viruses

MDPI

Article

Assessing Bias and Reproducibility of Viral Metagenomics Methods for the Combined Detection of Faecal RNA and DNA Viruses

Rik Haagmans [1,2], Oliver J. Charity [1], Dave Baker [3], Andrea Telatin [3], George M. Savva [3], Evelien M. Adriaenssens [1,4], Penny P. Powell [2,*] and Simon R. Carding [1,2,*]

1 Food, Microbiome, and Health Research Programme, Quadram Institute Bioscience, Norwich Research Park, Norwich NR4 7UQ, UK; rik.haagmans@quadram.ac.uk (R.H.); oliver.charity@quadram.ac.uk (O.J.C.); evelien.adriaenssens@quadram.ac.uk (E.M.A.)
2 Norwich Medical School, University of East Anglia, Norwich NR4 7TJ, UK
3 Core Science Resources, Quadram Institute Bioscience, Norwich NR4 7UQ, UK; david.baker@quadram.ac.uk (D.B.); andrea.telatin@quadram.ac.uk (A.T.); george.savva@quadram.ac.uk (G.M.S.)
4 Microbes and Food Science Research Programme, Quadram Institute Bioscience, Norwich Research Park, Norwich NR4 7UQ, UK
* Correspondence: p.powell@uea.ac.uk (P.P.P.); simon.carding@quadram.ac.uk (S.R.C.)

Abstract: Whole transcriptome amplification (WTA2) and sequence-independent single primer amplification (SISPA) are two widely used methods for combined metagenomic sequencing of RNA and DNA viruses. However, information on the reproducibility and bias of these methods on diverse viruses in faecal samples is currently lacking. A mock community (MC) of diverse viruses was developed and used to spike faecal samples at different concentrations. Virus-like particles (VLPs) were extracted, nucleic acid isolated, reverse-transcribed, and PCR amplified using either WTA2 or SISPA and sequenced for metagenomic analysis. A bioinformatics pipeline measured the recovery of MC viruses in replicates of faecal samples from three human donors, analysing the consistency of viral abundance measures and taxonomy. Viruses had different recovery levels with VLP extraction introducing variability between replicates, while WTA2 and SISPA produced comparable results. In comparing WTA2- and SISPA-generated libraries, WTA2 gave more uniform coverage depth profiles and improved assembly quality and virus identification. SISPA produced more consistent abundance, with a 50% difference between replicates occurring in ~20% and ~10% of sequences for WTA2 and SISPA, respectively. In conclusion, a bioinformatics pipeline has been developed to assess the methodological variability and bias of WTA2 and SISPA, demonstrating higher sensitivity with WTA2 and higher consistency with SISPA.

Keywords: viral metagenomics; mock community; eukaryotic viruses; bacteriophages

Academic Editor: Leyi Wang

Received: 25 November 2024
Revised: 17 January 2025
Accepted: 19 January 2025
Published: 23 January 2025

Citation: Haagmans, R.; Charity, O.J.; Baker, D.; Telatin, A.; Savva, G.M.; Adriaenssens, E.M.; Powell, P.P.; Carding, S.R. Assessing Bias and Reproducibility of Viral Metagenomics Methods for the Combined Detection of Faecal RNA and DNA Viruses. *Viruses* **2025**, *17*, 155. https://doi.org/10.3390/v17020155

1. Introduction

The human gastrointestinal (GI) virome consists of diverse viruses of prokaryotes (phages) and eukaryotes and plays a major role in human health and disease, with phage–host dynamics affecting the structure and function of the GI microbiome, and eukaryotic viruses directly infecting human cells [1–3]. Metagenomics methods enable the collective analysis of microbial genomes (i.e., including bacterial and viral genomes: the microbiome) and have been used for gastrointestinal (GI) virome analysis since the first study of the

human GI virome [4]. The majority of metagenomics-based virome studies to date have focused almost exclusively on double-stranded DNA (dsDNA) bacteriophages (phages) [5]. However, recent metatranscriptomic studies [6,7] have shown that the abundance of RNA phages has been underestimated [8] and a significant portion of human viruses have RNA genomes. Therefore, comprehensive metagenomic analysis of GI RNA and DNA viruses is important for human health as it improves our understanding of the interactions between phages and bacteria, as well as the surveillance and monitoring of human GI viruses [9]. Standardised methods for which bias and reproducibility have been assessed are important, particularly for longitudinal studies, to account for any methodological variability [10,11].

Viromes derived from faecal samples are commonly used as a proxy for the GI virome. Metagenomic protocols employ nucleic acid extracted from bulk faecal homogenates or from samples enriched for virus-like particles (VLPs) to reduce the amount of non-viral nucleic acid. While VLP samples contain reduced amounts of non-viral sequences, the VLP enrichment process can affect the recovery of viruses and therefore reduce accuracy and introduce bias [5,12–17]. For example, short-term storage at 4 °C and ambient temperatures, and long-term storage at −80 °C have negligible effects on phage virome composition [17]. On the other hand, operator bias can have a significant effect on between-sample variation in virome composition [17]. As a general principle, minimal sample processing is the optimal approach to achieve optimal virus recovery. Some widely used VLP extraction and purification methods including ultracentrifugation, CsCl density gradient centrifugation, ultrafiltration, and tangential flow filtration can lead to the loss of specific viruses [12–16]. Once extracted, direct ligation of sequencing adapters to recovered nucleic acid without subsequent amplification produces the least bias [16,18,19] but requires quantities of input material (0.5–5 g) and/or concentration (e.g., polyethylene glycol precipitation), which is time consuming [16,17,20]. This approach is also not possible using smaller biomass faecal samples, which require amplification to achieve sufficient quantities of nucleic acid for sequencing libraries.

Popular amplification methods include multiple displacement amplification (MDA) and sequence-independent single primer amplification (SISPA), both of which can include a reverse transcription step to produce complementary DNA (cDNA) for sequencing of RNA genomes [17,21]. MDA is known to introduce amplification bias of circular genomes due to rolling circle amplification [15,22,23], and introduces other amplification biases [24,25] that cannot be overcome by pooling samples [26]. SISPA randomly amplifies DNA and RNA virus genomes using a reverse transcriptase and random primers with a 5′ universal sequence [27,28]. While SISPA produces better representation of community relative abundances, both the MDA and SISPA libraries have higher GC bias and non-uniform genome coverage compared to non-amplified libraries [15]. Whole transcriptome approaches such as Whole Transcriptome Amplification (Sigma WTA2), as implemented in the NetoVir protocol, have shown high reproducibility and good correlation of read numbers with qPCR-based virus abundance [12]. The WTA2 approach is less time consuming than SISPA, although it is costly and has been—in our experience—susceptible to supply chain issues.

A common approach to evaluating viromics pipelines is the use of specific viruses or mixtures of viruses (mock virus communities, MCs) [12–15,23,29] that are added to (spiked) faecal samples to assess the recovery and loss of viral sequences across the pipeline [13,16,17,19,30,31]. However, most of these studies exclude RNA viruses [13,16,19,23] and use different combinations of VLP extraction and amplification methods. Moreover, each study focused on distinct aspects of bias and reproducibility such as the recovery bias of MC and faecal viruses, sequencing and GC bias, over- or underrepresentation of different virus sequences, genome coverage and genome coverage depth, diversity of faecal viruses, and taxonomic variation. While WTA2 can produce a good correlation of read

numbers to qPCR-based abundance of MC viruses [12], a comprehensive evaluation of other biases and reproducibility characteristics and its performance with faecal samples and its comparison with other amplification methods is lacking.

To address these issues, we have constructed a virus MC consisting of equal numbers of diverse viruses, including the dsDNA phages T5, Det7, and P22, the ssDNA phage M13, the dsDNA murid gammaherpesvirus-68 (MHV-68), dsRNA Rotavirus A (RV-A), and ssRNA viruses bovine viral diarheavirus-1 (BVDV-1). The MC was added to three faecal samples at two different concentrations, and VLPs were extracted from these spiked samples, as well as from an untreated aliquot of the sample. VLP nucleic acid was then processed using WTA2 and SISPA and sequenced. Using a custom bioinformatics pipeline, various bias and reproducibility metrics were assessed. This included recovery, sequencing depth uniformity, GC bias, and accuracy of de novo assembly-based relative abundance measurement of MC viruses. Additionally, the precision of relative abundance measurements of the faecal virome was as assessed by comparing the relative abundance between replicates and for different taxonomic clades.

2. Materials and Methods

2.1. Samples and Ethics

Three faecal samples were obtained for the study "Autoimmunity in ME/CFS" (AI-ME/CFS), registered on ClinicalTrials.gov with number NCT03254823 [32] with ethical approval obtained from the National Research Ethics Service (NRES) Committee London Hampstead (17/LO/1102, IRAS ID 218545). The samples came from healthy control individuals enrolled in the trial and were collected between March 2018 and October 2019, and were homogenised by mixing, divided into 100 mg aliquots, and stored at −80 °C as described previously [32]. Participants provided informed consent for the use of samples in subsequent research. All research was performed in accordance with the Declaration of Helsinki [33], and the International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)-Good Clinical Practice (ICH-GCP) guidelines. Data were handled following the European Union General Data Protection Regulation (GDPR) and United Kingdom Data Protection Act 2018.

2.2. Construction of the Mock Virus Community

A virus MC was constructed using virus stocks consisting of tailed phages T5, Det7, and P22, the filamentous phage M13, and the mammalian viruses MHV-68, BVDV-1 strain Ky1203nc [34], and RV-A strain simian rotavirus SA/11 (Table 1). Virus titres were determined by epifluorescence microscopy using a protocol adapted from [35,36]. Briefly, a 13 mm Al2O3 Anodisc 0.02 µm filter membrane (Cytiva, Little Chalfont, UK, ref 6809-7003) was placed in a Swinnex filter holder (Millipore, Darmstadt, Germany, ref SX0001300) fitted onto a glass tube protruding through a rubber stopper into a Büchner flask. The flask was connected to a Millivac Maxi vacuum pump (Millipore, Darmstadt, Germany, ref SD1P014M04). After the pump was switched on, 400 µL deionised water or PBS was added to the filter holder to confirm the inlet was sealed. The stained sample was added gradually using a 1 mL micropipette. The filter was then washed with 400 µL of deionised water or PBS and any remaining liquid was aspirated for one minute. The filter was then placed sample-side-up onto Whatman filter paper (Whatman, Marlborough, MA, USA, ref 1004-055) to dry for 5 min. A 7.5 µL drop of Fluoromount G (Invitrogen, Waltham, MA, USA, ref 00-4958-02) was placed on a glass microscopy slide (VWR, Leicestershire, UK, ref 631-0117) and the filter placed on top with 7.5 µL of Fluoromount G added to the filter and covered with a glass cover slip (VWR, ref 631-0125). The slide was then left in the dark at 21 °C for at least 2 h to allow the mountant to set. The slides were imaged on an

Axio Imager.M2 upright fluorescence microscope (Zeiss, Cambourne, UK). Samples were illuminated using a HAL 100 illuminator with a quartz collector 43 (Zeiss, Cambourne, UK, ref 423000-9901-000) and a 65HE Alexa 488 filter. For titre measurements, a 100X EC Plan-Neofluar oil immersion objective (Zeiss, Cambourne, UK, ref 420496-990-000) was used. Images were captured on an ICX 285 CCD monochrome camera (Sony, Surrey, UK). Between 15 and 20 images were taken of each filter at random locations.

Table 1. Characteristics of the viral mock community.

Host	Virus		Genome			Virion		MC	
	Name	Species	Type	Top.	Size (kb)	Env.	Size (nm)	HI (VLP [1])	LO (VLP [1])
Prokar-yotic	Det7	*Kuttervirus Det7*	dsDNA	linear	157.5		90 [C]/110 [T]	1.0×10^7	1.0×10^5
	M13	*Inovirus M13*	ssDNA	circular	6.4		6.5 [D]/860 [L]	1.0×10^7	1.0×10^5
	P22	*Lederbergvirus P22*	dsDNA	linear	41.7		60	1.0×10^7	1.0×10^5
	T5	*Tequintavirus T5*	dsDNA	linear	121.7		90 [C]/160 [T]	1.0×10^7	1.0×10^5
Eukar-yotic	BVDV-1	*Pestivirus bovis*	ssRNA	linear	12.5	Yes	50	1.0×10^7	1.0×10^5
	MHV-68	*Rhadinovirus muridgamma4*	dsDNA	linear	119.5	Yes	220	2.5×10^6	2.5×10^4
	RV-A	*Rotavirus alphagastroenteritidis*	dsRNA	linear (11 Sg)	18.6		80	1.0×10^7	1.0×10^5

Env.: envelope, Sg.: genome segments, Top.: topology, [1] VLP per spike-in, [C] capsid size, [T] tail length, [D] virion diameter, [L] virion length.

To prepare the MC, each virus stock was diluted in PBS to obtain the same final concentration, except for MHV-68, for which the concentration was 25% of the concentration of the other viruses. A high concentration mock community (HI) was constructed such that 35 µL of the final MC contained 0.25×10^7 particles of MHV-68 and 1×10^7 of each of the other viruses, for a total of 6.25×10^7 virus particles. A lower amount of MHV-68 was used due to limited availability of stock. As the estimated number of virus particles in human faeces is up to ~10^9 particles/g [36], the total number of virus particles in 35 µL HI roughly corresponded to the maximum expected number of virus particles in 50 mg faeces, for a final concentration of 1.25×10^9 MC particles per gram of faeces. A low concentration MC (LO) was produced by 100-fold dilution of HI in PBS, with 35 µL of LO added to a 50 mg faecal sample corresponding to 1.25×10^7 MC particles per gram of faeces. MCs were stored at 4 °C.

2.3. Spiking of Faecal Samples

For each faecal sample, 1.9–3.5 g was diluted to 10% (*w*/*v*) in PBS and homogenised using a Stomacher 400 Circulator Lab Blender (Seward, Worthing, UK) set to 260 RPM for 3 min. To 1 mL aliquots of S06, S07, and S08, 35 µL of HI MC was added, producing samples S06-HI, S07-HI, and S08-HI, respectively (Figure 1A). A second 1 mL aliquot of each sample was spiked with LO MC, producing samples S06-LO, S07-LO, and S08-LO, with a third control aliquot without spiking designated S06-No, S07-No, and S08-No. Additionally, 35 µL MC HI and 35 µL MC LO were each added to 1 mL PBS to produce the blank samples SMC-HI and SMC-LO, respectively, and a blank PBS sample was included, designated SBL-NO.

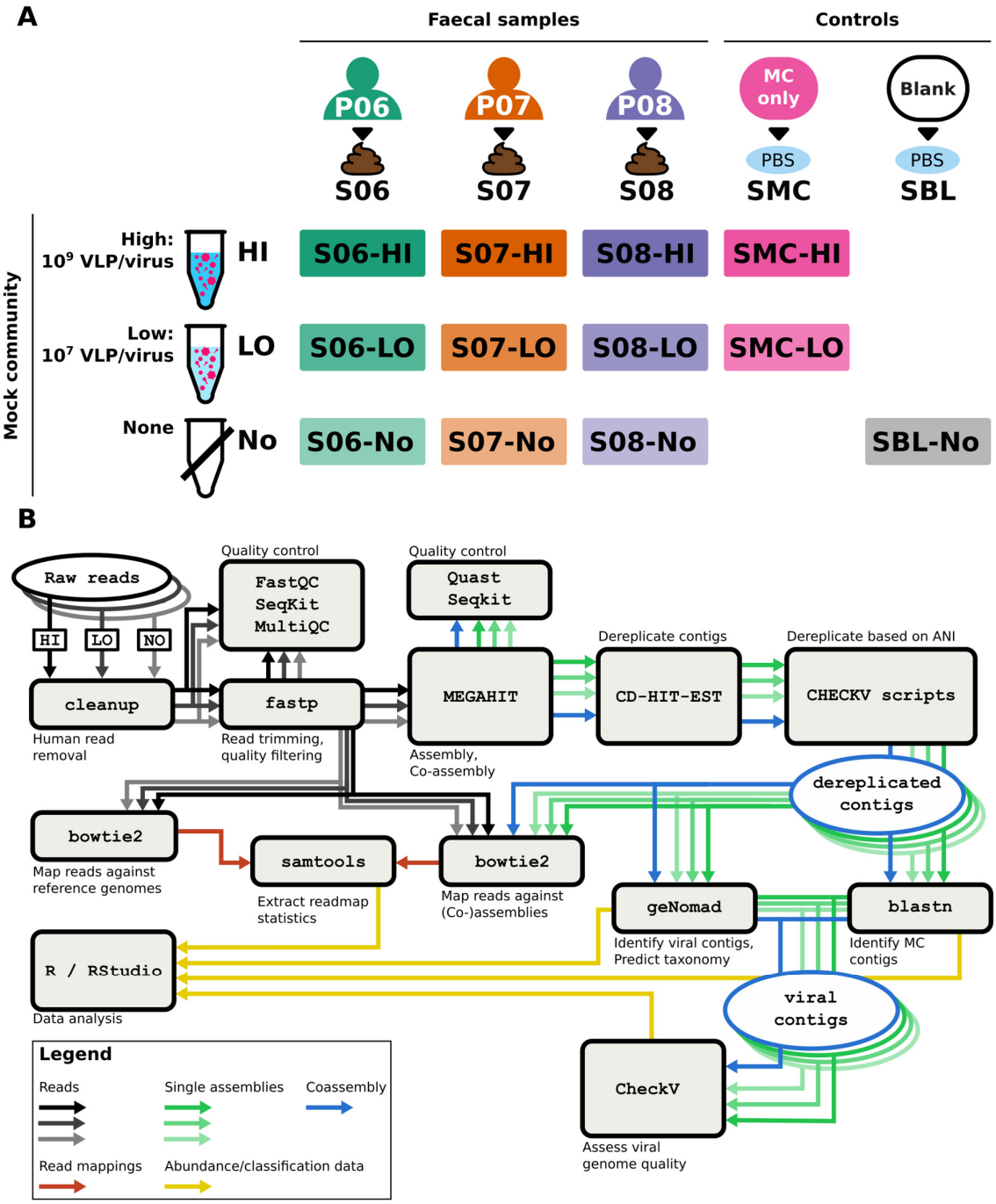

Figure 1. Overview of experimental design and methodology. (**A**) Stool samples (S06, S07, and S08) were obtained from three donors (P06, P07, and P08, respectively). After homogenisation, three aliquots were obtained from each stool sample. To one aliquot, a high-concentration MC (HI) was added, to the second aliquot a low-concentration MC (LO) was added, and the third aliquot was untreated (No). Additionally, an MC-only sample (SMC) of HI and LO and a blank PBS sample (SBL-No) were processed. (**B**) Bioinformatics workflow. Black and grey arrows depict flow of read data from raw reads to assembler and read mapping. Green and blue arrows indicate flow of assemblies of single samples and co-assembly of pooled samples, respectively, from the same stool sample from assembler to contig classification and read mapping. Red arrows indicate the flow of read mapping data, and yellow arrows indicate the flow of abundance and classification data to the data analysis stage.

2.4. VLP Nucleic Acid Extraction

Samples were centrifuged at 16,000× *g* for 3 min at 21 °C, and the supernatant syringe filtered using a 0.45 µm syringe filter unit (Starlab, Milton Keynes, UK, ref E4780-1456) to extract VLPs. The VLP extract was stored at −80 °C for 16 h after which 700 µL was treated with a nuclease cocktail consisting of 1 µL Benzonase (Millipore, Gillingham, UK, ref E1014-5KU), 4 µL RNAse I (Thermo Fisher Scientific, Loughborough, UK, ref AM2294), 16 µL DNAse I (Thermo Scientific, Loughborough, UK, ref EN0521), 40 µL TURBO DNase (Thermo Fisher Scientific, Loughborough, UK, ref AM2238), and 125 µL of the respective 10X buffers, and incubated at 37 °C for 2 h to digest unprotected nucleic acid. The samples were then incubated at 75 °C for 1 h and 10 µL 0.5 mM EDTA (pH = 8.0) (Thermo Fisher Scientific, Loughborough, UK, ref AM9260G) was added, to deactivate nucleases. Viral nucleic acid was extracted using the QIAmp Viral RNA Mini Kit (QIAGEN, Manchester, UK, ref 52904) according to the manufacturer's instructions, without the addition of carrier RNA. Nucleic acid extracts were stored at −20 °C.

2.5. Sequencing Libraries

Two approaches were tested, one using the Complete Whole Transcriptome Amplification (WTA2) kit (Sigma-Aldrich, Gillingham, UK, ref WTA2-10RXN) [12] and another using SISPA [37].

2.5.1. WTA2

The dsDNA pool was prepared as described previously [12]. Briefly, 2.82 µL of the sample was added to 0.5 µL of Library Synthesis Solution containing universal sequence-tagged quasi-random hexamer primers. The sample was denatured at 95 °C for 2 min and then cooled to 18 °C to prime RNA and DNA. Then, 1.68 µL library synthesis master mix was added, containing 0.5 µL Library Synthesis Buffer, 0.78 µL of RNAse-free water, and 0.4 µL Library Synthesis Enzyme and incubated in a thermocycler set to 18 °C for 10 min, 25 °C for 10 min, 37 °C for 30 min, 42 °C for 10 min, and 70 °C for 20 min, to produce a dsDNA library (Figure 2A). From each dsDNA library, 5 µL was added to 69.95 µL amplification master mix, containing 7.5 µL Amplification Mix with universal sequence primers, 60.2 µL nuclease-free water, 1.5 µL WTA2 dNTP mix, and 0.75 µL Amplification Enzyme. The sample was incubated in a thermocycler set to 94 °C for 2 min, followed by 17 cycles of 94 °C for 2 min and 70 °C for 5 min, to amplify the dsDNA fragments. DNA was then extracted using the QIAquick PCR Purification Kit (QIAGEN, Manchester, UK) according to the manufacturer's instructions and stored at −20 °C.

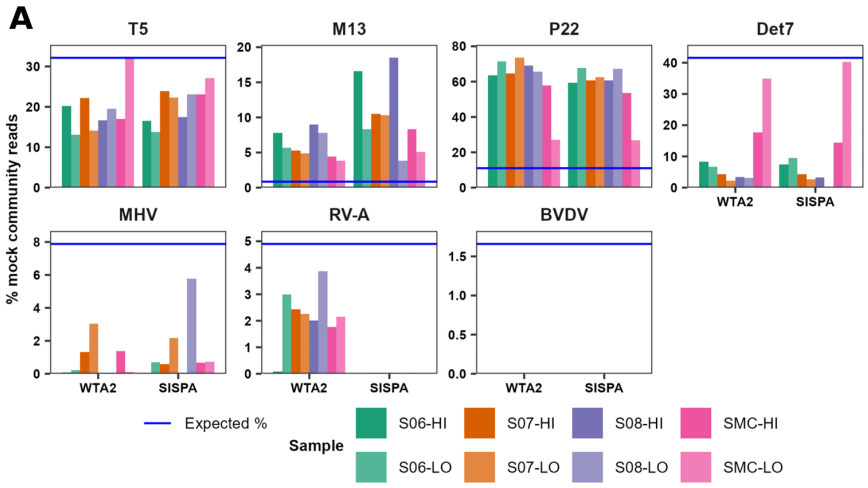

Figure 2. *Cont.*

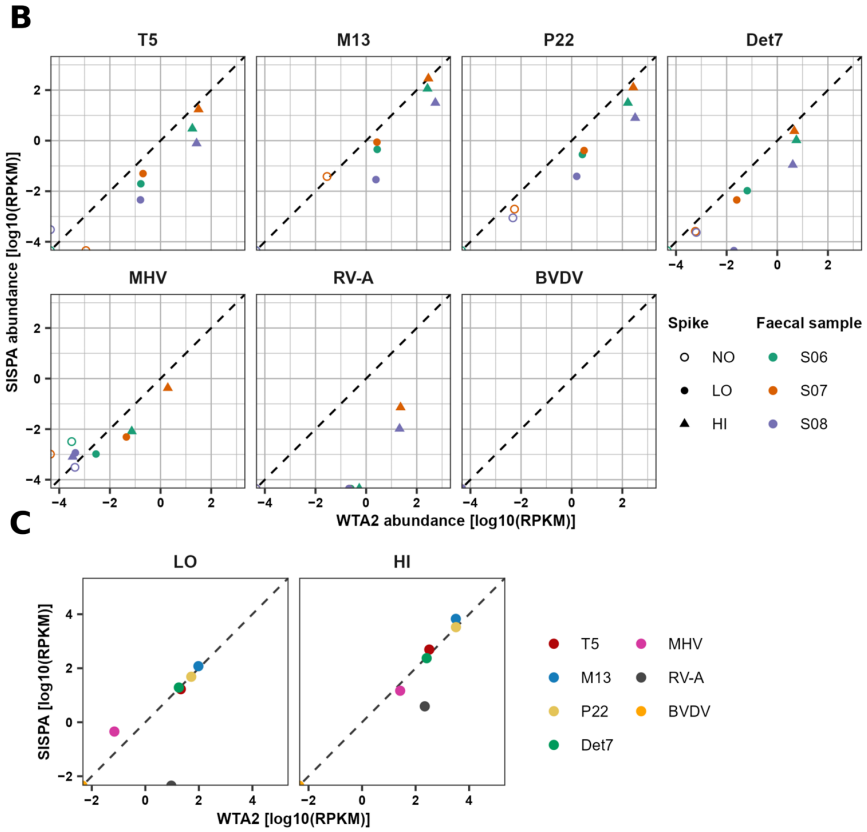

Figure 2. Bias and consistency of MC virus recovery in WTA2 and SISPA libraries. (**A**) Percentage of MC reads mapping to each MC virus, comparing samples processed using WTA2 to SISPA. The blue horizontal line shows the expected fraction of reads mapping to each virus based on the number of particles added, and the total genome length and number of strands. (**B**) Comparison of the abundance of each virus in the two spiked and one non-spiked aliquots of the three stool samples, generated using the WTA2 and SISPA methods. The diagonal dashed line depicts a perfect correlation. (**C**) Comparison of MC virus abundance in the MC HI control sample processed using WTA2 and SISPA. Abundance is measured in reads per kilobases of genome per million reads (RPKM).

2.5.2. SISPA

SISPA was used to produce a dsDNA pool for sequencing from the same nucleic acid extracts processed using the WTA2 kit. Samples were processed as described previously [37]. Briefly, 4 µL of the sample was added to 9 µL master mix containing 1 µL RNasin (40 U/µL) (Promega, Chilworth Southampton, UK, ref N261A), 1 µL 10 mM dNTPs (New England Biolabs, Hitchin, UK, ref N0447S), 1 µL 20 µM primer D2_8N (5′ AAGCTAAGACGGCGGTTCGGNNNNNNNN-3′), and 6.5 µL nuclease-free water (Thermo Fisher Scientific, Loughborough, UK, ref AM9937), incubated at 65 °C and then cooled to 4 °C. Then, 6 µL master mix containing 4 µL 5X first strand buffer, 1 µL 0.1 mM dithiothreitol, and 1.0 µL SuperScript III reverse transcriptase (200 U/µL) (Thermo Fisher Scientific, Loughborough, UK, ref 18080044) was added and incubated in a thermocycler at 50 °C for 1 h for first strand synthesis. Then, 1.5 µL master mix containing 0.85 µL nuclease-free water, 0.15 µL 10X Klenow buffer, and 0.5 µL DNA polymerase I large (Klenow) fragment (5 U/µL) (New England Biolabs, Hitchin, UK, ref M0212L) was added, and the sample was incubated at 37 °C for 1 h for second strand synthesis, followed by incubation at 75 °C for 10 min to inactivate the enzyme. Free primers and nucleotides were then digested and dephosphorylated by incubation with 20 µL master mix containing 17 µL nuclease-free water, 1.0 µL 10X Shrimp Alkaline Phosphatase (SAP) buffer, 1.0 µL Exonuclease I (20 U/µL) (New England Biolabs, Hitchin, UK, ref M0293S), and SAP (1 U/µL) (New England Biolabs, Hitchin, UK, ref M0371S), respectively, at 37 °C for 1 h, followed

by 15 min at 75 °C to inactivate the enzymes. The sample was then frozen at −20 °C for 16 h. The following day, the dsDNA library was generated and amplified from 8 µL of the reverse-transcribed sample, by adding it to 42 µL master mix containing 26 µL nuclease-free water, 5 µL 10X PCR buffer, 6 µL 25 mM MgCl2, 1.5 µL 10mM dNTP, 3 µL 20 µM primer D2 (5′ AAGCTAAGACGGCGGTTCGG-3′), and 0.5 µL AmpliTaq Gold (5 U/µL) (Thermo Fisher Scientific, Loughborough, UK, ref 10685095) DNA polymerase by incubation on a thermocycler. The thermocycler program consisted of: (1) denaturation for 5 min at 95 °C, (2) 5 cycles of denaturation at 95 °C for 1 min, annealing at 55 °C for 1 min, and extension at 72 °C for 1:30 min, (3) 25 cycles of denaturation at 95 °C for 30 s, annealing at 55 °C for 30 s, and extension at 72 °C for 1:30 min, adding 2 s to the extension time every cycle, and (4) final extension at 72 °C for 10 min. Samples were kept on ice between incubation steps throughout. The PCR product was loaded onto a 2% agarose (Melford, Ipswich, UK, ref 3913900099) gel in 0.5X TBE buffer (Thermo Fisher Scientific, Loughborough, UK, ref J62788.K2), and inspected for a smear between 200 and 500 bp. DNA was then extracted using the gDNA Cleanup & Concentrator-10 kit (Zymo Research, Freiburg im Breisgau, Germany, ref D4010) following manufacturer's instructions. Purified DNA of the WTA2 and SISPA methods was quantified using a Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Loughborough, UK, ref Q32851) on a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Loughborough, UK, ref Q33216) and samples were normalised to 5 ng/µL prior to sequencing library preparation.

2.6. Illumina Library Preparation and Sequencing

For preparation of the Illumina sequencing library, 0.5 µL of Tagmentation Buffer 1 was mixed with 0.5 µL Bead Linked Transposomes (Illumina, Cambridge, UK, ref 20018704) and 4 µL nuclease-free water in a master mix, and 5 µL added to a 96-well plate. A total of 2 µL of DNA normalised to 5 ng/µL was pipette-mixed with 5 µL of the Tagmentation mix and heated to 55 °C for 15 min. A PCR master mix was made using 10 µL KAPA 2G Fast Hot Start Ready Mix (Merck, Gillingham, UK, ref KK5601) and 2 µL PCR grade water per sample. Of the PCR master mix, 12 µL was added to each well to be used in a 96-well plate, and 1 µL of 10 µM primer mix containing both P7 and P5 Illumina barcodes [38] were added to each well. For the WTA2 samples, custom 9 bp dual barcodes were used, while for the SISPA samples, custom 10 bp unique dual index barcodes were used. Finally, 7 µL Tagmentation mix was added and mixed. The PCR was run at 72 °C for 3 min, 95 °C for 1 min, then 14 cycles of 95 °C for 10 s, 55 °C for 20 s, and 72 °C for 3 min. The libraries were quantified using the Promega QuantiFluor dsDNA System (Promega, Chilworth Southampton, UK, ref E2670) and measured on a GloMax Discover Microplate Reader (Promega, Chilworth Southampton, UK, ref GM3000). Libraries were pooled following quantification in equal quantities. The final pool was size selected using solid phase reversible immobilisation (SPRI) beads at 0.5X concentration, followed by SPRI beads size selection at 0.7X concentration, using Illumina DNA Prep, (M) Tagmentation sample purification beads (Illumina, Cambridge, UK, ref 20060059). The final pool was quantified on a Qubit 3.0 Fluorometer and run on a D5000 ScreenTape (Agilent, Stockport, UK, ref 5067-5579) using the Agilent TapeStation 4200 (Agilent, Stockport, UK) to calculate the final library pool molarity. The WTA2 pool was run at a final concentration of 1.8 pM on an Illumina NextSeq 500 instrument with a high-output 300-cycle flow cell (Illumina, Cambridge, UK, ref 20024908). The SISPA pool was run at a final concentration of 750 pM on an Illumina NextSeq 20000 instrument using a P3 300-cycle flow cell (Illumina, Cambridge, UK, ref 20040561). Each were run following the Illumina recommended denaturation and loading recommendations and included a 1% PhiX Control v3 spike-in (Illumina, Cambridge, UK, ref FC-110-3001).

2.7. Bioinformatics Analysis

An overview of the bioinformatics pipeline used is depicted in Figure 1B. The pipeline was composed in Nextflow v.24.04.3 [39]. The pipeline is available at https://github.com/RHaagmans/mc-spike (accessed 27 November 2024).

2.7.1. Host Read Removal

Before running the analysis pipeline, host reads were removed using the cleanup v1.4-11-gf422ea8 Nextflow pipeline (available from https://github.com/telatin/cleanup, accessed on 17 July 2024). Briefly, host reads were removed using Kraken2 against a custom masked human reference genome [40], SARS-CoV-2 genome, and phage PhiX174 genome using Kraken2 [41] followed by adapter trimming using fastp v.0.23.4 [42,43] and a quality report generated using MultiQC [44].

2.7.2. Read Quality Control

Read quality before and after quality filtering was analysed using FastQC v.0.12.1 [45]. Adapter sequences were trimmed, and reads were filtered based on quality using Fastp with the default settings and the options automatic adapter detection, base correction for pair-ended data, and cutting of the tail sequence. Fastp and FastQC outputs were gathered and visualised in MultiQC v.1.23. Duplicate reads were detected using the HTStream v1.3.3 module SuperDeduper, with default length set to 100, and the minimum read quality score set to 1. Read statistics were extracted using SeqKit v.2.8.2 [46] stats command.

2.7.3. Reference-Based Relative Abundance of MC Viruses

To determine the relative abundance of MC viruses in the samples, reads were mapped against an index of mock community reference genomes (Table 2). First, reference genomes were downloaded from the National Center for Biotechnology Information (NCBI) nucleotide database using the reference genome accession numbers with NCBI Entrez Direct program efetch v.22.1, and an index was built using Bowtie 2 v.2.5.4 command bowtie2-build with default settings. Reads were then mapped to the reference index using Bowtie 2. The number of reads mapping to respective viruses was calculated using the SAMtools v.1.20 [47] idxstats command. To calculate the expected percentage of MC reads that map to each individual virus, first the number of bases each virus contributes to the sample was calculated, by multiplying the genome length by the number of added particles and the number of strands in the genome (single or double). The expected percentage of reads for each virus was calculated as a fraction of the cumulative number of bases for all viruses. MC virus abundance was calculated as the number of reads mapped to the respective genomes, per 1000 bases of genome per 1 million reads in the samples (RPKM).

Table 2. Overview of MC reference sequences used.

Virus	NCBI Accession	Sequence Name	Genome	GC%
Abbreviation	Accession Nr.		Bases	
BVDV	NC_001461.1	Bovine viral diarrhoea virus 1, complete genome	12,573	45.79
Det7	NC_027119.1	Salmonella phage Det7, complete genome	157,498	44.60
M13	NC_003287.2	Enterobacteria phage M13, complete genome	6407	40.74
MHV	NC_001826.2	Murine herpesvirus 68 strain WUMS, complete genome	119,451	47.23
P22	NC_002371.2	Salmonella phage P22, complete genome	41,724	47.09
RV-A	NC_011507.2	Rotavirus A segment 1, complete genome	3302	33.74
RV-A	NC_011506.2	Rotavirus A segment 2, complete genome	2693	32.97
RV-A	NC_011508.2	Rotavirus A segment 3, complete genome	2591	28.91
RV-A	NC_011510.2	Rotavirus A segment 4, complete genome	2362	34.67
RV-A	NC_011500.2	Rotavirus A segment 5, complete genome	1614	31.16

Table 2. *Cont.*

Virus	NCBI Accession	Sequence Name	Genome	GC%
Abbreviation	Accession Nr.		Bases	
RV-A	NC_011509.2	Rotavirus A segment 6, complete genome	1356	38.57
RV-A	NC_011501.2	Rotavirus A segment 7, complete genome	1105	33.81
RV-A	NC_011502.2	Rotavirus A segment 8, complete genome	1059	33.30
RV-A	NC_011503.2	Rotavirus A segment 9, complete genome	1062	35.78
RV-A	NC_011504.2	Rotavirus A segment 10, complete genome	751	40.21
RV-A	NC_011505.2	Rotavirus A segment 11, complete genome	667	38.53
T5	NC_005859.1	Enterobacteria phage T5, complete genome	121,750	39.27

2.7.4. Mock Community Coverage Depth

Coverage depth and breadth of MC viruses were calculated using the SAMtools depth command with option '-aa'. Coverage breadth was calculated as the fraction of bases with a depth > 0. To compare sequencing depth across samples, the normalised sequencing depth was calculated as a fraction of the total number of bases sequenced for each virus in each sample.

2.7.5. GC-Content of MC Reads and Genomes

The GC-content of MC reads was calculated using the SAMtools view command to extract read sequences and a custom python script to calculate the GC fraction by dividing the number of GCs in all reads mapping to a virus genome by the total number of bases mapping to the genome. Genome GC-content was calculated using the SeqKit fx2tab command.

2.7.6. Assembly of Reads

High-quality reads corresponding to the same faecal sample were pooled. Pooled reads and reads from individual samples were co-assembled and assembled, respectively, using MEGAHIT v.1.2.9 [48,49] with default settings. Contigs were dereplicated in two steps. First, CD-HIT-EST v.4.8.1 [50,51] used cluster contigs based on ≥95% sequence identity, and ≥80% alignment coverage of the shorter sequence. Using BLAST v.2.15.0, a database of CD-HIT-EST was then built, and an all-versus-all alignment was performed using blastn [52]. Using the accessory scripts "anicalc.py" and "aniclust.py" from CheckV v.1.0.3 [52], contigs were again clustered, at 95% average nucleotide identity and 85% coverage of the shortest sequence to remove circularly permutated redundant sequences. Contigs were filtered from the original assembly using the list of contig IDs using the SeqKit grep command. (Co)assemblies before and after dereplication were analysed for quality control using Quast v.5.2.0 [53] and SeqKit.

2.7.7. Identification and Classification of Viral and MC Sequences

For identification and taxonomic annotation of viral contigs, geNomad v.1.8.0 [54] was used with geNomad database v.1.7. Dereplicated contigs were analysed using the end-to-end pipeline in geNomad with default settings. For downstream analysis, only non-provirus contigs with a virus score > 0.8 were used. Contigs derived from MC viruses were identified using BLAST. A BLAST database was built of MC virus reference genomes and dereplicated contigs were then aligned against the MC database. Contigs that aligned with ≥98% nucleotide identity and ≥95% coverage of the contig were marked as MC sequences. Assembly genome coverage was calculated by summing the lengths of contigs matching individual viruses. Viral sequence quality was evaluated using the CheckV end-to-end pipeline with database v.1.5.

2.7.8. Virus Abundance and Relative Abundance

To determine faecal viral abundances, reads were mapped against the dereplicated contigs using Bowtie 2 with default settings. Using SAMtools command idxstats, the number of reads mapping to each contig were extracted. Contig abundance was calculated on the sample and virome levels, by normalising the number of reads mapping to each contig by the contig length in kb and reads or total number of reads mapping to viral contigs, in millions (RPKM), respectively. The virome was defined as those contigs identified as viral by geNomad, excluding mock community sequences to maintain consistency between replicates with different spike-in treatments. Relative abundance was then calculated as a fraction of the cumulative total and virome abundance, respectively.

2.7.9. Variation in Relative Abundance and Rank

To calculate the variation in abundance, relative abundance, and abundance ranking, reads from each of the individual replicates were mapped to the respective co-assemblies. Abundance and relative abundance were calculated on the sample and virome levels as described above. Contigs with at least one read in each of the replicates were included in the analysis. The range in abundance was calculated as the log2-transformed ratio of the highest and lowest abundance of each contig in three replicates. Abundance rank was calculated by sorting contigs by their abundance from highest to lowest in each replicate, and the range was determined as the difference between the highest and lowest ranking of a contig among the three replicates. The coefficient of variation in relative abundance was calculated as the coefficient of variation of the relative abundance of the contig in the three replicates.

2.7.10. Virome Taxonomic and Diversity Analysis

Alpha diversity based on assemblies and co-assemblies was calculated using the R package phyloseq v.1.48 [55]. Beta diversity was calculated using the R package vegan v.2.6-6.1, by calculating the Bray–Curtis dissimilarity between pairs of the HI, LO, and NO replicates of each faecal sample based on the co-assembly data.

2.7.11. Data Analysis and Visualization

Data were analysed in R v.4.4.1 with RStudio v.2023.06.1. Data were handled using tidyverse v.2.0.0 [56] packages, including dplyr v.1.1.4, readr v.2.1.5, forcats v.1.0.0, stringr v.1.5.1, tibble v.3.2.1, tidyr v.1.3.1, and purr v.1.0.2. Data were visualised using ggplot2 v.3.4.4 [57], ggExtra v.0.10.1 [58], ggpubr v.0.6.0 [59], ggsci v.3.0.0 [60], ggh4x v0.2.6 [61], and gt v.0.11.0 [62].

3. Results

To assess differences in recovery efficiency between viruses, compare recovery of individual viruses between samples, determine an appropriate MC spike-in concentration, and assess sequence bias and faecal virome variation, three faecal samples were split into three aliquots each. One aliquot of each sample was spiked with HI, another was spiked with LO, and the last aliquot was not spiked. Viral nucleic acid extracts of each sample were processed using the WTA2 and SISPA methods for RT and PCR amplification of nucleic acids and sequenced. Sequencing of WTA2 and SISPA samples returned an average of 12 million reads per sample (Table S1). A custom bioinformatics pipeline was employed to analyse the sequencing data, including removal of host sequences, quality filtering of reads, assembly of reads into contigs, and classification of contigs (Figure 1B). After quality control, an average 3.6% of reads were removed (Figure S1). WTA2 and SISPA returned equivalent numbers of reads for most samples. However, mapping high-quality reads to MC reference

genomes (Table 2) in spiked faecal samples showed a 2- to 40-fold higher fraction of MC virus reads in WTA2 samples compared to SISPA (Table 3). In faecal samples spiked with HI, an average of 3.1% of reads in the WTA2 samples and an average of 0.72% reads in the SISPA samples mapped to MC viruses. In the HI- and LO-only samples, the percentage of MC reads was similar between WTA2 and SISPA libraries. In the blank sample (SBL), ~0.5% of total reads mapped to MC reference sequences, with >98% of those reads mapping to the four phages in WTA2 and SISPA libraries. Mapping of reads of the blank samples to the 2024-09-04 PlusPFP database (https://benlangmead.github.io/aws-indexes/k2 (accessed on 3 November 2024)) resulted in 17.8% and 23.8% of reads classified, with 83.8% and 87.1% of those reads mapping to bacteria in the WTA2 and SISPA library, respectively.

Table 3. MC virus reads in WTA2 and SISPA libraries.

Sample	Method	Total Reads (Million)			Mock Community Reads			Mock Community Reads per 10,000		
		HI	LO	No	HI	LO	No	HI	LO	No
S06	WTA2	14.43	10.38	13.51	313,522	3209	1	217	3.09	0.00
	SISPA	15.34	12.12	15.63	67,672	422	12	44	0.35	0.01
S07	WTA2	13.28	11.88	10.90	463,340	4168	12	349	3.51	0.01
	SISPA	10.38	11.98	12.27	182,285	649	9	176	0.54	0.01
S08	WTA2	12.34	9.66	10.06	478,236	1936	7	387	2.00	0.01
	SISPA	10.37	10.85	13.55	11,274	52	4	11	0.05	0.00
SMC	WTA2	9.23	4.20	-	4,271,871	67,654	-	4631	160.91	-
	SISPA	10.12	11.97	-	5,236,318	180,410		5174	150.75	-
SBL	WTA2	-	-	7.96	-	-	38,327	-	-	48.17
	SISPA	-	-	11.59	-	-	59,757	-	-	51.55

HI: high-concentration MC (1.25×10^9 particles/g faeces), LO: low-concentration MC (1.25×10^7 particles/g faeces), No: not-spiked sample, SISPA: sequence-independent single primer amplification, WTA2: whole transcriptome amplification kit, SMC: MC-only samples, SBL: blank samples.

3.1. Recovery Bias and Consistency of Mock Virus Community

To assess the recovery bias of individual MC viruses, the percentage of mapped MC reads that map to each of the individual viruses was calculated and compared to the expected percentage of reads based on the number of particles of each virus added, its total genome length, and number of strands. Phage levels were closest to the expected level in the SMC (MC with PBS) samples, particularly in the SMC-LO sample (Figure 2A). The smallest fold difference between the expected and mapped fraction of reads in WTA2 and SISPA SMC samples was 0.00, 3.52, 1.45, and 0.03, whereas the smallest fold difference in spiked samples was 0.25, 3.55, 4.39, and 0.77 for phages T5, M13, P22, and Det7, respectively. Nonetheless, there were large differences between the observed and expected levels of all viruses in all samples. Phages M13 and P22 were detected up to 22 and 7 times higher, relative to the rest of the community, than expected, while all other viruses were estimated to have a relative abundance lower than expected. Levels of the enveloped eukaryotic viruses MHV-68 and BVDV-1 were lowest, particularly BVDV-1, for which no reads were detected. The lower titre of MHV-68 in the MC may have contributed to the variation observed between samples. Although MC virus levels relative to the total MC abundance were generally the same between samples and when comparing WTA2 and SISPA, there was no systematic effect visible of the faecal samples and variation is likely due to sampling and measurement error. Exceptions were M13, which was increased, and RV-A, which was virtually absent, in SISPA libraries.

To investigate the contribution of library preparation methods to the observed variation, the abundance of MC viruses in WTA2 and SISPA libraries was compared. Overall, abundance relative to the total number of reads was higher in WTA2 libraries than in SISPA

libraries, and variation between samples was lower (Figure 2B). Comparing the abundance of the viruses in replicates spiked with HI and LO indicated a linear relationship and around a 100-fold difference between the two MCs. Some virus reads were also detected in the non-spiked aliquots, possibly indicating erroneous mapping, or mapping of reads from conserved regions of related faecal viruses. Comparing the abundances of the MC viruses in the HI-only sample WTA2 and SISPA libraries showed that abundances of all viruses, except RV-A and BVDV-1, were in close agreement (Figure 2C). In the SISPA library, RV-A abundance was lower, and BVDV-1 was not detected. Discrepancies between stock virus genome sequences and the reference sequences could have contributed to differences between expected and observed virus levels. However, substitution of reference sequences for metagenome assembled sequences did not have any meaningful effect.

3.2. Sequencing Bias in WTA2 and SISPA Libraries

Library preparation methods differ in the uniformity of genome coverage, and higher coverage uniformity reduces the sequencing depth required to recover a full genome. Differences in uniformity may therefore affect detection of viruses, particularly when depending on de novo assembly. Thus, the coverage uniformity of MC virus genomes was investigated in samples for which at least 80% of the genome was covered. For T5, P22, Det7, and M13, greater than 80% coverage was obtained in the WTA2 and SISPA libraries of at least two samples. Additionally, greater than 80% coverage was obtained for MHV-68 and segment 1 of RV-A in the WTA2 libraries for two samples. In both the WTA2 (Figure S3) and SISPA (Figure S4) libraries, the coverage depth profile, although highly non-uniform, was consistent between replicates. Coverage depth profiles of WTA2 and SISPA showed some similarity in the high-depth regions, but higher peaks were observed in the SISPA libraries. (Figure 3A). This was reflected in the coefficient of variation (CV) of the coverage depth of MC virus genomes in the WTA2 and SISPA libraries, with a mean CV for Det7, P22, T5, and M13 in all samples of 87% in the WTA2 libraries and 160% in the SISPA libraries (Figure 3B). The CV was higher for WTA2 than SISPA only for RV-A segments 6 and 10 in sample S06.

As various genomic features including GC-content can affect sequencing efficiency, the average GC-content of virus reads was compared to the genomic GC-content in samples with >20% genome coverage, as above that threshold the read GC-content did not depend on coverage (Figure S5). For DNA viruses T5 and M13 with a genomic GC-content of <41%, the mean read GC-content was higher than the virus genome GC-content (Figure 3C), while for phages P22 and Det7, the GC-content was close to the actual genomic GC-content, and for MHV-68, with the highest genomic GC-content (47%), the read GC-content was below the genomic GC-content. For all DNA viruses, the read GC-content was more consistent between replicates for the SISPA libraries than the WTA2 libraries. For the dsRNA virus RV-A, the genomic GC-content of genome segments varied from 28.9% to 40.2%. In the case of the WTA2 libraries, for the segments with a GC-content <36%, the read GC-content in most replicates was equal to or higher than the genomic GC-content (Figure S6). On the other hand, for the segments with a GC-content >36%, the read GC-content was equal to or lower than the genomic GC-content in most samples. In the case of the SISPA libraries, only two segments were sequenced to sufficient coverage and no consistent pattern could be determined.

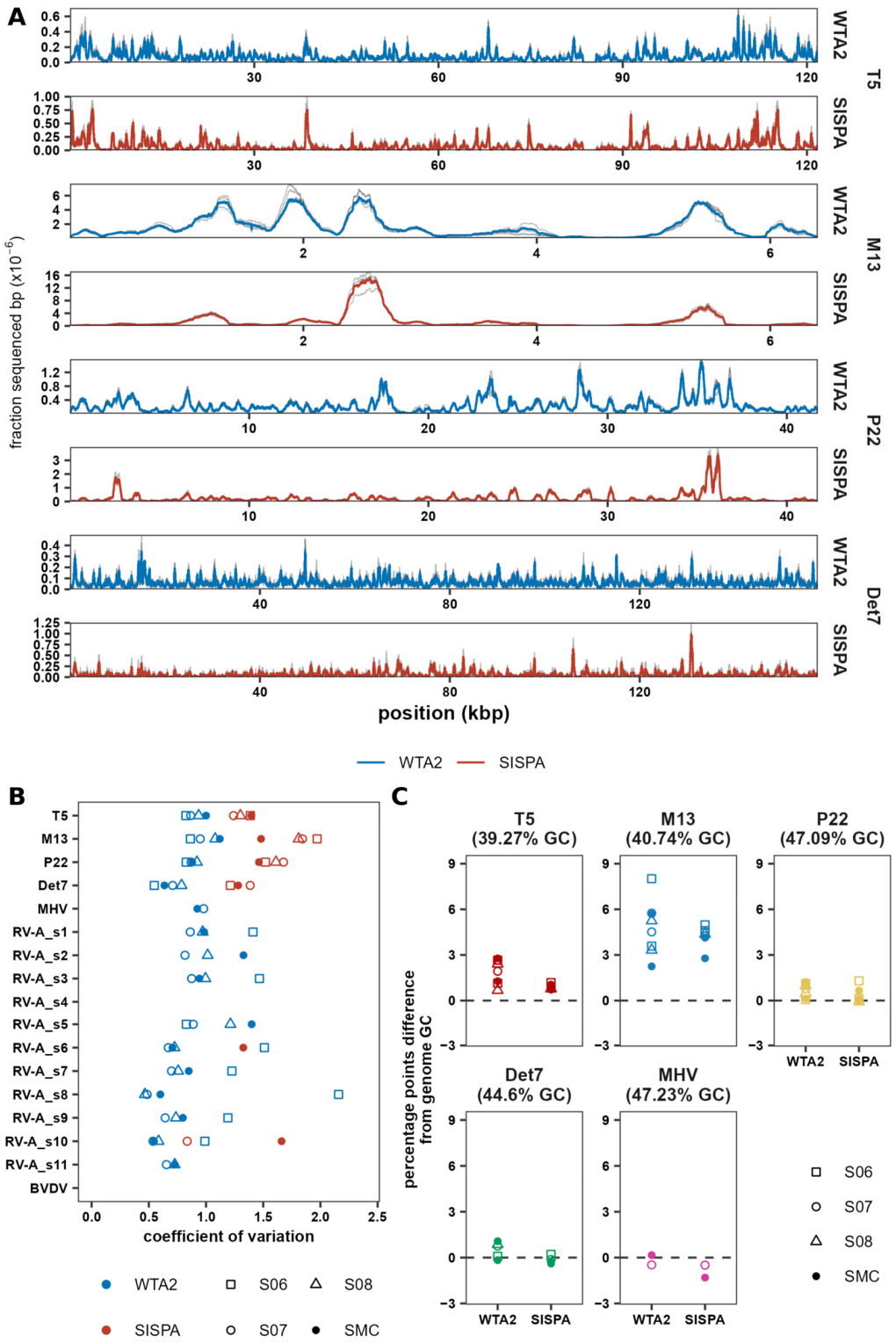

Figure 3. Sequencing depth uniformity and GC bias differ between WTA2 and SISPA. (**A**) Comparison of the normalised sequencing depth of T5, M13, P22, and Det7 in WTA2 and SISPA libraries. The grey lines show the normalised depth of samples in which at least 80% coverage was achieved. Thick blue and red lines depict the means of those samples. (**B**) Coefficient of variation of sequencing depth in SMC-HI and samples spiked with MC HI. (**C**) Difference in the average GC-content of reads mapping to reference genomes of viruses with a non-segmented genome, compared to the actual GC-content of the virus genome, for samples with >20% genome coverage.

3.3. Assembly Quality of WTA2 and SISPA Libraries

To determine differences in assembly quality between WTA2 and SISPA, reads from individual libraries were assembled. Assembly of WTA2 libraries consistently yielded more contigs overall and more contigs >10 kb (Figure 4A). The number of contigs in the assemblies of the spiked sample SISPA libraries was between 22% and 81% lower than the WTA2 library co-assemblies. The N50 was higher in every case for WTA2 than for SISPA library assemblies, indicating a higher overall quality of assemblies (Table S2). Additionally, the largest contig in samples S06-HI and S08-HI were 125 kb and 100 kb in the WTA2 co-assemblies and 77 kb and 33 kb in the SISPA assemblies, respectively. Cumulatively, WTA2 assemblies contained 8.8–58.7 Mbases, while SISPA assemblies contained 1.4–26.0 Mbases and the cumulative length of SISPA assemblies was 37–88% lower than WTA2 libraries (Figure 4B). The average length of the 250 longest contigs in the WTA2 assemblies was between 1.9 and 4.9 times longer than SISPA assemblies, suggesting increased fragmentation and reduced diversity in the SISPA library assemblies. On the other hand, the mean coverage depth of contigs in the SISPA assemblies was between 1.6 and 18.0 times higher than WTA2 library assemblies and the overall GC-content of assemblies was more consistent in SISPA replicate libraries of the same faecal sample than in WTA2 libraries (Table S2).

Next, geNomad [54] was used to identify viral sequences, while MC sequences were identified by aligning sequences to MC reference genomes. In assemblies and co-assemblies of both WTA2 and SISPA libraries, a large fraction of contigs could not be identified as viral, indicating the presence of non-viral nucleic acid and potentially false negatives. Classification of reads mapping to these contigs using Kraken2 with the 2024-09-04 PlusPFP database determined an enrichment of bacterial reads, suggesting a bacterial origin for these contigs, with the fraction of bacterial reads increased in SISPA libraries compared to WTA2 libraries for each of the samples (Figure S8). Across faecal samples and replicates, and ignoring mock community contigs, 3.4% and 3.5% of contigs were classified as viral in WTA2 and SISPA assemblies, respectively. The relative abundance of the virus sequences was calculated by mapping reads of individual samples to the contigs of the respective assemblies. Faecal viruses contributed on average 33.1% and 6.3% of the relative abundance in the WTA2 and SISPA libraries, respectively. While SISPA assemblies had a higher proportion of MC virus contigs, the cumulative length of MC contigs was lower, indicating increased fragmentation of MC virus genomes. The collective abundance of MC contigs relative to all viral contigs was similar for both WTA2 and SISPA in samples spiked with HI, of between 0.7% and 4.6%. On the other hand, the relative abundance of contigs not identified as viral was higher in the SISPA (93%) than the WTA2 (83%) assemblies. Finally, while WTA2 (co-) assemblies had a higher fraction of viral contigs, a higher fraction of co-assembly contigs were of medium quality or higher as well for samples S06 and S07 (Figure 4C). Although a higher fraction of viral sequences in the SISPA co-assembly of sample S08 were complete genomes, the total number of complete genomes was lower in the SISPA assembly (n = 6) than the WTA2 assembly (n = 21). Together, these data suggest increased fragmentation of SISPA assemblies leading to reduced identification of virus sequences.

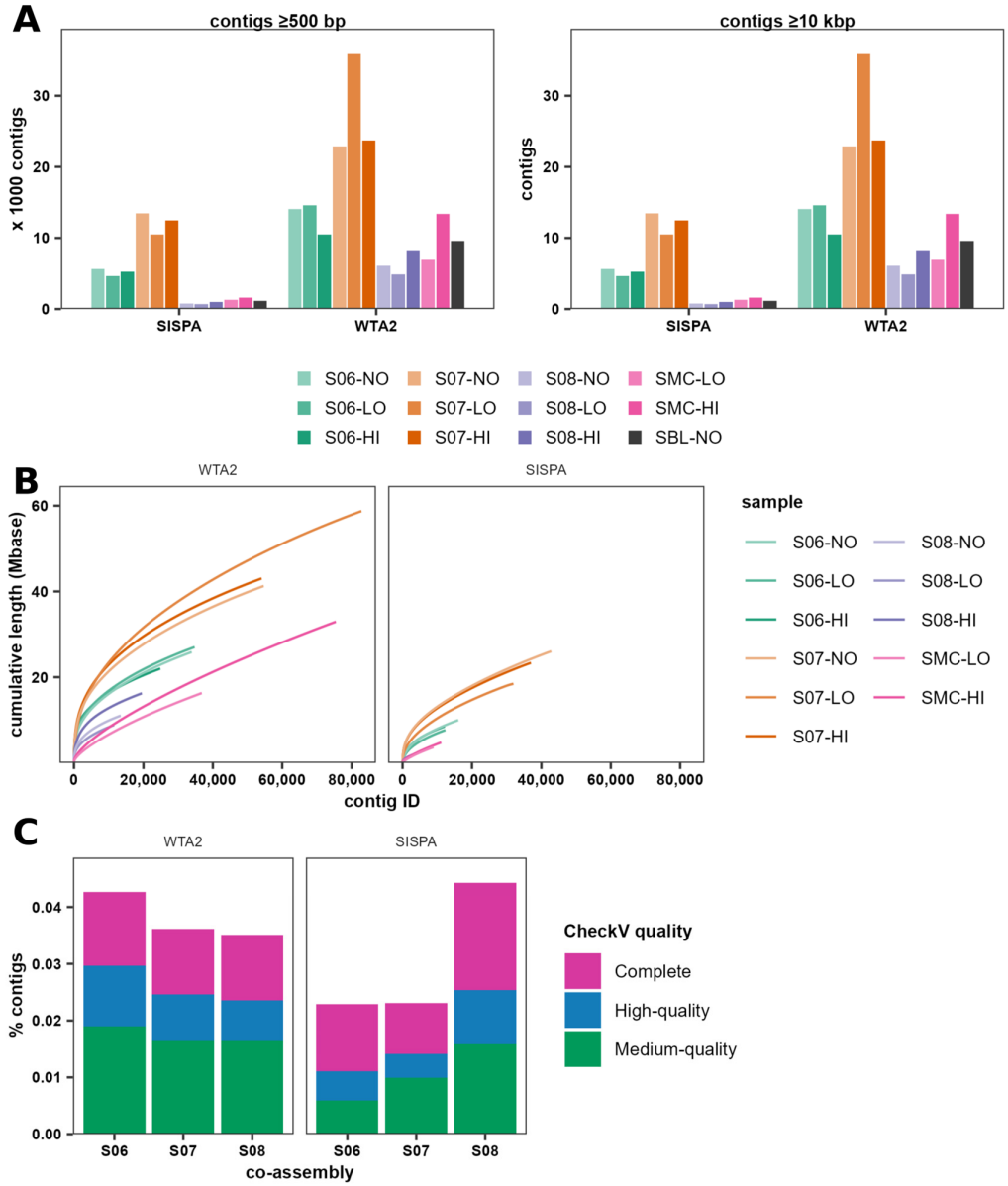

Figure 4. WTA2 samples yield more and larger contigs than SISPA. (**A**) Total number of contigs and number of contigs ≥10 kbp in single assemblies of each of the samples after removing redundant sequences. (**B**) Cumulative assembly length of single assemblies. (**C**) Percentage of contigs in WTA2 and SISPA co-assemblies with representing complete, high-quality, and medium-quality genomes according to CheckV.

3.4. Accuracy of Assembly-Based Virus Abundance

To investigate the accuracy of relative abundance estimates based on de novo assembly, the relative abundance of MC contigs was compared to the MC virus relative abundances calculated using the reference genome sequence. In the assemblies of WTA2 libraries, samples spiked with HI yielded between 1 and 38 MC contigs for phages Det7, P22, and M13 (Figure 5A), all collectively covering >98% of the reference genome, while phage T5 contigs covered 15%, 63%, and 25% of the genome in samples S06, S07, and S08, respectively. For phages T5, Det7, and P22, the SISPA libraries yielded more contigs, while genome coverage of contigs was lower for T5 and Det7. Coverage for Det7 varied between 1.0% and 65%, indicating reduced fragmentation in the WTA2 library assemblies.

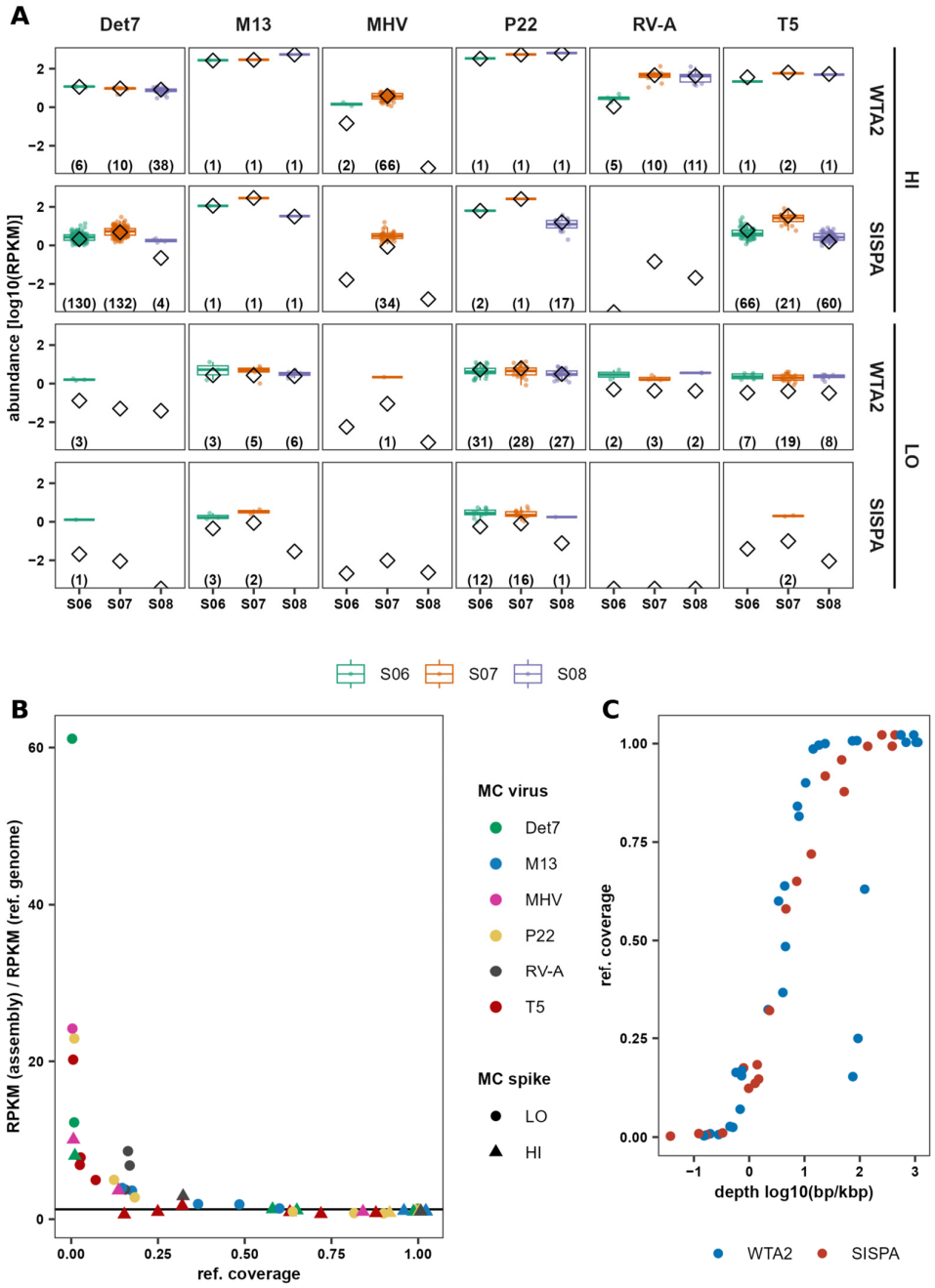

Figure 5. Assembly-based abundance is overestimating actual abundance when contig coverage of the genome is low. (**A**) The boxplots show the abundance of MC virus contigs in the assembly calculated as RPKM, which is compared to the calculated virus abundance based on reads mapping to the reference genomes in RPKM (diamonds). Numbers along the *x*-axis represent the number of contigs in the assemblies. (**B**) Fold difference between the weighted average abundance of MC virus contigs [RPKM (assembly)] and the MC virus abundance based on read mapping to the reference genome [RPKM (ref. genome)], compared to the total genome coverage of MC virus contigs (ref. coverage). The horizontal line shows a ratio of 1.25. (**C**) Percentage of the reference genome covered by genome fragments in the assemblies (ref. coverage) as a function of the sequencing depth. The depth is calculated from the number of sequenced bases per thousand base pairs of genomes.

In the samples spiked with LO, assemblies contained more contigs for each virus, together covering a lower portion of the genome, showing higher fragmentation of the genome in these assemblies. Additionally, several samples did not yield any contigs for Det7, M13, MHV-68, RV-A, and T5 in the WTA2 and SISPA libraries. This suggests reliable assembly-based detection of viruses using this method requires >10^5 particles in

50 mg of a faecal sample. For both the WTA2 and SISPA libraries, contigs of LO sample assemblies collectively covered lower portions of the reference genomes than those of HI sample assemblies.

Among genome fragments matching the same genome, there were considerable differences in relative abundance. Across libraries, there was up to a 24-fold difference between the fragment with the highest and lowest abundance. Calculating the virus abundance using collections of genome fragments ($RPKM_{assembly}$) only produced estimates close to the reference genome-based abundance ($RPKM_{ref}$) if the fragments collectively covered a sufficient portion of the reference genome (Figure 5B). When contigs collectively covered ≥80% of the reference genome, the $RPKM_{assembly}$ was close to the $RPKM_{ref}$ with on average 6% higher estimated abundance compared to $RPKM_{ref}$. However, abundance was considerably overestimated when contigs covered a smaller portion of the reference genome (Figure 5B). At a maximum overestimation of 100%, 50%, and 10%, contigs collectively covered at least 32%, 58%, and 63%, respectively, of the genome.

As the sequencing depth of the virus genomes was not uniform (Figure 3A), some areas of the genome have a higher probability of producing reads and thus assembling into contigs. Indeed, WTA2 libraries have a more uniform sequencing depth and have a slightly higher coverage for the same number of sequenced bases than SISPA libraries (Figure 5C). Thus, for a given sequencing depth, a more uniform depth profile will lead to increased coverage of the genome, whereas for a more heterogeneous depth profile, reads will accumulate on a shorter area of the genome (Figure S7), which inflates the estimated abundance. This is also supported by the greater number of duplicate reads in SISPA libraries compared to WTA2, with at most 14% of reads duplicated in WTA2 libraries, against 29% in SISPA (Figure S9).

3.5. Consistency of Assembly-Based Faecal Virus Abundance

We next sought to investigate the consistency of the recovery of faecal viruses. For this, co-assemblies were generated of the three aliquots (spiked with HI, LO, and untreated) of each of the faecal samples. For this analysis, these aliquots were regarded as technical replicates as the spiking treatment was assumed to not affect the faecal virome. The abundance of faecal virus contigs in each of the replicates was determined by mapping the reads of the individual replicates to the respective co-assemblies. Comparing the abundance of viral contigs between HI and LO replicates, variation in estimated abundance decreased with increasing abundance (Figure 6A), as would be expected. In samples S06 and S07, a large group of unclassified contigs were present at 10- to 100-fold higher concentrations in the LO replicate. This difference was particularly pronounced in the SISPA libraries of sample S07. MC virus relative abundance in the HI replicate was around 100-fold higher than in the LO replicate, consistent with the difference in concentration between the MC spike-ins. To analyse the variation in abundance between replicates, only viral contigs, but not MC contigs, were subsequently used.

To analyse the variation in estimated abundance for each contig between samples, the abundance of contigs that were present in all three replicates was compared between the HI, LO, and No replicates and the fold difference between the highest and lowest abundance was calculated. While longer contigs tended to have higher coverage depth, the variation in abundance was limited mostly by coverage depth, not contig length (Figure S9). Nearly all contigs with at least a 4-fold difference between the highest and lowest abundance had a mean coverage depth of <10X. For contigs with ≥10X coverage, the median ratio between the highest and lowest abundance was 1.7 and 1.4 for the WTA2 and SISPA co-assemblies, compared to 2.5 and 2.4 for contigs with <10X coverage, respectively. Contigs with a mean

coverage depth <10X will be referred to as low coverage depth (LCD), while contigs with a mean coverage depth ≥10X will be referred to as high coverage depth (HCD).

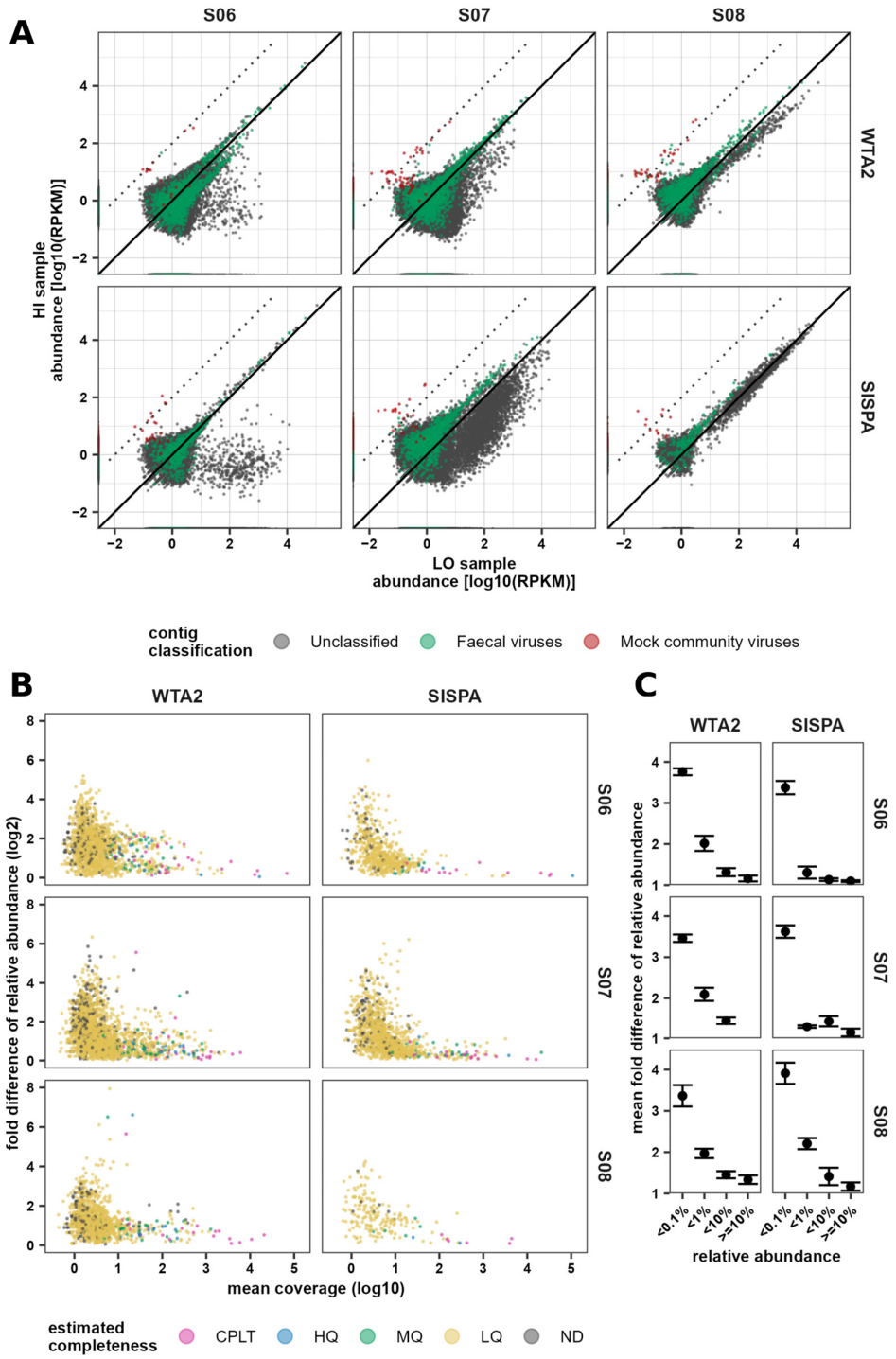

Figure 6. Concordance in virus abundance estimates between replicates within faecal samples is highest for high abundance contigs and medium- to high-quality genomes. (**A**) Comparison of normalised read counts (RPKM) of stool sample spikes with MC HI and LO. The black solid line depicts perfect agreement. The black dotted line indicates a 100-fold increase in the HI sample over the LO sample. (**B**) Log2 fold difference between the highest and lowest relative abundance of viral contigs across the three replicates, compared to mean coverage depth. (**C**) Variation in relative abundance measured as the mean fold difference between the highest and lowest relative abundance of contigs binned by a mean relative abundance <0.1%, between ≥0.1% and <1% (<1%), between ≥1% and <10% (<10%), and ≥10% (≥10%).

Analysing the fold difference in relative abundance between the highest and lowest abundance of contigs between replicates showed that 83% and 74% of contigs were LCD contigs in WTA2 and SISPA co-assemblies, respectively. Comparing the fold difference of LCD and HCD contigs showed a median of 2.5 for LCD contigs in WTA2 and SISPA samples, compared to 1.7 and 1.5 HCD contigs of WTA2 and SISPA samples, respectively (Figure 6B). Thus, while WTA2 co-assemblies produced more HCD contigs than SISPA co-assemblies, HCD contigs on average had higher variation in WTA2 library co-assemblies. Ranking contigs based on abundance and comparing the ranking between replicates showed a similar pattern, with the difference between highest and lowest ranking greatly increased for LCD contigs (Figure S10).

While variation was largest for LCD contigs, they collectively only contributed between 0.37% and 2.4% of faecal viral reads across all libraries, corresponding to between 1.00% and 10.5% of relative abundance. Across replicates, LCD contigs represented 3.1%, 9.6%, and 5.4% viral relative abundance in WTA2 libraries and 1.2%, 5.1%, and 5.2% in SISPA libraries for samples S06, S07, and S08, respectively. Additionally, all contigs with a relative abundance $\geq$0.1% were HCD contigs. Correspondingly, the range in abundance-based ranking was at most four places in both the WTA2 and SISPA libraries for contigs with an average relative abundance $\geq$1%, except WTA2 sample S07. Additionally, there was a median 1.5- and 1.3-fold difference between the highest and lowest relative abundance in the WTA2 and SISPA libraries, respectively, for contigs with $\geq$1% average relative abundance (Figure 6C). This means that for the higher abundance contigs, the abundance ranking of contigs with a relative abundance $\geq$1% is consistent across replicates. Nonetheless, there was up to a 50% difference between the highest and lowest relative abundance, from only three replicates, indicating considerable uncertainty in quantitative abundance estimates even in the best case. Most contigs that qualified as complete, high quality, or medium quality were HCD contigs, with 99%, 94%, and 81% contigs, respectively, having HCD. In SISPA co-assemblies, a higher fraction of medium-quality contigs were HCD contigs (88%) than in WTA2 co-assemblies (79%). Nonetheless, even among contigs of medium quality and above, the fold difference of some contigs was high, particularly in the WTA2 co-assemblies. Only 50%, 42%, and 37% of complete, high-quality, and medium-quality contigs, respectively, had at most a 1.5-fold difference in WTA2 samples, versus 92%, 93%, and 59% of those contigs, respectively, in SISPA samples. Taking all HCD contigs, two randomly selected relative abundances from the three replicates were at most 50% apart only 73.0% and 83.4% of the time in the WTA2 and SISPA libraries, respectively (Table S3), and measurements were at most 10% apart for only 15.7% and 31.1% of the time in the WTA2 and SISPA libraries, respectively. This suggests that if a difference of more than 50% abundance is measured between two samples, this could be due to sampling and measurement error alone around 30% and 20% of the time, respectively, for WTA2 and SISPA. Furthermore, the observed difference in abundance was 140% for WTA2 and 110% for SISPA in fewer than 5% of measurement pairs, suggesting that differences greater than this can be more reliably attributed to genuine differences between samples instead of measurement error.

3.6. Taxonomic Analysis and Sample Diversity

Lastly, the taxonomic content of the individual sample assemblies (Figure 7A) of the WTA2 and SISPA libraries was determined using the taxonomic classifications generated by geNomad, excluding MC sequences. Virtually all contigs that were identified as viral could be classified at the class level, although the classification rate at lower clades was considerably lower, with on average only 10.7% and 10.1% of viral sequences classified at the order and family levels, respectively. Overall, the phyla of *Phixviricota*, *Uroviricota*,

Kitrinoviricota, *Cressdnaviricota*, and *Pisuviricota* were the five most abundant among all samples in both libraries. The phylum *Phixviricota* contains phages with small protein capsids and small circular ssDNA genomes, while *Uroviricota* are tailed phages with large dsDNA genomes. While phages of the phylum *Phixviricota* were the most abundant viruses in samples S06 and S08 for both WTA2 and SISPA libraries, abundance of *Phixviricota* was greater in the SISPA libraries than WTA2 libraries. For sample S07, the difference between the libraries was much greater, with a mean relative abundance of 22.7% and 58.4% in the WTA2 and SISPA libraries, respectively (Table S4). Members of the phyla *Pisuviricota* (ssRNA(+) genomes), *Cressdnaviricota* (circular ssDNA genomes), and *Kitrinoviricota* (ssRNA) infect eukaryotic cells. In sample S07, *Kitrinoviricota* and *Cressdnaviricota* were more abundant in the SISPA library than the WTA2 library. *Kitrinoviricota* and *Cressdnaviricota* had 4.1% and 4.3% relative abundances, respectively, in the SISPA library, while the WTA2 libraries of sample S07 contained 0.28% and 0.55% of these phyla, respectively. For *Cressdnaviricota*, the pattern was reversed in samples S06 and S08, with 0.067% and 0.16% relative abundances in the WTA2 library, and 0.014% and 0.028% in the SISPA libraries, respectively. Such large differences were not found for *Kitrinoviricota* in the other two faecal samples. Lastly, the phylum *Pisuviricota* had a consistently higher relative abundance in the WTA2 libraries than the SISPA libraries of all three faecal samples.

The ratio between the highest and lowest relative abundance was lowest for the exclusively prokaryotic virus phyla, with 1.05–1.42 and 1.02–1.71 for *Phixviricota* and 1.09–1.40 and 1.51–2.10 for *Uroviricota* in the WTA2 and SISPA libraries, respectively. For the largely eukaryotic virus phyla *Pisuviricota*, *Cressdnaviricota*, and *Kitrinoviricota*, the ratio was higher, and the ratio could not always be determined due to their absence in at least one of the replicates. The relative abundance of these phyla was <0.1% in most samples, and the variation was between 0.0027 and 4.3 percentage points. Thus, the relative abundance of the prokaryotic phyla *Phixviricota* and *Uroviricota*, which typically represent the bulk of the virome, are robust, compared to *Pisuviricota*, *Cressdnaviricota*, and *Kitrinoviricota*, which is likely due to the low relative abundance of these taxa in the samples.

To determine differences in virome richness between WTA2 and SISPA, and variation between replicates, the Chao1 index was used and showed a greater richness in the WTA2 libraries than the SISPA libraries (Figure 7B). The mean CV of species richness in the WTA2 and SISPA libraries was 16% and 20%, respectively. Alpha diversity was measured by the Shannon and Simpson indices with WTA2 libraries having greater evenness than the SISPA libraries. Using the co-assemblies of each the WTA2 and SISPA libraries of the base faecal samples, the Bray–Curtis similarity index was calculated between each of the replicates to determine the variation in overall sample composition between replicates of each faecal sample (Figure 7C). For all three faecal samples, similarity was higher in the SISPA libraries than in the WTA2 libraries, while the samples with the highest and lowest similarity were the same in the WTA2 and SISPA co-assemblies. Interestingly, the level of similarity corresponded to the percentage of LCD contigs in each co-assembly.

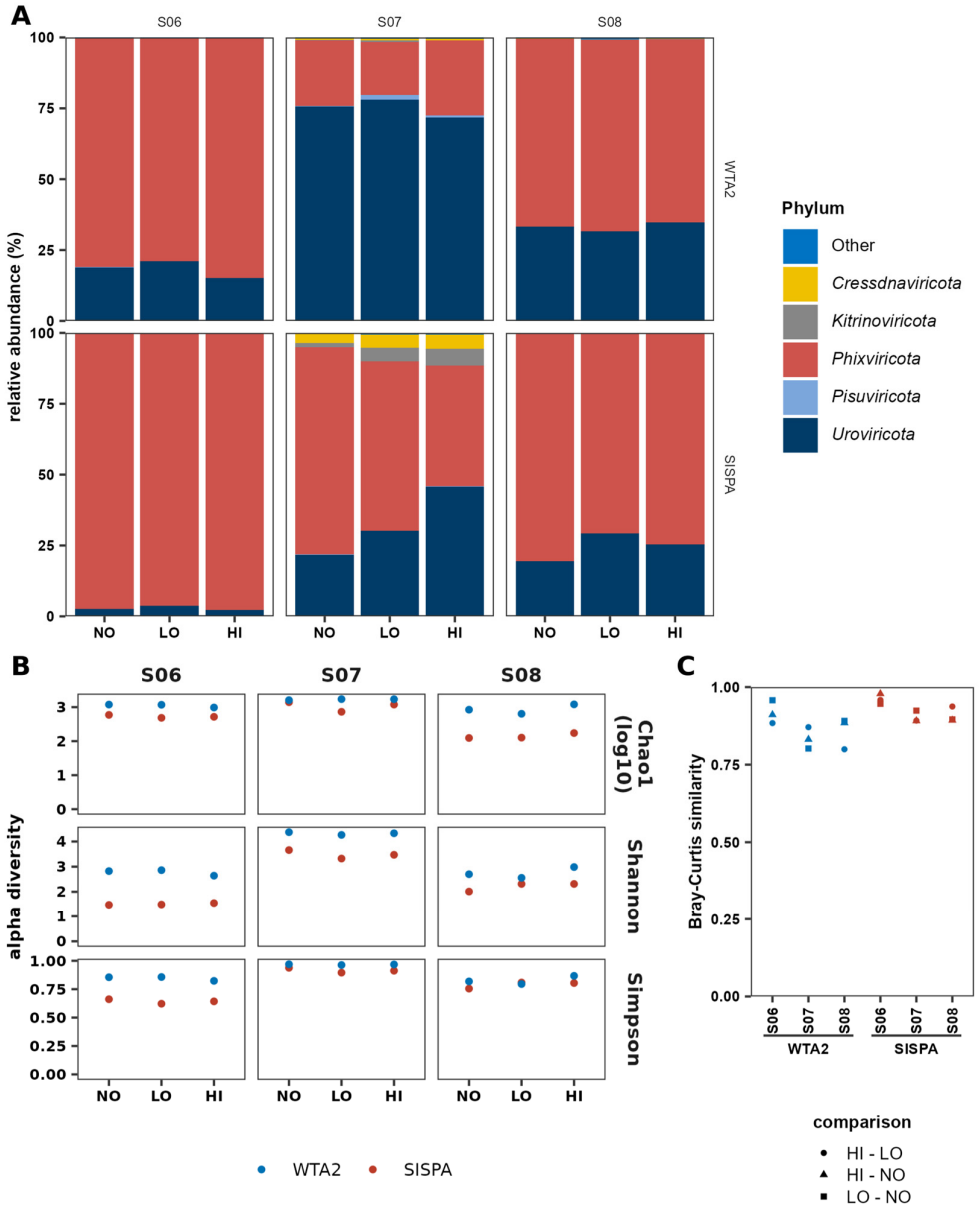

Figure 7. Difference in taxonomy and diversity between WTA2 and SISPA. (**A**) The five phyla with the highest relative abundance in any of the samples are shown. (**B**) Sample alpha diversity, as calculated by the Chao1 index for species richness and the Simpson and Shannon indices for species evenness. (**C**) Beta diversity, calculated through the Bray–Curtis similarity index of the replicates of each stool sample.

4. Discussion

We have assessed the bias and reproducibility of virus metagenomes obtained using two widely used methods of transcriptome amplification, WTA2 and SISPA, using an MC of diverse viruses to spike human faecal samples. In comparing the recovery of individual viruses, genome coverage depth profiles, contigs and assemblies, coverage depth and relative abundance of viral sequences, and taxonomy across replicates, WTA2 provides higher sensitivity, at the cost of higher variability, whereas the SISPA method, by comparison, is less sensitive and accurate but is more consistent. While WTA2 is a proprietary kit, SISPA uses commonly used reagents and is thus less costly and less sensitive to supply chain issues with reagents being substituted more easily.

4.1. Recovery of Individual Spiked Viruses from a Mock Community

The level of recovery of individual MC viruses varied between samples, which is larger for Det7 and MHV-68 than for T5 and P22. Overall, patterns were comparable between the SISPA and WTA2 libraries. Differential recovery across faecal samples of MC viruses could be due to differences in the faecal sample composition. For example, some phage capsids contain Ig-like domains that help attach to the intestinal mucus [63], including phages T5 [64] and P22 [65]. Differences in, for example, mucus content, could produce different interactions between viruses and faecal material that affect downstream recovery VLPs. Moreover, inherent differences in individual faecal samples may affect aggregation or adherence to plasticware and filter membranes, which is suggested when comparing the percentage of Det7 and P22 reads of the MC-only samples to the spiked samples (Figure 1A), with the levels in the MC-only samples being much closer to the expected value. Additionally, differences between replicates may be exaggerated in Figure 1A, as variation in the percentage of reads mapping to one virus will automatically affect the percentage of reads mapping to all others. Moreover, for the samples spiked with LO, around 100-fold fewer reads mapped to MC viruses and were more likely to be affected by noise.

In the WTA2 libraries of spiked samples, the difference between MC phage relative abundances in samples spiked with HI and LO reflected the difference in concentration of those viruses in HI and LO, in line with previous results [12]. However, the number of reads corresponding to eukaryotic viruses in the WTA2 libraries and all viruses in LO-spiked sample SISPA libraries was low and relative abundances did not reflect the difference between HI and LO in these cases. This is likely a result of increased noise associated with low read counts.

Individual viruses also differ in their recovery efficiency, with the abundance of M13 and P22 far exceeding expected levels. Phage concentrations were determined using epifluorescence microscopy, as in a previous study by Kleiner et al. [13]. They also found an increased recovery of P22 [13], suggesting a higher nucleic acid extraction efficiency of this phage. Differences in nucleic acid extraction efficiency between viruses and nucleic acid extraction kits have been observed before using serum, respiratory, faecal, and environmental samples [66–70]. Indeed, the extraction efficiency of some dsDNA phages is only 24–30% [23]. Structural proteins of the RNA viruses RV-A (VC2) and BVDV-1 (core) may inhibit extraction, which can be mitigated by the inclusion of proteases in the extraction kits [71,72]. The kit used here has no proteinases listed as a component of the lysis buffer and nucleic acid extraction kit [66], making it possible that the presence of viral core proteins could reduce the efficiency of viral nucleic acid recovery. Experimentally determined recovery efficiencies can be used to correct virus abundance in sequencing data [23], although based on our results, the question remains to what degree the recovery efficiency of one virus can be extrapolated to other viruses. The low recovery of eukaryotic viruses, particularly enveloped viruses MHV-68 and BVDV-1, may be due to the formation of aggregates and loss during centrifugation and/or due to the adsorption of faecal material to plasticware and filters used in their isolation leading to the loss of virus particles [14]. Indeed, two enveloped eukaryotic viruses, herpesvirus and coronavirus, were sensitive to centrifugation [12]. Another possible source of bias is the use of nucleases during the VLP extraction process, as some virus particles are susceptible to RNase [73]. DNA library preparation methods may also affect the representation of individual viruses, as WTA2 and SISPA show different relative abundances for several ssDNA and RNA virus taxa. The viruses of the phyla *Phixviricota* and *Cressdnaviricota*, as well as MC virus M13 (*Hofneiviricota*), are viruses with circular ssDNA genomes, and both M13 and faecal *Phixviricota* viruses have higher relative abundance in the SISPA than the WTA2 libraries of all three samples,

as well as a large increase in *Cressdnaviricota* in one sample. Another striking difference is the nearly complete absence of RV-A in the SISPA libraries. While the SISPA libraries were produced several months after the WTA2 libraries, and thus, long-term storage of the nucleic acid extracts should be considered, this suggests that amplification of RV-A dsRNA by SISPA is less efficient than WTA2.

4.2. Effects of Sequence Bias

Sequence bias is a phenomenon of many sequencing methods [74]. Although both WTA2 and SISPA produce non-uniform coverage depths, the profiles are highly consistent in the WTA2 and SISPA libraries. While there are similarities in the depth profiles of WTA2 and SISPA, they ultimately produce different profiles, with higher peaks in some regions leading to reduced uniformity in SISPA (Figure 3A). This likely contributes to the higher assembly quality of the WTA2 libraries. With equivalent numbers of reads in the SISPA and WTA2 libraries, WTA2 yields more and longer contigs in the WTA2 library assemblies. Together with the fact that MC phages had more contigs in the MC HI SISPA libraries than the WTA2 libraries, with equal or lower coverage, this is indicative of higher fragmentation, which has been reported previously for SISPA [14]. Decreased uniformity of the sequence depth profiles means that certain regions of the genome are more likely to produce reads and contigs, and higher sequencing depth is required to fully recover the genome [75]. This may explain a higher fraction of contigs being identified as viral by geNomad in the WTA2 assemblies. Thus, a reduction in sequence bias and an increase in coverage depth uniformity will produce higher quality assemblies and enhance virus discovery.

Sequence bias is introduced by several factors that lead to preferential amplification of certain genomic regions. For instance, GC-content affects various DNA polymerases to various degrees [76], tag sequences and the random nucleotide length of tagged random primers and bases downstream of the binding site influence amplification [77,78], DNA polymerases have different sequence preferences [79], and the transposase used in the Illumina DNA Prep kit has a preferred sequence motif [80]. Optimisation of the WTA2 protocol is not straightforward due to the proprietary nature of kit components. Use of longer random nucleotide sequences in the tagged primers and using multiple primers with different tag sequences may increase uniformity of the coverage depth [77]. However, incorporation of multiple primers with twelve random nucleotides did not improve coverage depth uniformity in an oral virome study [15]. Our results suggest WTA2 produces more uniform profiles despite using random hexamers, while random octamers were used for SISPA. Nevertheless, optimisation of SISPA primers could be explored. WTA2 included 17 cycles of PCR amplification, while SISPA had 30 cycles, possibly amplifying any bias inherent in the method. For WTA2, increasing PCR cycles past 17 has been previously shown not to improve the yield of viruses [12] and a recent benchmarking study showed a reduced total assembly length when using 30 amplification cycles, compared to 15 or fewer cycles [30]. Thus, while it may require higher input volumes, a reduced number of cycles should be considered for SISPA.

4.3. Precision and Reliability of Abundance Measures

While most faecal viral sequences show large differences between the lowest and highest abundance across three replicates, most of these contigs had an individual relative abundance <0.1% and collectively made up between 1% and 10% of relative abundance in samples. This is likely because sequences with lower relative abundance have lower read counts and low-count data are inherently more noisy. Thus, high variability in these methods is restricted to a minority of sequences, and using a minimum coverage of 10X removed most of these sequences. Nonetheless, even for contigs with high coverage, some

had >4-fold difference between the highest and lowest abundance across three replicates, and for contigs with at least 100X coverage, the median fold difference was still 1.5 and 1.2 on average across the three faecal samples in the WTA2 and SISPA libraries, respectively. SISPA libraries produced the most consistent abundance.

Relative abundance of virus sequences at the class level was consistent, except for *Phixviricota* in the SISPA library of sample S07. Only a small fraction of viral sequences could be assigned an order or family, preventing assessment of variability for those ranks. Although WTA2 assemblies have reduced fragmentation and increased diversity, the rate of classification at the order and family levels was similar for both SISPA and WTA2.

MC virus genomes were represented in the assemblies by several contigs that can have up to 30-fold difference in abundance. Theoretically, contigs in an assembly have a coverage of at least 1X, whereas the actual genome coverage can be much lower, leading to an overestimation of the abundance of the virus. Provided that contigs together sufficiently cover the virus genome ($\geq$63%), the abundance is overestimated by up to only 10%, whereas a coverage of <32% leads to a doubling of the calculated abundance. The accuracy of the calculated virus abundance therefore benefits from grouping of contigs into genome bins that belong to the same virus, as well as estimation of the genome completeness. The former can be accomplished by virus genome binning tools like PHAMB [81], vRhyme [82], MetaBAT 2 [83,84], CoCoNet [85], and Phables [86], and the assembly contiguity improvement tool COBRA [87], whereas the latter can be performed by CheckV [52].

4.4. MC Spiking Concentration

The use of a virus MC has provided insights into the consistency of virus recovery across samples, biases in virus recovery, sequence bias, and accuracy and precision of calculated abundance. The spike concentration of MC HI was sufficient to recover full genomes for most dsDNA viruses in the assemblies of WTA2 libraries. However, SISPA libraries contained fewer reads, particularly for the smaller RNA viruses. Therefore, a higher spiking concentration is recommended, in the order of 1×10^8 particles of each virus per 50 mg faeces. Alternatively, an MC in which virus concentrations are adjusted for the nucleotide content of the genome should in theory produce equal read numbers for each virus.

4.5. Recommendations

The data indicate that while both WTA2 and SISPA each have their biases, these biases are systematic and thus will allow valid comparisons of samples. Nonetheless, care should be taken when comparing samples, as more than two-fold differences between abundance measurements of sequences across three replicates were observed in more than 5% of contigs. For optimal precision of abundance measurement, contigs with a mean depth of at least 10X should be used, and sequences with at least 50% completeness are required for accurate estimation of abundance. Completeness can be achieved using various binning tools, with Phables previously producing the fewest chimeric and duplicate sequences, compared to VAMB [88] and vRhyme [89]. The choice between SISPA and WTA2 depends on the requirements of the researcher. Based on our results, SISPA is more cost effective and provides higher precision, while WTA2 provides higher assembly quality and diversity and thereby allows for improved virus discovery. During the writing of this manuscript, an alternative method was published based on the direct ligation of sequencing adapters to cDNA [29], which showed more accurate representation of DNA phages from a DNA phage MC, compared to MDA. Future comparison of this method to WTA2 and/or SISPA on DNA as well as RNA viruses will be valuable.

4.6. Limitations

While the WTA2 and SISPA libraries were generated from the same viral nucleic acid extracts, other sources of variation need to be considered. These include sample storage times, which for the SISPA libraries were longer, with Illumina libraries being prepared on separate occasions and sequenced on different sequencing platforms. An effect of each of these differences on the final data cannot be excluded. A sizeable portion of the sequencing data was attributable to sequences of unknown origin. These were especially pronounced in the SISPA libraries. Reads mapping to these contigs were enriched for bacterial reads, although a subset of these contigs may be viral sequences that were not recognised by geNomad. Additional virus detection tools, improved assembly contiguity and genome binning of contigs, and comparing non-recognised sequences to microbial and other sequence databases can help identify the source of these sequences. This would show whether SISPA libraries contain more contaminations, or whether the increased number of unclassified sequences is due to increased fragmentation hampering identification of virus sequences.

5. Conclusions

Using a set of faecal samples analysed in triplicate and spiked with a mock viral community, the reproducibility and bias of a virome sequencing method has been assessed, including the comparison of two methods for the reverse transcription and amplification of viral nucleic acid. The results show that individual viruses have different recovery efficiencies, and that recovery of individual viruses varies between replicates. Our comparison of WTA2 and SISPA shows that the WTA2 kit provides higher sensitivity, at the cost of higher variability. SISPA on the other hand is less sensitive and accurate but more consistent. While WTA2 is a proprietary kit, SISPA uses commonly used reagents and is thus less costly and less sensitive to supply chain issues since reagents can be substituted more easily. Future improvements to the SISPA protocol should be investigated to increase sensitivity, particularly to RNA viruses, as this would make SISPA a competitive alternative to WTA2. Incorporation of a mock viral community was key in this analysis and is a valuable control to be included in metagenomic studies [90,91]. The bioinformatics pipeline used in this study is available online.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/v17020155/s1, Figure S1: WTA2 and SISPA yield equivalent numbers of high-quality reads; Figure S2: SISPA libraries have a higher rate of duplicate reads than WTA2 libraries; Figure S3: Consistency of the sequencing depth profiles of virus genomes in spiked and MC-only samples processed using WTA2; Figure S4: Consistency of the sequencing depth profiles of virus genomes in spiked and MC-only samples processed using SISPA; Figure S5: Above 20% coverage, no relationship between difference in GC-content and coverage is detectable; Figure S6: Difference between GC-content of RV-A reads and genome segments differs by segment; Figure S7: A more uniform sequencing depth profile in WTA2 leads to improved genome coverage and lower overestimation of relative abundance; Figure S8: Contigs of unknown origin are enriched for bacterial reads; Figure S9: Most contigs with high variation are short contigs with low coverage; Figure S10: Variation in abundance-based ranking of contigs; Table S1: Number of reads per sample; Table S2: Assembly statistics for the WTA2 and SISPA libraries of spiked stool samples and MC-only samples; Table S3: Percentage of contigs for which the relative abundance (RPKM) of two randomly selected replicates falls within the a given percentage; Table S4: Variation in relative abundance at the phylum level between replicates of stool samples, based on individual replicate assemblies.

Author Contributions: Conceptualization, S.R.C., P.P.P., E.M.A., R.H. and O.J.C.; methodology, R.H., P.P.P., D.B. and O.J.C.; software, R.H.; formal analysis, R.H., A.T. and G.M.S.; investigation, R.H. and O.J.C.; data curation, R.H.; writing—original draft preparation, R.H.; writing—review and editing,

P.P.P. and S.R.C.; visualization, R.H.; supervision, P.P.P. and S.R.C.; project administration, S.R.C. and P.P.P.; funding acquisition, S.R.C. and P.P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the BBSRC Institute Strategic Programme Grant Gut Microbes and Health BB/R012490/1 and its constituent projects BBS/E/F/000PR10353, BBS/E/F/000PR10355, and BBS/E/F/000PR10356 (S.R.C., E.M.A, O.J.C., A.T.). E.M.A was additionally funded by the BBSRC Institute Strategic Programme Food Microbiome and Health BB/X011054/1 and its constituent projects BBS/E/F/000PR13631 and BBS/E/F/000PR13633; and by the BBSRC Institute Strategic Programme Microbes and Food Safety BB/X011011/1 and its constituent projects BBS/E/F/000PR13634, BBS/E/F/000PR13635, and BBS/E/F/000PR13636 (E.M.A). The QIB Sequencing facility and D.B. and G.M.S. were funded by BBSRC Core Capability Grant BB/CCG1860/1. R.H. was supported by PhD studentships jointly funded by Invest in ME Research (UK Charity number 1153730) and the Faculty of Medicine and Health, University of East Anglia.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the National Research Ethics Service (NRES) Committee London Hampstead (17/LO/1102, IRAS ID 218545, 29 June 2017).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The virome metagenomic sequencing data cleaned using the cleanup pipeline are freely available under NCBI BioProject number PRJNA1189570. The code for the analysis pipeline used in this paper is available as a Nextflow pipeline at https://github.com/RHaagmans/mc-spike (accessed on 27 November 2024).

Acknowledgments: The authors would like to thank Edward Mee (National Institute for Biological Standards and Control, Potters Bar, UK), Joe Brownlie (Royal Veterinary College, London, UK), and James Stewart (University of Liverpool, Liverpool, UK) for generously providing the stocks of RV-A, BVDV-1, and MHV-68, respectively.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Mirzaei, M.K.; Maurice, C.F. Ménage à Trois in the Human Gut: Interactions between Host, Bacteria and Phages. *Nat. Rev. Microbiol.* **2017**, *15*, 397–408. [CrossRef]
2. Li, Y.; Handley, S.A.; Baldridge, M.T. The Dark Side of the Gut: Virome–Host Interactions in Intestinal Homeostasis and Disease. *J. Exp. Med.* **2021**, *218*, e20201044. [CrossRef]
3. Liang, G.; Bushman, F.D. The Human Virome: Assembly, Composition and Host Interactions. *Nat. Rev. Microbiol.* **2021**, *19*, 514–527. [CrossRef]
4. Breitbart, M.; Hewson, I.; Felts, B.; Mahaffy, J.M.; Nulton, J.; Salamon, P.; Rohwer, F. Metagenomic Analyses of an Uncultured Viral Community from Human Feces. *J. Bacteriol.* **2003**, *185*, 6220–6223. [CrossRef]
5. Gregory, A.C.; Zablocki, O.; Zayed, A.A.; Howell, A.; Bolduc, B.; Sullivan, M.B. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **2020**, *28*, 724–740.e8. [CrossRef]
6. Callanan, J.; Stockdale, S.R.; Shkoporov, A.; Draper, L.A.; Ross, R.P.; Hill, C. Expansion of Known ssRNA Phage Genomes: From Tens to over a Thousand. *Sci. Adv.* **2020**, *6*, eaay5981. [CrossRef]
7. Neri, U.; Wolf, Y.I.; Roux, S.; Camargo, A.P.; Lee, B.; Kazlauskas, D.; Chen, I.M.; Ivanova, N.; Zeigler Allen, L.; Paez-Espino, D.; et al. Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages. *Cell* **2022**, *185*, 4023–4037.e18. [CrossRef]
8. Callanan, J.; Stockdale, S.R.; Shkoporov, A.; Draper, L.A.; Ross, R.P.; Hill, C. Biases in Viral Metagenomics-Based Detection, Cataloguing and Quantification of Bacteriophage Genomes in Human Faeces, a Review. *Microorganisms* **2021**, *9*, 524. [CrossRef]
9. Santiago-Rodriguez, T.M.; Hollister, E.B. Unraveling the Viral Dark Matter through Viral Metagenomics. *Front. Immunol.* **2022**, *13*, 1005107. [CrossRef]
10. Flores, G.E.; Caporaso, J.G.; Henley, J.B.; Rideout, J.R.; Domogala, D.; Chase, J.; Leff, J.W.; Vázquez-Baeza, Y.; Gonzalez, A.; Knight, R.; et al. Temporal Variability Is a Personalized Feature of the Human Microbiome. *Genome Biol.* **2014**, *15*, 531. [CrossRef]

11. Shkoporov, A.N.; Hill, C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe* **2019**, *25*, 195–209. [CrossRef]
12. Conceição-Neto, N.; Zeller, M.; Lefrère, H.; De Bruyn, P.; Beller, L.; Deboutte, W.; Yinda, C.K.; Lavigne, R.; Maes, P.; Ranst, M.V.; et al. Modular Approach to Customise Sample Preparation Procedures for Viral Metagenomics: A Reproducible Protocol for Virome Analysis. *Sci. Rep.* **2015**, *5*, 16532. [CrossRef]
13. Kleiner, M.; Hooper, L.V.; Duerkop, B.A. Evaluation of Methods to Purify Virus-like Particles for Metagenomic Sequencing of Intestinal Viromes. *BMC Genom.* **2015**, *16*, 7. [CrossRef] [PubMed]
14. Li, L.; Deng, X.; Mee, E.T.; Collot-Teixeira, S.; Anderson, R.; Schepelmann, S.; Minor, P.D.; Delwart, E. Comparing Viral Metagenomics Methods Using a Highly Multiplexed Human Viral Pathogens Reagent. *J. Virol. Methods* **2015**, *213*, 139–146. [CrossRef]
15. 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]
16. d'Humières, C.; Touchon, M.; Dion, S.; Cury, J.; Ghozlane, A.; Garcia-Garcera, M.; Bouchier, C.; Ma, L.; Denamur, E.; Rocha, E.P.C. A Simple, Reproducible and Cost-Effective Procedure to Analyse Gut Phageome: From Phage Isolation to Bioinformatic Approach. *Sci. Rep.* **2019**, *9*, 11331. [CrossRef]
17. Shkoporov, A.N.; Ryan, F.J.; Draper, L.A.; Forde, A.; Stockdale, S.R.; Daly, K.M.; McDonnell, S.A.; Nolan, J.A.; Sutton, T.D.S.; Dalmasso, M.; et al. Reproducible Protocols for Metagenomic Analysis of Human Faecal Phageomes. *Microbiome* **2018**, *6*, 68. [CrossRef]
18. Karlsson, O.E.; Belák, S.; Granberg, F. The Effect of Preprocessing by Sequence-Independent, Single-Primer Amplification (SISPA) on Metagenomic Detection of Viruses. *Biosecur. Bioterror. Biodefense Strategy Pract. Sci.* **2013**, *11*, S227–S234. [CrossRef]
19. Hsieh, S.-Y.; Tariq, M.A.; Telatin, A.; Ansorge, R.; Adriaenssens, E.M.; Savva, G.M.; Booth, C.; Wileman, T.; Hoyles, L.; Carding, S.R. Comparison of PCR versus PCR-Free DNA Library Preparation for Characterising the Human Faecal Virome. *Viruses* **2021**, *13*, 2093. [CrossRef]
20. Stockdale, S.R.; Shkoporov, A.N.; Khokhlova, E.V.; Daly, K.M.; McDonnell, S.A.; O' Regan, O.; Nolan, J.A.; Sutton, T.D.S.; Clooney, A.G.; Ryan, F.J.; et al. Interpersonal Variability of the Human Gut Virome Confounds Disease Signal Detection in IBD. *Commun. Biol.* **2023**, *6*, 221. [CrossRef]
21. Reyes, G.R.; Kim, J.P. Sequence-Independent, Single-Primer Amplification (SISPA) of Complex DNA Populations. *Mol. Cell Probes* **1991**, *5*, 473–481. [CrossRef]
22. 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]
23. 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] [PubMed]
24. Abulencia, C.B.; Wyborski, D.L.; Garcia, J.A.; Podar, M.; Chen, W.; Chang, S.H.; Chang, H.W.; Watson, D.; Brodie, E.L.; Hazen, T.C.; et al. Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments. *Appl. Environ. Microbiol.* **2006**, *72*, 3291–3301. [CrossRef]
25. Yilmaz, S.; Allgaier, M.; Hugenholtz, P. Multiple Displacement Amplification Compromises Quantitative Analysis of Metagenomes. *Nat. Methods* **2010**, *7*, 943–944. [CrossRef]
26. Marine, R.; McCarren, C.; Vorrasane, V.; Nasko, D.; Crowgey, E.; Polson, S.W.; Wommack, K.E. Caught in the Middle with Multiple Displacement Amplification: The Myth of Pooling for Avoiding Multiple Displacement Amplification Bias in a Metagenome. *Microbiome* **2014**, *2*, 3. [CrossRef]
27. Froussard, P. A Random-PCR Method (rPCR) to Construct Whole cDNA Library from Low Amounts of RNA. *Nucleic Acids Res.* **1992**, *20*, 2900. [CrossRef]
28. 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]
29. Zhai, X.; Gobbi, A.; Kot, W.; Krych, L.; Nielsen, D.S.; Deng, L. A Single-Stranded Based Library Preparation Method for Virome Characterization. *Microbiome* **2024**, *12*, 219. [CrossRef]
30. Wang, G.; Li, S.; Yan, Q.; Guo, R.; Zhang, Y.; Chen, F.; Tian, X.; Lv, Q.; Jin, H.; Ma, X.; et al. Optimization and Evaluation of Viral Metagenomic Amplification and Sequencing Procedures toward a Genome-Level Resolution of the Human Fecal DNA Virome. *J. Adv. Res.* **2023**, *48*, 75–86. [CrossRef]
31. Soria-Villalba, A.; Pesantes, N.; Jiménez-Hernández, N.; Pons, J.; Moya, A.; Pérez-Brocal, V. Comparison of Experimental Methodologies Based on Bulk-Metagenome and Virus-like Particle Enrichment: Pros and Cons for Representativeness and Reproducibility in the Study of the Fecal Human Virome. *Microorganisms* **2024**, *12*, 162. [CrossRef]

32. Seton, K.A.; Defernez, M.; Telatin, A.; Tiwari, S.K.; Savva, G.M.; Hayhoe, A.; Noble, A.; de Carvalho-KoK, A.L.S.; James, S.A.; Bansal, A.; et al. Investigating Antibody Reactivity to the Intestinal Microbiome in Severe Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS): A Feasibility Study. *Int. J. Mol. Sci.* **2023**, *24*, 15316. [CrossRef]
33. World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* **2013**, *310*, 2191–2194. [CrossRef]
34. Howard, C.J.; Brownlie, J.; Clarke, M.C. Comparison by the Neutralisation Assay of Pairs of Non-Cytopathic and Cytopathic Strains of Bovine Virus Diarrhoea Virus Isolated from Cases of Mucosal Disease. *Vet. Microbiol.* **1987**, *13*, 361–369. [CrossRef]
35. Holmfeldt, K.; Odić, D.; Sullivan, M.B.; Middelboe, M.; Riemann, L. Cultivated Single-Stranded DNA Phages That Infect Marine Bacteroidetes Prove Difficult to Detect with DNA-Binding Stains. *Appl. Environ. Microbiol.* **2012**, *78*, 892–894. [CrossRef]
36. Hoyles, L.; Mccartney, A.L.; Neve, H.; Gibson, G.R.; Sanderson, J.D.; Heller, K.J.; Sinderen, D.V. Characterization of Virus-like Particles Associated with the Human Faecal and Caecal Microbiota. *Res. Microbiol.* **2014**, *165*, 803–812. [CrossRef]
37. Kramná, L.; Cinek, O. Virome Sequencing of Stool Samples. In *The Human Virome*; Moya, A., Pérez Brocal, V., Eds.; Methods in Molecular Biology; Springer: New York, NY, USA, 2018; Volume 1838, pp. 59–83. ISBN 978-1-4939-8681-1.
38. Perez-Sepulveda, B.M.; Heavens, D.; Pulford, C.V.; Predeus, A.V.; Low, R.; Webster, H.; Dykes, G.F.; Schudoma, C.; Rowe, W.; Lipscombe, J.; et al. An Accessible, Efficient and Global Approach for the Large-Scale Sequencing of Bacterial Genomes. *Genome Biol.* **2021**, *22*, 349. [CrossRef]
39. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow Enables Reproducible Computational Workflows. *Nat. Biotechnol.* **2017**, *35*, 316–319. [CrossRef]
40. Handley, S.A. *Virus+ Sequence Masked Human Reference Genome (Hg19)*; Zenodo: Genève, Switzerland, 2020. [CrossRef]
41. Wood, D.E.; Lu, J.; Langmead, B. Improved Metagenomic Analysis with Kraken 2. *Genome Biol.* **2019**, *20*, 257. [CrossRef]
42. Chen, S. Ultrafast One-Pass FASTQ Data Preprocessing, Quality Control, and Deduplication Using Fastp. *iMeta* **2023**, *2*, e107. [CrossRef]
43. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
44. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef]
45. Braham Informatics. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 1 May 2023).
46. Shen, W.; Le, S.; Li, Y.; Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **2016**, *11*, e0163962. [CrossRef]
47. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve Years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008. [CrossRef]
48. Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph. *Bioinformatics* **2015**, *31*, 1674–1676. [CrossRef]
49. Li, D.; Luo, R.; Liu, C.-M.; Leung, C.-M.; Ting, H.-F.; Sadakane, K.; Yamashita, H.; Lam, T.-W. MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler Driven by Advanced Methodologies and Community Practices. *Methods* **2016**, *102*, 3–11. [CrossRef]
50. 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]
51. 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]
52. Nayfach, S.; Camargo, A.P.; Schulz, F.; Eloe-Fadrosh, E.; Roux, S.; Kyrpides, N.C. CheckV Assesses the Quality and Completeness of Metagenome-Assembled Viral Genomes. *Nat. Biotechnol.* **2021**, *39*, 578–585. [CrossRef]
53. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
54. Camargo, A.P.; Roux, S.; Schulz, F.; Babinski, M.; Xu, Y.; Hu, B.; Chain, P.S.G.; Nayfach, S.; Kyrpides, N.C. Identification of Mobile Genetic Elements with geNomad. *Nat. Biotechnol.* **2023**, *42*, 1303–1312. [CrossRef]
55. McMurdie, P.J.; Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **2013**, *8*, e61217. [CrossRef] [PubMed]
56. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the Tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [CrossRef]
57. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016; ISBN 978-3-319-24277-4.
58. Attali, D.; Baker, C. ggExtra 2023. Available online: https://github.com/daattali/ggExtra (accessed on 12 August 2024).
59. Kassambara, A. Ggpubr. Available online: https://rpkgs.datanovia.com/ggpubr/ (accessed on 1 September 2023).
60. Xiao, N. Ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for "Ggplot2". Available online: https://nanx.me/ggsci/ (accessed on 1 September 2023).

61. van den Brand, T. Ggh4x: Hacks for "Ggplot2". Available online: https://teunbrand.github.io/ggh4x/ (accessed on 1 October 2023).
62. Iannone, R.; Cheng, J.; Schloerke, B.; Hughes, E.; Lauer, A.; Seo, J.; Brevoort, K.; Roy, O. Gt: Easily Create Presentation-Ready Display Tables 2024. Available online: https://gt.rstudio.com (accessed on 19 August 2024).
63. Barr, J.J.; Auro, R.; Furlan, M.; Whiteson, K.L.; Erb, M.L.; Pogliano, J.; Stotland, A.; Wolkowicz, R.; Cutting, A.S.; Doran, K.S.; et al. Bacteriophage Adhering to Mucus Provide a Non–Host-Derived Immunity. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10771–10776. [CrossRef]
64. Fraser, J.S.; Yu, Z.; Maxwell, K.L.; Davidson, A.R. Ig-Like Domains on Bacteriophages: A Tale of Promiscuity and Deceit. *J. Mol. Biol.* **2006**, *359*, 496–507. [CrossRef]
65. Parent, K.N.; Khayat, R.; Tu, L.H.; Suhanovsky, M.M.; Cortines, J.R.; Teschke, C.M.; Johnson, J.E.; Baker, T.S. P22 Coat Protein Structures Reveal a Novel Mechanism for Capsid Maturation: Stability without Auxiliary Proteins or Chemical Crosslinks. *Structure* **2010**, *18*, 390–401. [CrossRef]
66. Read, S.J. Recovery Efficiencies of Nucleic Acid Extraction Kits as Measured by Quantitative LightCyclerTM PCR. *Mol. Pathol.* **2001**, *54*, 86–90. [CrossRef] [PubMed]
67. Sathiamoorthy, S.; Malott, R.J.; Gisonni-Lex, L.; Ng, S.H.S. Selection and Evaluation of an Efficient Method for the Recovery of Viral Nucleic Acids from Complex Biologicals. *NPJ Vaccines* **2018**, *3*, 31. [CrossRef]
68. Zhang, D.; Lou, X.; Yan, H.; Pan, J.; Mao, H.; Tang, H.; Shu, Y.; Zhao, Y.; Liu, L.; Li, J.; et al. Metagenomic Analysis of Viral Nucleic Acid Extraction Methods in Respiratory Clinical Samples. *BMC Genom.* **2018**, *19*, 773. [CrossRef]
69. Klenner, J.; Kohl, C.; Dabrowski, P.W.; Nitsche, A. Comparing Viral Metagenomic Extraction Methods. *Curr. Issues Mol. Biol.* **2017**, *24*, 59–70. [CrossRef]
70. Iker, B.C.; Bright, K.R.; Pepper, I.L.; Gerba, C.P.; Kitajima, M. Evaluation of Commercial Kits for the Extraction and Purification of Viral Nucleic Acids from Environmental and Fecal Samples. *J. Virol. Methods* **2013**, *191*, 24–30. [CrossRef] [PubMed]
71. Neill, J.D. Molecular Biology of Bovine Viral Diarrhea Virus. *Biologicals* **2013**, *41*, 2–7. [CrossRef]
72. Patton, J.T. Structure and Function of the Rotavirus RNA-Binding Proteins. *J. Gen. Virol.* **1995**, *76*, 2633–2644. [CrossRef]
73. Adriaenssens, E.M.; Farkas, K.; Harrison, C.; Jones, D.L.; Allison, H.E.; McCarthy, A.J. Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. *mSystems* **2018**, *3*, e00025-18. [CrossRef]
74. Sabina, J.; Leamon, J.H. Bias in Whole Genome Amplification: Causes and Considerations. In *Whole Genome Amplification: Methods and Protocols*; Kroneis, T., Ed.; Methods in Molecular Biology; Springer: New York, NY, USA, 2015; pp. 15–41. ISBN 978-1-4939-2990-0.
75. Smits, S.L.; Bodewes, R.; Ruiz-Gonzalez, A.; Koopmans, M.P.; Schürch, A.C. Assembly of Viral Genomes from Metagenomes. *Front. Microbiol.* **2014**, *5*, 714. [CrossRef]
76. Dabney, J.; Meyer, M. Length and GC-Biases during Sequencing Library Amplification: A Comparison of Various Polymerase-Buffer Systems with Ancient and Modern DNA Sequencing Libraries. *BioTechniques* **2012**, *52*, 87–94. [CrossRef] [PubMed]
77. Rosseel, T.; Borm, S.V.; Vandenbussche, F.; Hoffmann, B.; van de Berg, T.; Beer, M.; Höper, D. The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing. *PLoS ONE* **2013**, *8*, e76144. [CrossRef] [PubMed]
78. Pan, W.; Byrne-Steele, M.; Wang, C.; Lu, S.; Clemmons, S.; Zahorchak, R.J.; Han, J. DNA Polymerase Preference Determines PCR Priming Efficiency. *BMC Biotechnol.* **2014**, *14*, 10. [CrossRef] [PubMed]
79. Hansen, K.D.; Brenner, S.E.; Dudoit, S. Biases in Illumina Transcriptome Sequencing Caused by Random Hexamer Priming. *Nucleic Acids Res.* **2010**, *38*, e131. [CrossRef]
80. Gunasekera, S.; Abraham, S.; Stegger, M.; Pang, S.; Wang, P.; Sahibzada, S.; O'Dea, M. Evaluating Coverage Bias in Next-Generation Sequencing of Escherichia Coli. *PLoS ONE* **2021**, *16*, e0253440. [CrossRef]
81. Johansen, J.; Plichta, D.R.; Nissen, J.N.; Jespersen, M.L.; Shah, S.A.; Deng, L.; Stokholm, J.; Bisgaard, H.; Nielsen, D.S.; Sørensen, S.J.; et al. Genome Binning of Viral Entities from Bulk Metagenomics Data. *Nat. Commun.* **2022**, *13*, 965. [CrossRef] [PubMed]
82. Kieft, K.; Adams, A.; Salamzade, R.; Kalan, L.; Anantharaman, K. vRhyme Enables Binning of Viral Genomes from Metagenomes. *Nucleic Acids Res.* **2022**, *50*, e83. [CrossRef] [PubMed]
83. Kang, D.D.; Froula, J.; Egan, R.; Wang, Z. MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities. *PeerJ* **2015**, *3*, e1165. [CrossRef]
84. Kang, D.D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies. *PeerJ* **2019**, *7*, e7359. [CrossRef]
85. Arisdakessian, C.G.; Nigro, O.D.; Steward, G.F.; Poisson, G.; Belcaid, M. CoCoNet: An Efficient Deep Learning Tool for Viral Metagenome Binning. *Bioinformatics* **2021**, *37*, 2803–2810. [CrossRef] [PubMed]
86. Mallawaarachchi, V.; Roach, M.J.; Decewicz, P.; Papudeshi, B.; Giles, S.K.; Grigson, S.R.; Bouras, G.; Hesse, R.D.; Inglis, L.K.; Hutton, A.L.K.; et al. Phables: From Fragmented Assemblies to High-Quality Bacteriophage Genomes. *Bioinformatics* **2023**, *39*, btad586. [CrossRef]

87. Chen, L.; Banfield, J.F. COBRA Improves the Completeness and Contiguity of Viral Genomes Assembled from Metagenomes. *Nat. Microbiol.* **2024**, *9*, 737–750. [CrossRef]
88. Nissen, J.N.; Johansen, J.; Allesøe, R.L.; Sønderby, C.K.; Armenteros, J.J.A.; Grønbech, C.H.; Jensen, L.J.; Nielsen, H.B.; Petersen, T.N.; Winther, O.; et al. Improved Metagenome Binning and Assembly Using Deep Variational Autoencoders. *Nat. Biotechnol.* **2021**, *39*, 555–560. [CrossRef]
89. Cook, R.; Telatin, A.; Hsieh, S.-Y.; Newberry, F.; Tariq, M.A.; Baker, D.J.; Carding, S.R.; Adriaenssens, E.M. Nanopore and Illumina Sequencing Reveal Different Viral Populations from Human Gut Samples. *Microb. Genom.* **2024**, *10*, 001236. [CrossRef]
90. Knight, R.; Vrbanac, A.; Taylor, B.C.; Aksenov, A.; Callewaert, C.; Debelius, J.; Gonzalez, A.; Kosciolek, T.; McCall, L.-I.; McDonald, D.; et al. Best Practices for Analysing Microbiomes. *Nat. Rev. Microbiol.* **2018**, *16*, 410–422. [CrossRef] [PubMed]
91. Boers, S.A.; Jansen, R.; Hays, J.P. Understanding and Overcoming the Pitfalls and Biases of Next-Generation Sequencing (NGS) Methods for Use in the Routine Clinical Microbiological Diagnostic Laboratory. *Eur. J. Clin. Microbiol. Infect. Dis.* **2019**, *38*, 1059–1070. [CrossRef] [PubMed]