An Agent-Based Viral Venture: Adaptive Tool Selection for Scalable Genomics

Naomi Kolodisner [1] Alok Kamatar (Advisor) [2] Greg Pauloski (Advisor)[2] Kyle Chard (Advisor) [2]

[1]University of Arizona [2]University of Chicago

Abstract

Fecal Microbial Transplant (FMT) is an effective procedure for restoring gut microbiome balance in patients with *Clostridioides difficile* infection by introducing healthy donor microbes. Tracking viral genomes during FMT provides insight into microbial community transfer and recovery. We developed a viral detection pipeline that processes metagenomic samples to identify, dereplicate, cluster, and annotate viral sequences using GeNomad, CheckV, MMseqs2, and BLAST. The pipeline links viral sequences to donor and patient samples, enabling longitudinal tracking. Traditionally, such pipelines run sequential workflows with predefined tools and steps. Here, we implement an agent-based system using Academy that autonomously selects the optimal viral detection tool through an epsilon-greedy strategy, dynamically balancing sequence quality and database match scores. Scaling experiments show that parallelizing the pipeline using Parsl reduces runtime by over 50% with minimal variability. Tool comparison demonstrates trade-offs in speed, quality, and match ratio, demonstrating the benefits of adaptive, agent-driven workflows for scalable viral detection in microbiome studies.

FMT Viral Detection

FMT transfers stool from a healthy donor into the intestinal tract of a recipient. It is primarily used to treat infections caused by *Clostridioides difficile* (*C. diff*), a bacterium that disrupts the gut microbiome. This disruption allows *C. diff* to overgrow and cause illness. Introducing healthy donor microbes helps restore the recipient's gut microbiome balance.

Viral Detection Pipeline

- **Dataset:** Smillie et al., *Cell Host Microbe* (2018) 23(2):229–240. DOI: 10.1016/j.chom.2018.01.003
- Tracks viruses as they move from donor to patient, and over time during recovery.
- Uses a metagenomic approach to capture both lytic and temperate viruses.

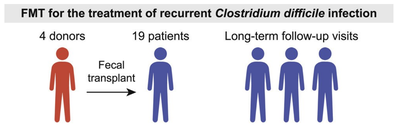


Figure 1. FMT Transplant

- **Identification/Annotation Database:** Aggregated Gut Viral Catalogue (AVrC) from Lugli et al. (*PLOS Computational Biology*, 2025; 21(5):e1012268. doi:10.1371/journal.pcbi.1012268)
- Pipeline takes assembled contigs (FASTA) as input, identifies viruses, and annotates them with sample IDs for patient/donor tracking.

Parsl

- Python library for building parallel and distributed workflows
- Simplifies parallelism by wrapping Python functions or external apps as "apps" that run concurrently
- Automatically manages task dependencies using implicit dataflow and dynamic task graphs
- Seamlessly scales from laptops to HPC clusters and supercomputers without modifying code

Academy

- Python library for building modular, stateful agents that run autonomously and interact asynchronously
- Supports agent-based design through @action and @loop decorators for clear separation of behavior and control logic
- Enables distributed and federated deployment across heterogeneous computing environments
- Handles inter-agent communication via message-passing and mailbox abstractions for robust coordination

Viral Detection Pipeline

- **Step 1: Viral Detection (GeNomad)** – Run GeNomad on all metagenomic samples. Uses machine learning and homology-based methods.
- **Step 2: Quality Assessment (CheckV)** – Evaluate completeness and quality of predicted viral genomes.
- **Step 3: Dereplication (MMseqs2)** – Remove duplicate or highly similar viral sequences across samples.

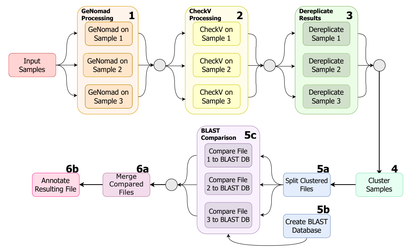


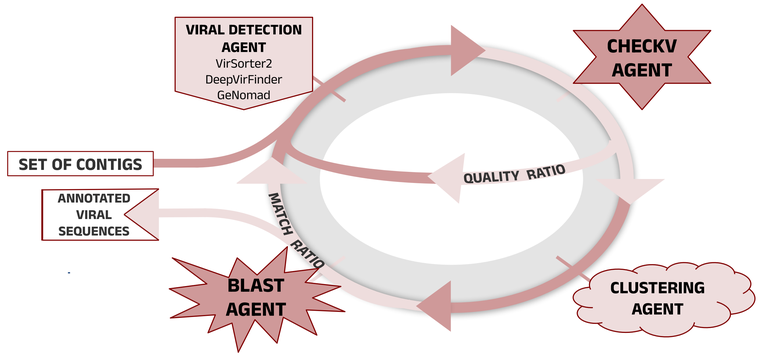
Figure 2. Linear Viral Detection Pipeline

- **Step 4: Clustering (MMseqs2)** – Combine dereplicated sequences into one file and cluster to select representative sequences.
- **Step 5: Identification (BLAST)** – **5a:** Split representative sequences into smaller files for efficient BLAST processing; **5b*:** Create a BLAST database from AVrC (*one-time setup*); **5c:** Compare representative sequences against the AVrC database using BLAST.
- **Step 6: Consolidation and Annotation** – **6a:** Merge all BLAST output files into one comprehensive file; **6b:** Collect annotation info from AVrC and link sequences back to original sample IDs.

Adaptive Workflow

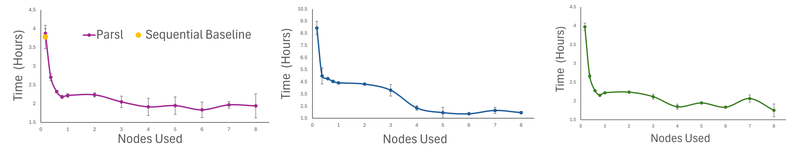
To introduce adaptability into viral detection, we implemented an adaptive workflow where tool selection is guided by an epsilon-greedy strategy with epsilon = 0.6. At each iteration, the system selects one of three viral detection tools (VirSorter2, DeepVirFinder, GeNomad) based on performance metrics while balancing exploration and optimization. Two agents provide key feedback for performance evaluation:

- **CheckV Agent:** Assigns a quality score based on the proportion of detected viral sequences classified as high quality.
- **BLAST Agent:** Assigns a match score based on the proportion of viral sequences matching the reference AVrC database.



Scaling Experiments

Before adding agent-based adaptation, the viral detection pipeline used Parsl for parallel execution, allowing independent tasks to run concurrently and reducing runtime. Scaling tests from 0.17 nodes (16 cores of a 94-core node) to 8 nodes showed execution time dropped from 4 hours to 2 hours up to 4 nodes, with minimal gains beyond that for both the linear and adaptive workflows. This plateau occurs because the workload is insufficiently large to fully utilize more than four nodes. With 12 samples, performance improved up to 6 nodes, showing that larger workloads better utilize additional resources.



(a) Linear Workflow Scaling with 3 Samples (b) Linear Workflow Scaling with 12 Samples (c) Adaptive Workflow Scaling with 3 Samples

Figure 4. Workflow Scaling Comparison

Tool Comparison

Methodology	VirSorter2	DeepVirFinder	GeNomad
Homology Based	☒	☐	☒
Machine Learning Based	☐	☒	☒

Table 1. Tool Methodology Comparison

Each tool in the adaptive workflow shows different strengths for the FMT metagenomic data. The homology-based tool VirSorter2 delivers the highest quality scores, while GeNomad performs best at finding viral matches to the AVrC database. Overall, the Epsilon-Greedy tool selection approach achieves a quality ratio comparable to GeNomad and a high match ratio. In terms of speed, DeepVirFinder is the fastest, with GeNomad close behind, while VirSorter2 takes the longest to run.

(a) Comparison of Match Ratio (b) Comparison of Quality Ratio (c) Time Comparison

Figure 5. Performance and Efficiency Tool Comparison

References

[1] Yadu Babuji, Anna Woodard, Zhuozhao Li, Daniel S. Katz, Ben Clifford, Rohan Kumar, Lukasz Lacinski, Ryan Chard, Justin M. Wozniak, Ian Foster, Michael Wilde, and Kyle Chard. Parsl: Pervasive parallel programming in python. *HPDC '19: Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, pages 25–36, 2019. doi:10.1145/3307681.3325400.

[2] Antonio Pedro Camargo, Simon Roux, Frederik Schulz, Michal Babinski, Yan Xu, Bin Hu, Patrick S. G. Chain, Stephen Nayfach, and Nikos C. Kyrpides. Identification of mobile genetic elements with genomad. *Nature Biotechnology*, 42:1303–1312, 2024. doi:10.1038/s41587-023-01953-y.

[3] Anastasia Galperina, Gabriele Andrea Lugli, Christian Milani, Willem M. De Vos, Marco Ventura, Anne Salonen, Bonnie Hurwitz, and Alise Jany Ponsero. The aggregated gut viral catalogue (avrc): A unified resource for exploring the viral diversity of the human gut. *PLOS Computational Biology*, 21(5), 2025. doi:10.1371/journal.pcbi.1012268.

[4] J. Gregory Pauloski, Yadu Babuji, Ryan Chard, Mansi Sakarvadia, Kyle Chard, and Ian Foster. Empowering scientific workflows with federated agents. *arXiv:2505.05428v2*, 2025. doi:10.48550/arXiv.2505.05428.

[5] Kenneth E. Schackart, Jessica B. Graham, Alise J. Ponsero, and Bonnie L. Hurwitz. Evaluation of computational phage detection tools for metagenomic datasets. *Frontiers in Microbiology*, 14, 2023. doi:10.3389/fmicb.2023.1076760.

[6] Christopher S. Smillie, Jenny Sauk, Dirk Gevers, Jonathan Friedman, Jaeyun Sung, Ilan Youngster, Elizabeth L. Hohmann, Christopher Staley, Alexander Khoruts, Michael J. Sadowsky, Jessica R. Allegretti, Mark B. Smith, Ramnik J. Xavier, and Eric J. Alm. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host & Microbe*, 23(2):229–240, 2018. doi:10.1016/j.chom.2018.01.003.

Funded by NSF Grant #2127548

kolodisner@arizona.edu