Integrated host-microbe metagenomics for sepsis diagnosis in critically ill adults

*Katrina Kalantar[1], *Lucile Neyton[2], Mazin Abdelghany[3], Eran Mick[3], Alejandra Jauregui[2], Saharai Caldera[3], Paula Hayakawa Serpa[3], Rajani Ghale[2,3], Jack Albright[4], Aartik Sarma[2], Alexandra Tsitsiklis[3], Aleksandra Leligdowicz[4], Stephanie Christenson[2], Kathleen Liu[5], Kirsten Kangelaris[6], Carolyn Hendrickson[3], Pratik Sinha[7], Antonio Gomez[8], Norma Neff[9], Angela Pisco[9], Sarah Doernberg[3], Joseph L. Derisi[9,10], Michael A. Matthay[2], †Carolyn S. Calfee[2], †Charles R. Langelier[3,9].

*†equal contributions

Affiliations:

[1]Chan Zuckerberg Initiative, San Francisco, CA, USA

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

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

[4]Department of Critical Care Medicine, Western University, London, Ontario, Canada

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

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

[7]Washington University, St Louis, St. Louis, MO, USA

[8]Department of Medicine, Zuckerberg San Francisco General Hospital, San Francisco, CA, USA

[9]Chan Zuckerberg Biohub, San Francisco, CA, USA

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

Keywords: sepsis, metagenomics, transcriptional profiling, classifier, machine learning, plasma

Abstract

Sepsis is a leading cause of death, and improved approaches for disease diagnosis and detection of etiologic pathogens are urgently needed. Here, we carried out integrated host and pathogen metagenomic next generation sequencing (mNGS) of whole blood (n=221) and plasma RNA and DNA (n=138) from critically ill patients following hospital admission. We assigned patients into sepsis groups based on clinical and microbiological criteria: 1) sepsis with bloodstream infection ($Sepsis^{BSI}$), 2) sepsis with peripheral site infection but not bloodstream infection ($Sepsis^{non\text{-}BSI}$), 3) suspected sepsis with negative clinical microbiological testing; 4) no evidence of infection (No-Sepsis), and 5) indeterminant sepsis status. From whole blood gene expression data, we first trained a bagged support vector machine (bSVM) classifier to distinguish $Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$ patients from No-Sepsis patients, using 75% of the cohort. This classifier performed with an area under the receiver operating characteristic curve (AUC) of 0.81 in the training set (75% of cohort) and an AUC of 0.82 in a held-out validation set (25% of cohort). Surprisingly, we found that plasma RNA also yielded a biologically relevant transcriptional signature of sepsis which included several genes previously reported as sepsis biomarkers (e.g., *HLA-DRA, CD-177*). A bSVM classifier for sepsis diagnosis trained on RNA gene expression data performed with an AUC of 0.97 in the training set and an AUC of 0.77 in a held-out validation set. We subsequently assessed the pathogen-detection performance of DNA and RNA mNGS by comparing against a practical reference standard of clinical bacterial culture and respiratory viral PCR. We found that sensitivity varied based on site of infection and pathogen, with an overall sensitivity of 83%, and a per-pathogen sensitivity of 100% for several key sepsis pathogens including *S. aureus, E. coli, K. pneumoniae* and *P. aeruginosa*. Pathogenic bacteria were also identified in 10/37 (27%) of patients in the No-Sepsis group. To improve detection of sepsis due to viral infections, we developed a secondary RNA host transcriptomic classifier which performed with an AUC of 0.94 in the training set and an AUC of 0.96 in the validation set. Finally, we combined host and microbial features to develop a proof-

of-concept integrated sepsis diagnostic model that identified 72/73 (99%) of microbiologically confirmed sepsis cases, and predicted sepsis in 14/19 (74%) of suspected, and 8/9 (89%) of indeterminate sepsis cases. In summary, our findings suggest that integrating host transcriptional profiling and broad-range metagenomic pathogen detection from nucleic acid may hold promise as a tool for sepsis diagnosis.

Introduction

Sepsis causes 20% of all deaths globally and contributes to 20-50% of hospital deaths in the United States alone[1,2]. Early diagnosis and identification of the underlying microbial pathogens is essential for timely and appropriate antibiotic therapy, which is critical for sepsis survival[3,4]. Yet in over 30% of cases, no etiologic pathogen is identified[5], reflecting the limitations of current culture-based microbiologic diagnostics[6]. Adding additional complexity is the need to differentiate sepsis effectively from non-infectious systemic illnesses, which often appear clinically similar at the time of hospital admission.

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. Similarly, clinicians often continue empiric antimicrobials despite negative microbiologic testing for fear of harming patients in the setting of falsely negative results. Both scenarios lead to antimicrobial overuse and misuse, which contributes to treatment failures, opportunistic infections such as *C. difficile* colitis, and the emergence of drug-resistant organisms[7].

With the introduction of culture-independent methods such as metagenomic next generation sequencing (mNGS), limitations in sepsis diagnostics may be overcome[8,9]. Recent advancements in plasma cell-free DNA sequencing have expanded the scope of metagenomic diagnostics by enabling minimally invasive detection of circulating pathogen nucleic acid originating from diverse anatomical sites of infection[9]. The clinical impact of plasma DNA metagenomics has been questioned, however, due to frequent identification of microbes of uncertain clinical significance, inability to detect RNA viruses that cause pneumonia, and limited utility in ruling-out presence of infection[10,11].

Whole blood transcriptional profiling offers the potential to mitigate these limitations by capturing host gene expression signatures that distinguish infectious from non-infectious conditions, and viral from bacterial infections[12,13]. However, because transcriptional profiling exclusively captures the host response to infection, it does not provide precise taxonomic

identification of sepsis pathogens, which limits the utility of this approach when performed alone. Further, transcriptional profiling has traditionally required isolating peripheral-blood mononuclear cells, or stabilizing whole blood in specialized collection tubes, and it has remained unknown whether a simple plasma specimen could yield informative data for host-based infectious disease diagnosis.

In recent work, a single-sample metagenomic approach combining host transcriptional profiling with unbiased pathogen detection was developed to improve lower respiratory tract infection diagnosis[14]. Sepsis, defined as, "life-threatening organ dysfunction from a dysregulated host response to infection[15]," provides an additional clear use case for this integrated host-microbe metagenomics approach. Here, we study a prospective cohort of critically ill adults to develop a novel sepsis diagnostic assay that combines host transcriptional profiling with broad-range pathogen identification. By applying machine learning to high dimensional mNGS data, we evaluate host and microbial features that distinguish microbiologically confirmed sepsis from non-infectious critical illness. We then demonstrate that plasma nucleic acid can be used to profile both host and microbe for precision sepsis diagnosis.

Results

Clinical features of study cohort

We conducted a prospective observational study of critically ill adults admitted from the Emergency Department (ED) to the Intensive Care Unit (ICU) at two tertiary care hospitals (**Figure 1**). Patients were categorized into five subgroups based on sepsis status (**Methods**). These included patients with: 1) clinically adjudicated sepsis and a microbiologically confirmed bacterial bloodstream infection ($Sepsis^{BSI}$), 2) clinically adjudicated sepsis and a microbiologically confirmed non-bloodstream infection ($Sepsis^{non\text{-}BSI}$), 3) suspected sepsis with negative clinical microbiologic testing ($Sepsis^{suspected}$), 4) patients with no evidence of sepsis and a clear alternative explanation for their critical illness (No-Sepsis), or 5) patients of indeterminant

112 status (Indeterm). The most common diagnoses in the No-Sepsis group were cardiac arrest,
113 overdose/poisoning, heart failure exacerbation, and pulmonary embolism. The majority of
114 patients, regardless of subgroup, required mechanical ventilation and vasopressor support
115 (**Supplementary Table 1**). Patients with microbiologically proven sepsis ($Sepsis^{BSI}$ + $Sepsis^{non-BSI}$) did not differ from No-Sepsis patients in terms of age, gender, race, ethnicity,
117 immunocompromise, APACHEIII score, maximum white blood cell count, intubation status, or
118 28-day mortality (**Figure 1**, **Supplementary Table 1**). All but one patient (in the No-Sepsis
119 group) exhibited ≥ 2 systemic inflammatory response syndrome (SIRS) criteria[16].

120 Host transcriptional signature of sepsis from whole blood

121 We first assessed transcriptional differences between patients with clinically and
122 microbiologically confirmed sepsis ($Sepsis^{BSI}$, $Sepsis^{non-BSI}$) versus those without evidence of
123 infection (No-Sepsis) by performing RNA sequencing (RNA-seq) on whole blood specimens (n
124 = 221 total) to obtain a median of 5.8 x 10^7 (95% CI 5.3-6.3 x 10^7) reads per sample. 5,807
125 differentially expressed (DE) genes were identified at an adjusted P value < 0.1 (**Figure 2a,**
126 **Supplementary Data 1**). Gene set enrichment analysis (GSEA), a method that identifies groups
127 of genes within a dataset sharing common biological functions[17], demonstrated upregulation of
128 genes related to neutrophil degranulation and innate immune signaling in the patients with
129 sepsis, with concomitant downregulation of pathways related to translation and rRNA
130 processing (**Figure 2b, Supplementary Data 2**).

131 To further characterize differences between sepsis patients with bloodstream versus
132 peripheral site (e.g., respiratory, urinary tract) infections, we performed differential gene
133 expression (DE) analysis between the $Sepsis^{BSI}$ and $Sepsis^{non-BSI}$ groups, which identified 5,227
134 genes (**Supplementary Data 3)**. GSEA demonstrated enrichment in genes related to CD28
135 signaling, immunoregulatory interactions between lymphoid and non-lymphoid cells, and other
136 functions in the $Sepsis^{non-BSI}$ patients, while the $Sepsis^{BSI}$ group was characterized by

enrichment in genes related to antimicrobial peptides, defensins, G alpha signaling and other pathways (**Supplementary Data 4**).

Host transcriptional classifier for sepsis diagnosis from whole blood

Given the practical necessity to identify sepsis in both Sepsis^{BSI} and Sepsis$^{non\text{-}BSI}$ patients, we constructed a 'universal' sepsis diagnostic classifier based on whole blood gene expression signatures. After dividing the cohort (n=221) into independent training (75% of data, n=165) and validation (25% of data, n=56) groups, we employed a bagged support vector machine learning approach (bSVM) to select genes that most effectively distinguished patients with sepsis (Sepsis^{BSI} and Sepsis$^{non\text{-}BSI}$) from those without (No-Sepsis). We elected to use a bSVM model due to better performance compared to random forest and gradient boosted trees, which were also tested (**Supplementary Table 2**). The bSVM model achieved an average cross-validation AUC of 0.81 (standard deviation (SD) 0.05) over 10 random splits within the training dataset (75% of data, n=165). In the held-out validation set (25% of data, n=56), an AUC of 0.82 was obtained. Additionally, an AUC of 0.85 (SD 0.02) was obtained over 10 randomly-generated validation sets (**Figure 2c, Supplementary Data 5**).

Host transcriptional classifier for sepsis diagnosis from plasma RNA

Sequencing of plasma DNA has emerged as a preferred strategy for culture-independent detection of bacterial pathogens in the bloodstream[9]. It remained unknown, however, whether plasma RNA could provide meaningful and biologically relevant gene expression data, as sepsis transcriptional profiling studies have historically relied on isolation of PBMCs or collection of whole blood.

To test this, we sequenced RNA from patients with available plasma specimens matched to the whole blood samples, and obtained a median of 2.3×10^7 (95% CI $2.2\text{-}2.5 \times 10^7$) reads per sample. Calculation of input RNA mass (**Methods**) demonstrated that samples with transcript counts below our QC cutoff (< 50,000) had a lower average input mass than those

with sufficient counts (65.2 pg versus 85.8 pg, respectively, p <0.0001, **Supplementary Data 8**). After filtering to retain samples with ≥ 50,000 transcripts (n=138), we performed DE analysis to assess whether a biologically plausible signal could be observed between patients with sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$, n=73) and those without (No-Sepsis, n=37), and found 62 genes at an adjusted P value < 0.1 (**Supplementary Data 6**), 28 of which were also significant in the whole blood analysis (**Supplementary Figure 1**). Remarkably, several of the top differentially expressed genes were previously reported sepsis biomarkers (e.g., elevated *CD177,* suppressed *HLA-DRA*),[18–21] suggesting a biologically relevant transcriptomic signature from plasma RNA (**Figure 2d, Supplementary Data 6**).

We then asked whether a host transcriptional sepsis diagnostic classifier could be constructed using plasma RNA transcriptomic data by dividing the cohort into independent training (75% of data, n=82) and validation groups (n=28), and employing the same bSVM approach to select genes that most effectively distinguished $Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$ patients from No-Sepsis patients. This approach yielded a classifier that achieved an average cross-validation AUC of 0.97 (SD 0.03) over 10 random splits within the training dataset (75% of data, n=82). In the held-out validation set (25% of data, n=28), an AUC of 0.77 was obtained. An AUC of 0.90 (SD 0.06) was obtained over 10 randomly-generated validation sets (**Figure 2e, Supplementary Data 7**).

<u>Detection of bacterial sepsis pathogens from plasma nucleic acid</u>

We began microbial metagenomic analyses by assessing DNA microbial mass (**Methods**), which was significantly lower in negative control water samples, but did not differ between adjudicated sepsis groups (**Figure 3a**, **Supplementary Data 8**). We next carried out bacterial pathogen detection using the IDseq pipeline[22] for taxonomic alignment followed by a previously developed rules-based model (RBM)[14] that identifies established sepsis pathogens overrepresented in mNGS data compared to less abundant commensal or contaminating

187 microbes[14] (**Methods, Figure 3b**).

188 We then asked how well the metagenomic RBM pathogen predictions agreed with
189 bacterial blood culture data. Polymicrobial blood cultures of ≥ 3 organisms were excluded (n=2)
190 given their unclear clinical significance, leaving a total of 40 blood culture-positive cases
191 available for comparison (**Supplementary Data 9**). Sensitivity versus blood culture as a
192 reference standard was 83%, and varied by pathogen, ranging from 0% (e.g., *C. difficile*) to
193 100% (e.g., *E. coli, S. aureus/argenteus* **Figure 3c**). Pathogens were called by the RBM in
194 10/37 (27%) of patients in the No-Sepsis group, equating to a specificity of 73%.

195 Detection of sepsis pathogens from peripheral sites using plasma nucleic acid

196 Plasma DNA mNGS identified 2/25 (8%) of culture-confirmed bacterial lower respiratory
197 tract infection (LRTI) pathogens in the Sepsis$^{non\text{-}BSI}$ group and 3/8 (38%) culture-confirmed
198 bacterial urinary tract infection (UTI) pathogens (**Figure 3d, Supplementary Data 9**). mNGS
199 did not identify *C. difficile* in any of the three patients with severe colitis from this organism.
200 Additional putative bacterial pathogens not detected by culture were detected in 8/73 (11%) of
201 patients with microbiologically confirmed sepsis (**Supplementary Data 9**).

202 Identification of viral infections using host transcriptional profiling of RNA and whole blood

203 Only one of 13 (8%) respiratory viruses identified by clinical testing could be detected by
204 plasma RNA mNGS (**Supplementary Data 9**). Recognizing that an alternative approach would
205 be needed, we asked whether host response could instead be used to identify viral sepsis by
206 carrying out differential gene expression analysis of patients with or without clinically confirmed
207 viral sepsis within the Sepsis^{BSI} and Sepsis$^{non\text{-}BSI}$ groups, using whole blood (**Supplementary**
208 **Data 10**) or plasma (**Supplementary Data 11**) transcriptomic data. GSEA demonstrated that
209 pathways related to interferon signaling and genes important for antiviral immunity were
210 enriched in samples from patients with viral sepsis versus those with bacterial sepsis, in data
211 derived from both whole blood (**Figure 4a, Supplementary Data 12a**) and plasma (**Figure 4b,**

Supplementary Data 12b) datasets.

We then leveraged this host signature to build a secondary bSVM diagnostic classifier for viral sepsis selecting differentially expressed genes as potential predictors, which on whole blood samples achieved an average cross-validation AUC of 0.90 (SD 0.07) over 10 random splits within the training dataset (75% of data, n=96). In the held-out validation set (25% of data, n=33), an AUC of 0.79 was obtained. An AUC of 0.87 (SD 0.04) was obtained over 10 randomly-generated validation sets (**Figure 4c, Supplementary Data 13**). Slightly better performance was obtained when building a classifier using plasma RNA-seq data, with an average cross-validation AUC of 0.94 (SD 0.09) over 10 random splits within the training dataset (75% of data, n=54). In the held-out validation set (25% of data, n=19), an AUC of 0.96 was obtained. An AUC of 0.94 (SD 0.07) was obtained over 10 randomly-generated validation sets (**Figure 4d, Supplementary Data 14**). Incorporation of the host-based viral sepsis classifier improved the sensitivity versus clinical respiratory viral PCR testing to 12/13 (92%), and predicted viral infection in one additional Sepsis$^{non\text{-}BSI}$ patient who didn't undergo viral PCR testing (**Supplementary Data 15**).

Integrated host-microbe sepsis diagnostic model using plasma nucleic acid

Given the relative success of each independent host and pathogen model, we considered whether combining them could enhance diagnosis, and potentially serve as a sepsis rule-out tool. To test this possibility, we developed a proof-of-concept integrated host + microbe model based on simple rules. It returned a sepsis diagnosis based on either host criteria: `[host sepsis classifier probability > 0.5]` or microbial criteria: `[(pathogen detected by RBM) AND (microbial mass > 20 pg)] OR [host viral classifier probability > 0.9]`. Applying these rules enabled detection of 42/42 (100%) of cases in the Sepsis^{BSI} group and 30/31 (97%) of cases in the Sepsis$^{non\text{-}BSI}$ subjects, for an overall sensitivity of 72/73 (99%) (**Figures 5a, 5b**). This proof-of-concept model yielded a specificity of 29/37 (78%) within the No-Sepsis subjects

(**Figure 5c, Supplementary Data 15**).

Application of the integrated model to suspected and indeterminant sepsis cases

Next, we asked whether patients with clinically adjudicated sepsis, but negative in-hospital microbiologic testing ($Sepsis^{suspected}$) would be predicted to have sepsis using the integrated host-microbe plasma mNGS model. 14/19 (74%) were classified as having sepsis, (**Figure 5d**), eight of which had a putative bacterial pathogen identified. Two additional patients had viral host classifier probabilities > 0.5, but did not meet the threshold for sepsis-positivity in the integrated model. With respect to the indeterminate group, the integrated host + microbe model classified 8/9 (89%) as sepsis-positive (**Figure 5e, Supplementary Data 15**). Of these, two had a putative bacterial pathogen identified and one had a putative viral infection identified by the viral host classifier.

Comparison against clinical variable models for sepsis diagnosis

Lastly, we asked how host/microbe mNGS compared against sepsis diagnostic models derived exclusively from clinical metrics that would be available at the time of initial evaluation in the ED. We tested three different machine learning methods to distinguish Sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) from No-Sepsis patients, using 34 clinical variables as input (**Supplementary Table 3a**). The data were split into training (75%) and validation (25%) sets, and model performance was evaluated on the latter. The greatest average AUC achieved was 0.62 (SD 0.04) using a random forest model (**Supplementary Table 3b**). We then computed the AUC using the qSOFA score, a widely used clinical score for sepsis diagnosis[15]. The qSOFA achieved an average AUC of 0.48 (SD 0.02).

Discussion

Sepsis is defined as a dysregulated host response to infection[15], yet existing diagnostics

have focused exclusively on either detecting pathogens or assessing features of the infected host. Here, we combined host transcriptional profiling with broad-range pathogen detection to accurately diagnose sepsis in critically ill patients upon hospital admission. Further, we demonstrate that an integrated host-microbe metagenomics approach can be performed on circulating RNA and DNA from plasma, a widely available clinical specimen type with previously unrecognized utility for host-based infectious disease diagnosis.

Identifying an etiologic pathogen is critical for optimal treatment of sepsis. We found that concordance between pathogen detection by plasma mNGS and traditional bacterial blood culture varied by organism. For instance, mNGS sensitivity for detecting *S. aureus* and *E. coli*, two of the most globally important sepsis pathogens[5], was 100%. In contrast, mNGS missed several important but less common sepsis pathogens, such as *S. pyogenes*. We noted that in all false-negative cases, the patients had received antibiotics prior to mNGS sample collection, which may have reduced the abundance of circulating bacterial DNA available for detection. Furthermore, mNGS and blood cultures were performed on different samples, with research specimens collected up to 24 hours after blood cultures, which may have resulted in lower concordance than if samples had been collected contemporaneously.

Several of the microbes missed by mNGS were organisms that in many contexts exist as commensals (e.g., Fusobacterium, Gemella and Streptococcus species). It is unclear whether these organisms were truly etiologic sepsis pathogens or commensals translocated to the blood in the setting of critical illness, and incidentally identified in culture. Clostridium species are frequently found as contaminants in blood cultures[23], which is perhaps why mNGS also did not detect an unspeciated Clostridium isolated in blood culture from one patient.

With respect to non-BSI sepsis, our findings suggest that plasma mNGS may be most useful for identifying UTI-associated pathogens, although we also observed some utility for respiratory pathogen detection, in line with a prior report[24]. mNGS failed to detect *C. difficile* in any patients with colitis from this pathogen, although this is not surprising given that the

organism is rarely associated with bacteremia[25].

Within the No-Sepsis group, 10/37 (27%) of patients had a pathogen detected by mNGS, each of which was validated to be present at statistically significantly higher levels than in background controls. Notably, 9/10 (90%) of the pathogens were gram negative enteric organisms, which may reflect gastrointestinal translocation of microbes, a well described phenomenon during critical illness[26]. In addition, all 10 of these patients had received antibiotics in the first day of study enrollment, so it is possible that sequences derived from non-viable organisms unable to grow in culture.

Plasma RNA sequencing alone performed poorly for detecting sepsis-associated respiratory viruses. Incorporation of a host-based viral classifier, however, markedly improved detection of clinically confirmed viral LRTI. The viral classifier predicted previously unrecognized viral infections in three patients with sepsis who did not undergo viral PCR testing during their hospitalizations. Prior work has demonstrated that different viral species elicit distinct host transcriptional signatures in the peripheral blood[27], suggesting that future studies could extend the RNA host viral classifier to identify specific viral pathogens, such as influenza or SARS-CoV-2, for which therapeutics exist. Future studies could additionally explore the use of targeted enrichment methods[28,29] to enhance detection of viral sepsis pathogen nucleic acid.

In line with prior reports[13], we found that viral sepsis has a unique host transcriptional signature characterized by expression of interferon and other signaling pathways. We also observed transcriptional differences based on whether sepsis was due to a bloodstream versus peripheral site infection, which was less expected. The host response in Sepsis^{BSI} patients was characterized by lower expression of genes related to CD28 signaling and T cell activation, and greater expression of genes related to antimicrobial peptides, defensins and G alpha signaling, compared to Sepsis$^{non\text{-}BSI}$ patients.

We found that detection of a pathogen alone was in many cases insufficient for sepsis diagnosis, but when combined with a host transcriptional profile, had promising diagnostic utility

and potential as a tool for infection rule-out. In addition to defining host signatures of sepsis from whole blood, we also found biologically relevant host transcripts in plasma. This may have direct clinical applications given that plasma mNGS is increasingly being used in hospitals for pathogen detection in patients with sepsis and other infectious diseases, with turnaround times of ≤48 hours.

Inappropriate antimicrobial use is a major challenge in the management of critical illness, and is often driven by the inability to rule-out infection in patients with systemic inflammatory diseases. Indeed, we found that clinical variables alone, including the qSOFA score, were unable to accurately distinguish patients with sepsis from those with non-infectious critical illnesses at the time of initial evaluation in the ED. In contrast, our proof-of-concept assessment of the integrated host + microbe mNGS model demonstrated 99% sensitivity across patients with microbiologically confirmed sepsis, and 78% specificity within the No-Sepsis group, which was comprised almost entirely of patients meeting the clinical definition of systemic inflammatory response syndrome[16].

Host/microbe mNGS may facilitate precision antimicrobial stewardship by discriminating sepsis from diverse types of non-infectious febrile inflammatory syndromes, ranging from autoimmune diseases to macrophage activation syndrome. We envision this assay being used at the time of ED presentation for all suspected sepsis patients, as an adjunct to blood cultures and other traditional microbiological testing.

Distinguishing true sepsis pathogens from environmental contaminants or human commensals (e.g., Gemella), is a challenge for both mNGS and traditional culture-based microbiologic methods. Concomitant assessment of a host-based metric offers an opportunity to determine whether the detected pathogen exists in the context of an immunological state consistent with infection. Considering this, host/microbe mNGS diagnostic classification could theoretically be more difficult in immunocompromised patients. Arguing against this, however, is prior work demonstrating accurate performance of a host/microbe mNGS pneumonia diagnostic

in an ICU cohort with a 40% prevalence of immunocompromised individuals[14].

Our study has several strengths, including the novel use of plasma RNA transcriptomics for sepsis diagnosis, development of the first sepsis diagnostic combining host and microbial mNGS data, detailed clinical phenotyping, and a large prospective cohort of critically ill adults with systemic illnesses. It also has some limitations. First, as noted above, mNGS and blood cultures were performed on different samples collected at different times, so the observed concordance with clinical microbiological testing may be an underestimate. Second, a significant fraction of plasma samples had insufficient host transcripts to permit gene expression analyses, leading to a smaller sample size for the plasma versus the whole blood cohorts. This limitation may be addressable in future studies by increasing the input plasma volume and thus RNA mass.

The host immune response during sepsis is dynamic, and thus the stage of infection at which gene expression is measured may influence accuracy of the classifier. While our study was cross-sectional in design, we attempted to control for this by sampling at a consistently early stage of critical illness, within the first 24 hours of ICU admission. Lastly, because we did not have access to any other sepsis studies with either plasma gene expression data or paired host and microbial mNGS data from blood, additional studies in an independent cohort will be needed to validate these findings.

In conclusion, we report that combining host gene expression profiling and metagenomic pathogen detection from plasma nucleic acid enables accurate diagnosis of sepsis. Future studies are needed to validate and test the clinical impact of this culture-independent diagnostic approach.

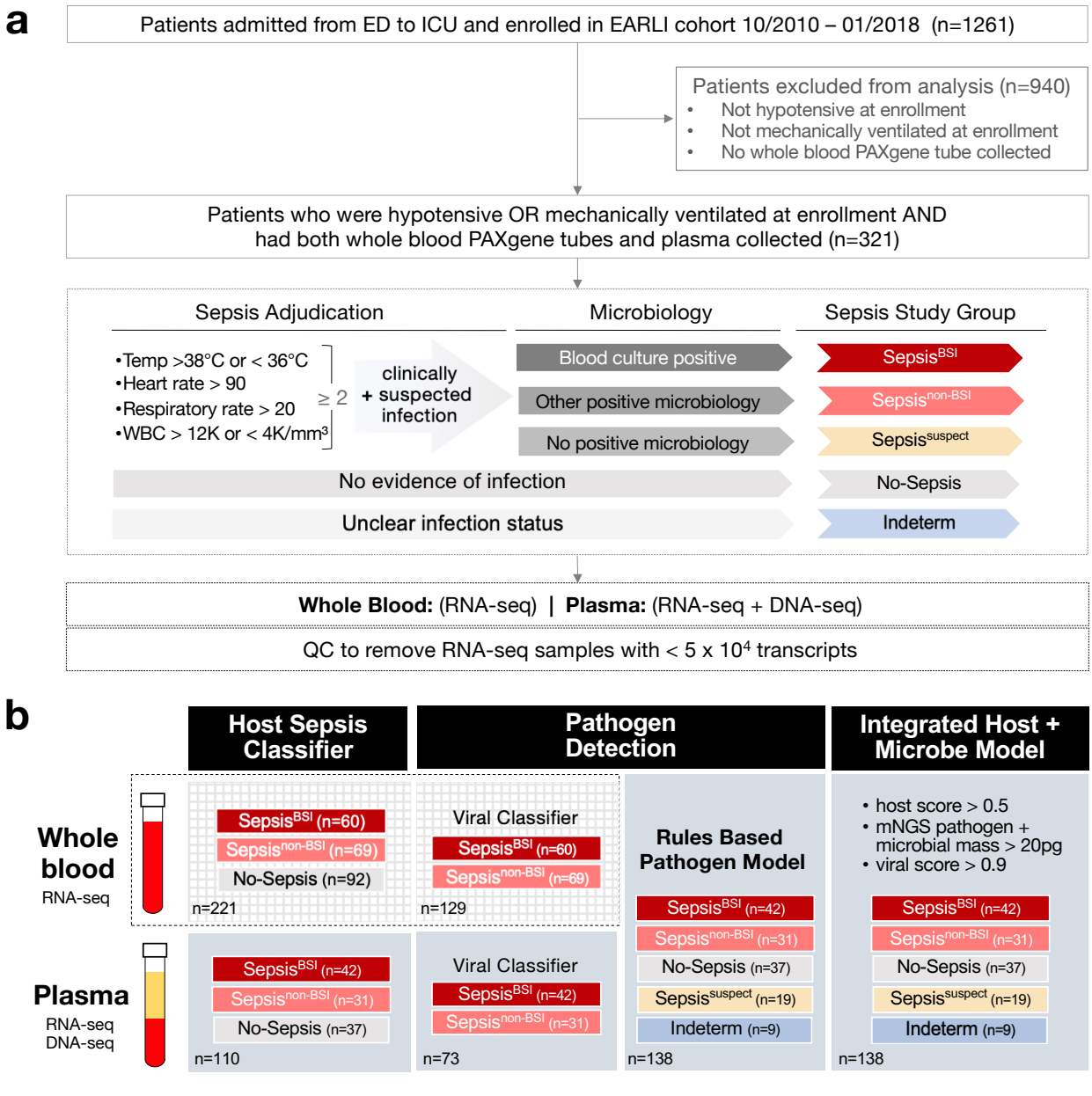
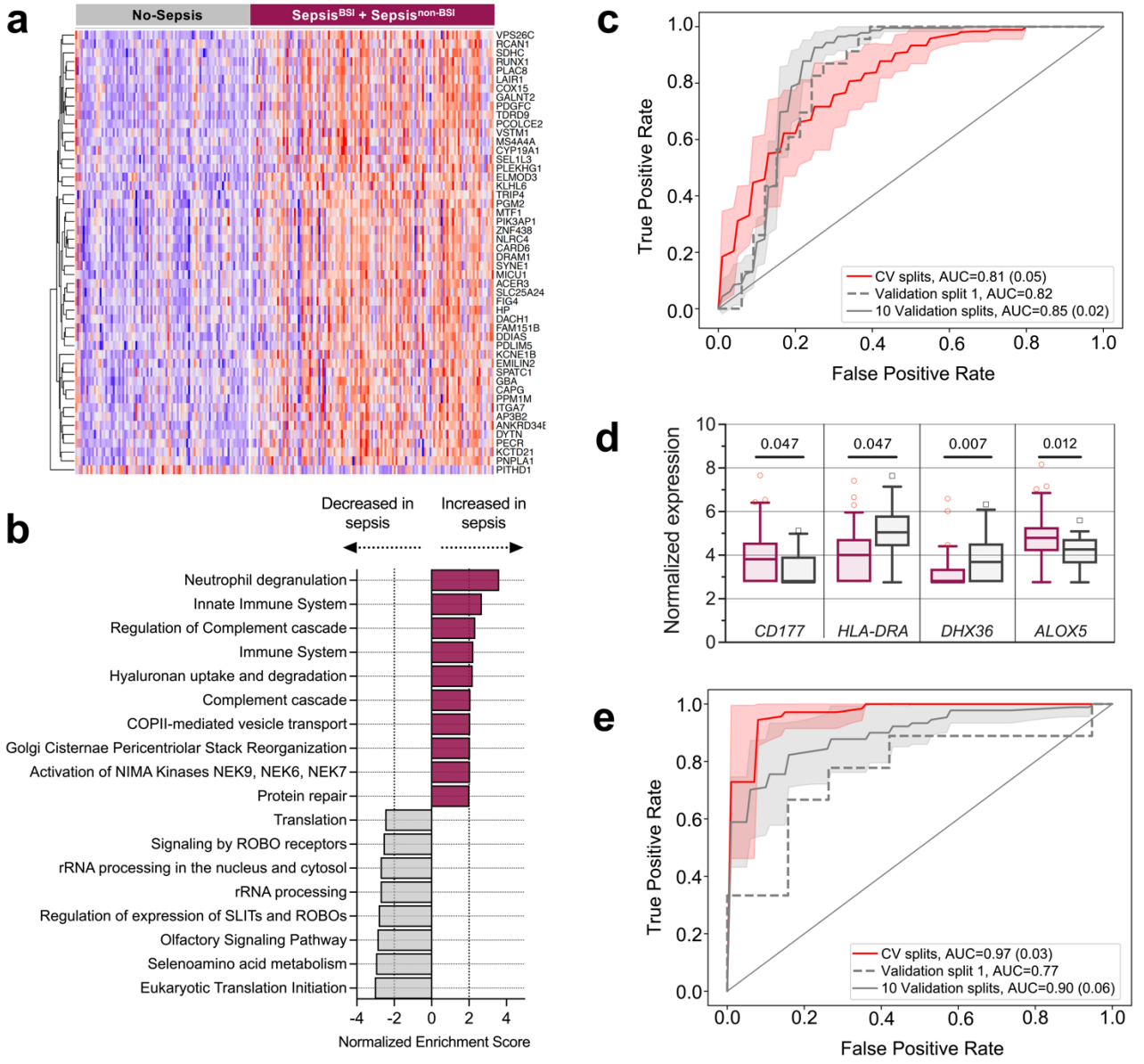


Figure 1. **a)** Study flow diagram. Patients studied were enrolled in the Early Assessment of Renal and Lung Injury (EARLI) cohort. Sepsis adjudication performed following hospital discharge was based on ≥ 2 or systemic inflammatory response syndrome (SIRS) criteria plus clinical suspicion of infection, and was used to delineate 5 patient subgroups. Following quality control (QC), whole blood underwent RNA-seq and plasma underwent RNA-seq and DNA-seq. **b)** Analytic approaches. Host transcriptional sepsis diagnostic classifiers were trained and tested on RNA-seq data from whole blood (n=221) and plasma (n=110), with a goal of differentiating patients with microbiologically confirmed sepsis ($Sepsis^{BSI}$ + $Sepsis^{non\text{-}BSI}$) from those without clinical evidence of infection (No-Sepsis). Viral infections were identified via a secondary host transcriptomic classifier. Sepsis pathogens were detected from plasma nucleic acid using metagenomic next generation sequencing (mNGS) followed by a rules-based bioinformatics model (RBM). Finally, an integrated host + microbe model for sepsis diagnosis was developed and evaluated.



375

Figure 2. Host gene expression differentiates patients with sepsis from those with non-infectious critical illnesses. **a)** Heatmap of top 50 differentially expressed genes from whole blood transcriptomics comparing patients with microbiologically confirmed sepsis ($Sepsis^{BSI}$ + $Sepsis^{non\text{-}BSI}$) versus those without evidence of infection (No-Sepsis). **b)** Gene set enrichment analysis of the differentially expressed genes with the top 10 up- and down-regulated pathways ($P < 0.05$) highlighted. **c)** Receiver operating characteristic (ROC) curve demonstrating performance of bagged support vector machine (bSVM) classifier for sepsis diagnosis from whole blood transcriptomics (n=221). The area under the ROC curve (AUC) and standard deviation (SD, in parentheses, when applicable) are listed in the figure panel for cross validation (CV) in the training set (red line: average over 10 random splits; red shaded area: ±1SD), the held-out validation set (dashed grey line), and over 10 randomly-generated validation sets (solid grey line: average; grey shaded area: ±1SD). **d)** Plasma RNA-seq expression differences of selected differentially expressed genes previously identified as sepsis biomarkers, with Sepsis patients in maroon, and No-Sepsis patients in grey. Adjusted P value provided above boxplot. Boxes represent the 25-75 percentiles and whiskers represent the 5-95 percentiles. **e)** ROC curve demonstrating performance of bSVM classifier for sepsis diagnosis from plasma RNA (n=110). The AUC and SD are listed in the figure panel for CV in the training set (red line: average over 10 random splits; red shaded area: average ±1SD), the held-out validation set (dashed grey line), and over 10 randomly-generated validation sets (solid grey line: average; grey shaded area: average ±1SD).

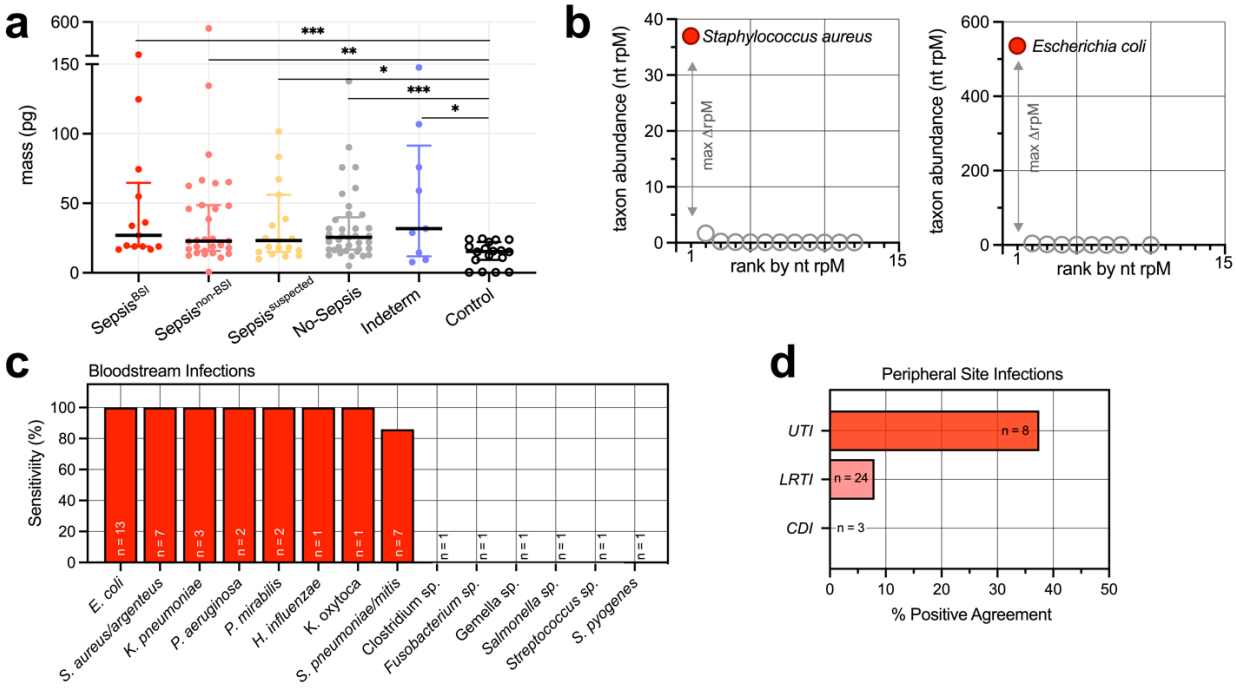


Figure 3. Plasma metagenomic next generation sequencing (mNGS) for detecting sepsis pathogens. **a)** Microbial plasma DNA mass differences between sepsis groups. Black bars represent median and error bars represent the interquartile range. **b)** Graphical depiction of the rules-based model (RBM) for sepsis pathogen detection from two different exemplary cases. The RBM identifies established pathogens with disproportionately high abundance compared to other commensal and environmental microbes in the sample. **c)** Concordance between plasma DNA mNGS for detecting bacterial pathogens in Sepsis^{BSI} patients with bacterial bloodstream infections compared to a gold standard of culture. **d)** Sensitivity of plasma nucleic acid mNGS for detecting pathogens in Sepsis$^{non\text{-}BSI}$ patients with sepsis from non-bloodstream, peripheral sites of infection. Legend: LRTI = lower respiratory tract infection; UTI = urinary tract infection; CDI = *Clostridium difficile* infection. Mass data are tabulated in Supplementary Data 8. Clinical microbiology and metagenomics data are tabulated in Supplementary Data 9. *** = P ≤ 0.001, ** = P ≤ 0.01, * = P ≤ 0.05.

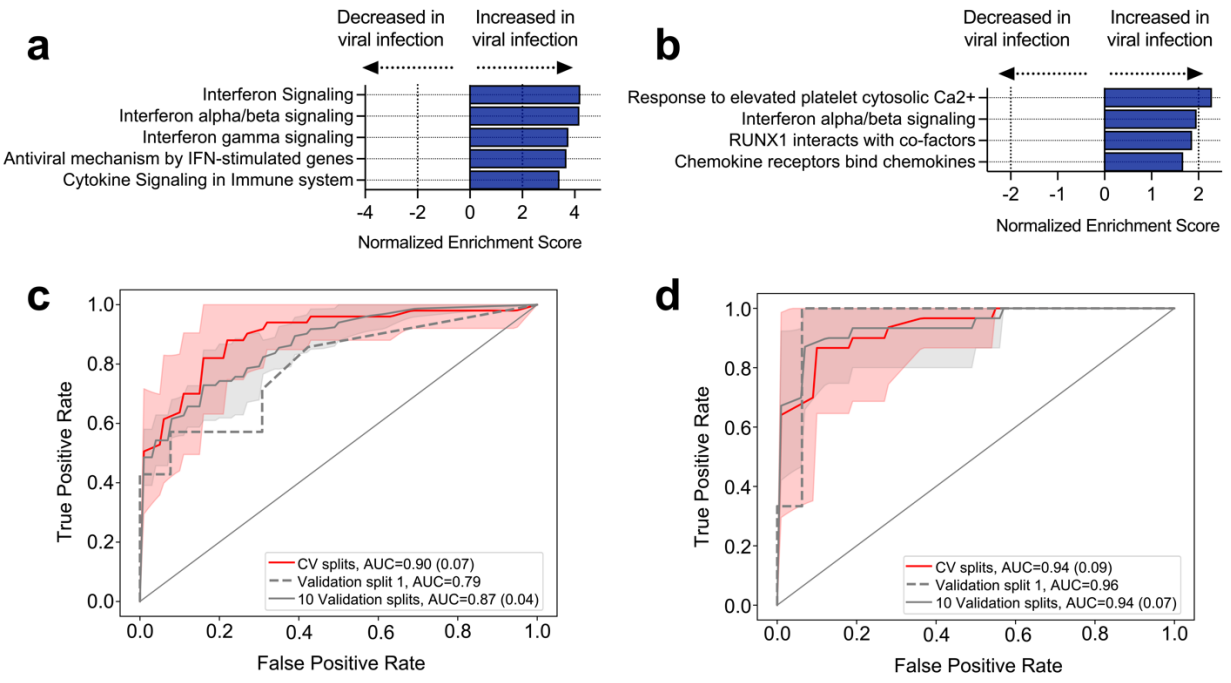
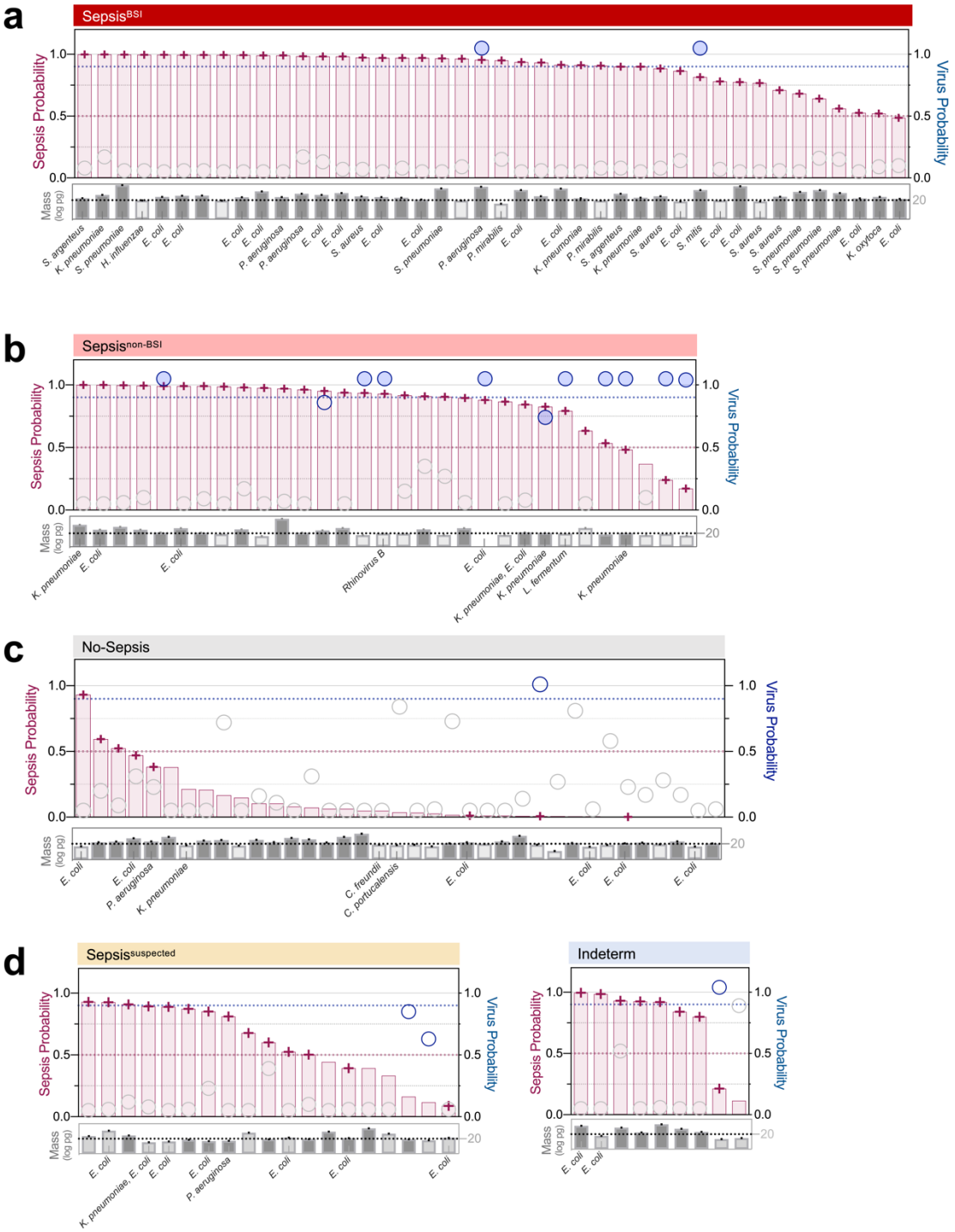


Figure 4. Detection of viral sepsis based on host gene expression. **a)** GSEA of differentially expressed genes from whole blood RNA-seq (n=129) demonstrating pathways enriched in patients with viral sepsis. Gene sets with $P < 0.05$ included. **b)** GSEA of differentially expressed genes from plasma RNA-seq (n=73) demonstrating pathways enriched in patients with viral sepsis. Gene sets with $P < 0.05$ included. **c)** ROC curve demonstrating performance of bagged support vector machine (bSVM) classifier for detecting viral sepsis from whole blood RNA-seq (n=129). The area under the ROC curve (AUC) and standard deviation (SD, in parentheses, when applicable) are listed in the figure panel for cross validation (CV) in the training set (red line: average over 10 random splits; red shaded area: ±1SD), the held-out validation set (dashed grey line), and over 10 randomly generated validation sets (solid grey line: average; grey shaded area: ±1SD). **d)** ROC curve demonstrating performance of bSVM classifier for detecting viral sepsis from plasma RNA-seq (n=73). The AUC and SD are listed in the figure panel for CV in the training set (red line: average over 10 random splits; red shaded area: ±1SD), the held-out validation set (dashed grey line), and over 10 randomly generated validation sets (solid grey line: average; grey shaded area: ±1SD).



425

Figure 5. Integrated host-microbe mNGS model for sepsis diagnosis from plasma. Host criteria for positivity can be met by a sepsis transcriptomic classifier probability > 0.5 (maroon bars, dotted line). Microbial criteria can be met based on either: 1) detection of a pathogen by mNGS *and* a sample microbial mass > 20 pg (grey bars), or 2) viral transcriptomic classifier probability > 0.9 (blue circles, dotted line). Host and microbial metrics are highlighted for patients with sepsis due to **a)** bloodstream infections ($Sepsis^{BSI}$), **b)** peripheral infection ($Sepsis^{non\text{-}BSI}$), **c)** patients with non-infectious critical illness (No-Sepsis), and **d)** patients with suspected sepsis but negative microbiological testing ($Sepsis^{suspected}$) and patients with indeterminant sepsis status (Indeterm). Maroon cross: sepsis positive based on model. Blue circles: virus predicted from plasma RNA secondary viral host classifier. Filled blue circles = virus also detected by clinical respiratory viral PCR. Cases with < 20pg microbial mass indicated by lighter grey shading. Samples with mNGS-detected pathogens have the microbe(s) listed below the sample microbial mass. Raw values for plots and original training/test split assignments are tabulated in Supplementary Data 16.

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 in the Early Assessment of Renal and Lung Injury (EARLI) cohort at the University of California, San Francisco (UCSF) or Zuckerberg San Francisco General Hospital between 10/2010 and 01/2018 (**Supplementary Table 1**). The study was approved by the UCSF Institutional Review Board (IRB) under protocol 10-02852, 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[30,31].

For the parent EARLI cohort, the inclusion criteria are: 1) age ≥ 18, 2) admission to the ICU from the ED, and 3) enrollment in the ED or within the first 24 hours of ICU admission. For this study, we selected patients for whom PAXgene whole blood tubes and matched plasma samples from the time of enrollment were available. PAXgene tubes were collected on patients enrolled in EARLI during the time period listed above who were hypotensive and/or mechanically ventilated at the time of enrollment. The main exclusion criteria for the EARLI study are: 1) exclusively neurological, neurosurgical, or trauma surgery admission, 2) goals of care decision for exclusively comfort measures, 3) known pregnancy, 4) legal status of prisoner, and 5) anticipated ICU length of stay < 24 hours. Enrollment in EARLI began in 10/2008 and continues.

Sepsis adjudication

Clinical adjudication of sepsis groups was carried out by study team physicians (MA, CL, AL, KL, PS, CH, AG, CC, KK, MM) using the sepsis-2 definition[32] (≥ 2 SIRS criteria + suspected infection) and incorporating all available clinical and microbiologic data from the entire ICU admission, with blinding to mNGS results. Each patient was reviewed by at least four

physicians. Disagreements were handled by discussion with the most senior physicians (CC, MM) in the phenotyping panel. Patients were categorized into five subgroups based on sepsis status (**Figure 1a, Supplementary Figure 1**). Patients with clinically adjudicated sepsis and a bacterial culture-confirmed bloodstream infection ($Sepsis^{BSI}$), sepsis due to a microbiologically confirmed primary infection at a peripheral site other than the bloodstream ($Sepsis^{non\text{-}BSI}$), suspected sepsis with negative clinical microbiologic testing ($Sepsis^{suspected}$), patients with no evidence of sepsis and a clear alternative explanation for their critical illness (No-Sepsis), or patients of indeterminant status (Indeterm). Clinical and demographic features of patients are summarized in (**Supplementary Table 1**) and tabulated in (**Supplementary Data 16 and 17**).

Metagenomic sequencing

Following enrollment, whole blood and plasma were collected in PAXgene and EDTA tubes, respectively. Whole blood PAXgene tubes were processed and stored at -80C according to manufacturer's instructions, and plasma was frozen at -80C within two hours. To evaluate host gene expression and detect microbes, RNA-seq was performed on the whole blood and plasma specimens, and DNA-seq was performed on plasma specimens. RNA was extracted from whole blood using the Qiagen RNeasy kit and normalized to 10ng total input per sample. Total plasma nucleic acid was extracted by first clarifying 300uL of plasma via maximum-speed centrifugation for five minutes at 21,300 x g, and then employing the Zymo Pathogen Magbead Kit on the supernatant following manufacturer's instructions. 10ng of total nucleic acid underwent DNA-seq using the NEBNext Ultra II DNA Kit. Samples with at least 10ng of remaining total nucleic acid were treated with DNAse (Qiagen) to recover RNA, and then underwent RNA-seq library preparation using the NEBNext Ultra II RNA-seq Kit as described below.

For RNA-seq library preparation, human cytosolic and mitochondrial ribosomal RNA and globin RNA was first depleted using FastSelect (Qiagen). For the purposes of background

contamination correction (see below) and to enable estimation of input microbial mass, we included negative water controls as well as positive controls (spike-in RNA standards from the External RNA Controls Consortium (ERCC))[33]. RNA was then fragmented and underwent library preparation using the NEBNext Ultra II RNA-seq Kit (New England Biolabs) according to described methods . Finished libraries underwent 146 nucleotide paired-end Illumina sequencing on an Illumina Novaseq 6000 instrument.

Index swapping can lead to read misassignment with Illumina sequencing. Dual indexing, that is adding barcode index sequences on both ends of the molecule, reduces the rate at which this misassignment occurs by requiring concordance between the two barcode sequences. The frequency of index-swapped reads has been estimated to be more than 35X lower when using dual vs single indexing[35]. Because we used dual indexing and because the RBM for pathogen detection operates by only identifying pathogen sequences disproportionately abundant in a sample versus the other sequences, our methods would not be expected to be negatively influenced by index swapping, which would only be anticipated to misassign low abundance reads irrelevant to the RBM.

Host differential expression and pathway analysis

Following demultiplexing, sequencing reads were aligned with STAR[36] 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. Samples retained in the dataset had a total of at least 50,000 counts associated with transcripts of protein coding genes.

Differential expression analysis was performed using DESeq2 and including covariates for age and gender. Significant genes were identified using an independent-hypothesis-weighted, Benjamini-Hochberg false discovery rate (FDR) < 0.1[38,39]. We generated heatmaps of the top 50 differentially expressed genes by absolute $\log_2$-fold change. To evaluate signaling

pathways from gene expression data, we employed gene set enrichment analysis using WebGestalt[40] on all ranked differentially expressed genes with a P value < 0.1. Significant pathways and upstream regulators were defined as those with a gene set P value < 0.05.

Pathogen detection

Detection of microbes leveraged the open-source IDseq pipeline[22] which incorporates subtractive alignment of the human genome (NCBI GRC h38) using STAR[36], quality and complexity filtering, and subsequent removal of cloning vectors and phiX phage using Bowtie2[22]. The identities of the remaining microbial reads are determined by querying the NCBI nucleotide (NT) database using GSNAP-L[22,41]. After background correction (see below), retained non-viral taxonomic alignments in each sample were aggregated at the genus level, and sorted in descending order by abundance measured in reads per million (rpM), independently for each sample. A previously validated rules based model (RBM)[14] was then utilized to identify disproportionately abundant bacteria and fungi in each sample, and flag them as pathogens. The RBM, originally developed to identify pathogens from respiratory mNGS data, detects outlier organisms within a sample by identifying the greatest gap in abundance between the top 15 sequentially ranked microbes in each sample. All microbes present in a reference index of established pathogens above this gap are then called by the RBM.

We adapted the original RBM specifically for sepsis pathogen detection, in which outlier organisms are sometimes present in low abundance, by incorporating a sepsis (as opposed to respiratory) pathogen reference index (**Supplementary Data 18**) and requiring that the species called by the RBM both be present in the reference index and detected at an abundance of > 1 rpM. Given the potential for respiratory viruses to cause sepsis, the RBM also identified human pathogenic respiratory viruses derived from a reference list of LRTI pathogens[14], present in the plasma RNA-seq data at an abundance of > 1 rpM. Sensitivity and specificity were calculated based on detection of reference index sepsis pathogens in each of the sepsis adjudication

groups.

The reference index (**Supplementary Data 18**) was established *a. priori* and no data from the enrolled patients were used to inform the distinction between pathogens and commensals. The index consisted of the most prevalent bloodstream infection pathogens reported by both the National Healthcare Safety Network (NHSN)[42] and a recent multicenter surveillance study of healthcare-associated infections[43]. These studies reported multiple species of Bacteriodes, Candida, Citrobacter, Enterobacter, Enterococcus, Klebsiella, Lactobacillus, Morganella, Prevotella, Proteus, Serratia, Stenotrophomonas and Streptococcus as common sepsis pathogens, and thus the reference index contains all species within these genera, yielding > 1000 total species detectable by the model based on current NCBI taxonomy.

Identification and mitigation of environmental contaminants

Negative control samples consisting of only double-distilled water (n=24) were processed alongside plasma DNA samples, which were sequenced in a single batch. Negative control samples enabled estimation of the number of background reads expected for each taxon[44]. A previously developed negative binomial model[44] was employed to identify taxa with NT sequencing alignments present at an abundance significantly greater compared to negative water controls. This was done by modeling the number of background reads as a negative binomial distribution, with mean and dispersion fitted on the negative controls. For each taxon, we estimated the mean parameter of the negative binomial by averaging the read counts across all negative controls. We estimated a single dispersion parameter across all taxa, using the functions glm.nb() and theta.md() from the R package MASS[45]. Taxa that achieved an adjusted P value <0.01 (Benjamini & Hochberg multiple test correction) were carried forward to the above-described RBM for pathogen detection.

563 Microbial mass calculations

564 Microbial mass was calculated based on the ratio of microbial reads in each sample to
565 total reads aligning to the External RNA Controls Consortium (ERCC) RNA standards spiked
566 into each sample[46]. The following equation was utilized for this calculation: [ERCC input
567 mass]/[microbial input mass] = [ERCC reads]/[microbial reads], where the ERCC input mass
568 was 25pg.

569 Host transcriptional classifiers for sepsis and viral infection diagnosis

570 To build classifiers that differentiated patients with sepsis ($Sepsis^{BSI}$, $Sepsis^{non\text{-}BSI}$) from
571 those with non-infectious critical illness (No-Sepsis), and distinguished viral from non-viral
572 sepsis, we built a Support Vector Machine (SVM)-based classifier[47] with the scikit-learn[48]
573 (v0.23.2) library in Python (v3.8.3). We tested several machine learning approaches (bagged
574 SVM, random forest and gradient boosted trees) and selected a bSVM classifier with a linear
575 kernel based on best performance (**Supplementary Table 2**). Each classifier used a
576 bootstrapped set of samples and a random subset of features.

577 We evaluated samples with ≥ 50,000 plasma gene counts and genes with more than
578 20% non-zero counts in that sample subset. Only differentially expressed genes, identified using
579 DESeq2 (v1.28.1) in the training set, were considered as potential predictors and included in
580 machine learning models, with FDR thresholds of 0.1 (whole blood), 0.2 (plasma, viral) and 0.3
581 (plasma, sepsis) chosen based on cross-validation. Age and sex were included as covariates in
582 the models. We used Z-score-scaled transformed (variance stabilizing transformation) gene
583 counts. 75% of the data was selected to train the model, and the rest was used as a held-out
584 set to test the final model. The training set was subsequently randomly split ten times for cross-
585 validation, using 75% of each as intermediate training sets, and the remaining 25% as their
586 associated testing sets.

On each one of those intermediate training sets, we carried out feature selection and parameters optimization using nested 5-fold cross-validations. We optimized three parameters: the regularization parameter, the maximum number of features considered for each classifier, and the total number of classifiers to use for bagging. For each parameters optimization fold, a recursive feature elimination (RFE) strategy was adopted, dropping 10% of the remaining least important features at each iteration. A bSVM classifier with default parameters was built at each iteration. We defined feature importance as the average squared weight across all estimators. To maximize interpretability, we restricted the maximum number of predictors to 100 genes.

We estimated model performances using the Area Under the Receiver Operating Characteristic Curve (AUC) values. To obtain a single set of features, we fitted a model, using the aforementioned strategy, to the initial training set. This model was then tested on the held-out set to obtain a final performance value and a single set of predictors.

Comparison of plasma nucleic acid mNGS against clinician-ordered diagnostic testing

Clinical microbiological testing was carried out based on decisions from the primary medical team during the patient's hospital admission at the UCSF and ZSFG clinical microbiology laboratories Tests utilized included bacterial culture from blood, lower respiratory tract and urine which were carried out according to previously described protocols[14]. Clinical testing for viral respiratory pathogens was performed from nasopharyngeal swabs and/or bronchioalveolar lavage using the Luminex XTag multiplex viral PCR assay. Polymicrobial blood cultures with ≥ 3 bacteria (n=2) were excluded from pathogen concordance given their unclear clinical significance and potential that some organisms reflected contamination.

Integrated host + microbe sepsis diagnosis and rule-out model

We developed a simple integrated host + microbe model that returned a sepsis diagnosis based on either host criteria `[host sepsis classifier probability > 0.5]` or microbial criteria: `[(pathogen detected by RBM) AND (microbial mass > 20 pg)] OR [host viral classifier`

612 `probability > 0.9]`. Combined metrics (**Supplementary Data 16**) including sepsis assignment
613 based on this model are depicted in Figure 5. Sensitivity was calculated in the ($Sepsis^{BSI}$) and
614 ($Sepsis^{non\text{-}BSI}$) groups, and specificity in the (No-Sepsis) group.
615
616 Clinical variable models for sepsis diagnosis
617 We tested the ability of clinical variables (**Supplementary Table 3a**) available at the
618 time of initial patient assessment to predict sepsis using three machine learning methods. These
619 included SVM using the e1071 package[49], random forest using the randomForest package[50]
620 and regularized logistic regression using the glmnet[51] package in R version 4.2.0[52]. Specifically,
621 we built models to classify Sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) versus No-Sepsis using 34
622 clinical variables that would be available at the time of ED evaluation. The data were split into
623 training (75%) and test (25%) sets and model performance (AUC) was evaluated on the test set.
624 This was repeated for a total of 10 randomized splits with the AUC computed at each iteration.
625 AUC was also computed for the qSOFA score (systolic blood pressure < 100 mmHg, respiratory
626 rate > 22 breaths/minute, Glasgow Coma Scale < 13). Results are tabulated in (**Supplementary**
627 **Table 3b**).

628 Statistics and reproducibility
629 Statistical tests utilized for each analysis are described in the figure legends and in
630 further detail in each respective methods section. The number of patient samples analyzed for
631 each comparison are indicated in the figure legends. Data were generated from single
632 sequencing runs without technical replicates.

633 Data availability
634 Source data are provided with this paper. The processed genecount data are available
635 from the National Center for Biotechnology Information Gene Expression Omnibus database
636 under accession code GSE189403. The raw sequencing data are protected due to data privacy

restrictions from the IRB protocol governing patient enrollment, which protects the release of raw genetic sequencing data from those patients enrolled under a waiver of consent. To honor this, researchers who wish to obtain raw fastq files for the purposes of independently generating genecounts can contact the corresponding author (chaz.langelier@ucsf.edu) and request to be added to the IRB protocol. The raw fastq files with microbial sequencing reads are available from the Sequence Read Archive under BioProject ID: PRJNA783060.

Code availability

Code for the differential expression, classifier development and RBM can be found at: (https://github.com/lucile-n/plasma_classifiers).

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. *The Lancet* **395**, 200–211 (2020).
2. Liu, V. *et al.* Hospital Deaths in Patients With Sepsis From 2 Independent Cohorts. *JAMA* **312**, 90 (2014).
3. Paul, M. *et al.* Systematic Review and Meta-Analysis of the Efficacy of Appropriate Empiric Antibiotic Therapy for Sepsis. *Antimicrob Agents Chemother* **54**, 4851–4863 (2010).
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*. *Critical Care Medicine* **42**, 1749–1755 (2014).
5. Novosad, S. A. *et al.* Vital Signs: Epidemiology of Sepsis: Prevalence of Health Care Factors and Opportunities for Prevention. *MMWR Morb. Mortal. Wkly. Rep.* **65**, 864–869 (2016).

6. Lamy, B., Roy, P., Carret, G., Flandrois, J. & Delignette-Muller, M. L. What Is the Relevance of Obtaining Multiple Blood Samples for Culture? A Comprehensive Model to Optimize the Strategy for Diagnosing Bacteremia. *CLIN INFECT DIS* **35**, 842–850 (2002).
7. Baur, D. *et al.* Effect of antibiotic stewardship on the incidence of infection and colonisation with antibiotic-resistant bacteria and Clostridium difficile infection: a systematic review and meta-analysis. *The Lancet Infectious Diseases* doi:10.1016/S1473-3099(17)30325-0.
8. Wilson, M. R. *et al.* Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N. Engl. J. Med.* **380**, 2327–2340 (2019).
9. Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nature Microbiology* **4**, 663–674 (2019).
10. Lee, R. A., Al Dhaheri, F., Pollock, N. R. & Sharma, T. S. Assessment of the Clinical Utility of Plasma Metagenomic Next-Generation Sequencing in a Pediatric Hospital Population. *J Clin Microbiol* **58**, (2020).
11. Hogan, C. A. *et al.* Clinical Impact of Metagenomic Next-Generation Sequencing of Plasma Cell-Free DNA for the Diagnosis of Infectious Diseases: A Multicenter Retrospective Cohort Study. *Clinical Infectious Diseases* **72**, 239–245 (2021).
12. Sweeney, T. E. *et al.* A community approach to mortality prediction in sepsis via gene expression analysis. *Nat Commun* **9**, 694 (2018).
13. Tsalik, E. L. *et al.* Host gene expression classifiers diagnose acute respiratory illness etiology. *Science Translational Medicine* **8**, 322ra11-322ra11 (2016).
14. Langelier, C. *et al.* Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. *Proc Natl Acad Sci USA* 201809700 (2018) doi:10.1073/pnas.1809700115.
15. Singer, M. *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 801 (2016).

16. Kaukonen, K.-M., Bailey, M., Pilcher, D., Cooper, D. J. & Bellomo, R. Systemic inflammatory response syndrome criteria in defining severe sepsis. *N. Engl. J. Med.* **372**, 1629–1638 (2015).
17. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).
18. Demaret, J. *et al.* Identification of CD177 as the most dysregulated parameter in a microarray study of purified neutrophils from septic shock patients. *Immunology Letters* **178**, 122–130 (2016).
19. Tang, B. M. *et al.* Neutrophils-related host factors associated with severe disease and fatality in patients with influenza infection. *Nat Commun* **10**, 3422 (2019).
20. Cajander, S. *et al.* Preliminary results in quantitation of HLA-DRA by real-time PCR: a promising approach to identify immunosuppression in sepsis. *Crit Care* **17**, R223 (2013).
21. Leijte, G. P. *et al.* Monocytic HLA-DR expression kinetics in septic shock patients with different pathogens, sites of infection and adverse outcomes. *Crit Care* **24**, 110 (2020).
22. Kalantar, K. L. *et al.* IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *Gigascience* **9**, (2020).
23. Hall, K. K. & Lyman, J. A. Updated Review of Blood Culture Contamination. *Clin Microbiol Rev* **19**, 788–802 (2006).
24. Langelier, C. *et al.* Detection of Pneumonia Pathogens from Plasma Cell-Free DNA. *Am. J. Respir. Crit. Care Med.* (2019) doi:10.1164/rccm.201904-0905LE.
25. Libby, D. B. & Bearman, G. Bacteremia due to Clostridium difficile—review of the literature. *International Journal of Infectious Diseases* **13**, e305–e309 (2009).
26. Frencken, J. F. *et al.* Associations Between Enteral Colonization With Gram-Negative Bacteria and Intensive Care Unit–Acquired Infections and Colonization of the Respiratory Tract. *Clinical Infectious Diseases* **66**, 497–503 (2018).

27. Mudd, P. A. *et al.* Distinct inflammatory profiles distinguish COVID-19 from influenza with limited contributions from cytokine storm. *Sci. Adv.* **6**, eabe3024 (2020).
28. Deng, X. *et al.* Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. *Nature Microbiology* (2020) doi:10.1038/s41564-019-0637-9.
29. Quan, J. *et al.* FLASH: a next-generation CRISPR diagnostic for multiplexed detection of antimicrobial resistance sequences. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz418.
30. Auriemma, C. L. *et al.* Acute respiratory distress syndrome-attributable mortality in critically ill patients with sepsis. *Intensive Care Med* **46**, 1222–1231 (2020).
31. Agrawal, A. *et al.* Plasma angiopoietin-2 predicts the onset of acute lung injury in critically ill patients. *Am J Respir Crit Care Med* **187**, 736–742 (2013).
32. Levy, M. M. *et al.* 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Crit Care Med* **31**, 1250–1256 (2003).
33. Pine, P. S. *et al.* Evaluation of the External RNA Controls Consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnology* **16**, 54 (2016).
34. Mick, E. *et al.* Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses. *Nature Communications* **11**, (2020).
35. Wilson, M. R. *et al.* Multiplexed Metagenomic Deep Sequencing To Analyze the Composition of High-Priority Pathogen Reagents. *mSystems* **1**, e00058-16 (2016).
36. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
38. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* **13**, 577–580 (2016).

39. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

40. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research* **47**, W199–W205 (2019).

41. Zhao, Y., Tang, H. & Ye, Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**, 125–126 (2012).

42. Weiner-Lastinger, L. M. *et al.* Antimicrobial-resistant pathogens associated with adult healthcare-associated infections: Summary of data reported to the National Healthcare Safety Network, 2015–2017. *Infect. Control Hosp. Epidemiol.* **41**, 1–18 (2020).

43. Magill, S. S. *et al.* Changes in Prevalence of Health Care–Associated Infections in U.S. Hospitals. *New England Journal of Medicine* **379**, 1732–1744 (2018).

44. Mick, E. *et al.* Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses. *Nature Communications* **11**, 5854 (2020).

45. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer-Verlag, 2002). doi:10.1007/978-0-387-21706-2.

46. Pine, P. S. *et al.* Evaluation of the External RNA Controls Consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnol* **16**, 54 (2016).

47. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).

48. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

49. Meyer, D. *et al.* Package 'e1071'. https://cran.r-project.org/web/packages/e1071/e1071.pdf (2022).

50. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).

762 51. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models
763 via Coordinate Descent. *Journal of Statistical Software, Articles* **33**, 1–22 (2010).
764 52. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for
765 Statistical Computing, 2020).
766

Supplementary Materials

Supplementary Table 1a. Summary of clinical and demographic features of patients evaluated in whole blood gene expression analyses (n=221). These include patients with microbiologically confirmed sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) and those with non-infectious critical illnesses (No-Sepsis). Source data are tabulated in (Supplementary Data 16).

Whole Blood		**$Sepsis^{BSI}$** (n=60)	**$Sepsis^{non\text{-}BSI}$** (n=69)	**No-Sepsis** (n=92)	**Sepsis** vs **No-Sepsis** P value*
Age (median, Q1-Q3/%)		63.5 (50.8-73.3)	68 (58-80)	65.5 (54-74.5)	0.51
Gender (median, %)	Male	41 (68.3%)	40 (58%)	53 (57.6%)	0.49
	Female	19 (31.7%)	28 (40.6%)	39 (42.4%)	
	Transgender	0 (0%)	1 (1.4%)	0 (0%)	
Race (n, %)	Caucasian	20 (33.3%)	29 (42%)	37 (40.2%)	0.63
	Asian	17 (28.3%)	21 (30.4%)	24 (26.1%)	
	African American	10 (16.7%)	11 (15.9%)	17 (18.5%)	
	Other	12 (20%)	5 (7.2%)	14 (15.2%)	
	Unknown	0 (0%)	3 (4.3%)	0 (0%)	
	Native American	1 (1.7%)	0 (0%)	0 (0%)	
Ethnicity (n, %)	Non-LatinX	51 (85%)	60 (87%)	79 (85.9%)	0.76
	LatinX	9 (15%)	6 (8.7%)	12 (13%)	
	NA	0 (0%)	3 (4.3%)	1 (1.1%)	
Temp Max (median, Q1-Q3%)		38.1 (37.1-38.9)	37.6 (37-38.4)	37.3 (36.1-37.9)	< 0.001
WBC Max (median, Q1-Q3%)		15.5 (10.1-23.8)	13.9 (9.5-19.2)	13.15 (9.5-18.9)	0.24
APACHEIII (median, Q1-Q3%)		69 (53-85)	50 (46-66)	64 (50-78)	0.17
SIRS	4	39 (65%)	26 (37.7%)	27 (29.3%)	0.01
	3	16 (26.7%)	30 (43.5%)	42 (45.7%)	
	2	5 (8.3%)	12 (17.4%)	22 (23.9%)	
	1	0 (0%)	0 (0.0%)	1 (1.1%)	
Bacterial infection‡		60 (100.0%)	50 (72.5%)	0 (0.0%)	< 0.001
Viral +/- Bacterial infection‡		3 (5.0%)	21 (30.4%)	0 (0.0%)	< 0.001
28-day mortality (n, %)		23 (38.3%)	15 (21.7%)	32 (34.8%)	0.49
Intubated (n, %)		47 (78.3%)	59 (85.5%)	86 (93.5%)	0.02
Vasopressors (n, %)		54 (90%)	51 (73.9%)	59 (64.1%)	< 0.01
Immunocompromised# (n, %)		7 (11.7%)	6 (8.7%)	6 (6.5%)	0.49
Antibiotics+ (n, %)		59 (98.3%)	66 (95.7%)	73 (79.3%)	< 0.001

*$Sepsis^{BSI}$ + $Sepsis^{non\text{-}BSI}$ vs No-Sepsis P value calculated by Mann-Whitney (continuous) or chi-squared (categorical)
‡Based on clinical microbiology testing
#Immunocompromise was defined as: history of solid organ transplantation, bone marrow transplantation, HIV/AIDS with CD4 < 200, leukemia or other hematologic malignancy, autoimmune inflammatory disease, or primary immunodeficiency.
+Antibiotics administered on or before the first day of study enrollment.

Supplementary Table 1b. Summary of clinical and demographic features of patients with plasma RNA-seq data, evaluated in all analyses (n=138). All sepsis adjudication groups represented. Source data are tabulated in (Supplementary Data 17).

Plasma		**Sepsis^{BSI} (n=42)** Median/n (Q1-Q3/%)	**Sepsis$^{non-BSI}$ (n=31)** Median/n (Q1-Q3/%)	**No-Sepsis (n=37)** Median/n (Q1-Q3/%)	P value* Sepsis$^{BSI+non-BSI}$ v No-Sepsis	**Sepsis^{suspected} (n=19)** Median/n (Q1-Q3/%)	**Indeterm (n=9)** Median/n (Q1-Q3/%)	Overall P value†
Age (median, Q1-Q3/%)		63.5 (51.25-72)	69 (58-78.5)	66 (55-79)	0.45	69 (55-80)	72 (63-80)	0.12
Gender (median, %)	Male	28 (66.7%)	17 (54.8%)	21 (56.8%)	0.77	9 (47.4%)	3 (33.3%)	0.24
	Female	14 (33.3%)	14 (45.2%)	16 (43.2%)		9 (47.4%)	6 (66.7%)	
	Transgender	0 (0%)	0 (0%)	0 (0%)		1 (5.3%)	0 (0%)	
Race (n, %)	Caucasian	13 (31%)	12 (38.7%)	14 (37.8%)	0.86	4 (21.1%)	7 (77.8%)	0.50
	Asian	12 (28.6%)	10 (32.3%)	13 (35.1%)		9 (47.4%)	1 (11.1%)	
	African American	8 (19%)	6 (19.4%)	7 (18.9%)		3 (15.8%)	1 (11.1%)	
	Other	8 (19%)	2 (6.5%)	3 (8.1%)		3 (15.8%)	0 (0%)	
	Native American	1 (2.4%)	0 (0%)	0 (0%)		0 (0%)	0 (0%)	
	Unknown	0 (0%)	1 (3.2%)	0 (0%)		0 (0%)	0 (0%)	
Ethnicity (n, %)	Non-LatinX	36 (85.7%)	28 (90.3%)	35 (94.6%)	0.48	16 (84.2%)	8 (88.9%)	0.62
	LatinX	6 (14.3%)	2 (6.5%)	2 (5.4%)		3 (15.8%)	1 (11.1%)	
	NA	0 (0%)	1 (3.2%)	0 (0%)		0 (0%)	0 (0%)	
Temp Max (median, Q1-Q3/%)		38.05 (37.1-38.9)	37.8 (37.1-38.5)	37.1 (36.1-38)	< 0.01	37 (36.6-38.2)	35.8 (34.9-37.2)	< 0.001
WBC Max (median, Q1-Q3/%)		16.05 (6.5-26.6)	17.4 (11.2-22.8)	13 (9.9-20.6)	0.40	19.3 (12.8-26.1)	15.3 (9.7-19.7)	0.60
APACHEIII (median, Q1-Q3/%)		72.5 (52-85)	54 (46-69)	72 (57-81)	0.20	70 (61-89)	72 (65-93)	0.04
SIRS	4	28 (66.7%)	13 (41.9%)	15 (40.5%)	0.19	12 (63.2%)	5 (55.6%)	0.52
	3	11 (26.2%)	15 (48.4%)	15 (40.5%)		5 (26.3%)	3 (33.3%)	
	2	3 (7.1%)	3 (9.7%)	6 (16.2%)		2 (10.5%)	1 (11.1%)	
	1	0 (0%)	0 (0%)	1 (2.7%)		0 (0%)	0 (0%)	
Bacterial infection‡		42 (100.0%)	24 (77.4%)	0 (0.0%)	< 0.001	0 (0.0%)	0 (0.0%)	1.00
Viral +/- Bacterial infection‡		2 (4.8%)	11 (35.5%)	0 (0.0%)	< 0.001	0 (0.0%)	0 (0.0%)	1.00
28-day mortality (n, %)		20 (47.6%)	8 (25.8%)	18 (48.6%)	0.41	11 (57.9%)	7 (77.8%)	0.04
Intubated (n, %)		32 (76.2%)	28 (90.3%)	35 (94.6%)	0.13	18 (94.7%)	9 (100%)	0.05
Vasopressors (n, %)		40 (95.2%)	23 (74.2%)	31 (83.8%)	0.95	18 (94.7%)	8 (88.9%)	0.08
Immunocompromised# (n, %)		6 (14.3%)	5 (16.1%)	4 (10.8%)	0.75	5 (26.3%)	2 (22.2%)	0.63
Antibiotics+ (n, %)		41 (97.6%)	31 (100.0%)	31 (83.8%)	< 0.01	19 (100.0%)	8 (88.9%)	0.02

*Mann-Whitney (continuous) /chi-squared (categorical) †Kruskal-Wallis (continuous) / chi-squared (categorical)
‡Based on clinical microbiology testing
#Immunocompromise was defined as: history of solid organ transplantation, bone marrow transplantation, HIV/AIDS with CD4 < 200, leukemia or other hematologic malignancy, autoimmune inflammatory disease, or primary immunodeficiency.
+Antibiotics administered on or before the first day of study enrollment.

Supplementary Table 2. Comparison of machine learning models for host-based sepsis classification. Area under the receiver operator characteristic curve (AUC) for three different machine learning models assessed for classifier construction. The AUC for cross-validation in the training set (standard deviation in parentheses) is listed first, and the AUC for the first validation split is listed second, after the vertical bar. The bagged support vector machine (bSVM) model performed best overall.

Classifier	Bagged Support Vector Machine	Random Forest	Gradient Boosted Tree
Whole blood - sepsis	0.81 (0.05) | 0.82	0.80 (0.06) | 0.86	0.79 (0.05) | 0.84
Plasma - sepsis	0.97 (0.03) | 0.77	0.76 (0.05) | 0.79	0.70 (0.09) | 0.82
Whole blood - viral	0.90 (0.07) | 0.79	0.83 (0.08) | 0.75	0.78 (0.07) | 0.69
Plasma - viral	0.94 (0.09) | 0.96	0.77 (0.09) | 0.81	0.72 (0.19) | 0.60

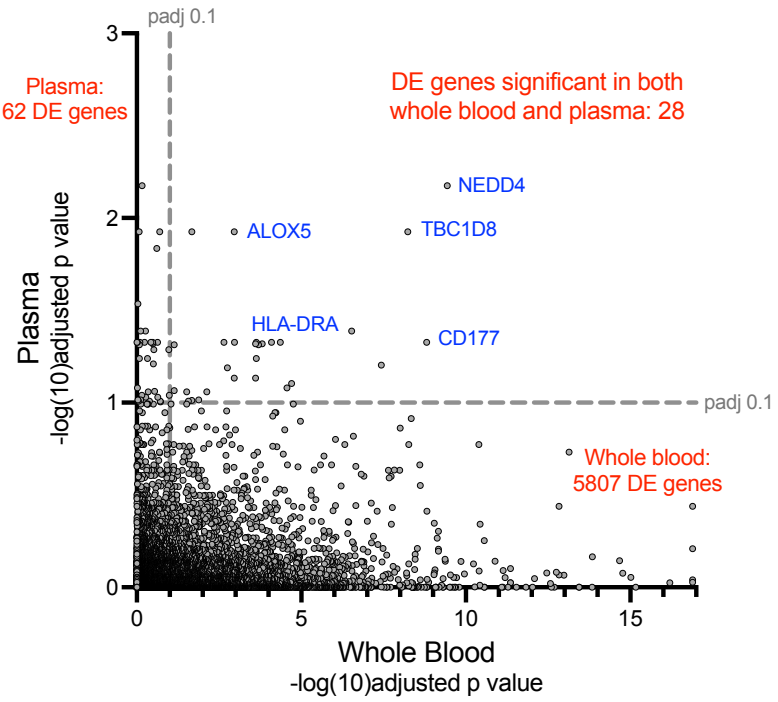
Supplementary Table 3a. Clinical variables used for classifier construction. Fever: > 38C, anemia: hemoglobin < 7, thrombocytopenia: platelets < 50, acute renal failure: creatinine > 2.0.

Anemia	Temp Max	Creatinine Max
Hyperkalemia	Fever	Creatinine Min
Hypokalemia	Temp Min	Acute Renal Failure
Hypernatremia	WBC Max	Platelets Min
Hyponatremia	WBC Min	Thrombocytopenia
Hypercalcemia	HR Max	Requirement for Intubation
Hypocalcemia	HR Min	Glasgow Coma Scale
Hyperthyroid	RR Max	Immunocompromise
Hypothyroid	RR Min	Chest Pain
Adrenal Insufficiency	SIRS total	Volume Overload
Hyperglycemia	SBP Min	Creatinine Max
Hypoglycemia	SBP Max	Creatinine Min

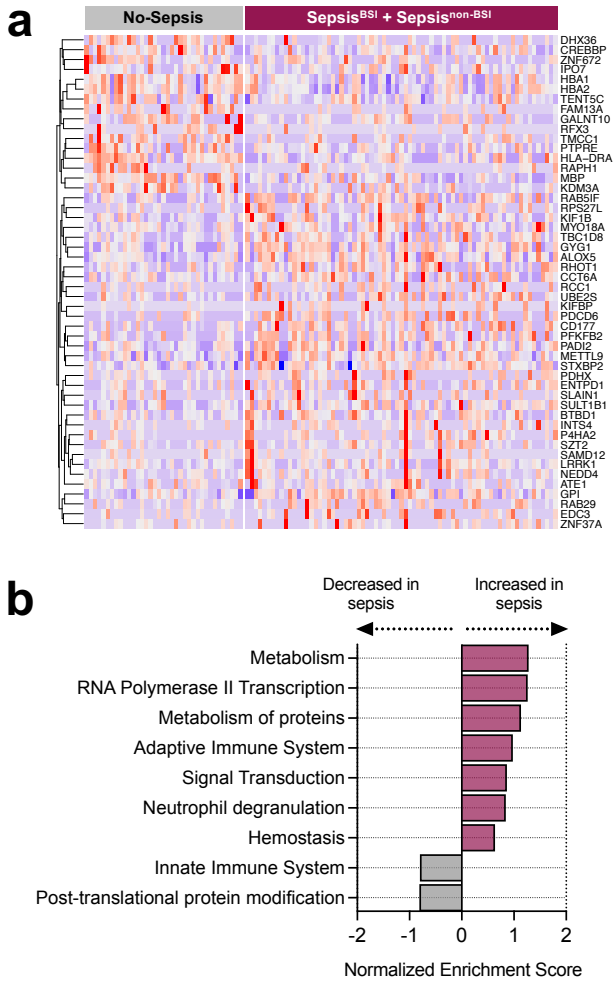
Supplementary Table 3b. Average classifier AUC over 10 iterations of training and test from different machine learning classifiers to distinguish Sepsis from No-Sepsis patients using clinical features alone. qSOFA score positive for sepsis if systolic blood pressure < 100 mmHg, respiratory rate > 22 breaths/minute and Glasgow Coma Scale < 13.

Method	AUC, mean (std)
Support vector machine	0.57 (0.04)
Random forest	0.62 (0.04)
Regularized logistic regression	0.57 (0.07)
qSOFA score	0.48 (0.02)

Supplementary Figures



Supplementary Figure 1. Overlap of significant genes in the differential expression analyses between the Sepsis and No-Sepsis groups for whole blood and plasma samples. Scatter plot of -log10(adjusted p-value) for individual genes from the differential expression analyses comparing patients with microbiologically confirmed sepsis ($Sepsis^{BSI}$ + $Sepsis^{non\text{-}BSI}$) versus those without evidence of infection (No-sepsis), from whole blood (x-axis) and plasma (y-axis). P-values derive from Benjamini-Hochberg adjustment. Dashed gray lines indicate the threshold of adjusted p-value < 0.1. Selected, significant, differentially expressed genes highlighted in blue.



Supplementary Figure 2. Plasma host gene expression differentiates patients with sepsis from those with non-infectious critical illnesses. **a)** Heatmap of top 50 differentially expressed genes from whole blood transcriptomics comparing patients with microbiologically confirmed sepsis ($Sepsis^{BSI}$ + $Sepsis^{non\text{-}BSI}$) versus those without evidence of infection (No-sepsis). **b)** Gene set enrichment analysis of the differentially expressed genes. All gene sets included.

Supplementary Data Files

Supplementary Data 1. Differentially expressed genes (adjusted P value < 0.1) between patients with microbiologically confirmed ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) and those with non-infectious critical illnesses (No-Sepsis), from whole blood RNA-seq.

Supplementary Data 2. Gene set enrichment analysis of differentially expressed genes between patients with microbiologically confirmed sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) and those with non-infectious critical illnesses (No-Sepsis). Data from whole blood RNA-seq. The top 10 positively and negatively enriched pathways by P value are included in table.

Supplementary Data 3. Differentially expressed genes (adjusted P value < 0.1) between patients with sepsis due to bloodstream infections ($Sepsis^{BSI}$) versus peripheral infections ($Sepsis^{non\text{-}BSI}$). Data from whole blood RNA-seq.

Supplementary Data 4. Gene set enrichment analysis of differentially expressed genes between patients with sepsis due to bloodstream infections ($Sepsis^{BSI}$) versus peripheral infections ($Sepsis^{non\text{-}BSI}$). Data from a) whole blood RNA-seq and b) plasma RNA-seq. The top 10 positively and negatively enriched pathways by P value are included in table.

Supplementary Data 5. a) Area under the receiver operating characteristic curve (AUC) values for 10 independent training set models for a whole blood gene expression support vector machine classifier to distinguish patients with microbiologically confirmed ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) from those with non-infectious critical illnesses (No-Sepsis). b) Composite list of all genes selected by each classifier model.

Supplementary Data 6. Differentially expressed genes (adjusted P value < 0.1) between patients with microbiologically confirmed ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) and those with non-infectious critical illnesses (No-Sepsis), from plasma RNA-seq.

Supplementary Data 7. a) AUC values for 10 independent training set models for a plasma gene expression support vector machine classifier to distinguish patients with microbiologically confirmed ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) from those with non-infectious critical illnesses (No-Sepsis). b) Composite list of all genes selected by each classifier model.

Supplementary Data 8. Mass (pg) of microbial DNA in each sample, calculated based on spiked-in 25 pg ERCC positive controls.

Supplementary Data 9. Sepsis pathogens detected by standard of care clinical microbiology versus plasma mNGS, using the rules-based model.

Supplementary Data 10. Differentially expressed genes (adjusted P value < 0.1) between patients with microbiologically confirmed viral sepsis and those with non-viral sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$ groups), from whole blood RNA-seq.

Supplementary Data 11. Differentially expressed genes (adjusted P value < 0.1) between patients with microbiologically confirmed viral sepsis and those with non-viral sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$ groups), from plasma RNA-seq.

Supplementary Data 12. Gene set enrichment analysis of differentially expressed genes between patients with viral versus non-viral causes of sepsis amongst the $Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$ patients. a) Data from whole blood RNA-seq. b) Data from plasma RNA-seq.

Supplementary Data 13. a) AUC values for 10 independent training set models for a whole blood gene expression support vector machine classifier to distinguish patients with microbiologically confirmed viral versus non-viral sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$), from whole blood RNA-seq. b) Composite list of all genes selected by each classifier model.

Supplementary Data 14. a) AUC values for 10 independent training set models for a plasma gene expression support vector machine classifier to distinguish patients with microbiologically confirmed viral versus non-viral sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$), from plasma RNA-seq. b) Composite list of all genes selected by each classifier model.

Supplementary Data 15. Complete integrated host-microbe mNGS dataset. This includes: per-sample classifier predictions for all patients with plasma sequencing data (n=138), including the sepsis diagnostic classifier and the viral sepsis classifier; pathogens detected by clinical diagnostics and by mNGS; and microbial mass per sample.

Supplementary Data 16. Clinical and demographic features of patients evaluated in whole blood gene expression analyses only (n=221). These include patients with microbiologically confirmed sepsis ($Sepsis^{BSI}$ and $Sepsis^{non\text{-}BSI}$) and those with (No-Sepsis).

Supplementary Data 17. Clinical and demographic features of patients with plasma RNA-seq data, evaluated in all analyses (n=138). All sepsis adjudication groups represented.

Supplementary Data 18. Reference index of established sepsis pathogens derived from the top 20 most prevalent sepsis pathogens reported by both the US CDC/ National Healthcare Safety Network[1] and a point prevalence survey of healthcare-associated infections[2].

Supplementary References

1. Weiner-Lastinger, L. M. *et al.* Antimicrobial-resistant pathogens associated with adult healthcare-associated infections: Summary of data reported to the National Healthcare Safety Network, 2015–2017. *Infect. Control Hosp. Epidemiol.* **41**, 1–18 (2020).
2. Magill, S. S. *et al.* Changes in Prevalence of Health Care–Associated Infections in U.S. Hospitals. *New England Journal of Medicine* **379**, 1732–1744 (2018).