Title of research project:
Comprehensive detection of viruses by metagenomic analysis of sewage and their host estimation by machine learning

Names of project members (including affiliation):
Low Kyle Young, Universiti Malaya

Purposes:

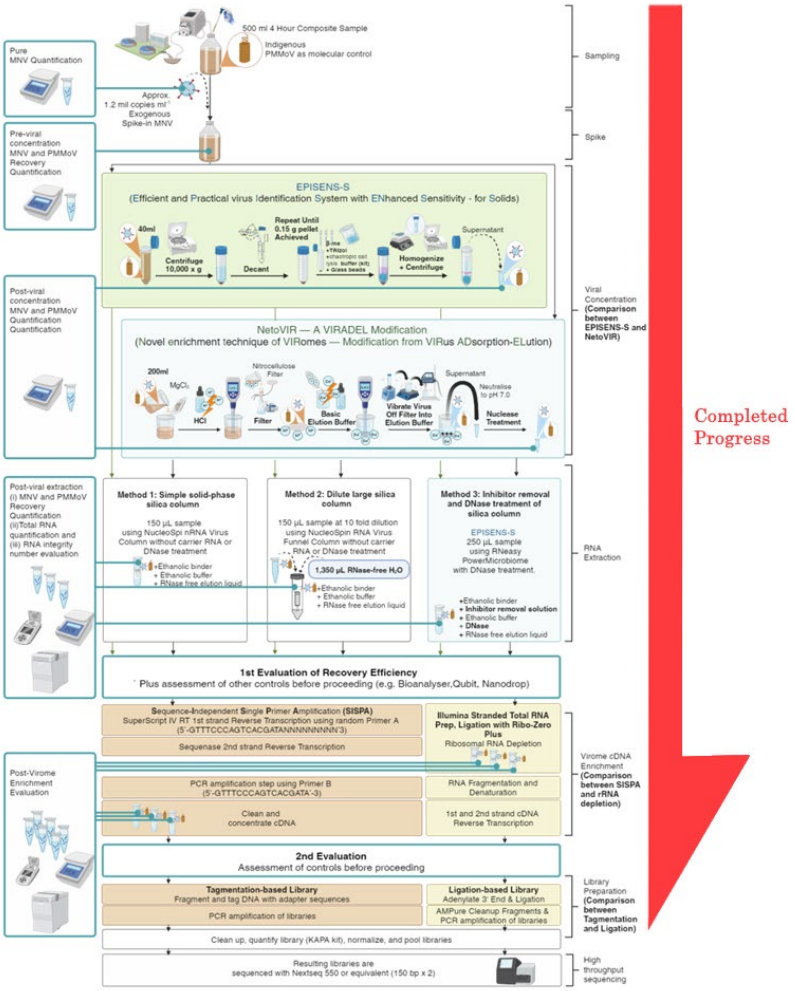
1. Analyze diversity optimised for RNA viromes in wastewater treatment plants from varied human demographics and dietary habits across different regions.
2. Develop supervised and semi-supervised machine learning models to predict virus-host associations for known and novel viral sequences identified in wastewater environments.
3. Compare viral diversity across unfiltered vMAGs, filtered metavirome, and integrated vMAG-metavirome samples.

Methods:

- Sample Collection:
    - 54 virome-concentrated wastewater samples were collected from wastewater treatment centre, Matsuyama, Japan

- Sample Processing:
    - Optimized RNA viral extraction using commercial silica-based membrane column kit for viral metagenome construction. Six combinatorial treatment types with varying levels of virus concentration and inhibitor removal efficiencies were explored: VIRADEL 1,2,3 and EPISENSE-S 1, 2, and 3.
    - Conducted cDNA synthesis for RNA extracts, tagmentation and ligation for library construction.
    - Conducted two parallel qPCR assays as process controls to evaluate inhibitory effects: spiked murine norovirus (MNV) against endogenous pepper mild mottle virus (PMMoV)

- Flow chart of a single viral metagenome run with ancillary qPCR assay



Results:

Virome extraction optimisation is in progress. High-quality sequences require purity, sufficient total RNA concentration, and proper fragment range. The following contains preliminary results on a single run of viral library sequence preparation.

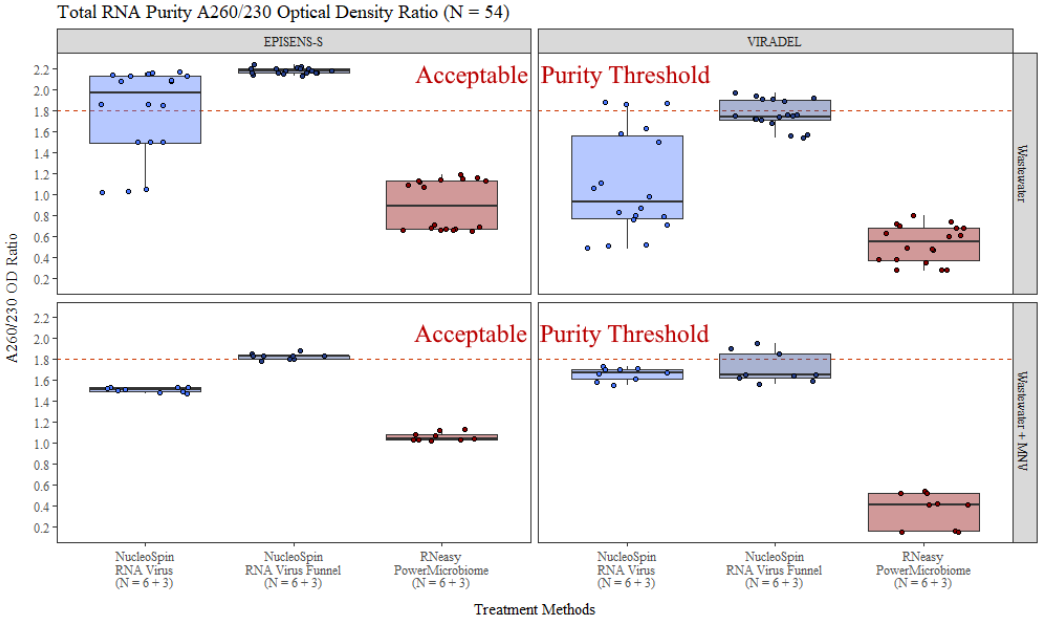


Figure 1. Purity by viral treatment method. MNV-Spiked and unspiked samples were tested.

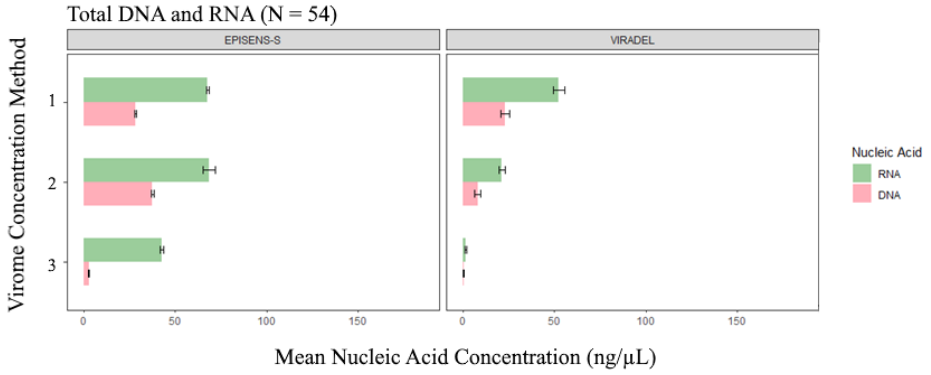


Figure 2. Total nucleic acid concentration. Virome extraction method 1: nucleospin RNA virus, 2: nucleospin virus funnel (10x dilution), and 3: RNeasy powermicrobiome with DNAse treatment.

For extraction kits, Nucleospin diluted treatment had the highest purity, followed by Nucleospin and PowerMicrobiome while viral concentration method EPISENS-S outperformed VIRADEL (**Fig. 1**). Total RNA for Nucleospin was relatively high for both EPISENS-S and VIRADEL, but DNA contamination was higher in PowerMicrobiome (**Fig. 2**). EPISENS-S contained more total RNA compared to VIRADEL.

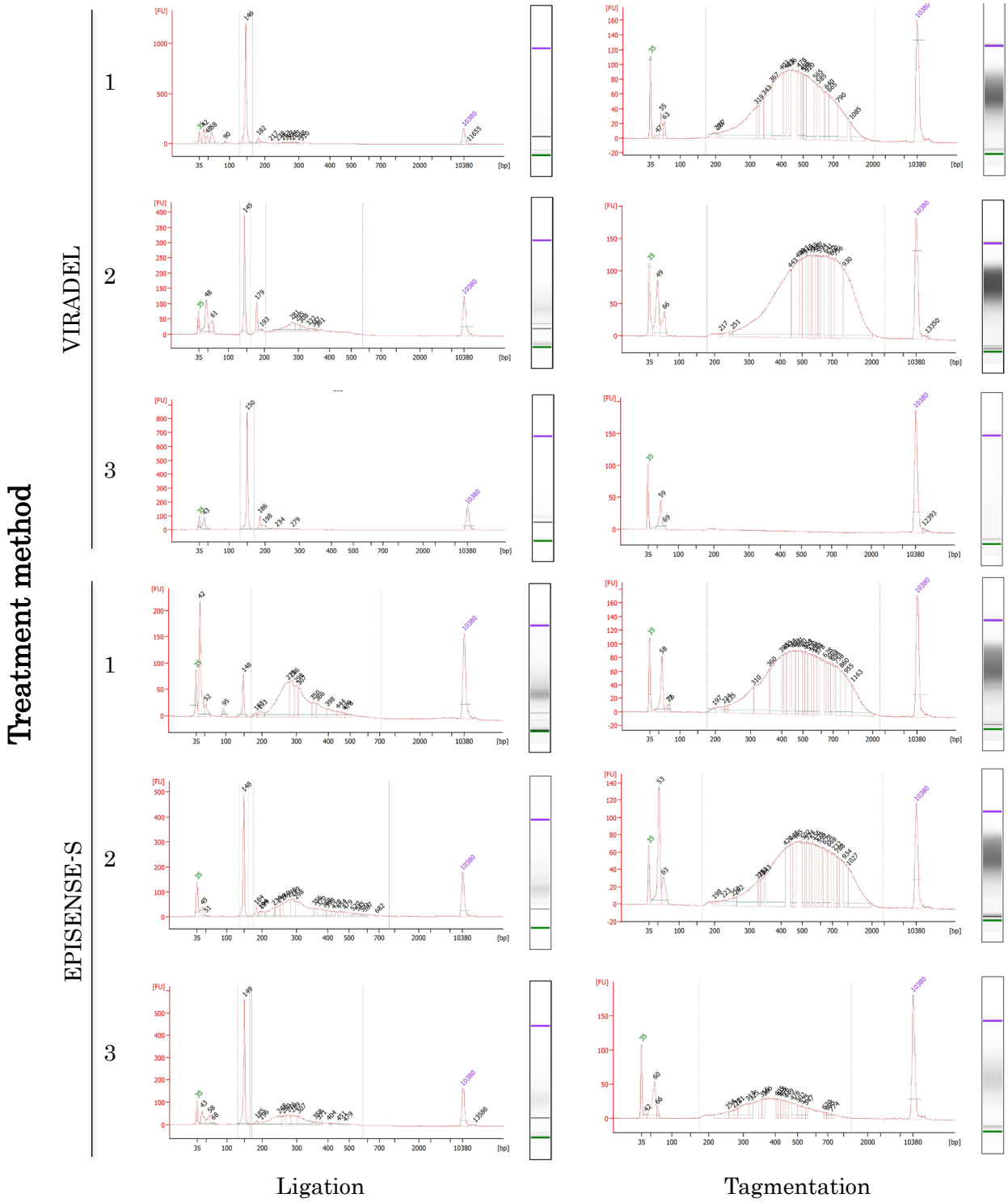


Figure 3. Examples of library fragment size range (n=12) according to library construction strategies. Virome extraction method 1: nucleospin RNA virus, 2: nucleospin virus funnel (10x dilution), and 3: RNeasy powermicrobiome with DNAse treatment.

Libraries from ligation kits ranged from approximately 197–682 bp, with adapter dimers observed at ~ 149 bp for all samples, while tagmentation libraries ranged from 197–2000 bp (**Fig.3**). Alarmingly, no libraries were constructed when treated with VIRADEL when combined with PowerMicrobiome for both ligation and tagmentation library strategies.

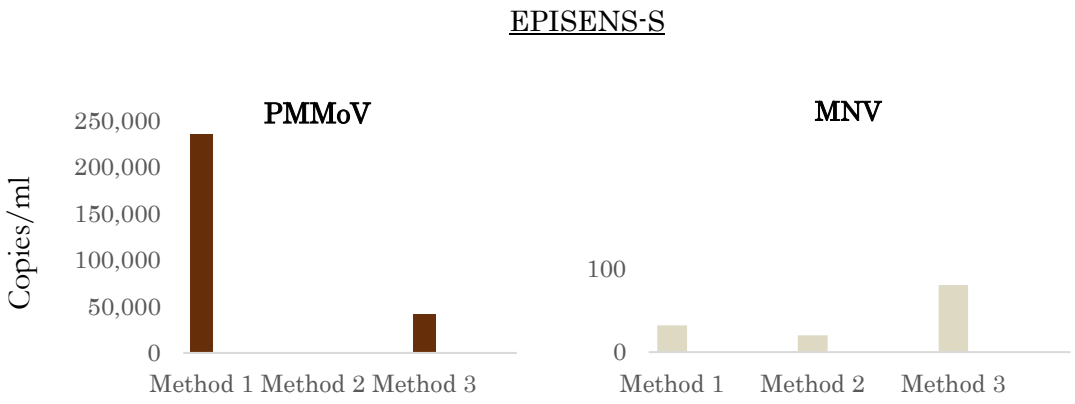


Figure 4. Ancillary Spike experiment post extraction for EPISENS-S. Virome concentration method 1: nucleospin RNA virus, 2: nucleospin virus funnel (10x dilution), and 3: RNeasy powermicrobiome with DNAse treatment.

PMMoV was found high when treated with method 1 for EPISENS-S. Recovery of MNV was possible in certain sampling dates but was inconsistent.

Future challenges:

The lack of standardized approaches and appropriately detailed reporting limits study design comparison and replication, Underlining the need for a systematic approach to data collection and analysis. The initial plan of executing an optimized virus metagenomic protocol on multiple sewage treatment plants across Asia was challenged by variable impurities inhibiting PCR performance during library construction. Furthermore, no libraries were generated using VIRADEL paired with PowerMicrobiome extraction kit, likely due to excessive enzymatic reaction, necessitating further testing. Considering virus feature prediction tools are reliant on reference databases, sequences should attempt to closely reflect natural reality to avoid bias. This study's sequence data may inform future machine learning models, but optimal library conditions must be achieved first. Further optimization is required.