3.25 Unclassified Viral Families and the Specificity of Hidden Markov Model Profiles

##

I generated all figures and tables presented (except Figure X) and performed preprocessing and calculations of summary statistics in the results section using Virusparies, the R package I created for this master's thesis. Contigs were grouped into different viral families on the y-axis of each figure by extracting the virus family from the `ViralRefSeq_taxonomy` column in both VirusHunter and VirusGatherer hittables. This extraction was handled internally by Virusparies functions.

Values in this column consist of a taxid and taxonomic information about the detected virus sequence, derived from the alignment of query sequences against the viral database, and are separated by the "|" character (e.g., `taxid:1354498|Hepacivirus colobi|Hepacivirus|Flaviviridae|Amarillovirales|Flasuviricetes|Kitrinoviricota|Orthornavirae|Riboviria`).

Occasionally, the VirusHunter and VirusGatherer pipelines fail to extract information about the viral family, resulting in contigs being grouped into the 'unclassified' category or 'unclassified' followed by the phylum name if phylum information is still present. I observed that approximately one-third of all contigs provided no information about family or phylum, while 3.35% provided information about the phylum but not the virus family. Only 2,894 out of the 6,209 viral contigs provided information about the family.

##

VirusHunter was employed to screen for viral sequences within unprocessed data by utilizing profiles of conserved proteins characteristic of specific virus families. The identified contigs were subsequently assembled using VirusGatherer. A total of eight profiles were applied (see Materials and Methods), and the results were consolidated into a single output file.

When running VirusHunter independently, it became apparent that the profiles used were not exclusively specific to their corresponding virus families. For example, when VirusHunter was run with SRA experiments aligned solely against the RNA-dependent RNA polymerase (RdRp) of the Flaviviridae family (Flavi_RdRp), 14 other families were identified within my VirusHunter hittable. Notably, some of these identified profiles—such as NidoAstro_RdRp, Nido_NiRAN, and Negative_Bunya-Arena_RdRp—correspond to the eight profiles utilized in this study. This suggests that the run for Flavi_RdRp also contributed to the assembly of contigs belonging to families associated with the other profiles.