PNAS
www.pnas.org

Supplementary Information for:

Combined Host and Microbe Next Generation Sequencing for Lower Respiratory Tract Infection Diagnosis in Critically Ill Adults

Authors: Charles Langelier[1]*, Katrina L Kalantar[2]* et al.

*equal contributions

Affiliations:

[1]Division of Infectious Diseases, Department of Medicine, University of California, San Francisco, CA, USA

[2]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA

[3]Division of Pulmonary, Critical Care, Sleep and Allergy, Department of Medicine, University of California, San Francisco, USA

[4]Chan Zuckerberg Biohub, San Francisco CA

[5]Department of Medicine, University of California, San Francisco, CA, USA

[6]Division of Critical Care Medicine, Department of Pediatrics, University of California, San Francisco - School of Medicine, San Francisco, CA, USA

[7] Department of Laboratory Medicine, University of California, San Francisco - School of Medicine, San Francisco, CA, USA

www.pnas.org/cgi/doi/10.1073/pnas.1809700115

[8]Weill Institute for Neurosciences, Department of Neurology, University of California, San Francisco, San Francisco, CA, USA

[9]Gladstone Institutes, San Francisco, CA 94158, USA

[10]Department of Epidemiology and Biostatistics and Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA, USA

Paste corresponding author name here

Email: joe@derisilab.ucsf.edu

This PDF file includes:

SI Appendix, Supplementary Methods

SI Appendix, Figures S1 to S5

SI Appendix, Dataset legends

SI Appendix, References

Other SI Appendix materials for this manuscript include the following:

SI Appendix Datasets S1 to S9

SI Appendix Methods

Identification of Subjects with LRTI

Subjects with LRTI were identified by two-physician adjudication, as described in the main text. 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.

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 logistic regression model 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…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 S5A)**. 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.

Differential Expression of Viral versus Bacterial LRTI

Gene count data were analyzed using the Bioconductor package DESeq2 (v 1.16.1) (69) 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 ToppGene (44) to evaluate for functional pathway enrichment.

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) (1). 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.

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 0.1.1(2) 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.

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 (**Figures S5B**). Learning curves are a widely used and robust approach to determine optimal training set sample size in machine learning analyses(3, 4). 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.

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.

SI Appendix Figures

Microbe identified by clinical diagnostics and in reference list of established LRTI pathogens

Microbe identified by clinical diagnostics but not in reference list of established LRTI pathogens

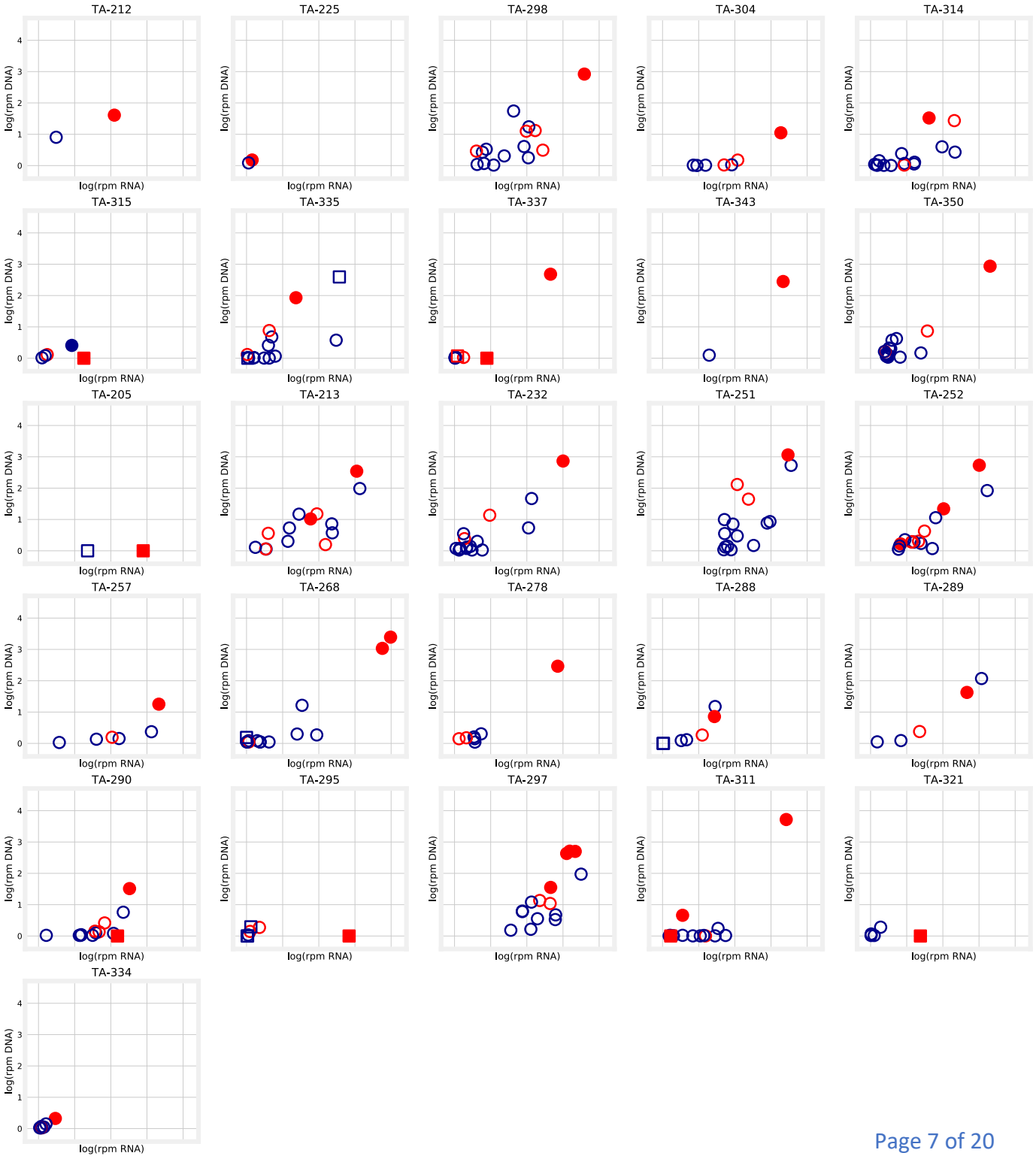
Microbe NOT identified by clinical diagnostics but present in reference list of established LRTI pathogens

Microbe NOT identified by clinical diagnostics and NOT present in reference list of established LRTI pathogens

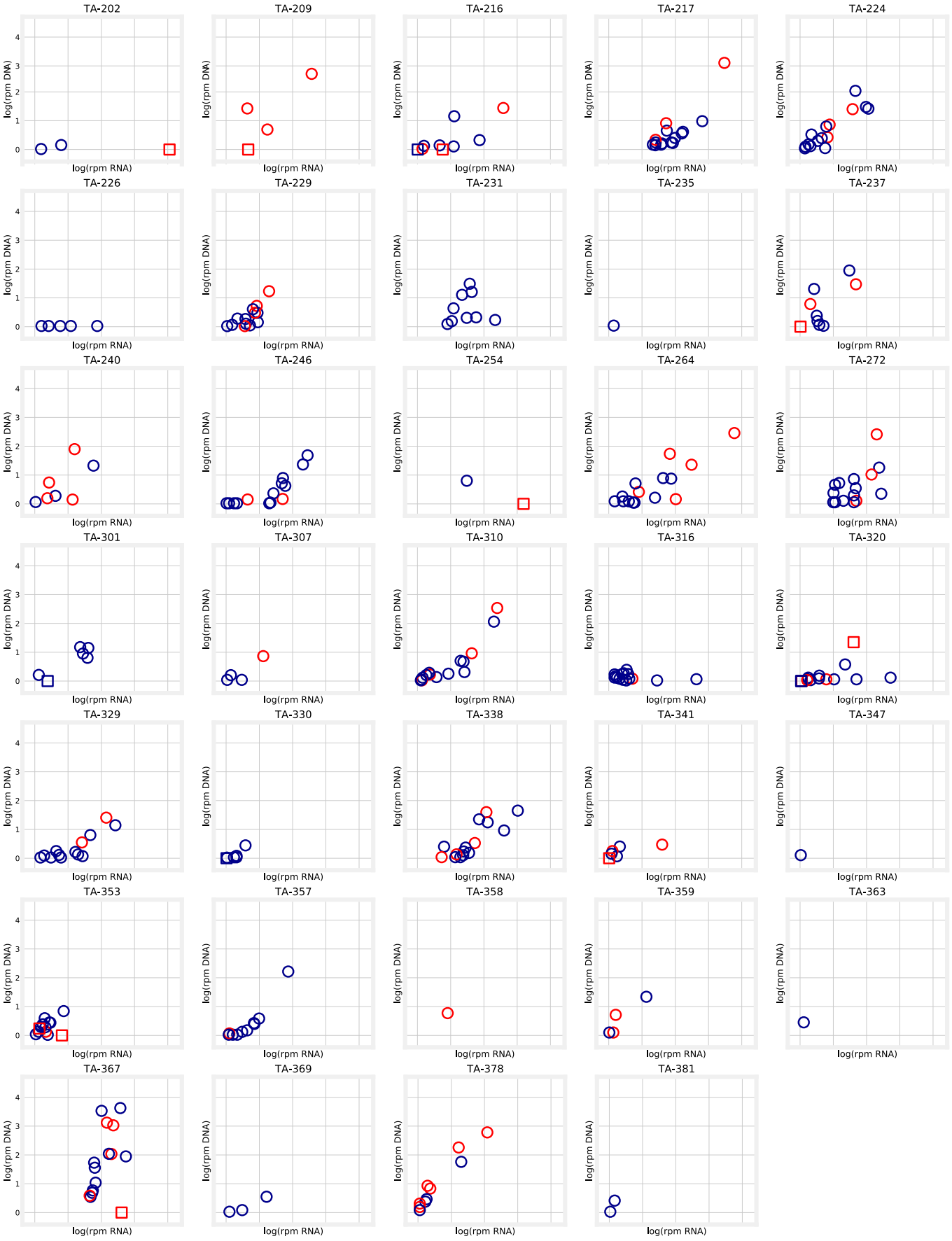
Viruses

Bacteria or Fungi

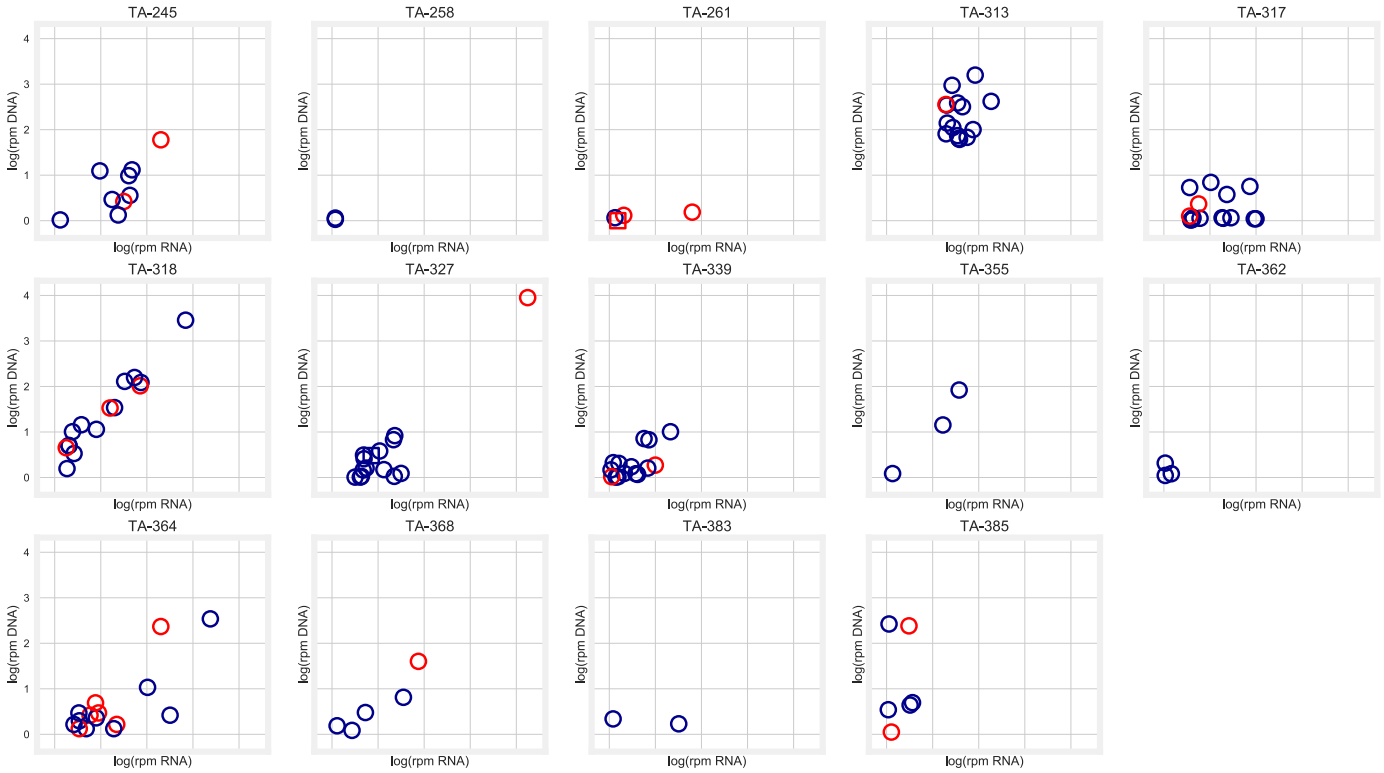
LRTI Identified by Clinical Criteria and Standard Clinical Microbiology (LRTI^{+C+M})



LRTI Identified by Clinical Criteria (LRTI^{+C})



LRTI Status Unknown (unk-LRTI)



No LRTI (no-LRTI)

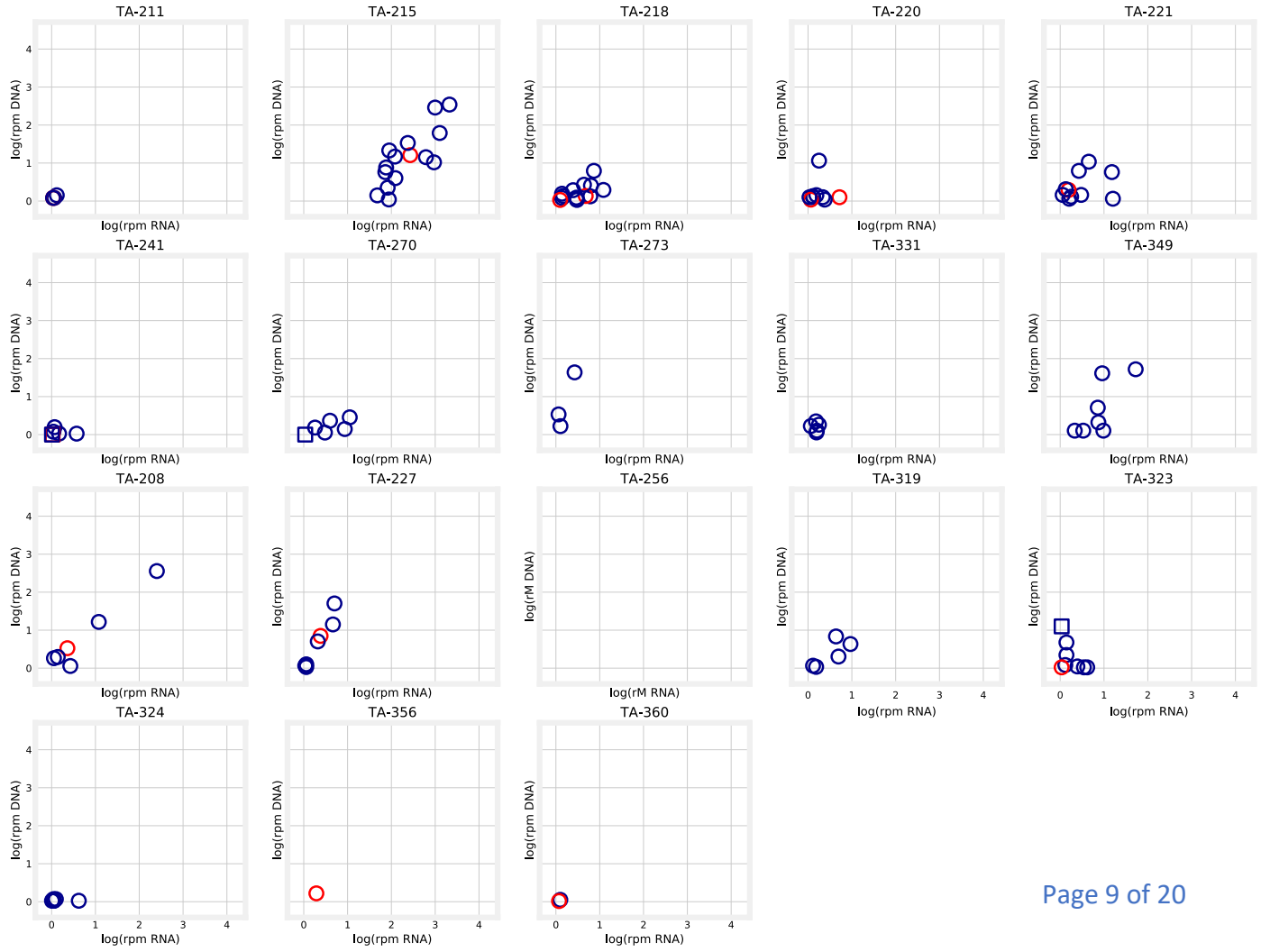


Figure S1.

Distribution of NGS-identified microbes by relative abundance denoting those identified by standard clinical microbiology.

Microbes plotted by log(RNA rpM) versus log(DNA 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.

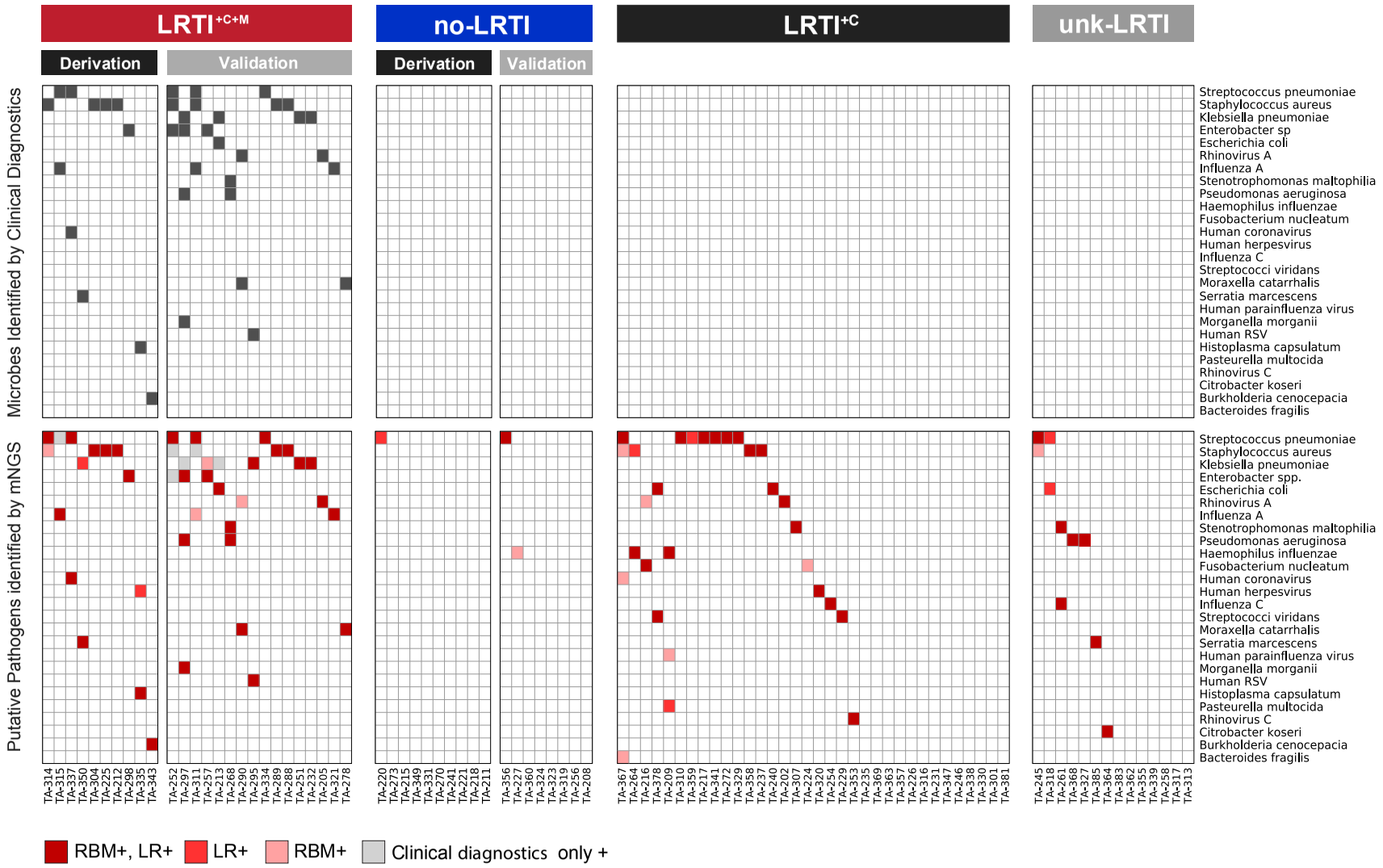

LRTI+C+M
no-LRTI
LRTI+C
unk-LRTI
Derivation
Validation
Derivation
Validation
Microbes Identified by Clinical Diagnostics
Putative Pathogens identified by mNGS
Streptococcus pneumoniae
Staphylococcus aureus
Klebsiella pneumoniae
Enterobacter sp
Escherichia coli
Rhinovirus A
Influenza A
Stenotrophomonas maltophilia
Pseudomonas aeruginosa
Haemophilus influenzae
Fusobacterium nucleatum
Human coronavirus
Human herpesvirus
Influenza C
Streptococci viridans
Moraxella catarrhalis
Serratia marcescens
Human parainfluenza virus
Morganella morganii
Human RSV
Histoplasma capsulatum
Pasteurella multocida
Rhinovirus C
Citrobacter koseri
Burkholderia cenocepacia
Bacteroides fragilis
Streptococcus pneumoniae
Staphylococcus aureus
Klebsiella pneumoniae
Enterobacter spp.
Escherichia coli
Rhinovirus A
Influenza A
Stenotrophomonas maltophilia
Pseudomonas aeruginosa
Haemophilus influenzae
Fusobacterium nucleatum
Human coronavirus
Human herpesvirus
Influenza C
Streptococci viridans
Moraxella catarrhalis
Serratia marcescens
Human parainfluenza virus
Morganella morganii
Human RSV
Histoplasma capsulatum
Pasteurella multocida
Rhinovirus C
Citrobacter koseri
Burkholderia cenocepacia
Bacteroides fragilis
TA-314 TA-315 TA-337 TA-350 TA-304 TA-225 TA-212 TA-298 TA-335 TA-343
TA-252 TA-297 TA-311 TA-257 TA-213 TA-268 TA-290 TA-295 TA-334 TA-289 TA-288 TA-251 TA-232 TA-205 TA-321 TA-278
TA-220 TA-273 TA-215 TA-349 TA-331 TA-270 TA-241 TA-221 TA-218 TA-211
TA-356 TA-227 TA-360 TA-324 TA-323 TA-319 TA-256 TA-208
TA-367 TA-264 TA-216 TA-378 TA-209 TA-310 TA-359 TA-217 TA-341 TA-272 TA-329 TA-358 TA-237 TA-240 TA-202 TA-307 TA-224 TA-320 TA-254 TA-229 TA-353 TA-235 TA-369 TA-363 TA-357 TA-226 TA-316 TA-231 TA-347 TA-246 TA-338 TA-330 TA-301 TA-381
TA-245 TA-318 TA-261 TA-368 TA-327 TA-385 TA-364 TA-383 TA-362 TA-355 TA-339 TA-258 TA-317 TA-313
RBM+, LR+
LR+
RBM+
Clinical diagnostics only +

Figure S2. Microbial pathogens identified by clinician-ordered diagnostics, compared to those identified by mNGS rules-based (RBM) and logistic regression (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+).

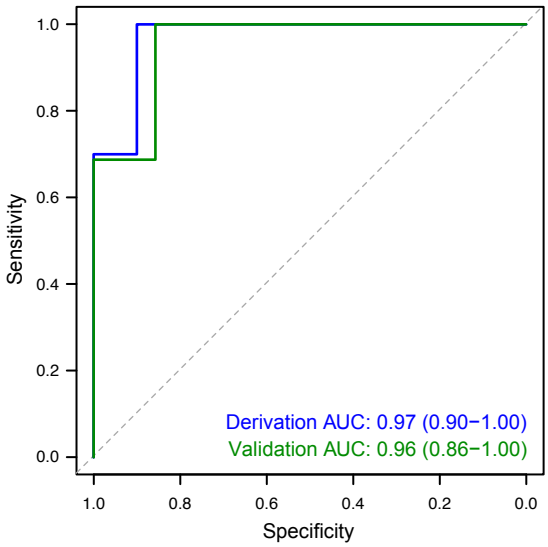


Figure S3. Performance of LRM-derived top microbe probability score for differentiating patients with LRTI from those with non-infectious causes of acute respiratory failure.

The top microbe probability score per patient from the LR model 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 AUC of 0.97 (95% CI = 0.90 - 1.00) in the derivation cohort and AUC of 0.93 (95% CI = 0.82-1.00) in the validation cohort.

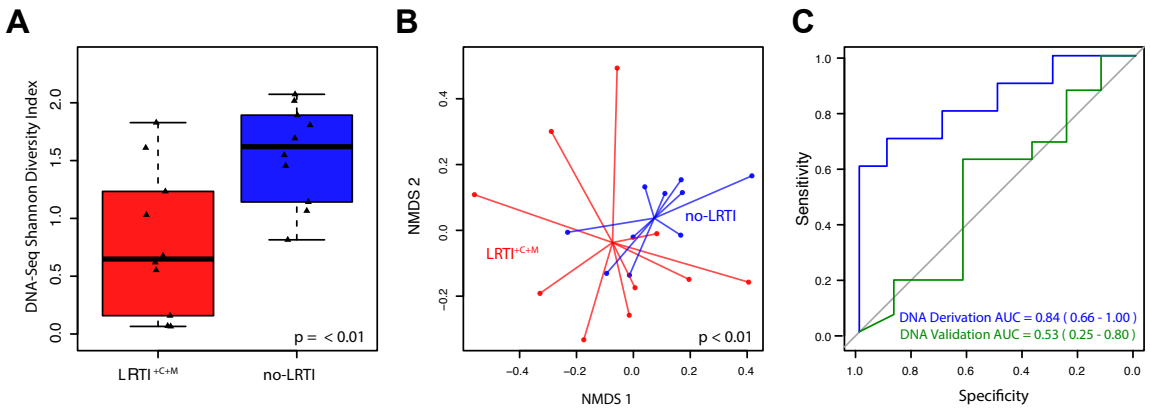


Figure S4. Performance DNA-Seq microbiome diversity assessment for differentiating patients with LRTI from those with non-infectious causes of acute respiratory failure.

A) DNA-Seq Shannon Diversity Index (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.** 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. The AUC curves for derivation and validation cohorts are blue and green, respectively.

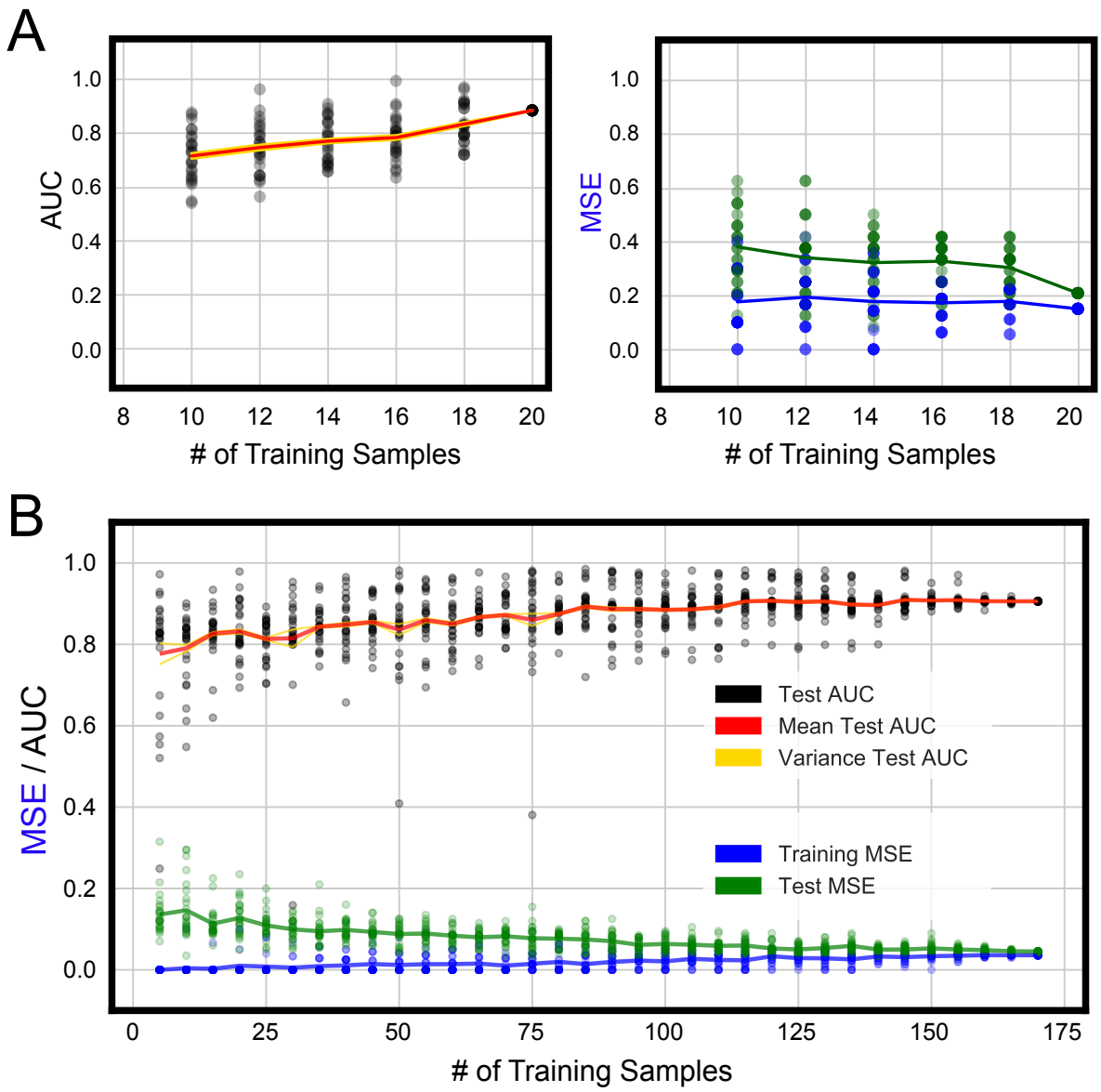


Figure S5. Learning Curve analyses for pathogen versus commensal and host gene classifier models.

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 LR model 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.

Additional Datasets S1 – S9

Dataset S1. 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.

Dataset S2. A) United States Centers for Disease Control/National Healthcare Surveillance Network surveillance definition of pneumonia used for adjudication in this study. **B)** 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(5–9).

Dataset S3. A) 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 LR model 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 mNGS-identified virus. +Sample orthogonally tested by 16S rRNA gene sequencing to confirm one or more bacterial results. **B)** Feature weights for the LRM determined by machine learning in the derivation cohort.

Dataset S4. 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.

Dataset S5. Diversity metrics of the transcriptionally active and total fractions of the lung microbiome assessed by RNA-Seq and by shotgun DNA sequencing, respectively. Wilcoxon Rank Sum statistical significance and AUC of Simpsons Diversity Index, Shannon Diversity Index, richness (number of genera), microbial sequence abundance (total microbial alignments by genus normalized per million reads sequenced) and Bray-Curtis index calculated on RNA-Seq and DNA-Seq datasets.

Dataset S6. 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 Biological Processes with significant enrichment in either $LRTI^{+C+M}$ or no-LRTI subjects.

Dataset S7. LRTI host transcriptional classifier specifics. **A)** 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. **B)** 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. **C)** CIBERSORT(1) was utilized to predict cell type proportions for each patient. M2 Macrophages were identified to have significant differences in estimated proportions (Wilcoxon Rank Sum $p = 0.03$).

Dataset S8. 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.

Dataset S9. Human Transcript 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.

Supplemental References

1. Dobbin KK, Zhao Y, Simon RM (2008) How Large a Training Set is Needed to Develop a Classifier for Microarray Data? *Clin Cancer Res* 14(1):108–114.

2. Newman AM, et al. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 12(5):453–457.

3. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH (2012) Predicting sample size required for classification performance. *BMC Med Inform Decis Mak* 12:8.

4. Meek C, Thiesson B, Heckerman D The Learning-Curve Sampling Method Applied to Model-Based Clustering. 22.

5. Jain S, et al. (2015) Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. *N Engl J Med* 373(5):415–427.

6. Magill SS, et al. (2014) Multistate point-prevalence survey of health care-associated infections. *N Engl J Med* 370(13):1198–1208.

7. Mandell LA, et al. (2007) Infectious Diseases Society of America/American Thoracic Society Consensus Guidelines on the Management of Community-Acquired Pneumonia in Adults. *Clin Infect Dis* 44(Supplement_2):S27–S72.

8. Sethi S, Murphy TF (2008) Infection in the Pathogenesis and Course of Chronic Obstructive Pulmonary Disease. *N Engl J Med* 359(22):2355–2365.

9. Fishman JA (2017) Infection in Organ Transplantation. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg* 17(4):856–879.