Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

A blood-based host gene expression assay for early detection of respiratory viral infection: an index-cluster prospective cohort study

Micah T McClain, Florica J Constantine, Bradly P Nicholson, Marshall Nichols, Thomas W Burke, Ricardo Henao, Daphne C Jones, Lori L Hudson, L Brett Jaggers, Timothy Veldman, Anna Mazur, Lawrence P Park, Sunil Suchindran, Ephraim L Tsalik, Geoffrey S Ginsburg, Christopher W Woods

Lancet Infect Dis 2021; 21: 396–404

Published Online September 24, 2020 https://doi.org/10.1016/S1473-3099(20)30486-2

See Comment page 305

Center for Applied Genomics and Precision Medicine (M T McClain MD, F J Constantine MS, M Nichols MS, T W Burke PhD, R Henao PhD, L L Hudson PhD, T Veldman PhD, A Mazur MS, S Suchindran PhD, E L Tsalik MD, G S Ginsburg MD, C W Woods MD) **and Division of Infectious Diseases** (M T McClain, L B Jaggers MD, L P Park PhD, E L Tsalik, C W Woods)**, Duke University Medical Center, Durham, NC, USA; Durham VA Medical Center, Durham, NC, USA** (M T McClain, D C Jones MD, L P Park, E L Tsalik, C W Woods)**; and Institute for Medical Research, Durham, NC, USA** (B P Nicholson PhD)

Correspondence to: Dr Micah T McClain, Division of Infectious Diseases, Duke University Medical Center, Durham, NC 27710, USA **micah.mcclain@duke.edu**

Summary

Background Early and accurate identification of individuals with viral infections is crucial for clinical management and public health interventions. We aimed to assess the ability of transcriptomic biomarkers to identify naturally acquired respiratory viral infection before typical symptoms are present.

Methods In this index-cluster study, we prospectively recruited a cohort of undergraduate students (aged 18–25 years) at Duke University (Durham, NC, USA) over a period of 5 academic years. To identify index cases, we monitored students for the entire academic year, for the presence and severity of eight symptoms of respiratory tract infection using a daily web-based survey, with symptoms rated on a scale of 0–4. Index cases were defined as individuals who reported a 6-point increase in cumulative daily symptom score. Suspected index cases were visited by study staff to confirm the presence of reported symptoms of illness and to collect biospecimen samples. We then identified clusters of close contacts of index cases (ie, individuals who lived in close proximity to index cases, close friends, and partners) who were presumed to be at increased risk of developing symptomatic respiratory tract infection while under observation. We monitored each close contact for 5 days for symptoms and viral shedding and measured transcriptomic responses at each timepoint each day using a blood-based 36-gene RT-PCR assay.

Findings Between Sept 1, 2009, and April 10, 2015, we enrolled 1465 participants. Of 264 index cases with respiratory tract infection symptoms, 150 (57%) had a viral cause confirmed by RT-PCR. Of their 555 close contacts, 106 (19%) developed symptomatic respiratory tract infection with a proven viral cause during the observation window, of whom 60 (57%) had the same virus as their associated index case. Nine viruses were detected in total. The transcriptomic assay accurately predicted viral infection at the time of maximum symptom severity (mean area under the receiver operating characteristic curve [AUROC] 0·94 [95% CI 0·92–0·96]), as well as at 1 day (0·87 [95% CI 0·84–0·90]), 2 days (0·85 [0·82–0·88]), and 3 days (0·74 [0·71–0·77]) before peak illness, when symptoms were minimal or absent and 22 (62%) of 35 individuals, 25 (69%) of 36 individuals, and 24 (82%) of 29 individuals, respectively, had no detectable viral shedding.

Interpretation Transcriptional biomarkers accurately predict and diagnose infection across diverse viral causes and stages of disease and thus might prove useful for guiding the administration of early effective therapy, quarantine decisions, and other clinical and public health interventions in the setting of endemic and pandemic infectious diseases.

Funding US Defense Advanced Research Projects Agency.

Introduction

Acute viral infections are one of the most common reasons for visits to primary care physicians in high-income countries.[1] Annually, influenza affects 5–20% of the US population, results in more than 400 000 hospital admissions, and causes up to 61 000 deaths.[2,3] Outbreaks of viral infection continue to affect countries worldwide, including outbreaks of measles, global pandemics such as the 2009 pandemic influenza A H1N1 outbreak, and emergence of novel coronaviruses such as severe acute respiratory syndrome coronavirus (SARS-CoV) in 2003, Middle East respiratory syndrome coronavirus in 2012, and most recently severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[4–7] However, available molecular tools that can contribute to outbreak investigation or directing early clinical management are insufficient and scarce. The usefulness of traditional pathogen-focused diagnostic methods for viral infection (eg, culture, serology, antigen detection, and PCR) is limited by the fact they can be slow, costly, and restricted in terms of breadth of pathogens detected; can require a-priori knowledge of the pathogen being tested; might not detect emerging strains (eg, the 2009 influenza A H1N1 and current SARS-CoV-2 pandemics[8]); and are often incapable of

Research in context

Evidence before this study

We searched PubMed up to Feb 20, 2020, for studies using host gene expression for the diagnosis of acute viral infections using the search terms "viral infection" AND ("gene expression" OR "transcriptome" OR "transcriptomic") AND ("diagnosis" OR "classifier") without language restrictions. Classifiers for the diagnosis of acute viral infections or to differentiate acute viral infections from acute bacterial infections based on host-gene expression have been successfully developed by more than 12 groups globally, with classifier sizes ranging from a single gene to hundreds of genes. The largest prospectively validated transcriptomic classifier had 100% sensitivity (95% CI 85–100) and 96% specificity (89–100) for viral infection in 370 children at the time of severe illness. However, no studies of naturally acquired viral infection have investigated the use of such technologies in the context of presymptomatic diagnosis or outbreak investigation.

Added value of this study

To our knowledge, this study is the first to show in a real-world setting that a blood-based host-gene expression assay can accurately predict respiratory viral infection before typical symptoms are present. Our data show that transcriptomic biomarkers of viral infection are present and detectable before clinical disease develops and thus could form the basis of novel approaches to early treatment and management of emerging viral outbreaks and pandemics.

Implications of all the available evidence

The findings of research describing the use of host transcriptomic classifiers for the diagnosis of acute infections are robust and rapidly expanding. With the addition of our data, the literature supports the potential use of such technologies for a wide spectrum of viral illnesses from an early, presymptomatic stage to clinical presentation and eventual disease resolution. When combined with emerging nucleic acid detection platforms that offer sample-to-answer times of less than 1 h, these approaches offer the potential to transform clinical approaches for the diagnosis of viral diseases.

distinguishing between active infection and colonisation. Therefore, there is increasing interest in using host-derived biomarkers to ascertain the presence or type of infection in at-risk hosts, including single-analyte biomarkers such as procalcitonin and composite, multiplex biomarker panels, which have been included in an increasing number of gene expression-based studies.[9–22]

Previous research on diagnostics for naturally acquired infection has focused on identifying symptomatic individuals at the time of clinical presentation for medical care,[20] which is often late in the time course of many viral infections. Identification of infectious causes in earlier, presymptomatic phases of illness provides an opportunity to optimise and deliver timely, and thus more effective, therapy, refine prophylaxis decisions, and guide public health interventions such as isolation and quarantine.[14,23] However, identifying individuals who have been exposed to infectious viruses but do not yet have symptoms of clinical illness is logistically challenging, reflected by the paucity of available data and tools to address these problems.

We have previously shown that during experimental influenza infection of healthy volunteers, the host genomic response is robust and detectable before typical symptoms become apparent.[17,23,24] Much of the genomic response seems to be driven by early innate responses at the site of infection that drive signalling cascades, resulting in the expression of interferon-response genes in peripheral blood leucocytes.[17,25] Other studies have confirmed that these transcriptomic markers in blood can be indicative of early viral infection in experimental human challenge models and animal models of disease.[26–29] However, challenge infections are contrived rather than natural, typically use laboratory-adapted viral strains, and result in a clinical illness that does not entirely resemble that seen in naturally acquired infections, all of which can limit the broad applicability of findings from human-challenge studies.[17,23,30–33] To date, the nature of the host response to infection during the presymptomatic phase of naturally occurring disease has not been defined. We aimed to assess the ability of transcriptomic biomarkers to identify naturally acquired respiratory viral infection before typical symptoms are present.

Methods

Study design and participants

We did an index-cluster, prospective cohort study in undergraduate students (aged 18–25 years) at Duke University (Durham, NC, USA), most of whom lived in dormitories on campus, over a period of 5 academic years (appendix p 10). Information leaflets inviting students to enrol were distributed around the university campus and students were approached and recruited by study staff during onboarding activities at the start of each academic year. At the time of enrolment, we collected baseline biospecimens and asked students to complete standardised questionnaires to ascertain demographic, behavioural, and other medical characteristics. After enrolment, participating students completed a web-based survey once a day for the duration of the academic year (September to May, excluding holidays) to describe the presence and severity of eight symptoms (cough, fever, headache, malaise, nasal congestion, nasal discharge, sneezing, and sore throat) of respiratory tract infection (appendix p 2).

See Online for appendix

Each symptom was scored on a scale of 0–4, with 0 indicating not present, 1 indicating mild symptoms, 2 indicating moderate symptoms, 3 indicating severe symptoms, and 4 indicating very severe symptoms. Study staff monitored the surveys daily for a 6-point increase in an individual's cumulative daily symptom score, which triggered an email notification or phone call and site visit by study staff for sample and data collection from a potential index case[9,17,32] (appendix p 2). Study staff then assessed suspected index cases to confirm the presence of reported symptoms of illness and collected biospecimens. On the basis of these sentinel cases of respiratory tract infection with suspected viral cause, we identified a prospective observational cohort of asymptomatic but potentially exposed close contacts (ie, people living in the same dormitory or identified by the index case as a close contact), who were then asked to complete the daily web-based survey to monitor the development of any symptoms and sampled by study staff who collected biospecimen samples at their place of residence or on campus for up to 5 days (defined as the observation window). This design permitted enrichment for, and collection of, samples from close contacts during the timeframe between exposure (which occurred at an unknown time) and subsequent development of symptomatic disease (which occurred under observation). The study was approved by the Institutional Review Board of Duke University in accordance with the Declaration of Helsinki. All participants provided written informed consent.

Sample and data collection

Blood (20 mL) and nasopharyngeal swab samples were collected daily by study staff from confirmed index cases at the time of illness identification. The nasopharyngeal samples were tested for the presence of viruses using commercial multiplex PCR assays (ResPlex II Panel v2.0 [Qiagen, Hilden, Germany], xTAG respiratory viral panel [Luminex, Austin, TX, USA], or Biofire FilmArray Respiratory Panel [BioFire Diagnostics, Salt Lake City, UT, USA]; appendix p 3). Asymptomatic close contacts identified by index cases provided blood and nasopharyngeal samples for up to 5 consecutive days and were monitored for symptomatic conversion (indicated by a 6-point increase in cumulative daily symptom score; appendix p 2) and viral shedding using multiplex PCR assays.

For detection of transcriptomic responses, we collected peripheral blood in PAXgene Blood RNA tubes (PreAnalytiX, Qiagen, Hilden, Germany), and we extracted total RNA using the PAXgene Blood miRNA Kit (Qiagen). cDNA was synthesised using SuperScript VILO Master Mix (Invitrogen, Carlsbad, CA, USA), according to the manufacturer's instructions. We did RT-PCR using custom-built TaqMan Low Density Array 384-well microfluidic cards with TaqMan Gene Expression Master Mix (Applied Biosystems, Foster City, CA, USA), run on a ViiA7 Real-Time PCR System (Applied Bio systems). The genes included in the signature are shown in the appendix (p 9).

RT-PCR

We previously developed a host response-derived transcriptomic signature of symptomatic viral infection using microarray data.[32] Since microarray and bulk RNA sequencing platforms are impractical for rapid clinical testing, we migrated our previous so-called pan-viral signature to an RT-PCR platform with potential for more direct translation to point-of-care testing.[9,10] Using sparsity-imposing techniques (appendix p 3), we selected 36 pre-designed TaqMan probes representing genes comprising the acute respiratory viral signature (and normalisation controls) to be used on a TaqMan Low Density Array platform.[9] We then used a regularised (2-norm) logistic regression model to determine coefficients for each gene in the signature,[34] such that the output of the model is a probability that a given individual will go on to have a symptomatic, viral PCR-positive event during the observation window. The performance of the RT-PCR signature for classification of symptomatic individuals was assessed using a repeated random

	2009–10 (n=70)	2010–11 (n=104)	2011–12 (n=57)	2013–14* (n=12)	2014–15 (n=62)	Overall (n=305)
Adenovirus	1 (1%)	1 (1%)	0	1 (8%)	2 (3%)	5 (2%)
Bocavirus	0	0	0	0	2 (3%)	2 (1%)
Coronavirus	11 (16%)	23 (22%)	8 (14%)	3 (25%)	9 (15%)	54 (18%)
Coxsackie or echovirus†	21 (30%)	38 (37%)	12 (21%)	0	0	71 (23%)
Human metapneumovirus	2 (3%)	1 (1%)	3 (5%)	4 (33%)	1 (2%)	11 (4%)
Influenza	9 (13%)	7 (7%)	1 (2%)	1 (8%)	1 (2%)	19 (6%)
Parainfluenza virus	3 (4%)	10 (10%)	1 (2%)	0	7 (11%)	21 (7%)
Respiratory syncytial virus	3 (4%)	1 (1%)	3 (5%)	0	3 (5%)	10 (3%)
Rhinovirus or enterovirus	20 (29%)	23 (22%)	29 (515)	3 (25%)	37 (60%)	112 (37%)

Data are number of times the virus was detected (% of total viral infections per year). Data are shown for 150 index cases, 106 close contacts, and 27 asymptomatic shedders who had a respiratory viral infection confirmed by multiplex PCR; 22 participants tested positive for more than one virus (ie, they had co-infections). *Truncated academic year with spring enrolment only. †Not detectable by 2013–15 testing assays.

Table 1: **Viral causes of respiratory tract infection in the student cohort by academic enrolment year**

	Cough	Fever	Headache	Malaise	Nasal congestion	Nasal discharge	Sneezing	Sore throat	Total symptom score
Influenza	3·2 (1·5)	1·9 (1·0)	2·4 (1·3)	3·0 (0·8)	2·5 (0·8)	2·4 (1·2)	1·8 (1·1)	2·7 (1·1)	20·0 (6·2)
Respiratory syncytial virus	2·1 (1·1)	0·7 (0·7)	1·6 (0·9)	2·9 (0·7)	2·6 (0·4)	2·6 (0·9)	1·5 (0·5)	2·6 (1·0)	16·6 (6·4)
Human metapneumovirus	2·6 (1·1)	0·8 (0·7)	1·5 (1·0)	2·5 (0·9)	2·4 (0·7)	2·5 (1·1)	1·9 (0·9)	2·2 (0·9)	16·3 (4·0)
Rhinovirus or enterovirus	2·2 (1·1)	0·5 (0·4)	1·1 (1·0)	2·1 (1·1)	2·3 (0·9)	2·3 (0·9)	1·5 (1·0)	2·1 (1·1)	14·0 (6·1)
Coxsackie or echovirus	2·1 (1·2)	0·5 (0·4)	1·1 (1·1)	2·2 (1·1)	2·4 (1·1)	2·4 (1·1)	1·6 (0·9)	1·8 (1·1)	14·0 (6·8)
Coronavirus	1·8 (1·0)	0·7 (0·6)	1·0 (0·9)	2·1 (1·5)	2·3 (0·9)	2·5 (0·8)	1·6 (0·4)	1·8 (0·8)	13·8 (5·3)
Parainfluenza virus	2·3 (1·1)	0·5 (0·7)	1·2 (1·0)	2·3 (0·9)	1·8 (0·8)	1·7 (0·7)	1·3 (0·9)	2·3 (1·0)	13·4 (4·9)
Adenovirus	2·2 (0·8)	1·5 (0·9)	1·4 (0·9)	2·2 (0·9)	1·6 (0·8)	1·6 (1·0)	0·2 (0·1)	2·2 (1·1)	12·9 (6·2)
Bocavirus	1·0 (0·8)	0	0	1·0 (0·7)	2·0 (1·1)	2·0 (1·1)	1·0 (0·7)	1·0 (0·7)	8·0*

Data are mean symptom score (SD). *Calculation of SD was not possible because only two samples were positive for bocavirus.

Table 2: **Clinical characteristics of viral infection at time of maximum symptom severity for index cases and their close contacts**

subsampling validation strategy on relevant participants across all study years.

Statistical analysis

For RT-PCR data the cycle of quantification values were transformed into normalised relative quantities using reference genes (ie, those with the most stable combined coefficient of variation across all samples) and assays were run in two experimental batches: *FPGS* and *TRAP1* were used for the first batch, and *DECR1* and *TRAP1* were used for the second batch. Data were normalised as previously described.[9,10] We used a regularised (2-norm) logistic regression model to discriminate symptomatic shedders (cases) from asymptomatic nonshedders (controls; appendix pp 2–5). The data were split into training and testing sets 25 times using repeated random subsampling, and each of the 25 models was trained using data from all timepoints. We assessed the performance of the model at each individual timepoint using receiver operating characteristic (ROC) curves and calculation of area under the ROC curves (AUROCs). All model analyses were implemented in Python (version 3.7.4) and Tensorflow (version 1.14.0).[35] Accuracies were calculated using the point on the ROC curve that maximises the Youden index,[36] while other comparisons of model performance (eg, the true positive rate) were calculated by selecting operating points along the ROC curves that correspond to a threshold or fixed value (eg, fixing a desired false positive rate; appendix p 5).

The Mann-Whitney *U* test was used for the comparison of means and Spearman's rank correlation coefficient was used to describe the association between variables where indicated. Z scores were calculated using normalised, log-transformed relative gene expression values, as previously described.[9]

Since the exact time of exposure or transmission during naturally acquired infection is unknown, for the purpose of analysis, close contact days were aligned to the time of maximum symptom severity (ie, day of highest symptom score; time [T]) for each individual in an attempt to place individuals on broadly similar time scales in terms of the host response to infection. For each individual close contact, 1 day after peak symptoms was defined as T1, 1 day before the highest symptom score was defined as T minus 1 (T–1), 2 days before as T minus 2 (T–2), 3 days before as T minus 3 (T–3), and 4 days before as T minus 4 (T–4).

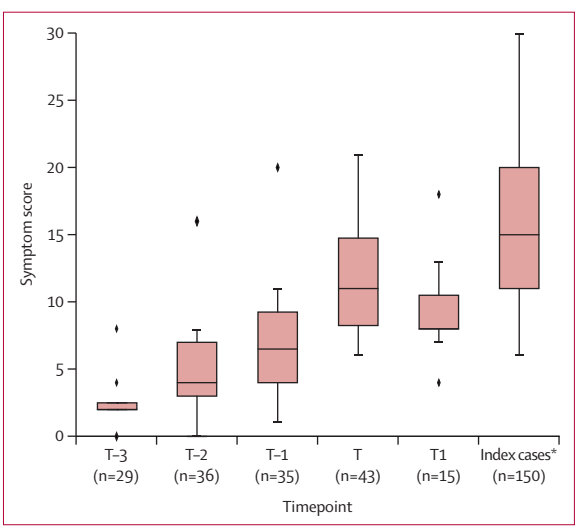

Figure 1: **Median symptom scores of index cases and close contacts**
Symptom scores of index cases and close contacts who developed PCR-proven viral respiratory tract infection under observation. For close contacts, T represents the day of maximum symptom severity, T1 represents the day after maximum symptom severity, T–1 represents 1 day before maximum illness, T–2 represents 2 days before maximum illness, and T–3 represents 3 days before maximum illness. Box and whisker plots show medians (lines) and IQRs (boxes); upper and lower whiskers indicate 1·5 × IQR and diamonds indicate outliers. *Index cases were only sampled at one timepoint.

Role of the funding source

The funder of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript. The corresponding author

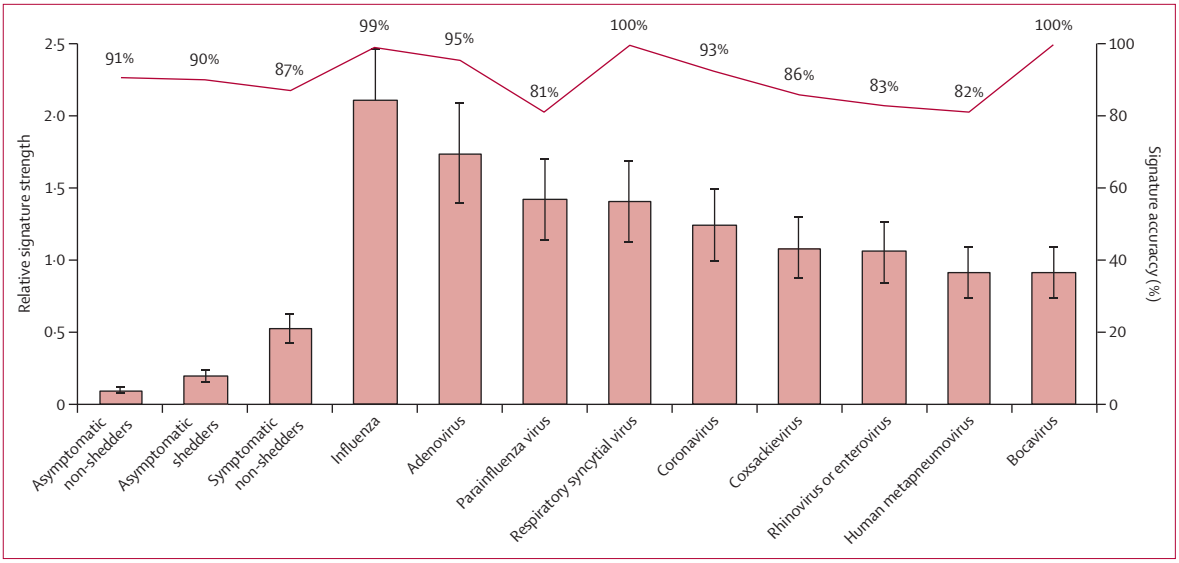

Figure 2: **Variation in genetic signature strength and accuracy across viral infections and control groups**
Quantitative transcriptomic signature strength (bars) and accuracy (red line) at the timepoint of maximum symptom severity for each type of viral infection and for asymptomatic non-shedders, asymptomatic shedders, and symptomatic non-shedders. For asymptomatic non-shedders, the timepoint T was unclear due to an absence of symptoms, thus timepoints were chosen at random from the observation window, or the date of shedding for shedders was used. Vertical black lines on bars indicate 95% CIs.

had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Between Sept 1, 2009, and April 10, 2015, we enrolled 1465 participants. Biospecimens were collected from 264 index cases with clinical illness, of whom 150 had a respiratory viral cause confirmed by multiplex PCR testing of nasopharyngeal samples. Of the 555 close contacts enrolled and sampled, 162 developed symptoms of respiratory tract infection during the observation window, of whom 106 had concomitant confirmatory viral PCR. 60 (57%) of 106 close contacts were infected with the same virus as their associated index case.

Overall, nine different viral pathogens were identified, and viral causes varied substantially across academic years (table 1). Co-infection with multiple viruses in a single individual was uncommon (22 [7%] of 305 participants), and most of these individuals (16 [73%] of 22 participants) had a combination of coxsackievirus or echovirus and rhinovirus or enterovirus infection, which could represent cross-reactivity. Participants with PCR-confirmed viral infection had many common symptoms of upper respiratory tract infection at the time of peak symptom severity, with clinical variation among viruses (table 2).

For all 106 close contacts with confirmed viral infection, at the time of maximum symptom severity (T), the median total symptom score was 11·0 (IQR 8·25–14·75; figure 1), although symptoms varied by virus (table 2). The median total symptom score for symptomatic shedders was 6·5 (IQR 4·00–9·25) at T–1, 4·0 (3·00–7·00) at T–2, and 2·0 (2·00–2·50) at T–3. A daily symptom score of 6 represented mild clinical illness and was the threshold for defining a symptomatic day. At the T–2 timepoint and earlier, most close contacts had subclinical illness or were asymptomatic (ie, they had a total symptom score of less than 6; 26 [90%] of 29 participants were asymptomatic at T–3 and 28 [78%] of 36 participants at T–2).

The RT-PCR-based host-gene expression assay distinguished symptomatic, virus-infected individuals (n=183; index cases and close contacts) from asymptomatic, viral PCR-negative controls (n=152 samples from 35 individuals) with 90% accuracy. However, mean accuracy varied depending on the virus, ranging from 100% accuracy for respiratory syncytial virus and bocavirus infections, to 99% for influenza infection, 93% for coronavirus infection, and 82% for human metapneumovirus infection (figure 2).

Analysis of individual gene components of the model showed that many individual genes had similar AUROC values to the aggregate model across all iterations (appendix pp 11,12). 18 (50%) of 36 individual gene targets had mean AUROC values of higher than 0·90 at time T; the three top performing genes at time T were *IFIT3* (AUROC 0·96 [95% CI 0·95–0·97]), *HERC5* (0·95 [0·94–0·96]), and *RSAD2* (0·96 [0·95–0·97]). The ten top performing genes are shown in the appendix (p 11).

The RT-PCR gene signature correctly identified symptomatic individuals with 92% accuracy on the day of maximal symptoms (appendix p 16). Accuracy of the gene signature decreased to 88% at the T–1 timepoint and to 83% at the T–2 timepoint (appendix p 16). Although accuracy of the assay varied between viruses at

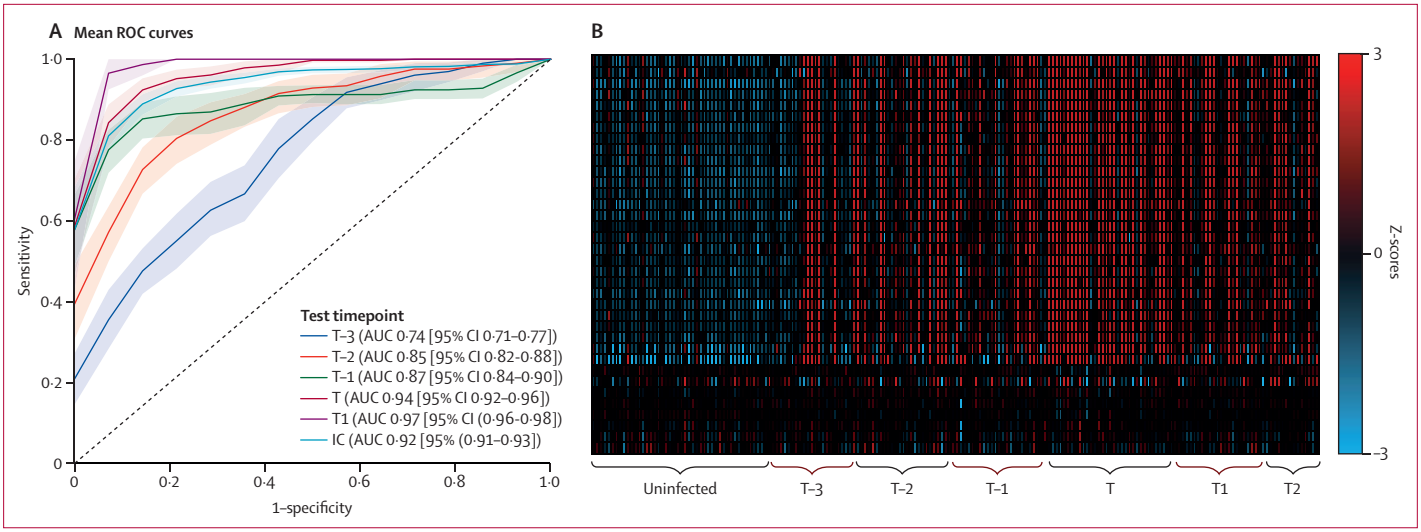

Figure 3: **RT-PCR-based 36-gene signature performance over time**

(A) Ability of the model based on the 36-gene RT-PCR assay to detect naturally acquired respiratory viral infection at various stages of infection in index cases and close contacts. (B) Heatmap of Z scores for each gene in the RT-PCR assay at each timepoint in close contacts with proven symptomatic viral infection or healthy controls, where T represents the day of maximum symptom severity for each participant. T1 represents the day after maximum symptom severity, T–1 represents 1 day before maximum illness, T–2 represents 2 days before maximum illness, and T–3 represents 3 days before maximum illness. ROC=receiver operating characteristic. AUC=area under the curve. IC=index case.

time T (figure 2), no significant differences in performance were identified between viruses at earlier timepoints (appendix p 5). For example, only one individual with influenza infection had presymptomatic samples available for all timepoints, although the signature correctly classified that individual at each timepoint (T–1, T–2, T–3, and T–4).

To investigate how signature strength correlates with overall symptom severity, we assessed the correlation between model-derived probabilities of infection and symptom scores at each individual timepoint (appendix p 5). At each timepoint, gene expression levels were similar across symptomatic individuals despite variable symptom severity. Thus, gene signature performance varied by timepoint (figure 3) but was not affected by symptom severity within each timepoint (Spearman's rank correlation coefficient at time T, ρ=0·06 [appendix p 13]). When comparing gene signature performance with clinical diagnosis (symptom-based), the model outperformed clinical diagnosis at all timepoints with the exception of timepoint T and for index cases (where symptom-based diagnosis by definition has 100% sensitivity; appendix p 14). 24 (82%) of 29 individuals who developed symptomatic respiratory tract infection had no detectable viral shedding at T–3, and 25 (69%) of 36 individuals were negative for any virus by routine clinical testing at time T–2.

Asymptomatic shedders (n=27) had significantly higher gene expression than asymptomatic non-shedders (n=35; p<0·0001, Mann-Whitney U test), but markedly lower levels of signature expression than symptomatic shedders (p<0·0001; figure 2). Many symptomatic non-shedders had predicted probabilities of viral infection that mirrored those of either symptomatic shedders or asymptomatic non-shedders (data not shown). Overall, the viral signature in the 199 symptomatic non-shedders was significantly higher than in asymptomatic non-shedders (p<0·0001) but significantly lower than in symptomatic shedders (p=0·04).

Discussion

In this study, we have used an innovative index-cluster study design with a focus on serial sampling of real-world, close contacts of infected individuals to enrich for cases of naturally occurring infection during these early, post-exposure but presymptomatic timepoints. This unique design has permitted real-world validation of a transcriptomic signature in peripheral blood that is capable of accurately identifying exposed but apparently healthy individuals who will go on to develop symptomatic viral infection. In the majority of participants (62%), the gene signature was present even before viral shedding was detected. The promise of this approach is highlighted by the discriminative ability of the genomic signature 2–3 days before maximum illness, when most individuals had minimal or no symptoms (90% and 78% of participants were asymptomatic or had subclinical illness at the T–3 and T–2 timepoints, respectively). Furthermore, mild symptoms at early timepoints were clinically vague and typical of seasonal allergies, mild chronic obstructive pulmonary disorder flares, or even symptoms due to sequelae of chronic smoking.[37] Thus, genomic analysis to classify viral infection among asymptomatic individuals or those with common, non-specific upper respiratory symptoms would be valuable. The accuracy of the transcriptomic signature across nine different respiratory viruses, each with specific incubation times and variable

clinical progression and duration, also highlights the potential of this approach. Additionally, the gene signature used represents conserved antiviral response signalling pathways that are active, and thus discriminatory, across the spectrum of illness, not only at early timepoints. This broad applicability is vital because in a real-world setting, the location of an individual along the continuum of infection at the moment of testing would be unknown.

Although our transcriptomic signature accurately classified a wide spectrum of viral respiratory tract infections, it had the highest accuracy in a subset of viruses, including influenza and coronavirus. Influenza is one of the most important viruses to identify early because of its contribution to pandemics, in which existing diagnostics might perform poorly (especially for emerging strains as observed in 2009[38]), questions of triage and quarantine are paramount, and effective therapeutic options are readily available and most effective when provided early in disease. Similarly, another global outbreak of a novel, highly virulent coronavirus—SARS-CoV-2—is ongoing, and it is evident that tools that can aid in assessment of individuals with potential exposure, especially tools that can diagnose infection before detectable viral shedding, would be a valuable to existing epidemiological and molecular approaches in outbreak settings. Our experience suggests that the ability to accurately detect exposure to these types of emerging infections will only become increasingly important over time.

A number of other studies have described interferon response-based transcriptomic signatures of acute respiratory infections that vary by target population, pathogens, modelling approaches, and the number of genes required to discriminate disease states.[10,11,14–19,32] Some of these findings have also been validated, combined, or expanded upon by secondary ex-vivo analysis of publicly available data.[12,13,39] Although the signature described in this study clearly represents a limited subset of all discriminating genes, some observations regarding simplified, interferon-driven signatures can be made. Our data confirm that that some genes identified by previous studies (eg, *IFI44L* and *IFI27*[11,40]) perform well as classifiers to distinguish between individuals with viral illness and those who are healthy (mean AUROC for *IFI44L* and *IFI27* at timepoint T were 0·92 [95% CI 0·90–0·94] and 0·79 [0·76–0·82], respectively [appendix p 12]), with strong upregulation observed in cases of viral infection. Additionally, we show that several other genes also perform well as classifiers, including *RSAD2* (AUROC 0·96 [95% CI 0·95–0·97]), *IFIT3* (0·96 [0·95–0·97]), and *IFI44* (0·93 [0·91–0·95]; appendix p 12). On the basis of our experience and the published literature, it seems that some diagnostic classification tasks, in which the biological responses are highly conserved, unique in character, and of great magnitude, might be reasonably addressed with pauci-analyte biomarker panels.[11,13,41–43] However, as population complexity increases by adding greater interhost variability (such as immunosuppression), discriminating clinically similar groups with more varied underlying pathophysiology, or discriminating multiple clinical mimics at once, more intricate biomarker panels will probably be required to maintain high fidelity in real-world settings.[10,15,44]

Our study has several limitations. First, this study was designed to validate the predictive power of a small set of genes that already had proven discriminatory capability during maximal illness, and thus was not designed to ascertain whether additional, untested gene sets could offer improved discriminatory performance at early timepoints in the course of illness. Studies using RNA sequencing on whole blood, single-cell sequencing, or other methods are needed to more completely define early genomic changes in the naturally infected host. Second, our patient cohort was comprised of otherwise young, healthy individuals with community-acquired viral respiratory tract infection of mild to moderate severity. Therefore, application of our findings to patient populations from other age groups, people with immunological comorbidities, or people with more severe viral infections (such as SARS-CoV-2) will need to be assessed. Finally, we did not identify any confirmed infections due to bacterial or other pathogens. As a result, although these transcriptomic markers show great promise for differentiating early infection and pre-symptomatic viral exposure from non-infectious states, we did not directly validate the ability of these markers to differentiate viral exposure from early forms of other types of infection.[10,11,15,45] As the clinical use of transcriptomic biomarkers continues to be assessed, consideration of broad applicability and performance across a wide array of clinical settings will continue to be paramount.

To the best of our knowledge, this is the first study to define the presymptomatic and temporal dynamics of transcriptomic biomarkers characterising the host response to naturally acquired viral infections in humans. Clinic-ready platforms capable of operationalising PCR-based signatures of the size tested herein already exist, some of which are approved with point-of-care functionality, offering a proximal pathway to clinical application of these findings.[46] Thus, analysis of the evolution of gene expression-based biomarkers over time shows promise for driving development of diagnostics for early detection of viral exposure and infection that could prove useful for guiding administration of early effective therapy, quarantine decisions, and other clinical and public health interventions in the setting of endemic and pandemic infectious disease.

Contributors

MTM, BPN, LBJ, TWB, TV, GSG, and CWW helped conceive and implement the study. All authors helped acquire, analyse, or interpret data. Statistical analyses were done by FJC, RH, SS, MN, and LPP. MTM, ELT, FJC, GSG, and CWW drafted the manuscript, which was critically revised by all remaining authors.

Declaration of interests

MTM reports grants from the Defense Advanced Research Projects Agency (DARPA), National Institutes of Health (NIH); and has patents

pending on Methods to Diagnose and Treat Acute Respiratory Infections. TWB reports grants from DARPA; consultancy fees from Predigen; and has a patent pending on Methods to Diagnose and Treat Acute Respiratory Infections. ELT reports grants from DARPA, NIH/Antibacterial Resistance Leadership Group (ARLG); consultancy fees from bioMerieux; is a cofounder of Predigen; and has patents pending on Biomarkers for the Molecular Classification of Bacterial Infection and Methods to Diagnose and Treat Acute Respiratory Infections. GSG is a cofounder of Predigen; has patents pending on Molecular Classification of Bacterial Infection and Gene Expression Signatures Useful to Predict or Diagnose Sepsis and Methods of Using the Same, and has patents issued on Methods to Diagnose and Treat Acute Respiratory Disease, and Methods of Identifying Infectious Disease and Assays for Identifying Infectious Disease. CWW reports grants from DARPA, NIH/ARLG, Predigen, and Sanofi; received consultancy fees from bioMerieux, Roche, Biofire, Giner, and Biomeme; and has patents pending on Molecular Biomarkers of Acute Infection. All other authors declare no competing interests.

Acknowledgments

We thank Debra Freeman, Sara Hoffman, Olga M Better, Mert Aydin, Stephanie Dobos, Kyle R Breitschwert, and Ashlee M Valente for enrolling participants, processing samples and data, and engaging with the undergraduate campus. This study was funded by the US Defense Advanced Research Projects Agency (projects N66001-07-C-2024 and W911NF1410052). The statements and conclusions herein are those of the authors and do not reflect the position or policy of the US Government.

References

1 Hong CY, Lin RT, Tan ES, et al. Acute respiratory symptoms in adults in general practice. *Fam Pract* 2004; **21:** 317–23.
2 Shrestha SS, Swerdlow DL, Borse RH, et al. Estimating the burden of 2009 pandemic influenza A (H1N1) in the United States (April 2009–April 2010). *Clin Infect Dis* 2011; **52** (suppl 1): S75–82.
3 Garten R, Elal AIA, Alabi N, et al. Update: influenza activity in the United States during the 2017–18 season and composition of the 2018–19 influenza vaccine. *MMWR Morb Mortal Wkly Rep* 2018; **67:** 634–42.
4 de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. *Nat Rev Microbiol* 2016; **14:** 523–34.
5 Arabi YM, Balkhy HH, Hayden FG, et al. Middle East Respiratory Syndrome. *N Engl J Med* 2017; **376:** 584–94.
6 Angelo KM, Gastañaduy PA, Walker AT, et al. Spread of measles in Europe and implications for US travelers. *Pediatrics* 2019; **144:** e20190414.
7 Perlman S. Another decade, another coronavirus. *N Engl J Med* 2020; **382:** 760–62.
8 Dawood FS, Jain S, Finelli L, et al. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med* 2009; **360:** 2605–15.
9 Zaas AK, Burke T, Chen M, et al. A host-based RT-PCR gene expression signature to identify acute respiratory viral infection. *Sci Transl Med* 2013; **5:** 203ra126.
10 Tsalik EL, Henao R, Nichols M, et al. Host gene expression classifiers diagnose acute respiratory illness etiology. *Sci Transl Med* 2016; **8:** 322ra11.
11 Herberg JA, Kaforou M, Wright VJ, et al. Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial *vs* viral infection in febrile children. *JAMA* 2016; **316:** 835–45.
12 Andres-Terre M, McGuire HM, Pouliot Y, et al. Integrated, multi-cohort analysis identifies conserved transcriptional signatures across multiple respiratory viruses. *Immunity* 2015; **43:** 1199–211.
13 Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. *Lancet Respir Med* 2016; **4:** 213–24.
14 Liu X, Speranza E, Muñoz-Fontela C, et al. Transcriptomic signatures differentiate survival from fatal outcomes in humans infected with Ebola virus. *Genome Biol* 2017; **18:** 4.
15 Suarez NM, Bunsow E, Falsey AR, Walsh EE, Mejias A, Ramilo O. Superiority of transcriptional profiling over procalcitonin for distinguishing bacterial from viral lower respiratory tract infections in hospitalized adults. *J Infect Dis* 2015; **212:** 213–22.
16 Kaforou M, Wright VJ, Oni T, et al. Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLoS Med* 2013; **10:** e1001538.
17 Woods CW, McClain MT, Chen M, et al. A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2. *PLoS One* 2013; **8:** e52198.
18 Mejias A, Dimo B, Suarez NM, et al. Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. *PLoS Med* 2013; **10:** e1001549.
19 Hu X, Yu J, Crosby SD, Storch GA. Gene expression profiles in febrile children with defined viral and bacterial infection. *Proc Natl Acad Sci USA* 2013; **110:** 12792–97.
20 Zhai Y, Franco LM, Atmar RL, et al. Host transcriptional response to influenza and other acute respiratory viral infections—a prospective cohort study. *PLoS Pathog* 2015; **11:** e1004869.
21 Roe J, Venturini C, Gupta RK, et al. Blood transcriptomic stratification of short-term risk in contacts of tuberculosis. *Clin Infect Dis* 2020; **70:** 731–37.
22 Afroz S, Giddaluru J, Abbas MM, Khan N. Transcriptome meta-analysis reveals a dysregulation in extra cellular matrix and cell junction associated gene signatures during dengue virus infection. *Sci Rep* 2016; **6:** 33752.
23 McClain MT, Nicholson BP, Park LP, et al. A genomic signature of influenza infection shows potential for presymptomatic detection, guiding early therapy, and monitoring clinical responses. *Open Forum Infect Dis* 2016; **3:** ofw007.
24 Huang Y, Zaas AK, Rao A, et al. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza A infection. *PLoS Genet* 2011; **7:** e1002234.
25 Iwasaki A, Pillai PS. Innate immunity to influenza virus infection. *Nat Rev Immunol* 2014; **14:** 315–28.
26 Shurtleff AC, Whitehouse CA, Ward MD, Cazares LH, Bavari S. Pre-symptomatic diagnosis and treatment of filovirus diseases. *Front Microbiol* 2015; **6:** 108.
27 Caballero IS, Yen JY, Hensley LE, Honko AN, Goff AJ, Connor JH. Lassa and Marburg viruses elicit distinct host transcriptional responses early after infection. *BMC Genomics* 2014; **15:** 960.
28 Duy J, Koehler JW, Honko AN, et al. Circulating microRNA profiles of Ebola virus infection. *Sci Rep* 2016; **6:** 24496.
29 Wang K, Langevin S, O'Hern CS, et al. Anomaly detection in host signaling pathways for the early prognosis of acute infection. *PLoS One* 2016; **11:** e0160919.
30 Memoli MJ, Czajkowski L, Reed S, et al. Validation of the wild-type influenza A human challenge model H1N1pdMIST: an A(H1N1) pdm09 dose-finding investigational new drug study. *Clin Infect Dis* 2015; **60:** 693–702.
31 Han A, Poon JL, Powers JH 3rd, Leidy NK, Yu R, Memoli MJ. Using the Influenza Patient-reported Outcome (FLU-PRO) diary to evaluate symptoms of influenza viral infection in a healthy human challenge model. *BMC Infect Dis* 2018; **18:** 353.
32 Zaas AK, Chen M, Varkey J, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe* 2009; **6:** 207–17.
33 Sobel Leonard A, McClain MT, Smith GJ, et al. Deep sequencing of influenza A virus from a human challenge study reveals a selective bottleneck and only limited intrahost genetic diversification. *J Virol* 2016; **90:** 11247–58.
34 Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biom J* 2010; **52:** 70–84.
35 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33:** 1–22.
36 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3:** 32–35.
37 Stanton N, Hood K, Kelly MJ, et al. Are smokers with acute cough in primary care prescribed antibiotics more often, and to what benefit? An observational study in 13 European countries. *Eur Respir J* 2010; **35:** 761–67.
38 LeBlanc JJ, Li Y, Bastien N, Forward KR, Davidson RJ, Hatchette TF. Switching gears for an influenza pandemic: validation of a duplex reverse transcriptase PCR assay for simultaneous detection and confirmatory identification of pandemic (H1N1) 2009 influenza virus. *J Clin Microbiol* 2009; **47:** 3805–13.

39 Bongen E, Vallania F, Utz PJ, Khatri P. KLRD1-expressing natural killer cells predict influenza susceptibility. *Genome Med* 2018; **10:** 45.

40 Tang BM, McLean AS, Dawes IW, Huang SJ, Lin RC. Gene-expression profiling of peripheral blood mononuclear cells in sepsis. *Crit Care Med* 2009; **37:** 882–88.

41 Warsinske HC, Rao AM, Moreira FMF, et al. Assessment of validity of a blood-based 3-gene signature score for progression and diagnosis of tuberculosis, disease severity, and treatment response. *JAMA Netw Open* 2018; **1:** e183779.

42 Tang BM, Shojaei M, Parnell GP, et al. A novel immune biomarker *IF127* discriminates between influenza and bacteria in patients with suspected respiratory infection. *Eur Respir J* 2017; **49:** 1602098.

43 Herberg JA, Kaforou M, Wright VJ, et al. Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. *JAMA* 2016; **316:** 835–45.

44 Maslove DM, Wong HR. Gene expression profiling in sepsis: timing, tissue, and translational considerations. *Trends Mol Med* 2014; **20:** 204–13.

45 Ramilo O, Allman W, Chung W, et al. Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* 2007; **109:** 2066–77.

46 Tsalik EL, Henao R, Aydin M, et al. 2012. FilmArray measurement of host response signatures rapidly discriminates viral, bacterial, and non-infectious etiologies of illness. *Open Forum Infect Dis* 2018; **5** (suppl 1)**:** S586.