Mining public metagenomic databases for the presence of AAV2

Student name: Lukas Visockas

Student ID: 2267636

Course: MSc Bioinformatics

Supervisor: Dr. Richard Orton

A report submitted in partial fulfilment of the requirements for the MSc

Bioinformatics Degree at The University of Glasgow

August, 2023

Summary:

In early 2022, over 1000s cases of acute severe hepatitis of unknown origin in children have been reported worldwide, with first few cases originating in Scotland. The initial routine blood tests excluded common causes of hepatitis, such as hepatitis A-E and human herpes viruses. Using untargeted metagenomic sequencing, several clinical studies found high levels of adeno-associated virus 2 (AAV2) in blood and liver samples that were later verified by RT-qPCR tests. Low levels of human adenovirus (HAdV) and human herpesvirus were also found in tested cases. There was no known prior association for AAV2 to be pathogenic, however it was established that it requires co-infection with a helper virus to replicate in humans. Understanding, how common AAV2 is in the general population, its genomic diversity and how often it co-occurs with other viruses can help to better prepare for such outbreaks. Hence, in this study a framework was developed to mine and analyse publicly available UnXplore and Serratus metagenomic datasets, to analyse how prevalent AAV2 as well as HAdV-F is in the human population. A simple bash script was implemented to create a local nucleotide BLAST database from UnXplore assembled contigs dataset, which was searched for the virus of interest. For Serratus dataset mining and analysis, a collection of custom-made bash and python scripts were created that would download and parse BAM and summary SRA sample files for the given viral GenBank ID, generate summary count matrix for co-occurrence analysis, visualise it in R, generate and filter consensus sequences and employ them for phylogenetic tree analysis. Out of available 963 human samples in UnXplore dataset, 0% (N=0) contained AAV2 and 1.45% (N=14) contained HAdV-F. For the Serratus analysis out of available 5,696,598 samples, 525 AAV2 sequences and 565 HAdV-F sequences were obtained and parsed, indicating 0.0027% (N=45) and 0.0033% (N=55) prevalence in the human sample population, respectively. The co-occurrence analysis results showed that in the AAV2 containing samples, *Parvoviridae* family was strongly clustering with *Adenoviridae*, as well as *Polyomaviridae* and *Herpesviridae* families, whereas no clustering was observed between *Adenoviridae* and other families in HAdV-F containing samples. Phylogenetic tree analysis of the AAV2 containing samples identified three interesting sequences that were clustering with sequences obtained from Scottish hepatitis cases, and one outlier sequence that looked very different from the rest, whereas no clustering was observed with HAdV-F containing samples. The results obtained from this study indicate that both AAV2 and HAdV-F are not that prevalent in the human population, however, they should be interpreted with caution as UnXplore sample size was quite small (N=963), whereas majority of Serratus samples did not contain full metadata information which would allow to identify human

samples. The framework developed in this study is not limited to AAV2 and HAdV-F only but can also be used to mine and analyse data on any virus of interest.

Acknowledgments:

I would like to take this opportunity to thank my supervisor Dr. Richard Orton, for granting me the opportunity to pursue this project and immensely helping me throughout the project.

I would also like to thank my families and friends for constant support. Special thanks to my wonderful flatmate Alexandra Cusmano for constat moral support and encouragement.

Abbreviations:

AAV2 – Adeno-Associated Virus 2.

HAdV – Human Adeno Virus.

SRA – Sequence Read Archive.

NGS – Next Generation Sequencing.

HHV – Human Herpes Virus.

Introduction and Aims:

In spring 2022, the first cases of acute severe hepatitis of unknown origin were reported in children in Scotland (Ho et al., 2023). This phenomenon shortly started to occur worldwide and as of 8th of July 2022, the World Health Organisation reported over 1000 of such cases across 35 different countries, with 272 cases located in the United Kingdom (World Health Organisation, 2022). The symptoms included abdominal pain, fatigue, vomiting, and in severe cases children required a liver transplant. A few fatalities have been reported as well. In order to identify the plausible cause, several independent studies were carried out in the UK and the USA.

Initially, human adenovirus F41, a member of the *Adenoviridae* family *Human mastadenovirus F* species (Benko et al., 2022), was suspected to be the cause of hepatitis in children due to the primary hospital PCR test results (Ho et al., 2023,). To further investigate the cause of hepatititis, clinical data was obtained from 32 infected children in Ho et al., 2023 study. Initially, routine blood tests for viral hepatitis were done, however they came negative, thus excluding common hepatitis causes such as hepatitis A-E viruses, acute Epstein-Barr virus, herpes simplex virus, human herpes viruses 6 and 7. Then metagenomic Next Generation Sequencing (NGS) was performed on the clinical samples as well as controls to identify both RNA and DNA virus that were present. The untargeted metagenomic sequencing analysis revealed that 26 out of 32 cases contained high levels of adeno-associated virus 2 (AAV2) in plasma and liver samples, that were further verified by RT-qPCR, in contrast to the unaffected control group where it was detected only in 5 out of 74 cases. In addition to that, human adenovirus (HAdV) F41, B and C as well as human herpesvirus (HHV) 6B were found in some of the samples, but at lower levels and not consistently. Another two studies conducted by Morfopoulou et al., 2023 in the UK and Servellita et al., 2023 in the USA, obtained similar results with patient samples containing high levels of AAV2, as well as lower levels of HAdV and HHV6. However, not much was known about AAV2 being pathogenic, as it has been widely used as a viral vector in gene therapy (Wang, Tai and Gao, 2019). Therefore, to better prepare for such outbreaks in the future, it is important understand how widespread and diverse AAV2 is in the human population, and its co-occurrence with other viruses.

Adeno-associated virus 2 is a small, non-enveloped, single stranded DNA satellite virus that belongs to *Parvoviridae* family *Dependoparvovirus* genus (Cotmore et al.,2019). Its genomic size is 4679 nucleotides long and its linear DNA contains ORFs for both structural and non-structural proteins located on the same strand (Figure 1). Its 3 promoter sequences P5, P19 and P40 are

transcribed by the host proteins through a rolling-hairpin mechanism, producing 9 different mRNAs that get translated into VP1, VP2 and VP3 capsid proteins. AAV2, like other *Dependoparvovirus* species, requires co-infection with a helper virus from *Adenoviridae* or *Herpesviridae* families for replication and growth (Meier et al., 2020). It infects many vertebrates, including up to 80% of humans, however, in adults it is usually asymptomatic (Li et al., 2011).

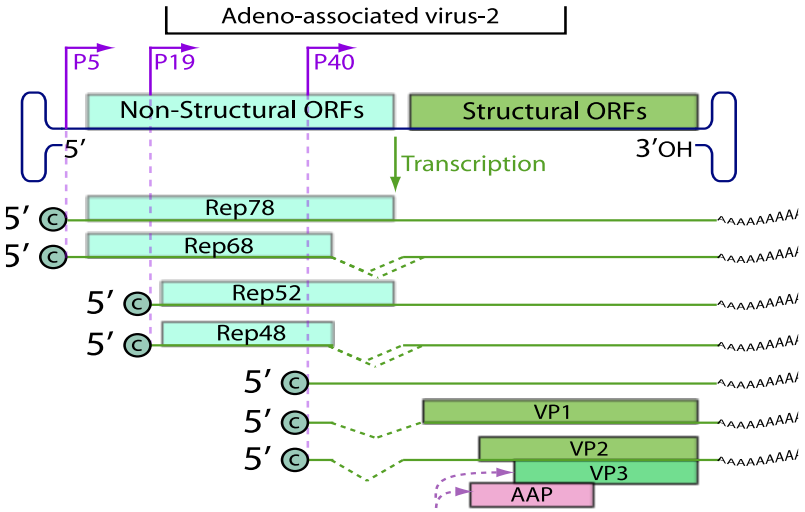


Figure 1: Schematic of the AAV2 genome.
The AAV2 genome is 4679bp long, containing both non-structural and structural ORFs on the same linear strand. It contains three promoter sequences P5, P19 and P40 that are being transcribed by the host proteins to produce VP1, VP2 and VP3 proteins. (Image taken from https://viralzone.expasy.org/226?outline=all_by_species).

Amid the rise of unprecedented viral infections, such as COVID-19 or AAV2-induced hepatitis, it becomes very useful to know how widespread these viruses are in general population. However, to individually test huge number of people for a specific virus would be cost and time expensive. That is why publicly available metagenomics datasets could be used to answer such biological questions in a cost-effective way, as they contain entire nucleotide sequences isolated from a specific microbiome.

A study conducted by Modha et al., 2022, used publicly available raw datasets to create a metagenomic analysis workflow for identifying and characterising unknown viral sequences in different microbiome samples. They used 963 human samples that originated from 40 different studies, covering 10 different microbiomes. The raw sequence data was *de-novo* assembled, generating a dataset of contigs longer than 300 nucleotides that could be easily searched for a

virus of interest. The framework they developed was called UnXplore and its generated dataset is publicly available and will be used in this study.

Another research conducted by Edgar et al., 2022, created a cloud computing infrastructure called Serratus that enabled high-throughput sequence alignment analysis at the petabase scale. They obtained over 5.7 million biologically diverse samples from SRA database and aligned them to a pangenome of Coronavirus sequences as well as all known RefSeq vertebrate viruses. The 7.3 terabytes of results were deposited into an open-access database www.serratus.io which can be explored via graphical interface (Figure 2). The database can be searched by entering either a Family name, GenBank ID or SRA Run. For example, a search of AAV2 outputs SRA runs with corresponding score, alignment identity and number of reads, which can be downloaded in a CSV format. In addition to exploration and visualisation at the family level, selecting individual SRA Run displays a coverage heatmap for each viral family that it is being mapped to. This information can be downloaded in a text summary file, as well as its corresponding BAM file that contains all the reads within the sample that were mapped to any of the reference sequences available on Serratus (N=13386). Overall, the free publicly available organised Serratus database structure can be easily searched and is a perfect resource for data mining and investigation on how widespread a virus of interest is in the population.

The main aim of this study was to create a set of scripts that would mine AAV2 and HAdV-F sequences from the publicly available UnXplore and Serratus metagenomic datasets. Then, using obtained sequences and available host information to analyse AAV2 and HAdV-F prevalence in the human population, their co-occurrence with other viral families, and investigate how genomically similar obtained sequences are to the first reported Scottish cases in children with hepatitis from Ho et al., 2023 study.

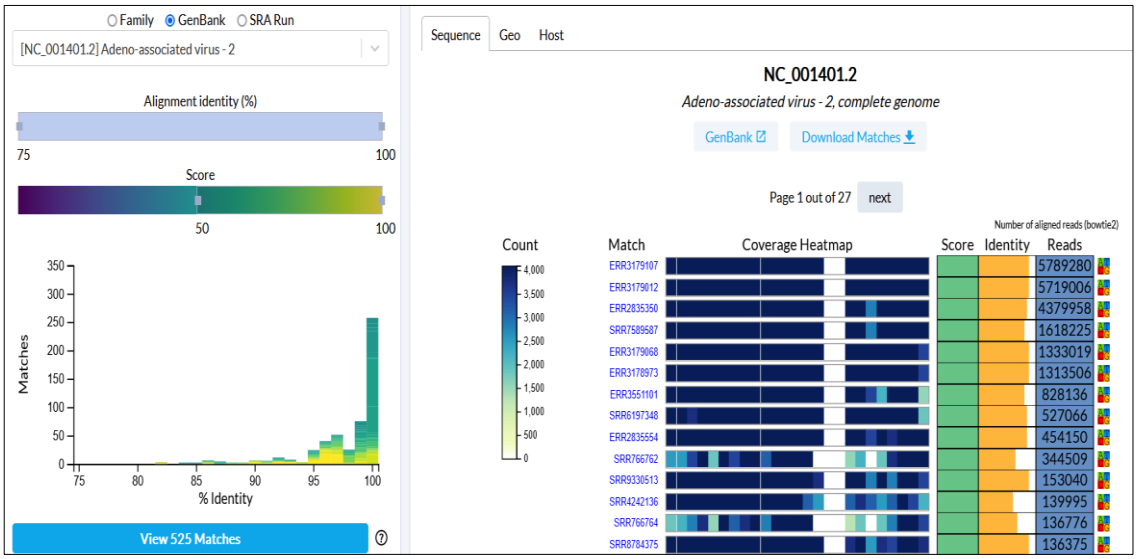


Figure 2: Screenshot of the Serratus web interface.
A screenshot of Serratus interface available at www.serratus.io. Virus of interest can be searched at three different levels: Family name, GenBank ID or SRA Run. Searching by a specific GenBank ID, it displays information about each SRA sample that query is found to, along with its coverage heatmap, score, alignment identity and a number of aligned reads to the query. Selecting individual SRA Runs display further information about how they map to various reference sequences that are available on Serratus. Finally, geographical location (Geo) and host information (Host) can also be viewed by selecting a corresponding tab located at the top. This screenshot displays an output example of search for "Adeno-associated virus 2".

Methods:

The overall analysis of this study consists of two parts. Analysis of the UnXplore dataset was performed using one custom made bash script, whereas Serratus dataset was analysed using three bash, three python and one R script. All scripts, detailed documentation and some input data is available at https://github.com/lukasv740/AAV2. The software and packages used for this study are displayed in Table 1.

Software	**Version**	**Use**
Bash	5.0.17(1)	Command Line Interpreter
Makeblastdb	2.11.0+	Software for Blast database creation
Blastn	2.11.0+	Software for blast database search
wget	1.20.3	Command for data scrapping
samtools	1.6	Software used for index, idxstats and mpileup commands
ivar	1.3.1	Software used for consensus sequence generation
mafft	7.475	Multiple alignment program for nucleotide sequences
iqtree2	2.1.3	Software for generating phylogenetic trees
python	3.10.4	A programming language
pandas	2.0.3	Library for data frame creation
biopython	1.81	Python module for sequence analysis
Rstudio	2023.06.1	A programming language for statistical analysis
ggplot2	3.4.2	R library for data visualisation
dplyr	1.1.2	R library for data manipulation
gridExtra	2.3	R library for graphics
devEMF	4.4-1	R library for graphics output
devtools	2.4.5	R library for developing R tools
ComplexHeatmaps	2.15.4	R library for heatmap creation
circlize	0.4.15	R library for circular visualisation
FigTree	1.4.4.	A tree figure drawing software

Table 1: A comprehensive list of software that was used for the data mining and analysis.

UnXplore:

The UnXplore dataset was provided by Dr. Richard Orton, which is also publicly available in the Modha et al., 2022 paper. The dataset consists of two FASTA files with assembled contigs that were longer than 300 bases (b) and 1 kilobase (Kb), respectively. It contains the contigs that were assembled from human metagenomic data coming from 963 samples from 40 different studies, across 10 unique microbiomes. To analyse this dataset, a custom bash script UnXplore.sh was created, which takes two positional arguments and is executed as follows:

bash UnXplore.sh <query> <database>

<query> - a FASTA file of genomic sequence that will be blasted (i.e., AAV2 sequence).

<database> - a FASTA file of assembled contigs to create a local blast database (i.e., 1Kb contig file).

The working principle of the script is that it first generates a local BLAST nucleotide database from the given assembled contig file by using 'makeblastdb' command from the NCBI BLAST suite of programs. Once the database is created, the nucleotide sequence query is searched using 'blastn' command from NCBI BLAST suite. The results are displayed in a .txt file in a default BLASTN tabular output format 6. Both results and generated database are saved in '../UnXplore' directory, which is created if it does not exit. Rerunning the script with the same contig file does not recreate the database, if it already exists.

Serratus:

The overall workflow for analysing Serratus database is outlined in Figure 3. It consists of downloading a Serratus summary file for an entered GenBank ID, followed by an extraction of the SRA sample accession IDs which are used to download corresponding SRA summary and BAM files. These files contain reads that were mapped to the pangenome of viral sequences as well as a text overview of number of reads mapping to different viral taxonomic families. These SRA summary files are used to create a family count summary matrix, which are subsequently analysed and visualised. Other steps include neighbourhood filtering, generation of consensus sequences, low quality sequences filtering, their alignment to study sequences and phylogenetic tree analysis. Documentation how to run each script, how it works, and recommended order is described below:

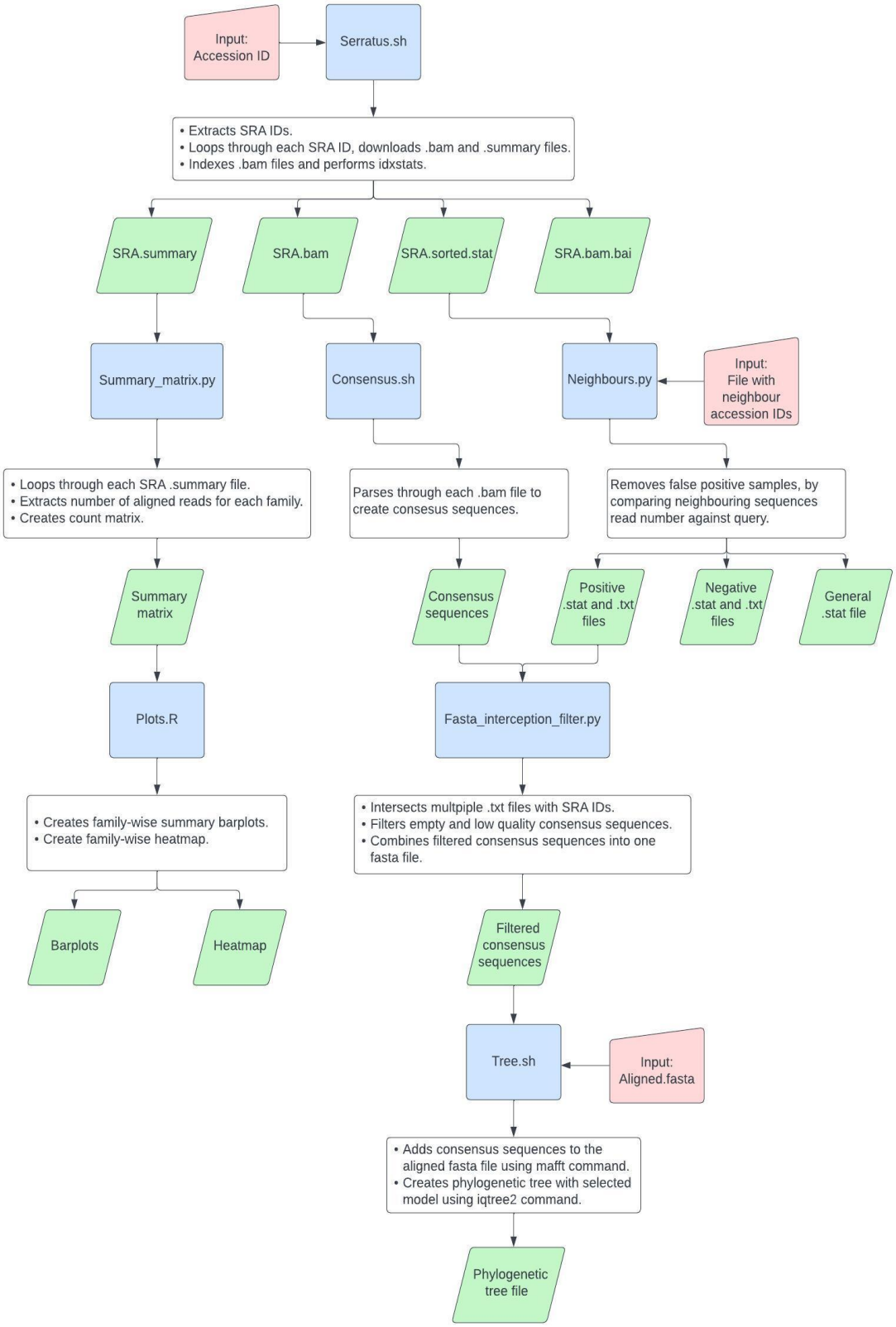


Figure 3: The overall Serratus analysis workflow.
The overall framework for mining and analysing sequences from publicly available Serratus database. Each script must be run separately, according to the outlined order and with specific user inputs, as defined in the method section.

bash Serratus.sh <Accession ID>

<Accession ID> -is a positional argument that intakes GenBank ID (i.e., NC_001401.2).

The main purpose of this script is to download Serratus BAM and summary files, according to a given genome sequence GenBank ID that is available in Serratus database. The script downloads a general summary file, which contains SRA matches which have alignment identity higher than 75% and score higher than 50%. Then, for each SRA match, it downloads a corresponding BAM and its summary file using 'wget' command. BAM files get indexed with 'samtools index' command and then 'samtools idxstats' is used to generate .sorted.stat file that displays number of hits to each viral genome, which is sorted in descending order. All the results are saved in '../Serratus/<Accession ID>/Data/' directory, which is created automatically.

python3 Summary_matrix.py <Accession ID> -output [output] -filter [filter] -sra [sra]

<Accession ID> - a positional argument that specifies working directory (i.e., NC_001401.2).

[output] – an optional argument for specifying output prefix name (default='Summary_matrix).

[filter] – an optional argument for specifying cutoff filter for number of aligned reads (default=100).

[sra] -an optional argument that intakes a .txt file with a list of SRA IDs (uses all SRA matches by default).

The script parses through BAM summary files to generate viral family read count matrix that is later used for co-occurrence analysis. The user specifies GenBank ID to access its BAM summary files, which are used to extract the number of hits to each viral family and are parsed to create a summary count matrix. In addition, it includes only those families that have a higher number of reads than the specified filter value.

R --vanilla --slave --args <Accession ID> <Virus name> [Filter value] [Summary file] < Plots.R

<Accession ID> - a positional argument that specifies working directory (i.e., NC_001401.2).

<Virus name> - a positional argument that specifies the name of the virus that is used for labelling purposes (i.e., AAV2).

[Filter value] – an optional positional argument that specifies a cutoff filter value for number of aligned reads (default=100).

[Summary file] – an optional positional argument that specifies summary file location, starting from specified working directory (default=Summary_matrix.csv).

The script uses generated summary count matrix for co-occurrence analysis. It creates two barplots using 'ggplot2' that show number of samples and total number of reads for each family. It also generates a heatmap using 'ComplexHeatmap' library that is used for visual representation of the matrix, as well as for family clustering analysis.

bash Consensus.sh <Accession ID>

<Accession ID> - a positional argument that specifies working directory (i.e., NC_001401.2).

The script generates consensus sequences for each SRA sample for the user provided GenBank ID. It works by generating text pileup output for each SRA BAM file using 'samtools mpileup' command, which is then piped into 'ivar consensus' command which generates consensus sequence that are saved in ../Serratus/<Accession ID>/Consensus/ directory.

python3 Neighbours.py <Accession ID> -query <query> -output <output> -neighbours_file <neighbours_file>

<Accession ID> - a positional argument that specifies working directory (i.e., NC_001401.2).

<query> – a required argument that intakes GenBank ID as a query (i.e., NC_001401.2).

<neighbours_file> - a required argument that intakes a .txt file with neighbouring GenBank IDs.

<output> - a required argument that intakes output prefix name (i.e., AAV2).

Due to the nature of sequencing error, biological artifacts and mis-mapping, a small fraction of reads is assigned to related sequences in Serratus database, which sometimes have higher read count than the original sequence. This script filters them out, by parsing through each SRA .sorted.stat file and comparing if the query sequence has a higher read count than the neighbouring sequences. For this script to work, the user must manually create a .txt file that contains GenBank IDs of the neighbouring sequences. For AAV2 the viral neighbour sequences are other adeno-associated viruses AAV1, AAV3, AAV4, AAV5, AAV6, AAV7 and AAV8, while for the HAdV-F the neighbour sequences are those closely related to the human adenoviruses HAdV-A, HAdV-B, HAdV-C, HAdV-D, HAdV-E and HAdV-54. The script generates five output files – a general .stat file that displays information on how many samples passed and failed the filter, as well as neighbouring sequence IDs that were used, .stat and .txt files with SRA IDs for positive samples that passed the filter, and .stat and .txt files with SRA IDs for negative samples that did not pass the filter.

python3 Fasta_intersection_filter.py <Accession ID> -files <files> -perc [perc] -output [output]

<Accession ID> - a positional argument that specifies working directory (i.e., NC_001401.2).

<files> - a required argument that intakes .txt files with SRA IDs for intersection.

[perc] – an optional argument that takes percentage value for filtering consensus sequences (default=50).

<output> - a required argument that intakes output prefix name.

The script intakes one or more .txt files with SRA IDs to create their intersection, which is used for filtering out low quality consensus sequences that will be later used for phylogenetic tree analysis. In this case, a positive SRA ID .txt file was used that was generated after filtering out neighbouring sequences. The script functionality allows user to input multiple .txt files, in which case their SRA IDs are intersected and used for further filtering. If only one file is provided, the script uses it to filter out low quality consensus sequences, which are defined as empty sequences or those that have a higher percentage of missing bases than specified by the user. All filtered consensus sequences are combined into one FASTA file.

bash tree.sh <Accession ID> <fasta_file> <alignment_file> <output> <model>

<Accession ID> - a positional argument that specifies working directory (i.e., NC_001401.2).

<fasta_file> - a positional argument that intakes FASTA file with filtered consensus sequences.

<alignment file> - a positional argument that intakes aligned FASTA files.

<output> - a positional argument that specifies output name.

<model> a positional argument that specifies model ('TIM+F+R3' for AAV2, 'K2P+R2' for HAdV-F).

The FASTA alignment files were provided by Dr. Richard Orton that were obtained from the Ho et al., 2022 study. The script takes consensus sequences and aligns to the alignment file using 'mafft' command. The final alignment file is then used by 'iqtree2' to generate a phylogenetic tree with a specified model. The phylogenetic tree was analysed using FigTree software.

Summary:

The Serratus workflow was performed for GenBank IDs NC_001401.2 and NC_001454.2, which correspond to Adeno-associated virus 2 (AAV2) and Human adeno virus F (HAdV-F), respectively. The neighbour files for AAV2 and HAdV-F with their neighbour sequence GenBank IDs are available in supplementary material (Named 'aav2_neighbours.txt' and 'hadvf_neighbours.txt', respectively) as well as alignment sequences from the studies used for phylogenetic tree generation (Named 'aav2_align.fasta' and 'hadv41_align.fasta' respectively). All the scripts create required subdirectories automatically and inform user through the terminal where all the data is saved. Finally, the script is versatile and should work with any other sequences of interest that are available on Serratus database.

Results:

UnXplore:

An analysis of the nucleotide sequences from the >300b and >1kB UnXplore contig datasets revealed that there were in total 7,155,623 and 28,837,029 assembled sequences incorporated into the generated local BLAST databases, respectively, with longest contig being 1,380,230 bases long in both cases. The nucleotide BLAST search for AAV2 sequences revealed 0 hits in both databases, whereas for HAdV-F there were 9 unique hits in 1kB contig database, and 14 unique hits in 300b contig database (Table 2). The size of contigs that were hit varied from 449 nucleotides to 6057 nucleotides long, meaning that only a small portion of the whole ~34Kb long HAdV-F genome was hit. Therefore, out of 963 human samples, there was 0% prevalence of AAV2 and 1.45% prevalence of HAdV-F, suggesting that AAV2 is not that widespread in the human population. Initially the analysis was performed only on >1kB contig dataset, however, since there were no hits for AAV2, the >300b contigs dataset was included for more thorough analysis.

	Total sequences	**AAV2 hits**	**HAdV-F hits**	**AAV2 %**	**HAdV-F %**
1Kb database	7155623	0	9	0	0.000126
300b database	28837029	0	14	0	0.000049

Table 2: Summary of UnXplore analysis.
The table shows a total number of contig sequences that were available in the UnXplore dataset, and the number of BLAST hits for AAV2 and HAdV-F, which are also represented as a percentage of all sequences.

Serratus:

The initial analysis of the Serratus database summary file for the AAV2 genome sequence (GenBank ID: NC_001401.2) revealed that there were 525 AAV2 samples with the alignment identity and score higher than 75% and 50%, respectively, from which 86 samples were of human origin. After filtering out samples were number of aligned reads to the neighbouring viral sequences (AAV1, AAV3, AAV4, AAV5, AAV6, AAV7 and AAV8) was higher than for AAV2 itself, 384 samples remained which were used to obtain AAV2 genome consensus sequences using 'ivar consensus' command on the Serratus BAM files. The final filtration of low-quality consensus sequences (removing sequences with high proportions of N bases) resulted in 157 samples remaining, out of which only 45 were of human origin. Therefore, out of 1,658,685 human samples available on Serratus, 0.0052% (86) had an indication of containing AAV2 while only 157 samples were used to generate reliable AAV2 genome consensus sequences (Table 3). However,

it should be noted that results might be underreported as many Serratus samples did not contain the host metadata or it was incomplete (labelled as generic 'metagenomics'), making it impossible to identify if samples were of human origin.

	Samples	Human samples	% Total human
Serratus	5696598	1658685	100.00
AAV2	525	86	0.0052
AAV2 neighbours	384	74	0.0045
AAV2 consensus	157	45	0.0027

Table 3: Summary of AAV2 Serratus analysis.
The table shows total number of samples and a number of human samples that were defined by the Serratus metadata. AAV2 stands for total samples that were initially obtained, AAV2 neighbours stands for number of samples after neighbour filtration, AAv2 consensus stands for samples after low quality consensus filtering.

Due to the nature of how Serratus metagenomic data was aligned against a pangenome of Coronavirus sequences as well as RefSeq vertebrate viruses, it was possible to obtain alignment information on family level. Therefore, a summary count matrix was created to capture possible co-occurrence viral families in each of the AAV2 containing samples. This is important as AAV2 (as well as other adeno-associated viruses) rely on co-infection with a helper virus in order to replicate in the host cells, so analysing co-occurrence might identify those families. Among the 525 AAV2 containing samples there were 24 unique families, with a filter value of at least 100 aligned reads. The filter value includes only those families that have a number of aligned reads higher or equal to the set filter value. The most dominant family *Parvoviridae*, to which AAV2 belongs to, was found in all 525 samples, followed by three major *Herpesviridae*, *Adenoviridae* and *Polyomaviridae* families that were found in 331, 321 and 276 samples, respectively (Figure 4). Additionally, out of 525 AAV2 containing samples, 65% (N=341) contained at least one of the aforementioned families. Interestingly, even though *Adenoviridae* was found in only 2/3 of all samples, it contained the highest total number of reads (higher than *Parvoviridae* itself), most likely due to its larger genome size which led to higher number of aligned reads per sample. Changing the filter value to include only samples with more than 1000 reads showed quite similar family grouping pattern, with *Parvoviridae* being the dominant one, followed by *Polyomaviridae*, *Adenoviridae* and *Hepadnaviridae*, and with considerably reduced number *Coronaviridae* positive samples (Appendix 1).

Figure 4: Barplots displaying viral family co-occurrence in AAV2 and their total number of reads. *The top barplot displays count of samples that contained corresponding family with the number of aligned reads higher than 100. The bottom one shows total number of reads for each family, displayed in log(10) scale.*

In order to determine the possible co-occurrence between the families, a clustered heatmap was produced using the same summary count matrix with filter value higher than 100 reads (Figure 5). Here a major cluster of *Parvoviridae*, *Adenoviridae*, *Polyomaviridae* and *Herpesviridae* families was observed, with *Parvoviridae* clustering the strongest with *Adenoviridae* family, suggesting possible occurrence relationship. Changing the filter value to 1000 showed similar clustering pattern, except for *Parvoviridae* clustering the strongest with *Polyomaviridae* family. This pattern change might contribute to the fact that some families like *Adenoviridae* was found in many samples, but with some samples containing very high levels of reads per sample (Appendix 2).

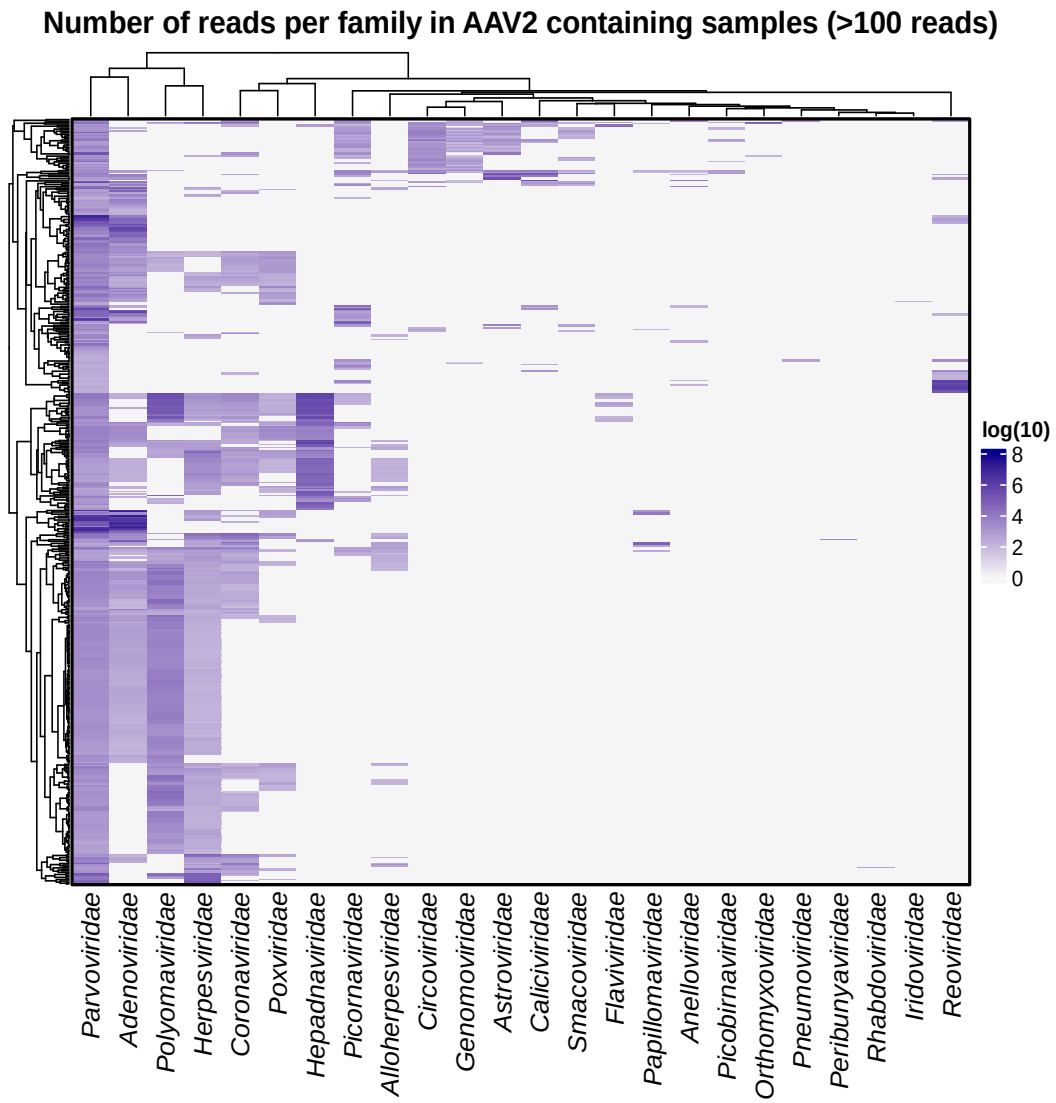


Figure 5: A heatmap of AAV2 count matrix.
A heatmap displaying family co-occurrence in AAV2 samples (N=525). Samples containing at least 100 reads are coloured in blue, more intense colour indicating higher number of reads. One major cluster of Parvoviridae, Adenoviridae, Polyomaviridae and Herpesviridae is observed.

After filtering out neighbour samples and removing low quality consensus sequences, a phylogenetic tree was built, using 157 filtered consensus sequences that were aligned against sequences used in the Ho et al., 2022 study (Figure 6). Here sequences from Serratus database are labelled as 'Consensus" followed by a corresponding SRA ID, whereas sequences from the study are labelled with GenBank ID. In addition to that, 9 study sequences that came from Scotland are coloured in red and labelled as Case1 to Case9. The phylogenetic tree was rooted against neighbouring adeno-associated virus sequences AAV1, AAV3, AAV3B and AAV6. Four interesting samples ERR2200449, ERR2298125, ERR2204077 and SRR4242136 were identified and are coloured in blue. The first three were observed to cluster with Scottish cases, indicating

their possible sequence similarity. Interestingly, upon further analysis on Serratus website, it was determined that all three of them came from the same global sewage sequencing study. The last interesting sample, SRR4242136, looked like an outlier and upon checking its origin it was revealed that it came from the megagenomic analysis of fecal virome of non-human primates from zoos in Eastern China.

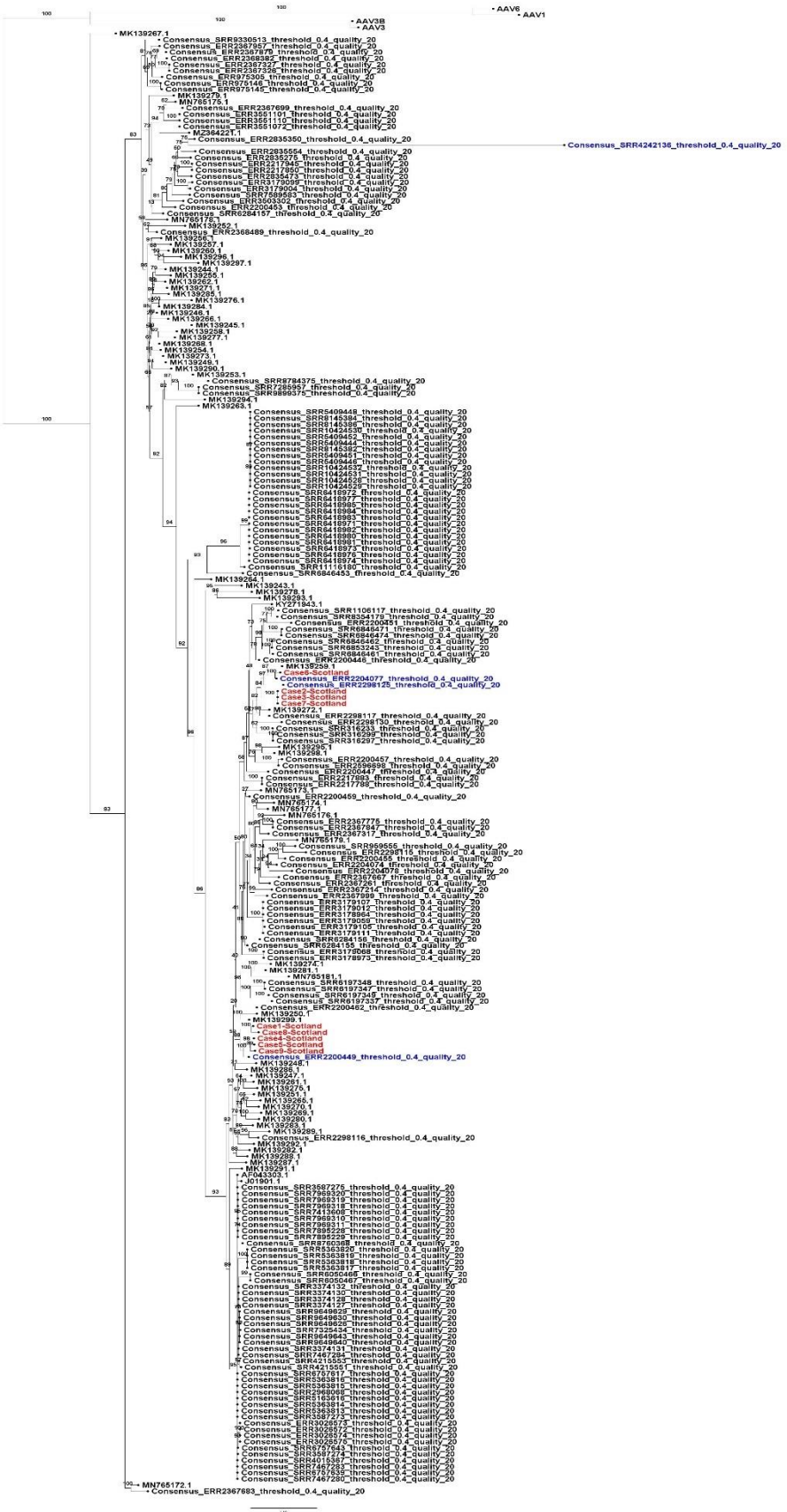


Figure 6: Phylogenetic tree of AAV2.

Samples that came from Serratus database labelled as 'Consensus', where samples from the study contain GenBank IDs. In addition to that, Scottish study cases are coloured. Four interesting samples were observed and are coloured in blue (Built using TIM+F+R3 model).

Similar analysis was conducted for HAdV-F, where initially 565 samples were obtained out of which 201 were of human origin. After filtering out neighbour sequences (HAVA, HAVB1, HAVB2, HAVC, HAVD, HAVE and HAV54) and removing low quality consensus sequences, 107 HAdV-F samples remained, out of which 55 belonged to human. Therefore, the HAdV-F prevalence was 0.0033% in all the human samples that were available on the Serratus database. Again, it should be noted that this analysis is dependent of the host metadata available for Serratus, which is often generic with terms such as 'metagenome'. However, it is clear that relatively few samples are positive for HAdvF in Serratus as a whole on Serratus, and it is therefore not a common virus within the data of the SRA.

	Samples	**Human samples**	**% Total human**
Serratus	5696598	1658685	100.00
HAdV-F	565	201	0.0121
HAdV-F neighbours	481	174	0.0105
HAdV-F consensus	107	55	0.0033

Table 4: Summary of HAdV-F Serratus analysis.
The table shows total number of samples and a number of human samples that were defined by the Serratus metadata. HAdV-F stands for total samples that were initially obtained, HAdV-F neighbours stands for number of samples after neighbour filtration, HAdV-F consensus stands for samples after low quality consensus filtering.

As before, summary count matrix was generated to analyse viral family co-occurrence with aligned read values of over 100. Here, among 565 HAdV-F samples there were 23 unique families identified (Figure 7). The most dominant family *Adenoviridae* was found in all 565 samples and contained the largest number of total reads. It was followed by *Coronaviridae*, *Reoviridae* and *Parvoviridae* families that were found in 120, 74 and 69 samples, respectively. Interestingly, the difference in number of samples between the first and second families is way bigger than it was for AAV2 samples. Changing the filter value to include only families with aligned reads over 1000, did not make substantial difference – *Adenoviridae* still being the most dominant one, followed by *Parvoviridae*, *Picomaviridae* and *Caliciviridae* (Appendix 3).

To visualise family co-occurrence a heatmap was constructed using the same HAdV-F summary count matrix as before (Figure 8). Here it was observed that *Adenoviridae* family does not cluster with other families, suggesting that there is no relationship between HAdV-F occurrence with other viruses.

Similarly, a phylogenetic tree was built using 107 filtered consensus sequences that were aligned to HAV41 sequences obtained from Ho et al., 2022 study (Figure 9). The tree was rooted at the midpoint, which lead to two major branches. Here, all the sequences that were obtained from

Serratus database are coloured either in magenta or blue, for top and bottom branch, respectively. Samples from study are labelled with corresponding GenBank ID and there was one Scottish case that is coloured in red. Upon further BLAST analysis of some sequences from two top and bottom branches it was determined that this separation occurs between two human adenovirus F serotypes – HAV40 and HAV41, respectively.

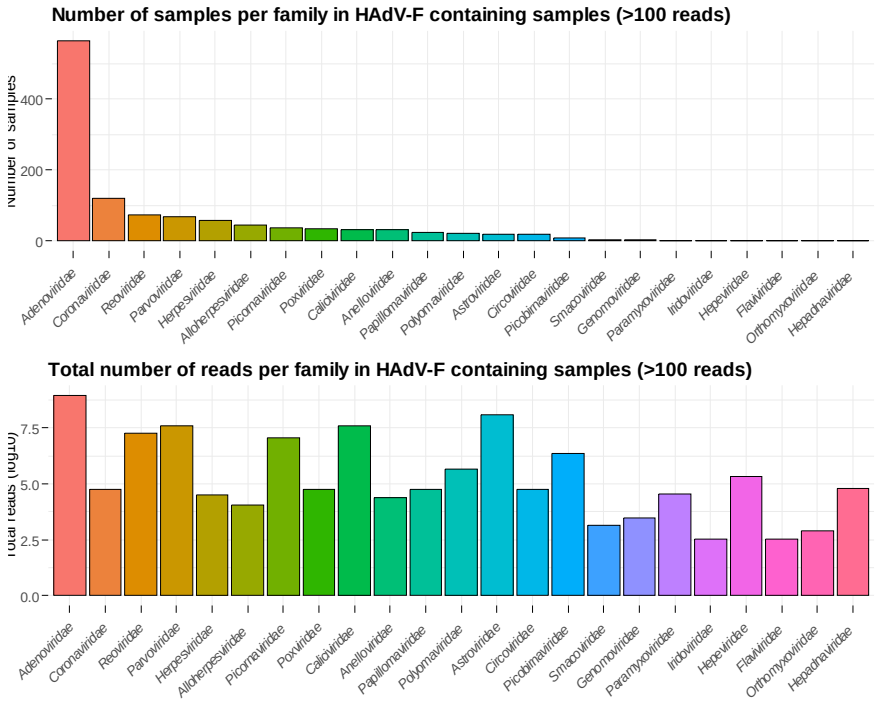


Figure 7: Barplots displaying viral family co-occurrence in HAdV-F and their total number of reads. *The top barplot displays count of samples that contained corresponding family with the number of aligned reads higher than 100. The bottom one shows total number of reads for each family, displayed in log(10) scale.*

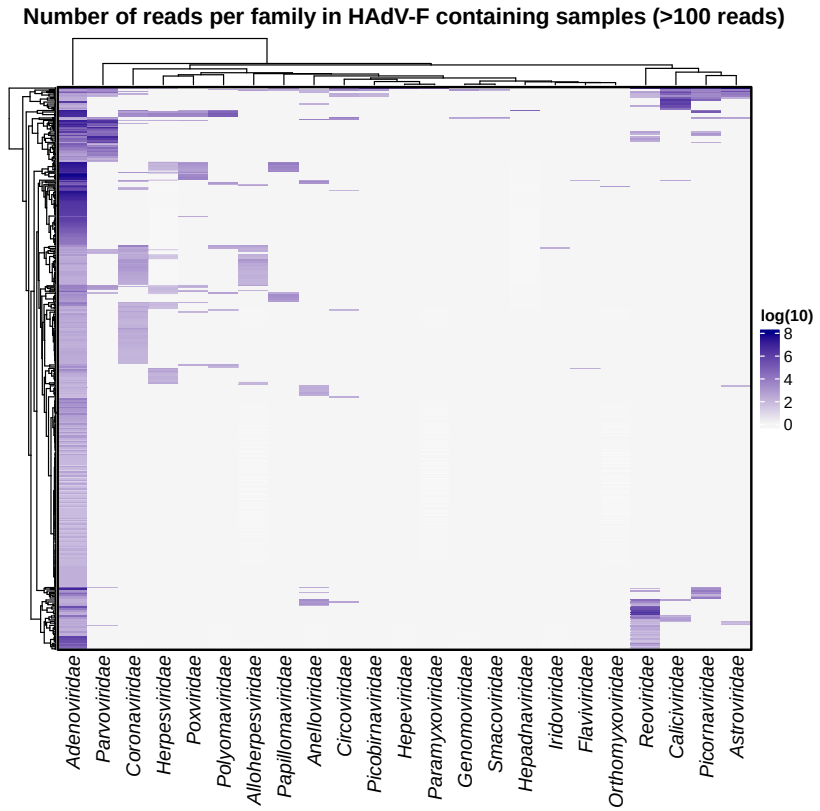


Figure 8: A heatmap of HAdV-F count matrix.
A heatmap displaying family co-occurrence in HAdV-F samples (N=565). Number of reads is displayed in a gradient blue colour – the more intense the colour, the higher number of reads. Adenoviridae does not cluster with other families.

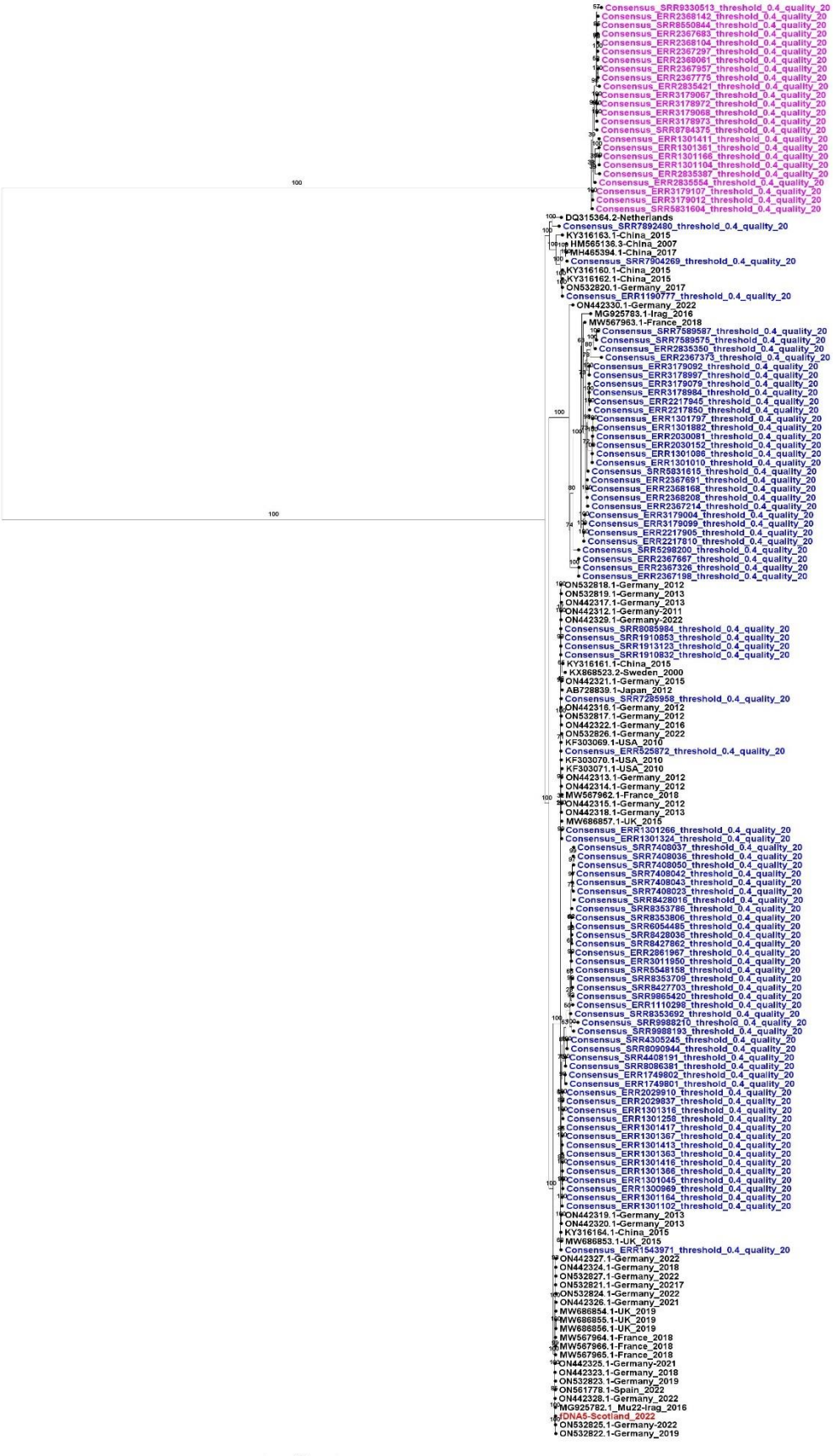


Figure 9: Phylogenetic tree of HAdV-F.

The main root was split at the midpoint, separating HAV40 and HAV41 serotypes for top and bottom branches, respectively. Red samples correspond to the Scottish case, purple samples represent HAdV-F40, while blue samples represent HAdV-F41 serotype. (Build using K2P+R2 model).

Discussion:

Computational overview:

In this study, a collection of scripts was created to mine publicly available metagenomic UnXplore and Serratus databases for any given query of the virus sequence. For the UnXplore dataset analysis, a simple bash script was created that intakes a file containing assembled contig sequences, from which local nucleotide BLAST database is generated. This database is then searched for the given virus sequence and the results are outputted into a nice statistical summary. For the Serratus dataset analysis, a set of scripts were created that search and download BAM and summary files for the SRA runs corresponding to the query virus accession ID. These files are parsed to create various statistics at different levels, R plots for data visualisation, generation and filtration of consensus sequences and phylogenetic tree analysis. All the scripts were created in such a way that they would not be limited to this study on AAV2 and HAdV-F analysis alone but should work for any provided virus sequence that is available on Serratus database. However, the user should be noted that some datasets can be large (i.e., both AAV2 and HAdV analysis required over 80GB of space), therefore a considerable size of a disk and fast internet connection is required.

In addition, the 'leaky alignment' problem was addressed where due to the sequencing errors and similarities between viral genomes, some reads are mistakenly mapped to the neighbouring viral sequences (i.e., AAV2 and AAV3), thus creating false positives in Serratus database. Usually, these reads are insignificant as they fall way below noise level, however, in some cases, especially in the samples with high viral read-count they outcompete the query sequence. The custom-made python script was able to successfully filter them out, however the user must manually define the neighbouring sequences GenBank IDs.

Finally, the script design was implemented in a way to maximise automation and minimise human error. The script creates all the required directories, and saves files in the organised manner, outputting saved path to the terminal. The user just provides an accession ID and sometimes a file name, and if something is not right (i.e., misspelling the accession ID or incorrect file path) the script outputs a corresponding error message to the terminal, without creating unnecessary directories or files. Scripts that create summary count matrix and R plots are also customisable, in case user wants to change the output file name or filter value for the number of aligned reads. Given more time, it should be possible to combine all scripts into a single pipeline, for even better automation and lower level of human errors.

Statistical overview:

The results obtained from UnXplore dataset analysis were quite unexpected. Out of 963 different human samples, none of them contained any hits for AAV2, and only 14 unique hits for HAdV-F, indicating that their prevalence in human population was 0% and 1.45%, respectively. This might have happened, either due to AAV2 actually not being that prevalent in the human population, or due to a small number of samples (N=963) that were used in UnXplore study, or a combination of both.

In contrast, Serratus database analysis yielded more interesting results. Out of 5696598 samples available on Serratus, 525 were mapped to AAV2 out of which 86 belonged to humans. After filtration of false positives and low-quality sequences, it was determined that AAV2 prevalence was 0.0027% (N=45) in human population, meaning it is not widespread in humans. Out of 157 filtered consensus sequences, only 3 samples, that came from wastewater metagenomic analysis, clustered with Scottish cases indicating their similarity. Co-occurrence analysis for AAV2 showed clustering with *Adenoviridae* and *Herpesviridae*, which reaffirmed the hypothesis that AAV2 requires helper virus from those families (Meier et al., 2020). In addition to that, these results correlate with the findings from other studies, where AAV2 was found in the presence with HAdV-F and HHV6B viruses (Ho et al., 2023). Interestingly, AAV2 was observed to cluster with *Polyomaviridae* family, and there are no reports describing it as AAV2 helper virus. However, previously *Polyomaviridae* and *Papillomaviridae*, a known AAV2 helper virus family, belonged to a single viral family called *Papovaviridae*, indicating their possible similarities and thus explaining its clustering with AAV2 (Chapter 10 Parvo, 1985).

The Serratus analysis on HAdV-F, where initially 565 samples were obtained, showed a prevalence of 0.0033% (N=55) in human population, which is way lower than the one obtained from UnXplore analysis. On one hand, Serratus contained over 1700 times more human samples than UnXplore dataset, meaning results should be more reliable, but on the other hand, it lacked information about the sample origin, thus possibly missing out on potential human samples. As expected, co-occurrence analysis showed no clustering with other families, as it does not require helper virus, and phylogenetic tree analysis displayed no clustering with the one Scottish case sequence.

Limitations:

One of the biggest limitations in analysing metagenomic data was availability of its metadata. Serratus database was built on samples obtained from Sequence Read Archive, many of which

were lacking or had incomplete information about where samples came from or how they were processed. For example, some AAV2 samples on Serratus were labelled as general metagenomic data, and there is no way of knowing whether they were of human origin or not, so the results might be underreported. Therefore, the metagenomic analysis results should be interpreted with a bit of caution.

Moreover, a better way of analysing sequences would be obtaining their raw FASTQ files from SRA and performing *de novo* assembly. This method produces more reliable contigs than generating consensus sequences, as it is less susceptible to false positive samples as well as genomic differences in viral sequences. However, this method takes a lot of time and is computationally expensive, as it requires many powerful processors and a lot of disk space.

In conclusion, publicly available metagenomic datasets are powerful assets for the bioinformatic analysis, allowing researchers to answer specific biological questions in a cost and time effective manner. The scripts created in this project offer a fast and efficient way to mine and analyse necessary data from Serratus database, and even though it was focused on AAV2 and HAdV-F viruses only, it could be used for any other virus of interest.

References:

Benkő, M., Aoki, K., Arnberg, N., Davison, A.J., Echavarría, M., Hess, M., Jones, M.S., Kaján, G.L., Kajon, A.E., Mittal, S.K., Podgorski, I.I., San Martín, C., Wadell, G., Watanabe, H. and Harrach, B. (2022). ICTV Virus Taxonomy Profile: Adenoviridae 2022. Journal of General Virology, 103(3). doi:https://doi.org/10.1099/jgv.0.001721.

Cotmore, S.F., Agbandje-McKenna, M., Canuti, M., Chiorini, J.A., Eis-Hubinger, A.-M., Hughes, J., Mietzsch, M., Modha, S., Ogliastro, M., Pénzes, J.J., Pintel, D.J., Qiu, J., Soderlund-Venermo, M., Tattersall, P. and Tijssen, P. (2019). ICTV Virus Taxonomy Profile: Parvoviridae. Journal of General Virology, 100(3), pp.367–368. doi:https://doi.org/10.1099/jgv.0.001212.

Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., Banfield, J.F., de la Peña, M., Korobeynikov, A., Chikhi, R. and Babaian, A. (2022). Petabase-scale sequence alignment catalyses viral discovery. Nature, 602(7895), pp.142–147. doi:https://doi.org/10.1038/s41586-021-04332-2.

Ho, A., Orton, R., Tayler, R., Asamaphan, P., Herder, V., Davis, C., Tong, L., Smollett, K., Manali, M., Allan, J., Rawlik, K., McDonald, S.E., Vink, E., Pollock, L., Gannon, L., Evans, C., McMenamin, J., Roy, K., Marsh, K. and Divala, T. (2023). Adeno-associated virus 2 infection in children with non-A–E hepatitis. Nature, [online] pp.1–11. doi:https://doi.org/10.1038/s41586-023-05948-2.

Li, C., Narkbunnam, N., Samulski, R.J., Asokan, A., Hu, G., Jacobson, L.J., Manco-Johnson, M.J. and Monahan, P.E. (2011). Neutralizing antibodies against adeno-associated virus examined prospectively in pediatric patients with hemophilia. Gene Therapy, 19(3), pp.288–294. doi:https://doi.org/10.1038/gt.2011.90.

Meier, A.F., Fraefel, C. and Seyffert, M. (2020). The Interplay between Adeno-Associated Virus and Its Helper Viruses. Viruses, 12(6), p.662. doi:https://doi.org/10.3390/v12060662.

Modha, S., Robertson, D.L., Hughes, J. and Orton, R.J. (2022). Quantifying and Cataloguing Unknown Sequences within Human Microbiomes. mSystems. doi:https://doi.org/10.1128/msystems.01468-21.

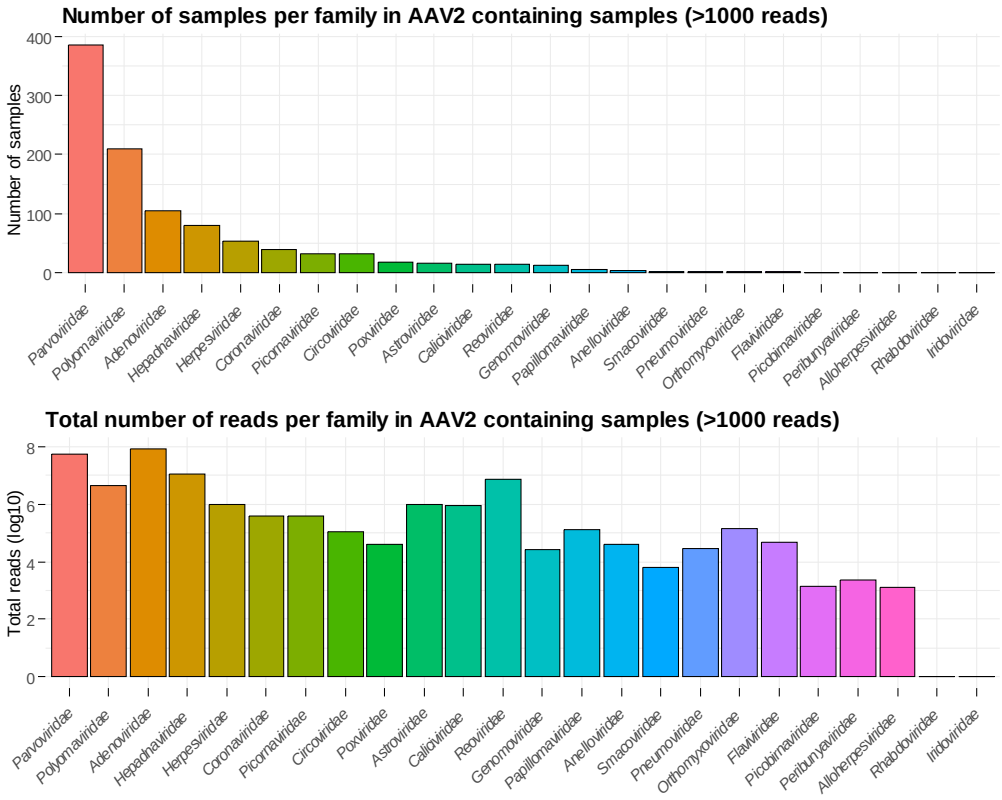
Morfopoulou, S., Buddle, S., Enrique, O., Atkinson, L., José Afonso Guerra-Assunção, Mahdi Moradi Marjaneh, Riccardo Zenezini Chiozzi, Storey, N., Campos, L., J. Ciaran Hutchinson, Counsell, J.R., Pollara, G., Roy, S., Venturini, C., Antinao, J.F., Siam, A., Tappouni, L.J., Zeinab Asgarian, Ng, J. and Hanlon, K.S. (2023). Genomic investigations of unexplained acute hepatitis in children. Nature, 617(7961), pp.564–573. doi:https://doi.org/10.1038/s41586-023-06003-w.

Servellita, V., Gonzalez, A.S., Lamson, D.M., Foresythe, A., Huh, H.J., Bazinet, A.L., Bergman, N.H., Bull, R.L., Garcia, K.Y., Goodrich, J.S., Lovett, S.P., Parker, K., Radune, D., Hatada, A., Pan, C.-Y., Rizzo, K., Bertumen, J.B., Morales, C., Oluniyi, P.E. and Nguyen, J. (2023). Adeno-associated virus type 2 in US children with acute severe hepatitis. Nature, [online] pp.1–3. doi:https://doi.org/10.1038/s41586-023-05949-1.

Wang, D., Tai, P.W.L. and Gao, G. (2019). Adeno-associated virus vector as a platform for gene therapy delivery. Nature Reviews Drug Discovery, [online] 18(5), pp.358–378. doi:https://doi.org/10.1038/s41573-019-0012-9.

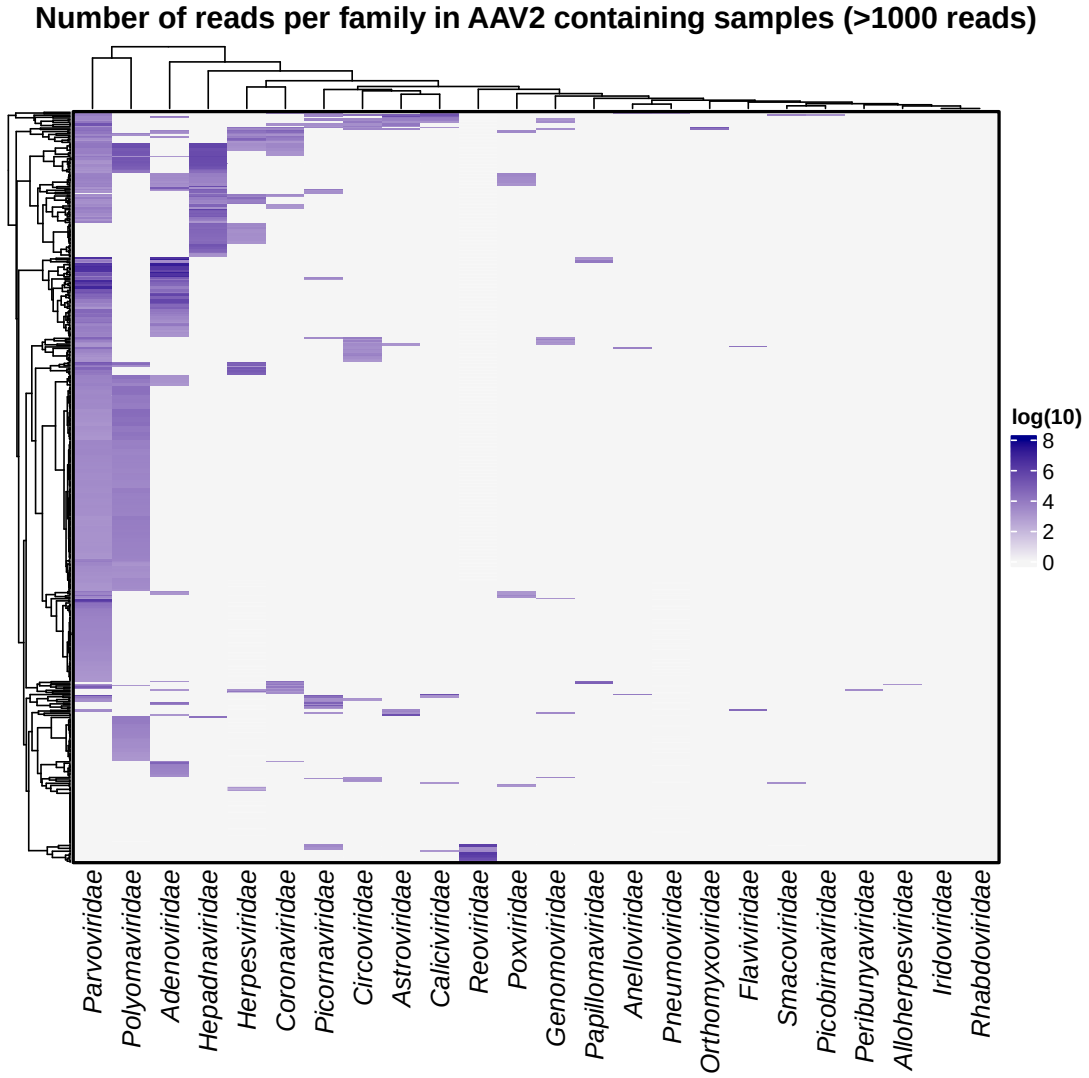
World Health Organization (12 July 2022). Disease Outbreak News; Acute hepatitis of unknown aetiology in children - Multi-country.

Appendix:



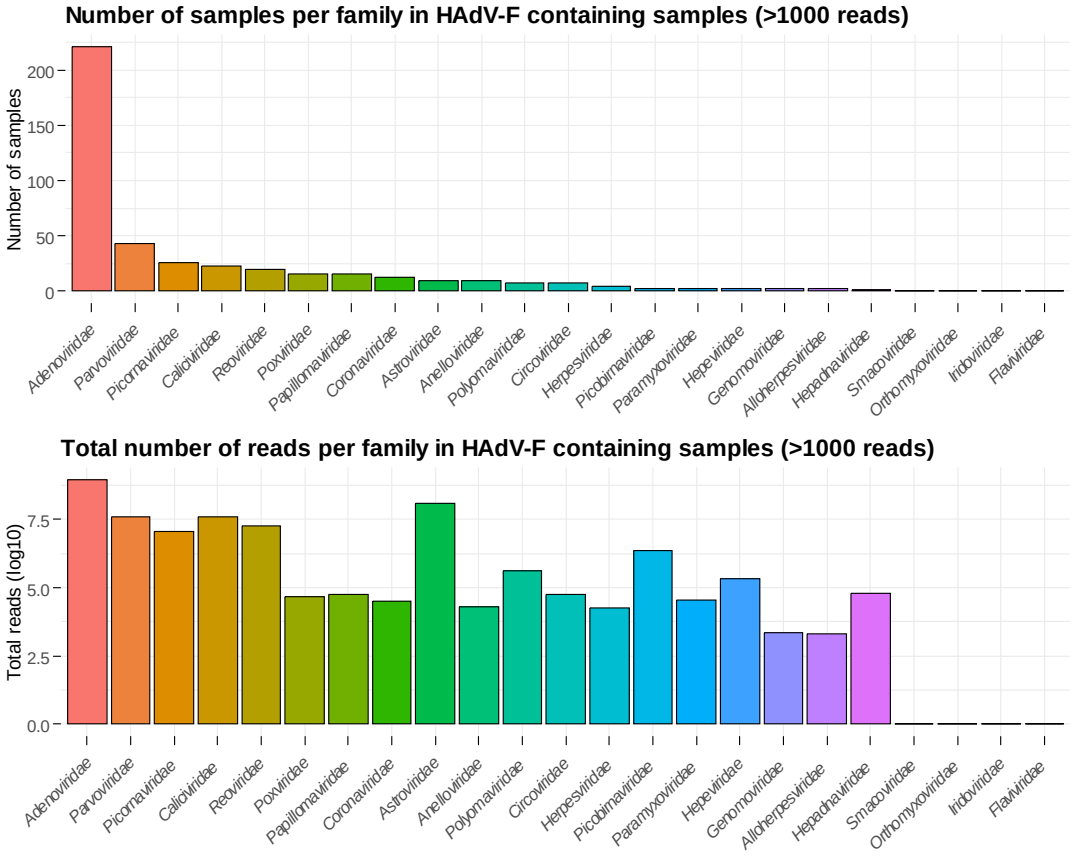
Appendix 1: Barplots displaying viral family co-occurrence in AAV2 and their total number of reads.

The top barplot displays count of samples that contained corresponding family with the number of aligned reads higher than 1000. The bottom one shows total number of reads for each family, displayed in log(10) scale.



Appendix 2:: A heatmap of AAV2 count matrix.

A heatmap displaying family co-occurrence in AAV2 samples (N=525). Samples containing at least 1000 reads are coloured in blue, more intense colour indicating higher number of reads.



Appendix 3: Barplots displaying viral family co-occurrence in HAdV-F and their total number of reads.

The top barplot displays count of samples that contained corresponding family with the number of aligned reads higher than 1000. The bottom one shows total number of reads for each family, displayed in log(10) scale.