ML_final_project

Integrating host response and unbiased microbe detection for precision sepsis diagnosis in critically ill adults

Introduction

Sepsis is a major threat to public health. It causes nearly 20% of all deaths globally and contributes to one in every two to three hospital deaths in the United States (U.S.). Furthermore, it is the most expensive condition treated in the U.S., costing hospitals more than $38 billion dollars annually [1-3]. Sepsis mortality increases by 8% every hour that appropriate antibiotic therapy is delayed [4], making the rapid and accurate identification of the causative pathogen in bacterial sepsis critical to saving patients' lives. Unfortunately, existing clinical infectious disease diagnostics are largely limited to antiquated, low-yield techniques [5, 6], and therefore, the causative pathogen is rarely identified. As a result, antibiotic treatment often remains empiric rather than pathogen-targeted, with clinical decision-making based on epidemiological information rather than individual patient data [7].

Despite decades of research to improve infection diagnostics, current methods are limited by their reliance on a culturable pathogen, the kinetics of microorganism growth, and clinician-guided testing schema. With the advent of culture-independent methods, these challenges may be overcome. Metagenomic next generation sequencing (mNGS) is an unbiased approach to microbial identification, which can be applied directly to clinical samples [8]. This technique has been utilized for pathogen detection in complex clinical cases posing diagnostic dilemmas due to the failure of traditional diagnostics [9, 10]. mNGS also allows for the identification of host gene expression signatures that distinguish infectious and non-infectious insults [11, 12]. Sepsis, a "life-threatening organ dysfunction due to a dysregulated host response to infection," [13] provides a well-suited opportunity to harness these strengths of mNGS and capture both host and microbe pathophysiology simultaneously.

Methods

Study design, clinical cohort and ethics statement

We conducted a prospective observational study of patients with acute critical illnesses admitted from the ED to the ICU. We studied patients who were enrolled at the University of California, San Francisco (UCSF) and Zuckerberg San Francisco General Hospital. The study was approved by the UCSF Institutional Review Board, which granted a waiver of initial consent for blood sampling. Informed consent was subsequently obtained from patients or their surrogates for continued study participation, as previously described.

For this analysis, inclusion criteria were: (1) admission to the intensive care unit for mechanical ventilation for ARDS or airway protection, (2) age ≥ 18 years, (3) availability of while blood RNA-seq data with 106 protein-coding reads collected within five days of intubation and (4) availability of matched plasma samples to undergo DNA-seq. Exclusion criteria were those patients identified as outliers in the DESeq2 analysis using the variance stabilizing transformation (VST) and principle components analysis (PCA). Subject charts were reviewed by study authors (MA, CL, AL) to confirm a diagnosis of sepsis. Sepsis was adjudicated using the Sepsis-3 definition.

Metagenomic sequencing

Whole blood and plasma were collected into Paxgene and standard tubes, and frozen within 2 hours. To evaluate host gene expression and detect microbes, metagenomic next generation sequencing RNA sequencing (RNA-seq) was performed on whole blood specimens and DNA-seq on plasma. RNA was extracted using the Qiagen RNEasy kit and normalized to 10ng total input per sample. Human cytosolic and mitochondrial ribosomal RNA and globin RNA was depleted using FastSelect (Qiagen). To control for background contamination, we included negative controls (water and HeLa cell RNA) as well as positive controls (spike-in RNA standards from the External RNA Controls Consortium (ERCC). RNA was then fragmented and underwent library preparation using the NEBNext Ultra II RNAseq Kit (New England Biolabs). DNA was extracted from 300uL of plasma using the Zymo Pathogen Magbead Kit before undergoing DNA-seq using the NEBNext Ultra II kit. Libraries underwent 146 nucleotide paired-end Illumina sequencing on an Illumina Novaseq 6000 instrument.

Host differential expression and pathway analysis

Following demultiplexing, sequencing reads were pseudo-aligned with kallisto27 (v. 0.46.1; including bias correction) to an index consisting of all transcripts associated with human protein coding genes (ENSEMBL v. 99), cytosolic and mitochondrial ribosomal RNA sequences, and the sequences of ERCC RNA standards. Gene-level counts were generated from the transcript-level abundance estimates using the R package `txImport28` , with the `scaledTPM` method.

Differential expression analysis was performed using `DESeq2` . We modeled the expression of individual genes using the design formula `~ Group` . Significant genes were identified using an independent-hypothesis-weighted, Benjamini-Hochberg false discovery rate (FDR) less than 0.1. Empirical Bayesian adaptive shrinkage estimators for $\log_2$-fold change were fit using `ashr` . We generated heatmaps of the top 50 differentially expressed genes by absolute $\log_2$-fold change. visualization, gene expression was normalized using the VST, centered, and z-scaled. Heatmaps were generated using the `pheatmap` package. Unsupervised hierarchical clustering was performed for patients using Euclidean distance and for genes using Manhattan distance. Differentially expressed genes (FDR < 0.05) were analyzed using `Webgestalt` .

Sepsis classifier using random forest

RNA sequencing reads that were used for differential gene expression analyses as outlined above were then used to perform sepsis classification using a random forest model. The dataset with only Group_1 and Group_4 was used with outliers identified by VST transformed gene count heatmaps and PCA analysis included. The dataset was split into training and validation sets at 70%/30%. The `caret` package was used to perform 10-fold cross-validation to fit the random forest model on the training set. The best random forest model that was identified by cross-validation was then used on the validation set to assess model performance. Model performance was assessed using area under the receiver operating characteristic curve (AUC) using the packages `pROC` and `Epi` . A confusion matrix also presented.

Results

Patient enrollment and sample collection

We conducted a prospective observational study of the EARLI cohort(refs) of critically ill adults admitted from the Emergency Department (ED) to the Intensive Care Unit (ICU). Patients were enrolled at two tertiary care hospitals in San Francisco, California under a research protocol approved by the University of California San Francisco Institutional Review Board (Methods). Following enrollment, whole blood and plasma were collected in the ED and stored at -80C prior to RNA and DNA extraction, respectively. RNA sequencing (RNA-seq) was carried out to assay the host response, and DNA-seq to identify microbes.

Patients with sepsis were clinically adjudicated with blinding to mNGS results using the Sepsis-3 definition [17], and categorized into five subgroups of sepsis status: patients with sepsis due to clinically identified bloodstream infection Group_1 (SepisBldCx+), sepsis due to a microbiologically confirmed primary infection at a site other than the bloodstream Group_2 (SepsisOthCx+), suspected sepsis with negative clinical microbiologic testing Group_3 (Suspected), patients with no evidence of sepsis and a clear alternative explanation for their critical illness Group_4 (No-Sepsis), or patients of indeterminant status Group_5 (Indeterm). Patient demographics are presented in Table 1.

	1_Sepsis+BldCx+	2_Sepsis+OtherCx+	3_Sepsis+Cx-	4_NO_Sepsis	5_Unclear	p	test
n	64	67	68	94	33		
Age (mean (SD))	61.80 (15.33)	67.52 (14.86)	64.37 (15.29)	63.70 (15.46)	70.24 (12.32)	0.049	
APACHEII (mean (SD))	33.05 (10.47)	29.67 (7.33)	32.19 (9.34)	32.05 (9.90)	31.06 (8.22)	0.285	
APACHEIII (mean (SD))	66.42 (22.36)	55.72 (16.34)	64.74 (19.52)	64.26 (18.58)	63.18 (18.83)	0.014	
TempMaxSAPS (mean (SD))	38.07 (1.11)	37.53 (1.43)	37.60 (1.16)	36.92 (1.49)	36.90 (1.43)	<0.001	
TempMinSAPS (mean (SD))	35.46 (2.09)	35.04 (2.10)	35.35 (1.60)	34.49 (2.02)	34.91 (2.10)	0.020	
WBCMaxSAPS (mean (SD))	17.80 (12.13)	15.92 (9.35)	16.11 (9.29)	16.44 (17.29)	12.25 (5.92)	0.352	
WBCMinSAPS (mean (SD))	12.51 (10.46)	12.11 (6.47)	12.01 (7.86)	12.16 (14.66)	10.13 (5.43)	0.870	
HRMaxSAPS (mean (SD))	125.20 (24.33)	109.49 (22.57)	112.19 (24.75)	106.18 (25.64)	103.61 (21.58)	<0.001	
HRMinSAPS (mean (SD))	75.55 (23.71)	66.39 (19.55)	68.24 (20.41)	58.76 (17.78)	61.70 (16.48)	<0.001	
RRMaxSAPS (mean (SD))	35.27 (5.83)	33.00 (8.63)	35.03 (9.24)	34.90 (11.57)	34.48 (10.67)	0.652	
RRMinSAPS (mean (SD))	14.89 (5.95)	12.99 (4.02)	13.60 (5.44)	12.83 (5.73)	12.18 (3.62)	0.074	
SBPMaxSAPS (mean (SD))	155.52 (31.83)	162.30 (32.09)	155.82 (26.33)	167.16 (33.63)	168.52 (29.21)	0.053	
SBPMinSAPS (mean (SD))	70.88 (15.30)	76.96 (16.64)	71.94 (14.22)	76.82 (18.78)	78.45 (16.58)	0.047	
CreatinineMaxSAPS (mean (SD))	2.19 (1.88)	2.12 (2.44)	2.46 (2.92)	2.68 (3.53)	1.50 (1.12)	0.258	
CreatinineMinSAPS (mean (SD))	1.75 (1.42)	1.66 (1.80)	1.91 (2.01)	2.01 (2.35)	1.34 (1.09)	0.453	
PlateletsMinSAPS (mean (SD))	135.72 (115.16)	175.85 (68.11)	173.63 (108.09)	177.30 (80.15)	206.73 (96.79)	0.007	
Gender (%)						0.243	
Male	45 (70.3)	37 (55.2)	38 (55.9)	54 (57.4)	14 (42.4)		
Female	19 (29.7)	29 (43.3)	29 (42.6)	40 (42.6)	19 (57.6)		
Transgender	0 (0.0)	1 (1.5)	1 (1.5)	0 (0.0)	0 (0.0)		
Race (%)						0.443	
Caucasian	22 (34.4)	28 (41.8)	24 (35.3)	37 (39.4)	15 (45.5)		
African American	10 (15.6)	11 (16.4)	12 (17.6)	18 (19.1)	10 (30.3)		
Asian	18 (28.1)	21 (31.3)	23 (33.8)	25 (26.6)	6 (18.2)		
Native American	1 (1.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)		
Other	12 (18.8)	5 (7.5)	9 (13.2)	14 (14.9)	2 (6.1)		
Unknown	1 (1.6)	2 (3.0)	0 (0.0)	0 (0.0)	0 (0.0)		
HospWithin30d = 1 (%)	7 (10.9)	7 (10.4)	9 (13.2)	17 (18.1)	8 (24.2)	0.282	
PNA_community = 1 (%)	8 (12.5)	9 (13.4)	16 (23.5)	5 (5.3)	2 (6.1)	0.009	
PNA_healthcare = 1 (%)	7 (10.9)	9 (13.4)	10 (14.7)	2 (2.1)	5 (15.2)	0.041	
COPD2 = 1 (%)	10 (15.6)	15 (22.4)	12 (17.6)	14 (14.9)	14 (42.4)	0.010	

CongestiveHeartFailure = 1 (%)	12 (18.8)	11 (16.4)	16 (23.5)	29 (30.9)	16 (48.5)	0.005
CardiacArrest = 1 (%)	5 (7.8)	7 (10.4)	4 (5.9)	7 (7.4)	3 (9.1)	0.901
PtSepsis = 1 (%)	27 (42.2)	31 (46.3)	32 (47.1)	5 (5.3)	4 (12.1)	<0.001
28d death = 1 (%)	23 (35.9)	14 (20.9)	28 (41.2)	32 (34.0)	12 (36.4)	0.142
60d death = 1 (%)	27 (42.2)	16 (23.9)	31 (45.6)	34 (36.2)	13 (39.4)	0.094
Hospital death = 1 (%)	24 (37.5)	15 (22.4)	28 (41.2)	30 (31.9)	11 (33.3)	0.192
MechVent = 1 (%)	49 (76.6)	59 (88.1)	60 (88.2)	90 (95.7)	31 (93.9)	0.005
SIRS_HR = 1 (%)	58 (90.6)	53 (79.1)	57 (83.8)	70 (74.5)	22 (66.7)	0.032
SIRS_temp = 1 (%)	51 (79.7)	51 (76.1)	54 (79.4)	71 (75.5)	24 (72.7)	0.915
SIRS_RR = 1 (%)	64 (100.0)	64 (95.5)	68 (100.0)	92 (97.9)	33 (100.0)	0.157
SIRS_WBC = 1 (%)	53 (82.8)	45 (67.2)	48 (70.6)	54 (57.4)	18 (54.5)	0.008
SIRS_total (%)						0.014
1	0 (0.0)	1 (1.5)	0 (0.0)	1 (1.1)	1 (3.0)	
2	6 (9.4)	11 (16.4)	11 (16.2)	22 (23.4)	9 (27.3)	
3	18 (28.1)	30 (44.8)	23 (33.8)	42 (44.7)	14 (42.4)	
4	40 (62.5)	25 (37.3)	34 (50.0)	29 (30.9)	9 (27.3)	
Intubated = 1 (%)	49 (76.6)	58 (86.6)	62 (91.2)	87 (92.6)	31 (93.9)	0.020
OnPressorsSAPS = 1 (%)	57 (89.1)	49 (73.1)	60 (88.2)	61 (64.9)	21 (63.6)	<0.001
virusPresent = 1 (%)	5 (7.8)	22 (32.8)	0 (0.0)	0 (0.0)	0 (0.0)	<0.001
Immunocompromised = 1 (%)	8 (12.5)	5 (7.5)	11 (16.2)	6 (6.4)	6 (18.2)	0.157
SOT = 1 (%)	3 (4.7)	0 (0.0)	3 (4.4)	1 (1.1)	1 (3.0)	0.296
HTN = 1 (%)	16 (25.0)	31 (46.3)	31 (45.6)	44 (46.8)	12 (36.4)	0.045
Cirrhosis = 1 (%)	8 (12.5)	4 (6.0)	10 (14.7)	6 (6.4)	2 (6.1)	0.247
CKD = 1 (%)	12 (18.8)	15 (22.4)	15 (22.1)	29 (30.9)	7 (21.2)	0.438
Malignancy = 1 (%)	13 (20.3)	14 (20.9)	14 (20.6)	20 (21.3)	7 (21.2)	1.000
HIV = 1 (%)	7 (10.9)	4 (6.0)	3 (4.4)	7 (7.4)	0 (0.0)	0.281
Diabetes = 1 (%)	18 (28.1)	22 (32.8)	17 (25.0)	31 (33.0)	13 (39.4)	0.599

Host signature of blood culture positive sepsis

We assessed transcriptional differences between patients with sepsis due to bloodstream infection Group_1 (SepisBldCx+) versus those with no evidence of infection Group_4 (No-Sepsis). More than 5000 differentially expressed genes were identified at an adjusted p value (padj) < 0.05 and unsupervised hierarchical clustering revealed clear separation of groups (Figure 1) using the top 500 differentially expressed genes. Gene set enrichment analysis (GSEA) demonstrated upregulation in pathways related to cytokine signaling and innate immune defense, consistent with acute infection (Figure 2).

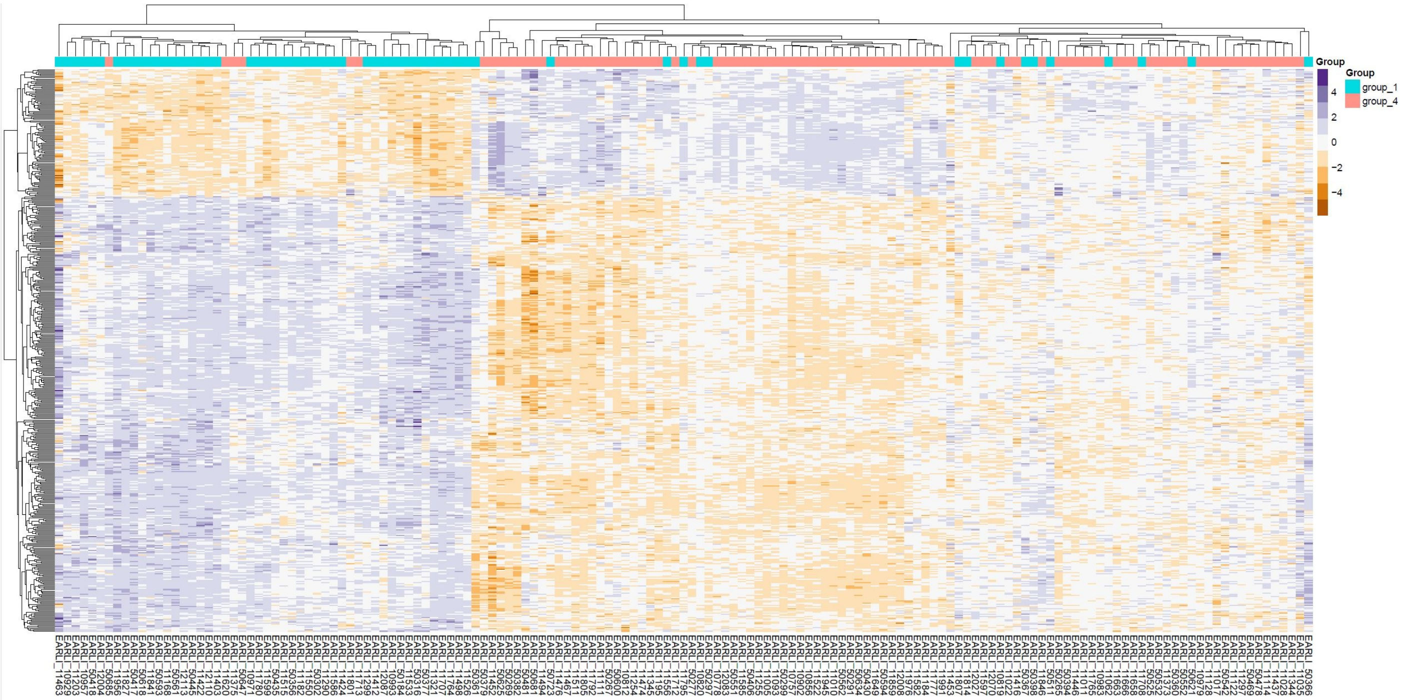

Figure 1

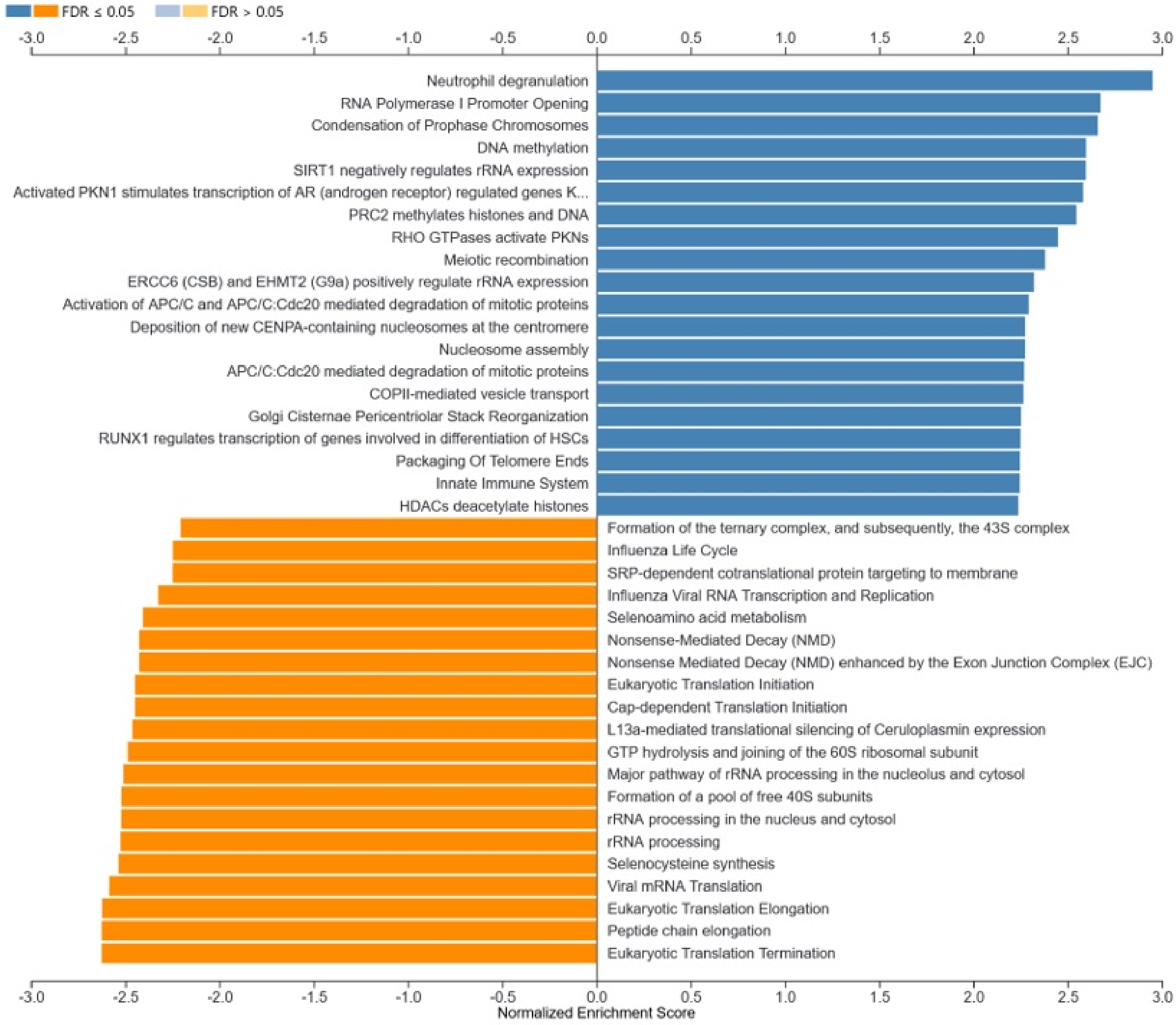

Figure 2

Host transcriptional classifier for bloodstream infection-related sepsis

After characterizing the biological pathways represented in the sepsis host transcriptional signature, we next sought to leverage this signature for clinical diagnosis. DE analysis suggested that BSI were characterized by greater uniformity of host expression compared to sepsis originating from diverse peripheral anatomical location, so we first focused on distinguishing Group_1 (SepisBldCx+) and Group_4 (No-Sepsis). After dividing the cohort into an independent training and validation set using an 70%/30% split, we employed a random forest model. 10-fold cross-validation was performed on the training set and model performance was assessed using a confusion matrix and area under the receiver operating characteristic curve (AUC).

This yielded a BSI classifier that performed with an accuracy of 0.93 (95% CI 0.82 - 0.99) (Table 2) and an AUC of 0.92 (95% CI 0.83 – 1.0) (Figure 3). The confidence interval for the AUC was calculated using 2000 boostrap samples.

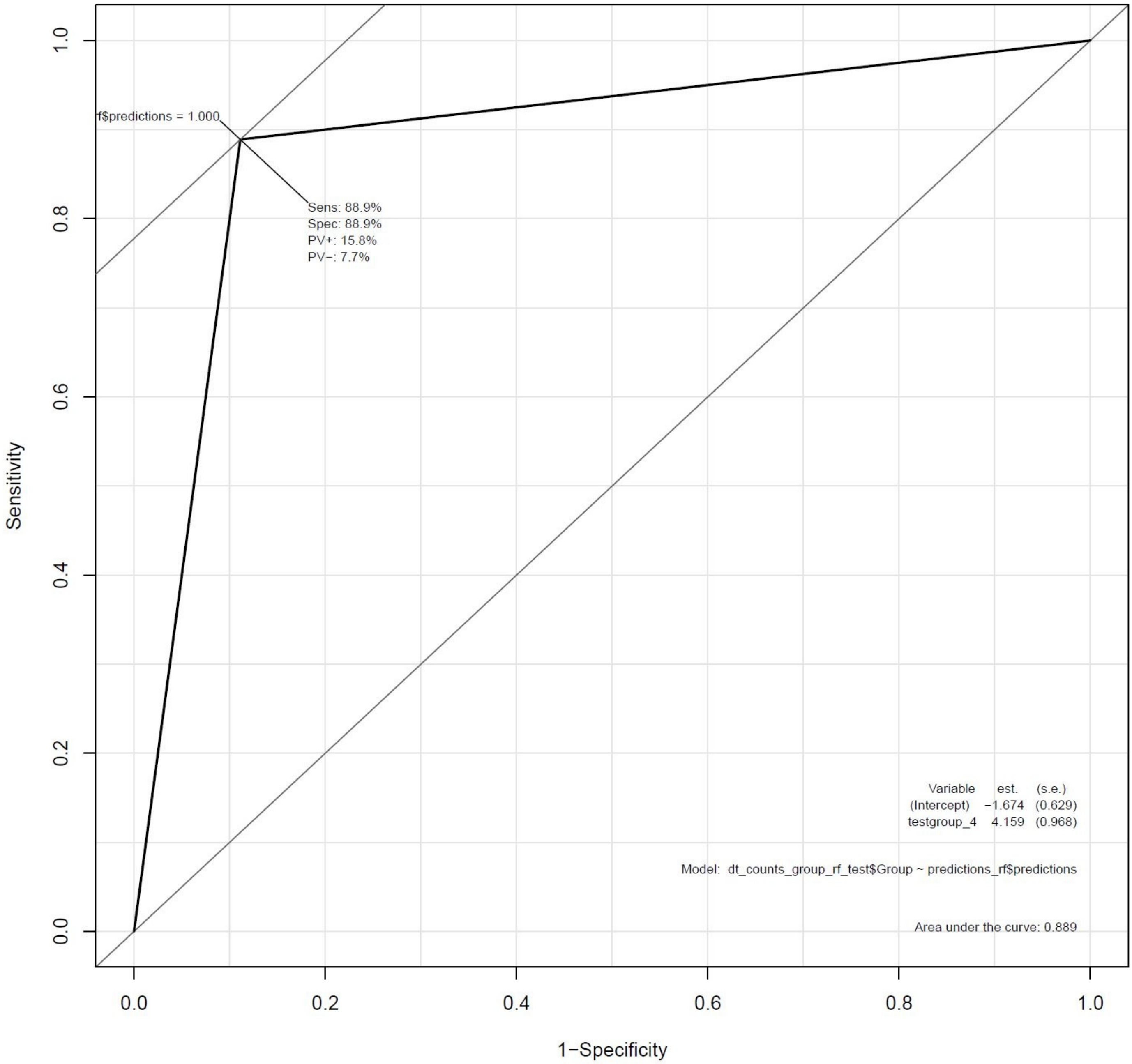

Figure 3

Table 2a - Confusion matrix: bolded labels are the true class

	group_1	**group_4**
group_1	15	0
group_4	3	27

Table 2b - Accuracy, kappa value, and 95% confidence interval of accuracy

	x
Accuracy	0.9333333
Kappa	0.8571429
AccuracyLower	0.8173155
AccuracyUpper	0.9860349

Table 2c - Sensitivity, specificity, positive predictive value, and negative predictive value

	x
Sensitivity	0.8333333
Specificity	1.0000000

Pos Pred Value	1.0000000
Neg Pred Value	0.9000000

Discussion

Sepsis is defined as a dysregulated host response to infection yet no existing clinical diagnostics evaluate both critical aspects of the disease. Here we present a novel approach to integrating host and microbial metrics using metagenomic sequencing. We demonstrate that this approach identified important differences in gene expression between bacteremic sepsis patients and critically ill patients without sepsis. Unsupervised hierarchical clustering was able to distinguish these two patient groups. Further, a random forest model was able to distinguish between bacteremic sepsis and critical illness without sepsis at an AUC of 0.90. These results set the stage for a transformation in sepsis diagnostics. Metagenomic sequencing can be performed at the bedside in order to help clinicians decide on triage and adminsitration of antibiotics. Next steps involve integrating the microbial information from metagenomic seqencing into the host response data presented here.

References

1. Rudd, K.E., et al., Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. Lancet, 2020. 395(10219): p. 200-211.
2. Liu, V., et al., Hospital deaths in patients with sepsis from 2 independent cohorts. JAMA, 2014. 312(1): p. 90-2.
3. Liang, L.A., Moore B (IBM Watson Health), Soni A (AHRQ). National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2017. HCUP Statistical Brief #261 2020 7/2/2020 [cited 2020 Sept. 2]; Available from: www.hcup-us.ahrq.gov/reports/statbriefs/sb261-Most-Expensive-Hospital-Conditions-2017.pdf.
4. Ferrer, R., et al., Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Crit Care Med, 2014. 42(8): p. 1749-55.
5. Lamy, B., et al., What is the relevance of obtaining multiple blood samples for culture? A comprehensive model to optimize the strategy for diagnosing bacteremia. Clin Infect Dis, 2002. 35(7): p. 842-50.
6. Aronson, M.D. and D.H. Bor, Blood cultures. Ann Intern Med, 1987. 106(2): p. 246-53.
7. Timbrook, T.T., et al., The Effect of Molecular Rapid Diagnostic Testing on Clinical Outcomes in Bloodstream Infections: A Systematic Review and Meta-analysis. Clin Infect Dis, 2017. 64(1): p. 15-23.
8. Bibby, K., Metagenomic identification of viral pathogens. Trends Biotechnol, 2013. 31(5): p. 275-9.
9. Wilson, M.R., et al., Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med, 2014. 370(25): p. 2408-17.
10. Wilson, M.R., et al., Diagnosing Balamuthia mandrillaris Encephalitis With Metagenomic Deep Sequencing. Ann Neurol, 2015. 78(5): p. 722-30.
11. Sweeney, T.E., H.R. Wong, and P. Khatri, Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Sci Transl Med, 2016. 8(346): p. 346ra91.
12. Tsalik, E.L., et al., Host gene expression classifiers diagnose acute respiratory illness etiology. Sci Transl Med, 2016. 8(322): p. 322ra11.
13. Singer, M., et al., The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 2016. 315(8): p. 801-10.