

SCIENTIFIC REPORTS



OPEN

A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms

R. Barral-Arca^{1,2,3,4}, J. Pardo-Seco^{1,2,3,4}, F. Martín-Torres^{3,4} & A. Salas^{1,2,3,4}

Recently, a biomarker signature consisting of 2-transcript host RNAs was proposed for discriminating bacterial from viral infections in febrile children. We evaluated the performance of this signature in a different disease scenario, namely a cohort of Mexican children ($n = 174$) suffering from acute diarrhea of different infectious etiologies. We first examined the admixed background of the patients, indicating that most of them have a predominantly Native American genetic ancestry with a variable amount of European background (ranging from 0% to 57%). The results confirm that the RNA test can discriminate between viral and bacterial causes of infection (t-test; P -value = 6.94×10^{-11} ; AUC = 80%; sensitivity: 68% [95% CI: 55%–79%]; specificity: 84% [95% CI: 78%–90%]), but the strength of the signal differs substantially depending on the causal pathogen, with the stronger signal being that of *Shigella* (P -value = 3.14×10^{-12} ; AUC = 89; sensitivity: 70% [95% CI: 57%–83%]; specificity: 100% [95% CI: 100%–100%]). The accuracy of this test improves significantly when excluding mild cases (P -value = 2.13×10^{-6} ; AUC = 85%; sensitivity: 79% [95% CI: 58%–95%]; specificity: 78% [95% CI: 65%–88%]). The results broaden the scope of previous studies by incorporating different pathogens, variable levels of disease severity, and different ancestral background of patients, and add confirmatory support to the clinical utility of these 2-transcript biomarkers.

There are increasing efforts aimed at developing host biomarkers allowing to distinguish viral from bacterial infections in febrile children¹. The interest stems not only from the need to discriminate potentially life-threatening bacterial infections from viral infections, but also to avoid the unnecessary prescription of empirical antibiotic therapy irrespective of severity. The overuse of antibiotics worldwide is accelerating antimicrobial resistance (AMR), which is recognized as one of the greatest threats to human health worldwide². Furthermore, AMR research now constitutes one of the main programs of the World Health Organization (WHO; <http://www.who.int/antimicrobial-resistance/>).

There is an increasing interest in discovering diagnostic biomarkers from whole blood gene expression able to distinguish viral from bacterial diseases showing similar initial clinical phenotypes³. Two promising approaches to this issue based on a RNA signature have been published recently. The study by Mahajan *et al.*⁴ detected a 66-transcript signature in blood that allowed to differentiate bacterial from viral infection in young febrile infants. Herberg *et al.*⁵ have recently identified a host whole blood RNA transcriptomic signature based on only two genes, *FAM89A* and *IFLI44L*, and these results were recently validated by Kaforou and colleagues⁶, again in neonates. The possibility of using only two biomarkers to accurately differentiate viral from bacterial infections makes this approach particularly attractive for the design of a test in a simple device for use in health institutions or as a point of care test⁵. However, despite of the encouraging results of these studies, further effort is required to evaluate the accuracy and clinical utility of these biomarkers in different clinical settings before they can be turned into a clinically applicable test.

¹Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia, Spain. ²GenPoB Research Group, Instituto de Investigacións Sanitarias (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain. ³Translational Pediatrics and Infectious Diseases, Hospital Clínico Universitario de Santiago, Santiago de Compostela, Spain. ⁴GENVIP Research Group (www.genvip.org), Instituto de Investigación Sanitaria de Santiago (IDIS), Galicia, Spain. Correspondence and requests for materials should be addressed to A.S. (email: antonio.salas@usc.es)

At the same time, the implications of the ancestral background of patients in transcription patterns is very poorly approached in the literature. However, some studies demonstrated that gene expression can depend on the ancestry background owing to existing differential frequencies at regulatory polymorphisms in populations⁷. Taking the current evidence into account, we propose here a new approach that allows to infer the ancestry from transcriptomic data without the need for an *ad hoc* test. The inferred ancestry can then be used to evaluate the possible dependence of expression on genetic background.

The present study differs from previous efforts⁵ in that: (i) the performance of the 2-transcript host RNA signature for discriminating bacterial from viral infection is evaluated in a real-life scenario of children suffering from acute diarrhea, (ii) the accuracy of the 2-transcript test is evaluated for acute diarrhea caused by different pathogens, (iii) the discriminatory capacity of this test is measured for different levels of disease severity, and (iv) the dataset employed is from Mexico, thus allowing to test the accuracy of these biomarkers in a non-European cohort.

Material and Methods

Samples and data. We retrieved from the GEO database⁸ an RNA expression dataset obtained from a cohort ($n = 174$) of Mexican children, aged less than 10 years old, suffering from acute diarrhea associated with a single viral or bacterial pathogen, and without systemic complications (GEO accession number: GSE69529).

Ancestral analysis from RNA data. In order to investigate the ancestral background of the biological samples analyzed, RNA data were preliminary processed for the extraction of DNA variant information. We used Opossum to pre-process RNA-seq reads prior to variant calling⁹. Subsequently, Platypus software was used as variant caller tool¹⁰. A total of 68,327 SNPs could be initially retrieved from the RNA data.

Analysis of ancestral background of patients was carried out using standard procedures (e.g.^{11,12}). Inferring ancestral characteristics of the donors requires the use of reference continental populations. Genome data from reference populations were retrieved from two SNP genome repositories. From The 1000 Genomes Project¹³ (<http://www.internationalgenome.org>), we used the European-CEU and the African Yoruban-YRI datasets. From Reich *et al.*¹⁴ we used different population samples (Aymara, Maya, Pima, Quechua, etc) that represent the Native American ancestry, as published previously¹⁵; only the Native American variation of these datasets was used for the analysis (masked data).

We then intersected the SNPs retrieved from the RNA data and the two reference DNA datasets, resulting in a final set of 416 SNPs. It has been shown that a few hundred SNPs can provide accurate population estimates of continental ancestry¹⁶.

Next, we computed identity-by-state (IBS) values from SNP data using PLINK¹⁷. With the aim of exploring for clusters of genetic variation in the population sets analyzed, a Multidimensional Scaling (MDS) plot was built using a matrix of pairwise individual IBS values. MDS was performed using the function *cmdscale* (library *stats*) from R (<http://www.r-project.org>). In addition, we obtained maximum likelihood estimations of individual ancestries from multi-locus SNP data using ADMIXTURE software¹⁸.

RNA data processing. First of all, we performed a quality control of the raw data obtained from the Sequence Read Archive (SRA) using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) to ensure that there were no problems or biases in our data which might affect the downstream analysis. Subsequently, we used MultiQC¹⁹ to aggregate the results from FastQC across many samples into a single plot. We eliminated several samples from our analysis [CA234, CA273, CA251, CA287, CA352, CA450, CA68, 8CA864, CA236] because they had either an incorrect G-C content or many overrepresented sequences²⁰. After the quality control of the raw sequencing data, we selected 175 whole blood samples including children infected by rotavirus ($n = 53$), enteroaggregative *Escherichia coli* (EAEC) ($n = 18$), enteropathogenic *Escherichia coli* (EPEC) ($n = 10$), diffuse-adhering *Escherichia coli* (DAEC) ($n = 21$), *Salmonella* ($n = 36$), and *Shigella* ($n = 36$). A group of healthy controls was also available in the GEO database, but it was processed only here for the aim of ancestry analysis (because the comparisons of virus *versus* bacterial transcription patterns does not need a group of controls⁵).

The whole transcriptome reads were mapped against the version of the human genome provided by Ensembl (version GRCh38_r90/release 90) using the ultrafast universal RNA-seq aligner STAR²¹. The aligned reads were reported in BAM format. We also used STAR for counting the number of reads that map to each gene.

Using the Integrative Genomics Viewer (IGV)²² we visualized where the mapped reads were positioned in the genes of interest. Last, we normalized the reads using several methods, including RPKM (Reads per million mapped reads), TMM (Trimmed mean of M-values), CQN (Conditional quantile normalization), and Deseq2 implemented in the *Deseq2* package of the R statistical software (<https://www.r-project.org>). All the methods yielded virtually the same result; Deseq2 was finally chosen.

Boxplots were built using the R package *beeswarm* (<https://cran.r-project.org/web/packages/beeswarm/>) to represent the Disease Risk Score (DRS)⁵ according to the type of infection and pathogen. Given that the 2-transcript signature in our dataset follows a normal distribution (Shapiro test; P -value > 0.05), we used the T-test for independent samples to evaluate the statistical significance of the differential expression in bacterial *vs.* viral infected patients. In addition, to evaluate the predictive accuracy of the 2-transcript signature we constructed Receiver Operating Characteristic (ROC) curves and the area under the curves (AUC) using the *pROC* package in R. The threshold value, defined as the point on the ROC curve that maximized sensitivity and specificity, was calculated using the R package *OptimalCutPoints*²³. The calculation of the confidence intervals for sensitivity and the specificity was based on a stratified bootstrap resampling.

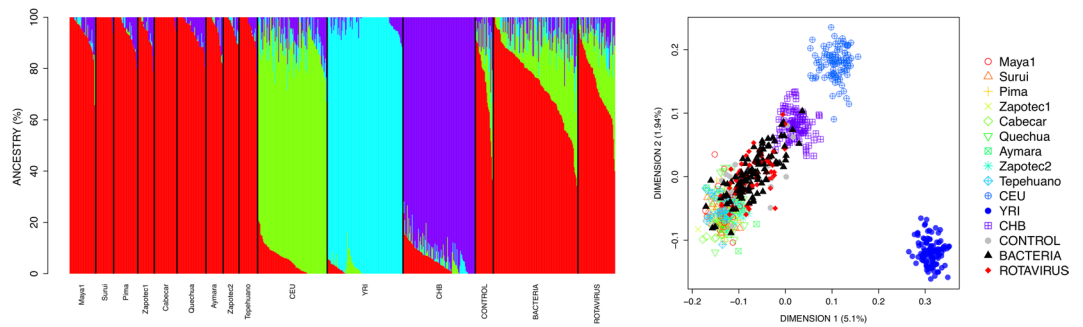


Figure 1. Ancestry analysis of RNA samples based on SNP data. **(A)** Barplot of maximum likelihood estimations of individual ancestries, indicating that most patients have a main Native American ancestry, while European ancestry is present in a few of them with values ranging from 0% to 57.7%. **(B)** MDS plot of pairwise individual IBS values showing that the genome profile of patients and controls (labeled in the figure as control, bacteria and rotavirus) falls close to the Native America cluster (identified by several population samples of Native American ethnic origin), and some of the samples slightly displaced towards the European pole represented by CEU in Dimensions 1 and 2.

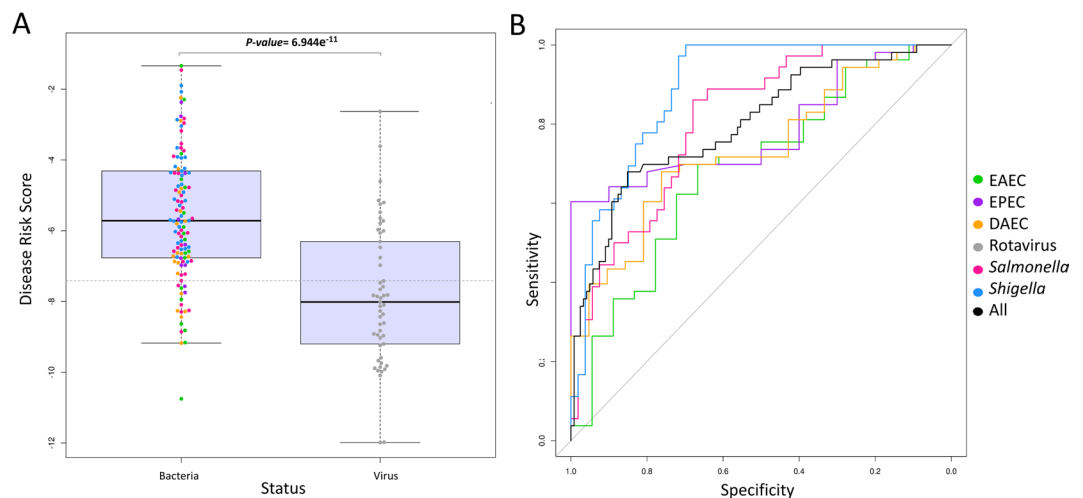


Figure 2. Classification performance based on the 2-transcript DRS combined as $[\log_2(FAM89\text{expression}) - \log_2(ILF44L\text{expression})]$. **(A)** Box and whisker plot of DRS: the horizontal lines in the boxes indicate the median of each group; the lower and upper edges of boxes reflect interquartile ranges, and the whiskers are <1 times the interquartile range; the horizontal grey line is the DRS threshold that maximizes the AUC when comparing patients with viral and bacterial infections. **(B)** ROC curves of different bacteria compared to rotavirus; the different colors identify the type of bacteria according to the inset legend. AUC values are provided in Table 1.

Results

Ancestry analysis. Admixture analysis of the RNA data in virus, bacteria, and control cohorts indicates that the predominant ancestry is clearly Native American (average values: virus = 76.8% [range = 44.7–96.8], bacteria = 76.9% [range = 33.4–100], controls = 77.7 [range = 46–93.7]); the three datasets contain substantial levels of European ancestry too (average values: virus = 19.4%, bacteria = 19.4%; controls = 17.4%) (Fig. 1A). On average, Native American and European ancestries in patients are 77.0% and 19.2%, respectively.

The MDS plot of virus, bacteria, and control groups, together with the three main continental reference populations is in good concordance with the admixture analysis (Fig. 1B). There are three main poles of ancestry in the MDS plots, coinciding with the main three continental groups: Native Americans, Europeans, and Africans. Patients and controls fall closely related to the Native American cluster; a few samples are slightly projected the European cluster, indicating a more marked European ancestry in their genomes.

The 2-transcript RNA signature in virus versus bacteria. We first examined if patients clustered according to their disease status (bacterial vs. viral infection) when applying the DRS. We generated a one-dimensional scatter plot with closely-packed but non-overlapping points (Fig. 2A), which shows a statistically significant difference in the DRS of children affected by bacterial diarrhea compared to those suffering from viral diarrhea ($P\text{-value} = 6.94 \times 10^{-11}$). The results are consistent with those obtained by Herberg *et al.*⁵ in that a higher DRS indicates bacterial infection, whereas a lower DRS indicates viral infection.

	<i>n</i>	AUC	Sensitivity	Specificity	DRS	<i>P</i> -value
All patients						
Rotavirus	53	—	—	—	—	—
EAEC	18	68.7 (54.0–83.3)	69.8 (56.6–81.1)	66.7 (44.4–88.9)	−6.76	2.160×10^{-2}
EPEC	10	78.6 (66.1–90.9)	60.4 (47.2–73.6)	90.0 (70.0–100.0)	−7.74	8.879×10^{-3}
DAEC	21	73.6 (61.7–85.5)	67.9 (54.7–81.1)	76.2 (57.1–90.5)	−7.21	$2.494 \times 10^{-3*}$
<i>Salmonella</i>	36	80.8 (71.9–89.6)	67.9 (54.7–79.2)	83.3 (69.4–94.4)	−7.41	$6.927 \times 10^{-7*}$
<i>Shigella</i>	36	89.0 (82.5–95.6)	69.8 (56.6–83.0)	100.0 (100.0–100.0)	−6.88	$3.140 \times 10^{-12*}$
All Bacteria	121	80.0 (72.6–87.4)	67.9 (54.7–79.2)	84.3 (77.7–90.1)	−7.41	$6.944 \times 10^{-11*}$
Severe & Moderate						
Rotavirus	19	—	—	—	—	—
EAEC	6	58.8 (28.4–89.2)	79.0 (57.9–94.7)	33.3 (0.0–66.7)	−6.25	6.102×10^{-1}
EPEC	3	84.2 (52.3–100.0)	100.0 (100.0–100.0)	66.7 (0.0–100.0)	−4.37	1.780×10^{-1}
DAEC	6	86.0 (66.6–100.0)	89.5 (73.7–100.0)	66.7 (33.3–100.0)	−5.80	4.057×10^{-2}
<i>Salmonella</i>	12	83.3 (69.0–97.6)	68.4 (47.4–89.5)	83.3 (58.3–100.0)	−7.41	$5.879 \times 10^{-3*}$
<i>Shigella</i>	22	93.5 (86.6–100.0)	73.7 (52.6–94.7)	100.0 (100.0–100.0)	−6.75	$2.219 \times 10^{-7*}$
All bacteria	49	85.3 (75.7–94.9)	79.0 (57.9–94.7)	77.6 (65.3–87.8)	−6.40	$2.126 \times 10^{-6*}$
Mild						
Rotavirus	34	—	—	—	—	—
EAEC	12	73.5 (57.6–89.5)	67.7 (47.1–82.4)	75.0 (0.5–100.0)	−6.76	1.818×10^{-2}
EPEC	7	74.8 (59.3–90.3)	67.7 (47.1–82.4)	100.0 (100.0–100.0)	−7.57	4.668×10^{-2}
DAEC	15	69.0 (53.9–84.2)	64.7 (47.1–79.4)	80.0 (60.0–100.0)	−7.78	3.601×10^{-2}
<i>Salmonella</i>	24	80.0 (68.6–91.4)	67.7 (53.0–82.4)	66.7 (83.3–95.8)	−7.55	$4.698 \times 10^{-5*}$
<i>Shigella</i>	14	84.2 (73.4–95.1)	67.7 (52.9–82.4)	100.0 (100.0–100.0)	−6.88	$1.157 \times 10^{-5*}$
All bacteria	72	77.0 (66.5–87.4)	67.7 (50.0–82.4)	86.1 (77.8–93.1)	−7.57	$8.154 \times 10^{-6*}$

Table 1. AUC values and DRS for different pathogens. Figures in round brackets show the CI 95% (2,000 bootstrap replicates). *P*-values evaluate the significance of the difference between bacterial pathogen vs. viral infection under a t-test; an asterisk. (*) indicates those values surpassing a Bonferroni correction for multiple tests. Note that the reference viral pathogen in our study is rotavirus (for which there is only very little information available in the literature; e.g.⁴³).

The diagnostic accuracy of the test to discriminate viral from bacterial diarrhea was evaluated using ROC analysis (Fig. 2B), and considering different scenarios: (i) all the bacterial infections against rotavirus, and (ii) each bacterial pathogen (EAEC, DAEC, EPEC, *Salmonella*, and *Shigella*) against rotavirus. For the first scenario, the ROC curve indicates that the accuracy of the test is ~80%, a value that is comparable to that reported by Herberg *et al.* (namely, AUC = ~90%). At the same time, the diagnostic performance of the 2-transcript test seems to be pathogen-dependent (Fig. 2B), with some bacterial infections showing more distinctive pattern than viral infections; for example, the AUC for *Shigella* is very high (AUC = 89.0%; 95% CI: 82.5–95.6), while this is more moderate for EAEC (AUC = 68.7%; 95% CI: 54.0–83.3). In most of the comparisons carried out considering the different groups of bacteria independently, the differential expression in bacterial vs. viral infections was highly statistically significant, surpassing the Bonferroni threshold; in cases not showing statistical significance, this could be due to small sample sizes, since most are close to the limit of the significance threshold; Table 1).

Most interesting is the observation that the level of severity affects the accuracy of the 2-biomarker test. Thus, patients with mild disease introduce noise in the capacity of the two biomarkers to characterize bacterial from viral infections; when these patients are excluded, the signal improves substantially as reflected in the boxplots in Fig. 3B,C; the difference with respect to viral infection is still statistically significant in both groups of patients, severe + moderate (P -value = 2.13×10^{-6}) and mild (P -value = 8.15×10^{-6}) cases (Table 1). This difference is also evident when examining the ROCs obtained by considering different levels of severity (Fig. 3A; AUC_{MILD} = 77.0% vs. AUC_{SEVERE+MODERATE} = 85.3%; AUC_{SEVERE} = 85.0%; AUC_{MODERATE} = 89.2%).

Finally, a correlation analysis between the 2-transcript test and the different ancestral coefficients indicate that expression of these two genes is not correlated with ancestral background (Fig. S1 from Supplementary Data).

Discussion

It is common practice in hospitals to administer antibiotics to febrile children until the culture results are available⁵. Antibiotics are generally prescribed out of fear of missing a bacterial infection, and such practice is contributing to the rise of drug-resistant pathogens all over the world¹. To give some examples, according to the data presented by Craig *et al.*²⁴ on a prospective cohort study of >15,700 febrile illnesses, antibiotics were prescribed acutely in 66% to 81% of children (depending on the type of infection); while 20% of patients without bacterial infection received also antibiotic treatment. According to another study on meningitis, no pathogen was identified in 50%–80% of the patients but most of them received antibiotics, and almost 50% were (probably unnecessarily) hospitalized for more than a week²⁵. Nijman *et al.*²⁶ stated that “a prediction model should discriminate between children at high and low risk of serious bacterial infections [...] This is helpful to identify

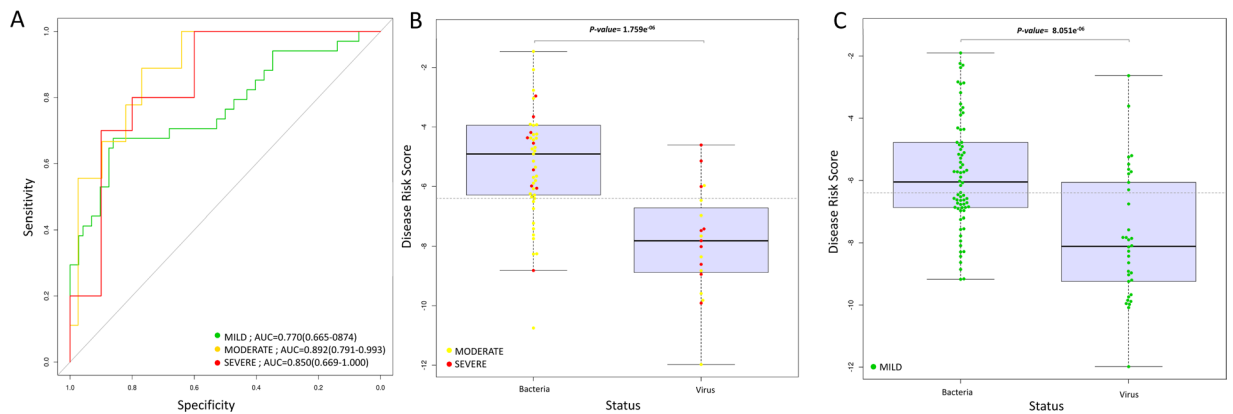


Figure 3. Classification performance based on the 2-transcript DRS combined as $[\log_2(FAM89\text{expression}) - \log_2(ILF44L\text{expression})]$ and considering different levels of severity. **(A)** ROC curves of different bacteria compared to rotavirus; AUC values by pathogen are provided in Table 1; and **(B)** and **(C)** box and whisker plots of DRS for moderate plus severe cases and mild cases, respectively.

children who may need further tests, such as additional blood tests or chest radiography, and guide the need for antibiotic prescribing". On the other hand, culture of bacteria from normally sterile sites may be negative in scenarios where the infection resides in inaccessible sites, when the patient received antibiotic previously, or when the culture media selected fail to grow certain bacteria^{5,27}. According to the meta-analysis carried out by Iroth Tam *et al.*²⁸, blood culture identifies organisms in only a very small percentage of patients (~5%) suffering pediatric community-acquired pneumonia (CAP) (predominantly *S. pneumoniae*); this result was also corroborated by Kwon *et al.*²⁹, who claimed the low utility of blood culture in CAP in an observational study considering >2,700 patients. In addition, while some authors defend the utility of MALDI-TOF as a new rapid technology for microbial identification³⁰, the results from conventional cultures of bacteria may take several days^{5,27,30}, so physicians commonly make the decision to administer preventive antibiotic treatment^{25,31}.

In this scenario, developing a cost-effective device to differentiate viral from bacterial infections in early stages is one of the biggest challenges in medical research. In 2016, Herberg *et al.*⁵, published a study that shed new light on this problem as they identified a simple blood 2-transcript RNA expression signature that distinguishes viral from bacterial infection in febrile children.

Our study adds further support to these findings and also broadens the scope of the initial research. First, the Herberg *et al.*⁵ study focused on a number of well-selected viral and bacterial febrile patients that needed intensive care or had severe infections (e.g. sepsis); whereas the data set investigated here consisted of children affected by diarrhea caused by several pathogens, and thus a cohort that might represent a frequent scenario in emergency departments or outpatient clinics. Our results indicate that the 2-transcript test performs differently in different infectious scenarios, with the best results shown for diarrhea caused by *Shigella*.

Second, the majority of the pathogens included in the Mexican database (e.g. *Shigella*, *Rotavirus*) had not been tested in the original study or, as in the case of *E. coli* and *Salmonella*, our study adds more cases and different strains. This supports the conclusion that the test seems to perform differently depending on the pathogen.

Third, the database explored here included children with mild, moderate and severe symptomatology, which allowed us to assess the 2-transcript signature ranked by levels of severity. The results show that the test performs better in patients suffering from moderate to severe symptoms of diarrhea.

Fourth, we measured the potential dependence of the 2-transcript test on the ancestral background of patients. To the best of our knowledge, this is the first time that genetic background is inferred from transcriptomic data aiming at evaluating its potential effect on gene expression. Commonly, self-reported ancestry is used to stratify the statistical analyses; e.g.³²; and in a number of studies, although well-differentiated ancestral groups were analyzed, there is no indication on how the ancestral background was assessed³³. The exception to this common practice is Serrano-Gómez *et al.*³⁴; this study used ancestry-informative markers (AIMs)¹⁶ to measure the ancestry as a potential modifier of gene expression in breast cancer. The original dataset used by Herberg *et al.*⁵ included patients of main European ancestry (UK, Spain, the Netherlands, and USA); while our sample represents a Mesoamerican population with predominantly Native American ancestry but variable levels of admixture. However, our results indicate that the 2-transcript test performs well in a different background to the one used for its discovery³⁵, and that there is no correlation between the expression signal and the ancestral background of patients.

Last but not least, the original study used data provided by different microarray platforms; the batch effects were targeted computationally. The entire Mexican dataset was generated using RNA-seq and the sequencer Illumina HiSeq2500, and therefore batch effects should be minimal.

There are several further considerations from the present study. The results indicate that the 2-biomarkers test performs much better in severe patients (more closely related to those examined by Herberg *et al.*⁵). We have exclusively analyzed cases with a single infecting pathogen, but it would most interesting to further explore the behavior of this 2-biomarker test in scenarios of coinfection³⁶⁻³⁸. The translation of this 2-biomarker test into a clinical setup still needs further validation; and this validation would require the analysis of more virus and

bacteria scenarios, as well as different time points in the course of the same infection, or the case of parasitic infections. Despite its limitations, the present study (together with the previous ones^{5,6}) represents a step forward towards a bench to bedside test that could help make decisions for antibiotic administration³. It is now possible to test transcription signatures using portable devices^{39–41}, so these techniques will be most helpful in small hospitals in rural and isolated areas, and in hospitals from developing countries. In addition, although the signal provided by *IFI44L* and *FAM89A* transcripts is not as strong in diarrhea patients as in more severe disease scenarios⁵, the evidence points to an important role of these two genes in the differential molecular mechanisms employed by virus and bacteria for infection. These observations should attract further research exploring the role of these genes and related pathways under the hypothesis that both could be key at understanding mechanisms of infections. To the best of our knowledge, there is very little information available for these two genes, but the *FAM89A* (interferon-induced protein 44-like) is known to exhibit antiviral activity⁴².

Summarizing, in all of the scenarios tested, our results suggest that these two biomarkers, *IFI44L* and *FAM89A* transcripts, provide a strong signal to differentiate bacterial from viral infections, a signal that is non-population dependent, and useful for discriminating a wide range of pathogens and different levels of severity. Further studies focusing on the time point of the disease, testing new (portable) devices for RNA typing, and targeting different populations worldwide and a wider spectrum of pathogens will be necessary to further confirm the accuracy of this biomarker test before it can be turned into a clinically applicable test.

References

- Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).
- Llor, C. & Bjerrum, L. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. *Ther. Adv. Drug. Saf.* **5**, 229–241 (2014).
- Gliddon, H. D., Herberg, J. A., Levin, M. & Kaforou, M. Genome-wide host RNA signatures of infectious diseases: discovery and clinical translation. *Immunology* **153**, 171–178 (2018).
- Mahajan, P. *et al.* Association of RNA biosignatures with bacterial infections in febrile infants aged 60 days or younger. *JAMA* **316**, 846–857 (2016).
- Herberg, J. A. *et al.* Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. *JAMA* **316**, 835–845 (2016).
- Kaforou, M., Herberg, J. A., Wright, V. J., Coin, L. J. M. & Levin, M. Diagnosis of bacterial infection using a 2-transcript host RNA signature in febrile infants 60 days or younger. *JAMA* **317**, 1577–1578 (2017).
- Spielman, R. S. *et al.* Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* **39**, 226–231 (2007).
- Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
- Oikkonen, L. & Lise, S. Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. *Wellcome Open Res.* **2**, 6 (2017).
- Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
- Salas, A. *et al.* Whole Exome Sequencing reveals new candidate genes in host genomic susceptibility to Respiratory Syncytial Virus Disease. *Sci. Rep.* **7**, 15888 (2017).
- Pardo-Seco, J., Gómez-Carballea, A., Amigo, J., Martín-Torres, F. & Salas, A. A genome-wide study of modern-day Tuscans: revisiting Herodotus's theory on the origin of the Etruscans. *PLoS One* **9**, e105920 (2014).
- The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
- Pardo-Seco, J., Heinz, T., Taboada-Echalar, P., Martín-Torres, F. & Salas, A. Mapping the genomic mosaic of two 'Afro-Bolivians' from the isolated Yungas valleys. *BMC Genomics* **17**, 207 (2016).
- Pardo-Seco, J., Martín-Torres, F. & Salas, A. Evaluating the accuracy of AIM panels at quantifying genome ancestry. *BMC Genomics* **30**, 543 (2014).
- Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
- Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016).
- Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
- López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C. & Gude-Sampedro, F. OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *J. Stat. Softw.* **61**, 1–36 (2014).
- Craig, J. C. *et al.* The accuracy of clinical symptoms and signs for the diagnosis of serious bacterial infection in young febrile children: prospective cohort study of 15 781 febrile illnesses. *BMJ* **340**, c1594 (2010).
- Sadarangani, M. *et al.* Childhood meningitis in the conjugate vaccine era: a prospective cohort study. *Arch. Dis. Child.* **100**, 292–294 (2015).
- Nijman, R. G. *et al.* Clinical prediction model to aid emergency doctors managing febrile children at risk of serious bacterial infections: diagnostic study. *BMJ* **346**, f1706 (2013).
- Riddle, M. S., DuPont, H. L. & Connor, B. A. ACG Clinical Guideline: Diagnosis, Treatment, and Prevention of Acute Diarrheal Infections in Adults. *Am. J. Gastroenterol.* **111**, 602–622 (2016).
- Iroh Tam, P. Y., Bernstein, E., Ma, X. & Ferrieri, P. Blood Culture in Evaluation of Pediatric Community-Acquired Pneumonia: A Systematic Review and Meta-analysis. *Hosp. Pediatr.* **5**, 324–336 (2015).
- Kwon, J. H. *et al.* Low utility of blood culture in pediatric community-acquired pneumonia: An observational study on 2705 patients admitted to the emergency department. *Medicine (Baltimore)* **96**, e7028 (2017).
- Singhal, N., Kumar, M., Kanaujia, P. K. & Viridi, J. S. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.* **6**, 791 (2015).
- Colvin, J. M. *et al.* Detection of viruses in young children with fever without an apparent source. *Pediatrics* **130**, e1455–1462 (2012).
- Cai, G. *et al.* Population effect model identifies gene expression predictors of survival outcomes in lung adenocarcinoma for both Caucasian and Asian patients. *PLoS One* **12**, e0175850 (2017).

33. Lukas, T. J. *et al.* Susceptibility to glaucoma: differential comparison of the astrocyte transcriptome from glaucomatous African American and Caucasian American donors. *Genome Biol.* **9**, R111 (2008).
34. Serrano-Gómez, S. J. *et al.* Ancestry as a potential modifier of gene expression in breast tumors from Colombian women. *PLoS One* **12**, e0183179 (2017).
35. Sandoval, K. *et al.* Linguistic and maternal genetic diversity are not correlated in Native Mexicans. *Hum. Genet.* **126**, 521–531 (2009).
36. Cebeý-López, M. *et al.* Viral Co-Infections in Pediatric Patients Hospitalized with Lower Tract Acute Respiratory Infections. *PLoS One* **10**, e0136526 (2015).
37. Cebeý-López, M. *et al.* Does Viral Co-Infection Influence the Severity of Acute Respiratory Infection in Children? *PLoS One* **11**, e0152481 (2016).
38. DeBerg, H. A. *et al.* Shared and organism-specific host responses to childhood diarrheal diseases revealed by whole blood transcript profiling. *PLoS One* **13**, e0192082 (2018).
39. Walter, M. C. *et al.* MinION as part of a biomedical rapidly deployable laboratory. *J Biotechnol* **250**, 16–22 (2017).
40. Russell, J. A. *et al.* Unbiased Strain-Typing of Arbovirus Directly from Mosquitoes Using Nanopore Sequencing: A Field-forward Biosurveillance Protocol. *Sci Rep* **8**, 5417 (2018).
41. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc* **12**, 1261–1276 (2017).
42. Schoggins, J. W. *et al.* A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature* **472**, 481–485 (2011).
43. Salas, A. *et al.* Strong down-regulation of glycoporphin genes: A host defense mechanism against rotavirus infection. *Infect. Genet. Evol.* **44**, 403–411 (2016).

Acknowledgements

We would like to thank the support and help of Myrsini Kaforou and Jethro A. Herberg (Imperial College London; UK). We would also like to gratefully acknowledge CESGA (Supercomputing Centre of Galicia, Santiago de Compostela, Spain) for the supercomputing availability, web hosting and support. This study received support from the Instituto de Salud Carlos III (Proyecto de Investigación en Salud, Acción Estratégica en Salud): project GePEM ISCIII/PI16/01478/Cofinanciado FEDER (to A.S.); project ReSVinext ISCIII/PI16/01569/Cofinanciado FEDER (to F.M.-T.); Consellería de Sanidade, Xunta de Galicia (RHI07/2-intensificación actividad investigadora, PS09749 and 10PXIB918184PR), Instituto de Salud Carlos III (Intensificación de la actividad investigadora 2007–2012, PI16/01569) (to F.M.-T.); Fondo de Investigación Sanitaria (FIS; PI070069/PI1000540) del plan nacional de I + D + I and ‘fondos FEDER’ (to F.M.-T.); 2016-PG071 Consolidación e Estructuración REDES 2016GI-1344 G3VIP (Grupo Gallego de Genética Vacunas Infecciones y Pediatría, ED341D R2016/021) (to A.S. and F.M.-T.).

Author Contributions

A.S., F.M.-T. and R.B.-A. conceived and designed the study. A.S., R.B.A. and J.P.-S. analyzed the data. A.S. and R.B.A. wrote the first draft of the manuscript and it was revised by F.M.-T. and J.P.-S. All the authors contributed to the final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-26239-1>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018