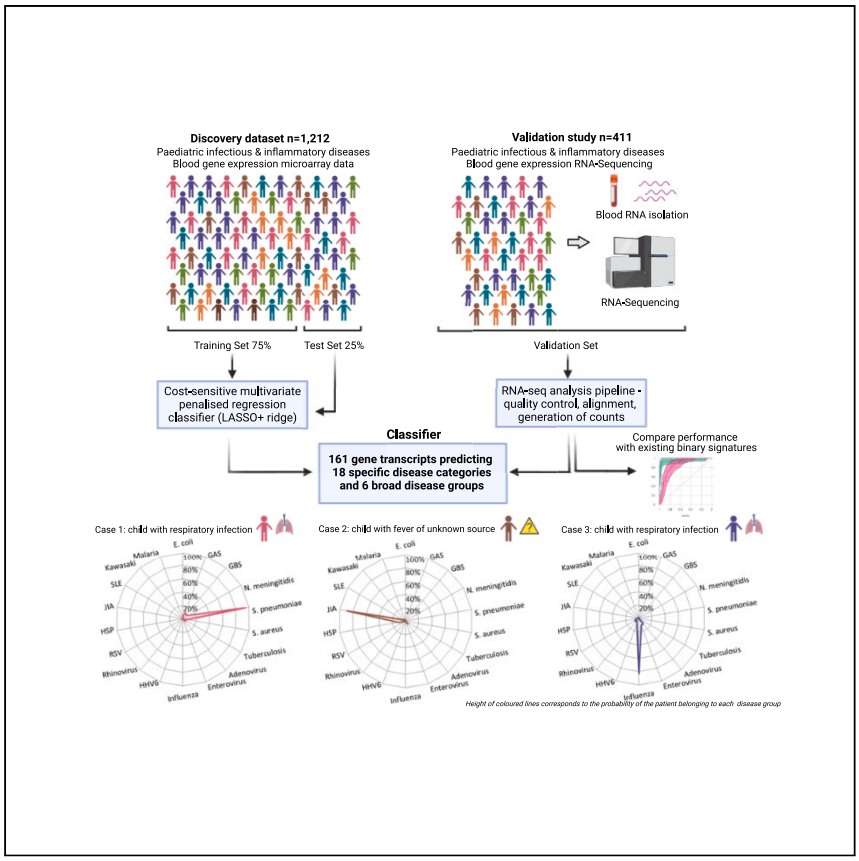
Med

CellPress
OPEN ACCESS

Clinical and Translational Article

Diagnosis of childhood febrile illness using a multi-class blood RNA molecular signature



A multi-class supervised machine-learning approach applied to whole-blood transcriptomics can classify 18 categories of pediatric infectious and inflammatory diseases, reflecting individual causative pathogen or specific disease. Habgood-Coote et al. identify a panel of 161 RNA transcripts in blood using gene expression microarrays, validate using RNA-sequencing, and benchmark against existing dichotomous RNA signatures.

Dominic Habgood-Coote, Clare Wilson, Chisato Shimizu, ..., Jethro A. Herberg, Michael Levin, Myrsini Kaforou

m.kaforou@imperial.ac.uk

Highlights

Multi-class supervised machine learning for infectious and inflammatory diseases

Host-blood RNA expression can discriminate multiple pediatric diseases simultaneously

Eighteen specific diseases or causative pathogens are distinguished using 161 transcripts

Translation to Patients

Habgood-Coote et al., Med 4, 635–654
September 8, 2023 © 2023 The Author(s).
Published by Elsevier Inc.
https://doi.org/10.1016/j.medj.2023.06.007

Med

CellPress
OPEN ACCESS

Clinical and Translational Article

Diagnosis of childhood febrile illness using a multi-class blood RNA molecular signature

Dominic Habgood-Coote,[1] Clare Wilson,[1] Chisato Shimizu,[2] Anouk M. Barendregt,[3] Ria Philipsen,[4] Rachel Galassini,[1] Irene Rivero Calle,[5,6] Lesley Workman,[7] Philipp K.A. Agyeman,[8] Gerben Ferwerda,[4] Suzanne T. Anderson,[9] J. Merlijn van den Berg,[3] Marieke Emonts,[10,11] Enitan D. Carrol,[12] Colin G. Fink,[13,14] Ronald de Groot,[4] Martin L. Hibberd,[15] John Kanegaye,[2] Mark P. Nicol,[25] Stéphane Paulus,[12,16] Andrew J. Pollard,[16] Antonio Salas,[5,6,17] Fatou Secka,[9] Luregn J. Schlapbach,[18,19] Adriana H. Tremoulet,[2] Michael Walther,[9] Werner Zenz,[20] the Pediatric Emergency Medicine Kawasaki Disease Research Group (PEMKDRG), the UK Kawasaki Genetics consortium, the GENDRES consortium, the EUCLIDS consortium, the PERFORM consortium, Michiel Van der Flier,[21,22] Heather J. Zar,[7] Taco Kuijpers,[3,23] Jane C. Burns,[2] Federico Martinón-Torres,[5,6] Victoria J. Wright,[1] Lachlan J.M. Coin,[24] Aubrey J. Cunnington,[1] Jethro A. Herberg,[1] Michael Levin,[1,26] and Myrsini Kaforou[1,26,27,*]

SUMMARY

Background: Appropriate treatment and management of children presenting with fever depend on accurate and timely diagnosis, but current diagnostic tests lack sensitivity and specificity and are frequently too slow to inform initial treatment. As an alternative to pathogen detection, host gene expression signatures in blood have shown promise in discriminating several infectious and inflammatory diseases in a dichotomous manner. However, differential diagnosis requires simultaneous consideration of multiple diseases. Here, we show that diverse infectious and inflammatory diseases can be discriminated by the expression levels of a single panel of genes in blood.

Methods: A multi-class supervised machine-learning approach, incorporating clinical consequence of misdiagnosis as a "cost" weighting, was applied to a whole-blood transcriptomic microarray dataset, incorporating 12 publicly available datasets, including 1,212 children with 18 infectious or inflammatory diseases. The transcriptional panel identified was further validated in a new RNA sequencing dataset comprising 411 febrile children.

Findings: We identified 161 transcripts that classified patients into 18 disease categories, reflecting individual causative pathogen and specific disease, as well as reliable prediction of broad classes comprising bacterial infection, viral infection, malaria, tuberculosis, or inflammatory disease. The transcriptional panel was validated in an independent cohort and benchmarked against existing dichotomous RNA signatures.

Conclusions: Our data suggest that classification of febrile illness can be achieved with a single blood sample and opens the way for a new approach for clinical diagnosis.

Funding: European Union's Seventh Framework no. 279185; Horizon2020 no. 668303 PERFORM; Wellcome Trust (206508/Z/17/Z); Medical Research Foundation (MRF-160-0008-ELP-KAFO-C0801); NIHR Imperial BRC.

CONTEXT AND SIGNIFICANCE

Infectious and inflammatory diseases are the most common causes of children seeking medical care in both hospital and community settings. It is a considerable challenge for clinical teams to reliably distinguish common viral infections, bacterial infections (which are potentially serious), and less common inflammatory diseases, with existing tests when children initially present at healthcare settings. Habgood-Coote et al. describe an approach for simultaneously distinguishing between 18 infectious and inflammatory diseases using the differences in the levels of expression of 161 genes in patients' blood. A future diagnostic test based on this approach could help provide the right treatment, to the right patient, at the right time, while optimizing antibiotic use and reducing lengthy time to diagnosis for inflammatory diseases.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

INTRODUCTION

Infectious and inflammatory diseases are the most common causes of children seeking medical care in both hospital and community settings.[1] It is a considerable challenge for clinical teams to identify and appropriately treat the small proportion of patients who have severe bacterial infection[2,3] or inflammatory conditions while avoiding over-treating the majority of patients who have self-limiting, usually viral, illness.

Conventional diagnostic tests cannot distinguish the multitude of potential etiologies with sufficient speed and accuracy to inform initial treatment.[4] Culture-based microbiological diagnosis is slow, and while molecular diagnostic techniques are faster, they are limited by the pathogens included in the panel and positive results may identify pathogens that are not the cause of the current illness, particularly for respiratory samples.[5] Infection can involve either a single causative pathogen or the interaction of multiple organisms, limiting the utility of viral pathogen detection.[6] Infections are frequently localized in inaccessible sites (such as the lungs) and, consequently, pathogen detection from accessible sites such as blood, urine, or cerebrospinal fluid is frequently negative, even when severe infections are present.

For most inflammatory disorders, there is currently no single test to confirm or refute the diagnosis, and therefore patients with conditions such as Kawasaki disease (KD) or juvenile idiopathic arthritis are often not diagnosed until after a long period of hospitalization, treatment for presumed infection, and numerous investigations.[7–9] As a result of the limitations of existing diagnostics, definitive final diagnoses are made for less than 50% of children attending an emergency department with fever, and in only half of children admitted to pediatric intensive care with suspected infection.[10,11] Given this diagnostic uncertainty, many patients without bacterial infection are unnecessarily treated with broad-spectrum antibiotics to mitigate the risks of missing severe bacterial infection, contributing to the growing problem of antimicrobial drug resistance.[12]

Gene expression microarrays and, more recently, RNA sequencing (RNA-seq) have revealed an alternative approach, in which infectious or inflammatory diseases are characterized by unique patterns of host gene expression in patients' blood, thus bypassing the need for direct pathogen detection. There is a growing literature documenting that specific infectious and inflammatory diseases can be distinguished from conditions with similar presenting features using sparse transcriptional signatures in whole blood, including discriminating between bacterial and viral infections,[13–19] malaria,[20,21] dengue virus,[22] respiratory syncytial virus,[23,24] rotavirus,[25,26] and tuberculosis (TB)[27,28–30] and diagnosing inflammatory conditions such as KD[31] and systemic lupus erythematosus (SLE).[32]

Previous studies using gene expression for diagnosis have focused on simplified binary distinctions, either one versus one (e.g., bacterial versus viral) or one versus all (e.g., tuberculosis versus other diseases). However, in clinical practice, there is a hierarchy of diagnostic categories, and many potential etiologies must be considered and prioritized according to the risks posed by each. We hypothesized that multiple infectious and inflammatory diseases could be simultaneously discriminated by a limited number of gene transcripts measured in patients' blood.

To investigate this hypothesis, we applied a multi-class feature selection and classification approach based on least absolute shrinkage and selection operator (LASSO) and Ridge regression to genome-wide RNA expression data from 1,212 febrile

[1]Section of Paediatric Infectious Disease and Centre for Paediatrics & Child Health, Department of Infectious Disease, Imperial College London, London, UK

[2]Department of Pediatrics, Rady Children's Hospital San Diego/University of California San Diego School of Medicine, La Jolla, CA, USA

[3]Department of Pediatric Immunology, Rheumatology and Infectious Diseases, Emma Children's Hospital, Amsterdam University Medical Center (AUMC), University of Amsterdam, Amsterdam, the Netherlands

[4]Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Department of Laboratory Medicine, Nijmegen, the Netherlands

[5]Pediatrics Department, Translational Pediatrics and Infectious Diseases Section, Santiago de Compostela, Spain

[6]Genetics- Vaccines- Infectious Diseases and Pediatrics Research Group GENVIP, Instituto de Investigación Sanitaria de Santiago (IDIS), Santiago de Compostela, Spain

[7]Department of Paediatrics & Child Health, Red Cross Childrens Hospital and SA-MRC Unit on Child & Adolescent Health, University of Cape Town, Cape Town, South Africa

[8]Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

[9]Medical Research Council Unit, Fajara, The Gambia at the London School of Hygiene and Tropical Medicine, MRCG at LSHTM Fajara, Banjul, The Gambia

[10]Great North Children's Hospital, Department of Paediatric Immunology, Infectious Diseases & Allergy and NIHR Newcastle Biomedical Research Centre, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, UK

[11]Translational and Clinical Research Institute, Newcastle University, Newcastle Upon Tyne, UK

[12]Department of Clinical Infection, Microbiology and Immunology, University of Liverpool Institute of Infection, Veterinary and Ecological Sciences, Liverpool, UK

[13]Micropathology Ltd Research and Diagnosis, Coventry, UK

[14]University of Warwick, Coventry, UK

[15]Department of Infection Biology, Faculty of Infectious and Tropical Disease, London School of Hygiene and Tropical Medicine, London, UK

[16]Oxford Vaccine Group, Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Centre, Oxford, UK

[17]Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), 15706 Galicia, Spain

Continued

CellPress
OPEN ACCESS

children in 12 publicly available gene expression microarray datasets, representing 18 disease categories, incorporating the clinical risks associated with incorrect diagnosis as a "cost weighting." We identified 161 transcripts that predicted the broad disease category (bacterial, viral, inflammatory, malaria, TB, and KD) with high confidence, as well as specific causative pathogens and diseases.

Cross-platform validation of the signature was performed in an independent cohort of patients (n = 411) for whom gene expression was instead measured in whole blood by RNA-seq. Our data provide proof of concept that the pattern of expression of a single set of transcripts in each patient's blood can be used to diagnose a wide range of infectious and inflammatory diseases.

RESULTS

Two datasets measuring whole-blood gene expression from pediatric patients with febrile illness were used to investigate the potential of a multi-class approach to biomarker discovery. A dataset composed of 12 publicly available gene expression microarray datasets (n = 1,212) was used for the discovery of a biomarker panel, which was then applied to a newly generated RNA-seq dataset (n = 411).

The discovery dataset

To explore the feasibility of using a limited number of RNA transcripts to classify febrile illness, we merged and analyzed publicly available microarray datasets. A comprehensive literature search, limited to Illumina Beadchip arrays, identified 12 datasets (GEO: GSE73464, GSE68004, GSE65391, GSE64456, GSE42026, GSE40396, GSE39941, GSE38900, GSE34404, GSE30119, GSE29366, GSE22098) that measured gene expression in whole-blood samples from both pediatric patients with acute febrile illnesses and appropriate controls (Table S1). The control samples in each dataset were used to perform batch correction with the COmbat CO-Normalization Using conTrols (COCONUT) method[19](Figure S1).

Patients with multiple potentially causative pathogens and disease groups with fewer than 10 cases were excluded, leaving 1,212 patients across 18 disease classes. Of these patients, 338 had bacterial infections caused by *Staphylococcus aureus* (n = 107), *Streptococcus pneumoniae* (n = 15), group A *Streptococcus* (GAS) (n = 38), group B *Streptococcus* (GBS) (n = 10), *Neisseria meningitidis* (n = 10), *Escherichia coli* (n = 58), or *Mycobacterium tuberculosis* (n = 100). There were 290 cases due to viral infections, including respiratory syncytial virus (RSV) (n = 61), rhinovirus (n = 12), human herpesvirus 6 (HHV6) (n = 10), influenza virus (n = 98), enterovirus (n = 57), and adenovirus (n = 52). The 487 cases of inflammatory disease included SLE (n = 204), juvenile idiopathic arthritis (JIA) (n = 98), Henoch-Schönlein purpura (HSP) (n = 18), and KD (n = 167). Malaria (n = 97) was the only parasitic infection present in the datasets. The merged and batch-corrected data were randomly split into subsets comprising 75% and 25% for training and testing respectively using stratified holdout to maintain class proportions.

Identification of a multi-class signature of febrile illness

In the discovery set, repeated cross-validation was performed in order to select the best method for deriving a multi-class signature of febrile illness (see STAR Methods). Of the five multivariate penalized regression methods compared, LASSO + Ridge derived the smallest models with good classification performance while allowing cost-sensitivity (Figure S2A). This is a two-stage method in which LASSO regression is used to perform feature selection followed by a Ridge regression to refit the coefficients and improve predictive performance.[33]

[18]Pediatric and Neonatal Intensive Care Unit, and Children's Research Center, University Children's Hospital Zurich, Zurich, Switzerland

[19]Child Health Research Centre, The University of Queensland, and Paediatric Intensive Care Unit, Queensland Children's Hospital, Brisbane, QLD, Australia

[20]University Clinic of Paediatrics and Adolescent Medicine, Department of General Paediatrics, Medical University of Graz, Graz, Austria

[21]Paediatric Infectious Diseases and Immunology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, the Netherlands

[22]Paediatric Infectious Diseases and Immunology Amalia Children's Hospital, Radboudumc, Nijmegen, the Netherlands

[23]Department of Blood Cell Research, Sanquin Blood Supply, Division Research and Landsteiner Laboratory of Amsterdam UMC (AUMC), University of Amsterdam, Amsterdam, the Netherlands

[24]Department of Microbiology and Immunology, University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

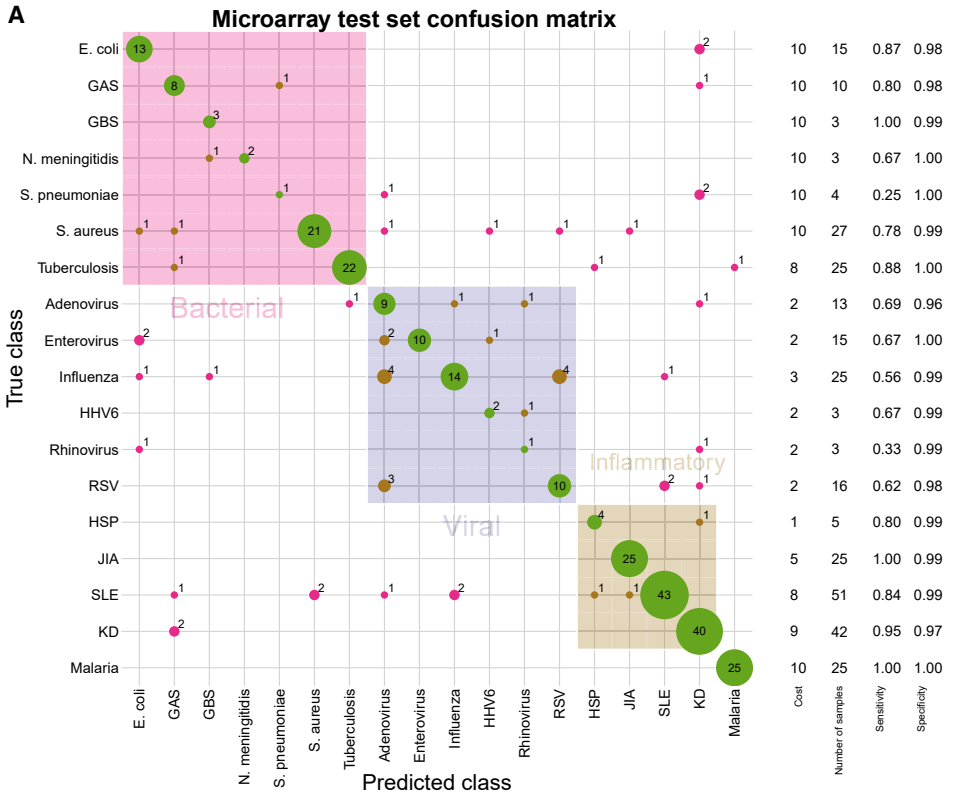
[25]Marshall Centre, School of Biomedical Sciences, University of Western Australia, Perth, Australia

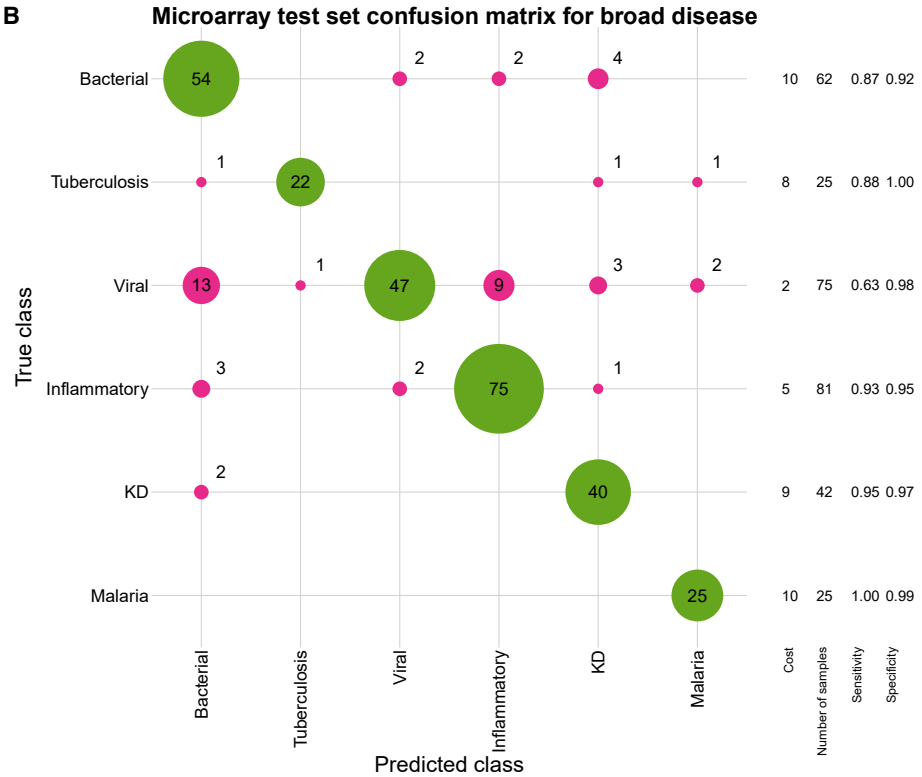
[26]These authors contributed equally

[27]Lead contact

*Correspondence: m.kaforou@imperial.ac.uk

https://doi.org/10.1016/j.medj.2023.06.007


A
Microarray test set confusion matrix
True class
Predicted class
Bacterial
Viral
Inflammatory
E. coli
GAS
GBS
N. meningitidis
S. pneumoniae
S. aureus
Tuberculosis
Adenovirus
Enterovirus
Influenza
HHV6
Rhinovirus
RSV
HSP
JIA
SLE
KD
Malaria
Cost
Number of samples
Sensitivity
Specificity


B
Microarray test set confusion matrix for broad disease
True class
Predicted class
Bacterial
Tuberculosis
Viral
Inflammatory
KD
Malaria
Cost
Number of samples
Sensitivity
Specificity

CellPress
OPEN ACCESS

Figure 1. Confusion matrix for the gene expression microarray test set predictions

(A and B) Performance of the 161-transcript signature in the 25% microarray test set over 18 specific disease classes (A) and over six broad disease classes (B). Confusion matrices show the numbers of each type of misclassification made where each sample is predicted to belong to the class with highest probability. Green corresponds to correct predictions (true positives), brown corresponds to incorrect predictions within the same broad disease category, and pink shows incorrect predictions for disease category (false positives/negatives). Cost-weighting, point estimates for sensitivity, and specificity for each prediction are shown on the right. E. coli, *Escherichia coli*; GAS, group A *Streptococcus*; GBS, group B *Streptococcus*; N. meningitidis, *Neisseria meningitidis*; S. pneumoniae, *Streptococcus pneumoniae*; S. aureus, *Staphylococcus aureus*; HHV6; human herpesvirus 6; RSV, respiratory syncytial virus; HSP, Henoch-Schönlein purpura; JIA, juvenile idiopathic arthritis; SLE, systemic lupus erythematosus; KD, Kawasaki disease.

An important consideration in the context of clinical diagnostics is the potential consequence of incorrect diagnosis. In order to incorporate the clinical consequences of misdiagnosis, we applied a "cost-sensitive learning" approach by performing example weighting. Class weights were assigned by the consensus judgment of five independent pediatric infectious disease specialists to reflect the risks posed by each disease if untreated, the speed of disease progression, and the availability of effective treatment. Weights were divided by the abundance of each class to offset the effect of class imbalance (Table S2). The effect of incorporating these weights in the training process is to bias the feature set and coefficients to reduce the false-negative error for high-risk groups at the expense of increasing the false-negative error of low-risk groups (Figure S2B).

We applied multinomial LASSO + Ridge penalized regression in the 75% discovery set to identify an RNA transcript panel consisting of 161 probes for the discrimination of 18 disease classes. This set of probes was selected from the LASSO regularization path at a value of lambda at which the cross-validated mean square error (weighted by cost and class imbalance) for the Ridge regression was within two standard errors of the minimum. A heatmap of the standardized expression of the selected probes is shown in Figure S3.

Test set predictions

Predicted probabilities for the 25% test set were used to derive a confusion matrix for which discrete class predictions for each sample were made by taking the class with highest predicted probability (Figure 1A). The ability of the classifier to separate disease groups was also assessed for pairwise (one-versus-one) and one-versus-all discrimination on the basis of predicted probabilities (Figure S4 and Table S3). While the model was able to reliably predict most diagnostic classes, the predictive performance was lower for groups with smaller number of samples in the training dataset (Figure S2C). While many of the samples could be assigned to specific disease classes, many of the misclassifications occur between classes from the same broad disease category (bacterial or viral) (Figures 1B and S5).

Broad clinical categories with immediate clinical implications

While the rapid identification of causative pathogens would be useful for optimal treatment and choice of antibiotics, clinical teams require a high degree of confidence in the broad disease category (i.e., viral, bacterial, or inflammatory) to ensure potentially life-threatening conditions are not missed and to direct empiric treatment and appropriate subsequent investigations.

We investigated whether the biomarker panel could also be used to make confident predictions of broad disease category. Refitting the coefficients for the 161 transcripts using multinomial Ridge regression allowed the panel to predict the broad disease categories: inflammatory disease, viral infection, bacterial infection, KD, malaria, and tuberculosis (Figures S2D and Table S3). Although tuberculosis is a bacterial disease, it was considered as a separate class, as it requires very different

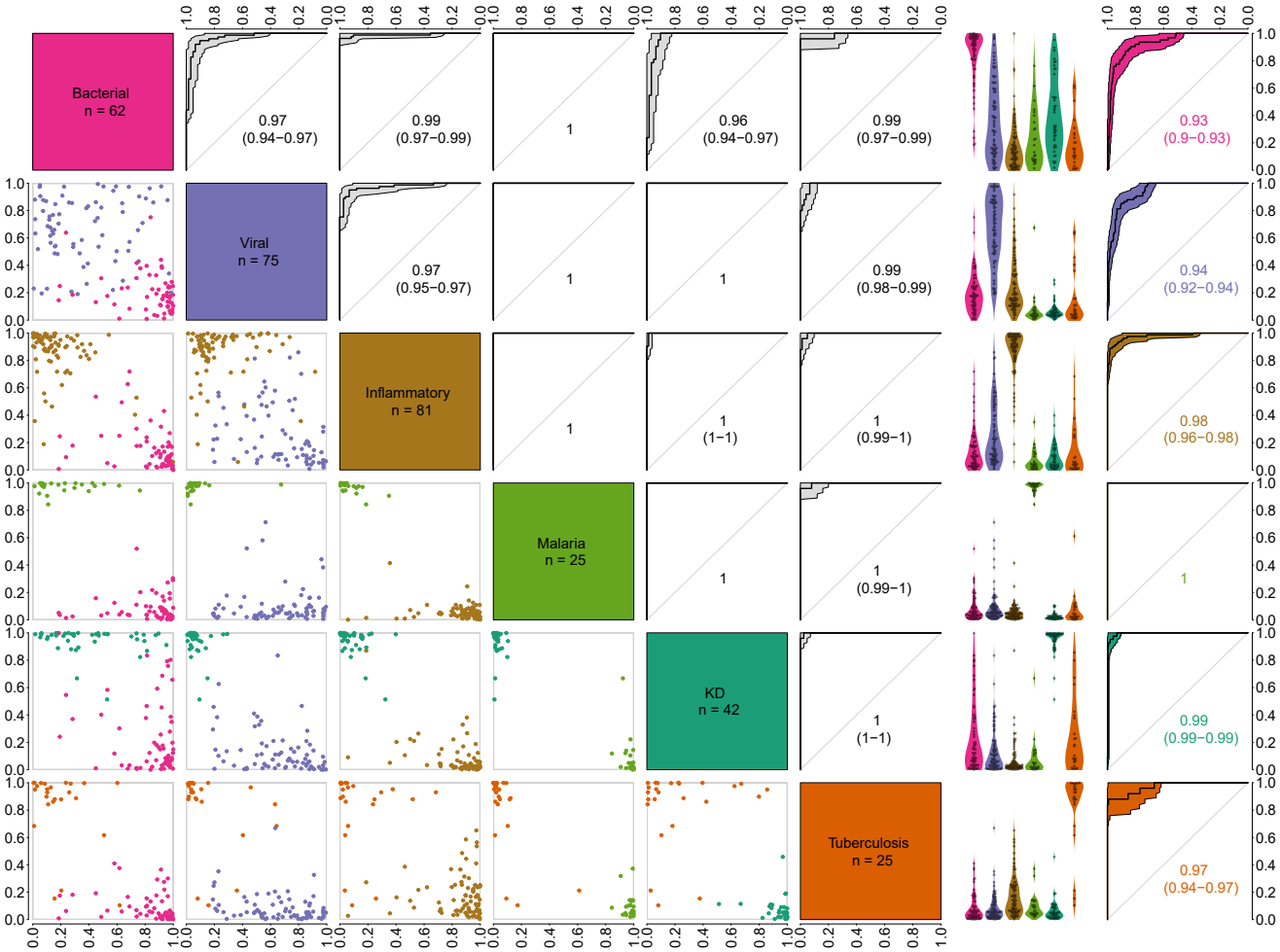


Figure 2. Microarray test set predictions of broad disease categories

Pairwise and one-versus-all discrimination of broad disease categories. Scatterplots and ROC curves are shown for pairs of disease categories (columns 1–6). Each scatterplot shows the predicted probabilities for patients with one of a pair condition; conditions for each scatterplot are given on the diagonal, above (x axis) and to the right (y axis) of each plot. ROC curves show the performance when distinguishing each pair of conditions; conditions for each ROC plot are given on the diagonal, below and to the left of each plot. Separation of the pair of diseases is performed using the predicted probabilities of both classes where the decision threshold is defined by varying the gradient of a line p(class1) = m × p(class2). The rightmost panels (columns 7–8) show the predicted probabilities for each class (left) and the one-versus-all ROC curve defined using only these probabilities to distinguish the class in a one-versus-all comparison. 95% confidence intervals are shown for all ROC curves except where they could not be calculated due to lack of overlap.

clinical management from the other bacterial infections, and also induces distinct transcriptional responses (Figure S2G). Similarly, KD, which also induced distinct transcriptional responses, was considered as a distinct class. Although epidemiological features suggest an infectious agent as the cause of KD, its etiology remains unknown and treatment is directed at immunomodulation.[34,35] The resulting model accurately predicted the presence of these six disease classes both when considering the most likely class for each patient (Figure 1B) and when considering classes independently (Figure 2 and Table S3). These predictions allow the model to reflect the diagnostic classification used in clinical decision making and simultaneously address multiple clinical questions. The clinical teams can be provided with the probabilities for each patient to belong in each class as an optimal input for decision making.

CellPress
OPEN ACCESS

Table 1. RNA-seq set demographics

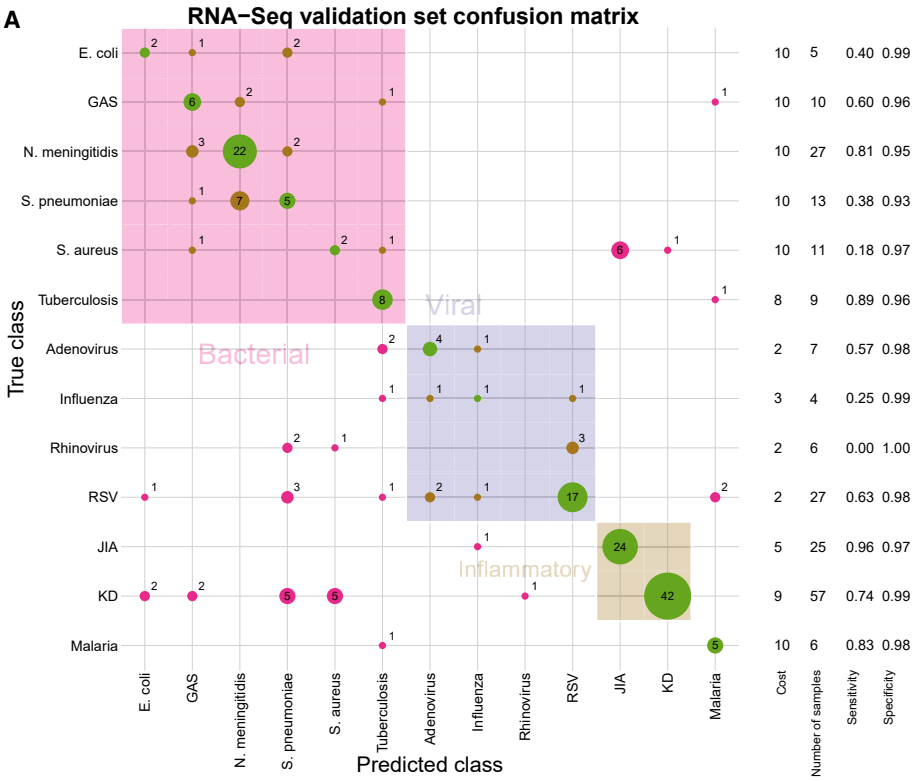
Characteristic		Bacterial	Viral	Inflammatory	TB	Malaria	KD
Number of patients		130	88	50	18	12	113
Age: months, median (IQR)		30 (9–65)	7 (2–20)	171 (132–200)	79 (43–93)	70 (51–93)	35 (18–56)
Male sex: no. (%)		72 (55)	58 (66)	11 (22)	10 (56)	6 (50)	68 (60)
Population group	African	25 (19.2)	5 (5.7)	0	1 (5.6)	12 (100)	7 (6.2)
	Asian	5 (3.9)	2 (2.3)	1 (2.0)	0	0	16 (14.2)
	European	85 (65.4)	62 (70.5)	49 (98.0)	0	0	27 (23.9)
	Latin American	1 (0.8)	8 (9.1)	0	1 (5.6)	0	37 (32.7)
	Mixed/other/ unknown	14 (10.8)	11 (12.5)	0	16 (88.9)	0	26 (23.0)
Days from symptoms: median (IQR)		2 (1–4)	5 (2–7)	264 (158–765)/ 877 (364–2,095)[a]	14 (7–30)	3 (2–3)	6 (5–7)
Intensive care: no. (%)		69 (53.1)	17 (19.3)	0	0	0	2 (1.8)
Deaths: no.		10	1	0	0	0	0
CRP (mg/L): median (IQR)		203 (111–281)	6 (3–18)	10 (3–44)	60 (51–69)	NA	72 (42–162)
Blood cell differential							
Neutrophil %: median (IQR)		75.0 (59.9–85.7)	29.0 (16.8–48.7)	51.3 (42.8–59.4)	NA	65.8 (56.6–74.4)	61.3 (52.5–76.9)
Lymphocytes %: median (IQR)		17.0 (9.3–27.9)	47.5 (28.5–60.3)	35.4 (29.8–45.0)	29.5 (25.0–35.5)	27.1 (18.8–34.4)	22.5 (11.8–30.2)
Monocyte %: median (IQR)		5.88 (3.0–8.0)	7.7 (5.0–11.1)	8.4 (7.0–10.7)	5.0 (5.0–6.9)	6.8 (4.7–8.5)	5.0 (3.0–8.0)
Clinical syndromes							
Soft tissue		10	0	0	0	0	0
Inflammatory		0	0	50	0	0	113
Gastrointestinal		3	1	0	0	0	0
Urinary tract infection		8	0	0	0	0	0
Upper respiratory/ear, nose, throat		3	25	0	0	0	0
Lower respiratory tract		17	62	0	17	0	0
Central nervous system involvement		38	0	0	0	0	1
Musculoskeletal		7	0	0	0	0	0
Other[c]		6	1	0	9	0	0
Pathogen specific[b]		3	0	0	0	12	0
Sepsis		76	0	0	0	0	0

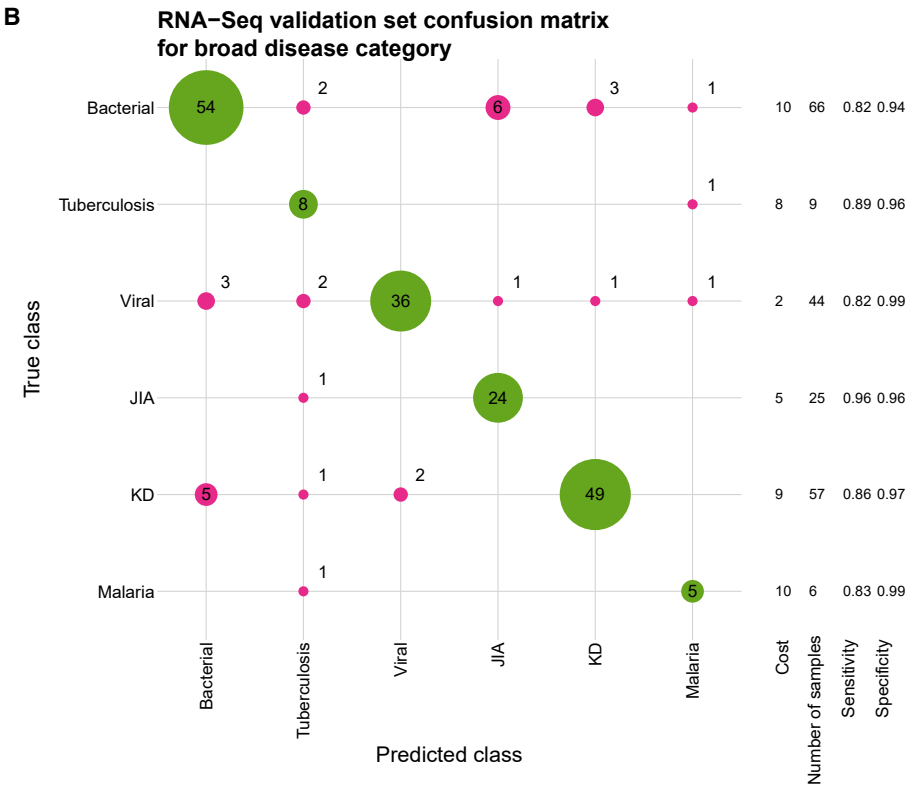
IQR, interquartile range; CRP, C-reactive protein; Ethnicity, self-reported ethnicity; TB, tuberculosis; KD, Kawasaki disease.
[a]Relative to initial symptom onset for the first episode or exacerbations respectively.
[b]Including scarlet fever, staphylococcal scalded skin, and malaria.
[c]Including central-line-associated bloodstream infection, endocarditis, extra-pulmonary TB, facial palsy, pericarditis, and status epilepticus. Patients could be affected by more than one syndrome at the same time.

Validation in an independent study using RNA-seq

We evaluated the performance of the diagnostic signature in an independent patient cohort and using a different RNA quantification platform. We used a newly generated dataset of whole-blood RNA-seq including 411 pediatric patients with a range of infectious or inflammatory diseases, covering all six broad diagnostic classes and 13 of the 18 specific diagnostic classes used in the discovery dataset (demographic and clinical details in Table 1 and study details in STAR Methods).

The 161 microarray probes were mapped uniquely to 155 genes, of which 10 did not have sufficient read counts in the RNA-seq dataset for reliable quantification, leaving 145 genes in the panel in the RNA-seq dataset (Table S4). Gene level read counts


A
RNA−Seq validation set confusion matrix
True class
Predicted class
E. coli
GAS
N. meningitidis
S. pneumoniae
S. aureus
Tuberculosis
Adenovirus
Influenza
Rhinovirus
RSV
JIA
KD
Malaria
Bacterial
Viral
Inflammatory
Cost
Number of samples
Sensitivity
Specificity


B
RNA−Seq validation set confusion matrix
for broad disease category
True class
Predicted class
Bacterial
Tuberculosis
Viral
JIA
KD
Malaria
Cost
Number of samples
Sensitivity
Specificity

CellPress
OPEN ACCESS

Figure 3. Performance of the 145-transcript panel in the validation cohort

(A and B) Multi-class confusion matrices for disease prediction in the RNA-seq validation set for specific disease classes (A) and for broad disease classes (B). Circle area and color correspond to number of patients and type of misclassification, respectively, where green is correct classification, pink is incorrect classification in a different broad disease group, and brown is incorrect classifications within the same broad disease class. Specificities and sensitivities for the detection of each class were derived from discrete class predictions. Cost weighting and point estimates for sensitivity and specificity for each prediction are shown on the right.

were normalized for sequencing depth with scaling factors calculated with DESeq2[36] followed by a log transformation. To account for the different quantification platform and smaller signature, the coefficients of the multi-class models for classifying both broad and specific disease class were refitted on a random selection of 50% of the dataset using multinomial Ridge regression with class weighting (Table S2). The performance in the remaining 50% is shown for discrete class predictions (Figure 3) using predicted probabilities for pairwise and one-versus-all comparisons (Figures 4, S2E, S2F, and S5 and Table S3) and for individual patients.

In addition, and although microarray and RNA-seq rely on very different quantification approaches, we assessed the performance of the broad disease classifier in the RNA-seq dataset without retraining the coefficients in the RNA-seq dataset. The coefficients were refitted using Ridge regression in the complete microarray dataset. This model was then used to make predictions on the RNA-seq dataset after applying limma voom transformation[37] to the DESeq2 depth normalized RNA-seq count data (Figure S6).

The utility of a diagnostic test is highly dependent on the prevalence of disease in the population on which it is being used; however, since a multi-class diagnostic panel could be applied in different clinical contexts, values for specificity, positive predictive value, and negative predicted value are shown for four illustrative scenarios of disease prevalence in different populations (Table S5).[38–41]

We performed differential expression analyses between each broad disease category and the other disease groups using DESeq2[36] and the enrichment analysis for the different comparisons using g:Profiler[42] of Gene Ontology (biological pathways) and Reactome terms (Figure S7).

Benchmarking with previously published one-versus-all signatures

There are no previously reported transcriptional panels that can simultaneously distinguish multiple causes of fever in children against which to benchmark performance. We therefore compared the performance of our multi-class biomarker panel to four previously reported binary classification signatures for the classification of pediatric febrile illness: tuberculosis,[27,30] KD,[31] and for distinguishing bacterial from viral infections[14](Table S6). Since some of the microarray datasets were used for binary signature derivation, performance was only compared in the RNA-seq dataset for comparison fairness. The coefficients of each linear model were refitted using the same 50% of the RNA-seq dataset and performance was evaluated using receiver operating characteristic (ROC) curves on the remaining 50%. There was no significant difference in the area under the ROC curve (AUC) between the Wright signature[31] and the Kawasaki component of the multi-class signature ($p = 0.2$, bootstrap test, Benjamini Hochberg corrected). The multi-class biomarker panel performed better in terms of AUC than the single-class Sweeney signature for tuberculosis[30] ($p = 0.03$) but there was not a significant difference to the AUC of the Anderson signature[27] ($p = 0.08$) (Figure 5 and Table S7). The improvement relative to the bacterial-viral signature was significant for the identification of viral infection (viral versus all; $p = 0.007$) and bacterial infection (bacterial versus all; $p = 0.007$)

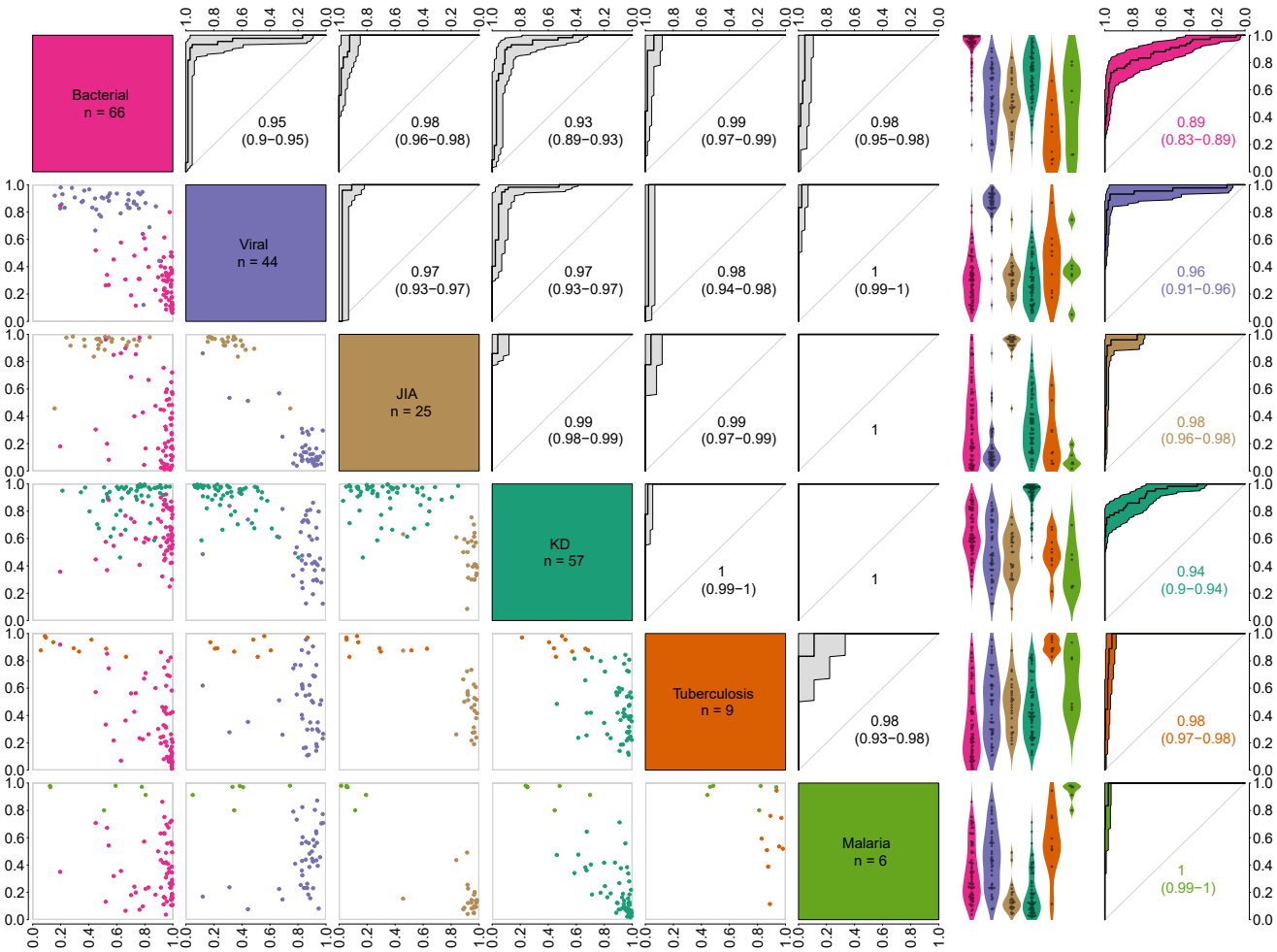


Figure 4. RNA-seq validation set predictions of broad disease classes

Pairwise and one-versus-all discrimination of broad disease categories. Scatterplots and ROC curves are shown for pairs of disease categories (columns 1–6). Each scatterplot shows the predicted probabilities for patients with one of a pair condition; conditions for each scatterplot are given on the diagonal, above (x axis) and to the right (y axis) of each plot. ROC curves show the performance when distinguishing each pair of conditions; conditions for each ROC plot are given on the diagonal, below and to the left of each plot. Separation of the pair of diseases is performed using the predicted probabilities of both classes where the decision threshold is defined by varying the gradient of a line p(class1) = m × p(class2). The rightmost panels (columns 7–8) show the predicted probabilities for each class (left) and the one-versus-all ROC curve defined using only these probabilities to distinguish the class in a one-versus-all comparison. 95% confidence intervals are shown for all ROC curves except where they could not be calculated due to lack of overlap.

but not for distinguishing bacterial from viral infection (bacterial versus viral; p = 0.3), reflecting that the inclusion of the additional disease groups on this study has only a minor impact on overall performance measured by AUC for the direct bacterial-viral comparison, but the use of a cost-sensitive approach improves the sensitivity to bacterial infection for the lower values of specificity.

DISCUSSION

We investigated whether multiple diseases could be distinguished simultaneously using a single whole-blood transcriptional panel. A multi-class machine-learning approach was applied to publicly available blood gene expression datasets to identify a set of 161 transcripts sufficient for accurate diagnosis of diverse causes of febrile illness in children. The 161-transcript panel can identify 18 specific inflammatory diseases and pathogen species and distinguish between six broad disease

CellPress
OPEN ACCESS

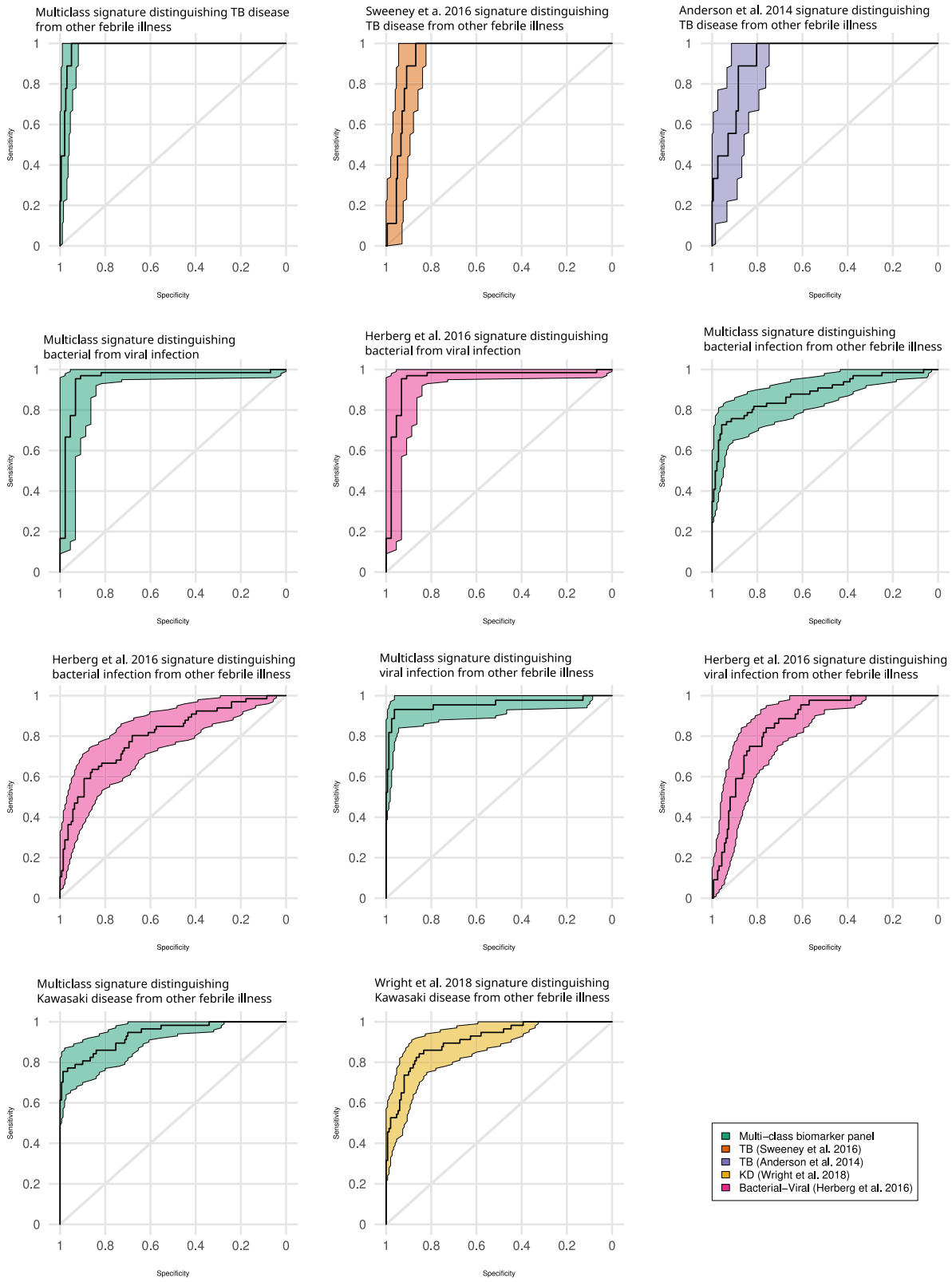


Figure 5. Comparison of the multi-class RNA signature to previously published signatures of infectious disease

(A–E) ROC curves and 95% confidence intervals of specificity are shown for the multi-class signature and previously reported signatures for tuberculosis (A), KD (B), and for distinguishing bacterial and viral infection (C–E). The comparison to the bacterial-viral signature is split by the formulation of the classification problem.

(C) A bacterial versus viral comparison where, for the multi-class classifier, the ratio of predicted probabilities for bacterial and viral infection are used.

(D and E) The problem as a viral versus all and bacterial versus all respectively with each using the corresponding component of the multi-class signature.

categories (bacterial infection, viral infection, inflammatory disease, tuberculosis, malaria, and KD). As some diagnostic errors carry severe consequences (such as failure to diagnose a life-threatening bacterial infection), while others have few adverse consequences (such as failing to diagnose a self-limiting viral infection for which there is no specific treatment), we used a cost-sensitive learning approach in our discovery pipeline by example weighting. Although derivation of a full cost matrix based on formally defined outcome measures would be necessary to fully recapitulate the clinical consequences of misclassifications, this is beyond the scope of this study. We instead used a weighting scheme based on expert consensus that could effectively prioritize the predictions in favor of diseases for which misdiagnosis carries the greatest consequence. The 161-transcript signature identified using gene expression microarray datasets was validated in a translated 145-gene form in an independent study of febrile children in whom gene expression levels were detected by RNA-seq, supporting the clinical validity, the robustness, and the reproducibility of the approach.

In order to incorporate a multi-class transcriptomic signature for febrile illness into clinical care, the panel of RNA transcripts needs to be translated into a diagnostic test suitable for use in hospitals or clinics, which would be able to measure the transcripts rapidly and at affordable cost. There is a rapidly expanding number of molecular methods and technologies for rapid, inexpensive, and high-throughput measurement of large numbers of targets, including customized arrays,[43] high-throughput PCR-based methods, NanoString technologies,[44] and electrochemical biosensor technology.[45,46]

The imperative need of a novel diagnostic platform capable of simultaneously identifying multiple pathogens and different analyte types to improve diagnosis and management of febrile patients (MAPDx) was highlighted by Médecins Sans Frontières (MSF), Foundation for Innovative New Diagnostics (FIND), and the World Health Organization (WHO)[47] in a foundation document for a fever-specific assay. A host RNA-based approach that could simultaneously identify different causes of fever would meet several of the target product profile (TPP) characteristics reported: single sample, kinetics of infection, and semi-open design to allow for relevant genes to be measured in different settings and to address different clinical questions. Further research needs to be conducted to optimize the presentation of results and level of detail that would be made available to the clinical team and to determine whether results are presented as most likely cause, probability of each cause, or enhanced with management suggestions taking pre-test probabilities into account.

This study provides a proof of principle that a single panel of RNA transcripts can be used to assign patients with fever and non-specific clinical and laboratory findings to a range of etiologies from a single whole-blood sample. Coupled with diagnostic technological advances able to measure RNA transcripts rapidly and at an affordable cost, a multi-class diagnostic test for febrile illness could circumvent lengthy clinical diagnostic processes and reduce delays to diagnoses, missed diagnoses, and unnecessary antibiotic treatment, having a significant impact on global health.

CellPress
OPEN ACCESS

Limitations of the study

While our study provides proof of principle that disease class assignment for a range of infectious and inflammatory diseases can be achieved using the pattern of gene expression in the blood of each patient, further development of the approach is currently limited by the availability of whole-blood gene expression datasets. Although the use of publicly available gene expression datasets allows more heterogeneity to be captured, particularly where a single disease is considered across multiple datasets,[30] the scope of the approach can be limited by the representativeness and completeness of datasets used for discovery, imposed by the research focus of prior studies. Further optimization of the transcript panel will require a large comprehensive blood gene expression dataset where a wide range of illnesses is considered simultaneously to ensure signatures are robust to the full range of potential etiologies.

In our study, class imbalance was taken into account; however, some of the pathogens and diseases for which a limited number of samples were available were not accurately identified by the transcript panel (e.g., *E. coli* in the RNA-seq data), as were pathogens whose clinical relevance is less well defined (such as viral pathogen detection in respiratory samples).[48] While a more stringent filtering of disease groups of small sample size would improve predictive performance, this would mean the omission of clinically relevant diseases from this proof-of-concept study. For some of these underrepresented disease groups in the discovery set, good performance could be achieved, such as *N. meningitidis*, which could be reliably detected in both discovery and validation datasets. Additionally, we have excluded patients with more than one potential cause of fever, but further work needs to be undertaken to benchmark the predictions for cases with more than one clinical diagnosis. While the discrimination of healthy control samples might be an advantage in a screening context, the populations targeted by a diagnostic test of this kind will consist of febrile patients with sufficient clinical concern to warrant a blood test.

Although we successfully performed cross-platform and cross-cohort validation moving from a microarray discovery cohort to an RNA-seq validation cohort, some of the originally discovered 161-microarray transcript set were not present in the RNA sequence data. Additionally, there were insufficient numbers of patients in the RNA-seq data to include five of the disease groups present in the microarray data (enterovirus, GBS, HHV6, HSP, and SLE). To ensure clinical utility, further development of the approach will require large prospective patient cohorts, with consistent, detailed, and accurate clinical phenotypes. By expanding the range of conditions included in the discovery of the transcript panels, it may be possible to improve the treatment of a large number of patients, particularly for rare and under-diagnosed conditions for which early detection and thus treatment could have a significant benefit. Similarly, given appropriate clinical cohorts and gene expression datasets, it may be possible to expand this principle to other populations such as adults, patients with co-morbidities, and populations affected by pathogens specific to certain geographic areas, such dengue, arbovirus infections, or zoonotic illnesses such as Lyme disease and typhus, which pose considerable diagnostic challenges.

CONSORTIA

The members of EUCLIDS (www.euclids-project.eu) are Michael Levin, Lachlan Coin, Stuart Gormley, Shea Hamilton, Jethro Herberg, Bernardo Hourmat, Clive Hoggart, Myrsini Kaforou, Vanessa Sancho-Shimizu, Victoria Wright, Amina Abdulla, Paul Agapow, Maeve Bartlett, Evangelos Bellos, Hariklia Eleftherohorinou, Rachel Galassini, David Inwald, Meg Mashbat, Stephanie Menikou, Sobia Mustafa, Simon Nadel,

Rahmeen Rahman, Hannah Shailes, Clare Thakker, S. Bokhandi, Sue Power, Heather Barham, N. Pathan, Jenna Ridout, Deborah White, Sarah Thurston, Saul Faust, Sanjay Patel, Jenni McCorkell, P. Davies, Lindsey Crate, Helen Navarra, Stephanie Carter, R. Ramaiah, Rekha Patel, Catherine Tuffrey, Andrew Gribbin, Sharon McCready, Mark Peters, Katie Hardy, Fran Standing, Lauren O'Neill, Eugenia Abelake, Akash Deep, Eniola Nsirim, Andrew Pollard, Louise Willis, Zoe Young, C. Royad, Sonia White, P.M. Fortune, Phil Hudnott, Federico Martinón-Torres, Antonio Salas, Fernando Álvez González, Ruth Barral-Arca, Miriam Cebey-López, María José Currás-Tuala, Natalia García, Luisa García Vicente, Alberto Gómez-Carballa, Jose Gómez Rial, Andrea Grela Beiroa, Antonio Justicia Grande, Pilar Leboráns Iglesias, Alba Elena Martínez Santos, Federico Martinón -Torres, Nazareth Martinón-Torres, José María Martinón Sánchez, Beatriz Morillo Gutiérrez, Belén Mosquera Pérez, Pablo Obando Pacheco, Jacobo Pardo-Seco, Sara Pischedda, Irene Rivero-Calle, Carmen Rodríguez-Tenreiro, Lorenzo Redondo-Collazo, Sonia Serén Fernández, María del Sol Porto Silva, Ana Vega, Lucía Vilanova Trillo, Susana Beatriz Reyes, María Cruz León León, Álvaro Navarro Mingorance, Xavier Gabaldó Barrios, Eider Oñate Vergara, Andrés Concha Torre, Ana Vivanco, Reyes Fernández, Francisco Giménez Sánchez, Miguel Sánchez Forte, Pablo Rojo, J.Ruiz Contreras, Alba Palacios, Cristina Epalza Ibarrondo, Elizabeth Fernández Cooke, Marisa Navarro, Cristina Álvarez Álvarez, María José Lozano, Eduardo Carreras, Sonia Brió Sanagustín, Olaf Neth, Ma del Carmen Martínez Padilla, Luis Manuel Prieto Tato, Sara Guillén, Laura Fernández Silveira, David Moreno, Ronald de Groot, A.M. Tutu van Furth, Michiel Van der Flier, N.P. Boeddha, G.J.A. Driessen, J.A. Hazelzet, T.W. Kuijpers, D. Pajkrt, E.A.M. Sanders, D. van de Beek, A. van der Ende, H.L.A. Philipsen, A.O.A. Adeel, M.A. Breukels, D.M.C. Brinkman, C.C.M.M. de Korte, E. de Vries, W.J. de Waal, R. Dekkers, A. Dings-Lammertink, R.A. Doedens, A.E. Donker, M. Dousma, T.E. Faber, G.P.J.M. Gerrits, J.A.M. Gerver, J. Heidema, J. Homan-van der Veen, M.A.M. Jacobs, N.J.G. Jansen, P. Kawczynski, K. Klucovska, M.C.J. Kneyber, Y. Koopman-Keemink, V.J. Langenhorst, J. Leusink, B.F. Loza, I.T. Merth, C.J. Miedema, C. Neeleman, J.G. Noordzij, C.C. Obihara, A.L.T. van Overbeek – van Gils, G.H. Poortman, S.T. Potgieter, J. Potjewijd, P.P.R. Rosias, T. Sprong, G.W. ten Tussher, B.J. Thio, G.A. Tramper-Stranders, M. van Deuren, H. van der Meer, A.J.M. van Kuppevelt, A.M. van Wermeskerken, W.A. Verwijs, T.F.W. Wolfs, Luregn J. Schlapbach, Philipp Agyeman, Christoph Aebi, Christoph Berger, Eric Giannoni, Martin Stocker, Klara M. Posfay-Barbe, Ulrich Heininger, Sara Bernhard-Stirnemann, Anita Niederer-Loher, Christian Kahlert, Paul Hasters, Christa Relly, Walter Baer, Enitan D. Carrol, Stéphane Paulus, M.Hannah Frederick, Rebecca Jennings, Joanne Johnston, Rhian Kenwright, Colin G. Fink, Elli Pinnock, Marieke Emonts, Rachel Agbeko, Suzanne Anderson, Fatou Secka, Kalifa Bojang, Isatou Sarr, Ngange Kebbeh, Gibbi Sey, Momodou Saidykhan, Fatoumata Cole, Gilleh Thomas, Martin Antonio, Werner Zenz, Daniela S. Kohlfürst, Alexander Binder, Nina A. Schweintzger, Manfred Sagmeister, Hinrich Baumgart, Markus Baumgartner, Uta Behrends, Ariane Biebl, Robert Birnbacher, Jan-Gerd Blanke, Carsten Boelke, Kai Breuling, Jürgen Brunner, Maria Buller, Peter Dahlem, Beate Dietrich, Ernst Eber, Johannes Elias, Josef Emhofer, Rosa Etschmaier, Sebastian Farr, Ylenia Girtler, Irina Grigorow, Konrad Heimann, Ulrike Ihm, Zdenek Jaros, Hermann Kalhoff, Wilhelm Kaulfersch, Christoph Kemen, Nina Klocker, Bernhard Köster, Benno Kohlmaier, Eleni Komini, Lydia Kramer, Antje Neubert, Daniel Ortner, Lydia Pescollderungg, Klaus Pfurtscheller, Karl Reiter, Goran Ristic, Siegfried Rödl, Andrea Sellner, Astrid Sonnleitner, Matthias Sperl, Wolfgang Stelzl, Holger Till, Andreas Trobisch, Anne Vierzig, Ulrich Vogel, Christina Weingarten, Stefanie Welke, Andreas Wimmer, Uwe Wintergerst, Daniel Wüller, Andrew Zaunschirm, Ieva Ziuraite, Veslava Žukovskaja, Martin L. Hibberd, Sonia Davila, and Isabel Delany.

CellPress
OPEN ACCESS

The members of PERFORM (www.perform2020.org) are Michael Levin, Aubrey Cunnington, Tisham De, Jethro Herberg, Myrsini Kaforou, Victoria Wright, Lucas Baumard, Evangelos Bellos, Giselle D'Souza, Rachel Galassini, Dominic Habgood-Coote, Shea Hamilton, Clive Hoggart, Sara Hourmat, Heather Jackson, Ian Maconochie, Stephanie Menikou, Naomi Lin, Samuel Nichols, Ruud Nijman, Ivonne Pena Paz, Priyen Shah, Hannah Shailes, Ortensia Vito, Clare Wilson, Amina Abdulla, Ladan Ali, Sarah Darnell, Rikke Jorgensen, Sobia Mustafa, Salina Persand, Molly Stevens, Eunjung Kim, Benjamin Pierce, Katy Fidler, Julia Dudley, Vivien Richmond, Emma Tavliavini, Ching-Fen Shen, Ching-Chuan Liu, Shih-Min Wang, Federico Martinón-Torres, Antonio Salas, Fernando Álvez González, Cristina Balo Farto, Ruth Barral-Arca, María Barreiro Castro, Xabier Bello, Mirian Ben García, Sandra Carnota, Miriam Cebey-López, María José Currás-Tuala, Carlos Durán Suárez, Luisa García Vicente, Alberto Gómez-Carballa, Jose Gómez Rial, Pilar Leboráns Iglesias, Nazareth Martinón-Torres, José María Martinón Sánchez, Belén Mosquera Pérez, Jacobo Pardo-Seco, Lidia Piñeiro Rodríguez, Sara Pischedda, Sara Rey Vázquez, Irene Rivero-Calle, Carmen Rodríguez-Tenreiro, Lorenzo Redondo-Collazo, Miguel Sadiki Ora, Antonio Salas, Sonia Serén Fernández, Cristina Serén Trasorras, Marisol Vilas Iglesias, Dace Zavadska, Anda Balode, Arta Bārzdiņa, Dārta Deksne, Dace Gardovska, Dagne Grāvele, Ilze Grope, Anija Meiere, Ieva Nokalna, Jana Pavāre, Zanda Pučuka, Katrīna Selecka, Aleksandra Sidorova, Dace Svile, Urzula Nora Urbāne, Effua Usuf, Kalifa Bojang, Syed M. A. Zaman, Fatou Secka, Suzanne Anderson, Anna RocaIsatou Sarr, Momodou Saidykhan, Saffiatou Darboe, Samba Ceesay, Umberto D'alessandro, Henriëtte A. Moll, Dorine M. Borensztajn, Nienke N. Hagedoorn, Chantal Tan, Clementien L. Vermont, Joany Zachariasse, W. Dik, Philipp Agyeman, Luregn J. Schlapbach, Christoph Aebi, Verena Wyss, Mariama Usman, Eric Giannoni, Martin Stocker, Klara M. Posfay-Barbe, Ulrich Heininger, Sara Bernhard-Stirnemann, Anita Niederer-Loher, Christian Kahlert, Giancarlo Natalucci, Christa Relly, Thomas Riedel, Christoph Aebi, Christoph Berger, Enitan D. Carrol, Stéphane Paulus, Elizabeth Cocklin, Aakash Khanijau, Rebecca Jennings, Joanne Johnston, Simon Leigh, Karen Newall, Sam Romaine, Maria Tsolia, Irini Eleftheriou, Maria Tambouratzi, Antonis Marmarinos Marietta Xagorari, Kelly Syggelou, Colin Fink, Marie Voice, Leo Calvo-Bado, Werner Zenz, Benno Kohlmaier, Nina A. Schweintzger, Manfred G. Sagmeister, Daniela S. Kohlfürst, Christoph Zurl, Alexander Binder, Susanne Hösele, Manuel Leitner, Lena Pölz, Glorija Rajic, Sebastian Bauchinger, Hinrich Baumgart, Martin Benesch, Astrid Ceolotto, Ernst Eber, Siegfried Gallistl, Gunther Gores, Harald Haidl, Almuthe Hauer, Christa Hude, Markus Keldorfer, Larissa Krenn, Heidemarie Pilch, Andreas Pfleger, Klaus Pfurtscheller, Gudrun Nordberg, Tobias Niedrist, Siegfried Rödl, Andrea Skrabl-Baumgartner, Matthias Sperl, Laura Stampfer, Volker Strenger, Holger Till, Andreas Trobisch, Sabine Löffler, Shunmay Yeung, Juan Emmanuel Dewez, Martin Hibberd, David Bath, Alec Miners, Ruud Nijman, Catherine Wedderburn, Anne Meierford, Baptiste Leurent, Ronald de Groot, Michiel Van der Flier, Marien I. de Jonge, Koen van Aerde, Wynand Alkema, Bryan van den Broek, Jolein Gloerich, Alain J. van Gool, Stefanie Henriet, Martijn Huijnen, Ria Philipsen, Esther Willems, G.P.J.M. Gerrits, M. van Leur, J. Heidema, L. de Haan, C.J. Miedema, C. Neeleman, C.C. Obihara, G.A. Tramper-Stranders, Andrew J. Pollard, Rama Kandasamy, Stéphane Paulus, Michael J. Carter, Daniel O'Connor, Sagida Bibi, Dominic F. Kelly, Meeru Gurung, Stephen Thorson, Imran Ansari, David R. Murdoch, Shrijana Shrestha, Zoe Oliver, Marieke Emonts, Emma Lim, Lucille Valentine, Karen Allen, Kathryn Bell, Adora Chan, Stephen Crulley, Kirsty Devine, Daniel Fabian, Sharon King, Paul McAlinden, Sam McDonald, Anne McDonnell, Ailsa Pickering, Evelyn Thomson, Amanda Wood, Diane Wallia, Phil Woodsford, Frances Baxter, Ashley Bell, Mathew Rhodes, Rachel Agbeko, Christine Mackerness, Bryan Baas, Lieke Kloosterhuis, Wilma Oosthoek, Tasnim Arif, Joshua Bennet, Kalvin

Collings, Ilona van der Giessen, Alex Martin, Aqeela Rashid, Emily Rowlands, Gabriella de Vries, Fabian van der Velden, Joshua Soon, Ulrich von Both, Laura Kolberg, Manuela Zwerenz, Judith Buschbeck, Christoph Bidlingmaier, Vera Binder, Katharina Danhauser, Nikolaus Haas, Matthias Griese, Tobias Feuchtinger, Julia Keil, Matthias Kappler, Eberhard Lurz, Georg Muench, Karl Reiter, Carola Schoen, François Mallet, Karen Brengel-Pesce, Alexandre Pachot, Marine Mommert, Marko Pokorn, Mojca Kolnik, Katarina Vincek, Tina Plankar Srovin, Natalija Bahovec, Petra Prunk, Veronika Osterman, Tanja Avramoska, Taco Kuijpers, Ilse Jongerius, J.M. van den Berg, D. Schonenberg, A.M. Barendregt, D. Pajkrt, M. van der Kuip, A.M. van Furth, Evelien Sprenkeler, Judith Zandstra, G. van Mierlo, and J. Geissler

The members of GENDRES (www.gendres.org) are Federico Martinón-Torres, Miriam Cebey-López, Jacobo Pardo-Seco, Ana Dacosta, Irene Rivero-Calle, Lorenzo Redondo-Collazo, Carmen Rodríguez-Tenreiro, Ruth Barral-Arca, Sara Pischedda, María José Currás-Tuala, Alberto Gómez-Carballa, Sandra Viz-Lasheras, Xabier Bello-Paderne, Antonio Salas Ellacuriaga, Nazareth Martinón-Torres, Alberto Gómez-Carballa, José María Martinón Sánchez, Belén Mosquera Pérez, María del Sol Porto Silva, Mirian Ben García, Patricia Regueiro Casuso, Blanca Díaz Estéban, José Peña Guitián, Carmen Curros Novo, Miriam Puente Puig, Rosaura Leis Trabazo, Nazareth Martinón-Torres, José María Martinón Sánchez, Máximo Francisco Fraga Rodríguez, José Ramón Antúnez, Laura Moreno Galarraga, Teresa González López, Ana López Fernández, Nuria Romero Pérez, José Antonio Couceiro Gianzo, Nazareth Fuentes Pérez, Miguel Sánchez Forte, Cristina Calvo Rey, Carlos Grasa, María Luz García García, Cristina Calvo Monge, Eider Oñate Vergara, Carmen Martínez Padilla, María de las Mercedes Martínez Rebollo, María Rocío Martín Moya, Susana Beatriz Reyes, María Cruz León León, Santiago Alfayate, Servicio de Pediatría, Hospital Universitario de Getafe, Spain: Andrés J. Alcaraz Romero, Roi Piñeiro, Elvira González Salas, Sira Fernández de Miguel, Belen Joyanes Abancens, Esther Aleo Lujan, Alfredo Tagarro García, María Luisa Herreros, Rut del Valle, Libertad Latorre Navarro, María Concepción Zazo Sanchindrán, Mariano Esteban, Marta González Lorenzo, Lorena Moreno Requena, Juan Luis Santos Pérez, César Gavilán Martín, Lucía González-Moro Azorín, María José Cilleruelo Ortega, Luz Golmayo, Francisca Portero Azorín, Andrés Concha Torre, Lucía Rodríguez García, Carlos Rodrigo Gonzalo de Liria, Andrés Antón Pagarolas, Maria Méndez, Critina Prat, Jesús López-Herce, Miriam García Samprudencio, Gema Manrique Martín, Paula García Casas, María Jesús Cabero, Marta Pareja, Miguel Lillo Lillo, Pablo Rojo, Cristina Epalza, Juan Ignacio Sánchez Díaz, Alba Palacios, Adriana Navas Carretero, Estefanía Barral, Miriam Herrera, Elvira Cobo Vazquez, Elena del Castillo Navío, Patricia Flores Perez, Raquel Jiménez, Paula García Casas, Javier Pilar Orive, Elva Rodriguez Merino, Ana Pérez Aragón, Francisco Giménez Sánchez, José Manuel Fernández Bustillo, Mercedes del Río Garma, Josefina Pena Nieto, Ma Elena Álvarez Garnelo, Ángel López-Silvarrey Varela, Patricia Vilas Rodríguez, Consuelo Alonso Alonso, Julio Regueiro Martínez, Natalia García Sánchez, Lucía Castro Paz, Juan Manuel Sánchez Lastres, Manuel Sampedro Campos, Matilde Sagrario Somoza Martín, Marta Carballal Mariño, Samuel Héctor Campuzano Martín, Clara García Cendón, María José Méndez Bustelo, Fernanda Crespo Vázquez, María Paz Vior Álvarez, María Elena Noya Beiroa, and Miguel Ángel San José González.

The members of UCSD Pediatric Emergency Medicine Kawasaki Disease Research Group are Naomi Abe, Lukas Austin-Page, Amy Bryl, J. Joelle Donofrio-Ödmann, Atim Ekpenyong, Michael Gardiner, David J. Gutglass, Scott Herskovitz, Paul Ishimine, Margaret B. Nguyen, Kristy Schwartz, Stacey Ulrich, Tatyana Vayngortin, and Elise Zimmerman.

The members of the UK Kawasaki Genetics consortium are M. Levin, R. Galassini, V. Wright, J. Herberg, Y. Singh, J. Bytham, J. Sharp, P. Bala, A. Kitching, S. Paulus, E. Carol, B. Larru, S. Wadeson, J. Johnstone, R. Jennings, A. Chickermane, K. Cotter, H. Jepps, T. Booth, R. Swingler, R. Tulloh, Karen Sheehan, M. Ahmed, S. Boswell, C. Backhouse, M. Olabi, K.U. Rahman, S. Kilroy, M. Home, T. Banerjee, G. Nyamugunduru, A. Cowton, D. Egginton, M. Jawad, L. Bailey, E. Menson, P. Brogan, Y. Glackin, C. Brunskill, S. Hackett, J. Daglish, S. Meyrick, E. Collins, D. Bolton, S. Gormley, S. Mustafa, P. Desai, L. Hunt, T. Chawatama, S. Luck, J. Crooks, T. O'Brien, S. O'Riordan, N. Balatoni, N. Maher, S. Chandrasekaran, N. Keenan, K. Davies, S. Kempson, C. Busby, E. Odeka, G. O'Connor, I. Haar, G. Osborne, H. Walker, A. Williams, L. Woodward, C. Rishton, V. Puthi, A. Pearson, P. Goodyear, C. Davidson, N. De Vere, G. Castle, M. Farrier, N. Pemberton, S. Misra, C. Fish, P. Graham, J. Henry, A. Olabi, K. Allison, A. Kannivelu, J. Jones, J. Alexander, E. Roe, R. Pringle, A. Cope, F. Shackley, R. Kumar, G. McIlhinney, S. Armstrong, P. Heath, E. Vitale, J. Stuart, N. Brennan, C. Cooper, S. Bennett, A. Petkar, C. Greenway, W. Hulse, K. Fidler, K. Moscovici, S. Sobowieck Kouman, F. Franklin, M. Bartsota, N. Venkata, A. Prendiville, O. Elmasry, H. Osborne, G. Craig, B. Bromage, M. Raman, P. Fitzell, H. Bearne, J. Palmer, S. Davidson, and N. Clements.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- STUDY PARTICIPANT DETAILS
  - Description of the validation study (RNA-Seq dataset)
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Microarray data pre-processing
  - Feature pre-filtering
  - Method selection
  - LASSO+Ridge hybrid
  - Cost and rescaling
  - Performance
  - RNA-sequencing analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.medj.2023.06.007.

ACKNOWLEDGMENTS

The authors acknowledge funding from European Union's Seventh Framework programme and the Horizon 2020 research and innovation programme under GA no. 279185 EUCLIDS and no. 668303 PERFORM. M.K. is supported from the Wellcome Trust (206508/Z/17/Z) and the Medical Research Foundation (MRF-160-0008-ELP-KAFO-C0801). D.H.-C., C.W., R.G., V.W., A.C., J.H., M.L., and M.K. have received support from the NIHR Imperial BRC.

AUTHOR CONTRIBUTIONS

Conceptualization, D.H.-C., A.J.C., J.A.H., M.L., and M.K.; methodology, all authors; software programming, D.H.-C. and C.W.; investigation, all authors;

resources, all authors; data curation management, all authors; writing – original draft, D.H.-C., V.J.W., L.J.M.C., A.J.C., J.A.H., M.L., and M.K.; writing – review & editing preparation, all authors; visualization, D.H.-C., V.J.W., L.M.J.C., A.J.C., J.A.H., M.L., and M.K.; supervision, S.T.A., J.M.v.d.B., M.E., E.D.C., C.G.F., R.d.G., M.L.H., M.P.N., A.J.P., A.S., L.J.S., W.Z., M.V.d.F., H.J.Z., T.K., J.C.B., F.M.-T., V.J.W., L.J.M.C., A.J.C., J.A.H., M.L., and M.K.; funding acquisition, P.K.A.A., S.T.A., M.E., E.D.C., C.G.F., R.d.G., M.L.H., M.P.N., S.P., A.J.P., A.S., L.J.S., W.Z., M.V.d.F., H.J.Z., T.K., J.C.B., F.M.-T., V.J.W., L.J.M.C., A.J.C., J.A.H., M.L., and M.K.; D.H.-C. performed and replicated statistical analyses; L.J.M.C. and M.K. oversaw statistical analyses; D.H.-C. and M.K. had unrestricted access to all data; D.H.-C., V.J.W., L.J.M.C., A.J.C., J.A.H., M.L., and M.K. prepared the first draft of the manuscript, reviewed it, and edited it; all authors agreed to submit the manuscript, read and approved the final draft, and take full responsibility of its content, including the accuracy of the data and the fidelity of the work and its statistical analysis.

DECLARATION OF INTERESTS

The authors declare that a patent application on the method described in this manuscript has been filed (2304229.4/GB/PRV, 23-03-2023).

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: November 15, 2021
Revised: June 8, 2023
Accepted: June 19, 2023
Published: August 19, 2023

REFERENCES

1. Sands, R., Shanmugavadivel, D., Stephenson, T., and Wood, D. (2012). Medical problems presenting to paediatric emergency departments: 10 years on. Emerg. Med. J. *29*, 379–382. https://doi.org/10.1136/emj.2010.106229.

2. Irwin, A.D., Grant, A., Williams, R., Kolamunnage-Dona, R., Drew, R.J., Paulus, S., Jeffers, G., Williams, K., Breen, R., Preston, J., et al. (2017). Predicting risk of serious bacterial infections in febrile children in the emergency department. Pediatrics *140*, e20162853. https://doi.org/10.1542/peds.2016-2853.

3. Nijman, R.G., Zwinkels, R.L.J., van Veen, M., Steyerberg, E.W., van der Lei, J., Moll, H.A., and Oostenbrink, R. (2011). Can urgency classification of the Manchester triage system predict serious bacterial infections in febrile children? Arch. Dis. Child. *96*, 715–722. https://doi.org/10.1136/adc.2010.207845.

4. Caliendo, A.M., Gilbert, D.N., Ginocchio, C.C., Hanson, K.E., May, L., Quinn, T.C., Tenover, F.C., Alland, D., Blaschke, A.J., Bonomo, R.A., et al. (2013). Better tests, better care: Improved diagnostics for infectious diseases. Clin. Infect. Dis. *57*, S139–S170. https://doi.org/10.1093/cid/cit578.

5. Schreiber, J., Nierhaus, A., Braune, S.A., de Heer, G., and Kluge, S. (2013). Comparison of three different commercial PCR assays for the detection of pathogens in critically ill sepsis patients. Med. Klin. Intensivmed. Notfmed. *108*, 311–318. https://doi.org/10.1007/s00063-013-0227-1.

6. Pneumonia Etiology Research for Child Health PERCH Study Group (2019). Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: The PERCH multi-country case-control study. Lancet *394*, 757–779. https://doi.org/10.1016/S0140-6736(19)30721-4.

7. Shiff, N.J., Tucker, L.B., Guzman, J., Oen, K., Yeung, R.S.M., and Duffy, C.M. (2010). Factors associated with a longer time to access pediatric rheumatologists in Canadian children with juvenile idiopathic arthritis. J. Rheumatol. *37*, 2415–2421. https://doi.org/10.3899/jrheum.100083.

8. Lewandowski, L.B., Watt, M.H., Schanberg, L.E., Thielman, N.M., and Scott, C. (2017). Missed opportunities for timely diagnosis of pediatric lupus in South Africa: A qualitative study. Pediatr. Rheumatol. *15*, 14. https://doi.org/10.1186/s12969-017-0144-6.

9. Minich, L.L., Sleeper, L.A., Atz, A.M., McCrindle, B.W., Lu, M., Colan, S.D., Printz, B.F., Klein, G.L., Sundel, R.P., Takahashi, M., et al. (2007). Delayed diagnosis of kawasaki disease: What are the risk factors? Pediatrics *120*, e1434–e1440. https://doi.org/10.1542/peds.2007-0815.

10. Martinón-Torres, F., Salas, A., Rivero-Calle, I., Cebey-López, M., Pardo-Seco, J., Herberg, J.A., Boeddha, N.P., Klobassa, D.S., Secka, F., Paulus, S., et al. (2018). Life-threatening infections in children in Europe (the EUCLIDS project): A prospective cohort study. Lancet. Child Adolesc. Health *2*, 404–414. https://doi.org/10.1016/S2352-4642(18)30113-5.

11. Boeddha, N.P., Schlapbach, L.J., Driessen, G.J., Herberg, J.A., Rivero-Calle, I., Cebey-López, M., Klobassa, D.S., Philipsen, R., de Groot, R., Inwald, D.P., et al. (2018). Mortality and morbidity in community-acquired sepsis in European pediatric intensive care units: A prospective cohort study from the European Childhood Life-threatening Infectious Disease study (EUCLIDS). Crit. Care *22*, 143. https://doi.org/10.1186/s13054-018-2052-7.

12. WHO. Data Global Action Plan on Antimicrobial Resistance. 2015. https://www.who.int/publications/i/item/9789241509763

13. Herberg, J.A., Kaforou, M., Gormley, S., Sumner, E.R., Patel, S., Jones, K.D.J., Paulus, S., Fink, C., Martinon-Torres, F., Montana, G., et al. (2013). Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes. J. Infect. Dis. *208*, 1664–1668. https://doi.org/10.1093/infdis/jit348.

14. Herberg, J.A., Kaforou, M., Wright, V.J., Shailes, H., Eleftherohorinou, H., Hoggart, C.J.,

Cebey-López, M., Carter, M.J., Janes, V.A., Gormley, S., et al. (2016). Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. JAMA, J. Am. Med. Assoc. *316*, 835–845. https://doi.org/10.1001/jama.2016.11236.

15. Kaforou, M., Herberg, J.A., Wright, V.J., Coin, L.J.M., and Levin, M. (2017). Diagnosis of bacterial infection using a 2-transcript host RNA signature in febrile infants 60 days or younger. JAMA, J. Am. Med. Assoc. *317*, 1577–1578. https://doi.org/10.1001/jama.2017.1365.

16. Mahajan, P., Kuppermann, N., Mejias, A., Suarez, N., Chaussabel, D., Casper, T.C., Smith, B., Alpern, E.R., Anders, J., Atabaki, S.M., et al. (2016). Association of RNA biosignatures with bacterial infections in febrile infants aged 60 days or younger. JAMA, J. Am. Med. Assoc. *316*, 846–857. https://doi.org/10.1001/jama.2016.9207.

17. Tsalik, E.L., Henao, R., Nichols, M., Burke, T., Ko, E.R., McClain, M.T., Hudson, L.L., Mazur, A., Freeman, D.H., Veldman, T., et al. (2016). Host gene expression classifiers diagnose acute respiratory illness etiology. Sci. Transl. Med. *8*, 322ra11. https://doi.org/10.1126/scitranslmed.aad6873.

18. O'Connor, D., Pinto, M.V., Sheerin, D., Tomic, A., Drury, R.E., Channon-Wells, S., Galal, U., Dold, C., Robinson, H., Kerridge, S., et al. (2020). Gene expression profiling reveals insights into infant immunological and febrile responses to group B meningococcal vaccine. Mol. Syst. Biol. *16*, e9888. https://doi.org/10.15252/MSB.20209888.

19. Sweeney, T.E., Wong, H.R., and Khatri, P. (2016). Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Sci. Transl. Med. *8*, 346ra91. https://doi.org/10.1126/SCITRANSLMED.AAF7165.

20. Colborn, J.M., Ylöstalo, J.H., Koita, O.A., Cissé, O.H., and Krogstad, D.J. (2015). Human gene expression in uncomplicated plasmodium falciparum malaria. J. Immunol. Res. *2015*, 162639. https://doi.org/10.1155/2015/162639.

21. Lee, H.J., Georgiadou, A., Walther, M., Nwakanma, D., Stewart, L.B., Levin, M., Otto, T.D., Conway, D.J., Coin, L.J., and Cunnington, A.J. (2018). Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. Sci. Transl. Med. *10*, eaar3619. https://doi.org/10.1126/scitranslmed.aar3619.

22. Loke, P., Hammond, S.N., Leung, J.M., Kim, C.C., Batra, S., Rocha, C., Balmaseda, A., and Harris, E. (2010). Gene expression patterns of dengue virus-infected children from Nicaragua reveal a distinct signature of increased metabolism. PLoS Negl. Trop. Dis. *4*, e710. https://doi.org/10.1371/journal.pntd.0000710.

23. Mejias, A., Dimo, B., Suarez, N.M., Garcia, C., Suarez-Arrabal, M.C., Jartti, T., Blankenship, D., Jordan-Villegas, A., Ardura, M.I., Xu, Z., et al. (2013). Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. PLoS Med. *10*, e1001549. https://doi.org/10.1371/journal.pmed.1001549.

24. Barral-Arca, R., Gómez-Carballa, A., Cebey-López, M., Bello, X., Martinón-Torres, F., and Salas, A. (2020). A meta-analysis of multiple whole blood gene expression data unveils a diagnostic host-response transcript signature for respiratory syncytial virus. Int. J. Mol. Sci. *21*, 1831. https://doi.org/10.3390/ijms21051831.

25. DeBerg, H.A., Zaidi, M.B., Altman, M.C., Khaenam, P., Gersuk, V.H., Campos, F.D., Perez-Martinez, I., Meza-Segura, M., Chaussabel, D., Banchereau, J., et al. (2018). Shared and organism-specific host responses to childhood diarrheal diseases revealed by whole blood transcript profiling. PLoS One *13*, e0192082. https://doi.org/10.1371/journal.pone.0192082.

26. Gómez-Carballa, A., Barral-Arca, R., Cebey-López, M., Currás-Tuala, M.J., Pischedda, S., Gómez-Rial, J., Habgood-Coote, D., Herberg, J.A., Kaforou, M., Martinón-Torres, F., and Salas, A. (2021). Host transcriptomic response following administration of rotavirus vaccine in infants' mimics wild type infection. Front. Immunol. *11*. https://doi.org/10.3389/fimmu.2020.580219.

27. Anderson, S.T., Kaforou, M., Brent, A.J., Wright, V.J., Banwell, C.M., Chagaluka, G., Crampin, A.C., Dockrell, H.M., French, N., Hamilton, M.S., et al. (2014). Diagnosis of childhood tuberculosis and host RNA expression in Africa. N. Engl. J. Med. *370*, 1712–1723. https://doi.org/10.1056/nejmoa1303657.

28. Berry, M.P.R., Graham, C.M., McNab, F.W., Xu, Z., Bloch, S.A.A., Oni, T., Wilkinson, K.A., Banchereau, R., Skinner, J., Wilkinson, R.J., et al. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature *466*, 973–977. https://doi.org/10.1038/nature09247.

29. Kaforou, M., Wright, V.J., Oni, T., French, N., Anderson, S.T., Bangani, N., Banwell, C.M., Brent, A.J., Crampin, A.C., Dockrell, H.M., et al. (2013). Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: A case-control study. PLoS Med. *10*, e1001538. https://doi.org/10.1371/journal.pmed.1001538.

30. Sweeney, T.E., Braviak, L., Tato, C.M., and Khatri, P. (2016). Genome-wide expression for diagnosis of pulmonary tuberculosis: A multicohort analysis. Lancet Respir. Med. *4*, 213–224. https://doi.org/10.1016/S2213-2600(16)00048-5.

31. Wright, V.J., Herberg, J.A., Kaforou, M., Shimizu, C., Eleftherohorinou, H., Shailes, H., Barendregt, A.M., Menikou, S., Gormley, S., Berk, M., et al. (2018). Diagnosis of kawasaki disease using a minimal whole-blood gene expression signature. JAMA Pediatr. *172*, e182293. https://doi.org/10.1001/jamapediatrics.2018.2293.

32. Mackay, M., Oswald, M., Sanchez-Guerrero, J., Lichauco, J., Aranow, C., Kotkin, S., Korsunsky, I., Gregersen, P.K., and Diamond, B. (2016). Molecular signatures in systemic lupus erythematosus: Distinction between disease flare and infection. Lupus Sci. Med. *3*, e000159. https://doi.org/10.1136/lupus-2016-000159.

33. Liu, H., and Yu, B. (2013). Asymptotic properties of lasso+mLS and lasso+ridge in sparse high-dimensional linear regression. Electron. J. Stat. *7*, 3124–3169. https://doi.org/10.1214/14-EJS875.

34. McCrindle, B.W., Rowley, A.H., Newburger, J.W., Burns, J.C., Bolger, A.F., Gewitz, M., Baker, A.L., Jackson, M.A., Takahashi, M., Shah, P.B., et al. (2017). Diagnosis, treatment, and long-term management of kawasaki disease: A scientific statement for health professionals from the American Heart Association. Circulation *135*, e927–e999. https://doi.org/10.1161/CIR.0000000000000484.

35. Burgner, D., and Harnden, A. (2005). Kawasaki disease: What is the epidemiology telling us about the etiology? Int. J. Infect. Dis. *9*, 185–194. https://doi.org/10.1016/j.ijid.2005.03.002.

36. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

37. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*. R29–R17. https://doi.org/10.1186/gb-2014-15-2-r29.

38. Nijman, R.G., Oostenbrink, R., Moll, H.A., Casals-Pascual, C., von Both, U., Cunnington, A., De, T., Eleftheriou, I., Emonts, M., Fink, C., et al. (2021). A novel framework for phenotyping children with suspected or confirmed infection for future biomarker studies. Front. Pediatr. *9*, 688272. https://doi.org/10.3389/fped.2021.688272.

39. Yoshizato, R., and Koga, H. (2020). Comparison of initial and final diagnoses in children with acute febrile illness: A retrospective, descriptive study: Initial and final diagnoses in children with acute fever. J. Infect. Chemother. *26*, 251–256. https://doi.org/10.1016/j.jiac.2019.09.015.

40. Antoon, J.W., Peritz, D.C., Parsons, M.R., Skinner, A.C., and Lohr, J.A. (2018). Etiology and resource use of fever of unknown origin in hospitalized children. Hosp. Pediatr. *8*, 135–140. https://doi.org/10.1542/hpeds.2017-0098.

41. D'Acremont, V., Kilowoko, M., Kyungu, E., Philipina, S., Sangu, W., Kahama-Maro, J., Lengeler, C., Cherpillod, P., Kaiser, L., and Genton, B. (2014). Beyond malaria — causes of fever in outpatient Tanzanian children. N. Engl. J. Med. *370*, 809–817. https://doi.org/10.1056/nejmoa1214482.

42. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). G:profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198. https://doi.org/10.1093/NAR/GKZ369.

43. Alessandrini, M., Kappala, S.S., and Pepper, M.S. (2017). AMLprofiler: A diagnostic and prognostic microarray for acute myeloid leukemia. Methods Mol. Biol. *1633*, 101–123. https://doi.org/10.1007/978-1-4939-7142-8_7.

44. Tsang, H.F., Xue, V.W., Koh, S.P., Chiu, Y.M., Ng, L.P.W., and Wong, S.C.C. (2017). NanoString, a novel digital color-coded

barcode technology: Current and future applications in molecular diagnostics. Expert Rev. Mol. Diagn. *17*, 95–103. https://doi.org/10.1080/14737159.2017.1268533.

45. Rodriguez-Manzano, J., Moniri, A., Malpartida-Cardenas, K., Dronavalli, J., Davies, F., Holmes, A., and Georgiou, P. (2019). Simultaneous single-channel multiplexing and quantification of carbapenem-resistant genes using multidimensional standard curves. Anal. Chem. *91*, 2013–2020. https://doi.org/10.1021/acs.analchem.8b04412.

46. Pennisi, I., Rodriguez-Manzano, J., Moniri, A., Kaforou, M., Herberg, J.A., Levin, M., and Georgiou, P. (2021). Translation of a host blood RNA signature distinguishing bacterial from viral infection into a platform suitable for development as a point-of-care test. JAMA Pediatr. *175*, 417–419. https://doi.org/10.1001/jamapediatrics.2020.5227.

47. Dailey, P.J., Kelly-Cirino, C., and Osborn, J. (2019). Landscape of Benchtop Immunoassay Platforms for Near Patient Testing: The MAPDx Program. https://www.finddx.org/wp-content/uploads/2022/12/20180405_lds_fever_immunoassay_FV_EN.pdf.

48. Berkley, J.A., Munywoki, P., Ngama, M., Kazungu, S., Abwao, J., Bett, A., Lassauniére, R., Kresfelder, T., Cane, P.A., Venter, M., et al. (2010). Viral etiology of severe pneumonia among Kenyan infants and children. JAMA, J. Am. Med. Assoc. *303*, 2051–2057. https://doi.org/10.1001/jama.2010.675.

49. Jaggi, P., Mejias, A., Xu, Z., Yin, H., Moore-Clingenpeel, M., Smith, B., Burns, J.C., Tremoulet, A.H., Jordan-Villegas, A., Chaussabel, D., et al. (2018). Whole blood transcriptional profiles as a prognostic tool in complete and incomplete kawasaki disease. PLoS One *13*, e0197858. https://doi.org/10.1371/JOURNAL.PONE.0197858.

50. Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M., Cepika, A.M., Acs, P., Turner, J., Anguiano, E., et al. (2016). Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. Cell *165*, 551–565. https://doi.org/10.1016/J.CELL.2016.03.008.

51. Hu, X., Yu, J., Crosby, S.D., and Storch, G.A. (2013). Gene expression profiles in febrile children with defined viral and bacterial infection. Proc. Natl. Acad. Sci. USA *110*, 12792–12797. https://doi.org/10.1073/PNAS.1302968110.

52. Idaghdour, Y., Quinlan, J., Goulet, J.P., Berghout, J., Gbeha, E., Bruat, V., Malliard, T.D., Grenier, J.C., Gomez, S., Gros, P., et al. (2012). Evidence for additive and interaction effects of host genotype and infection in malaria. Proc. Natl. Acad. Sci. USA *109*, 16786–16793. https://doi.org/10.1073/PNAS.1204945109/SUPPL_FILE/SD05.XLSX.

53. Banchereau, R., Jordan-Villegas, A., Ardura, M., Mejias, A., Baldwin, N., Xu, H., Saye, E., Rossello-Urgell, J., Nguyen, P., Blankenship, D., et al. (2012). Host immune transcriptional profiles reflect the variability in clinical disease manifestations in patients with staphylococcus aureus infections. PLoS One *7*, e34390. https://doi.org/10.1371/JOURNAL.PONE.0034390.

54. Hao, S., Jin, B., Tan, Z., Li, Z., Ji, J., Hu, G., Wang, Y., Deng, X., Kanegaye, J.T., Tremoulet, A.H., et al. (2016). A classification tool for differentiation of kawasaki disease from other febrile illnesses. J. Pediatr. *176*, 114–120.e8. https://doi.org/10.1016/j.jpeds.2016.05.060.

55. R Core Team (2018). R: A Language and Environment for Statistical Computing. https://www.R-project.org.

56. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: Archive for functional genomics data sets - update. Nucleic Acids Res. *41*, D991–D995. https://doi.org/10.1093/nar/gks1193.

57. Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., et al. (2015). ArrayExpress update-simplifying data submissions. Nucleic Acids Res. *43*, D1113–D1116. https://doi.org/10.1093/nar/gku1057.

58. Davis, S., and Meltzer, P.S. (2007). GEOquery: A bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics *23*, 1846–1847. https://doi.org/10.1093/bioinformatics/btm254.

59. Du, P., Kibbe, W.A., and Lin, S.M. (2008). Lumi: A pipeline for processing illumina microarray. Bioinformatics *24*, 1547–1548. https://doi.org/10.1093/bioinformatics/btn224.

60. Du, P., Kibbe, W.A., and Lin, S.M. (2007). NuID: A universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays. Biol. Direct *2*, 16. https://doi.org/10.1186/1745-6150-2-16.

61. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127. https://doi.org/10.1093/biostatistics/kxj037.

62. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

63. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22.

64. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. Nat. Methods *12*, 115–121. https://doi.org/10.1038/nmeth.3252.

65. Kang, C., Huo, Y., Xin, L., Tian, B., and Yu, B. (2019). Feature selection and tumor classification for microarray data using relaxed lasso and generalized multi-class support vector machine. J. Theor. Biol. *463*, 77–91. https://doi.org/10.1016/j.jtbi.2018.12.010.

66. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. Ann. Statist. *32*, 407–499. https://doi.org/10.1214/009053604000000067.

67. Zadrozny, B., Langford, J., and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In Proceedings - IEEE International Conference on Data Mining (ICDM), pp. 435–442. https://doi.org/10.1109/icdm.2003.1250950.

68. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: An open-source package for r and s+ to analyze and compare ROC curves. BMC Bioinf. *12*, 77. https://doi.org/10.1186/1471-2105-12-77.

69. Andrews, S. (2010). Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. URL: https://www.bioinformatics. babraham. ac. uk/projects/fastqc.

70. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

71. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

72. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

73. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

74. Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

75. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 20Q18. Nucleic Acids Res. *46*, D754–D761. https://doi.org/10.1093/nar/gkx1098.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
	Wright et al. (2018)[31]	GEO: GSE73464
	Jaggi et al. (2018)[49]	GEO: GSE68004
	Banchereau et al.*(2016)*[50]	GEO: GSE65391
	Mahajan et al. (2016)[16]	GEO: GSE64456
	Herberg et al. (2013)[13]	GEO: GSE42026
	Hu et al.(2013)[51]	GEO: GSE40396
	Anderson et al.(2014)[27]	GEO: GSE39941
	Mejias et al. (2013)[23]	GEO: GSE38900
	Idaghdour et al.*(2012)*[52]	GEO: GSE34404
	Banchereau et al.*(2012)*[53]	GEO: GSE30119
	NA	GEO: GSE29366
	Berry et al. (2010)[28]	GEO: GSE22098
	This paper	EBI: E-MTAB-11671
Human genome	ensembl	Gch38 version 89
Software and algorithms		
R	https://cran.r-project.org/	Version 3.4.4; RRID:SCR_001905
GEOquery	https://bioconductor.org/packages/release/bioc/html/GEOquery.html	RRID:SCR_000146
Lumi	https://bioconductor.org/packages/release/bioc/html/lumi.html	RRID:SCR_012781
lumiHumanIDMapping	https://bioconductor.org/packages/release/data/annotation/html/lumiHumanIDMapping.html	N/A
SVA	https://bioconductor.org/packages/release/bioc/html/sva.html	RRID:SCR_012836
Limma	https://bioconductor.org/packages/release/bioc/html/limma.html	RRID:SCR_010943
glmnet	https://cran.r-project.org/web/packages/glmnet/index.html	RRID:SCR_015505
pROC	https://cran.r-project.org/web/packages/pROC/index.html	N/A
STAR	https://github.com/alexdobin/STAR	RRID:SCR_004463
Samtools	http://www.htslib.org/	RRID:SCR_002105
FeatureCounts	https://subread.sourceforge.net/featureCounts.html	RRID:SCR_012919
DESeq2	https://bioconductor.org/packages/release/bioc/html/DESeq2.html	RRID:SCR_015687
G:Profiler	https://cran.r-project.org/web/packages/gprofiler2/index.html	N/A
Original code	Zenodo https://doi.org/10.5281/zenodo.7620205	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Myrsini Kaforou (m.kaforou@imperial.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Newly generated transcriptomic RNA-seq datasets have been uploaded to EBI ArrayExpress: E-MTAB-11671 and the programming code has been uploaded to github https://github.com/d-h-c/multiclass_fever_biomarkers and Zenodo https://doi.org/10.5281/zenodo.7620205. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

STUDY PARTICIPANT DETAILS

Description of the validation study (RNA-Seq dataset)

Patient recruitment. Patients were recruited as part of the European Union Childhood Life-threatening Infectious Disease Study (EUCLIDS https://www.euclids-project.eu), a prospective, multicentre, cohort study conducted in six countries in Europe. Patients aged 1 month to 18 years with sepsis (or suspected sepsis) or severe focal infections, admitted to 98 participating hospitals in the UK, Austria, Germany, Lithuania, Spain, Switzerland and the Netherlands were prospectively recruited between July 1, 2012, and Dec 31, 2015. Febrile patients were recruited additionally with similar criteria in Spain (GENDRES network, Santiago de Compostela), in the Netherlands (Virgo cohort, JIA cohort), in the USA (Rady Children's Hospital-San Diego as described previously),[54] and in Cape Town (Red Cross Children's hospital) between 2009 and 2013. Patients were recruited if they met the inclusion criteria of having febrile illness (temperature $\geq$38°C) of perceived sufficient severity to warrant blood testing or hospital admission and were <17 years of age. Participants' information on sex, age, and race was self-reported, or reported by parent/guardian. Information on gender and socioeconomic status was not collected. Patients were excluded if they had comorbidities or treatments likely to affect gene expression, including prior bone marrow transplant, immunodeficiency, or immunosuppressive treatment. Blood samples for RNA analysis were collected together with clinical blood tests at, or as close as possible to, presentation to hospital, irrespective of antibiotic use at the time of collection.

Diagnostic process. All patients underwent routine diagnostic investigations as part of clinical care in each hospital's microbiology and virology laboratories, including blood count and differential, C-reactive protein (CRP), blood chemistry, blood, and urine cultures, and cerebrospinal fluid (CSF) analysis where indicated. Throat swabs were cultured for bacteria, and viral diagnostics were undertaken on nasopharyngeal aspirates using multiplex PCR for common respiratory viruses. Chest radiographs and other tests were undertaken as clinically indicated. Patients were assigned to diagnostic groups using predefined criteria as described previously.[14,30] The Definite Bacterial group included only patients with bacteria identified in a sample from a sterile site, and the Definite Viral group included only patients with culture, PCR or Immunofluorescent test - confirmed viral infection. Children were recorded as having juvenile idiopathic arthritis, Kawasaki disease, tuberculosis disease and malaria in the respective studies. Children in whom definitive diagnosis was not established were not used in this study.

Study conduct and oversight. Clinical data and patient samples were identified only by study number. Assignment of patients to clinical groups was made independent of those managing the patient clinically by consensus of two experienced clinicians, after review of the investigation results and using previously agreed definitions. Statistical analysis was conducted after the RNA expression data and clinical assignment databases had been locked. Written, informed consent was obtained from parents or guardians at all sites using locally approved permissions (St Mary's Research Ethics Committee (REC 09/H0712/58); Ethical Committee of Clinical Investigation of Galicia (CEIC ref 2010/015); Amsterdam, the Netherlands (NL41023.018.12 and NL34230.018.10); the University of California San Diego (Human Research Protection Program 140220); The Gambia Govemment/MRC Joint Ethics (Committee reference L2013.07V2); Cape Town, South Africa (HREC No 389/2017 linked to No 045/2008); Cantonal Ethis Committee Bern (KEK-029)).

CellPress
OPEN ACCESS

Peripheral blood RNA sequencing. Whole blood was collected at the time of recruitment into PAXgene blood RNA tubes (PreAnalytiX, Germany), frozen, and later extracted. Library preparation and sequencing of 30 million 75 or 100 bp paired end reads was conducted using the Illumina's TruSeq RNA Sample Preparation Kit, ribosomal and globin RNA depletion was performed using the Illumina® Ribo-Zero Gold kit and HiSeq 4000 at The Wellcome Centre for Human Genetics.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analyses were performed in R version 3.4.4[55]

Microarray data pre-processing

We identified human Illumina gene expression micro-array datasets in National Institutes of Health Gene Expression Omnibus database[56] and ArrayExpress,[57] which included expression data from children with infectious and inflammatory diseases as well as healthy controls (Table S1). Only datasets where Illumina Beadchip arrays (V3, V4) were used to measure whole blood gene expression were included. Datasets were retrieved with GEOquery,[58] normalised using robust spline normalisation (RSN) from the lumi package[59] and log transformed independently prior to batch correction. Probes common to all datasets were identified using lumiHumanIDMapping to map probes to Illumina nuIDs.[60] Duplicate samples between datasets were identified using correlation structure and checked using patient characteristics. ComBat as was performed using the R package SVA[61] to correct for a batch internal to GSE72829 before using COCONUT,[19] which assumes that healthy controls are drawn from the same distribution, to correct for batch effects between experiments (Figure S1).

Disease groups with fewer than 10 patients were excluded from the discovery set, as were cases in which a single causative pathogen was not identified or the diagnosis was uncertain. Stratified holdout was used to select the 25% of the data to be used in testing.

Feature pre-filtering

Prior to feature selection pre-filtering was performed to reduce the size of the search space and remove probes with little or no association with any of the diseases considered. To this end a differential expression analysis was performed with limma[62] for all 153 pairwise disease comparisons. Probes with absolute log2 fold change below 0.5 were discarded and the remaining probes for each comparison were ranked by p value. Probes were selected from these lists in an iterative process until at least 2,000 probes were present. At each iteration the contribution of each probe was divided between the comparisons in which it was selected (i.e. a probe selected by 2 comparisons contributes a weight of 0.5 to each) in order that all comparisons were defined by similar numbers of discriminatory probes.

Method selection

In order to compare methods for performing the feature selection and classification, we used stratified 10-fold cross-validation in the microarray training set, this was repeated 10 times by changing seed values. We considered five different multivariate penalised regression methods, implemented here using glmnet[63]: one-vs-all LASSO, one-vs-all LASSO followed by multinomial Ridge regression over the selected feature set, multinomial-LASSO, multinomial-LASSO + Ridge and multinomial relaxed LASSO. Nested cross validation was used to select hyper-parameters; when performing feature selection, the 1SE method was used and when refitting coefficients, the parameters were selected to minimise error. Performance was

evaluated using mean weighted square error (MWSE) and mean size of the selected feature set. We concluded that the one-versus-all approach was not feasible due to the identification of very large gene signatures with more highly correlated and redundant features (Figure 1A). Of the multi-class approaches used, the LASSO+Ridge two-stage procedure obtained the smallest models with high predictive performance.

LASSO+Ridge hybrid

Penalised regression was performed on standardised expression values using the glmnet[63] package in Bioconductor.[64] Coefficients were grouped so that all coefficients for each feature were set to zero together. l1 and l2 penalised regression were combined into a two-stage procedure, referred to here as LASSO+Ridge,[33] for which the LASSO (l1 penalty) was used to perform feature selection followed by a Ridge regression (l2 penalty) to refit the coefficients for the resulting feature set. This method has similarities with the Relaxed LASSO[65] and LARS-OLS[66] methods, which use LASSO and ordinary least squares (OLS) for the second stage respectively. For the LASSO+Ridge procedure the tuning parameters of LASSO (λ) and Ridge fits (φ) were selected using nested cross validation. At each λ of the LASSO regularisation path, genes with non-zero coefficients are used as input for a Ridge regression. For each Ridge regression the tuning parameter φ was selected to minimise the MWSE. λ was then selected to minimise model size such that the MWSE was within two standard errors of the minimum (2SE). Relative to LASSO and Relaxed LASSO, the LASSO+Ridge hybrid method had lower MWSE for each feature set, which resulted in smaller signatures with similar predictive accuracy (Figure 1A).

Cost and rescaling

Example weighting[67] was used to bias the feature selection in order to prioritise the reduction of false negative error for diseases which are associated with greater immediate risk to the patient. These relative weights were defined for each disease class by a team of 5 paediatric infectious disease specialists to reflect: risk of negative outcome (e.g. death, organ damage), speed of disease progression and the availability of effective treatment (Table S2). The effect of adding class weights to a multinomial LASSO is to bias the feature set and weights towards reducing the false negative error for classes with worse potential outcomes, more rapid progression and available treatment; this also leads to an increase in the false positive error for these classes and the converse for diseases with smaller weights (Figure 1B).

Weights were also modified to counteract the bias induced by differences in the numbers of samples in each group (class imbalance), as there was a 20:1 ratio between most and least abundant classes. This was done by updating class weights by dividing costs by the number of patients in each class.

Performance

Classifier performance is shown using confusion matrices where discrete class predictions, for each patient, are the class with highest predicted probability. ROC curves were derived using the predicted probabilities for each class with the pROC package and trapezoidal calculation of AUC,[68] the bootstrap methods were used for to derive confidence intervals and compare AUCs. Pairwise ROC curves were derived using the ratio of the predicted probabilities of the two classes.

CellPress
OPEN ACCESS

RNA-sequencing analysis

The RNA-Seq analysis pipeline consisted of: quality control using FastQC,[69] MultiQC[70] and annotations modified with BEDTools,[71] alignment and read counting using STAR,[72] SAMtools,[73] FeatureCounts[74] and version 89 ensembl[75] GCh38 genome and annotation. Normalisation was performed using the DESeq2[36] method for estimating scale factors with a subsequent log transformation (Figures S1D and S1E). The 161 microarray probes were mapped uniquely using BLAST to 155 genes, 10 of which were removed due to low read counts (unnormalised counts >5 in fewer than 10 samples) (Table S4). Refitting of the coefficients of the model was performed using Ridge regression on 50% of the dataset, the remainder was used for performance evaluation. The same split was used when retraining and testing comparator signatures. Differential expression and enrichment analyses were performed using DESeq2 and g:Profiler[41] (Figure S6).