Chinese Medical Journal®

Uncovering novel human gut virome using ultra-deep metagenomic sequencing

Harry Cheuk Hay Lau, Jun Yu

Institute of Digestive Disease and Department of Medicine and Therapeutics, State Key Laboratory of Digestive Disease, Li Ka Shing Institute of Health Sciences, the Chinese University of Hong Kong, Hong Kong SAR, China.

Human gastrointestinal tract harbors trillions of commensal microbes including bacteria, viruses, and fungi to form an ecological community, commonly known as the gut microbiome. Being the second most dominant taxonomic kingdom in the gut (5.8% of total microbiome DNA), enteric viruses are increasingly acknowledged for their roles in shaping the overall composition of the gut microbiome and maintaining human health.[1,2] Recent advancements in metagenomic sequencing have brought breakthrough in deciphering the features and functions of the gut virome.[3] Notably, there are several factors restricting the application of this technology in virome study. For instance, the amount of isolated viral DNA from human samples is usually insufficient for metagenomic sequencing. To solve this issue, amplification techniques such as polymerase chain reaction and multiple displacement amplification are widely used to increase the amount of extracted viral DNA prior to sequencing, which in turn could introduce amplification bias, chimeric reads, and random mutations.[4] Meanwhile, given that most studies analyzed metagenomic reads at low depth (from 0.0003 GB to 7.7 GB),[5,6] there is currently lack of high quality assembled viral genomes, as well as failure in identifying viruses with low abundance in samples. In addition, Illumina sequencing that generates short reads, has been frequently used. However, short sequencing reads were reported to be inadequate for accurate assembly of viral genomes as they contain hypervariable sequences and repeating regions.[7]

All these issues have greatly obstructed virome investigations, as evidenced by the high portion (75–95%) of viral metagenomic reads from the human gut being unclassified.[8,9] A plausible solution is to conduct deep metagenomic sequencing as well as long-read sequencing (Pacific Biosciences [PacBio] or Oxford Nanopore), which could be an important complement to short-read sequencing.[1] The team in the Chinese University of Hong Kong previously reported the alteration of gut virome in patients with colorectal cancer (CRC),[10] while it is also of importance to investigate the complete profile of the gut virome in nondisease conditions.

To this end, the study by Zhao *et al*[11], published in June 2022 online issue of *Gastroenterology*, has performed ultra-deep amplification-free metagenomic sequencing on fecal samples of healthy individuals. First and foremost, sufficient amount of viral DNA is necessary for sequencing at such depth; hence, Zhao *et al* adopted and modified extraction protocols from multiple previous studies. For example, multi-steps filtration and lysozymes were used to hydrolyze bacterial cell walls and eliminate as many bacteria as possible.[12] To obtain viral DNA, viral capsids need to be disrupted and lysed which is commonly done by the treatment of phenol/chloroform/isoamyl alcohol.[13] However, the extraction yield of this method has been unsatisfactory due to the low efficiency in DNA lysis and precipitation. In comparison, previous studies reported that column-based DNA purification has greater recovery efficiency than the conventional lysis method.[14] Zhao *et al* therefore applied this column-based approach, which has a single lysis step with more lenient experimental conditions, and succeeded to extract an increased amount of viral DNA compared to previous protocols. Together, these modifications including differential lysis, depletion of cell-free DNA, and multiple-time filtration were capable of improving the extraction yield and depleting nonviral DNA in human fecal samples, thereby facilitating amplification-free library preparation prior to metagenomic sequencing.

To comprehensively evaluate the gut virome, Zhao *et al*[11] parallelly performed Illumina and PacBio High-Fidelity

Access this article online

Quick Response Code:

Website: www.cmj.org

DOI: 10.1097/CM9.0000000000002382

Correspondence to: Prof. Jun Yu, Institute of Digestive Disease and Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China
E-Mail: junyu@cuhk.edu.hk

Chinese Medical Journal 2022;135(20)

Received: 22-06-2022; **Online:** 07-11-2022 **Edited by:** Yuanyuan Ji

sequencing on the extracted viral DNA. While Illumina short-read sequencing has been used extensively, PacBio sequencing is a more recent technology that can generate long reads. In general, long reads can improve contig generation from genome regions that have high similarity or complexity, as well as prevent formation of chimeric contigs.[15] Once sequencing reads are generated, they need to be assembled as contigs for constructing genomes. As two different sequencing platforms were used, Zhao *et al* developed a strategy for *de novo* assembly to accurately identify viral genomes from raw long and short reads. Typically, a single assembler is employed for genome assembly.[16] To take advantage of each assembler, Zhao *et al*[11] used a combination of six assemblers to achieve better recovery of viral genomes. For instance, assemblers for short reads (e.g., megagit and metaSPAdes) are well established and proved to have great performance,[17] whereas assemblers for long reads are much more diverse probably due to the recent bloom of their development. Filtering is another critical process in distinguishing viral genomes from complex contigs. A series of filters was therefore established to recognize clean viral genomes with high precision and remove any bacterial or other nonviral sequences, eventually obtaining 1178 complete virus contigs. Hence, as demonstrated in the study by Zhao *et al*, a rigorous approach to processing raw reads is critical for analyzing viral genomes.

Among all assembled viral genomes, 1058 of them were newly identified in the study by Zhao *et al*[11] with subsequent validation in three published metagenomes and fecal samples from additional healthy individuals. Long reads are known to benefit the assembly of microbial genomes with long sequences. Indeed, 13 viral genomes with lengths greater than the longest phage sequence found in human samples (393 kb) were identified.[18] Particularly, two genomes were assembled (HugePhage1 and Hugephage2), which are even longer than the current reported largest phage sequence (735 kb).[19] Given that many viruses are newly identified, it is of interest to depict the evolutionary history of these viruses. Zhao *et al*[11] therefore constructed a phylogenetic tree to reveal uncharacterized viruses based on the hidden Markov model (HMM) alignment scores of four conserved viral proteins (major capsid protein [MCP], primase, terminase, and portal protein). Viruses with similar HMM alignment scores were clustered together to form nine clades (Hkyuvirus 1–9), and each clade has specific HMM alignment scores. For example, although MCP is the major component of viral capsids, it is absent in viruses from Hkyuvirus 4 and Hkyuvirus 7 clades. To date, a large portion of human enteric viruses is still undefined. In the current study by Zhao *et al*, the authors developed a strategy to assort viruses based on conserved viral proteins, of which the establishment of these protein-based clades could benefit the characterization of enteric viruses with unknown taxonomy.

Zhao *et al*[11] further incorporated their viral genomes into the NCBI RefSeq database, and this significantly improved the mapping percentage of published metagenomic datasets by up to 18.1 times. RefSeq has been heavily used for the alignment of viral metagenomic reads. However, it only includes a limited number of well-characterized viruses, while most gut commensal viruses remain unrecognized. The substantial insufficiency of viral genomes included in RefSeq was suggested to be the major cause of unmappable genomes, as some studies reported that $\leq$98% of sequencing reads could not be assigned to viral taxa when using RefSeq.[1] Similarly, Zhao *et al* found that only 40–45% of their raw reads were recognized by RefSeq. Notably, incorporating viral genomes identified by Zhao *et al* into RefSeq markedly increased the read mapping ratio to over 80%. The assembled viral genomes could therefore expand the reference genomes in RefSeq, enabling future studies to have a more complete mapping of sequencing reads.

Based on the identified viral genomes, Zhao *et al*[11] established a biomarker panel comprising of 14 newly identified viruses, and this panel demonstrated great performance in discriminating patients with CRC from healthy subjects. Among 11 enriched viruses, 8 belong to Podoviridae, Myoviridae, and Siphoviridae which are frequently reported as the most abundant viral families in the human gut.[6] The enrichment of *Lactobacillus* prophage phiadh was also observed in CRC patients. As probiotic *Lactobacillus* are known to be depleted in CRC, increased abundance of *Lactobacillus*-infecting phage/prophage may contribute to *Lactobacillus* depletion, indicating a negative correlation between gut bacteria and phage in CRC. In general, Zhao *et al* showed that even uncharacterized viruses could be potential biomarkers; thus, it is worth investigating whether incorporating enteric viruses into established biomarker panel could further improve diagnostic performance. Nevertheless, complete characterization and functional investigation are necessary before utilizing these newly identified viral genomes in clinical practice.

In conclusion, using the modified extraction protocol, Zhao *et al* obtained sufficient viral DNA that largely represent the human gut virome. Through performing ultra-deep metagenomic sequencing with *de novo* assembly and multistage filtering, Zhao *et al* succeeded to accurately classify the viral genomes. This enabled the identification of 1058 novel viral genomes including 13 long viral sequences, as well as improved read mapping ratio of published metagenomic datasets. Altogether, these findings will contribute to clinical diagnosis, current viral reference genome database and future in-depth investigation of the human gut virome. In addition, it is noteworthy to highlight that only double-stranded DNA viruses were considered in the study by Zhao *et al*, as the library preparation strategy used in amplification-free Illumina and PacBio sequencing could only identify double-stranded DNA.[20,21] This sequencing approach is also incapable of discriminating foodborne viruses from gut commensal viruses, as the majority of foodborne viruses is not double-stranded DNA viruses.[22] Future study with deep metagenomic sequencing is needed to investigate the roles of other virus types, including single-stranded DNA and RNA viruses in human gut virome.

Conflicts of interest

None.

References

1. Shkoporov AN, Hill C. Bacteriophages of the human gut: the "known unknown" of the microbiome. Cell Host Microbe 2019;25:195–209. doi: 10.1016/j.chom.2019.01.017.
2. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, *et al*. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe 2019;26:764–778. e5. doi: 10.1016/j.chom.2019.10.009.
3. Liang G, Zhao C, Zhang H, Mattei L, Sherrill-Mix S, Bittinger K, *et al*. The stepwise assembly of the neonatal virome is modulated by breastfeeding. Nature 2020;581:470–474. doi: 10.1038/s41586-020-2192-1.
4. 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. doi: 10.1186/s40168-018-0507-3.
5. Pérez-Brocal V, García-López R, Vázquez-Castellanos JF, Nos P, Beltrán B, Latorre A, *et al*. Study of the viral and microbial communities associated with Crohn's disease: A metagenomic approach. Clin Transl Gastroenterol 2013;4:e36. doi: 10.1038/ctg.2013.9.
6. Zuo T, Lu XJ, Zhang Y, Cheung CP, Lam S, Zhang F, *et al*. Gut mucosal virome alterations in ulcerative colitis. Gut 2019;68:1169–1179. doi: 10.1136/gutjnl-2018-318131.
7. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, *et al*. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ 2019;7:e6800. doi: 10.7717/peerj.6800.
8. Aggarwala V, Liang G, Bushman FD. Viral communities of the human gut: metagenomic analysis of composition and dynamics. Mob DNA 2017;8:12. doi: 10.1186/s13100-017-0095-y.
9. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife 2015;4:e08490. doi: 10.7554/eLife.08490.
10. Nakatsu G, Zhou H, Wu WKK, Wong SH, Coker OO, Dai Z, *et al*. Alterations in enteric virome are associated with colorectal cancer and survival outcomes. Gastroenterology 2018;155:529–541. e5. doi: 10.1053/j.gastro.2018.04.018.
11. Zhao L, Shi Y, Cheuk-Hay Lau H, Liu W, Luo G, Wang G, *et al*. Uncovering 1,058 novel human enteric DNA viruses through deep long-read third-generation sequencing and their clinical impact. Gastroenterology 2022;163:699–711. doi: 10.1053/j.gastro.2022.05.048.
12. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, *et al*. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 2010;466:334–338. doi: 10.1038/nature09199.
13. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, *et al*. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. Cell Host Microbe 2018;24:653–664. e6. doi: 10.1016/j.chom.2018.10.002.
14. Dairawan M, Shetty PJ. The evolution of DNA extraction methods. Am J Biomed Sci Res 2020;8:39–46. doi: 10.34297/AJBSR.2020.08.001234.
15. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ 2017;5:e3817. doi: 10.7717/peerj.3817.
16. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, *et al*. Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat Microbiol 2019;4:693–700. doi: 10.1038/s41564-018-0338-9.
17. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe 2020;28:724–740. e8. doi: 10.1016/j.chom.2020.08.003.
18. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, *et al*. Lytic to temperate switching of viral communities. Nature 2016;531:466–470. doi: 10.1038/nature17193.
19. Al-Shayeb B, Sachdeva R, Chen LX, Ward F, Munk P, Devoto A, *et al*. Clades of huge phages from across Earth's ecosystems. Nature 2020;578:425–431. doi: 10.1038/s41586-020-2007-4.
20. Pacbio. Pacbio SMRTbell Express Template Prep Kit 2.0. Available from: https://www.pacb.com/wp-content/uploads/SMRTbell-Express-TPK-2.0-Overview-Large-Insert-gDNA-Library-Prep-Customer-Training.pdf. Last accessed on June 22, 2022.
21. TruSeq D. TruSeq® DNA Library Prep Kits. Available from: https://www.illumina.com/documents/products/datasheets/datasheet_truseq_dna_sample_prep_kits.pdf. Last accessed on June 22, 2022.
22. D'Souza DH, Joshi SS. Caballero B, Finglas P, Toldrá F. Foodborne viruses of human health concern. Encyclopedia of Food and Health 2016; Academic Press, 87–93. doi: 10.1016/B978-0-12-384947-2.00727-3.

How to cite this article: Lau HCH, Yu J. Uncovering novel human gut virome using ultra-deep metagenomic sequencing. Chin Med J 2022;135:2395–2397. doi: 10.1097/CM9.0000000000002382