nature microbiology

Article

https://doi.org/10.1038/s41564-022-01237-2

Integrated host-microbe plasma metagenomics for sepsis diagnosis in a prospective cohort of critically ill adults

Received: 10 March 2022

Accepted: 23 August 2022

Published online: 20 October 2022

Check for updates

Katrina L. Kalantar[1,11], Lucile Neyton[2,11], 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 A. Christenson[2], Kathleen Liu[5], Kirsten N. Kangelaris[6], Carolyn Hendrickson[3], Pratik Sinha[7], Antonio Gomez[8], Norma Neff[9], Angela Pisco[9], Sarah B. Doernberg[3], Joseph L. Derisi[9,10], Michael A. Matthay[2], Carolyn S. Calfee[2,11] and Charles R. Langelier[3,9,11] ✉

We carried out integrated host and pathogen metagenomic RNA and DNA next generation sequencing (mNGS) of whole blood ($n = 221$) and plasma ($n = 138$) from critically ill patients following hospital admission. We assigned patients into sepsis groups on the basis of clinical and microbiological criteria. From whole-blood gene expression data, we distinguished patients with sepsis from patients with non-infectious systemic inflammatory conditions using a trained bagged support vector machine (bSVM) classifier (area under the receiver operating characteristic curve (AUC) = 0.81 in the training set; AUC = 0.82 in a held-out validation set). Plasma RNA also yielded a transcriptional signature of sepsis with several genes previously reported as sepsis biomarkers, and a bSVM sepsis diagnostic classifier (AUC = 0.97 training set; AUC = 0.77 validation set). Pathogen detection performance of plasma mNGS varied on the basis of pathogen and site of infection. To improve detection of virus, we developed a secondary transcriptomic classifier (AUC = 0.94 training set; AUC = 0.96 validation set). We combined host and microbial features to develop an integrated sepsis diagnostic model that identified 99% of microbiologically confirmed sepsis cases, and predicted sepsis in 74% of suspected and 89% of indeterminate sepsis cases. In summary, we suggest that integrating host transcriptional profiling and broad-range metagenomic pathogen detection from nucleic acid is a promising tool for sepsis diagnosis.

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 aetiologic 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

A full list of affiliations appears at the end of the paper. ✉e-mail: chaz.langelier@ucsf.edu

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]. However, the clinical impact of plasma DNA metagenomics has been questioned due to frequent identification of microbes of uncertain clinical relevance, 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 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 (Fig. 1). Patients were categorized into five subgroups on the basis of sepsis status (Methods). These included patients with: (1) clinically adjudicated sepsis and a microbiologically confirmed bacterial bloodstream infection (Sepsis$^{\mathrm{BSI}}$), (2) clinically adjudicated sepsis and a microbiologically confirmed non-bloodstream infection (Sepsis$^{\text{non-BSI}}$), (3) suspected sepsis with negative clinical microbiologic testing (Sepsis$^{\text{suspected}}$), (4) patients with no evidence of sepsis and a clear alternative explanation for their critical illness (No-sepsis) or (5) patients of indeterminate status (Indeterm). The most common diagnoses in the No-sepsis group were cardiac arrest, overdose/poisoning, heart failure exacerbation and pulmonary embolism. The majority of patients, regardless of subgroup, required mechanical ventilation and vasopressor support (Supplementary Tables 1a,b). Patients with microbiologically proven sepsis (Sepsis$^{\mathrm{BSI}}$ + Sepsis$^{\text{non-BSI}}$) did not differ from No-sepsis patients in terms of age, gender, race, ethnicity, immunocompromise, APACHEIII score, maximum white blood cell count, intubation status or 28 d mortality. All but one patient (in the No-sepsis group) exhibited ≥2 systemic inflammatory response syndrome (SIRS) criteria[16].

Host transcriptional signature of sepsis from whole blood

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

To further characterize differences between sepsis patients with bloodstream versus peripheral site (for example, respiratory, urinary tract) infections, we performed differential gene expression (DE) analysis between the Sepsis$^{\mathrm{BSI}}$ and Sepsis$^{\text{non-BSI}}$ groups, which identified 5,227 genes (Supplementary Data 3). GSEA demonstrated enrichment in genes related to CD28 signalling, immunoregulatory interactions between lymphoid and non-lymphoid cells, and other functions in the Sepsis$^{\text{non-BSI}}$ patients, while the Sepsis$^{\mathrm{BSI}}$ group was characterized by enrichment in genes related to antimicrobial peptides, defensins, G alpha signalling 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$^{\mathrm{BSI}}$ and Sepsis$^{\text{non-BSI}}$ patients, we constructed a 'universal' sepsis diagnostic classifier on the basis of 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 (bSVM) learning approach to select genes that most effectively distinguished patients with sepsis (Sepsis$^{\mathrm{BSI}}$ and Sepsis$^{\text{non-BSI}}$) from those without (No-sepsis). We elected to use a bSVM model due to better performance compared with random forest and gradient boosted trees, which were also tested (Supplementary Table 2). The bSVM model achieved an average cross-validation AUC (area under the receiver operating characteristic (ROC) curve) of 0.81 (s.d. 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 (s.d. 0.02) was obtained over 10 randomly generated validation sets (Fig. 2c and 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 remains 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 peripheral blood mononuclear cells 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 2.2×10^7–2.5×10^7) reads per sample. Calculation of input RNA mass (Methods) demonstrated that samples with transcript counts below our quality control (QC) cut-off (<50,000) had a lower average input mass than those with sufficient counts (65.2 pg versus

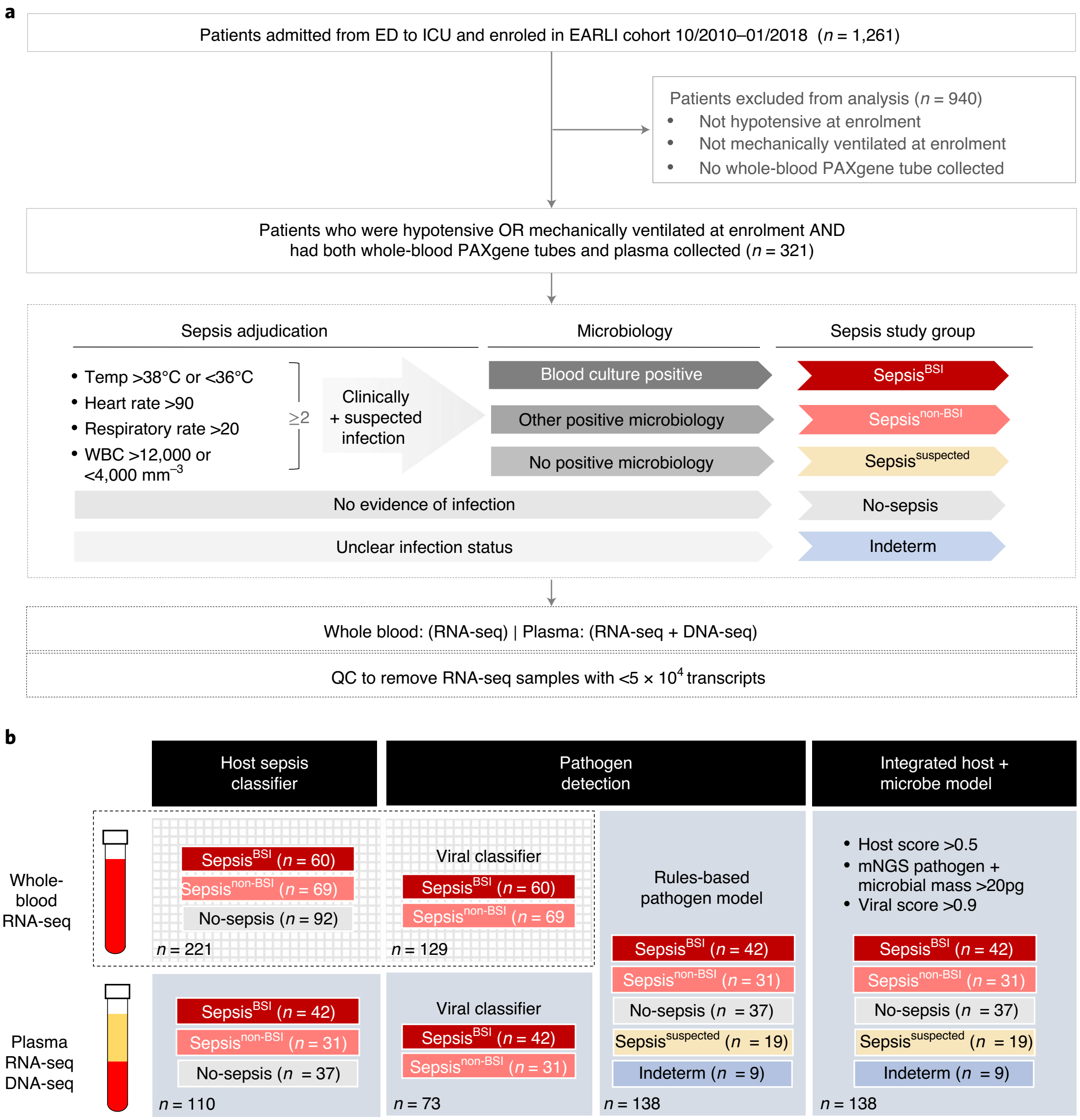


Fig. 1 | Study overview. a, Study flow diagram. Patients studied were enrolled in the EARLI cohort. Sepsis adjudication performed following hospital discharge was based on ≥2 SIRS criteria plus clinical suspicion of infection and was used to delineate five patient subgroups. Following QC, whole blood was subjected to RNA-seq, and plasma to RNA-seq and DNA-seq. WBC, white blood cell count. **b**, Analytic approaches. Host transcriptional sepsis diagnostic classifiers were trained and tested on RNA-seq data from whole blood ($n = 221$) or 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 mNGS followed by an RBM. Finally, an integrated host + microbe model for sepsis diagnosis was developed and evaluated.

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 $P < 0.1$ (Extended Data Fig. 1 and Supplementary Data 6), 28 of which were also significant in the whole-blood analysis (Extended Data Fig. 2). Remarkably, several of the top differentially expressed genes were previously reported sepsis biomarkers (for example, elevated *CD177*, suppressed *HLA-DRA*)[18–21], suggesting a biologically relevant transcriptomic signature from plasma RNA (Fig. 2d and 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 (s.d. 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 (s.d. 0.06) was obtained over 10 randomly generated validation sets (Fig. 2e and Supplementary Data 7).

Detection of bacterial sepsis pathogens from plasma nucleic acid

We began microbial metagenomic analyses by assessing DNA microbial mass (Methods), which was significantly higher in Sepsis^{BSI} compared with other groups, except for the Indeterminate group. Microbial mass was significantly lower in the negative control water samples compared with each group (Fig. 3a and 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

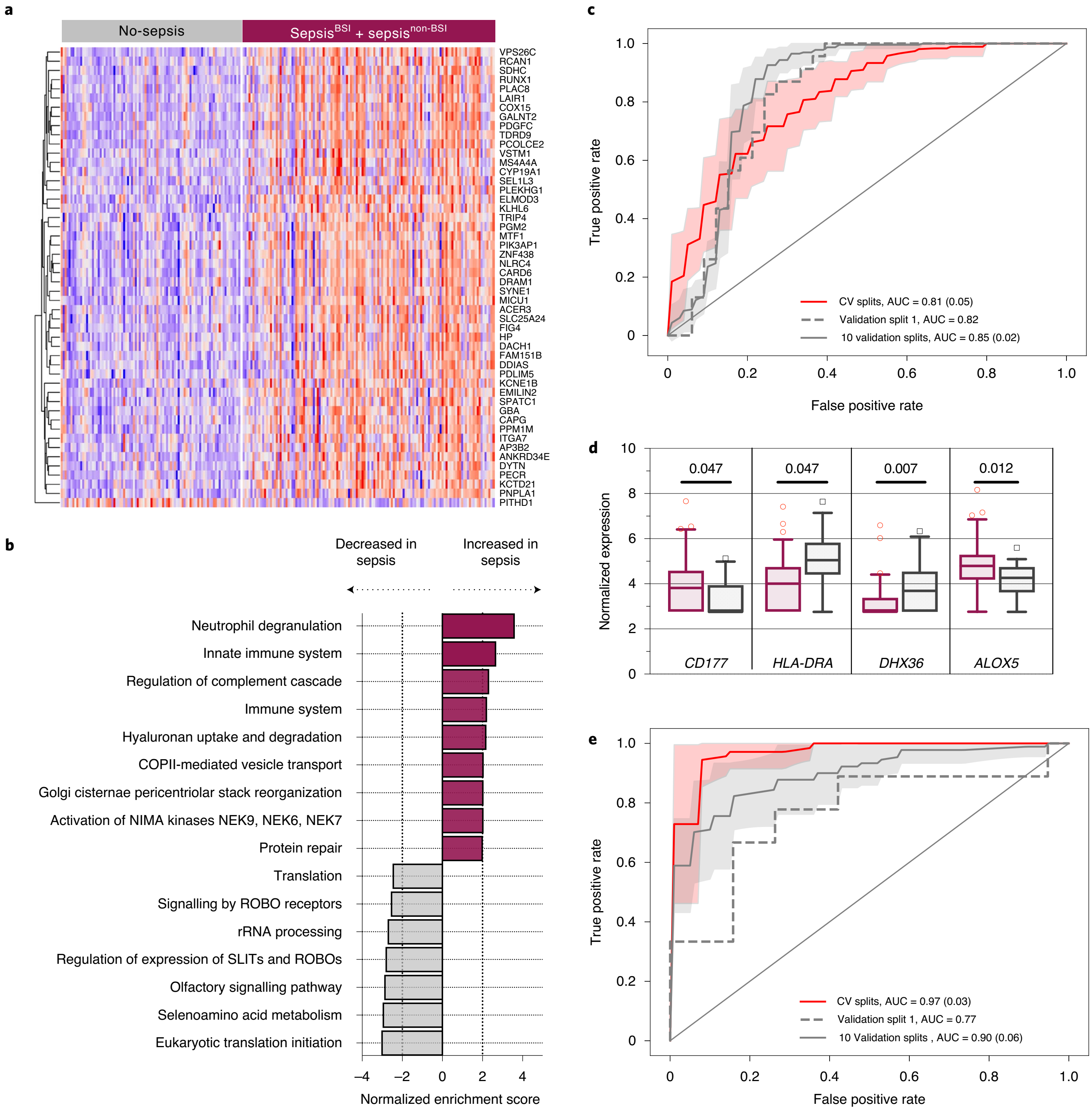


Fig. 2 | Host gene expression differentiates patients with sepsis from those with non-infectious critical illnesses. a, Heat map of top 50 differentially expressed genes from whole-blood transcriptomics comparing patients with microbiologically confirmed sepsis (Sepsis^{BSI} + Sepsis$^{non\text{-}BSI}$) versus those without evidence of infection (No-sepsis). The heatmap colour range represents the row *Z*-score of the normalized gene expression values ranging from +4 (red) to −4 (blue). **b**, GSEA of the differentially expressed genes demonstrating pathways enriched in patients with sepsis. Source data including enriched genes and pathway *P* values (hypergeometric test) are provided in Supplementary Data 2a and in the Source Data file. **c**, ROC curve demonstrating performance of the bSVM classifier for sepsis diagnosis from whole-blood transcriptomics ($n = 221$). The AUC and s.d. (in parentheses, when applicable) are listed for cross validation (CV) in the training set (red line: average over 10 random splits; red shaded area: ±1 s.d.), the held-out validation set (dashed grey line) and over 10 randomly generated validation sets (solid grey line: average; grey shaded area: ±1 s.d.). **d**, Plasma RNA-seq expression differences of selected differentially expressed genes previously identified as sepsis biomarkers, with Sepsis patients in maroon ($n = 73$) and No-sepsis patients in grey ($n = 37$). Adjusted *P* values (Benjamini–Hochberg method) from DESeq2 noted above boxplot. Expression data are presented as boxes extending from the 25th to the 75th percentiles, with whiskers extending to the 5th and 95th percentiles, and a central horizontal line at the median. Source data are provided in the Source Data file. **e**, ROC curve demonstrating performance of the bSVM classifier for sepsis diagnosis from plasma RNA ($n = 110$). The AUC and s.d. are listed for CV in the training set (red line: average over 10 random splits; red shaded area: average ± 1 s.d.), the held-out validation set (dashed grey line) and over 10 randomly generated validation sets (solid grey line: average; grey shaded area: average ± 1 s.d.).

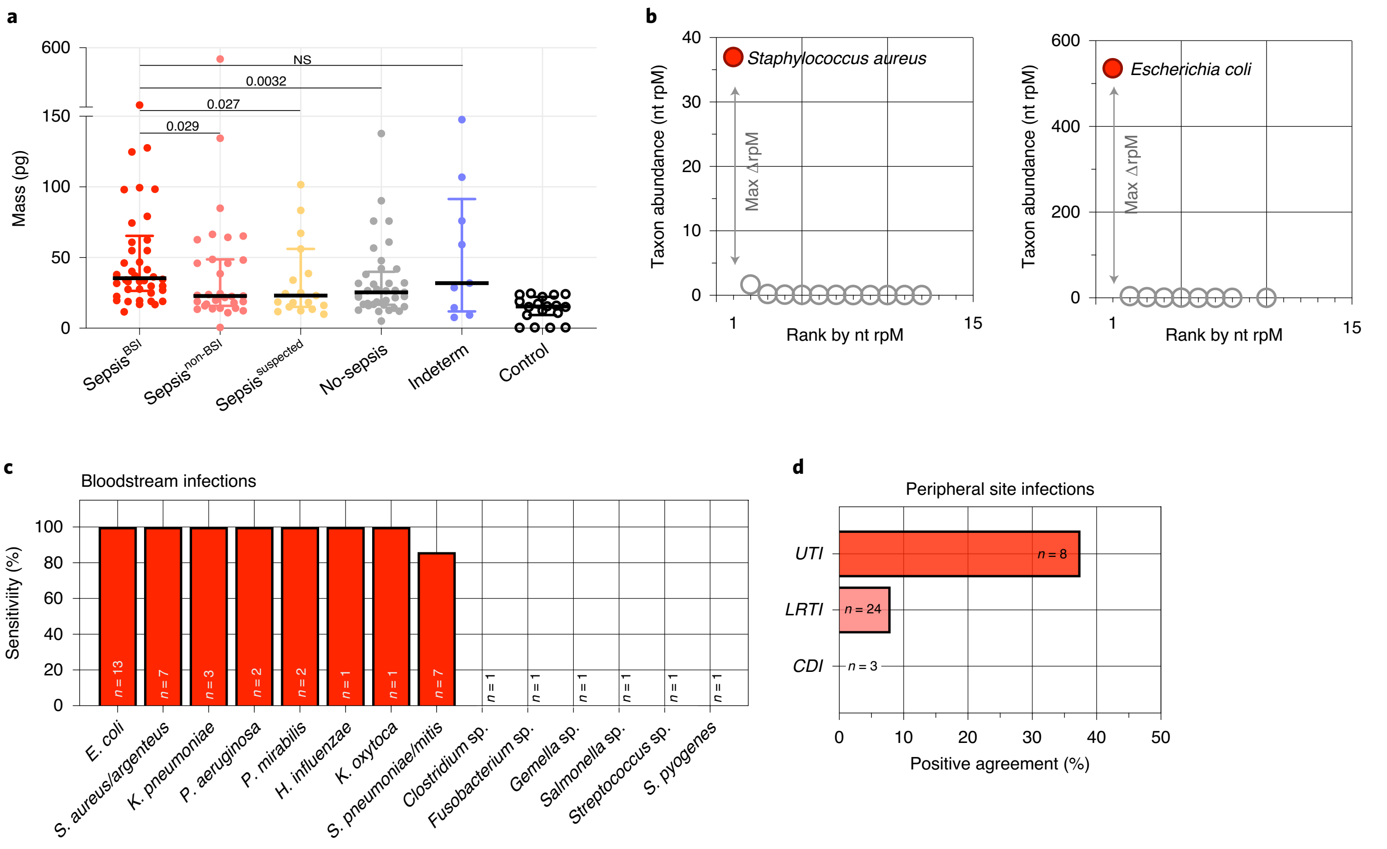


Fig. 3 | Plasma mNGS for detecting sepsis pathogens. a, Microbial plasma DNA mass differences between sepsis groups. Data are presented with a centre horizontal bar at the median, and error bars representing the interquartile ranges. Pairwise comparisons between groups were performed with a two-sided Mann–Whitney test. Sample sizes are as follows for each group: Sepsis^{BSI} $n = 42$, Sepsis$^{non\text{-}BSI}$ $n = 31$, Sepsis^{suspected} $n = 19$, Indeterminate $n = 9$, No-sepsis $n = 37$, Control $n = 18$. Source data and P values for comparisons between samples, including water controls, are provided in Supplementary Data 8 and in the Source Data file. **b**, Graphical depiction of the RBM for sepsis pathogen detection from two different exemplary cases. The RBM identifies established pathogens with disproportionately high abundance compared with other commensal and environmental microbes in the sample. **c**, Concordance between plasma DNA mNGS for detecting bacterial pathogens in Sepsis^{BSI} patients and bacterial bloodstream infections compared to a reference 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. LRTI = lower respiratory tract infection; UTI = urinary tract infection; CDI = *Clostridium difficile* infection. Clinical microbiology and metagenomics data are tabulated in Supplementary Data 9.

(RBM)[14] that identifies established sepsis pathogens overrepresented in mNGS data compared with less abundant commensal or contaminating microbes[14] (Methods and Fig. 3b).

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

Detection of sepsis pathogens from peripheral sites using plasma nucleic acid

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

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

Only 1 of 13 (8%) respiratory viruses identified by clinical testing could be detected by plasma RNA mNGS (Supplementary Data 9). Recognizing that an alternative approach would be needed, we asked whether host response could instead be used to identify viral sepsis by carrying out differential gene expression analysis of patients with or without clinically confirmed viral sepsis within the Sepsis^{BSI} and Sepsis$^{non\text{-}BSI}$ groups, using whole blood (Supplementary Data 10) or plasma (Supplementary Data 11) transcriptomic data. GSEA demonstrated that pathways related to interferon signalling and genes important for antiviral immunity were enriched in samples from patients with viral sepsis versus those with bacterial sepsis, in data derived from both whole blood (Fig. 4a and Supplementary Data 12a) and plasma (Fig. 4b and 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 (s.d. 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 (s.d. 0.04) was obtained over 10 randomly generated validation sets (Fig. 4c and Supplementary Data 13). Slightly better performance

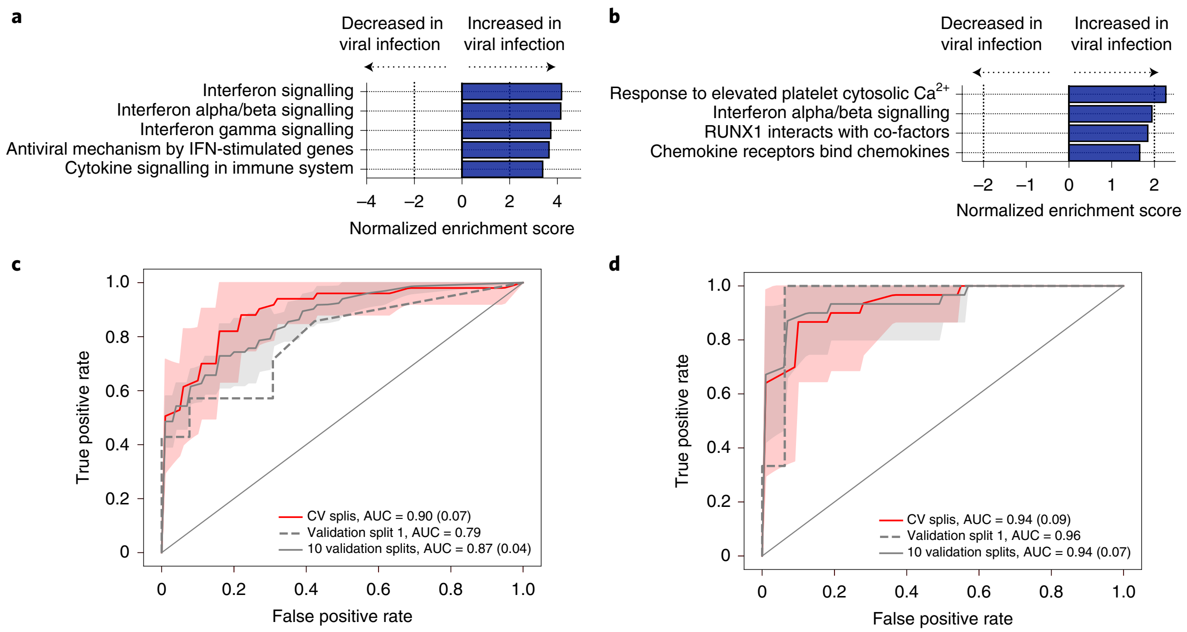


Fig. 4 | Detection of viral sepsis based on host gene expression. a, GSEA of differentially expressed genes from whole-blood RNA-seq ($n = 129$) demonstrating pathways enriched in patients with viral sepsis. The top five most significant pathways by P value (hypergeometric test) are plotted. Source data including enriched genes and pathway P values are provided in Supplementary Data 12a and in the Source Data file. **b**, GSEA of differentially expressed genes from plasma RNA-seq ($n = 73$) demonstrating pathways enriched in patients with viral sepsis. All identified pathways are plotted. Source data including enriched genes and pathway P values (hypergeometric test) are provided in Supplementary Data 12b and in the Source Data file. **c**, ROC curve demonstrating performance of the bSVM classifier for detecting viral sepsis from whole-blood RNA-seq ($n = 129$). The AUC and s.d. (in parentheses, when applicable) are listed for CV in the training set (red line: average over 10 random splits; red shaded area: ±1 s.d.), the held-out validation set (dashed grey line) and over 10 randomly generated validation sets (solid grey line: average; grey shaded area: ±1 s.d.). **d**, ROC curve demonstrating performance of the bSVM classifier for detecting viral sepsis from plasma RNA-seq ($n = 73$). The AUC and s.d. are listed for CV in the training set (red line: average over 10 random splits; red shaded area: ±1 s.d.), the held-out validation set (dashed grey line) and over 10 randomly generated validation sets (solid grey line: average; grey shaded area: ±1 s.d.).

was obtained when building a classifier using plasma RNA-seq data, with an average cross-validation AUC of 0.94 (s.d. 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 (s.d. 0.07) was obtained over 10 randomly generated validation sets (Fig. 4d and 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$^{\text{non-BSI}}$ patient who did not 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 on the basis of simple rules. It returned a sepsis diagnosis on the basis of 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%) cases in the Sepsis$^{\text{BSI}}$ group and 30/31 (97%) cases in the Sepsis$^{\text{non-BSI}}$ group, for an overall sensitivity of 72/73 (99%) (Fig. 5a,b). This proof-of-concept model yielded a specificity of 29/37 (78%) within the No-sepsis group (Fig. 5c and Supplementary Data 15).

Application of the integrated model to suspected and indeterminate sepsis cases

Next, we asked whether patients with clinically adjudicated sepsis but negative in-hospital microbiologic testing (Sepsis$^{\text{suspected}}$) would be predicted to have sepsis using the integrated host-microbe plasma mNGS model. Of the 19 patients, 14 (74%) were classified as having sepsis (Fig. 5d), 8 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 (Fig. 5e and 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$^{\text{BSI}}$ and Sepsis$^{\text{non-BSI}}$) from No-sepsis patients, using 34 clinical variables as input (Supplementary Table 3). 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 (s.d. 0.04) using a random forest model (Supplementary Table 4). We then computed the AUC using the qSOFA score, a widely used clinical score for identifying patients with sepsis in the emergency department[15]. The qSOFA achieved an average AUC of 0.48 (s.d. 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 aetiologic 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 before mNGS sample collection, and that research mNGS specimens collected up to 24 h after blood cultures 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 (for example, *Fusobacterium*, *Gemella* and *Streptococcus* species). It is unclear whether these organisms were truly aetiologic sepsis pathogens or commensals translocated to the blood in the setting of critical illness and incidentally identified in culture. 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 previous report[23]. 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[24].

Within the No-sepsis group, 10/37 (27%) patients had a pathogen detected by mNGS. Notably, 9/10 (90%) pathogens were Gram-negative enteric organisms, which may reflect gastrointestinal translocation of microbes, a well-described phenomenon during critical illness[25]. In addition, all 10 of these patients had received antibiotics in the first day of study enrolment, so it is possible that sequences were 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. Previous work has demonstrated that different viral species elicit distinct host transcriptional signatures in peripheral blood[26], suggesting that future studies could extend the RNA host viral classifier to identify specific virus, such as influenza or SARS-CoV-2

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

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 h.

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 composed 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 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 previous work demonstrating accurate performance of a host/microbe mNGS pneumonia diagnostic in an ICU cohort with a 40% prevalence of immunocompromised individuals[14].

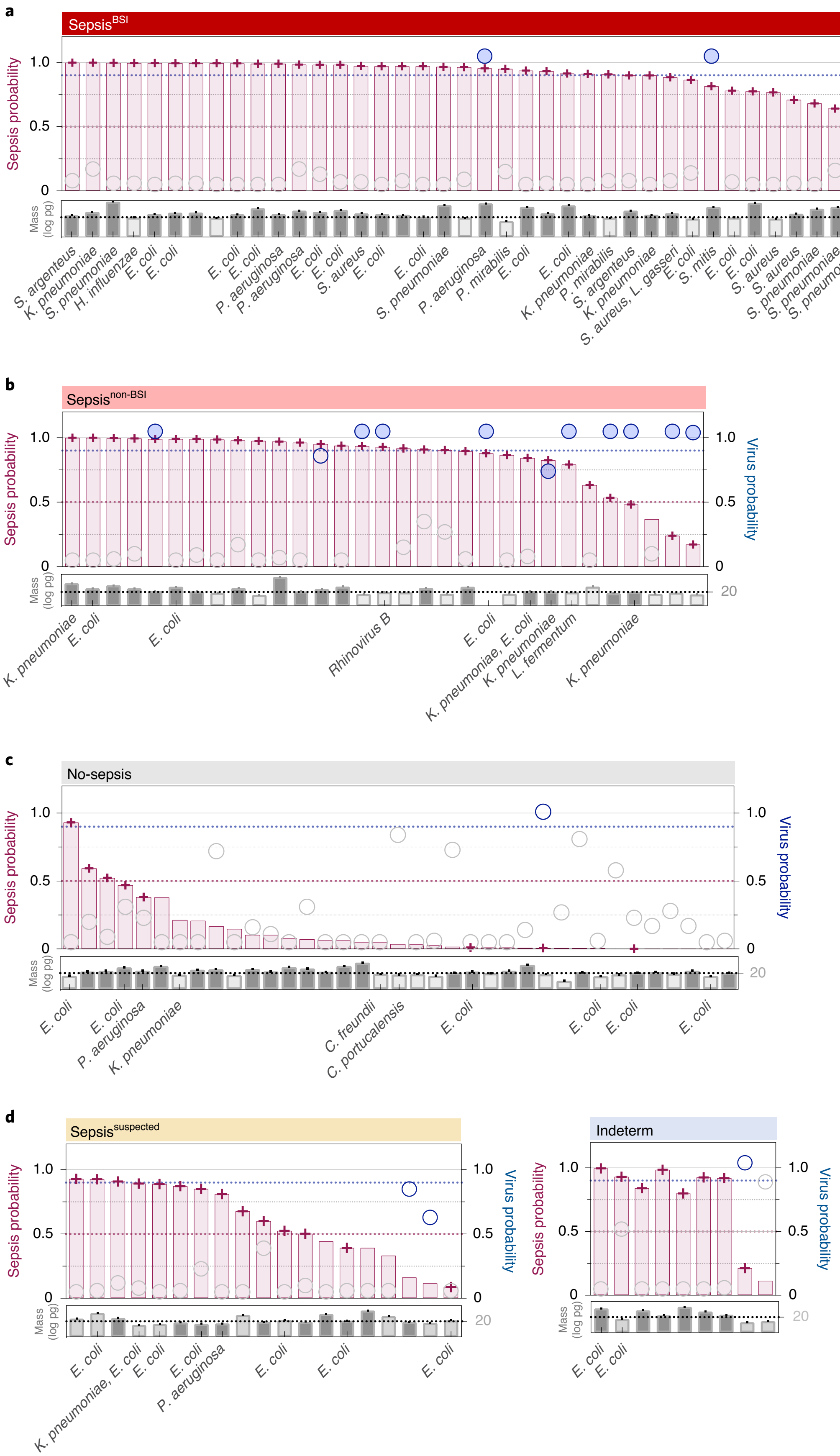
Our study has several strengths, including the innovative use of plasma RNA transcriptomics for sepsis diagnosis, development of a sepsis diagnostic combining host and microbial mNGS data, detailed clinical phenotyping and a large prospective cohort of critically ill adults with systemic illnesses. It 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, several 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 amount of plasma RNA.

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 h of ICU admission. Lastly, because we did not have access to any other sepsis studies with either

Fig. 5 | Integrated host-microbe model for sepsis diagnosis from plasma mNGS. a–**d**, Host criteria for positivity can be met by a sepsis transcriptomic classifier probability >0.5 (maroon bars, dotted line). Microbial criteria can be met on the basis of either: (1) detection of a pathogen by mNGS and a sample microbial mass (grey bars) >20 pg (dashed line), or (2) viral transcriptomic classifier probability >0.9 (blue circles, dotted line). Host and microbial metrics are highlighted for patients with sepsis due to bloodstream infections (Sepsis$^{\mathrm{BSI}}$) (**a**), peripheral infection (Sepsis$^{\mathrm{non\text{-}BSI}}$) (**b**), patients with non-infectious critical illness (No-sepsis) (**c**), patients with suspected sepsis but negative microbiological testing (Sepsis$^{\mathrm{suspected}}$) (**d**, left) and patients with indeterminate sepsis status (Indeterm) (**d**, right). 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 <20 pg microbial mass are 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 and provided in the Source Data file.

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.

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 (ZSFGH) between October 2010 and January 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[27,28].

For the parent EARLI cohort, the inclusion criteria are: (1) age ≥18, (2) admission to the ICU from the ED and (3) enrolment in the ED or within the first 24 h of ICU admission. For this study, we selected patients for whom PAXgene whole-blood tubes and matched plasma samples from the time of enrolment 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 enrolment. 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 h. Enrolment in EARLI began in October 2008 and continues. Study data were collected and managed using REDCap and Quesgen electronic data capture tools hosted at UCSF[29,30].

Sepsis adjudication

Clinical adjudication of sepsis groups was carried out by study team physicians (M.A., C.R.L., A.L., K.L., P.S., C.H., A.G., C.C., K.N.K., M.A.M.) using the sepsis-2 definition[31] (≥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 (C.S.C., M.A.M.) in the phenotyping panel. Patients were categorized into five subgroups on the basis of sepsis status (Fig. 1a): patients with clinically adjudicated sepsis and a bacterial culture-confirmed bloodstream infection (Sepsis$^{\mathrm{BSI}}$), sepsis due to a microbiologically confirmed primary infection at a peripheral site other than the bloodstream (Sepsis$^{\text{non-BSI}}$), suspected sepsis with negative clinical microbiologic testing (Sepsis$^{\text{suspected}}$), patients with no evidence of sepsis and a clear alternative explanation for their critical illness (No-sepsis), or patients of indeterminate status (Indeterm). Clinical and demographic features of patients are summarized in Supplementary Tables 1a,b and tabulated in Supplementary Data 16 and 17.

Metagenomic sequencing

Following enrolment, whole blood and plasma were collected in PAXgene and EDTA tubes, respectively. Whole-blood PAXgene tubes (Qiagen, 762165) were processed and stored at −80 °C according to the manufacturer's instructions, and plasma was frozen at −80 °C within 2 h. 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 (Qiagen, 74004) and normalized to 10 ng total input per sample. Total plasma nucleic acid was extracted by first clarifying 300 µl of plasma via maximum-speed centrifugation for 5 min at 21,300 × *g*, and then employing the Zymo Pathogen Magbead kit (Zymo Research, R2145) on the supernatant following the manufacturer's instructions. Total nucleic acid (10 ng) was subjected to DNA-seq using the NEBNext Ultra II DNA kit. Samples with at least 10 ng of remaining total nucleic acid were treated with DNAse (Qiagen) to recover RNA, and then subjected to RNA-seq library preparation using the NEBNext Ultra II RNA-seq kit (New England Biolabs, E7770S) as described below.

For RNA-seq library preparation, human cytosolic and mitochondrial ribosomal RNA and globin RNA were first depleted using FastSelect (Qiagen, 334385). For 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); ThermoFisher, 4456740)[32]. RNA was then fragmented and subjected to library preparation using the NEBNext Ultra II RNA-seq kit (New England Biolabs, E7770S) according to the manufacturer's instructions, with protocol optimization for a LabCyte Echo acoustic liquid handler[33]. 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 35× lower when using dual vs single indexing[34]. 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 to an index of the human genome (NCBI GRC h38) plus ERCC RNA standards using STAR (version 2.6.1)[35]. Samples retained in the dataset had a total of at least 50,000 counts associated with protein coding genes. Differential expression analysis was performed using DESeq2 (ref. [36]) and including covariates for age and gender. Significant genes were identified using an independent-hypothesis-weighted, Benjamini–Hochberg false discovery rate (FDR) < 0.1[37,38]. We generated heat maps of the top 50 differentially expressed genes by absolute $\log_2$-fold change. To evaluate signalling pathways from gene expression data, we employed gene set enrichment analysis using WebGestalt[39] on all ranked differentially expressed genes with 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 (v3.7, https://czid.org/), which incorporates subtractive alignment of the human genome (NCBI GRC h38) using STAR[35] (v2.5.3), quality and complexity filtering, and subsequent removal of cloning vectors and phiX phage using Bowtie2 (v2.3.4)[22]. The identities of the remaining microbial reads were determined by querying the NCBI nucleotide (NT) database using GSNAP-L[22,40] in the final steps of the IDseq pipeline. 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 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 a respiratory) pathogen reference index (Supplementary Data 18) and requiring that the species called by the RBM be both present in the reference index and detected at an abundance >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 on the basis of 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)[41] and a recent multicentre surveillance study of healthcare-associated infections[42]. These studies reported multiple species of *Bacteroides*, *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 >1,000 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[43]. A previously developed negative binomial model[43] (https://github.com/czbiohub/idseqr/) was employed to identify taxa with NT sequencing alignments present at an abundance significantly greater compared with negative water controls. This was done by modelling 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[44] (v7.3-51). Taxa that achieved an adjusted $P < 0.01$ (Benjamini and Hochberg multiple test correction) were carried forward to the above-described RBM for pathogen detection.

Microbial mass calculations

Microbial mass was calculated on the basis of the ratio of microbial reads in each sample to total reads aligning to the External RNA Controls Consortium (ERCC) RNA standards spiked into each sample[32]. The following equation was utilized for this calculation: (ERCC input mass)/(microbial input mass) = (ERCC reads)/(microbial reads), where the ERCC input mass was 25 pg.

Host transcriptional classifiers for viral sepsis diagnosis

To build classifiers that differentiated patients with sepsis (Sepsis$^{\mathrm{BSI}}$, Sepsis$^{\mathrm{non\text{-}BSI}}$) from those with non-infectious critical illness (No-sepsis), and distinguished viral from non-viral sepsis, we built a support vector machine (SVM)-based classifier[45] with the scikit-learn[46] (v0.23.2) library in Python (v3.8.3). We tested several machine learning approaches (bagged SVM, random forest and gradient boosted trees) and selected a bSVM classifier with a linear kernel based on best performance (Supplementary Table 2). Each classifier used a bootstrapped set of samples and a random subset of features.

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

On each one of those intermediate training sets, we carried out feature selection and parameter 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 parameter's optimization fold, a recursive feature elimination 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 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 mNGS against clinician-ordered testing

Clinical microbiological testing was carried out on the basis of decisions from the primary medical team during the patient's hospital admission at the UCSF and the ZSFGH clinical microbiology laboratories. Tests utilized included bacterial culture from blood, lower respiratory tract and urine, which were carried out in the clinical microbiology laboratories at each hospital as previously described[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 relevance 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 on the basis of 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). Combined metrics (Supplementary Data 16) including sepsis assignment based on this model are depicted in Fig. 5. Sensitivity was calculated in the Sepsis$^{\mathrm{BSI}}$ and Sepsis$^{\mathrm{non\text{-}BSI}}$ groups, and specificity in the No-sepsis group.

Clinical variable models for sepsis diagnosis

We tested the ability of clinical variables (Supplementary Table 3) available at the time of initial patient assessment to predict sepsis using three machine learning methods. These included SVM using the e1071 package[47] v1.7, random forest using the randomForest package[48] v4.7 and regularized logistic regression using the glmnet package[49] v4.1 in R v4.2.0[50]. Specifically, we built models to classify Sepsis (Sepsis$^{\mathrm{BSI}}$ and Sepsis$^{\mathrm{non\text{-}BSI}}$) versus No-sepsis using 34 clinical variables that were available at the time of ED evaluation. The data were split into training (75%) and test (25%) sets, and model performance (AUC) was evaluated on the test set. This was repeated for a total of 10 randomized splits, with the AUC computed at each iteration. AUC was also computed for the qSOFA score (systolic blood pressure <100 mmHg, respiratory rate >22 breaths per minute, Glasgow coma scale <13). Results are tabulated in (Supplementary Table 4).

Statistics and reproducibility

Statistical tests utilized for each analysis are described in the figure legends and in further detail in each respective Methods section. The numbers of patient samples analysed for each comparison are indicated in the figure legends. Data were generated from single sequencing runs without technical replicates.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The processed genecount data are available from the National Center for Biotechnology Information Gene Expression Omnibus database under accession code GSE189403. The raw sequencing data are protected due to data privacy restrictions from the IRB protocol governing patient enrolment, which protects the release of raw genetic sequencing data from those patients enrolled under a waiver of consent. To honour this, researchers who wish to obtain raw fastq files for the purposes of independently generating gene counts 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 IDs: PRJNA782906 and PRJNA782908. Source data are provided with this paper.

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. *Lancet* **395**, 200–211 (2020).
2. Liu, V. et al. Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA* **312**, 90–92 (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. *Crit. Care Med.* **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. *Lancet Infect. Dis.* **17**, 990–1001 (2017).
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. *Nat. Microbiol.* **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**, e00419–e00420 (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. *Clin. Infect. Dis.* **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. *Sci. Transl. Med.* **8**, 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* **115**, E12353–E12362 (2018).
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. *Proc. Natl Acad. Sci. USA* **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. *Immunol. Lett.* **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**, giaa111 (2020).
23. Langelier, C. et al. Detection of pneumonia pathogens from plasma cell-free DNA. *Am. J. Respir. Crit. Care Med.* **201**, 491–495 (2019).
24. Libby, D. B. & Bearman, G. Bacteremia due to *Clostridium difficile*—review of the literature. *Int. J. Infect. Dis.* **13**, e305–e309 (2009).
25. 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. *Clin. Infect. Dis.* **66**, 497–503 (2018).
26. Mudd, P. A. et al. Distinct inflammatory profiles distinguish COVID-19 from influenza with limited contributions from cytokine storm. *Sci. Adv.* **6**, eabe3024 (2020).
27. Auriemma, C. L. et al. Acute respiratory distress syndrome-attributable mortality in critically ill patients with sepsis. *Intensive Care Med.* **46**, 1222–1231 (2020).
28. 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).
29. Harris, P. A. et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
30. Harris, P. A. et al. The REDCap consortium: building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
31. Levy, M. M. et al. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Crit. Care Med.* **31**, 1250–1256 (2003).
32. 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).

33. Mayday, M. Y., Khan, L. M., Chow, E. D., Zinter, M. S. & DeRisi, J. L. Miniaturization and optimization of 384-well compatible RNA sequencing library preparation. *PLoS ONE* **14**, e0206194 (2019).
34. Wilson, M. R. et al. Multiplexed metagenomic deep sequencing to analyze the composition of high-priority pathogen reagents. *mSystems* **1**, e00058-16 (2016).
35. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
37. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13**, 577–580 (2016).
38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
39. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).
40. 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).
41. 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).
42. Magill, S. S. et al. Changes in prevalence of health care–associated infections in U.S. hospitals. *N. Engl. J. Med.* **379**, 1732–1744 (2018).
43. Mick, E. et al. Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses. *Nat. Commun.* **11**, 5854 (2020).
44. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer-Verlag, 2002); https://doi.org/10.1007/978-0-387-21706-2
45. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
46. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
47. Meyer, D. et al. Package 'e1071' (The Comprehensive R Archive Network, 2022); https://cran.r-project.org/web/packages/e1071/e1071.pdf
48. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
49. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
50. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).

Acknowledgements

This study was supported by the National Heart, Lung and Blood Institute (K23HL138461-01A1 (C.R.L.), R35 HL140026 (C.S.C.), F32 HL151117 (A.S.)) and the Chan Zuckerberg Biohub (C.R..L, J.L.D.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

C.R.L., K.L.K., C.S.C. and L.N. conceived and designed the study. P.H.S., S.C., R.G. and N.N. performed or supervised key aspects of sample processing, library preparation and/or sequencing. M.A., C.R.L., A.L., K.L., P.S., C.H., A.G., C.C., K.N.K. and M.A.M. carried out clinical adjudication of sepsis groups. C.S.C., M.A.M., K.L., C.H., K.N.K., A.L., A.G., P.S., R.G. and A.J. coordinated or contributed to clinical operations. K.L.K., L.N., C.R.L., E.M., J.A., A.P., A.T. and A.S. performed data analysis and interpretation. C.S.C., J.L.D., S.B.D., S.A.C. and M.A.M. provided guidance, advice and comments on the study design and manuscript. C.R.L., K.L.K., L.N. and C.S.C. wrote the manuscript with input from all authors.

Competing interests

C.R.L., K.L.K., L.N. and C.S.C. are inventors on a provisional patent (no. 63/342,528) related to the methodology. The authors declare no other competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-022-01237-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-022-01237-2.

Correspondence and requests for materials should be addressed to Charles R. Langelier.

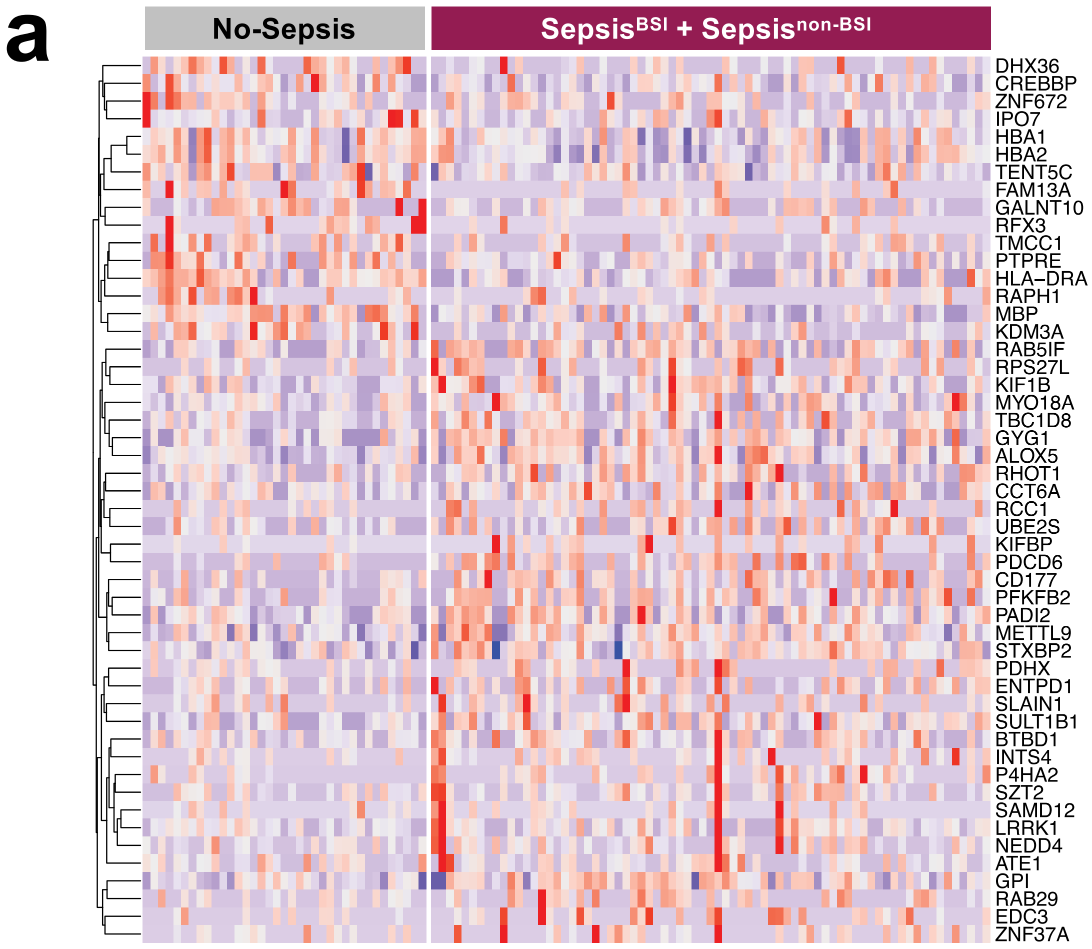
Peer review information *Nature Microbiology* thanks Vanya Gant, Samuel Yang, Cesar de la Fuente and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

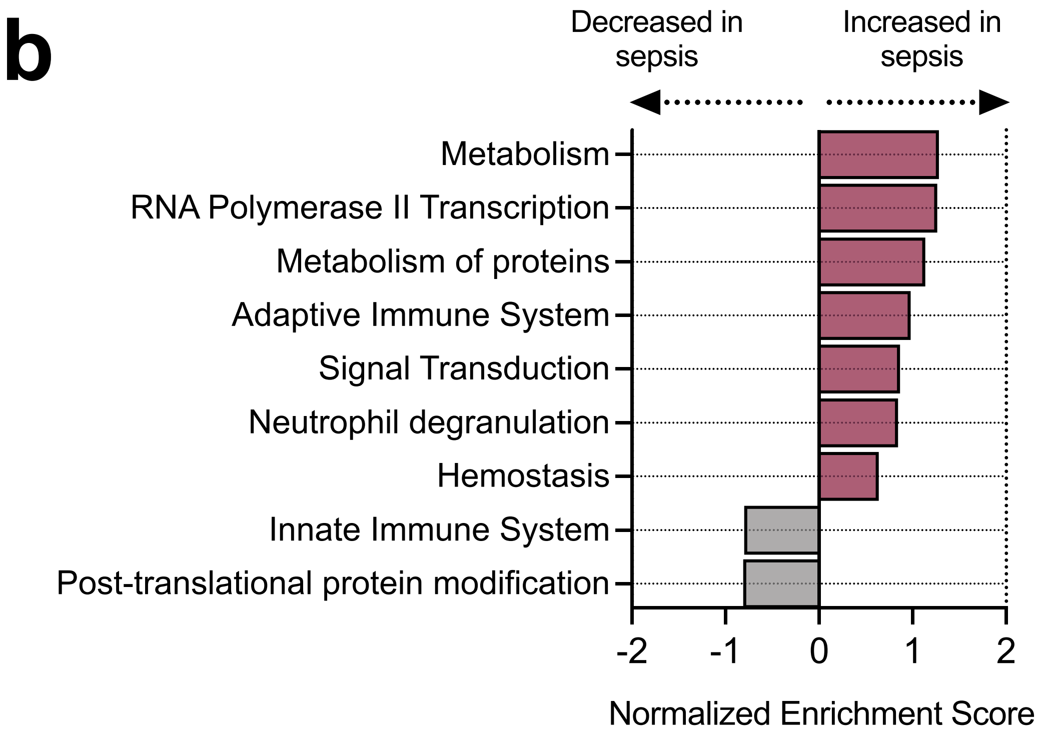
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional 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. [11]These authors contributed equally: Katrina L. Kalantar, Lucile Neyton, Carolyn S. Calfee, Charles R. Langelier. ✉e-mail: chaz.langelier@ucsf.edu

https://doi.org/10.1038/s41564-022-01237-2



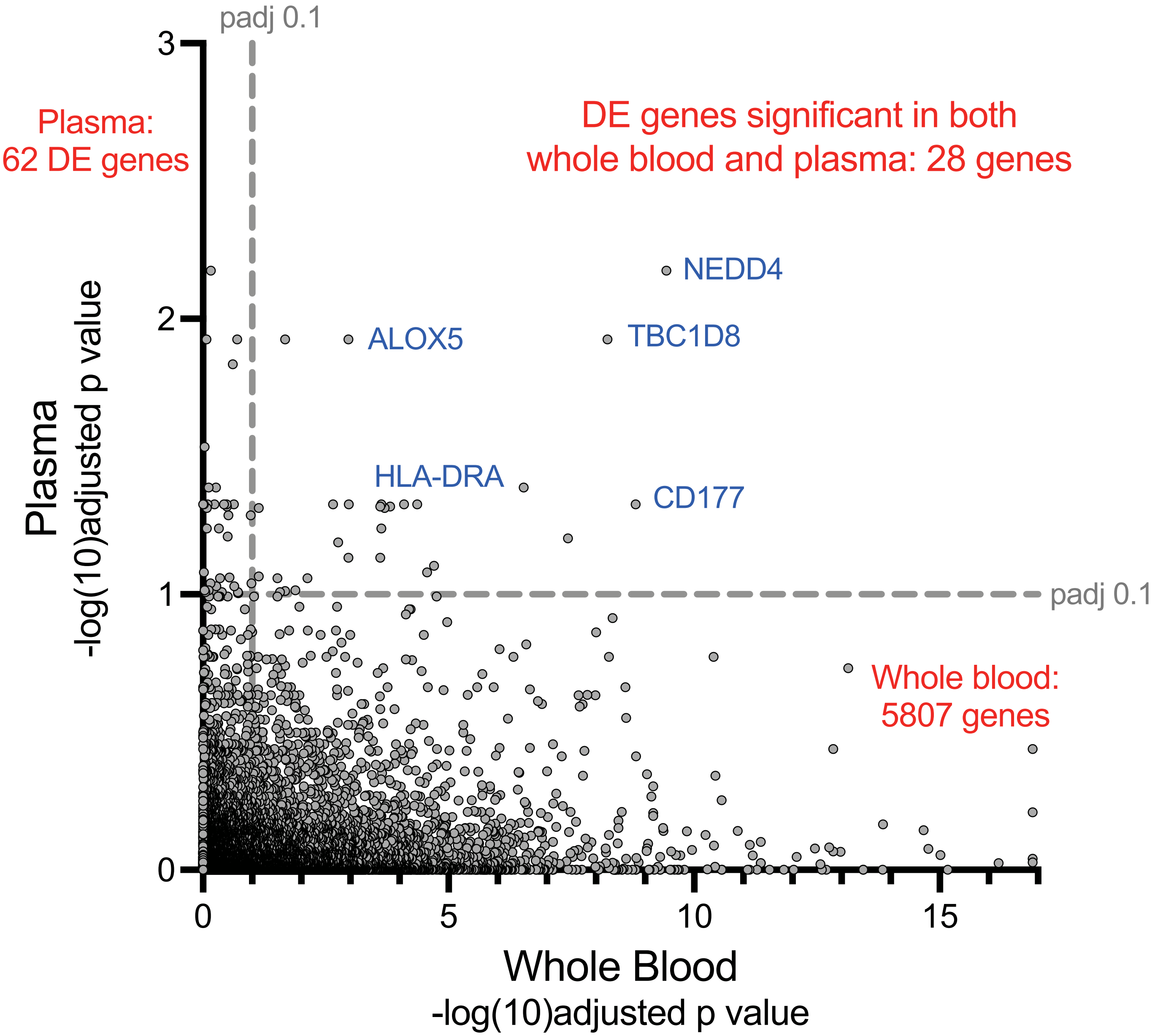


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41564-022-01237-2

Extended Data Fig. 1 | Plasma host gene expression differentiates patients with sepsis from those with non-infectious critical illnesses. 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 identified pathways are plotted. Source data including enriched genes and pathway P values (hypergeometric test) are provided in Supplementary Data 2b.

https://doi.org/10.1038/s41564-022-01237-2



Extended Data Fig. 2 | Overlap of significant genes in the differential expression analyses between the Sepsis and No-Sepsis groups for whole blood and plasma samples. 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 (two-sided) are derived from DESeq2 and incorporate Benjamini–Hochberg adjustment for multiple testing. Dashed grey lines indicate the threshold of adjusted p-value <0.1. Selected, significant, differentially expressed genes highlighted in blue.

nature portfolio

Corresponding author(s): Charles Langelier

Last updated by author(s): Aug 7, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Clinical data were collected and stored using QuesGen and REDCap databases

Data analysis

Bulk RNAseq and differential expression:
Following demultiplexing, sequencing reads were aligned with STAR 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.

Classifier construction:
To build gene expression classifiers that differentiated patients with sepsis from those with non-infectious critical illness, and distinguished viral from non-viral sepsis, we built a Support Vector Machine (SVM)-based classifier with the scikit-learn (v0.23.2) library in Python (v3.8.3). To build clinical variable classifiers we tested three different machine learning methods. These included SVM using the e1071 package v1.7, random forest using the randomForest package v4.7 and regularized logistic regression using the glmnetpackage v4.1 in R v4.2.0.

Pathogen detection:
Detection of microbes leveraged the open-source IDseq pipeline v3.7 (https://czid.org/) which incorporates subtractive alignment of the human genome (NCBI GRC h38) using STAR (v2.5.3), quality and complexity filtering, and subsequent removal of cloning vectors and phiX phage using Bowtie2 (v2.3.4). The identities of the remaining microbial reads are determined by querying the NCBI nucleotide (NT) database using GSNAP-L in the final steps of the IDseq pipeline.

Background correction:
Negative control samples enabled estimation of the number of background reads expected for each taxon. A previously developed negative

binomial model (https://github.com/czbiohub/idseqr/) 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 (v7.3-51).

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

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Study data were collected and managed using REDCap and Quesgen electronic data capture tools hosted at UCSF. Source data are provided with this paper. The processed genecount data are available from the National Center for Biotechnology Information Gene Expression Omnibus database under accession code GSE189403. The raw sequencing data are protected due to data privacy restrictions from the IRB protocol governing patient enrollment in this study, which protect 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 explicit purpose of independently generating genecounts for assessing gene expression can contact the corresponding author (chaz.langelier@ucsf.edu) and request to be added to the IRB protocol. Requests will be addressed within a timeframe of two weeks. The raw fastq files with microbial sequencing reads are available from the Sequence Read Archive under BioProject IDs: PRJNA782906 and PRJNA782908.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Samples were selected from an observational cohort. We used the RNASeqPower package for R to calculate the power of differential expression analysis, and determined that we had greater than 99% to detect a 2-fold change in expression at an FDR < 0.1 in our primary analysis.

Data exclusions

The main exclusion criteria for the cohort were: 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.

Replication

All analyses were performed in a single cohort of patients. We have made a concerted attempt to clearly indicate the number of patients analyzed in each comparator group (Sepis-BSI, Sepsis-non-BSI, Sepsis-suspected, No Sepsis) in the manuscript and figure legends. This is the first publicly available host/microbe sequencing dataset of sepsis patients, and there is therefore no dataset available for a replication analysis.

Randomization

N/A - observational study

Blinding

Investigators were blinded to group allocation during data collection. Investigators were blinded to any information about gene expression or metagenomic sequencing prior to chart review for sepsis adjudication. The sequencing and alignment pipeline did not have any information about the subject diagnosis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
□	☒	Human research participants
□	☒	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

We conducted a prospective observational study of adults with acute critical illnesses admitted from the ED to the ICU at the University of California, San Francisco (UCSF) or Zuckerberg San Francisco General Hospital between 10/2010 and 01/2018. We studied patients who were enrolled in the longstanding Early Assessment of Renal and Lung Injury (EARLI) cohort. Detailed demographic and clinical characteristics of the cohort and analyzed patient groups are provided in Supplementary Table 1.

Recruitment

We would like to note that in the manuscript, we reference 2 prior studies which describe recruitment in detail:
Auriemma, C. L. et al. Acute respiratory distress syndrome-attributable mortality in critically ill patients with sepsis. Intensive Care Med 46, 1222–1231 (2020).
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).

We would also like to provide a more comprehensive description here:

If a patient met inclusion criteria for the EARLI cohort, then a study coordinator or physician obtained written informed consent for enrollment from the patient or their surrogate. Patients or their surrogates were provided with detailed written and verbal information about the goals of the study, the data and specimens that would be collected, and the potential risks to the subject. Patients and their surrogates were also informed that there would be no benefit to them from being enrolled in the study and that they may withdraw informed consent at any time during the course of the study. All questions were answered, and informed consent documented by obtaining the signature of the patient or their surrogate on the consent document.

Many critically ill patients are unconscious at the time of intensive care unit (ICU) admission due to their underlying illness and/or are endotracheally intubated for airway management or acute respiratory failure. The patients who are not unconscious are often in pain and may have acute delirium due to critical illness and/or medications. For these reasons, many subjects are unable to provide informed consent at the time of enrollment. Because this study could not practically be done otherwise and was deemed to be minimal risk by the UCSF IRB, if a patient was unable and a surrogate was not available to provide consent, patients were enrolled with waiver of initial consent, including the collection of biological samples.

Specifically, for subjects who were unable to provide informed consent at the time of enrollment, our study team was permitted to collect biological samples as well as clinical data from the medical record obtained prior to consent. Surrogate consent was vigorously pursued for all patients; moreover, each patient was regularly examined to determine if and when s/he was able to consent for him/herself, and the nursing and ICU staff were contacted daily for information about surrogates' availability. For patients whose surrogates provided informed consent, follow-up consent was subsequently obtained from the patient if they survived their acute illness and regained the ability to consent. For subjects who died prior to the consent being obtained, a full waiver of consent was approved by the UCSF IRB for both cohort studies.

Lack of a surrogate to provide consent is common in critically ill patients. To address this, the UCSF IRB also approved a full waiver of consent for subjects who remained unable to provide informed consent and had no contactable surrogate identified within 28 days. Before utilizing this waiver, we made and documented at least three separate attempts to identify and contact the patient or surrogate over a month-long period. No personally identifiable information has been included as part of this manuscript for any enrolled patients.

Lastly, we would like to note that patients with more severe disease (e.g., mechanical ventilation, hypotension) were preferentially selected for inclusion, and thus our study population may not be representative of every patient transferred from the ED to ICU.

Ethics oversight

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 above and in the following references:

Auriemma, C. L. et al. Acute respiratory distress syndrome-attributable mortality in critically ill patients with sepsis. Intensive Care Med 46, 1222–1231 (2020).
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).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJEguidelines for publication of clinical research and a completedCONSORT checklist must be included with all submissions.

Clinical trial registration	N/A
Study protocol	N/A
Data collection	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.
Outcomes	The primary outcome was diagnosis of sepsis using host +/- microbial metagenomics. Secondary outcomes included pathogen detection by metagenomics and host-based identification of viral sepsis.