Camparison of current methods for Viral Detection in Tumor RNA Sequencing

Huiming Zhou
Student ID: 1327380
Program: MDS

March 4, 2025

Topic Description and Motivations

During my paper review, I selected VirHunter as my research focus because I was particularly interested in novel virus detection technologies. Currently, I want to explore how deep learning can improve viral detection in clinical settings. This led me to discover a recent study published in Nature titled ”A deep learning approach reveals unexplored landscape of viral expression in cancer” , which introduces the viRNAtrap method for efficient viral identification in tumors. The most challenging problem :

- VirHunter has only been tested on plant viruses, and its performance on human cancer data remains unverified
- Findinag a benchmark for comparison of methods including viRNAtrap, VirHunter, DeepVirFinder

This project will address these challenges through:

- Direct performance comparison between VirHunter (adapted for human virus data) and viRNAtrap
- Validation using genome-wide sequencing (GWS) as the gold standard for known viruses (e.g., HPV)
- Optional structural validation with AlphaFold-predicted viral proteins if time permits

Objectives and Research Methods

The primary goal of this study is to benchmark deep learning-based tools (viRNAtrap, DeepVirFinder) against traditional machine learning methods (VirHunter) for viral detection in tumor RNA-seq data. Key objectives include:

- Identifying superior performance patterns among tools (e.g., viRNAtrap for low-abundance viruses vs. DeepVirFinder for novel strains)
- Establishing minimum performance thresholds through ROC curve analysis to define clinical usability standards
- Testing VirHunter’s adaptability to human oncology data through transfer learning

Implementation strategies focus on efficiency and reproducibility:

- **Pretrained Models**: Direct deployment of viRNAtrap (GitHub) and DeepVirFinder (BioMedInfor) using default parameters, with VirHunter fine-tuned on human viral sequences
- **Validation Framework**: BLASTn validation against NCBI RefSeq viruses
- **Optional Extension**: Integrating AlphaFold-predicted viral protein structures with viRNAtrap outputs through random forest classifier

Timeline and Plan

Week Tasks
Week 1 • Data Preparation • Download all models needed and run locally
Week 2 • VirHunter retraining • process the data for model input
Week 3 • Process all samples through viRNAtrap/DeepVirFinder/VirHunter • Compare the results and performance
Week 4 • Report writing and refining • Alphafold if time permits

References

1. Elbasir, A., Ye, Y., Schäffer, D.E. et al. A deep learning approach reveals unexplored landscape of viral expression in cancer. Nat Commun 14, 3456 (2023). https://doi.org/10.1038/s41467-023-36336-z
2. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
3. Rampelli, S., Soverini, M., Turroni, S. et al. VirHunter: A Deep Learning-Based Method for Plant Virus Identification. Front Bioinform 2, 867111 (2022). https://doi.org/10.3389/fbinf.2022.867111
4. Alharbi, F. Vakanski, A. Machine learning methods for cancer classification using gene expression data: A review. Bioengineering 10, 242 (2023). https://doi.org/10.3390/bioengineering10020242
5. Uffelmann, E., Huang, Q. Q., Munung, N. S. et al. Genome-wide association studies. Nat Rev Methods Primers 1, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9
6. Ren, J., Ahlgren, N.A., Lu, Y.Y. et al. DeepVirFinder: identifying viruses from metagenomic data using deep learning. *Bioinformatics* 36, 4646–4654 (2020). https://pubmed.ncbi.nlm.nih.gov/34084563/