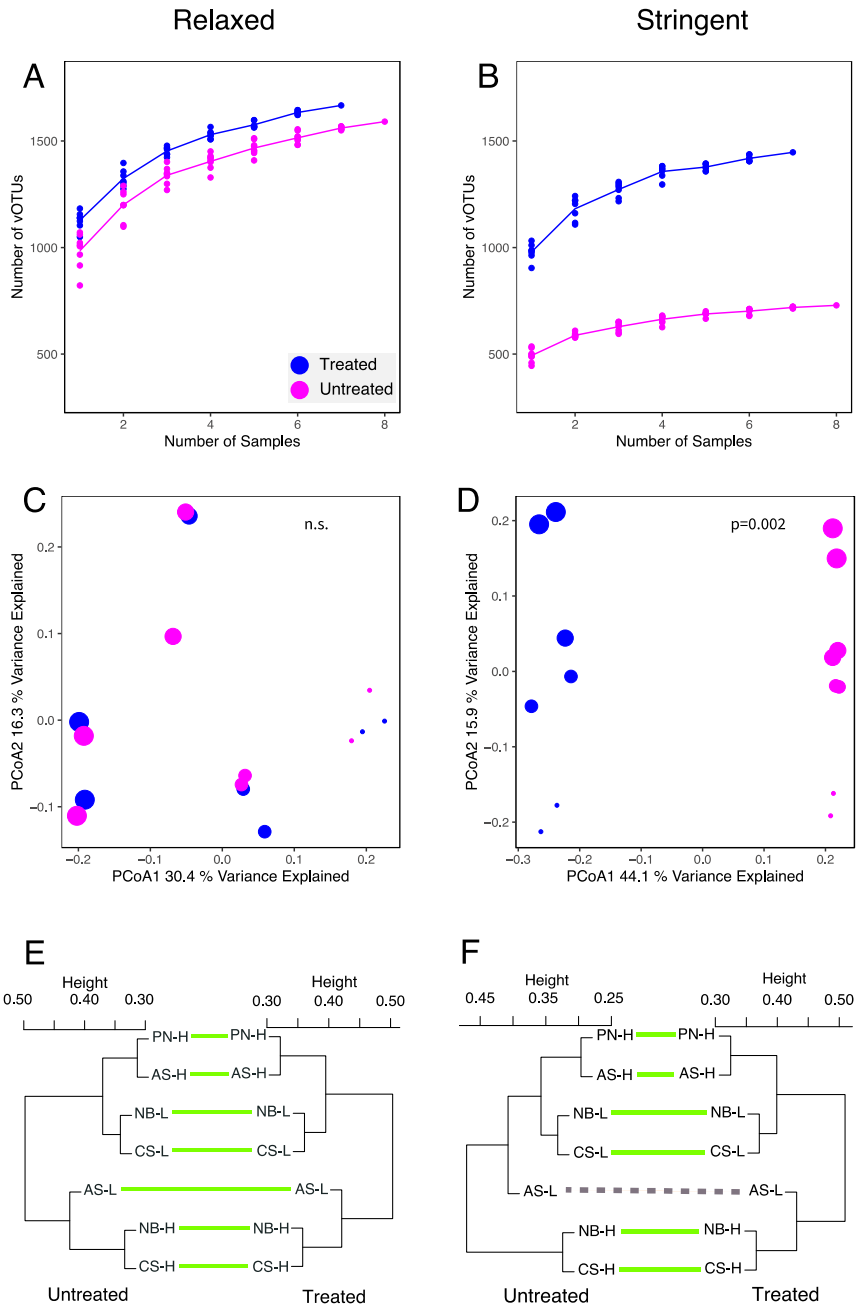


Figure 4 Comparisons of ecological properties across DNase treatment and vOTU detection criteria. (Left Panels) Results from the "Relaxed" vOTU detection criteria and (Right) results from the "Stringent" vOTU detection criteria ("Relaxed" detection allowed for read mapping to all vOTUs in the dataset, and "Stringent" detection allowed for mapping only to vOTUs derived from the same virome treatment group). (A&B) Accumulation curves showing the total number

785 of vOTUs detected within a DNase treatment group at different numbers of samples. (C&D)
786 Principal coordinates analyses of Bray-Curtis dissimilarities (each point is one virome), labeled
787 by DNase treatment (color) and location along the E-W axis of the sampled field (shape size,
788 largest symbols correspond to locations farthest East, with decreasing size along the E-W axis).
789 P-values show significance of DNase treatment on community structure using PerMANOVA
790 (E&F) Tanglegrams, each linking two sets of hierarchical clusters of viral community
791 composition (one per DNase treatment group). Green lines connect samples with congruent
792 clustering between the two treatment groups, and dashed lines connect samples with
793 discongruent clustering. Numbers in the middle of each tanglegram correspond to the plot's
794 location along the E-W axis of the sample field. Dendrograms were created using complete
795 linkage clustering with Bray-Curtis dissimilarities. In panels E and F, the untreated virome from
796 plot PN-L was removed, as it did not have a paired DNase-treated virome. As a result, there
797 was only one virome per treatment (from plot AS-L) at that particular E-W location within the
798 field; all other E-W locations were represented by two viromes per treatment.
799
800