viruses

MDPI

Article

Critical Assessment of Whole Genome and Viral Enrichment Shotgun Metagenome on the Characterization of Stool Total Virome in Hepatocellular Carcinoma Patients

Fan Zhang [1], Andrew Gia [1], Guowei Chen [2], Lan Gong [1], Jason Behary [1,3], Georgina L. Hold [1], Amany Zekry [1,3], Xubo Tang [2], Yanni Sun [2], Emad El-Omar [1,3] and Xiao-Tao Jiang [1,*]

[1] Microbiome Research Centre, St George and Sutherland Clinical School, University of New South Wales, Sydney, NSW 2217, Australia
[2] Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China
[3] Department of Gastroenterology, St George Hospital, Sydney, NSW 2217, Australia
* Correspondence: xiaotao.jiang@unsw.edu.au

Abstract: Viruses are the most abundant form of life on earth and play important roles in a broad range of ecosystems. Currently, two methods, whole genome shotgun metagenome (WGSM) and viral-like particle enriched metagenome (VLPM) sequencing, are widely applied to compare viruses in various environments. However, there is no critical assessment of their performance in recovering viruses and biological interpretation in comparative viral metagenomic studies. To fill this gap, we applied the two methods to investigate the stool virome in hepatocellular carcinoma (HCC) patients and healthy controls. Both WGSM and VLPM methods can capture the major diversity patterns of alpha and beta diversities and identify the altered viral profiles in the HCC stool samples compared with healthy controls. Viral signatures identified by both methods showed reductions of Faecalibacterium virus Taranis in HCC patients' stool. Ultra-deep sequencing recovered more viruses in both methods, however, generally, 3 or 5 Gb were sufficient to capture the non-fragmented long viral contigs. More lytic viruses were detected than lysogenetic viruses in both methods, and the VLPM can detect the RNA viruses. Using both methods would identify shared and specific viral signatures and would capture different parts of the total virome.

Keywords: metagenome; total virome; hepatocellular carcinoma; deep sequencing

Citation: Zhang, F.; Gia, A.; Chen, G.; Gong, L.; Behary, J.; Hold, G.L.; Zekry, A.; Tang, X.; Sun, Y.; El-Omar, E.; et al. Critical Assessment of Whole Genome and Viral Enrichment Shotgun Metagenome on the Characterization of Stool Total Virome in Hepatocellular Carcinoma Patients. *Viruses* **2023**, *15*, 53. https://doi.org/10.3390/v15010053

Academic Editor: Philippe Gallay

Received: 30 November 2022
Revised: 19 December 2022
Accepted: 22 December 2022
Published: 24 December 2022

1. Introduction

Viruses are ubiquitous and are the most abundant entities on earth with an estimated number of 1031 particles infecting their hosts [1]. Viruses are associated with many facets including biogeochemical cycles [2], infectious diseases including COVID 19 [3], direct immune activation [4] and non-communicable diseases such as inflammatory bowel diseases [5], colorectal cancer [6,7] and non-alcoholic fatty liver disease [8]. High throughput culture-independent next-generation sequencing technologies have enabled major progress in understanding the role of viruses in these diseases.

To investigate viruses in an omics way, three strategies have been applied: (1) Viral like particles enriched metagenome which investigates a collection of viruses in a viral purified biological sample including both known and novel viruses [9,10]; (2) Viruses are captured with designed probes on a chip and then undergoing deep sequencing [11,12]; (3) Mining the viruses directly from the whole genome shotgun metagenome of the total nucleotide content in a sample, which had been applied in numerous studies to investigate their roles in the healthy and diseased human gut [7,13,14]. Moreover, thousands of novel viral genomes have been assembled with either metagenome or virome and catalogued into viral databases, greatly expanding the size of the resources [14–16]. The viral-capture

method can only capture viruses that have a certain similarity to the probes in the database and hence will not be covered in this study.

In a biological specimen, the genetic content includes bacteria, archaea, eukaryotic and prokaryotic dsDNA ssDNA and RNA viruses, eukaryotic fungal and protozoal content, host genetic content and debris from dead cells (Figure S1). The experimental processes of WGSM and VLPM methods pose different impacts on the genetic contents of the samples, and hence finally influence the captured viral profiles. The WGSM extracted the total DNAs from a sample, hence will miss those RNAs. However, as there is no nucleotide digestion with DNA/RNA enzymes, thus, the WGSM data will include free extracellular nucleotides. The VLPM approach includes a low-speed centrifugation step, DNase/RNase digestion process and cDNA synthesis for RNA viruses, Hence, it captures both DNA and RNA viruses. For VLMP, the pore size of filtration, DNase/RNase digestion, precipitation and gradient centrifugation can all influence the outcome [17]. Additionally, different nucleic acid extraction methods, the use of nuclease treatment to digest host DNA [18], random amplification such as multiple displacement amplification, random hexamers and single-primers amplification (SISPA) methods all have an impact [18,19]. The size fractionation to enrich viral-like particles prior to DNA extraction was found to better capture soil viruses [20], but this is not clear for other environments. The phages can be divided into lytic and lysogenic life forms. Lysogenic phages are those viruses integrated into their bacterial host genome under normal conditions and lytic phages are external phages that infect the host and lytic the host cell directly [21]. Hence, the experimental difference between the two methods may influence the recovery of both types of viruses. Shotgun metagenome generally sequenced about 3 or 5 Gb [22–24] data to investigate the microbiome including virome, however, there is no assessment on whether this depth of sequencing is sufficient to capture all the viruses in a biological sample. During the bioinformatics analysis, the choice of the assembler will significantly influence the final viral annotations [25].

Despite the popularity of mining viral signals directly from shotgun metagenome samples to study their roles in health and disease [7,13,26] and VLPM [5,8,14,20], there has been no critical assessment of how these two methods compare with regard to viral diversity and viral biomarker identification. In this study, the VLPM and WGSM were systematically compared on total virome in patients with HCC versus healthy controls. Furthermore, two samples were selected to perform ultra-deep sequencing in both methods to a depth of 305 million reads (averagely 36 Gb from 29 to 45 Gb) to assess their performance on viral alpha and beta diversity interpretation. The capture of lytic and lysogenic bacteriophage between the two methods was also compared.

2. Materials and Methods

2.1. Sample Collections

Six patients who were diagnosed with NAFLD-HCC and 6 healthy controls were recruited at St George Hospital, Sydney. The study was approved by Sydney Local Health District Human Research Ethics Committee, New South Wales Health. Informed consent was obtained from all study participants. Patients with HCC were recruited to the study at the Liver Clinic, St George Hospital, Sydney. Age, gender, BMI and etiology of two groups were included in (Table S1). HCC was diagnosed according to international guidelines, integrating history, physical examination, biochemistry, and imaging techniques obtained by multiphasic CT, and/or dynamic contrast-enhanced MRI. Diagnosis of HCC was further confirmed by histopathological examination of surgical resection specimens. Total nucleotides were extracted to perform WGSM and enriched VLPM sequencing, respectively (Figure S1). Samples of two subjects (i.e., one HCC patient SLN_09 and one healthy control SCO_10) were selected for ultradeep sequencing to investigate the influence of sequencing depth in detecting the viral signals. Together, 12 WGSM and 12 VLPM samples were generated (Table S1). Fecal samples were collected using Stratec PSP stool sampling tube by ColOff® stool collection sleeve and stored at −80 °C within 48 h after sample collection.

2.2. Shotgun Metagenome DNA Extraction

According to manufacturer instructions, total DNA was extracted with PSP Spin Stool DNA Plus Kit (Stratech, Invitek Molecular, Robert-Roessle-Str. 10 D-13125 Berlin).

2.3. Viral-like Particles Enrichment

To enrich both DNA and RNA viruses, a low-speed centrifuge and filtration and DNase/RNase digestion process were applied based on an inflammatory bowel disease virome study [5]. Firstly, 0.2 g of sample was mixed with 400 µL SM Buffer and applied low-speed centrifuge at speed of 2000× *g* for 10 min, the supernatant was kept and filtered through a 1 mL Luer-lok syringe and 13 mm diameter 0.45 µm filters. The filtered liquid was then incubated with 70 µL Lysozyme/DNase mix (27 µL Turbo DNase buffer, TurboDNase 5 µL of 2 U/µL, Baseline zero 1 µL of 1 U/µL, Lysozyme 20 µL of 100 mg/mL and 17 µL H_2O) at 37 °C for 1 h to digest free nucleotides. Then, PureLink virus RNA/DNA mini extraction kit (Thermo Fisher, 1/4 Talavera Rd, North Ryde NSW 2113) was used to extract both DNA and RNA from the sample. To synthesis cDNA, 1 µL ofextraction including both RNA and DNA was then reverse transcribed using Superscript IV Reverse Transcriptase (Invitrogen, 1/4 Talavera Rd, North Ryde NSW 2113).

2.4. DNA Sequencing

The extracted total DNA and VLP cDNA were quantified by Qubit 3.0 Fluorometer (Invitrogen) with sequencing performed by Ramacciotti Sequencing Centre. The Nextera library preparation KIT was used to prepare the sequencing library and pooled for Novaseq S6000 to generate pair-end 150 bps short reads. The viral extraction was not amplified to avoid duplicate PCR as the Nextera library prep has 5 cycles of PCR reaction. Two samples were deeply sequenced to evaluate the influence of sequencing depth.

2.5. Sequencing Data Processing

Raw sequencing data were deduplicated using BBmap Clumpify (last modified 30 October 2019) [27], decontaminated host reads employing Minimap2 (version 2.18-r1015) [28] with hg38 reference genome, removed low quality reads with fastp (version 0.20.1) [29], excluded all ribosomal RNA utilizing SortMeRNA (version 4.3.2) [30] only for the VLPM to pre-process the sequencing data. Clean reads were next mix-assembled with MEGAHIT (version 1.2.9) [31] and the generated contigs clustered by CD-HIT (version 4.8.1; with a sequence identity of 0.95 and length coverage of 0.9) [32]. The representative contigs for each cluster were then taken as the reference to which the clean reads were mapped using Bowtie 2 (version 2.4.2) [33]. This generated a feature table where rows and columns indicated representative contigs or OVU (Operational Viruses Unit) and samples, respectively. A contig was considered to be present in a sample if 75% or more of the contig has a coverage $\geq$1 to more accurately quantify the presence of viruses in a specimen [34]. To identify all the potential viral contigs, three methods were applied: (1) used VIBRANT (version 1.2.1) [35] to get the bacteria and archaea DNA and RNA viruses. VIBRANT applied a hybrid machine learning and protein similarity searching approach to annotate viruses. It uses a neural network and developed a v-score metric to differentiate the lytic and lysogenic viruses. VIBRANT outperformed other similar tools such as VirSorter1, VirFinder and MARVEL by lower false-positive and higher recovery rates. (2) applied VirSorter2 [36] to identify potential RNA viruses; (3) used VirBot [37] to detect and annotate RNA viruses from all the contig length over 500 bps. For the viral contig predicted by VIBRANT and VirSorter2, the Kraken 2 (version 2.1.1) run [38] with Metagenomics Virus Database [39] which was built from 700 K metagenomics viruses from JGI IMG/VR [40] was used to give taxonomical annotation. A combined taxonomical annotation was then generated. The abundance of the predicted viral contigs was normalized into FPKM (Fragments Per Kilobase of transcript per Million mapped reads) for differential abundance analysis. To scale the samples to create the heatmap while preserving each sample's distribution, z-score normalization was applied to the FPKMs of each sample. The relative abundance of

a genus (or a family) was calculated by dividing the sum of FPKMs assigned to this genus (or this family) by the total FPKMs of this sample. The viral contigs identified by VIBRANT were classified into lytic or lysogenic viruses. Lysogenic viruses which integrated into the host bacterial genome were characterized by viral contigs with bacterial fragments on both sides. The remaining contigs were annotated as lytic viruses.

2.6. Statistical Analysis

R (version 4.0.4) packages vegan [41] and phyloseq [42] were employed for rarefaction species richness analysis and diversities analysis, respectively. Mann–Whitney–Wilcoxon test was applied in the comparison of the means of the Alpha diversities between different groups. To access the significance of disease and sequencing methods effects between two distance matrices in the Beta diversity analysis, adonis (Permutational Multivariate Analysis of Variance Using Distance Matrices) was used to permute the distance matrix 999 times to yield *p*-values and ESS. In the identification of disease-associated viral signatures, MaAsLin2 [43] (Microbiome Multivariable Association with Linear Models) was applied to determine associations between case–control metadata and viral signals. The STORMS checklist has been completed (Table S2) [44].

3. Results

3.1. Deep Sequencing Contributes to the Identification of Long Viral Contigs

Figure 1 draws the experimental design to compare the performance of the two methods on virome investigations. First, the number of predicted viral contigs was associated with a list of different thresholds of the contig length and coverage rate to determine an optimal setting for the two parameters. In general, given a certain coverage rate (e.g., a contig was present in a sample if 75% or more of the contig has a coverage $\geq$1), the number of viral contigs dropped sharply when the contig length increased from 1 kb to 7 kb, then this number decreased slowly with the contig length further increasing from 8 kb to 15 kb (Figure 2a). To balance the contig length and the number of retained contigs to be analyzed, 3 kb, 5 kb, 8 kb and 10 kb were selected as cut-offs. The rarefaction curves showed that for both sequencing methods, deeper sequencing captured more viral signals at contig length $\geq$10 kb (Figure 2b). The observed number of viral contigs was increased from 3 Gb sequencing depth to 5 Gb depths, which were the commonly used sequencing depths in previous studies [45,46]. After 5 Gb depth, the curve quickly reached a plateau, and as a result, the sequencing depth showed a limited effect on the observed viral contigs, demonstrating that with the coverage rate = 75%, 3 Gb sequencing depth is sufficient to cover all these contigs. The total number of captured viral contigs in the two ultradeep sequenced subjects surpassed all the other subjects suggesting that deep sequencing played an important role in obtaining long viral contigs by assembling the low abundant viruses. Similar conclusions can be drawn from the cut-offs of 3 kb, 5 kb and 8 kb (Figure S2a–c). Next, the coverage rate was varied from 10%, 25%, 50%, 75% to 90% given the contig length $\geq$10 kb (Figures 2b and S3a–d). The enrichment procedure facilitated the assembly of viral contigs with high coverage. Moreover, the higher the coverage rate, the quicker the rarefaction curve reaches the plateau.

To remove the major non-digested viruses generated only by the WGSM protocol which is mainly composed of free dead viral genetic fragments, the 48 WGSM specific viral contigs (Figure 2c), which were detected only from WGSM samples, were discarded, and the viral contig list used in the downstream analysis consisted of 1320 + 107 = 1427 contigs.

Alpha diversity indices in terms of the two methods were next compared under the abovementioned contig length and coverage rate cut-offs. When assessing all samples collectively, no significant difference was observed in terms of capturing viral contigs and Shannon diversity between the two methods, regardless of the parameter thresholds (Figures 2d, S2d–g and S3e–h). A similar conclusion can be drawn when stratifying the cohort by sample type, i.e., HCC patients have no significant difference from healthy controls (Figures 2e, S2d–g and S3e–h). Therefore, different parameter cut-offs do not

impact the biological interpretation, and as such, 10 kb and coverage rate = 75% were selected as the thresholds consistent with previous literature [30,31,34].

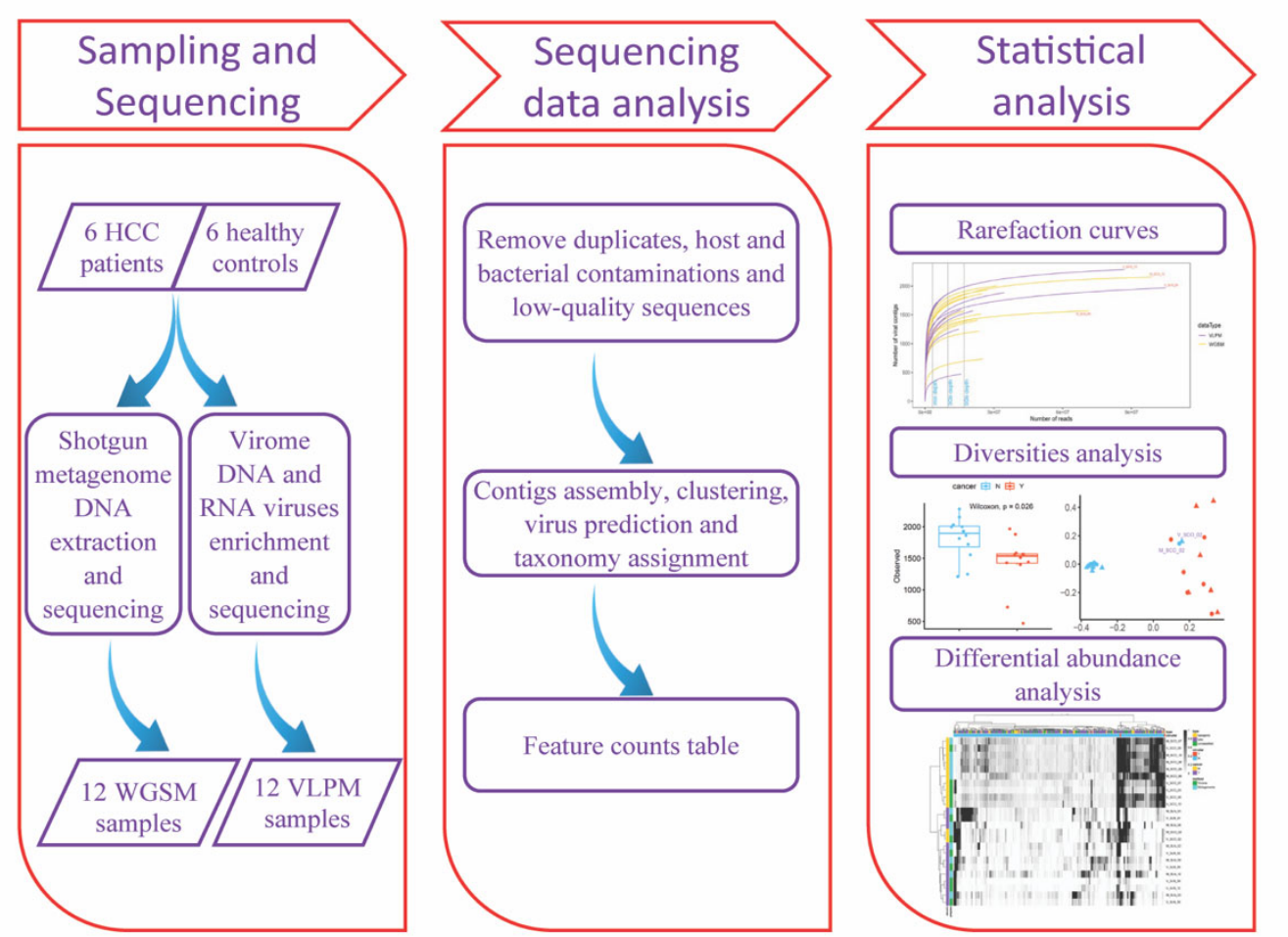

Figure 1. The experimental and analytical flowchart of whole genome shotgun metagenome (WGSM) and enriched viral-like particle metagenome (VLPM) sequencing.

The ratio of reads mapped to virus contigs to those mapped to all the assembled contigs was then compared between the two sequencing methods (Figure S3i), demonstrating that viral-like particles were successfully enriched in the VLPM samples.

3.2. VLPM and WGSM Perform Similarly in Beta Diversity Interpretation

The principal coordinates analysis (PCoA) revealed an altered viral profile between HCC patients and the healthy controls, with similar distributions seen for both WGSM and VLPM sample sets (Figures 2f,g and S4a,b). When investigating the healthy controls alone, subject SCO_02 was a clear outlier (Figures 2h and S4c) whose variance contributed most to the first principal component (Axis_1). The second principal component (Axis_2 in Figure 2h), on the other hand, mainly captured the variances between the two sequencing methods. While the healthy WGSM samples were clustered near the upper left corner, the healthy VLPM samples scattered along the y-axis, indicating that the VLPM method could detect more subject-specific viral signals (Figure 2h). On the contrary, the HCC patients mostly revealed individual patterns which could not be separated by the two methods (Figures 2i and S4d). When combining all the samples together, the disease-associated differences were also clearly captured (Figures 2j and S4e). The healthy controls were more tightly clustered demonstrating that healthy people had similar viral profiles while the HCC patients had unique individual profiles (Figure 2j). Not surprisingly, the differences between the two methods were not significant (Figure 2j). The phenotype difference was

larger than those of the methods, demonstrating that phenotype is a bigger impacting factor than the methods.

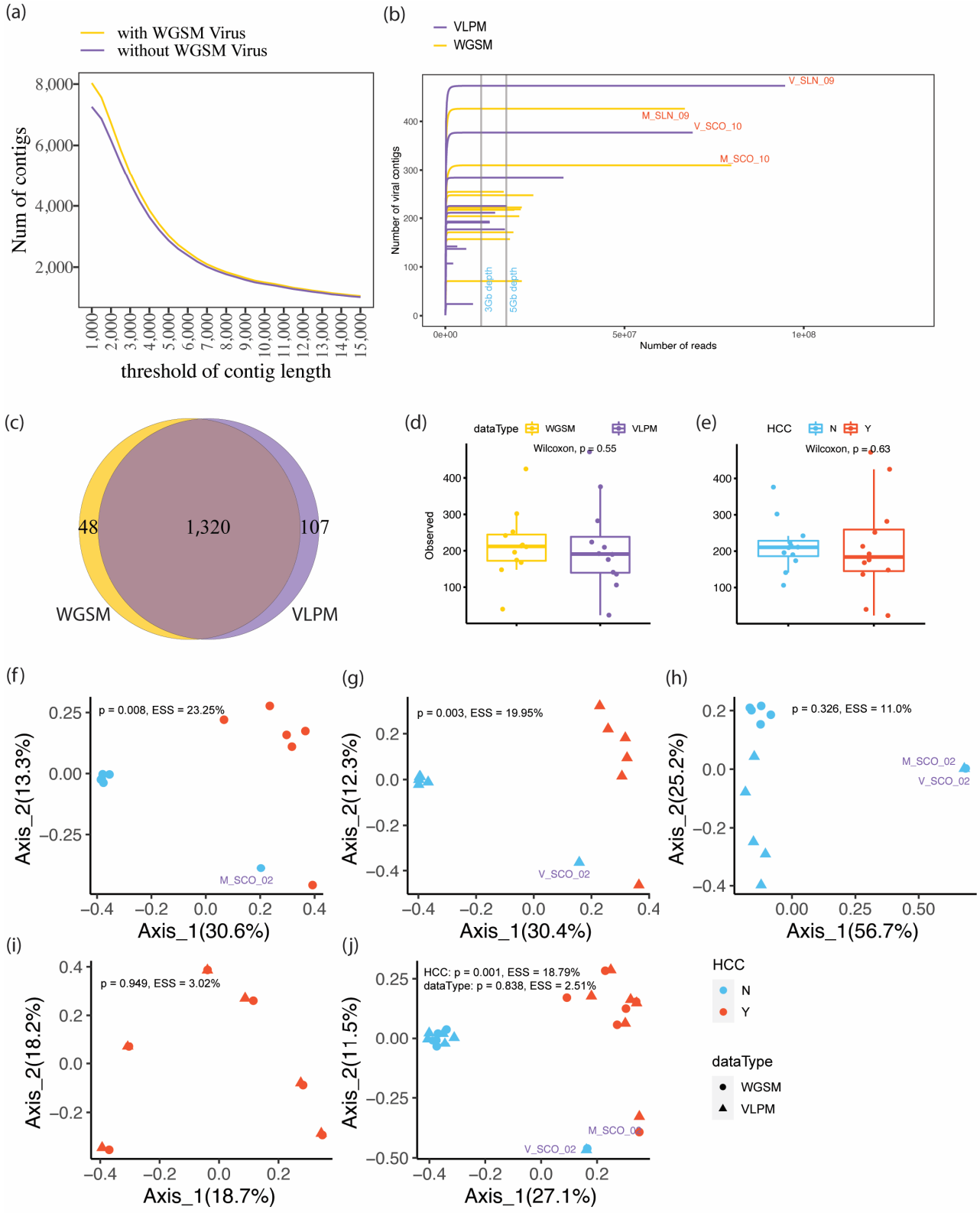

Figure 2. (**a**) The association between the length cut-offs of the predicted viral contigs and the number of retained contigs. The purple curve excluded the WGSM specific viral contigs. (**b**) The rarefaction

curves of the WGSM and the VLPM samples given contig length ≥ 10 kb and coverage rate ≥ 75%. The two samples, SLN_09 and SCO_10, were selected to perform ultra-deep sequencing. SLN and SCO indicate HCC patients and healthy controls, respectively. M and V represent WGSM and VLPM samples, respectively. Two vertical auxiliary lines indicate 3 Gb and the 5 Gb sequencing depth, respectively. (**c**) The Venn diagram of viral contigs detected from the WGSM and the VLPM samples (contig length ≥ 10 kb and coverage rate ≥ 75%). (**d**) The alpha diversities (observed features) between the two sequencing methods. Samples from the patients and the healthy controls were considered together. (**e**) The alpha diversities (observed features) between the HCC and the healthy samples. Samples from the two sequencing methods were considered together. The Mann–Whitney–Wilcoxon test was applied in the comparison of the means. The principal coordinates analysis plots based on (**f**) the WGSM samples, (**g**) the VLPM samples, (**h**) the healthy controls, (**i**) the HCC patients and (**j**) the combination of all the samples, respectively. The plots used Bray–Curtis dissimilarities. The outputs of adonis analysis, i.e., *p*-values and the explained sum of squares (ESS), were attached to the plots.

3.3. Comparison of Viral Compositions and Disease-Associated Signatures

Next, a compositional analysis was conducted to compare HCC cases and healthy controls. Given a taxonomical level (contigs with unclassified taxa were excluded), e.g., in the top 10 genus level (Figure 3a, family level in Figure S4f), crAss-like viruses had a higher proportion in the HCC samples, while *Taranisvirus* contigs were more abundant in healthy controls. When comparing the two methods, a higher proportion of *Brigitvirus* and *Oengusvirus* were seen in the WGSM samples (Figure 3a and Table S3). This demonstrated the preferential enrichment of different viruses in the two methods which originated from the experimental stage. To determine the viral features that are most likely to explain the differences between the HCC patients and the healthy controls by both methods while control age, gender and BMI, MaAsLin2 assessment was applied (Table 1). A virus infecting commensal bacteria *Faecalibacterium* virus Taranis (Figure 3b) was enriched in the healthy controls by both methods. Given the 10 kb length threshold, no RNA viruses were detected. By loosening the length cut-off to 500 bp, 14 RNA viruses were detected with most of them were plant viruses including cucumber green mottle mosaic virus, Tobacco mild green mosaic virus, and Pepper mild mottle virus, and some *Picobirnaviridae* sp. (Figure 3c and Table S4). The difference between the two methods in capturing RNA viruses was evaluated by comparing the abundance of predicted RNA virus contigs (Figure 3d). The VLPM samples have an extra portion of RNA viruses detected.

Table 1. Enriched viral signatures identified by MaAsLin2 in healthy subjects and HCC patients for both WGSM and VLPM methods (with age, gender and BMI controlled).

Group	Viral Signature	*p*-Value	Order	Family	Genus
WGSM + Health	*Faecalibacterium virus Taranis*	0.0001	Caudovirales	Myoviridae	*Taranisvirus*
VLPM + Cancer	*Faecalibacterium virus Epona*	0.015	Caudovirales	Myoviridae	*Eponavirus*
VLPM + Health	*Faecalibacterium virus Lugh*	0.049	Caudovirales	Siphoviridae	*Lughvirus*
	Faecalibacterium virus Toutatis	0.038	Caudovirales	Myoviridae	*Toutatisvirus*
	Faecalibacterium virus Taranis	0.019	Caudovirales	Myoviridae	*Taranisvirus*

To further investigate the differences in the captured viral signals between the two sequencing methods, the top 500 contigs (because the maximum number of non-zero abundance viral contigs in each sample is 457) with the highest median absolute deviations after z-score normalization were selected and clustered in a heatmap (Figure 4a). The healthy controls (except SCO_02) were clearly separated from HCC patients which is consistent with the Beta diversity results (Figures 2j and S4e). The healthy controls were further grouped based on the bioinformatics approach, demonstrating that the healthy subjects were similar to each other in terms of the viral signal patterns and the variances of the viral features were impacted mainly by the methods. In contrast, HCC samples

were clustered by the patient (Figures 2i, S4d and 4a). For example, the WGMS and the VLPM samples from HCC patient SLN_03 were grouped together away from other HCC samples. Further comparisons between the two methods on their ability to capture lytic and lysogenic viruses were undertaken (Figure 4b). It was found that both methods obtained more lytic than lysogenic viruses. This indicated that there is a great portion of lysogenic viruses which are only included within bacterial genomes that are retained in the bacterial cell filtration based VLPM method. The two ultradeep sequenced subjects clearly detected more lytic and lysogenic viruses than others reconfirming that deep sequencing significantly improved the performance in detecting long viral features.

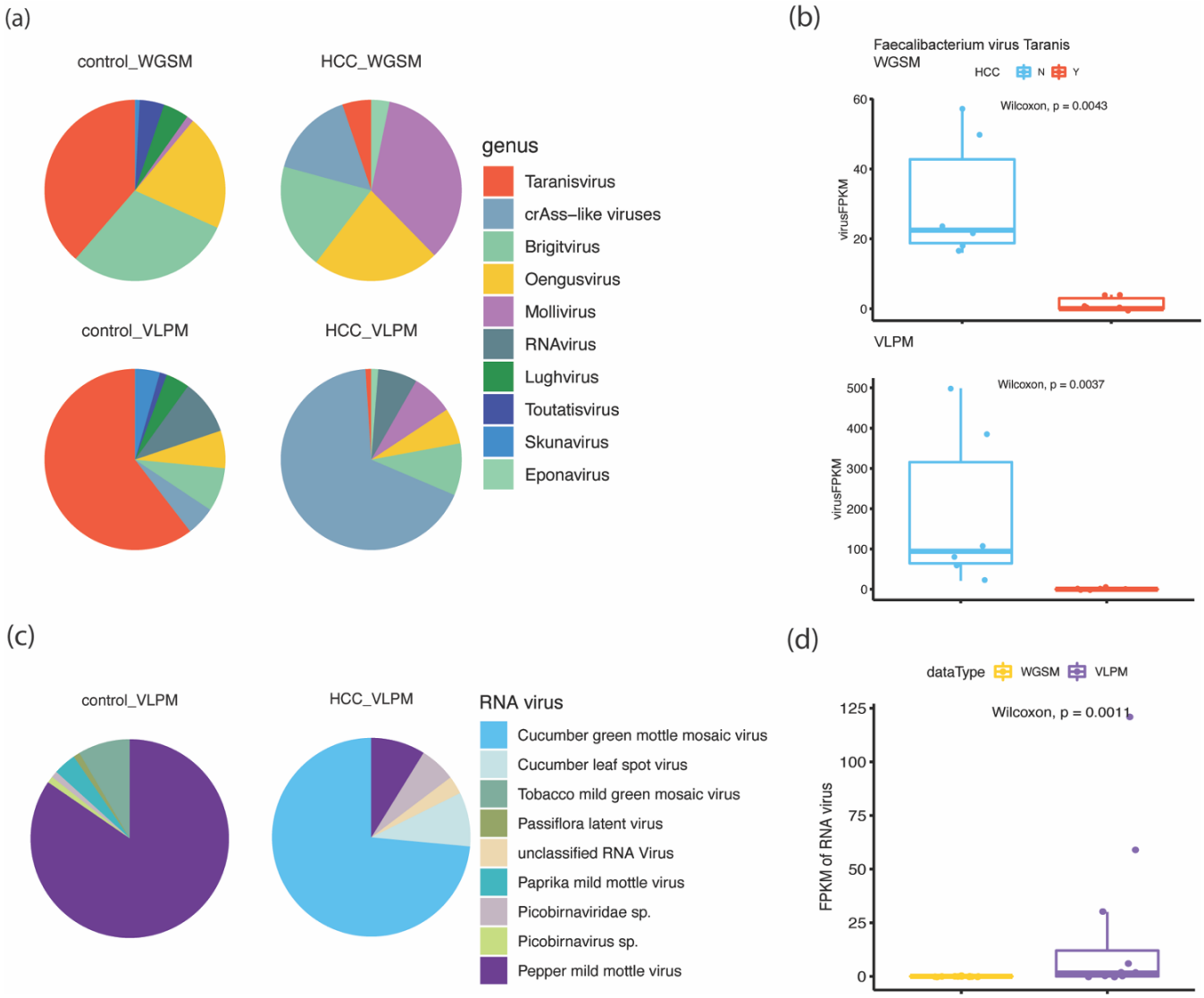

Figure 3. (**a**) Viral compositions at the genus level (top 10 genera). Contigs with unclassified taxa were excluded. Length cut-off for RNA virus was 500 bp. (**b**) Comparison of the abundance (FPKM) of Faecalibacterium virus Taranis between HCC patients and healthy controls within each sequencing method. (**c**) RNA virus compositions. (**d**) Comparison of the abundance (FPKM) of the RNA virus between the WGSM and the VLPM samples.

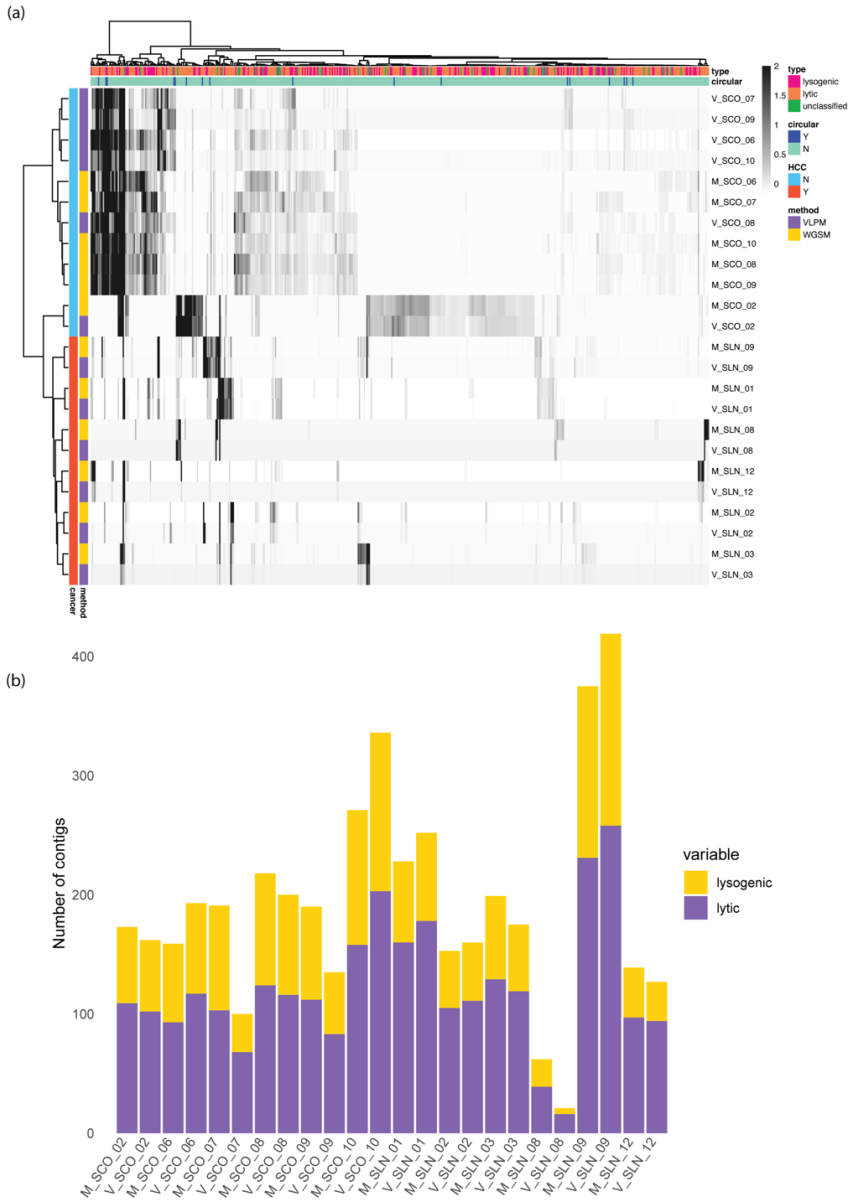

Figure 4. (**a**) Heatmap of the normalized abundances of the contigs. The top 500 contigs with the highest median absolute deviations were included. Clustering methods ward.D2 was applied. (**b**) The lysogenic-lytic distribution of all the samples.

4. Discussion

In this study, the performance of two methods of VLPM and WGSM was compared on virome in patients with hepatocellular carcinoma (HCC) versus healthy controls. Whether the traditional 3 Gb or 5 Gb sequencing depth is sufficient to mine viruses for comparative analysis was tested. Moreover, the important, yet often ignored, factor of choice of contig length, coverage rate, lytic and lysogenic viruses recovering were investigated.

The impacts of contig length and coverage rate for viral analysis have generally not been evaluated in real samples, and simulation studies have suggested keeping relatively longer contigs at around 5 kb to even 10 kb and coverage rate $\geq$ 75% [34] to retain high fidelity [47]. However, this conservative decision might cause low sensitivity to virus detection. This study showed that the numbers of viral-like contigs were dramatically reduced when the cut-offs of contig length increased from 1 to 7 kb (extending to 15 kb). This suggests that fragmental viral contigs from low abundant viruses are potentially missed with stringent cut-offs. Studies have investigated the impact of sequencing depth

on bacteria characterization [48]. The two deeply sequenced samples results showed that as depth increased, viral alpha diversity increased in both WGSM and VLPM methods. The abundance pattern captured at 3 Gb and 5 Gb in all samples was not different as there were no cross curves in the rarefaction analysis. Hence, 3 Gb in both WGSM and VLPM was good enough to capture the viral alpha diversity pattern.

The two commonly applied methods were assessed on the biological interpretations of alpha and beta diversities and differential abundant analysis. The alpha diversity results for both comparison methods and disease phenotype were influenced by the selection of contig length and coverage rate cut-offs. However, the results for both methods were comparable at a certain cut-off. Beta diversity analysis of viral profile showed consistent alternation in cancer patients for both methods. Diversity analyses give insight into whether the viral profiles are associated with disease phenotypes, which is an important ecological standard to interpret the data. The results demonstrated that both sequencing methods are satisfactory if the aim of the research is to investigate whether the viral community is associated, or not, with certain disease phenotypes. Moreover, WGSM inherently do not capture RNA viruses, whereas the VLPM includes RNA viruses. As a result, the viral profiles from the two methods were different, indicating that each method had its own advantages in capturing specific viral signals. These specificities are very likely due to the different experimental processing where the VLPM performed size fraction, DNase/RNase digestion of free nucleotides and the WGSM sequenced the total DNA including eDNA.

The HCC-associated viral signatures analysis has further confirmed this discovery by identifying that only a subset of the enriched viral signatures from the two methods was shared. This demonstrated that each method could capture part of the virome but not the full picture. Therefore, combining different experimental methods is recommended when the purpose is to collect as much information as possible. Moreover, the outcomes in our study have demonstrated the limitations of enrichment-based methods, which have also been studied previously, and that low-speed centrifuge and filtering could lose some viruses. Furthermore, the prophages, which are located inside the bacterial genome, are less captured. Our study did not show large differences between the two methods as Christian et al.'s work on soil virome [20], which detected 2961 viral clusters in total with only 94 shared and three specific viral clusters in the total metagenome. Of the total 1427 viral contigs over 10 Kb, 1320 (92.5%) were shared. This indicated that different types of samples (i.e., soil or stool) also have an influence on the viral profile captured by the two methods. Hence, it is demonstrated that although WGSM can capture diversity variances, the profiles captured are only a fraction of the total virome at the current sequencing depth. On the other hand, VLPM has bias due to the enrichment methods applied but can capture more viruses than WGSM at a similar sequencing depth. The influence of methods on the capture of lysogenic and lytic viruses was further distinguished. Consistent with previous literature [49,50], most of the viruses in the stool samples were lytic viruses. Other viruses such as RNA viruses or other eukaryotic viruses were less detected in the stool, with only 14 detected RNA viruses in the stool sample. The limited RNA viruses might be because the samples are fecal. Altogether, the results indicated that the influence of methods is dependent on parameter settings (e.g., contig length and coverage rate selections), phenotypic information of samples and sample types.

We acknowledge that it is unrealistic to consider all factors relevant to the investigation of viruses in omics datasets. Hence, we controlled the other steps in the bioinformatics analysis and applied a widely adopted strategy for VLP enrichment. Still, there are limitations. Firstly, there is no golden standard to investigate the virome in a biological sample since a real sample contains different viruses and bacteria, fungi, and protozoa. The in silico mock viromics with only several types of viruses evaluated elsewhere was never a real sample estimation. Secondly, we only tried on widely applied viral particle enrichment methods and did not investigate other methods. Finally, the sample size here is relatively small to capture all the differences between cancer phenotype and healthy controls indicating that a

larger cohort study is needed to further validate the biological/clinical importance of our methodological findings.

This study critically assessed the efficiency of two widely applied methods, WGSM and VLPM, in characterizing viruses. The significant impact of virus investigation by the two methods, length cut-offs and coverage rate of the assembled contig, phenotypes of samples and even sample types were identified in this study. Although both methods can identify the alpha and beta diversity patterns and most viral signatures in comparative experiential design, it should be noted that each method preferentially detects certain viral signatures. Hence, where completeness of the virome signature is the aim of the research a comprehensive technical routine by combining various methods will aid in capturing total viruses in the specimen.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/v15010053/s1. Figure S1. (a) The schematic representation of this study. Faecal samples of 6 HCC patients and 6 healthy controls were collected and then sequenced using WGSM and VLPM, respectively. (b) Study design. (c) Evaluation of sequencing depth. Genetic contents flow was depicted. Figure S2. The rarefaction curves of the samples given (a) contig length $\geq$ 3 kb, (b) contig length $\geq$ 5 kb and (c) contig length $\geq$ 8 kb. The coverage rate remains $\geq$ 75%. (d) The alpha diversities (Shannon index) given contig length $\geq$ 10 kb and coverage rate $\geq$ 75%. The alpha diversities (observed features) given (e) contig length $\geq$ 3 kb, (f) contig length $\geq$ 5 kb and (g) contig length $\geq$ 8 kb. The coverage rate remains$\geq$ 75%. Figure S3. The rarefaction curves of the samples given (a) coverage rate remains $\geq$10%, (b) coverage rate remains$\geq$ 25%, (c) coverage rate remains$\geq$ 50% and (d) coverage rate remains$\geq$ 90%. The contig length remains$\geq$ 10 kb. The alpha diversities (observed features) given (e) coverage rate remains$\geq$ 10%, (f) coverage rate remains$\geq$ 25%, (g) coverage rate remains$\geq$ 50% and (h) coverage rate remains$\geq$ 90%. The contig length remains$\geq$ 10 kb. (i) The ratio of reads mapped to virus contigs to those mapped to all the assembled contigs. Figure S4. The principal coordinates analysis plots based on (a) the WGSM samples, (b) the VLPM samples, (c) the healthy controls, (d) the HCC patients and (e) the combination of all the samples, respectively. The plots used Jaccard dissimilarities. (f) Virus compositions at the family level. Table S1. QC table of the sequencing data. The number of reads from raw data and those after deduplicates, host contamination removal, low-quality reads and removal of rRNA reads. Table S2. The STORMS checklist. Table S3. Enriched viral signatures from MaAsLin2. Abundance of contigs with same annotations were merged into one representative record. Table S4. The FPKM of the detected RNA Viruses.

Author Contributions: X.-T.J. and E.E.-O. conceived the project. X.-T.J. and F.Z. designed the analysis. A.Z., J.B. and L.G. diagnosed the patients and collected the samples. G.L.H. and A.G. contributed to the wet lab. G.C., X.T. and Y.S. contribute to the RNA viruses analysis. X.-T.J. and F.Z. wrote the manuscript. E.E.-O. and all authors contributed and approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by grants from the Federal Government through the St George and Sutherland Medical Research Foundation, and Hong Kong Research Grants Council (RGC) General Research Fund (GRF) 11206819 and Hong Kong Innovation and Technology Fund (ITF) MRP/071/20X.

Informed Consent Statement: Not applicable.

Data Availability Statement: Raw sequencing data are deposited at NCBI SRA with accession number PRJNA755142.

Acknowledgments: X.-T.J. and F.Z. acknowledge the support of Microbiome Research Centre bioinformatics initiate funding from state government Australia and the UNSW IT.

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

References

1. Whitman, W.B.; Coleman, D.C.; Wiebe, W.J. Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6578–6583. [CrossRef] [PubMed]
2. Suttle, C.A. Marine viruses—Major players in the global ecosystem. *Nat. Rev. Microbiol.* **2007**, *5*, 801–812. [CrossRef] [PubMed]

3. Cao, J.; Wang, C.; Zhang, Y.; Lei, G.; Xu, K.; Zhao, N.; Lu, J.; Meng, F.; Yu, L.; Yan, J.; et al. Integrated gut virome and bacteriome dynamics in COVID-19 patients. *Gut. Microbes* **2021**, *13*, 1–21. [CrossRef]
4. Neil, J.A.; Cadwell, K. The Intestinal Virome and Immunity. *J. Immunol.* **2018**, *201*, 1615–1624. [CrossRef] [PubMed]
5. Norman, J.M.; Handley, S.A.; Baldridge, M.T.; Droit, L.; Liu, C.Y.; Keller, B.C.; Kambal, A.; Monaco, C.L.; Zhao, G.; Fleshner, P.; et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **2015**, *160*, 447–460. [CrossRef]
6. Zheng, D.W.; Dong, X.; Pan, P.; Chen, K.W.; Fan, J.X.; Cheng, S.X.; Zhang, X.Z. Phage-guided modulation of the gut microbiota of mouse models of colorectal cancer augments their responses to chemotherapy. *Nat. Biomed. Eng.* **2019**, *3*, 717–728. [CrossRef]
7. Nakatsu, G.; Zhou, H.; Wu, W.K.K.; Wong, S.H.; Coker, O.O.; Dai, Z.; Li, X.; Szeto, C.H.; Sugimura, N.; Lam, T.Y.; et al. Alterations in Enteric Virome Are Associated with Colorectal Cancer and Survival Outcomes. *Gastroenterology* **2018**, *155*, 529–541.e5. [CrossRef]
8. Lang, S.; Demir, M.; Martin, A.; Jiang, L.; Zhang, X.; Duan, Y.; Gao, B.; Wisplinghoff, H.; Kasper, P.; Roderburg, C.; et al. Intestinal virome signature associated with severity of nonalcoholic fatty liver disease. *Gastroenterology* **2020**, *159*, 1839–1852. [CrossRef]
9. 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]
10. Rosario, K.; Breitbart, M. Exploring the viral world through metagenomics. *Curr. Opin. Virol.* **2011**, *1*, 289–297. [CrossRef]
11. Briese, T.; Kapoor, A.; Mishra, N.; Jain, K.; Kumar, A.; Jabado, O.J.; Lipkin, W.I. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *Mbio* **2015**, *6*, e01491-15. [CrossRef] [PubMed]
12. Wylie, T.N.; Wylie, K.M.; Herter, B.N.; Storch, G.A. Enhanced virome sequencing through solution-based capture enrichment. *Genome Res.* **2015**, *32*, gr-191049.
13. Waller, A.S.; Yamada, T.; Kristensen, D.M.; Kultima, J.R.; Sunagawa, S.; Koonin, E.V. Bork Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME J.* **2014**, *8*, 1391–1402. [CrossRef] [PubMed]
14. 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]
15. Camarillo-Guerrero, L.F.; Almeida, A.; Rangel-Pineros, G.; Finn, R.D.; Lawley, T.D. Massive expansion of human gut bacteriophage diversity. *Cell* **2021**, *184*, 1098–1109.e9. [CrossRef]
16. Paez-Espino, D.; Eloe-Fadrosh, E.A.; Pavlopoulos, G.A.; Thomas, A.D.; Huntemann, M.; Mikhailova, N.; Rubin, E.; Ivanova, N.N.; Kyrpides, N.C. Uncovering Earth's virome. *Nature* **2016**, *536*, 425–430. [CrossRef]
17. 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]
18. Lewandowska, D.W.; Zagordi, O.; Geissberger, F.D.; Kufner, V.; Schmutz, S.; Böni, J.; Metzner, K.J.; Trkola, A.; Huber, M. Optimization and validation of sample preparation for metagenomic sequencing of viruses in clinical samples. *Microbiome* **2017**, *5*, 94. [CrossRef]
19. 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]
20. Santos-Medellin, C.; Zinke, L.A.; Ter Horst, A.M.; Gelardi, D.L.; Parikh, S.J.; Emerson, J.B. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.* **2021**, *15*, 1956–1970. [CrossRef]
21. Mukhopadhya, I.; Segal, J.P.; Carding, S.R.; Hart, A.L.; Hold, G.L. The gut virome: The 'missing link'between gut bacteria and host immunity? *Therap. Adv. Gastroenterol* **2019**, *12*, 1756284819836620. [CrossRef] [PubMed]
22. Quince, C.; Walker, A.W.; Simpson, J.T.; Loman, N.J.; Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **2017**, *35*, 833–844. [CrossRef] [PubMed]
23. Pal, C.; Bengtsson-Palme, J.; Kristiansson, E.; Larsson, D.G. The structure and diversity of human, animal and environmental resistomes. *Microbiome* **2016**, *4*, 54. [CrossRef] [PubMed]
24. Noyes, N.R.; Weinroth, M.E.; Parker, J.K.; Dean, C.J.; Lakin, S.M.; Raymond, R.A.; Rovira, P.; Doster, E.; Abdo, Z.; Martin, J.N.; et al. Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. *Microbiome* **2017**, *5*, 142. [CrossRef] [PubMed]
25. Sutton, T.D.S.; Clooney, A.G.; Ryan, F.J.; Ross, R.P.; Hill, C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **2019**, *7*, 12. [CrossRef]
26. Tisza, M.J.; Buck, C.B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2023202118. [CrossRef]
27. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner. No. LBNL-7065E*; Ernest Orlando Lawrence Berkeley National Laboratory: Berkeley, CA, USA, 2014.
28. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [CrossRef]
29. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
30. Kopylova, E.; Noé, L.; Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **2012**, *28*, 3211–3217. [CrossRef]
31. 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]
32. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]

33. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]
34. Roux, S.; Emerson, J.B.; Eloe-Fadrosh, E.A.; Sullivan, M.B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **2017**, *5*, e3817. [CrossRef] [PubMed]
35. Kieft, K.; Zhou, Z.; Anantharaman, K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **2020**, *8*, 90. [CrossRef] [PubMed]
36. Guo, J.; Bolduc, B.; Zayed, A.A.; Varsani, A.; Dominguez-Huerta, G.; Delmont, T.O.; Pratama, A.A.; Gazitúa, M.C.; Vik, D.; Sullivan, M.B.; et al. VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **2021**, *9*, 37. [CrossRef] [PubMed]
37. Chen, G. VirBot: An RNA Viral Contig Detector for Metagenomic Data. Available online: https://github.com/GreyGuoweiChen/RNA_virus_detector (accessed on 20 June 2022).
38. Wood, D.E.; Lu, J.; Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **2019**, *20*, 257. [CrossRef]
39. Garcia, B.J.; Simha, R.; Garvin, M.; Furches, A.; Jones, P.; Hyatt, P.D.; Schadt, C.; Pelletier, D.; Jacobson, D. *Kraken2 Metagenomic Virus Database*; Oak Ridge National Lab: Oak Ridge, TN, USA, 2020. [CrossRef]
40. Nordberg, H.; Cantor, M.; Dusheyko, S.; Hua, S.; Poliakov, A.; Shabalov, I.; Smirnova, T.; Grigoriev, I.V.; Dubchak, I. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids. Res.* **2014**, *42*, D26–D31. [CrossRef]
41. Oksanen, J. Vegan: Community Ecology Package. R Package Version 1.17-9. 2011. Available online: http://cran.r-project.org/package=vegan (accessed on 19 August 2020).
42. 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]
43. Mallick, H.; Rahnavard, A.; McIver, L.J.; Ma, S.; Zhang, Y.; Nguyen, L.H.; Tickle, T.L.; Weingart, G.; Ren, B.; Schwager, E.H.; et al. Multivariable Association Discovery in Population-scale Meta-omics Studies. *PLoS Comput. Biol.* **2021**, *17*, e1009442. [CrossRef]
44. Mirzayi, C.; Renson, A.; Furlanello, C.; Sansone, S.-A.; Zohra, F.; Elsafoury, S.; Geistlinger, L.; Kasselman, L.J.; Eckenrode, K.; van de Wijgert, J.; et al. Genomic Standards Consortium. et al. Reporting guidelines for human microbiome research: The STORMS checklist. *Nat. Med.* **2021**, *27*, 1885–1892. [CrossRef]
45. Gweon, H.S.; Shaw, L.P.; Swann, J.; De Maio, N.; AbuOun, M.; Niehus, R.; Hubbard, A.T.M.; Bowes, M.J.; Bailey, M.J.; Peto, T.E.A.; et al. The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environ. Microbiome* **2019**, *14*, 7. [CrossRef] [PubMed]
46. Pereira-Marques, J.; Hout, A.; Ferreira, R.M.; Weber, M.; Pinto-Ribeiro, I.; van Doorn, L.J.; Knetsch, C.W.; Figueiredo, C. Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Front. Microbiol.* **2019**, *10*, 1277. [CrossRef] [PubMed]
47. 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] [PubMed]
48. Zaheer, R.; Noyes, N.; Ortega Polo, R.; Cook, S.R.; Marinier, E.; Van Domselaar, G.; Belk, K.E.; Morley, P.S.; McAllister, T.A. Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci. Rep.* **2018**, *8*, 5890. [CrossRef]
49. Shkoporov, A.N.; Hill, C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe* **2019**, *25*, 195–209. [CrossRef]
50. Minot, S.; Bryson, A.; Chehoud, C.; Wu, G.D.; Lewis, J.D.; Bushman, F.D. Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 12450–12455. [CrossRef]