Machine-learning Analysis of Host Phenotypic Signatures in High-Content Images to Determine Virus Identity in Infected Human Cells

Eric Nguyen[1,2], Joshua Overcash[1,2*], Rahul Harish[1,3#], Nikia Brown[1,2], Xiaoli Chi[1], Rajini Mudhasani[1], LTC Jeffrey Froude[1], Rekha Panchal[1]*

1 – USAMRIID, Frederick, MD. 2 – Chenega Professional and Technical Services. 3 – AEOP, Reston, VA.

Abstract

High-content imaging techniques, Machine Learning Models (MLMs) and Language Learning Models (LLMs) extracted quantifiable phenotypic data from single cells, generated fingerprint signatures that characterized host-virus interactions, and identified viral infections. An Opera-Phenix high-content imager captured 63X-magnification confocal images of MRC5 human lung fibroblasts infected with either Ebola virus (EBOV), Marburg Virus (MARV) or Lassa Virus (LASV). This imaging generated training data necessary for random-forest classification models to evaluate critical cell features associated with infection and identify viruses in blinded synthetic data. Continued development in characterizing host-virus interactions and validating data consistency will be critical for assessing the identity and severity of novel pathogenic threats to the warfighter.

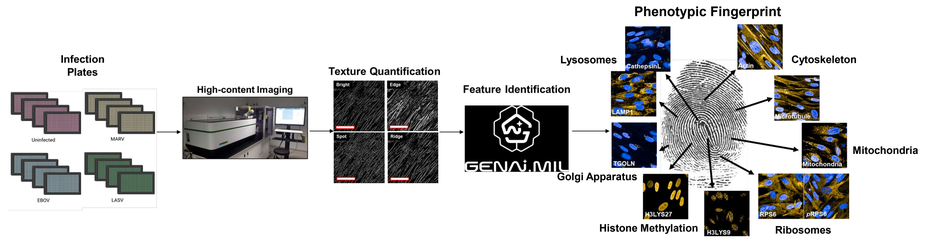
Introduction

High-content imaging tools can characterize cell phenotypic changes that occur in response to pathogenic insults (1-3). Current biosurveillance needs require the ability to precisely detect, analyze, and describe novel pathogenic threats to accelerate the process of treating and protecting the warfighter. Imaging data of cell and organelle morphology can contribute to assessments of infectivity, pathogenic severity and mechanistic activity of novel pathogenic threats when used in complement with genomic sequencing data (3). Here, we used an Opera Phenix high-content imager and Google Gemini to compile human host cell marker texture information from multiple cell organelles and created phenotypic fingerprints of EBOV, MARV, and LASV.

Materials and Methods

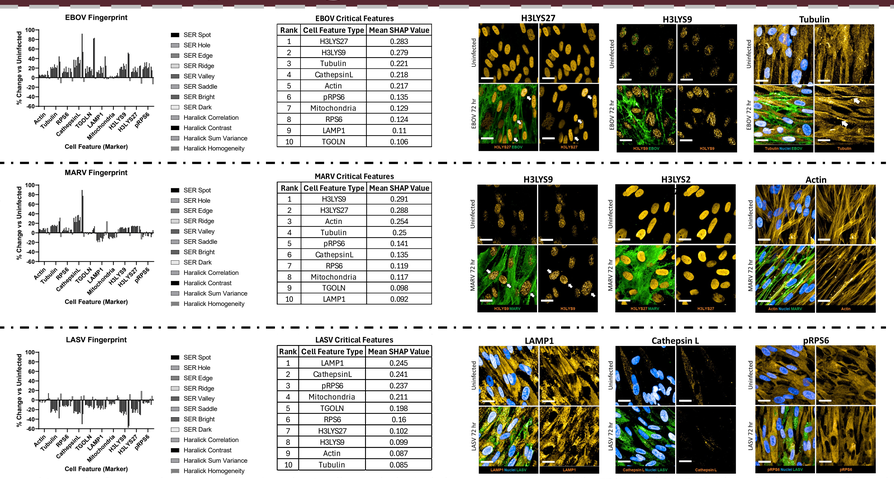
MRC5 human lung fibroblasts were cultured in 384-well imaging plates and infected with EBOV (2.0 MOI), MARV (0.2 MOI) or LASV (0.1 MOI) for 72 hours. After formalin fixation and virus inactivation, the cells were labeled for viral antigen expression via immunofluorescent staining to quantify virus infection. Once viral antigen detection was confirmed positive in > 50% of infected cells and infections resulted in < 20% cell loss, cells were fluorescently labeled for host cell markers related to the cytoskeleton, ribosomes, lysosomes, mitochondria, Golgi apparatus and methylated nuclear histones. The cells were imaged using an Opera Phenix high-content confocal imaging platform using a 63X water immersion objective, and the images were filtered to quantify Spot/Edge/Ridge (SER) and Haralick texture data for each marker within infected and uninfected cells. After compiling an overall training dataset of >720,000 cells, average percent change in host marker texture properties relative to uninfected cells was compiled across all markers to generate preliminary fingerprint signatures for each virus. Host marker texture data was further analyzed using Random Forest Classifier models in Google Gemini (GenAI.Mil) to rank key changes in host marker textures that are associated with each virus type. Additionally, the Random Forest Classifier model was adapted to identify the specific virus infecting cell populations in blinded synthetic infection test plates.

Workflow, Phenotypic Fingerprints



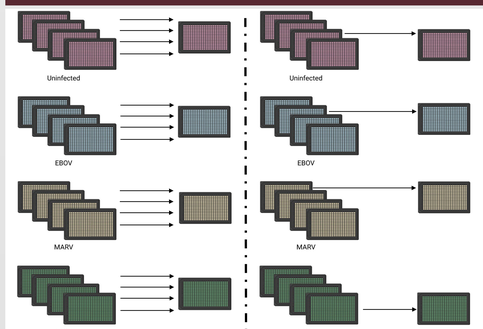
MRC5 human lung fibroblasts were seeded as monolayers in 384-well plates and infected with one of three BSL-4 viruses. The viruses were inactivated after 72 hours of incubation and labeled for one of ten host cell markers to gauge viral phenotypic activity. The Opera Phenix high-content imaging system imaged the cells at 63X confocal magnification and Revvity's Harmony image analysis software extracted quantitative texture data on a single-cell basis. Training data was separated between cells in uninfected wells versus cells in infected cells and uploaded to Google Gemini via GenAI.mil to discern cell fingerprint signatures of the viral infections. *Scale Bar: 100 µm. Blue: Nuclei. Orange: Host Cell Marker. (Fingerprint Image: Wikipedia.org)*

Results: Unique fingerprints of viral infections



MRC5 cells in wells exposed to viral infections demonstrated unique host cell morphological changes in texture. While EBOV and MARV had similar texture profiles related to several markers such as actin, tubulin, and cathepsin L, LASV generated a more divergent signature. This is potentially due to familial similarities and differences between filoviruses (EBOV and MARV) and arenavirus (LASV) replication. Google Gemini ranked critical features that significantly affected Random Forest classification. During ranking, the positive and negative contributions to the Random Forest model's decisions were quantified as SHapley Additive exPlanations (SHAP) values (4). Several outcomes, such as spot appearance in H3Lys27 and H3Lys9 nuclear staining (H3Lys27, H3Lys9, arrowheads), Tubulin and Actin bundle formation (Tubulin, arrowheads), and LAMP1 distribution can be confirmed by visual inspection of the images. *Scale Bar: 20 µm.*

Results: Virus Identification



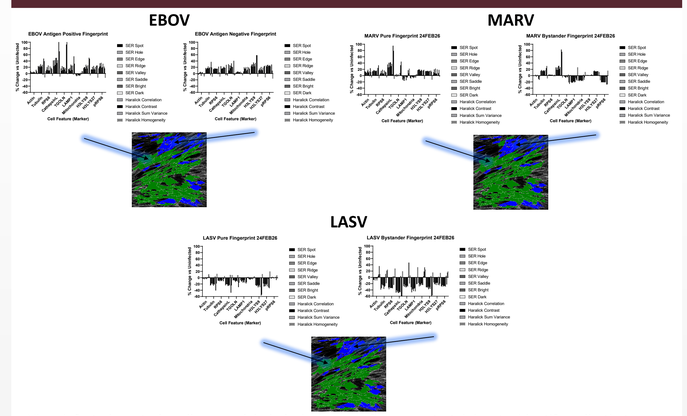
Method A: Randomly sample ~25% of cells from full, combined dataset and make a **synthetic test plate** for each infection model.

Method B: Take one of the four training data plates and make a **synthetic test plate** for each infection model

Ground Truth	Final Prediction	Result	EBOV Prob.	LASV Prob.	MARV Prob.	Uninfected Prob.	
EBOV	EBOV	Correct	98.20%	0.10%	1.60%	0.10%	Method A
MARV	MARV	Correct	1.60%	0.10%	98.40%	0.00%	
Uninfected	Uninfected	Correct	0.10%	0.10%	0.30%	99.50%	
LASV	LASV	Correct	0.20%	99.20%	0.40%	0.10%	
LASV	LASV	Correct	0.20%	99.20%	0.40%	0.10%	Method B
LASV	LASV	Correct	0.20%	99.20%	0.50%	0.10%	
LASV	LASV	Correct	0.30%	99.20%	0.50%	0.10%	
LASV	Uninfected	Incorrect	0.30%	39.80%	0.50%	58.50%	
EBOV	EBOV	Correct	98.20%	0.10%	1.60%	0.10%	
EBOV	EBOV	Correct	98.20%	0.10%	1.80%	0.00%	
EBOV	EBOV	Correct	98.10%	0.10%	1.80%	0.00%	
EBOV	Uninfected	Incorrect	39.40%	0.20%	1.80%	59.00%	
MARV	MARV	Correct	1.60%	0.10%	98.40%	0.00%	
MARV	MARV	Correct	1.70%	0.10%	98.40%	0.00%	
MARV	MARV	Correct	1.80%	0.10%	98.30%	0.00%	
MARV	MARV	Correct	1.80%	0.10%	98.30%	0.00%	

Training on reference datasets (N=4 infection plates) and identification of blinded synthetic plates resulted in an overall 87.5% accuracy of virus identification. Synthetic datasets were generated either by sampling approximately 25% of a pooled dataset of single cells (Method A), or sampling replicate plates to account for plate-to-plate variation in training data (Method B). In cases of incorrect guesses, the second-plate guess was the correct ground-truth virus identity by probabilistic distribution. (*Schematics: Biorender.*)

Results: Fingerprints without Antigen Expression



Virus-infected control wells were labeled via immunofluorescent staining to identify cells that were positive or negative for viral antigen. Emerging fingerprint signatures were evident even in cells that were negative for viral antigen labelling, suggesting that critical host information for virus identification may be obtainable even if viral antigen transcription and translation occurred below the limit of detection by specific antibodies. *(Green: Representative antigen positive cells. Blue: Representative antigen-negative cells)*

Conclusions

Future implementation of these characterization techniques across multiple time points, varying virus clades, and multiple human cell types will generate mechanistic insights of known and novel emerging biothreats that will lead to rapid identification of countermeasures. These techniques may be applied as practical methods for assessing the infectivity and threat level of novel emerging threats and vetting mechanistic activity of potential medical countermeasures against pathogenic threats when used in complement with genomic sequencing.

References

(1) Bray, M.A., et al., Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc, 2016. 11(9): p. 1757-74.

(2) Seal, S., et al., Cell Painting: a decade of discovery and innovation in cellular imaging. Nat Methods, 2025. 22(2): p. 254-268.

(3) Ng, T.L., et al., High-Content Screening and Computational Prediction Reveal Viral Genes That Suppress the Innate Immune Response. mSystems, 2022. 7(2): p. e0146621.

(4) Lundberg, S.M. and S.I. Lee, A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (Nips 2017), 2017. 30.

Acknowledgements

LTC Jeffrey Kugelman, USAMRIID Molecular Biology Division

This research was funded by the Defense Threat Reduction Agency under project numbers CB11074, CB11485, and the Medical Research & Development Command Military Infectious Diseases Program under project number MI240143.

*Contractor – this does not constitute an endorsement by the US Government of this or any other contractor.

Summer Student – AEOP

The use of either trade or manufacturers' names in this report does not constitute an official endorsement of any commercial products. This report may not be cited for purposes of advertisement.

To learn more, visit:
https://usamriid.health.mil/