

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Combined host and microbial metagenomic next-generation sequencing: Applying integrated analysis approaches for a comprehensive evaluation of infectious disease response to inform diagnosis, surveillance, and treatment

Permalink

<https://escholarship.org/uc/item/3x7795xq>

Author

Kalantar, Katrina Louise

Publication Date

2019

Peer reviewed|Thesis/dissertation

Combined host and microbial metagenomic next-generation sequencing:
Applying integrated analysis approaches for a comprehensive evaluation of infectious
disease response to inform diagnosis, surveillance, and treatment
by

Katrina Kalantar

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

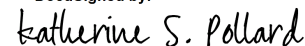
Approved:

DocuSigned by:

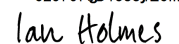
50AE848EEDE2449...

Joe DeRisi

Chair

DocuSigned by:


Katherine S. Pollard

DocuSigned by:

CF5F2DDAA9CB478...

Ian Holmes

Committee Members

Copyright 2019
by
Katrina Louise Kalantar

ACKNOWLEDGEMENTS

First and foremost, I must thank Dr. Joe DeRisi, my graduate advisor. His support throughout the course of my graduate career has been essential. Joe has built an incredible network that includes researchers with diverse expertise, and I am fortunate to have had the opportunity to work with and learn from all these colleagues. Thank you for connecting me with people and projects unlike those available anywhere else. Finally, Joe's consistent enthusiasm for science is admirable and has filled even the most challenging times with hope for the larger impact. I must also thank my thesis committee members, Drs. Katie Pollard and Ian Holmes, for their mentorship, support, and guidance on technical and career-related aspects throughout my graduate career.

Thank you to Dr. Chaz Langelier, my close collaborator and mentor. I am forever grateful for the opportunity to have worked so closely on such exciting projects. Thank you for your persistent confidence in my work and for supporting my ideas and research goals. You have played a considerable role in helping me to grow as a scientist and to understand the academic process. Your continued commitment to improving patient care by tackling the challenging and important questions in infectious diseases is an inspiration. This will undoubtedly continue to generate insightful contributions.

Throughout my graduate career, I have had the wonderful privilege of working with and learning from some fantastic female mentors. Thank you to my unofficial mentors, Stephanie Christenson and Carolyn Calfee. I have learned so much through your technical expertise and leadership.

Thank you to all the wonderful members of the DeRisi Lab and associated collaborators. In particular, thank you to Akshaya Ramesh for always being available on

Slack with the most enthusiastic emoji sets; Hanna Retallack for always answering lab-related questions, providing great photography décor, and for your inspiring curiosity; Valentina Garcia and Kristeene Knopp, for your patience and willingness to answer questions during the times that I was in the wet lab. Thank you to Brittany Worth, Manny DeVera, and Jenn Mann, all whom have played an instrumental role in ensuring the lab is functioning and enabling much-needed meetings with Joe. Finally, thank you to all the clinical coordinators who make these clinical translational studies possible, especially Thomas Diess and the Calfee Lab team.

To Patricia Anderson for your support. You will forever be my graduate school partner, a title earned through your constant support and willingness to listen to unsolicited descriptions of my research progress – both the ups and the downs. To Emily and Megan, for always being willing to listen and sending a text every once in a while.

And finally, thank you to my parents, Clare and Dan Kalantar, for their love and support – throughout graduate school and life in general. I owe a great deal of my ambition to the examples that you both set and values you instilled in me.

CONTRIBUTIONS

The following two chapters consist of reprints of published manuscripts and may deviate from the final published manuscripts.

Chapter 2 of this dissertation is a reprint of the material as it appears in:

Kalantar, K.¹, Moazed, F.², Christenson, S.C.², Wilson J.³, Deiss, T.², Belzer, A.², Vessel, K.², Caldera, S.⁴, Jauregui, A.², Bolourchi, S.², DeRisi, J.L.^{1,4}, Calfee, C.S.², Langelier, C.² *A Metagenomic Comparison of Tracheal Aspirate and Mini-Bronchial Alveolar Lavage for Assessment of Respiratory Microbiota*. Am J Physiol Lung Cell Mol Physiol. 2019.

Chapter 3 of this dissertation is a reprint of the material as it appears in:

Langelier, C.^{2*}, Kalantar, K.^{1.*}, Moazed, F.², Wilson, M.R.⁵, Crawford, E.D.^{1,4}, Deiss, T.², Belzer, A.², Bolourchi, S.², Caldera, S.⁴, Fung, M.², Jauregui, A.², Malcolm, K.², Lyden, A.⁴, Khan, L.¹, Vessel, K.², Quan, J.^{1,4}, Zinter, M.⁶, Chiu, C.Y.⁷, Chow, E.D.¹, Wilson, J.³, Miller, S.⁷, Matthay, M.A.², Pollard, K.S.^{4,8,9}, Christenson, S.², Calfee, C.S.², DeRisi, J.L.^{1,4} *Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults*. Proc Natl Acad Sci. 2018. *equal contributions.

Chapter 4 contains an assortment of applications and discussion of related work, which are each individually published manuscripts or preprints. Emphasis has been placed on describing analyses performed by Katrina Kalantar, but details on sample selection,

preparation, and confirmatory testing (performed by collaborators) have been included where appropriate. The original manuscripts are referenced in their respective section and collaborators are listed here in full. Author contributions and full detailed methods can be found in the original manuscripts and preprints.

Section 4.1 contains work published in the following manuscript:

Langelier, C.^{*2}, Zinter, M.^{*6}, Kalantar, K.¹, Yanik, G.A.¹⁰, Christenson, S.C.², O'Donovan, B.¹, White, C.¹, Wilson, M.R.⁵, Sapru, A.⁶, Dvorak, C.C.⁶, Miller, S.⁷, Chiu, C.⁷, DeRisi, J.L.^{1,4} *Metagenomic Sequencing Detects Respiratory Pathogens in Hematopoietic Cellular Transplant Patients*. AJRCCM. 2018. *equal contributions

Specifically, all sequencing libraries were prepared by C.L. and M.Z..

Section 4.2 contains work published in the following manuscript:

Zinter, M.S.⁶, Dvorak, C.C.⁶, Mayday, M.Y.⁶, Iwagana, K.⁶, Ly, N.P.⁶, McGarry, M.E.⁶, Church, G.D.⁶, Faricy, L.E.¹¹, Rowan, C.M.¹², Hume, J.R.¹³, Steiner, M.E.¹³, Crawford, E.D.^{1,4}, Langelier, C.², Kalantar, K.¹, Chow, E.D.¹, Miller, S.⁷, Shimano, K.⁶, Melton, A.⁶, Yanik, G.A.¹⁰, Sapru, A.⁶, DeRisi, J.L.^{1,4} *Pulmonary Metagenomic Sequencing Suggests Missed Infections in Immunocompromised Children*. Clin Infect Dis. 2018.

Specifically, all sequencing libraries were prepared by M.Z. and M.M. The outlier detection analysis approach for pathogen identification was developed and applied by M.Z.

Section 4.3 contains work published in the following manuscript:

Saha, S.^{14,15}, Ramesh, A.⁵, Kalantar, K.¹, Malaker, R.¹⁴, Hasanuzzaman Md.¹⁴,
Khan, L.¹, Mayday M.⁶, Sajib, M.S.I.¹⁴, Li, L.⁴, Langelier, C.², Hafizur R.¹⁴,
Crawford, E.D.^{1,4}, Tato, C.M.⁴, Islam, M.¹⁴, CZI IDSeq Team¹⁶, Wilson, M.R.⁵,
Saha, S.K.*¹⁴, DeRisi, J.L.^{1,4} *. *Metagenomic Survey Identifies Chikungunya
Meningitis outbreak and Other Unrealized Bacterial and Viral Pathogens in
Pediatric Meningitis cases in Bangladesh*. Biorxiv. 2019. *equal contributions

Specifically, sample collection was performed by S.S.. Clinical testing and patient follow-up were performed by S.S. and the team in Dhaka, Bangladesh. All sequencing libraries were prepared by L.K. and M.M. and confirmatory testing was performed by L.K. and A.R..

Section 4.4 contains work published in the following manuscript:

Zinter, M.S.⁶, Barrows, B.D.¹⁷, Ursell, P.C.¹⁷, Kowalek, K.⁶, Kalantar, K.¹,
Cambroner, N.¹⁸, DeRisi, J.L.^{1,4}, Oishi, P.⁶, Dvorak, C.C.⁶ *Extracorporeal life
support survival in a pediatric hematopoietic cellular transplant recipient with
presumed GvHD-related fulminant myocarditis*. Bone Marrow Transplantation.
2017

Specifically, all sequencing libraries were prepared by M.Z..

Section 4.5 contains work published in the following manuscript:

Langelier, C.², Graves, M.*¹⁹, Kalantar, K.*¹, Caldera, S.⁴, Durrant, R.^{19,20},

Fisher, M.^{19,20}, Backman, R.¹⁹, Tanner, W.¹⁹, DeRisi, J.L.^{1,4}, Leung, D.T.¹⁹

Microbiome and Antimicrobial Resistance Gene Dynamics in International

Travelers. Biorxiv. 2018. *equal contributions

Specifically, all sequencing libraries were prepared by C.L. and S.C. and the SRST2 AMR analysis was performed by C.L.

¹Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA

²Department of Medicine, University of California, San Francisco, San Francisco, CA, USA

³Department of Emergency Medicine, Stanford University, Stanford, CA, USA

⁴Chan Zuckerberg Biohub, San Francisco, CA, USA

⁵Department of Neurology, University of California, San Francisco, San Francisco, CA, USA

⁶Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA

⁷Department of Laboratory Medicine, University of California, San Francisco, San Francisco, CA, USA

⁸Gladstone Institutes, San Francisco, CA, USA

⁹Department of Epidemiology and Biostatistics, Institute for Human Genetics, Quantitative Biology Institute, and Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, CA, USA

¹⁰University of Michigan, Ann Arbor, MI, USA

¹¹Department of Pediatrics, University of Vermont School of Medicine, Burlington, VT, USA

¹²Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN , USA

¹³Department of Pediatrics, University of Minnesota School of Medicine, Minneapolis, MN, USA

¹⁴Child Health Research Foundation, Department of Microbiology, Dhaka Shishu Hospital, Sher-E-Bangla Nagar, Dhaka, Bangladesh

¹⁵Department of Infectious Diseases, Stanford University School of Medicine, Stanford, CA, USA

¹⁶Chan Zuckerberg Initiative, CA, USA

¹⁷Department of Pathology, University of California, San Francisco, San Francisco, CA, USA

¹⁸Department of Surgery, University of California, San Francisco, San Francisco, CA, USA

¹⁹University of Utah School of Medicine, Salt Lake City, UT, USA

²⁰ARUP Laboratories, Salt Lake City, UT, USA

**Combined host and microbial metagenomic next-generation sequencing:
Applying integrated approaches for a comprehensive evaluation of infectious
disease response to inform diagnosis, surveillance, and treatment**

Katrina Louise Kalantar

ABSTRACT

Infectious diseases are a leading cause of morbidity and mortality worldwide. Despite significant advancement in our understanding of infectious disease biology, existing microbiologic diagnostic tests often fail to identify etiologic pathogens in cases of suspected infection. Metagenomic next-generation sequencing (mNGS) offers the potential for a universal pathogen detection method, but analysis and interpretation of findings are challenging. This is especially true for lower respiratory tract infections (LRTIs) where mNGS data interpretation is complicated by the existence of a respiratory microbiome composed of pathobionts present in both health and disease.

To address the need for improved LRTI diagnostics, we first compared two fluid types commonly used for diagnosis of LRTI, showing that despite moderate microbiome differences, both mini-bronchioalveolar lavage (mBAL) and tracheal aspirate (TA) samples are suitable for identification of pathogens in the context of an infection. Then, we evaluated the utility of mNGS as a diagnostic for LRTI in a cohort of 92 TA samples from adults with acute respiratory failure. We developed methods for sifting putative pathogens from commensal microbiota as well as pathogen, microbiome diversity, and host gene expression metrics to identify LRTI-positive patients and differentiate them

from critically ill controls with noninfectious acute respiratory illnesses. We applied the models developed for evaluation of LRTI status to several other cohorts and disease contexts to show their broad applicability.

The low sensitivity of existing clinical diagnostics results in an imperfect gold standard, complicating the development of mNGS-based biomarkers. We explored the impact of label noise on host gene expression classifiers and methods for circumventing the issue. First, we tested whether label-noise robust logistic regression approaches could improve classifier performance by enabling the use of a larger training set. Then, we tested whether variational autoencoders, an unsupervised dimensionality reduction approach, could generate novel insight from combined host and microbial mNGS data. Altogether, this work suggests that a single streamlined protocol offering an integrated genomic portrait of pathogen, microbiome, and host transcriptome may hold promise as a tool for diagnosis of infections and contextualization of patient response.

TABLE OF CONTENTS

1	INTRODUCTION.....	1
1.1	UNDERSTANDING OF INFECTIOUS DISEASES THROUGH HISTORY	1
1.2	BIOLOGY OF LOWER RESPIRATORY TRACT INFECTIONS	2
1.3	INFECTIOUS DISEASE DIAGNOSTICS ARE BEHIND-THE-TIMES	3
1.4	METAGENOMIC NEXT-GENERATION SEQUENCING AS A DIAGNOSTIC ...	5
1.5	TECHNICAL CHALLENGES IN MACHINE LEARNING FOR DIAGNOSTICS...	6
1.6	SUMMARY OF APPROACH AND FINDINGS	8
1.7	ACRONYMS AND ABBREVIATIONS	10
2	A METAGENOMIC COMPARISON OF TRACHEAL ASPIRATE AND MINI- BRONCHIAL ALVEOLAR LAVAGE FOR ASSESSMENT OF RESPIRATORY MICROBIOTA	11
2.1	ABSTRACT	11
2.2	INTRODUCTION	12
2.3	METHODS	12
2.4	RESULTS	14
2.4.1	<i>Comparison of mBAL to TA across all patient groups</i>	<i>14</i>
2.4.2	<i>Comparison of mBAL and TA as a function of pneumonia status</i>	<i>15</i>
2.5	DISCUSSION.....	17
2.6	FIGURES	19
2.7	TABLES	22

3	INTEGRATING HOST RESPONSE AND UNBIASED MICROBE DETECTION FOR LOWER RESPIRATORY TRACT INFECTION DIAGNOSIS IN CRITICALLY ILL ADULTS.....	27
3.1	ABSTRACT	27
3.2	INTRODUCTION	28
3.3	RESULTS	29
3.3.1	<i>Pathogen Detection</i>	<i>30</i>
3.3.2	<i>LRTI Prediction Based on Pathogen.</i>	<i>35</i>
3.3.3	<i>LRTI Prediction Based on Lung Microbiome Diversity.</i>	<i>35</i>
3.3.4	<i>LRTI Prediction Based on Host Response.</i>	<i>36</i>
3.3.5	<i>Evaluation of a Combined LRTI Metric.</i>	<i>39</i>
3.4	DISCUSSION.....	39
3.5	METHODS	43
3.5.1	<i>Study Design and Subjects.....</i>	<i>43</i>
3.5.2	<i>Clinical Microbiologic Testing.</i>	<i>44</i>
3.5.3	<i>Definitions and Clinical Adjudication of LRTI.</i>	<i>44</i>
3.5.4	<i>Identification of Subjects with LRTI.....</i>	<i>45</i>
3.5.5	<i>Host/Microbe mNGS.....</i>	<i>46</i>
3.5.6	<i>Pathogen Detection Bioinformatics.....</i>	<i>46</i>
3.5.7	<i>Statistical Analysis.</i>	<i>47</i>
3.5.8	<i>Pathogen Versus Commensal Models.</i>	<i>47</i>
3.5.9	<i>LRTI Prediction Based on Pathogen.</i>	<i>51</i>
3.5.10	<i>Lung Microbiome Diversity Analysis.</i>	<i>51</i>

3.5.11	<i>Host Gene Expression Analysis.</i>	52
3.5.12	<i>Host Gene Expression Classifier for LRTI Prediction.</i>	53
3.5.13	<i>Classifier Combination.</i>	56
3.5.14	<i>Identification and Mitigation of Environmental Contaminants.</i>	56
3.5.15	<i>Estimation of Antibiotic Use Reduction.</i>	57
3.5.16	<i>Data Availability.</i>	57
3.6	FIGURES	58
3.7	TABLES	65
3.8	SUPPLEMENTAL FIGURES	67
3.9	SUPPLEMENTAL TABLES	76
4	APPLICATION NOTES	85
4.1	EARLY INTERROGATION OF LRTI THROUGH MNGS	85
4.1.1	<i>Application Goal.</i>	85
4.1.2	<i>Introduction</i>	86
4.1.3	<i>Methods</i>	86
4.1.4	<i>Results</i>	88
4.1.5	<i>Discussion and Conclusion</i>	89
4.2	PATHOGEN DETECTION FOR LRTI DIAGNOSIS IN A DISTINCT COHORT	91
4.2.1	<i>Application Goal.</i>	91
4.2.2	<i>Introduction</i>	91
4.2.3	<i>Methods</i>	92
4.2.4	<i>Results</i>	93
4.2.5	<i>Discussion</i>	94

4.2.6	<i>Conclusion and Opportunity for Future Work</i>	95
4.3	PATHOGEN DETECTION FOR MENINGITIS ETIOLOGY SURVEILLANCE..	96
4.3.1	<i>Application Goal.....</i>	96
4.3.2	<i>Introduction</i>	96
4.3.3	<i>Methods</i>	98
4.3.4	<i>Results.....</i>	100
4.3.5	<i>Discussion</i>	101
4.3.6	<i>Conclusion</i>	102
4.4	SEPARATION OF HOST FROM HOST: DONOR VS. RECIPIENT IN MNGS OF HCT PATIENTS	103
4.4.1	<i>Application Goal.....</i>	103
4.4.2	<i>Introduction</i>	104
4.4.3	<i>Methods</i>	105
4.4.4	<i>Results.....</i>	106
4.4.5	<i>Discussion</i>	107
4.4.6	<i>Conclusion</i>	107
4.5	EXPANDING MNGS TO UNDERSTAND THE FLOW OF ANTIMICROBIAL RESISTANCE GENES AS A FUNCTION OF GLOBAL TRAVEL	108
4.5.1	<i>Application Goal.....</i>	108
4.5.2	<i>Introduction</i>	108
4.5.3	<i>Methods</i>	109
4.5.4	<i>Results.....</i>	110
4.5.5	<i>Discussion and Conclusion.....</i>	112

4.6	FIGURES	113
4.7	TABLES	119
5	CLASSIFICATION OF COMBINED HOST AND MICROBIAL MNGS DATA IN THE CONTEXT OF AN IMPERFECT GOLD STANDARD.....	123
5.1	SUPERVISED CLASSIFICATION: HOST GENE EXPRESSION CLASSIFIERS WITH NOISY LABELS	123
5.1.1	<i>Introduction</i>	123
5.1.2	<i>Methods</i>	126
5.1.3	<i>Results</i>	129
5.1.4	<i>Discussion and Conclusion</i>	131
5.2	UNSUPERVISED ANALYSIS OF COMBINED HOST AND MICROBIAL MNGS DATA USING VARIATIONAL AUTOENCODERS	133
5.2.1	<i>Introduction</i>	133
5.2.2	<i>Methods</i>	135
5.2.3	<i>Results</i>	138
5.2.4	<i>Discussion and Conclusion</i>	140
5.3	FIGURES	142
5.4	TABLES	148
6	CONCLUSIONS AND FUTURE DIRECTIONS	152
6.1	FUTURE WORK FOR DEVELOPING LRTI DIAGNOSTICS.....	153
6.2	FUTURE DIRECTIONS FOR COMBINED HOST AND MICROBE MNGS AS A DIAGNOSTIC FOR INFECTIONS	155

6.3	FUTURE EXPANSION OF METHODS FOR CLASSIFICATION WITH NOISY LABELS	156
6.4	CONCLUDING REMARKS	157
7	APPENDIX OF EXTERNAL ELECTRONIC RESOURCES	159
	BIBLIOGRAPHY	162

LIST OF FIGURES

FIGURE 2.1 Fraction of total microbial sequencing reads in mBAL and TA samples....	19
FIGURE 2.2 Shannon diversity index as a PNA biomarker in mBAL versus TA.....	20
FIGURE 2.3 Sample-type differences in Richness and Shannon diversity index	21
FIGURE 3.1 Study overview and novel analysis workflow	58
FIGURE 3.2 Distinguishing LRTI pathogens from commensal respiratory microbiota using an algorithmic approach	60
FIGURE 3.3 Distribution of respiratory pathogens identified in patients using clinician- ordered diagnostics versus mNGS	61
FIGURE 3.4 Diversity of the transcriptionally active lung microbiome in patients with LRTI versus non-infectious respiratory illnesses	62
FIGURE 3.5 Host transcriptional profiling distinguishes patients with acute LRTI from those with non-infectious acute respiratory illness	63
FIGURE 3.6 Combined LRTI prediction metric integrating pathogen detection and host gene expression	64
FIGURE 3.7 Distribution of mNGS-identified microbes per patient, by relative abundance	70
FIGURE 3.8 Microbial pathogens identified by clinician-ordered diagnostics, compared to those identified by mNGS RBM and LRM models	72
FIGURE 3.9 LRM probability score differentiates patients with LRTI from those with non-infectious causes of acute respiratory failure	73

FIGURE 3.10 Performance DNA-Seq microbiome diversity for differentiating patients with LRTI from those with non-infectious causes of acute respiratory failure	74
FIGURE 3.11 Learning Curve analyses for pathogen vs commensal and host gene expression classifier models	75
FIGURE 4.1 Simpson diversity index assessed for HCT patients with confirmed vs unlikely or uncertain pathogens	113
FIGURE 4.2 Host gene metric assessed for HCT patients with confirmed vs unlikely or uncertain pathogens	114
FIGURE 4.3 Comparison of pathogen identification methods in pediatric HCT recipients	115
FIGURE 4.4 Application of the LRM to determine etiology of meningitis in Bangladesh	117
FIGURE 4.5 Application of the LRM to identify CHKV meningitis cases in Bangladesh	118
FIGURE 5.1 Performance of LR and rLR on simulated data in the context of label noise	142
FIGURE 5.2 Impact of label noise on a priori feature selection	143
FIGURE 5.3 Learning curves computed to test rLR and LR	144
FIGURE 5.4 VAE training performance and latent representation	145
FIGURE 5.5 TSNE projection of zero-one scaled dataset	146
FIGURE 5.6 VAE latent dimension activations for each sample	147

LIST OF TABLES

TABLE 2.1	Demographics and clinical characteristics of mBAL vs TA study cohort.....	22
TABLE 2.2	Microbial community metrics and oropharyngeal contaminants.....	23
TABLE 2.3	Patient diagnoses and clinical microbiology testing	24
TABLE 2.4	Top five most abundant microbes per sample	26
TABLE 3.1	Demographics and clinical characteristics of study cohort.....	65
TABLE 3.2	Expanded clinical and microbiologic data	76
TABLE 3.3	United States CDC/NHSN surveillance definition of pneumonia.....	76
TABLE 3.4	Reference index of established respiratory pathogens	77
TABLE 3.5	Microbial pathogens identified by clinician-ordered diagnostics or predicted by pathogen identification model	79
TABLE 3.6	LRM feature weights	79
TABLE 3.7	Microbes identified in no-LRTI patients.....	80
TABLE 3.8	Diversity metrics assessed for patients with LRTI versus those with non- infectious causes of acute respiratory failure.....	81
TABLE 3.9	Differentially expressed genes between LRTI ^{+C+M} and no-LRTI patients....	82
TABLE 3.10	LRTI host transcriptional classifier specifics.....	82
TABLE 3.11	Covariates controlled for in the host gene expression classifier	83
TABLE 3.12	Estimated cell type proportions	83
TABLE 3.13	Most abundant genera in water controls	84
TABLE 3.14	Human Transcriptome Counts	84
TABLE 4.1	Composite host immune gene metric for HCT patients.....	119
TABLE 4.2	Host gene counts from mNGS study of adult HCT recipients	120

TABLE 4.3 Comparison of pathogen identification methods in pediatric HCT recipients	120
TABLE 4.4 Fold change in abundance of AMR genes	122
TABLE 5.1 Simulated and publicly available datasets used for label noise analyses .	148
TABLE 5.2 Test conditions for LR versus rLR learning curve	148
TABLE 5.3 Datasets used for VAE analysis	149
TABLE 5.4 GO Biological Processes significantly enriched in VAE latent dimensions	150

1 INTRODUCTION

1.1 UNDERSTANDING OF INFECTIOUS DISEASES THROUGH HISTORY

Infectious diseases have long played a role in influencing human history¹, but our understanding of the underlying microbiology has continued to evolve over the past several centuries of scientific advancement. Dating back to the era of Hippocrates, infectious diseases were generally identified by *fluid dysregulation* – characterized by changes in clinical presentation as a function of the human immune response, such as fever, swelling, etc.². However, invention of the microscope in the 1600s enabled direct observation of micro-organisms. This led eventually to an understanding that pathogens were causative agents in infections³. For many decades following these discoveries, it was thought, based on the germ theory presented by Louis Pasteur and further refined by Robert Koch⁴, that microbes played a purely antagonistic role in human health. However, recent developments in molecular assays have enabled the discovery and appreciation of the microbiome – complex communities of microbes found ubiquitously in nature – as an essential component of the human body⁴. The microbiome plays a role in maintaining human health through symbiotic relationships with the host by carrying out a variety of essential metabolic processes⁵. At present, there is an increasing appreciation of human health as a balance between the host and microbial communities and disruption of the microbiota is now viewed as a key element in infectious disease pathogenesis.

1.2 BIOLOGY OF LOWER RESPIRATORY TRACT INFECTIONS

Lower respiratory tract infections (LRTI; infections of the lower airway and lungs, including pneumonia, bronchitis, and bronchiolitis) are a leading cause of morbidity and mortality world-wide⁶. In the United States, they cause the majority of infectious-disease related deaths⁷. However, the burden of LRTI is not equally distributed⁸. In low and middle income countries (LMICs), pneumonia remains the leading cause of childhood mortality and the most common cause of hospitalization among adults⁹.

Lower respiratory tract infections can be caused by a variety of pathogens – including viruses, bacteria, and fungi. At the microbiologic level, molecular pathogenesis is characterized by pathogen invasion of the airway and subsequent replication, often in the airway epithelial cells. Traditionally, LRTIs have been attributed to a limited number of pathogens, including *Streptococcus pneumoniae* and *Haemophilus influenza*¹⁰. However, since the initial discovery of the microbiome in the alimentary tract, recent work re-examining body sites that were previously considered to be free of microbes (so-called “sterile sites”), has expanded our appreciation of the lung microbiome⁴. Through application of molecular assays to study the lung microbiome, it has become clear that many of the traditional pathogens causing LRTIs are found as commensal residents of the airway (both in the upper and lower respiratory tracts)¹⁰. This shifts our understanding of the microbiology of infection, suggesting that the balanced airway microbiome may provide protective effects, while perturbations in the microbiome, caused by environmental factors or rapid transmission of potentially pathogenic microbes (pathobionts), may result in over-growth and subsequent presentation of acute respiratory infection^{10,11}.

The observed clinical presentation of an acute respiratory tract infection is a function of the host response. At the microbiologic level, host innate defense mechanisms, including the mucociliary escalator, antimicrobial proteins secreted by airway epithelial cells, and alveolar macrophages, provide initial defense against the pathogen. However, once the pathogen gains hold, a subsequent inflammatory response recruits innate and adaptive immune cells to the site of the infection. Mounting evidence supports the understanding that infection severity is largely influenced by the host response. A prolonged and exaggerated immune response may be the source of more damage than protection^{12,13}. The large-scale symptoms of LRTI, resulting from inflammatory response and airway damage, include fever, cough, and fluid leakage into the airway causing difficulty breathing. These symptoms often overlap with those of colds or the flu, making diagnosis of LRTI challenging¹⁴.

1.3 INFECTIOUS DISEASE DIAGNOSTICS ARE BEHIND-THE-TIMES

Despite major advancements in our understanding of LRTI disease biology, there has been little change in relative mortality due to LRTI over the past several decades^{15–17}. Factors contributing to this may include the diversity of pathogens known to cause LRTIs¹⁵ as well as the relatively static landscape of available diagnostic tests.

Candidate-based approaches for pathogen identification include microbial culture and microscopy (both of which originated in the 19th century), as well as more recent antigen and PCR-based assays introduced in the 1980's and 2000's, respectively. Culture-based methods are limited to detecting only microbes capable of growing in culture conditions and often fail following administration of prophylactic antibiotics¹⁸. Even

for microbes amenable to culture-based assays, growth may require several days, precluding rapid diagnosis. Similarly, antimicrobial therapy and inadequate sputum collection lead to low diagnostic yields via microscopy¹⁹. Antigen testing provides the added benefit of detecting pathogens even after they are rendered non-viable due to antibiotic administration, but requires relatively high levels of antigen and may result in false-positives due to cross-reactivity. Finally, PCR-based assays require a defined set of primers and will fail to identify pathogens whose sequences differ, such as occurs for rapidly-evolving RNA-viruses. All candidate-based approaches rely on clinicians to order the correct tests based on clinical presentation, so atypical clinical presentation may result in failure to run the correct tests to identify etiology even for common pathogens. Standard diagnostics profiling host factors, including white blood cell (WBC) count, chest x-ray, and more recently, c-reactive protein and procalcitonin, are used to assess inflammatory state. However, they have low sensitivity, especially among critically ill patients where other inflammatory processes may confound results²⁰, and cannot provide a precise microbial diagnosis to inform treatment.

Altogether, the current diagnostics for LRTI fail to identify etiology in 62% of cases, making precise and informed treatment and patient care challenging²¹. Further complicating diagnosis of LRTIs is the prevalence of non-infectious etiologies of lower respiratory disease with similar symptoms²⁰. These patterns extend beyond LRTIs, as many of the same diagnostic tests are run for various infectious diseases. For example, etiology of meningitis goes undetected in upwards of 50% of cases^{22–24}. Diagnostic challenges are compounded for LMICs, where the landscape of available diagnostic tests is limited¹⁷. The need for assays that can consider the recent developments in our

understanding of infectious disease biology and assay multiple aspects of infection risk (host factors, pathogen identification, and microbiome) is clear.

1.4 METAGENOMIC NEXT-GENERATION SEQUENCING AS A DIAGNOSTIC

As the cost of sequencing continues to decline²⁵, next-generation sequencing (NGS)-based methods have been increasingly explored for evaluating host and microbial constituents of infection. One such approach, metagenomic next-generation sequencing (mNGS), has emerged as a promising universal pathogen detection method²⁶. mNGS is a culture-independent method that enables sequencing of nucleic acid obtained directly from an environmental sample without prior amplification, thus providing the most direct and unbiased assay of microbial communities in the sample²⁷. One challenge with mNGS for pathogen detection is the incredible sensitivity²⁶. Previous work has developed methods to filter putative pathogens from laboratory contaminants²⁸. However, these methods have largely been developed in “sterile” fluids, such as cerebral spinal fluid or ocular fluid, where no background microbiota exists^{28,29}. In the context of LRTIs, there is an added challenge to distinguish between pathogens and commensal microbiota^{26,30}. New methods are required to address the outstanding challenges with interpretation of mNGS data.

For evaluating the host response to infection, transcriptomic profiling via microarrays and RNA-sequencing (RNA-Seq) has gained traction as a promising alternative to existing low-sensitivity host biomarkers, such as procalcitonin^{31–33}. Machine learning methods have been applied to derive gene expression signatures indicative of particular disease states. Recent studies have demonstrated the ability to distinguish

between patients with Respiratory syncytial virus LRTI versus those with other viral LRTI (Human Rhinovirus and Influenza) with sensitivity and specificity greater than 95%³⁴. Other studies have developed host gene expression signatures that predict the presence of a bacterial or viral infection with an area under the receiver operating curve (AUC) above 0.9 across multiple external validation cohorts³¹. While transcriptional signatures may identify the presence of particular classes of infection, the overlap in host immune response to pathogens limits the granularity of this approach for etiology detection. A method with the ability to assay the host response while providing specific pathogen-identification would increase granularity for pathogen detection while contextualizing putative pathogen findings within the host response.

1.5 TECHNICAL CHALLENGES IN MACHINE LEARNING FOR DIAGNOSTICS

A variety of machine learning methods are commonly applied to derive transcriptional signatures of disease, including regression-based methods, random forests, and support vector machines^{31,32,34,35}. While the specifics of each algorithm may vary, all algorithms require labeled data from cohorts of known disease-positive and -negative samples. Models are trained on labeled data and then validated with an external test set. Additionally, validation of any new diagnostic test requires comparison to the current gold standard to prove sensitivity and specificity^{36,37}. In the context of infectious disease diagnosis, the gold standard is imperfect at best and a negative test result cannot rule out infection. Thus, the risk of sample mislabeling is high. This phenomenon is increasingly appreciated in a range of biological applications^{38,39}.

The presence of errors in labelling of training data has been shown to reduce the performance of derived models⁴⁰. For diagnostic development, one method for reducing the potential for sample mislabeling is to focus the derivation of models on only the most definitively positive or negative cases. These can be identified by expert review considering the full patient history and all clinical test results. But, when nearly 50% of cases with suspected infectious disease lack an etiologic assignment by current standard diagnostics^{21–24}, a significant portion of data may be excluded. Given the variable nature of biological data, especially from diverse patient groups, ensuring that even a simple model is robust requires relatively large cohorts. The challenges and cost associated with recruiting patients and preparing samples often impose limits on cohort size, such that removal of 50% of samples would be detrimental to model performance.

While the cost of data collection is exceptionally high for clinical studies, the cost of data collection and label curation in other machine learning applications is also non-trivial. To reduce data collection cost, recent work in machine learning has begun to examine methods for combating label noise. These methods attempt to either filter out potentially noisy samples prior to model training or model the likelihood of mislabeling during model training^{41–43}. Label noise robust classification methods may be useful for deriving host gene expression classifiers in the context of an imperfect gold standard. Alternatively, unsupervised classification methods may provide a complementary avenue for further insight into disease states in the context of label-free data.

1.6 SUMMARY OF APPROACH AND FINDINGS

Towards the goal of developing an improved molecular diagnostic for LRTI, we evaluate mNGS as a diagnostic tool. In Chapter 2, I present a comparison of the microbiome from two distinct sample types commonly used for LRTI diagnosis – tracheal aspirate (TA) and mini-bronchial alveolar lavage (mBAL). I show that in the context of LRTI, any sample-type specific differences in microbiome content are rendered insignificant for the purposes of diagnosis.

In Chapter 3, I present the development and validation of bioinformatics approaches for interpretation of mNGS data for diagnosis of LRTI. I consider host, pathogen, and microbiome metrics and demonstrate how combining these into one assay shows promise for improving our understanding of patient condition and informing precise treatment decisions. Specifically, I derive an improved method for sifting pathogens from background microbiota, generate a host gene expression classifier for distinguishing patients with LRTI from those without, and evaluate changes in the microbiome and a function of LRTI.

In Chapter 4, I provide a series of *Application Notes*. First, I outline a pilot study that preceded the development of the approaches derived in Chapter 3. Then, I apply these approaches to two distinct cohorts with differing disease characteristics to demonstrate broad applicability. The first cohort consists of pediatric hematopoietic cell transplantation (HCT) patients with possible LRTI. There, I test the diagnostic performance in a similar fluid type where the associated comorbidities of the patient population shift the distribution of pathogens to include more fungal infections. The second cohort consists of pediatric patients with suspected meningitis in Dhaka,

Bangladesh, a LMIC. There, I highlight how mNGS may provide utility for infectious disease surveillance. Finally, I provide two more downstream analyses indicating directions in which this technology may extend in the future - first, for identifying host sequences from two distinct hosts after transplantation and second, for understanding global transmission of antimicrobial resistance (AMR) genes.

In Chapter 5, I discuss an exploration of machine learning methods for combating label noise in the context of mNGS infection diagnostics. First, I use both simulated and publicly available data to evaluate the use of a label-noise robust logistic regression algorithm⁴³ for improving classification performance in the presence of label noise. Then, I evaluate variational autoencoders (VAEs), an unsupervised learning approach, for their ability to learn biologically relevant features from combined host and microbial mNGS data without relying on potentially-noisy labels.

Finally, in Chapter 6 I summarize the contributions made through this work and discuss opportunities for future work.

1.7 ACRONYMS AND ABBREVIATIONS

AMR	Antimicrobial resistance
AUC	Area under the receiver operating curve
BAL	Bronchial alveolar lavage / Bronchoalveolar lavage
CHKV	Chikungunya virus
CSF	Cerebral spinal fluid
Ct	Cycle threshold
DNA-Seq	DNA sequencing
ERCC	External RNA Controls 103 Consortium
ESBL	Extended spectrum beta lactamases
ESBL-PE	ESBL-producing E. coli
GO	Gene Ontology
GvHD	Graft versus host disease
HCT	Hematopoietic cell transplantation
IQR	Interquartile range
LD	Latent dimension
LMICs	Low and middle income countries
LR	Standard logistic regression, specifically defined in Chapter 5
LRM	Logistic regression model, specifically defined in Chapter 3
LRTI	Lower respiratory tract infection
mBAL	Mini-bronchial alveolar lavage / mini-bronchoalveolar lavage
mNGS	Metagenomic next-generation sequencing
MSE	Mean squared error
NGS	Next-generation sequencing
PCA	Principal components analysis
PNA	Pneumonia
RBM	Rules-based method
ROC	Receiver operator characteristic
rpm	Reads per million
rLR	Robust logistic regression
RNA-Seq	RNA sequencing
SDI	Shannon diversity index
SRA	Sequence read archive
TA	Tracheal aspirate
VAE	Variational autoencoder
WBC	White blood cell

2 A METAGENOMIC COMPARISON OF TRACHEAL ASPIRATE AND MINI-BRONCHIAL ALVEOLAR LAVAGE FOR ASSESSMENT OF RESPIRATORY MICROBIOTA

2.1 ABSTRACT

Accurate and informative microbiologic testing is essential for guiding diagnosis and management of pneumonia in critically ill patients. Sampling of tracheal aspirate is less invasive compared to mini-bronchoalveolar lavage and is now recommended as a frontline diagnostic approach in mechanically ventilated patients, despite the historical belief that TA was suboptimal due to contamination from oral microbes. Advancements in mNGS now permit assessment of airway microbiota without a need for culture, and as such provide an opportunity to examine differences between mBAL and TA at a resolution previously unachievable. Here, we engaged shotgun mNGS to quantitatively assess the airway microbiome in matched mBAL and TA specimens from a prospective cohort of critically ill adults. We observed moderate differences between sample types across all subjects, however we found significant compositional similarity in subjects with bacterial pneumonia, whose microbial communities were characterized by dominant pathogens. In contrast, in patients with non-infectious acute respiratory illnesses, significant differences were observed between sample types. Our findings suggest that TA sampling provides a similar assessment of airway microbiota as more invasive testing by mBAL in patients with pneumonia.

2.2 INTRODUCTION

Pneumonia causes more deaths each year in the United States than any other type of infectious disease⁷. The ability to accurately detect etiologic pathogens and distinguish them from background commensal microbiota is essential for guiding optimal antimicrobial treatment. In patients requiring mechanical ventilation, less invasive TA sampling has historically been considered inferior to specimen collection by mini-bronchoalveolar lavage/telescoping catheter due to the potential for oropharyngeal microbiota contamination^{44,45}. This idea has been challenged, however, by studies demonstrating a lack of clinically significant differences between sample types^{44–47}, and a greater acceptance of TA sampling is now reflected in recent updates to clinical practice guidelines⁴⁸. Despite the broad potential implications of this shift in diagnostic sampling approach, relatively little information exists regarding microbial composition differences between mBAL and TA specimens and the potential implications of such differences for both clinical diagnostic testing and airway microbiome studies. To address this gap in knowledge, we employed shotgun mNGS to evaluate the microbial compositions of matched mBAL and TA specimens.

2.3 METHODS

We enrolled 52 adults who were intubated within 72 hours of intensive care unit admission with acute respiratory failure according University of California San Francisco protocol 10-02701. Demographic and clinical characteristics of the 52 study subjects are summarized in (TABLE 2.1). Two-physician adjudication based on retrospective medical record review (blinded to mNGS results) and the United States Centers for Disease

Control surveillance case definition of pneumonia was used to identify 15 subjects with culture-confirmed bacterial pneumonia (PNA-pos)⁴⁹. Adjudication also identified 12 subjects with a clear alternative non-infectious etiology of acute respiratory failure (PNA-neg), and 25 subjects with acute respiratory illnesses of indeterminate etiology (PNA-ind), which included those with negative bacterial cultures but suspected pneumonia based on clinical criteria alone. Subjects with PCR-confirmed viral etiologies were also included in the PNA-ind group because occult bacterial co-infection could not be excluded.

Excess mBAL and TA specimens collected on the same day and within 72 hours of patient intensive care unit admission underwent DNA extraction and sequencing library preparation according to previously described methods^{28,50}. Following paired-end Illumina sequencing, we employed a previously reported bioinformatics pipeline to detect and profile the airway microbiome. Briefly, this incorporated subtractive alignment of the human genome (NCBI GRCh38) using STAR⁵¹ followed by quality filtering using PRICESeqfilter⁵². Additional filtering to remove Pan troglodytes (UCSC PanTro4) and non-fungal eukaryotes, cloning vectors and phiX phage was performed using Bowtie2⁵³. The identities of the remaining microbial reads were determined by querying the NCBI nucleotide database using GSNAP-L^{28,50}. We sequenced no-template water control samples and restricted analyses to taxa present at > 1% of the microbial population by abundance, as previously described^{45,50}. No microbe was universally present in every sample, suggesting that systematic contamination across TA or mBAL sampling methods was unlikely. Microbial community composition metrics were calculated using the vegan R package version 2.5.2⁵⁴. P-values were computed using Wilcoxon rank sums. When

evaluating community richness, one outlier (> 3 standard deviation above the mean) was identified and removed prior to computing significance.

2.4 RESULTS

2.4.1 Comparison of mBAL to TA across all patient groups

To compare the microbial community compositions of matched mBAL and TA specimens across all patients in the cohort, we first calculated the Bray-Curtis dissimilarity index, which revealed no significant differences ($p = 0.31$ by PERMANOVA). We next asked whether within-subject diversity of the respiratory microbial communities differed by specimen type, and also did not observe a significant difference in Shannon's Diversity Index (SDI) (1.05 (0.71-1.55) versus 1.45 (0.74-2.05) for TA and mBAL, respectively, $p = 0.057$, TABLE 2.2A), although the p value approached significance. Community richness (total number of different genera identified in each sample), however, was higher in mBAL samples than in TA ($p = 0.046$) (TABLE 2.2A). Calculation of Spearman correlation between matched mBAL and TA specimens across all subjects revealed moderate differences, with a mean correlation of 0.41 (Interquartile range (IQR): 0.03 – 0.87) (FIGURE 2.1).

Because oropharyngeal microbiota have historically been suspected to compromise TA specimens, we next evaluated for differences in the abundance of common oropharyngeal microbiota⁴⁵. Surprisingly, we found no statistically significant differences between mBAL and TA specimens with respect to *Prevotella*, *Veillonella*, *Streptococcus*, *Fusobacterium*, *Rothia*, or *Neisseria* abundance (TABLE 2.2B). Assessment of microbial relative abundance (total genus alignments per million reads sequenced) also revealed

no significant inter-specimen type differences (Median = 43.30 (IQR: 6.69 – 327.93) versus 26.89 (5.66 - 167.63), $p = 0.66$).

2.4.2 Comparison of mBAL and TA as a function of pneumonia status

We reasoned that differences in microbial composition between mBAL and TA specimens would be most clinically significant if they impacted diagnostic accuracy in patients with pneumonia, and thus assessed taxonomic similarity between the PNA-pos and PNA-neg groups. We found significantly greater correlation in PNA-pos subjects as compared to the PNA-neg subjects (pairwise Spearman correlation of 0.75 (0.67 – 1.00) versus 0.19 (-0.22 – 0.55), $p = 1.62 \times 10^{-3}$), suggesting that pathogen dominance of the lung microbiome during infection may drive compositional similarity^{50,55}.

For both sample types, a culture-confirmed pathogen was the most abundant microbe detected by mNGS in 14 (93%) of PNA-pos subjects, and the second most abundant in the remaining subject (TABLE 2.3, TABLE 2.4). Gram-negative pathogens were cultured from a relatively high percentage of PNA-pos patients (65%) as compared to prior surveillance studies²¹. mNGS of mBAL specimens detected all 23 culture-identified microbes, while mNGS of TA samples identified 22. The discrepant microbe was from a polymicrobial culture and was detected by mNGS in the TA specimen but present at < 1% relative abundance and therefore indistinguishable from background using our bioinformatic approach (TABLE 2.3, TABLE 2.4).

Reduced alpha diversity of the human respiratory microbiome has been described as an ecological marker of infection^{50,55}, and thus we next asked whether Shannon's Diversity Index differed by specimen type. We found that SDI differed significantly

between PNA-pos and PNA-neg subjects within mBAL samples, and a similar, albeit less significant difference, was also observed within TA samples ($p = 5.2 \times 10^{-6}$ and 4.7×10^{-2} , respectively) (FIGURE 2.2, TABLE 2.2A). Community richness (total number of different genera identified in each sample) was decreased in PNA-pos compared to PNA-neg subjects when assessed by mBAL and trended towards significance when assessed by TA ($p = 1.2 \times 10^{-3}$ and 6.5×10^{-2} , respectively TABLE 2.2A). In addition, we calculated SDI and richness for patients in the PNA-ind group with culture-negative suspected pneumonia. Unlike the PNA-pos subjects, we did not observe significant differences in terms of SDI and richness compared to the PNA-neg patients for either fluid type, ($p = 0.211$ and $p = 0.679$ for SDI of mBAL and TA, respectively; $p = 0.156$ and $p = 0.756$ for richness of mBAL and TA, respectively).

To further explore differences in sample type as a function of pneumonia status, we compared mBAL versus TA across each patient subgroup (PNA-pos, PNA-ind, and PNA-neg) with respect to Shannon Diversity and Richness (FIGURE 2.3). While overall, mBAL samples trended towards increased community richness, we observed that this was driven largely by differences in the PNA-neg group. In PNA-neg subjects, significant differences between sample types were observed with respect to SDI and richness ($p = 5.0 \times 10^{-3}$ and 4.0×10^{-2} , respectively), as well as by Spearman correlation (0.19 (-0.22 – 0.55)). In contrast, in PNA-pos subjects, no differences were observed between fluid types based on SDI or richness ($p = 0.46$ and 0.88 , respectively).

2.5 DISCUSSION

Advances in genome sequencing have revealed that the lung, previously considered sterile, supports diverse microbial communities that play a role in both health and disease⁴⁵. Using shotgun mNGS, we compared the microbial compositions of matched mBAL and TA samples from critically ill adults. Across all patient groups, moderate differences were observed based on Spearman correlation, differences approached significance with respect to alpha diversity (SDI, $p = 0.057$), and richness was significantly higher in mBAL samples ($p = 0.046$). In contrast, we did not find systematic differences in the abundance of oropharyngeal microbes or in beta diversity, measured by Bray-Curtis index. Notably, however, we found that fluid type differences became inconsequential in the setting of clinically identified pneumonia and became more pronounced in patients with non-infectious acute respiratory illnesses.

Prior studies using 16S rRNA amplicon sequencing have observed differences in community richness in the setting of pneumonia^{56,57}, and although we only found significant differences by mBAL, those for TA trended towards significance and may have demonstrated an association with a larger sample size. Lastly, despite the historical assumption that TA specimens are compromised by oropharyngeal contamination, we found that abundance of oropharyngeal microbiota did not significantly differ by sample type.

Reflective of current practices in the ICU, the majority of patients in this study received broad-spectrum antibiotics prior to sample collection. As such, the possibility that antibiotic exposure may have driven compositional similarity between fluid types must

be considered. The observation that a greater fraction of PNA-neg versus PNA-pos patients received antibiotics, however, suggests this may be less likely.

Together, our data indicate that from a metagenomic perspective, TA sampling is an effective alternative to more invasive mBAL testing for patients with pneumonia, a conclusion consistent with findings of prior clinical studies^{46,47} and the Clinical Practice Guidelines from the Infectious Diseases Society of America and the American Thoracic Society⁴⁸. Future studies with a larger sample size may clarify trends in diversity differences that approached, but did not reach, significance. These results may help inform both culture-independent clinical microbiologic testing, and research on the lung microbiome.

Data availability: Data are available publicly via BioProject Accession ID PRJNA445982. TABLE 2.4 and analyses of primary sequencing data including pre-processed data files and an R markdown file with documentation are available at: (dx.doi.org/10.17504/protocols.io.wqnfdve).

2.6 FIGURES

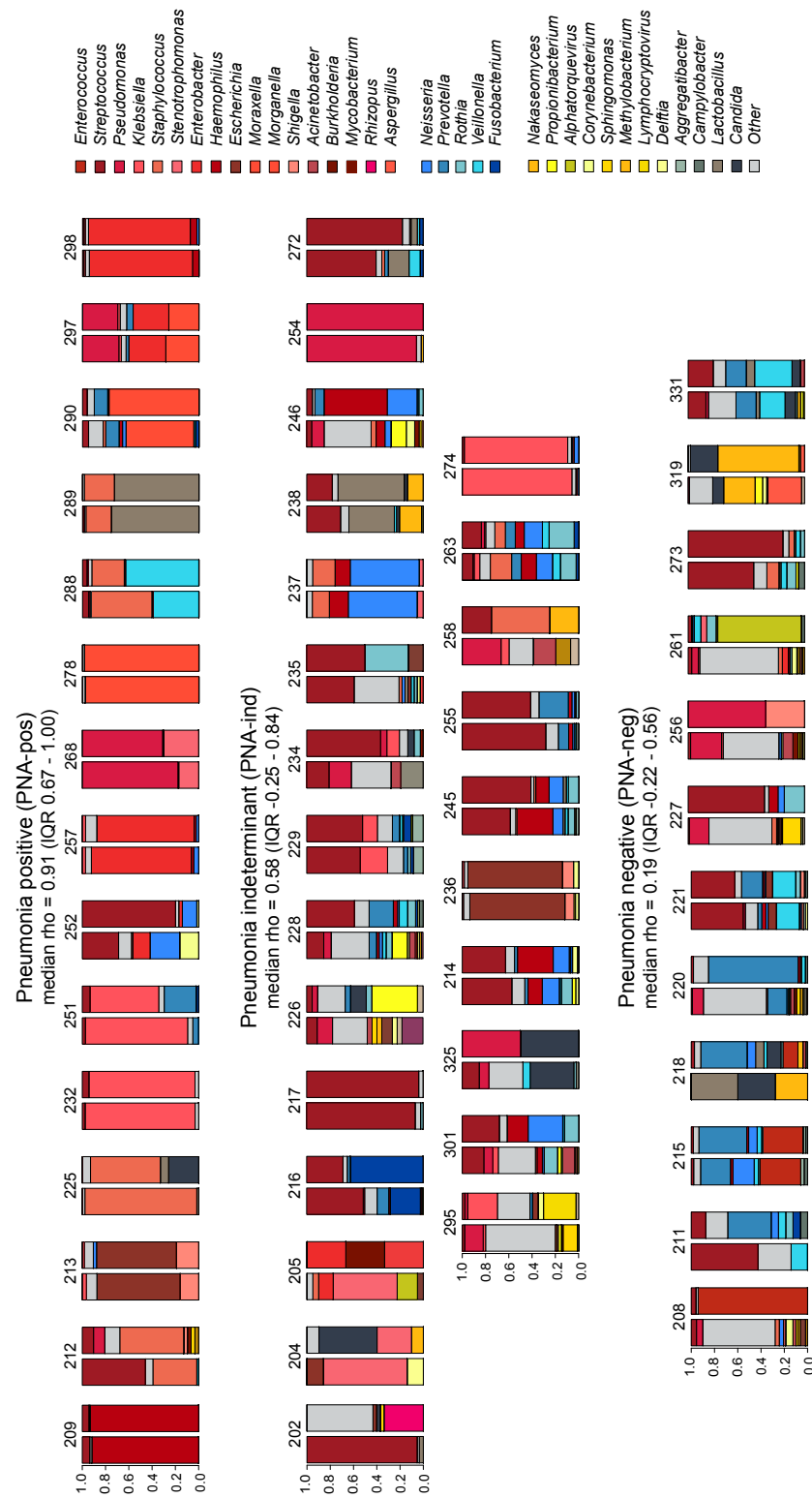


FIGURE 2.1 Fraction of total microbial sequencing reads in mBAL and TA samples

Fraction of total microbial sequencing reads represented by each genus in matched mini-bronchial alveolar lavage (mBAL, left column) and tracheal aspirate (TA, right column) specimens. A summary of the spearman correlation ρ values for pairwise comparisons in each pneumonia (PNA) group is listed under each header. Legend colors correspond to the top most abundant microbes across all samples, with red shading indicating microbes with established respiratory pathogenicity as recently defined¹¹⁶, blue indicating genera known to be common oropharyngeal microbiota, and yellow/grey indicating other microbial genera. Category "Other" refers to all other microbes identified and those identified at abundance < 1%.

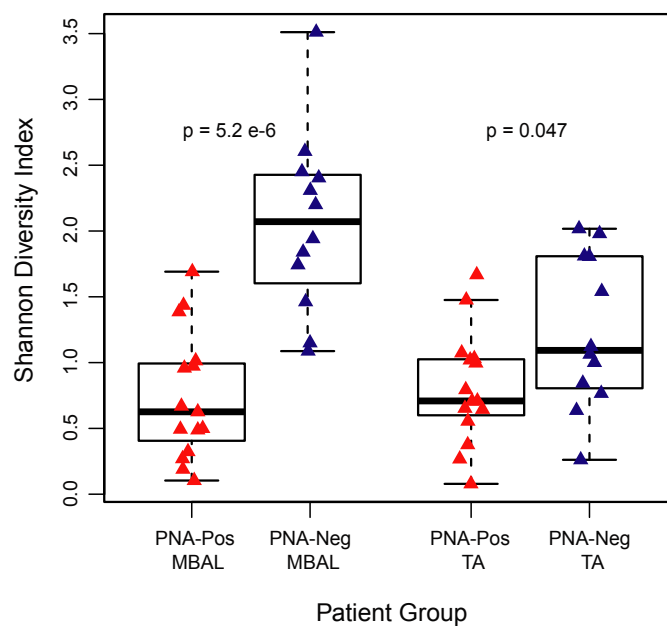


FIGURE 2.2 Shannon diversity index as a PNA biomarker in mBAL versus TA

Shannon diversity index (SDI) in pneumonia-positive subjects (PNA-pos, red) versus pneumonia-negative subjects (PNA-neg, blue) by specimen type (mBAL, left; TA, right). PNA-pos subjects had lower diversity compared to PNA-neg subjects when assessed by either specimen type.

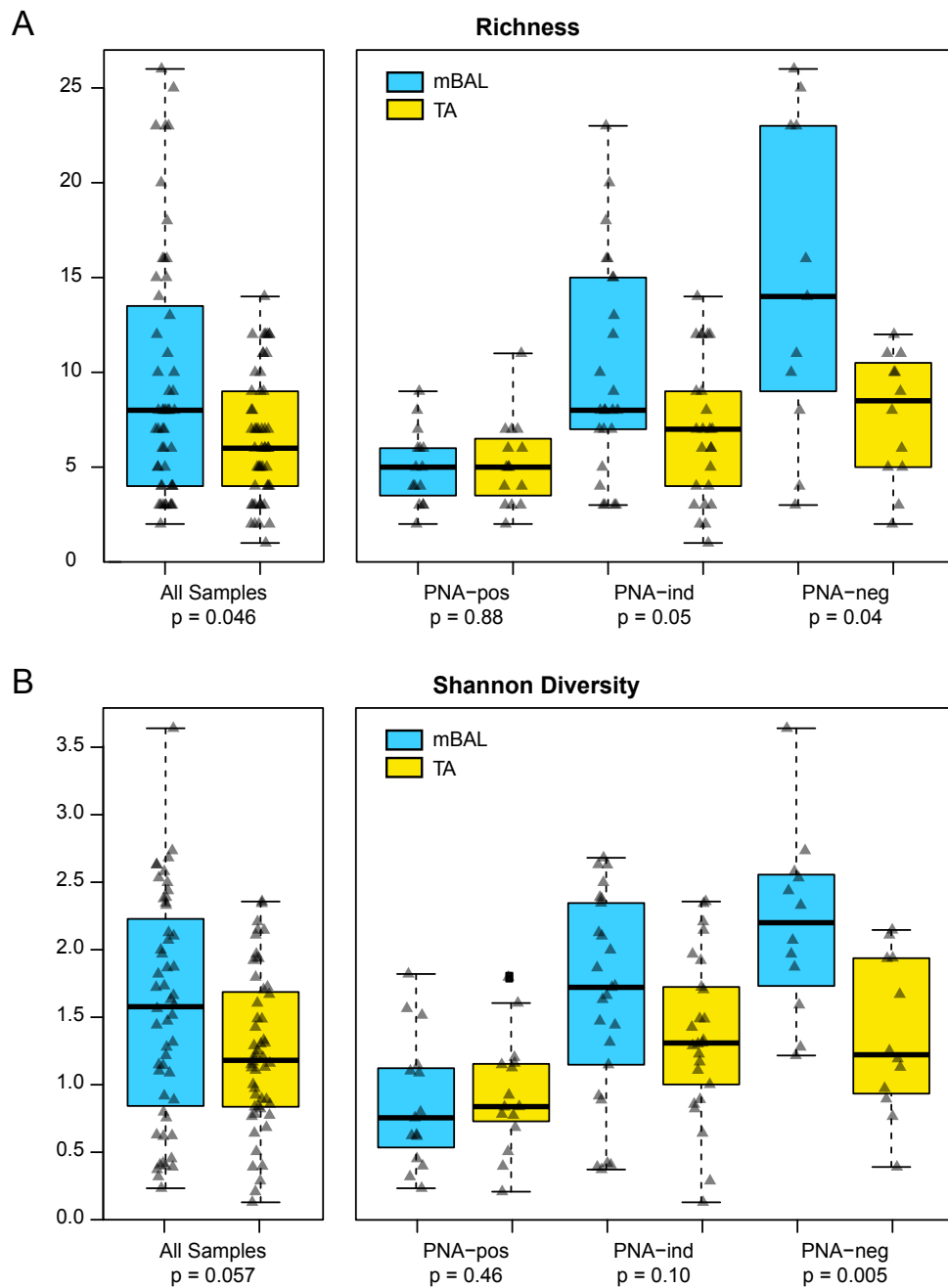


FIGURE 2.3 Sample-type differences in Richness and Shannon diversity index

A) Sample-type differences in Richness as a function of pneumonia status, with overall comparison of the entire cohort (left box) and comparisons within each group (right box, PNA-pos, PNA-ind, and PNA-neg, respectively). P-values, evaluated by Wilcoxon rank sums, are shown below each comparison. **B)** Sample-type differences in Shannon diversity index as a function of pneumonia status.

2.7 TABLES

TABLE 2.1 Demographics and clinical characteristics of mBAL vs TA study cohort

Demographics and clinical characteristics of the study cohort are shown. Values are calculated for the whole cohort (Total) as well as each of the three groups used for analysis (PNA-pos, PNA-neg, and PNA-ind).

	Total (n)	PNA-pos (n)	PNA-neg (n)	PNA-ind (n)
Total enrolled	52	15	12	25
Age, average years (range)	63	60	64	64
Female gender	17 (33%)	3 (20%)	7 (58%)	7 (28%)
Race:				
African American	3 (6%)	1 (7%)	0 (0%)	2 (8%)
Asian	14 (27%)	4 (27%)	2 (15%)	8 (32%)
Caucasian	27 (52%)	8 (53%)	8 (67%)	11 (44%)
Other	4 (8%)	0 (0%)	2 (17%)	2 (8%)
Hispanic Ethnicity	4 (8%)	2 (13%)	0 (0%)	2 (8%)
Suspected Pneumonia Type:				
Community acquired pneumonia	22 (42%)	8 (53%)	0 (0%)	14 (56%)
Hospital acquired pneumonia	8 (15%)	5 (33%)	0 (0%)	3 (12%)
Ventilator associated pneumonia	2 (4%)	2 (13%)	0 (0%)	0 (0%)
Indeterminant pneumonia status	8 (15%)	0 (0%)	0 (0%)	8 (32%)
Immunosuppression	18 (35%)	6 (40%)	4 (33%)	8 (32%)
Prior antibiotic use	43 (83%)	11 (73%)	12 (100%)	20 (80%)
Bacteremia	12 (23%)	4 (27%)	2 (17%)	6 (25%)
Mortality (30 day)	14 (27%)	4 (27%)	2 (17%)	8 (32%)

TABLE 2.2 Microbial community metrics and oropharyngeal contaminants

A) Microbial community metrics - Richness and Shannon diversity index (SDI) - in tracheal aspirate (TA) and mini-bronchial alveolar lavage (mBAL) samples, including differences between subjects with pneumonia (PNA-pos) and those with respiratory failure due to etiologies other than respiratory infection (PNA-neg). **B)** Abundance differences (reads per million reads mapped, rpm) for common oropharyngeal microbes by genus rpm between all TA and mBAL samples, irrespective of PNA group, calculated using Wilcoxon rank sum.

Legend: IQR: Interquartile range.

A)	TA		mBAL		mBAL vs. TA (all samples)
	Metric	Median (IQR)	PNA-pos vs. PNA-neg	Median (IQR)	PNA-pos vs. PNA-neg
	Richness	6.00 (4.00-9.00)	$p = 6.5 \times 10^{-2}$	8.00 (4.00-13.50)	$p = 1.2 \times 10^{-3}$
	Shannon Diversity	1.05 (0.71-1.55)	$p = 4.7 \times 10^{-2}$	1.45 (0.74-2.05)	$p = 5.2 \times 10^{-6}$
B)	TA		mBAL		mBAL vs. TA
	Genus	mean rpm (IQR)		mean rpm (IQR)	
	<i>Prevotella</i>	0.07 (0.00 - 0.06)		0.02 (0.00 - 0.03)	$p = 0.15$
	<i>Veillonella</i>	0.03 (0.00 - 0.02)		0.03 (0.00 - 0.02)	$p = 0.99$
	<i>Streptococcus</i>	0.20 (0.00 - 0.33)		0.18 (0.00 - 0.34)	$p = 0.88$
	<i>Fusobacterium</i>	0.02 (0.00 - 0.00)		0.01 (0.00 - 0.00)	$p = 0.53$
	<i>Rothia</i>	0.03 (0.00 - 0.03)		0.01 (0.00 - 0.00)	$p = 0.31$
	<i>Neisseria</i>	0.04 (0.00 - 0.02)		0.03 (0.00 - 0.02)	$p = 0.72$

TABLE 2.3 Patient diagnoses and clinical microbiology testing

Patient diagnoses and results of clinical microbiology testing for bacterial and viral pathogens, separated by pneumonia group. **A)** PNA-pos patients; patients with clinically identified pneumonia with a positive bacterial culture. For patients with positive clinical microbiology results, the rank of each bacterial genus based on abundance (rpm) of sequencing reads in matched mBAL or TA specimens is listed. Samples for which clinical diagnostics returned polymicrobial cultures contain multiple lines and are noted as *polymicrobial*. Staphylococcus in sample 252 (marked by *) was present at < 1% relative abundance. **B)** PNA-neg patients, subjects with a clear alternative non-infectious etiology of acute respiratory failure. **C)** PNA-ind patients, patients with acute respiratory illnesses of indeterminate etiology including patients with negative bacterial cultures but suspected bacterial pneumonia based on clinical criteria alone.

A)

ID	PNA Group	Diagnosis	Single vs. Polymicrobial	Microbe(s) detected by Culture	Rank, TA	Rank, mBAL
209	PNA-pos	Community acquired pneumonia	<i>polymicrobial</i>	<i>Heamophilus</i>	1	1
				<i>Streptococcus</i>	2	2
212	PNA-pos	Hospital acquired pneumonia	<i>polymicrobial</i>	<i>Staphylococcus</i>	2	1
213	PNA-pos	Ventilator associated pneumonia	<i>polymicrobial</i>	<i>Escherichia</i>	1	1
				<i>Klebsiella</i>	4	5
225	PNA-pos	Community acquired pneumonia	<i>single microbe</i>	<i>Staphylococcus</i>	1	1
232	PNA-pos	Community acquired pneumonia	<i>single microbe</i>	<i>Klebsiella</i>	1	1
251	PNA-pos	Community acquired pneumonia	<i>single microbe</i>	<i>Klebsiella</i>	1	1
252	PNA-pos	Community acquired pneumonia	<i>polymicrobial</i>	<i>Streptococcus</i>	1	1
				<i>Enterobacter</i>	4	3
				<i>Staphylococcus</i>	7	*
257	PNA-pos	Hospital acquired pneumonia	<i>single microbe</i>	<i>Enterobacter</i>	1	1
268	PNA-pos	Community acquired pneumonia	<i>polymicrobial</i>	<i>Pseudomonas</i>	1	1
				<i>Stenotrophomonas</i>	2	2
278	PNA-pos	Hospital acquired pneumonia	<i>single microbe</i>	<i>Moraxella</i>	1	1
288	PNA-pos	Community acquired pneumonia	<i>single microbe</i>	<i>Staphylococcus</i>	1	2
289	PNA-pos	Community acquired pneumonia	<i>single microbe</i>	<i>Staphylococcus</i>	2	2
290	PNA-pos	Community acquired pneumonia	<i>single microbe</i>	<i>Moraxella</i>	1	1
297	PNA-pos	Hospital acquired pneumonia	<i>polymicrobial</i>	<i>Enterobacter</i>	1	1
				<i>Morganella</i>	2	2
				<i>Klebsiella</i>	3	3
				<i>Pseudomonas</i>	5	5
298	PNA-pos	Community acquired pneumonia	<i>single microbe</i>	<i>Enterobacter</i>	1	1

Table continued below.

B)

ID	PNA Group	Diagnosis	Microbe(s) detected by Culture or PCR
208	PNA-neg	Septic shock due to <i>Enterococcus</i> bacteremia	
211	PNA-neg	Acute myocardial infarction	
215	PNA-neg	Craniotomy for resection of arterio venous malformation	
218	PNA-neg	Small bowel obstruction, pancreatitis, hypoxic respiratory failure	
220	PNA-neg	Elective craniotomy for aneurysm bypass surgery	
221	PNA-neg	Seizure	
227	PNA-neg	Hemorrhagic shock and pulseless electrical activity cardiac arrest	
256	PNA-neg	Status post heart transplant with suspected drug-related fever	
261	PNA-neg	Volume overload secondary to constrictive pericarditis	
273	PNA-neg	Hemorrhagic stroke	
319	PNA-neg	Status post balloon angioplasty of hepatic artery	
331	PNA-neg	Acute liver failure secondary to portal vein thrombosis	

C)

ID	PNA Group	Diagnosis	Microbe(s) detected by Culture or PCR
202	PNA-ind	MSSA bacteremia and septic shock, suspected pneumonia	
204	PNA-ind	Osteomyelitis and hypercarbic respiratory failure	
205	PNA-ind	Community acquired pneumonia	<i>Rhinovirus</i>
214	PNA-ind	Intracranial hemorrhage	
216	PNA-ind	COPD exacerbation with hypercarbic respiratory failure, suspected pneumonia	
217	PNA-ind	Hypoxic respiratory failure, suspected pneumonia	
226	PNA-ind	Hypoxic respiratory failure, suspected pneumonia	
228	PNA-ind	Community acquired pneumonia	<i>RSV</i>
229	PNA-ind	Altered mental status and septic shock, suspected pneumonia	
234	PNA-ind	Severe sepsis and hypoxia with urinary tract infection	
235	PNA-ind	Acute renal failure, shock and hypoxic respiratory failure, suspected pneumonia	
236	PNA-ind	Aspiration secondary to altered mental status	
237	PNA-ind	Seizure and hypernatremia, suspected pneumonia	
238	PNA-ind	Hypoxemic and hypercarbic respiratory failure following PEA arrest	
245	PNA-ind	Seizure, altered mental status, suspected aspiration	
246	PNA-ind	Hypoxic respiratory failure, suspected pneumonia	
254	PNA-ind	Respiratory failure and shock, suspected pneumonia	
255	PNA-ind	Status post aortic valve repair	
258	PNA-ind	Subarachnoid hemorrhage	
263	PNA-ind	Altered mental status, hypoxic respiratory failure	
272	PNA-ind	Sepsis with possible aspiration pneumonia	
274	PNA-ind	Altered mental status following perioperative intubation	
295	PNA-ind	Community acquired pneumonia	<i>RSV</i>
301	PNA-ind	Acute respiratory distress syndrome	
325	PNA-ind	Hypoxic respiratory failure	

TABLE 2.4 Top five most abundant microbes per sample

The top five most abundant microbes by reads per million reads mapped for each sample type (mBAL and TA) for each patient. Due to the large table size, only the first five patients are shown below, but the full supplemental table is available at dx.doi.org/10.17504/protocols.io.wqnfvdve

Patient ID	Top 5 Microbes Detected by mBAL	mBAL rpm	Top 5 Microbes Detected by TA	TA rpm
209	<i>Haemophilus</i>	148.49	<i>Haemophilus</i>	421.5
	<i>Streptococcus</i>	10.35	<i>Streptococcus</i>	25.61
	<i>Betatorquevirus</i>	0.92	<i>Pasteurella</i>	3.94
	<i>Pasteurella</i>	0.57	NA	NA
	<i>Neisseria</i>	0.23	NA	NA
212	<i>Streptococcus</i>	5.37	<i>Staphylococcus</i>	39.79
	<i>Staphylococcus</i>	3.73	<i>Pseudomonas</i>	7.02
	<i>Veillonella</i>	0.16	<i>Streptococcus</i>	7.02
	<i>Mycoplasma</i>	0.11	<i>Azospirillum</i>	2.34
	<i>Haemophilus</i>	0.08	<i>Burkholderia</i>	2.34
213	<i>Escherichia</i>	59.15	<i>Escherichia</i>	346.7
	<i>Shigella</i>	13.16	<i>Shigella</i>	96.06
	<i>Serratia</i>	3.59	<i>Serratia</i>	14.05
	<i>Klebsiella</i>	2.92	<i>Neisseria</i>	13.76
	<i>Citrobacter</i>	1.22	<i>Klebsiella</i>	9.42
225	<i>Staphylococcus</i>	20.2	<i>Staphylococcus</i>	0.51
	<i>Lactobacillus</i>	0.34	<i>Candida</i>	0.22
	<i>Pseudomonas</i>	0.11	<i>Collimonas</i>	0.06
	<i>Citrobacter</i>	0.06	<i>Lactobacillus</i>	0.06
	<i>Corynebacterium</i>	0.06	NA	NA
232	<i>Klebsiella</i>	719.7	<i>Klebsiella</i>	732.44
	<i>Streptococcus</i>	18.25	<i>Streptococcus</i>	45.54
	<i>Citrobacter</i>	13.4	<i>Citrobacter</i>	12.67
	<i>Enterobacter</i>	4.27	<i>Enterobacter</i>	4.39
	<i>Yersinia</i>	1.85	<i>Yersinia</i>	2.51

3 INTEGRATING HOST RESPONSE AND UNBIASED MICROBE DETECTION FOR LOWER RESPIRATORY TRACT INFECTION DIAGNOSIS IN CRITICALLY ILL ADULTS

3.1 ABSTRACT

Lower respiratory tract infections lead to more deaths each year than any other infectious disease category. Despite this, etiologic LRTI pathogens are infrequently identified due to limitations of existing microbiologic tests. In critically ill patients, noninfectious inflammatory syndromes resembling LRTIs further complicate diagnosis. To address the need for improved LRTI diagnostics, we performed mNGS on tracheal aspirates from 92 adults with acute respiratory failure and simultaneously assessed pathogens, the airway microbiome, and the host transcriptome. To differentiate pathogens from respiratory commensals, we developed a rules-based model (RBM) and logistic regression model (LRM) in a derivation cohort of 20 patients with LRTIs or noninfectious acute respiratory illnesses. When tested in an independent validation cohort of 24 patients, both models achieved accuracies of 95.5%. We next developed pathogen, microbiome diversity, and host gene expression metrics to identify LRTI-positive patients and differentiate them from critically ill controls with noninfectious acute respiratory illnesses. When tested in the validation cohort, the pathogen metric performed with an AUC of 0.96 (95% CI, 0.86–1.00), the diversity metric with an AUC of 0.80 (95% CI, 0.63–0.98), and the host transcriptional classifier with an AUC of 0.88 (95% CI, 0.75–1.00).

Combining these achieved a negative predictive value of 100%. This study suggests that a single streamlined protocol offering an integrated genomic portrait of pathogen, microbiome, and host transcriptome may hold promise as a tool for LRTI diagnosis.

3.2 INTRODUCTION

Lower respiratory tract infections are a leading cause of mortality worldwide^{6,58,59}. Early and accurate determination of acute respiratory disease etiology is crucial for implementing effective pathogen-targeted therapies but is often not possible due to the limitations of current microbiologic tests in terms of sensitivity, speed, and spectrum of available assay targets²¹. For instance, even with the best available clinical diagnostics, a contributory pathogen can be detected in only 38% of adults with community acquired pneumonia, due to the low sensitivity and time requirements of culture, and the limited number of microbes detectable by serologic and PCR assays^{21,60}.

In the absence of a definitive microbiologic diagnosis, clinicians may presume symptoms are due to a noninfectious inflammatory condition and initiate empiric corticosteroids, which can exacerbate an occult infection⁶¹. Furthermore, even with negative microbiologic testing, providers often continue empiric antibiotics due to concerns of falsely negative results, a practice that drives emergence of antibiotic resistance and increases risk of *Clostridium difficile* infection⁶². In the intensive care unit (ICU), LRTI diagnosis is particularly complex due to a high prevalence of noninfectious inflammatory conditions with overlapping clinical features²⁰ and a patient demographic that includes severely immunocompromised individuals who may exhibit atypical presentations of pulmonary infections.

Advancements in genome sequencing hold promise for overcoming these diagnostic challenges by affording culture-independent assessment of microbial genomes from microliter volumes of clinical samples^{63,64}. Recent work has highlighted the utility of mNGS for rapid and

actionable diagnosis of complicated infections^{61,65–67}. While these results are encouraging, most mNGS computational pipelines have been developed for analysis of sterile fluids or cultured bacterial isolates and have limited capacity to identify pathogens amid the complex background of commensal microbiota present in respiratory specimens^{28,67,68}.

Host transcriptional profiling from peripheral blood has emerged as a promising alternative to pathogen-based diagnostics that can distinguish viral from bacterial LRTIs as well as differentiate between patients with acute respiratory infections versus those with noninfectious illnesses^{31,32,60}. This approach, while highly promising, has not been well studied in ICU patients with respiratory failure or in severely immunocompromised subjects. Furthermore, host transcriptional profiling has not yet been coupled with simultaneous detection of pulmonary pathogens^{60,69}, which could improve diagnostic accuracy and more precisely inform optimal antimicrobial treatment.

mNGS can extend both host gene expression assays and current microbe-based diagnostics by simultaneously detecting pathogens, the airway microbiome, and transcriptional biomarkers of the host's immune response. Here, we address the need for better LRTI diagnostics by developing an mNGS-based method that integrates host response and unbiased microbe detection. We then evaluate the performance of this approach in a prospective cohort of critically ill patients with acute respiratory failure.

3.3 RESULTS

We prospectively enrolled 92 adults admitted to the ICU with acute respiratory failure and collected tracheal aspirate samples within 72 h of intubation (TABLE 3.1). Patients underwent testing with clinician-ordered standard of care microbiologic diagnostics at the University of California, San Francisco, Moffitt–Long Hospital, a tertiary-care referral center. Subjects with LRTI were identified by two-physician adjudication

using US Centers for Disease Control/National Healthcare Safety Network (CDC/NHSN) surveillance case definitions and retrospective electronic medical record review, with blinding to mNGS results (TABLE 3.3)⁴⁹. Using this approach, patients were assigned to one of four groups: (i) LRTI defined by both clinical and microbiologic criteria (LRTI^{+C+M}, n = 26); (ii) no evidence of LRTI and a clear alternative explanation for acute respiratory failure (no-LRTI, n = 18); (iii) LRTI defined by clinical criteria alone with negative conventional microbiologic testing (LRTI^{+C}, n = 34); and (iv) respiratory failure due to unclear cause, infectious or noninfectious (unk-LRTI, n = 14).

From extracted nucleic acid samples, we performed both metagenomic shotgun DNA-sequencing (DNA-Seq) as well as RNA-sequencing. We first developed computational algorithms to sift respiratory pathogens from background commensal flora in an effort to enhance detection of LRTI etiology. To differentiate patients with LRTI from those with noninfectious critical respiratory illnesses, we next developed metrics of LRTI probability based on pathogen, airway microbiome diversity, and host gene expression (FIGURE 3.1). To assess assay performance, we focused on the most unambiguously LRTI-positive and -negative subjects (LRTI^{+C+M} and no-LRTI) by randomly dividing them into independent derivation (n = 20, used for model training) and validation cohorts (n = 24, used for model testing). Each metric (pathogen, microbiome, and host) was evaluated independently and then in combination.

3.3.1 Pathogen Detection

While many NGS platforms utilize only one nucleic acid type, we combined both RNA-Seq and DNA-Seq. This approach allowed for simultaneous host transcriptional

profiling, permitted detection of RNA viruses, and enriched for actively transcribing microbes (versus latent or nonviable taxa). In addition, requiring concordant detection of microbes across both nucleic acid types reduced spurious alignments derived from reagent contaminants intrinsic to the library preparations of each nucleic acid type (20). From each TA sample, we generated a mean of 19.6 and 32.6 million paired-end sequencing reads, from DNA-Seq and RNA-Seq, respectively, of which the median fraction of microbial reads was 0.04% (IQR, 0.01–0.16%). Raw reads were analyzed using a rapid computational pipeline that aligns and classifies microbial taxa by nucleotide and peptide translation using the National Center for Biotechnology Information (NCBI) NT and NR databases, respectively^{50,67}. RNA-Seq yielded a greater abundance of sequences compared with DNA-Seq for 78% of identified microbes, with a median of 2.2 times more reads per microbe.

We and others have previously developed NGS methodologies for “sterile site” clinical fluids such as cerebrospinal fluid^{28,67}. The lung, however, is not a sterile environment and in fact harbors microbial communities during states of both health and disease^{29,45,70,71}. Asymptomatic carriage of potentially pathogenic organisms is common^{72,73}, and only in a subset of cases do these microbes overtake airway microbial communities and precipitate LRTI⁷⁴. As such, distinguishing legitimate pathogens from commensal or colonizing microbiota is a central challenge for LRTI diagnostics and adds complexity to the interpretation of metagenomic sequencing data. To this point, while we detected all 38 pathogens identified from clinician-ordered microbiologic tests in the 26 LRTI^{+C+M} patients using mNGS (TABLE 3.5), a 10-fold greater number of airway commensals were also identified. The most prevalent microbes in the no-LRTI patient

group included well-known commensal taxa (TABLE 3.7). Thus, to distinguish probable pathogens from airway commensals, we developed two complementary algorithms: (i) an RBM optimized for detecting well-established respiratory pathogens, and (ii) a more flexible LRM that also permitted novel pathogen detection (FIGURE 3.1).

The goal of both models was to correctly identify pathogens amid abundant and heterogeneous populations of commensals. Microbes identified by clinician-ordered diagnostics plus all viruses with established respiratory pathogenicity in the LRTI^{+C+M} group were categorized as pathogens (n = 12 in derivation cohort and n = 26 in validation cohort; TABLE 3.2). Any additional microbes identified by mNGS were considered commensals (n = 155 in derivation cohort; n = 174 in validation cohort). We accepted that this “practical” gold standard would provide an attenuated estimate of performance due to the sensitivity limitations of microbial culture in the setting of antibiotic preadministration²¹.

In the RBM, respiratory microbes from each patient were assigned an abundance score based on the sum of log(RNA-Seq) and log(DNA-Seq) genus reads per million reads mapped (rpm) (TABLE 3.5). After ranking microbes by this abundance score, the greatest score difference between sequentially ranked microbes was identified and used to distinguish the group of highest-scoring microbes within each patient (FIGURE 3.2A and FIGURE 3.7). These high-scoring microbes plus all RNA viruses detected at a conservative threshold of >0.1 rpm were indexed against an *a priori* developed table of established lower respiratory pathogens derived from landmark surveillance studies and clinical guidelines (TABLE 3.4) and, if present, were identified as putative pathogens by the RBM^{21,48,75,76}.

The RBM achieved an accuracy for pathogen detection of 98.8% and 95.5% in the derivation and validation cohorts, respectively (TABLE 3.5). In subjects whose respiratory cultures grew three or more different bacteria, mNGS was able to detect each of the species. In most cases, however, their abundance differed by several 100-fold, which confounded detection of the lower abundance taxa (TABLE 3.5). Given the unclear significance of single species in such polymicrobial cases with respect to pathogenicity⁷⁷, we performed a secondary analysis in which only the most abundant microbe was considered a pathogen, and this approach yielded an accuracy of 98.4%.

While the RBM performed well for identifying microbes with established pulmonary pathogenicity, we recognized the need to also detect novel or atypical species. We thus employed machine learning to distinguish respiratory pathogens from commensals using an LRM trained on microbes detected in the derivation cohort patients (n = 20) using the predictor variables of RNA-Seq rpm, DNA-Seq rpm, rank by RNA-Seq rpm, established LRTI pathogen (yes/no), and virus (yes/no). These features were selected to preferentially favor highly abundant organisms with established pathogenicity in the lung, but still permit detection of uncommon taxa that could represent putative pathogens.

To evaluate LRM performance in the derivation cohort, we performed leave-one-patient-out cross-validation, in which all microbes from a single patient were held out in each round of cross-validation. This yielded an AUC of 0.90 (95% CI, 0.76–0.99). A final model was trained on all microbes from derivation cohort patients, and this achieved an AUC of 0.91 (95% CI, 0.83–0.97) for pathogen identification in the validation cohort (FIGURE 3.2B and TABLE 3.5 and 3.6). At an optimized probability threshold of 0.36 (Methods), this translated to an accuracy of 96.4% and 95.5% in the derivation and

validation cohorts, respectively. As with the RBM, LRM performance suffered in polymicrobial culture cases with species that differed by several magnitudes in abundance when assessed by mNGS. As such, when only the most abundant microbe identified by clinical microbiologic diagnostics per LRTI^{+C+M} patient was considered as the etiologic pathogen, the AUC increased to 0.997 (95% CI, 0.99–1.00) in the validation cohort.

Combining the RBM and LRM identified more putative pathogens than either model alone and revealed a potential LRTI etiology in 62% (n = 21) of the LRTI^{+C} patients with clinically adjudicated LRTI but negative microbiologic testing (FIGURE 3.3, FIGURE 3.8, and TABLE 3.5). Compared with clinician-ordered diagnostics, this permitted a microbiologic diagnosis in a greater number of LRTI-positive subjects (78% vs. 43%; $P < 1.00 \times 10^{-4}$ by McNemar's test; FIGURE 3.3). Putative new pathogens in a representative subset of the LRTI^{+C} group patients (n = 11; 32%) were orthogonally confirmed by clinical multiplex respiratory virus PCR, influenza C PCR⁷⁸, or by 16S bacterial rRNA gene sequencing (TABLE 3.5).

Putative pathogens identified in the unk-LRTI group (n = 6, 42%) may have represented atypically presenting respiratory infections or incidental carriage in the respiratory tract (FIGURE 3.8 and TABLE 3.5). Microbes identified in the no-LRTI group (n = 3; 17%) were present at lower abundance compared with microbes in LRTI^{+C+M} subjects ($P < 0.01$ by Wilcoxon rank sum), LRTI^{+C} ($P < 0.01$), and unk-LRTI subjects ($p = 0.02$), and included contextual pathogens such as *Streptococcus pneumoniae* and *Haemophilus influenzae* that colonize the airways of 20–50% of healthy individuals^{77,79,80}. Together, these findings highlighted the reality of asymptotic carriage of potentially

pathogenic species, emphasizing the need to contextualize microbial detection with respect to other key elements of an airway infection, in particular the airway microbiome and the host's immune response^{72,81}. We thus undertook further analytical development to predict LRTI status by calculating combined metrics based on pathogen, microbiome, and host transcriptional response.

3.3.2 *LRTI Prediction Based on Pathogen.*

We recognized that the highest per-patient LRM pathogen versus commensal probability value differed significantly between LRTI^{+C+M} and no-LRTI subjects ($p = 3.8 \times 10^{-4}$ by Wilcoxon rank sum). As such, we hypothesized that this value might have utility not only for pathogen versus commensal prediction, but also for LRTI prediction in general. Testing this idea, we found that the maximum per patient LRM probability value predicted LRTI status with an AUC of 0.97 (95% CI, 0.90–1.00) in the derivation cohort and 0.96 (95% CI, 0.86–1.00) in the validation cohort (FIGURE 3.9).

3.3.3 *LRTI Prediction Based on Lung Microbiome Diversity.*

Several studies have demonstrated reduced diversity of the airway microbiome in the setting of LRTI^{50,55,82,83}. We measured intrapatient (α) diversity of airway genera using the Shannon diversity index and found that LRTI^{+C+M} subjects had significantly lower SDI compared with no-LRTI subjects when assessed by both RNA-Seq (FIGURE 3.4A; $p = 1.3 \times 10^{-4}$) and DNA-Seq (FIGURE 3.10A; $p = 8.9 \times 10^{-3}$) (TABLE 3.8). We next examined interpatient (β) diversity (40) using the Bray–Curtis Index⁸⁴ and found that this also differed between LRTI^{+C+M} and no-LRTI subjects, with assessment by RNA-Seq

again yielding a more significant difference versus DNA-Seq [$p = 5 \times 10^{-3}$ versus $p = 9 \times 10^{-3}$ by permutation analysis of variance (PERMANOVA), respectively; FIGURE 3.4B, FIGURE 3.10B]. We then tested whether diversity alone might predict LRTI and found that RNA-Seq SDI differentiated LRTI^{+C+M} from no-LRTI subjects with an AUC of 0.96 (95% CI, 0.89–1.00) in the derivation cohort and an AUC of 0.80 (95% CI, 0.63–0.96) in the validation cohort (FIGURE 3.4C). DNA-Seq SDI did not perform as well, with AUCs of 0.84 (95% CI, 0.66–1.00) and 0.53 (95% CI, 0.25–0.80) in the derivation and validation cohorts, respectively (FIGURE 3.10C). These findings suggested that genus diversity assessed by RNA-Seq was a useful, albeit imperfect, biomarker of LRTI.

3.3.4 LRTI Prediction Based on Host Response.

In the setting of critical illness, systemic inflammatory responses due to diverse physiologic processes can make true LRTI clinically indistinguishable from noninfectious respiratory failure or severe extrapulmonary infection. Consistent with this, we found that the systemic inflammatory response syndrome (SIRS) criteria (temperature, white blood cell count, heart rate, respiratory rate) had limited utility for LRTI detection despite being widely used for infection assessment (TABLE 3.2). We thus hypothesized that transcriptional profiling, which has emerged as a promising and accurate host-based approach for assessing infection, might provide diagnostic insight in settings when clinical rules are uninformative^{31,60,85}.

As such, we examined differential gene expression between LRTI^{+C+M} and no-LRTI subjects in the derivation cohort to define a host transcriptional signature of LRTI in patients with critical illness. Using a false-discovery rate (FDR) of <0.05 , we identified a

total of 882 differentially expressed genes, 414 of which were up-regulated in LRTI^{+C+M} subjects (TABLE 3.9A). Gene set enrichment analysis³³ identified up-regulation of pathways related to innate immune responses, NF- κ B signaling, cytokine production, and the type I IFN response in LRTI^{+C+M} subjects. In comparison, gene expression pathways in the no-LRTI group were enriched for oxidative stress responses and MHC class II receptor signaling (TABLE 3.9B). A sub analysis evaluating differences between viral and bacterial infections in known LRTI^{+C+M} patients identified four differentially expressed genes (RSAD2, OAS3, CXCL2, DUSP2). Genes up-regulated in viral cases (RSAD2, OAS3) were related to the type-1 IFN and antiviral responses, reflecting biologically relevant differences in host response indicative of pathogen type, despite a relatively limited sample size within a heterogeneous cohort and high proportion of immune-compromising conditions in the majority of patients with detected viruses.

We next sought to construct an airway-specific host transcriptional classifier that could differentiate LRTI^{+C+M} patients from no-LRTI subjects by employing machine learning (Methods). Elastic net regularized regression in the derivation cohort identified a 12-gene classifier that was then used to score patients based on a weighted sum of scaled expression values (FIGURE 3.5 A and B and TABLE 3.10). We found that predictive classifier genes up-regulated in LRTI^{+C+M} patients compared with no-LRTI patients included NFAT-5, which plays a role in T-cell function and inducible gene transcription during immune responses⁸⁶; ZC3H11A, which encodes a zinc-finger protein involved in the regulation of cytokine production and immune cell activation⁸⁷; and PRRC2C, which functions in RNA binding and may play a role in hematopoietic progenitor cell differentiation in response to infection⁸⁸. Genes up-regulated in no-LRTI patients

compared with LRTI^{+C+M} patients included the following: CD36, which encodes a macrophage phagocytic receptor involved in scavenging dying/dead cells and oxidized lipids^{89,90}; BLVRB, which is involved in oxidative stress responses⁹¹; EDF1, which contributes to the regulation of nitric oxide release in endothelial cells⁹²; and ENG, an integral membrane glycoprotein receptor that may modulate inflammation and angiogenesis⁹³.

Classifier performance assessed by leave-one-out cross-validation demonstrated an AUC of 0.90 (95% CI, 0.75–1.00) in the derivation cohort and an AUC of 0.88 (95% CI, 0.75–1.00) in the validation cohort (FIGURE 3.5C). Covariates for immune suppression, concurrent nonpulmonary infection, antibiotic use, age, and gender were iteratively incorporated into the regression model, but none was significant enough to be maintained when sparsity was added by elastic net (TABLE 3.11). We tested whether differences in host gene expression could be attributed to enrichment of specific cell types using CIBERSORT⁹⁴ (TABLE 3.12) and found that only M2 macrophages were enriched in the no-LRTI group ($p = 0.03$ by Wilcoxon rank sum).

Finally, given our modest sample size, we tested the statistical power of our host classifier by computing learning curves (Methods). We observed that even with subsampling, the 12 classifier genes were continually represented. While the derivation cohort sample size approached the limit required for robust performance assessment, the analysis suggested that additional patients might lead to further improvement (FIGURE 3.11A). A similar analysis for the pathogen versus commensal LRM indicated that performance metrics had converged with the given microbial sample size, indicating robust performance assessment and sufficient training data (FIGURE 3.11B).

3.3.5 *Evaluation of a Combined LRTI Metric.*

Given the relative success of each independent metric (pathogen, microbiome, and host) for discerning the presence of infection, we asked whether combining them could enhance LRTI detection. We recognized the potential of mNGS to empower a data-driven assessment of a patient's LRTI status during the critical time frame following ICU admission. As such, we developed a readily interpretable compilation of host and pathogen mNGS metrics in a rule-out model designed to maximize LRTI diagnostic sensitivity. This process, which involved optimizing intrametric LRTI positivity thresholds in the derivation cohort and calling positivity based on either the host or pathogen scores (Methods), achieved a sensitivity and specificity of 100% and 87.5%, respectively, in the validation cohort, equating to a negative predictive value of 100% (FIGURE 3.6B). Despite the limitations of a small cohort, we investigated the potential utility of the rule-out model for curbing broad-spectrum antibiotic overuse in the ICU by performing a theoretical calculation in the no-LRTI group to estimate the potential impact of mNGS result availability at 48-h postenrollment. This estimate suggested that a significant reduction in unnecessary empiric antibiotic use could have been possible (78 versus 50 d of therapy; $p = 0.03$).

3.4 DISCUSSION

Of all infectious disease categories, LRTIs impart the greatest mortality both worldwide and in the United States (1). Contributing to this is the rising rate of treatment failure due to antibiotic resistance⁹⁵ and the limited performance of existing diagnostics

for identifying respiratory pathogens^{21,96}. In this prospective cohort study, we describe the use of unbiased mNGS for respiratory infectious disease diagnosis in the ICU. We develop methods that advance pathogen-based genomic diagnostics as well as existing host transcriptional classifier platforms by simultaneously assessing respiratory pathogens, the airway microbiome, and the host transcriptome in a single test to predict LRTI and identify disease etiology. We find that host/pathogen mNGS accurately detects LRTI in patients with acute respiratory failure and can provide a microbiologic diagnosis in cases due to unknown etiology.

Host transcriptional profiling has gained attention as a promising approach to LRTI diagnosis^{31,97} but is understudied in critically ill and immunocompromised patients, who may be the most likely to benefit from this technology. We addressed this gap by interrogating airway gene expression in a critically ill cohort with 45% immunocompromised patients to develop an accurate host transcriptional classifier. Unlike existing classifiers, host–microbe mNGS offers the advantage of simultaneous species-level microbial identification.

The role of commensal lung microbiota in health and disease is an area of active investigation. We corroborated prior findings demonstrating microbiome differences between subjects with respiratory infections and those with noninfectious airway disease^{50,82}. More specifically, we found that LRTI was associated with reduced inpatient α diversity of the airway microbiome and that, collectively, patients with LRTI differed significantly from those without in terms of β diversity and microbial sequence abundance. This diversity difference was more pronounced when assessed by RNA-Seq, potentially due to inclusion of RNA viruses and transcripts from actively replicating

pathogens in infected patients. As a biomarker, RNA-Seq SDI had moderate utility for predicting LRTI; however, it did not enhance performance in combination with the other metrics, perhaps due to negative correlation with microbe score ($r = -0.84$ in the derivation cohort).

Discriminating respiratory pathogens from background commensal microbiota is a key challenge for LRTI diagnostics and is particularly relevant for sensitive molecular assays⁹⁶. We directly addressed this by developing two complementary algorithms (RBM and LRM) that parsed putative pathogens from airway commensals. When combined, these models enabled a microbiologic diagnosis in significantly more patients with LRTI compared with clinician-ordered diagnostics. The fact that the a priori selected model features successfully differentiated pathogens from commensals validated the underlying model assumptions related to pathogen dominance resulting in disruption of α diversity. Notably, both models also proved useful despite widespread antibiotic use before airway sampling (90% of subjects), a practice that occurs commonly and that can sterilize microbial cultures³¹.

The capacity for mNGS to detect pathogens unidentifiable by standard clinical diagnostics was highlighted in several cases, including that of subject 254, who developed rapidly worsening respiratory failure and fever during a prolonged postsurgical admission. He was treated empirically for hospital acquired pneumonia with linezolid, aztreonam, and metronidazole. Lower respiratory cultures returned negative, but mNGS identified influenza C, which is not available on most clinical multiplex viral PCR assays. Notably, 12% of subjects were found to have undetected and potentially transmissible respiratory viruses despite strict precautionary respiratory contact policies at the study

site, a finding that suggests the potential value of mNGS for hospital infection control. Several cases also highlighted the potential for mNGS to enhance antibiotic stewardship, and we estimated that theoretical implementation of the rule-out model within 48 h could have reduced antibiotic days of therapy by 36% in the no-LRTI validation cohort patients.

Since at the time of ICU admission it is often difficult to distinguish infectious from noninfectious acute respiratory disease, a theoretical workflow for host/microbe mNGS could involve first employing the rule-out model to assess LRTI probability and complement clinical decision-making regarding discontinuation of empiric antimicrobials. In cases where LRTI was ultimately suspected, a microbiologic diagnosis could then be obtained using a combination of the RBM and LRM to accurately screen for both well-established and uncommon respiratory pathogens. A principal advantage of mNGS is that all potential infectious agents can be simultaneously assessed, which avoids the need for ordering multiple individual tests for each different pathogen of concern. Future studies in a larger validation cohort can help optimize host and microbe LRTI rule-out thresholds and further assess test performance before deployment in a clinical setting.

Some limitations of host/microbe mNGS were apparent and included false-positive detection of pathobionts such as *H. influenzae* and *S. pneumoniae* in the no-LRTI group, and false positivity of the host-response metric in subjects including patient 349, who was diagnosed with α -1 antitrypsin deficiency-associated pulmonary disease. The relatively small sample size of our derivation and validation cohorts increased the potential for data overfitting and was a limitation of our study. Learning curve estimates, however, indicated that the sample size was optimal for pathogen versus commensal prediction, and adequate for the host classifier, consistent with the estimate from an established sample

size prediction tool for high-dimensional classifiers⁹⁸. Nonetheless, a larger cohort will be necessary to improve the robustness of model performance estimates and better assess synergy resulting from combining host and microbial metrics.

Strengths of this study include an innovative bioinformatics approach, detailed patient phenotyping, and a study population reflective of the true heterogeneity of ICU patients, including severely immunocompromised subjects and patients receiving broad-spectrum antibiotics. Future studies in a larger cohort can further validate these findings, strengthen the utility of these models, and assess the impact of mNGS on clinical outcomes. In summary, we report a multifaceted approach to LRTI diagnosis that integrates three central elements of airway infections: the pathogen, airway microbiome, and host's response.

3.5 METHODS

3.5.1 Study Design and Subjects.

This prospective observational study evaluated adults with acute respiratory failure requiring mechanical ventilation who were admitted to the University of California, San Francisco (UCSF) Moffitt–Long Hospital ICUs. Subjects were enrolled sequentially between July 25, 2013, and October 17, 2017, within the first 72 h of intubation for respiratory failure. The UCSF Institutional Review Board approved an initial waiver consent for obtaining excess respiratory fluid, blood, and urine samples, and informed consent was subsequently obtained from patients or their surrogates for continued study participation according to CHR protocol 10-02701. For patients whose surrogates

provided informed consent, follow-up consent was then obtained if patients survived their acute illness and regained the ability to consent. For subjects who died before consent being obtained, a full waiver of consent was approved. For all surviving subjects, if consent was not eventually obtained from either patient or surrogate, all specimens were discarded.

3.5.2 Clinical Microbiologic Testing.

During the period of study enrollment, subjects received standard of care microbiologic testing ordered by the treating clinicians. Respiratory testing from TA, bronchial alveolar lavage (BAL), or mini-BAL included the following: bacterial and fungal stains and semiquantitative cultures (n = 90); AFB stains and cultures (n = 8); 12-target clinical multiplex PCR (Luminex) for influenza A/B, respiratory syncytial virus (RSV), human metapneumovirus (HMPV), human rhinovirus (HRV), adenovirus (ADV), and parainfluenza viruses (PIV) 1–4 (n = 23); Legionella culture (n = 1); Legionella pneumophila PCR (n = 4); cytomegalovirus (CMV) culture (n = 4); and cytology for *Pneumocystis jirovecii* (n = 4). Other microbiologic testing included blood culture (n = 89); urine culture (n = 87); serum cryptococcal antigen (n = 4); serum galactomannan (n = 1); and serum β -D-glucan (n = 1).

3.5.3 Definitions and Clinical Adjudication of LRTI.

Because admission diagnoses made by treating clinicians at the time of study enrollment were by necessity based on incomplete clinical, microbiologic, and treatment outcome information, a post hoc adjudication approach was carried out to enhance

accuracy of LRTI diagnosis. For this, two attending physicians [one from infectious disease (C.L.) and one from pulmonary medicine (F.M.)] blinded to mNGS results, retrospectively reviewed each patient's medical record following hospital discharge or death to determine whether they met the CDC/NHSN surveillance definition of pneumonia, with respect to clinical and/or microbiologic criteria (TABLE 3.2)⁴⁹. Chart review consisted of in-depth analysis of complete patient histories, including laboratory and radiographic results, inpatient notes, and postdischarge clinic notes. Using this approach, subjects were assigned to one of four groups, consistent with a recently described approach³¹: (i) LRTI defined by both clinical and laboratory criteria (LRTI^{+C+M}); (ii) no evidence of respiratory infection and with a clear alternative explanation for respiratory failure (no-LRTI); (iii) LRTI defined by clinical criteria only (LRTI^{+C}); and (iv) unknown, LRTI possible (unk-LRT). A determination of noninfectious etiology was made only if an alternative diagnosis could be established and results of standard clinical microbiological testing for LRTI were negative.

3.5.4 Identification of Subjects with LRTI

Subjects with LRTI were identified by two-physician adjudication, as described above. The Cohen's kappa for physician adjudication was 0.86 (95% CI = 0.77 – 0.93). Disagreement was resolved by discussion involving focused review of each subject's clinical and microbiologic evidence as related to the CDC definition of pneumonia. A third adjudicator was available (CC) in the event that disagreements could not be resolved, however this was not needed.

3.5.5 *Host/Microbe mNGS.*

Excess TA was collected on ice, mixed 1:1 with DNA/RNA Shield (Zymo), and frozen at -80°C . RNA and DNA were extracted from 300 μL of patient TA using bead-based lysis and the Allprep DNA/RNA kit (Qiagen). RNA was reverse transcribed to generate cDNA and used to construct sequencing libraries using the NEBNext Ultra II Library Prep Kit (New England Biolabs). DNA underwent adapter addition and barcoding using the Nextera library preparation kit (Illumina) as previously described⁵⁰. Depletion of abundant sequences by hybridization (DASH) was employed to selectively deplete human mitochondrial cDNA, thus enriching for both microbial and human protein coding transcripts⁹⁹. The final RNA-Seq and DNA-Seq libraries underwent 125-nt paired-end Illumina sequencing on a HiSeq 4000.

3.5.6 *Pathogen Detection Bioinformatics.*

Detection of host transcripts and airway microbes leveraged a custom bioinformatics pipeline⁵⁰ that incorporated quality filtering using PRICESeqfilter⁵² and alignment against the human genome (NCBI GRC h38) using the STAR⁵¹ aligner to extract genecounts. To capture respiratory pathogens, additional filtering to remove Pan troglodytes (UCSC PanTro4) was performed using STAR and removal of nonfungal eukaryotes, cloning vectors, and phiX phage was performed using Bowtie2¹⁰⁰. The identities of the remaining microbial reads were determined by querying the NCBI nucleotide (NT) and nonredundant protein (NR) databases using GSNAP-L and RAPSEARCH2, respectively.

Microbial alignments detected by RNA-Seq and DNA-Seq were aggregated to the genus-level and independently evaluated to determine genus α diversity as described below. The sequencing reads comprising each genus were then evaluated for taxonomic assignment at the species level based on species relative abundance as previously described⁵⁰. For each patient, the top 15 most abundant taxa by RNA rpm were identified and evaluated under the requirement that all bacteria, fungi, and DNA viruses had concordant detection of their genomes by DNA-Seq and concordant alignments in NR and NT. RNA viruses did not require concordant DNA-Seq reads (FIGURE 3.2 and TABLE 3.5). To differentiate putative pathogens from commensal microbiota, we developed RBM and LRM methods and benchmarked each on sequencing data from LRTI^{+C+M} and no-LRTI subjects.

3.5.7 *Statistical Analysis.*

Statistical significance was defined as P less than 0.05, using two-tailed tests of hypotheses. Categorical data were analyzed by χ^2 test and nonparametric continuous variables were analyzed by Wilcoxon rank sum. For statistical validation in the pathogen versus commensal and LRTI prediction metrics, 10 LRTI^{+C+M} and 10 no-LRTI cases were randomly assigned to create a derivation cohort. Model performance was assessed in an independent validation cohort consisting of 16 LRTI^{+C+M} and 8 no-LRTI cases.

3.5.8 *Pathogen Versus Commensal Models.*

We found that all clinically confirmed LRTI pathogens were present within the top 15 most abundant microbes by RNA-Seq rpm, which on average represented 99% of

reads across all samples. We thus limited analysis to the 15 most abundant NGS-detected genera in each sample. For both models, microbes identified using clinician-ordered diagnostics and all viruses with established respiratory pathogenicity in the derivation cohort subjects were considered “pathogens.” Any additional microbes identified by mNGS in these subjects were considered “commensals”. This equated to 12 “pathogens” and 155 “commensals” in the 20 derivation cohort patients, and 26 “pathogens” and 174 “commensals” in the 24 validation cohort patients.

Rules-based model; RBM

This model leveraged previous findings demonstrating that microbial communities in patients with LRTI are characterized by one or more dominant pathogens present in high abundance^{50,83}. Using either RNA-Seq rpm alone (RNA-viruses) or the combination of RNA-Seq and DNA-Seq rpm (all others), this model identified the subset of microbes with the greatest relative abundance in each sample, which consisted of single microbes in cases of a dominant pathogen and also identified coinfections where several microbes were present within a similar range. All viruses detected by RNA-Seq at >0.1 rpm and present within the a priori-developed reference index of established respiratory pathogens were considered putative pathogens in the model. The remaining taxa (bacteria, fungi, and DNA viruses) were then aggregated at the genus level, assigned an abundance score based on $[\log(\text{RNA-Seq rpm}) + \log(\text{DNA-Seq rpm})]$, and sorted in descending order by this score. The greatest change in abundance score between sequentially ranked microbes was identified, and all genera with an abundance score greater than this threshold were then evaluated at the species level, by identifying the most abundant

species within each genus. If the species was present within the a priori-developed reference index of established respiratory pathogens, it was selected as a putative pathogen by the model (FIGURE 3.2).

Logistic regression model; LRM

This model employed the Python (version 3.6.1) sklearn (version 0.18.1) package to train on distinguishing between “pathogen” and “commensals” using the following five input features: log(RNA-Seq rpm), log(DNA-Seq rpm), per-patient RNA-Seq abundance rank, and two binary variables indicating whether the microbe could be identified in the established index of respiratory pathogens or was a virus. These features were selected in alignment with the observation that the pathogens identified in the LRTI^{+C+M} group were more abundant and within the top-ranked microbes. Moreover, the individual features were significantly different between the pathogens and commensals: (RNA-Seq rpm, $p = 2.44 \times 10^{-4}$; DNA-Seq rpm, $p = 3.55 \times 10^{-3}$; scoring rank, $p = 3.51 \times 10^{-6}$). Model performance was estimated in the derivation and validation cohorts and learning curves were computed. For identification of etiologic pathogens reported (FIGURE 3.3 and TABLE 3.4 and 3.5) the threshold of 0.36 was used for consistency between the LRM for pathogen identification and LRTI detection.

Pathogen versus Commensal LRM Performance Evaluation.

To evaluate LRM performance, we first performed 1000 rounds of cross-validation in which we randomly sub-divided the derivation cohort into training (70%) and test (30%) sets during each round, which yielded an average AUC of 0.93 +/- 0.08 standard

deviations. This assessed model variability as a function of the input training data. However, to obtain microbe predictions for all microbes in patients in the derivation cohort, while mitigating the potential for microbes within a single patient to disproportionately impact model performance, we performed leave-one-patient-out (LOPO) cross validation. In each round of LOPO-CV, all microbes from a single patient were left out, the model was trained on the microbes from all remaining patients, and prediction probabilities were calculated for the microbes in the left-out patient. This was repeated for all LRTI^{+C+M} and no-LRTI patients in the derivation cohort. Finally, the logistic regression model trained on microbes from patients in the derivation cohort (12 “pathogens” and 155 “commensals”) was applied to all microbes from validation cohort patients (26 “pathogens” and 174 “commensals”).

Learning Curves for Pathogen versus Commensal Model.

To evaluate the LRM performance as a function of derivation cohort size, learning curves were computed using randomized subsets of microbes from the derivation set ($n = 5, 10, 15 \dots 165$ total training microbes). The training and test mean square error (MSE) were computed along with the AUC for the test set at each iteration. This process was repeated over 25 rounds and the mean learning curve was computed (FIGURE 3.11B). The results indicate that the training set has saturated model performance, suggesting adequate sample size for the aforementioned analyses. We note that balanced classes may be of benefit but are unrealistic given the distribution of pathogens amongst the lung microbiome.

3.5.9 LRTI Prediction Based on Pathogen.

Outside of identifying putative LRTI pathogens, we evaluated whether LRM microbial score alone could be used to classify subjects as LRTI positive or LRTI negative. To do so, we used the top LRM-derived pathogen probability score per patient and evaluated the performance of this value alone to predict likelihood of infection in the LRTI^{+C+M} versus no-LRTI subjects.

3.5.10 Lung Microbiome Diversity Analysis.

The α diversity of the respiratory microbiome for each subject was assessed by SDI and Simpson diversity index at the genus level using NT rpm and the Vegan (version 2.4.4)⁵⁴ package in R (version 3.4.0)¹⁰¹. Richness (total number of genera) and genus-specific library sequence abundance (total number of microbial reads normalized per million reads sequenced) were also evaluated. Viral, bacterial, and fungal microbes were included in all diversity analyses, computed independently for RNA- and DNA-Seq samples without requiring that taxa be concordant on both nucleic acids. Diversity values were then compared between patients with clinically adjudicated LRTI (LRTI^{+C+M}) and those with respiratory failure due to noninfectious causes (no-LRTI) using the nonparametric Wilcoxon rank sum test. Evaluation of α diversity for prediction LRTI status was performed using the SDI value. The β diversity was evaluated using the Bray–Curtis dissimilarity metric calculated at the genus level using NT rpm and the Vegan package in R. Statistical significance of the β diversity between LRTI^{+C+M} and no-LRTI patients was assessed using PERMANOVA (999 permutations), and the results were visualized using nonmetric multidimensional scaling.

3.5.11 Host Gene Expression Analysis.

Following quality filtration with PRICESeqfilter⁵², RNA transcripts were aligned to the ENSEMBL CRCh38 human genome build using STAR. Subsequently, genes were filtered to include only protein-coding genes that were expressed in at least 50% of patients. All samples used for host transcriptome analysis (both derivation and validation sets) ultimately included more than 95,000 protein-coding genes with an average of 734,844 transcripts per patient.

Differential Expression Analysis.

Gene count data were analyzed using the Bioconductor package DESeq2 (version 1.16.1)¹⁰² in R statistical programming environment. To avoid batch-related confounding and class imbalance, we limited our differential expression analysis to the derivation cohort of 10 LRTI^{+C+M} and 10 no-LRTI samples, sequenced in the same batch. Differentially expressed genes with FDR <0.05 were used as input to ToppGene⁸⁶ to evaluate for functional pathway enrichment.

Differential Expression of Viral versus Bacterial LRTI.

Gene count data were analyzed using the Bioconductor package DESeq2 (v 1.16.1)¹⁰² in R statistical programming environment. To ensure adequate sample size, we extended differential expression analysis as a function of pathogen type to include all LRTI^{+C+M} patients (both derivation and validation cohorts) with known bacterial (n = 17) or viral (n = 3) infections. Cases of co-infection (n = 6) were left out of the analysis.

Differentially expressed genes with $FDR < 0.05$ were used as input to ToppGene⁸⁶ to evaluate for functional pathway enrichment.

In Silico Analysis of Cell Type Proportions.

Cell-type proportions were estimated from bulk host transcriptome data using the CIBERSORT algorithm implemented in R package EpiDISH version 0.1.1⁹⁴ and the LM22 reference dataset for distinguishing 22 human hematopoietic cell phenotypes. The cell types estimated with this reference cover all expected cell types in the TA sample, however the LM22 matrix was derived from microarray data. The estimated proportions were compared between LRTI^{+C+M} and no-LRTI patients within the derivation cohort using the Wilcoxon rank sum test.

3.5.12 Host Gene Expression Classifier for LRTI Prediction.

The derivation cohort was independently normalized using DESeq2 and log-transformed. The values for each gene in the derivation cohort were then scaled and centered by z score. A classifier was built using the elastic net regularized regression model implementation from the glmnet package (version 2.0.13) in the R Statistical Programming Language (version 3.4.0). Regularization parameter $\alpha = 0.5$ was selected using leave-one-out cross-validation and optimizing for AUC. To account for heterogeneity in the cohort, the model included covariates of concurrent bloodstream infection, immunosuppression, and gender. No significant difference was seen in these parameters between LRTI^{+C+M} and no-LRTI (TABLE 3.11). These covariates were reduced to zero in the model-fitting stage. Genes with nonzero weights were used for

classification. To obtain a single-value score for each patient, genes selected by the elastic net were evaluated for their correlation with each of the two groups. Genes for which the mean expression was greater in the LRTI^{+C+M} were assigned a weight of 1, and those with mean expression greater in no-LRTI were assigned a weight of -1. The normalized, scaled, expression values for each patient were multiplied by the weight vector and summed across all genes. The total sum was used as a representative score, and the AUC was calculated. Given the importance of sensitivity in the context of diagnostics, the threshold selected for analysis of the test cohort and combined metrics (scores, -4) was chosen as the threshold which provided 100% sensitivity in the derivation cohort. The host gene expression classifier was then validated on the validation set, and learning curves were used to estimate the reliability of the performance metrics.

Sample Size Calculations for Host Gene Expression Classifier

To estimate the sample size required to develop a binary classifier from high-dimensional data with performance within a tolerance of .05 of the best possible classifier, we employed a sample size calculator available from the National Cancer Institute which incorporates standardized log fold change (3.46), number of genes (11,918), and class prevalence (0.5)⁹⁸. To compute standardized fold change, the maximum absolute value log fold change value was obtained from DESeq2 for the 12 classifier genes (logFC = 3.07 for gene BLVRB). The within-class standard deviation for this gene (0.71) was computed and the suggested scaling factor of 0.8 was used. The number of genes (11,918) was based on the total number of genes that met QC thresholds for the classifier

analysis. At a tolerance of 0.05 the calculator indicated that the derivation cohort would require nine subjects in each group (LRTI^{+C+M} and no-LRTI).”

Validation of Host Gene Expression Classifier

To evaluate the performance of the classifier on the independent validation cohort (16 LRTI^{+C+M} and eight no-LRTI samples), genes from the validation cohort were independently normalized using DESeq2 and subsequently scaled and centered according to the scaling parameters derived from the derivation cohort. Then, the scaled counts were multiplied by the weights, values were summed, and AUC computed.

Learning Curves for Host Gene Expression Classifier

To assess the power of our host classifier given the limited derivation cohort size, learning curves were generated (FIGURE 3.11A). Learning curves are a widely used and robust approach to determine optimal training set sample size in machine learning analyses^{104,105}. A learning curve is computed by evaluating the performance of a model at varying training set sizes and relies on the observation that beyond a certain sample size threshold the performance of a model has diminishing improvements as a function of the size of training data.

For each learning curve, the derivation cohort was subsampled randomly at size $n = 10, 12, 14, 16, 18,$ and 20 patients. The training and test MSE were computed along with the AUC for the test set at each iteration. Finally, for each iteration, the genes identified by regularized regression were tallied. These mean squared errors and AUCs

were plotted as a function of training set size. After repeating this process 25 times, the mean learning curve was computed.

3.5.13 Classifier Combination.

To generate a readily interpretable compilation of host and microbial mNGS metrics that could enable a data-driven assessment of LRTI, the rule-out model was developed. In the rule-out model, we identified score thresholds from the pathogen and host metrics required to achieve 100% sensitivity in the derivation cohort (pathogen > 0.36, and host > -4) and applied these to the validation cohort to predict LRTI using the following combinatorial rule: LRTI = (Host) positive OR (Microbe) positive.

3.5.14 Identification and Mitigation of Environmental Contaminants.

To minimize inaccurate taxonomic assignments due to environmental contaminants, we processed negative water controls with each group of samples that underwent nucleic acid extraction, and included these, as well as positive control clinical samples, with each sequencing run. We directly subtracted alignments to those taxa in water control samples detected by both RNA-Seq and DNA-Seq analyses (TABLE 3.13) from the raw rpm values in all samples. To account for selective amplification bias of contaminants in water controls resulting from PCR amplification of metagenomic libraries to a fixed standard concentration across all samples, before direct subtraction we scaled taxa rpms in the water controls to the median percent microbial reads present across all samples (0.04%). In addition, we confirmed reproducibility of results by sequencing 10% of samples in triplicate and evaluated discrepancies between mNGS and standard

diagnostics in a random subset of LRTI⁺C patients using clinically validated 16S bacterial rRNA gene sequencing and/or viral PCR testing, as described above.

3.5.15 Estimation of Antibiotic Use Reduction

Days of therapy for each antibiotic administered to each subject in the no-LRTI group was tracked until date of ICU discharge or for up to seven days. Subjects in this group received empiric antibiotics because either: 1) a non-infectious etiology of respiratory failure was unapparent to the treating clinicians during the time of ICU admission but evident upon *post-hoc* adjudication based on review of the medical record, or 2) the patient had a non-pulmonary infection. Changes in total days of therapy per antibiotic were estimated based on theoretical use of mNGS *rule-out model* results 48 hours post-study enrollment to inform discontinuation of antibiotics empirically prescribed for LRTI. Standard of care prophylactic antibiotics for immunocompromised patients prescribed prior to admission and antibiotics prescribed for non-pulmonary infections were excluded from the analysis. The Wilcoxon rank sum test was used to determine the significance of the differential days of antibiotic therapy.

3.5.16 Data Availability.

Raw microbial sequences are available via the Sequence Read Archive (SRA) BioProject accession ID SRP139967. Host transcript counts are tabulated in TABLE 3.12. Scripts for the classification algorithms are available on GitHub at: <https://github.com/DeRisi-Lab/Host-MicrobeLRTI> .

3.6 FIGURES

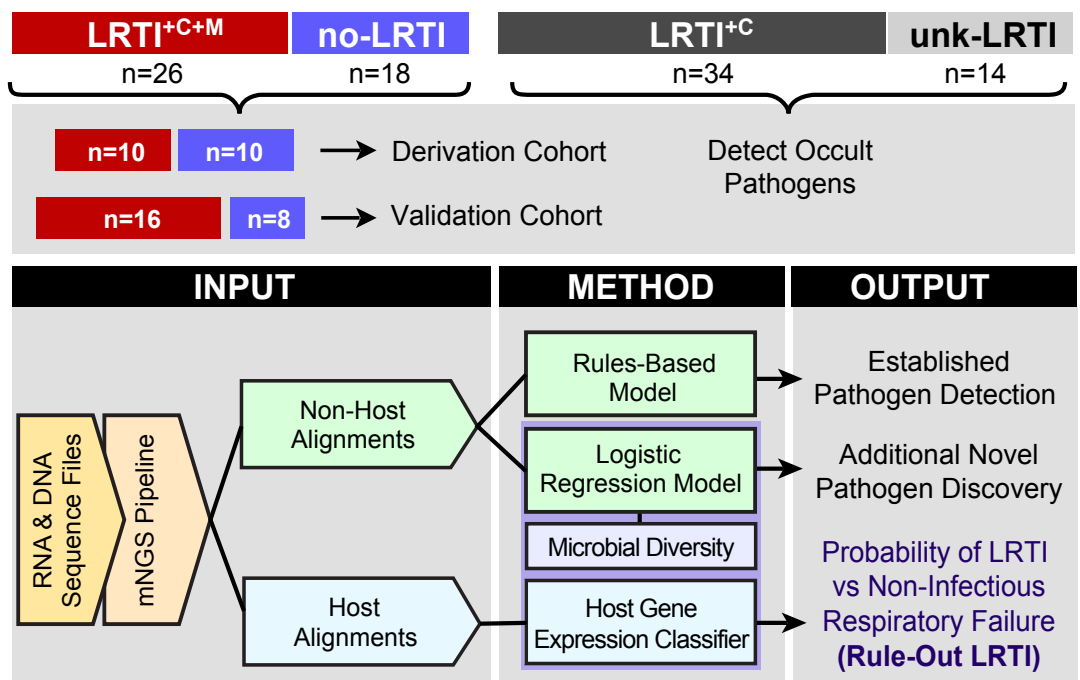


FIGURE 3.1 Study overview and novel analysis workflow

Patients with acute respiratory failure were enrolled within 72 hours of ICU admission and TA samples were collected and underwent both RNA sequencing (RNA-Seq) and shotgun DNA sequencing (DNA-Seq). Post-hoc clinical adjudication, blinded to metagenomic next-generation sequencing (mNGS) results, identified patients with lower respiratory tract infections (LRTIs) defined by clinical and microbiologic criteria (LRTI^{+C+M}); LRTI defined by clinical criteria only (LRTI^{+C}); patients with non-infectious reasons for acute respiratory failure (no-LRTI); and respiratory failure due to unknown cause (unk-LRTI). The LRTI^{+C+M} and no-LRTI groups were divided into derivation and validation cohorts. To detect pathogens and differentiate them from a background of commensal microbiota we developed two models - a rules-based model (RBM) and a logistic regression model (LRM). LRTI probability was next evaluated with i) a pathogen metric, ii) a lung microbiome diversity metric, and iii) a 12-gene host transcriptional classifier. Models were then combined and optimized for LRTI rule-out.

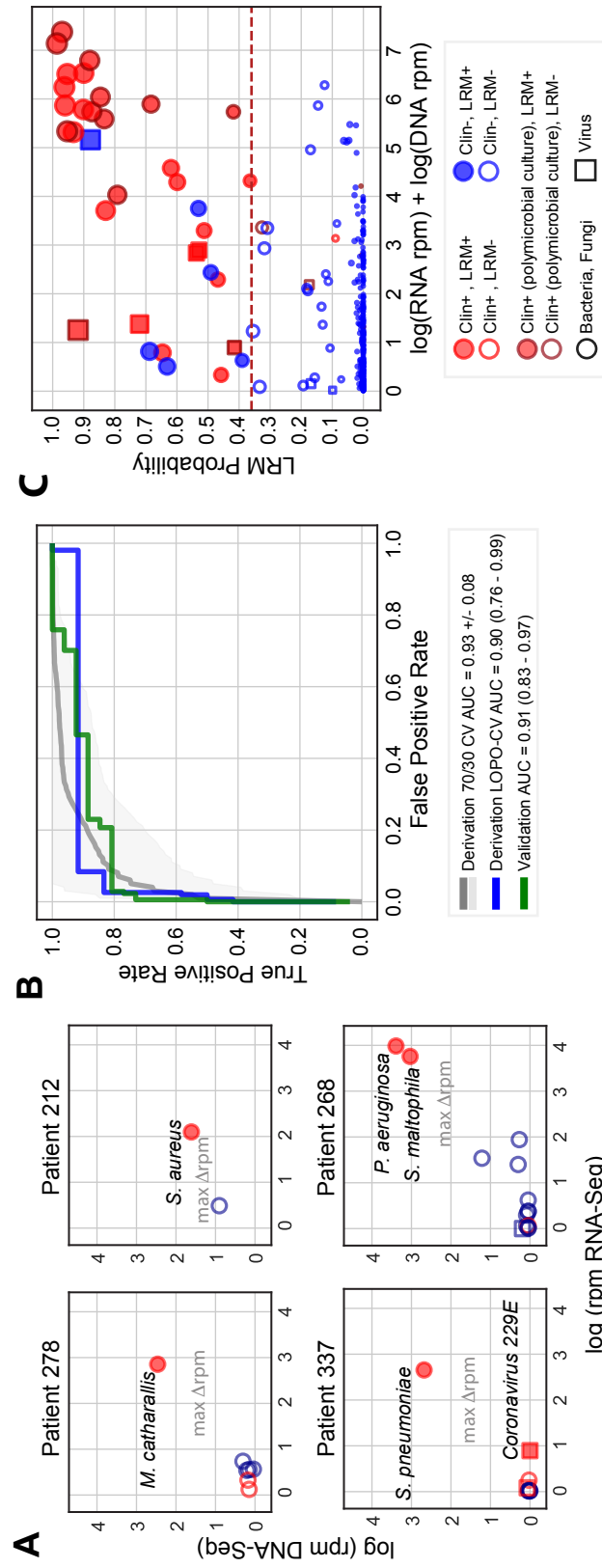


FIGURE 3.2 Distinguishing LRTI pathogens from commensal respiratory microbiota using an algorithmic approach

A) Projection of microbial relative abundance in log rpm (reads per million reads sequenced) by RNA-sequencing (RNA-Seq, X axis) versus DNA-sequencing (DNA-Seq, Y axis) for representative cases. In the LRTI^{+C+M} group, pathogens identified by standard clinical microbiology (filled shapes) had higher overall relative abundance as compared to other taxa detected by sequencing (open shapes). The largest score differential between ranked microbes (max Δ rpm) was used as a threshold to identify high-scoring taxa, distinct from the other microbes based on abundance (line with arrows). Red indicates taxa represented in the reference list of established LRTI pathogens. **B)** Receiver operator characteristic (ROC) curve demonstrating LRM performance for detecting pathogens versus commensal microbiota in both the derivation and validation cohorts. The grey ROC curve and shaded region indicate results from 1000 rounds of training and testing on randomized sets from the derivation cohort. The blue and green lines indicate predictions using leave-one-patient-out cross-validation (LOPO-CV) on the derivation and validation on the validation cohort, respectively. **C)** Microbes predicted by the LRM to represent putative pathogens. The X axis represents combined RNA-Seq and DNA-Seq relative abundance; the Y axis indicates pathogen probability. The dashed line reflects the optimized probability threshold for pathogen assignment.

Legend: *Red filled circles:* microbes predicted by LRM to represent putative LRTI pathogens that were also identified by conventional microbiologic tests. *Blue filled circles:* microbes predicted to represent putative LRTI pathogens by LRM only. *Blue open circles:* microbes identified by NGS but not predicted by the LRM to represent putative pathogens. *Red open circles:* microbes identified using NGS and by standard microbiologic testing but not predicted to be putative pathogens. *Dark red outlined circles:* microbes detected as part of a polymicrobial culture.

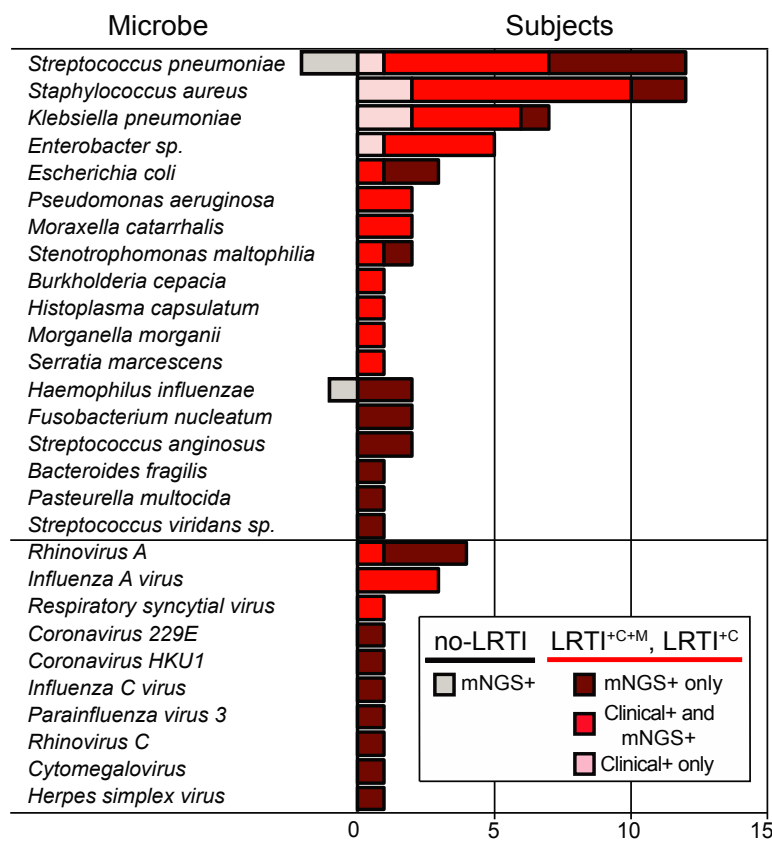


FIGURE 3.3 Distribution of respiratory pathogens identified in patients using clinician-ordered diagnostics versus mNGS

Number of subjects in whom each respiratory microbe was detected. All microbes detected by clinician-ordered diagnostics were detected by mNGS, however pink bars indicate microbes misclassified as negative by either the rules-based or logistic regression models. Notably, all microbes identified by clinician-ordered diagnostics and misclassified by either the rules-based or logistic regression models (pink bars) were found in polymicrobial cultures, highlighting the presence of dominant pathogens by NGS that are not captured in the polymicrobial culture results. Red bars indicate microbes detected by clinician-ordered diagnostics and also predicted as pathogens by either the rules based or logistic regression models. More detail on which model identified each microbe can be found in FIGURE 3.8. Dark red bars (LRTI^{+C+M} and LRTI^{+C} subjects) and grey bars (no-LRTI subjects) indicate number of cases with microbes detected only by mNGS.

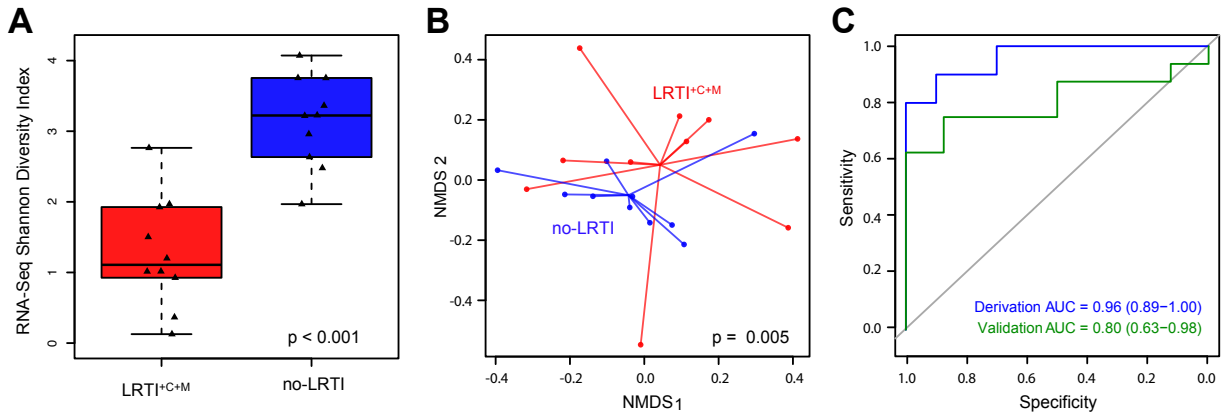


FIGURE 3.4 Diversity of the transcriptionally active lung microbiome in patients with LRTI versus non-infectious respiratory illnesses

A) Box plots show SDI of the lung microbiome assessed by RNA-Seq at the genus level for LRTI^{+C+M} and no-LRTI patients (in the derivation cohort). RNA-Seq SDI was found to be significantly different between LRTI^{+C+M} and no-LRTI patients ($p < 0.001$) **B)** Beta diversity assessed by PERMANOVA on Bray-Curtis dissimilarity values in the derivation cohort differed between LRTI^{+C+M} and no-LRTI patients ($p < 0.01$). **C)** ROC curve demonstrating performance of RNA-Seq SDI for distinguishing between LRTI^{+C+M} from no-LRTI groups (blue = derivation cohort, green = validation cohort). RNA-Seq SDI differentiated LRTI^{+C+M} from no-LRTI patients with an AUC of 0.96 (0.89 – 1.0) in the derivation cohort and 0.80 (0.63 – 0.98) in the validation cohort.

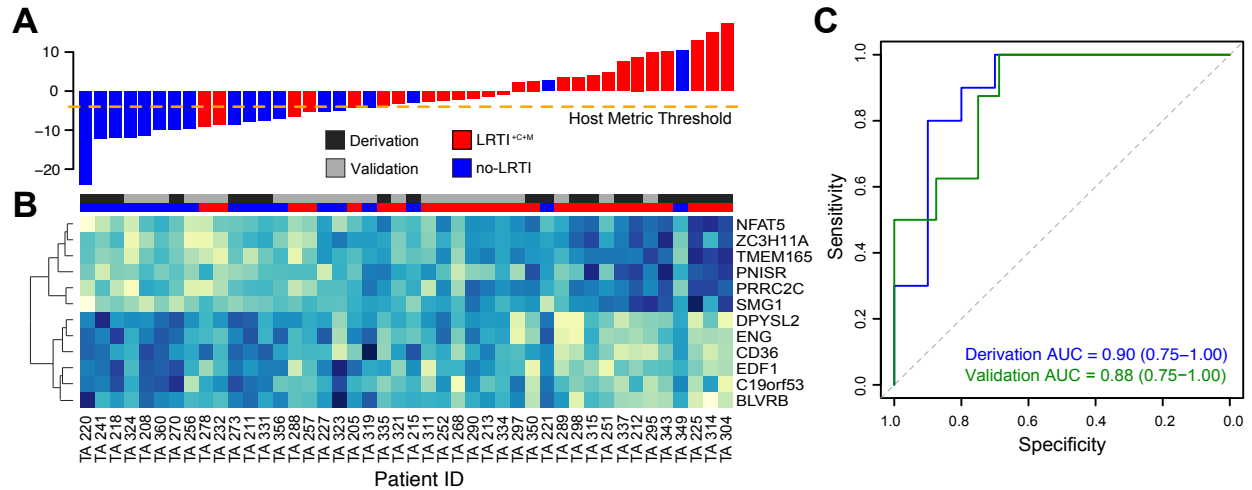


FIGURE 3.5 Host transcriptional profiling distinguishes patients with acute LRTI from those with non-infectious acute respiratory illness

A) Host classifier scores for all patients in the derivation and validation cohorts, each bar indicates a patient score and is colored as follows: LRTI^{+C+M} = red, no-LRTI = blue. Orange dotted line indicates the host classifier threshold (score = -4) that achieved 100% sensitivity in the training set and was used to classify the test set samples. **B)** Normalized expression levels, arranged by unsupervised hierarchical clustering, reflect over-expression (blue) or under-expression (turquoise) of classifier genes (rows) for each patient (columns). 12 genes were identified as predictive in the derivation cohort and subsequently applied to predict LRTI status in the validation cohort. Column colors above the heatmap indicate whether a patient belonged to the derivation cohort (dark grey) or validation cohort (light grey) and whether they were adjudicated to have LRTI^{+C+M} (red) or no-LRTI (blue). **C)** ROC curves demonstrating host classifier performance for derivation (blue) and validation (green) cohorts.

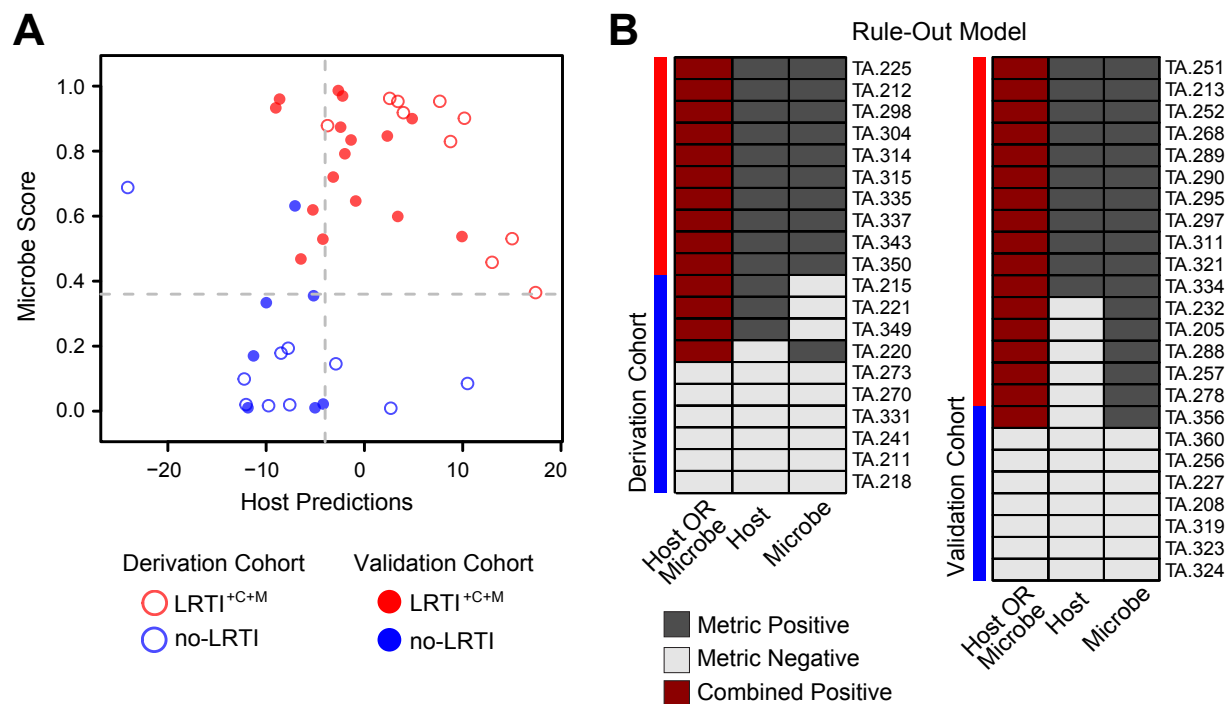


FIGURE 3.6 Combined LRTI prediction metric integrating pathogen detection and host gene expression

A) Scores per patient for each of the two components of this *LRTI rule-out model* are displayed in a scatterplot (X axis represents the host metric; Y axis represents the microbe score). The thresholds optimized for sensitivity in the derivation cohort are indicated in grey dashed line. Each point represents one patient – open circles represent those that were in the derivation cohort and solid circles represent those that were in the validation cohort. Red indicates LRTI^{+C+M} and blue indicates no-LRTI subjects. **B)** LRTI rule-out model results for each patient are shown for both the derivation and validation cohorts, with study subjects shown in rows and metrics in columns. Dark grey indicates a metric exceeded the optimized LRTI threshold, light grey indicates it did not. Dark red indicates the subject was positive for both pathogen + host metrics, and thus was classified as having LRTI.

3.7 TABLES

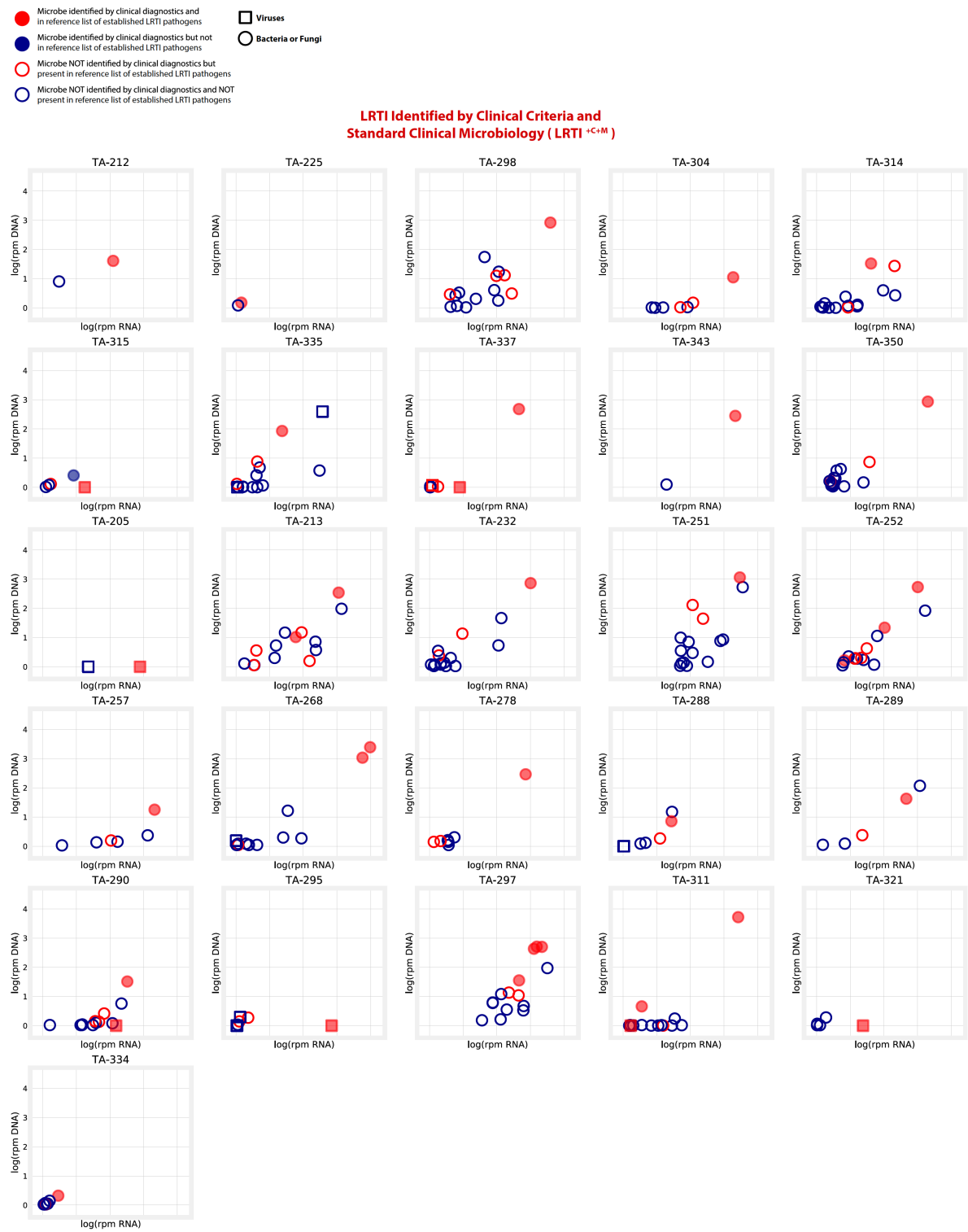
TABLE 3.1 Demographics and clinical characteristics of study cohort

Demographics and clinical characteristics of the study cohort containing patients with acute respiratory failure who were enrolled within 72 hours of ICU admission. **Legend:** LRTI^{+C+M} = subjects who met both clinical and microbiologic criteria for lower respiratory tract infection (LRTI). no-LRTI = subjects with a non-infectious etiology of acute respiratory failure. SIRS = systemic inflammatory response syndrome, defined as two or more abnormalities in white blood cell count (>12,000 cells/μL or <4,000 cells/μL), temperature (>38°C or < 36°C), heart rate (>90 beats per minute) or respiratory rate (> 20 breaths per minute). APACHEIII score predicts mortality and disease severity for critically ill patients . Pneumonia severity index score estimates mortality for adult patients with community-acquired pneumonia¹⁰³. COPD = chronic obstructive pulmonary disease, WBC = white blood cell. *Chi-squared test. *Wilcoxon rank sum test.

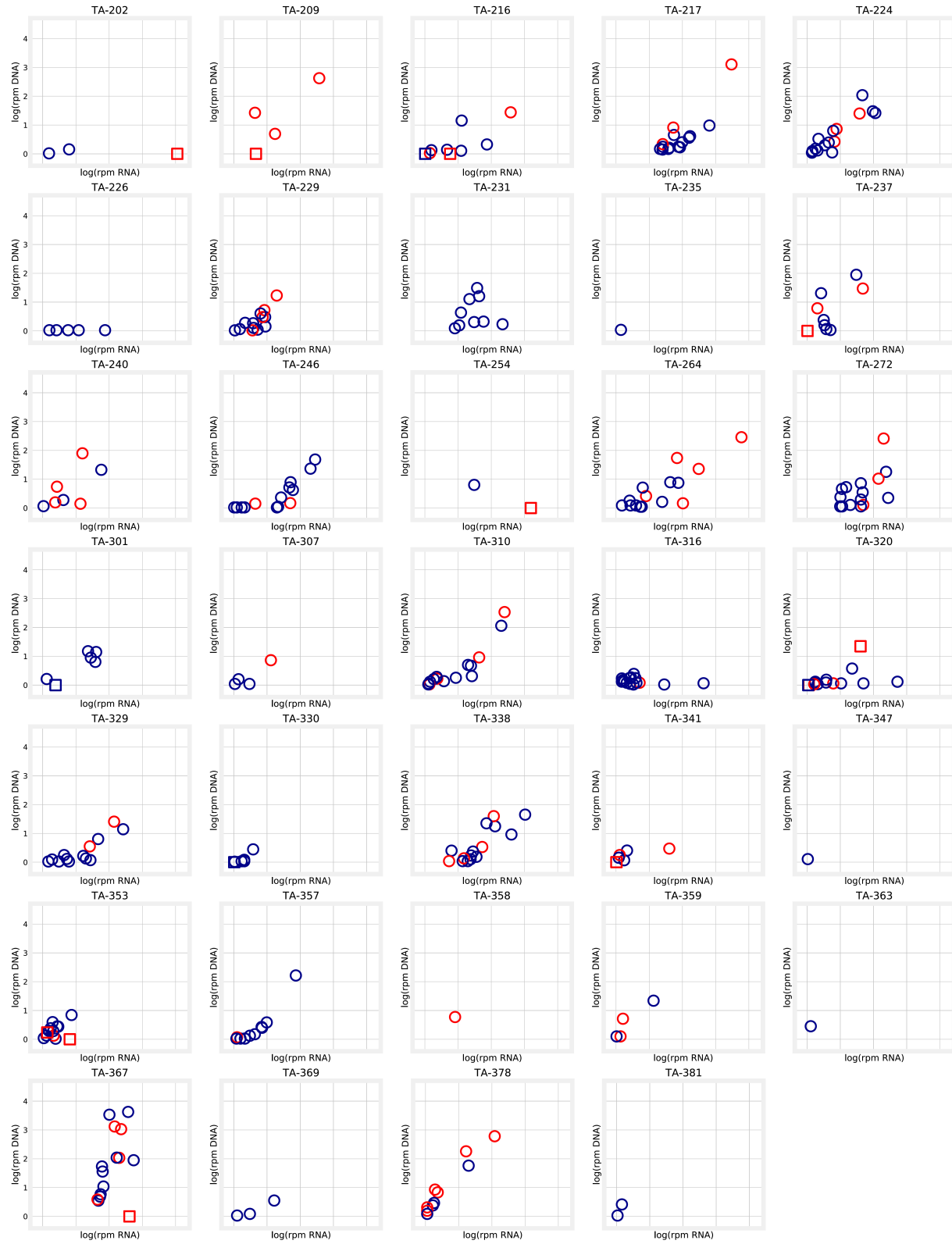
Patient Characteristics	Cohort Overall		LRTI ^{+C+M}	no-LRTI	p [*]
Total enrolled	92		26	18	-
Age, years (mean, range)	62	(21-85+)	61	63	0.80
Female gender	31	34%	6	9	0.13
Race:					
African American	5	5%	2	1	0.82
Asian	26	28%	7	5	
Caucasian	50	46%	15	9	
Other	11	12%	2	3	
Hispanic Ethnicity	8	9%	3	1	0.88
Comorbidities and Outcomes			LRTI ^{+C+M}	no-LRTI	p [*]
Bacteremia	21	23%	6	3	0.90
Non-pulmonary infections	29	32%	9	4	0.58
COPD	12	13%	3	0	0.37
Diabetes mellitus	6	7%	1	3	0.36
Congestive heart failure	7	8%	1	1	1.00
Current smoker	12	13%	5	1	0.39
Immune suppression	41	45%	10	9	0.65
Solid organ transplantation	13	14%	1	5	0.07
Prior antibiotic use	84	91%	22	18	0.23
Community acquired pneumonia	42	46%	18	-	-
Hospital acquired pneumonia	13	14%	5	-	-
Ventilator associated pneumonia	3	3%	3	-	-
30-day mortality	18	20%	6	1	0.25

Clinical Metrics		LRTI^{+C+M}	no-LRTI	p*
Max Temperature °C	37.8	38.1	38.0	0.33
Max WBC count (10 ⁶ cells/μL)	14.3	13.8	12.8	0.58
Max Heart Rate (bpm)	110	111	107	0.50
Max Respiratory Rate (breaths/min)	36	35	35	0.74
SIRS Criteria (mean)	3	3	3	0.54
APACHE III score (mean)	97	101	94	0.62
Pneumonia Severity Index (mean)	151	148	137	0.65

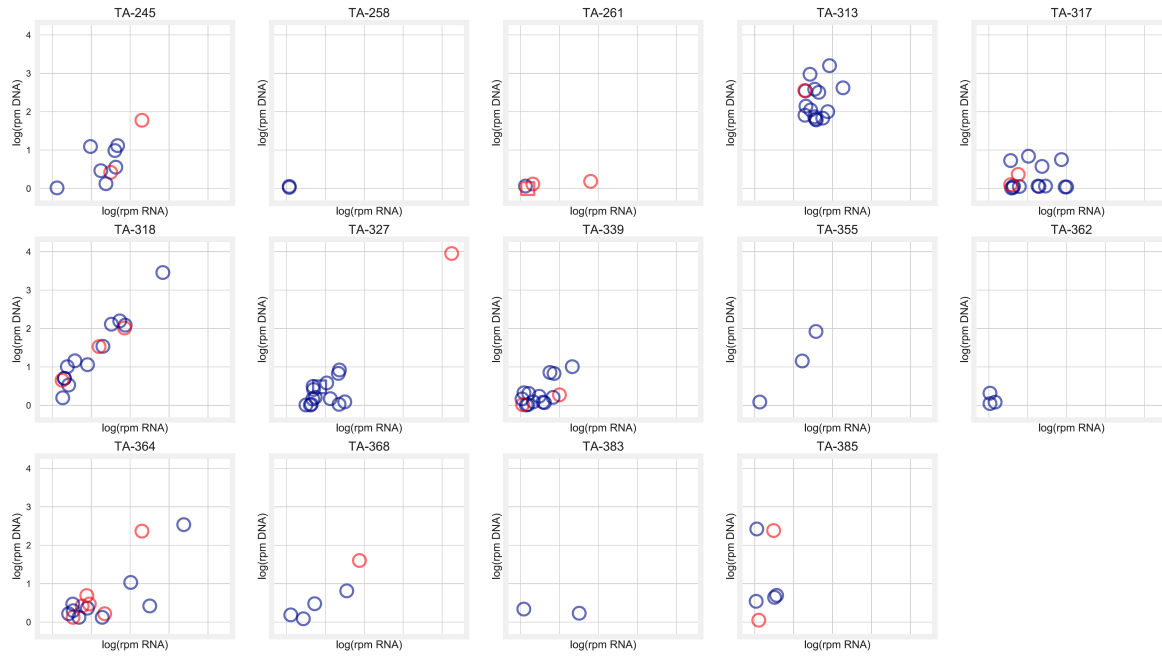
3.8 SUPPLEMENTAL FIGURES



LRTI Identified by Clinical Criteria (LRTI^{LC})



LRTI Status Unknown (unk-LRTI)



No LRTI (No-LRTI)



FIGURE 3.7 Distribution of mNGS-identified microbes per patient, by relative abundance

Microbes plotted by $\log(\text{RNA-Seq rpm})$ versus $\log(\text{DNA-Seq rpm})$ demonstrate the microbial community composition for each patient.

Legend: circles represent bacteria or fungi, squares represent viruses. Filled markers: microbes identified by conventional microbiologic tests. Red filled: microbes indexed in the reference established respiratory pathogens. Blue filled: microbes with uncertain respiratory pathogenicity, not present in the reference list. Open circles: microbes identified by mNGS, but not identified by conventional clinical microbiology.

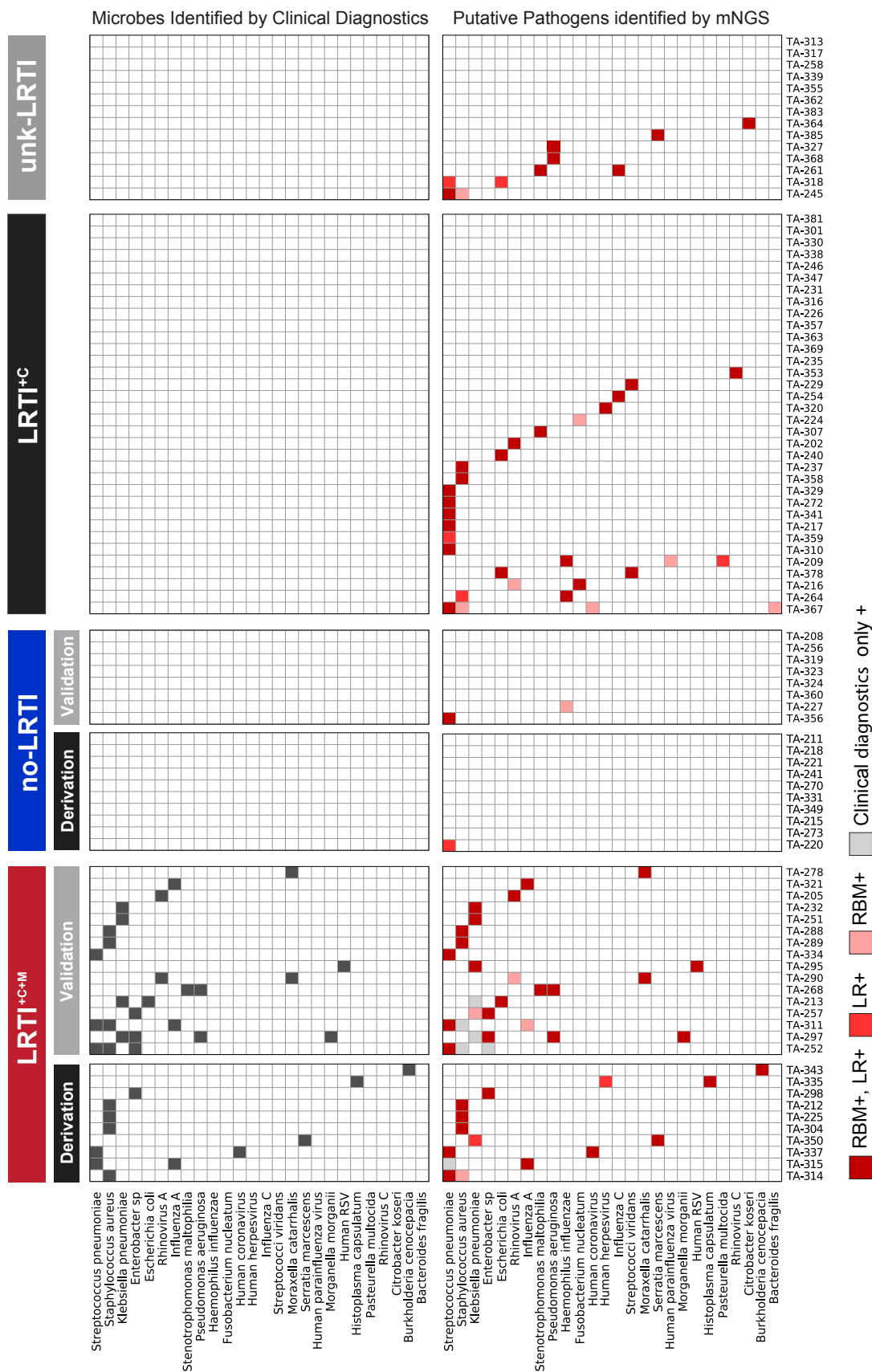


FIGURE 3.8 Microbial pathogens identified by clinician-ordered diagnostics, compared to those identified by mNGS RBM and LRM models

Microbial pathogens identified by standard clinical microbiologic diagnostics (upper panel) versus those identified by mNGS (lower panel). Patients are grouped by LRTI adjudication: 1) LRTI^{+C+M} = LRTI defined by both clinical and microbiologic criteria; 2) No-LRTI = no evidence of LRTI with a clear alternative explanation for acute respiratory failure; 3) LRTI^{+C} = LRTI defined by clinical criteria only with negative conventional diagnostic testing; 4) unk-LRTI = respiratory failure due to unknown cause. LRTI^{+C+M} and no-LRTI patient groups are further divided into derivation and validation cohorts. Microbes are depicted in rows, ordered by prevalence within the cohort, and patients in columns.

Legend: color shading indicates whether the microbe was identified by conventional diagnostics (gray, Clin+); the rules based model (light red RBM+), the logistic regression model (medium red, LRM+), both the rules based model and logistic regression models (dark red, RBM+, LRM+).

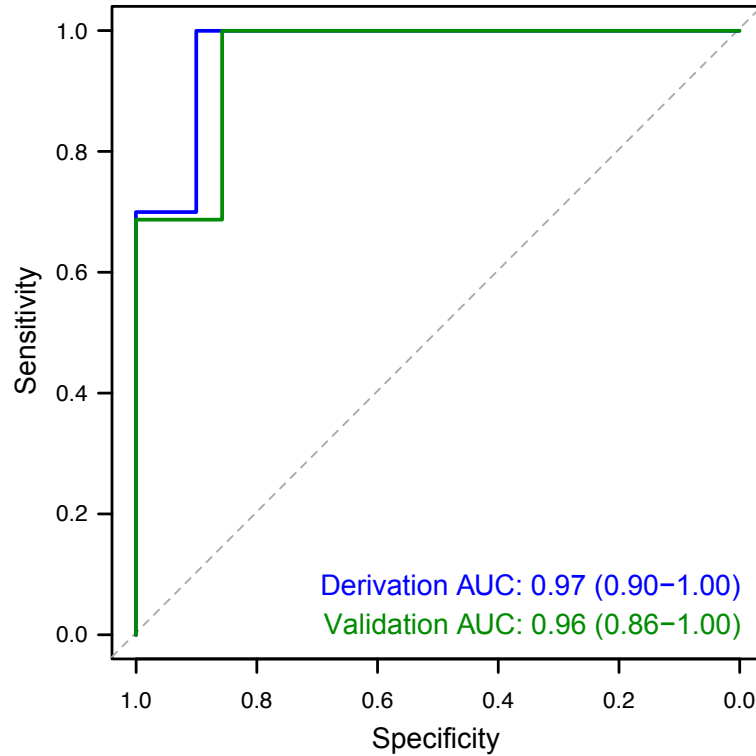


FIGURE 3.9 LRM probability score differentiates patients with LRTI from those with non-infectious causes of acute respiratory failure

The top microbe probability score per patient from the LRM was significantly higher in LRTI^{+C+M} subjects versus the no-LRTI subjects ($p = 3.8 \times 10^{-4}$ in the derivation cohort). This value predicted LRTI with an area under the receiver operator curve (AUC) of 0.97 (95% CI = 0.90 to 1.00) in the derivation cohort and AUC of 0.93 (95% CI = 0.82 to 1.00) in the validation cohort.

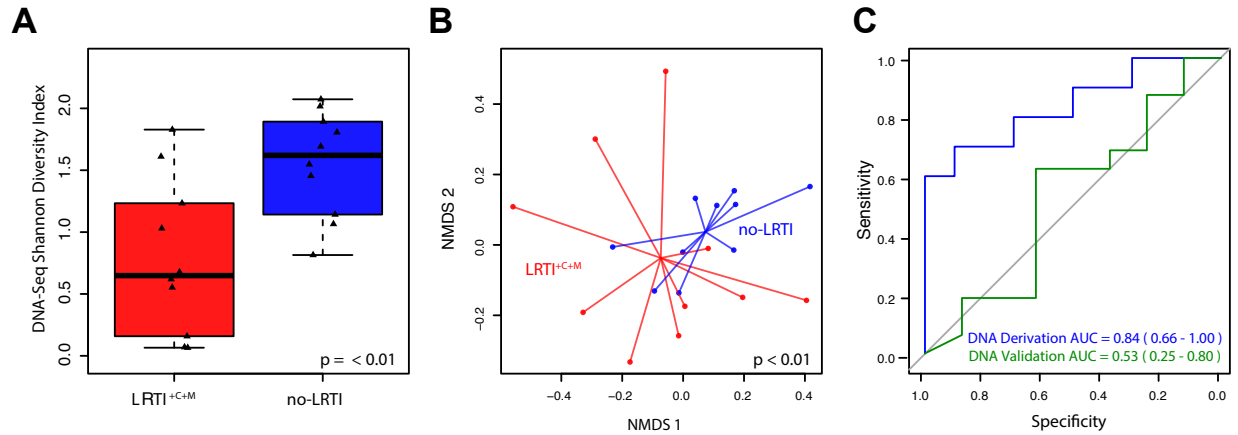


FIGURE 3.10 Performance DNA-Seq microbiome diversity for differentiating patients with LRTI from those with non-infectious causes of acute respiratory failure

A) Box plots show SDI of the lung microbiome assessed by DNA-Seq at the genus level for LRTI^{+C+M} and no-LRTI patients (in the derivation cohort). DNA-Seq SDI was found to be significantly different between LRTI^{+C+M} and no-LRTI patients ($p = < 0.01$) **B)** Beta diversity assessed by PERMANOVA on Bray-Curtis dissimilarity values in the derivation cohort differed between LRTI^{+C+M} and no-LRTI patients ($p < 0.01$). **C)** ROC curve demonstrating performance of DNA-Seq SDI for distinguishing between LRTI^{+C+M} from no-LRTI groups (blue = derivation cohort, green = validation cohort). DNA-Seq SDI differentiated LRTI^{+C+M} from no-LRTI patients with an AUC of 0.84 (0.66 – 1.0) in the derivation cohort and 0.53 (0.25 - 0.80) in the validation cohort.

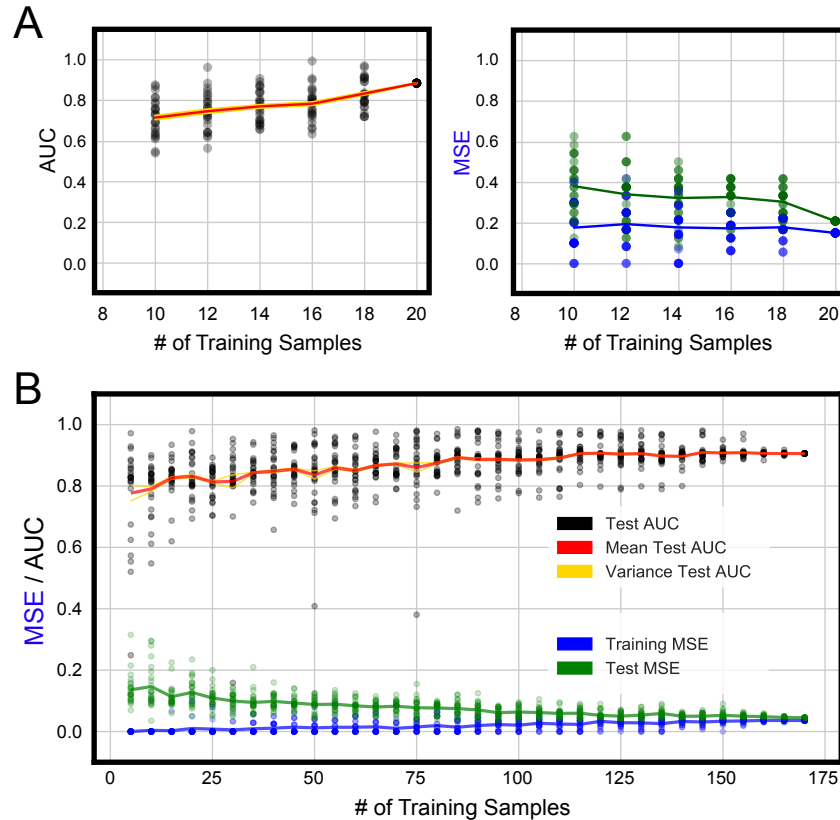


FIGURE 3.11 Learning Curve analyses for pathogen vs commensal and host gene expression classifier models

A) Learning curve analyses of the host gene expression classifier model indicated that $n=20$ samples in the derivation cohort approached model saturation. **B)** Learning curve analyses for the pathogen versus commensal LRM demonstrated convergence of the derivation cohort (blue) and validation cohort (green) mean squared error (MSE) for each of 25 iterations, with increasing training set size. The mean MSE is plotted as a solid line. The validation cohort AUC is overlaid, with individual values plotted in black, the mean plotted as a red line, and variance shown in gold. The AUC increased with increasing training size, but plateaued at training size of $n = 125$, indicating adequate sample size.

3.9 SUPPLEMENTAL TABLES

TABLE 3.2 Expanded clinical and microbiologic data

Expanded clinical and microbiologic data including: age (years), race, gender, temperature (°C), maximum heart rate, maximum respiratory rate, maximum white blood cell count (cells/uL), number of systemic inflammatory response (SIRS) criteria met, bacteremia, concurrent non-pulmonary infection, immune suppression, clinically adjudicated diagnosis, standard of care microbiologic testing and antimicrobial use.

The full dataset can be downloaded as part of Dataset S1 here:

<https://www.pnas.org/content/115/52/E12353/tab-figures-data>

TABLE 3.3 United States CDC/NHSN surveillance definition of pneumonia

United States Centers for Disease Control/National Healthcare Surveillance Network surveillance definition of pneumonia used for adjudication in this study.

US CDC/NHSN Surveillance Definition of Pneumonia
1. Clinical Criteria:
<ul style="list-style-type: none">- More than 2 serial chest imaging results with new, progressive or persistent infiltrate, consolidation, or cavitation AND at least one of the following: <ul style="list-style-type: none">- Fever ($>38.0^{\circ}\text{C}$ or $>100.4^{\circ}\text{F}$)- Leukopenia (≤ 4000 WBC/mm³) or leukocytosis ($>12,000$ WBC/mm³)- For adults >70 years old, altered mental status with no other recognized cause AND at least two of the following: <ul style="list-style-type: none">- New purulence or change in sputum character, increased secretions or suctioning requirements- New onset/worsening cough, dyspnea, tachypnea- Rales or bronchial breath sounds- Worsening gas exchange, increased oxygen requirements, or increased ventilation demands
2. Microbiologic criteria:
<ul style="list-style-type: none">- Positive blood culture unrelated to another source- Positive lower respiratory or pleural fluid culture- Positive viral PCR- Positive antibody or antigen testing

TABLE 3.4 Reference index of established respiratory pathogens

Reference index of established respiratory pathogens derived from epidemiologic surveillance studies, clinical guidelines from the Infectious Diseases Society of America and American Thoracic Society and systematic reviews^{21,76,77,106,107}.

Reference Index of Established Lower Respiratory Tract Pathogens	
Bacteria and Fungi	Ref
Acinetobacter baumannii	1
Bacteroides fragilis	2,3
Bordatella pertussis	3
Burkholderia cepacia	3
Burkholderia pseudomallei	3
Chlamydomphila pneumoniae	2
Chlamydomphila psittaci	2
Citrobacter freundii	1,2
Citrobacter koseri	1,2
Coxiella burnetii	3
Enterobacter aerogenes	1,2
Enterobacter cloacae	1,2
Escherichia coli	1,2
Francisella tularemia	3
Fusobacterium necrophorum	2,3
Fusobacterium nucleatum	2,3
Haemophilus influenzae	2,3
Klebsiella oxytoca	1
Klebsiella pneumoniae	1
Legionella pneumophila	2,3
Moraxella catarrhalis	3
Morganella morganii	2
Mycobacterium tuberculosis	2,3
Mycoplasma pneumoniae	2,3
Nocardia spp.	5
Pasteurella multocida	3
Proteus mirabilis	1
Pseudomonas aeruginosa	1-4
Serratia marcescens	1,2
Stenotrophomonas maltophilia	1
Streptococcus pneumoniae	1-4
Streptococcus pyogenes	2
"Streptococcus viridans group (S. anginosus, S. intermedius, S. constellatus)"	2,3
Staphylococcus aureus	1-4
Viruses	
Cytomegalovirus	5
Human adenovirus	2,3
Human coronavirus	2,3
Human metapneumovirus	2,3
Human parainfluenza virus	2,3
Human respiratory syncytial virus	2,3
Human rhinovirus	2,3
"Influenza virus (Influenza A, Influenza B, Influenza C)"	2,3
Respiratory syncytial virus	2,3

Table continued below.

Reference Index of Established Lower Respiratory Tract Pathogens	
Fungi	Ref
Aspergillus flavus	5
Aspergillus fumigatus	5
Aspergillus niger	5
Aspergillus terreus	5
Cryptococcus gatii	5
Cryptococcus neoformans	5
Coccidioides immitis	2,3
Coccidioides posadisii	2,3
Histoplasma capsulatum	2,3
Blastomyces dermatitidis	2,3
Pneumocystis jirovecii	5
Rhizopus, Mucor and Rhizomucor spp.	5
References	
1. N Engl J Med 2014;370:1198-208.	
2. N Engl J Med. 2015 Jul 30;373(5):415-27.	
3. Clin Infect Dis (2007) 44 (Supplement 2): S27-S72.	
4. N Engl J Med. 2008 Nov 27;359(22):2355-55.	
5. Am J Transplant. 2017 Apr;17(4):855-879.	

TABLE 3.5 Microbial pathogens identified by clinician-ordered diagnostics or predicted by pathogen identification model

Microbial pathogens detected by clinician-ordered microbiologic diagnostics (Clin+) or predicted using the rules-based model (RBM+) and/or the logistic regression model (LRM+). For each microbe listed, the values of the LRM features (RNA-Seq rpm, DNA-Seq rpm, rank by RNA-Seq rpm, established LRTI pathogen (yes/no), and virus (yes/no)) are listed. *Sample orthogonally tested by viral PCR to validate metagenomic next-generation sequencing (mNGS)-identified virus. *Sample orthogonally tested by 16S rRNA gene sequencing to confirm one or more bacterial results.

The full dataset can be downloaded as part of Dataset S3 here:

<https://www.pnas.org/content/115/52/E12353/tab-figures-data>

TABLE 3.6 LRM feature weights

LRM features and associated weights determined by machine learning in the derivation cohort.

LRM Feature	LRM Weight
RNA-Seq rpm	-0.546
DNA-Seq rpm	1.558
is_pathogen	4.609
is_virus	1.346
ranks	-1.031

TABLE 3.7 Microbes identified in no-LRTI patients

Microbes identified in no-LRTI patients. The top 10 most prevalent genera concordant by DNA- and RNA-Seq across all no-LRTI patients are listed alongside the relative distribution of species for each.

Genus, species	# of Patients
<i>Streptococcus</i>	12
<i>S. pneumoniae</i>	3
<i>S. salivarius</i>	3
<i>S. thermophilus</i>	2
<i>S. constellatus</i>	1
<i>S. oralis</i>	1
<i>S. sp. I-P16</i>	1
<i>S. sp. VT</i>	1
<i>Veillonella</i>	8
<i>V. parvula</i>	8
<i>Lactobacillus</i>	7
<i>L. rhamnosus</i>	2
<i>L. crispatus</i>	1
<i>L. fermentum</i>	1
<i>L. mucosae</i>	1
<i>L. sakei</i>	1
<i>L. sp. wkB8</i>	1
<i>Prevotella</i>	7
<i>P. melaninogenica</i>	7
<i>Neisseria</i>	6
<i>N. meningitidis</i>	3
<i>N. elongata</i>	1
<i>N. gonorrhoeae</i>	1
uncultured <i>Neisseria</i> sp.	1
<i>Campylobacter</i>	5
<i>C. concisus</i>	4
<i>C. insulaenigrae</i>	1
<i>Fusobacterium</i>	5
<i>F. nucleatum</i>	5
<i>Haemophilus</i>	5
<i>H. parainfluenzae</i>	4
<i>H. influenzae</i>	1
<i>Candida</i>	5
<i>C. albicans</i>	3
<i>C. glabrata</i>	2
<i>C. dubliniensis</i>	1
<i>Rothia</i>	5
<i>R. mucilaginosa</i>	5

TABLE 3.8 Diversity metrics assessed for patients with LRTI versus those with non-infectious causes of acute respiratory failure

Diversity metrics of the transcriptionally active and total fractions of the lung microbiome assessed by **A)** RNA-Seq and by **B)** shotgun DNA sequencing, respectively. Wilcoxon rank sum statistical significance and area under the receiver operator curve (AUC) are calculated on both RNA-Seq and DNA-Seq datasets for each of the following metrics: Simpsons diversity index, Shannon diversity index, richness (number of genera), microbial sequence abundance (total microbial alignments by genus normalized per million reads sequenced).

A)

Diversity Metric	RNA-Seq	
Simpsons	p = 1.3e-04	
	Derivation AUC = 0.96 (95% CI: 0.89 - 1.00)	Validation AUC = 0.77 (95% CI: 0.57 - 0.97)
Shannon	p = 1.3e-04	
	Derivation AUC = 0.96 (95% CI: 0.89 - 1.00)	Validation AUC = 0.80 (95% CI: 0.63 - 0.98)
Richness	p = 0.029	
	Derivation AUC = 0.79 (95% CI: 0.56 - 1.00)	Validation AUC = 0.83 (95% CI: 0.66 - 0.99)
Microbial Sequence Abundance	p = 0.353	
	Derivation AUC = 0.63 (95% CI: 0.36 - 0.89)	Validation AUC = 0.91 (95% CI: 0.80 - 1.00)
Bray-Curtis Permanova	p = 0.005	

B)

Diversity Metric	DNA-Seq	
Simpsons	p = 2.88e-03	
	Derivation AUC = 0.88 (95% CI: 0.73 - 1)	Validation AUC = 0.51 (95% CI: 0.24 - 0.78)
Shannon	p = 8.93e-03	
	Derivation AUC = 0.84 (95% CI: 0.66 - 1)	Validation AUC = 0.52 (95% CI: 0.25 - 0.80)
Richness	p = 0.130	
	Derivation AUC = 0.71 (95% CI: 0.44 - 0.97)	Validation AUC = 0.61 (95% CI: 0.34 - 0.88)
Microbial Sequence Abundance	p = 0.190	
	Derivation AUC = 0.68 (95% CI: 0.42 - 0.94)	Validation AUC = 0.66 (95% CI: 0.43 - 0.90)
Bray-Curtis Permanova	p = 0.009	

TABLE 3.9 Differentially expressed genes between LRTI^{+C+M} and no-LRTI patients

A) Differentially expressed genes in the derivation cohort with an adjusted P value of < 0.05 between LRTI^{+C+M} and no-LRTI patients. **B)** Gene Ontology (GO) Biological Processes with significant enrichment in either LRTI^{+C+M} or no-LRTI subjects.

The full dataset can be downloaded as part of Dataset S6 here:

<https://www.pnas.org/content/115/52/E12353/tab-figures-data>

TABLE 3.10 LRTI host transcriptional classifier specifics

12 genes were identified as highly predictive for differentiating LRTI^{+C+M} and no-LRTI subjects in the derivation cohort, and these were subsequently applied to the validation cohort.

Ensembl Gene ID	HGNC Gene ID	Weight	Selected in Every Round of Training / Test Split
ENSG00000117523	PRRC2C	1	X
ENSG00000058673	ZC3H11A	1	
ENSG00000134851	TMEM165	1	
ENSG00000132424	PNISR	1	
ENSG00000135218	CD36	-1	X
ENSG00000092964	DPYSL2	-1	
ENSG00000106991	ENG	-1	
ENSG00000107223	EDF1	-1	X
ENSG00000157106	SMG1	1	
ENSG00000102908	NFAT5	1	
ENSG00000104979	C19orf53	-1	
ENSG00000090013	BLVRB	-1	

TABLE 3.11 Covariates controlled for in the host gene expression classifier

Covariates for immune suppression, concurrent non-pulmonary infection, antibiotic use, age, and gender were not significantly different between LRTI^{+C+M} and no-LRTI patients. For the covariate *age*, the inter-quartile range is provided.

Covariate	Training Data			All Data		
	LRTI ^{+C+M}	LRTI-NEG	p-val	LRTI ^{+C+M}	LRTI-NEG	p-val
Bloodstream Infection	20%	20%	1	25%	13%	1
Any Immune Suppression	40%	40%	1	38%	63%	0.47
Antibiotics Prior	90%	100%	1	81%	100%	0.51
Age	65.5 (60.5 - 71.8)	62.5 (54 - 67.75)	0.34	59.0 (45.0 - 72.25)	66.0 (60.3 - 76.0)	0.91
Gender (Male)	60%	40%	0.65	88%	50%	0.13

TABLE 3.12 Estimated cell type proportions

CIBERSORT⁹⁴ was utilized to predict cell type proportions for each patient. M2 Macrophages were identified to have significant differences in estimated proportions (Wilcoxon rank sum $p = 0.03$).

The full dataset can be downloaded as part of Dataset S7C here:

<https://www.pnas.org/content/115/52/E12353/tab-figures-data>

TABLE 3.13 Most abundant genera in water controls

To mitigate the impact of ubiquitous environmental contaminants, no-template water controls were sequenced alongside each batch of samples that underwent nucleic acid extraction (n = 5 for DNA-Seq, n = 6 for RNA-Seq). The 10 most abundant genera, concordant across both RNA- and DNA-Seq water controls are listed.

Genus	NCBI TaxID
<i>Propionibacterium</i>	1743
<i>Pseudomonas</i>	286
<i>Xanthomonas</i>	338
<i>Rhodococcus</i>	1827
<i>Cupriavidus</i>	106589
<i>Pseudoxanthomonas</i>	83618
<i>Caulobacter</i>	75
<i>Burkholderia</i>	32008
<i>Escherichia</i>	561
<i>Variovorax</i>	34072

TABLE 3.14 Human Transcriptome Counts

Gene counts obtained using alignment against the ENSEMBL GRCh38 human genome build are listed. Genes and associated ENSEMBL ID are listed in rows and subjects are grouped by columns.

The full dataset can be downloaded as part of Dataset S9 here:

<https://www.pnas.org/content/115/52/E12353/tab-figures-data>

4 APPLICATION NOTES

Chapter 3 developed methods to address outstanding challenges with interpretation of mNGS data for use as a clinical diagnostic for LRTIs. First, we developed algorithms for distinguishing pathogens from the background of commensal microbiota. Then, we evaluated pathogen, host, and microbiome-based metrics for distinguishing patients with infections from those without. While LRTIs are unique in some ways, the disruption of the relationship between pathogens, host factors, and the microbiome, is increasingly viewed as the hallmark of a broad range of infections. Thus, we hypothesized that these algorithms would have utility in a broader range of settings. Through this *Application Notes* section, we seek to demonstrate the broad utility of the approaches developed in Chapter 3. Additionally, we discuss two analytical extensions that may provide additional detail on the host and microbial condition - first for determining the presence of donor cells after transplant, and second for identifying antimicrobial resistance profiles associated with microbes. We expect that these analyses will further advance the potential for precision diagnostics.

4.1 EARLY INTERROGATION OF LRTI THROUGH mNGS

4.1.1 *Application Goal*

Previous studies laid the foundation for methods derived in the preceding two chapters. Here, we outline the findings from a pilot study investigating the utility of mNGS for enhanced diagnostic capability in LRTIs. Highlighting lessons learned serves to

demonstrate key elements of preceding studies that led to the development of these algorithms.

4.1.2 Introduction

Hematopoietic stem cell transplantation treats hematologic malignancies. Donor cells establish marrow and immune functions in a recipient whose malignant immune cells have been ablated¹⁰⁸. But, ablation of the recipient's immune system before transplantation¹⁰⁹ leaves post-HCT patients at higher risk for LRTIs from a diverse array of pathogens, including opportunistic pathogens^{110,111}. HCT recipients thus represent an important cohort for evaluation of novel diagnostics. This study provides an initial evaluation of mNGS for identifying pathogens from adult HCT recipients, with preliminary consideration of host and microbiome metrics as biomarkers for infection.

4.1.3 Methods

Sample Collection and Processing

Bronchial alveolar lavage (BAL) samples were collected from 22 adult HCT recipients hospitalized for acute respiratory illness between January 25, 2012 and May 20, 2013, under University of Michigan protocol HUM00043287. Both RNA and DNA were extracted from each sample and sequenced. Notably, RNA was processed using the Ovation RNAseq system (Nugen). Raw sequencing data was processed using the previously described pipeline²⁸.

Microbial Analysis

The existing state-of-the-art method to evaluate the significance of microbial sequences with respect to no-template water controls was used to generate z-scores for each microbe²⁸. The z-score was then evaluated against known positives (pathogens identified by standard clinical microbiology) to generate a ranking score optimized for concordance with known positives. Via the resulting scoring scheme, microbes were scored by nucleotide reads aligned per million reads sequenced (rpm) multiplied by the sum of the nt and nr Z-score [score = rpm_{nt} × (Z_{nt} + Z_{nr})]⁵⁰. Samples were grouped into the following three categories with respect to their clinical and mNGS results 1) confirmed pathogens 2) new potential pathogens and 3) unlikely or uncertain pathogens as described in Langelier and Zinter *et al*⁵⁰.

Diversity Calculations

The Simpsons diversity index was used to assess alpha diversity of microbial taxa present in BAL fluid from all subjects. The Simpsons diversity index formula is $D = 1 - ([\sum n(n-1)] / [N(N-1)])$, where n= the total reads for each single organism and N = the total number of reads for all organisms combined¹¹². Genus rpm were used for the calculations.

Host Gene Expression Analysis

On the raw sequencing data, we applied quality filtration and human genome alignment using STAR as described above¹¹³. The average percent of reads uniquely mapped to the human transcriptome was 2.26% (mean 550,178 human reads per

sample). Subjects with fewer than 250 unique genes indexed in the HGNC database and with ENSEMBL gene biotype of protein coding (patients 19, 24, 34, and 37) were excluded from analysis. Read count normalization was performed by cumulative sum scaling¹¹⁴ and genes expressed in fewer than 30% of samples or as outliers in only 10% of samples were removed from analysis. This normalization method was used to provide more stable estimates of expression at low read counts in the context of zero inflated data.

Pathways related to immune function, with over 70% of their genes expressed by at least 10% of the samples, were selected from the molecular signatures database¹¹⁵. The sum of CSS normalized, scaled expression values for all genes in the pathway was taken for each sample, summarized in TABLE 4.1. Raw gene counts are provided in TABLE 4.2. Pathway expression values were compared between groups (subjects with established LRTI pathogens, potential new mNGS-identified pathogens, microbes of unclear significance and pathogen negative groups) using a nonparametric Wilcoxon rank sum test.

4.1.4 Results

The pathogen scoring method above identified all seven of the pathogens identified by standard-of-care clinical diagnostics (100%), as well as pathogens in six patients with previously negative testing. To put the mNGS microbiologic findings in context, this study tested the significance of community diversity as a biomarker for infection. Previous studies profiling 16S ribosomal RNA have noted reduced diversity in the context of infection⁵⁵. Here, the patients with confirmed pathogens had significantly

reduced Simpson's diversity as compared to the patients with unlikely or uncertain pathogens (Wilcox rank sum $p = 0.017$, FIGURE 4.1). This supports the hypothesis that alpha diversity is reduced in the presence of an infection, even via mNGS.

To further contextualize microbial findings with transcriptomic signatures of host immune response, we evaluated an *a priori* selected gene set related to innate and adaptive immune responses from the Molecular Signatures Database¹¹⁵. We found significantly increased expression in patients with confirmed LRTI pathogens versus those without, as shown in FIGURE 4.2 (median, 94.9 [IQR, 93.8–105.6] vs. 33.1 [IQR, 20.7–75.1], Wilcoxon rank sum $p = 0.022$).

4.1.5 Discussion and Conclusion

Three major limitations with this study serve as the foundation for future investigation. First, this cohort was small, limiting the ability to draw robust statistical conclusions. This was especially relevant after sub-dividing the cohort into groups by infection likelihood. Chapter 3 addressed this issue by recruiting a larger cohort of 92 patients. Yet, more patients may provide added benefit.

Secondly, the dataset contained a high prevalence of zeros for transcriptome features. We attribute this to low sequencing depth and the use of the Nugen kit for RNA processing. Zero inflation coupled with the small cohort size, led to reduced statistical power for identifying significant differences on a per-gene level using standard differential expression and machine learning methods. To circumvent these challenges, we used models developed for zero-inflated data and only evaluated a multi-gene metric. To more adequately describe how to leverage host response and improve diagnostic capacity,

future studies would require larger cohorts and enhanced transcriptome coverage. Chapter 3 improved statistical power for host gene expression metrics by using a different library preparation method and greater sequencing depth.

Finally, the study was limited by the lack of subjects with exclusively clinical diagnoses of noninfectious airway disease. Studies benchmarking diagnostic performance need more robust negative control groups. Chapter 3 addressed this limitation by including a no-LRTI patient group. This group consisted of patients without suspected infection who had clear alternative explanations for respiratory failure.

Despite several limitations, this study provided two key take-aways. First, it indicated that the presence of the background microbiota in respiratory samples challenged the previously described z-score method²⁸. Thus, a new ranking score was optimized for concordance with known positives. But, without comparison to true negative cases, the scoring metric had limited applicability. Work with cohorts with a larger number of negative controls (Chapter 3) was required to tackle the challenge of creating a robust pathogen identification method. Secondly, this study provided a key exploratory analysis. It demonstrated the increased sensitivity of mNGS for pathogen identification. It also showed the potential benefits of coupling pathogen detection with simultaneous profiling of the host response and microbiome structure.

4.2 PATHOGEN DETECTION FOR LRTI DIAGNOSIS IN A DISTINCT COHORT

4.2.1 *Application Goal*

mNGS has now been applied for infection diagnosis in several clinical settings including diagnosis of central nervous system infections, ocular infections, and respiratory infections^{26,28,29,50,67}. Yet, there is currently no standardization of methods for sequencing and data analysis. A variety of tools exist for mNGS data analysis, and methods for interpretation of mNGS data in the clinical context are being expanded^{28,56,116,117}. To enhance standardization and adoption of analytical methods across disease settings, it is necessary to prove their performance in many contexts. Here, we apply the automated pathogen detection algorithms derived in Chapter 3 to a separate, but related, cohort.

4.2.2 *Introduction*

The RBM and LRM were derived in a cohort composed of adults (ages 21 – 85+, mean age = 62). Their conditions represent the array of comorbidities and distribution of pathogens common in the adult ICU¹¹⁶, but patient characteristics can vary widely across locations, ages, and diseases. We sought to benchmark the models derived for LRTI diagnosis in Chapter 3 against another cohort of respiratory infections (same tissue type), but with different patient characteristics. Here, we tested the RBM and LRM in a cohort of patients collected as part of a study by Zinter *et al*¹¹⁷. There are two key distinctions between this cohort and the cohort in which the models were derived. First, the patients are much younger (median age = 11.2 (IQR, 4.3-16.2)). Secondly, due to a majority of

immune compromised patients, there are significantly more opportunistic fungal infections.

As noted in the previous application note, prior mNGS studies have applied various methods for pathogen ranking^{28,50,56,116}. One recent method, proposed by Zinter *et al*, relies on a cohort-centric approach for evaluating pathogens. Their approach leverages the observation that the majority of taxa are present at low abundance and in similar quantities across all samples. They describe each microbe as a function of two criteria, 1) abundance relative to other microbes in the sample and 2) abundance relative to the same microbe across all samples in the cohort. They define “outlier organisms” as those with an abundance of ≥ 10 rpm and a z-score ≥ 2 ¹¹⁷. We sought to compare the results of the RBM and LRM to the outlier identification approach proposed by Zinter *et al*.

4.2.3 Methods

Sample Collection / processing

Zinter *et al*. collected 41 BAL specimens from 34 immunocompromised children being evaluated for pulmonary disease at 3 children’s hospitals (University of California San Francisco Benioff Children’s Hospital, Indiana University Riley Hospital for Children, and the University of Minnesota Masonic Children’s Hospital) between 2014 and 2016. Samples were processed for RNA- and DNA-sequencing and raw mNGS data was processed using the metagenomics pipeline previously described¹¹⁷.

Applying LRM and RBM

Zinter *et al.* focused their analysis primarily on RNA-Seq data after showing that RNA-Seq yielded a 10-fold higher lower limit of detection for recovery of nucleic acid from hardy fungal isolates (*Aspergillus niger*) in a series of spike-in experiments¹¹⁷. We downloaded processed files from Zinter *et al.*'s RNA-Seq samples and applied both the RBM and LRM. The RBM showed reduced specificity as compared to clinical results, so the LRM was used for all subsequent analyses. The DNA-Seq rpm feature was discarded from the model, but weights for the remaining four features were maintained. Pathogens identified by the LRM were compared to those identified by clinical testing and then to those identified by Zinter *et al* (TABLE 4.3).

4.2.4 Results

Of the 23 pathogens identified by standard clinical microbiology, Zinter *et al.* identified 16, 70% and LRM identified 17, 74% (FIGURE 4.3). The two methods were concordant for 15, 65% of the clinically-confirmed cases; the LRM identified two cases of PCR-positive Rhinovirus that went undetected by Zinter *et al.*, while Zinter *et al.* identified one case of Enterobacter that was not significant by the LRM. Two cases of PCR-identified CMV were not selected as potential pathogens by either Zinter *et al.* or the LRM. Additionally, two cases of *Aspergillus* were identified only by standard diagnostics – one case was identified by galactose and the other had rare growth and was not considered clinically relevant.

Both methods identified more pathogens than were identified by clinical microbiology; Zinter *et al* identified 37 total pathogens and the LRM identified 40 total

pathogens. Of the pathogens identified by LRM, 25, 63% overlapped with those identified by Zinter *et al.* Of those that were concordant with Zinter *et al.*, and not identified by standard diagnostics, the majority (60%) were viral pathogens. Both methods identified some unique pathogens (not concordant with the other NGS method) that could not be identified by standard diagnostics. The pathogens identified uniquely by Zinter *et al.* were enriched for fungal taxa (5 of 11, 45% fungal) while those identified by the LRM were enriched for bacterial taxa (11 of 13, 85% bacterial), with particular enrichment of *Streptococcus* (6 of 13, 46%).

4.2.5 Discussion

As expected, both mNGS methods identified pathogens missed by standard clinical testing (especially viruses). Both methods identified three cases of coronavirus, two cases of influenza, and one parainfluenza. Standard diagnostics require independent tests for each suspected pathogen. The clinical notes state that in several of these cases, viral PCRs were not sent. This highlights a critical advantage of unbiased mNGS. The two cases of *Aspergillus* missed by mNGS may be expected, given the challenge with isolating nucleic acid from hardy fungi^{118,119}. In one case, this pathogen was identified clinically only by galactose testing and the second had rare growth in culture. It is reasonable to expect that the rare growth sample may have been below the RNA-Seq limit of detection.

When comparing to the method used by Zinter *et al.*, we highlight a few important cohort-specific differences. First, the LRM was trained with a cohort containing few fungal infections. Zinter *et al.* showed, through a series of dilution experiments, that fungal

nucleic acid is difficult to isolate for sequencing, even when highly abundant¹¹⁷. Thus, it is reasonable that the model weight for abundance skews performance for detecting low-abundance fungal pathogens. Secondly, many of these fungi are opportunistic pathogens – they only cause infections in immune compromised hosts (i.e. *Candida*, *Cladosporium*, *Fusarium*). Therefore, they are not on the list of common pathogens. The pervasiveness of immunocompromised status in the pediatric cohort may warrant inclusion of opportunistic pathogens on the list of known pathogens. Finally, the LRM identifies more instances of common respiratory pathogens, including Rhinovirus and Streptococcus. This could be an artifact of the approach used by Zinter *et al.* that relied on an intra-cohort z-score. By nature of this method, highly prevalent pathogens will skew the z-score. Rhinovirus and Streptococcus are two of the most common respiratory tract infection pathogens^{30,120,121}. So, assuming that their distribution remains the same in this cohort, the intra-cohort z-score would have reduced sensitivity in these cases.

4.2.6 Conclusion and Opportunity for Future Work

We have shown that the LRM for pathogen identification enhances diagnosis of LRTI etiology in an independent cohort with distinct patient characteristics. Yet, in cases of opportunistic fungal infection, the performance is reduced as compared to the outlier detection approach developed by Zinter *et al.* Future work may profit from developing a model through meta-analysis of cohorts with diverse characteristics. And, with enough data, one could envision training a model specific for fungal vs. bacterial vs. viral infections. This may be further enhanced by incorporating differences in host response to various types of infection^{33,122}.

4.3 PATHOGEN DETECTION FOR MENINGITIS ETIOLOGY SURVEILLANCE

4.3.1 *Application Goal*

The purpose of this application is two-fold. First, to evaluate the utility of the LRM for identifying pathogens in a different disease context and tissue type. And second, as a case study testing whether we can use an automated data interpretation approach to expedite mNGS-based surveillance for informing policy decisions in low- and middle-income countries.

4.3.2 *Introduction*

Results shown in Chapter 3 and Section 4.2 have demonstrated the utility of mNGS and the LRM for LRTI diagnosis. Previous studies have shown that mNGS improves sensitivity for identification of meningitis etiology^{28,67}. While most studies on meningitis have relied on expert interpretation of mNGS results with respect to clinical covariates^{28,67}, we hypothesize that the LRM could automate pathogen detection in the context of meningitis.

Meningitis is a disease characterized by inflammation of the membranes (meninges) surrounding the brain and spinal cord¹²³. Globally, there are 10.6 million cases each year, resulting in 288,000 deaths¹²⁴. There are multiple causes of meningitis – infectious (bacterial, viral, fungal), as well as non-infectious autoimmune causes¹²³. In greater than 50% of meningitis cases, standard clinical diagnostics fail to identify the etiology^{22–24}. The burden of meningitis is especially high in LMICs, where diagnostic test availability is limited¹⁷. A lack of surveillance data on possible meningitis etiologies

contributes to this burden¹⁷. Increased surveillance can inform policy decisions for allocating funding towards preventative measures and additional diagnostics^{8,17}, with the goal of reducing idiopathic meningitis and improving treatment plans.

Related studies have shown the utility of mNGS for pathogen surveillance^{125,126}. But, interpretation of mNGS data can be challenging due to the high sensitivity for detecting nucleic acid present in the reagents, laboratory, and clinical environments. Data interpretation is consistently listed as an outstanding challenge with mNGS^{26,126}. One challenge with expanding mNGS for disease surveillance in LMICs is the paradox whereby most expertise in mNGS data analysis is centralized in wealthy countries, whereas the greatest infectious disease burden and opportunity for impact is in LMICs¹²⁷. As mNGS expands into regions with less mNGS expertise, an interpretable model for pathogen ranking is essential. One appeal of the LRM for pathogen detection is the rigorous and automated nature.

Here, we apply the LRM to a cohort of pediatric patients in Dhaka, Bangladesh, with meningitis due to known infectious, non-infectious, and idiopathic causes. We elaborate on minor modifications to the algorithm and demonstrate its performance. We then discuss how the findings in this surveillance study can translate into low-cost diagnostic assays.

4.3.3 Methods

Sample collection

Cerebral spinal fluid (CSF) samples were collected from pediatric patients at Dhaka Shishu Hospital (DSH) who met World Health Organization-defined criteria for meningitis. Laboratory tests were run to identify meningitis etiology and CSF samples were stored. Positive and negative controls were identified, as well as several idiopathic meningitis samples, where mNGS could be applied to identify unknown disease etiologies. Positive samples included those for which an infectious etiology was identified through culture, serology, antigen-testing, and/or qPCR. Negative samples were randomly chosen from the set where meningitis was suspected but not attributed in the final diagnosis, patients were discharged within 6 days after hospitalization, and CSF samples contained ≤ 6 WBC/uL and ≤ 30 ug/dl of protein. Idiopathic meningitis samples included those with a suspected infection (≥ 20 WBC/uL, at least 40% polymorphonuclear cells, and ≥ 40 ug/dl of protein), but no etiology detected by standard diagnostics. An additional set of environmental controls was included, consisting of water samples exposed to the entire diagnostic pipeline.

RNA was extracted from all samples and sequencing was performed according to previously published methods¹²⁸. External RNA Controls 103 Consortium (ERCC) spike-in controls were added to the sequencing libraries to enable back-calculation of total input RNA. Libraries were sequenced on NovaSeq 6000 to generate 150bp paired-end reads.

To benchmark a reduced-cost, single-nucleic acid test for surveillance purposes, we used only RNA-Seq data. Thus, we modified the LRM to remove the DNA-Seq rpm

feature. The model was re-trained on the original derivation cohort from Chapter 3 using only the following three features: 1) RNA-Seq reads per million, 2) rank amongst all detected microbes within the sample, 3) a binary variable indicating whether this microbe is known to cause meningitis.

Pathogen detection by LRM

Microbial pathogens were identified from raw sequencing reads using the IDseq portal (v1.8)¹²⁹, a cloud-based, open-source bioinformatics platform designed for detection of microbes from metagenomic data. Similar to prior methods²⁸, a z-score metric was used to distinguish microbes from both ubiquitous environmental contaminants and commensal flora. The z-score was computed for each genus relative to background distribution derived from the set of non-infectious CSF. Microbes with a z-score less than 1.0 were removed from the analysis due to overrepresentation in the background. Then, the modified LRM was applied to classify microbes and assign etiological candidates in each sample. Given limited documentation on meningeal coinfections, in cases where two potential pathogens were identified for one patient, only the top scoring pathogen was considered. Based on template-free (“water only”) controls, a minimum calculated RNA input threshold of 3.0 pg was required for pathogen prediction. The average RNA input of the set of non-infectious CSF samples was 2.2 pg.

Chikungunya virus pathogen prediction

Patient follow-up was conducted for all resolved idiopathic meningitis cases. Following data analysis, 472 additional CSF samples were tested for *Chikungunya virus*

(CHKV) using a qPCR assay with previously published primers¹³⁰. Samples with CHKV present in the CSF by qPCR were sequenced in a second batch using the same protocol as the initial set of samples. In the subsequent set of CHKV viral infections, the LRM was again applied as described above.

4.3.4 Results

Pathogen Detection

Specimens of known etiology: The set of positive controls included samples in which a combination of standard lab diagnostics (culture, qPCR, antigen testing) had previously identified pathogens. The LRM correctly identified pathogens in 7 of 8 samples that were culture positive (FIGURE 4.4). Taking into account all specimens that were culture, PCR and/or antigen/serology positive (n=36) regardless of cycle threshold (Ct), a total of 25 (69.4%) specimens were classified as containing a potentially pathogenic microbe (FIGURE 4.4). However, we were able to identify pathogens in 24 of 27 (88.8%) samples with RT-PCR Ct < 32.

Negative specimens: Only four of the negative control samples contained input RNA masses greater than the 3.0 pg threshold identified from the water controls. The LRM did not identify any potential pathogens in these samples.

Idiopathic meningitis specimens: Of the 25 idiopathic specimens, the LRM identified potential etiologies in ten cases – four bacterial pathogens (*Salmonella enterica*, *Stenotrophomonas maltophilia*, *Bacillus cereus*, *Mycobacterium tuberculosis*) in four cases, and three viral pathogens (Mumps virus, Enterovirus B, and Chikungunya virus) distributed across six cases.

Follow-up and confirmatory testing

Follow-up and confirmatory testing provided support for the pathogens identified in idiopathic cases. Based on the detection of three CHKV cases, a qPCR assay was developed and used to screen 472 additional CSF samples. This confirmed the presence of Chikungunya meningitis in 17 more cases during a previously identified febrile CHKV outbreak. Nucleic acid aligning to CHKV was detected in all 17 of these samples. But, using the LRM, CHKV was only predicted as a pathogen in 10 of 11 samples with Ct \leq 35.1 (FIGURE 4.5).

4.3.5 Discussion

Application of the LRM can facilitate pathogen identification as part of large-scale surveillance efforts in LMICs. The LRM relies on the assumption of an abundance of nucleic acid in the context of an infection. Thus, the LRM had difficulty identifying true positive cases of *Streptococcus pneumoniae* with Ct above 32. A high cycle threshold indicates low amounts of nucleic acid – potentially due to clearing of the initial infection, antibiotic use, or sample degradation over time. Applying the LRM to this cohort, with well-characterized RT-PCR data, informs us of the limitations in the context of low-abundance microbes.

This study demonstrates how mNGS surveillance can translate into actionable point-of-care diagnostic tests. Even in a small cohort, we managed to identify patterns in meningitis disease etiology. Development of a low-cost PCR assay after identifying three *Chikungunya virus* cases shows how future surveillance efforts could use similar data to

influence policy and direct funds towards new diagnostics. Idiopathic meningitis cases were all attributed to pathogens not covered by the standard diagnostics, further highlighting the need for surveillance to identify trends in pathogen prevalence and inform addition of new tests.

4.3.6 Conclusion

This study demonstrates how we can use mNGS to inform important policy decisions worldwide. It further demonstrates, as a pilot, that it is possible to generate mNGS data in diverse settings. Increasing the diversity of locations sampled by mNGS will enable conclusions on a global scale.

4.4 SEPARATION OF HOST FROM HOST: DONOR VS. RECIPIENT IN mNGS OF HCT PATIENTS

4.4.1 *Application Goal*

As described in Section 4.1.2, HCT refers to the transplantation of hematopoietic stem cells to treat hematologic malignancies. Donor cells establish marrow and immune functions in recipients whose immune cells have been ablated via conditioning regimens including chemotherapy and radiation¹⁰⁸. Novel infusion and myeloablative conditioning strategies have recently expanded allogenic HCT availability^{131,132}. Yet, there remain many potential complications associated with allogenic HCT. These include increased risk of infections, disease relapse, and graft-versus-host disease (GvHD)¹³³.

One particularly challenging complication is GvHD. GvHD occurs when donor cells (the graft) recognize the recipient's healthy tissues (the host) as foreign, and attack them¹³⁴. Risk factors for GvHD include age, disease, HLA-mismatch, unrelated donors, and sex-mismatch^{133,135–137}. Current trends suggest increasing transplantation from unrelated and sex-mismatched donors¹³², contributing to increased risk of GvHD. One aspect making GvHD challenging to address is that its' clinical presentation mimics that of infection^{134,138}. As discussed previously, standard diagnostics for infection are already limited. Thus, ruling out an infection in favor of a GvHD diagnosis is challenging. However, adoption of highly sensitive molecular assays, such as mNGS for pathogen identification, is being explored for this particularly challenging patient population⁵⁰.

From mNGS data, we can also extract more information that may assist in the diagnosis of GvHD. Here, we explore one computational approach to tease out host and

donor (graft) sequences from a single mNGS case study. As transplantation techniques continue to advance, this is one foreseeable application of combined host and metagenomic sequencing.

4.4.2 Introduction

GvHD occurs when transplanted immune progenitor cells recognize the recipient's (host) cells as foreign. The resulting immune response degrades healthy tissue. One risk factor for GvHD is a sex-mismatched donor. In the context of sex-mismatched donor, one set of cells will contain two copies of the X-chromosome (female) and one set of cells will contain a single copy of each X- and Y- chromosomes (male). Here, we comment on bioinformatic methods applied to review the case of a 15 -year-old female (XX) who received an allogeneic HCT from a male donor (XY) for treatment of pre-B ALL (acute lymphocytic leukemia).

This case is particularly interesting due to the possibility of cardiotropic GvHD. Specific details of the case are provided in Zinter *et al*¹³⁹. The patient had a variety of early complications prior to being discharged on transplant day +40. However, on transplant day +75 she was re-admitted with a ventricular fibrillation. She received conventional supportive care while undergoing exhaustive diagnostic testing for infectious, inflammatory, toxin-mediated, and other common etiologies of fulminant myocarditis¹³⁹. Eventually, a myocardial biopsy was taken. Imaging revealed marked CD8+ lymphocytic infiltration of the tissue, which is often indicative of viral myocarditis, but no viruses were identified via immunohistochemical staining. The biopsy tissue was further processed for mNGS to test for any missed pathogens (Methods, below).

mNGS, as discussed previously, captures sequences belonging to both host and microbes. We hypothesized that the proportion of sequencing reads belonging to the Y-chromosome could indicate the relative abundance of donor versus recipient cells. However, significant portions of the Y-chromosome are composed of repetitive sequences or homologous sequences to the X chromosome¹⁴⁰. This makes alignment of short read sequences to the reference genome challenging. It is not uncommon for a single sequence to map to both X- and Y- chromosomes or to multiple locations within the Y-chromosome¹⁴¹. Thus, to evaluate the relative abundance of donor versus recipient cells based on proportions of sequencing reads belonging to the Y-chromosome, we must take into account the baseline proportion of reads uniquely mappable to the Y-chromosome.

4.4.3 Methods

Sample Processing

DNA was extracted from the formalin-fixed paraffin-embedded myocardial biopsy block and sequenced on the HiSeq 4000 platform to generate 1.5×10^8 125 base-pair (bp) sequencing reads.

Microbial Bioinformatics

Raw .fastq files were first processed for pathogen detection using the previously defined pipeline²⁸. Automated methods for pathogen-calling were not applied, but

microbial alignment counts were considered by clinician review with respect to the single patient case.

Host Bioinformatics

To establish the baseline for unique mappability of the Y- chromosome, Y- chromosome reference sequences were computationally chopped into 125 bp sequences offset by 1 bp each. We aligned the sequences to HG38 using STAR¹¹³ (v 2.5.2b) and then determined the proportion of uniquely mappable reads. This process was repeated from chromosomes 1 and 2 to establish a normalization factor. Raw .fastq sequences from the patient sample were then aligned to HG38 using STAR¹¹³. The ratio of XY to XX cells was computed from the number of uniquely mapped sequences to the Y- chromosome, after normalizing for the unique mappability.

4.4.4 Results

No microbial sequences were identified above one read per million reads sequenced, ruling out the potential for DNA viral or other non-viral infection of the tissue. Since mNGS was limited to DNA-Seq, it was not possible to rule out the possibility of RNA-viral infection.

The Y-chromosome was found to be 27.16% uniquely mappable, as compared to 87.11% and 95.85% uniquely mappable sequences from chromosomes 1 and 2, respectively. 0.40% of the human DNA in the sample aligned uniquely to the Y- chromosome, suggesting a ratio of XY to XX cells of greater than 10:1 (after scaling by the uniquely mappable scaling factors identified by simulation experiments)¹³⁹. This

finding is consistent with lymphocyte:myocyte ratio in the biopsy, as measure by visual inspection at 40x magnification.

4.4.5 Discussion

The lack of microbial sequences and the large ratio of donor (XY) cells as compared to host (XX) cells suggests infiltration of donor cells into the tissue, consistent with GvHD. The case-study presented is significant for two clinically-appreciable reasons, as outlined by Zinter *et al.* First, the patient survived 17 days of venoarterial extracorporeal life support. Secondly, this represents a case of cardiotropic GvHD, for which few other case studies exist. Further, we have demonstrated a preliminary evaluation of the use of X- and Y-chromosome sequences to evaluate the likelihood of GvHD, taking into account challenges in bioinformatic approaches for repetitive sequences.

4.4.6 Conclusion

It is reasonable to expect that as techniques for allogeneic HCT increase the procedure prevalence and as other therapies involving transplantation of donor or CRISPR-modified host cells become more widely adopted^{142,143}, the number of cases in which teasing apart host from non-host (both human) for a variety of clinical interpretations will grow. This analytical extension on the standard mNGS pipeline used throughout these applications demonstrates another way to gain insight based on the host fraction of the sequencing data.

4.5 EXPANDING mNGS TO UNDERSTAND THE FLOW OF ANTIMICROBIAL RESISTANCE GENES AS A FUNCTION OF GLOBAL TRAVEL

4.5.1 *Application Goal*

The success of unbiased mNGS as a diagnostic derives from its ability to identify pathogens without *a priori* knowledge of the sample contents. While microbe-level information can improve diagnosis of infections and pathogen surveillance, probing deeper to understand the microbial gene content has major implications for public health. In particular, the prevalence antimicrobial resistance genes, microbial genes which confer resistance to some of the front-line antibiotics used to treat infections, presents a growing global health threat¹⁴⁴. mNGS provides the ability to assess both gut microbiota composition and the antimicrobial resistome with a single assay. Here we provide a case study using mNGS to evaluate both the microbiome and AMR profiles.

4.5.2 *Introduction*

As mNGS gains utility in diverse settings, we can further understand global trends in disease transmission. Beyond understanding the transmission of pathogens in a more globally connected world, one point of interest for improved public health policy is tracking the flow of AMR genes. International travel is a known contributor to the spread of AMR^{145–147}. Resistant microbes acquired during travel may persist asymptomatically well beyond the return from travel, resulting in transmission into the environment and susceptible populations¹⁴⁸. It is hypothesized that changes in the intestinal microbiota

during travel may underlie the acquisition of AMR bacteria during travel, though this phenomenon is not completely understood¹⁴⁹.

To better understand acquisition and global exchange of AMR bacteria, this study applied mNGS to samples collected longitudinally from healthcare workers traveling internationally. A particular focus has been placed on characterizing the prevalence of plasmid-encoded extended spectrum beta lactamases (ESBL), enzymes that confer resistance to most beta-lactam antibiotics including penicillins and cephalosporins^{150,151}. ESBL-producing Enterobacteriaceae are of particular concern due to their worldwide prevalence and typical resistance to multiple other antibiotic classes^{151,152}. While the goal differs from that of a diagnostic test, if we consider infection to be a dysbiosis of the normal microbiome, we can use similar techniques to evaluate microbiome diversity and prevalence of particular taxa over time.

4.5.3 Methods

Study design and sample collection

Healthcare workers with planned travel to Asia or Africa were recruited between March 2016 and February 2018. Stool samples were collected by participants and deposited into vials with either RNAProtect (Qiagen) or Cary-Blair (CB) media. These were then submitted alongside surveys pre-travel (PRE), post-travel (PST), 30 days post-travel (30D), and 6 months post-travel (6MO). From these samples, RNA and DNA were extracted and sequenced using the Qiagen Power fecal kit¹⁵³. Meanwhile phenotypic ESBL resistance was tested by microbial culture, as described in Langelier *et al.*¹⁵³.

Bioinformatics processing

RNA- and DNA-Seq .fastq files were processed via the previously described bioinformatics pipeline²⁸. Before downstream analysis, microbial alignments were aggregated at the genus-level. To control for potential background contaminants, no-template water control samples were sequenced, and the rpm values associated with taxa found in water samples were directly subtracted from the samples. The short-read sequence typing tool, SRST2¹⁵⁴, was used with the Argannot2 database¹⁵⁵ to identify AMR genes. Those with allele coverage of at least 20% were used for downstream analysis. Downstream analysis of microbial community structure was performed using the Vegan⁵⁴ R package. Alpha and beta diversity metrics were computed and compared across time points, pre- and post-travel. Metrics were also compared across groups of patients with persistent ESBL-producing *E. coli* (ESBL-PE) carriage and those without. Statistical significance was computed using Wilcoxon rank sum test.

Data availability Raw sequences are publicly available on the SRA, with accession number: SUB4474900.

4.5.4 Results

Nine of 10 subjects were culture-positive for ESBL-PE upon return from travel. Of those, one individual was found to have been colonized prior to departure (T3). Three subjects had persistent ESBL-PE carriage at 30-days (T2, T3, T5), with two continuing

carriage at six months (T3, T5). Four subjects experienced diarrhea during travel, but only one had persistent diarrheal symptoms at six months (T5).

We first examined changes in the gut microbiome by alpha diversity and found that following international travel, SDI did not significantly differ upon return or at 30-days post-travel ($p = 0.674$ and 0.25 , respectively). We then evaluated whether microbial community composition (beta diversity) differed across all subjects post-travel, but found no difference (Bray Curtis Index⁸⁵, $p = 0.23$ by PERMANOVA). Even though global composition and diversity of the gut microbiota didn't change significantly following travel, there were significant differences in the abundance of discrete microbial genera. *Enterobacteriaceae* showed the greatest fold change in abundance post-travel, with the genus *Escherichia* being the most differentially increased ($p < 0.001$). Comparing between individuals with persistent carriage of ESBL-PE at 30 days and those without indicated no difference in SDI ($p = 0.56$, by t-test). Similarly, Bray-Curtis beta diversity measured pre- or post- travel did not differ between subjects with persistent ESBL-PE carriage at 30 days versus those without ($p=0.32$ by PERMANOVA).

The antimicrobial resistome was characterized by an increase in identified AMR genes on both RNA- and DNA-sequencing following return from travel ($p = 0.03$ and $p < 0.01$, respectively). ESBL and/or *AmpC* encoding genes were identified in 100% of samples with ESBL phenotype determined by culture, and 14% of samples without. Beta-lactam resistance genes were increased post-travel, as well as qnr plasmid-mediated quinolone resistance genes, trimethoprim, sulfa, macrolide and aminoglycoside resistance genes (TABLE 4.2).

4.5.5 Discussion and Conclusion

Application of mNGS to study the enteric microbiota and resistome profiles of returned travelers revealed a significant increase in AMR genes associated with an increase in the proportion of *Escherichia*. Yet, global trends in the diversity of the enteric microbiota were preserved. Interestingly, mNGS identified ESBL and/or AmpC encoding genes in all patients with phenotypic evidence of ESBL-PE. This suggests that mNGS may be a promising alternative to culture-based assays resistome profiling. The identification of several other resistance genes post-travel supports the utility of mNGS for unbiased detection of AMR genes. This study is limited by small sample size, which may have resulted in failure to identify significant changes. However, it demonstrates the utility of mNGS for paired evaluation of the microbiome and resistome. Future studies may use similar methods for assessing global flow of AMR genes.

4.6 FIGURES

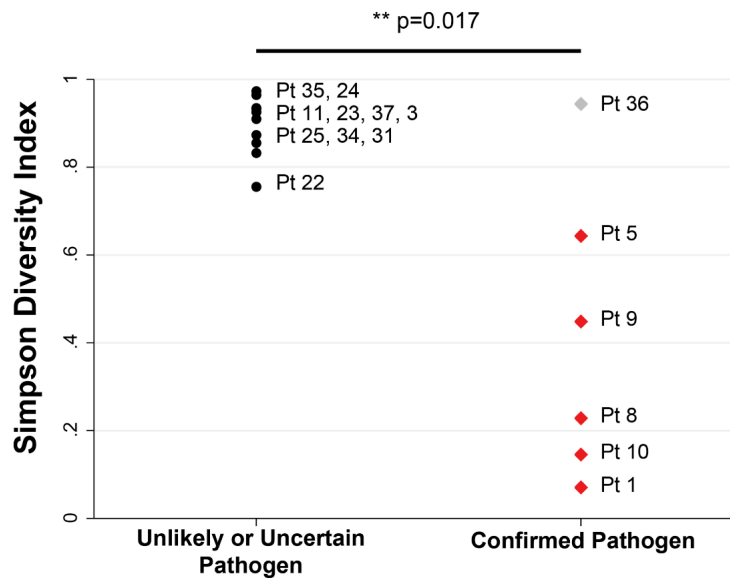


FIGURE 4.1 Simpson diversity index assessed for HCT patients with confirmed vs unlikely or uncertain pathogens

Bronchoalveolar lavage microbial diversity is inversely associated with presence of a transcriptionally active respiratory pathogen. Each data point represents a single patient (Pt) for whom the Simpson diversity index is plotted on the y-axis. Subjects are grouped according to confirmed pathogen (red diamonds) vs. unlikely or uncertain pathogen (black circles). Patients with confirmed pathogens had significantly lower diversity relative to patients with only microbes of unlikely pathogenicity (median, 0.34 [interquartile range (IQR), 0.15–0.64; $n = 6$] vs. 0.92 [IQR, 0.86–0.93; $n = 10$]; $P = 0.017$).

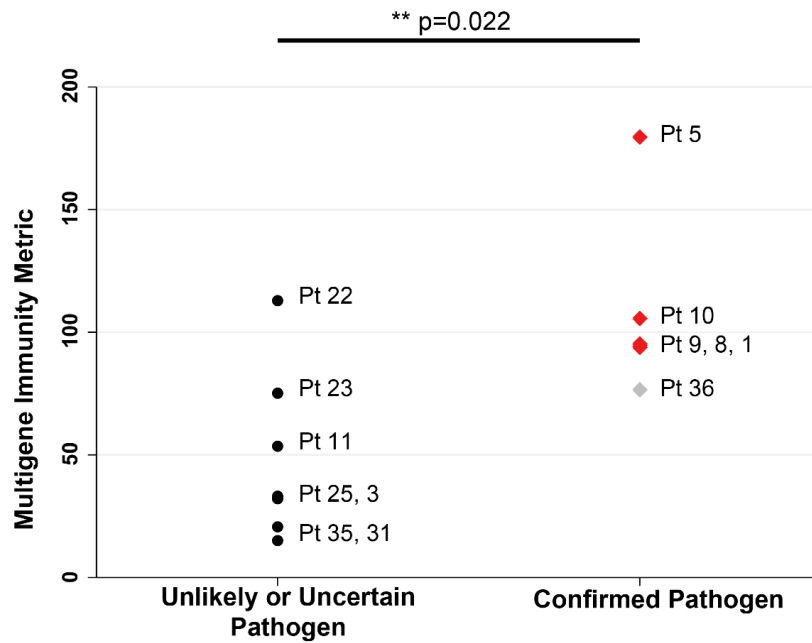


FIGURE 4.2 Host gene metric assessed for HCT patients with confirmed vs unlikely or uncertain pathogens

Expression of a host immune response multi-gene metric correlates with detection of LRTI pathogens. Each data point represents a single patient for whom the composite immune response gene metric is plotted on the y-axis. Subjects are grouped according to confirmed pathogen (red triangles) vs. unlikely or uncertain pathogen (black circles). Patients with confirmed pathogens had significantly higher multi-gene metric expression relative to patients with only microbes of unlikely pathogenicity (33.1, IQR 20.7-75.1, $n=7$ vs. 94.9, IQR 93.8-105.6, $n=6$, $p=0.022$).

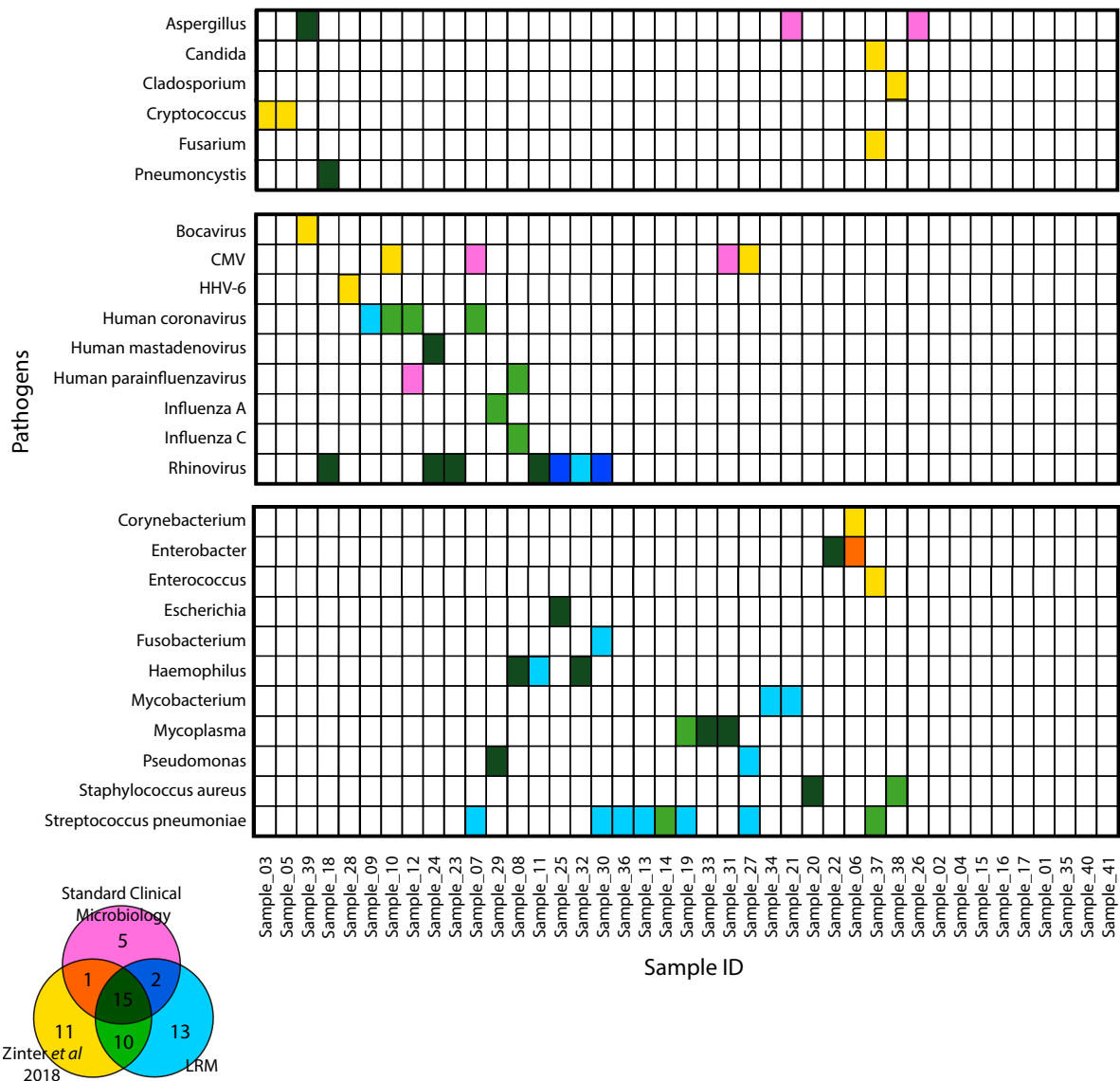


FIGURE 4.3 Comparison of pathogen identification methods in pediatric HCT recipients

Applying the LRM to the Zinter *et al.* 2018 dataset results in largely concordant findings. Both mNGS approaches identified more viral pathogens than standard diagnostics, but Zinter *et al.* identified more fungal pathogens and the LRM identified more bacterial and common pathogens. Pathogens are displayed in rows, separated into the categories of Fungal (top), Viral (middle), and Bacterial (bottom). Samples are listed in columns. Heatmap colors correspond to the concordance of results between the two approaches. The Venn diagram serves as a legend, the colors corresponding to which method or overlap of methods a particular pathogen was identified by. The number of pathogens in each section of the Venn diagram are listed.

Legend: Dark green: concordant across clinical diagnostics, the Zinter *et al.* approach, and the LRM, Light green: concordant between LRM and Zinter *et al.*, Dark blue: concordant between LRM and standard clinical microbiology, Orange: concordant between Zinter *et al.* and standard clinical microbiology, Light Blue: identified only by LRM, Yellow: identified only by Zinter *et al.*, Pink: identified only by standard clinical diagnostics.

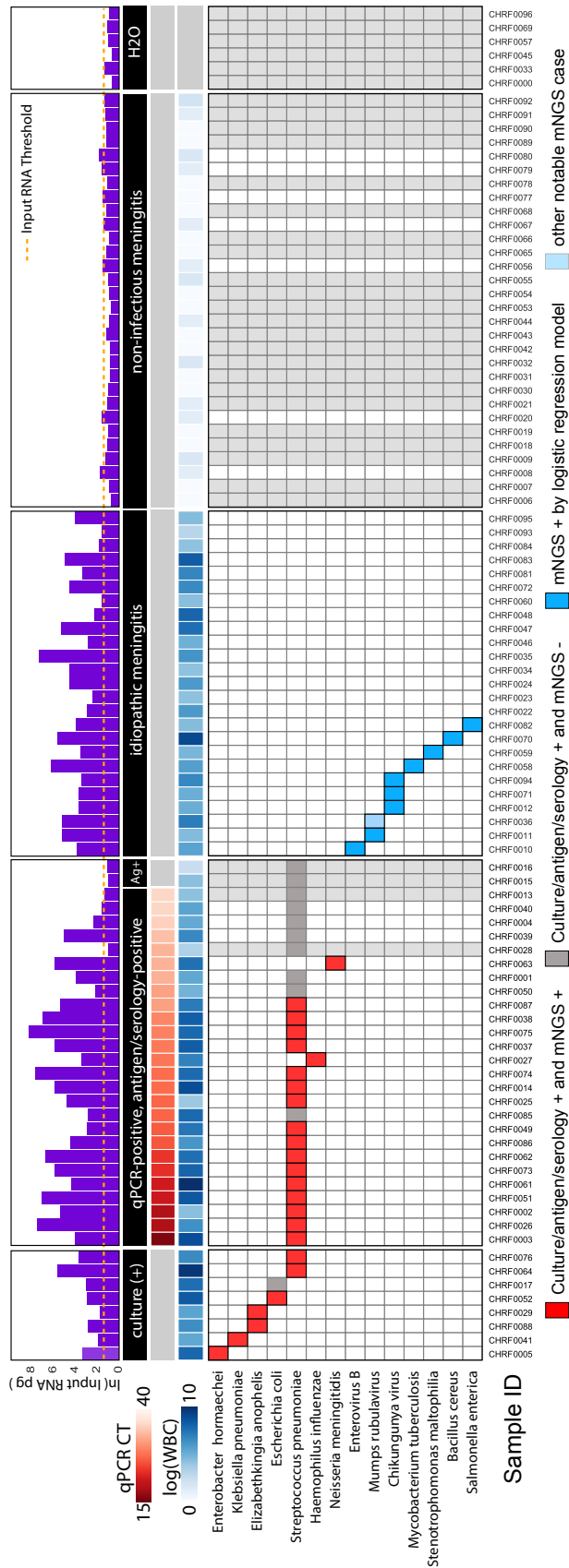


FIGURE 4.4 Application of the LRM to determine etiology of meningitis in Bangladesh

Summary of LRM pathogen identification applied to a cohort of pediatric meningitis cases in Dhaka, Bangladesh. Samples are shown in columns, ordered in sections from left to right as follows: specimens of known etiology (identified by culture, qPCR, antigen/serology, or antigen only), idiopathic meningitis specimens, negative specimens, water controls. **(Top)** Total input RNA (log pg.) is shown in the bar chart for all samples. Pathogen predictions were not made for samples with less input RNA than the maximum in the water samples (threshold indicated by the orange dotted line). **(Middle)** The white blood cell (WBC) counts obtained by the clinical lab are plotted beneath the x-axis as a heatmap in blue. These values indicate strong correlation between the total input RNA level and the WBC count. Samples with known etiology are ordered by increasing cycle threshold (Ct) and the qPCR Ct is indicated by the red heatmap. **(Bottom)** The predicted pathogens (listed in rows) for all samples are shown as filled-in squares. Samples that failed to meet the input RNA threshold for pathogen prediction are shown as filled grey columns. Dark grey squares indicate samples which were considered positive by clinical diagnostics, but for whom no pathogen was detected by the LRM using mNGS data. Red boxes indicate concordant findings and blue boxes indicate new putative pathogens identified by mNGS data that were not identified by standard clinical methods. The light blue squares indicate pathogens that were not picked up by the logistic regression method but were flagged as potentially interesting by manual review and followed up as if detected.

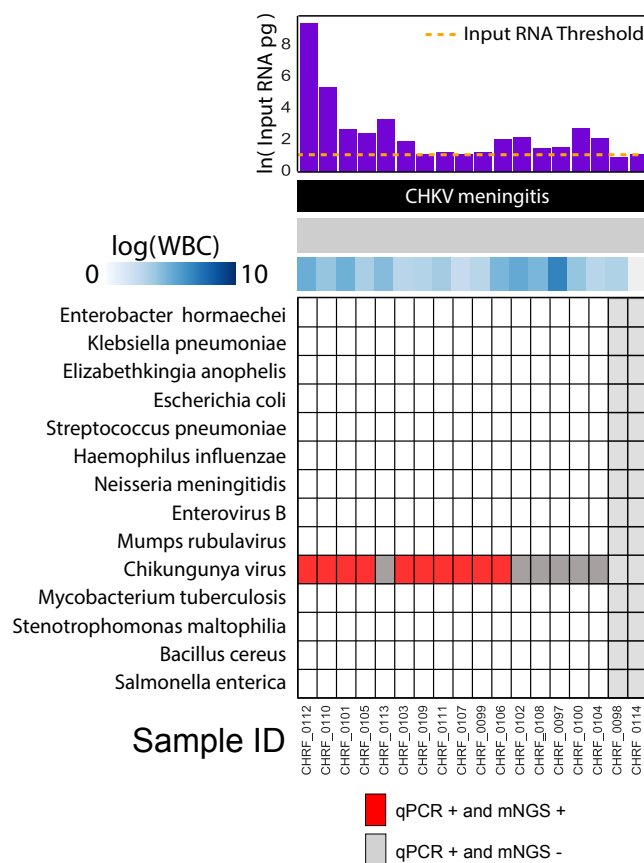


FIGURE 4.5 Application of the LRM to identify CHKV meningitis cases in Bangladesh

Summary of LRM pathogen identification applied to additional CHKV meningitis cases in Dhaka, Bangladesh. Samples are shown in columns, ordered by increasing Ct value. **(Top)** Total input RNA (log pg.) is shown in the bar chart for all samples. Pathogen predictions were not made for samples with less input RNA than the maximum in the water samples (threshold indicated by the orange dotted line). **(Middle)** The WBC counts obtained by the clinical lab are plotted beneath the x-axis as a heatmap, indicating strong correlation between the total input RNA level and the WBC count. **(Bottom)** The predicted pathogens (listed in rows) for all samples are shown as filled-in squares. Samples that failed to meet the input RNA threshold for pathogen prediction are shown as filled grey columns. Dark grey squares indicate samples which were considered positive by clinical diagnostics, but for whom no pathogen was detected by the LRM using mNGS data. Red boxes indicate concordant findings and blue boxes indicate new putative pathogens identified by mNGS data that were not identified by standard clinical methods. The light blue squares indicate pathogens that were not picked up by the logistic regression method but were flagged as potentially interesting by manual review and followed up as if detected. CHRF_0114 is a water sample.

4.7 TABLES

TABLE 4.1 Composite host immune gene metric for HCT patients

Summed expression values for genes in each of six *a priori* selected immunity pathways from the molecular signatures database as well as a composite multigene metric consisting of all genes represented in the immune-related gene sets. Group median expression levels were compared using the nonparametric Wilcoxon rank sum test. Patients 19, 24, 34, and 37 were excluded due to insufficient human transcripts. The original supplemental table, including lists of genes represented in the composite host immune response metric, can be found here: <https://datadryad.org/resource/doi:10.5061/dryad.800tj>

ID	Antiviral	IFN-alpha	IFN-beta	IFN-gamma	IL-6/JAK/STAT5	Adaptive Immunity	Composite
Established Pathogens							
1	33.01	53.37	38.96	63.15	25.14	4.61	93.82
5	47.66	54.85	84.63	101.71	52.67	14.63	179.58
8	38.22	22.24	20.44	51.68	31.70	3.10	94.64
9	12.50	17.22	35.88	59.12	34.39	13.06	95.24
10	12.71	28.14	46.13	61.72	25.36	8.98	105.61
36	11.87	10.09	26.84	48.80	17.39	12.86	76.54
Median (IQR)	22.9 (12.5-38.2)	25.2 (17.2-53.4)	37.4 (26.8-46.1)	60.4 (51.7-63.1)	28.5 (25.1-34.4)	10.9 (4.6-13.1)	94.9 (93.8-105.6)
Potential Pathogens Identified by NGS							
6	33.58	42.37	47.06	87.09	30.18	18.87	107.53
7	33.01	53.37	38.96	63.15	25.14	4.61	93.82
13	24.69	25.86	36.09	62.48	33.43	1.22	105.22
14	33.31	50.60	50.20	75.77	19.51	4.37	51.01
18	4.36	8.82	4.46	4.45	11.33	0	11.33
19	--	--	--	--	--	--	--
Median (IQR)	33.3 (24.7-33.6)	28.0 (25.9-42.4)	36.1 (27.2-47.1)	62.5 (51.8-75.8)	19.5 (12.4-30.2)	4.4 (1.2-5.8)	69.6 (51.0-105.2)
Microbes of Uncertain Pathogenicity							
3	4.36	4.36	16.86	4.37	15.23	0	32.09
11	0.04	14.33	17.10	33.19	16.16	1.20	53.56
22	23.14	31.36	30.48	81.13	39.30	6.48	112.86
23	38.87	39.31	29.28	97.54	38.43	9.27	75.08
24	--	--	--	--	--	--	--
25	8.22	8.72	16.18	33.27	5.94	0	33.09
31	4.31	3.18	9.76	15.61	1.24	0	15.03
34	--	--	--	--	--	--	--
37	--	--	--	--	--	--	--
Median (IQR)	8.2 (4.3-23.1)	14.0 (4.4-31.4)	16.9 (14.4-29.3)	33.2 (15.6-81.1)	15.2 (4.9-38.4)	0.1 (0-6.5)	33.1 (20.7-75.1)
Significance	p=0.063	p=0.116	p=0.015	p=0.116	p=0.116	p=0.022	p=0.022

TABLE 4.2 Host gene counts from mNGS study of adult HCT recipients

Raw host gene counts from Langelier and Zinter *et al.* can be found, along with additional supplemental data, here: <https://datadryad.org/resource/doi:10.5061/dryad.800tj>

TABLE 4.3 Comparison of pathogen identification methods in pediatric HCT recipients

Pathogens identified by standard clinical microbiology in the Zinter *et al.* 2018 cohort of pediatric HCT recipients with respiratory failure are compared to pathogens identified by the LRM/RBM and to those identified by the Zinter *et al.* outlier detection method.

Patient ID	Clinical Test Result	Zinter <i>et al</i> Pathogen	RBM/LRM Pathogen
Sample 03	None	Cryptococcus	None
Sample 05	None	Cryptococcus	None
Sample 39	Aspergillus	Aspergillus, Human bocavirus	Aspergillus
Sample 18	Pneumocystis, Rhinovirus	Pneumocystis, Rhinovirus	Pneumocystis, Rhinovirus
Sample 28	None	HHV6	None
Sample 09	None	None	Coronavirus
Sample 10	CMV	Coronavirus	Coronavirus
Sample 12	Parainfluenzavirus	Coronavirus	Coronavirus
Sample 24	Adenovirus, Rhinovirus	Adenovirus, Rhinovirus	Adenovirus, Rhinovirus
Sample 23	Rhinovirus	Rhinovirus	Rhinovirus
Sample 07	CMV	Coronavirus	Coronavirus, Streptococcus
Sample 29	Pseudomonas	Pseudomonas, Influenza A	Pseudomonas, Influenza A
Sample 08	Haemophilus	Haemophilus, Parainfluenza, Influenza C	Haemophilus, Parainfluenza, Influenza C
Sample 11	Rhinovirus	Rhinovirus	Rhinovirus, Haemophilus
Sample 25	Escherichia, Rhinovirus	Escherichia	Escherichia, Rhinovirus
Sample 32	Haemophilus	Haemophilus	Haemophilus, Rhinovirus
Sample 30	Rhinovirus	None	Rhinovirus, Fusobacterium, Streptococcus
Sample 36	None	None	Streptococcus

Table continued below.

Patient ID	Clinical Test Result	Zinter <i>et al</i> / Pathogen	RBM/LRM Pathogen
Sample 13	None	None	Streptococcus
Sample 14	None	Streptococcus	Streptococcus
Sample 19	None	Mycoplasma	Mycoplasma, Streptococcus
Sample 33	Mycoplasma	Mycoplasma	Mycoplasma
Sample 31	Mycoplasma, CMV	Mycoplasma	Mycoplasma
Sample 27	None	CMV	Streptococcus, Pseudomonas
Sample 34	None	None	Mycobacterium
Sample 21	Aspergillus	None	Mycobacterium
Sample 20	Staphylococcus	Staphylococcus	Staphylococcus
Sample 22	Enterobacter	Enterobacter	Enterobacter
Sample 06	Enterobacter	Enterobacter, Corynebacterium	None
Sample 37	None	Candida, Fusarium, Streptococcus, Enterococcus	Streptococcus
Sample 38	None	Staphylococcus, Cladosporium	Staphylococcus
Sample 01	None	None	None
Sample 02	None	None	None
Sample 04	None	None	None
Sample 15	None	None	None
Sample 16	None	None	None
Sample 17	None	None	None
Sample 26	Aspergillus	None	None
Sample 35	None	None	None
Sample 40	None	None	None
Sample 41	Aspergillus (not believed to be clinically important)	None	None

TABLE 4.4 Fold change in abundance of AMR genes

Fold change in abundance of antimicrobial resistance (AMR) genes detected by DNA-Seq with at least 20% allele coverage at the pre-travel and post-travel visits, listed by drug class or AMR gene class.

Legend: BLA: beta-lactamase, SUL: sulfa, GLY: glycopeptide, FLQ: fluoroquinolone, DFR: dihydrofolate reductase, AGL: aminoglycoside, MLS: macrolide, lincosamide, streptogramin, Tet: tetracycline, ESBL: extended-spectrum beta-lactamase.

Antibiotic class or AMR gene	Fold Change Pre- vs. Post-Travel	
	DNA-Seq	RNA-Seq
BLA	6	9
- <i>AmpC</i>	>100	>100
- <i>AmpH</i>	>100	>100
- <i>CTX</i>	>100	>100
- <i>MrdA</i>	>100	>100
- <i>OXA</i>	2	1
- <i>SHV</i>	>100	>100
- <i>TEM</i>	>100	>100
FLQ	>100	>100
TMP	>100	>100
SMX	21	29
MLS	2	7
TET	<1	<1
GLY	<1	<1
AGL	2	5

5 CLASSIFICATION OF COMBINED HOST AND MICROBIAL MNGS DATA IN THE CONTEXT OF AN IMPERFECT GOLD STANDARD

5.1 SUPERVISED CLASSIFICATION: HOST GENE EXPRESSION CLASSIFIERS WITH NOISY LABELS

5.1.1 Introduction

Challenges arising from poor performance of standard infection diagnostics extend beyond clinical decision-making. Studies applying machine learning methods to identify biomarkers for various disease conditions^{31,32,116} must rely on an imperfect gold standard, thus increasing the likelihood of sample mislabeling. Previous studies have shown that mislabeled samples can reduce model performance⁴⁰. And even in the strict case, where potentially-mislabeled samples are discarded from model training, the number of available samples is reduced. This has the potential to affect performance and applicability of the derived models¹⁵⁶.

The challenges imposed by an imperfect gold standard for machine learning methods are seen in Chapter 3. There, labeled data were used to generate models for predicting the presence of LRTI using the microbial or host fractions of mNGS data¹¹⁶. One strength of the study included the rigorous phenotyping performed by two-physician review of each case. This process was intended to enhance the quality of the dataset used to train the models for predicting infection. However, 52% of the cohort remained

without a diagnosis (LRTI^{+C} and unk-LRTI cases). With the remaining 48% of samples (LRTI^{+C+M} and no-LRTI cases), learning curves demonstrated that the host classifier was only approaching performance saturation. This indicated that additional samples may have added value for the classifier performance. (FIGURE 3.11). Additionally, while expert review may strengthen adjudications, even expert opinions are subject to deviation¹¹⁶ (Cohen's kappa for inter-rater agreement was 0.86). Finally, there were still cases in which mNGS suggested potential deviation from the assigned label, including false positive identification of pathobionts such as *H. influenzae* and *S. pneumoniae* in the no-LRTI group.

Differences in the available diagnostics at different medical centers further complicates the issue of an imperfect gold standard when extending biomarker studies across multiple-center cohorts. As mNGS expands for use as a surveillance tool in LMICs where available diagnostics are limited¹⁷, standard of care practices may differ, and minimal clinical phenotyping may be available, the issue is further compounded. One way to address the challenges posed by an imperfect gold standard, is to use methods that are robust to the impact of sample mislabeling. Such methods may improve models generated in the context of noisy labels while simultaneously allowing use of the whole dataset.

Machine Learning in the context of noisy labels

Machine learning refers to the creation and evaluation of algorithms for pattern recognition, classification, and prediction¹⁵⁷. The field of machine learning has undergone rapid expansion over the past 10 years¹⁵⁸. New methods for classification have proven

utility in a variety of fields¹⁵⁹. Supervised learning, in particular, has become a popular method for development of biomarkers from high-dimensional genomic and transcriptomic datasets^{31,34,35}. These methods rely on labeled data to identify features that can distinguish between one or more groups.

One of the limiting factors for machine learning is the need for large amounts of data. Generating large, labeled datasets can be expensive and time consuming, especially in the biomedical field. Recent work in machine learning has investigated methods for classification in the context of noisy labels. The goal is to source large datasets from public repositories where there exist increased potential for mislabeled samples^{41–43,160}. The existing literature on such methods can be broadly categorized into two types of methods. First, noise resistant methods and second, filter methods¹⁶¹. Briefly, noise resistant methods make use of the entire dataset, modeling the probability that a sample was mislabeled alongside the model for classification^{43,162}. Meanwhile, filter methods attempt to identify potentially-mislabeled samples and remove them from the training process.

Existing work regarding label noise in genomic data

Most exploration of classification methods for use with noisy labels has relied on benchmark datasets¹⁶³ and image data⁴¹. But, recently Bootkrajang *et al.* investigated their utility for analysis of microarray data^{162,164}. In particular, Bootkrajang *et al.* proposed a robust logistic regression (rLR) method that models the ambiguity of the label assignments directly as it builds the logistic regression classifier⁴³. They applied it to

cancer microarray data and showed success with recovering performance in the context of injected label noise¹⁶².

Chapter 3 demonstrated success in using the microbial fraction of mNGS reads for identification of patients with LRTI versus those without (FIGURE 3.9)¹¹⁶. Thus, we hypothesized that we could improve performance of host gene expression classifiers by leveraging the microbial fraction of mNGS data as an initial set of “noisy labels” across all samples and subsequently apply noise-resistant classification approaches. To test this, we first evaluated whether a simple pathogen metric could indicate the likelihood of an infection. Then, using simulated and publicly available datasets, we benchmarked the performance of standard logistic regression (LR) and rLR approaches for classification in the context of injected label noise.

5.1.2 Methods

Evaluation of simple pathogen-based metric

The top per-patient pathogen probability, output by the LRM, can predict LRTI status (Chapter 3). However, deriving the LRM required the use of well-labeled clinical microbiology data. To check the performance of a simple pathogen-based metric, we leveraged the previous observation that during an infection, there is generally a dominant microbe¹¹⁶. We hypothesized that the total sum of rpm aligning to microbes known to be potentially pathogenic^{21,76,77,106,107} may provide reasonable initial labels. On two datasets (the 92-sample tracheal aspirate dataset from Chapter 3 and the Bangladesh dataset from Chapter 4.3), we evaluated the sum of microbial rpm associated with 1) potentially-

pathogenic bacterial species and 2) viruses. We evaluated performance against standard clinical microbiology as reported in the respective studies (Chapter 3, Chapter 4.3).

Evaluation of host gene classifier performance in the context of label noise

Previous studies have shown reduced performance for classification in the context of label noise^{40,165}. To benchmark the impact of label noise on gene expression data, we used both simulated and publicly available data, summarized in TABLE 5.1. We evaluated the performance of logistic regression for classifying samples by infection type. Data was randomly split into training (70%) and test (30%) sets and noise was injected randomly into each class within the training set at pre-defined levels (0%, 5%, 10%... 50%). Using the mislabeled training data, a logistic regression model was trained using the R glmnet package¹⁶⁶. The AUC was computed and averaged over ten repetitions with different training and test sets.

Application of existing label-noise resistant logistic regression approaches

To test the utility of label-noise-robust logistic regression (rLR) methods for recovering model performance despite label noise, datasets were simulated using MATLAB scripts¹⁶⁷. Simulation parameters are shown in TABLE 5.1. For each simulated dataset, both standard logistic regression (LR) and rLR were applied to datasets with injected label noise. The accuracy was computed on a held-out test set and averaged over ten repetitions with different training and test sets.

Evaluating impact of feature filtration in the context of label noise.

We sought to determine whether a *priori* feature selection to reduce the set of input genes would be affected by the presence of label noise. At multiple levels of injected label noise, we applied feature selection. We evaluated the Jaccard Similarity between features selected on noisy data and features selected on the correctly labeled data. This was repeated for the top N = 100, 500, 1,000, 2,500, 5,000, and 10,000 ranked differentially variable genes between classes. Features were ranked using the Wilcoxon rank sum test and rankFeatures function in MATLAB¹⁶⁸. This function ranks samples by the absolute value of the standardized u-statistic for a two-sample unpaired Wilcoxon rank sum test.

Evaluating rLR performance as a function of sample size

Learning curves enable performance characterization as a function of input training set size¹⁰⁵. To test the assumption that increasing the sample size would increase classifier performance, we generated learning curves. To begin, we used the GSE60244 dataset and flipped 20% of the labels from each class. The dataset was subsampled to iterate over training set sizes with n = 6, 11, 16...71 samples. For each training set size, three classification methods were evaluated (TABLE 5.2). First, to evaluate the baseline performance of logistic regression without label noise, we ran LR using the full subsampled dataset with no injected label noise (Exp 1). Then, to test the impact of feature selection on performance, we ran a *priori* feature selection (selecting the top 500 genes by Wilcox rank sum) on the subset of samples with “correct labels” and then evaluated the performance of either applying LR to just this subset (Exp 2) or by applying rLR to the full dataset using only the features selected from the subset (Exp 3). The AUC

for performance on the left-out test set was calculated for each training size subset over seven iterations.

5.1.3 Results

The percentage of total rpm that are pathogenic, specifically $\text{argmax}((\text{pathogenic bacterial rpm} / \text{total bacterial rpm}), (\text{pathogenic viral rpm} / \text{total viral rpm}))$ resulted in an AUC = 0.88 for distinguishing between LRTI^{+C+M} and no-LRTI patients. At a threshold of 25% this metric would produce 8 false negatives (5/26 = 19%) and 1 false positive (5.8%). Meanwhile, we achieve AUC = 0.68 in the Bangladesh cohort, with only one false positive label. The challenge with this cohort comes from the high-Ct value Streptococcus cases where the LRM fails to identify pathogens. These results suggest that a simple metric for pathogen presence can be used to generate noisy labels on an unlabeled cohort.

Consistent with previous reports, we observed reduced classification performance in the context of label noise in both publicly available and simulated datasets. In Dataset GSE60244, injection of 50% label noise in one label class reduced average performance from AUC of 0.90 to 0.76. In a simulated dataset with 100 samples, each with 100 features, flipping 50% of the labels from one class reduced the mean accuracy from 100% to 72% (regardless of which class was flipped) (FIGURE 5.1). Meanwhile, rLR maintained greater than 95% accuracy at the same level of noise. Across all noise levels tested, rLR significantly improved accuracy. However, on a simulated dataset comprised of only 50 samples, with 1000 features each, the performance of rLR suffers. Here, rLR results were comparable to, or worse than, LR. Finally, on a dataset with 50 samples and 100 features each, the superior performance of the rLR was recovered. This suggests that the rLR

approach suffers in the context of high dimensional data. Differences in performance were also influenced by the data simulation method. Specifically, the performance boost for rLR was magnified when simulated data had a larger class distribution but suffered when the two classes were simulated from similar distributions.

Given the reduced performance of rLR in the context of high dimensional data, we sought to evaluate methods for *a priori* feature selection. We observe that label noise does impact feature selection (FIGURE 5.2). For example, with two classes, when 30% of the labels for one class are flipped while the labels for the other class remain unchanged, we observe 18% and 40% overlap with the top 100 features identified by the true labels. While the overlap improves amongst the top 10,000 features (68% and 78%, respectively), the difference is notable. The trend remains constant for all proportions of flipped samples tested (0, 10%, 20%, 30% for each class). Samples where labels from both classes were flipped were most affected.

The learning curve for each of the three tested methods (FIGURE 5.3) shows a trend towards increasing performance with increasing training set size across all three tested conditions. We had hypothesized that Exp 3 would outperform Exp 2, given the larger available training set size. This would provide justification for using the rLR to improve performance. However, there were no significant differences ($p < 0.05$, by Wilcoxon rank sum) observed between performance of the three methods at any of the subsampled training set sizes. We conclude that feature selection performed by ranking of genes using a subset of the data has negative impact on the performance generalizability across the cohort. Any potential positive effect of using rLR is rendered insignificant at larger training set sizes, where Exp 2, which uses the subset of correctly

labelled samples and LR, outperforms the rLR method using the full dataset, including noisy labels.

5.1.4 Discussion and Conclusion

We had hypothesized that using a rLR method would enable incorporation of all data for model training and that this would improve performance. But, we first discovered that the rLR had reduced performance when the number of features (p) was significantly larger than the number of input samples (n). These so-called “large p , small n ” problems are common in bioinformatics applications¹⁵⁷. One common method to address challenges with large dimensionality is to reduce the number of features prior to model training. There are a number of methods for reducing the feature space prior to applying machine learning algorithms. We hypothesized that *a priori* feature selection would enable the rLR to improve classifier performance. We were interested in maintaining gene-level features, as opposed to applying methods such as principal components analysis (PCA) which reduce dimensionality through linear transformation of the input data. So, we evaluated feature selection methods that rank input features. Unfortunately, we observed that the features selected by Wilcox rank sum were affected by label noise.

To circumvent the feature selection issue, we hypothesized that we could use only the subset of data for which we have confident labels and apply feature selection on those samples. Then, we could generate “noisy metagenomic labels” to the remaining “unknown” samples and use the same features identified with the subset. The underlying hypothesis was that the increase in sample size enabled through use of the noisy labeled samples and the rLR method would increase performance. However, we observed no

significant difference between performance of models derived from the known subset of the data using LR and models derived from the full (noisy labelled) dataset using the rLR approach.

Altogether, the approaches evaluated here suggest several challenges with using label-noise robust algorithms for transcriptomic biomarker development. However, we propose several limitations may be addressed in the future. First, it is possible that the algorithms tested here are too numerical in nature. Small perturbations in labelling can induce large effects on feature rankings when using Wilcox rank sum test. Other approaches allowing for fuzzy feature selection or semi-supervised classification would be more amenable to this type of problem¹⁶⁹. Second, this study only evaluated algorithm performance on a limited number of datasets. Trends not observed in this particular dataset may be revealed with application to more datasets or classification over different covariates (beyond viral versus bacterial infection). As the amount of available data grows, we may identify cases where these and other algorithms for classification in the context of noisy labels would be well-suited. In the meantime, we explore other methods for analysis of combined host and metagenomic data. Chapter 5.2 reframes the label noise problem and discusses the use of unsupervised learning approaches.

5.2 UNSUPERVISED ANALYSIS OF COMBINED HOST AND MICROBIAL mNGS DATA USING VARIATIONAL AUTOENCODERS

5.2.1 Introduction

While supervised classification algorithms have dominated the biomarker discovery field, unsupervised learning is commonly used in exploratory analyses of genomic data. Unsupervised learning refers to a variety of methods for identifying structure within datasets without the use of *a priori* defined labels. In the context of noisy labels, it is possible to ignore the labels and reframe the analysis as an unsupervised learning problem.

For high-dimensional datasets, methods for dimensionality reduction (including PCA, t-SNE, etc.) are often applied prior to unsupervised cluster analysis. PCA, which collapses high-dimensional datasets into components of maximal variance, has been widely adopted in the field^{73,170,171}. Recently, with the development of single-cell RNA-sequencing and increasingly large datasets, several new algorithms have been established^{172–174}. But, most of these still rely on linear combinations of the input features.

Autoencoders are a type of artificial neural network that can be used to learn data encodings in an unsupervised manner. In the field of deep learning, autoencoders have recently been employed to improve the performance of deep neural networks for a variety of classification tasks^{175,176}. Autoencoders are trained to reconstruct the original input data using a reduced dimensionality latent representation. What distinguishes them from the aforementioned linear dimensionality reduction methods, is their use of nonlinear feature reduction to generate the low-dimensional latent representation. If we used only linear

activations, then the optimal solution would be strongly related to PCA^{177,178}. One challenge in autoencoder training is the potential to learn an identity function for the training dataset. Such a model would be overfit and have limited generalizability to external datasets. Several variations on autoencoders have been developed to ensure that the autoencoder doesn't learn an identity function, but rather captures generalizable features. These include denoising autoencoders (DAE)¹⁷⁹, sparse autoencoders¹⁸⁰, and variational autoencoders¹⁸¹.

The small size of many biological datasets has limited the application of deep learning approaches in the field. Neural network models generally need thousands of instances for training. Most studies involving NGS and transcriptional profiles include on the order of tens to hundreds of samples. As the number of publicly available datasets grows, work to merge the recent advances in deep learning to biological data is ongoing¹⁸². Several recent studies have investigated the utility of autoencoder architectures (DAEs and VAEs) for extracting biological insight from large transcriptome datasets. VAEs are well suited for analysis of large-dimensional datasets. They enable automatic engineering of non-linear features as well as a learning a reduced dimension manifold of the expression space. This can be used to simulate data and evaluate transitions between states¹⁸¹. Additionally, constraints requiring that the VAE's feature activations be normally distributed help regularize the model and make the manifold more interpretable.

Way *et al.*¹⁸³ developed a VAE, named Tybalt, to test whether a VAE could model cancer gene expression data. They used Tybalt to compress the 5000 most variable genes across 10,459 samples from 33 cancer subtypes within the Cancer Genome

Atlas¹⁸⁴ into 100 latent dimensions (LDs). Through gene set enrichment analysis, they interpreted the gene weights associated with each LD from the trained model. This showed that learned features represent known biological patterns¹⁸³. Meanwhile, the latent representation constructed by Tybalt separated cancer subtypes similarly to the full dataset, indicating that the VAE can learn low-dimensional structure. Here, we explore whether the Tybalt VAE can generate novel insight from combined host and microbial data from mNGS data. We first provide a recommendation on how host and microbial data features may be engineered before VAE analysis. We then provide some initial benchmarking data using this approach with a limited mNGS dataset.

5.2.2 Methods

Recommendation on method structure for VAE

In Chapters 3 and 5.1, we had separated host and microbial fractions of mNGS datasets and analyzed them as independent components. It is possible that integrating the data may reveal novel biological interactions or lend to improved performance and generalizability. One way to integrate the datasets would be to concatenate all the host features (counts per gene), along with all the microbial features (microbial alignments per million reads), into one large matrix. However, we know that the microbial fraction of the data can be sparse¹¹⁴, even at the genera level. To reduce sparsity in the matrix, we can compute the microbial content at multiple levels of the phylogeny. Specifically, we can calculate rpm values for every genus, then walk up the taxonomic tree to the family level and recompute the rpm values. We can repeat this process for the Order, Class, and

Phylum levels. Using regularization algorithms or feature selection, we can then reduce the feature space after computing the phylogenetic counts.

Neural networks have gained increasing popularity for their ability to make sense of high-dimensional datasets^{159,182}. Thus, we expect them to be reasonable candidates for use with this high-dimensional host-microbe data matrix. But, the amount of mNGS data is still limited. Here, we combined a series of mNGS datasets to test the utility of VAEs for dimensionality reduction and feature engineering.

Bioinformatics pipeline for generating input host and microbe count data

Microbial alignments were performed using IDseq, a cloud-based pipeline for mNGS analysis based on a previously described pipeline²⁸. Taxonomic counts per sample were aggregated at the genus level. Then, for each sample, the taxonomic tree was walked back to the family level and rpm values were summed across all members of the family. Similarly, rpm values summed across categories were calculated for virus, bacteria, eukaryote, and archaea. Host gene counts for each sample were generated using STAR¹¹³ alignment to HG38. Microbial rpm features were concatenated to the gene counts for each sample.

Datasets

To evaluate the VAE, we combined datasets across multiple tissue types and disease states previously discussed. The dataset included 1201 samples, with 73,295 features (including host and microbial features, computed as described above). TABLE 5.3 outlines the datasets from which the samples were derived. To improve data quality

and consistency across cohorts, all values were log-transformed, and the mean number of counts was computed across the cohort. Samples with their total sum of log-transformed counts less than 2 standard deviations from the mean were removed from the analysis (mean = 5.86, standard deviation = 0.50). This left 1143 samples distributed across five independent studies (TABLE 5.3). The top 5000 most variable features (genes and microbial taxa) were selected by median absolute deviation.

VAE Structure

We inherit the VAE architecture from Tybalt¹⁸³. We then initialized the VAE architecture based on the parameters identified by Way *et al.* in a parameter sweep. In particular, the learning rate was set to 0.0005, the batch size was set to 50, and the number of epochs to 100. After each training epoch, validation loss was computed in a test set that consisted of 10% of the samples. The model was initially trained using 100 latent features. However, given that reducing the number of latent features is one method for reducing overfitting¹⁸⁵, we subsequently evaluated the performance of a model trained with 50 and 25 latent variables.

Evaluating model features

To interpret the latent features learned by the model, we adopted the approach proposed by Way *et al.* We selected genes with significant weight (> 2.5 standard deviations above the mean) in each latent dimension. Some latent features had no genes that met this qualification. Genes identified as significantly weighted in each latent feature were then input into WebGestalt¹⁸⁶ for overrepresentation pathway analysis against the

Gene Ontology (GO) biological processes database¹⁸⁷. We used the genes from the set of 5000 highly variable genes and microbial taxonomic IDs input into the VAE as the background for statistical testing. Latent variables with significant pathway enrichment (after Benjamini-Hochberg FDR adjustment¹⁸⁸) are discussed below.

5.2.3 Results

One goal of unsupervised analysis is to identify clusters of related samples. Through visual inspection, the clusters visualized by TSNE of the VAE latent representation (FIGURE 5.4) generally mimic those in the TSNE representation of the zero-one scaled data (FIGURE 5.5). This suggests that the VAE is capturing much of the information content from the full dataset within the reduced set of latent dimensions; even at 25 latent dimensions, the large scale clustering patterns persist. We observe that MBAL and VAP datasets cluster together, while UGD is closely related, but not overlapping. This is reasonable given that both the VAP and MBAL datasets are derived from mNGS of tracheal aspirate samples, while the UGD study is composed of BAL samples, a similar, though not identical, fluid type (see Chapter 2). The serum and CSF fluids are also more similar to each other than to the other sample types. These clusters highlight the tissue type and cohort-specific differences as characterized by both their microbial profiles as well as known gene expression differences.

To apply the VAE for downstream applications, such as feature reduction for classification or interpretation of latent features, it is important to ensure that the model has learned robust features and not overfit the training data. The learning curves (FIGURE 5.4) show that the training and test error have plateaued over 100 epochs of

training, suggesting ample training. However, we observe a critical difference in the magnitude of error for the training and validation sets, indicating that the VAE model is overfit. One method to reduce overfitting is to limit the number of latent features in the VAE¹⁸⁵. Yet, even with only 25 latent features, the model remained overfit. We conclude that more training data may be required to train a robust model. This is not surprising, given that we have 10-fold fewer samples than the original dataset Way *et al*/ used to train Tybalt¹⁸³.

Despite being overfit to the training data, an initial analysis of the latent dimensions revealed biologically relevant features. Analyzing the model trained with 50 latent dimensions enabled manual interpretation of latent encodings through pathway analysis. WebGestalt overrepresentation pathway analysis identified six LDs with genes significantly overrepresented in GO biological pathways (TABLE 5.4). Pneumoviridae, Orthopneumovirus, and the category Viruses were both significantly associated with the LDs 3 and 14, while LD3 shown GO enrichment for response to type I interferon, and LD14 showed enrichment for response to virus. LD29 was significantly associated with response to bacterium and response to lipopolysaccharide. Finally, LDs 11, 34, and 41 were enriched for transmembrane and ion transport processes.

When considering which samples had high activation of each latent dimension (FIGURE 5.6), there are some clear trends. First, the two latent dimensions related to interferon response and viruses (LD3, LD14) were more highly active in the VAP samples. The VAP cohort was dominated by *Orthopneumovirus* viral infections and the samples highly active in LD3 and LD14 show higher levels of *Orthopneumovirus*. The samples with high activation of LD29, enriched for pathways relating to response to bacteria, are

a distinct but overlapping set. Meanwhile, we observe samples with similar activation of LD11 and LD34, but a distinct subset of samples with high activation of LD41, despite all three of these dimensions being significantly enriched for transmembrane and ion transport.

In the majority of cases (six of eight total) where microbial taxa were selected as being significantly associated with a particular latent dimension (> 2.5 standard deviations from the mean), multiple microbial taxa came from the same lineage. For example, in LDs 3, 14, and 21, both the genus *Orthopneumovirus* and its family *Pneumoviridae* were selected. Meanwhile, in LDs 9, 41, and 48 both the genus *Providencia* and its family *Morganellaceae* were selected. However, many unrelated taxa were identified as significantly associated with LD23 and LD45.

5.2.4 Discussion and Conclusion

Here, we show that we can train a VAE on combined host and microbial data to achieve a latent representation of the original dataset. But, with the available input size, the model was overfit, even when reducing to only 25 latent dimensions. This suggests that, as expected in comparison to other studies using VAEs, additional samples (on the order of 10,000) may be useful for this type of analysis. Despite the sample size limitation, we interpreted the latent features learned by the model with 50 latent dimensions, using methods described by Way et al. The interpretation suggests that, if fully trained, the model may be able to learn biologically relevant features. Combining host and microbial features has demonstrated some expected trends with respect to types of pathogen and host response (i.e. viruses and response to virus in LD3 and LD14). However,

interpretation of these features in this context will require more research and will benefit from training the model on additional data.

5.3 FIGURES

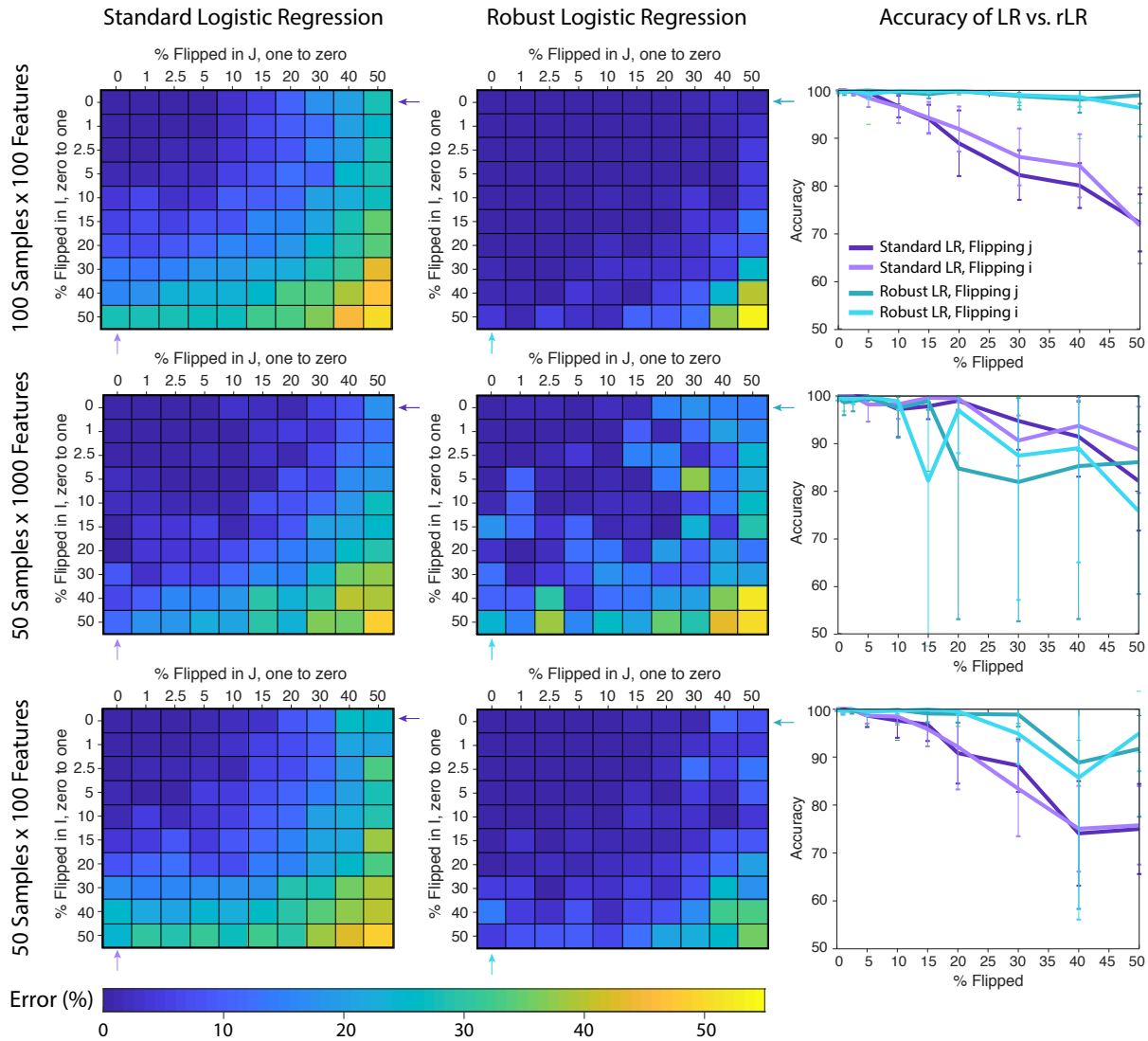


FIGURE 5.1 Performance of LR and rLR on simulated data in the context of label noise

Label noise was injected into simulated datasets of varying sizes and classifier performance was evaluated on a left-out test set using either standard logistic regression (LR) or robust logistic regression (rLR) methods; 100 samples, by 100 features (top row), 50 samples, by 1000 features (middle row), and 50 Samples, by 100 features (bottom row). Heatmaps show the error rate for each method, LR (left column) and rLR (middle column), with injected label noise in the positive class (x-axis) and negative class (y-axis). When leaving labels from one class constant and flipping only the labels from the other class, significant differences are observed between the methods (line plots, right column). rLR, in blue shades, shows significant reduction in the error (as compared to LR, in purple) at relatively high levels of label noise when there is a limited set of features (top and bottom rows). However, with many more features than samples (middle row), the performance is reduced.

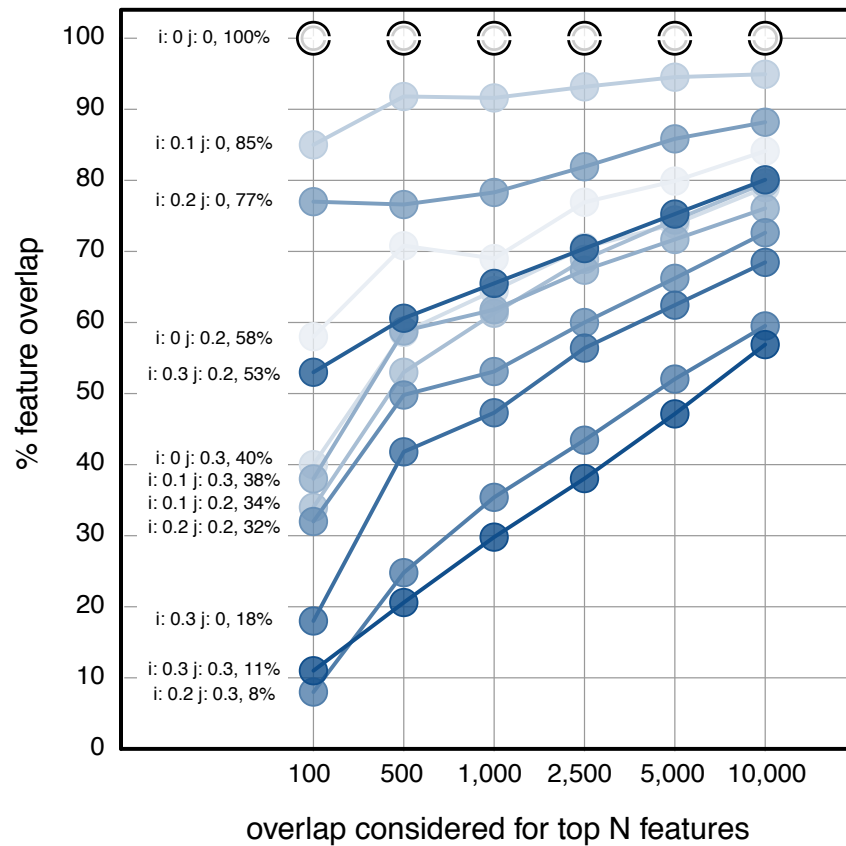


FIGURE 5.2 Impact of label noise on *a priori* feature selection

Label noise impacts the *a priori* selected features, therefore potentially influencing downstream classifier performance. *A priori* feature selection was performed with different levels of injected label noise. Each set of blue shaded points indicates one permutation, with the proportion of flipped samples indicated to the left of the points; i indicates flipping labels from *zero* to *one*, while j indicates labels flipped from *one* to *zero*. The percentage of features overlapping between those selected on the noisy dataset and those selected on the true dataset is indicated on the y-axis. For each dataset, we considered the overlap in the top 100, 500, ... 10,000 features (x-axis). At higher levels of noise (i and j between 0.2 and 0.3) less than 20% of the features selected on the noisy dataset overlap with those from the true labeled dataset. Meanwhile, across all noise levels, the percentage of overlapping features increases with increasing numbers of features considered.

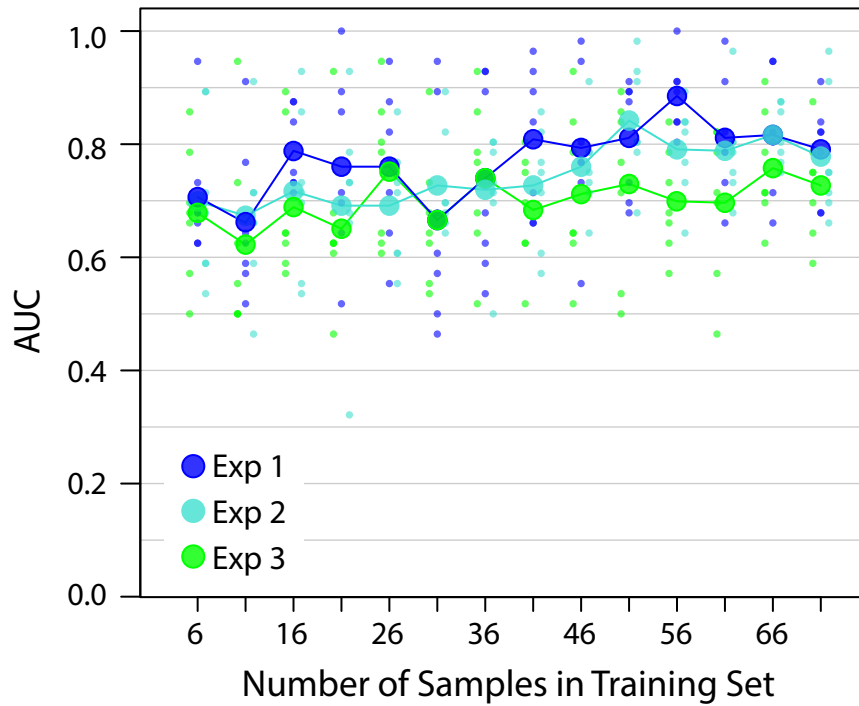


FIGURE 5.3 Learning curves computed to test rLR and LR

Learning curves were computed according to the three test conditions. Exp 1 (dark blue) is the positive control and indicates the performance using the full dataset, Exp 2 (light blue) is the negative control using only a subset of the available samples with true labels, and Exp 3 (green) is the test condition using rLR on the full dataset containing mislabeled samples. For each training size subset (x-axis), the AUC was computed (y-axis) over multiple iterations. The values at each iteration are indicated by small circles, while the mean value across all iterations is shown by the large circles.

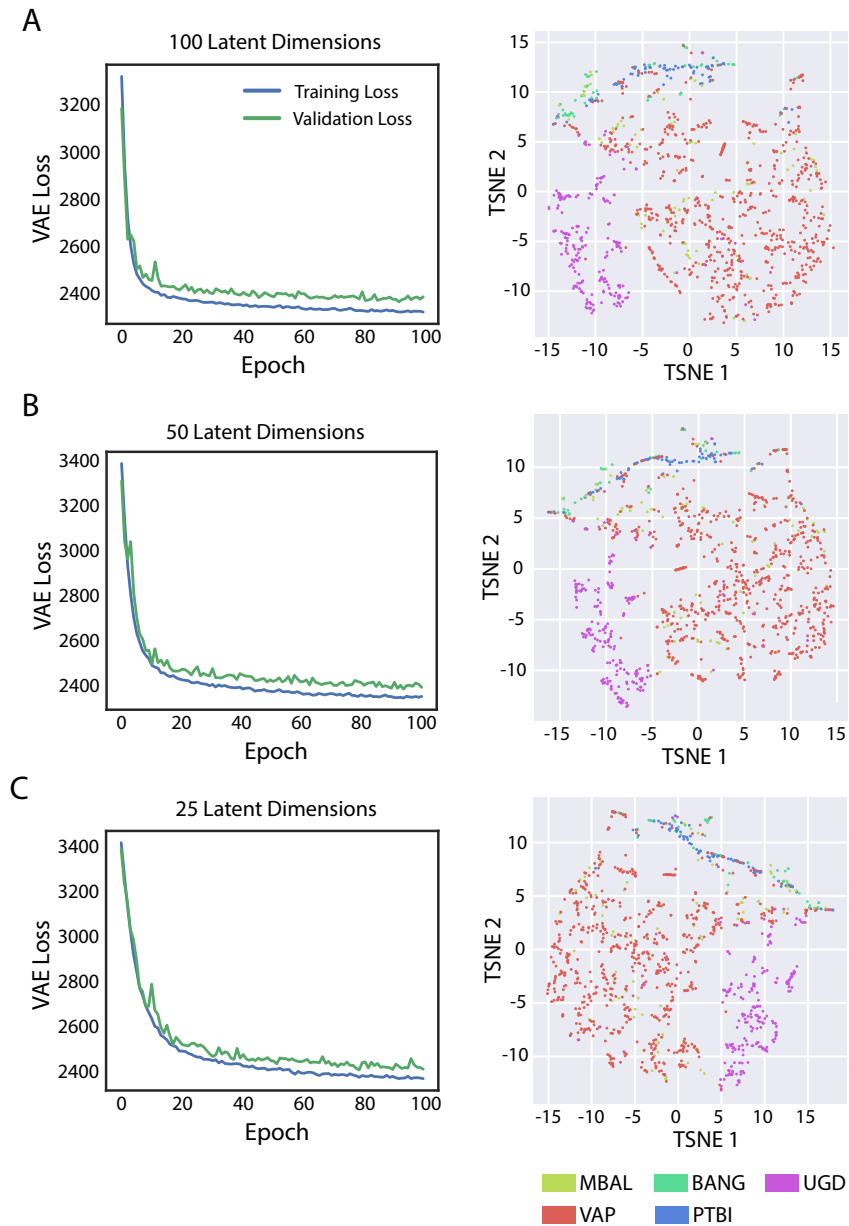


FIGURE 5.4 VAE training performance and latent representation

The variational autoencoder (VAE) was trained with **A**) 100 latent dimensions (LDs), **B**) 50 latent dimensions, and **C**) 25 latent dimensions. The line plots indicate the model loss computed on the training data (blue) and the test data (green) for each epoch of VAE training. In all three cases, the loss values plateau before 100 epochs, indicating that the model has reached maximal performance. However, in all cases, the loss on the test data is higher than the loss in the training data, suggesting that the model is overfit to the training data and would benefit from additional samples. For each of the three models (100, 50, and 25 latent dimensions), the TSNE projection of the VAE latent representation is shown (right). The samples cluster based on study and sample type, indicating that the model is learning relevant signal despite being overfit to the training data.

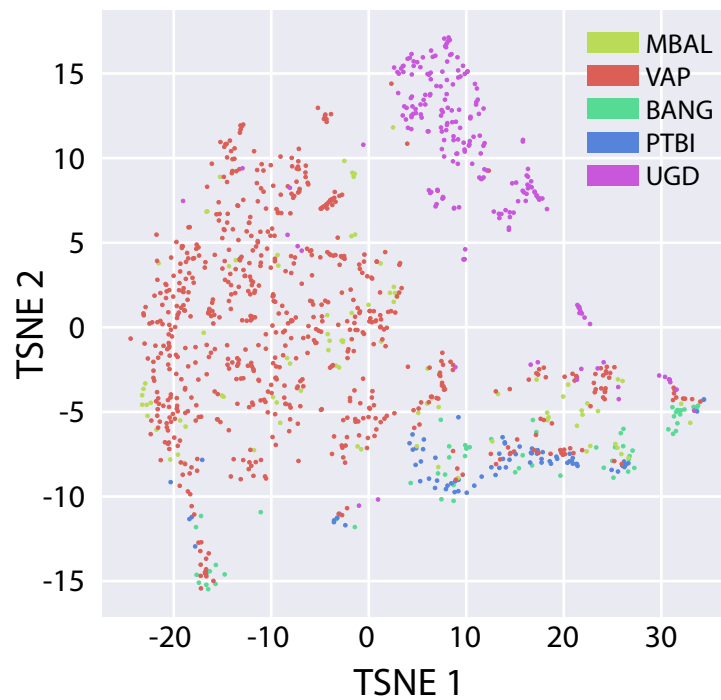


FIGURE 5.5 TSNE projection of zero-one scaled dataset

TSNE representation of the full zero-one scaled dataset including host and microbial features. The studies cluster based on fluid type. Notably, MBAL (yellow) samples are overlapping with VAP (red) samples, both of which are studies on tracheal aspirate samples. Meanwhile, UGD (pink) is a closely related, but distinct cluster comprised of bronchoalveolar lavage (BAL) samples. BANG (green) and PTBI (blue) samples include cerebral spinal fluid (CSF) and serum, respectively.

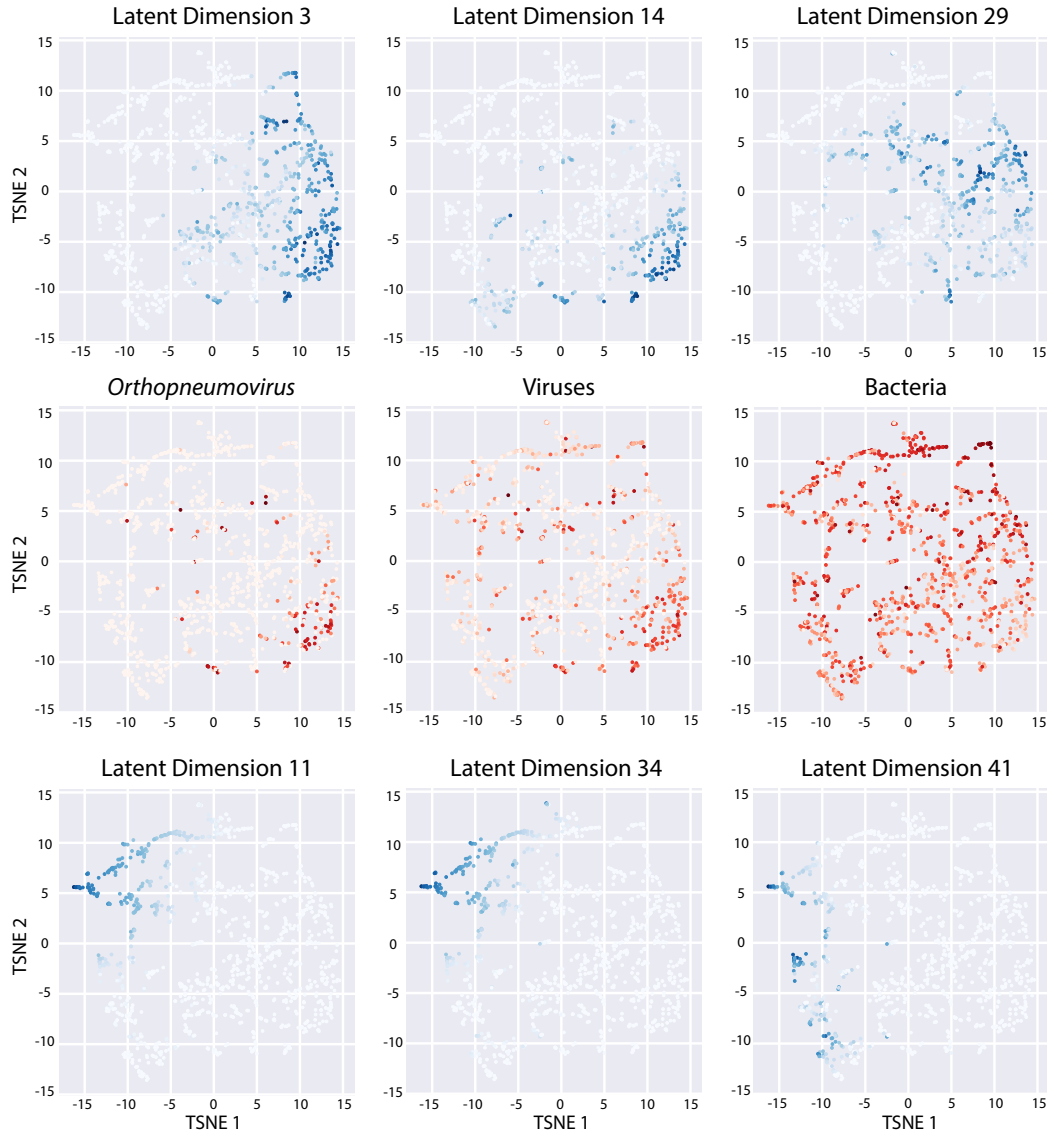


FIGURE 5.6 VAE latent dimension activations for each sample

TSNE plots show the reduced dimensionality representation from the VAE trained with 50 latent dimensions (FIGURE 5.4B). Each point indicates one sample and samples are colored based on the activation of each LD (blue, top and bottom rows) or pathogen counts (red, middle row). Lower activation or pathogen count values are shown in white and higher activations scale to dark blue or red, respectively. We show the activation of LDs with enrichment for Gene Ontology (GO) Biological Processes. In particular, LD 3 and LD 14 were enriched for GO BPs related to response to viruses. Samples with highly abundant *Orthopneumovirus*, or Virus category (red) overlap with the samples with high activation of LD 3 and LD 14. These samples are largely from the VAP study, which was known to be enriched for pediatric patients with *Orthopneumovirus* infections. LD 29 was enriched for response to bacteria and includes an overlapping but distinct set of samples with high activation. LD 11, 34, and 41 were all enriched for ion transport. However, LD 41 is strongly activated in a different set of samples than LD 11 and 34, which appear to be overlapping.

5.4 TABLES

TABLE 5.1 Simulated and publicly available datasets used for label noise analyses

Simulated and publicly available datasets were used to understand the impact of label noise on high dimensional gene expression data. The datasets are described and referenced.

Dataset	Summary	Reference
Dataset #1	Simulated 100 samples, 50 positive / 50 negative, 100 features, generative model parameter = 2.5	Simulated via Bootkrajang <i>et al.</i>
Dataset #2	Simulated 50 samples, 25 positive / 25 negative, 1000 features, generative model parameter = 2.5	Simulated via Bootkrajang <i>et al.</i>
Dataset #3	Simulated 50 samples, 25 positive / 25 negative, 100 features, generative model parameter = 2.5	Simulated via Bootkrajang <i>et al.</i>
GSE33341	Peripheral blood microarray data from mice challenged with <i>S. aureus</i> , <i>E. coli</i> , or no infection.	Ahn, SH., <i>et al.</i>
GSE60244	Whole blood microarray samples from patients with bacterial, viral, bacterial-viral coinfection, and healthy controls	Suarez, NM., <i>et al.</i>

TABLE 5.2 Test conditions for LR versus rLR learning curve

Learning curves were generated to test the hypothesis that using the robust logistic regression (rLR) approach with an increased sample size including noisy labels would improve model performance over that of a model trained on the subset of true labeled data. The specific conditions evaluated in the learning curve (FIGURE 5.3) are listed. LR: standard logistic regression.

Condition	Label noise?	Samples used for feature selection	Samples used for training	Algorithm
FULL (LR), Exp1	No	All samples	All samples	LR
SUB FILT (LR), Exp 2	Yes	Subset of samples with "correct labels"	Subset of samples with "correct labels"	LR
FULL FILT (rLR), Exp3	Yes	Subset of samples with "correct labels"	All samples, including mislabeled samples	Robust LR

TABLE 5.3 Datasets used for VAE analysis

To obtain a large dataset for training the VAE, several unpublished datasets were concatenated together. Broad descriptions of the datasets are provided. The number of samples in the dataset after filtering are shown, along with the number of samples in the original dataset (shown in parenthesis). CSF: Cerebral spinal fluid.

Dataset	# of Samples	Description	IDseq version	Published?
MBAL	91 (92)	TA samples collected from adult patients with respiratory failure due to infectious or non-infectious causes	2.7	Yes ¹¹⁶
VAP	688 (699)	TA samples collected from pediatric patients with respiratory failure due to infectious or non-infectious causes	3.2	No
BANG	70 (98)	CSF samples collected from pediatric patients with suspected meningitis	3.1	No
PTBI	81 (94)	Serum samples collected from pregnant women at 20 weeks	3.2	No
UGD	213 (218)	BAL samples collected from patients with suspected Tuberculosis in Uganda	3.2	No

TABLE 5.4 GO Biological Processes significantly enriched in VAE latent dimensions

For each latent dimension (LD) of the VAE (containing 50 total latent dimensions), genes with significant activation (> 2.5 standard deviations from the mean) were evaluated for pathway overrepresentation using WebGestalt. The LDs for which WebGestalt identified significantly enriched GO Biological Pathways are shown. All pathways with $FDR < 0.01$ are shown. Additionally, microbial taxa identified as significantly activated are included.

Latent Dimension ID	GO BP ID / NCBI Tax ID	GO Biological Process / Taxonomy Name
LD 3	GO:0034340 GO:0060337 GO:0071357 GO:0019221 GO:0071345 GO:0034097 GO:0045071 GO:0045069 GO:1903901 GO:0048525	response to type I interferon type I interferon signaling pathway cellular response to type I interferon cytokine-mediated signaling pathway cellular response to cytokine stimulus response to cytokine negative regulation of viral genome replication regulation of viral genome replication negative regulation of viral life cycle negative regulation of viral process
	11244 1868215	Pneumoviridae Orthopneumovirus Viruses
LD 11	GO:0055085 GO:0003008 GO:0006811 GO:0030001 GO:0015672 GO:0035637 GO:0008016 GO:0003015 GO:0060047 GO:0006812	transmembrane transport system process ion transport metal ion transport monovalent inorganic cation transport multicellular organismal signaling regulation of heart contraction heart process heart contraction cation transport
LD 14	GO:0006952 GO:0009607 GO:0009615 GO:0019221 GO:0034340 GO:0043207 GO:0045087 GO:0051607 GO:0051707 GO:0060337	defense response response to biotic stimulus response to virus cytokine-mediated signaling pathway response to type I interferon response to external biotic stimulus innate immune response defense response to virus response to other organism type I interferon signaling pathway
	11244 1868215	Pneumoviridae Orthopneumovirus Viruses

Table continued below.

Latent Dimension ID	GO BP ID / NCBI Tax ID	GO Biological Process / Taxonomy Name
LD 29	GO:0002376 GO:0009617 GO:0032496 GO:0002237 GO:0006950 GO:0051707 GO:0043207 GO:0009605 GO:0006955 GO:0009607	immune system process response to bacterium response to lipopolysaccharide response to molecule of bacterial origin response to stress response to other organism response to external biotic stimulus response to external stimulus immune response response to biotic stimulus
LD 34	GO:0055085 GO:0030199 GO:0034220 GO:0003008	transmembrane transport collagen fibril organization ion transmembrane transport system process
LD 41	GO:0055085	transmembrane transport

6 CONCLUSIONS AND FUTURE DIRECTIONS

Metagenomic next-generation sequencing is increasingly viewed as a universal pathogen detection method. But, interpretation of mNGS data has remained an outstanding challenge²⁶. The extreme sensitivity for nucleic acid detection complicates pathogen identification and analysis in the context of a background microbiome. This is especially true for the lower respiratory tract, where there exist pathobionts capable of coexisting in a healthy microbiome or causing an infection¹⁰. Additionally, previous studies using mNGS have focused on identifying pathogens, without considering the majority of sequencing reads that map to the host.

Here, we first evaluated the utility of mNGS as an improved molecular diagnostic for LRTI. To this end, we evaluated microbial differences between two sample types commonly used for LRTI diagnosis (TA and mBAL). Then, we developed methods for combined host and microbial analysis of mNGS data for LRTI diagnosis. We addressed the outstanding challenge of pathogen identification in the context of a background microbiota by developing one method for distinguishing likely pathogens from commensals. Then, we benchmarked biomarkers of LRTI using microbiome and transcriptome metrics. By combining analysis of pathogens with microbiome and host transcriptome biomarkers, we achieved a more comprehensive evaluation of the patient status.

After showing the utility of the combined host and metagenomic analysis for the diagnosis of LRTIs, we applied the methods to other disease contexts – including different cohorts of LRTIs as well as meningitis. We demonstrated that the methods are broadly applicable, but also identified a number of interesting considerations and opportunities for

future work to improve upon the existing algorithms. Limitations and future directions are outlined below.

6.1 FUTURE WORK FOR DEVELOPING LRTI DIAGNOSTICS

In Chapter 2, we presented a comparison of the microbiome from two distinct sample types commonly used for LRTI diagnosis – TA and mBAL. We showed that in the context of LRTI, any sample-type specific differences in microbiome content are rendered insignificant for the purposes of diagnosis. However, this analysis was limited to DNA-Seq only. Future work duplicating the analysis for RNA-Seq may be useful for informing development on single-nucleic acid tests. In Chapters 4.2 and 4.3 we demonstrated that RNA-Seq alone could achieve reasonable diagnostic performance. Therefore, characterizing the microbial differences by RNA-Seq may be important for ensuring diagnostics for LRTI use the fluid type that achieves the greatest sensitivity.

Additionally, the work presented in Chapter 2 doesn't account for host transcriptional differences between the fluids. It is possible that the cell types present in the lower airway sample (mBAL) would differ from those that accumulate in the tracheal samples. This may modulate the transcriptional signatures obtained from each sample type. Given our success in developing a host gene expression classifier for LRTI from TA samples in Chapter 3, we do expect that either fluid type would be suitable for host classification. The challenge to be addressed would be for the case where a model trained on one sample type (i.e. TA) is applied to classify a different sample type (i.e. mBAL). Characterizing possible limitations around transcriptional profiles between the two fluid types would inform their application.

As discussed in Chapter 3, the cohort samples size used for model derivation was a significant limitation. The cohort size was suitable for a proof-of-principle exploration of combined host and microbial mNGS as a diagnostic for LRTI. However, future studies will benefit from including more patients. Specifically, the small sample size precluded robust characterization of the host gene expression classifier developed in Chapter 3 against a distinct validation cohort. The learning curve analysis suggested that including more samples in the training set may further improve the model performance. Undoubtedly, inclusion of more samples would enhance the performance of the host classifier in distinct cohorts.

The combined metric presented in FIGURE 3.6 shows the promising ability to rule out infections, but this metric also inherits all the challenges of a small cohort. Future studies with larger samples sizes may enable a more rigorous analysis of methods for combining pathogen, host, and microbiome-based metrics into a comprehensive evaluation of patient status. Additionally, future studies may also begin to assess factors that are common between patients for whom the pathogen-based metrics of infection differ from their host transcriptome or microbiome-based metrics (FIGURE 3.6). This could inform stronger recommendations for ruling out infection. Given the high stakes for removing antimicrobial treatment in the face of uncertainty, a greater ability to rule out infection would enable more rapid adoption of these methods into clinical practice.

6.2 FUTURE DIRECTIONS FOR COMBINED HOST AND MICROBE mNGS AS A DIAGNOSTIC FOR INFECTIONS

By applying the methods derived in Chapter 3 to a variety of disease contexts (Chapter 4), we observed several opportunities for improving the algorithms in future studies. With regards to the LRM, it is possible that future work on feature engineering strategies may provide additional benefit by simplifying or improving the performance of the model. For example, the initial model was derived using both RNA and DNA sequencing¹¹⁶. But, we have shown that similar results can be achieved using a simplified model with only RNA (Chapter 4.3). As sequencing technologies continue to develop, analytical methods may adapt to incorporate new information. The addition of ERCCs to the library preparation step¹²⁸ in Chapter 4.3 enabled the back-calculation of total input RNA and showed that CSF from infected samples has significantly greater input RNA than from samples without an infection. For the purposes of Chapter 4.3, we implemented a simple threshold based on the total input in the water controls. Future studies could incorporate this feature directly into the LRM, potentially improving the ability to distinguish between patients with infection versus those without.

Throughout the *Application Notes*, we focused mostly on applying the LRM for pathogen identification. However, applying the host classifier and microbiome biomarkers derived in Chapter 3 to these datasets may produce additional insight into patient status. The differences in fluid type will likely influence the host transcriptional signatures. Thus, the model derived in Chapter 3 may not be directly applicable, but future studies testing this model would be required to verify the assumption. In the case that the host classifier

is not directly applicable, future studies deriving host gene expression signatures for each fluid type may be useful for contextualizing the pathogens identified by the LRM.

Investigation of host response in each of the cohorts evaluated in the *Application Notes* section would also enable a more rigorous analysis of methods for combining pathogen, host, and microbiome-based metrics into a comprehensive evaluation of patient status across diseases. While the model specifics may differ across diseases (i.e. the genes associated with meningeal infections may differ from those associated with respiratory infections), the methods for integrating these factors would likely have applications across a broad range of infectious diseases. We have already observed this by applying models derived for LRTI diagnosis to identify pathogens implicated in meningeal infections (Chapter 4.3).

6.3 FUTURE EXPANSION OF METHODS FOR CLASSIFICATION WITH NOISY LABELS

Throughout the process of developing methods for mNGS data interpretation, the importance of an accurate gold standard for labeling data became exceedingly clear. However, the poor sensitivity of current infectious disease diagnostics is the driving motivation for developing new diagnostics, thus creating a catch-22 situation. We investigated two methods for combating label noise caused by the lack of a robust gold standard – first, through label-noise robust classification algorithms and second, through unsupervised analysis using variational autoencoders.

The large number of features and small sample sizes characteristic of transcriptomic datasets challenged the label-noise robust approaches. It is possible that the methods

investigated here for robust classification and *a priori* feature selection are heavily influenced by label noise and that other approaches would be better suited for this purpose. More work investigating feature selection techniques may improve the performance. But, evaluation with larger datasets would likely provide a similar benefit to model performance.

We again observed the need for more training data in our exploration of variational autoencoders. Initial analyses suggest VAEs may provide a promising approach for direct integration of host and microbial data, specifically with the goal of learning about the biology of infections. However, the model requires more training data to reach its full potential. As mNGS technology expands to a greater diversity of regions, integration of datasets will yield the power required to provide a more in-depth analysis of the features learned by the VAE.

6.4 CONCLUDING REMARKS

In summary, my thesis work has contributed to the interpretation of mNGS data by developing algorithms that take into account many aspects of infectious disease response. Altogether, these methods enable a more complete view of a patient's response to an infection or other inflammatory process. In particular, I have evaluated the use of combined host and microbial mNGS for diagnosis of LRTIs. I have then extended the use of these algorithms by applying them in a variety of other disease contexts. These include diagnosis of LRTI in different patient cohorts with respiratory failure, diagnosis of meningitis, and a few small extensions for analysis of host-of-origin and antimicrobial resistance profiles. Each of these aspects will be critical for understanding infectious

diseases globally. The role of an imperfect gold standard in the development of a diagnostic algorithm became clear through the development process. As mNGS technology spreads to regions where diagnostic resources are limited, the differences in gold standards are important to consider. Based on these challenges, I provide initial exploration of algorithms that may be used to extend mNGS data biomarker development in the context of noisy labels. It is my hope, that this work may be part of the foundation for globally integrated diagnostics for infections.

7 Appendix of External Electronic Resources

Below is a reference of all external datasets, raw sequencing data, code, and supplementary tables available in public repositories. Resources are ordered according to their relevant chapters and subsections are indicated where necessary.

Chapter 2

Description	Source / Filename
Raw DNA-Seq .fastq files from paired mBAL and TA samples, filtered to contain only the microbial sequences.	NCBI BioProject Accession ID PRJNA445982
Supplemental table , listed as TABLE 2.4 in the above manuscript. The top five most abundant microbes by reads per million reads mapped for each sample type (mBAL and TA) for each patient.	dx.doi.org/10.17504/ protocols.io.wqnf-dve Filename: Table S1
Code used for analysis of mBAL versus TA microbiome differences in subjects with and without pneumonia. The code is contained in an R markdown file.	dx.doi.org/10.17504/ protocols.io.wqnf-dve Filename: mBALvTA_analysis_datafiles_R_markdown.zip

Chapter 3

Description	Source / Filename
Raw RNA- and DNA-Seq .fastq files for 92 TA samples from patients with infectious or non-infectious causes of acute respiratory failure. Raw .fastq files were filtered to contain only the microbial sequences.	NCBI SRA Accession ID SRP139967
Supplementary tables referenced in TABLES 3.2 through 3.14. For detailed descriptions of table identity to PNAS Dataset identity, see Chapter 3.	https://www.pnas.org/content/115/52/E12353/tab-figures-data Filename: Dataset S01 through S09

Description	Source / Filename
Code used for combined host and microbe analysis, including .py and .R scripts. The .py scripts develop the LRM and RBM for distinguishing putative pathogens from commensals. The .R scripts implement the host gene expression classifier and microbiome analyses.	https://github.com/DeRisi-Lab/Host-MicrobeLRTI

Chapter 4

Description	Source / Filename
Section 4.1 Raw host transcriptome gene counts after STAR alignment, from mBAL samples collected from HCT patients with acute respiratory failure.	https://datadryad.org/resource/doi:10.5061/dryad.800tj Filename: Table S6.csv
Section 4.2 Raw RNA-Seq .fastq files from BAL samples collected from pediatric HCT recipients with acute respiratory failure.	dbGAP https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001684.v1.p1
Section 4.3 Raw RNA-Seq .fastq files for CSF samples from patients with infectious or non-infectious causes of meningitis, in Dhaka Bangladesh. Raw .fastq files were filtered to contain only the microbial sequences.	NCBI SRA Accession ID PRJNA516582
Section 4.4 Raw DNA-Seq .fastq files from heart biopsy tissue in single-patient case study.	dbGAP http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001336.v1.p1
Section 4.5 Raw RNA- and DNA-Seq .fastq files; stool samples	NCBI SRA Accession ID SUB4474900

Chapter 5

Description	Source / Filename
Publicly available test dataset published by Ahn, SH. et al. Peripheral blood microarray data from mice challenged with <i>S. aureus</i> , <i>E. coli</i> , or no infection.	GEO Database https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33341
Publicly available test dataset published by Suarez, NM., et al. Whole blood microarray samples from patients with bacterial, viral, bacterial-viral coinfection, and healthy controls	GEO Database https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60244
The original code for rLR dataset simulation and benchmarking as published in manuscripts by Bootkrajang et al.	http://www.cs.science.cmu.ac.th/person/jakramate/#code
Code adapting the rLR algorithms for data simulation and evaluation of the approach in the context of host gene expression analysis. The algorithms implemented in the above MATLAB code were adapted for use with R.	https://github.com/katrinakalantar/NMLHC
The original code from the Greene Lab Tybalt VAE. This code was used for initial benchmarking and then modified for application to combined host and microbial mNGS data.	https://github.com/greenelab/tybalt
Code adapted to apply Tybalt VAE for analysis of host and microbial mNGS data. The jupyter notebooks for the initial Tybalt analysis were modified. The first script generates the combined data matrix from host gene counts and microbial taxonomic data, which is then normalized in the second script. This data is then used for training the VAE. Finally, there are two scripts for evaluating the model and its learned features.	https://github.com/katrinakalantar/ApplyTybaltVAE

Bibliography

1. Cunha BA. Historical aspects of infectious diseases, part I. *Infect Dis Clin North Am*. 2004;18(1):XI-V. doi:10.1016/S0891-5520(03)00098-9
2. Pappas G, Kiriakou IJ, Falagas ME. Insights into infectious disease in the era of Hippocrates. *Int J Infect Dis*. 2008;12(4):347-350. doi:10.1016/J.IJID.2007.11.003
3. Brachman PS. Infectious diseases—past, present, and future. *Int J Epidemiol*. 2003;32(5):684-686. doi:10.1093/ije/dyg282
4. Young VB. The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ*. 2017;356:j831. doi:10.1136/bmj.j831
5. Tremaroli V, Bäckhed F. Functional interactions between the gut microbiota and host metabolism. *Nature*. 2012;489(7415):242-249. doi:10.1038/nature11552
6. The top 10 causes of death. World Health Organization. <http://www.who.int/mediacentre/factsheets/fs310/en/>. Published 2017. Accessed March 20, 2017.
7. Hansen V, Oren E, Dennis LK, Brown HE. Infectious Disease Mortality Trends in the United States, 1980-2014. *Jama*. 2016;316(20):2149. doi:10.1001/jama.2016.12423
8. GBD 2015 LRI Collaborators C, Forouzanfar M, Rao PC, et al. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Infect Dis*. 2017;17(11):1133-1161. doi:10.1016/S1473-3099(17)30396-1

9. Zar HJ, Madhi SA, Aston SJ, Gordon SB. Pneumonia in low and middle income countries: progress and challenges. *Thorax*. 2013;68(11):1052-1056.
doi:10.1136/THORAXJNL-2013-204247
10. de Steenhuijsen Piters WAA, Sanders EAM, Bogaert D. The role of the local microbial ecosystem in respiratory health and disease. *Philos Trans R Soc Lond B Biol Sci*. 2015;370(1675). doi:10.1098/rstb.2014.0294
11. Bogaert D, de Groot R, Hermans P. Streptococcus pneumoniae colonisation: the key to pneumococcal disease. *Lancet Infect Dis*. 2004;4(3):144-154.
doi:10.1016/S1473-3099(04)00938-7
12. Newton AH, Cardani A, Braciale TJ. The host immune response in respiratory virus infection: balancing virus clearance and immunopathology. *Semin Immunopathol*. 2016;38(4):471-482. doi:10.1007/s00281-016-0558-0
13. Russell CD, Unger SA, Walton M, Schwarze J. The Human Immune Response to Respiratory Syncytial Virus Infection. *Clin Microbiol Rev*. 2017;30(2):481-502.
doi:10.1128/CMR.00090-16
14. Pneumonia. National Heart, Lung, and Blood Institute (NHLBI).
<https://www.nhlbi.nih.gov/health-topics/pneumonia>. Accessed January 7, 2019.
15. Mizgerd JP. Lung infection--a public health priority. *PLoS Med*. 2006;3(2):e76.
doi:10.1371/journal.pmed.0030076
16. Marshall DC, Goodson RJ, Xu Y, et al. Trends in mortality from pneumonia in the Europe union: a temporal analysis of the European detailed mortality database between 2001 and 2014. *Respir Res*. 2018;19(1):81. doi:10.1186/s12931-018-0781-4

17. Raymond Zunt J, Kassebaum NJ, Blake N, et al. Global, regional, and national burden of meningitis, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 2018;17(12):1061-1082. doi:10.1016/S1474-4422(18)30387-9
18. Skvarc M, Stubljär D, Rogina P, Kaasch AJ. Non-culture-based methods to diagnose bloodstream infection: Does it work? *Eur J Microbiol Immunol (Bp).* 2013;3(2):97-104. doi:10.1556/EuJMI.3.2013.2.2
19. Song JY, Eun BW, Nahm MH. Diagnosis of pneumococcal pneumonia: current pitfalls and the way forward. *Infect Chemother.* 2013;45(4):351-366. doi:10.3947/ic.2013.45.4.351
20. Ranzani OT, Prina E, Menéndez R, et al. New Sepsis Definition (Sepsis-3) and Community-acquired Pneumonia Mortality. A Validation and Clinical Decision-Making Study. *Am J Respir Crit Care Med.* 2017;196(10):1287-1297. doi:10.1164/rccm.201611-2262OC
21. Jain S, Self WH, Wunderink RG, et al. Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. *N Engl J Med.* 2015;372(9):1507-1514. doi:10.1056/NEJMoa1500245
22. Bloch KC, Glaser C. Diagnostic approaches for patients with suspected encephalitis. *Curr Infect Dis Rep.* 2007;9(4):315-322. <http://www.ncbi.nlm.nih.gov/pubmed/17618552>. Accessed January 10, 2019.
23. Glaser CA, Gilliam S, Schnurr D, et al. In Search of Encephalitis Etiologies: Diagnostic Challenges in the California Encephalitis Project, 1998–2000. *Clin Infect Dis.* 2003;36(6):731-742. doi:10.1086/367841

24. Glaser CA, Honarmand S, Anderson LJ, et al. Beyond Viruses: Clinical Profiles and Etiologies Associated with Encephalitis. *Clin Infect Dis*. 2006;43(12):1565-1577. doi:10.1086/509330
25. Wetterstrand KA. DNA Sequencing Costs: Data - National Human Genome Research Institute (NHGRI). <https://www.genome.gov/27541954/dna-sequencing-costs-data/>. Accessed January 9, 2019.
26. Simner PJ, Miller S, Carroll KC. Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. *Clin Infect Dis*. 2018;66(5):778-788. doi:10.1093/cid/cix881
27. Bragg L, Tyson GW. Metagenomics Using Next-Generation Sequencing. In: Humana Press, Totowa, NJ; 2014:183-201. doi:10.1007/978-1-62703-712-9_15
28. Wilson MR, O'Donovan BD, Gelfand JM, et al. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. *JAMA Neurol*. 2018;75(8):947. doi:10.1001/jamaneurol.2018.0463
29. Doan T, Wilson MR, Crawford ED, et al. Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens. *Genome Med*. 2016;8(1):90. doi:10.1186/s13073-016-0344-6
30. Pavia AT. Viral infections of the lower respiratory tract: old viruses, new viruses, and the role of diagnosis. *Clin Infect Dis*. 2011;52 Suppl 4(Suppl 4):S284-9. doi:10.1093/cid/cir043
31. Tsalik EL, Henao R, Nichols M, et al. Host gene expression classifiers diagnose acute respiratory illness etiology. *Sci Transl Med*. 2016;8(322):322ra11-322ra11. doi:10.1126/scitranslmed.aad6873

32. Suarez NM, Bunsow E, Falsey AR, Walsh EE, Mejias A, Ramilo O. Superiority of transcriptional profiling over procalcitonin for distinguishing bacterial from viral lower respiratory tract infections in hospitalized adults. *J Infect Dis.* 2015;212(2):213-222. doi:10.1093/infdis/jiv047
33. Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Sci Transl Med.* 2016;8(346):346ra91. doi:10.1126/scitranslmed.aaf7165
34. Mejias A, Dimo B, Suarez NM, et al. Whole Blood Gene Expression Profiles to Assess Pathogenesis and Disease Severity in Infants with Respiratory Syncytial Virus Infection. *PLoS Med.* 2013;10(11). doi:10.1371/journal.pmed.1001549
35. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics.* 2008;9(Suppl 1):S13. doi:10.1186/1471-2164-9-S1-S13
36. Saunders N, Zambon M, Sharp I, et al. Guidance on the development and validation of diagnostic tests that depend on nucleic acid amplification and detection. *J Clin Virol.* 2013;56(3):344-354. doi:10.1016/J.JCV.2012.11.013
37. Gold R, Reichman M, Greenberg E, et al. Developing a new reference standard: is validation necessary? *Acad Radiol.* 2010;17(9):1079-1082. doi:10.1016/j.acra.2010.05.021
38. Zhang C, Wu C, Blanzieri E, et al. Methods for labeling error detection in microarrays based on the effect of data perturbation on the regression model. *Bioinformatics.* 2009;25(20):2708-2714. doi:10.1093/bioinformatics/btp478
39. Malossini A, Blanzieri E, Ng RT. Detecting potential labeling errors in microarrays

- by data perturbation. *Bioinformatics*. 2006;22(17):2114-2121.
doi:10.1093/bioinformatics/btl346
40. Malossini A., Blanzieri E. NRT. Assessment of SVM reliability for microarray data analysis. In: *Dutch–Belgian Conference on Machine Learning*. ; 2005.
 41. Xiao T, Xia T, Yang Y, Huang C, Wang X. *Learning from Massive Noisy Labeled Data for Image Classification*.
<http://www.ee.cuhk.edu.hk/~xgwang/papers/xiaoXYHWcvpr15.pdf>. Accessed January 15, 2019.
 42. Raykar VC, Yu S, Zhao LH, et al. *Learning From Crowds 1. Supervised Learning From Multiple Annotators/Experts*. Vol 11.; 2010. <https://www.mturk.com>. Accessed January 15, 2019.
 43. Bootkrajang J, Kabán A. Label-noise robust logistic regression and its applications. In: *Lecture Notes in Computer Science*. Vol 7523. ; 2012:143-158.
doi:10.1007/978-3-642-33460-3_15
 44. Boersma WG, Erjavec Z, van der Werf TS, de Vries-Hosper HG, Gouw ASH, Manson WL. Bronchoscopic diagnosis of pulmonary infiltrates in granulocytopenic patients with hematologic malignancies: BAL versus PSB and PBAL. *Respir Med*. 2007;101(2):317-325. doi:10.1016/j.rmed.2006.04.021
 45. Dickson RP, Erb-Downward JR, Freeman CM, et al. Bacterial Topography of the Healthy Human Lower Respiratory Tract. *MBio*. 2017;8(1):e02287-16.
doi:10.1128/mBio.02287-16
 46. Berton DC, Kalil AC, Teixeira PJZ. Quantitative versus qualitative cultures of respiratory secretions for clinical outcomes in patients with ventilator-associated

- pneumonia. *Cochrane Database Syst Rev*. October 2014.
doi:10.1002/14651858.CD006482.pub4
47. Group TCCCT. A Randomized Trial of Diagnostic Techniques for Ventilator-Associated Pneumonia. *N Engl J Med*. 2006;355(25):2619-2630.
doi:10.1056/NEJMoa052904
 48. Kalil AC, Metersky ML, Klompas M, et al. Management of Adults With Hospital-acquired and Ventilator-associated Pneumonia: 2016 Clinical Practice Guidelines by the Infectious Diseases Society of America and the American Thoracic Society. *Clin Infect Dis*. 2016;63(5):e61-e111. doi:10.1093/cid/ciw353
 49. CDC, Nceid, DHQP. *CDC/NHSN Surveillance Definitions for Specific Types of Infections.*; 2019.
https://www.cdc.gov/nhsn/pdfs/pscmanual/17pscnosinfdef_current.pdf. Accessed January 2, 2019.
 50. Langelier C, Zinter MS, Kalantar K, et al. Metagenomic Sequencing Detects Respiratory Pathogens in Hematopoietic Cellular Transplant Patients. *Am J Respir Crit Care Med*. 2018;197(4):524-528. doi:10.1164/rccm.201706-1097LE
 51. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
 52. Ruby JG, Bellare P, Derisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)*. 2013;3(5):865-880. doi:10.1534/g3.113.005967
 53. Langmead. Bowtie2. 2013;9(4):357-359. doi:10.1038/nmeth.1923.Fast
 54. Oksanen J, Blanchet FG, Friendly M, et al. *Title Community Ecology Package.*;

2018. <https://cran.ism.ac.jp/web/packages/vegan/vegan.pdf>. Accessed January 2, 2019.
55. Abreu NA, Nagalingam NA, Song Y, et al. Sinus microbiome diversity depletion and *Corynebacterium tuberculo*stearicum enrichment mediates rhinosinusitis. *Sci Transl Med*. 2012;4(151):151ra124. doi:10.1126/scitranslmed.3003783
56. Kitsios GD, Fitch A, Manatakis D V., et al. Respiratory Microbiome Profiling for Etiologic Diagnosis of Pneumonia in Mechanically Ventilated Patients. *Front Microbiol*. 2018;9:1413. doi:10.3389/fmicb.2018.01413
57. Kelly BJ, Imai I, Bittinger K, et al. Composition and dynamics of the respiratory tract microbiome in intubated patients. *Microbiome*. 2016;4:7. doi:10.1186/s40168-016-0151-8
58. FastStats - Leading Causes of Death. <https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm>. Accessed January 2, 2019.
59. el Bcheraoui C, Mokdad AH, Dwyer-Lindgren L, et al. Trends and Patterns of Differences in Infectious Disease Mortality Among US Counties, 1980-2014. *JAMA*. 2018;319(12):1248. doi:10.1001/jama.2018.2089
60. Zaas AK, Garner BH, Tsalik EL, Burke T, Woods CW, Ginsburg GS. The current epidemiology and clinical decisions surrounding acute respiratory infections. *Trends Mol Med*. 2014;20(10):579-588. doi:10.1016/J.MOLMED.2014.08.001
61. Wilson MR, Naccache SN, Samayoa E, et al. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *N Engl J Med*. 2014;370(25):2408-2417. doi:10.1056/NEJMoa1401268
62. Leffler DA, Lamont JT. *Clostridium difficile* Infection. Longo DL, ed. *N Engl J Med*.

- 2015;372(16):1539-1548. doi:10.1056/NEJMra1403772
63. Bibby K. Metagenomic identification of viral pathogens. *Trends Biotechnol.* 2013;31(5):275-279. doi:10.1016/J.TIBTECH.2013.01.016
64. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis.* 2012;6(2):e1485. doi:10.1371/journal.pntd.0001485
65. Fischer N, Indenbirken D, Meyer T, et al. Evaluation of Unbiased Next-Generation Sequencing of RNA (RNA-seq) as a Diagnostic Method in Influenza Virus-Positive Respiratory Samples. *J Clin Microbiol.* 2015;53(7):2238-2250. doi:10.1128/JCM.02495-14
66. Graf EH, Simmon KE, Tardif KD, et al. Unbiased Detection of Respiratory Viruses by Use of RNA Sequencing-Based Metagenomics: a Systematic Comparison to a Commercial PCR Panel. *J Clin Microbiol.* 2016;54(4):1000-1007. doi:10.1128/JCM.03060-15
67. Wilson MR, Shanbhag NM, Reid MJ, et al. Diagnosing Balamuthia mandrillaris Encephalitis With Metagenomic Deep Sequencing. *Ann Neurol.* 2015;78(5):722-730. doi:10.1002/ana.24499
68. Naccache SN, Federman S, Veeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* 2014;24(7):1180-1192. doi:10.1101/gr.171934.113
69. Tsalik EL, McClain M, Zaas AK. Moving toward prime time: Host signatures for

- diagnosis of respiratory infections. *J Infect Dis.* 2015;212(2):173-175.
doi:10.1093/infdis/jiv032
70. Panzer AR, Lynch S V., Langelier C, et al. Lung Microbiota Is Related to Smoking Status and to Development of Acute Respiratory Distress Syndrome in Critically Ill Trauma Patients. *Am J Respir Crit Care Med.* 2018;197(5):621-631.
doi:10.1164/rccm.201702-0441OC
 71. Morris A, Beck JM, Schloss PD, et al. Comparison of the Respiratory Microbiome in Healthy Nonsmokers and Smokers. *Am J Respir Crit Care Med.* 2013;187(10):1067-1075. doi:10.1164/rccm.201210-1913OC
 72. Segal LN, Clemente JC, Tsay J-CJ, et al. Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. *Nat Microbiol.* 2016;1(5):16031. doi:10.1038/nmicrobiol.2016.31
 73. Heinonen S, Jartti T, Garcia C, et al. Rhinovirus Detection in Symptomatic and Asymptomatic Children: Value of Host Transcriptome Analysis. *Am J Respir Crit Care Med.* 2015;193(7):rccm.201504-0749OC. doi:10.1164/rccm.201504-0749OC
 74. Wertheim HF, Melles DC, Vos MC, et al. The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect Dis.* 2005;5(12):751-762.
doi:10.1016/S1473-3099(05)70295-4
 75. McCullers JA. The co-pathogenesis of influenza viruses with bacteria in the lung. *Nat Rev Microbiol.* 2014;12(4):252-262. doi:10.1038/nrmicro3231
 76. Magill SS, Edwards JR, Bamberg W, et al. Multistate Point-Prevalence Survey of Health Care–Associated Infections. *N Engl J Med.* 2014;370(13):1198-1208.

doi:10.1056/NEJMoa1306801

77. Mandell LA, Wunderink RG, Anzueto A, et al. Infectious Diseases Society of America/American Thoracic Society Consensus Guidelines on the Management of Community-Acquired Pneumonia in Adults. *Clin Infect Dis*. 2007;44(Supplement_2):S27-S72. doi:10.1086/511159
78. Cillóniz C, Civljak R, Nicolini A, Torres A. Polymicrobial community-acquired pneumonia: An emerging entity. *Respirology*. 2016;21(1):65-75. doi:10.1111/resp.12663
79. Pabbaraju K, Wong S, Wong A, May-Hadford J, Tellier R, Fonseca K. Detection of influenza C virus by a real-time RT-PCR assay. *Influenza Other Respi Viruses*. 2013;7(6):954-960. doi:10.1111/irv.12099
80. Dewhirst FE, Chen T, Izard J, et al. The human oral microbiome. *J Bacteriol*. 2010;192(19):5002-5017. doi:10.1128/JB.00542-10
81. Chen C, Shen T, Tian F, et al. New microbiota found in sputum from patients with community-acquired pneumonia. *Acta Biochim Biophys Sin (Shanghai)*. 2013;45(12):1039-1048. doi:10.1093/abbs/gmt116
82. Ichinohe T, Pang IK, Kumamoto Y, et al. Microbiota regulates immune defense against respiratory tract influenza A virus infection. *Proc Natl Acad Sci U S A*. 2011;108(13):5354-5359. doi:10.1073/pnas.1019378108
83. Dickson RP, Erb-Downward JR, Prescott HC, et al. Analysis of culture-dependent versus culture-independent techniques for identification of bacteria in clinically obtained bronchoalveolar lavage fluid. *J Clin Microbiol*. 2014;52(10):3605-3613. doi:10.1128/JCM.01028-14

84. Birtel J, Walser J-C, Pichon S, Bürgmann H, Matthews B. Estimating Bacterial Diversity for Ecological Studies: Methods, Metrics, and Assumptions. Larsen P, ed. *PLoS One*. 2015;10(4):e0125356. doi:10.1371/journal.pone.0125356
85. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol Monogr*. 1957;27(4):325-349. doi:10.2307/1942268
86. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*. 2009;37(Web Server):W305-W311. doi:10.1093/nar/gkp427
87. Macian F. NFAT proteins: key regulators of T-cell development and function. *Nat Rev Immunol*. 2005;5(6):472-484. doi:10.1038/nri1632
88. Fu M, Blakeshear PJ. RNA-binding proteins in immune regulation: a focus on CCH zinc finger proteins. *Nat Rev Immunol*. 2017;17(2):130-143. doi:10.1038/nri.2016.129
89. Biswas K, Wagner Mackenzie B, Waldvogel-Thurlow S, et al. Differentially Regulated Host Proteins Associated with Chronic Rhinosinusitis Are Correlated with the Sinonasal Microbiome. *Front Cell Infect Microbiol*. 2017;7:504. doi:10.3389/fcimb.2017.00504
90. Stewart CR, Stuart LM, Wilkinson K, et al. CD36 ligands promote sterile inflammation through assembly of a Toll-like receptor 4 and 6 heterodimer. *Nat Immunol*. 2010;11(2):155-161. doi:10.1038/ni.1836
91. Cohen TS, Jones-Nelson O, Hotz M, et al. *S. aureus* blocks efferocytosis of neutrophils by macrophages through the activity of its virulence factor alpha toxin. *Sci Rep*. 2016;6:35466. doi:10.1038/srep35466

92. Baranano DE, Rao M, Ferris CD, Snyder SH. Biliverdin reductase: a major physiologic cytoprotectant. *Proc Natl Acad Sci U S A*. 2002;99(25):16093-16098. doi:10.1073/pnas.252626999
93. Leidi M, Mariotti M, Maier JAM. EDF-1 contributes to the regulation of nitric oxide release in VEGF-treated human endothelial cells. *Eur J Cell Biol*. 2010;89(9):654-660. doi:10.1016/J.EJCB.2010.05.001
94. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2016;21(3):193-201. doi:10.1016/j.molmed.2014.11.008.Mitochondria
95. Currie CJ, Berni E, Jenkins-Jones S, et al. Antibiotic treatment failure in four common infections in UK primary care 1991-2012: longitudinal analysis. *BMJ*. 2014;349:g5493. doi:10.1136/BMJ.G5493
96. Jain S, Williams DJ, Arnold SR, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med*. 2015;372(9):835-845. doi:10.1056/NEJMoa1405870
97. Zaas a. K, Chen M, Varkey J, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe*. 2009;6(3):207–217. doi:10.1016/j.chom.2009.07.006.Gene
98. Dobbin KK, Zhao Y, Simon RM. How Large a Training Set is Needed to Develop a Classifier for Microarray Data? *Clin Cancer Res*. 2008;14(1):108-114. doi:10.1158/1078-0432.CCR-07-0443
99. Gu W, Crawford ED, O'Donovan BD, et al. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species

- in sequencing libraries and molecular counting applications. *Genome Biol.* 2016;17(1):41. doi:10.1186/s13059-016-0904-5
100. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-359. doi:10.1038/nmeth.1923
 101. R Core Team. No Title. R: A language and environment for statistical computing.
 102. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550. doi:10.1186/s13059-014-0550-8
 103. Fine MJ, Auble TE, Yealy DM, et al. A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia. *N Engl J Med.* 1997;336(4):243-250. doi:10.1056/NEJM199701233360402
 104. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak.* 2012;12(1):8. doi:10.1186/1472-6947-12-8
 105. Meek C, Thiesson B, Heckerman D. *The Learning-Curve Sampling Method Applied to Model-Based Clustering.* Vol 2.; 2002.
<http://www.jmlr.org/papers/volume2/meek02a/meek02a.pdf>. Accessed January 3, 2019.
 106. Sethi S, Murphy TF. Infection in the Pathogenesis and Course of Chronic Obstructive Pulmonary Disease. *N Engl J Med.* 2008;359(22):2355-2365. doi:10.1056/NEJMra0800353
 107. Fishman JA. Infection in Organ Transplantation. *Am J Transplant.* 2017;17(4):856-879. doi:10.1111/ajt.14208

108. Hematopoietic Cell Transplantation | Cancer Network.
<http://www.cancernetwork.com/cancer-management/hematopoietic-cell-transplantation>. Accessed January 8, 2019.
109. Gyurkocza B, Sandmaier BM. Conditioning regimens for hematopoietic cell transplantation: one size does not fit all. *Blood*. 2014;124(3):344-353.
doi:10.1182/blood-2014-02-514778
110. Srinivasan A, McLaughlin L, Wang C, et al. Early infections after autologous hematopoietic stem cell transplantation in children and adolescents: the St. Jude experience. *Transpl Infect Dis*. 2014;16(1):90-97. doi:10.1111/tid.12165
111. Servais S, Lengline E, Porcher R, et al. Long-Term Immune Reconstitution and Infection Burden after Mismatched Hematopoietic Stem Cell Transplantation. *Biol Blood Marrow Transplant*. 2014;20(4):507-517. doi:10.1016/j.bbmt.2014.01.001
112. SIMPSON EH. Measurement of Diversity. *Nature*. 1949;163(4148):688-688.
doi:10.1038/163688a0
113. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA -seq aligner. *Bioinformatics*. 2013;1-7. doi:doi: 10.1093/bioinformatics/bts635
114. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods*. 2013;10(12):1200-1202.
doi:10.1038/nmeth.2658
115. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst*. 2015;1(6):417-425. doi:10.1016/j.cels.2015.12.004
116. Langelier C, Kalantar KL, Moazed F, et al. Integrating host response and

- unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. *Proc Natl Acad Sci*. 2018;115(52):E12353-E12362.
doi:10.1073/PNAS.1809700115
117. Zinter MS, Dvorak CC, Mayday MY, et al. Pulmonary Metagenomic Sequencing Suggests Missed Infections in Immunocompromised Children. *Clin Infect Dis*. September 2018. doi:10.1093/cid/ciy802
 118. Bittinger K, Charlson ES, Loy E, et al. Improved characterization of medically relevant fungi in the human respiratory tract using next-generation sequencing. *Genome Biol*. 2014;15(10):487. doi:10.1186/s13059-014-0487-y
 119. White PL, Bretagne S, Klingspor L, et al. Aspergillus PCR: one step closer to standardization. *J Clin Microbiol*. 2010;48(4):1231-1240. doi:10.1128/JCM.01767-09
 120. Khan S, Priti S, Ankit S. Bacteria Etiological Agents Causing Lower Respiratory Tract Infections and Their Resistance Patterns. *Iran Biomed J*. 2015;19(4):240-246. doi:10.7508/IBJ.2015.04.008
 121. Seo S, Waghmare A, Scott EM, et al. Human rhinovirus detection in the lower respiratory tract of hematopoietic cell transplant recipients: association with mortality. *Haematologica*. 2017;102(6):1120-1130.
doi:10.3324/haematol.2016.153767
 122. Andres-Terre M, McGuire HM, Pouliot Y, et al. Integrated, Multi-cohort Analysis Identifies Conserved Transcriptional Signatures across Multiple Respiratory Viruses. *Immunity*. 2015;43(6):1199-1211. doi:10.1016/j.immuni.2015.11.003
 123. Hoffman O, Weber RJ. Pathophysiology and treatment of bacterial meningitis.

- Ther Adv Neurol Disord.* 2009;2(6):1-7. doi:10.1177/1756285609337975
124. GBD 2017 Causes of Death Collaborators GA, Abate D, Abate KH, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet (London, England)*. 2018;392(10159):1736-1788. doi:10.1016/S0140-6736(18)32203-7
 125. Ramesh A, Nakielnny S, Hsu J, et al. Etiology of fever in Ugandan children: identification of microbial pathogens using metagenomic next-generation sequencing and IDseq, a platform for unbiased metagenomic analysis. *bioRxiv*. August 2018:385005. doi:10.1101/385005
 126. Nieuwenhuijse DF, Koopmans MPG. Metagenomic Sequencing for Surveillance of Food- and Waterborne Viral Diseases. *Front Microbiol.* 2017;8:230. doi:10.3389/fmicb.2017.00230
 127. Evans JA, Shim J-M, Ioannidis JPA. Attention to local health burden and the global disparity of health research. *PLoS One*. 2014;9(4):e90147. doi:10.1371/journal.pone.0090147
 128. Mayday MY, Khan LM, Chow ED, Zinter MS, DeRisi JL. Miniaturization and optimization of 384-well compatible RNA sequencing library preparation. Thomas T, ed. *PLoS One*. 2019;14(1):e0206194. doi:10.1371/journal.pone.0206194
 129. IDSeq Portal. www.idseq.net.
 130. Lanciotti RS, Kosoy OL, Laven JJ, et al. Chikungunya virus in US travelers returning from India, 2006. *Emerg Infect Dis.* 2007;13(5):764-767. doi:10.3201/eid1305.070015

131. Welniak LA, Blazar BR, Murphy WJ. Immunobiology of Allogeneic Hematopoietic Stem Cell Transplantation. *Annu Rev Immunol*. 2007;25(1):139-170.
doi:10.1146/annurev.immunol.25.022106.141606
132. Ferrara JL, Levine JE, Reddy P, Holler E. Graft-versus-host disease. *Lancet*. 2009;373(9674):1550-1561. doi:10.1016/S0140-6736(09)60237-3
133. Nakasone H, Remberger M, Tian L, et al. Risks and benefits of sex-mismatched hematopoietic cell transplantation differ according to conditioning strategy. *Haematologica*. 2015;100(11):1477-1485. doi:10.3324/haematol.2015.125294
134. Graft-Versus-Host Disease | Leukemia and Lymphoma Society.
<https://www.lls.org/treatment/types-of-treatment/stem-cell-transplantation/graft-versus-host-disease>. Accessed January 8, 2019.
135. Harris AC, Ferrara JLM, Levine JE. Advances in predicting acute GVHD. *Br J Haematol*. 2013;160(3):288-302. doi:10.1111/bjh.12142
136. Flowers MED, Inamoto Y, Carpenter PA, et al. Comparative analysis of risk factors for acute graft-versus-host disease and for chronic graft-versus-host disease according to National Institutes of Health consensus criteria. *Blood*. 2011;117(11):3214-3219. doi:10.1182/blood-2010-08-302109
137. Jagasia M, Arora M, Flowers MED, et al. Risk factors for acute GVHD and survival after hematopoietic cell transplantation. *Blood*. 2012;119(1):296-307.
doi:10.1182/blood-2011-06-364265
138. Jacobsohn DA, Vogelsang GB. Acute graft versus host disease. *Orphanet J Rare Dis*. 2007;2:35. doi:10.1186/1750-1172-2-35
139. Zinter MS, Barrows BD, Ursell PC, et al. Extracorporeal life support survival in a

- pediatric hematopoietic cellular transplant recipient with presumed GvHD-related fulminant myocarditis. *Bone Marrow Transplant*. 2017;52(9):1330-1333.
doi:10.1038/bmt.2017.114
140. Smith KD, Young KE, Talbot CC, Schmeckpeper BJ. Repeated DNA of the human Y chromosome. *Development*. 1987;101 Suppl:77-92.
<http://www.ncbi.nlm.nih.gov/pubmed/2846258>. Accessed January 8, 2019.
 141. Legato MJ. *Principles of Gender-Specific Medicine : Gender in the Genomic Era*.
<https://www.sciencedirect.com/book/9780128035061/principles-of-gender-specific-medicine>. Accessed January 8, 2019.
 142. Trounson A, McDonald C. Stem Cell Therapies in Clinical Trials: Progress and Challenges. *Cell Stem Cell*. 2015;17(1):11-22. doi:10.1016/J.STEM.2015.06.007
 143. Zhang Z, Zhang Y, Gao F, et al. CRISPR/Cas9 Genome-Editing System in Human Stem Cells: Current Status and Future Prospects. *Mol Ther Nucleic Acids*. 2017;9:230-241. doi:10.1016/j.omtn.2017.09.009
 144. Appelbaum PC. 2012 and beyond: potential for the start of a second pre-antibiotic era? *J Antimicrob Chemother*. 2012;67(9):2062-2068. doi:10.1093/jac/dks213
 145. Arcilla MS, van Hattem JM, Haverkate MR, et al. Import and spread of extended-spectrum β -lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study. *Lancet Infect Dis*. 2017;17(1):78-85. doi:10.1016/S1473-3099(16)30319-X
 146. Kuenzli E, Jaeger VK, Frei R, et al. High colonization rates of extended-spectrum β -lactamase (ESBL)-producing Escherichia coli in Swiss Travellers to South Asia—a prospective observational multicentre cohort study looking at epidemiology,

- microbiology and risk factors. *BMC Infect Dis*. 2014;14(1):528. doi:10.1186/1471-2334-14-528
147. Hassing RJ, Alisma J, Arcilla MS, van Genderen PJ, Stricker BH, Verbon A. International travel and acquisition of multidrug-resistant Enterobacteriaceae: a systematic review. *Euro Surveill*. 2015;20(47). doi:10.2807/1560-7917.ES.2015.20.47.30074
 148. van Duijkeren E, Wielders CCH, Dierikx CM, et al. Long-term Carriage of Extended-Spectrum β -Lactamase–Producing *Escherichia coli* and *Klebsiella pneumoniae* in the General Population in The Netherlands. *Clin Infect Dis*. 2018;66(9):1368-1376. doi:10.1093/cid/cix1015
 149. David LA, Materna AC, Friedman J, et al. Host lifestyle affects human microbiota on daily timescales. *Genome Biol*. 2014;15(7):R89. doi:10.1186/gb-2014-15-7-r89
 150. Extended-spectrum beta-lactamases - UpToDate.
<https://www.uptodate.com/contents/extended-spectrum-beta-lactamases>.
Accessed January 19, 2019.
 151. Schultsz C, Geerlings S. Plasmid-Mediated Resistance in Enterobacteriaceae. *Drugs*. 2012;72(1):1-16. doi:10.2165/11597960-000000000-00000
 152. Carattoli A. Plasmids and the spread of resistance. *Int J Med Microbiol*. 2013;303(6-7):298-304. doi:10.1016/j.ijmm.2013.02.001
 153. Langelier C, Graves M, Kalantar KL, et al. Microbiome and Antimicrobial Resistance Gene Dynamics in International Travelers. *bioRxiv*. December 2018:506394. doi:10.1101/506394
 154. Inouye M, Dashnow H, Raven L-A, et al. SRST2: Rapid genomic surveillance for

- public health and hospital microbiology labs. *Genome Med.* 2014;6(11):90.
doi:10.1186/s13073-014-0090-6
155. Gupta SK, Padmanabhan BR, Diene SM, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother.* 2014;58(1):212-220. doi:10.1128/AAC.01310-13
 156. Kim S-Y. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics.* 2009;10(1):147.
doi:10.1186/1471-2105-10-147
 157. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S. Machine Learning and Its Applications to Biology. *PLoS Comput Biol.* 2007;3(6):e116.
doi:10.1371/journal.pcbi.0030116
 158. Shoham Y, Perrault R, Brynjolfsson E, Openai JC. *AI Index.*; 2017.
<http://aiindex.org/2017-report.pdf>. Accessed January 15, 2019.
 159. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science.* 2015;349(6245):255-260. doi:10.1126/science.aaa8415
 160. Nagelkerke N, Fidler V. Estimating a Logistic Discrimination Functions When One of the Training Samples Is Subject to Misclassification: A Maximum Likelihood Approach. Fernandez-Reyes D, ed. *PLoS One.* 2015;10(10):e0140718.
doi:10.1371/journal.pone.0140718
 161. Frénay B, Verleysen M. Classification in the presence of label noise: A survey. *IEEE Trans Neural Networks Learn Syst.* 2014;25(5):845-869.
doi:10.1109/TNNLS.2013.2292894
 162. Bootkrajang J, Kabán A. Classification of mislabelled microarrays using robust

- sparse logistic regression. *Bioinformatics*. 2013;29(7):870-877.
doi:10.1093/bioinformatics/btt078
163. Dua, D. and Karra Taniskidou E. UCI Machine Learning Repository. niversity of California, School of Information and Computer Science.
<http://archive.ics.uci.edu/ml>. Published 2017.
 164. Lu X, Chen D. Cancer classification through filtering progressive transductive support vector machine based on gene expression data. 2017;1864:30133.
doi:10.1063/1.4992918
 165. Lachenbruch PA. Discriminant Analysis When the Initial Samples Are Misclassified. *Technometrics*. 1966;8(4):657. doi:10.2307/1266637
 166. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1-22.
<http://www.ncbi.nlm.nih.gov/pubmed/20808728>. Accessed January 17, 2019.
 167. Bootkrajang J, Chaijaruwanich J. Towards instance-dependent label noise-tolerant classification: a probabilistic approach. Pattern Analysis and Applications.
<http://www.cs.science.cmu.ac.th/person/jakramate/#code>. Published November 28, 2018. Accessed October 1, 2018.
 168. Rank key features by class separability criteria - MATLAB rankfeatures.
<https://www.mathworks.com/help/bioinfo/ref/rankfeatures.html#References>. Accessed January 19, 2019.
 169. He S, Chen H, Zhu Z, et al. Robust twin boosting for feature selection from high-dimensional omics data with label noise. *Inf Sci (Ny)*. 2015;291(C):1-18.
doi:10.1016/j.ins.2014.08.048

170. Taguchi YH. Principal components analysis based unsupervised feature extraction applied to gene expression analysis of blood from dengue haemorrhagic fever patients. *Sci Rep.* 2017;7(March):1-14.
doi:10.1038/srep44016
171. Chung NC, Storey JD. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics.* 2015;31(4):545-554.
doi:10.1093/bioinformatics/btu674
172. Wang B. Visualization and analysis of single-cell RNA-seq data by kernel- based similarity learning. 2016;2.
173. Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single cell RNA-Seq data. *bioRxiv.* 2016:068775.
doi:10.1101/068775
174. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 2015;16(1):241. doi:10.1186/s13059-015-0805-z
175. Masci J, Meier U, Cireşan DC, Schmidhuber J. *Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction.*
<http://people.idsia.ch/~cireşan/data/icann2011.pdf>. Accessed January 28, 2019.
176. Erhan D, Bengio Y, Courville A, Ca P-AM, Ca PV, Com B. *Why Does Unsupervised Pre-Training Help Deep Learning? Pierre-Antoine Manzagol Pascal Vincent Samy Bengio.* Vol 11.; 2010.
<http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf>. Accessed January 28, 2019.

177. Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biol Cybern.* 1988;59(4-5):291-294.
doi:10.1007/BF00332918
178. Chicco D, Sadowski P, Baldi P. Deep Autoencoder Neural Networks for Gene Ontology Annotation Predictions. doi:10.1145/2649387.2649442
179. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A. *Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion*. Vol 11.; 2010.
http://delivery.acm.org/10.1145/1960000/1953039/11-3371-vincent.pdf?ip=73.70.168.134&id=1953039&acc=OPEN&key=4D4702B0C3E38B35.4D4702B0C3E38B35.4D4702B0C3E38B35.6D218144511F3437&__acm__=1547930970_c5eb06d17d459413b67ed388750c6a07. Accessed January 19, 2019.
180. Makhzani A, Frey B. k-Sparse Autoencoders. December 2013.
<http://arxiv.org/abs/1312.5663>. Accessed January 19, 2019.
181. Kingma DP, Welling M. *Auto-Encoding Variational Bayes*.
http://www.cs.columbia.edu/~blei/seminar/2016_discrete_data/readings/KingmaWelling2013.pdf. Accessed January 19, 2019.
182. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. 2016. doi:10.1021/acs.molpharmaceut.5b00982
183. Way GP, Greene CS. Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders. doi:10.1101/174474
184. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-

- Cancer analysis project. *Nat Genet.* 2013;45(10):1113-1120. doi:10.1038/ng.2764
185. Dai B, Wang Y, Aston J, Wipf D. *Hidden Talents of the Variational Autoencoder*.
<https://arxiv.org/pdf/1706.05148.pdf>. Accessed January 19, 2019.
186. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* 2017;45(W1):W130-W137. doi:10.1093/nar/gkx356
187. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet.* 2000;25(1):25-29. doi:10.1038/75556
188. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med.* 1990;9(7):811-818. doi:10.1002/sim.4780090710

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Katrina Kalantar

Author Signature

February 19, 2019
Date