Understanding the Human Virome in Time and Space: Methods Development

Jiayi Duan[1], Andrew Marques[1], Matt Hogenauer[1], Young Hwang[1], Julia Malnak[1], Scott Daniel[3], Ronald G. Collman[2], Frederic D. Bushman[1]

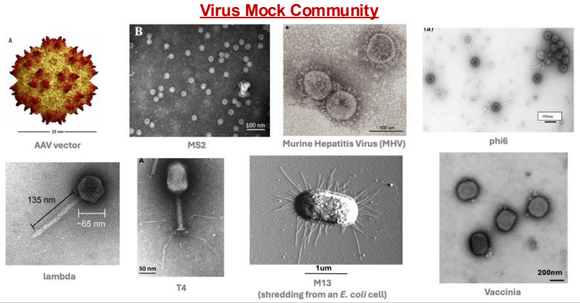
[1]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[3]Division of Gastroenterology, Hepatology, and Nutrition, Children's Hospital of Philadelphia, Pennsylvania, USA

Introduction

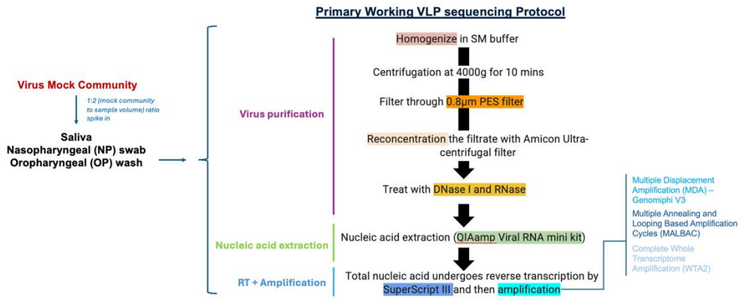
The human virome is vast and diverse, with approximately 10^{13} virus particles per individual. Virus populations are heterogenous and vary across different body sites. Despite being a major player in the human body, much remains unknown about these viruses. This is largely because of the current gaps in virome studies, including the lack of a universal marker gene for viruses, absence of an optimized protocol for virus-like particles isolation applicable to all viruses, difficulty in efficiently annotating and comprehensively cataloguing discovered viruses, and unknown or previously neglected factors like DNA modifications potentially affecting study results.

In response to the NIH Human Virome Program's initiative to identify and catalog the healthy human virome, we are optimizing methods to capture viral diversity across human sample types. We prepared a mock viral community consisting of eight known viruses, including *E. coli* phages T4, lambda, M13, and MS2, *Pseudomonas* virus phi6, an adenovirus-associated vector (AAV), murine hepatitis virus (MHV), and vaccinia virus (VV). Using the mock community, we examined different aspects of common methods used by current virome studies for their effect on a range of viruses varying in sizes and structures.

Virus Mock Community



Method



Our main approach in examining the current virus enrichment protocol (viromeprep) is spiking the virus mock community of known quantities of each virus into various sample types, including saliva, OP wash, and NP swab, and then passing the samples through viromeprep for Illumina sequencing.

To investigate whether PCR amplification skews sequencing result, we sequenced post-reverse transcription (RT) samples that were unamplified or amplified with 1) Genomiphi, 2) MALBAC, or 3) one of the three different PCR cycle thresholds (7, 12, and 17) of WTA2.

To investigate the potential effect of DNA modifications on sequencing detection of viruses, we mixed known quantities of phage lambda with either phage T4 with glucosyl hydroxymethylated cytosine (T4ghmC), T4 with hydroxymethylated cytosine (T4hmC), or T4 with unmodified cytosine (T4C), and then extracted nucleic acids followed by RT for Illumina Nextseq sequencing.

To determine the effect of nuclease treatment on removal of free nucleic acids and recovery of viruses, we treated the mock community with a titration (none, 0.1X, 1X, 10X) of DNase I and RNase, where 1X consists of 20U DNase I and 0.5ug RNase (Roche). We then extracted nucleic acids followed by RT for Illumina Nextseq sequencing.

To determine the optimal storage condition for viruses in human samples, we added the mock community into NP swabs resuspended in virus transport media (VTM) and into saliva spiked with an oral cocktail comprised of 10% glycerol, 2ug/mL aprotinin, 1mM PMSF, 10mM Mg^{2+}, and 10mM DTT. Samples were stored at -80°C overnight and, after one freeze-thaw, either underwent Trizol extraction with/without an initial centrifugation or enriched for VLPs, and finally extracted for nucleic acids followed by RT for Illumina Nextseq sequencing.

Post sequencing, reads were mapped to the reference genome of each virus of the mock community and calculated for reads per kilobase per million reads (RPKM). Relative abundance of each virus based on RPKM was then compared with that based on genome copies determined by performing qPCR on each sample. To evaluate the enrichment of viruses, percentage of viral reads (% in total reads) is calculated by dividing the sum of reads mapped to reference genomes of all viruses by the total number of reads for the sample.

Result

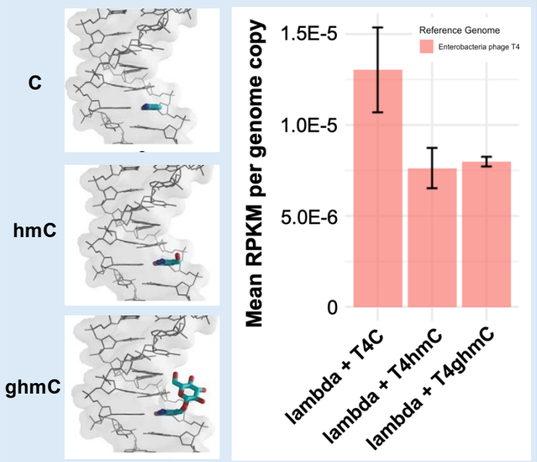


Figure 1. Left: DNA modifications in phage T4 showing C-containing DNA (top), hmC-containing DNA (middle), and ghmC-containing DNA (bottom). Right: mean RPKM of T4 per genome copy based on the qPCR result. Both T4ghmC and T4C showed lower RPKM (determined by sequencing) per genome copy (determined by qPCR) compared to that of T4C. Nevertheless, the difference is moderate, and all T4 variants can be detected by sequencing despite the base modification.

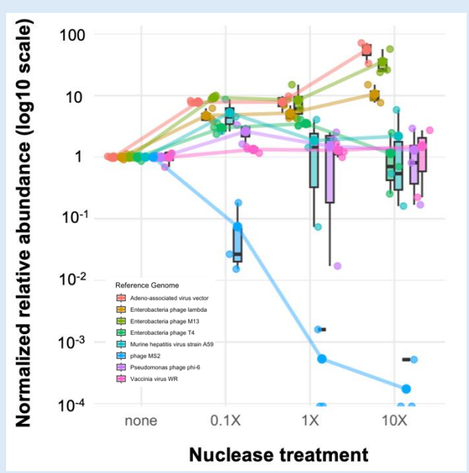


Figure 2. Normalized relative abundance of each virus in the mock community after different nuclease treatments. The relative abundance of phage MS2 drops dramatically as the concentration of nucleases increases, accompanied by a corresponding increase of other viruses in the mock community, specifically T4 at 0.1X and 1X nuclease treatment, and AAV, lambda and M13 at 10X nuclease treatment.

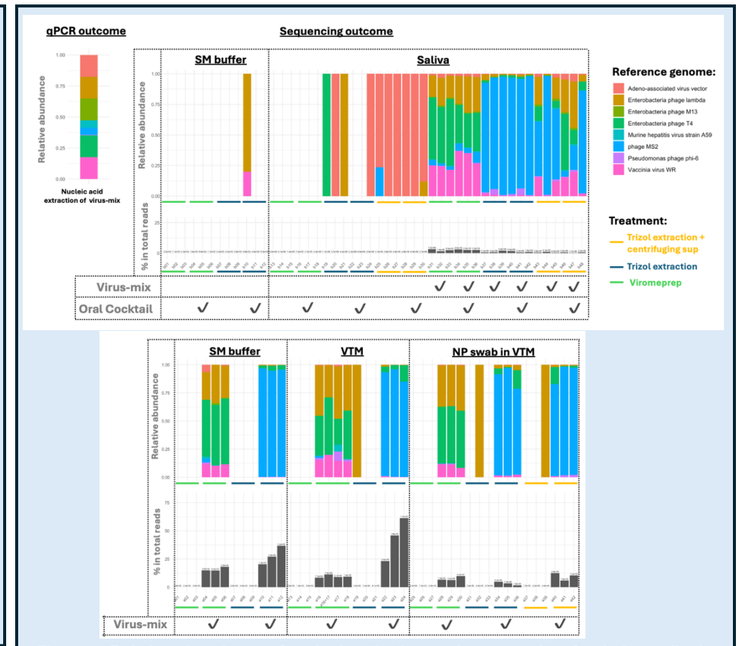


Figure 4. Relative abundance of each virus in the mock community after being spiked into different sample types in different storage buffers and undergoing viromeprep or Trizol extraction with or without initial centrifugation. Numbers on the top of each bar in % viral reads represent the total number of reads in the sample. Viruses survived well in VTM and oral cocktail, generally reflecting the virus community composition determined by qPCR. Samples undergoing Trizol extraction greatly differs in virus community composition from those undergoing viromeprep, specifically in MS2 recovery. No detectable contaminating viruses introduced in VTM or oral cocktail [data not shown].

Figure 3. Relative abundance of each virus in the mock community after being spiked into different sample types and undergoing viromeprep or direct nucleic acid extraction. Genomiphi showed great amplification efficiency, but introduced huge biases, greatly favoring small circular single-stranded DNA viruses like M13. In contrast, MALBAC and WTA2 recovered viruses with a relative abundance more similar to the input mock community. WTA2 also preferentially amplifies for RNA viruses present in low abundance, shown by increased recovery of phi6 and MS2 in WTA2-amplified samples. Vaccinia passed through the virus enrichment protocol relatively well despite its large size particle size. Samples having undergone viromeprep consistently shows a moderately higher percentage of viral reads compared to the samples that have undergone only nucleic acid extraction. Novel virus recovery was more robust from unspiked saliva (13.69%) compared to the recovery from OP wash (0.24%) and NP swabs (3.82%) [data not shown].

Key Takeaways

- Virus enrichment methods can significantly influence virome profiling outcomes, with strong biases introduced by MDA-based techniques such as Genomiphi.
- Current viromeprep protocol enriches for viruses compared to direct nucleic acid extraction.
- Saliva samples yielded the most consistent virus recovery, other sample types presented challenges.
- DNA modifications such as methylation, hydroxy-methylation, and glucosyl hydroxymethylation did not significantly impede viral detection by Illumina sequencing; the effect of other viral DNA modifications remain to be explored.
- VTM and oral cocktail showed no negative impact on recovery of virions for short-term storage, but whether they protect the virions in long-term remains to be determined.
- Nuclease treatment can lead to significant loss of RNA phage like MS2, potentially due to the structure of the virus.

Future Directions

- More spike-in experiments using different VLP enrichment methods for samples like stool, buccal swab and broncho alveolar lavage.
- Exploring the unknown viruses in different sample types and the effect of different virus enrichment strategies on them.

References

Liang, G. et al, 2020. The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature*.

Liang, G., Bushman, F.D., 2021. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology*.

Conceicao-Neto, N. et al, 2015. Modular approach to customize sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Nature Scientific Reports*.

Bryson, A.L., et al. 2015. Covalent Modification of Bacteriophage T4 DNA Inhibits CRISPR-Cas9. *mBio*.