Adaptive Tool Selection in a Scalable Genomics Pipeline

Naomi Kolodisner

CICF 2025, Globus Labs

Mentors: Alok Kamatar, Greg Pauloski, and Kyle Chard

USC Viterbi School of Engineering Information Sciences Institute
INDIANA UNIVERSITY
UNIVERSITY OF NOTRE DAME
THE UNIVERSITY OF UTAH
TEXAS TECH UNIVERSITY
renci

My Project

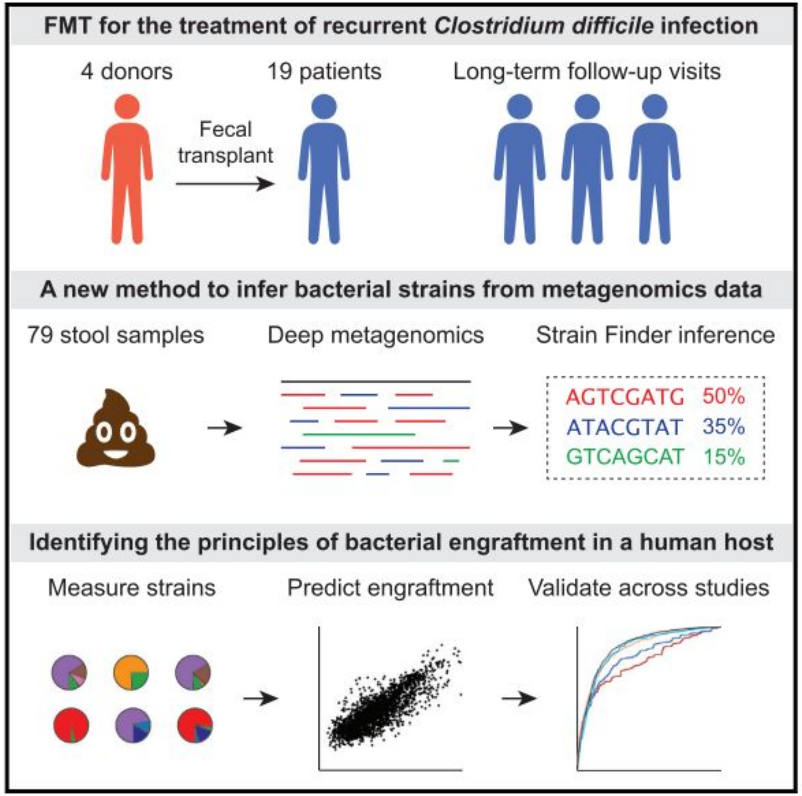
Created FMT viral detection pipeline as part of my undergraduate research

Use Parsl make my original pipeline more efficient and scalable

Apply adaptive tool selection with Academy

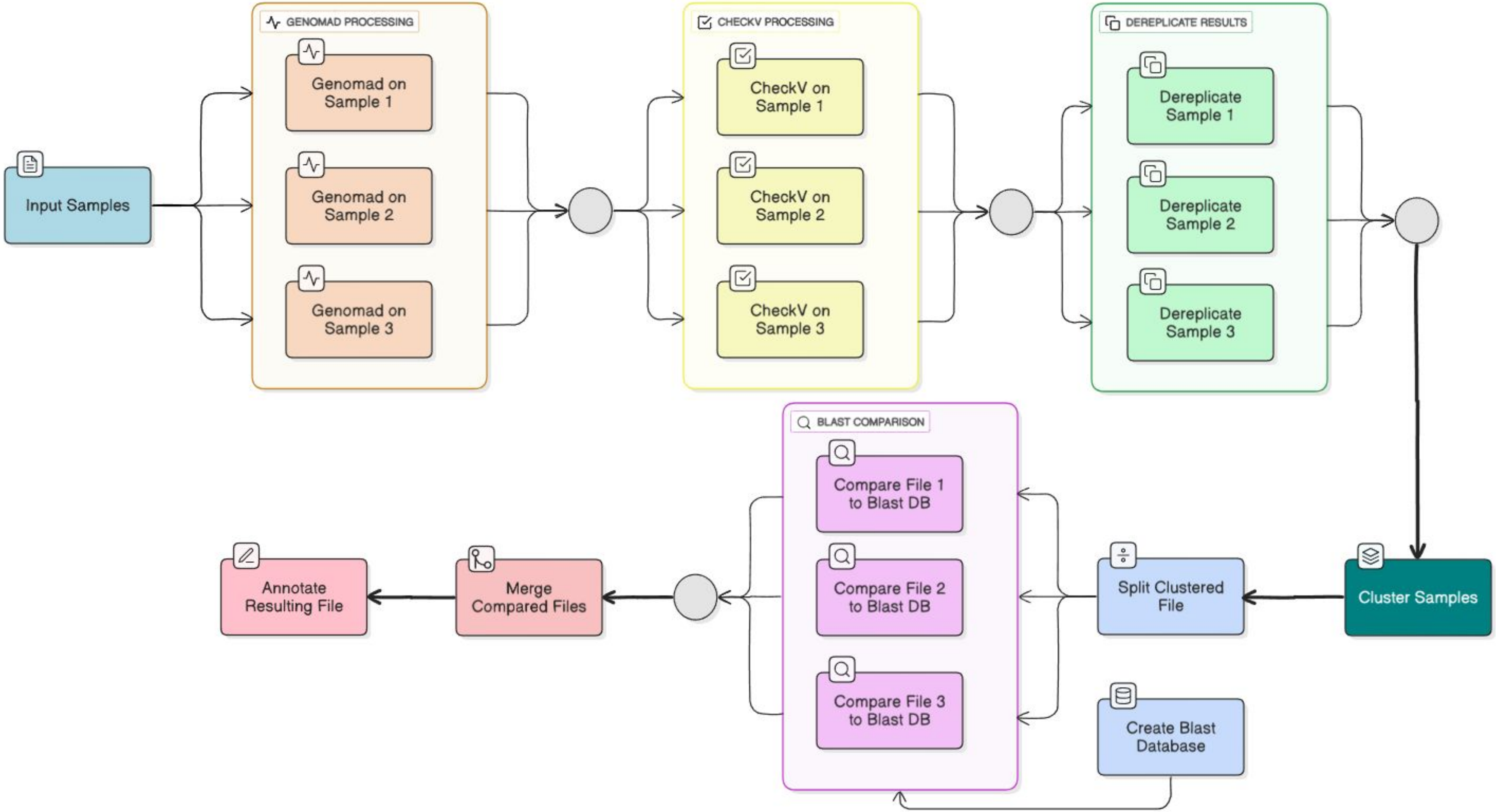
FMT Viral Detection

- FMT for C. diff
- Tracking viruses
- From donor to patient, and over time
- Metagenomic approach
- Assembled contigs → annotated viruses



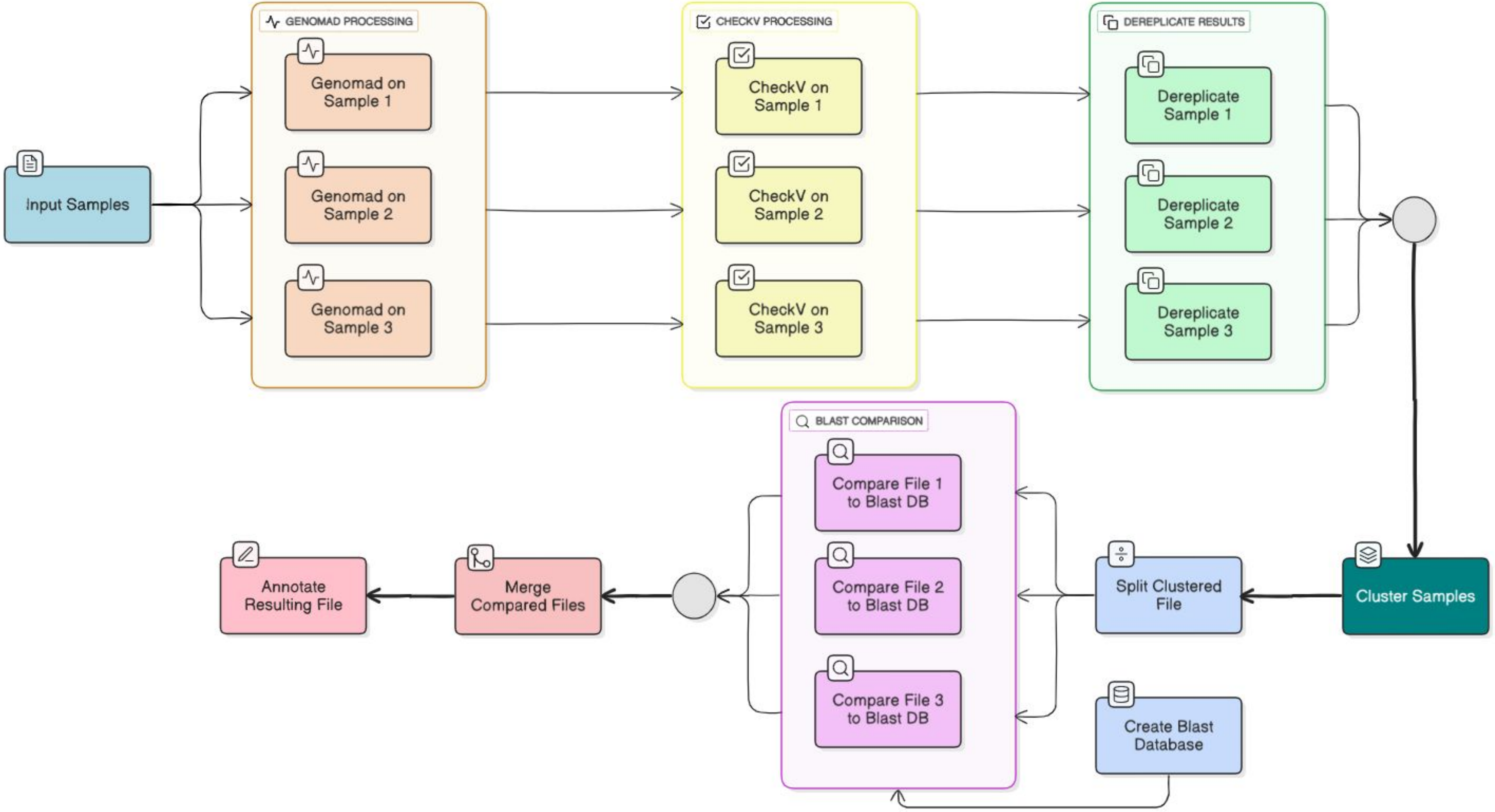
From: Smillie et al., *Cell Host Microbe* 2018;23(2):229–240.
DOI: 10.1016/j.chom.2018.01.003.

Funded by the U.S. National Science Foundation, Grant #2127548



Parsl

- Apps run concurrently
- Parsl creates a dynamic graph of tasks and their dependencies
- Tasks executed when their dependencies are met
- “Natural parallel programming!”



Input Samples

GENOMAD PROCESSING

Genomad on Sample 1

Genomad on Sample 2

Genomad on Sample 3

CHECKV PROCESSING

CheckV on Sample 1

CheckV on Sample 2

CheckV on Sample 3

DEREPLICATE RESULTS

Dereplicate Sample 1

Dereplicate Sample 2

Dereplicate Sample 3

Cluster Samples

Split Clustered File

Create Blast Database

BLAST COMPARISON

Compare File 1 to Blast DB

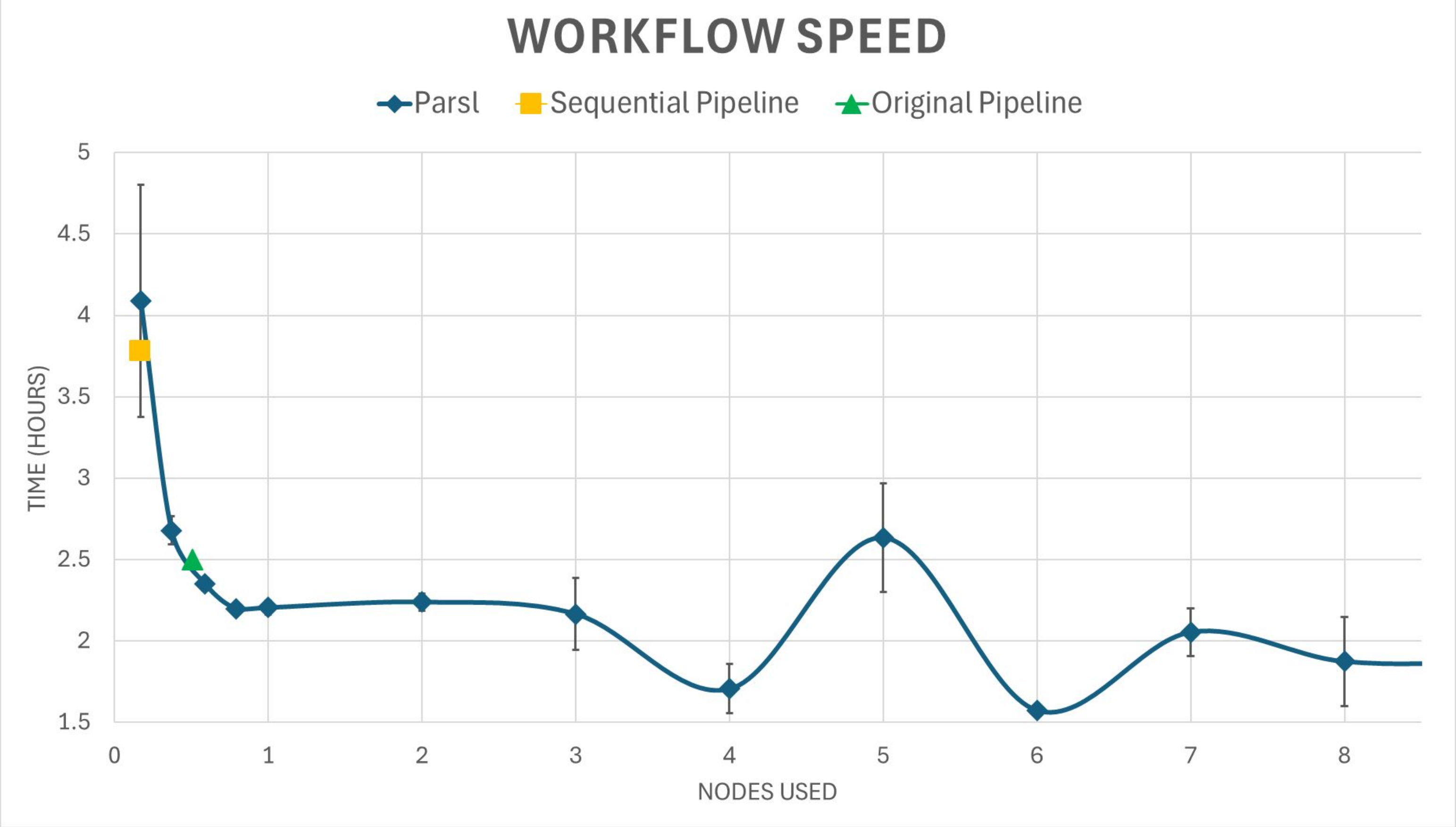
Compare File 2 to Blast DB

Compare File 3 to Blast DB

Merge Compared Files

Annotate Resulting File

Scaling the Workflow

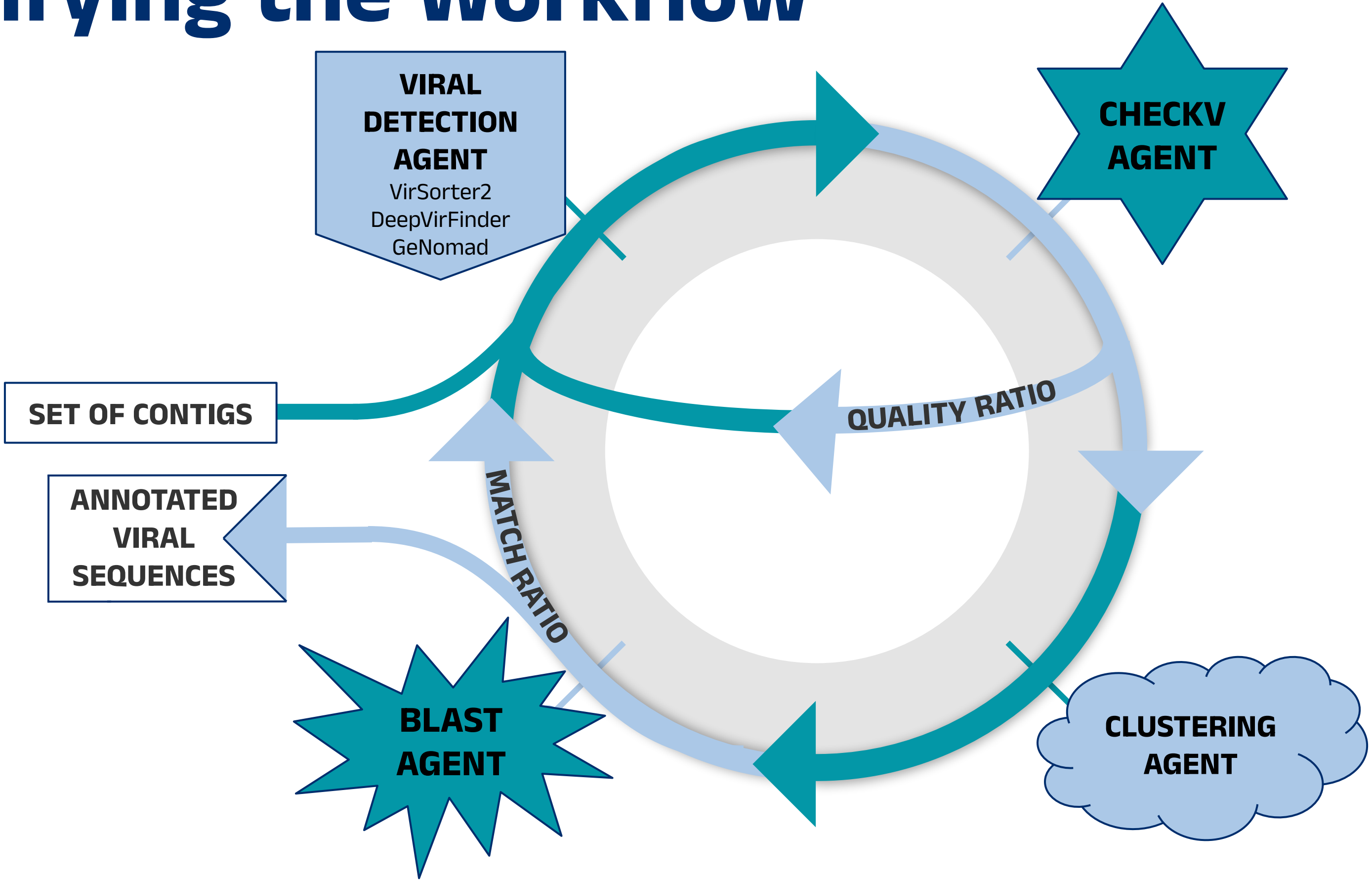


Academy

- Modular middleware for building autonomous agents
- Manage inter-agent communication
- “Deploy agents across distributed, federated, and heterogeneous resources”

Funded by the U.S. National Science Foundation, Grant #2127548

Agentifying the Workflow



VIRAL DETECTION AGENT
VirSorter2
DeepVirFinder
GeNomad

CHECKV AGENT

SET OF CONTIGS

QUALITY RATIO

ANNOTATED VIRAL SEQUENCES

MATCH RATIO

BLAST AGENT

CLUSTERING AGENT

Funded by the U.S. National Science Foundation, Grant #2127548

Skills Learned and Takeaways

1. Understanding of HPC and resource management
2. Understanding of workflow management (Parsl, Academy, etc)
3. Learned about autonomous agents and how to agentify a workflow
4. Data collection and documentation
5. Creativity in a workflow

Funded by the U.S. National Science Foundation, Grant #2127548

Future of the Project

Submitting poster to the ACM Student Research Competition at SC25

Expanding to other datasets to understand and explore tool selection process

Exploring possibilities of agents (annotation, clustering, etc.)

Acknowledgments

Thank you to my mentors Alok Kamatar, Greg Pauloski, and Kyle Chard!

Thank you to Globus Labs and NSF CI Compass, funded by the U.S. National Science Foundation, Grant #2127548 for supporting me and making this opportunity possible!

Check out my code!

https://github.com/globus-labs/academy-fmt-viral-workflow