Hecatomb: An End-to-End Research Platform for Viral Metagenomics

Author Names: Michael Roach (1), Sarah Beecroft (2), Kathie A Mihindukulasuriya (3,4), Leran Wang (3,4), Lais Farias Oliveira Lima (5), Elizabeth A. Dinsdale (1), Robert A. Edwards (1), Scott A. Handley (3,4)

Author Affiliations:

1. Flinders Accelerator for Microbiome Exploration, Flinders University, Adelaide, SA, Australia

2. Harry Perkins Institute of Medical Research, Perth, WA, Australia

3. Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO, USA

4. The Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St. Louis, MO, USA

5. Biology Department, San Diego State University, San Diego, CA, USA

Keywords: virome, virus discovery, bioinformatic workflow, metagenomics

ABSTRACT

Analysis of viral diversity using modern nucleic acid sequencing technologies presents several unique challenges. Foremost being that virus detection requires a non-targeted, random (shotgun) approach. This process collects sequences not only from the viral fraction of the sample, but also from other biological sources. Annotation and enumeration of collected sequences requires rigorous quality control, effective search strategies against relevant reference sequence databases and statistical and visualisation strategies to evaluate results. Here we introduce *hecatomb*, a bioinformatics platform enabling end-to-end virome sequence analysis. Hecatomb enables both read and contig based analysis and integrates query information from both amino acid and nucleotide reference sequence databases. Hecatomb prioritizes integration of data collected throughout the workflow as well as with external viral data sources. This process results in a rich, high-dimensional data which can be used by researchers to rigorously evaluate their results.

INTRODUCTION

Viruses play important roles in every ecosystem they are found. Not only are they extraordinarily abundant, (with current estimates as high as 10^{31} viral particles, making them the most dominant entity on the planet (Hendrix et al. 1999; Mushegian 2020)), but they exert significant influence on their surrounding environment. This is in part due to their omnipresence in all cellular life forms (excluding some intracellular bacteria) (Koonin et al. 2020). Viruses parasitize host cell molecular processes and as a result alter host cell physiology. This interaction may only alter a single cell, but often results in changes to organismal or ecosystem biology.

Unfortunately, viral impact on organismal physiology often manifests in the form of infectious disease. Perhaps the most acute example being the ongoing pandemic caused by the emergence and spread of the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) in late 2019. SARS-CoV-2 was identified in patients exhibiting acute respiratory infection symptoms, with some patients quickly developing acute respiratory failure and other serious complications (Li et al. 2021). The devastating SARS-CoV-2 pandemic resulted in the death of millions and disruption to all aspects of daily life.

While global pandemics alter the health and lives of millions on the planet, the mechanisms and impact of viral infectious diseases are incredibly diverse. This ranges from well-known viral diseases such as the "common cold" which can be attributed to rhinoviruses, coronaviruses, adenoviruses and enteroviruses (Heikkinen and Järvinen 2003) to emerging examples such as the potential role for Epstein-Barr virus (EBV) in multiple sclerosis (Bjornevik et al. 2022). Consequential virus - host interactions are also not limited to humans. For example, Geminivirus infection of plants has resulted in catastrophic crop loss such as the regular 100% crop loss of tomatoes in Italy and the Dominican Republic (Picó et al. 1996) and a nearly US$2 billion loss in African cassava production (Patil and Fauquet 2009). Similarly, animal viruses such as foot-and-mouth disease virus (FMDV), a highly contagious disease of cloven-hoofed animals, is widespread in Africa with an annual US$2.3 billion impact (Casey-Bryars et al. 2018). While FMDV rarely infects humans, the impact of virus 'spillover' infection from animal to human (zoonotic viruses) is unfortunately a regular event for many other viruses (Prempeh 2001). Some zoonotic viruses are well studied, including Ebola and Zika viruses, however concern for future impactful spillover is also merited (Grange et al. 2021). Knowledge of what viruses are in vertebrate and invertebrate organisms is key to identifying virus - disease associations as well as developing preparatory strategies to reduce their economic and health impacts.

Viruses also have influence over global ecosystems. This has been extensively studied in aquatic environments. One example of this is intracellular iron release from bacteria following lytic bacteriophage

(herein called phage) infection. The liberated cellular iron is taken advantage of by phytoplankton improving growth rates (Poorvin et al. 2004). Many other aquatic nutrient cycling and biogeochemical processes have also been attributed, both directly and indirectly, to how phages modify prokaryotic and protistan assemblages (Wilhelm and Suttle 1999; Fuhrman 1999; Wommack and Colwell 2000; Weinbauer 2004; Suttle 2005). Terrestrial environments are also shaped by interactions between phage and their host bacteria with emerging data suggesting a role for soil viral communities in carbon and nutrient cycling (Emerson et al. 2018; Trubl et al. 2018; Starr et al. 2019).

While the infectious disease examples described above were associated with single viruses and the ecological examples were from viral assemblages there is a growing body of evidence to suggest that viral assemblages are also important in human health and disease. Disease associated viral assemblages, regularly referred to as 'viromes' have been identified in a variety of human diseases (Liang and Bushman 2021). For example, stool samples from patients with inflammatory bowel disease (IBD) have expanded numbers of phage from the Order Caudovirales (Norman et al. 2015; Pérez-Brocal et al. 2015; Fernandes et al. 2019; Clooney et al. 2019; Liang et al. 2020). Expansion of viruses that infect vertebrate cells are also regularly associated with disease in the gut. Enteric vertebrate virus expansion has also been reported in both primates and humans with lentivirus induced acquired immunodeficiency syndrome (AIDS). With the enteric burden of picornaviruses and adenoviruses increasing in macaques with Simian Immunodeficiency Virus (SIV) induced AIDS and similarly the enteric *Adenovirus* burden expanding in humans with severe Human immunodeficiency virus (HIV) induced immunodeficiency (Handley et al. 2016; Monaco et al. 2016). Thus, similar to environmental ecosystems, human health also appears to be influenced by viral assemblages and not just single infectious viruses.

Understanding viromes as key components of organismal and environmental microbiomes is crucial to the future of human, animal, and environmental health. Metagenomic sequencing offers a powerful tool to uncover viral diversity and identify novel viruses. However, there are currently many challenges associated with viral metagenomics. For example, despite viruses being the most abundant and diverse biological entity on the planet, they represent a minority of reference genomes in GenBank, largely due to difficulties associated with studying them (Krishnamurthy and Wang, 2017). Recent efforts to populate new viral genomes into reference database have yielded 10s to 100s of thousands of novel phage metagenome-assembled genomes (MAGs) (Camarillo-Guerrero et al. 2021; Nayfach et al. 2021; Nayfach et al. 2021; Soto-Perez et al. 2019; Gregory et al. 2020; Paez-Espino et al. 2019; Tisza and Buck 2021). Other efforts have yielded many new high-quality viral genomes by combining the laborious and time-consuming experimental work with student-learning outcomes (Hanauer et al. 2017), but there is still much to learn about the viral genetic content of ecosystems.

This vast viral dark matter poses a significant barrier to annotating viral sequences. Our most efficient approach for sequence annotation relies on sequence similarity searches using popular tools such as BLAST or RAPSearch2 (Altschul et al. 1990; Zhao et al. 2012; Zhao et al. 2017; Kalantar et al. 2020). While the search algorithm may differ between approaches, if an appropriate reference is absent from reference databases annotation efforts will be futile.

Reference databases also vary greatly in size. As of January of 2022 there were over 224 million proteins and 40 million nucleotide entries in NCBI RefSeq compared to only 11,561 viral protein 14,742 viral nucleotide entries (https://www.ncbi.nlm.nih.gov/refseq/statistics/) (O'Leary et al. 2016). The difference between querying all of RefSeq (all-vs-all) or a RefSeq viral only database would require computation across a database 3-4 orders of magnitude larger than a virus-only RefSeq sequence database. When querying the millions of sequences obtained from modern instruments this can pose significant computational challenges. The benefit of the all-vs-all strategy is that you can annotate sequences using information from all domains of life. This is the strategy employed by the IDSeq platform which has been

successfully used to annotate viral and non-viral sequences, but requires a large compute infrastructure (Kalantar et al. 2020). Alternatively, a 'tiered-approach' as implemented by VirusSeeker offers an interesting alternative which can be run on commodity hardware (Zhao et al. 2017). The tiered approach first queries all sequences against a virus-only reference database to identify sequences that are potentially viral. To confirm that they are likely viral in origin, a secondary query of the potentially viral sequences against a database containing all sequences is done. This 'divide and conquer' approach minimizes the secondary search to the larger databases by minimizing the number of sequences in the primary search to those that are most likely to be viral in origin. A major drawback of this approach is that it will only classify viral sequences due to the primary search, but the approach can greatly decrease the computational requirements. Of note, issues regarding database content and size are amplified when deciding to query target sequences against protein or nucleotide database. Choosing to query only one type of reference database (amino acid or nucleotide) may influence what organisms you are able to detect based on the content of each reference database and should be taken into consideration.

While database content and search strategies complicate positively identifying true-positive viral hits with confidence, a potentially greater challenge is that viral metagenomes are often plagued with false positive classifications (Rosseel et al. 2014; Skewes-Cox et al. 2014; Ponsero and Hurwitz 2019). False positives occur for many reasons. Viruses often share sequence homology with other domains of life. For instance, as 'stolen' genes from their hosts' genomes, or viruses may contain repetitive or low-complexity regions that are also found on other organisms such as insertion elements or transposons (Rosseel et al. 2014; Skewes-Cox et al. 2014; Ponsero and Hurwitz 2019). Regardless of the providence of false-positive classifications, they can greatly influence data interpretation. At worst, false-positive classification could mis-identify a viral pathogen in a clinical sample. Or for virome analysis increased false-positive results would greatly influence estimates of diversity leading to spurious conclusions.

Here we present Hecatomb, a bioinformatic platform designed to address many of these issues. Hecatomb performs rigorous quality control followed by tiered taxonomic assignment using MMSeqs2 against customised viral and secondary confirmation databases enabling analysis on commodity hardware. Hecatomb also performs metagenomic assembly and contig classification enabling simultaneous analysis of both read and contig based annotations. Importantly, while hecatomb provides pre-compiled databases and recommended settings it is easily customizable. The primary output of Hecatomb is a comprehensive annotation table containing data generated throughout the workflow that can be easily merged with sample data. This annotation table can then be used by investigators to investigate the properties of virome sequence data using plotting and statistical analysis.

METHODS

Hecatomb pipeline overview and implementation. An overview of the Hecatomb pipeline is shown in Figure 1. In brief, Hecatomb processes reads through 4 key modules (Fig 1A). First, reads are preprocessed to remove low-quality or contaminating sequences. Next, preprocessed reads are passed through both a read-based analysis and an assembly module (modules 2 and 3). The final and fourth module combines information obtained from both the read-based and assembly modules. Results are stored throughout each module, primarily as tab-separated value (tsv) files for easy interrogation using commonly used software approaches (e.g. python, R, Bash). Emphasis is placed on data preservation at each stage to provide analysts with as much detail as possible to perform analysis and decision making.

Hecatomb consists of the main analysis pipeline, as well as several accessory scripts for installing reference databases. The Hecatomb pipeline is written in python using Snakemake (https://doi.org/10.12688/f1000research.29032.2). There are several features that are designed to facilitate installing and running Hecatomb on a high-performance computing (HPC) environment (Fig 1B).

Hecatomb is installed via Conda and makes liberal use of Conda environments to ensure portability and ease of installation of dependencies. Conda environments for jobs are created automatically by Snakemake in the Hecatomb installation directory and dependencies are downloaded and installed as needed. The use of isolated Conda environments for Hecatomb and the individual pipeline jobs minimises package version conflicts, minimises overhead when rebuilding environments for updated dependencies, and allows maintenance and customisation of different versions of Hecatomb and its dependencies without interacting with installed programs and modules on the system.

A launcher script is included to make running the pipeline as simple as possible. Accessory scripts are also available from the launcher. The launcher provides the required file paths and default configuration and offers a convenient way to modify parameters and customise options. The Snakemake command generated and run by the launcher is printed to the terminal window and saved to the run log file for the user's reference.

Hecatomb is able to be deployed on a HPC and has been designed to make use of Snakemake profiles for cluster job schedulers (e.g. Slurm, SGE, etc.). The use of profiles allows Snakemake to submit pipeline jobs to the job scheduler and monitor their progress. Although optional, using the scheduler is highly recommended as it allows for more efficient use of HPC resources compared to submitting the whole Hecatomb pipeline to run as a single job. Profiles can be created manually. Example profiles are included for users to use as templates for their own HPC environment. Hecatomb has been designed with the available Cookiecutter (https://github.com/cookiecutter/cookiecutter) profiles in mind (https://github.com/Snakemake-Profiles/doc) with the aim of making Hecatomb compatible with existing profiles, and to make setting up new profiles as straight-forward as possible.

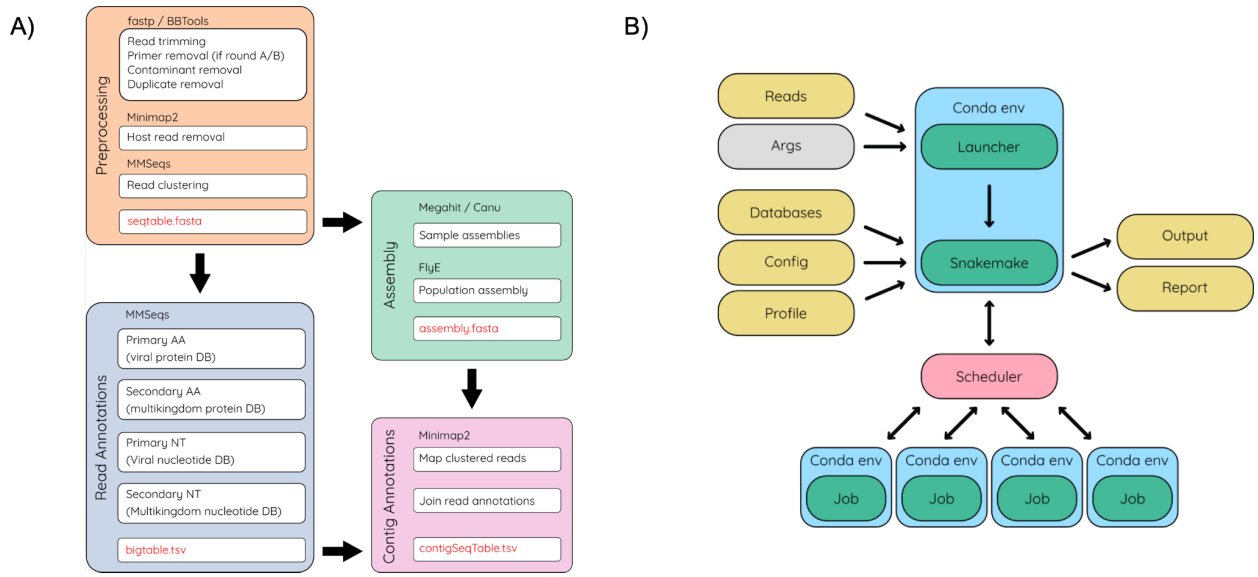

Figure 1: Hecatomb pipeline and implementation. (**A**) The Hecatomb pipeline is divided into four modules. Sequencing reads for each sample undergo preprocessing and clustering (*orange*); quality trimmed reads for each sample undergo assembly and assemblies for each sample are coalesced into a single assembly (*green*); clustered reads undergo annotation using viral and multi-kingdom protein databases and clustered reads not annotated by the protein search are annotated using viral and multi-kingdom nucleotide databases (*blue*); the read-based annotations are combined with the assembly to provide contig annotations (*pink*). The assembly stages–*green* and *pink*–can optionally be skipped. **(B)** Hecatomb takes in command line arguments, data, configuration parameters and outputs both results for

analysis and information on the run. It can interact with the job scheduler in HPC environments, which distributes individual tasks to the job queue. Boxes are coloured like so: Command-line arguments, *grey*; files, *yellow*; Conda environments, *blue*; scripts/programs, *green*; workload manager, *pink*.

Sequence data preprocessing. Hecatomb is capable of processing reads obtained using a variety of commonly used library preparation and sequencing platforms. This includes processing of both single and paired-end Illumina reads as well as long-read technology from PacBio and Oxford NanoPore platforms. Hecatomb is capable of processing sequences obtained from other library types with minor modifications to the Hecatomb configuration file and by supplying library specific adapters or primer sequences to remove during preprocessing. Hecatomb is also designed to process reads obtained from the round A/B protocol (Finkbeiner *et al.*, 2009). This protocol enables library production from all types of viral genomes (single and double stranded RNA and DNA viruses) but in doing so requires the use of a combination of phased PCR primers which can result in the integration of several non-biological sequence contaminants. Hecatombs' preprocessing module removes all types of non-biological sequence contaminants, ensuring the removal of non-biological sequences which may contaminate assemblies or result in mis-leading sequence similarity queries.

As an additional precaution, Hecatomb removes common laboratory sequence contaminants. This includes the removal of ΦX174 bacteriophage genome sequences and sequences that map to the NCBI UniVec database (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/). ΦX174 is normally spiked-in to Illumina runs to improve sequence quality scores by increasing sequence heterogeneity. This may remove "real viral sequence" belonging to lambda-like phages in a sample. To limit this we require an exact match to the ΦX174 reference genome. However, if a project is specifically interested in lamba-like phages with high-degrees of sequence similarity to ΦX174, then this step can be removed. The number ΦX174 and UniVec sequences that are removed are tabulated and can be reviewed by analysts to evaluate the number and types of sequences detected and removed by Hecatomb. Low-quality sequence is also trimmed from the ends of each read. Importantly, quality filtering rules (e.g. minimum read length, low-quality PHRED score threshold) can be easily adjusted in the hecatomb configuration file to suit individual project needs.

For host-associated samples (e.g. stool, saliva, skin swabs) Hecatomb implements a host-sequence removal strategy. Host sequences are removed using Minimap2 and a host reference genome specifically optimised to avoid removing potentially viral sequences using the following strategy. First, all viral genomes from NCBI viral assembly (https://www.ncbi.nlm.nih.gov/assembly/?term=viruses) are downloaded and computationally 'shred' into short fragments with an average length of 85 bases sharing a 30 base overlap using shred.sh from the BBTools suite (https://sourceforge.net/projects/bbmap/). Shredded viral sequences are then mapped and masked from host-reference genomes using bbmap.sh requiring a minimum identity of 90% and at most, 2 insertions and deletions. In addition, low-entropy sequences are masked from host genomes (entropy = 0.7) using bbmask.sh. This process results in a set of host-associated reference genomes that have had 'virus-like' and low-entropy sequences masked, limiting the likelihood that a real viral sequence will be removed from downstream analysis at this step. Pre-computed masked reference genomes for the following host genomes: human, mouse, rat, camel, *Caenorhabditis elegans*, dog, cow, macaque, mosquito, pig, rat and tick. These are available directly within Hecatomb using the *–host* flag and scripts are provided to generate new host genomes as necessary.

At this stage Hecatomb has generated high-quality, contaminant/host cleaned sequences. For the final stage of the module, Hecatomb removes sequence redundancy by clustering each sample using linclust (Steinegger and Söding 2018). The purpose of clustering sequences is to reduce the number of

sequences requiring taxonomic classification to a single, representative sequence from a clutter of similar sequences. Prior to sequence clustering, exact duplicates are removed. Sequences are clustered requiring a minimum sequence identity of 97% and 80% alignment coverage of target sequence to the representative sequence (--min-seq-id 0.97 -c 0.8 --cov-mode 1). These settings can be customised. Importantly, Hecatomb maintains the size of each cluster in the annotation table as well as the counts normalised to the total number of high-quality reads per each individual sample (Counts Per Million (CPM)). These counts and normalised counts can be used by analysts as a measure of cluster size for each representative sequence. Similar to preprocessing settings, clustering settings are also easily adjustable in the Hecatomb configuration file.

At the end of this process sequences should be clean from non-biological (reagent) contaminants, free from likely host-sequence and redundancy reduced. These high-quality sequences are then used for read-based taxonomic classification and *de novo* assembly.

Taxonomic assignment. Taxonomy is assigned to high-quality, representative sequences using an iterative query strategy against both nucleotide and amino acid reference databases (Fig. 2). This strategy is designed to provide the fewest possible false-positive viral annotations while maintaining sensitivity and runtime performance. All queries are carried out using mmseqs2 (Steinegger and Söding 2017). The strategy starts with a translated query of all sequences against a database of all viral (taxonomy ID: 10239) amino acid sequences in UniProtKB (UniProt Consortium 2021) clustered at 99% identity to reduce redundancy and target database size. Any sequence assigned a tophit virus taxonomy is subsequently cross-checked by being queried against the full multi-kingdom UniClust50 amino acid database (Mirdita et al. 2017). This two-step query strategy captures all potentially viral reads in the first step, reducing the number of queries required in the secondary confirmatory step against the larger trans-kingdom database. This effectively reduces the computational burden and reduces the probability of false-positive viral sequence classifications and is similar to the approach employed by other virus sequence annotation pipelines such as VirusSeeker (Zhao et al. 2017).

Sequences not identified as viral using translated queries are subject to a similar iterative search strategy using untranslated queries against a viral nucleotide sequence database consisting of all viral sequences in GenBank clustered at 100% identity to remove redundancy followed by a secondary confirmatory query against a customised polymicrobial nucleotide database containing representative RefSeq genomes from Bacteria (n = 14,933), Archaea (n = 511), Fungi (n = 423), Protozoa (n = 90) and plant (n = 145) genomes. This trans-kingdom reference database is 6 orders of magnitude smaller than GenBank nt ($5.0x10^6$ vs $1.3x10^{12}$) which is used by other computational intensive platforms such as VirusSeeker and IDSeq (Zhao et al. 2017) (Kalantar et al. 2020)). This enables the sequence searches to be run on commodity hardware while still having representation of a broad diversity of non-viral kingdoms.

Following the secondary translated and untranslated searches Hecatomb will annotate sequences using the 2bLCA approach (Hingamp et al. 2013). This approach provides conservative taxonomic assignments at lower-nodes of the tree when similarity is found across a heterogeneous collection of taxonomies. However, the LCA algorithm fails when crossing certain taxonomic levels. For example, sequences with similarity to both bacterial and viral taxa have a LCA of “root” in the NCBI tree (Fig 2B). This can also be an issue within the viral tree wherein sequences will be assigned to “virus root” (Fig 2C). Hecatomb detects these instances and instead of classifying them to the assigned root lineages, the sequences are refactored (LCA-refactoring) to the top-hit. While this sometimes results in the reclassification of sequences to a non-virus lineage (e.g. if the tophit was bacterial) it provides additional information about sequences with ambiguous taxonomic assignments which can then be used as additional sequence

annotation information for an analyst instead of discarding the sequence with the uninformative “root” or “virus root” lineage information.

These MMSeqs searches are the most time-consuming stage of Hecatomb. Users have the option of using the default high-sensitivity parameters (--start-sens 1 --sens-steps 3 -s 7 --lca-mode 2) for identifying all known viruses including distantly-related viruses, or specifying fast search parameters (-s 4.0 --lca-mode 2) for identifying known and less distantly-related viruses, but with greatly improved runtime performance.

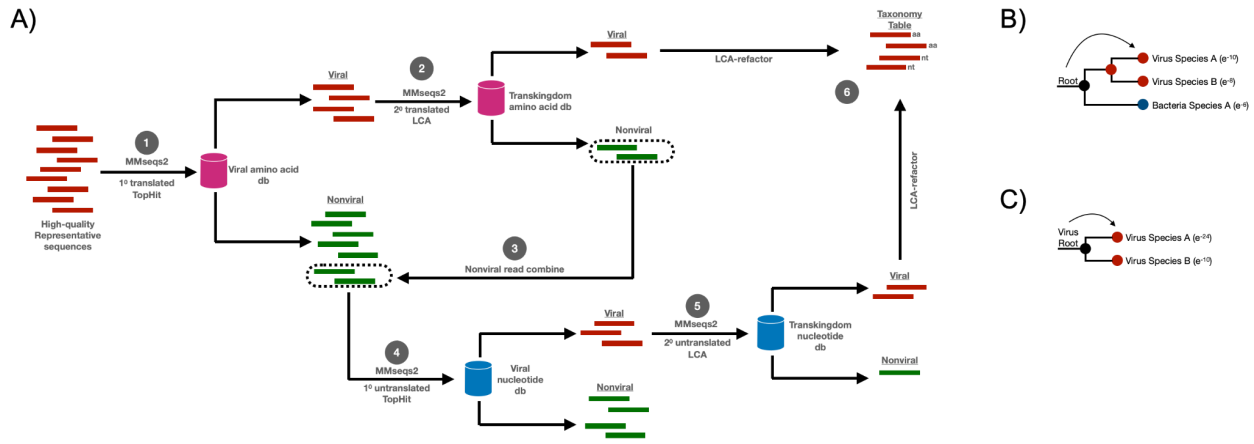

Figure 2: Iterative Taxonomic Assignment Strategy. (A) All queries are completed using MMSeqs2. (1) High-quality representative sequences are queried against a viral amino acid (aa) sequence database. (2) Potentially viral sequences are subjected to a secondary, confirmatory query against a transkingdom aa sequence database. (3) Sequences classified as viral during the translated queries are combined and subjected to an untranslated query to a viral nucleic acid sequence database (nt) (4) followed by a secondary, confirmatory query against a transkingdom nt database (5). Sequences classified as viral in either the translated (aa database) or untranslated (nt database) queries are combined into a final taxonomy table. **(B)** Example of lowest-common ancestor (LCA) refactoring when a sequence is LCA classified across kingdoms and collapses to ‘root’. **(C)** Example of LCA refactoring when a sequence is collapsed to the ‘virus root’ and is refactored to the tophit. For both **(B)** and **(C)** Hecatomb reclassified the sequence to the top-hit (sequence with the lowest e-value).

Assembly and contig abundance calculation. A key feature of Hecatomb is that it completes both a read-based and an assembly-based analysis. For short-reads obtained through technology like Illumina, the first step of the assembly module is individual sample assemblies using MEGAHIT (Li et al. 2015) (Fig 3). Long-reads are not amenable to using short-read assemblers such as megahit and are therefore individual samples are assembled using the Canu assembler (10.1101/gr.215087.116). Contigs resulting from individual sample assemblies are subsequently assembled into a unique set of contigs (herein called a contig dictionary) across the population of samples using Flye (Kolmogorov et al. 2020) . Per sample contig abundance is then calculated by mapping individual sample reads to the contig dictionary. Read counts are normalised to library size and contig length as Sequences Per Million (SPM). This is exactly the same calculation to Transcripts Per Kilobase Million (TPM) other than that the sequences are not assumed to be transcripts, thus the change in nomenclature. Taxonomic assignment of contigs in the contig dictionary is accomplished using MMseqs2 (Steinegger and Söding 2017) queries against all viral genomes in the NCBI Assembly Database (taxID: 10239) (Kitts et al. 2016). Contig properties (e.g.

length, GC-content) are combined with taxonomic assignments and sample abundance estimates (SPM) into a final table. This table is also merged with data obtained through the read based analysis to supplement mapping data with read-based taxonomic assignments and individual read properties (discussed more below in Data Structure).

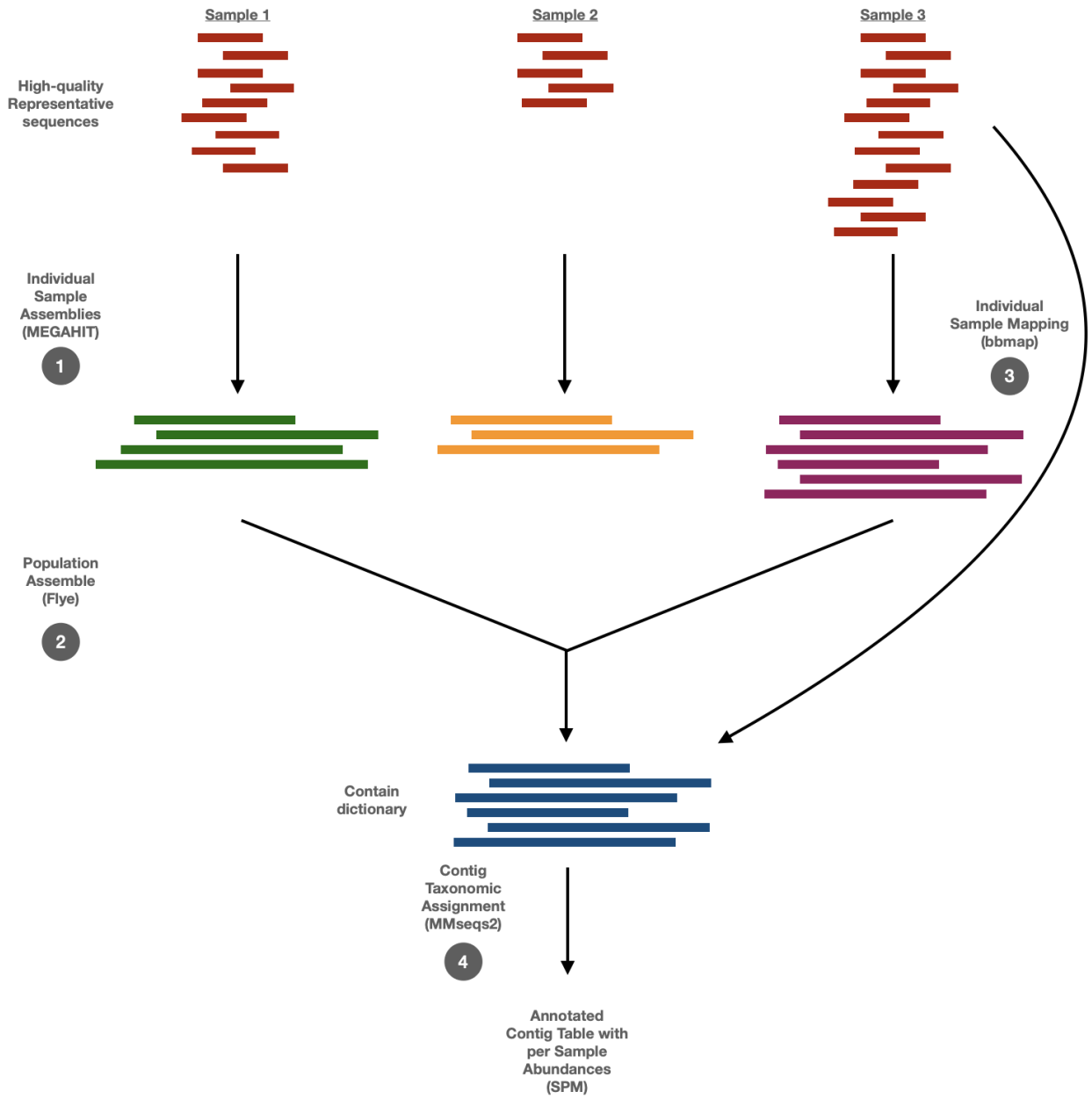

Figure 3: Metavirome Assembly. (1) High-quality representative sequences are from individual samples are assembled using MEGAHIT. **(2)** Individual sample assemblies are assembled into a contig dictionary (a unique set of contigs across a collection of samples) using Flye. **(3)** Sequences from individual samples are mapped to the contig dictionary to calculate Sequences Per kilobase Million (SPM). **(4)** Taxonomy is assigned to contigs via MMSeqs2 queries against a viral genome database.

Data Structure. Hecatomb data structure is designed to acquire, preserve and organise data obtained throughout the pipeline with both study specific sample information as well as external data sources. This combines data obtained through the read-based analysis with data obtained through the assembly processing to make a unified data set accessible to analysis. The tabular format should be accessible by analysts familiar with commonly used research software such as Python, R or Bash and is easily merged with data from external sources, such as viral Baltimore classifications based on Family name or to other external sources.

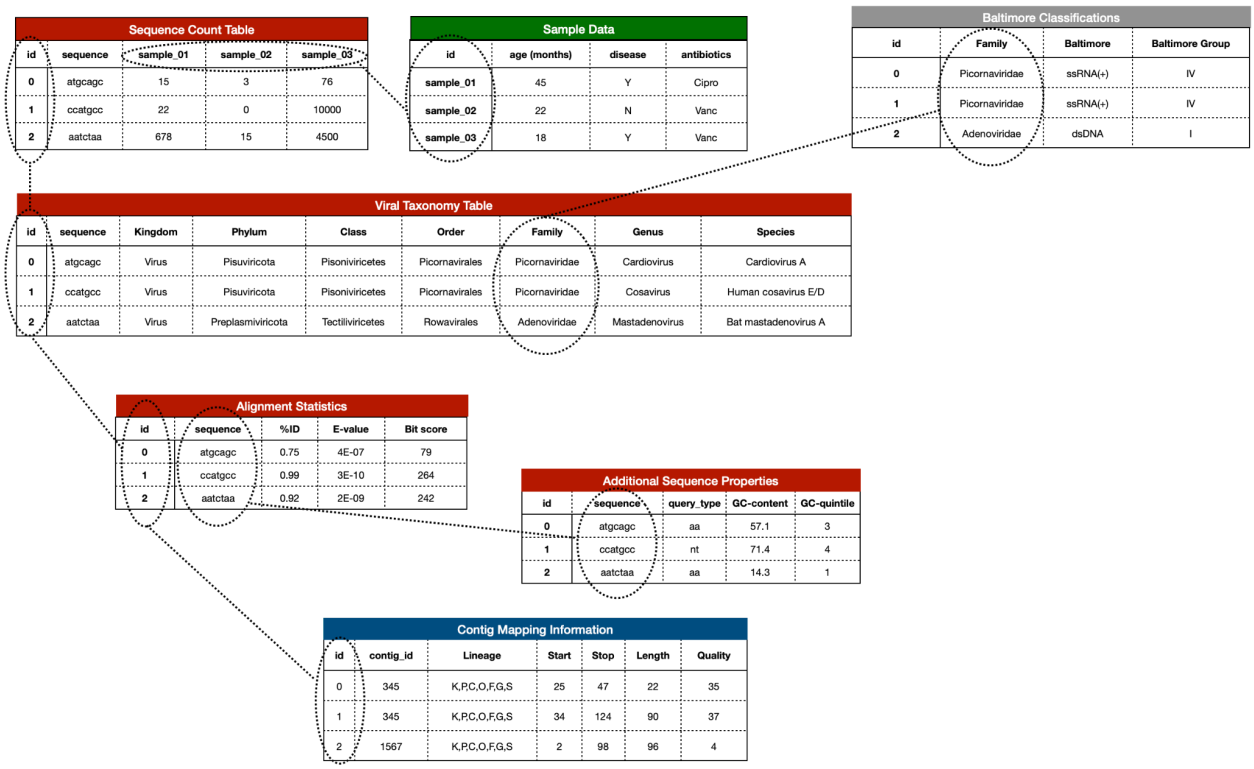

Figure 4: Hecatomb Data Structure. Sequence counts, taxonomy, alignment statistics and additional sequence properties (red) are joined with project sample data (green). Sequence properties, including read-based taxonomic assignments, are joined with mapping information from the assembly pipeline (blue). External data sources can be joined given a common key value (grey).

Availability. Hecatomb is open-source with the project files hosted on GitHub at github.com/shandley/hecatomb, with full support available using GitHub *issues*. Documentation and training vignettes are available at hecatomb.readthedocs.io. Documentation covers installing and configuring the software; detailed information including databases, and output files; advanced options and configuration; an FAQ; and a tutorial covering some example analyses of the results. Hecatomb is

available for installation via Bioconda (Grüning *et al.*, 2018) and is distributed under a permissive MIT licence. Bioconda package information for Hecatomb is available at anaconda.org/bioconda/hecatomb.

RESULTS

Data Assessment Using Taxonomic Binning of Sequence Properties. Hecatomb's data structure (Fig 4) integrates a large amount of information about individual sequences including taxonomic lineage assignments, alignment statistics (e.g. e-values, percent identity, alignment length) and data from external virus information resources (e.g. Baltimore classification). To assess how this data structure can be used to evaluate the content of a complex virome we reanalysed a previously published data set obtained from stool samples collected from simian immunodeficiency virus (SIV) infected rhesus macaques (*Macaca mulatta*) (Handley *et al.*, 2016). This data set was selected as it contains sequences from viruses from multiple Baltimore classifications (RNA and DNA genomes) that infect a variety of cell types (e.g. animal and plant). In addition, the original study identified differences in enteric virus abundances associated with SIV infection, enabling a comparative quantitative benchmark to evaluate Hecatomb with previously published results.

For the reevaluation study, Hecatomb was run using default parameters on sequence data obtained from the full previously published data (95 samples). Sequence data were generated using the Illumina MiSeq 2×250 paired-end protocol on libraries on total nucleic acid (DNA and cDNA to enable detection of both RNA and DNA viruses) extracted from stool samples. Hecatomb's taxonomic assignment approach classified sequences into phylogenetically diverse groups (Fig 5). Bacteriophage from the order Caudovirales, the Siphoviridae, Myoviridae and Podoviridae, were the most abundantly classified viral sequence in the study. Sequences not classified at family level (unclassified Virus family) were also highly abundant. Hecatomb also classified sequences belonging to a diverse set of animal, plant and protist infecting viruses. This data structure enables easy and rapid selection of different viral types for deeper investigation by analysts (Figure S1).

Hecatomb assigns taxonomy using MMseqs2 to query metagenomic sequences to relevant reference sequence databases. Taxonomy is assigned when statistically significant amounts of similarity between query and reference sequence are identified. There are a number of factors that can be used to define 'statistically significant sequence similarity'. For example, e-values can be used to filter potentially false-positive classifications. Low e-value thresholds are more stringent, resulting in fewer false-positive classifications at the sacrifice of increasing false-negatives. High e-values thresholds are more permissive allowing for more potential false-positives while decreasing the number of false-negatives. In practice, setting a low e-value will enable taxonomic assignment to query sequences with high-levels of similarity to sequence in the reference database. Setting a higher, more permissive e-value threshold will still assign taxonomy of query sequences with a high amount of similarity to reference sequences, but will also assign taxonomy to query sequences with low levels of similarity. However, due to the decreased stringency, higher e-value thresholds are more likely to result in false-positive taxonomic assignments to query sequences. Thus an investigator interested in more distant sequence relationships has no option but to select a permissive e-value at the expense of accepting an increased false-positive rate. Hecatomb's approach to dealing with this increased false-positive rate is to collect as much data as possible (e.g. e-values, percent identity, alignment length, etc.) during the taxonomic assignment procedure and organise these data so that an analyst can evaluate both true and false-positive taxonomic classifications.

We evaluated how evaluation of the percent identity and alignment lengths of query sequences to reference sequences can be used to evaluate true and false positive taxonomic assignments for the 4 viral families identified in the original study (Circoviridae, Picornaviridae, Adenoviridae and Parvoviridae)

(Fig 5B). Hecatomb also identified these viral families and assigned taxonomy to a large number of sequences to each family (Fig 5B). Sequences were assigned using both translated queries to amino acid (aa) databases and untranslated queries to nucleotide (nt) databases.

Evaluation of the percent identities and alignment lengths of each individual query in a taxonomic context reveals some important information. A quadrant system can be used to evaluate individual per family (or other taxonomic level) taxonomic assignments. Sequences in the upper two quadrants are highly similar to sequences in the reference databases over short (upper left, Q1) or long (upper right, Q2) while sequences in the lower two quadrants have low similarity over short (lower left, Q3) or long (lower right, Q4) alignment lengths. For this analysis we arbitrarily selected 70% identity to represent the cut-off between low and high-identity for translated (aa database) and 90% identity for untranslated (nt database) alignments. Short and long is arbitrarily set at and 150 bases for both translated and untranslated alignment lengths (Fig 5B). Using this framework it is clear that there are a large number of query sequences with high-identity (both short and long alignments) to sequences in both the aa and nt reference databases for the 4 families of previously identified animal viruses (Fig S2). This indicates that there is evidence for some sequences in the query data set with high-identity to sequences in the reference databases

There were also a large number of query sequences classified as having statistically significant sequence similarity to reference sequences from viruses of protists (Fig 5B). Phycodnaviridae and Mimiviridae are both dsDNA viruses with large genomes. Mimiviridae have been shown to infect *Acanthamoeba*, and Phycodnaviridae infect algae (Sun et al. 2020). While it is conceivable that these viruses may exist in the stool samples of rhesus macaques via water or food, using the quadrant framework it is clear that there is little or no evidence of high-identity alignments to any sequence in either the aa or nt databases (Fig 5B, Fig S2). Importantly, hecatomb does not automatically remove sequences from these families. There is evidence for short and long low identity alignments (quadrant 4) to both Phycodnaviridae and Mimiviridae reference sequences. Thus, if an analyst so chooses they could continue to analyse these sequences using additional metrics (i.e. e-values, abundance across samples, etc.) to determine if these represent potentially novel viral sequences. This would not have been possible using stringent e-value filtering prior to data analysis.

Hecatomb also quantifies the normalised number of sequences (Counts per Million (CPM)) at each taxonomic depth. CPM per sample can be evaluated as the number of sequences assigned to a taxonomy per sample enabling statistical comparisons. The original study found evidence for 4 families of animal viruses ((1) Circoviridae, 2) Picornaviridae, 3) Adenoviridae and 4) Parvoviridae) in stool samples obtained from macaques infected with SIV or uninfected controls. The abundance of sequences from each viral family were similar between SIV-infected and uninfected macaques early in the study, but the abundance increased significantly in SIV-infected animals as infection progressed while remaining the same in uninfected control animals. Evaluation of the CPM for each of these 4 viral families using Hecatomb revealed similar results (Fig 5C).

Interestingly, there were several viral families represented only using untranslated alignment to Hecatombs nucleotide database. One of these, the Herpesviridae, was selected for additional analysis (Fig 6). All of the sequences assigned to the Herpesviridae aligned to only three target GenBank accessions (Fig 6B). One accession dominated, but all three were assigned a taxonomy with very low e-values suggesting statistically significant alignments (Fig 6C). However, all three of these accessions belong to a single type of virus, Stealth virus 1 clone 3B43 (Martin 1999). Stealth virus 1 was originally described as having obtained a number of bacterial sequences resulting in the GenBank genome entry containing sections of both bacterial and viral nucleic acid. Of note, the 3 Stealth virus accessions with significant sequence similarity to the query sequences align to regions with similarity to bacterial

segments when queried against the NCBI nt database (Fig 6D). These sequences are classified as viral in the NCBI and Hecatomb database resulting in their initial classification as viral. Taken together these data suggest that the sequences are more likely bacterial contamination and that there is no evidence of Herpesviridae in any sample.

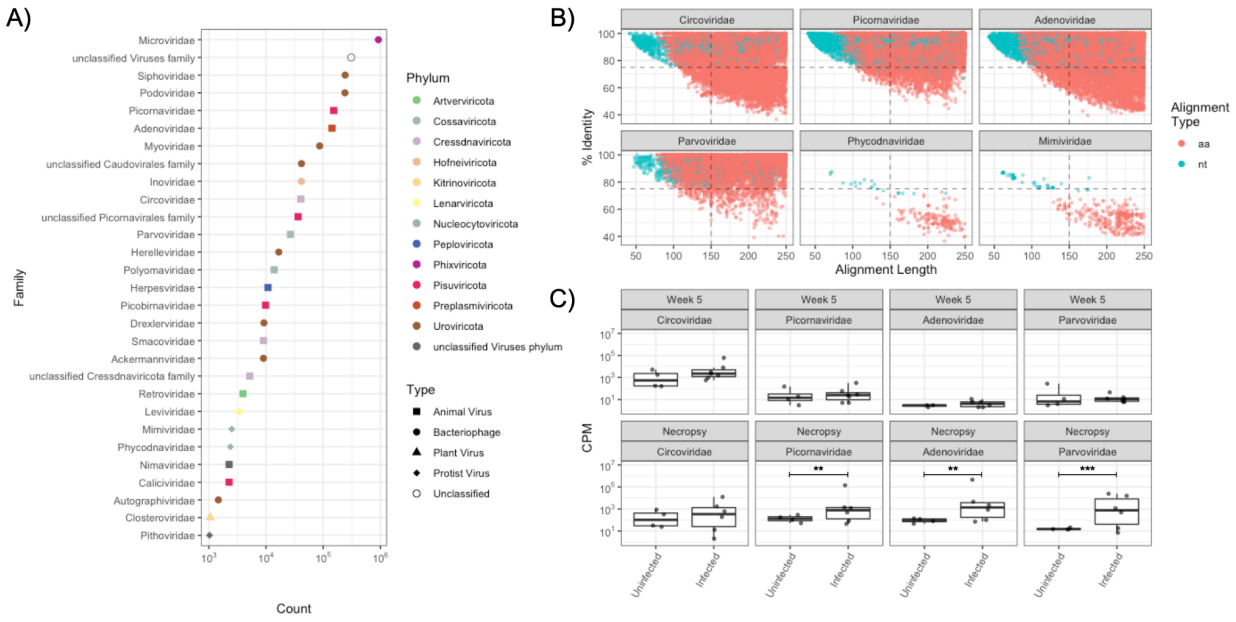

Figure 5: Reanalysis of rhesus macaque stool viromes. (A) Abundance of reads classified by viral Phylum (colour) and Type (shape). Phyla represented by fewer than 1,000 reads were excluded. **(B)** Percent identity and alignment lengths of all sequences classified for the 4 animal viruses identified in the previous study and two viruses of protists. Horizontal (70% identity) and vertical (150 base alignment length) dashed lines indicate a user-defined quadrant space. Each point represents an individual sequence colored by classification method (aa = classified via a translated search to an amino acid database, nt = classified via an untranslated search to a nucleotide database). Panels A and B represent data obtained from all 95 samples in the study. **(C)** Comparison of the number of sequences in SIV-infected and uninfected samples. Significance determined by the Wilcoxon signed-rank test. * = $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, **** $P \leq 0.0001$.

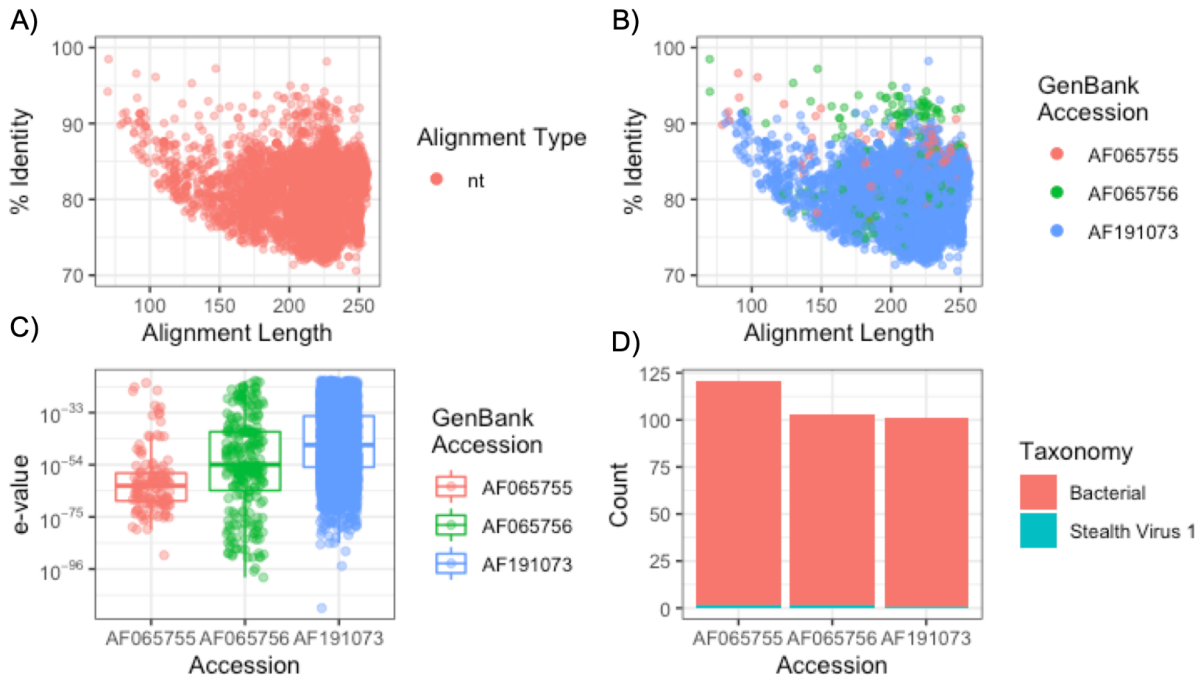

Figure 6: Ambiguous classification of bacterial sequences as Herpesviridae. (A) % identity and alignment length of all sequences assigned to the Herpesviridae. Note, there are no reads that were assigned using a translated search to an amino acid (aa) database. **(B)** Representation of GenBank accessions assigned to the Herpesviridae. **(C)** Summary of e-values for the 3 Herpesviridae accessions. **(D)** Summary counts of the taxonomic hits using BlastN to the NCBI nucleotide (nt) database for each accession.

Analysis of environmental viromes The Hecatomb pipeline is able to process sequence data obtained from any environment for virome analysis. We assessed Hecatomb's ability to analyse non-host associated viromes by processing a previously studied coral reef dataset (Lima *et al.*, 2021) (Lima *et al.*, 2020). This dataset consists of WGS metagenomics sequencing of seawater from inner and outer sections of three different reef systems. The original studies identified significant differences in bacterial compositions between groups.

Hecatomb was run using the fast parameters (--fast) on all samples. A majority of sequences classified as viral over all samples aligned to *Myoviridae*, *Phycodnaviridae*, unclassified Viruses and *Caudovirales* phages (Fig 7A). This is not surprising as bacteria and algae are commonly associated with coral reef and aquatic communities. Using the contigs output from Hecatomb's assembly pipeline we first compared the viromes between inner and outer reef samples (Fig 7C). The previous study found significant differences in bacterial communities between these two sites. Comparison of the viromes (as measured by per sample contig abunandaces) revealed that the viromes of inner and outer reef samples are distinct from one another (Fig 7C). Similarly, the original study also determined that bacterial communities were also significantly different between coral mucosa and the surrounding reef water. Analysis of viral contigs generated by Hecatomb revealed similar differences between coral mucosal and surrounding water viromes (Fig 7D).

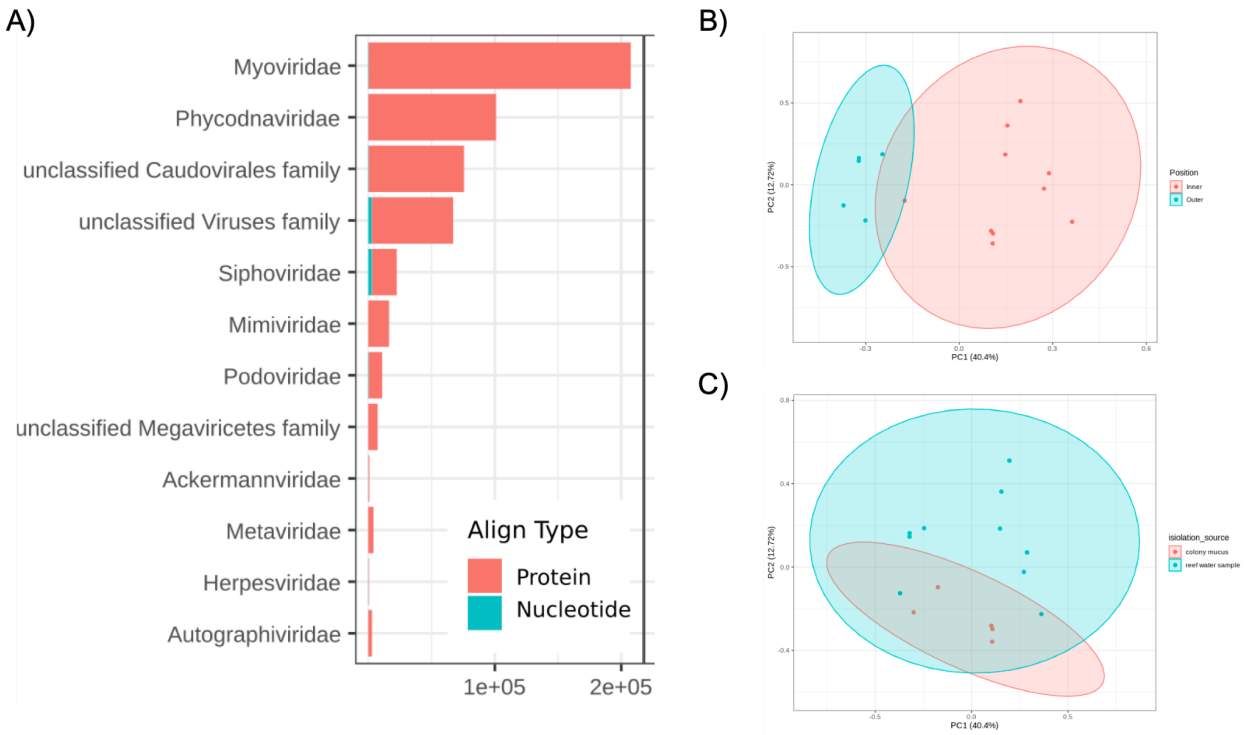

Figure 7: Analysis of environmental viromes. (A) Viral families with more than 2 x 10^3 sequences classified via translated queries to amino acid (aa) or nucleotide (nt) databases. in all coral mucus samples. (**B)** Comparison of viromes in Inner or outer reef samples. **(C)** Comparison of coral mucus to surrounding reef water viromes. Significance was determined using Permutational Analysis of Variance (PERMANOVA). * = P ≤ 0.05, ** P ≤ 0.01, *** P ≤ 0.001, **** P ≤ 0.0001.

DISCUSSION

Virome sequencing has become a commonly used approach to evaluate the viral content of both host-derived and environmental samples. In the broadest terms, virome sequencing is used to answer two related but distinct questions: 1) What viruses are represented in a sample or set of samples or 2) How does virome composition compare between groups of samples. These questions are related in that virome composition is defined by the viruses represented in a sample. However, the answers to these questions can be used to evaluate two completely different biological questions. For example, knowing what viruses are represented in a sample can be useful for identifying etiological agents of infectious disease or novel viruses. In contrast, the compositional analysis of viromes can be used to determine if viral community structure is associated with a disease or environmental state. Both types of studies are dependent on effective computational tools not only to identify and classify viral reads within a metagenome, but also to assist in interpretation of complex virome data in association with sample data.

Virome analysis is almost completely reliant on sequence similarity searches. There are two primary approaches to classifying viral sequences from metagenomic data. The first is ‘brute force’ wherein all unclassified sequences are queried against a comprehensive, multi-kingdom reference sequence database (e.g. NCBI nt or nr). This approach relies on the search algorithm (e.g. Blast, DIAMOND) to pick the best or lowest-common ancestor of a group of hits (defined by statistical thresholds) to provide a final taxonomic assignment to an unknown query sequence. Hecatomb takes a different approach by first

capturing all ‘potentially viral’ sequences by querying sequences against a relatively small viral sequence database. These ‘potentially viral’ sequences typically represent only a small fraction of the full metagenomic data. To confirm taxonomic provenance, all potentially viral sequences are cross-checked against a clustered and relatively small transkingdom reference database. Hecatomb completes this iterative search approach using translated searches against amino acid databases as well as untranslated searches against nucleotide databases, combining the results of each. This iterative search strategy uses databases orders of magnitude smaller than comprehensive, trans-kingdom databases making it available to run on commodity hardware.

The design philosophy of Hecatomb recognizes that there are no ‘perfect’ databases or search algorithms. Both the brute force and iterative search approaches are valid and result in different rates of true/false positives/negatives. Instead, Hecatomb relies on providing a compiled and rich set of data for analyst use to evaluate search results. We used this strategy to reassess the virome composition of SIV-infected and uninfected macaques (Handley et al. 2016). The original study used an iterative approach, but relied on comprehensive, transkingdom databases (NCBI nt and nr) and identified associations between 4 families of animal viruses (Circoviridae, Picornaviridae, Adenoviridae and Parvoviridae) and SIV-infection. Hecatomb identified the same four viral families as well as the relationship to SIV mediated disease (Fig 5C). Similar to our analysis of these samples using Hecatomb, the original study also classified a number of sequences to the Mimiviridae and Phycodnaviridae. Statistical comparison of these sequences between groups (e.g. SIV-infected vs. uninfected) did not reveal any significant associations thus they were not discussed further. However, evaluation of results from Hecatomb indicates that these were likely false-positive classifications (Fig 5B). This highlights how coordinated data such as alignment statistics and taxonomy can be a powerful tool for virome evaluation.

Importantly, we were also able to evaluate the viromes of environmental (non-host associated) viromes. This analysis was primarily designed to identify compositional changes in viromes between reef types (inner or outer) and within coral mucosa or the surrounding water from a previously published metagenomic data set (Fig 7) (Lima et al. 2020; Lima et al. 2021). The previous study demonstrated that the metagenomes of coral surface mucus layer largely reflects the composition of the surrounding reef water. However, when samples taken from inner patch reefs (where large thermal fluctuations occur) were compared to outer patch reefs (which have greater thermal stability), significant differences were observed.This dataset serves as an interesting benchmark for reanalysis with Hecatomb, as we can explore how the viral populations in these samples compare to their microbial profiles. The viral families that were most abundant in the coral dataset were consistent with the identified core virome of corals described in previous studies (Correa et al. 2016; Wood-Charlson et al. 2015). Compositional analysis of viral contigs created and classified by hecatomb indicate significant compositional differences between inner and outer reef viral communities as well as between coral colony mucosal regions and the surrounding reef water.

Virome analysis, regardless if the goals of the study are to identify known or novel viruses or to characterise viral compositional changes between groups requires extensive computational and analytical work. Hecatomb was designed to provide an accessible ‘front end’ to virome analysis. This front end relies on reduced representation sequence databases and an iterative search strategy to classify viral sequences within a metagenome. Data emerging from these iterative searches is maintained and coordinated into easily accessible output tables for detailed analysis which can be used to easily parse data as well as to evaluate true and false positive viral taxonomic classifications.

DATA AVAILABILITY

The reanalysis with Hecatomb utilised pre-existing datasets which are available under the NCBI BioProject accessions PRJEB9503 for the macaque SIV dataset (Handley et al. 2016) and PRJNA595374 for the coral reef dataset (Lima et al. 2021). The Hecatomb annotations are available at doi.org/10.5281/zenodo.6388251, and and all commands used for analysing the results are available at https://gist.github.com/beardymcjohnface/58b127ea7ed536f20daa0ff11a224096 and https://gist.github.com/beardymcjohnface/e6b228572ccfbe89f0d9938f58ee2817.

FUNDING SUPPORT

Research reported in this publication was supported by the National Institute Of Diabetes And Digestive And Kidney Diseases of the National Institutes of Health under Award Number RC2DK116713. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

ACKNOWLEDGEMENTS

The authors would like to thank Chandni Desai, Barry Hykes and Anne Paredes for their thoughtful commentary regarding the design philosophy of Hecatomb.

REFERENCES

Barouch, D.H. *et al.* (2015) 'Protective efficacy of adenovirus/protein vaccines against SIV challenges in rhesus monkeys', *Science*, 349(6245), pp. 320–324.

Bhaijee, F. *et al.* (2011) 'Human immunodeficiency virus-associated gastrointestinal disease: common endoscopic biopsy diagnoses', *Pathology research international*, 2011, p. 247923.

Bolourian, A. and Mojtahedi, Z. (2018) 'Streptomyces, shared microbiome member of soil and gut, as "old friends" against colon cancer', *FEMS microbiology ecology*, 94(8). doi:10.1093/femsec/fiy120.

Brown, P.S. *et al.* (2014) 'Directly sampling the lung of a young child with cystic fibrosis reveals diverse microbiota', *Annals of the American Thoracic Society*, 11(7), pp. 1049–1055.

Bushnell, B. (2014) *BBMap: A fast, accurate, splice-aware aligner*. LBNL-7065E. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). Available at: https://www.osti.gov/biblio/1241166 (Accessed: 11 November 2021).

Camarillo-Guerrero, L.F. *et al.* (2021) 'Massive expansion of human gut bacteriophage diversity', *Cell*, 184(4), pp. 1098–1109.e9.

Chen, S. *et al.* (2018) 'fastp: an ultra-fast all-in-one FASTQ preprocessor', *Bioinformatics* , 34(17), pp. i884–i890.

Feeney, A. *et al.* (2014) 'Cronobacter sakazakii: stress survival and virulence potential in an opportunistic foodborne pathogen', *Gut microbes*, 5(6), pp. 711–718.

Fernandes, M.A. *et al.* (2019) 'Enteric Virome and Bacterial Microbiota in Children With Ulcerative Colitis and Crohn Disease', *Journal of pediatric gastroenterology and nutrition*, 68(1), pp. 30–36.

Finkbeiner, S.R. *et al.* (2009) 'Human stool contains a previously unrecognized diversity of novel astroviruses', *Virology journal*, 6, p. 161.

Grüning, B. *et al.* (2018) 'Bioconda: sustainable and comprehensive software distribution for the life

sciences', *Nature methods*, 15(7), pp. 475–476.

Hanauer, D.I. *et al.* (2017) 'An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning', *Proceedings of the National Academy of Sciences of the United States of America*, 114(51), pp. 13531–13536.

Handley, S.A. *et al.* (2012) 'Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome', *Cell*, 151(2), pp. 253–266.

Handley, S.A. *et al.* (2016) 'SIV Infection-Mediated Changes in Gastrointestinal Bacterial Microbiome and Virome Are Associated with Immunodeficiency and Prevented by Vaccination', *Cell host & microbe*, 19(3), pp. 323–335.

Hendrix, R.W. *et al.* (1999) 'Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage', *Proceedings of the National Academy of Sciences of the United States of America*, 96(5), pp. 2192–2197.

Hoyle, N. *et al.* (2018) 'Phage therapy against Achromobacter xylosoxidans lung infection in a patient with cystic fibrosis: a case report', *Research in microbiology*, 169(9), pp. 540–542.

Jaradat, Z.W. *et al.* (2014) 'Cronobacter spp.--opportunistic food-borne pathogens. A review of their virulence and environmental-adaptive traits', *Journal of medical microbiology*, 63(Pt 8), pp. 1023–1037.

Kankwatira, A.M., Mwafulirwa, G.A.K. and Gordon, M.A. (2004) 'Non-typhoidal salmonella bacteraemia--an under-recognized feature of AIDS in African adults', *Tropical doctor*, 34(4), pp. 198–200.

Kolmogorov, M. *et al.* (2019) 'Assembly of long, error-prone reads using repeat graphs', *Nature biotechnology*, 37(5), pp. 540–546.

Koren, S. *et al.* (2017) 'Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation', *Genome research*, 27(5), pp. 722–736.

Köster, J. and Rahmann, S. (2012) 'Snakemake—a scalable bioinformatics workflow engine', *Bioinformatics* [Preprint]. Available at: https://academic.oup.com/bioinformatics/article-abstract/28/19/2520/290322.

Krishnamurthy, S.R. and Wang, D. (2017) 'Origins and challenges of viral dark matter', *Virus research*, 239, pp. 136–142.

Lamps, L.W. (2007) 'Infective disorders of the gastrointestinal tract', *Histopathology*, 50(1), pp. 55–63.

LaVergne, S. *et al.* (2018) 'Phage Therapy for a Multidrug-Resistant Acinetobacter baumannii Craniectomy Site Infection', *Open forum infectious diseases*, 5(4), p. ofy064.

Li, D. *et al.* (2015) 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics* , 31(10), pp. 1674–1676.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* , 25(16), pp. 2078–2079.

Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics* , 34(18), pp. 3094–3100.

Lima, L.F.O. *et al.* (2020) 'Modeling of the Coral Microbiome: the Influence of Temperature and Microbial Network', *mBio*, 11(2). doi:10.1128/mBio.02691-19.

Lima, L.F.O. *et al.* (2021) 'Coral and Seawater Metagenomes Reveal Key Microbial Functions to Coral

Health and Ecosystem Functioning Shaped at Reef Scale'.

Lim, Y. *et al.* (2012) 'The Microbiome and Virome of CF', *Pediatric pulmonology*, 47, pp. 321–322.

Marchiori, E. *et al.* (2005) 'Rhodococcus equi pneumonia in AIDS: high-resolution CT findings in five patients', *The British journal of radiology*, 78(933), pp. 783–786.

Muhlebach, M.S. *et al.* (2018) 'Initial acquisition and succession of the cystic fibrosis lung microbiome is associated with disease progression in infants and preschool children', *PLoS pathogens*, 14(1), p. e1006798.

Mushegian, A.R. (2020) 'Are There 1031 Virus Particles on Earth, or More, or Fewer?', *Journal of bacteriology*, 202(9). doi:10.1128/JB.00052-20.

Nayfach, S. *et al.* (2021) 'Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome', *Nature microbiology*, 6(7), pp. 960–970.

Nkenfou, C.N. *et al.* (2019) 'Evaluation of the effect of HIV virus on the digestive flora of infected versus non infected infants', *The Pan African medical journal*, 34, p. 24.

Ponsero, A.J. and Hurwitz, B.L. (2019) 'The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes', *Frontiers in microbiology*, 10, p. 806.

Roach, M. *et al.* (2021) 'No Evidence Known Viruses Play a Role in the Pathogenesis of Onchocerciasis-Associated Epilepsy. An Explorative Metagenomic Case-Control Study', *Pathogens*, 10(7). doi:10.3390/pathogens10070787.

Rosseel, T. *et al.* (2014) 'False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis', *Transboundary and emerging diseases*, 61(4), pp. 293–299.

Sartor, R.B. and Wu, G.D. (2017) 'Roles for Intestinal Bacteria, Viruses, and Fungi in Pathogenesis of Inflammatory Bowel Diseases and Therapeutic Approaches', *Gastroenterology*, 152(2), pp. 327–339.e4.

Shen, W. *et al.* (2016) 'SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation', *PloS one*, 11(10), p. e0163962.

Shen, W. and Ren, H. (2021) 'TaxonKit: A practical and efficient NCBI taxonomy toolkit', *Journal of genetics and genomics = Yi chuan xue bao*, 48(9), pp. 844–850.

Skewes-Cox, P. *et al.* (2014) 'Profile hidden Markov models for the detection of viruses within metagenomic sequence data', *PloS one*, 9(8), p. e105067.

Steinegger, M. and Söding, J. (2017) 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nature biotechnology*, 35(11), pp. 1026–1028.

Vandeplassche, E. *et al.* (2019) 'Influence of the lung microbiome on antibiotic susceptibility of cystic fibrosis pathogens', *European respiratory review: an official journal of the European Respiratory Society*, 28(152). doi:10.1183/16000617.0041-2019.

Wallace, M.R. and Brann, O.S. (2000) 'Gastrointestinal manifestations of HIV infection', *Current gastroenterology reports*, 2(4), pp. 283–293.

Wang, W. *et al.* (2015) 'Metagenomic analysis of microbiome in colon tissue from subjects with inflammatory bowel diseases reveals interplay of viruses and bacteria', *Inflammatory bowel diseases*, 21(6), pp. 1419–1427.

Waters, E.M. *et al.* (2017) 'Phage therapy is highly effective against chronic lung infections with

Pseudomonas aeruginosa', *Thorax*, 72(7), pp. 666–667.

Zou, Y. *et al.* (2019) '1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses', *Nature biotechnology*, 37(2), pp. 179–185.

SUPPLEMENTAL INFORMATION

Databases. All Hecatomb databases are hosted in the cloud with Amazon AWS. Downloading and installing them is completely managed by the Hecatomb launcher following initial installation. The most up to date descriptions of the databases can be obtained from Hecatomb's documentation at hecatomb.readthedocs.io.

1. **Contaminants.** This database consists of a collection of NEBNext and TruSeq sequencing adapters, primers, and vector contaminants from UniVec. The contaminants database is used exclusively during sequence preprocessing.
2. **Hosts.** A collection of host genomes that have been preprocessed to mask viral-like and low-entropy sequences. Host genomes are used for host read removal during preprocessing. Hecatomb comes with several common host genomes ready to use, and users can add their own host genomes to the database via the Hecatomb launcher.
3. **AA and NT.** These are the amino acid (AA) and nucleotide (NT) databases used for sequence annotation of both reads and contigs. For each of the AA and NT databases, there is a primary viral database used for identifying reads that match a known virus, and a secondary multi-kingdom database which is used for assigning taxonomy to either reads and contigs. The primary AA database includes all UniProt viral protein entries clustered at 99% identity. The secondary AA database consists of the Uniclust50 database (Mirdita et al. 2017) (https://doi.org/10.1093/nar/gkw1081) supplemented with the primary AA database. The primary NT database consists of all viral sequences in GenBank clustered at 100% identity to remove redundancy. The secondary NT database consists of a customised polymicrobial nucleotide database containing representative RefSeq genomes from Bacteria (n = 14,933), Archaea (n = 511), Fungi (n = 423), Protozoa (n = 90) and plant (n = 145) genomes.
4. **Tax.** This is NCBI's taxonomy database used for translating taxonomy IDs into full taxonomic lineages.
5. **Tables.** These are additional classification tables such as the ICTV 2019 Baltimore groups for viruses, which are used to supplement the output annotation tables.

Outputs. The most up to date descriptions of all the pipeline output files can be obtained from Hecatomb's documentation at hecatomb.readthedocs.io. A tutorial is included to demonstrate typical analyses based on these output files using both R or Python. The main output files are described:

1. **Report**. Hecatomb utilises a Snakemake feature to generate a report (*report.html*) of the pipeline run. The report includes the default rule graph of the pipeline, details about resource usage and runtime for each rule, and the configuration file used. The report also includes read counts generated following every preprocessing and annotation step and renders this as a Sankey diagram to visualise the fate of all reads.
2. **Seqtable**. The sequence table (*seqtable.fasta*) contains all the representative sequences for the clustered reads for all samples. The sample names and cluster counts are incorporated into the sequence IDs whilst remaining in a standard sequential multi-fasta format.
3. **Read annotations**. The read annotations table (*bigtable.tsv*) forms the basis for the majority of analyses that would be performed following analysis with Hecatomb. This table combines the sequence and sample IDs, cluster counts, alignment metrics, and taxonomic annotations into one table and is designed to expedite downstream analysis of the read annotations.
4. **Assembly**. The population assembly (*assembly.fasta*) is the output of Flye during the assembly process. Resulting contigs are annotated using MMSeqs queries against the NCBI Viral Assembly Database.

5. **Contig annotations**. The contig annotations (*contigSeqTable.tsv*) are generated by combining the read annotations with their mapping coordinates to the assembly. This format allows the user to fine tune the final annotations of contigs, as well as assign taxonomy to reads that were not originally annotated based upon the contig annotations.

Dependencies. Software dependencies are summarised in Table S1. While the pipeline depends on all of these packages, users are only required to manually install Conda. Once Conda is installed, Hecatomb and its dependencies can be installed and updated via Conda.

Table S1. Software packages used by Hecatomb.

Package name	**Current Version**	**URL**	**Reference**
Conda	4.10.3	conda.io	
Mamba	0.17.0	github.com/mamba-org/mamba	
Python3	3.10.0	python.org	
Snakemake	6.4.1	snakemake.github.io	(Köster and Rahmann, 2012)
fastp	0.23.2	github.com/OpenGene/fastp	(Chen *et al.*, 2018)
PySam	0.17.0	github.com/pysam-developers/pysam	
BBTools	38.90	jgi.doe.gov/data-and-tools/bbtools	(Bushnell, 2014)
Megahit	1.2.9	github.com/voutcn/megahit	(Li *et al.*, 2015)
Canu	2.2	github.com/marbl/canu	(Koren *et al.*, 2017)
Flye	2.7.1	github.com/fenderglass/Flye	(Kolmogorov *et al.*, 2019)
Minimap2	2.20	github.com/lh3/minimap2	(Li, 2018)
Samtools	1.13	github.com/samtools/samtools	(Li *et al.*, 2009)
MMSeqs2	12.113e3	github.com/soedinglab/MMseqs2	(Steinegger and Söding, 2017)
TaxonKit	0.8.0	github.com/shenwei356/taxonkit	(Shen and Ren, 2021)
Plotly	5.3.1	plotly.com	
Kaleido	0.2.1	pypi.org/project/kaleido	
SeqKit	0.16.1	github.com/shenwei356/seqkit	(Shen *et al.*, 2016)

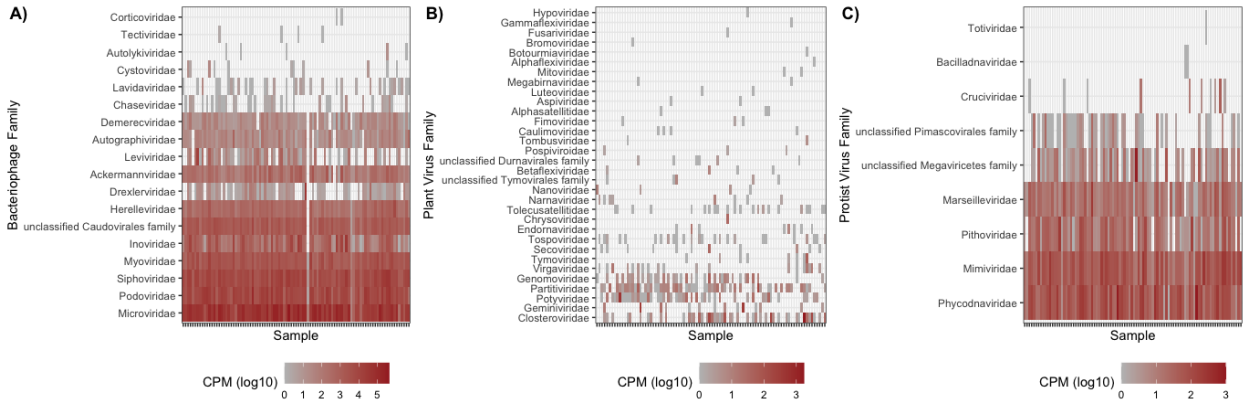

Figure S1: Taxonomic subsets of virus types. Viral families present in the 95-sample SIV reanalysis study **(A)** Bacteriophage, **(B)** Plant viruses, and **(C)** Protist viruses.

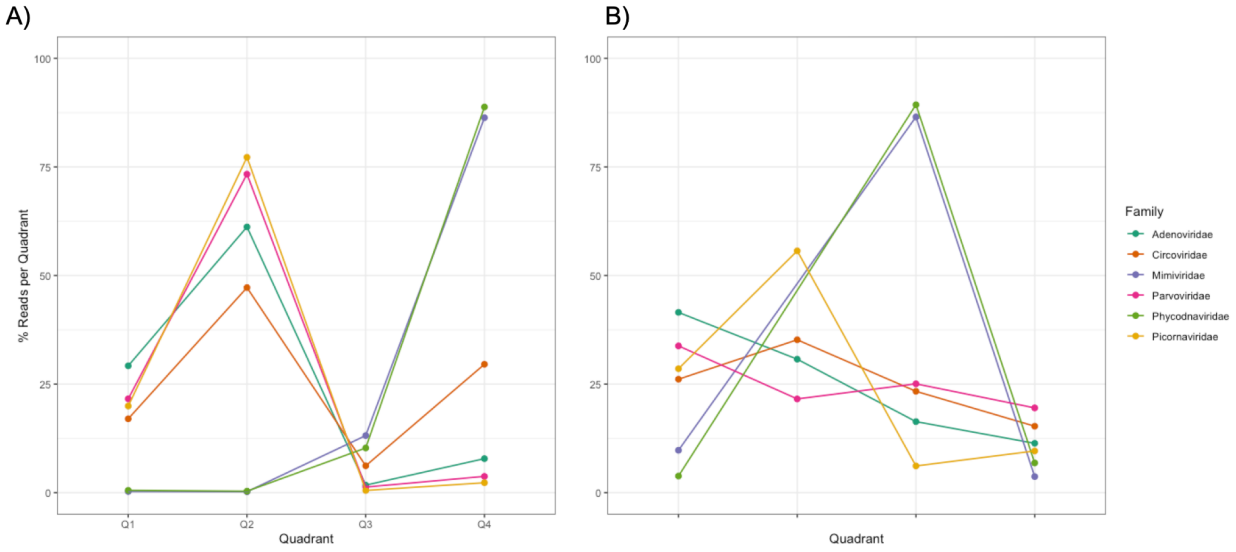

Figure S2: Sequence per Quadrant Evaluation. Percentage of reads per quadrant in Figure 5. **(A)** translated (aa reference database) and **(B)** untranslated (nt reference database)