Pacific Northwest
NATIONAL LABORATORY

PNNL-34828

Threat Agnostic Virulence Assessment of Pathogens

September 2023

Becky M Hess
Lisa Bramer
Trinidad Alfaro

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-34828

Threat Agnostic Virulence Assessment of Pathogens

September 2023

Becky M Hess
Lisa Bramer
Trinidad Alfaro

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Virulence assessment of new, emerging, and engineered pathogens is critical to mounting an appropriate response to a biothreat agent. The capacity of the pathogen to colonize human and harm tissues must be characterized to understand pathogenicity pathways and optimize diagnosis and treatment of resulting disease. Respiratory pathogens are of interest because they can have high transmissibility rates, as observed with the SARS-CoV-2 virus, the causative agent of Covid-19. Current technologies are insufficient to assess threats due to their reliance on systems with only one cell type and on sequencing the pathogen. However, it is known that sequence is not an accurate predictor of function, and sequencing can be unreliable for newly emerged or engineered pathogens. An ideal system would consist of relevant epithelial cell types and an assay sensitive enough to detect changes in host responses that do not rely on DNA sequencing. We chose a system consisting of host lung epithelial cells that can be used to assess the virulence of unknown respiratory pathogens. We interrogated pathogens using this model and assess features of pathogenicity. Our objective is to leverage PNNLs strengths in tissue engineering and proteomics capabilities to build a multiple reaction monitoring (MRM) or parallel reaction monitoring (PRM) liquid chromatography-tandem mass spectrometry assay for human host cell proteins whose abundance is influenced by infection. These responses can were then assessed for relative virulence using pathogen agnostic signatures. When confronted with a pathogen, cells activate dedicated signaling pathways, typically through phosphorylation of regulatory proteins and downstream activation of host cell networks. As a model system, we used two different pathogens that can cause lung disease: the bacteria, Pseudomonas aeruginosa, and one virus, influenza A H1N1. As nonpathogenic controls we included commensal bacteria. Our main hypothesis was that pathogens elicit a differential early host response than nonpathogens within 6 to 18 hours post infection. We also expected that we will be able to differentiate between bacterial and viral pathogens given that the pathogen activated molecular pathways vary between bacterial and viral pathogens. To establish the appropriate time frame for taking measurements for proteomics, we ran imaging assays to determine the onset of cell death, and selected time points upstream of these extreme cytotoxic effects. Expected outcomes from the project include the development of proteomics assays that enable assessment of whether or not a respiratory pathogen has the potential to be virulent in the human airway. Our approach is agent agnostic in that any agent that influenced response pathways in human cells could be characterized. This platform could form a key part of a response to a novel/emerging pathogen and the molecular basis of its pathogenicity would be directly revealed. By understanding the basis of pathogenesis, specific and rapid mitigation response and targeted treatments could be deployed on a shorter timescale compared to traditional methods. The final outcome of this work will be the development of a framework for rapid virulence assessment of bacterial and viral pathogens that can be translated to other tissue types.

Acknowledgments

This research was supported by the **National Security Mission Seed**, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). Part of the research was performed using the Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility sponsored by the DOE's Office of Biological and Environmental Research and located at PNNL. PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL0-1830.

Contents

1.0 Introduction

Two of the major goals of the National Biodefense Strategy are: 1) Enable risk awareness to inform decision making across the biodefense enterprise; 2) Rapidly respond to limit the impacts of biological incidents. Virulence assessment of new, emerging, and engineered pathogens is critical to mounting an appropriate response to a biothreat agent. Respiratory pathogens are of interest because they can have high transmissibility rates. When confronted with a pathogen, cells activate dedicated signaling pathways, typically through phosphorylation of regulatory proteins and downstream activation of host cell networks. These changes can be measured through proteomics analysis to determine the abundance of change in these proteins and provide confirmation of the pathogen's virulence (Figure 1). Current technologies are insufficient to assess threats due to their reliance on on genome sequencing of the pathogen. However, it is known that genomes are not an accurate predictor of function (Alberts et al). In contrast, a targeted multiplexed proteomics assay could specifically detect activation of over 100 target proteins, allowing for readouts from multiple pathways.

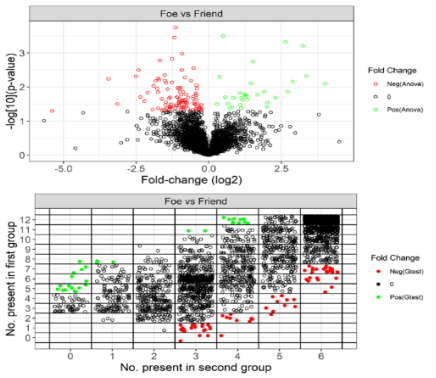


Figure 1. Differentially expressed pathogen activated molecular pathway proteins expressed in a "friendly" bacterium, S. epidermidis, compared to a "foe" bacterium, P. aeruginosa, in the lung co-culture system at 10 hours post infection.

2.0 Objective and Approach

2.1 Objective

We will leverage existing PNNL capabilities to develop a threat agnostic proteomics assay for detecting respiratory pathogens and predicting their virulence. Our central hypothesis is that pathogens will elicit a differential early host response compared to nonpathogens within 6 to 48 hours of infection, enabling a rapid threat response. Biomarkers are measurable indicators of biological processes such as normal metabolism, disease-state, infection, or an environmental exposure. The identification of host molecular biomarkers in response to pathogen exposure, especially in the period before clinical signs and symptoms are observable, is essential to develop rapid diagnostics and therapeutics for disease. Mass spectrometry (MS) based proteomic approaches are one of the most powerful tools to identify potential disease specific molecular biomarkers (e.g. enzymes, transporters) in complex biological samples from in vitro cell-based assays to in vivo samples from hosts. MS-based biomarker proteomics can be divided into two main steps: discovery and targeted proteomics. The “discovery phrase” uses global proteomic approaches to identify potential protein targets of interest. Targeted proteomics is a hypothesis-driven approach that monitors specific protein targets with high specificity and sensitivity for accurate quantitation across many samples.

In the discovery phase, global/shot-gun proteomic approaches will be used such as iTRAQ and TMT which use isobaric tags to label peptides within a sample with a specific “barcode” allowing multiplexing of up to 12 different samples enabling parallel analysis across multiple proteomes (Figure 2). In this approach identical peptides with different isobaric tags co-migrate during MS separations enabling downstream analysis of differential protein expressions between samples using existing methods developed by the Chemical and Biological Signatures group. By examining differential protein expression between samples exposed to either a pathogen or non-pathogen a candidate list of potential protein biomarkers for host responses to pathogen infection can be determined; general inflammatory markers will be carefully reviewed to ensure a general response to a microbe does not serve as a confounding factor in analysis. In addition to performing in-house experiments, potential protein targets and their peptide signatures will also be gleaned from literature-based resources including primary literature and targeted biomarker databases including SRMAtlas (Kusebauch et al., 2016) and PASSEL (Farrah et al., 2012).

Targeted proteomic approaches include selected reaction monitoring (SRM; also known as multiple reaction monitoring: MRM) and parallel reaction monitoring (PRM). Targeted proteomics provides sensitive and accurate quantitation of up to hundreds of specific proteins from samples with detection down to the low nanograms per mL. This selectivity and sensitivity are not achievable by current global proteomics approaches. Both SRM and PRM utilize isotope-labeled synthetic peptides that are spiked into the sample as internal standards and quantitation is based on the ratio of intensities of the specific peptides in the sample and their matching internal standards. Part of developing an SRM assay is developing and optimizing of these isotope-labeled synthetic peptides (often needing 1 to 6) for each protein of interest. Some of this development can be leveraged by using internal peptide standards from biomarker databases (i.e. SRMAtlas) or previous work. Data processing and analysis of SRM proteomics is well established with the use of tools such as Skyline (MacLean et al., 2010). Once protein targets are quantified, statistical approaches including regressions can be used to determine differential expression of the candidate biomarkers. We will implement our established mass spectrometry (MS)-based proteomics quantitative data analyses pipelines including outlier

sample detection, normalization, and differential statistics, which are implemented in standardized workflows in R (Stratton, et al.). Univariate differential expression discovery, of friend vs foe samples, will fit models independently for each protein using generalized linear mixed models. These results can be used to quantify baseline discriminatory efficacy. Classification models will be used for the identification of biomolecule signatures that differentiate friend vs foe samples. For the biosignature discovery, statistical classification approaches such as elastic nets (Zou et al. and Friedman et al.) coupled with cross-validation (Friedman et al) will be used to discriminate friend vs foe. These methods also select the features (i.e. proteins) that most strongly contribute to the predicted class and identify the optimal set of proteins for predicting friend vs foe for unseen samples. Classification metrics such as accuracy, F1 score, and area under the curve (AUC) will be used to measure models' predictive performance (Friedman et al). These classifiers can be used to predict disease state in other samples. If successful, the proposed co-culture system/targeted proteomics platform would provide a method to characterize the threat posed by novel or engineered pathogens. More advanced machine learning methods can also be explored to develop classifiers based on expression patterns of different proteins.

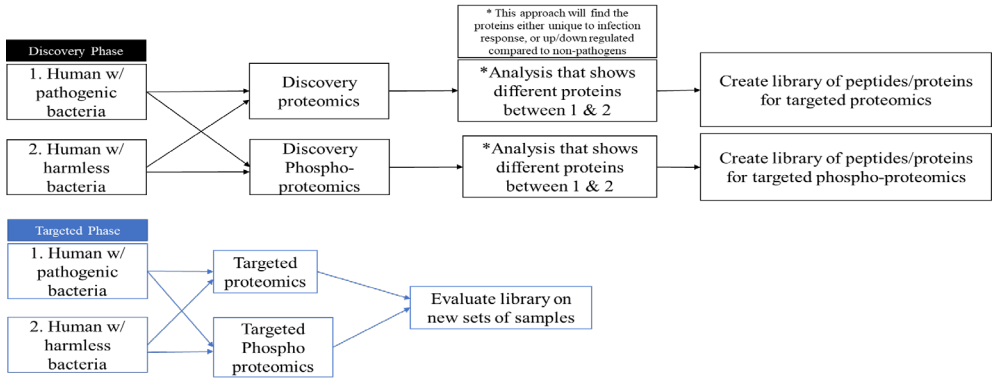


Figure 2. Summary of discovery and targeted phase of proteomics and phospho-proteomics analysis with replicate analysis planned in each phase.

2.2 Proposed Approach

We propose to interrogate respiratory pathogens using a lung cell culture model (A-549 cells) and assess features of pathogenicity (i.e., the ability to harm host cells and niche find). As a model system in Year 1, we will use two different pathogens that can cause human lung disease: the bacteria, *Pseudomonas aeruginosa*, and a virus such as influenza A H1N1. As non-pathogenic controls we will include the commensal bacteria, *S. epidermidis*, that do not confer pathogenicity. Once we have established signatures using pathogenic controls, we will expand our data set to include both high pathogenicity and low pathogenicity viruses. The predictive value of the assay will be critically dependent on testing a range of pathogen types to ensure a broad set of signatures is obtained using this approach. Examples of "highly pathogenic" bacteria include *P. aeruginosa, Y. pseudotuberculosis, S. aureus,* and *K. pneumoniae*. With respect to viruses, "low pathogenic" viruses include HCoV-NL63 and "highly pathogenic" viruses include H5N1 influenza.

During the discovery phase, samples will be taken over an infection time course and analyzed using a TMT proteomics assay to identify and measure signaling pathway activity associated with pathogenesis. We will use the lung culture system to identify the interplay between host cells and the pathogens that contribute to pathogenesis. This approach builds on PNNL's experience and successful execution in the DARPA Friend or Foe program, as well as our expertise in targeted mass spectrometry. PNNL will develop a targeted proteomics approach

using multiple reaction monitoring (MRM) or parallel reaction monitoring (PRM) liquid chromatography-tandem mass spectrometry assay for human host cell proteins whose abundance levels are indicative of pathogenicity. MRM enables targeted and sensitive monitoring of the abundance of hundreds of protein targets from samples with detection down to low nanograms per mL; a more sensitive approach compared to global proteomics methods. We will leverage this approach based upon a growing understanding of how human cells respond to bacterial and viral pathogens. One aspect that MRM will elucidate is the temporal regulation of post translation modifications (PTMs) such as phosphorylation in the host proteome response. These modifications of existing host proteins, as opposed to protein replication, may shed light on the earliest stages of infection. A targeted proteomics assay will be developed to detect activation of these host cell networks. A multiplexed assay could specifically detect activation of over 100 targets for early host cell response proteins, allowing for readouts from multiple pathways. By comparing target protein abundance in human cell samples that have been exposed to infectious and non-infectious agents, we will develop an analysis pipeline that will lead to a set of decision criteria for classifying organisms as infectious/virulent or avirulent. There is also a potential that this project will identify suites of protein biomarkers that can differentiate between bacterial and viral infections. This method would be applicable to any respiratory pathogen, even if it is unknown or engineered. **Our approach would be agent agnostic in that any agent that influenced the targeted immune response pathways in human cells could be characterized.** This platform could form a key part of a response to a novel pathogen—in addition to identification, the molecular basis of its pathogenicity would be directly revealed. **By understanding the basis of pathogenesis, specific and rapid mitigation response and targeted treatments could be deployed on a shorter timescale compared to current methods.**

3.0 Results and Discussion

3.1 Time Point Selection for Challenge Assays

3.1.1 Bacterial Challenge Assays, Time Point Selection for Proteomics Analysis

We used a high throughput imaging system (the Cytation5 coupled to a BioSpa automated incubator) to determine the time points for proteomics analysis of host cells. The hours post infection (hpi) for collection of the host cells following the bacterial challenge with pathogenic *Pseudomonas aeruginosa* were determined by manual review of lung cell phenotype over time. The goal was to collect healthy cells (i.e., uninfected controls), challenged cells (early in infection), and stressed cells (mid infection). As shown in Figure 3, the 0 hpi, 2 hpi, and 6 hpi samples, respectively, met these criteria. At 0 hpi, the cells have tight junctions and are adherent, which are all indicators of healthly cell culture. At 2 hpi, the cells begin to lose tight junctions, which is an indicator of early infection, and at 6 hpi, the cells are rounding and beginning to lose adherence, which indicates active infection.

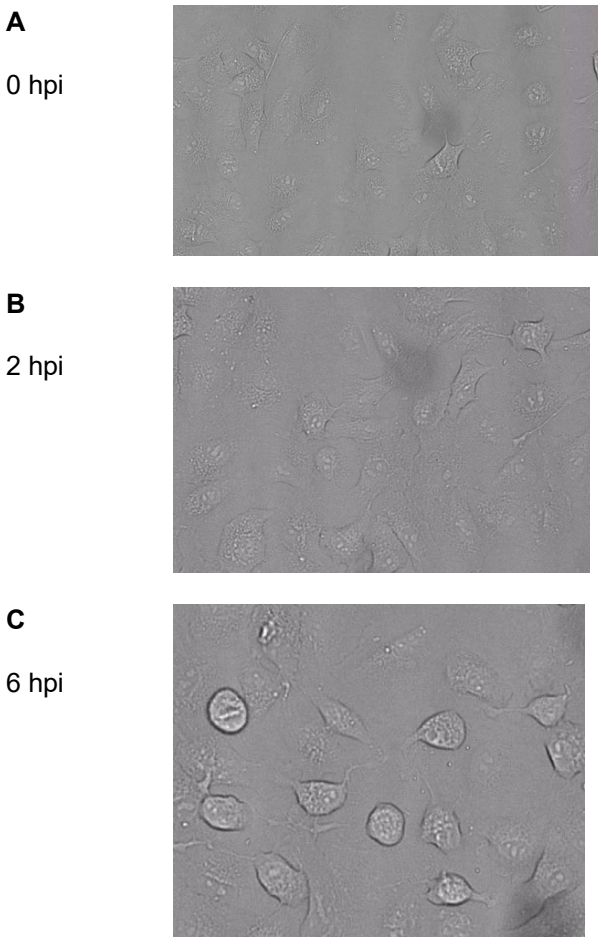


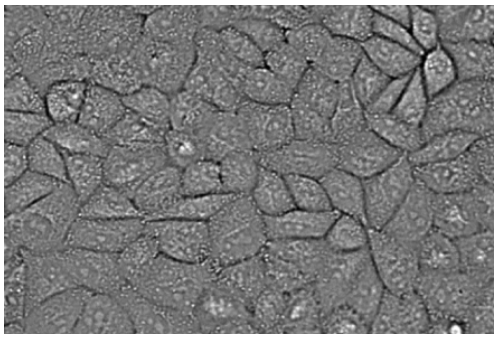
Figure 3. Lung cells (A-549) exposed to the pathogenic bacteria, *P. aeruginosa* at A) 0 hpi (uninfected), B) 2 hpi, and C) 6 hpi. Over time, the cells begin to lose tight junctions between each other, begin to round, and ultimately lose adherence.

3.1.2 Viral Challenge Assays, Time Point Selection for Proteomics Analysis

The hours post infection (hpi) for collection of the host cells following the viral challenge with pathogenic Influenza A were determined by manual review of lung cell phenotype over time. The goal was to collect healthy cells (i.e., uninfected controls), challenged cells (early in infection), and stressed cells (mid infection). As shown in Figure 4, the 0 hpi, 5 hpi, and 18 hpi samples, respectively, met these criteria. At 0 hpi, the cells have tight junctions and are adherent, which are all indicators of healthly cell culture. At 5 hpi, the cells begin to lose tight junctions, which is an indicator of early infection, and at 18 hpi, the cells are rounding and beginning to lose adherence, which indicates active infection.

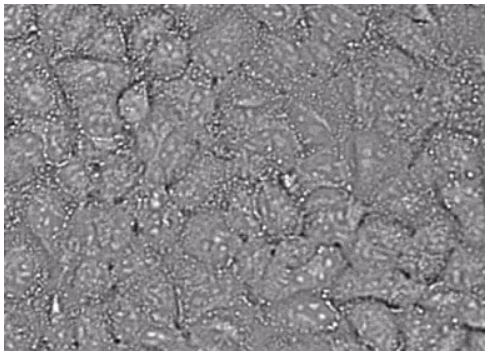
A

0 hpi



B

5 hpi



C

18 hpi

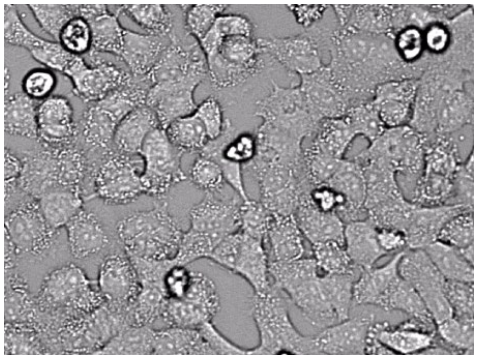


Figure 4. Lung cells (A-549) exposed to the pathogenic virus, Influenza A at A) 0 hpi (uninfected), B) 5 hpi, and C) 18 hpi. Over time, the cells begin to lose tight junctions between each other, begin to round, and ultimately lose adherence.

3.2 Global Proteomics Analysis

We first ran global proteomics analysis of the lung cells challenged with the commensal bacteria, *S. epidermidis* (SE) and the pathogenic bacteria, *P. aeruginosa* (PA). We also included

an uninfected control (C). We prepared four biological replicates of each sample type. The normalized peptide data from these assays is provided in Figure 5.

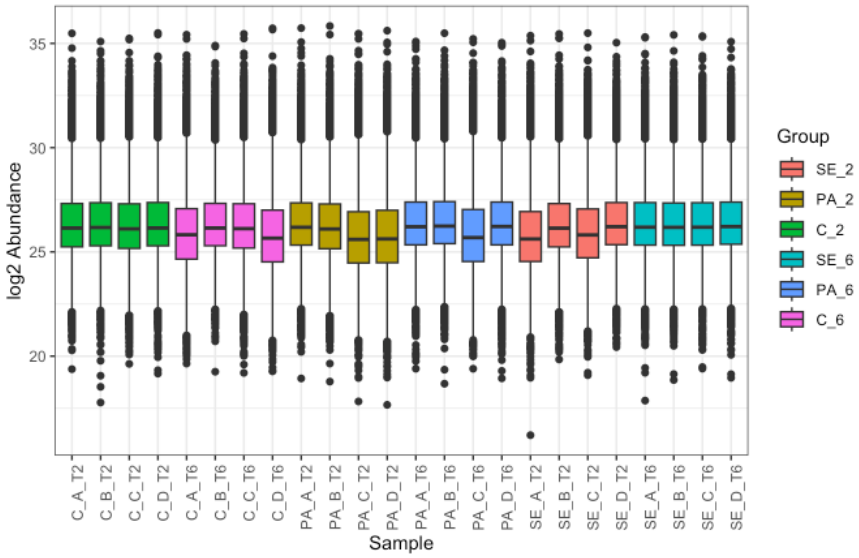


Figure 5. Boxplots of reference standardized normalized peptide data ordered by group.

After normalization, we conducted pairwise comparisons at the 2 hpi and 6 hpi time points for the commensal and pathogenic bacteria (SE vs. PA). As shown in Figure 6, four proteins were highly varied in the two time points: SODM, HDHD1, GLYC, and IGBP1.

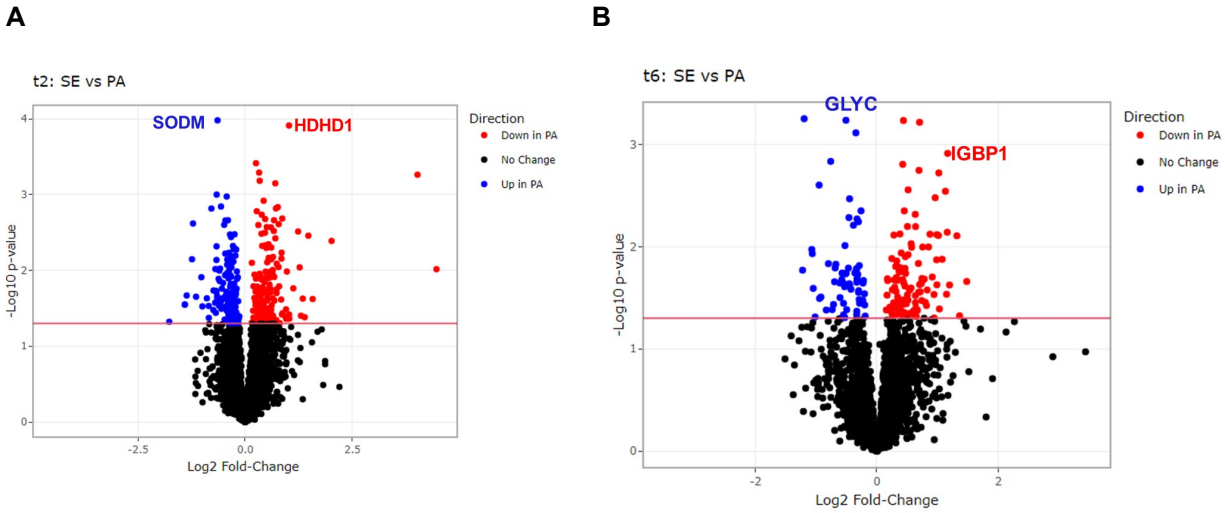
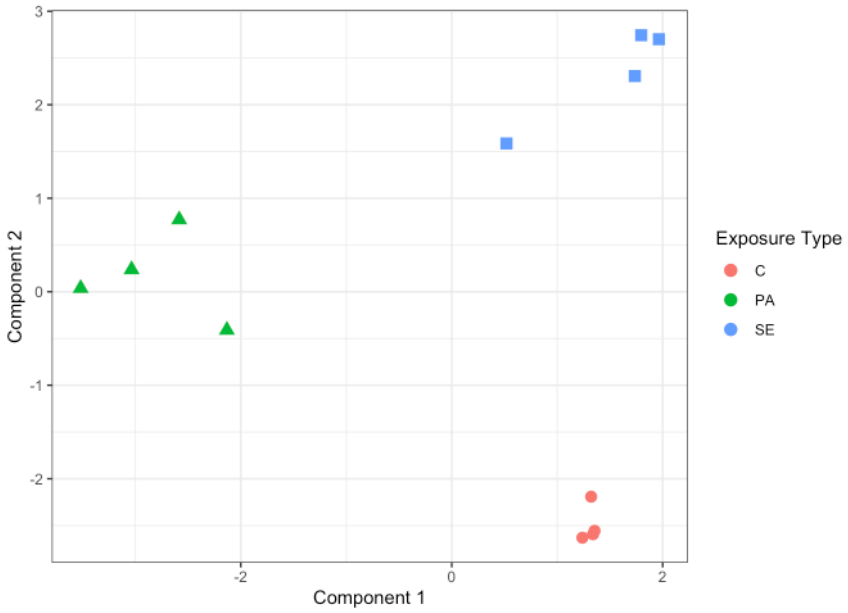


Figure 6. Volcano plots of the pairwise comparison of the commensal bacteria, *S. epidermidis* (SE) and the pathogenic bacteria, *P. aeruginosa* (PA) at A) 2 hpi and B) 6 hpi.

Based on the number of significant proteins identified at each time point in the pairwise comparison, we chose to conduct machine learning using sparse partial least squares discriminant analysis (sPLS-DA). This approach enables the identification of a multivariate signature of proteins to discriminate samples exposed to pathogenic bacteria versus commensal bacteria versus the uninfected control samples (Figures 7 and 8).

A



B

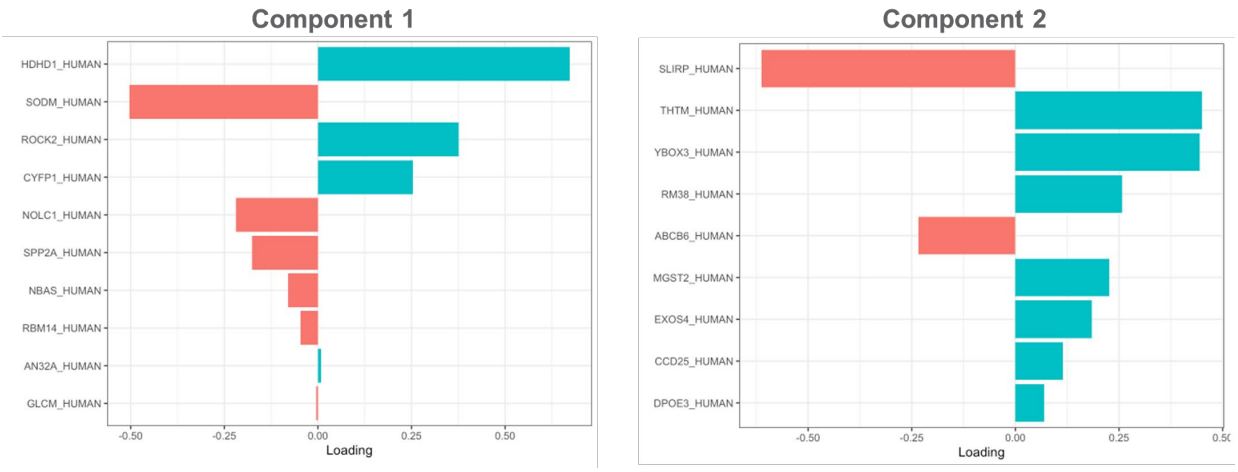
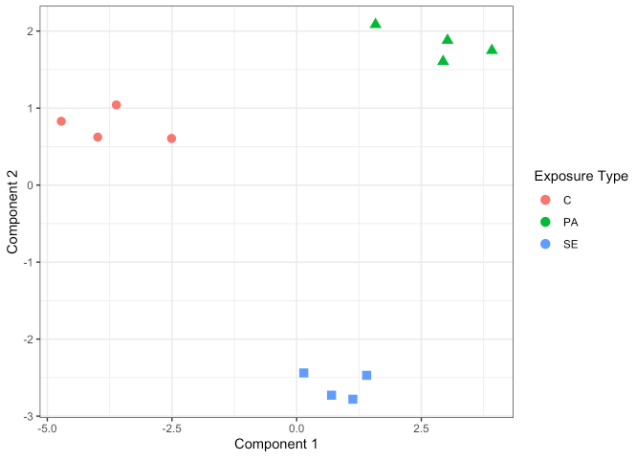


Figure 7. Multivariate comparison of the control to SE to PA from global proteomics data at 2 hpi. The most critical proteins responsible for the observed clustering (A) provided (B).

As shown in Figures 7 and 8, at 2 hpi and 6 hpi the exposure time results in significant variability with respect to expression of specific proteins. These data indicate that the kinetics of infection are critical to determining appropriate biomarkers for a targeted assay.

A



B

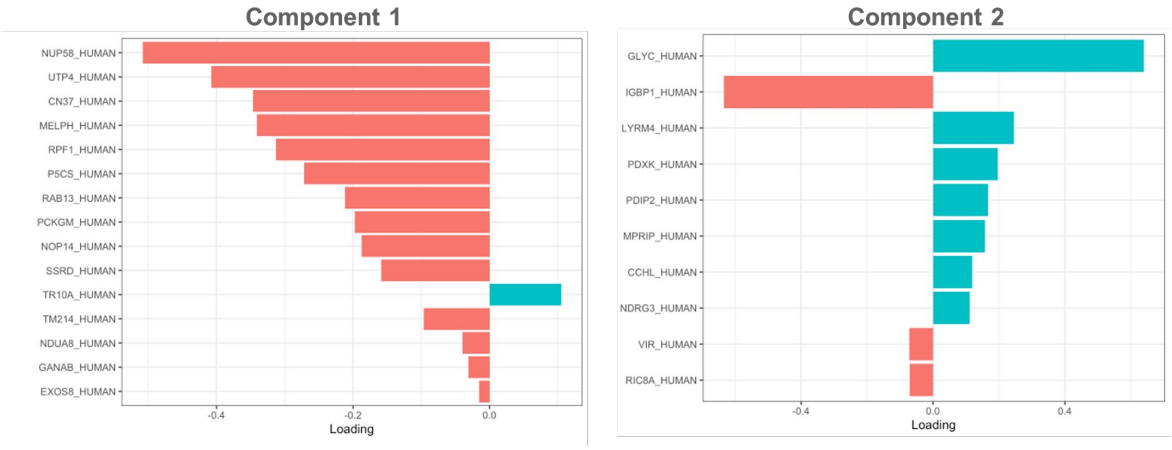


Figure 8. Multivariate comparison of the control to SE to PA from global proteomics data at 6 hpi. The most critical proteins responsible for the observed clustering (A) provided (B).

The biological relevance of each of these proteins was investigated to determine which pathways are involved in responding to infection at early and mid-stages of the challenge assay. As shown in Figures 9 and 10, the function of each protein identified in the multivariate analysis is identified and a short analysis is provided given the collection of the proteins.

Time Point	Protein	Function
2 hpi	HDHD1	RNA Degradation
2 hpi	SODM	Destroys superoxide anion radicals
2 hpi	ROCK2	Stress fiber and focal adhesion formation
2 hpi	CYFP1	Translational repression
2 hpi	SLIRP	Nuclear receptor corepressor
2 hpi	THTM	Acts as an antioxidant
2 hpi	YBOX3	Binds to the GM-CSF promoter and seems to act as a repressor
2 hpi	RM38	Mitochondrial translation

- What does this mean?
  - Cells infected with PA are responding to free radicals, repressing gene expression and protein expression, and beginning signaling for immune cell support
  - Results are consistent with early infection

Figure 9. Biological relevance of the critical proteins identified by machine learning at 2 hpi.

The variability in biological relevance is evident between early and mid-infection based on the host response.

Time Point	Protein	Function
6 hpi	TR10A	Receptor for the cytotoxic ligand TNFSF10/TRAIL
6 hpi	NUP58	A complex required for the trafficking across the nuclear membrane
6 hpi	UTP4	Ribosome biogenesis factor
6 hpi	CN37	May participate in RNA metabolism
6 hpi	GLYC	Interconversion of serine and glycine
6 hpi	IGBP1	Associated to surface IgM-receptor; may be involved in signal transduction
6 hpi	LYRM4	First step of the mitochondrial iron-sulfur protein biogenesis
6 hpi	PDXK	Cofactor metabolism

- What does this mean?
  - Late infection corresponds to iron scavenging by the pathogen
  - Metabolism is disrupted
  - Death-inducing signaling complex

Figure 10. Biological relevance of the critical proteins identified by machine learning at 6 hpi.

3.3 Phospho-proteomics Analysis

The samples used for phosphor-proteomics analysis were identical to those used for global proteomics analysis. The goal of the phospho-proteomics data is to gain an understanding of the pathways that are activated by kinases or deactivated by phosphatases. Similar to the global proteomics analysis, we first normalized the peptide data for this data set (Figure 11).

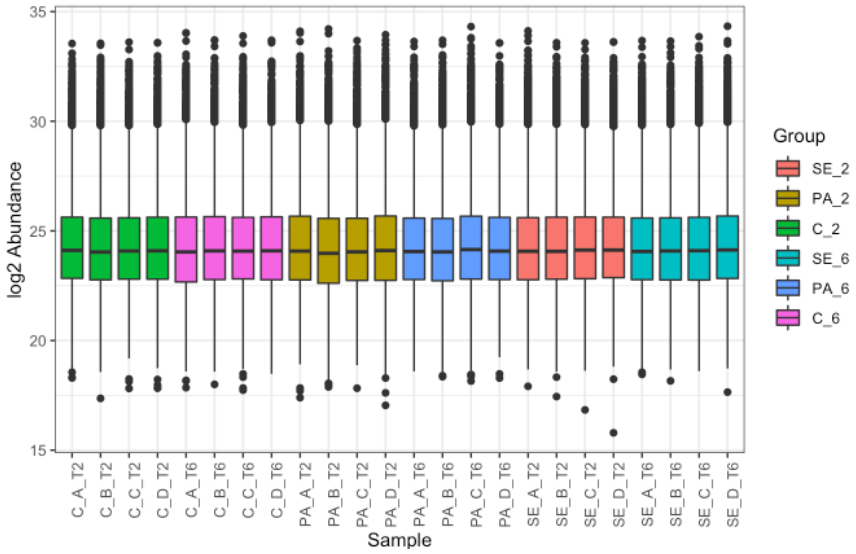


Figure 11. Boxplots of reference standardized normalized peptide data ordered by group.

After normalization, we conducted pairwise comparisons at the 2 hpi and 6 hpi time points for the commensal and pathogenic bacteria (SE vs. PA). As shown in Figure 12, multiple proteins were highly varied in the two time points.

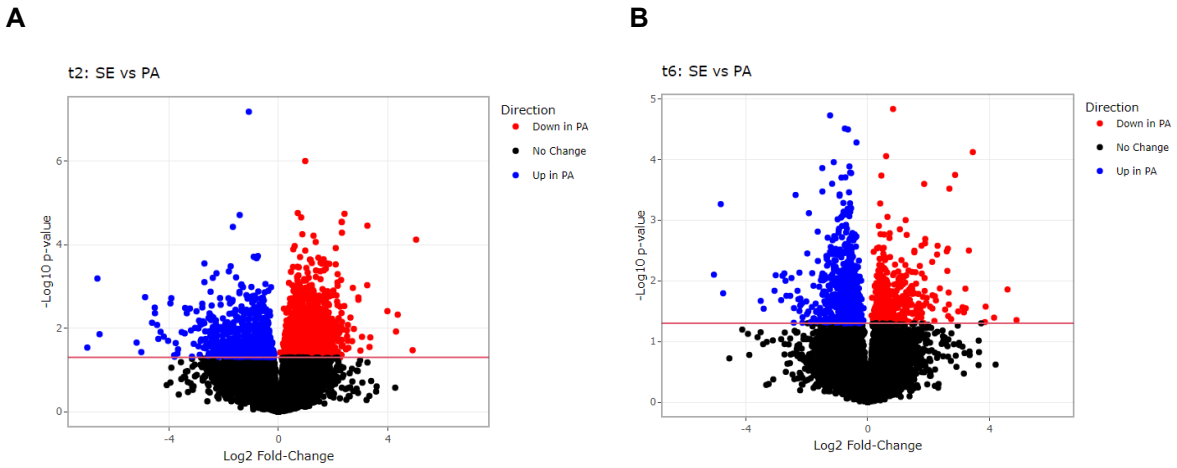


Figure 12. Volcano plots of the pairwise comparison of the commensal bacteria, *S. epidermidis* (SE) and the pathogenic bacteria, *P. aeruginosa* (PA) at A) 2 hpi and B) 6 hpi.

We conducted pathway analysis from these data and identified four critical pathways that require investigation: Interleukin-7 signaling, defective pyroptosis, RHO GTPase activation of protein kinase networks, and oxidative stress induced senescence.

Based on these promising data, we conducted additional challenges assays with lung cells challenged with viruses and additional commensal and pathogenic bacteria to improve the robustness of the data sets. Unfortunately, these samples were lost during the proteomics preparation process and require them to be reprepared. Due to this issue, this LDRD project is continuing into FY24. Future directions are provided in the Appendix.

4.0 References

Alberts B, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Studying Gene Expression and Function. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26818/
Hess BM, et al. (2019) An integrated experimental-computational approach for predicting virulence in New Zealand white rabbits and humans following inhalation exposure to *Bacillus anthracis* spores. PLOS ONE 14: e0219160.
Farrah, T., et al. (2012). PASSEL: The Peptide Atlas SRM experiment library. *PROTEOMICS, 12*(8), 1170-1175.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.
Friedman, J., et al. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.
Kusebauch, U., et al. (2016). Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome. *Cell, 166*(3), 766-778.
MacLean, B., et al. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics, 26*(7), 966-968.
Stratton, K. G., et al. (2019). Pmartr: Quality control and statistics for mass spectrometry-based biological data. *Journal of proteome research*, *18*(3), 1418-1425.
Zou, H., et al. (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. *JR Stat Soc Ser B*, *67*, 301-20.

Appendix A – Future Directions

In order to improve the robustness of the data already collected in this effort, we plan to conduct challenge assays followed by global and phospho-proteomics with Influenza A, HCoV-NL63, RSV, *K. pneumoniae,* and *P. putida*.

Once these data have been generated, we will use the same machine learning approach to identify the critical nodes in the pathways that will serve as biomarkers for developing a targeted mass spectrometry assay for threat agnostic virulence assessment.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov