Machine Learning Enables Viral Genome-Agnostic Classification of RNA Virus Infections from Host Transcriptomes

Benjamin Kaza*[1,2], Amandine Gamble[1], Stephanie N. Seifer[3] & Timothée Poisot[2]

[1]Department of Public and Ecosystem Health, Cornell University, Ithaca, NY, USA, [2]Department de Sciences Biologiques, Université de Montréal, Montréal, QC, Canada, [3] Paul G. Allen School for Global Health, Washington State University, Pullman, WA, USA

*Corresponding author

ABSTRACT:

Targeted PCR diagnosis of RNA viruses is sequence dependent, meaning that the accuracy of the assay depends on the identity of the viral sequence. However, sequence-targeted assays can miss novel or divergent viruses. We test whether the host transcriptome alone can classify RNA virus infections without using viral sequences. Using publicly available data on Huh7 and Calu-3 cells experimentally infected with diverse negative-sense RNA, we evaluate two host-derived feature sets (differential expression dataset of identified genes and alignment-free nucleotide k-mer spectra) and compare hierarchical clustering with Random Forests. Across datasets and timepoints (12 to 24 hpi), Random forests accurately distinguished cells infected with different viruses. Controls with label permutation and read shuffling across experimental conditions established non-random performance. Influenza A infections exhibited the strongest, most distinct signatures, whereas Ebola and Lassa virus responses were subtler yet still classifiable. These results show that host-only transcriptomics encodes virus-specific, complex signals that machine learning can exploit, enabling genome-agnostic classification of infection statuses. This approach could aid early outbreak triage when primer based detection methods fail or viral genomes are unknown and complements sequence-based discovery.

INTRODUCTION

Zoonotic viruses—those that spill over from animal reservoirs into human populations—pose a growing threat to global health. Anthropogenic global change and biodiversity loss are accelerating the frequency of spillover events by altering ecosystems, disrupting host-pathogen dynamics, and increasing contact between humans and wildlife[1,2]. While only a subset of animal viruses known to be are capable of infecting humans, those that do are disproportionately likely to cause severe disease and, in some cases, global pandemics[3–7]. Despite this risk, the vast majority of viruses circulating in wildlife remain undescribed[8–10]. Most have never been isolated, sequenced, or even detected[9–11]. This knowledge gap severely limits our ability to assess which viruses might pose a threat to public health. In addition, once in humans, many viruses may cause subclinical or misdiagnosed diseases (such as undiagnosed febrile diseases that could resemble relatively common diseases like malaria or influenza), further complicating efforts to identify them before they spread[12,13].

Targeted PCR-based assays face similar challenges. These methods depend on primers designed to match conserved regions of known viral genomes. If a virus's genome diverges too much from the expected sequence, primer binding may be inefficient or fail entirely, resulting in false negatives[14–17]. Even assays designed with degenerate primers—such as pan-family PCRs—can fail to detect divergent strains. For example, a second genotype of Hendra virus (HeV-g2) circulated undetected in a surveilled population of bats due to mismatches with primers designed for the original Hendra virus genome[18]. Similarly, a recombinant coronavirus (CCoV-HuPn-2018) that caused febrile illness in children in Malaysia and Haiti escaped detection by coronavirus-specific RT-PCR[19,20]. This is particularly concerning because Henipaviruses and Coronaviruses are both considered WHO priority pathogens that are flagged for their potential to cause pandemics with high fatality rates, and in these two (and potentially many more) cases they were missed[21].

Modern virus discovery relies heavily on sequence-based methods (such as genome amplification by targeted PCR for detection and sequencing), particularly high-throughput sequencing and metagenomics[22,23]. These approaches enable the recovery of viral genome fragments directly from clinical or environmental samples without prior knowledge of the virus's identity[24,25]. However, sequence-based discovery typically requires some level of similarity between the unknown virus and existing sequences in reference databases[26–28]. Highly divergent viruses may go undetected if they lack conserved genes or recognizable sequence motifs[29,30]. This is a particularly relevant problem for RNA viruses, which tend to have high mutation rates, genetic variability, and extreme expression differences between genes making it difficult to pick a good PCR target across families—and even within genera[31]. For example, transcription in Ebola virus (EBOV) follows a 3' – 5' gradient with the RNA-dependent RNA polymerase being the least expressed gene[32]. As such, there is an urgent need for virus detection strategies that do not depend on prior sequence information.

One alternative is to detect infection by measuring changes in the host cell, rather than detecting the virus directly. Viral infection can induce characteristic transcriptional responses in host cells[32,33], which may serve as indirect markers of infection. Host-based diagnostic methods are already used clinically to distinguish between viral and bacterial infections by analyzing expression patterns of 29 genes that are up- or downregulated during infection[34,35]. Some studies have used short k-mers from viral genomes to train machine learning models that can classify viral infections from total RNA-seq data, without relying on alignment to reference genomes to determine the presence or absence of one type of viral infection or to discriminate between bacterial or viral infection[36–39].

However, few studies have systematically tested whether host transcriptomic responses alone—without any viral sequence—can be used to classify infections across different viral families. It remains unclear whether the host response provides enough resolution to infer the identity or family of the infecting virus, especially across diverse cell types and timepoints.

In this study, we test the hypothesis that viral species can be inferred from host transcriptomic responses alone. We evaluate two feature types (short nucleotide k-mer spectra from total RNA and differential gene expression) and assess their utility for classifying viral infections using machine learning and hierarchical clustering. We focus on negative-sense RNA viruses infecting mammalian cell lines, using publicly available transcriptomics datasets[40,41]. Our panel includes experimental infections of Huh7 cells (a human hepatocyte immortalized cell line) and Calu-3 cells (a human lung epithelial immortalized cell line) with viruses including Influenza H1N1, H5N1, H7N7, H7N9, EBOV, Marburg virus (MARV), Respiratory Syncytial Virus (RSV), Nipah virus (NiV), Lassa virus (LASV), Rift Valley Fever virus (RVFV), and Sandfly Fever Sardinia Virus (SFSV). Across both cell types, we show that different viruses elicit distinct, classifiable host responses—supporting the possibility of sequence-independent viral classification using host transcriptional data alone.

RESULTS

Gene expression patterns elicited by infection vary in magnitude and differentiability

To test the hypothesis that each viral infection results in a differentiable response compared to other viruses, we performed hierarchical clustering of differential gene expression profiles from experimentally infected Huh7 cells at 12 and 24 hours post infection. In short, hierarchical clustering is a method for grouping similar data points into clusters by progressively merging the

closest pairs to form a tree-like structure called a dendrogram. We used the Euclidean distance, which is a measure of the straight-line distance between two points in space, calculated as the square root of the sum of the squared differences between their corresponding coordinates[42].

The magnitude and specificity of the host transcriptional response varied across viral infections, resulting in distinct patterns of gene expression. Hierarchical clustering using the top 100

differentially expressed genes effectively separated infected from uninfected samples (Figure 1). While some viral infections elicited sharply divergent transcriptomic profiles, others were more subtle and clustered closer to uninfected controls. Notably, infections with Influenza A viruses (H1N1 and H5N1) induced the most distinct expression patterns, clustering farthest from the uninfected samples and from other viruses.

Several other viruses—including NiV, RSV, SFSV, RVFV, and MARV—also produced expression profiles distinguishable from uninfected controls, as their transcriptomes were placed outside of the subtree containing most mock-infected samples. In contrast, infections with EBOV and LASV produced host responses that were less distinguishable, with infected samples

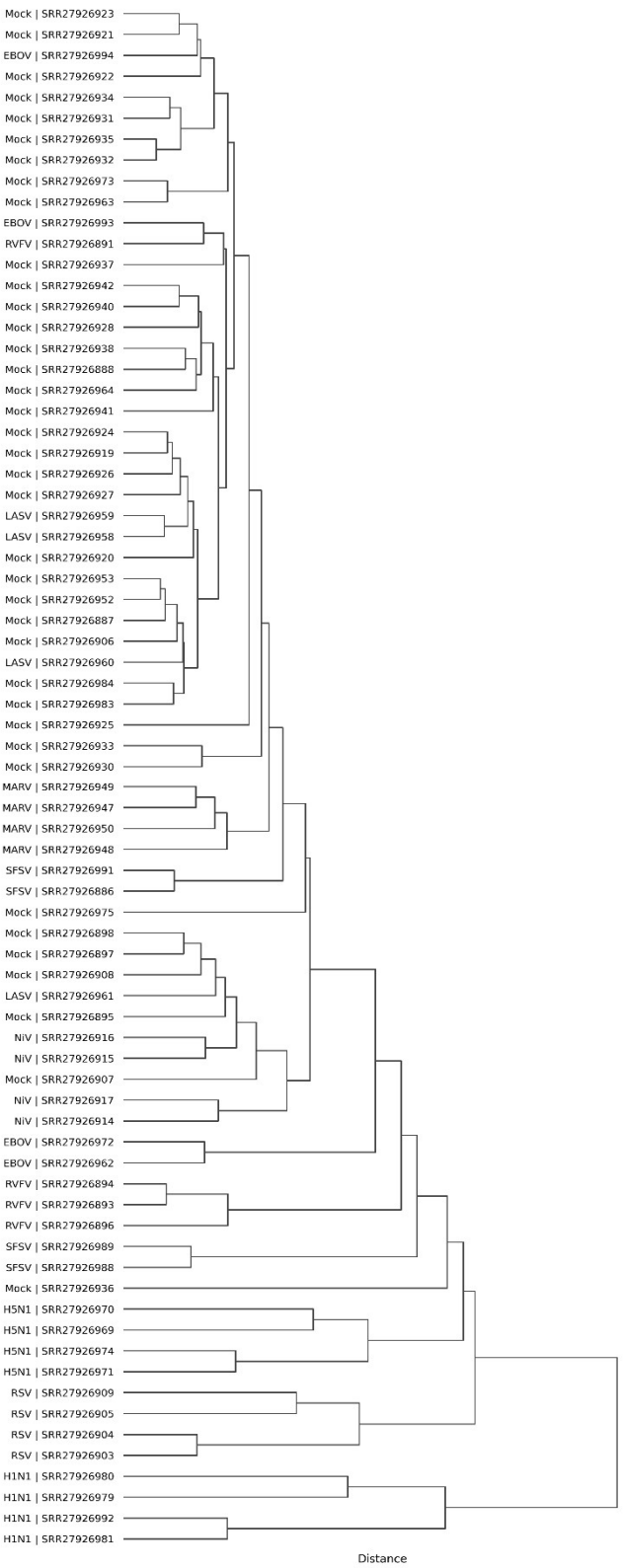

Figure 1: Hierarchical clustering of gene expression profiles from Huh7 cells 12 hours post infection

clustering within or adjacent to the subtree of uninfected controls (Figure 1).

The genes driving these distinctions were primarily involved in translation, immune response, and chromatin remodeling, followed by genes associated transport, cell signaling, and metabolism (Figure 2).

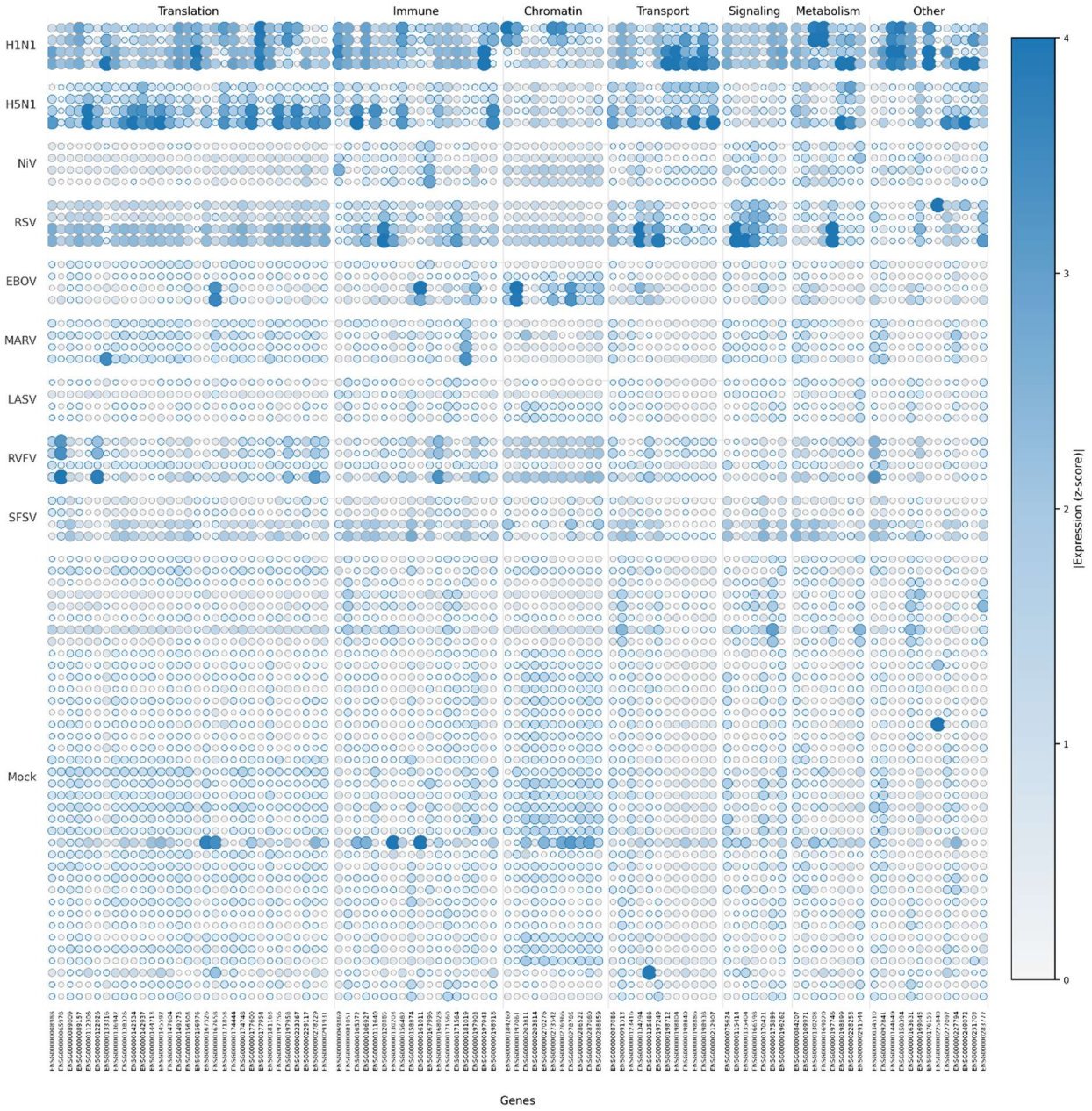

Figure 2: Z scores of top 100 differentially expressed genes in Huh7 cells infected with various negative-sense RNA viruses after 12 hours post infection and mock

A consistent feature across most infections was the downregulation of genes involved in translation, a known host antiviral response[43–45]; however, this trend was reversed in H5N1 and H1N1, and to a lesser extent RVFV, where translation-associated genes were upregulated relative to uninfected samples. Furthermore, only H5N1, H1N1, and MARV exhibited clustering patterns so distinct that their replicates grouped into clearly defined, virus-specific subtrees.

The patterns of gene expression elicited by closely related Influenza A viruses are specific and differentiable by hierarchical clustering and random forest classification

To test whether Influenza A viruses consistently induced differentiable responses in a different cell type and across viral genotypes, we used experimental infection micro-array gene expression data from Calu-3 cells experimentally infected with H1N1, H5N1, H7N7, and H7N9 viruses[41]. We found that all of the four viruses were able to be classified by both Hierarchical clustering and random forest classification at 12 and 24 hours post infection (Figure 3).

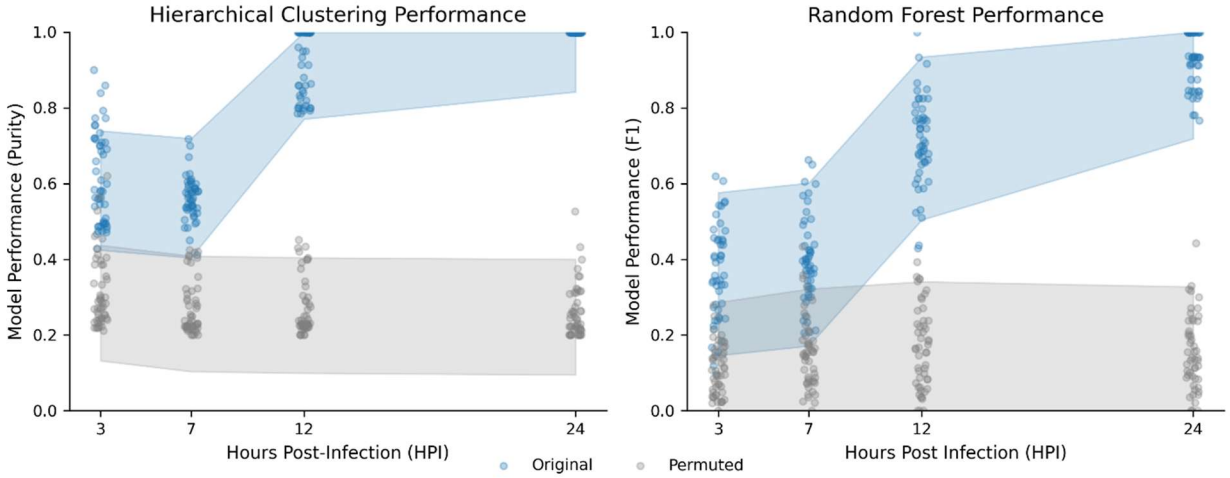

Figure 3: Hierarchical clustering Tree Purity and Random forest Classification F1 score of Differential Gene Expression Profiles originating from Calu-3 cells infected with various Influenza A viruses or mock at 3, 7, 12, and 24 hours post infection

The distribution and composition of nucleotides in an infected cell varies in a virus-specific manner

To test whether the nucleotide composition of an infected cell can be used to predict which virus was infecting the cell, we compared experimental data to simulated parallel datasets where the nucleotides were randomly shuffled within each sequencing read or a dataset where the infection status label was randomly assigned to each kmer abundance profile.

Infection status was differentiable for kmer lengths 8 to 12 above random chance across all of the viruses tested via hierarchical clustering using Euclidean distance. Virally infected transcriptomes were differentiable to random forest classification above randomly assigned data for the entire k-mer range, and was differentiable to the model in contrast with nucleotide-shuffled data when the k-mer length was greater than 4 quantified by purity and F1 score which measure the accuracy of hierarchical clustering or random forests to group samples of the same category together, respectively (Figure 4).

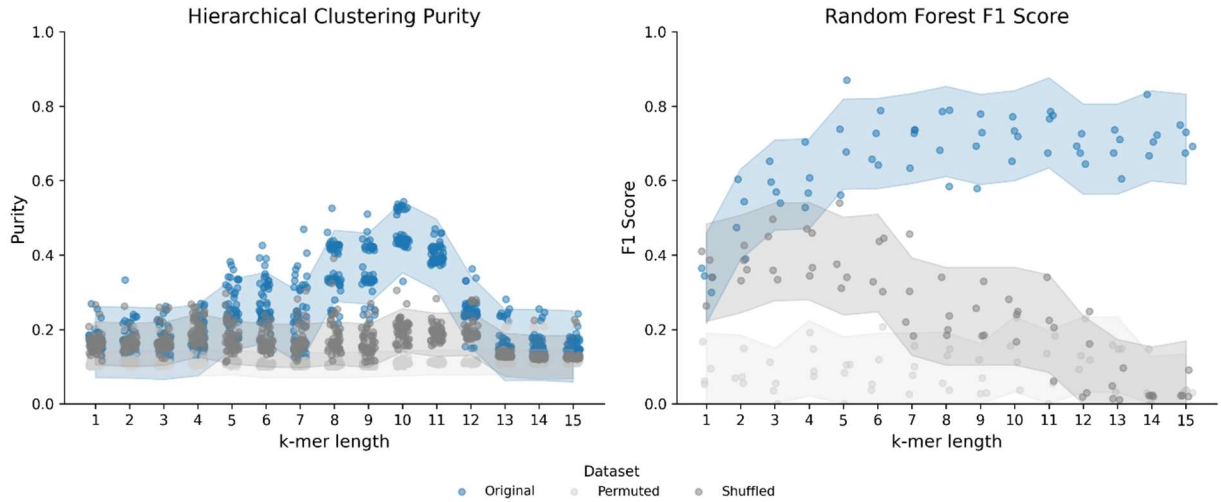

Figure 4: Performance of Hierarchical clustering measured by Tree Purity (a), and Random forest F1 Score (b), Recall (c), and Precision (d) on infected and uninfected Huh7 transcriptomes after 12 hours post infection and synthetic datasets where all of the reads in the original transcriptomes were randomly shuffled, or where each k-mer abundance profile was randomly assigned a virus (ie. permuted)

The contribution of nucleotides from negative-sense RNA viruses drives the differentiability of k-mer abundance profiles in hierarchical clustering but not random forest classification

To test if the viral RNA itself drives the differentiability of the transcriptomes in Random forest classification and Hierarchical Cluster, we created a synthetic dataset where the viral reads were depleted.

In transcriptomes where reads originating from each virus were depleted, at no point in the kmer range were the transcriptomes differentiable above the shuffled or permuted data. In Random forest classification, kmers above length 2 were able to inform the model above random chance, but not to consistently high accuracy (Figure 5).

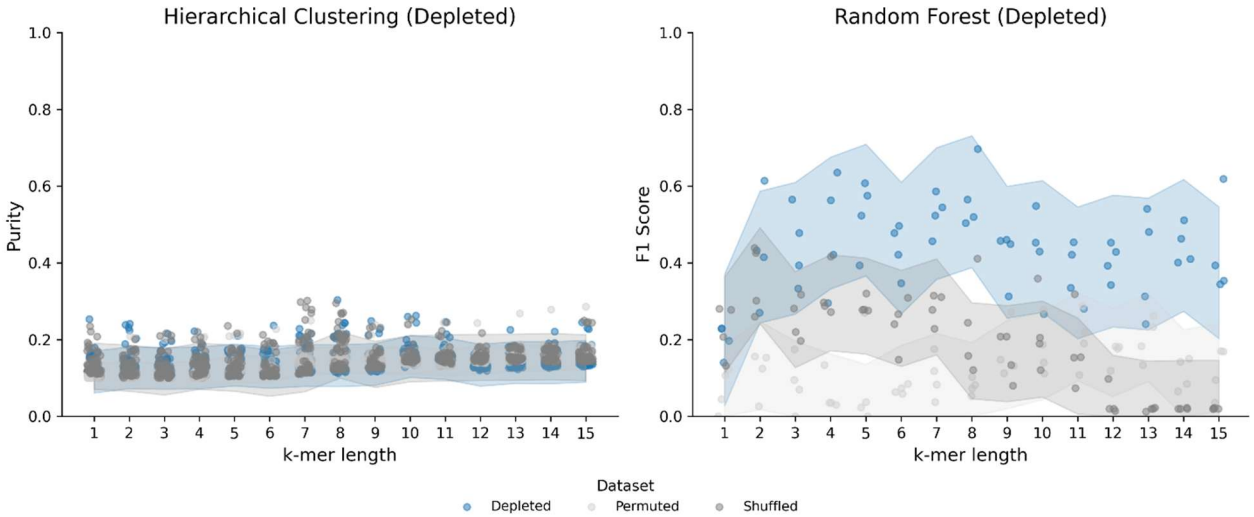

Figure 5: Performance of hierarchical clustering measured by tree purity (a) and random forest F1 score (b) on infected and uninfected Huh7 transcriptomes at 12 and 24 hours post infection that were computationally depleted of viral reads

Enrichment of viral RNA in transcriptomic data enhances the differentiability to random forest classification and hierarchical clustering

When viral reads were doubled in the dataset, the resulting differentiability by hierarchical clustering and random forest classification increased. For hierarchical clustering, the data was differentiable above random chance over the k-mer range 4 – 12 and was differentiable by random forest classification for all k-mer lengths tested (Figure 6).

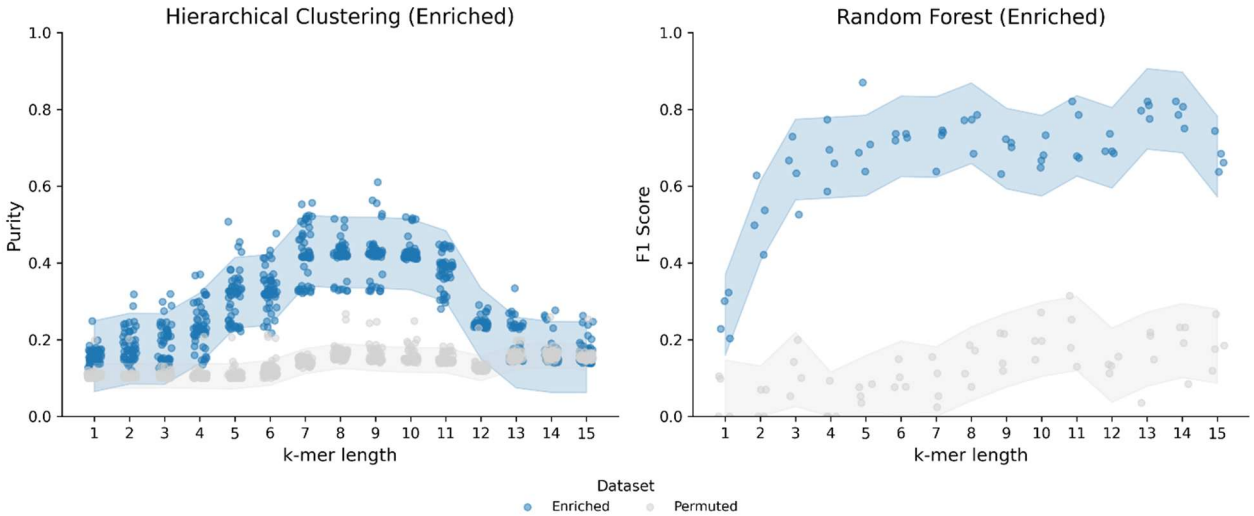

Figure 6: Performance of Hierarchical clustering measured by Tree Purity (a) and Random forest F1 Score (b) on infected and uninfected Huh7 transcriptomes at 12 and 24 hours post infection that were computationally enriched for viral reads

DISCUSSION

To replicate and spread, all viruses must manipulate host cell biology[46]. At the cellular level, this often involves subverting host gene expression and regulatory networks to promote viral replication, suppress immune responses, and evade detection[46]. Viruses systemically manipulate the host's immune system, inhibit antiviral signaling pathways, rewire host metabolism, or disrupt intercellular communication to maximize their replication fitness[47–49].

Despite their compact genomes, RNA viruses encode multifunctional proteins that carry out both essential replication tasks and host antagonism[46]. These dual roles are a product of strong selective pressures: with limited coding space, many RNA viruses have evolved proteins and genome structures that both facilitate replication and interfere with host defenses. Viral replication itself often hijacks or suppresses core host processes at every stage—transcription, translation, signaling, and transport[47,50–52].

In turn, host cells have evolved sophisticated and dynamic genetic mechanisms to detect and restrict viral infection. Mammalian cells encode hundreds of innate immune genes that are activated in response to viral invasion, though the specific pathways engaged can vary depending on the virus and host cell type[33,50–53]. Effective immune responses must be both rapid and specific: activating the appropriate antiviral programs without triggering unnecessary or energetically costly responses is critical to controlling infection without collateral damage. These diverse, dynamic responses are reflected in the transcriptome, making them an attractive substrate for computational classification.

A clear example of this functional specialization is seen in the differential response to DNA versus RNA viruses. For instance, infection with a DNA virus such as mpox (monkeypox) activates cytosolic DNA sensors like cGAS, DNA-PK, and IFI16, which initiate downstream signaling to induce interferon-β and other antiviral genes[13,14]. In contrast, RNA viruses such as SARS-CoV-2 or Influenza A are recognized primarily through double-stranded RNA intermediates by RIG-I and MDA5[52]. These sensors engage the mitochondrial adapter MAVS to trigger antiviral transcriptional programs[52].

Transcriptomic studies have shown that even among related viral strains, host responses can differ markedly. For instance, in the 2022 outbreak of mpox (Clade IIb), several genes were significantly upregulated (e.g., SLC2A3, ATP2B1, VEGFA) or downregulated (MAP3K8, IL1A, SGK1), reflecting a distinct host signature[53]. Similarly, SARS-CoV-2 variants induce variant-specific responses: the Delta variant was associated with upregulation of EGR1 and IFIT, while Alpha triggered genes like SYVN1, CH25H, VIPR1, and others[54].

In this study, we show that host transcriptomic data—whether represented as normalized gene expression values or alignment-free k-mer frequency spectra—contain sufficient information to

classify the infecting virus using machine learning. For viruses like Influenza A, which contribute substantial amounts of viral RNA to the transcriptome and elicit strong transcriptional changes, both hierarchical clustering and Random forest classification achieved high performance. In contrast, even for viruses with more subtle effects on host gene expression and lower transcriptomic viral burden, Random forest classifiers were still able to differentiate viral genotypes with high accuracy. This likely reflects the ability of Random forests to capture non-linear patterns in complex, high-dimensional data[55]. Where hierarchical clustering failed to separate infected from uninfected samples—especially in cases where the host response was mild or variable—Random forest models consistently identified the correct viral infection. This suggests that while traditional clustering methods detect global patterns of perturbation, machine learning models can exploit finer-grained, distributed signals that are not apparent through unsupervised methods.

This study has several important limitations. Our analyses are based on infections in two immortalized human cell lines (Huh7 and Calu-3) under controlled laboratory conditions, which may not capture the complexity of in vivo or clinical infections. Host responses in primary cells, across tissues, or in whole organisms can differ in both timing and magnitude, potentially affecting classification performance. The datasets are further constrained to mid-stage infection (12–24 hpi); earlier or later responses may be weaker, noisier, or biologically distinct. In addition, our work is limited to negative-sense RNA viruses, leaving open how well this approach generalizes to positive-sense RNA or DNA viruses with different replication dynamics. Finally, clinical samples often present additional challenges—mixed infections, inter-individual variability, and technical artifacts such as batch effects or sequencing depth—that may reduce robustness. Although our depletion and enrichment experiments begin to address the role of viral

read abundance, the minimal data requirements and cross-platform reproducibility remain unresolved.

Several directions could strengthen the translational relevance of this framework. Extending benchmarking to additional cell types, primary cultures, and eventually clinical samples will be essential for testing whether host-only signatures generalize beyond controlled in vitro settings. Longitudinal sampling across broader post-infection windows could help determine the earliest time points at which responses become diagnostic and the duration for which they remain informative..

Despite these caveats, our findings provide a clear proof-of-concept. We show that host transcriptomes alone are sufficient to distinguish between infections with diverse RNA viruses, even when viral reads are scarce and when unsupervised clustering fails. This underscores that host responses are not merely generic markers of infection but encode virus-specific, nonlinear signatures that can be captured by machine learning. In practice, such genome-agnostic signals could support early diagnostic triage when primers are unreliable or viral genomes are unknown, complementing sequence-based approaches. These results align with a growing body of work on transcriptomic diagnostics in bacterial sepsis and respiratory infections, suggesting a broader paradigm in which host responses serve as sensitive biosensors of pathogen class. By demonstrating consistent classification across multiple viral families, we add to evidence that systems-level host responses can bridge key gaps between pathogen discovery, outbreak surveillance, and clinical diagnostics.

DATA AVAILABILITY

The data used in this study were obtained from the Sequencing Read Archive (PRJNA1074963) and Gene Expression Omnibus (GSE49840). From PRJNA1074963, we created synthetic permuted, shuffled, enriched, and depleted derivative datasets.

MATERIALS AND METHODS

Data sources.

We analyzed publicly available RNA-seq datasets from human hepatocellular carcinoma (Huh7) and lung adenocarcinoma (Calu-3) cell lines infected with a panel of negative-sense RNA viruses, including influenza A subtypes (H1N1, H5N1, H7N7, H7N9), Ebola virus (EBOV), Marburg virus (MARV), Nipah virus (NiV), Rift Valley fever virus (RVFV), respiratory syncytial virus (RSV), Sandfly fever Sicilian virus (SFSV), and Lassa virus (LASV). Metadata on infection time point (3, 6, 7, 12 or 24 hpi), replicate, and sequencing depth were extracted from SRA and GEO project accessions (PRJNA1074963, GSE49840). We used the first 5,000 reads from each .fastq file, because we saw that the nucleotide k-mer spectra converged around 1000-5000 reads (Suppl. Fig. S1).

Feature generation.

We derived two complementary feature sets from host transcriptomes: (i) normalized differential gene expression (TPM) values and (ii) alignment-free k-mer frequency spectra (k = 1–15). Viral reads were excluded or enriched as controls by re-mapping and sub-sampling.

Classification and clustering.

To evaluate separability of host responses, we applied hierarchical clustering (Euclidean distance, Ward linkage) and Random Forest classifiers. For Random Forests, we used 200 trees, class-balanced stratified 4-fold cross-validation (smallest class n = 4), and default scikit-learn parameters unless otherwise specified.

Controls.

We implemented strict controls, including label permutation, read shuffling, viral read depletion, and viral read enrichment, to confirm that observed signal reflected biologically meaningful host responses.

Reproducibility.

All analyses were performed in Python (v3.11) using pandas, scikit-learn, Biopython, and SciPy. Data accessions are provided above; analysis code will be publicly archived on Zenodo before publication. All of the code is hosted on the following GitHub repository: https://github.com/blatuscaspot/multivirus-classification

Supplemental Material

Acknowledgements

We acknowledge the input of Dr. John Parker at Cornell University in suggesting to use shuffled nucleotides. This research was enabled in part by support provided by Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

Funding

BK is supported by a Canadian Institutes for Health Research Doctoral Foreign Study Award: *Machine Learning to Identify Viral Pathogens through Host Cell Gene Expression Patterns,* and was supported by a gift from the Griffin Foundation, as well as a Fellows-In-Residence Award from Verena (an NSF Biology Integration Institute). TP was funded through award 223764/Z/21/Z from the Wellcome Trust. TP, BK, and SNS are supported by the U.S. National Science Foundation (NSF DBI 2213854).

References

1. Carlson, C. J. *et al.* Pathogens and planetary change. *Nat. Rev. Biodivers.* **1**, 32–49 (2025).
2. Carlson, C. J. *et al.* Climate change increases cross-species viral transmission risk. *Nature* **607**, 555–562 (2022).
3. Plowright, R. K. *et al.* Ecological countermeasures to prevent pathogen spillover and subsequent pandemics. *Nat. Commun.* **15**, 2577 (2024).
4. Gurley, E. S. & Plowright, R. K. A Roadmap of Primary Pandemic Prevention Through Spillover Investigation - Volume 31, Number 8—August 2025 - Emerging Infectious Diseases journal - CDC. doi:10.3201/eid3108.250442.
5. Ruiz-Aravena, M. *et al.* Ecology, evolution and spillover of coronaviruses from bats. *Nat. Rev. Microbiol.* **20**, 299–314 (2022).
6. Guth, S., Visher, E., Boots, M. & Brook, C. E. Host phylogenetic distance drives trends in virus virulence and transmissibility across the animal-human interface. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **374**, 20190296 (2019).
7. Guth, S. *et al.* Bats host the most virulent-but not the most dangerous-zoonotic viruses. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2113628119 (2022).

8. Kawasaki, J., Kojima, S., Tomonaga, K. & Horie, M. Hidden Viral Sequences in Public Sequencing Data and Warning for Future Emerging Diseases. *mBio* **12**, e01638 (2021).

9. Anthony, S. J. *et al.* A Strategy To Estimate Unknown Viral Diversity in Mammals. *mBio* **4**, e00598-13 (2013).

10. Kiselev, D. *et al.* Current Trends in Diagnostics of Viral Infections of Unknown Etiology. *Viruses* **12**, 211 (2020).

11. Lipkin, W. I. The changing face of pathogen discovery and surveillance. *Nat. Rev. Microbiol.* **11**, 133–141 (2013).

12. A Zoonotic Henipavirus in Febrile Patients in China | New England Journal of Medicine. https://www.nejm.org/doi/full/10.1056/NEJMc2202705.

13. Bartekwa, J. W., Abene, E. E., Luka, P. D., Yilgwan, C. S. & Shehu, N. Y. COVID-19 Subclinical Infection and Immunity: A Review. *Niger. J. Med. J. Natl. Assoc. Resid. Dr. Niger.* **30**, 631–636 (2021).

14. Chow, C.-K., Qin, K., Lau, L.-T. & Cheung-Hoi Yu, A. Significance of a Single-Nucleotide Primer Mismatch in Hepatitis B Virus Real-Time PCR Diagnostic Assays. *J. Clin. Microbiol.* **49**, 4418–4419 (2011).

15. Holbrook, M. G. *et al.* Updated and Validated Pan-Coronavirus PCR Assay to Detect All Coronavirus Genera. *Viruses* **13**, 599 (2021).

16. Whiley, D. M. & Sloots, T. P. Sequence variation in primer targets affects the accuracy of viral quantitative PCR. *J. Clin. Virol. Off. Publ. Pan Am. Soc. Clin. Virol.* **34**, 104–107 (2005).

17. Khan, K. A. & Cheung, P. Evaluation of the Sequence Variability within the PCR Primer/Probe Target Regions of the SARS-CoV-2 Genome. *Bio-Protoc.* **10**, e3871 (2020).

18. Taylor, J. *et al.* Novel variant Hendra virus genotype 2 infection in a horse in the greater Newcastle region, New South Wales, Australia. *One Health* **15**, 100423 (2022).

19. Lednicky, J. A. *et al.* Isolation of a Novel Recombinant Canine Coronavirus From a Visitor to Haiti: Further Evidence of Transmission of Coronaviruses of Zoonotic Origin to Humans. *Clin. Infect. Dis.* **75**, e1184–e1187 (2022).

20. Zehr, J. D. *et al.* Recent Zoonotic Spillover and Tropism Shift of a Canine Coronavirus Is Associated with Relaxed Selection and Putative Loss of Function in NTD Subdomain of Spike Protein. *Viruses* **14**, 853 (2022).

21. Pathogens prioritization: a scientific framework for epidemic and pandemic research preparedness. https://www.who.int/publications/m/item/pathogens-prioritization-a-scientific-framework-for-epidemic-and-pandemic-research-preparedness.

22. Ho, T. & Tzanetakis, I. E. Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **471–473**, 54–60 (2014).

23. Woolhouse, M., Scott, F., Hudson, Z., Howey, R. & Chase-Topping, M. Human viruses: discovery and emergence. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 2864–2871 (2012).

24. Chin, P.-J. *et al.* Virus detection by short read high throughput sequencing in a high virus low cellular background. *Npj Vaccines* **10**, 61 (2025).

25. Minicka, J., Zarzyńska-Nowak, A., Budzyńska, D., Borodynko-Filas, N. & Hasiów-Jaroszewska, B. High-Throughput Sequencing Facilitates Discovery of New Plant Viruses in Poland. *Plants* **9**, 820 (2020).

26. Edgar, R. C. *et al.* Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).

27. Ulhuq, F. R. *et al.* Analysis of the ARTIC V4 and V4.1 SARS-CoV-2 primers and their impact on the detection of Omicron BA.1 and BA.2 lineage-defining mutations. *Microb. Genomics* **9**, mgen000991 (2023).

28. Nesti, D. R. *et al.* Development of a semicomprehensive detection method for paramyxoviruses and its validation using Indonesian bats. *Sci. Rep.* **15**, 19154 (2025).

29. Phan, M. V. T. *et al.* Identification of missed viruses by metagenomic sequencing of clinical respiratory samples from Kenya. *Sci. Rep.* **12**, 202 (2022).

30. Kamau, E. *et al.* Recent sequence variation in probe binding site affected detection of respiratory syncytial virus group B by real-time RT-PCR. *J. Clin. Virol.* **88**, 21–25 (2017).

31. Duffy, S. Why are RNA virus mutation rates so damn high? *PLoS Biol.* **16**, e3000003 (2018).

32. Brauburger, K., Boehmann, Y., Krähling, V. & Mühlberger, E. Transcriptional Regulation in Ebola Virus: Effects of Gene Border Structure and Regulatory Elements on Gene Expression and Polymerase Scanning Behavior. *J. Virol.* **90**, 1898–1909 (2016).

33. Merchant, M. *et al.* SARS-CoV-2 variants induce increased inflammatory gene expression but reduced interferon responses and heme synthesis as compared with wild type strains. *Sci. Rep.* **14**, 25734 (2024).

34. Tong-Minh, K. *et al.* A 29-mRNA host response test to identify bacterial and viral infections and to predict 30-day mortality in emergency department patients with suspected infections: A prospective observational cohort study. *Diagn. Microbiol. Infect. Dis.* **111**, 116599 (2025).

35. Sampson, D. *et al.* Blood transcriptomic discrimination of bacterial and viral infections in the emergency department: a multi-cohort observational validation study. *BMC Med.* **18**, 185 (2020).

36. Riquier, S. *et al.* Kmerator Suite: design of specific k-mer signatures and automatic metadata discovery in large RNA-seq datasets. *NAR Genomics Bioinforma.* **3**, lqab058 (2021).

37. Alam, Md. N. U. & Chowdhury, U. F. Short k-mer abundance profiles yield robust machine learning features and accurate classifiers for RNA viruses. *PLoS ONE* **15**, e0239381 (2020).

38. Allesøe, R. L. *et al.* Automated download and clean-up of family-specific databases for kmer-based virus identification. *Bioinformatics* **37**, 705–710 (2020).

39. Mouratidis, I. *et al.* kmerDB: A database encompassing the set of genomic and proteomic sequence information for each species. *Comput. Struct. Biotechnol. J.* **23**, 1919–1928 (2024).

40. Hofmann, N. *et al.* Distinct negative-sense RNA viruses induce a common set of transcripts encoding proteins forming an extensive network. *J. Virol.* **98**, e0093524 (2024).

41. Josset, L., Zeng, H., Kelly, S. M., Tumpey, T. M. & Katze, M. G. Transcriptomic characterization of the novel avian-origin influenza A (H7N9) virus: specific host response and responses intermediate between avian (H5N1 and H7N7) and human (H3N2) viruses and implications for treatment options. *mBio* **5**, e01102-01113 (2014).

42. Ultsch, A. & Lötsch, J. Euclidean distance-optimized data transformation for cluster analysis in biomedical data (EDOtrans). *BMC Bioinformatics* **23**, 233 (2022).

43. Livingstone, M. *et al.* Assessment of mTOR-Dependent Translational Regulation of Interferon Stimulated Genes. *PLoS ONE* **10**, e0133482 (2015).

44. Karousis, E. D., Schubert, K. & Ban, N. Coronavirus takeover of host cell translation and intracellular antiviral response: a molecular perspective. *EMBO J.* **43**, 151–167 (2024).

45. Kroczynska, B., Mehrotra, S., Arslan, A. D., Kaur, S. & Platanias, L. C. Regulation of Interferon-Dependent mRNA Translation of Target Genes. *J. Interferon Cytokine Res.* **34**, 289–296 (2014).

46. Kaza, B. & Aguilar, H. C. Pathogenicity and virulence of henipaviruses. *Virulence* **14**, 2273684 (2023).

47. Rosenke, K. *et al.* UK B.1.1.7 variant exhibits increased respiratory replication and shedding in nonhuman primates. *BioRxiv Prepr. Serv. Biol.* 2021.06.11.448134 (2021) doi:10.1101/2021.06.11.448134.

48. Hansen, F. *et al.* SARS-CoV-2 reinfection prevents acute respiratory disease in Syrian hamsters but not replication in the upper respiratory tract. *Cell Rep.* **38**, 110515 (2022).

49. Ithinji, D. G. *et al.* Multivalent viral particles elicit safe and efficient immunoprotection against Nipah Hendra and Ebola viruses. *NPJ Vaccines* **7**, 166 (2022).

50. Welcome to Interferome. https://interferome.org/interferome/home.jspx.

51. Lu, Y. & Zhang, L. DNA-Sensing Antiviral Innate Immunity in Poxvirus Infection. *Front. Immunol.* **11**, 1637 (2020).

52. Rehwinkel, J. & Gack, M. U. RIG-I-like receptors: their regulation and roles in RNA sensing. *Nat. Rev. Immunol.* **20**, 537–551 (2020).

53. Debnath, J. P. *et al.* Identification of potential biomarkers for 2022 Mpox virus infection: a transcriptomic network analysis and machine learning approach. *Sci. Rep.* **15**, 2922 (2025).

54. SARS-CoV-2 variants induce increased inflammatory gene expression but reduced interferon responses and heme synthesis as compared with wild type strains | Scientific Reports. https://www.nature.com/articles/s41598-024-76401-1.

55. Ryo, M. & Rillig, M. C. Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere* **8**, e01976 (2017).