Simulating Virus Detection with Nanopore Sequencing, BLAST, and the Hypergeometric Distribution

Alexander Li[1], Raeuf Roushangar[2], Anthony Sinskey[2], Stacy Springs[2]; Thesis Advisor: Lorin Crawford
[1]Brown University, Providence, RI
[2]Center for Biomedical Innovation, MIT, Cambridge, MA

BROWN
School of Public Health

Overview

We utilized simulations to prove the feasibility of metagenomic virus detection as a methodology in biomanufacturing and compared the results against a statistical model.

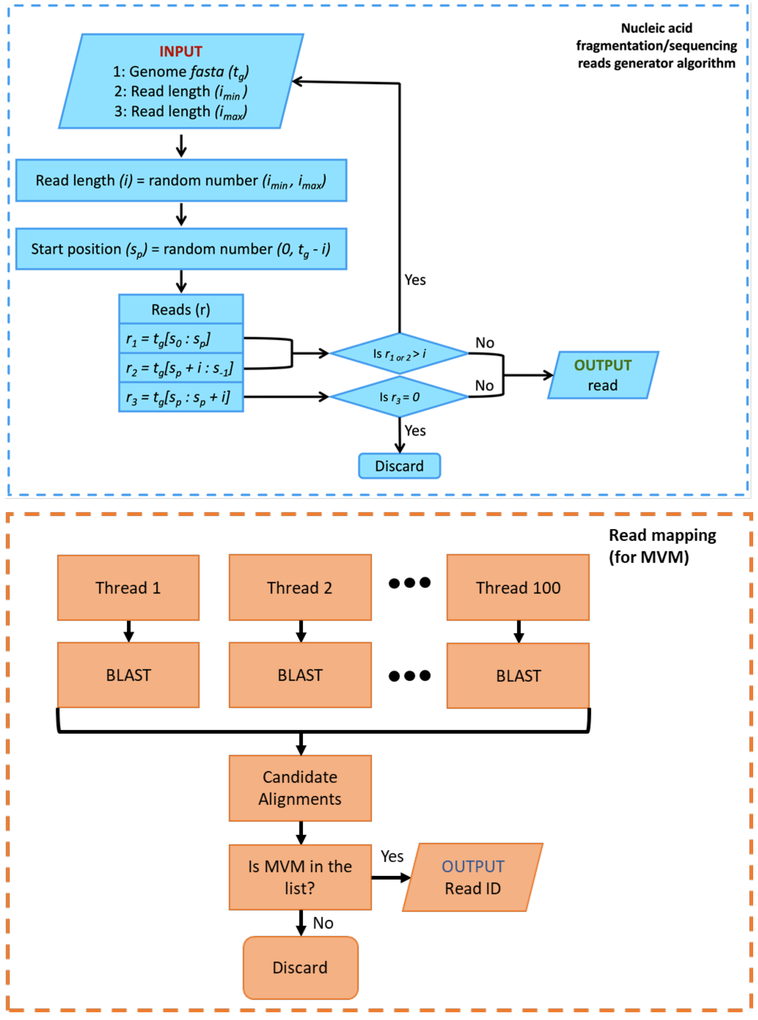
Background

- Recent advancements in biomanufacturing have resulted in an increase in demand for methodologies to improve the safety and efficacy of the biomanufacturing process
- Virus detection through metagenomic sequencing can detect genomic material from any contaminant
- Oxford Nanopore Technology (ONT) sequencers provide high-throughput while decreasing cost and turnaround time.
- Simulations enable low-cost benchmarking of bioinformatics tools
- Statistical models can estimate sensitivity and turnaround time for future virus detection experiments, allowing for more widespread adoption of metagenomic virus detection

Methodology

- ONT sequencing was simulated by fragmenting and modifying copies of the reference genomes for Chinese Hamster Ovary (CHO) and Minute Virus of Mice (MVM)
- Reads were classified by mapping against the Reference Viral Database (RVDB) using parallelized BLAST
- Three sets of experiments were conducted:
  - Positive control samples containing only MVM were used to measure classification accuracy
  - Detection time was measured for multiple ratios of MVM to CHO
  - Sensitivity and speed were measured through repeated simulations with a single ratio of MVM to CHO
- ONT sequencing was modeled using the hypergeometric distribution
- Virus detection sensitivity in the simulations was compared against sensitivity predictions from the hypergeometric distribution

Simulation Workflow



Results

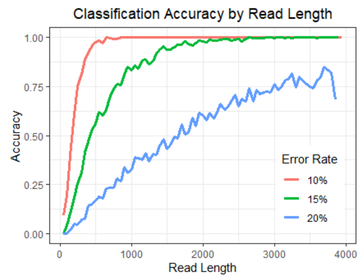


Fig. 1. Plot of classification accuracy by read length and read error rate:10% error rate (red), 15% error rate (green), and 20% error rate (blue).

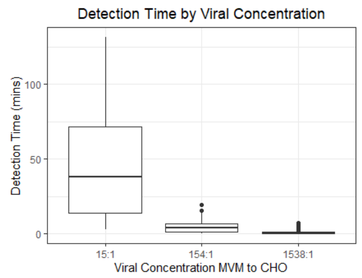


Fig. 2. Boxplot of runtime for virus detection at different viral concentrations.

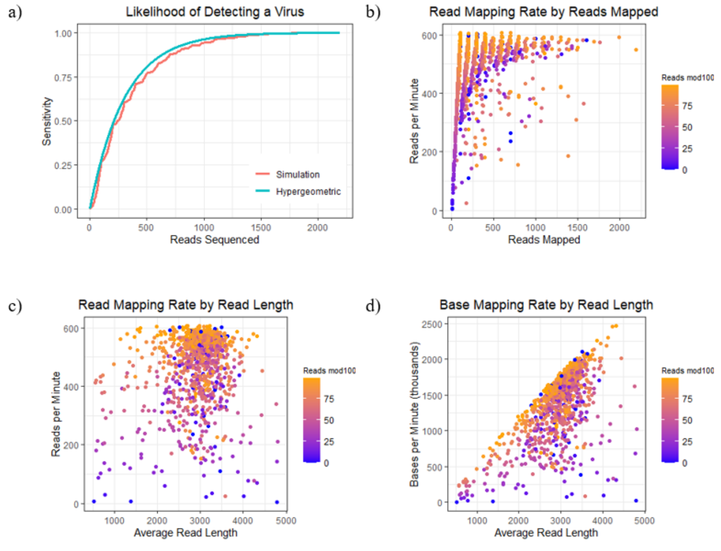


Fig. 3. Metagenomic virus detection sensitivity and runtime. a) Likelihood of detecting a virus read after mapping x reads. Observed results from 1,000 repeated simulations (red) and predicted likelihood from the hypergeometric model (cyan) shown. b) Plot of number of reads mapped per minute against reads mapped for each of the 1,000 simulations. The color coding denotes the number of reads mapped mod 100. c) Plot of reads mapped per minute against average read length in 1,000 simulations. d) Plot of number of bases mapped per minute against the average read length in 1,000 simulations.

Conclusions

- Metagenomic virus detection will likely become more attractive in biomanufacturing as further research is conducted
- Basic read mapping algorithms are capable of high speed and accuracy for read classification
- Understanding ONT sequencing from the perspective of sampling theory will provide much needed clarity around the expected sensitivity and turnaround time of metagenomic virus detection