Leveraging the pulmonary immune response and microbiome for improved lower respiratory tract infection diagnosis in critically ill children

Authors

Eran Mick[1,2,3,*], Alexandra Tsitsiklis[2,*], Jack Kamm[1,^,*], Katrina Kalantar[4], Saharai Caldera[1,2], Amy Lyden[1], Michelle Tan[1], Angela M. Detweiler[1], Norma Neff[1], Christina M. Osborne[5], Kayla M. Williamson[6], Victoria Soesanto[6], Matthew Leroue[5], Aline B. Maddux[5], Eric A. F. Simões[5], Todd Carpenter[5], Brandie D. Wagner[5,6], Joseph L. DeRisi[1,7], Lilliam Ambroggio[5], Peter M. Mourani[5,8,–], Charles R. Langelier[1,2,–,+]

Affiliations

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

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

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

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

[5] Department of Pediatrics, University of Colorado and Children's Hospital Colorado, Aurora, CO, USA

[6] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado, Aurora, CO, USA

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

[8] Department of Pediatrics, University of Arkansas for Medical Sciences and Arkansas Children's Research Institute, Little Rock, AR, USA

[^] Present address: Genentech, Inc., South San Francisco, CA, USA

* These authors contributed equally

[–] These authors jointly supervised this work

[+] Correspondence to: chaz.langelier@ucsf.edu

ABSTRACT

BACKGROUND. Lower respiratory tract infection (LRTI) is a leading cause of death in children worldwide. LRTI diagnosis is challenging since non-infectious respiratory illnesses appear clinically similar and existing microbiologic tests are often falsely negative or detect incidentally-carried microbes common in children. These challenges result in antimicrobial overuse and adverse patient outcomes. Lower airway metagenomics has the potential to detect host and microbial signatures of LRTI. Whether it can be applied at scale and in a pediatric population to enable improved diagnosis and precision treatment remains unclear.

METHODS. We used tracheal aspirate RNA-sequencing to profile host gene expression and respiratory microbiota in 261 children with acute respiratory failure. We developed a random forest gene expression classifier for LRTI by training on patients with an established diagnosis of LRTI (n=117) or of non-infectious respiratory failure (n=50). We then developed a classifier that integrates the: i) host LRTI probability, ii) abundance of respiratory viruses, and iii) dominance in the lung microbiome of bacteria/fungi considered pathogenic by a rules-based algorithm.

RESULTS. The host classifier achieved a median AUC of 0.967 by 5-fold cross-validation, driven by activation markers of T cells, alveolar macrophages and the interferon response. The integrated classifier achieved a median AUC of 0.986 and significantly increased the confidence of patient classifications. When applied to patients with an uncertain diagnosis (n=94), the integrated classifier indicated LRTI in 52% of cases and nominated likely causal pathogens in 98% of those.

CONCLUSIONS. Lower airway metagenomics enables accurate LRTI diagnosis and pathogen identification in a heterogeneous cohort of critically ill children through integration of host, pathogen, and microbiome features.

INTRODUCTION

Lower respiratory tract infection (LRTI) causes more deaths each year than any other type of infection and disproportionately impacts children[1,2]. The ability to accurately determine whether LRTI underlies or contributes to respiratory failure in the intensive care unit and to identify the etiologic pathogens is critical for effective and targeted treatments. However, LRTI diagnosis is challenging since non-infectious respiratory conditions can appear clinically similar. Moreover, no microbiologic diagnosis is obtained in many cases of suspected LRTI[3,4] since standard tests (such as bacterial culture) suffer from a narrow spectrum of targets and limited sensitivity[5]. At the same time, children are especially susceptible to false positive diagnoses due to frequent incidental carriage of potentially pathogenic microbes[3,4,6]. As such, LRTI treatment is often empirical, leading to antimicrobial overuse, selection for resistant pathogens, and adverse outcomes[7,8].

Profiling host gene expression in the blood has shown promise as an innovative modality for diagnosing LRTI in hospitalized adult patients[9–11]. However, this approach remains unproven in the diagnostically challenging critically ill pediatric population[9,10,12,13]. Moreover, while blood gene expression can in some cases distinguish between the response to viral and bacterial infection[9], it cannot pinpoint the specific pathogens active in the respiratory tract, which is critical for optimal antimicrobial therapy.

Metagenomic next generation sequencing (mNGS) of the lower airway (tracheal aspirate, TA) has the potential to detect pathogens and host gene expression signatures of LRTI[14]. Whether such an approach can be successfully applied at scale remains unclear. The applicability of this approach in a pediatric population has also never been examined despite well-established age-related differences in LRTI epidemiology[3,6], rates of incidental pathogen carriage[3,4,6], and the immune response to infection[15,16]. Furthermore, to our knowledge, no metagenomic approach for LRTI diagnosis thus far integrates host and microbial features into a single diagnostic output, a crucial step toward streamlined clinical application.

Here, we perform metagenomic RNA sequencing of TA in a prospective cohort of 261 children with acute respiratory failure requiring mechanical ventilation. We develop a host gene expression classifier for LRTI by training on patients with an LRTI diagnosis supported by clinical microbiologic testing and patients with respiratory failure due to non-infectious causes. We then develop a classifier that integrates host, pathogen, and microbiome features to accurately diagnose LRTI and identify the likely causal pathogens, including in cases with negative clinical microbiologic testing. Our results demonstrate the feasibility of lower airway metagenomics for improved LRTI diagnosis in a large and heterogeneous cohort, and reveal the importance of profiling both the pulmonary immune response and microbiome in a pediatric population.

RESULTS

Patient cohort and LRTI adjudication

We enrolled children with acute respiratory failure requiring mechanical ventilation at eight hospitals in the United States between February 2015 and December 2017, as previously described[6,17]. Tracheal aspirate (TA) was collected within 24 hours of intubation and underwent metagenomic next generation sequencing (mNGS) of RNA to assay host gene expression and detect respiratory microbiota (**Fig. 1**). High-quality host gene expression and microbial data was obtained for n=261 patients.

Adjudication of LRTI status incorporated the final clinical diagnoses of the treating physicians as well as standard-of-care microbiologic tests performed during the admission (nasopharyngeal swab viral PCR and/or TA culture), and was blinded to mNGS results. Patients were assigned to one of four LRTI status groups, according to a recently described approach[6,9,14]: i) Definite, where clinicians established an LRTI diagnosis and the patient had clinical microbiologic findings (n=117); ii) Suspected, where clinicians established an LRTI diagnosis but the microbiologic testing performed returned negative (n=57); iii) Indeterminate,

where it remained uncertain whether LRTI was a contributing factor to respiratory failure due to conflicting clinical and microbiologic findings (n=37); and iv) No Evidence, where clinicians identified a clear non-infectious cause of acute respiratory failure and no clinical or microbiologic suspicion of LRTI arose (n=50) (**Fig. 1**).

The patients in the Definite and No Evidence groups were used to develop the metagenomic classifiers and evaluate their performance by cross-validation due to the high degree of confidence in their clinical diagnoses (**Fig. 1**). Nevertheless, we recognized that retrospective clinical adjudication, while the only practical approach, was subject to variability in clinical and microbiologic practices across study sites and to the known limitations of standard microbiologic diagnostics. Moreover, comprehensive microbiologic testing was not always performed in the No Evidence group in the absence of clinical suspicion.

The patients in the Definite group were 39% female with a median age of 0.5 years (IQR 0.2-1.8) while the patients in the No Evidence group were 50% female with a median age of 6.5 years (IQR 1.5-12.9) (**Table 1**; **Supp. Fig. 1**). The difference in the age distribution of the two groups ($p<0.001$, Wilcoxon rank-sum test) reflected recognized epidemiological distinctions in the conditions typically leading to respiratory failure in very young versus older children[4,6].

Within the Definite group, clinical microbiologic testing identified viral infection alone in 46% of patients, bacterial infection alone in 14% of patients, and viral/bacterial co-infection in 40% of patients (**Table 1**). The most common pathogens were respiratory syncytial virus (RSV) and *Haemophilus influenzae*, which frequently co-occurred[6]. Diagnoses in the No Evidence group included trauma, neurological conditions, cardiovascular disease, airway abnormalities, ingestion of drugs/toxins, and sepsis that was clearly unconnected to LRTI. Nevertheless, most patients received antibiotic treatment by the time of TA sample collection in both the Definite (96%) and No Evidence (84%) groups (**Table 1**).

Classification of LRTI status based on TA host gene expression features

We first compared TA host gene expression between the Definite and No Evidence groups to determine whether it could distinguish patients based on LRTI status, regardless of the underlying cause of infection. We identified 4,718 differentially expressed genes at a Benjamini-Hochberg adjusted p-value < 0.05 (**Fig. 2a**; **Supp. Data 2**). As expected, gene set enrichment analysis (GSEA)[18,19] identified elevated expression of pathways involved in the immune response to infection in the Definite group (**Fig. 2b**; **Supp. Data 3**). Pathways related to the interferon response, a hallmark of anti-viral innate immunity, were most strongly upregulated, consistent with the high prevalence of viral infections in the Definite group. Additional immune pathways upregulated in this group included toll-like receptor signaling, cytokine signaling, inflammasome activation, neutrophil degranulation, antigen processing, and B cell and T cell receptor signaling. Conversely, pathways with reduced expression in the Definite group included translation, cilium assembly and lipid metabolism (**Fig. 2b**; **Supp. Data 3**).

Since we observed a clear host signature of infection, we proceeded to develop a classification approach to distinguish the Definite and No Evidence patients based on gene expression and evaluated its performance by 5-fold cross-validation. In each cross-validation loop, we: i) used lasso logistic regression on the samples in the training folds to select a parsimonious set of informative genes, ii) trained a random forest classifier using the selected genes, and iii) applied it to the samples in the test fold to obtain a host probability of LRTI.

Our approach yielded a median AUC of 0.967 (range: 0.953-0.996) (**Fig. 2c**), with the number of genes selected for use in the classifier ranging from 11 to 25 across the five train/test splits (**Supp. Table 1**). Using a 50% out-of-fold probability threshold to classify a patient as suffering from LRTI (LRTI+), the classifier assigned 92% of Definite patients and 80% of No Evidence patients according to their clinical LRTI adjudication (**Fig. 2d**).

Having validated the performance of our approach by cross-validation, we then applied lasso logistic regression to all the Definite and No Evidence patients to select a final set of genes (n=14) for later classification of patients with Suspected or Indeterminate LRTI status (**Fig. 2e**; **Supp. Table 2**). As expected, the genes in the final classifier set that were assigned high absolute regression coefficients were also repeatedly selected in the cross-validation procedure (**Supp. Table 2**).

The selected genes with the most positive regression coefficients, corresponding to higher expression in the Definite group, were: *GNLY*, encoding an anti-bacterial peptide present in cytolytic granules of cytotoxic T cells and natural killer cells[20]; *SLC38A2*, encoding a glutamine transporter upregulated in CD28-stimulated T cells[21,22]; *FFAR3*, encoding a G protein-coupled receptor activated by short-chain fatty acids that is induced by alveolar macrophages upon infection[23]; and the interferon-stimulated genes *PSMB8*, *ISG15* and *IRF1* (**Fig. 2e**; **Supp. Table 2**).

The selected genes with the most negative regression coefficients, corresponding to lower expression in the Definite group, were: *FABP4*, encoding a fatty acid-binding protein considered a marker of alveolar macrophages, whose expression in the lung decreases in patients with LRTI[24,25] (including COVID-19[26]); and *RBP4*, encoding a retinol-binding protein, whose expression in the lung has also been shown to sharply decrease following onset of LRTI[24] and whose expression by macrophages *in vitro* is depressed by inflammatory stimuli[27] (**Fig. 2e**; **Supp. Table 2**).

We examined the expression of the final classifier genes as a function of patient age to confirm that their selection was not influenced by the different age distributions of the Definite and No Evidence groups (**Supp. Fig. 2**). Reassuringly, we found no significant difference in the expression of the 14 genes when comparing No Evidence patients under the age of four (n=23) and over the age of four (n=27) (**Supp. Table 3a**). Further, we found that expression of 12 of

the genes remained significantly different when comparing only children under the age of four in the Definite (n=100) and No Evidence (n=23) groups (**Supp. Table 3b**).

Detection of pathogens by mNGS and definition of microbial classification features

We proceeded to analyze the microbial mNGS data to nominate likely pathogens whose features could be integrated into the LRTI classifier to increase confidence in the results and whose identity could be used to guide treatment. We processed the TA samples alongside water controls through the CZ-ID pipeline[28] to obtain a count matrix of microbial taxa. The water controls allowed us to generate a background count distribution for each taxon, which modeled the contribution of contamination by microbes present in the laboratory environment or reagents[6,29].

Viruses with known ability to cause LRTI[14] that were present at an abundance statistically exceeding their background distribution were considered probable pathogens. By this criterion, we detected viruses in the lungs of 107/117 (91%) Definite patients, with RSV the most prevalent (**Fig. 3a**). Among No Evidence patients, 8/50 (16%) also had viruses detected by mNGS, which were probably missed clinically in the absence of characteristic symptoms. We defined the summed abundance of all pathogenic viruses detected in a patient, measured in reads-per-million (rpM), as the patient's 'viral score' for later use in an integrated host/microbe classifier (**Fig. 3b**).

Because most Definite patients had a positive nasopharyngeal (NP) swab viral PCR test, we could compare the viruses detected by PCR and mNGS (**Supp. Data 4**). The comparison was complicated, however, by the fact that PCR was performed on upper airway samples, so a virus detected by PCR was not necessarily causal of LRTI. Bearing this in mind, we found that 99/101 (98%) Definite patients with a viral PCR hit also had a virus detected by mNGS, and both approaches detected at least one virus in common in 91 (92%) of those patients (**Supp. Fig. 3a**). Most cases where NP swab PCR detected a virus but mNGS did not involved

adenovirus (**Supp. Fig. 3b**). mNGS alone detected viruses in 8/16 (50%) Definite patients lacking a viral PCR hit (**Supp. Fig. 3a**). In a subset of Definite patients where we performed viral PCR on the same TA samples subjected to mNGS (n=21), 96% of PCR hits were detected by mNGS (**Supp. Table 4**).

Bacterial and fungal taxa in the mNGS data also underwent background filtering to retain only those present at an abundance statistically exceeding their background distribution based on water controls. Because asymptomatic carriage of pathogenic bacteria is common in children[4,6,30,31], we additionally applied a previously published algorithm to distinguish possible pathogens from commensals, called the rules-based model (RBM)[6,14]. The RBM identifies bacteria and fungi with known pathogenic potential that are relatively dominant in a sample (**Fig. 3c, d**), based on the principle that uncontrolled growth of a pathogen leads to reduced lung microbiome alpha diversity in the context of LRTI[14,32–34] (**Supp. Fig. 3c, d**).

The RBM identified possible bacterial/fungal pathogens in 78/117 (66%) Definite patients, with the most common being *H. influenzae*, *Moraxella catarrhalis* and *Streptococcus pneumoniae* (**Fig. 3e**). The RBM also identified potential bacterial/fungal pathogens in 17/50 (34%) No Evidence patients. Patients in the Definite group with an RBM-identified pathogen exhibited markedly lower bacterial alpha diversity compared to Definite patients without an RBM-identified pathogen and compared to No Evidence patients (**Supp. Fig. 3d**). In contrast, No Evidence patients with an RBM-identified pathogen did not typically exhibit a loss of bacterial alpha diversity (**Supp. Fig. 3d**), and in such cases the RBM-identified species was far less dominant (**Fig. 3f**). We therefore defined the patient's 'bacterial score' for use in an integrated host/microbe classifier as the proportion of the RBM-identified pathogens out of the non-host counts, a measure of relative dominance (**Fig. 3f**).

We next sought to compare the bacterial and fungal pathogens identified by mNGS with those found by culture of TA samples (**Supp. Data 4**). Importantly, mNGS can detect organisms that are challenging to grow in culture or are inhibited by previous antibiotic treatment[5], and the

RBM identifies the likeliest pathogen based on a global view of the microbiome. Despite these inherent differences, we found that in 44/63 (70%) Definite patients who had a positive culture, at least one pathogen identified by the RBM was also found by culture (**Supp. Fig. 3e**). In the remaining 19 patients, the RBM identified a different species than culture (n=7), or no pathogen at all (n=12). Even in these cases, the species grown in culture was usually present in the mNGS data but other species were more dominant (**Supp. Fig. 3e**). The RBM also identified a potential pathogen in 27/54 (50%) Definite patients lacking a positive culture (**Supp. Fig. 3e**). Most cases where the species grown in culture was absent from the mNGS data after background filtering involved *Staphylococcus aureus*, *Streptococcus* species other than *S. pneumoniae*, and *Escherichia coli* (**Supp. Fig. 3f**).

Host gene expression differences between viral and bacterial LRTI

Overall, mNGS identified viral and/or bacterial pathogens in 114/117 (97%) Definite patients. Having established by mNGS which Definite patients had an exclusively bacterial infection (n=7), an exclusively viral infection (n=36), or a viral/bacterial co-infection (n=71), we went back and examined how effectively the top host classifier genes captured these different scenarios (**Supp. Fig. 4a**). As expected, some of the interferon-stimulated genes (e.g., *ISG15*) provided much more discriminating power for Definite patients with a viral infection as compared to those with a purely bacterial infection. Reassuringly, however, several other classifier genes behaved similarly regardless of the underlying infection type.

We then asked more broadly whether host gene expression differed between patients with any bacterial LRTI (including viral co-infection) and patients with purely viral LRTI. We identified 108 differentially expressed genes at an adjusted p-value < 0.05 (**Supp. Fig. 4b**; **Supp. Data 2**), and found that genes related to neutrophil degranulation and cytokine signaling were enriched in patients with any bacterial LRTI (**Supp. Fig. 4c**; **Supp. Data 3**). These results

suggest the potential for developing, in future work, a rule-out classifier for bacterial infection that could be used to limit unnecessary antibiotic usage.

Classification of LRTI status based on integration of host and microbial features

Next, we asked whether integrating the host and microbial features could improve the performance of metagenomic LRTI classification. We fit a logistic regression model on the following three features: i) the LRTI probability output of the host classifier, ii) the summed abundance, measured in reads-per-million (rpM), of any pathogenic viruses present after background filtering ('viral score'), and iii) the proportion of the potentially pathogenic bacteria/fungi identified by the RBM out of all non-host read counts ('bacterial score') (**Fig. 4a**). As expected, the host and microbial features were correlated across most samples, but some notable exceptions were observed (**Supp. Fig. 5**).

The integrated classifier achieved an AUC of 0.986 (range: 0.953-1.000) when assessed by 5-fold cross-validation (**Fig. 4b**), applying the same train/test splits used in the host-only cross-validation. Using an out-of-fold probability threshold of 50%, the integrated classifier assigned 109/117 (93%) Definite patients as LRTI+ and 44/50 (88%) No Evidence patients as LRTI- (**Fig. 4c**). Compared to the host-only classifier, a net of five additional patients were now classified according to their clinical adjudication and the confidence of patient classifications significantly increased, as reflected by more extreme output probabilities (**Fig. 4d**). We note that at a much lower out-of-fold probability threshold of 15%, the integrated classifier's sensitivity for LRTI in the Definite group rose to >98%, suggesting a use-case as a rule-out test for LRTI.

Finally, we trained the integrated host/microbe classifier on all the Definite and No Evidence patients and then applied it to the Suspected and Indeterminate patients, whose clinical diagnosis was less certain. The integrated classifier indicated 37/57 (65%) Suspected patients were LRTI+ compared with 12/37 (32%) Indeterminate patients (**Fig. 5a**), consistent with the stronger clinical suspicion of LRTI in the former group. Across all n=52 patients classified as LRTI+ in these two groups, likely pathogens (viral, bacterial, or fungal) were identified in 51 patients (98%). Pathogens detected included common (e.g., rhinovirus, *H.*

influenzae), uncommon (e.g., bocavirus, parechovirus), and difficult to culture (e.g., *Mycoplasma pneumoniae*) microbes (**Fig. 5b**). We also designed a visual summary incorporating all three inputs of the integrated classifier and its output LRTI probability (**Fig. 5c**).

DISCUSSION

LRTI involves a dynamic relationship between pathogen, lung microbiome and host response that is not captured by existing clinical diagnostic tests. Here, we demonstrate that mNGS of lower respiratory samples enables accurate LRTI diagnosis based on features of each of these key elements in critically ill children, a demographic facing a high burden of LRTI. We build on proof-of-concept work in adults[14] to develop the first fully integrated host/microbe LRTI diagnostic classifier, and validate its performance in a large, multicenter prospective cohort.

Incidental carriage of pathogens in the respiratory tract is common in children[6,35]. Consistent with this, detection of a pathogen by mNGS was in many cases insufficient for accurate LRTI diagnosis in our cohort. Among No Evidence patients, 40% had potentially pathogenic microbes detected by mNGS even after application of the RBM (for bacteria and fungi). This is notably different from adults, for whom prior metagenomic studies have demonstrated much lower rates of incidental pathogen carriage[14,36]. Profiling the host response is thus particularly important for pediatric LRTI diagnosis, as it provides evidence of an immune response to infection.

Remarkably, an LRTI diagnostic classifier based on host gene expression performed very well on its own, with a median AUC of 0.967 by cross-validation. The host signature was driven by activation markers of T cells, alveolar macrophages and the interferon response, and successfully captured cases of viral infection, bacterial infection or co-infection. This performance suggests the gene signature could be incorporated into a clinical PCR assay as a standalone rapid diagnostic. It is likely that an even more parsimonious signature than the one

used in the mNGS classifier would suffice, as six genes exhibited the most discriminating power.

The integrated host/microbe classifier achieved a median AUC of 0.986 by cross-validation. The incorporation of microbial features significantly increased the confidence of LRTI classification, as reflected by more extreme output probabilities compared to the host-only classifier, even though relatively few patients switched their assigned diagnosis. It is likely that the integrated classification approach will prove even more valuable in settings where the host signature may not perform as well on its own (e.g., immune-compromised patients), and will generalize better to future cohorts. Moreover, it can provide clinicians with a unified framework both for LRTI diagnosis and pathogen identification.

Unlike for host gene expression, the microbial features in the integrated classifier were not automatically selected by training on identified taxa and their features in the Definite and No Evidence groups. Such an approach was not feasible given the sparse presence of individual respiratory pathogens across patients in the cohort, especially in the No Evidence group. As larger datasets are generated, it may be possible to use machine learning approaches to capture the ‘null distribution’ of incidentally carried pathogens in the lower respiratory tract and identify outlier cases that signal LRTI. Even then, designating a specific microbe as a ‘true’ causal pathogen for training purposes would be non-trivial, especially in cases of co-infection. Instead, we defined summary viral and bacterial scores motivated by accumulated clinical and microbiologic knowledge. For bacteria and fungi in particular, we took advantage of the collapse of lung microbiome diversity in the setting of pathogen dominance, an established feature of LRTI.

Comparison of mNGS and clinical microbiologic testing was complicated by inherent differences in the anatomical site of testing (upper respiratory viral PCR vs. lower respiratory mNGS) or the question addressed (growth in culture vs. dominance by mNGS), as well as by heterogeneity in microbiologic practices among study sites. Nevertheless, when clinical testing

identified a microbe, it was in most cases present in the mNGS data. A notable exception was adenovirus, which was consistently missed by mNGS compared to NP swab PCR. This could reflect sensitivity limitations of RNA-sequencing for a DNA virus or true absence in the lower airway. Our secondary analysis revealing higher concordance of PCR and mNGS when performed on the same lower respiratory specimens, however, argues for the latter possibility. Future work could examine targeted enrichment strategies[37,38] to improve detection of this or any other pathogen that proves challenging to detect by mNGS. Regardless, our findings highlight the value of concomitant assessment of the host response, which can accurately inform LRTI status even when pathogens are not detected.

A key advantage of mNGS is the capacity to provide a microbiologic diagnosis when traditional clinical testing returns negative, as in an estimated 30-60% of suspected community- or hospital-acquired pneumonia cases[3,39]. Indeed, the integrated mNGS classifier confirmed LRTI in 65% of children with suspected infection but negative clinician-ordered testing in our cohort, and in 32% of patients with respiratory failure of indeterminate etiology. It also provided a microbiologic diagnosis in all but one of these patients, highlighting the potential to inform pathogen-targeted versus empirical treatment.

Acute respiratory illnesses are a leading contributor to inappropriate antimicrobial use, a practice driven by challenges distinguishing LRTI from non-infectious causes of respiratory failure or distinguishing bacterial from viral LRTI. Reflecting this is the observation that 90% of children in our cohort received empiric antimicrobials by the time of sample collection, including 84% in the No Evidence group. Host/microbe mNGS offers an opportunity for improved antimicrobial stewardship, particularly in clinically uncertain cases, by providing a probability of infection and by nominating the likely pathogen. In fact, we found that the integrated classifier could be tuned to achieve >98% sensitivity for LRTI detection, highlighting its potential use as a rule-out test to help exclude the need for antimicrobials. Moreover, our host gene expression analysis revealed potential for development of a host classifier specifically for bacterial infection.

Metagenomic sequencing assays are increasingly being used in hospitals for microbe detection in patients with sepsis and meningitis, with turnaround times of ≤48 hours[40,41]. We envision the approach for LRTI diagnosis by host/microbe mNGS outlined in this study being used at the time of intubation for critically ill children with acute respiratory failure, as a complement to traditional culture and PCR-based microbiologic testing. Our approach would need to be further validated and its impact on clinical outcomes evaluated in a randomized clinical trial before deployment in the hospital. Future studies could be designed to overcome some of the limitations of the present work by: i) maintaining more closely matched age distributions among comparator groups; ii) including a larger fraction of purely bacterial infections in the cohort; and iii) applying uniform and comprehensive clinical microbiologic testing on lower airway samples across all patients, including those not suspected of LRTI. Finally, future work should examine the generalizability of our mNGS classifiers in a similarly large and heterogeneous adult cohort.

In summary, we report that combining host gene expression and unbiased microbial profiling from lower airway mNGS enables accurate LRTI diagnosis and pathogen identification in critically ill children. Future studies in a larger cohort can validate these findings, strengthen the utility of these classifiers, and assess the impact of mNGS on clinical outcomes.

METHODS

Study cohort

We conducted a secondary analysis of a prospective cohort study of mechanically ventilated children admitted to eight Pediatric Intensive Care units in the National Institute of Child Health and Human Development's Collaborative Pediatric Critical Care Research Network (CPCCRN) from February 2015 to December 2017[6,17]. The primary study was approved by the single Collaborative Pediatric Critical Care Research IRB at the University of Utah (protocol

#00088656). Informed consent was obtained from parents or other legal guardians, which included permission for collected specimens and data to be used in future studies.

We enrolled children aged 31 days to 18 years who were expected to require mechanical ventilation (MV) via endotracheal tube (ETT) for at least 72 hours. Exclusion criteria included inability to obtain a tracheal aspirate (TA) sample from the subject within 24 hours of intubation; presence of a tracheostomy tube or plans to place one; any condition in which deep tracheal suctioning was contraindicated; previous episode of MV during the hospitalization; family/team lack of commitment to aggressive intensive care as indicated by 'do not resuscitate' orders and/or other limitation of care; or previous enrollment into this study. Some patients were ultimately excluded from the present analysis based on sequencing metrics, as described in the following.

Parents or other legal guardians of eligible patients were approached for consent by study-trained staff as soon as possible after intubation. Waiver of consent was granted for TA samples to be obtained from standard-of-care suctioning of the ETT until the parents or guardians could be approached for informed consent.

Prospectively collected clinical data were recorded in a web-based research database maintained by the CPCCRN data coordinating center at the University of Utah.

Clinical microbiologic diagnostics

Enrolled patients received standard-of-care clinical respiratory microbiologic diagnostics, as ordered by treating clinicians at each study site. These diagnostics included nasopharyngeal (NP) swab respiratory viral testing by multiplex PCR and/or tracheal aspirate (TA) bacterial and fungal semi-quantitative cultures. Clinical diagnostic tests on samples obtained within 48 hours of intubation were included in the analyses. Microbes reported by the clinical laboratory as representing laboratory, skin or environmental contaminants, or reported as mixed upper respiratory flora, were excluded.

Adjudication of LRTI status

Adjudication of LRTI status was based on the final diagnoses reported by treating clinicians at each study site, who typically reviewed chest x-ray findings as part of this process, and any standard-of-care clinical respiratory microbiologic diagnostics performed during the admission.

Patients were ultimately assigned into one of four groups by study team physicians who were blinded to the mNGS results, according to a recently described approach[6,9,14]: i) Definite, where clinicians established an LRTI diagnosis and the patient had positive microbiologic findings; ii) Suspected, where clinicians established an LRTI diagnosis but the microbiologic testing performed returned negative; iii) Indeterminate, where it remained uncertain whether LRTI was a contributing factor to respiratory failure due to conflicting clinical and microbiologic findings; and iv) No Evidence, where clinicians identified a clear non-infectious cause of acute respiratory failure and no clinical or microbiologic suspicion of LRTI arose. Not all No Evidence patients underwent clinical microbiologic testing due to the absence of clinical suspicion.

Sample collection, processing and mNGS

Tracheal aspirate (TA) was collected within 24 hours of intubation, mixed 1:1 with DNA/RNA Shield (Zymo) and frozen at -80°C. RNA was extracted from 300µl of patient TA using bead-based lysis and the Allprep DNA/RNA kit (Qiagen), which included a DNase treatment step. RNA was reverse transcribed to generate cDNA, and sequencing library preparation was performed using the NEBNext Ultra II Library Prep Kit. RNA-Seq libraries underwent 150 nucleotide paired-end sequencing on an Illumina Novaseq 6000 instrument.

Host gene expression analysis

Following de-multiplexing, sequencing reads were pseudo-aligned with kallisto[42] (v. 0.46.1; including bias correction) to an index consisting of all transcripts associated with human protein coding and long non-coding RNA genes (ENSEMBL v.99). We excluded samples with less than 500,000 estimated counts associated with transcripts of protein-coding genes. Gene-

level counts were generated from the transcript-level abundance estimates using the R package tximport[43], with the scaledTPM method.

Genes were retained for differential expression (DE) analysis if they had at least 10 counts in at least 20% of the samples included in the analysis. DE analyses were performed with the R package limma[44], using quantile normalization and the voom method. Full DE results comparing: i) Definite and No Evidence patients, and ii) Definite patients with any bacterial LRTI and with purely viral LRTI are provided as **Supp. Data 2**.

Gene set enrichment analyses (GSEA)[18] were performed using the fgseaMultilevel function in the R package fgsea[19] on REACTOME[45] pathways with a minimum size of 10 genes and a maximum size of 1,500 genes. All genes from the respective DE analysis were included as input, pre-ranked by the DE test statistic. The gene sets shown in the figures were manually selected to reduce redundancy and highlight diverse biological functions from among those with a Benjamini-Hochberg adjusted p-value < 0.05. Full GSEA results are provided as **Supp. Data 3**.

Classification of LRTI status based on host gene expression features

Genes with at least 10 counts in at least 20% of the Definite (n=117) and No Evidence (n=50) patients were used as input for the host-based LRTI classification (n=13,323). We applied a variance-stabilizing transformation to the gene counts, as implemented in the R package DESeq2[46].

We implemented a 5-fold cross-validation procedure such that in each loop, we i) used lasso logistic regression on the samples in the training folds for feature (gene) selection, ii) trained a random forest classifier on the samples in the training folds using only the selected features, and iii) applied the random forest classifier to the samples in the test fold to obtain an out-of-fold host probability of LRTI. We required at least 9 No Evidence patients in each of the folds to ensure sufficient negative samples in each test set.

Simple lasso logistic regression was fit using the cv.glmnet(family='binomial') function from the R package glmnet[47], leaving all other parameters at their defaults. We used the 1se criterion for selecting the tuning parameter, which picks the sparsest value of the tuning parameter that lies within 1 standard error of the optimum. When evaluating test error, we selected the tuning parameter via nested cross-validation within the training set only.

Random forest was implemented using the R package randomForest[48]. We used 10,000 trees and left all parameters at their defaults.

Detection of microbes by mNGS and background filtering

We processed patient TA samples alongside water controls through the open-source CZ-ID (formerly called IDSeq) metagenomic analysis pipeline[28]. The pipeline performs subtractive alignment of the human genome and then reference-based alignment of the remaining reads at both the nucleotide and amino acid level against sequences in the National Center for Biotechnology Information (NCBI) nucleotide (NT) and non-redundant (NR) databases, respectively. This is followed by assembly of the reads matching each taxon. Taxa with ≥5 read counts in the NT alignment and an average assembly nucleotide alignment ≥70bp were retained for downstream analysis.

Water controls enabled estimation of the number of background reads expected for each taxon, as previously described[6,29]. 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 batch (sequencing run) and taxon, we estimated the mean parameter of the negative binomial distribution by averaging the read counts across all negative controls after normalizing by the total non-host reads, slightly regularizing this estimate by including the global average (across all batches) as an additional sample. We estimated a single dispersion parameter across all taxa and batches using the functions glm.nb() and theta.md() from the R package MASS[49]. Taxa were then tested for whether they exceeded the count expected from the background distribution, and a Holm-Bonferroni correction was applied to all tests performed in

the same sample. Taxa were considered present in a sample if they achieved an adjusted p-value < 0.05.

Any virus with known ability to cause LRTI, based on a previously conducted literature curation[14], that was present in a patient sample after background filtering was considered a probable pathogen.

Rules-based model (RBM)

For bacteria and fungi that were present after background filtering, we additionally applied a rules-based algorithm for distinguishing potential pathogens from likely commensals, which was slightly adapted from its previously published version[14].

We retained only the most abundant bacterial/fungal species from each genus. In case a less abundant species in the genus had known ability to cause LRTI[14], we retained it as well. We then ranked these species from greatest to least overall abundance in the sample and retained, at most, the top 15. The largest drop in abundance between the ranked species in the sample was identified. All species with an abundance above the largest drop were selected. If any species above the largest drop had known ability to cause LRTI, it was deemed a potential pathogen.

Analysis of microbiome diversity

The Shannon diversity index was calculated using either all viral and bacterial taxa, or only bacterial taxa, that were present after background filtering using the R package Vegan[50]. Mann-Whitney tests with Bonferroni correction were used to evaluate statistical significance of group differences.

Classification of LRTI status based on integration of host and microbial features

For the integrated host/microbe LRTI classifier, we fit a logistic regression model on the following features: i) the host LRTI probability; ii) the summed abundance, measured in reads-per-million (rpM), of any pathogenic viruses present after background filtering ('viral score'); and iii) the proportion of any potentially pathogenic bacteria/fungi identified by the RBM out of all non-host read counts ('bacterial score').

The integrated classifier was evaluated using 5-fold cross-validation using the same train/test splits from the host classifier cross-validation and the same per-fold host classifiers. To avoid any leakage from the test set affecting the host probabilities, we used the out-of-bag random forest votes as the host probabilities of the training samples.

Before fitting the integrated classifier, we applied a logistic (log-odds) transformation to the host probabilities. In order to apply this transformation, we slightly regularized the raw probabilities. For the viral/bacterial scores, we applied a $\log_{10}$ transformation. In order to avoid taking the log of 0, we added a small uniform quantity to the scores of all the samples, which was calculated by taking the minimum non-zero viral or bacterial score, respectively, in the corresponding training set and dividing it by 10.

Data availability

Source data are provided with this paper. Processed gene counts are available in the NCBI Gene Expression Omnibus (GEO) database under accession GSE212532. FASTQ files subtracted of reads aligning to the human genome due to patient privacy concerns are available under NCBI BioProject ID PRJNA748764.

Code availability

Code for the background correction algorithm is available at: (https://github.com/czbiohub/idseqr/). Code for the differential expression, classifier development and RBM will be made available upon publication.

Acknowledgements

We thank all subjects and their families for participating in this study. We acknowledge the contributions of Tammara L. Jenkins, MSN, RN, and Robert F. Tamburro, MD, from Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD. The following is a summary of Performance Sites, Principal Investigators (PI), Co-Investigators (CI), Research Coordinators (RC), and Allied Research Personnel. Children's Hospital of Colorado, Aurora, CO: Peter Mourani (PI); Todd Carpenter (CI); Yamila Sierra (RC); Katheryn Malone (RC), Diane Ladell (RC); Kimberly Ralston (RC); Kevin Van (RC). Children's Hospital of

Michigan, Detroit, MI: Kathleen L. Meert (PI); Sabrina Heidemann (CI); Ann Pawluszka (RC); Melanie Lulic (RC). Children's Hospital of Philadelphia, Philadelphia, PA: Robert A Berg (PI); Athena Zuppa (CI); Carolann Twelves (RC); Mary Ann DiLiberto (RC). Children's National Medical Center, Washington, DC: Murray Pollack (PI); David Wessel (PI); Randall Burd (CI); Elyse Tomanio (RC); Diane Hession (RC); Ashley Wolfe (RC). Nationwide Children's Hospital, Columbus, OH: Mark Hall (PI); Andrew Yates (CI); Lisa Steele (RC); Maggie Flowers (RC); Josey Hensley (RC). Mattel Children's Hospital, University of California Los Angeles, Los Angeles, CA: Anil Sapru (PI); Rick Harrison (CI), Neda Ashtari (RC); Anna Ratiu (RC). Children's Hospital of Pittsburgh, University of Pittsburgh Medical Center, Pittsburgh, PA: Joe Carcillo (PI); Ericka Fink (CI); Leighann Koch (RC); Alan Abraham (RC). Benioff Children's Hospital, University of California, San Francisco, San Francisco, CA: Patrick McQuillen (PI); Anne McKenzie (RC); Yensy Zetino (RC). University of Utah; Data Coordinating Center, Salt Lake City, Utah: Mike Dean (PI); Richard Holubkov (PI), Juhee Peterson, Melissa Bolton, Whit Coleman, and Stephanie Dorton.

This study was supported in part by the following cooperative agreements from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) and the National Heart, Lung and Blood Institute (NHLBI): UG1HD083171 (Dr. Mourani), 1R01HL124103 (Drs. Mourani and Sontag), UG1HD049983 (Dr. Carcillo), UG01HD049934 (Drs. Reeder, Locandro, and Dean), UG1HD083170 (Dr. Hall), UG1HD050096 (Dr. Meert), UG1HD63108 (Dr. Zuppa), UG1HD083116 (Dr. Sapru), UG1HD083166 (Dr. McQuillen), UG1HD049981 (Dr. Pollack), K23HL138461-01A1 and 5R01HL155418-02 (Dr. Langelier). The study was also supported by funding from the Chan Zuckerberg Biohub (Dr. Langelier). Study sponsors were not involved in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the report for publication.

References

1. World Health Organization. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.
2. U.S. Centers for Disease Control and Prevention. Leading Causes of Death. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm (2022).
3. Jain, S., Finelli, L., & CDC EPIC Study Team. Community-acquired pneumonia among U.S. children. *N. Engl. J. Med.* **372**, 2167–2168 (2015).
4. O'Brien, K. L. *et al.* Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. *The Lancet* **394**, 757–779 (2019).
5. Jain, S. *et al.* Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. *N. Engl. J. Med.* **373**, 415–427 (2015).
6. Tsitsiklis, A. *et al.* Lower respiratory tract infections in children requiring mechanical ventilation: a multicentre prospective surveillance study incorporating airway metagenomics. *Lancet Microbe* **3**, e284–e293 (2022).
7. Baggs, J., Fridkin, S. K., Pollack, L. A., Srinivasan, A. & Jernigan, J. A. Estimating National Trends in Inpatient Antibiotic Use Among US Hospitals From 2006 to 2012. *JAMA Intern. Med.* **176**, 1639–1648 (2016).
8. Tamma, P. D., Avdic, E., Li, D. X., Dzintars, K. & Cosgrove, S. E. Association of Adverse Events With Antibiotic Use in Hospitalized Patients. *JAMA Intern. Med.* **177**, 1308–1315 (2017).
9. Tsalik, E. L. *et al.* Host gene expression classifiers diagnose acute respiratory illness etiology. *Sci. Transl. Med.* **8**, 322ra11 (2016).
10. Zaas, A. K. *et al.* The current epidemiology and clinical decisions surrounding acute respiratory infections. *Trends Mol. Med.* **20**, 579–588 (2014).
11. Sweeney, T. E., Wong, H. R. & Khatri, P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Sci. Transl. Med.* **8**, 346ra91 (2016).
12. Suarez, N. M. *et al.* Superiority of Transcriptional Profiling Over Procalcitonin for Distinguishing Bacterial From Viral Lower Respiratory Tract Infections in Hospitalized Adults. *J. Infect. Dis.* **212**, 213–222 (2015).
13. Zaas, A. K. *et al.* Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe* **6**, 207–217 (2009).
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. U. S. A.* **115**, E12353–E12362 (2018).
15. Mick, E. *et al.* Upper airway gene expression shows a more robust adaptive immune response to SARS-CoV-2 in children. *Nat. Commun.* **13**, 3937 (2022).

605 16. Molony, R. D. *et al.* Aging impairs both primary and secondary RIG-I signaling for interferon
606 induction in human monocytes. *Sci. Signal.* **10**, eaan2392 (2017).

607 17. Mourani, P. M. *et al.* Temporal airway microbiome changes related to ventilator associated
608 pneumonia in children. *Eur. Respir. J.* (2020) doi:10.1183/13993003.01829-2020.

609 18. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
610 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550
611 (2005).

612 19. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021)
613 doi:10.1101/060012.

614 20. Krensky, A. M. & Clayberger, C. Biology and clinical relevance of granulysin. *Tissue Antigens*
615 **73**, 193–198 (2009).

616 21. Nakaya, M. *et al.* Inflammatory T Cell Responses Rely on Amino Acid Transporter ASCT2
617 Facilitation of Glutamine Uptake and mTORC1 Kinase Activation. *Immunity* **40**, 692–705
618 (2014).

619 22. Carr, E. L. *et al.* Glutamine Uptake and Metabolism Are Coordinately Regulated by
620 ERK/MAPK during T Lymphocyte Activation. *J. Immunol.* **185**, 1037–1044 (2010).

621 23. Liu, Q., Tian, X., Maruyama, D., Arjomandi, M. & Prakash, A. Lung immune tone via gut-lung
622 axis: gut-derived LPS and short-chain fatty acids' immunometabolic regulation of lung IL-1β,
623 FFAR2, and FFAR3 expression. *Am. J. Physiol.-Lung Cell. Mol. Physiol.* **321**, L65–L78 (2021).

624 24. Pathak, K. V. *et al.* Molecular Profiling of Innate Immune Response Mechanisms in
625 Ventilator-associated Pneumonia. *Mol. Cell. Proteomics* **19**, 1688–1705 (2020).

626 25. Perea, L. *et al.* Reduced airway levels of fatty-acid binding protein 4 in COPD: relationship
627 with airway infection and disease severity. *Respir. Res.* **21**, 21 (2020).

628 26. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-
629 19. *Nat. Med.* **26**, 842–844 (2020).

630 27. Broch, M. *et al.* Macrophages are novel sites of expression and regulation of retinol binding
631 protein-4 (RBP4). *Physiol. Res.* **59**, 299–303 (2010).

632 28. Kalantar, K. L. *et al.* IDseq—An open source cloud-based pipeline and analysis service for
633 metagenomic pathogen detection and monitoring. *GigaScience* **9**, (2020).

634 29. Mick, E. *et al.* Upper airway gene expression reveals suppressed immune responses to SARS-
635 CoV-2 compared with other respiratory viruses. *Nat. Commun.* **11**, 5854 (2020).

636 30. S, H. *et al.* Rhinovirus Detection in Symptomatic and Asymptomatic Children: Value of Host
637 Transcriptome Analysis. *Am. J. Respir. Crit. Care Med.* **193**, (2016).

638 31. Hf, W. *et al.* The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect. Dis.*
639 **5**, (2005).

640 32. Dickson, R. P. *et al.* Analysis of culture-dependent versus culture-independent techniques
641 for identification of bacteria in clinically obtained bronchoalveolar lavage fluid. *J. Clin.*
642 *Microbiol.* **52**, 3605–3613 (2014).

643 33. Abreu, N. A. *et al.* Sinus microbiome diversity depletion and *Corynebacterium*
644 *tuberculostearicum* enrichment mediates rhinosinusitis. *Sci. Transl. Med.* **4**, 151ra124
645 (2012).

646 34. Langelier, C. *et al.* Metagenomic Sequencing Detects Respiratory Pathogens in
647 Hematopoietic Cellular Transplant Patients. *Am. J. Respir. Crit. Care Med.* **197**, 524–528
648 (2018).

649 35. Schlaberg, R. *et al.* Viral Pathogen Detection by Metagenomics and Pan-Viral Group
650 Polymerase Chain Reaction in Children With Pneumonia Lacking Identifiable Etiology. *J.*
651 *Infect. Dis.* **215**, 1407–1415 (2017).

652 36. Kalantar, K. L. *et al.* Metagenomic comparison of tracheal aspirate and mini-bronchial
653 alveolar lavage for assessment of respiratory microbiota. *Am. J. Physiol. Lung Cell. Mol.*
654 *Physiol.* **316**, L578–L584 (2019).

655 37. Deng, X. *et al.* Metagenomic sequencing with spiked primer enrichment for viral diagnostics
656 and genomic surveillance. *Nat. Microbiol.* **5**, 443–454 (2020).

657 38. Quan, J. *et al.* FLASH: a next-generation CRISPR diagnostic for multiplexed detection of
658 antimicrobial resistance sequences. *Nucleic Acids Res.* **47**, e83 (2019).

659 39. Magill, S. S. *et al.* Changes in Prevalence of Health Care–Associated Infections in U.S.
660 Hospitals. *N. Engl. J. Med.* **379**, 1732–1744 (2018).

661 40. Wilson, M. R. *et al.* Clinical Metagenomic Sequencing for Diagnosis of Meningitis and
662 Encephalitis. *N. Engl. J. Med.* **380**, 2327–2340 (2019).

663 41. Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA
664 sequencing test for infectious disease. *Nat. Microbiol.* **4**, 663–674 (2019).

665 42. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
666 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

667 43. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level
668 estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2015).

669 44. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and
670 microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

671 45. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472-477
672 (2014).

673 46. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
674 RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

675 47. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models
676 via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

48. Liaw, A. & Wiener, M. Classification and Regression by randomForest. **2**, 5 (2002).

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

50. vegan: Community Ecology Package version 2.3-5 from R-Forge. https://rdrr.io/rforge/vegan/.

51. Simon, T. D., Haaland, W., Hawley, K., Lambka, K. & Mangione-Smith, R. Development and Validation of the Pediatric Medical Complexity Algorithm (PMCA) Version 3.0. *Acad. Pediatr.* **18**, 577–580 (2018).

Table 1: Demographic and clinical cohort characteristics.

	Definite (n=117)	No Evidence (n=50)	P-value*	Suspected (n=57)	Indeterminate (n=37)
Female, n (%)	45 (38.5%)	25 (50.0%)	0.18	26 (45.6%)	16 (43.2%)
Age, median [IQR]	0.5 [0.2, 1.8]	6.5 [1.5, 12.9]	<0.001	1.7 [0.5, 6.0]	1.5 [0.6, 10.8]
Race, n (%)					
White	69 (59.0%)	30 (60.0%)	0.99	33 (57.9%)	20 (54.1%)
Black/African American	26 (22.2%)	7 (14.0%)	0.29	11 (19.3%)	10 (27.0%)
Asian	5 (4.3%)	6 (12.0%)	0.088	2 (3.5%)	2 (5.4%)
American Indian or Alaskan Native	1 (0.9%)	1 (2.0%)	0.99	1 (1.8%)	0 (0.0%)
Native Hawaiian/Other Pacific Islander	1 (0.9%)	0 (0.0%)	0.99	0 (0.0%)	0 (0.0%)
More than one race	3 (2.6%)	1 (2.0%)	0.99	1 (1.8%)	1 (0.03%)
Unknown	12 (10.3%)	5 (10.0%)	0.99	9 (15.8%)	4 (10.8%)
Hispanic or Latino, n (%)	17 (14.5%)	6 (12.0%)	0.81	14 (24.6%)	7 (18.9%)
Comorbidities (CCC)†, n (%)	38 (32.5%)	26 (52.0%)	0.024	34 (59.7%)	14 (37.8%)
Immunosuppressed, n (%)	3 (2.6%)	7 (14.0%)	0.0085	5 (8.8%)	6 (16.2%)
Admission category, n (%)					
Medical	117 (100.0%)	28 (56.0%)	p<0.001	57 (100.0%)	29 (78.4%)
Surgical	0 (0.0%)	15 (30.0%)	p<0.001	0 (0.0%)	3 (8.1%)
Trauma	0 (0.0%)	7 (14.0%)	p<0.001	0 (0.0%)	4 (10.8%)
Time from hospital admission to intubation (hours), median [IQR]	4.8 [0.0, 23.6]	3.5 [0.0, 20.9]	0.60	2.6 [0.0, 15.9]	1.0 [0.0, 26.5]
Abx on or before sample date‡, n(%)	112 (95.7%)	42 (84.0%)	0.022	51 (89.5%)	30 (69.8%)

*Comparing Definite and No Evidence patients. Wilcoxon rank sum test used for all continuous variables. Fisher's exact test used for all categorical variables.
†Complex Chronic Conditions[51].
‡Antibiotic treatment started on or prior to the date of sample collection.

Supplementary Tables

Supplementary Table 1: Host genes selected in each of the 5 cross-validation loops.

Gene ID	Coefficient	Gene symbol	Fold	AUC
(Intercept)	-2.0280997	NA	1	0.996
ENSG00000115523	0.40435424	*GNLY*	1	
ENSG00000114737	0.29736499	*CISH*	1	
ENSG00000185897	0.2334893	*FFAR3*	1	
ENSG00000134294	0.23225034	*SLC38A2*	1	
ENSG00000204264	0.14737796	*PSMB8*	1	
ENSG00000133101	0.13698008	*CCNA1*	1	
ENSG00000125347	0.08788657	*IRF1*	1	
ENSG00000152766	0.01618487	*ANKRD22*	1	
ENSG00000103356	-0.005208	*EARS2*	1	
ENSG00000272168	-0.0129449	*CASC15*	1	
ENSG00000163710	-0.0141204	*PCOLCE2*	1	
ENSG00000196139	-0.0145091	*AKR1C3*	1	
ENSG00000163735	-0.0202633	*CXCL5*	1	
ENSG00000164631	-0.0296153	*ZNF12*	1	
ENSG00000162929	-0.031873	*KIAA1841*	1	
ENSG00000102962	-0.0470626	*CCL22*	1	
ENSG00000196305	-0.0541444	*IARS1*	1	
ENSG00000259094	-0.0642868	*AC013457.1*	1	
ENSG00000203865	-0.0785439	*ATP1A1-AS1*	1	
ENSG00000115414	-0.0963792	*FN1*	1	
ENSG00000132300	-0.1094826	*PTCD3*	1	
ENSG00000273173	-0.1236364	*SNURF*	1	
ENSG00000138207	-0.1867265	*RBP4*	1	
ENSG00000170323	-0.2260672	*FABP4*	1	
ENSG00000182141	-0.2407744	*ZNF708*	1	
(Intercept)	-0.4183313	NA	2	0.953
ENSG00000134294	0.21115084	*SLC38A2*	2	
ENSG00000185897	0.17219249	*FFAR3*	2	
ENSG00000115523	0.15221918	*GNLY*	2	
ENSG00000204264	0.0974774	*PSMB8*	2	
ENSG00000152766	0.06800369	*ANKRD22*	2	
ENSG00000272821	0.05245633	*U62317.2*	2	
ENSG00000187608	0.03953433	*ISG15*	2	
ENSG00000162929	-0.0224986	*KIAA1841*	2	
ENSG00000196139	-0.0991982	*AKR1C3*	2	
ENSG00000170323	-0.1720699	*FABP4*	2	
ENSG00000253729	-0.2010562	*PRKDC*	2	
ENSG00000138207	-0.2940594	*RBP4*	2	

(Intercept)	0.27758103	NA	3	0.986
ENSG00000185507	0.29760005	*IRF7*	3	
ENSG00000133101	0.22848885	*CCNA1*	3	
ENSG00000115523	0.20275662	*GNLY*	3	
ENSG00000185897	0.09319602	*FFAR3*	3	
ENSG00000134294	0.05523801	*SLC38A2*	3	
ENSG00000272821	0.03959668	*U62317.2*	3	
ENSG00000196189	0.03111383	*SEMA4A*	3	
ENSG00000136231	0.01793655	*IGF2BP3*	3	
ENSG00000114737	0.01066616	*CISH*	3	
ENSG00000117143	-0.0043278	*UAP1*	3	
ENSG00000163735	-0.0087013	*CXCL5*	3	
ENSG00000113068	-0.0087115	*PFDN1*	3	
ENSG00000149021	-0.0088169	*SCGB1A1*	3	
ENSG00000232629	-0.0158939	*HLA-DQB2*	3	
ENSG00000164631	-0.0351792	*ZNF12*	3	
ENSG00000273173	-0.0502438	*SNURF*	3	
ENSG00000196139	-0.0532949	*AKR1C3*	3	
ENSG00000170324	-0.0607793	*FRMPD2*	3	
ENSG00000259094	-0.0637096	*AC013457.1*	3	
ENSG00000272660	-0.1024448	*AC090425.2*	3	
ENSG00000008226	-0.1276673	*DLEC1*	3	
ENSG00000272168	-0.1499889	*CASC15*	3	
ENSG00000170323	-0.4076148	*FABP4*	3	
(Intercept)	-5.181784	NA	4	0.954
ENSG00000175073	0.50212823	*VCPIP1*	4	
ENSG00000115523	0.23413097	*GNLY*	4	
ENSG00000135604	0.23226792	*STX11*	4	
ENSG00000168394	0.13854581	*TAP1*	4	
ENSG00000185897	0.13639215	*FFAR3*	4	
ENSG00000185885	0.07866599	*IFITM1*	4	
ENSG00000133106	0.05899777	*EPSTI1*	4	
ENSG00000158769	-0.0055638	*F11R*	4	
ENSG00000163710	-0.0206461	*PCOLCE2*	4	
ENSG00000149021	-0.0263191	*SCGB1A1*	4	
ENSG00000182141	-0.1006603	*ZNF708*	4	
ENSG00000138207	-0.1413799	*RBP4*	4	
ENSG00000170323	-0.2005039	*FABP4*	4	
(Intercept)	0.5726104	NA	5	0.967
ENSG00000204264	0.27278452	*PSMB8*	5	
ENSG00000115523	0.0762971	*GNLY*	5	
ENSG00000175073	0.05579582	*VCPIP1*	5	
ENSG00000185897	0.04184826	*FFAR3*	5	

ENSG00000188820	0.03110363	*CALHM6*	5	
ENSG00000133106	0.02410902	*EPSTI1*	5	
ENSG00000187608	0.00687972	*ISG15*	5	
ENSG00000163710	-0.0191247	*PCOLCE2*	5	
ENSG00000151914	-0.029031	*DST*	5	
ENSG00000163735	-0.1149234	*CXCL5*	5	
ENSG00000170323	-0.389078	*FABP4*	5	

Supplementary Table 2: Genes selected for the final host classifier by lasso logistic regression applied to all the Definite and No Evidence patients, with their regression coefficients. The number of times the gene was selected across the 5 cross-validation (CV) loops is also indicated.

Gene ID	Gene symbol	Gene product	Coefficient	Times selected in CV
ENSG00000115523	*GNLY*	Granulysin	0.257	5
ENSG00000204264	*PSMB8*	Proteasome subunit beta 8	0.249	3
ENSG00000185897	*FFAR3*	Free fatty acid receptor 3	0.224	5
ENSG00000134294	*SLC38A2*	Solute carrier family 38 member 2	0.214	3
ENSG00000187608	*ISG15*	ISG15 ubiquitin-like modifier	0.070	2
ENSG00000125347	*IRF1*	Interferon regulatory factor 1	0.027	1
ENSG00000162929	*KIAA1841* (also known as *SANBR*)	SANT and BTB domain regulator of class switch recombination	-0.014	2
ENSG00000272660	*AC090425.2*	Long non-coding RNA, antisense to *ACTL6A*	-0.016	1
ENSG00000196139	*AKR1C3*	Aldo-keto reductase family 1 member C3	-0.019	3
ENSG00000163735	*CXCL5*	C-X-C motif chemokine ligand 5	-0.019	3
ENSG00000080546	*SESN1*	Sestrin 1	-0.033	0
ENSG00000163710	*PCOLCE2*	Procollagen C-endopeptidase enhancer 2	-0.033	3
ENSG00000138207	*RBP4*	Retinol binding protein 4	-0.167	3
ENSG00000170323	*FABP4*	Fatty acid binding protein 4	-0.297	5
(Intercept)			-3.112	

Supplementary Table 3: Results for the n=14 final classifier genes of differential expression analyses comparing **a)** No Evidence patients under four (n=23) versus over four years old (n=27); and **b)** Definite (n=100) and No Evidence (n=23) patients under four years old.

a

Gene symbol	Log_2 fold change	P-value	Adjusted P-value
GNLY	-1.14	0.01	0.72
PSMB8	-0.33	0.21	0.82
FFAR3	-0.43	0.33	0.87
SLC38A2	-0.57	0.07	0.75
ISG15	-1.68	0.01	0.72
IRF1	-0.30	0.25	0.83
KIAA1841	0.14	0.59	0.94
AC090425.2	1.48	0.07	0.76
AKR1C3	0.81	0.10	0.76
CXCL5	0.24	0.67	0.95
SESN1	0.27	0.47	0.91
PCOLCE2	0.54	0.21	0.81
RBP4	-0.04	0.95	0.99
FABP4	-0.56	0.45	0.90

b

Gene symbol	Log_2 fold change	P-value	Adjusted P-value
GNLY	2.73	2.11E-08	1.46E-06
PSMB8	1.11	7.78E-08	4.22E-06
FFAR3	2.93	9.70E-12	4.25E-09
SLC38A2	0.60	2.50E-05	4.33E-04
ISG15	3.49	3.71E-09	3.89E-07
IRF1	1.52	5.18E-12	2.88E-09
KIAA1841	-0.89	2.84E-05	4.76E-04
AC090425.2	-0.18	7.50E-01	8.56E-01
AKR1C3	-2.48	2.24E-12	1.42E-09
CXCL5	-2.63	4.62E-09	4.52E-07
SESN1	-0.53	4.92E-02	1.38E-01
PCOLCE2	-2.06	2.60E-09	2.93E-07
RBP4	-3.64	1.51E-17	1.00E-13
FABP4	-5.54	4.02E-26	5.36E-22

Supplementary Table 4: Comparison of mNGS viral detection in TA samples with PCR viral detection in nasopharyngeal (NP) swabs or in the same TA samples in a subset of patients.

Definite patients with matched NP swab and TA viral PCR testing (n=21)	Agreement of mNGS with NP swab PCR	Congruence of mNGS and NP swab PCR	Agreement of mNGS with TA PCR	Congruence of mNGS and TA PCR
All viruses	22/34 = 64.7%	22/37 = 59.5%	23/24 = 95.8%	23/26 = 88.5%
RSV	11/12 = 91.7%	11/12 = 91.7%	10/10 = 100%	10/11 = 90.9%
Rhinovirus	4/7 = 57.1%	4/10 = 40%	6/6 = 100%	6/7 = 85.7%
Adenovirus	0/6 = 0%	0/6 = 0%	0/0 = 100%	0/0 = 100%
Coronavirus	2/3 = 66.7%	2/3 = 66.7%	2/2 = 100%	2/2 = 100%
Human metapneumovirus	3/3 = 100%	3/3 = 100%	3/3 = 100%	3/3 = 100%
Parainfluenza virus	2/3 = 66.7%	2/3 = 66.7%	2/3 = 66.7%	2/3 = 66.7%

Agreement reflects the number of viruses detected by mNGS out of the total number of viruses detected by PCR.
Congruence reflects the number of viruses detected by both mNGS and PCR out of the total number of viruses detected by at least one method.
RSV=respiratory syncytial virus.

Figures

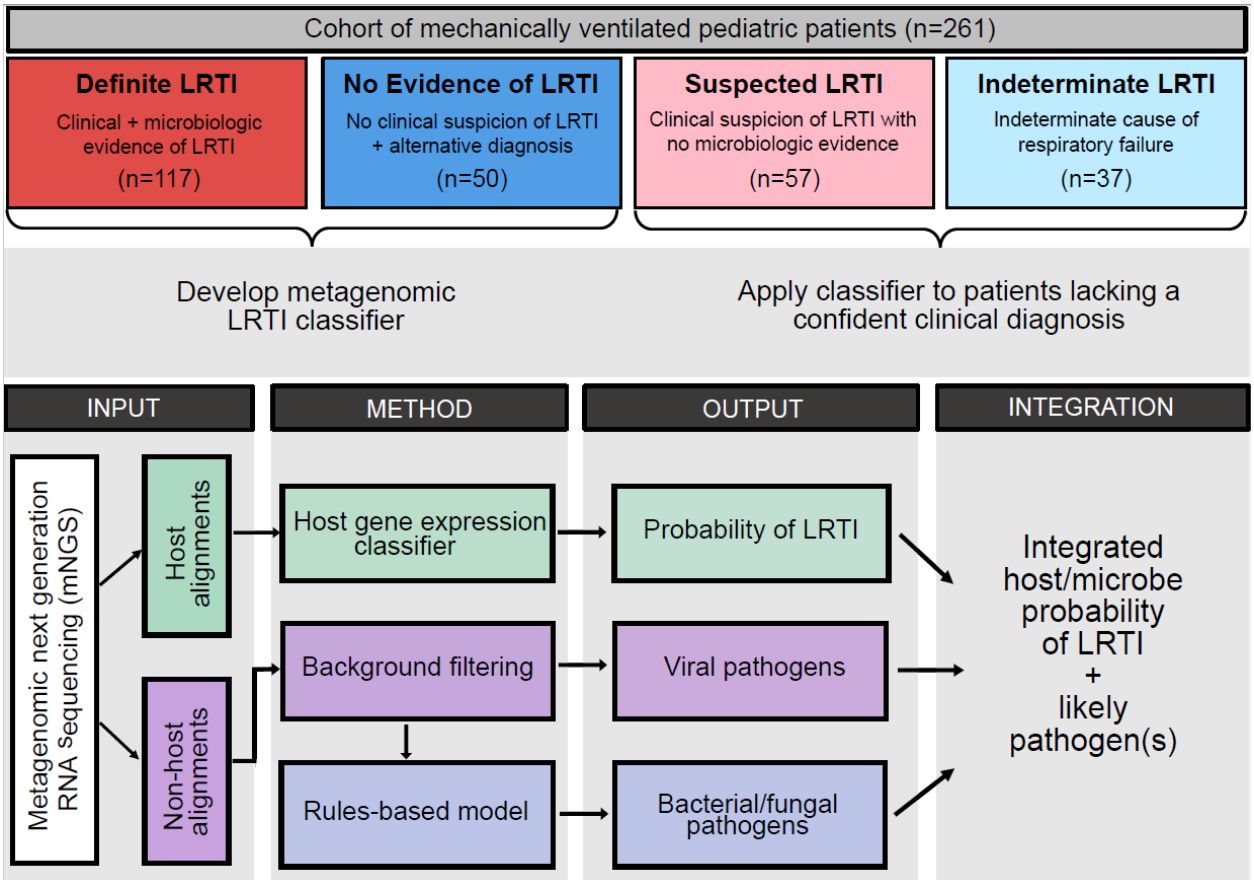


Figure 1: Study overview.

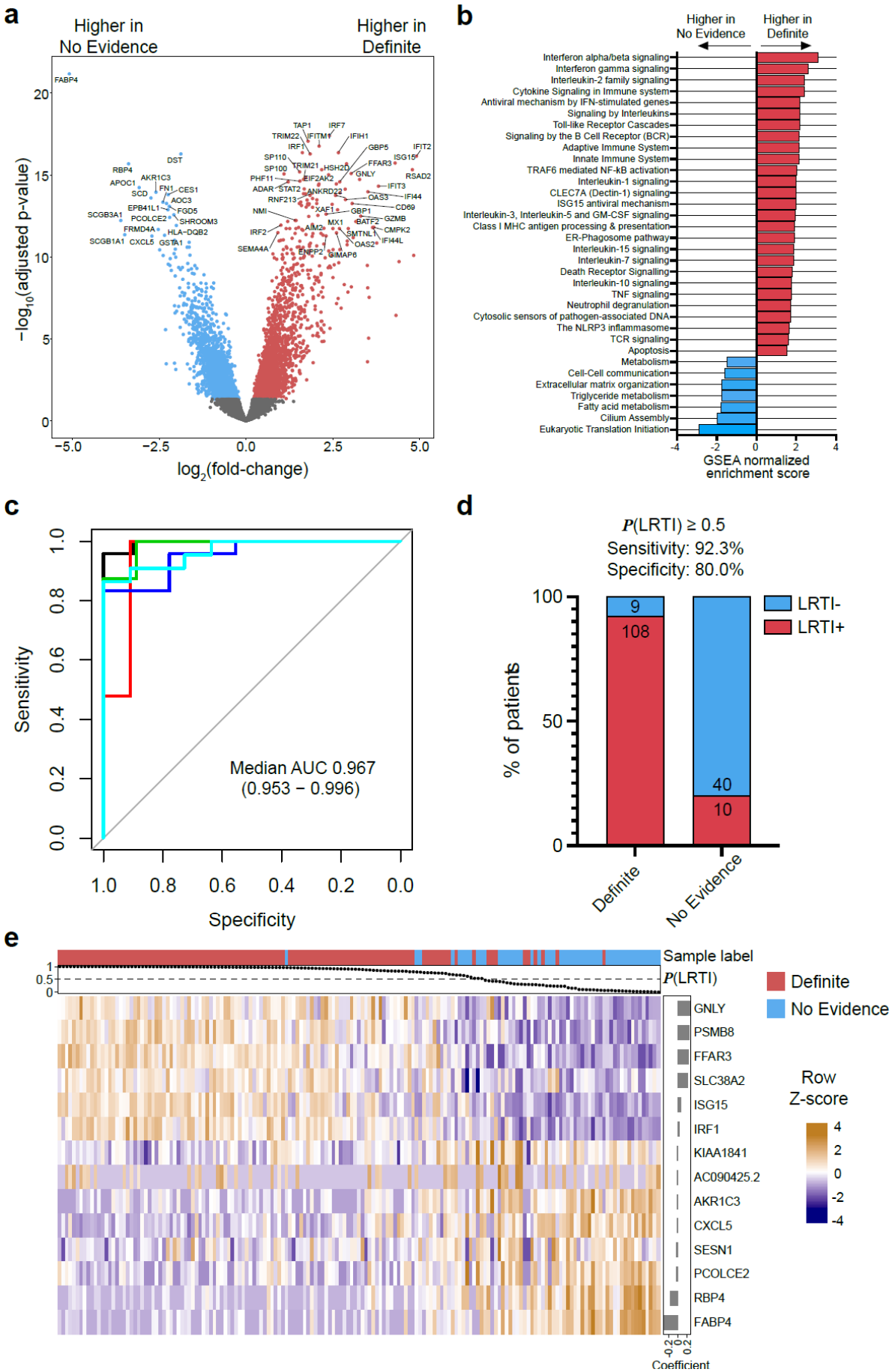

a
Higher in No Evidence
Higher in Definite
−log10(adjusted p-value)
log2(fold-change)
b
Higher in No Evidence
Higher in Definite
Interferon alpha/beta signaling
Interferon gamma signaling
Interleukin-2 family signaling
Cytokine Signaling in Immune system
Antiviral mechanism by IFN-stimulated genes
Signaling by Interleukins
Toll-like Receptor Cascades
Signaling by the B Cell Receptor (BCR)
Adaptive Immune System
Innate Immune System
TRAF6 mediated NF-kB activation
Interleukin-1 signaling
CLEC7A (Dectin-1) signaling
ISG15 antiviral mechanism
Interleukin-3, Interleukin-5 and GM-CSF signaling
Class I MHC antigen processing & presentation
ER-Phagosome pathway
Interleukin-15 signaling
Interleukin-7 signaling
Death Receptor Signalling
Interleukin-10 signaling
TNF signaling
Neutrophil degranulation
Cytosolic sensors of pathogen-associated DNA
The NLRP3 inflammasome
TCR signaling
Apoptosis
Metabolism
Cell-Cell communication
Extracellular matrix organization
Triglyceride metabolism
Fatty acid metabolism
Cilium Assembly
Eukaryotic Translation Initiation
GSEA normalized enrichment score
c
Sensitivity
Specificity
Median AUC 0.967
(0.953 − 0.996)
d
P(LRTI) ≥ 0.5
Sensitivity: 92.3%
Specificity: 80.0%
% of patients
LRTI-
LRTI+
Definite
No Evidence
e
Sample label
P(LRTI)
Definite
No Evidence
Row Z-score
GNLY
PSMB8
FFAR3
SLC38A2
ISG15
IRF1
KIAA1841
AC090425.2
AKR1C3
CXCL5
SESN1
PCOLCE2
RBP4
FABP4
Coefficient

Figure 2: Host gene expression classifier for LRTI diagnosis.
a) Volcano plot highlighting genes differentially expressed (DE) between Definite and No Evidence patients. Colored genes reached statistical significance (adjusted p-value < 0.05). **b)** Normalized enrichment scores of selected REACTOME pathways that reached statistical significance (adjusted p-value < 0.05) in the GSEA using DE genes between Definite and No Evidence patients. **c)** Receiver operating characteristic (ROC) curve showing the performance of a host gene expression classifier in each of the 5 cross-validation folds. **d)** Bar plot showing the number and percentage of Definite and No Evidence patients that were classified according to their clinical adjudication, using a 50% out-of-fold probability threshold. **e)** Heatmap showing the normalized expression across patients (columns) of the 14 final classifier genes (rows) selected when training on the complete Definite and No Evidence dataset. Also shown are the regression coefficients of each selected gene and the out-of-fold probability of LRTI assigned to each patient.

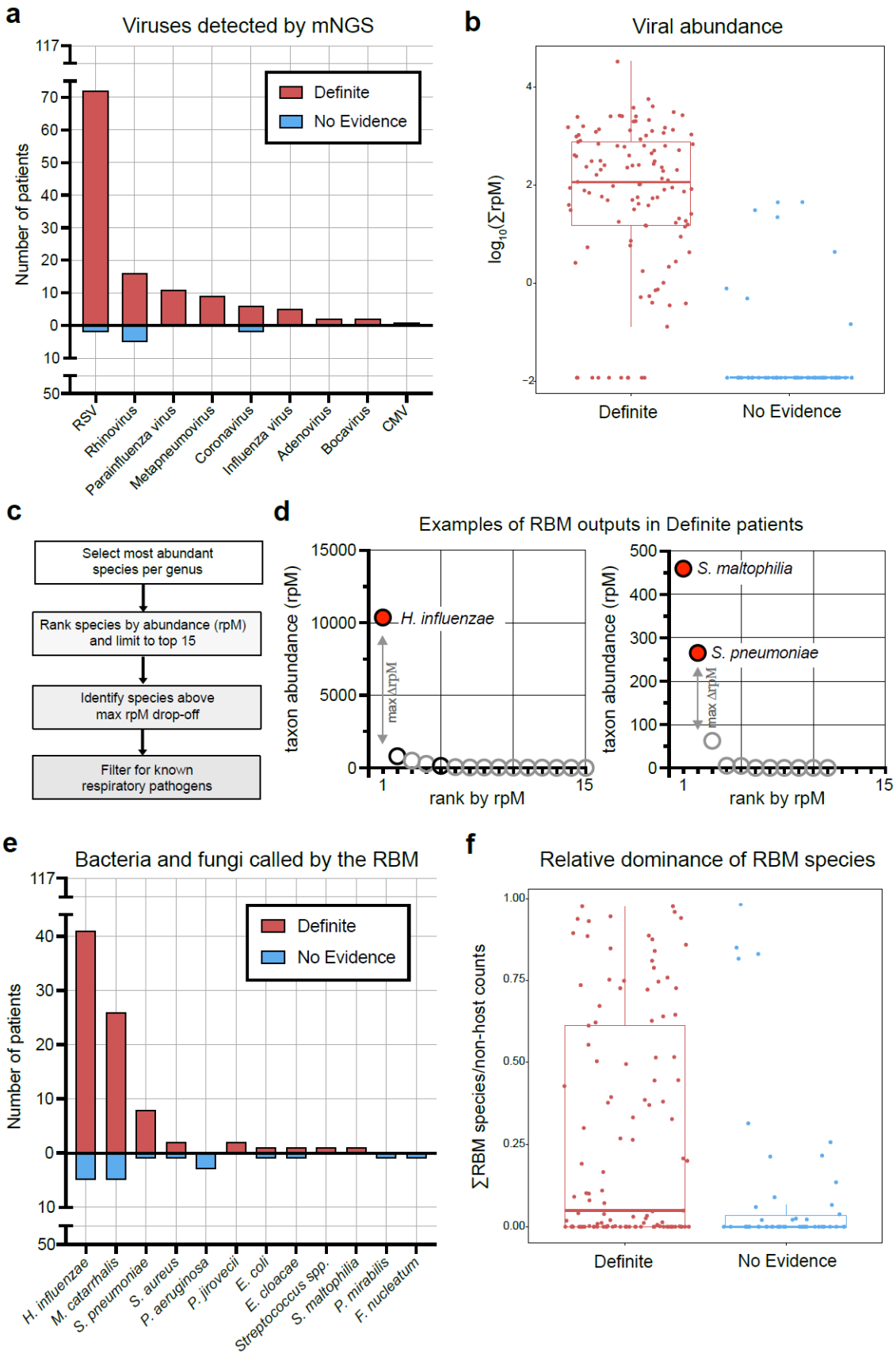

a
Viruses detected by mNGS
Definite
No Evidence
Number of patients
RSV
Rhinovirus
Parainfluenza virus
Metapneumovirus
Coronavirus
Influenza virus
Adenovirus
Bocavirus
CMV
b
Viral abundance
log10(∑rpM)
Definite
No Evidence
c
Select most abundant species per genus
Rank species by abundance (rpM) and limit to top 15
Identify species above max rpM drop-off
Filter for known respiratory pathogens
d
Examples of RBM outputs in Definite patients
taxon abundance (rpM)
H. influenzae
max ΔrpM
rank by rpM
S. maltophilia
S. pneumoniae
e
Bacteria and fungi called by the RBM
Number of patients
H. influenzae
M. catarrhalis
S. pneumoniae
S. aureus
P. aeruginosa
P. jirovecii
E. coli
E. cloacae
Streptococcus spp.
S. maltophilia
P. mirabilis
F. nucleatum
f
Relative dominance of RBM species
∑RBM species/non-host counts
Definite
No Evidence

Figure 3: Metagenomic identification of respiratory pathogens.
a) Bar plot showing the distribution of viruses detected by mNGS after background filtering in the Definite and No Evidence patients. RSV=respiratory syncytial virus; CMV=cytomegalovirus. **b)** Boxplot showing the $\log_{10}$-transformed summed abundance, measured in reads-per-million (rpM), of any pathogenic viruses detected in a patient, separated by group. **c)** Analysis steps applied as part of the rules-based model (RBM), a heuristic approach designed to identify potential bacterial/fungal pathogens in the context of LRTI. **d)** Graphical illustration of the RBM results obtained in two Definite patients. Each dot represents a bacterial/fungal species most abundant in its respective genus. A species on the list of known respiratory pathogens has a black outline, otherwise the outline is gray. A species above the maximum drop-off in rpM has a red fill, otherwise the fill is white. **e)** Bar plot showing the distribution of bacteria/fungi called as potential pathogens by the RBM in the Definite and No Evidence patients. *Streptococcus spp.=Streptococcus* species other than *S. pneumoniae*. **f)** Boxplot showing the proportion of the RBM-identified pathogens out of all non-host counts in each patient, separated by group.

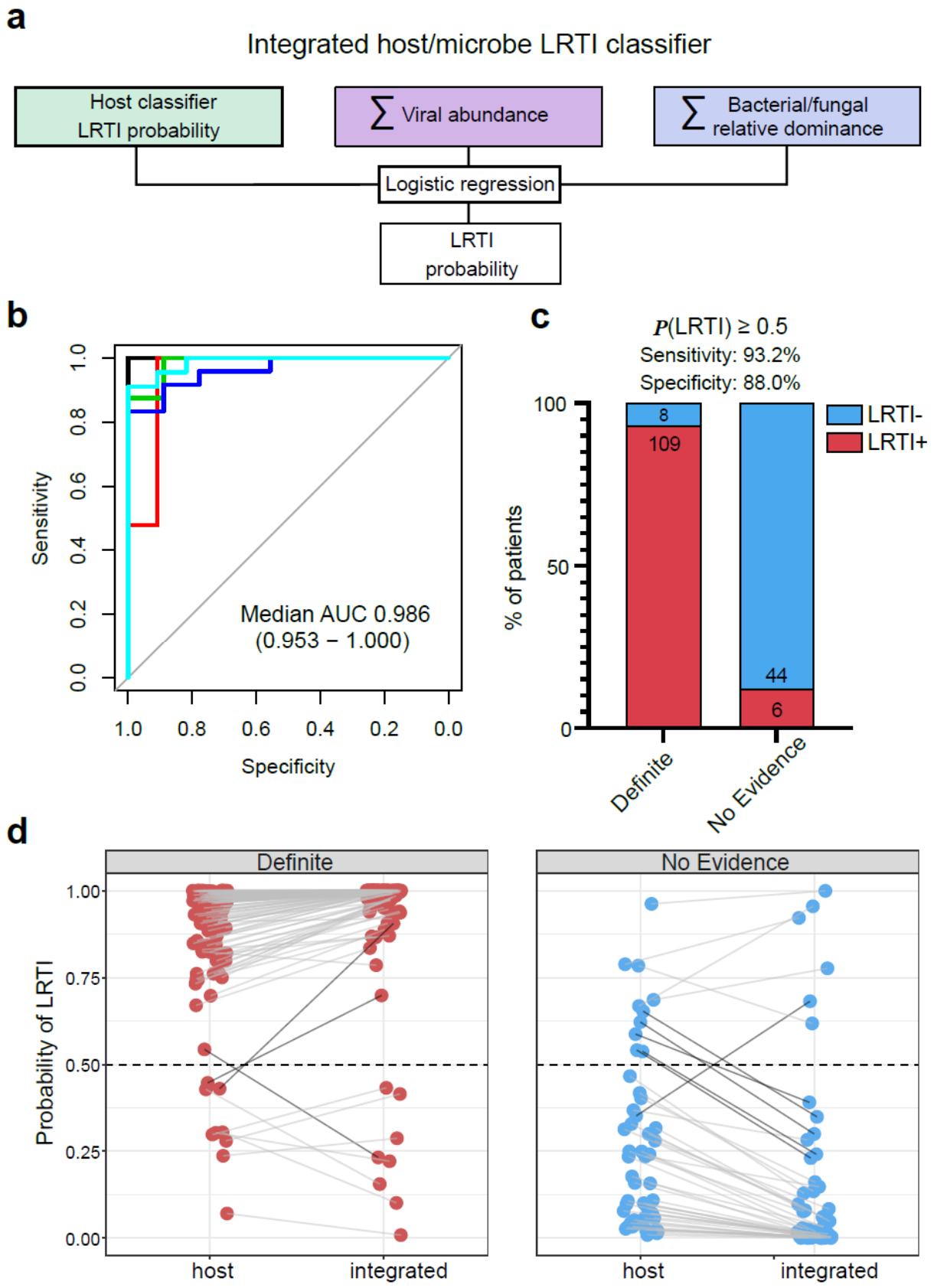

a
Integrated host/microbe LRTI classifier
Host classifier LRTI probability
∑ Viral abundance
∑ Bacterial/fungal relative dominance
Logistic regression
LRTI probability
b
Sensitivity
Specificity
Median AUC 0.986
(0.953 – 1.000)
c
P(LRTI) ≥ 0.5
Sensitivity: 93.2%
Specificity: 88.0%
% of patients
LRTI-
LRTI+
8
109
44
6
Definite
No Evidence
d
Definite
No Evidence
Probability of LRTI
host
integrated

Figure 4: Integrated host/microbe classifier for LRTI diagnosis.
a) Schematic of the integrated host/microbe classifier, which incorporates three input features into a logistic regression model in order to calculate the probability of LRTI. **b)** Receiver operating characteristic (ROC) curve showing the performance of the integrated classifier on Definite and No Evidence patients across the 5 cross-validation folds. **c)** Bar plot showing the number and percentage of Definite and No Evidence patients that were classified according to their clinical adjudication using a 50% out-of-fold probability threshold. **d)** Probability of LRTI derived from the host classifier and the integrated classifier for Definite (left panel) and No Evidence (right panel) patients. Dark connecting lines represent samples whose LRTI probability crossed the 50% threshold between the two classifiers.

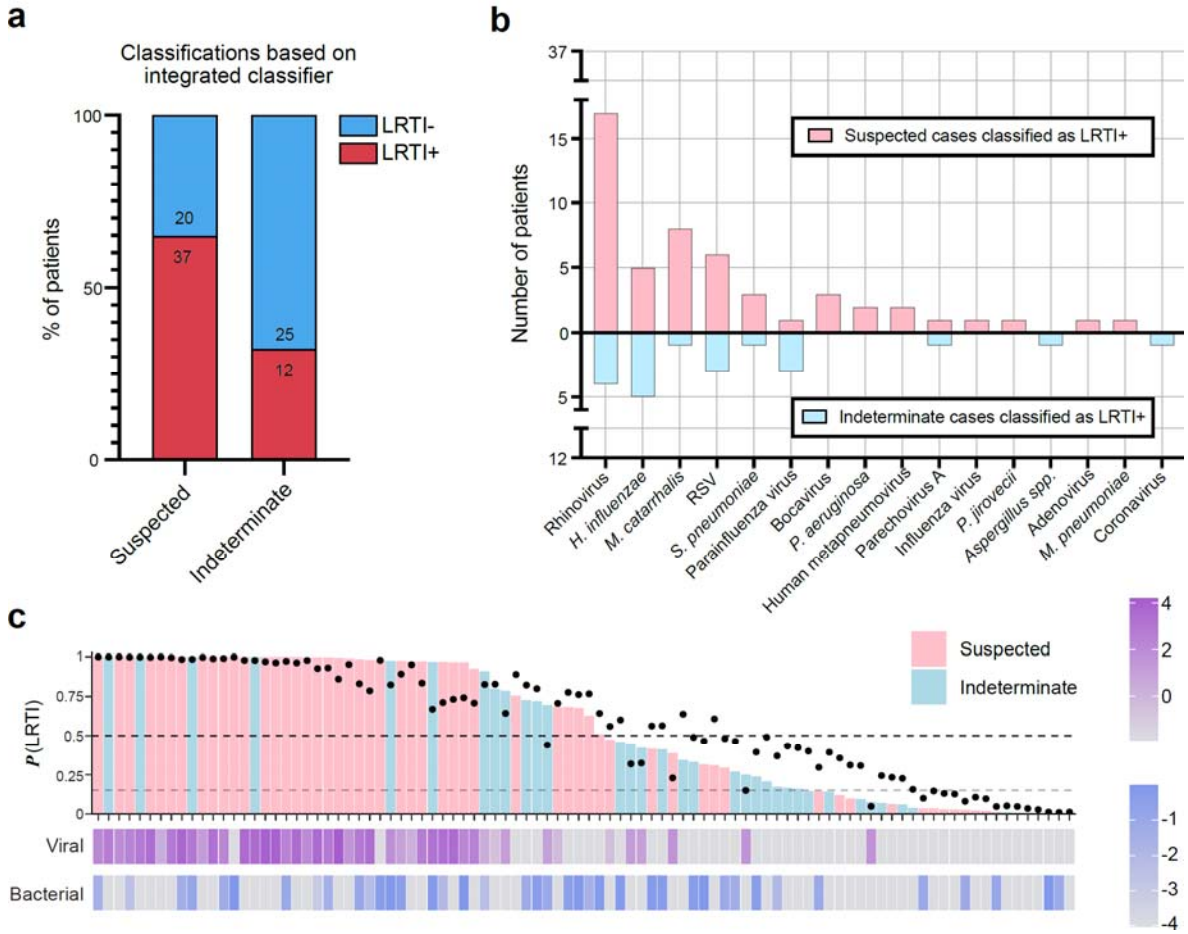
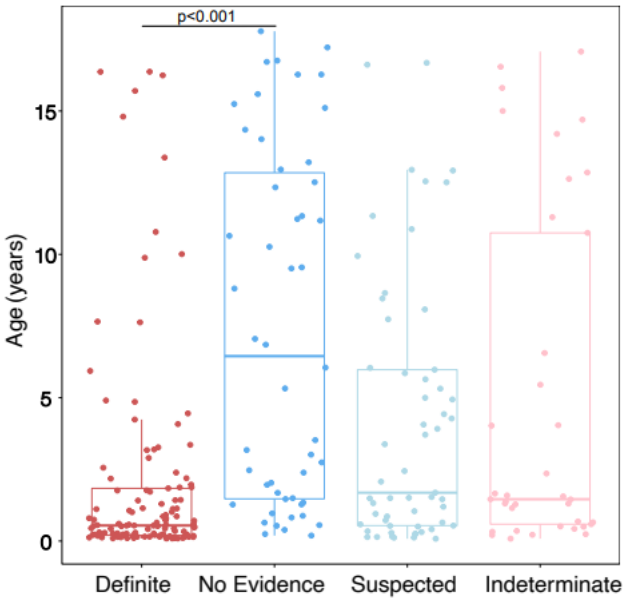
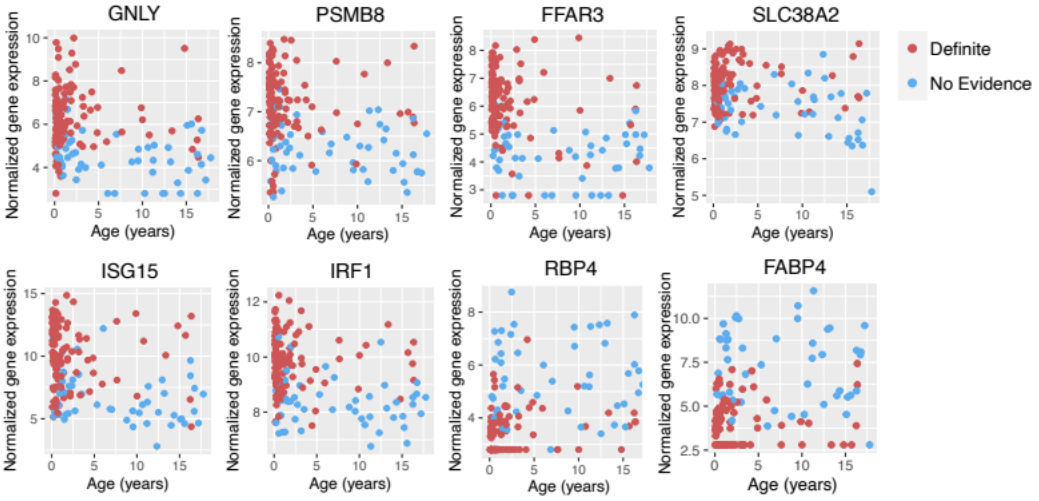


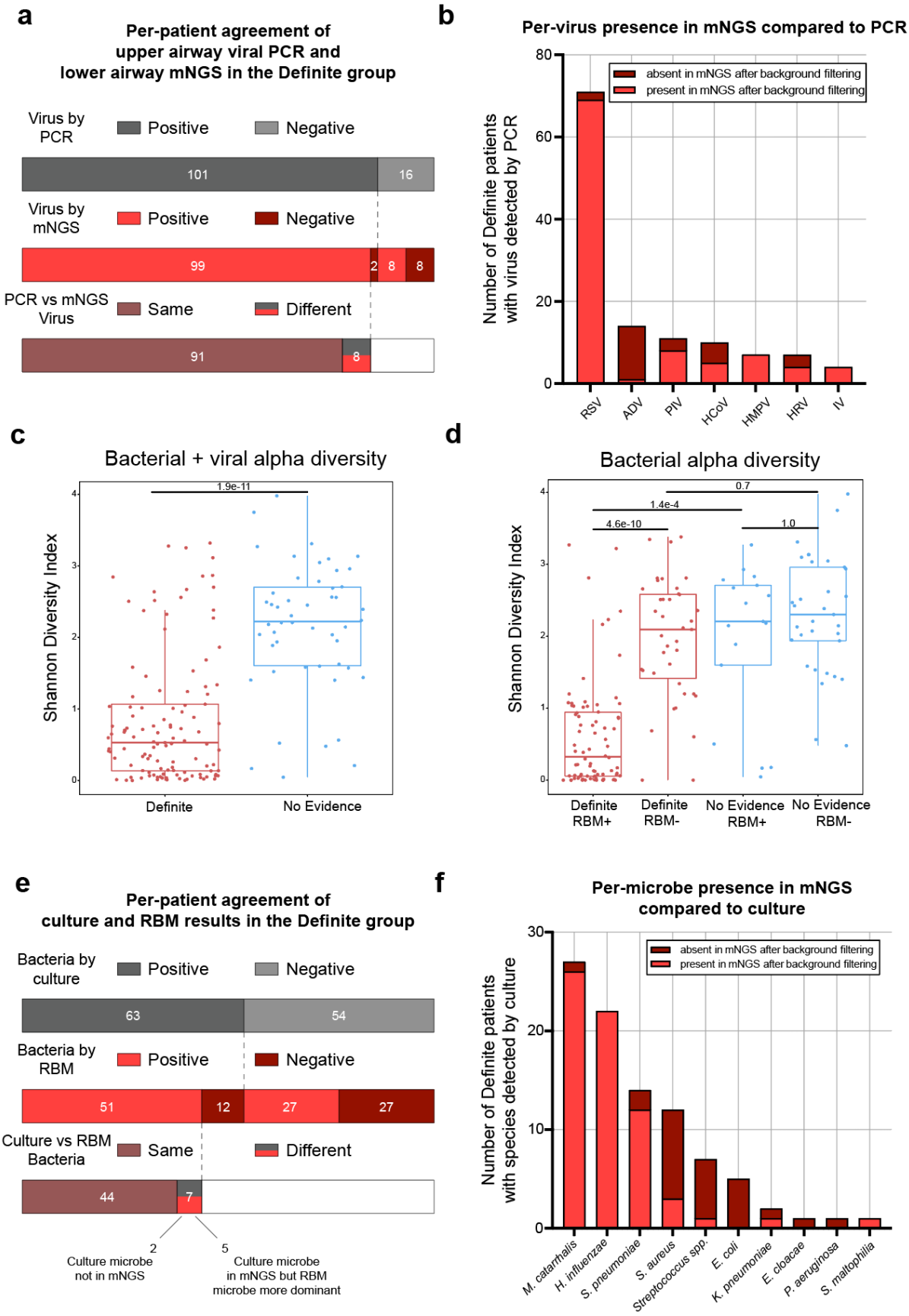
Figure 5: Application of the integrated classifier to Suspected and Indeterminate patients. **a)** Bar plot showing the number and percentage of Suspected and Indeterminate patients that were classified as LRTI+ by the integrated classifier using a 50% probability threshold. **b)** Viruses detected by mNGS and bacteria/fungi identified by the RBM across the patients classified as LRTI+. **c)** Overview of inputs and output of the integrated classifier for all Suspected and Indeterminate patients. Top bars denote the integrated probability of LRTI and are colored by patient group; black dots represent the input host LRTI probability; bottom vertical bars show the input $\log_{10}$-transformed viral and bacterial scores. Dashed lines indicate the 50% LRTI probability threshold and the 15% 'rule-out' threshold.



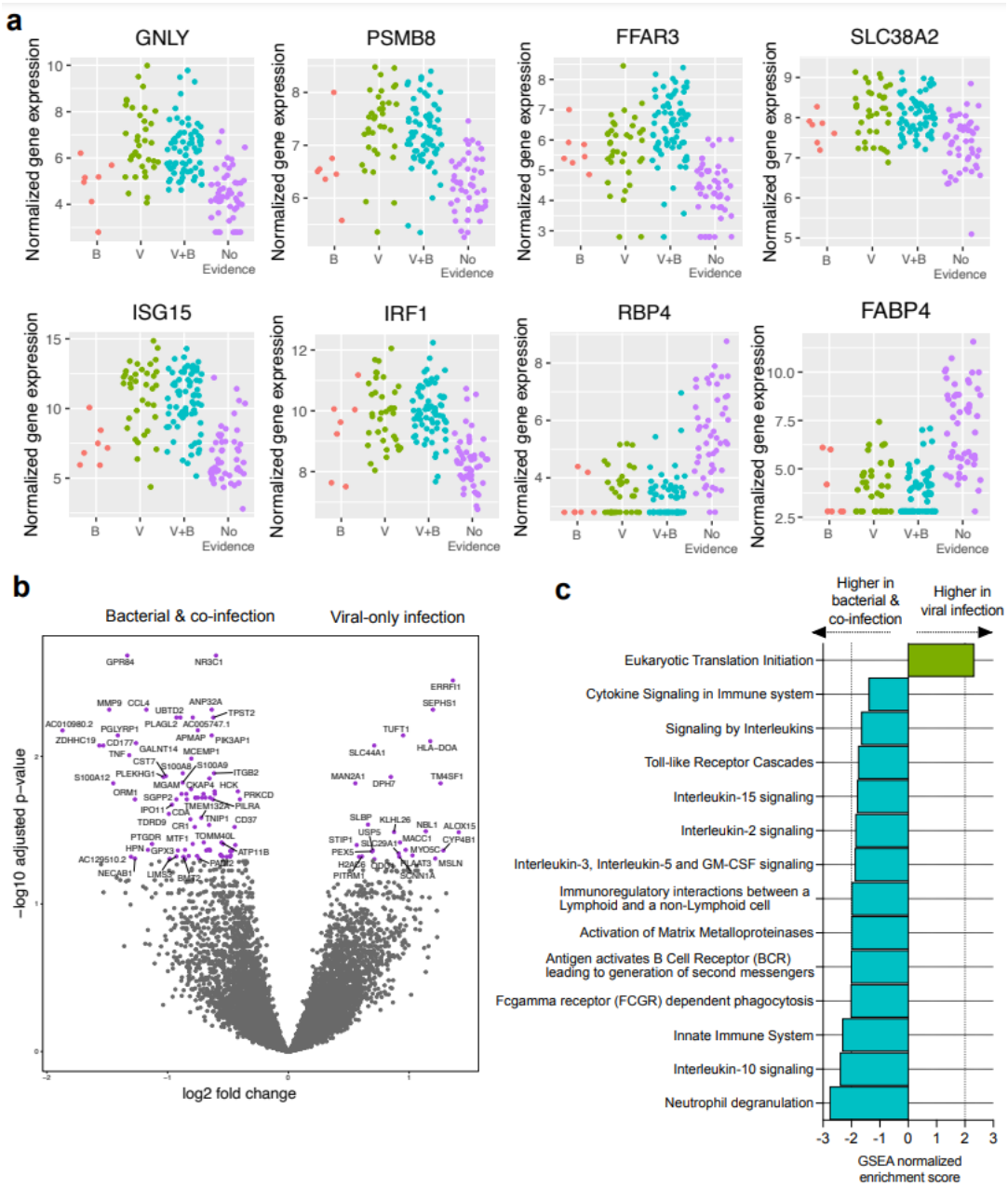
Supplementary Figure 1: Age distribution across the four LRTI status groups. P-value between Definite and No Evidence patients was calculated using a Mann-Whitney test.



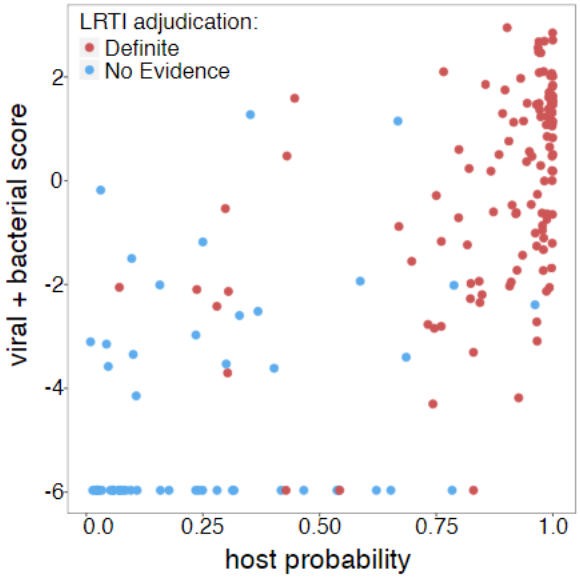
Supplementary Figure 2: Expression of eight host classifier genes as a function of age in Definite (red) and No Evidence (blue) patients.


a
Per-patient agreement of upper airway viral PCR and lower airway mNGS in the Definite group
Virus by PCR
Positive
Negative
101
16
Virus by mNGS
Positive
Negative
99
2
8
8
PCR vs mNGS Virus
Same
Different
91
8
b
Per-virus presence in mNGS compared to PCR
absent in mNGS after background filtering
present in mNGS after background filtering
Number of Definite patients with virus detected by PCR
RSV
ADV
PIV
HCoV
HMPV
HRV
IV
c
Bacterial + viral alpha diversity
1.9e-11
Shannon Diversity Index
Definite
No Evidence
d
Bacterial alpha diversity
0.7
1.4e-4
4.6e-10
1.0
Shannon Diversity Index
Definite RBM+
Definite RBM-
No Evidence RBM+
No Evidence RBM-
e
Per-patient agreement of culture and RBM results in the Definite group
Bacteria by culture
Positive
Negative
63
54
Bacteria by RBM
Positive
Negative
51
12
27
27
Culture vs RBM Bacteria
Same
Different
44
7
2
Culture microbe not in mNGS
5
Culture microbe in mNGS but RBM microbe more dominant
f
Per-microbe presence in mNGS compared to culture
absent in mNGS after background filtering
present in mNGS after background filtering
Number of Definite patients with species detected by culture
M. catarrhalis
H. influenzae
S. pneumoniae
S. aureus
Streptococcus spp.
E. coli
K. pneumoniae
E. cloacae
P. aeruginosa
S. maltophilia

Supplementary Figure 3: a) Diagram depicting the agreement at the patient level between viral clinical testing and mNGS detection after background filtering in the Definite group. Agreement between the two methods in a patient was defined as at least one virus identified by both. **b)** Bar plot showing the number of cases of each virus detected by clinical upper respiratory PCR testing and the proportion that was also present by mNGS after background filtering. RSV=respiratory syncytial virus; ADV=adenovirus; PIV=parainfluenza virus; HCoV=seasonal human coronavirus; HMPV=human metapneumovirus; HRV=human rhinovirus; IV=influenza virus. **c)** Boxplots of bacterial + viral microbiome alpha diversity measured by the Shannon index in Definite and No Evidence patients. **d)** Boxplots of bacterial-only alpha diversity measured by the Shannon index. Definite and No Evidence patients are split by whether a potential pathogen was identified by the RBM. P-values in c) and d) were calculated by a Mann-Whitney test with Bonferroni correction. **e)** Diagram depicting the agreement at the patient level between clinical culture and the results of the RBM in the Definite group. Agreement between the two methods in a patient was defined as at least one species identified by both. **f)** Bar plot showing the number of cases of each species detected by clinical culture and the proportion that was also present by mNGS after background filtering. *Streptococcus spp.=Streptococcus* species other than *S. pneumoniae*.



Supplementary Figure 4: a) Expression of eight host classifier genes in Definite patients with only bacterial pathogens identified by the RBM (B; n=7), only viral pathogens detected by mNGS (V; N=36), viral and bacterial pathogens (V+B; n=71), and the No Evidence patients (n=50) for comparison. Three patients from the Definite group are not shown because they did not have any pathogens identified by mNGS. One No Evidence sample in the plot of *SLC38A2* was omitted since it was an extreme outlier. **b)** Volcano plot highlighting genes differentially expressed (DE) between Definite patients with any bacterial infection (bacterial-only + co-infection) and viral-only infection. Genes colored in purple reached statistical significance (adjusted p-value < 0.05). **c)** Normalized enrichment scores of selected REACTOME pathways that reached statistical significance (adjusted p-value < 0.05) in the GSEA using the DE genes.



Supplementary Figure 5: Scatterplot of the host LRTI probability (x-axis) and the sum of the $\log_{10}$-transformed microbial features (y-axis) in the Definite and No Evidence patients.

Supplementary Data

Supp. Data 1. Basic sample metadata.

Supp. Data 2. DE analyses between: i) Definite and No Evidence patients; ii) Definite patients with any bacterial LRTI and with purely viral LRTI.

Supp. Data 3. GSEA results from the DE analyses between: i) Definite and No Evidence patients; ii) Definite patients with any bacterial LRTI and with purely viral LRTI.

Supp. Data 4. Pathogens identified in Definite patients by clinical testing and by mNGS.