Profiling of Human Gut Virome with Oxford Nanopore Technology

Jiabao Cao[1,2,#], Yuqing Zhang[1,2,#], Min Dai[3], Jiayue Xu[1], Liang Chen[1], Faming Zhang[3,4], Na Zhao[1,*], Jun Wang[1,*]

Author Affiliations:

1. CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China.

2. University of Chinese Academy of Sciences, Beijing 100049, China.

3. Medical Center for Digestive Diseases, the Second Affiliated Hospital of Nanjing Medical University, Nanjing 210011, China.

4. Key Lab of Holistic Integrative Enterology, Nanjing Medical University, Nanjing 210011, China.

[#] Jiabao Cao, and Yuqing Zhang contributed equally to this manuscript.

* Correspondence to:

Professor Jun Wang, CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, No. 1-3 Beichenxi Road, Chaoyang District, Beijing, China; junwang@im.ac.cn; Dr. Na Zhao, CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, No. 1-3 Beichenxi Road, Chaoyang District, Beijing, China; zhaona@im.ac.cn.

Abstract

Human gut virome play critical roles in maintaining gut microbial composition and functionality, as well as host physiology and immunology. Yet, there are insufficient amount of studies on this topic mainly due to methodological limitations, including enrichment of viruses (phages and host viruses) as well as short read-length from current sequencing technology. Here we developed a full working protocol for analyzing human gut virome using physical enrichment, reverse transcription and random amplification, and eventually the state-of-art single-molecule real-time sequencing (SMRT) platform of Oxford Nanopore Technology (ONT). We demonstrate that sequencing viral DNA directly, or viral DNA/cDNA after amplification using ONT achieves much longer reads and provides more information regarding virome diversity, many of the virome sequences do not have match in current databases. Moreover, direct DNA sequencing of virome provides first insights into the epigenetic modifications on phages, where signals of methylations can be directly detected. Our study demonstrates that progressing sequencing technology and bioinformatic improvements will bring more knowledge into virome composition, diversity and potentially their important functions.

Highlights:

1. Virus-like particles were enriched from human stool samples;
2. Viral nucleotides were sequenced with Oxford Nanopore Technology with and without amplification;
3. Gut virome in humans showed highly individualized composition;

4. Novel sequences and contigs were found to be the majority in the resulted sequences;

5. Epigenetic modifications were detected directly on virus genomes.

Keywords:

Human gut virome; Enrichement; Amplification; Oxford Nanopore Technology; Epigenomics

Introduction

The human gut is home to tremendous amount of microbes [1]. They inhabit different ecological niches in the gut, forming complex interaction networks between themselves and with the human cells, and the dynamic balance between gut microbiome and host is required for human health [2-6]. Studies in human cohort and mouse models, among others, have confirmed that gut microbial communities are associated with increasing number of some of metabolism diseases and infectious diseases, providing insights as well as potential targets for future monitoring and therapies [3, 7-14].

The gut microbiome contains bacteria, archaea, fungi, protozoa, and, lastly but not leastly virus. The most abundant cellular members of the microbiome are bacteria and archaea (account for more than 99% of biomass), and have received most attention in human microbiome studies over the years [15-17]. Yet, the advances in next-generation sequencing (NGS) technology and bioinformatic tools have also facilitated the development of human virome studies. Metagenomic analysis suggests that the gut of healthy humans harbors commensal virus, including phages, DNA virus and RNA virus [18-22]. Virome (phages and other host viruses) play roles in intestinal physiology, enteric immune system, host health and disease [23, 24]. The dynamic balance between the virome and the intestinal immune system is finely regulated by cytokines secreted by immune cells [20, 25]. For instance, virome changes in inflammatory bowel disease (IBD) (Crohn’s disease and ulcerative colitis) are disease specific [26]. Phages residing in mucosal surfaces can influence the host

by providing nonhost-derived immunity against bacterial infections [26, 27]. By inducing interferons (IFNs), commensal virus can protect from gut inflammation during tissue damage [28, 29]. However, using current short read sequencing technologies, such as Illumina, can only offer knowledge on gut virome that is both biased and fragmentary.

Oxford Nanopore Technologies (ONT) as one of the emerging single-molecule real-time sequencing technology (SMRT) has the advantage of rapid library preparation, ultra-long reads and real-time data acquisition [30-32]. For virome, ONT sequencing has the potential to acquire virus genome by producing genome-length reads that cover all of mutation within a single virus particle. In addition, biological nanopores are able to discriminate not only the genome but also single base modifications such as 5-methylation of cytosine (5mC for DNA and m5C for RNA) and 6-methylation of adenine (6mA for DNA and m6A for RNA) in native DNA/RNA [32]. Increasing evidence in last years suggests that DNA/RNA methylation can influence biological function, including regulation of DNA/RNA replication and repair, and gene expression [33-36]. Recently, *Oliveira* et al. reported that DNA methyltransterase in *Clostridioides difficile* has able to mediate sporulation, *C. difficile* disease transmission and pathogenesis [37]. *Xue* et al. reported viral N^6-methyladenosine could upregulate replication and pathogenesis of human respiratory syncytial virus [38]. These findings suggest that epigenetic regulation is also important for the pathogenesis of important pathogens.

To profile the gut virome in healthy adults, including identity as well as potential

93 epigenetic information in the virome, we developed a protocol combining physical
94 enrichment, optional reverse transcription and amplification of nucleotides, and
95 bioinformatic analytical pipelines, and firstly characterized the virome in five healthy
96 humans using the ONT PromethION platform. We were able to generate long reads
97 for virome up to tens of kilobases, resulting in many novel contigs that do not have
98 matches in the available databases, and also for abundant virus we could detect
99 epigenetic signals. These discoveries are instructive to future investigations into the
100 genomics, epigenomics and potential function of human gut virome.
101
102 *Material and Methods*
103 1. Enrichment and purification of virus-like particles (VLPs)
104 Each frozen faecel samples (approximately 1.5 g) from five individuals who
105 provided written informed consent were resuspended in 15 ml sterile Phosphate
106 Buffered Saline and homogenized thoroughly. The suspension was centrifuged at
107 4,500 rpm for 10 min at the 4 ℃ to remove large food residues (Beckman Coulter
108 Allegra™ X-22R). Transfer the supernatant to fresh tubes and centrifuged at 4,500
109 rpm for 10 min at the 4 ℃ again. The suspernatant was filtered through 0.45 μm
110 PVDF membrane (Millipore) to remove eukaryotic and bacterial cell-sized particles
111 before ultracentrifugating at 180,000× g for 3 hours at the 4 ℃ (Beckman Coulter
112 XP-100). The pellets were resuspended in 400 μl sterile Phosphate Buffered Saline
113 and treated with 8 U of TURBO DNase Ⅰ (Ambion) and 20 U of RNase A (Fermentas)
114 at 37 ℃ for 30 min. The viral nucleic acids (DNA and RNA) were extracted by using

115 QIAamp MinElute Virus Spin Kit (Qiagen) following the manufacturer's instructions
116 and eluted into RNase-free water. [39, 40]

117 2. Reverse transcription and random amplification

118 Viral first strand cDNA was synthesized in a 20 μl reaction mixture with 13 μl of
119 purified vial nucleic acids from each sample and 100 pmol of primer Rrm
120 (5'-GACCATCTAGCGACCTCCAC - NNNNNN-3'), as previously described
121 [39-41]. For the double-strand cDNA synthesis, 100 pmol of primer Rrm and Klenow
122 fragment (3.5 U/μl; Takara) were added. Random amplification was conducted with
123 8μl of the double-strand cDNA template in a final reaction volume of 200μl, which
124 contained 4μM primer Rm (5'-GCCGGAGCTCTGCAGAATTC-3'), 90 μM dNTPs
125 each, 80 μM Mg^{2+}, 10x Buffer and 1 U of KOD-Plus DNA polymerase (Toyobo). The
126 amplification product was purified by agarose gel electrophoresis and QIAquick Gel
127 Extraction Kit (Qiagen).

128 3. PromethION library preparation and sequencing

129 PromethION library preparation was performed according to the manufacturer's
130 instructions for the barcoding cDNA/DNA and native DNA (SQK-LSK109 and
131 EXP-NBD104). When multiplexing, all the samples were pooled together. ONT
132 MinKNOW software (v.19.10.1) was used to collect raw sequencing data, and Guppy
133 (v.3.2.4) was used for local base-calling of the raw data after sequencing runs were
134 completed. The PromethION was run for up to 96 h.

135 4. ONT sequence analysis and assembly

136 Qcat (Oxford Nanopore Technology), python command-line tool for

demultiplexing ONT reads from FASTQ files, was used to trim adaptor and barcode sequences. With genomeSize = 2k and default parameters, trimmed raw reads were analysis using Canu v1.9 [42] for virome genome de novo assembly, which includes read correction, read trimming and contig construction.

5. Matching to current database

Raw reads qcat trimmed were analysis using mimimap2 [43] to identify gut virome composition, which aligned reads to the reference genome in the National Center for Biotechnology Information (NCBI) virus genome database, including all known viruse genome sequences. To improve the accuracy of the viral taxonomy, two following criteria were adapted: (1) the depth of coverage of reference viral genome >= 5X; (2) the breadth of coverage of the reference viral genome >= 50%.

To assign the taxonomy of assembled contigs by Canu, two approaches and three databases were applied, including minimap2, blastn, NCBI virus genome database, The human gut virome database (GVD) and NCBI nucleotide database. The filter criteria of alignment results of contigs by minimap2 was the same with raw reads, while contig matched length >= 1000 bp with nucleic similarity >= 98% and e-value <= 10^{-5} was adapted as the identified criteria of viruses by blastn.

6. Identification and annotation of bacteriophage ORFs

Seeker [44], a new prediction tool via deep learning framework, was used to identify putative phages from contigs in amplified cDNA/DNA group with default parameters. According to the multiPhATE pipeline [45], ORFs in phages were identified by PHANOTATE [46], a tool to annotate phage genomes. Consequently,

amino sequences of ORFs were aligned to Phantome (http://www.phantome.org) and pVOGs [47] databases by blastp with parameters “percent of identity >= 60, e-value <= 0.01” and hmm searched to pVOGs by jackhmmer [48] with default parameters, respectly. ORFs repeated in two or more samples were clusterd by usearch [49] with percent of indentity >= 99. To validate these ORFs, we mapped our in-house NGS data using bowtie2 [50] with default parameters. Coverage of ORFs was calculated by weeSAM (https://github.com/centre-for-virus-research/weeSAM) and only ORFs with mapped reads >= 10 were counted.

7. Methylation analysis

Tombo v1.5 was used to detect the methylation states of nucleic acids from raw DNA samples [51]. A log-likelihood threshold of 2.5 was used to call methylation and the filter cutoff of methylation sites was defined by the estimated fraction of significantly modified reads >= 0.7 and coverage depth >= 10X. Different methylation sites of 5-methylcytosine (5mC) and N6-Methyladenine DNA Modification (6mA) were visualized by Integrative Genomics Viewer (IGV Version 2.5.3) [52] with default parameters and the putative methyltransferase recognition motifs were identified by MEME (Version 5.0.5) [53] with the following parameters: “-dna –mod zoops”, and webLogo (https://weblogo.berkeley.edu/logo.cgi) was used to plot the logo of the motifs we identified.

8. Accession number

The sequencing data were deposited at GSA (Genome Sequence Archive) under BioProject accession no. PRJCA002499. The full protocol is available at

https://github.com/caojiabao/VirPipeline.

Results

1. Virome separation, enrichment and sequencing

Since metagenomic sequencing using fecal DNA usually results in only minor fraction of virome sequences, and most of reads will be either from bacteria or archaea, enrichment of viruses are necessary for studying virome in human gut samples. Thus, we have combined a series of enrichment methods including filtration and super centrifugation, to enrich for virus-like-particles (VLPs) in the fecal samples (Figure 1). After VLPs were isolated, additional DNase/RNase treatment were used to remove any potential free-DNA that were not virus-originated. The left-over DNA/RNA were quantified, and half of the nucleotides were directly subjected to ONT DNA library preparation, to profile the DNA virus abundances as well as methylation; and the other half were first subjected to RNA reverse transcription and then amplificaiton with short random primers, to have RNA virus sequenced as cDNA, and improve the chance of low-abundance DNA viruses to be detected in the sequencing results.

In our study as a primary investigation, we have first profiled five healthy volunteers' fecal samples with our protocol. Virus-like particle (VLP) fractions of five individuals were enriched, and raw DNA, as well as enriched cDNA/DNA were sequenced using ONT PromthION platform. With one flowcell, the ONT PromthION yielded a total of 8.2 Gb raw data, with a median of 1.7 Gb per sample in amplified

group; and 452 Mb raw data, with a median of 67 Mb per sample in raw DNA group (Table S1).

2. Virome composition in healthy individuals revealed by ONT sequencing

With sequencing reads from amplified cDNA/DNA results, we have mapped the ONT reads onto NCBI virus genomes and then analyzed the composition across five individuals. Consistent with other studies, our result from amplified cDNA/DNA of virome showed that bacteriophage families were the most frequently detected and accounted for the majority of the intestinal virome in number. The final catalogue of bacteriophages included the *Caudovirals* order (families *siphovirdae*, *podoviridae*), family *Inoviridae* and family *Microviridae*. Meanwhile, the eukaryotic CRESS-DNA viruses (family *genomoviridae*) was also detected (Figure 2; Table S2). Of special note was the presence of the RNA plant viruses including family *Virgaviridae* and *Alphaflexiviridae*, which showed good agreement with the findings of Shkoporov *et al* [54]. We observed that individual specificity is probably a feature of the faecal viral communities (Figure 2), which had been already demonstrated by several previous studies [55-57]. However, uncultured phage WW-nAnB strain 3 belonging to family *inoviridae* was detected to be presented in amplified cDNA/DNA among five individuals.

To characterize the potential biases regarding virome composition caused by amplification, we have compared the results from amplified cDNA/DNA and raw virome DNA. We could not compare raw RNA due to the fact that it is still not yet possible to multiplex RNA samples on ONT platforms. The relative abundance of

virus each virome was defined using relative proportion of each virus in terms of breadth of coverage on the assembled genome, similar to the definition of bacterial abundances in metagenomic studies. As expected, numbers of viruses are higher in amplified cDNA/DNA results, except for individual 2 and individual 5 who remained the same in terms of virus diversity. Further, abundances of the common existing viruses between raw DNA and amplified cDNA/DNA showed essential shifts in all five individuals, demonstrated detectable bias of reverse transcription as well as random amplification approach we adapted.

3. Virome assembly using ONT sequences

We next focused on the assembly of nanopore sequencing reads in five individuals. Assembler Canu was used to assemble the virome sequences separately from raw DNA and amplified cDNA/DNA groups into contigs, which yielded a total of 1564 contigs, with a median of 15 and 347 contigs per sample for raw DNA and amplified cDNA/DNA group. Average length of contigs from raw DNA group was longer than those from amplified cDNA/DNA group (Figure S1). The contigs vary largely in length, ranging from 1kb to 53kb (Figure S2). Consequently, we obtained the identity of certain contigs by matching with NCBI virus genomes, human gut virus database (GVD) [58] and NCBI nucleotide databases, however all with very low matching rate, suggesting a large collection of potentially novel genomes in our results (Table S3). Thus, Seeker was used to identify bacteriophages in amplified cDNA/DNA group. As a result, more than 50% of contigs per sample were identified as phages in amplified cDNA/DNA group. To characterize these bacteriophages, we

247 performed phages ORFs prediction and functional annotation, which yielded average
248 7 ORFs per contig, ranging from 6 to 10, and average 3 ORFs per kb in length. In
249 addition, ORFs repeated in two or more samples were validated by mapping to
250 additional Illumina Hiseq sequencing data of metagenomics from same samples. We
251 founded that over half ORFs (59.3%, 35/59) can be matched, indicating that these
252 phages stably exist in our data (Table S4).

253 To examine the potential of covering full virus genome of long sequence
254 produced from the oxford nanopore technology sequencing, we mapped raw reads to
255 the contigs by canu separated from amplified cDNA/DNA and raw DNA. We
256 aggregated and counted the length of raw reads aligned to the longest contigs of our
257 choice from these two groups, and calculated the proportion of raw reads to contigs in
258 length (Figure 3). In amplified VLP cDNA/DNA group, the max value of reads length
259 in each sample was all more than 15% of contigs in length, the highest achieving 40%
260 of contigs' full length in individual 4 (Figure S3). In raw DNA group, this proportion
261 was higher in general , understandably due to the fact random amplification usually
262 can not reach full length available in the raw DNA. A small number of reads were
263 longer than the final contigs (Figure 3), for which the possible reason for this result
264 was that some long reads were trimmed by canu during assembly, due to the lower
265 quality of part of the sequences being abandoned by Canu.

266 4. Viral epigenomic detection using ONT

267 In addition, methylation states of bases in DNA from reads can be detected
268 directly by the Oxford Nanopore sequencing without extra laboratory techniques. In

this study, we have analyzed methylation signals on a few contigs that reached >10X coverage in raw DNA data in any of individuals. Methylation detection can not be carried out on cDNA and amplified DNA samples, for they will lose all the methylation signals. For the only one of the contigs (contig00000015) with known identity (Uncultured crAssphage), we detected in total 17 5mC and 120 6mA methylation sites covered in the 8kb genome (Figure 4). For 5mC and 6mA methylation, the nucleotide motifs YCHYTTA***C***TWMRECT (e-value = 1.2×10^{-2}) and motif MADWDTW***A***NADYYWW (e-value = 2.5×10^{-4}) were identified respectively, with the methylated nucleotide highlighted in bold italic.

We have also found another 4 contigs with >10X coverage, but they do not have detectable relative in the databases we have searched against. They are potentially novel viruses and very unlikely to be bacteria or archaea, as our protocol has removed most of the none VLPs, and bacteria/archaea contigs would most likely have matches in the databases we have searched. They have 5mC methylation sites ranging from 0.3 to 1.8 (% of genome) and 6mA methylation sites ranging from 0.7 to 2.5 (% of genome). The motifs for methylation are also extremely diverse, including NHHYYKG***C***DHNNN, WDWADDW***C***DWYNDDW, MNNNNTR***C***GBNNNND and HDNBYDV***C***VVVVNNH for 5mC, and WHWHNYD***A***HNNYYHTT, VNWWDWH***A***YBYNNNT, DRNVRKK***A***BBNDNNN and NRNARND***A***SYAHHNH for 6mA (Figure S4). Since most of the methylation studies are performed in eukaryotes, and only starting in bacteria with limited information available, it is yet difficult to compare the methylation profiles to understand its

underlying mechanisms.

Discussion

Our study combined the physical enrichment of VLPs in fecal samples, nucleotide amplification with the latest sequencing technology to establish a complete workflow of human gut virome profiling. With longer reads as well as richer information of additional epigenetic modifications, developments in sequencing technology could bring another round of revolution in metagenome as well as virome investigations. More importantly, despite the complex steps before sequencing were designed for maximum enrichment of VLPs, as well as removal of any DNA/RNA that were not of virus origin, and thus inevitably makes it relatively time-consuming, ONT could carry out sequencing and produce reads nearly simultaneously, making it possible to finish data generation from samples with five days of working time, and potentially even shorter if PromethION was run for less time, or data were analyzed during the course of being generated (real-time); comparatively, Illumina based platforms usually takes longer to generate enough read length for downstream analyses. There are several studies who have already utilized this property for fast pathogen detection in infectious diseases [59-61], there are cases of virome analysis that might require such time efficiency as well, and our protocol provides a feasible choice for virome studies that also requires short turn-around time.

We also compared the effect of reverse transcription and consequent amplifications on virome analysis with ONT. Such steps were used for several reasons.

Firstly, despite ONT is capable of sequencing DNA or RNA directly, multiplexing is still only possible for DNA libraries, and directly sequencing RNA is not yet cost-effective for virome analysis, while cDNA is a better alternative. Also, to utilize the capacity of sequencing on ONT, very high concentrations of DNA or RNA libraries are required to generate enough reads, yet this is also very difficult for virome DNA/RNA, who usually only reaches 10% of the required input from 1.5 grams of fecal samples. Amplification does lead to higher amount of reads and enables detection of low abundance viruses to be detected, as revealed in our study; but it also leads to on average shorter reads and certain biases in the estimations of virus abundances, resulting from both affinity to random primers as well as PCR produced artefacts. Lastly, amplified cDNA/DNA loses all the methylation modifications on the viral nucleotides and prevents investigations into this potentially vital epigenetic information. Thus, future investigator needed to balance the pros and cons of amplification processes, and could take advantage of our approach of using both raw DNA (and/or RNA) and amplified cDNA/DNA for sequencing, gaining complementary information within the same sequencing run.

The profiles of virome in our studied individuals suggest a highly diverse virome, with only small amount of “core” viruses shared in between. This core could shrink even further with increasing number of individuals while the total diversity of viruses increase, which we plan to investigate in the future. We found phages making the majority part of the virome in the human gut, plus a few host viruses; while the mystery of plant RNA viruses is again present with ONT data, many reads achieving >

1kb in length, whether they are left-overs from our plant-based food, or rather human viruses with their closest relatives in the plant-associated viruses, still call for more investigations, especially within functional experiments. Our data suggest that there are potentially high number of unknown, novel viruses in the human gut, as our assembled contigs have very low rate of matches in the current databases; we consider those not likely to be contaminations due to our vigorous depletion of any non-VLPs and non-viral nucleotides, plus the fact that they do not have any match in nucleotide collection of NCBI either.

It's also the first time we demonstrate that the phage genomes are methylated via direct sequencing. RSV viruses and influenza viruses are known to have m6A methylation on their RNA genome [38, 62], detected with more complex methods with low throughput, while DNA phages (or other DNA viruses) are not yet studied to our knowledge. In *E. coli* and *C. difficile* it is shown that 6mA is the main form of DNA methylation, while eukaryotes usually lack this form, and in our results phages also have 6mA as the main form of DNA methylation [37, 63, 64]. Since DNA methylations play an important role in bacterial defense against phages, how phage genome becomes methylated, and the consequent impact on phage life cycle and interactions with bacterial hosts remain to be explored with dedicated studies. Besides, it remains possible the motifs of methylation between phage and bacteria are intrinsically linked, and provide additional information to determine the host range of phages; this would require increasing the current knowledge on epigenetics of both bacteria and phages in the future.

Conclusions

To summarize, we developed and pilot-tested a thorough protocol for human gut virome analysis using the lastest ONT sequencing platform, and generated novel insights into the individuality, diversity of gut virome with new sequencing data. Our approach of course can be applied for other virome studies, including animal gut, soil and water virome etc., and accumulating both sequences as well as epigenetic information on those samples, have the long potential of opening up new directions in metagenomic, microbiological and medical researches.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (grant number 2018YFC2000500), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number XDB29020000), and the National Natural Science Foundation of China (grant number 31771481 and 91857101).

References

[1] Sender R, Fuchs S. Milo R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell 2016;164(3):337-40. https://doi.org/10.1016/j.cell.2016.01.013.

[2] Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level

379 analysis of gut microbiome variation. Science 2016;352(6285):560-4. https://
380 doi.org/10.1126/science.aad3503.

381 [3] Moschen AR, Gerner RR, Wang J, Klepsch V, Adolph TE, Reider SJ, et al. Lipocalin 2
382 Protects from Inflammation and Tumorigenesis Associated with Gut Microbiota Alterations. Cell
383 Host Microbe 2016;19(4):455-69. http://doi.org/10.1016/j.chom.2016.03.007.

384 [4] Wang J, Thingholm LB, Skieceviciene J, Rausch P, Kummen M, Hov JR, et al. Genome-wide
385 association analysis identifies variation in vitamin D receptor and other host factors influencing
386 the gut microbiota. Nat Genet 2016;48(11):1396-406. https:// doi.org/10.1038/ng.3695.

387 [5] Tschurtschenthaler M, Wang J, Fricke C, Fritz TMJ, Niederreiter L, Adolph TE, et al. Type I
388 interferon signalling in the intestinal epithelium affects Paneth cells, microbial ecology and
389 epithelial regeneration. Gut 2014;63(12):1921-31. https:// doi.org/10.1136/gutjnl-2013-305863.

390 [6] Wang J, Chen L, Zhao N, Xu XZ, Xu YK. Zhu BL. Of genes and microbes: solving the
391 intricacies in host genomes. Protein Cell 2018;9(5):446-61. https://
392 doi.org/10.1007/s13238-018-0532-9.

393 [7] Belkaid Y. Hand TW. Role of the Microbiota in Immunity and Inflammation. Cell
394 2014;157(1):121-41. https://doi.org/10.1016/j.cell.2014.03.011.

395 [8] Wang ZN, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, et al. Gut flora metabolism
396 of phosphatidylcholine promotes cardiovascular disease. Nature 2011;472(7341):57-U82.
397 https://doi.org/10.1038/nature09922.

398 [9] Ridaura VK, Faith JJ, Rey FE, Cheng JY, Duncan AE, Kau AL, et al. Gut Microbiota from
399 Twins Discordant for Obesity Modulate Metabolism in Mice. Science 2013;341(6150):1079-U49.
400 https://doi.org/10.1126/science.1241214.

401 [10] Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe
402 interactions have shaped the genetic architecture of inflammatory bowel disease. Nature
403 2012;491(7422):119-24. https://doi.org/10.1038/nature11582.
404 [11] Giongo A, Gano KA, Crabb DB, Mukherjee N, Novelo LL, Casella G, et al. Toward defining
405 the autoimmune microbiome for type 1 diabetes. Isme J 2011;5(1):82-91.
406 https://doi.org/10.1038/ismej.2010.92.
407 [12] Cox LM. Blaser MJ. Pathways in Microbe-Induced Obesity. Cell Metab 2013;17(6):883-94.
408 https://doi.org/10.1016/j.cmet.2013.05.004.
409 [13] Zhu WF, Gregory JC, Org E, Buffa JA, Gupta N, Wang ZN, et al. Gut Microbial Metabolite
410 TMAO Enhances Platelet Hyperreactivity and Thrombosis Risk. Cell 2016;165(1):111-24.
411 https://doi.org/10.1016/j.cell.2016.02.011.
412 [14] Wang ZN, Roberts AB, Buffa JA, Levison BS, Zhu WF, Org E, et al. Non-lethal Inhibition of
413 Gut Microbial Trimethylamine Production for the Treatment of Atherosclerosis. Cell
414 2015;163(7):1585-95. https://doi.org/10.1016/j.cell.2015.11.055.
415 [15] Zarate S, Taboada B, Yocupicio-Monroy M. Arias CF. Human Virome. Arch Med Res
416 2017;48(8):701-16. http://doi.org/10.1016/j.arcmed.2018.01.005.
417 [16] Zou SM, Caler L, Colombini-Hatch S, Glynn S. Srinivas P. Research on the human virome:
418 where are we and what is next. Microbiome 2016;4. http://doi.org/10.1186/s40168-016-0177-y.
419 [17] Scarpellini E, Ianiro G, Attili F, Bassanelli C, De Santis A. Gasbarrini A. The human gut
420 microbiota and virome: Potential therapeutic implications. Digest Liver Dis 2015;47(12):1007-12.
421 http://doi.org/10.1016/j.dld.2015.07.008.
422 [18] Liu L, Gong T, Tao WY, Lin BL, Li C, Zheng XS, et al. Commensal viruses maintain

423 intestinal intraepithelial lymphocytes via noncanonical RIG-I signaling. Nat Immunol
424 2019;20(12):1681-1691. https://doi.org/10.1038/s41590-019-0513-z.
425 [19] Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the faecal
426 microbiota of monozygotic twins and their mothers. Nature 2010;466(7304):334-U81.
427 https://doi.org/10.1038/nature09199.
428 [20] Virgin HW. The Virome in Mammalian Physiology and Disease. Cell 2014;157(1):142-50.
429 http://doi.org/10.1016/j.cell.2014.02.032.
430 [21] Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SWL, et al. RNA viral community in
431 human feces: Prevalence of plant pathogenic viruses. Plos Biol 2006;4(1):108-18.
432 https://doi.org/ARTN e310.1371/journal.pbio.0040003.
433 [22] Shi Y. Mu LB. An expanding stage for commensal microbes in host immune regulation. Cell
434 Mol Immunol 2017;14(4):339-48. https://doi.org/10.1038/cmi.2016.64.
435 [23] Handley SA. The virome: a missing component of biological interaction networks in health
436 and disease. Genome Med 2016;8. http://doi.org/10.1186/s13073-016-0287-y.
437 [24] Mukhopadhy I, Segal JP, Carding SR, Hart AL. Hold GL. The gut virome: the 'missing link'
438 between gut bacteria and host immunity? Ther Adv Gastroenter 2019;12.
439 http://doi.org/10.1177/1756284819836620.
440 [25] Foca A, Liberto MC, Quirino A, Marascio N, Zicca E. Pavia G. Gut Inflammation and
441 Immunity: What Is the Role of the Human Gut Virome? Mediat Inflamm 2015.
442 http://doi.org/10.1155/2015/326032.
443 [26] Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, et al. Disease-Specific
444 Alterations in the Enteric Virome in Inflammatory Bowel Disease. Cell 2015;160(3):447-60.

http://doi.org/10.1016/j.cell.2015.01.002.

[27] Manrique P, Dills M. Young MJ. The Human Gut Phage Community and Its Implications for Health and Disease. Viruses-Basel 2017;9(6). http://doi.org/10.3390/v9060141.

[28] Yang JY, Kim MS, Kim E, Cheon JH, Lee YS, Kim Y, et al. Enteric Viruses Ameliorate Gut Inflammation via Toll-like Receptor 3 and Toll-like Receptor 7-Mediated Interferon-beta Production. Immunity 2016;44(4):889-900. https://doi.org/10.1016/j.immuni.2016.03.009.

[29] Broggi A, Tan Y, Granucci F. Zanoni I. IFN-lambda suppresses intestinal inflammation by non-translational regulation of neutrophil function. Nat Immunol 2017;18(10):1084-1093. https://doi.org/10.1038/ni.3821.

[30] Deamer D, Akeson M. Branton D. Three decades of nanopore sequencing. Nat Biotechnol 2016;34(5):518-24. https://doi.org/10.1038/nbt.3423.

[31] Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, et al. Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol 2014;32(8):829-33. https://doi.org/10.1038/nbt.2950.

[32] Schatz MC. Nanopore sequencing meets epigenetics. Nat Methods 2017;14(4):347-8. https://doi.org/10.1038/nmeth.4240.

[33] Low DA, Weyand NJ. Mahan MJ. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. Infect Immun 2001;69(12):7197-204. https://doi.org/Doi 10.1128/Iai.69.12.7197-7204.2001.

[34] Casadesus J. Low D. Epigenetic gene regulation in the bacterial world. Microbiol Mol Biol R 2006;70(3):830-56. https://doi.org/10.1128/Mmbr.00016-06.

[35] Oliveira PH, Touchon M. Rocha EPC. Regulation of genetic flux between bacteria by

467 restriction-modification systems. P Natl Acad Sci USA 2016;113(20):5658-63.
468 https://doi.org/10.1073/pnas.1603257113.
469 [36] Cohen NR, Ross CA, Jain S, Shapiro RS, Gutierrez A, Belenky P, et al. A role for the
470 bacterial GATC methylome in antibiotic stress survival. Nat Genet 2016;48(5):581-6.
471 https://doi.org/10.1038/ng.3530.
472 [37] Oliveira PH, Ribis JW, Garrett EM, Trzilova D, Kim A, Sekulovic O, et al. Epigenomic
473 characterization of Clostridioides difficile finds a conserved DNA methyltransferase that mediates
474 sporulation and pathogenesis. Nat Microbiol 2019; https://doi.org/10.1038/s41564-019-0613-4.
475 [38] Xue MG, Zhao BS, Zhang ZJ, Lu MJ, Harder O, Chen P, et al. Viral N-6-methyladenosine
476 upregulates replication and pathogenesis of human respiratory syncytial virus. Nat Commun
477 2019;10 https://doi.org/ARTN 459510.1038/s41467-019-12504-y.
478 [39] Thurber RV, Haynes M, Breitbart M, Wegley L. Rohwer F. Laboratory procedures to generate
479 viral metagenomes. Nat Protoc 2009;4(4):470-83. https://doi.org/10.1038/nprot.2009.10.
480 [40] Ge XY, Li Y, Yang XL, Zhang HJ, Zhou P, Zhang YZ, et al. Metagenomic Analysis of Viruses
481 from Bat Fecal Samples Reveals Many Novel Viruses in Insectivorous Bats in China. J Virol
482 2012;86(8):4620-30. https://doi.org/10.1128/Jvi.06671-11.
483 [41] Froussard P. rPCR: a powerful tool for random amplification of whole RNA sequences.
484 Genome Research 1993;2185-90. https://doi.org/10.1101/gr.2.3.185.
485 [42] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH. Phillippy AM. Canu: scalable and
486 accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res
487 2017;27(5):722-36. https://doi.org/10.1101/gr.215087.116.
488 [43] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics

2018;34(18):3094-100. https://doi.org/10.1093/bioinformatics/bty191.

[44] Auslander N, Gussow AB, Benler S, Wolf YI. Koonin EV. Seeker: Alignment-free identification of bacteriophage genomes by deep learning. bioRxiv 2020;2020.04.04.025783. https://doi.org/10.1101/2020.04.04.025783.

[45] Ecale Zhou CL, Malfatti S, Kimbrel J, Philipson C, McNair K, Hamilton T, et al. multiPhATE: bioinformatics pipeline for functional annotation of phage isolates. Bioinformatics (Oxford, England) 2019;35(21):4402-4. https://doi.org/10.1093/bioinformatics/btz258.

[46] McNair K, Zhou C, Dinsdale EA, Souza B. Edwards RA. PHANOTATE: a novel approach to gene identification in phage genomes. Bioinformatics (Oxford, England) 2019;35(22):4537-42. https://doi.org/10.1093/bioinformatics/btz265.

[47] Grazziotin AL, Koonin EV. Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic acids research 2017;45(D1):D491-d8. https://doi.org/10.1093/nar/gkw975.

[48] Johnson LS, Eddy SR. Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 2010;11(1):431. https://doi.org/10.1186/1471-2105-11-431.

[49] Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ 2018;6e5030. https://doi.org/10.7717/peerj.5030.

[50] Langmead B, Wilks C, Antonescu V. Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics (Oxford, England) 2019;35(3):421-32. https://doi.org/10.1093/bioinformatics/bty648.

[51] Marcus Stoiber JQ, Rob Egan, Ji Eun Lee, Susan Celniker, Robert K. Neely, Nicholas Loman,

Len A Pennacchio, James Brown. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. bioRxiv 2017; https://doi.org/10.1101/094672.

[52] Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A. Mesirov JP. Variant Review with the Integrative Genomics Viewer. Cancer Res 2017;77(21):e31-e4. https://doi.org/10.1158/0008-5472.CAN-17-0337.

[53] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 2009;37(Web Server issue):W202-8. https://doi.org/10.1093/nar/gkp335.

[54] Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, et al. The human gut virome is highly diverse, stable and individual-specific. Cell Host & Microbe 2019; https://doi.org/10.1101/657528.

[55] Moreno-Gallego JL, Chou S-P, Di Rienzi SC, Goodrich JK, Spector TD, Bell JT, et al. Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins. Cell Host & Microbe 2019;25(2):261-72.e5. https://doi.org/10.1016/j.chom.2019.01.019.

[56] Garmaeva S, Sinha T, Kurilshikov A, Fu J, Wijmenga C. Zhernakova A. Studying the gut virome in the metagenomic era: challenges and perspectives. BMC Biol 2019;17(1):84. https://doi.org/10.1186/s12915-019-0704-y.

[57] Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, et al. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 2011;21(10):1616-25. https://doi.org/10.1101/gr.122705.111.

[58] Gregory AC ZO, Howell A, Bolduc B. Sullivan MB. The human gut virome database. bioRxiv 2019; https://doi.org/10.1101/655910.

[59] Gardy JL. Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 2018;19(1):9-20. https://doi.org/10.1038/nrg.2017.88.

[60] Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, et al. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat Commun 2019;10 https://doi.org/10.1038/s41467-019-08734-9.

[61] Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol 2019;37(7):783-792. https://doi.org/10.1038/s41587-019-0156-5.

[62] Courtney DG, Kennedy EM, Dumm RE, Bogerd HP, Tsai K, Heaton NS, et al. Epitranscriptomic Enhancement of Influenza A Virus Gene Expression and Replication. Cell Host Microbe 2017;22(3):377-386. https://doi.org/10.1016/j.chom.2017.08.004.

[63] Beaulaurier J, Schadt EE. Fang G. Deciphering bacterial epigenomes using modern sequencing technologies. Nat Rev Genet 2019;20(3):157-72. https://doi.org/10.1038/s41576-018-0081-3.

[64] Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing (vol 30, pg 1232, 2012). Nat Biotechnol 2013;31(6):566-9. https://doi.org/10.1038/nbt0613-565f.

551 *Figure Legends:*

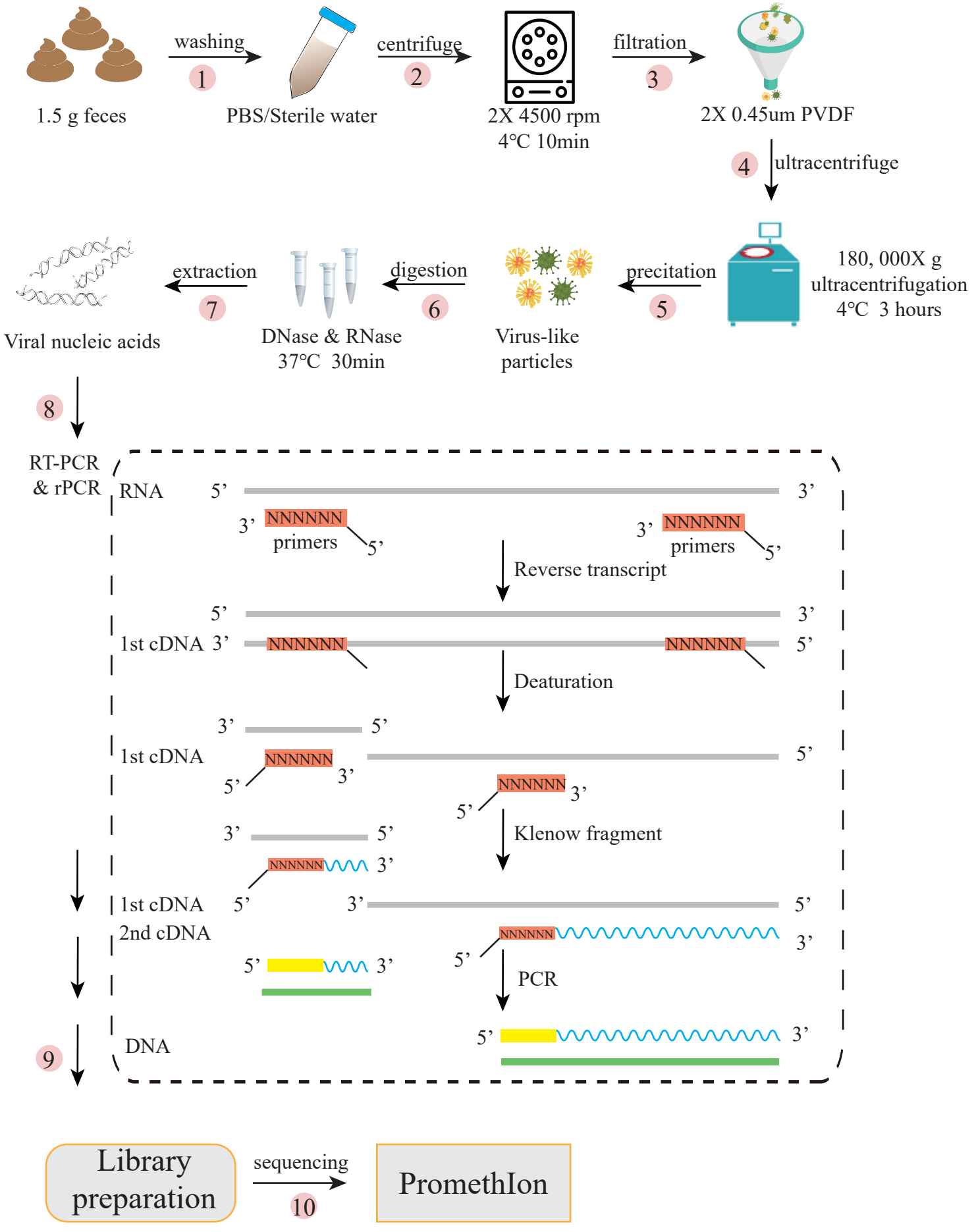
552 **Figure 1. An integrated novel workflow for enrichment of virus-like particles**
553 **(VLPs), extraction of nucleic acids and ONT sequencing.** The complete workflow
554 consists of four fragments: (1) Washing and filtration of faecal samples using
555 PBS/Sterile and PVDF membrane including step 1-3; (2) Precipitation of VLPs
556 including step 4-5; (3) Extraction, amplification and purification of viral nucleic acids
557 including step 6-8; (4) Construction of library and ONT sequencing including step
558 9-10.

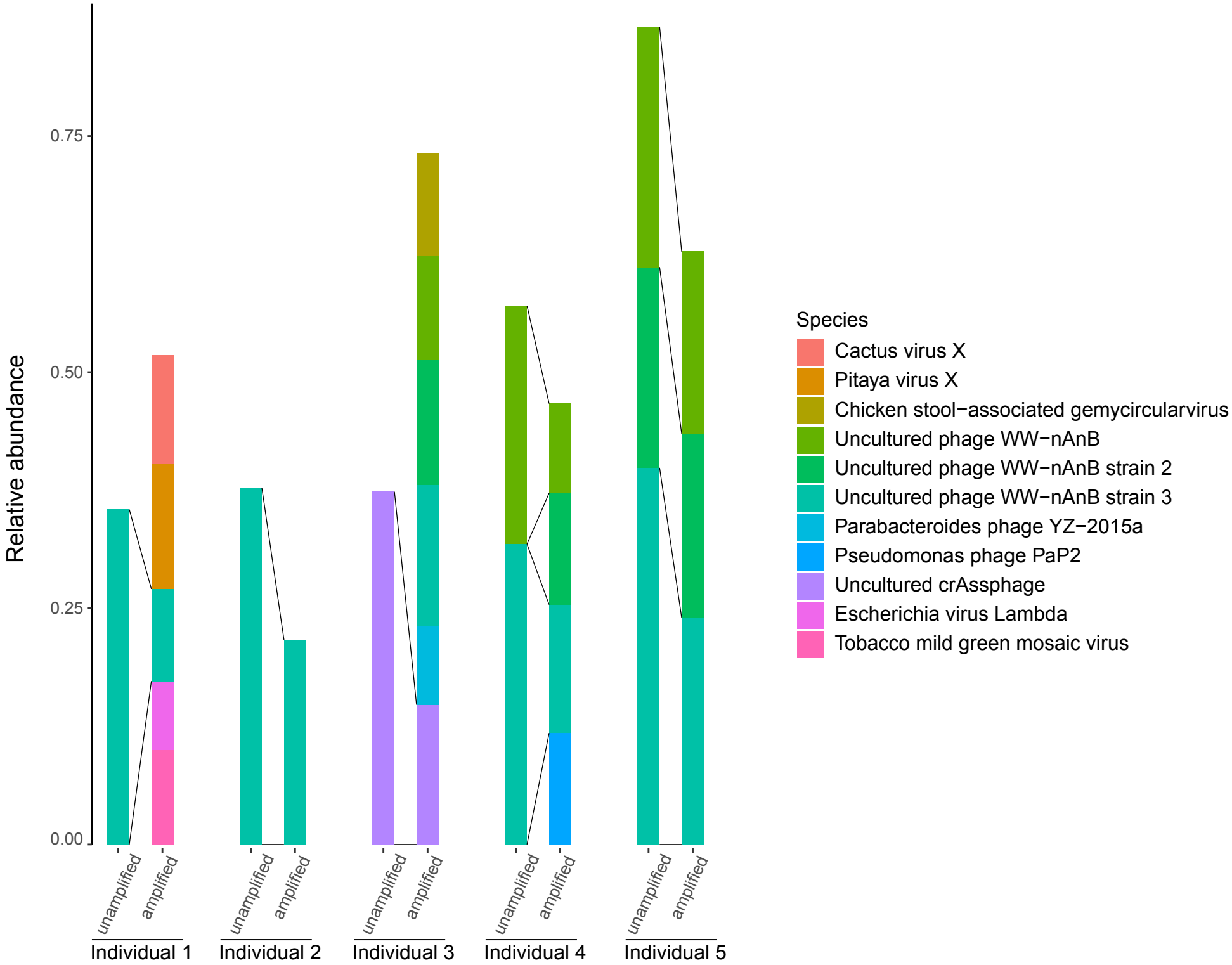
559 **Figure 2. Composition and relative abundance of viruses in each individual.**
560 Different color of the bar represents different viral species or strains. The line between
561 unamplified and amplified group in each individual represents common virus species
562 between two groups.

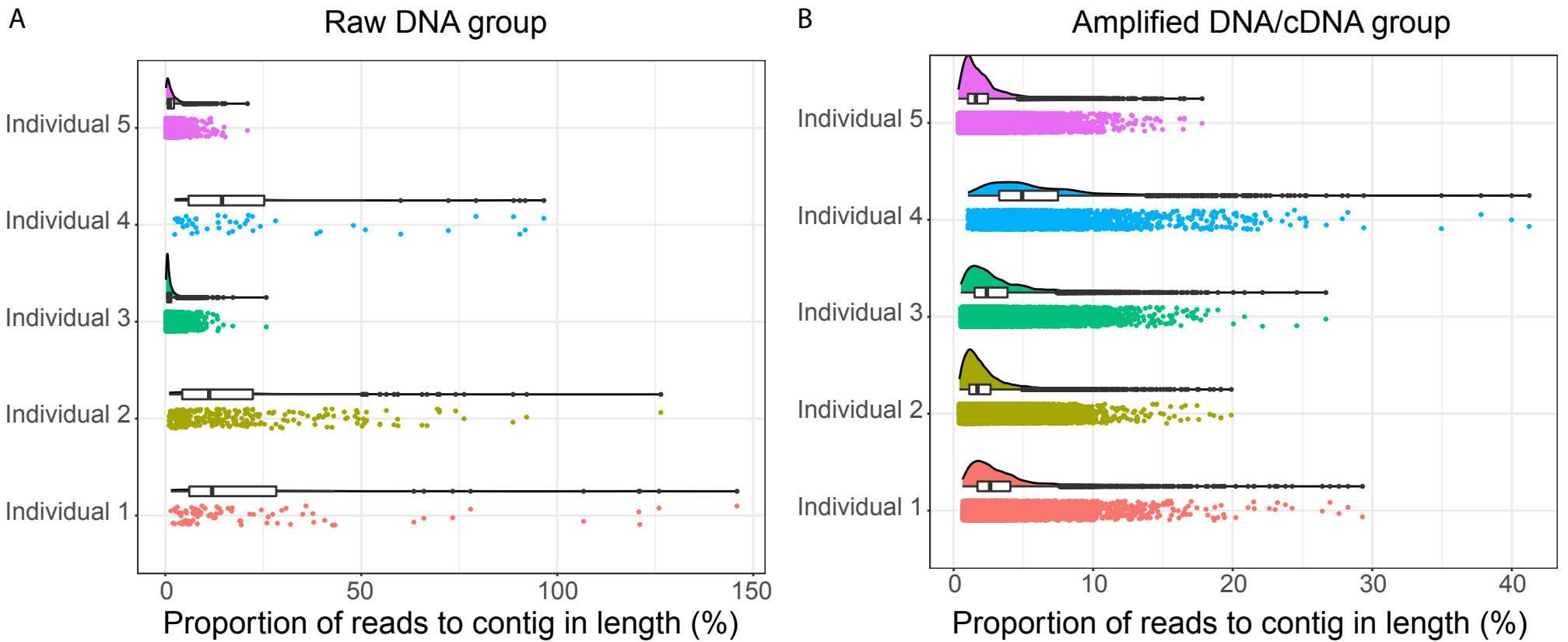
563 **Figure 3. Proportion of raw reads to contigs in length.** The proportion of raw reads
564 to contigs in length is shown by combination of scatter diagram, boxplot and violin
565 chart. Different color represents five different individuals. (A) raw DNA group, (B)
566 amplified DNA/cDNA group.

567 **Figure 4. Different methylation sites identification and motif recognition of viral**
568 **contig.** (A) the distribution of methylation sites (5mC and 6mA) in the
569 contig00000015; (B) 5mC motif; (C) 6mA motif.

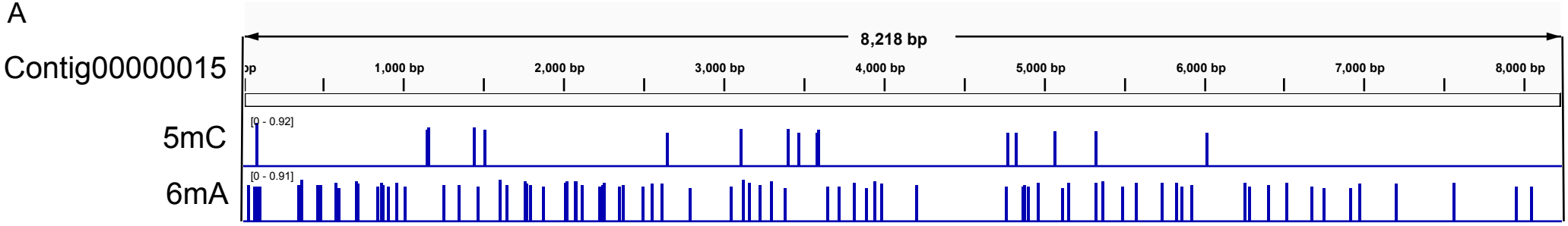
570


1.5 g feces
washing
1
PBS/Sterile water
centrifuge
2
2X 4500 rpm
4°C 10min
filtration
3
2X 0.45um PVDF
4
ultracentrifuge
180, 000X g
ultracentrifugation
4°C 3 hours
precitation
5
Virus-like
particles
digestion
6
DNase & RNase
37°C 30min
extraction
7
Viral nucleic acids
8
RT-PCR
& rPCR
RNA
5'
3'
NNNNNN
primers
Reverse transcript
1st cDNA
Deaturation
Klenow fragment
2nd cDNA
PCR
DNA
9
Library
preparation
sequencing
10
PromethIon


Relative abundance
0.75
0.50
0.25
0.00
unamplified
amplified
Individual 1
Individual 2
Individual 3
Individual 4
Individual 5
Species
Cactus virus X
Pitaya virus X
Chicken stool−associated gemycircularvirus
Uncultured phage WW−nAnB
Uncultured phage WW−nAnB strain 2
Uncultured phage WW−nAnB strain 3
Parabacteroides phage YZ−2015a
Pseudomonas phage PaP2
Uncultured crAssphage
Escherichia virus Lambda
Tobacco mild green mosaic virus

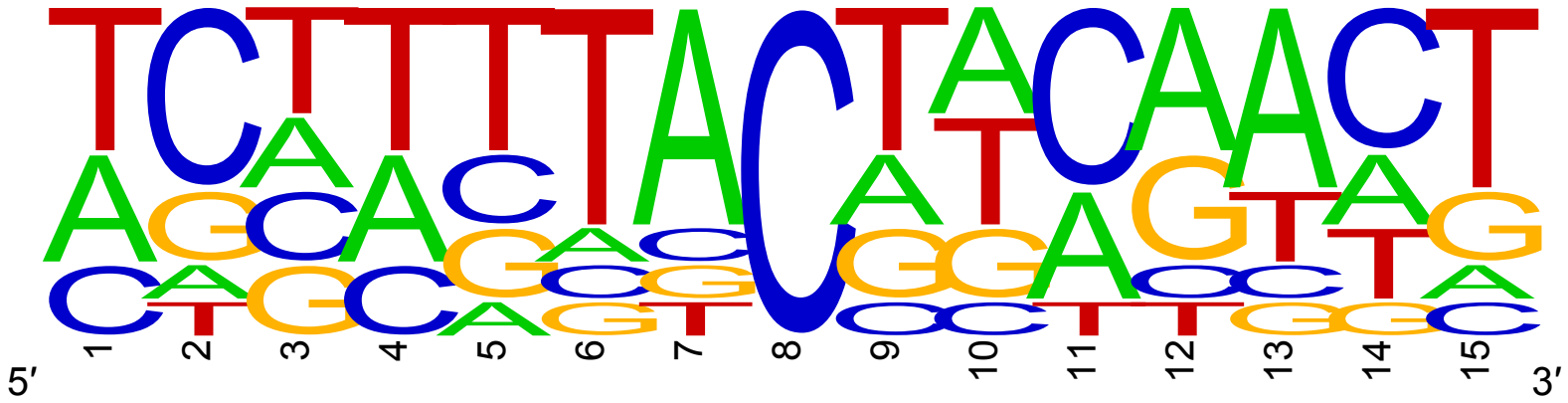

A
Raw DNA group
Individual 5
Individual 4
Individual 3
Individual 2
Individual 1
0
50
100
150
Proportion of reads to contig in length (%)
B
Amplified DNA/cDNA group
Individual 5
Individual 4
Individual 3
Individual 2
Individual 1
0
10
20
30
40
Proportion of reads to contig in length (%)

A



B

5mC motif



C

6mA motif