DiscoVir: A Comprehensive Automated Web-Based Virome Analysis Pipeline

Lauren E. Krausfeldt[1], Poorani Subramanian[1], Duc Doan[1], Kathryn McCauley[1], Michael Dolan[1], and Mariam Quiñones[1]

[1] Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA

Introduction

Viruses impact microbial diversity, abundance, fitness, phenotype, and microbial relationships while also influencing host evolution and gene transfer[1]. Thus, studying the viral component of a microbial community is crucial to understanding microbial community dynamics. Viral metagenomics is the primary method to study the virome of host-associated and environmental microbiomes.

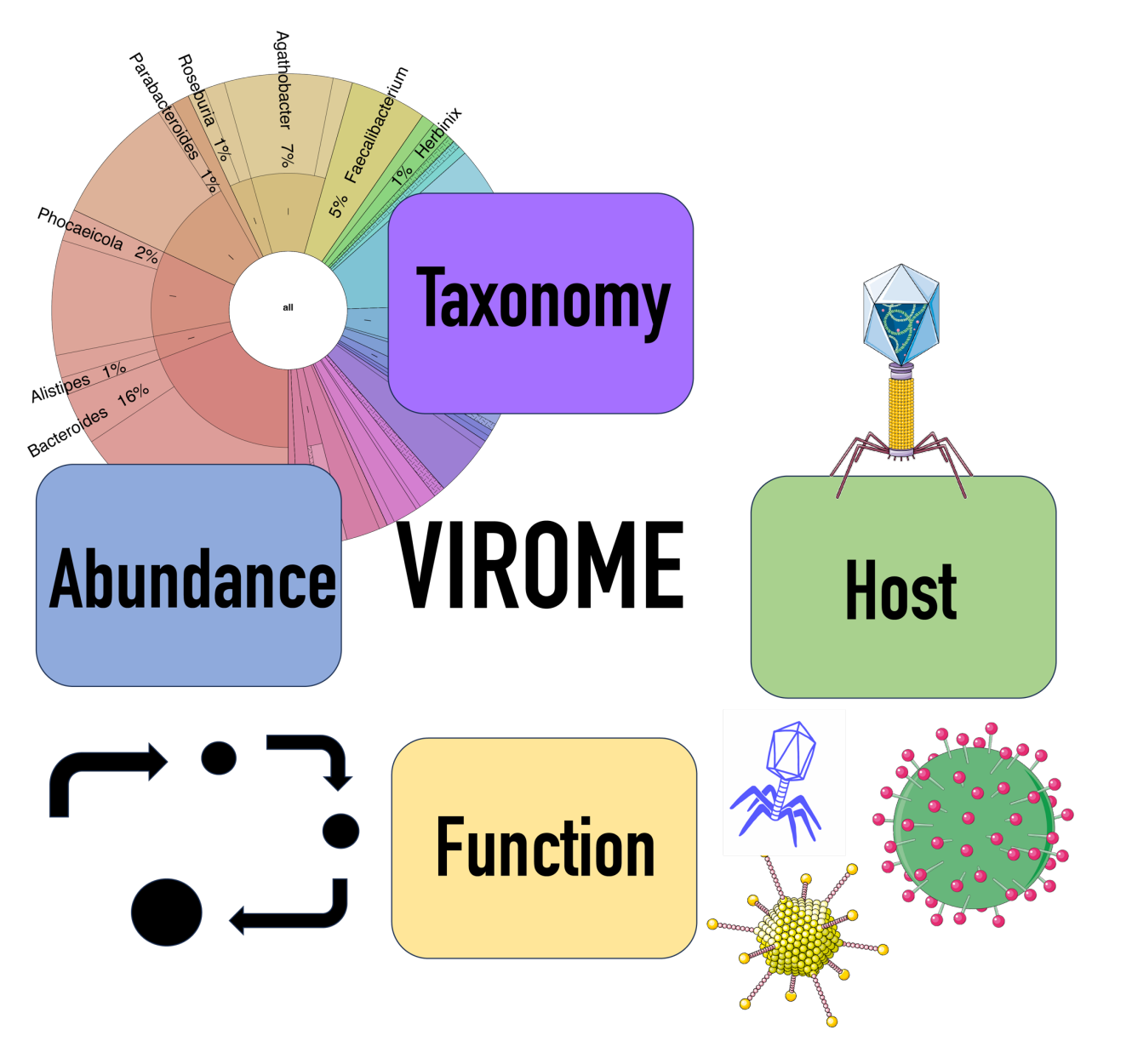

Figure 1. Overview of the scope and functionality of DiscoVir.

Numerous bioinformatics tools specific to viral analysis are available that enable viral discovery, taxonomic annotation, and functional classification of viral communities. However, these tools often require extensive computational resources and bioinformatics skills. DiscoVir makes viral metagenomics analysis easy for any researcher by combining the most comprehensive and popular viral analysis tools into one streamlined pipeline. DiscoVir is available in Nephele[2], NIAID's web application for microbial -omics analysis, which is free for public use! Nephele is easy to use and accessible for researchers with all levels of bioinformatics skills.

Methods

DiscoVir was written in Snakemake and includes viral analysis tools and custom python scripts for estimating abundances, generating heatmaps, and Krona plots. All code is publicly available so that it can be accessed locally and used on any HPC at github.com/niaid/virome-pipeline. The inputs to DiscoVir are metagenomic assemblies and .bam files. You can obtain these from external pipelines or by running your data through the WGSA2[3] pipeline in Nephele!

Scan here to check out Nephele!

nephele.niaid.nih.gov

The pipeline

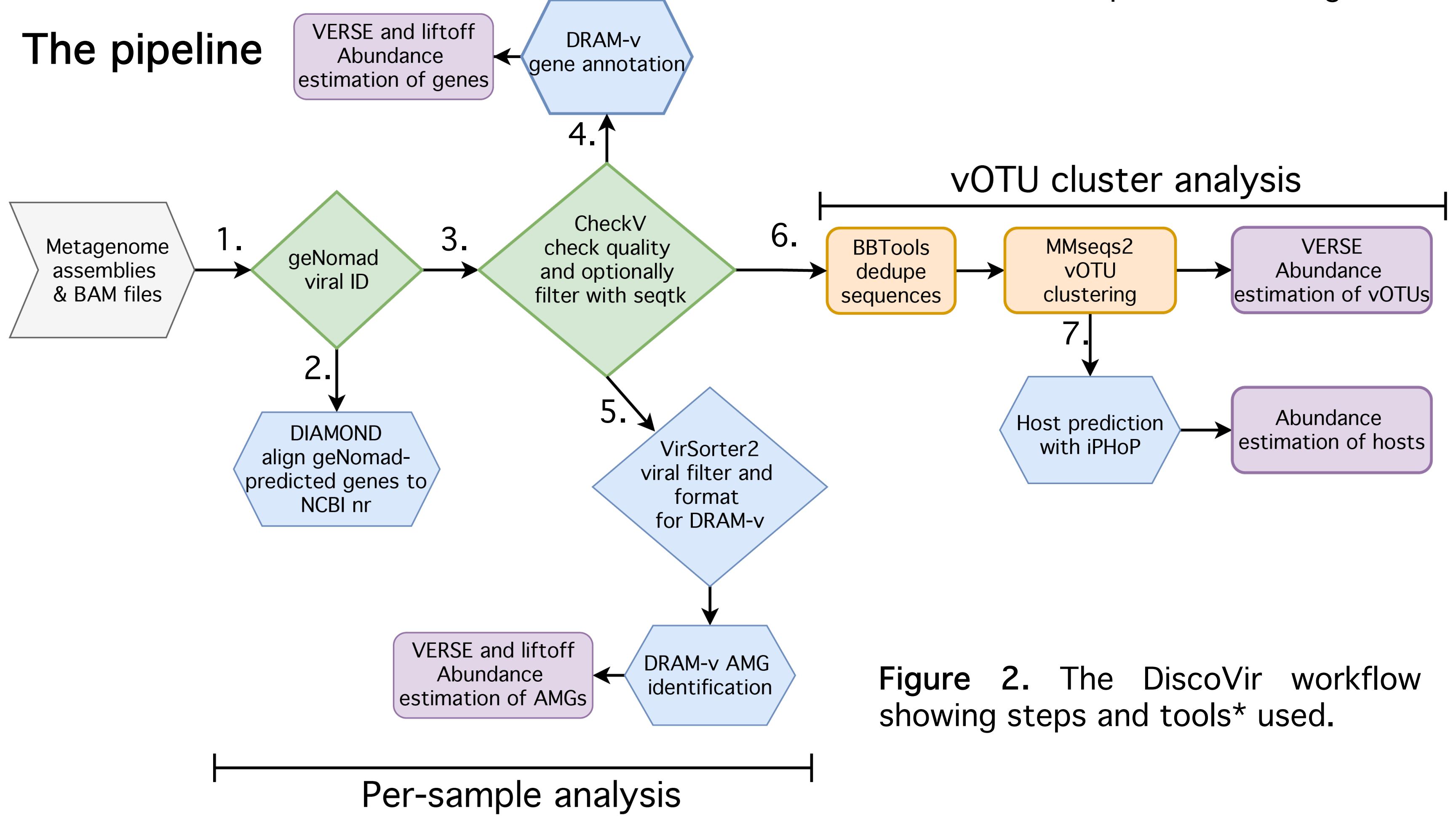
Metagenome assemblies & BAM files
1.
geNomad viral ID
2.
DIAMOND align geNomad-predicted genes to NCBI nr
3.
CheckV check quality and optionally filter with seqtk
4.
DRAM-v gene annotation
VERSE and liftoff Abundance estimation of genes
5.
VirSorter2 viral filter and format for DRAM-v
DRAM-v AMG identification
VERSE and liftoff Abundance estimation of AMGs
Per-sample analysis
6.
vOTU cluster analysis
BBTools dedupe sequences
MMseqs2 vOTU clustering
VERSE Abundance estimation of vOTUs
7.
Host prediction with iPHoP
Abundance estimation of hosts

Figure 2. The DiscoVir workflow showing steps and tools* used.

1. Assemblies are screened for viral genomes and prophage are trimmed for any bacterial contamination. Viral genomes are also taxonomically annotated.
2. All viral genes are annotated with NCBI's nr database.
3. Viral genomes are filtered for quality based on completeness.
4. Additional functional annotation is optionally performed on viral genes to obtain VOGIDs, Pfams, and KO IDs on quality filtered genomes and abundances are calculated.
5. Auxiliary metabolic genes (AMGs) are optionally identified and abundances are calculated.
6. vOTUs are generated by clustering viral genomes from all samples using MIUVIG[4] guidelines and abundances are calculated
7. Phage hosts are predicted and abundances are calculated.

Results

DiscoVir enables viral taxonomic and functional diversity analysis by producing abundance matrices for vOTUs (Fig. 3), hosts, functional genes, and auxiliary metabolic genes. Visualizations are also produced (Fig. 4).

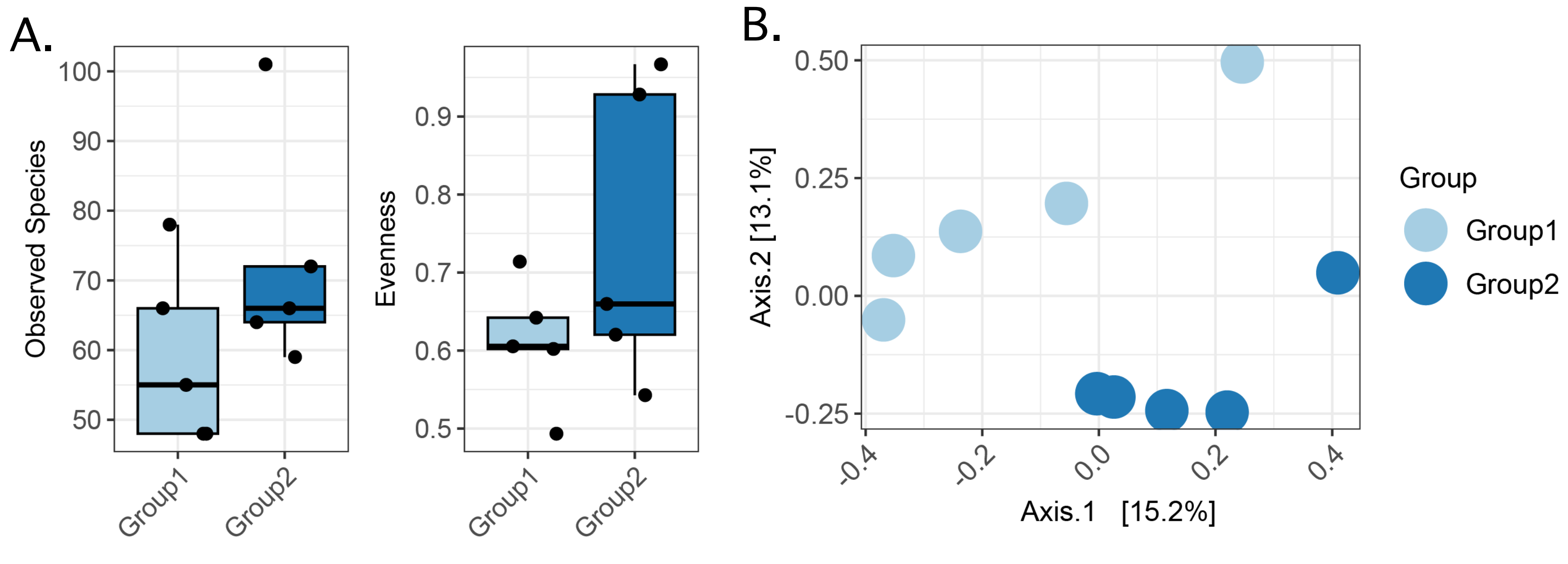

Figure 3. Example downstream analyses enabled by DiscoVir to explore diversity (A) and viral community composition (B).

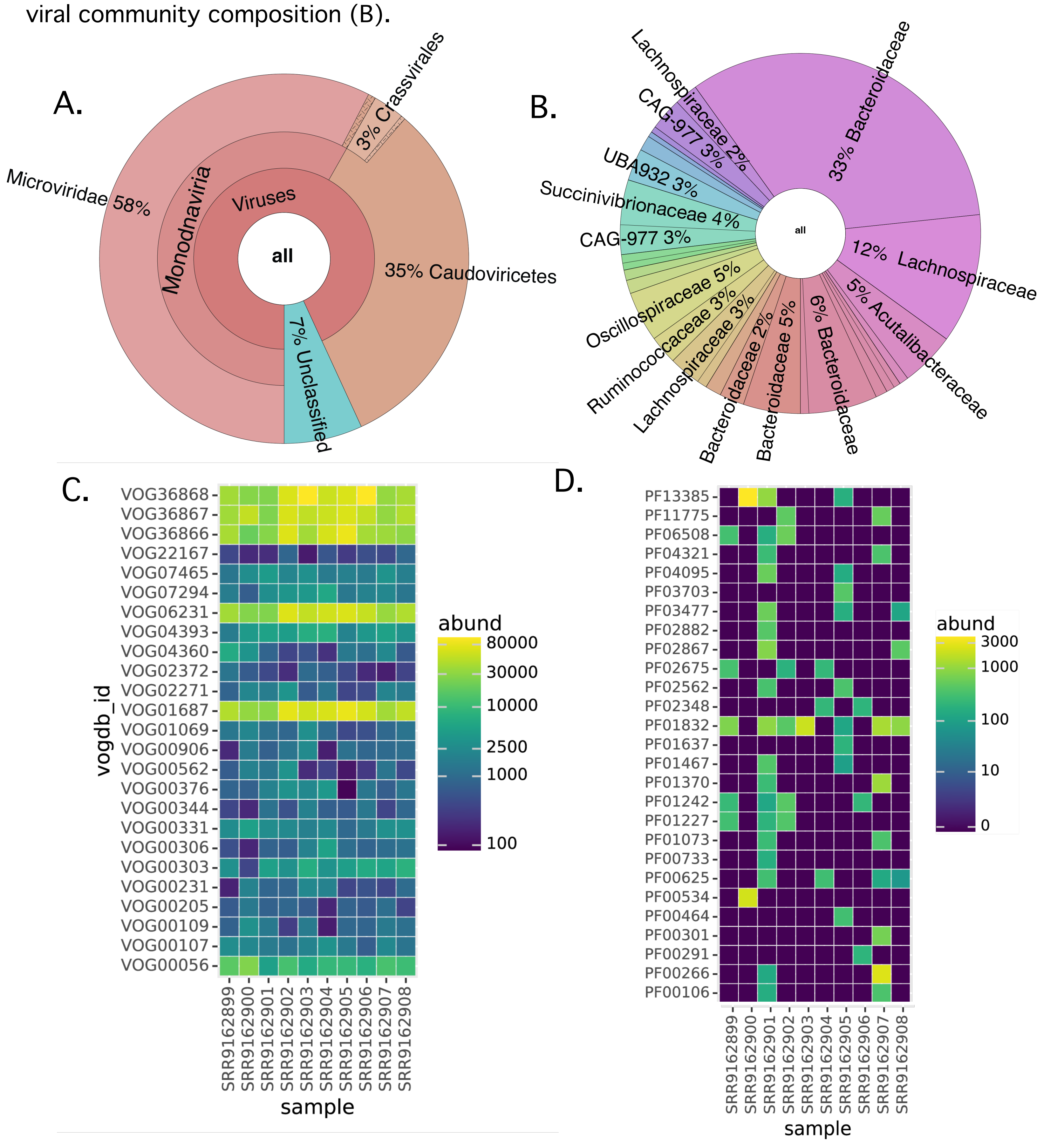

Figure 4. Outputs from DiscoVir. A. Krona plots showing viral taxonomy. B. Krona plot showing phage host taxonomy. C. Heatmap of VOGIDs identified in virome. D. Heatmap of AMGs identified in virome.

Benchmarking

Time and memory requirements were evaluated with different sized datasets in DiscoVir using default parameters (Table 1). Clustering methods were compared for efficiency and accuracy with mock viral sequences with alternative methods: dRep, BLAST, ClusterONE, and CD-HIT. Additionally, DiscoVir's functionality was compared to other viral pipelines in the literature and ViWrap[5], a comparable pipeline, was run and compared with DiscoVir.

Table 1. DiscoVir stats with different data sizes. Size represents total input.

	Type	Size (GB)	# of samples	Time (hh:mm)	Memory (GB)
Dataset 1	metagenome	0.64	10	4:00	155
Dataset 2	metatranscriptome	3.47	16	6:32	135
Dataset 3	metagenome	5.62	10	12:43	164
Dataset 4	metagenome	12.90	8	14:28	164
Dataset 5	metatranscriptome	13.14	16	6:09	113
Dataset 6	metagenome	43.44	24	40:40	163
Dataset 7	metagenome	114.6	49	94:44	164

Clustering

- *ClusterONE*: no clear equivalent to ANI or coverage to meet MIUVIG guidelines.
- *dRep:* Primary clustering with MASH considers Ns when calculating % identity. Considers 100% identical genomes as individuals if alignment is less than value assigned.
- *CD-HIT and vsearch*: too slow but work, not efficient for large datasets.
- *BLAST ANI method:* similar in efficiency and accuracy to DiscoVir's method.

Table 2. Comparison of DiscoVir to ViWrap.

	DiscoVir	ViWrap
Inputs	short read multiple assemblies	long and short read one assembly
Filtering by score/quality	yes	no
Binning contigs	no	yes
vOTUs	bbtools and mmseqs2	vConTACT2 and dRep
MIUVIG	yes	no
Function	geNomad + nr + DRAM-v + AMGs	geNomad or KO AMGs
Host prediction	yes	yes
Outputs	Abundances and visualizations	Abundances and visualizations
Abundances	reads per bp per M	read per 100M
Time (test data[5]; hh:mm)	01:44 (t=3-16)	05:15 (t=16)
Max memory (GB)	57GB	85GB
Available for HPC	yes	yes
Web app	yes!	no

* In Zhou et al (5), using Virsorter2 and Vibrant time was reported as 14 h with threads (t) = 20

Discussion

Tools/methodology

- More tool options in ViWrap for viral discovery, but more flexibility in parameters and filtering in DiscoVir.
- Functional annotations in DiscoVir are comprehensive
- Efficient and reproducible clustering method following MIUVIG guidelines.

Efficiency

- Inputs to DiscoVir less than ~20GB will finish in less than 24 hours.
- DiscoVir is ~ 3 times faster to run than ViWrap.
- Comprehensive and collated analysis of multiple samples in one run with DiscoVir.

Outputs

- DiscoVir provides abundance matrices and visualizations for vOTUs, functional genes, AMGs, and hosts.
- DiscoVir provides inputs to Downstream Analysis pipeline in Nephele, MicrobiomeDB, or R for diversity analysis.

User experience

- Only fully automated viral metagenomics pipeline available with a web application.

Future work and considerations

- Viral binning could be beneficial with metagenomic analysis.
- Modify to also allow for accurate long read abundance calculations
- Receive user feedback to enhance and improve!

Funding: This work has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services under BCBB Support Services Contract HHSN316201300006W/75N93022F00001 to MEDICAL SCIENCE & COMPUTING.

References: 1. Sommers, et al. (2021) https://doi.org/10.1146/annurev-virology-010421-053015; 2. Weber, et al. (2018) https://doi.org/10.1093/bioinformatics/btx617; 3. https://dx.doi.org/10.17504/protocols.io.n92ldm98xl5b/v1; 4. Roux, et al. (2019) https://doi.org/10.1038/nbt.4306; 5. Zhou, et al. (2023). https://doi.org/10.1002/imt2.118; All references for tools are at https://nephele.niaid.nih.gov/pipeline_details/discovir/