viruses

MDPI

Article

Evaluation of Sequencing Library Preparation Protocols for Viral Metagenomic Analysis from Pristine Aquifer Groundwaters

René Kallies [1,*], Martin Hölzer [2,3], Rodolfo Brizola Toscan [1], Ulisses Nunes da Rocha [1], John Anders [1,4], Manja Marz [2,3,5] and Antonis Chatzinotas [1,5]

[1] Helmholtz Centre for Environmental Research - UFZ, Department of Environmental Microbiology, 04318 Leipzig, Germany; rodolfo.toscan@ufz.de (R.B.T.); ulisses.rocha@ufz.de (U.N.d.R.); johnanders@posteo.de (J.A.); antonis.chatzinotas@ufz.de (A.C.)

[2] Friedrich Schiller University Jena, RNA Bioinformatics and High-Throughput Analysis, 07743 Jena, Germany; martin.hoelzer@uni-jena.de (M.H.); manja@uni-jena.de (M.M.)

[3] European Virus Bioinformatics Center, 07743 Jena, Germany

[4] Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University Leipzig, 04081 Leipzig, Germany

[5] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany

* Correspondence: rene.kallies@ufz.de; Tel.: +49-(0341)-235-1375

Received: 31 March 2019; Accepted: 27 May 2019; Published: 28 May 2019

Abstract: Viral ecology of terrestrial habitats is yet-to be extensively explored, in particular the terrestrial subsurface. One problem in obtaining viral sequences from groundwater aquifer samples is the relatively low amount of virus particles. As a result, the amount of extracted DNA may not be sufficient for direct sequencing of such samples. Here we compared three DNA amplification methods to enrich viral DNA from three pristine limestone aquifer assemblages of the Hainich Critical Zone Exploratory to evaluate potential bias created by the different amplification methods as determined by viral metagenomics. Linker amplification shotgun libraries resulted in lowest redundancy among the sequencing reads and showed the highest diversity, while multiple displacement amplification produced the highest number of contigs with the longest average contig size, suggesting a combination of these two methods is suitable for the successful enrichment of viral DNA from pristine groundwater samples. In total, we identified 27,173, 5,886 and 32,613 viral contigs from the three samples from which 11.92 to 18.65% could be assigned to taxonomy using blast. Among these, members of the *Caudovirales* order were the most abundant group (52.20 to 69.12%) dominated by *Myoviridae* and *Siphoviridae*. Those, and the high number of unknown viral sequences, substantially expand the known virosphere.

Keywords: viral metagenome; groundwater; aquifer; AquaDiva; sequencing library preparation

1. Introduction

Groundwater systems are important compartments of the global hydrological cycle. They donate about 30% of all freshwater sources [1] and provide important ecosystem services. For example, purification and storage of water, active biodegradation of anthropogenic contaminants and nutrient recycling [2]. Many of these services are directly linked to the presence of microorganisms [2,3]. Studies in particular in marine systems have significantly contributed to a better understanding of viruses and their impacts on the mortality, diversity and genetic landscape of their microbial hosts [4–6]. However, only recently, and only in a limited number of surveys, has the potential role of viruses been explored in terrestrial subsurface systems [7–11].

In theory, metagenomics enables the identification and genomic characterisation of all (micro)organisms present in a sample, including viruses [12]. However, the proportion of viral sequences within a metagenome is usually far lower than for other organisms, leading to limitations in their detection. Especially, pristine aquifers are characterised by low microbial biomass and low abundances of virus particles [9,13,14], which might make their detection even more difficult. Size filtration or density-based enrichment methods are therefore widely used to concentrate virus particles from environmental samples [15,16]. However, a significant obstacle in applying metagenomics for pristine aquifers is the still too low amount of DNA required for the direct sequencing of such samples, making amplification techniques mandatory to further enrich viral nucleic acids. It is however widely known that DNA amplification is a source of bias that may lead to inaccurate conclusions after sequence analysis [17]. Three amplification techniques are commonly used to enrich low amounts of DNA [17,18], i.e., (i) linker amplification shotgun libraries (LASL) [19,20]; (ii) sequence-independent, single-primer amplification (SISPA) [21,22]; and (iii) multiple displacement amplification (MDA) [23,24]. Each method has its own potential source of bias. LASL relies on DNA fragmentation and subsequent linker ligation to blunt-end repaired DNA molecules prior to amplification, using primer oligos that bind to the linker sequences [19,25]. Linker ligation efficiency might be one source of bias, especially for very low amounts of DNA [26]. However, previous studies demonstrated that as little as a few pg to ng of DNA is sufficient for low amplification biases [26,27]. LASL may, in addition, be inefficient in recovering ssDNA viruses due to the double-stranded nature of linker DNA molecules [28] though this has recently been overcome with an adapted LASL protocol [29]. SISPA is built upon the use of pseudo-degenerated primer oligonucleotides, containing a stretch of random nucleotides at their 3′-end and a defined sequence at their 5′-end [21], and has successfully been applied to recover both RNA and DNA virus sequences [22,30]. It has, however, been reported that SISPA has a strong amplification bias resulting in an uneven sequencing read distribution and hence overrepresentation of some genome parts while other parts were completely uncovered. In addition, SISPA negatively affects the detection of low abundant genomes [31]. MDA works under isothermal conditions [32] with very low amounts of input DNA, random hexamer primer oligonucleotides and high fidelity as well as strand displacement functions of the phi29 polymerase [23]. Several sources of bias have been identified for phi29 amplification, including chimera formation [33], discontinuous amplification of linear DNA molecules [34] and preferential amplification of circular ssDNA molecules [35]. Recent studies evaluated different library preparation protocols using low input-DNA to assess the reconstruction of microbial communities from metagenomes [36,37]. Similar studies have been performed for the identification of virus sequences from, for example, seawater and human samples [17,35]. Despite these advances, to our knowledge no study has to date assessed and benchmarked sequence library preparation protocols for the identification of viral sequences from pristine aquifer groundwaters.

The Hainich Critical Zone Exploratory (Hainich CZE) in central Germany is an infrastructure designed to, among others, investigate the diversity, identity and abundance of microorganisms in the Hainich aquifers. In addition, analysis of metabolic potential and activities of microorganisms will be linked to physico-chemical parameters in spatial and temporal scales [38]. Here we sampled three carbonate-rock aquifer assemblages of the Hainich CZE, which represent a pristine and uncontaminated aquifer [38]. One problem in obtaining viral sequences from groundwater samples is the low amount of DNA (usually a pico- to few nanograms) that was extracted from isolated virus particles. The aim of this project is therefore two-fold. The first aim is to evaluate different DNA amplification techniques that may offer a sufficient amount of DNA for high-throughput sequencing. In addition, these methods should have a low amplification bias to reflect the natural diversity of the analysed samples. The second aim consisted of evaluating different viral sequence recovery tools to provide first insights on which viruses are present in the Hainich CZE groundwater aquifers.

2. Materials and Methods

2.1. Sample Collection

Groundwater samples were taken from three Hainich CZE aquifer wells in Thuringia, Germany, within the framework of the Collaborative Research Centre AquaDiva (http://www.aquadiva.uni-jena.de) (CRC 1076) [38]. The sampling site was located in the agriculturally used midslope and footslope regions of the Hainich low-mountain range. The three wells were drilled to depths of 50 m (H53), 65 m (H52) and 88 m (H51). H53 and H52 reflect anoxic conditions while oxic conditions prevailed for H51. A detailed description of hydrochemical and geostructural parameters can be found elsewhere [39,40].

Ten liters of groundwater (with approximately 2.3×10^5 (SD: 1.2×10^4) viral particles per milliliter) were collected from each well during a sampling campaign in May 2015. Water was filtered through 200 nm pore filters using a cross-flow system (Sartorius, Göttingen, Germany). Samples were then enriched for viral particles by filtration through 35 kDa filters using the same system. Approximately 60 mL were retained and further concentrated by ultracentrifugation at 22,000× *g* for 2 h and 4 °C. The viral particle containing pellet was resuspended in 500 µL TM buffer (50 mM Tris HCl, 10 mM Magnesium sulfate at pH 7.5). One volume of chloroform was given to the samples to remove microsized prokaryotes. The upper phase, intended for DNA extraction, was treated with DNase I to remove free DNA.

2.2. DNA Extraction, Library Construction and Sequencing

Viral DNA was extracted as described previously [20]. Viral DNA concentration was determined using the Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) resulting in total DNA amounts of 31.8 ng (H51), 5.4 ng (H52) and 25.9 ng (H53). DNA was divided into four parts to prepare four libraries for each sample. Non-amplified shotgun libraries (NASL): using a Covaris ultrasonicator, DNA was sheared to approx. 350 bp fragments and libraries were prepared with a TruSeq DNA PCR-Free Library Prep Kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Linker amplification shotgun libraries (LASL): DNA was sheared to approx. 350 bp fragments as mentioned above and LASL was performed with a NEBNext Ultra DNA Lib Prep Kit (New England Biolabs, Ipswich, MA, USA) as recommended by the manufacturer including 12 PCR cycles to enrich adaptor-ligated DNA. Single-primer amplification (SISPA): PCR was performed by ten cycles using random octamer primers that were linked to a specific primer sequence followed by amplification using a 1:9 mixture of random octamers and a primer targeting the specific primer sequence as described previously [41]. Multiple displacement amplification (MDA): DNA was subjected to phi29 amplification at 25 °C for 8 h using the illustra GenomiPhi V2 DNA Amplification Kit (Thermo Fisher Scientific) as described in the manual. PCR amplicons for the latter two libraries were purified using the Sureclean reagent (Bioline, Luckenwalde, Germany), fragmented as described above and libraries were prepared as described for NASL. Sequencing was performed on one lane of an Illumina HiSeq 2500 system to generate 100-bp paired-end reads.

2.3. Sequencing Read Processing and Assembly

PhiX contaminants were removed, SISPA primer sequences were clipped and raw sequencing reads were quality checked using Trimmomatic [42] and low-quality bases were trimmed from both ends. Reads were screened with a 4-base wide sliding window until the remaining sequences had a Phred-score of at least 15 and a minimum length of 36 nt. Sequencing read redundancy was identified by clustering at 90% sequencing read identity using CD-hit v.4.6 [43,44].

Sequencing reads were independently assembled for each sampling site and library preparation using metaSPAdes [45,46] and SOAPdenovo-Trans [47]. In addition, cross-assemblies were performed for each sampling site including all reads from LASL, SISPA and MDA libraries. We used the transcriptome assembler SOAPdenovo-Trans in addition to SPAdes because recent analyses revealed this assembly tool as very efficient for the assembly of RNA virus genomes [48]. Further analyses suggested this might be also true for the assembly of DNA virus genomes.

2.4. Viral Contig Recovery

Three viral sequence identification tools were used to recover viral contigs, i.e., VirSorter [49], VirFinder [50] and VrAP (https://www.rna.uni-jena.de/research/software/vrap-viral-assembly-pipeline/). VirSorter is based on the identification of viral hallmark genes, enrichment in hypothetical proteins and other viral signatures [49]. Only contigs identified as VirSorter categories 1 and 2 (higher confidence predictions) were retained for further analysis. VirFinder is a kmer based tool for the identification of viral contigs from metagenomes with improvements especially for the detection of short viral contigs [50]. Contigs with a p-value < 0.01 were used for further analysis. These two detection tools were completed by using VrAP, a novel de novo genome assembly pipeline especially designed for viruses. The pipeline is able to assemble complete genomes of viruses representing new strains and species, as well as prototypes of new genera and families. VrAP is based on the genome assembler SPAdes [45] combined with an additional read correction [51,52] and several filter steps. The pipeline classifies the contigs to distinguish host from viral sequences by annotation and open reading frame (ORF) density scores. By applying the ORF density method we were able to identify potential novel viruses without any sequence homology to known references (manuscript in preparation).

2.5. Virome Diversity Measures and Comparison of Library Preparation Methods

Nonpareil [53–55] was used with default settings to estimate diversity and coverage of virome datasets. Viral reads present in one or more datasets reflecting LASL, SISPA and MDA per sampling site were identified as follows. Redundancy was removed for each dataset by CD-hit-est clustering at 95% identity. A database was created containing all viral contigs and, using Bowtie2 [56], read cluster per library preparation method and sampling site were mapped to the database. Mapped clusters were extracted, counted and overlapping information were generated using SAMtools [57]. Viral contigs were compared between sites by an all-versus-all clustering approach (95% identity) with CD-hit-est-2D [44].

Venn diagrams were computed in R [58] using the package "venneuler" (https://cran.r-project.org/web/packages/venneuler/index.html).

2.6. Viral Taxonomic Assignment

All viral contigs per sampling site, i.e., contigs identified from all virus identification tools and library preparation methods, were combined (resulting in three datasets) and redundancies were removed by clustering with CD-hit-est at 95% nt identity. Open reading frames (ORFs) were translated from these contigs using prodigal [59] and aligned to a viral RefSeq protein database (February 2019) using DELTA-BLAST [60] with an e-value cut off of 10^{-3}. Hits were sorted by e-value and bit score and ORFs with most significant hits were aligned to the respective contigs using an in-house python script (Supplementary Information), resulting in one hit per contig. Gene sharing networks based on shared protein clusters (PCs) between viral genomes were calculated with vConTACT2 [61,62] on the iVirus platform [63] and were displayed with Cytoscape [64]. DNA contamination from cellular organisms was determined using EMIRGE [65].

2.7. Data Availability

Sequence read raw data have been made available at Sequence Read Archive accession: PRJNA530103.

3. Results

3.1. Raw Sequencing Output Statistics

The first aim of this study was to evaluate different DNA amplification techniques that may result in a sufficient amount of viral DNA for high throughput sequencing. We therefore compared three DNA amplification methods, i.e., LASL, SISPA and MDA. NASL was used as control.

MDA produced highest (quality trimmed) sequencing read numbers followed by SISPA and LASL as compared to NASL that exhibited lowest read numbers (Table S1). Significant differences (ANOVA) in quality-trimmed sequencing output were observed between NASL-MDA, NASL-SISPA and SISPA-MDA (Table 1).

Table 1. *P*-values of analysis of variance (ANOVA) of raw sequencing read and read cluster numbers between the different library preparation methods.

		Number of Raw Reads		
Library Preparation	NASL	LASL	SISPA	MDA
NASL	n/a	>0.05	0.008	0.002
LASL		n/a	>0.05	0.023
SISPA			n/a	>0.05
MDA				n/a

				Clusters at 90% Read Identity				
	Relative proportion				Number of clusters			
Library Preparation	NASL	LASL	SISPA	MDA	NASL	LASL	SISPA	MDA
NASL	n/a	<0.001	>0.05	>0.05	n/a	0.018	>0.05	0.008
LASL		n/a	<0.001	<0.001		n/a	>0.05	>0.05
SISPA			n/a	>0.05			n/a	>0.05
MDA				n/a				n/a

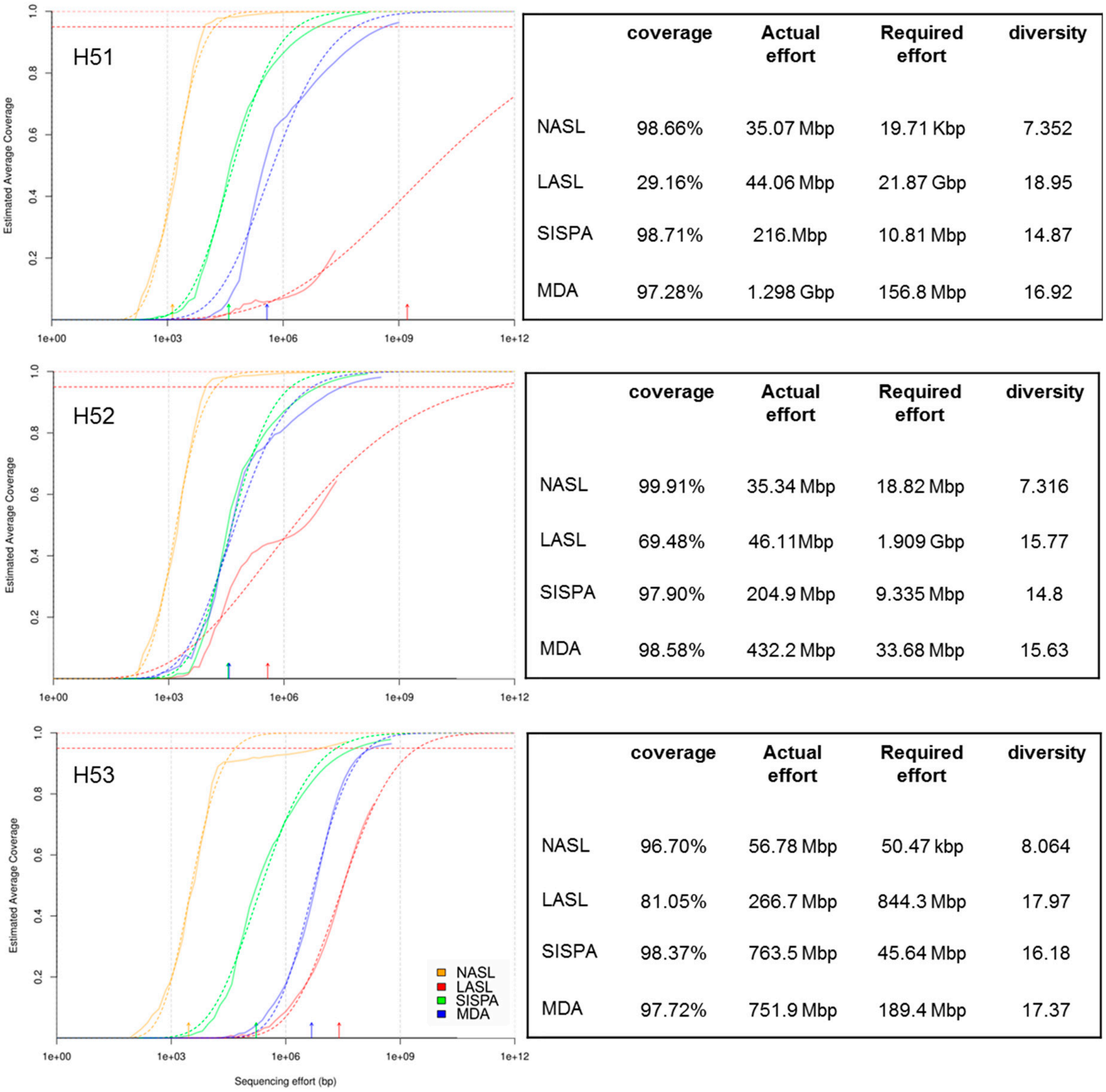
NASL: non-amplified shotgun library; LASL: linker amplification shotgun libraries; SISPA: single-primer amplification; MDA: multiple displacement amplification.

Read quality of all libraries was >97% except for libraries H51 LASL and H52 LASL for which 33.78% and 28.13% of the reads were discarded after quality trimming. However, no significant differences (ANOVA) in quality between any of the library preparation methods was observed.

PCR amplification bias may influence the evenness among sequencing reads. For example, GC-rich primers and primers with GC-stretches at their 3′-end, both present in a random primer mix, may anneal more efficiently to a target sequence than AT-rich primer oligos do. As a result, amplicons amplified from such target sequences may be favored during the amplification process what in turn leads to high numbers of identical or related DNA molecules. We therefore clustered all quality-trimmed sequencing reads with a 90% cut-off to remove this redundancy. LASL libraries produced the lowest redundancy (41.7 to 60.7% relative proportion of clusters to sequencing read numbers), with significant differences not only to non-amplified libraries (10.8 to 21.3 relative proportion of clusters to sequencing read numbers) but MDA libraries (9.1 to 17.0% relative proportion of clusters to sequencing read numbers) and SISPA libraries (5.9 to 7.4 relative proportion of clusters to sequencing read numbers) (Table 1). These data suggest an amplification bias during PCR with random primer oligomers. However, MDA libraries (together with LASL libraries) still resulted highest average numbers of read clusters (Table S1). The presence of many repetitive and homopolymeric sequencing reads (possibly sequencing artefacts) may explain the low proportion of clustered reads in NASL libraries.

3.2. Data Set Comparison

We used Nonpareil, i.e., a kmer based approach that examines the degree of overlap among individual sequence reads, to determine redundancy [53–55] among the individual reads to further assess the average coverage created from the different library data sets. NASL, SISPA and MDA libraries seem to reach a nearly full coverage while LASL libraries vary between ~20 to 80% coverage (Figure 1). However, diversity among libraries increased from NASL to SISPA, MDA and LASL, the latter being the most diverse libraries (Figure 1). These results strongly indicate the target discrimination of SISPA and MDA during PCR that results in uneven coverage of the viral metagenomes and in addition, may fail to target low abundant sequences. NASL sequencing reads dominantly consisted of repetitive and homopolymeric sequences (see also below), with most likely too low an input of DNA explaining the observed Nonpareil curve for these libraries.



	coverage	Actual effort	Required effort	diversity
NASL	98.66%	35.07 Mbp	19.71 Kbp	7.352
LASL	29.16%	44.06 Mbp	21.87 Gbp	18.95
SISPA	98.71%	216.Mbp	10.81 Mbp	14.87
MDA	97.28%	1.298 Gbp	156.8 Mbp	16.92

	coverage	Actual effort	Required effort	diversity
NASL	99.91%	35.34 Mbp	18.82 Mbp	7.316
LASL	69.48%	46.11Mbp	1.909 Gbp	15.77
SISPA	97.90%	204.9 Mbp	9.335 Mbp	14.8
MDA	98.58%	432.2 Mbp	33.68 Mbp	15.63

	coverage	Actual effort	Required effort	diversity
NASL	96.70%	56.78 Mbp	50.47 kbp	8.064
LASL	81.05%	266.7 Mbp	844.3 Mbp	17.97
SISPA	98.37%	763.5 Mbp	45.64 Mbp	16.18
MDA	97.72%	751.9 Mbp	189.4 Mbp	17.37

Figure 1. Comparison of Hainich groundwater viromes diversity and coverage as function of sequencing effort using Nonpareil curves [53–55]. Estimated coverage is shown as dotted lines, true coverage as solid lines. Estimated diversity is shown with arrows on the *x*-axis. Horizontal dotted line shows 95% coverage. Viral metagenome coverage, actual sequencing effort, required sequencing effort and kmer-based diversity for each library are shown in the right panel.

We were further interested in both, the number of viral reads that were exclusively detected by one of the library preparation methods and those reads that were identified from more than one library preparation method. For this, redundancy removed reads (i.e., reads that clustered at 90% identity) of LASL, SISPA and MDA libraries were independently mapped to viral contigs per individual sampling site (i.e., all viral contigs that were identified by the three virus identification tools and cross-assemblies) and counted. MDA libraries produced most reads (average: 350 k) followed by LASL (average: 143 k). Least reads were identified from SISPA libraries (average: 64 k). Overlapping information (reads found in more than one library) was rather low with 0.27 to 4.64% of reads present in all three libraries while 0.07 to 14.77% of reads were identified by two libraries (Figure 2). These data indicate target sequence discrimination between each of the library preparation methods.

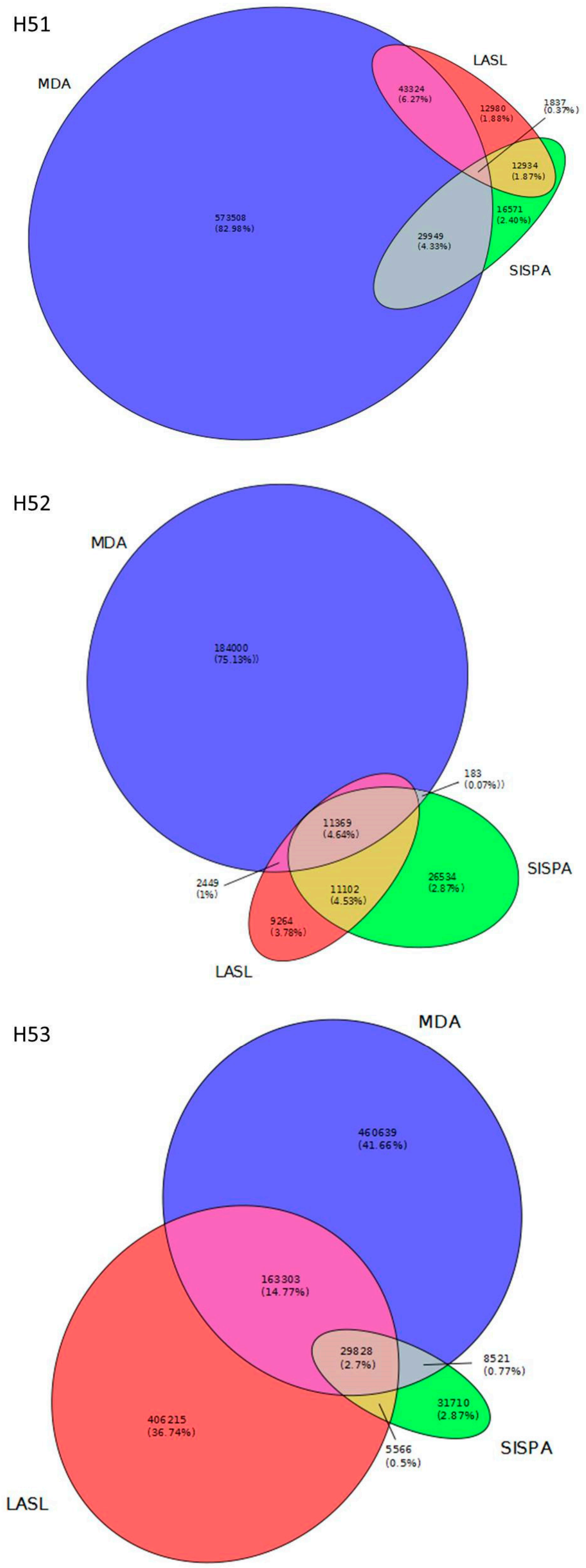


Figure 2. Overlap of sequencing read cluster (90% identity) information identified by library preparation methods, independently shown for each sampling site. Non amplified sequencing libraries (NASL) were not included in the analysis due to the homopolymeric and repetitive nature of sequences obtained from these libraries.

3.3. Assembly Statistics and Evaluation of Viral Identification Tools

Using both, SPAdes and SOAPdenovo-Trans, assemblies from non-amplified libraries completely failed due to repetitive and homopolymeric sequences. We therefore excluded these datasets from further analysis. Contig numbers tend to be higher for LASL and MDA libraries than for SISPA libraries (LASL-SISPA: $p = 0.062$, SISPA-MDA: $p = 0.087$; statistical test: one-way ANOVA) when assembled with SPAdes. Similar results were observed for SOAPdenovo-Trans assemblies (LASL-SISPA: $p = 0.052$, SISPA-MDA: $p = 0.028$; statistical test: one-way ANOVA). In addition, MDA library assemblies produced longer contigs (N50) when compared to LASL and SISPA ($p = 0.001$ (SPAdes), $p < 0.001$ (SOAPdenovo-Trans); statistical test: one-way ANOVA) (Figure 3, Table S2). A comparison (student's t-test) of the two assembly tools showed that SOAPdenovo-Trans may tend to produce longer contigs ($p = 0.084$), while there is no significant difference in the average contig size (N50) ($p = 0.2972$).

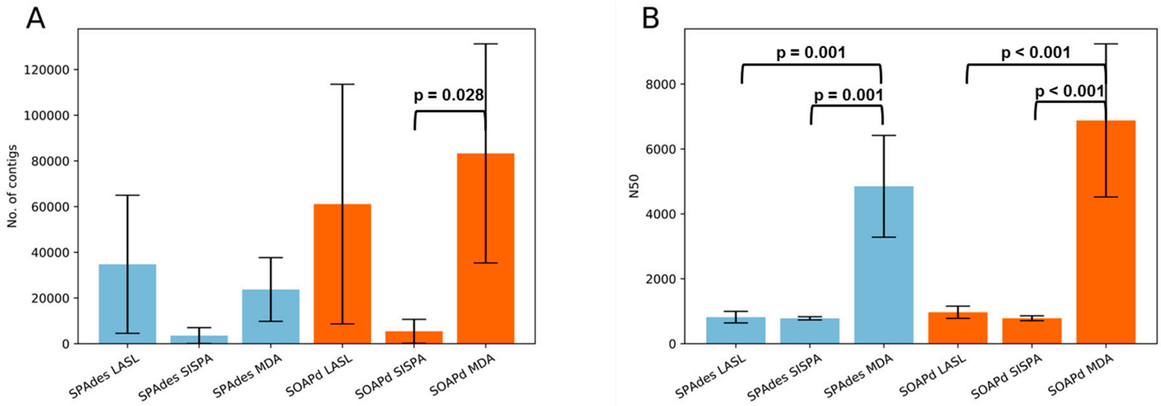


Figure 3. Number of contigs (**A**) and N50 (**B**) produced by sequence library preparation methods and assembly tools. Differences between library preparation methods were tested using analysis of variance (ANOVA). SOAPd: SOAPdenovo-Trans, NASL: non-amplified shotgun library, LASL: linker amplification shotgun libraries, SISPA: single-primer amplification, MDA: multiple displacement amplification.

Viral contigs (as identified by VirSorter, VirFinder and VrAP obtained from cross-assemblies) were clustered at 95% identity to determine a core set of sequences among the sampling sites. Only 37 contigs (0.5%) were shared by the three viromes indicating there is at least a minor common core set in the groundwater aquifers. The amount of shared contigs increased from 0.85% (H51 and H52) and 1.04% (H52 and H53) to 2.85% (H51 and H53) when two viromes were compared. However, the majority of viral contigs is exclusive for the respective virome (Figure 4). The overall viral contig number from H52 is rather low compared to H51 and H53, most likely due to the lower amount of DNA extracted from this sample. This might explain the lower contig overlap of H52 with H51 and H53, respectively, than the overlap of H51 and H53.

We used three different viral sequence identification tools that are based on the detection of viral hallmark genes (VirSorter), kmer distribution (VirFinder) and orf density (VrAP) (see more detailed description in the Materials and Methods section). VirFinder and VrAP significantly identified a higher number of viral contigs than VirSorter (One-way ANOVA $p < 0.001$). The size of viral contigs obtained by VirSorter and VirFinder were in contrast significantly higher than for VrAP (one-way ANOVA $p < 0.05$). However, each tool identified viral contigs that were not recognized by the other two revealing an advantage in the use of several identification tools for the recovery of viral sequences.

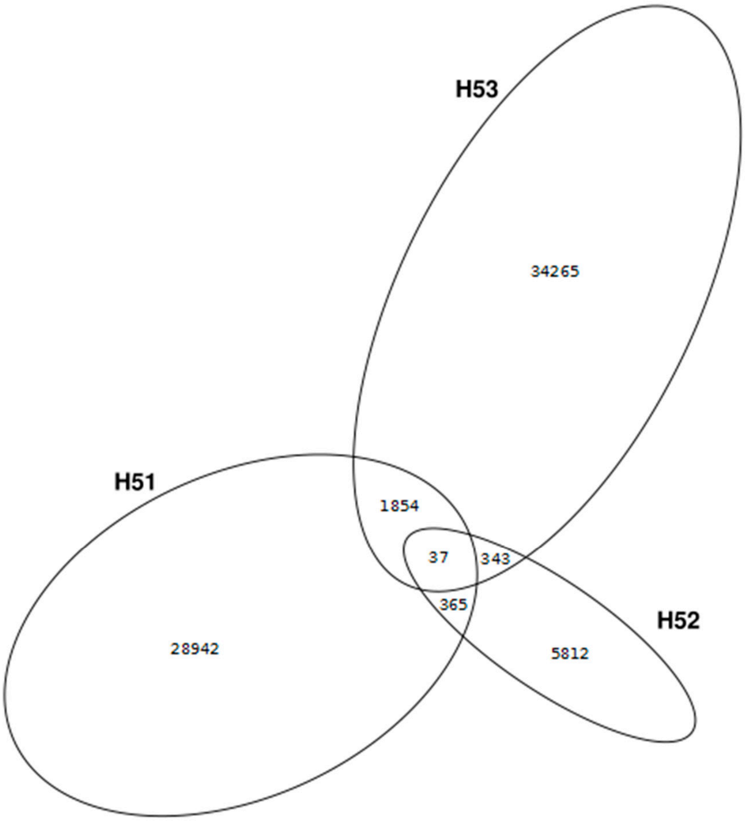


Figure 4. The venn diagram presents numbers of unique and shared viral contigs among the different viromes. Cross-assembled viral contigs (as identified by VirSorter, VirFinder and VrAP) were compared between sites by an all-versus-all clustering approach (95% identity) with CD-hit-est-2D [44].

3.4. First Insights into Viral Taxonomic Composition of Hainich Groundwater

Using cross-assembled contigs (assemblies including sequencing reads from LASL, SISPA and MDA per sampling site) and a set of three viral sequence recovery tools, we identified 27,173 (H51), 5,886 (H52) and 32,613 (H53) viral contigs from the Hainich groundwater samples (Figure 3; Table S3). These contigs were assembled from 31.19% (H51), 52.08% (H52) and 28.41% (H53) of the quality trimmed sequence reads. Among these reads we identified 19 Small subunit ribosomal ribonucleic acid sequences (8 bacterial 16S, 11 unclassified) demonstrating a low contamination with DNA from cellular organisms. Only 14.81% (H51), 18,65% (H52) and 11.92% (H53) of the viral populations could be assigned to taxonomy using delta-blast (Figure 5A). Most of them were assigned to dsDNA viruses dominated by the order *Caudovirales* (H51: 69.12%, H52: 58.20%, H53: 62.40%). Within the *Caudovirales,* members of the *Myoviridae* (40 to 46.5%) and *Siphoviridae* (40.7 to 41.9%) families were most abundant (Figure 5B, Figure S1). These findings are not surprising since *Caudovirales* have previously been presented as the most abundant group of viruses in environmental ecosystems [8,66,67]. Other identified dsDNA virus sequences belonged, for example, to the amoeba infecting giant virus families *Marseilleviridae* and *Mimiviridae*, to the algae infecting *Phycodnaviridae* family whose hosts has been shown to be present in groundwater [68], and invertebrate-infecting viruses such as *Iridoviridae* and *Poxviridae* (Figure 5B). Surprisingly, we identified only a small number of circular ssDNA viruses (Figure 5B). These viruses have been revealed as an abundant group in other environments [69,70]. We used Phi29 polymerase in MDA that preferentially amplifies circular ssDNA [35] and one could expect a bias towards overrepresentation of circular ssDNA genomes. Although this study is only a first snapshot into the Hainich groundwater virome we speculate that circular ssDNA viruses are rare in this environment. A small fraction of these DNA viromes was assigned to RNA viruses, most likely due to PCR errors and incomplete/erroneous virus reference databases.

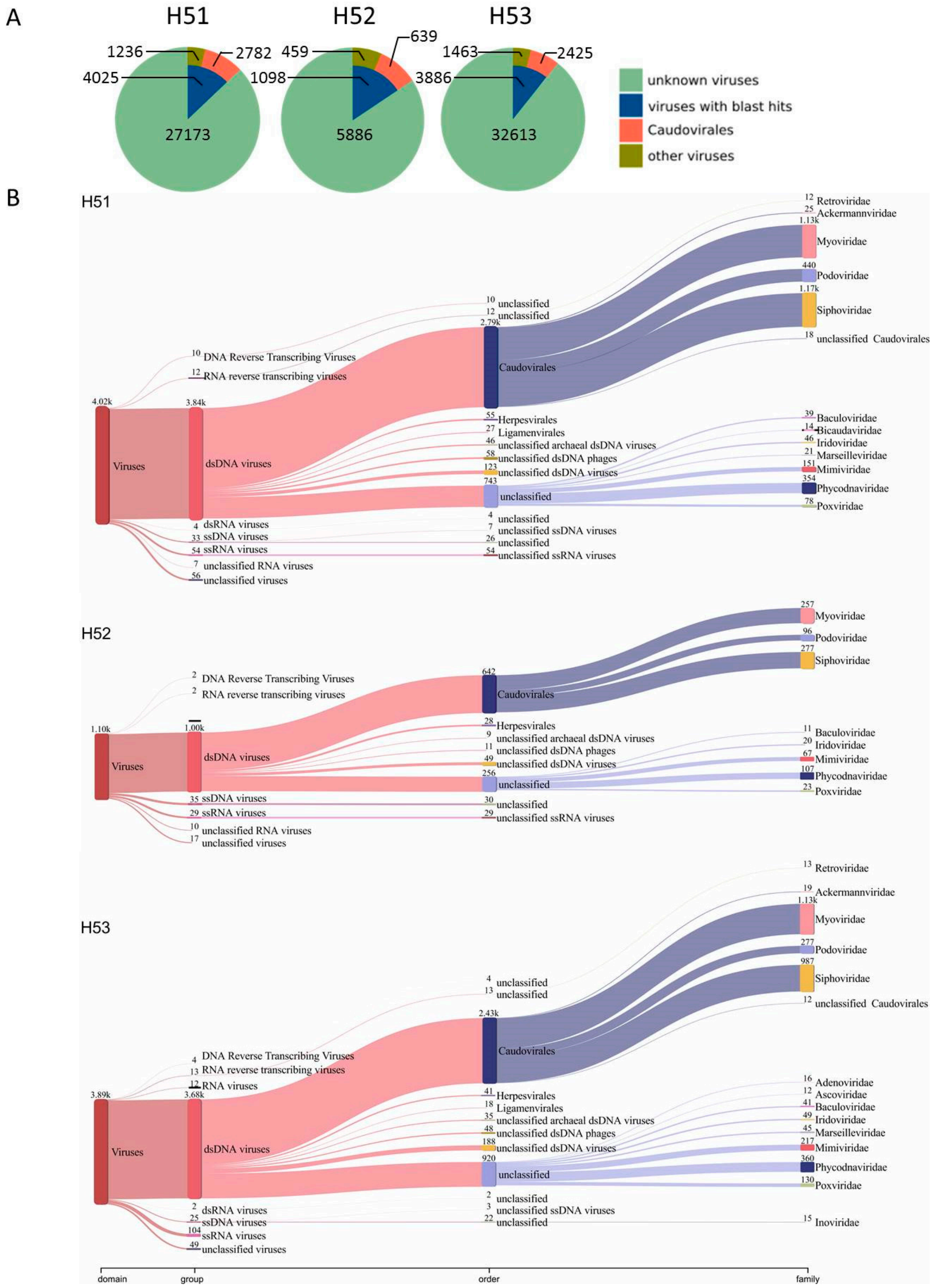


Figure 5. Taxonomic assignment of viral contigs identified from cross-assemblies. (**A**) pie charts present relative and absolute abundance of viral contigs after blastp analysis. (**B**) Taxonomic profile of viral contigs as classified by blastp (viral contigs with blast hits in figure **A**). Data were visualized with Pavian [71].

However, a high number of blast-based taxonomy assigned contigs could not be affiliated to deeper taxonomic levels but have similarity to unclassified viruses present in the viral RefSeq database (Figure 5B). These findings, together with the huge number of unknown viral contigs (without any blast hit) reveal substantial genomic and taxonomic diversity in Hainich groundwater viromes, as observed also in other environments [66,72]. To further investigate the similarity of Hainich groundwater viromes to viral RefSeq database, we used a genome-based network analysis of their shared protein content (Figure 6) [61,62]. This analysis groups viral contigs at the approximately genus level into viral clusters [61,62,73]. In total, 539 viral clusters were identified. Of those, Hainich viral contigs were found in 191 clusters, 183 of them were exclusive to Hainich viromes, among those 95 clusters exclusive to H51, 8 clusters to H52 and 23 clusters to H53. In addition, approximately 34% (H51), 64% (H52) and 63% (H53) of viral protein clusters were present in at least one other Hainich groundwater sample, suggesting some sequence conservation across these samples.

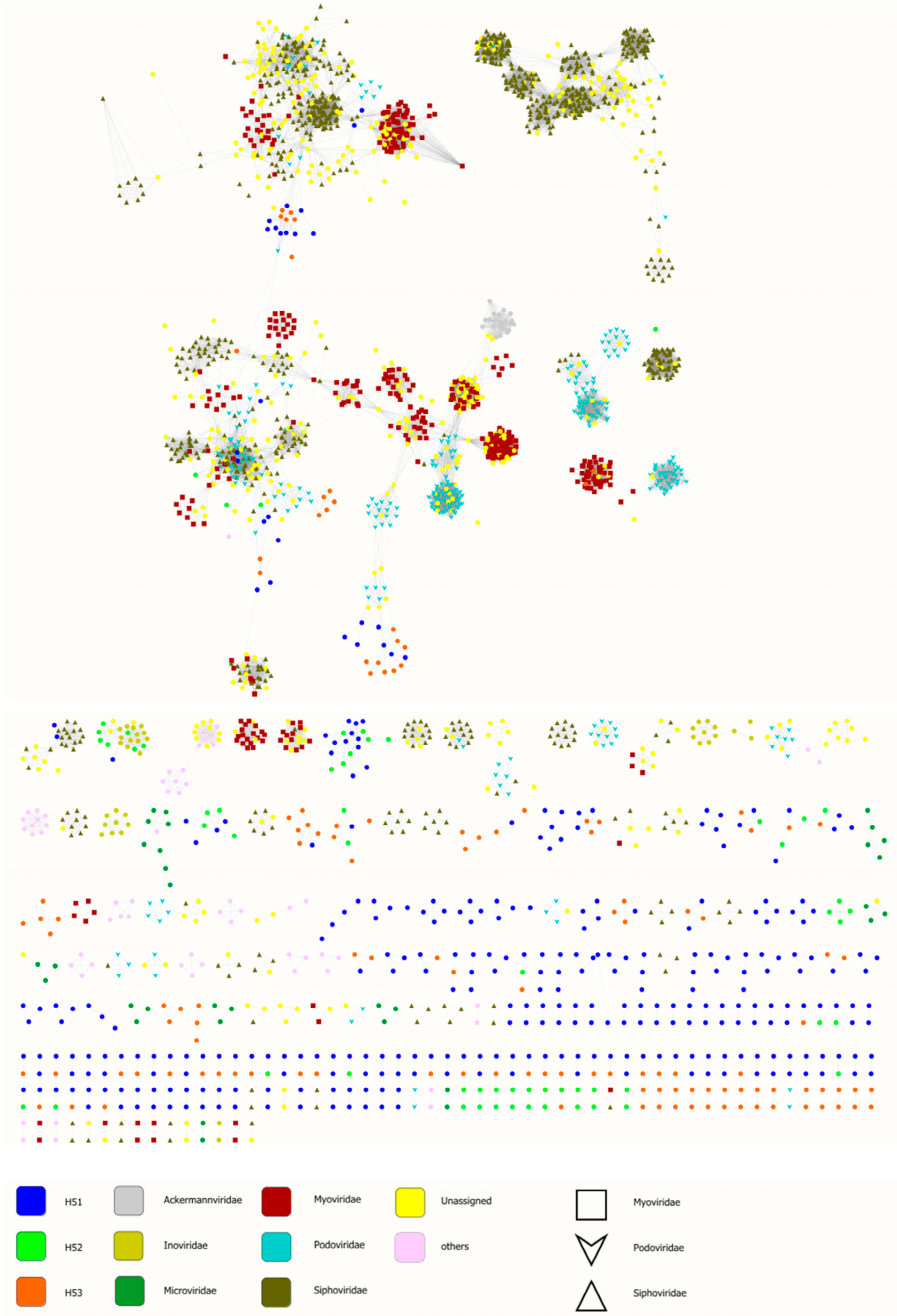


Figure 6. A network analysis of shared predicted protein content between viral RefSeq database and Hainich viral populations. Nodes (circles) indicate contigs and shared edges (lines) indicate shared protein content. Data were analysed using vConTACT2 [61,62] and displayed with cytoscape [64].

4. Discussion

Viruses play a key role in ecosystems, with most of them infecting microbes. They directly affect their hosts by lysis and horizontal gene transfer, and hence are responsible for changes in microbial community structure and composition what in turn has consequences on biogeochemical cycles and food web structures [4–6,74]. Viral metagenomics has been increasingly used to unravel viral community composition and interactions with their hosts from different ecosystems, such as marine environments and soil [66,67,72]. The terrestrial subsurface including groundwater ecosystems is at present yet underexplored [7,8,10,11,75]. A common problem is the relatively low biomass present in these difficult to obtain samples, which in return, results in only low amounts of DNA not sufficient for standard preparation of metagenome sequencing libraries [9,13,14]. Efforts have been undertaken to overcome this problem, including DNA enrichment using different DNA amplification techniques [17–19,22]. Each of these methods has its own advantages and limitations making it difficult to provide a standard protocol. Benchmark tests should therefore be performed when investigating new sample types.

Sampling procedure, virus particle isolation and nucleic acid extraction protocols are potential sources of bias [17] that have to be considered prior to sampling. Here, we focused on non-enveloped DNA viruses that passed a pore size of 200 nm after filtration and performed a benchmark study to find a method of choice to enrich viral DNA that is sufficient for sequencing. We furthermore intended to get a first snapshot of the viruses present in Hainich groundwater aquifers.

We used three DNA amplification methods, i.e., LASL, SISPA and MDA to compare one another and with NASL, using three groundwater samples. Although NASL resulted in some sequencing output none of the reads could be used for further analysis (assembly, virus sequence identification) due to their repetitive and homopolymeric nature; demonstrating that direct sequencing of NASL is not feasible with low DNA amounts. According to the Nonpareil curves, LASL was the method with the lowest amplification bias since the curves were located rightward in the plots indicating a higher diversity than for SISPA and MDA (Figure 1). Nonpareil curves for SISPA and MDA simulate a nearly full sequence coverage that emerge from redundant sequence information (Table 1, Table S1). False sequence coverage interpretation could be a result when data analysis exclusively rely on these library preparation methods. In addition, LASL resulted in the highest number of unique sequencing reads as compared to SISPA and MDA. MDA on the other hand outperformed LASL and SISPA in terms of viral contig numbers and their average contig size. In addition, MDA performed (at least in two samples) much better for taxonomic assignments in the case of *Caudovirales* families, which were dominant among the viral contigs with taxonomic affiliation (Figure S1). Considering the amount of unique viral reads per method and their low overlap (Figure 2), together with the results from cross-assemblies, it became apparent that none of the here tested DNA enrichments methods could completely detect viral sequences from pristine groundwater. However, SISPA even underperformed in terms of sequencing output, diversity and assembly statistics. Metagenomic benchmark studies using both, microbial mock communities and marine samples demonstrated the use of Mondrian and Illumina Nextera XT technologies produced high quality metagenomes from even femtogram-input DNA libraries [36,37]. These library preparation methods are comparable with the LASL protocol used in this study because all these methods use linker ligation on fragmented or tagmented DNA prior to amplification for generation of sequencing libraries. The low bias introduced by LASL on virus enriched groundwater samples from our work is consistent with these previous studies on prokaryotic metagenomes. In addition, other studies on viromes from marine and human samples showed substantial differences with respect to diversity, assembly output, types and ratio of viral sequences between LASL and MDA [18] and an outperformance of MDA over SISPA [17]. However, these studies observed an overrepresentation of circular sequences in MDA libraries as compared to LASL and SISPA. In contrast, our data identified only a few contigs that belong to circular ssDNA viruses (see also discussion below). We therefore suggest the combined use of LASL and MDA for future analysis of viral communities from pristine groundwater aquifers.

SOAPdenovo-Trans produced more contigs than SPAdes. However, average contig size was similar (Figure 3, Table S2). A combination of the assembly output seems to produce most comprehensive results but might also introduce unnecessary redundancy. Assembly for metagenomic data is already difficult, but appear to be more complex for viruses with their possibly more uneven genome coverage. Specialized tools are needed for the (de novo) assembly of viral sequences from metagenomic data [76]. The lower number of contigs for H52 could be a result of the lower amount of DNA extracted from this sample as compared to H51 and H53. Future studies will reveal whether there is a correlation between input DNA amount and contig numbers, including replicates and different yields of DNA input.

There is a high number of virus identification tools available, with all of them having their limitations [77]. We decided to use VirSorter [49], VirFinder [50] and VrAP. The latter two do not rely on database matches, increasing the chance to detect novel viruses not related to those present in public databases. Using our dataset, each tool exclusively identified some viral contigs demonstrating a combination of different virus identification tools increases the number of recovered viral contigs as also suggested previously [59,78]. However, the number of viral contigs was lower than the total number of contigs (compare Table S2 and Table S3). The experimental procedure included several steps to enrich virus particles, i.e., size filtration, chloroform treatment to remove most small-sized bacteria and digestion of free DNA that is not protected by a protein shell. Although some non-viral sequences might still be present after such methodology, one could assume the majority of the dataset consists of viral sequences and consequently includes a high number of viral contigs not recovered by one of the detection tools. Efforts should be undertaken, e.g., using machine learning, to overcome these likely limitations [78,79].

Like in many environmental studies, the taxonomy of most viral contigs remained unknown as demonstrated by blast and network analysis (Figure 5, Figure 6) [8,10,66]. Members of the order *Caudovirales* were dominating among viral contigs with taxonomic assignment. This group of tailed viruses infects a wide variety of bacteria and has been shown as one major group present in environmental ecosystems [8,66,72,80,81]. Another group of commonly highly abundant viruses, i.e., circular ssDNA viruses of the families *Microviridae* and *Circoviridae* [69,70,82], were almost entirely absent in our dataset. This is in contrast to previous results from groundwater aquifers where these viruses even dominated over dsDNA viruses among the classified sequences [10]. A technical bias seems to be unlikely since MDA is known for preferential amplification of these target sequences [35]. Future analyses including spatial and temporal variation will reveal whether these viruses are rare in pristine groundwater. We further identified viruses infecting algae, invertebrates and microeukaryotes, among the latter, contigs similar to giant viruses from the *Mimiviridae* family. These viruses should, by default, not be detected after 200 nm pore size filtration. A possible explanation could be sequence similarity of conserved mimivirus orfs, such as polymerases, to yet unknown viruses [83,84].

We show viral metagenome libraries can be produced from pristine aquifer groundwaters and suggest a combination of LASL and MDA to enrich viral DNA from these samples and to diminish an amplification bias that may occur during enrichment. We further identified new viral sequences that will help to understand the role of viruses in pristine groundwaters.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/11/6/484/s1, Figure S1: Bubble plot shows families of the Caudovirales order present in each virome., Table S1: Overview of raw sequencing read pair numbers, sequence quality and sequencing read clusters for each prepared sequencing library., Table S2: Assembly statistics for LASL, SISPA and MDA libraries per sampling site., Table S3: Overview of identified viral contigs as per virus identification tool, assembly software, sequencing library and sampling site., Supplementary information: Python script to assign orfs to contigs.

Author Contributions: Designed the experiments: R.K., M.H., M.M. and A.C.; performed the experiments: R.K. and M.H.; data analysis: R.K., M.H., R.B.T., U.N.d.R. and J.A.; prepared the manuscript: R.K., M.H., U.N.d.R., M.M. and A.C. All authors read and approved the final manuscript.

Funding: This research was funded by Deutsche Forschungsgemeinschaft, grant number CRC 1076 “AquaDiva” and grant number SPP-1596 “Ecology and species barriers in emerging viral diseases”.

Acknowledgments: We are very grateful to Nicole Steinbach Anett Heidtmann for excellent technical assistance, to Robert Lehmann, Friedrich-Schiller-Universität Jena, who helped with sampling and to Ben Bolduc, Ohio State University, for his help in using vConTACT2. We further thank Ivonne Görlich and Marco Groth from the Core Facility DNA sequencing of the Leibniz Institute on Aging - Fritz Lipmann Institute in Jena for their help with DNA sequencing. MH appreciates the support of the Joachim Herz Foundation by the add-on fellowship for interdisciplinary life science.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Danielopol, D.L.; Pospisil, P.; Rouch, R. Biodiversity in groundwater: A large-scale view. *Trends Ecol. Evol.* **2000**, *15*, 223–224. [CrossRef]
2. Griebler, C.; Avramov, M. Groundwater ecosystem services: A review. *Freshw. Sci.* **2015**, *34*, 355–367. [CrossRef]
3. Griebler, C.; Lueders, T. Microbial biodiversity in groundwater ecosystems. *Freshw. Biol.* **2009**, *54*, 649–677. [CrossRef]
4. Suttle, C.A. Viruses in the sea. *Nature* **2005**, *437*, 356. [CrossRef] [PubMed]
5. Suttle, C.A. Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol.* **2007**, *5*, 801. [CrossRef] [PubMed]
6. Breitbart, M. Marine Viruses: Truth or Dare. *Annu. Rev. Mar. Sci.* **2011**, *4*, 425–448. [CrossRef]
7. Daly, R.A.; Borton, M.A.; Wilkins, M.J.; Hoyt, D.W.; Kountz, D.J.; Wolfe, R.A.; Welch, S.A.; Marcus, D.N.; Trexler, R.V.; MacRae, J.D.; et al. Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nat. Microbiol.* **2016**, *1*, 16146. [CrossRef]
8. Daly, R.A.; Roux, S.; Borton, M.A.; Morgan, D.M.; Johnston, M.D.; Booker, A.E.; Hoyt, D.W.; Meulia, T.; Wolfe, R.A.; Hanson, A.J.; et al. Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat. Microbiol.* **2019**, *4*, 352–361. [CrossRef]
9. Kyle, J.E.; Eydal, H.S.C.; Ferris, F.G.; Pedersen, K. Viruses in granitic groundwater from 69 to 450 m depth of the Äspö hard rock laboratory, Sweden. *ISME J.* **2008**, *2*, 571. [CrossRef]
10. Smith, R.J.; Jeffries, T.C.; Roudnew, B.; Seymour, J.R.; Fitch, A.J.; Simons, K.L.; Speck, P.G.; Newton, K.; Brown, M.H.; Mitchell, J.G. Confined aquifers as viral reservoirs. *Environ. Microbiol. Rep.* **2013**, *5*, 725–730. [CrossRef]
11. Pan, D.; Watson, R.; Wang, D.; Tan, Z.H.; Snow, D.D.; Weber, K.A. Correlation between viral production and carbon mineralization under nitrate-reducing conditions in aquifer sediment. *ISME J.* **2014**, *8*, 1691–1703. [CrossRef]
12. Wooley, J.C.; Ye, Y. Metagenomics: Facts and Artifacts, and Computational Challenges. *J. Comput. Sci. Technol.* **2009**, *25*, 71–81. [CrossRef] [PubMed]
13. Wilhartitz, I.C.; Kirschner, A.K.T.; Brussaard, C.P.D.; Fischer, U.R.; Wieltschnig, C.; Stadler, H.; Farnleitner, A.H. Dynamics of natural prokaryotes, viruses, and heterotrophic nanoflagellates in alpine karstic groundwater. *Microbiol. Open* **2013**, *2*, 633–643. [CrossRef] [PubMed]
14. Roudnew, B.; Lavery, T.J.; Seymour, J.R.; Smith, R.J.; Mitchell, J.G. Spatially varying complexity of bacterial and virus-like particle communities within an aquifer system. *Aquat. Microb. Ecol.* **2013**, *68*, 259–266. [CrossRef]
15. Ruby, J.G.; Bellare, P.; Derisi, J.L. PRICE: Software for the targeted assembly of components of (Meta) genomic sequence data. *G3 Bethesda Md* **2013**, *3*, 865–880. [CrossRef] [PubMed]
16. Rose, R.; Constantinides, B.; Tapinos, A.; Robertson, D.L.; Prosperi, M. Challenges in the analysis of viral metagenomes. *Virus Evol.* **2016**, *2*, vew022. [CrossRef] [PubMed]
17. Parras-Moltó, M.; Rodríguez-Galet, A.; Suárez-Rodríguez, P.; López-Bueno, A. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* **2018**, *6*, 119. [CrossRef] [PubMed]
18. Kim, K.-H.; Bae, J.-W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* **2011**, *77*, 7663–7668. [CrossRef] [PubMed]
19. Breitbart, M.; Salamon, P.; Andresen, B.; Mahaffy, J.M.; Segall, A.M.; Mead, D.; Azam, F.; Rohwer, F. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 14250–14255. [CrossRef] [PubMed]

20. Thurber, R.V.; Haynes, M.; Breitbart, M.; Wegley, L.; Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **2009**, *4*, 470–483. [CrossRef]
21. Froussard, P. A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Res.* **1992**, *20*, 2900. [CrossRef] [PubMed]
22. Djikeng, A.; Halpin, R.; Kuzmickas, R.; Depasse, J.; Feldblyum, J.; Sengamalay, N.; Afonso, C.; Zhang, X.; Anderson, N.G.; Ghedin, E.; et al. Viral genome sequencing by random priming methods. *BMC Genom.* **2008**, *9*, 5. [CrossRef] [PubMed]
23. Dean, F.B.; Nelson, J.R.; Giesler, T.L.; Lasken, R.S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **2001**, *11*, 1095–1099. [CrossRef] [PubMed]
24. Angly, F.E.; Felts, B.; Breitbart, M.; Salamon, P.; Edwards, R.A.; Carlson, C.; Chan, A.M.; Haynes, M.; Kelley, S.; Liu, H.; et al. The marine viromes of four oceanic regions. *PLoS Biol.* **2006**, *4*, e368. [CrossRef] [PubMed]
25. Henn, M.R.; Sullivan, M.B.; Stange-Thomann, N.; Osburne, M.S.; Berlin, A.M.; Kelly, L.; Yandava, C.; Kodira, C.; Zeng, Q.; Weiand, M.; et al. Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PloS ONE* **2010**, *5*, e9083. [CrossRef] [PubMed]
26. Solonenko, S.A.; Ignacio-Espinoza, J.C.; Alberti, A.; Cruaud, C.; Hallam, S.; Konstantinidis, K.; Tyson, G.; Wincker, P.; Sullivan, M.B. Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genom.* **2013**, *14*, 320. [CrossRef] [PubMed]
27. Duhaime, M.B.; Deng, L.; Poulos, B.T.; Sullivan, M.B. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: A rigorous assessment and optimization of the linker amplification method. *Environ. Microbiol.* **2012**, *14*, 2526–2537. [CrossRef] [PubMed]
28. Székely, A.J.; Breitbart, M. Single-stranded DNA phages: From early molecular biology tools to recent revolutions in environmental microbiology. *FEMS Microbiol. Lett.* **2016**, *363*. [CrossRef]
29. Roux, S.; Solonenko, N.E.; Dang, V.T.; Poulos, B.T.; Schwenck, S.M.; Goldsmith, D.B.; Coleman, M.L.; Breitbart, M.; Sullivan, M.B. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **2016**, *4*, e2777. [CrossRef]
30. Drexler, J.F.; Corman, V.M.; Müller, M.A.; Maganga, G.D.; Vallo, P.; Binger, T.; Gloza-Rausch, F.; Cottontail, V.M.; Rasche, A.; Yordanov, S.; et al. Bats host major mammalian paramyxoviruses. *Nat. Commun.* **2012**, *3*, 796. [CrossRef]
31. Karlsson, O.E.; Belák, S.; Granberg, F. The Effect of Preprocessing by Sequence-Independent, Single-Primer Amplification (SISPA) on Metagenomic Detection of Viruses. *Biosecurity Bioterrorism Biodefense Strategy Pract. Sci.* **2013**, *11*, S227–S234. [CrossRef] [PubMed]
32. Blanco, L.; Bernad, A.; Lázaro, J.M.; Martín, G.; Garmendia, C.; Salas, M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **1989**, *264*, 8935–8940. [PubMed]
33. Lasken, R.S.; Stockwell, T.B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.* **2007**, *7*, 19. [CrossRef] [PubMed]
34. Zhang, K.; Martiny, A.C.; Reppas, N.B.; Barry, K.W.; Malek, J.; Chisholm, S.W.; Church, G.M. Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **2006**, *24*, 680. [CrossRef] [PubMed]
35. Kim, K.-H.; Chang, H.-W.; Nam, Y.-D.; Roh, S.W.; Kim, M.-S.; Sung, Y.; Jeon, C.O.; Oh, H.-M.; Bae, J.-W. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl. Environ. Microbiol.* **2008**, *74*, 5975–5985. [CrossRef] [PubMed]
36. Rinke, C.; Low, S.; Woodcroft, B.J.; Raina, J.-B.; Skarshewski, A.; Le, X.H.; Butler, M.K.; Stocker, R.; Seymour, J.; Tyson, G.W.; et al. Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *PeerJ* **2016**, *4*, e2486. [CrossRef]
37. Bowers, R.M.; Clum, A.; Tice, H.; Lim, J.; Singh, K.; Ciobanu, D.; Ngan, C.Y.; Cheng, J.-F.; Tringe, S.G.; Woyke, T. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genom.* **2015**, *16*, 856. [CrossRef]
38. Küsel, K.; Totsche, K.U.; Trumbore, S.E.; Lehmann, R.; Steinhäuser, C.; Herrmann, M. How Deep Can Surface Signals Be Traced in the Critical Zone? Merging Biodiversity with Biogeochemistry Research in a Central German Muschelkalk Landscape. *Front. Earth Sci.* **2016**, *4*, 32. [CrossRef]

39. Kohlhepp, B.; Lehmann, R.; Seeber, P.; Küsel, K.; Trumbore, S.E.; Totsche, K.U. Aquifer configuration and geostructural links control the groundwater quality in thin-bedded carbonate–siliciclastic alternations of the Hainich CZE, central Germany. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 6091–6116. [CrossRef]
40. Kumar, S.; Herrmann, M.; Thamdrup, B.; Schwab, V.F.; Geesink, P.; Trumbore, S.E.; Totsche, K.-U.; Küsel, K. Nitrogen Loss from Pristine Carbonate-Rock Aquifers of the Hainich Critical Zone Exploratory (Germany) Is Primarily Driven by Chemolithoautotrophic Anammox Processes. *Front. Microbiol.* **2017**, *8*, 1951. [CrossRef]
41. Palacios, G.; Quan, P.-L.; Jabado, O.J.; Conlan, S.; Hirschberg, D.L.; Liu, Y.; Zhai, J.; Renwick, N.; Hui, J.; Hegyi, H. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg. Infect. Dis.* **2007**, *13*, 73. [CrossRef] [PubMed]
42. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
43. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
44. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
45. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]
46. Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P.A. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **2017**, *27*, 824–834. [CrossRef]
47. Xie, Y.; Wu, G.; Tang, J.; Luo, R.; Patterson, J.; Liu, S.; Huang, W.; He, G.; Gu, S.; Li, S. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **2014**, *30*, 1660–1666. [CrossRef]
48. Hölzer, M.; Marz, M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience* **2019**, *8*, giz039. [CrossRef]
49. Roux, S.; Enault, F.; Hurwitz, B.L.; Sullivan, M.B. VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **2015**, *3*, e985. [CrossRef]
50. Ren, J.; Ahlgren, N.A.; Lu, Y.Y.; Fuhrman, J.A.; Sun, F. VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **2017**, *5*, 69. [CrossRef]
51. Song, L.; Florea, L.; Langmead, B. Lighter: Fast and memory-efficient sequencing error correction without counting. *Genome Biol.* **2014**, *15*, 509. [CrossRef] [PubMed]
52. Magoč, T.; Salzberg, S.L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **2011**, *27*, 2957–2963. [CrossRef] [PubMed]
53. Rodriguez-r, L.M.; Konstantinidis, K.T. Nonpareil: A redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **2013**, *30*, 629–635. [CrossRef] [PubMed]
54. Rodriguez, L.M.; Konstantinidis, K.T. Estimating coverage in metagenomic data sets and why it matters. *ISME J.* **2014**, *8*, 2349. [CrossRef] [PubMed]
55. Rodriguez-R, L.M.; Gunturu, S.; Tiedje, J.M.; Cole, J.R.; Konstantinidis, K.T. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. *MSystems* **2018**, *3*, e00039-18.
56. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
57. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
58. RStudio Team. RStudio: Integrated Development for R. Available online: http://www.rstudio.com (accessed on 27 May 2019).
59. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **2010**, *11*, 119. [CrossRef]
60. Boratyn, G.M.; Schäffer, A.A.; Agarwala, R.; Altschul, S.F.; Lipman, D.J.; Madden, T.L. Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **2012**, *7*, 12. [CrossRef]
61. Bolduc, B.; Jang, H.B.; Doulcier, G.; You, Z.-Q.; Roux, S.; Sullivan, M.B. vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **2017**, *5*, e3243. [CrossRef]
62. Bin Jang, H.; Bolduc, B.; Zablocki, O.; Kuhn, J.H.; Roux, S.; Adriaenssens, E.M.; Brister, J.R.; Kropinski, A.M.; Krupovic, M.; Lavigne, R.; et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **2019**. [CrossRef] [PubMed]

63. Bolduc, B.; Youens-Clark, K.; Roux, S.; Hurwitz, B.L.; Sullivan, M.B. iVirus: Facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *Isme J.* **2016**, *11*, 7. [CrossRef] [PubMed]
64. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]
65. Miller, C.S.; Baker, B.J.; Thomas, B.C.; Singer, S.W.; Banfield, J.F. EMIRGE: Reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol.* **2011**, *12*, R44. [CrossRef]
66. Emerson, J.B.; Roux, S.; Brum, J.R.; Bolduc, B.; Woodcroft, B.J.; Jang, H.B.; Singleton, C.M.; Solden, L.M.; Naas, A.E.; Boyd, J.A.; et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **2018**, *3*, 870–880. [CrossRef] [PubMed]
67. Roux, S.; Brum, J.R.; Dutilh, B.E.; Sunagawa, S.; Duhaime, M.B.; Loy, A.; Poulos, B.T.; Solonenko, N.; Lara, E.; Poulain, J.; et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **2016**, *537*, 689. [CrossRef] [PubMed]
68. Reisser, W. The Hidden Life of Algae Underground. In *Algae and Cyanobacteria in Extreme Environments*; Seckbach, J., Ed.; Springer Netherlands: Dordrecht, The Netherlands, 2007; pp. 47–58, ISBN 978-1-4020-6112-7.
69. Rosario, K.; Duffy, S.; Breitbart, M. Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J. Gen. Virol.* **2009**, *90*, 2418–2424. [CrossRef]
70. Tucker, K.P.; Parsons, R.; Symonds, E.M.; Breitbart, M. Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *Isme J.* **2010**, *5*, 822. [CrossRef]
71. Breitwieser, F.P.; Salzberg, S.L. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *bioRxiv* **2016**. [CrossRef]
72. Paez-Espino, D.; Eloe-Fadrosh, E.A.; Pavlopoulos, G.A.; Thomas, A.D.; Huntemann, M.; Mikhailova, N.; Rubin, E.; Ivanova, N.N.; Kyrpides, N.C. Uncovering Earth's virome. *Nature* **2016**, *536*, 425. [CrossRef]
73. Roux, S.; Hallam, S.J.; Woyke, T.; Sullivan, M.B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **2015**, *4*, e08490. [CrossRef] [PubMed]
74. Wegner, C.-E.; Gaspar, M.; Geesink, P.; Herrmann, M.; Marz, M.; Küsel, K. Biogeochemical Regimes in Shallow Aquifers Reflect the Metabolic Coupling of the Elements Nitrogen, Sulfur, and Carbon. *Appl. Environ. Microbiol.* **2019**, *85*, e02346-18. [CrossRef] [PubMed]
75. Anderson, R.E.; Brazelton, W.J.; Baross, J.A. Is the genetic landscape of the deep subsurface biosphere affected by viruses? *Front. Microbiol.* **2011**, *2*, 219. [CrossRef] [PubMed]
76. Hölzer, M.; Marz, M. Chapter Nine—Software Dedicated to Virus Sequence Analysis "Bioinformatics Goes Viral." In *Advances in Virus Research*; Beer, M., Höper, D., Eds.; Academic Press: Cambridge, MA, USA, 2017; Volume 99, pp. 233–257, ISBN 0065-3527.
77. Nooij, S.; Schmitz, D.; Vennema, H.; Kroneman, A.; Koopmans, M.P.G. Overview of Virus Metagenomic Classification Methods and Their Biological Applications. *Front. Microbiol.* **2018**, *9*, 749. [CrossRef] [PubMed]
78. Hurwitz, B.L.; Ponsero, A.; Thornton, J.; U'Ren, J.M. Phage hunters: Computational strategies for finding phages in large-scale 'omics datasets. *Virus Res.* **2018**, *244*, 110–115. [CrossRef] [PubMed]
79. Bzhalava, Z.; Tampuu, A.; Bała, P.; Vicente, R.; Dillner, J. Machine Learning for detection of viral sequences in human metagenomic datasets. *BMC Bioinform.* **2018**, *19*, 336. [CrossRef] [PubMed]
80. Wommack, K.E.; Colwell, R.R. Virioplankton: Viruses in Aquatic Ecosystems. *Microbiol. Mol. Biol. Rev.* **2000**, *64*, 69. [CrossRef] [PubMed]
81. Hurwitz, B.L.; Sullivan, M.B. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PLoS ONE* **2013**, *8*, e57355. [CrossRef]
82. Roux, S.; Krupovic, M.; Poulet, A.; Debroas, D.; Enault, F. Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS ONE* **2012**, *7*, e40418. [CrossRef]

83. Earl, P.L.; Jones, E.V.; Moss, B. Homology between DNA polymerases of poxviruses, herpesviruses, and adenoviruses: Nucleotide sequence of the vaccinia virus DNA polymerase gene. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 3659–3663. [CrossRef]

84. Villarreal, L.P.; DeFilippis, V.R. A Hypothesis for DNA Viruses as the Origin of Eukaryotic Replication Proteins. *J. Virol.* **2000**, *74*, 7079. [CrossRef] [PubMed]