EBioMedicine 67 (2021) 103352

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

VB_{10}, a new blood biomarker for differential diagnosis and recovery monitoring of acute viral and bacterial infections

Sathyabaarathi Ravichandran[a], Ushashi Banerjee[b], Gayathri Devi DR[f], Rooparani Kandukuru[f], Chandrani Thakur[b], Dipshikha Chakravortty[c,d], Kithiganahalli Narayanaswamy Balaji[d], Amit Singh[d,e], Nagasuma Chandra[a,b,c,*]

[a] IISc Mathematics Initiative, Indian Institute of Science, Bangalore 560012, India
[b] Department of Biochemistry, Indian Institute of Science, Bangalore 560012, India
[c] Centre for Biosystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India
[d] Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore 560012, India
[e] Centre for Infectious Disease Research, Indian Institute of Science, Bangalore 560012, India
[f] Department of Microbiology, M S Ramaiah Medical College, Bangalore 560054, Karnataka, India

ARTICLE INFO

Article History:
Received 4 December 2020
Revised 4 April 2021
Accepted 7 April 2021
Available online xxx

Keywords:
Acute infections
Antimicrobial resistance
Biomarker
Blood
Transcriptome
Systems biology
Classifier
Diagnostic score

ABSTRACT

Background: Precise differential diagnosis between acute viral and bacterial infections is important to enable appropriate therapy, avoid unnecessary antibiotic prescriptions and optimize the use of hospital resources. A systems view of host response to infections provides opportunities for discovering sensitive and robust molecular diagnostics.

Methods: We combine blood transcriptomes from six independent datasets (n = 756) with a knowledge-based human protein-protein interaction network, identifies subnetworks capturing host response to each infection class, and derives common response cores separately for viral and bacterial infections. We subject the subnetworks to a series of computational filters to identify a parsimonious gene panel and a standalone diagnostic score that can be applied to individual samples. We rigorously validate the panel and the diagnostic score in a wide range of publicly available datasets and in a newly developed Bangalore-Viral Bacterial (BL-VB) cohort.

Finding: We discover a 10-gene blood-based biomarker panel (Panel-VB) that demonstrates high predictive performance to distinguish viral from bacterial infections, with a weighted mean AUROC of 0.97 (95% CI: 0.96–0.99) in eleven independent datasets (n = 898). We devise a new stand-alone patient-wise score (VB_{10}) based on the panel, which shows high diagnostic accuracy with a weighted mean AUROC of 0.94 (95% CI 0.91–0.98) in 2996 patient samples from 56 public datasets from 19 different countries. Further, we evaluate VB_{10} in a newly generated South Indian (BL-VB, n = 56) cohort and find 97% accuracy in the confirmed cases of viral and bacterial infections. We find that VB_{10} is (a) capable of accurately identifying the infection class in culture-negative indeterminate cases, (b) reflects recovery status, and (c) is applicable across different age groups, covering a wide spectrum of acute bacterial and viral infections, including uncharacterized pathogens. We tested our VB_{10} score on publicly available COVID-19 data and find that our score detected viral infection in patient samples.

Interpretation: Our results point to the promise of VB_{10} as a diagnostic test for precise diagnosis of acute infections and monitoring recovery status. We expect that it will provide clinical decision support for antibiotic prescriptions and thereby aid in antibiotic stewardship efforts.

Funding: Grand Challenges India, Biotechnology Industry Research Assistance Council (BIRAC), Department of Biotechnology, Govt. of India.

1. Introduction

Infectious diseases pose a significant health concern and kill over 17 million people in a year globally according to the World Health Organization reports [1,2]. The current pandemic due to SARS-CoV-2

* Corresponding author at: IISc Mathematics Initiative, Indian Institute of Science, Bangalore 560012, India.
E-mail address: nchandra@iisc.ac.in (N. Chandra).

https://doi.org/10.1016/j.ebiom.2021.103352

Research in context

Evidence before this study

The treatment of infectious diseases, of late, has taken a new dimension globally due to the emergence of antimicrobial resistance (AMR). Inappropriate use of antibiotics is a major cause of this problem. A solution to this problem is to find new markers for precise differential diagnosis between bacterial and viral infections and thereby guide the physician to avoid unnecessary antibiotic prescriptions. The current diagnostic strategies rely mainly on pathogen-based detection techniques, which suffer from several limitations. A clear alternative to this is host-based markers. An example of this is Procalcitonin (PCT), which is increasingly used in the clinic to diagnose gram-negative bacterial infections from other bacterial and fungal infections in clinical settings. However, elevated levels of PCT are seen in many other clinical conditions as well, leading to its sub-optimal performance as a diagnostic marker. On the other hand, blood transcriptomes from different viral and bacterial infections have shown the host response to be distinct in viral and bacterial infections. A few studies report the use of such information to identify RNA - based biomarker panels for differentiating viral from bacterial infections. These clearly demonstrate the promise of RNA panels. The key enabling factors that will significantly aid in translating these biomarkers into the clinic are (a) improvement in sensitivity and specificity, (b) demonstrating sufficient generality – concerning the applicability across different populations, and (c) making it accessible as a simple readout to the clinician.

Added value of this study

We achieve all these factors by discovering a new robust 10-gene biomarker panel that exhibits improved diagnostic accuracy and applicability across a wide range of bacteria and viruses. To push it towards translation, we formulate a standalone diagnostic score and demonstrate our score's diagnostic utility with rigorous best practices in the field. We show that VB_{10} can be used as a blood test for precise differential diagnosis of viral and bacterial infections through an extensive analysis on a range of datasets. We demonstrate that VB_{10} exhibits high diagnostic accuracy across different age groups, different geographical locations, and across a broad spectrum of acute infection, including COVID-19. We also show that VB_{10} can monitor the recovery status, and moreover, as a clinical decision support tool.

Implication of all the available evidence

Our study demonstrates that VB_{10}, a new standalone diagnostic-score has high classification power for the differential diagnosis of acute viral and bacterial infections. It follows from this that VB_{10} could guide a clinician in choosing an optimal treatment plan, including deciding whether to prescribe antibiotics.

has shown that the mortality rate due to a viral infection can be alarmingly high [3]. A major challenge in treating them is in the accurate diagnosis of whether it is of viral or bacterial etiology because a wide variety of them present with common clinical manifestations. This often leads to misdiagnosis and consequently trial-and-error treatment plans [4,5]. Moreover, overuse of antibiotics leads to antimicrobial resistance (AMR), which is a significant threat to human health [6]. The highest mortality rate due to AMR in the world is recorded in India, with about 416.75 deaths per 100,000 persons [7]. Accurate discrimination between bacterial and viral infections will help enormously in guiding a clinician to select appropriate treatment strategies, to optimally deploy hospital resources and in the judicious use of antibiotics. In cases such as sepsis [8] and community-acquired pneumonia [9], the decision of whether to prescribe antibiotics can be a life-determining factor.

At present, the 'gold standard' diagnostic methods used in the clinic are based on pathogen detection techniques [10,11]. However, these methods suffer from several limitations, as they cannot be used to detect uncultivable or uncharacterized pathogens. They also cannot detect infections with low pathogen counts or discriminate between live and dead organisms. Instead, a more promising approach is to focus on host-based markers. Blood tests that measure the hemogram, erythrocyte sedimentation rate and C-reactive protein are often used as broad indicators of infection during a clinical examination [12,13]. However, they are at best only approximate indicators as they are seen to vary in a wide variety of diseases and lack both the sensitivity and specificity to discriminate between bacterial and viral infections. For example, procalcitonin is increasingly used as a marker for detecting bacterial infections in case of sepsis and lower respiratory tract infections, but its performance is limited due to suboptimal sensitivity and specificity and hence does not meet the requirement of an accurate actionable diagnostic test [14]. A reliable sensitive diagnostic test is needed to accurately determine the nature of the infection and obtain a quantitative picture of the disease burden. A need for such a diagnostic test has become even more acute considering the currently ongoing COVID-19 pandemic.

Several reports have indicated the promise of molecular diagnostics that are based on the host response to infections. The starting point for most of these studies is the host blood transcriptomes [15–17]. Blood, with its unique advantages of capturing the systemic effect of a given infection and being a highly accessible tissue, serves as an ideal source for obtaining transcriptome profiles from different patients. Blood transcriptomes from multiple studies have shown the host response to be distinct in viral and bacterial infections, which have led to identification of gene panels of different sizes capable of classifying samples with viral infections from those with bacterial infections in different clinical scenarios [18–25]. The best of the panels, while capable of sensitively distinguishing between viral and bacterial diseases, show low specificity, indicating the need for identifying improved panels. A key factor in translating the biomarkers into clinical use is to bring in improvement in specificity and applicability across a wide variety of acute viral and bacterial diseases.

Transcriptomes being unbiased genome-wide profiles, although recognized to contain a wealth of information about the conditions, present a huge challenge to identify minimal gene panels with high classification power. Multiple studies have deposited clinical transcriptomes in public repositories, making them available for independent analysis using different approaches [26,27]. Most studies so far have used statistical models to probe the data to identify distinguishing gene panels. Statistical models are known to be critically sensitive to the method adopted for applying correction factors to place different datasets on a comparable framework and hence suffer from the possibility of over-dependence and naive interpretation of the test procedure's p-value [28,29]. Heterogeneity in gene expression profiles due to differences in genetic and environmental backgrounds is a well-recognized problem in the biomarker discovery field [30,31]. Since the clinical transcriptome data is large and heterogeneous, it is important to interrogate the data with orthogonal methods to explore new panels with improved diagnostic power and generality. Network-based methods provide an excellent platform to address these issues [32,33].

In this work, we seek to identify a RNA signature for accurately differentiating viral from bacterial infections and formulating a diagnostic score to enable testing individual patient samples. To achieve

this, we configure a computational pipeline involving genome-wide protein-protein interaction networks and model the host response to viral and bacterial infections using the publicly available blood transcriptomes from multiple populations. We then apply a series of filters to discover a 10-gene panel that can robustly discriminate viral from bacterial infections. We then formulate a standalone diagnostic score which leads to a blood test to aid clinical decision-making for antibiotic prescriptions. We demonstrate that our test is capable of diagnosis in independent datasets as well as in a new pool of South Indian patients with high accuracy and specificity. We also show that our test is capable of accurately capturing disease recovery.

2. Methods

2.1. Systematic curation and preprocessing of publicly available transcriptomes

We performed a comprehensive search in Gene Expression Omnibus [26] and ArrayExpress [27] using defined keywords to identify transcriptome data containing blood samples from patients with viral or bacterial infections. Next, we systematically screened these transcriptome datasets and selected as per the guidelines defined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (Fig. S1). The raw data for the selected studies were downloaded and an appropriate preprocessing procedure was adopted using Bioconductor packages in R [34–37]. Affymetrix arrays were background corrected using Robust Multi-array Average (RMA), whereas Agilent and Illumina arrays were corrected using '*normexp*' followed by quantile normalization and $\log_2$ transformation. Preprocessed data were considered for the samples hybridized using custom arrays. Probes that were below the detection limit in >80% of the arrays were filtered out, and the rest were mapped onto the respective genes. Each dataset was preprocessed independently. Detailed information of the publicly available whole blood transcriptome datasets considered in the study is provided in (Table S1). We performed differential gene expression analysis using the limma package in R [38] by comparing 1) Viral vs. Healthy Control, 2) Bacterial vs. Healthy Control, and 3) Viral vs. Bacterial for each dataset in the discovery set independently.

2.2. Reconstruction of the human interactome

We constructed a knowledge-based genome-scale human protein-protein interaction network (hPPiN2), which is an improved version of a previous network hPPiN from our laboratory [39]. This network is built by considering experimentally determined structural and functional interactions incorporated from resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [40], OmniPath [41], Signalink 2.0 [42], Harmonizome [43], RegNetwork [44], HTRIdb [45], TRRUST [46] and TFCat [47]. In brief, the interactions from various primary resources capture 1) regulatory interactions between transcription factors and their targets, 2) metabolic enzyme-coupled interactions, 3) the kinome network, 4) protein-protein complexes, and 5) signaling interactions. The resultant of all interactions after removing redundancy from (1-5) yielded a network of 20,183 nodes that are interconnected by 255,486 edges. In total 215,206 were directed, and 40,280 were bidirected, corresponding to binding interactions. The nodes represent proteins and edges represent interactions among the corresponding proteins.

2.3. Generating context specific networks

We used a sensitive network mining approach developed earlier in our laboratory to generate context-specific networks, and mine the top-ranked perturbed interactions (333,847). In brief, the differential transcriptome computed for viral and bacterial samples with respect to the corresponding healthy controls was mapped on to the hPPiN2 in the form of node and edge weights. The top-ranked activated paths (TAP) and top-ranked repressed paths (TRP) were computed and combined to obtain a top perturbed network (TPN) for each condition. To generate an activated network, the node weight of node i in a diseased condition A was computed as:

$$N_i(A) = FC_i(A/B) \tag{1}$$

Where FC was the fold change of gene i in diseased condition A with respect to the reference condition B (antilog values were used to compute fold changes). To generate the repressed network, the node weight of node i in a diseased condition A was computed as:

$$N_i(A) = FC_i(B/A) \tag{2}$$

The edge weight $We_{ij}(A)$ in a given condition A for an edge e comprised of nodes $Ni(A)$ and $N_j(A)$ was calculated as

$$We_{ij}\ (A) = \frac{1}{\sqrt{\left(N_i(A) * N_j(A)\right)}} \tag{3}$$

Where $N_i(A)$ and $N_j(A)$ are the node weights of nodes i and j, respectively. Lower the edge weight, higher is the edge activity.

2.4. Computing top perturbed networks

We mined the weighted network as described before [33,39,48] to obtain top-active and top-repressed paths that were combined to obtain the top-perturbed network. The algorithm computes minimum weight shortest paths, in which each path begins from a source node and ends with a sink node, passing through interacting nodes in such a way that the least-cost edge is incorporated in every step. The shortest paths between all pairs of genes were computed using Dijkstra's algorithm implemented in the Zen library, Python2.7. For a path of length n, the path cost was calculated as a summation of the edge weights $\sum W_e(A)$ of all edges forming the path, normalized over the path length. All paths were sorted with respect to their path costs, with the least-cost paths ranked the highest. Subsequently, paths belonging to the top 0.05% were taken to constitute the top perturbed network. To dissipate the concern of overfitting and evaluate the sensitivity of the results with respect to the chosen threshold (i.e., 0.05), TPNs constructed based on the cutoffs in and around the threshold (i.e., 0.04 and 0.06) were evaluated. This analysis showed that the cores are relatively stable around the chosen threshold in terms of network size.

2.5. Network visualization and enrichment analysis

We visualized all networks in Allegro Spring-Electric layout using Cytoscape 3.2.0, and compute the network properties using NetworkAnalyzer plugin [49]. We used Reactome with default parameters for pathway enrichment analysis [50] and the resultant hits with q-value $\leq$ 0.01 were considered to be significant. The highly curated gene-disease association reported for viral (C0042769), and bacterial infection (C0004623) were retrieved from DisGeNET [51]. These genes were considered as a gold standard gene set (GSGS) to perform overlap analysis with the top perturbed networks. We used a hypergeometric test for computing the overlap significance [52].

2.6. Evaluation of classifier performance

The classification models were built using the discovery set and their predictive performance were tested on the validation meta cohorts using Logistic Regression (LR). The area under the receiver operating characteristic curve (AUROC) with confidence intervals (CI) (95%) was estimated using the DeLong method for each dataset using the pROC package in R [53]. For comparison with other signatures,

the weighted mean AUROC, sensitivity and specificity with 95% CI was calculated for each model [54,55]. The weighted mean AUROC was computed by calculating AUROC weighted by the number of samples in the respective dataset.

2.7. Ethics

Ethical approval for this study was obtained from the Institutional Ethics Committee at MS Ramaiah medical college, Bangalore, India (ECR/215/Inst/KA/2013/RR-16), and IISc (11-15032017), Bangalore, India. Written informed consent was obtained from all study participants before sample collection.

2.8. Bangalore – Viral Bacterial (BL-VB) cohort

This is an observational cohort on adults with acute infections (2018–2019) from MS Ramaiah medical college, Bangalore, India, and matched healthy controls from the health centre (primary care centre within the university), Indian Institute of Science (IISc), Bangalore, India.

Patients with acute infection-associated diseases, enrolled at an intensive care unit, MS Ramaiah Medical hospital were screened for bacterial and viral infections, and blood samples were collected. These patients were grouped into confirmed viral, confirmed bacterial and indeterminate infection groups based on clinical and microbiological investigation results prior to the targeted validation of the proposed signature panel. Briefly, patients with viral infections were diagnosed based on serological tests, and bacterial infections were diagnosed by bacterial culture tests. Patients with inconclusive diagnosis based on the microbiological investigations (culture and serology negative) were categorized as indeterminate infections. Age matched healthy controls were recruited from the Health Centre, IISc based on the following inclusion criteria: a) no febrile illness (within a month), b) not on medications (within a month) and c) no history of acute or chronic inflammatory diseases. Blood samples were then obtained from these healthy controls and screened for tuberculosis and HIV in addition to a routine hemogram. Table 1 provides the clinical characteristics of patient groups in Bangalore - Viral Bacterial (BL-VB) Cohort. Detailed information on the Clinical characteristics of patients recruited for BL-VB Cohort is presented in Table S2.

2.9. Signature validation

Whole blood samples (2 ml) were collected for targeted gene expression validation using nanostring and qRT-PCR. These samples were mixed with RNAlater (Thermo Fisher Scientific) and stored at -70 °C. Later, RNA was extracted from blood using RiboPure-Blood kit (ThermoFisher scientific) following the manufacturer's protocol, which is followed by DNase treatment and quantification using NanoDrop Light UV-Vis Spectrophotometer (Thermo Fisher Scientific). Ncounter based RNA quantification was performed based on the manufacturer's protocol to quantify gene expression using the custom-made codeset. This custom panel contained 13 genes (including internal housekeeping control genes - *ALAS1*, *POLR2A*, and *SDHA*), which showed expression level changes upon viral and bacterial infection. The counts were renormalized to housekeeping genes using nSolver software (nanostring technologies) (Data file S1). The expression of these genes in a subset of samples in the BL- VB cohort was independently validated using qRT-PCR. Towards this, first-strand cDNA synthesis was performed using 600 ng of total RNA with iScript cDNA synthesis kit (Bio-Rad). Gene expression was analyzed with real-time PCR using iTaq Universal SYBR Green Supermix (Bio-Rad) on the CFX384 instrument (BioRad). Calculation of ΔCt and Relative Copy Number (RCN) for all genes were performed using geometric mean of Ct values of the three control genes (ALAS1, POLR2A, and SDHA). The list of primers used for the experiment was provided in Table S4.

2.10. Statistical analysis

Genes with $\geq \pm 1.5$-fold change with q-value ≤ 0.01 computed using moderated t-statistics, followed by the False Discovery Rate (FDR) correction using the Benjamini–Hochberg method [56] were considered to be statistically significant differentially expressed genes (DEGs). For all two group comparisons, we used the Student's t-test for computing statistical significance and differences with p-value ≤ 0.05 of were considered to be significant. All statistical analyses were performed using R version 3.6.3.

2.11. Role of funders

The funders did not have any role in the study design, data collection, analysis, interpretation, writing or submission of the manuscript. The corresponding author had complete access to the data and hold final responsibility for the decision to submit for publication.

3. Results

3.1. Description of the blood transcriptome datasets used in the study

We have obtained 56 publicly available whole blood transcriptome datasets from 19 different countries, consisting of 4,259 samples belonging to patients with viral or bacterial infections and healthy controls (Table S1). Of these, seven datasets contained transcriptome profiles of follow-up patients. In all, six datasets that contained viral, bacterial, and matched healthy controls in the same experiment, which we selected for biomarker discovery (Discovery Set) (Fig. 1a) and the remaining 50 datasets were used for validation purposes. About eleven datasets that contain both viral and bacterial infections in the same experiment were considered in the Validation Set-1 (Fig. 1a). All other datasets containing either bacterial or viral samples were considered for independent validation (Validation Set-2). Further, we have used the datasets with follow-up information to study if our test could provide insights on disease recovery. We

Table 1
The clinical characteristics of patient groups in Bangalore - Viral Bacterial (BL-VB). IQR – Inter Quartile Range.

Clinical characteristics	Bacterial	Viral	Indeterminate	Healthy Controls
No. Of Samples	16	14	8	18
Age (Years)Median (IQR)	54 (46–59)	34.50 (27.75–52.75)	51 (46–54. 75)	30 (24.45–32)
Gender Male (M), Female (F)	10M, 6F	7M, 7F	4M, 4F	12M, 6F
Total Leucocyte Count (Cells/cu.mm) Median (IQR)	12500 (8400 - 15875)	5300 (3450 - 8525)	11300 (8800 - 12905)	6650 (5875–8075)
Neutrophils % Median (IQR)	72.9 (64.55–86.3)	65.2 (55.5–71.83)	76.2 (56.73–88.65)	57.15 (51.68–60.73)
Lymphocytes % Median (IQR)	15.2 (9.15–23)	22.65 (14.75–32.75)	18.05 (5.78–30.88)	32.8 (24.43–37.88)
Monocytes % Median (IQR)	6.8 (5.93–7.7)	8.95 (5.13–10)	5.85 (2.98–9.3)	6.95 (5.9–8.03)
Erythrocyte Sedimentation Rate in mm Median (IQR)	60 (43.75–90)	32 (20–39.75)	98 (45–110)	5.5 (4–8.75)

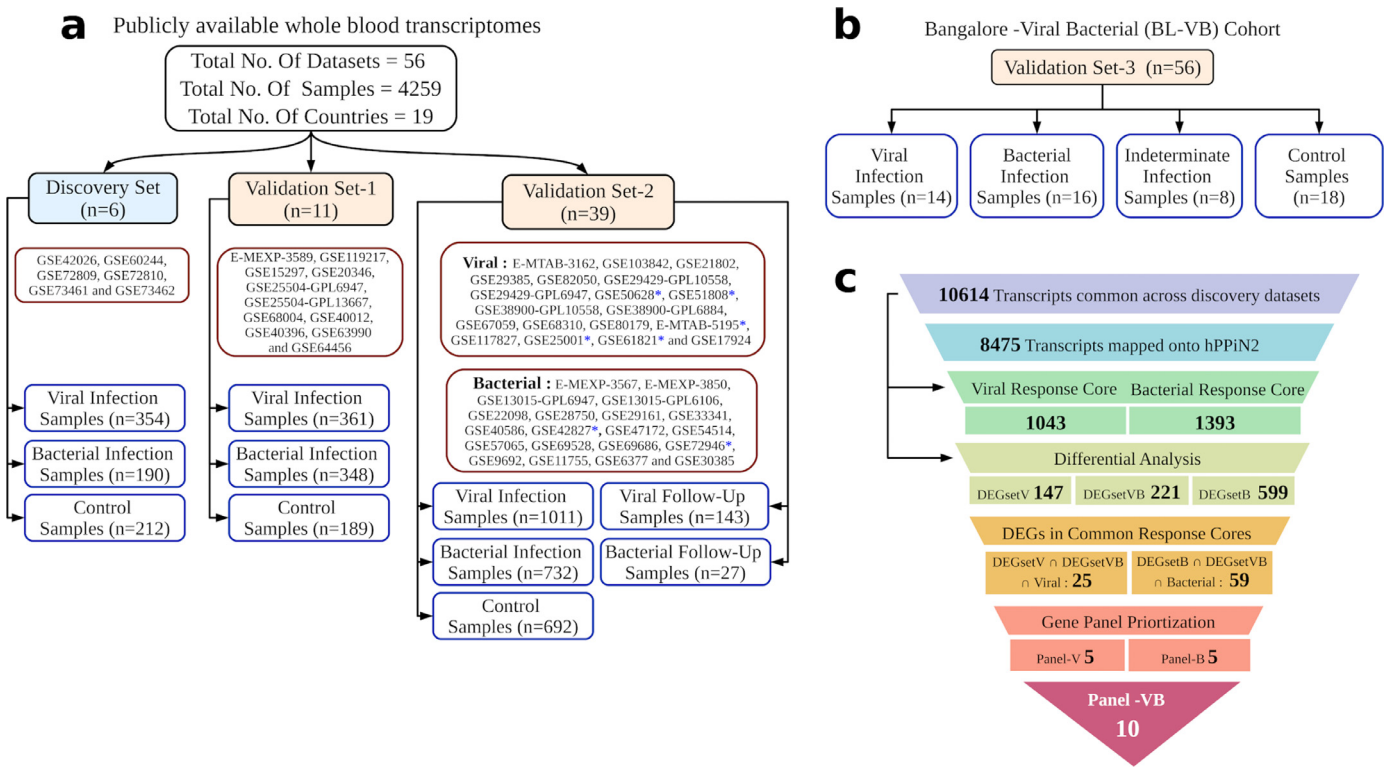


Fig. 1. (a) A flowchart describing the publicly available whole blood transcriptome datasets considered in this study. A total of 4259 whole blood samples belonging to 56 datasets from 19 different countries were considered in this study. Datasets with follow-up information are starred in blue. (b) A flowchart summarizing Bangalore – Viral Bacterial Cohort (BL-VB) generated in this study for external validation. (c) The biomarker discovery pipeline. A funnel describing multiple filters to discover a biomarker panel for accurate discrimination between viral and bacterial infections. The numbers in each step correspond to the number of genes that successfully pass the filter to finally yield a panel of 10 genes.

further evaluated the performance of the signature panel in a newly developed Bangalore-Viral Bacterial cohort (BL-VB) from a South Indian population (Validation Set-3). This cohort contains blood samples from 18 healthy controls and 38 patients belonging to 16 confirmed bacterial, 14 confirmed viral, and 8 indeterminate infection cases (Fig. 1b). Detailed information on the clinical characteristics of patients recruited for BL-VB Cohort is given in Table S2.

3.2. Discovery of a 10-gene panel (Panel-VB) to discriminate between viral and bacterial infections

Briefly, our computational pipeline consists of computing response networks, sensitively mining them to identify top-ranked perturbations and then a series of filters to identify a common viral subnetwork, a common bacterial subnetwork, and symmetric components between the two. Each step in the pipeline serves as a filter and retains only those genes that satisfies the criteria (Fig. 1c) and result in a biomarker signature that can distinguish viral from bacterial infections.

In applying the filters, our first goal was to identify the prominent host responses and to investigate the extent of their similarity in whole blood transcriptomes across different viral diseases, and separately among different bacterial diseases. Our discovery set contained whole blood transcriptomes of 354 patients with confirmed viral infections belonging to six different studies. Differential analysis by comparing the transcriptome profile of acute viral infection patients with their respective healthy controls in different datasets indicated that the number of Differentially Expressed Genes (DEGs) with Fold Change $\geq$ 1.5 & q-value $\leq$ 0.01 approximately ranged from 406 to 1750. Further, an overlap analysis identified 147 common DEGs (DEGsetV) among these datasets (Data file S2), suggestive of substantial similarity in the host response to individual viral infections. Similarly, for bacterial infections, our discovery set contained whole blood transcriptomes of 190 samples from the same six studies. The DEG (Fold Change $\geq$ 1.5 & q-value $\leq$ 0.01) analysis indicated the number of DEGs to be in the range of 1411–2603 for different bacterial infections and about 599 to be common DEGs (DEGsetB) among them (Data file S3), again indicative of commonalities in host response to bacterial infections. Further, to identify the host responses varying between bacterial and viral infection samples, dataset wise differential analysis was performed by comparing viral infection samples with respect to the dataset matched bacterial infection samples. This analysis resulted in DEGs (Fold Change $\geq$ 1.5 & q-value $\leq$ 0.01) ranging from 210 to 1095 for different bacterial vs viral comparisons and about 221 to be common DEGs (DEGsetVB) in at least 50 % of such comparisons in discovery datasets (Data file S4). A comparison between these three categories indicated that about 49 are common between DEGsetV and DEGsetVB, and 103 of them are common between DEGsetB and DEGsetVB (Fig. S2a). Hierarchical clustering of discovery datasets using the resultant of ((DEGsetV $\cap$ DEGsetVB) $\cup$ (DEGsetB $\cap$ DEGsetVB)), which yields 141 genes, is shown in Fig. S2b, indicating the transcriptome alterations to be sufficiently characteristic of each category.

Next, to prioritize the candidate biomarkers from the resultant 141 genes based on their biological relevance for the given disease, we apply our network analysis pipeline to each viral and bacterial disease. This requires (a) a comprehensive knowledge-based molecular interaction network, (b) a method to integrate the transcriptome data into the network, and (c) a sensitive network mining method to extract top-ranked perturbations that occur in different diseases. To address these, we first upgraded our previous human protein-protein interaction network (hPPiN) [39] through adding thousands of signaling and regulatory interactions, curating their directionality, and pruning the previous network to remove any redundant information. This resulted in construction of hPPiN2, which contains 20,183 nodes (proteins) and 255,486 edges (interactions among proteins) (Data file

S5). Using this as a base network, we then construct condition-specific networks by mapping the transcriptome data from the discovery datasets onto hPPiN2 in the form of node and edge-weights using the Eqs. (1)–(3) (described in methods). Our method then sensitively extracts the edge-sequences connecting the nodes (also known as paths) that show the highest alterations in each viral or bacterial disease to an appropriate healthy control cohort. A connected set of such alterations result in a response network which serves as an excellent model to describe the biological response in the host to the given disease [33,39]. The top-active and the top-repressed edges forming separate subnetworks together constitute the top-perturbed networks for each disease. An intersection of all top-perturbed networks across viral diseases yields a common viral response core (Fig. 2a, Data file S6) and likewise an intersection of all top-perturbed networks across bacterial diseases yields a common bacterial response core (Fig. 2b, Data file S7). A unique feature of these perturbed networks is that they contain the most influential DEGs and the genes bridging them directly or indirectly that include influential constitutively expressed genes. This viral response core was observed to contain 1,043 nodes, of which 62 belong to DEGsetV. Similarly, the bacterial response core was found to contain 1393 nodes, of which 287 belong to DEGsetB.

We tested whether the genes in the two response cores were reflecting the known host biology in these diseases by carrying out a pathway enrichment analysis. Towards this, we have identified a set of 215 pathways significantly (q-value $\leq$ 0.01) enriched in the viral response core (Data file S8) and 183 pathways enriched in the bacterial response core (Data file S9). DDX58(RIG-I)-mediated induction of interferon-alpha/beta, cytosolic sensors of pathogen-associated DNA, and antiviral response mediated by IFN-stimulated genes were some key active pathways in viral infections, while the pathways related to the host cell cycle, transcription and translation, surveillance machinery (Nonsense-Mediated Decay), and selenocysteine metabolism were enriched in the most repressed set. Further, the network analysis reveals that the viral core has a giant connected component containing *STAT1, ISG15, EIF2AK2, NOV(CCN3)*, and *LAP3*. On the other hand, the bacterial response core was centered around *STAT3, PPARG,* and *CEBPB* and was significantly enriched with inflammatory processes such as Toll-Like Receptor (TLR) Cascade, neutrophil degranulation, Interleukin-4, and Interleukin-13 signaling. At the same time, pathways such as Programmed cell Death 1 (PD-1) signaling, TCR signaling, Wnt, and Notch Signaling were enriched in the repressed set primarily centered around *LEF1* and *ETS1*. All of these are indeed known to be important in their respective categories, for which there are multiple lines of evidence in the literature. For example, the role of interferon-mediated host antiviral defense [57] and the gene expression changes in the host transcriptional and translational landscapes to subvert host immune response are some known host responses upon viral infections [58,59]. The role of TLRs in pathogen recognition [60,61], neutrophils on extracellular bacterial clearance [62,63], and PD-1 mediated T-cell impairment upon bacterial infection [64] are some known host immune mechanisms observed in bacterial infections. Our response networks correctly capture these known mechanisms in their respective cores.

We then tested specifically if the gold standard genes of viral and bacterial infections retrieved from DisGeNET are captured in the respective response networks and found that there is indeed a significant overlap between the gold standards and genes in the viral (Enrichment score of 2.9, *p*-value: 5.7E−041) and bacterial (Enrichment score of 3.2, *p*-value: 2.10E−23) response cores. The response networks are significantly more enriched with the gold standard genes as compared to the initial DEGsetV (Enrichment score of 2.1 & *p*-value: 9.3E−05) and DEGsetB (Enrichment score of 1.8 & *p*-value: 9.70E−03), illustrating the biological significance of the network models and their power to prioritize crucial DEGs. We thus establish that our response networks are good models to understand the host response to these infections and serve as excellent platforms to identify biomarkers.

From the above analysis, we retained those genes that are common to DEGsetV, DEGsetVB and the viral response core, which results in a set of 25 genes, of which we select top five genes (*IFI27, IFI44, ISG15, MX1, EPSTI1*, referred to as Panel-V), based on a statistical threshold for differential gene expression across all discovery datasets. Similarly, the next filter retains those genes that are common to DEGsetB, DEBsetVB and the bacterial response core to shortlist 59 genes, from which we select five genes (*MMP9, HK3, GYG1, DNMT1*, and *PRF1*, referred to as Panel-B), using the same statistical threshold as for the Panel-V. Finally, we combine Panel-V and Panel-B to obtain a 10-gene panel (Panel-VB) and rigorously test its classification performance. The filtering in this step selects those genes that satisfy the following criteria (a) significantly perturbed in bacterial or viral diseases as compared to their controls, (b) significantly perturbed between viral and bacterial diseases. The genes in the resulting panel (Panel-VB) have known direct or indirect associations with viral or bacterial diseases (Table S4), indicating their biological significance.

3.3. Performance evaluation of Panel-V, Panel-B and Panel-VB

First and foremost, we evaluated the performance of Panel-V and Panel-B to distinguish between (i) viral and healthy controls and (ii) bacterial and healthy controls in the discovery and independent validation datasets.

Panel-V showed a clear separation of viral and healthy controls with a weighted mean AUROC of 0.96 (95% CI: 0.95–0.98) (Fig. S3a) and Panel-B showed a clear separation of bacterial and healthy controls with a weighted mean AUROC of 0.98 (95% CI: 0.97–0.99) (Fig. S4a) in the discovery dataset. Next, we tested the performance of Panel-V in the three independent validation sets (Validation Set-1, Validation Set-2, and Validation Set-3) comprising 1,386 Viral and 580 matched controls and find the panel to have high classification power with a weighted mean AUROC of 0.95 (95% CI: 0.92–0.97) (Figs. S3b–d). Similarly, we tested the performance of Panel-B in Validation Set-1, Validation Set-2 and Validation Set-3 comprising 1,096 bacterial and 526 matched controls which showed a weighted mean AUROC of 0.96 (95% CI: 0.94–0.98) (Figs. S4b–d). This analysis clearly indicates that Panel-V and Panel-B are reflective of viral and bacterial infections and that the combined 10-gene panel (Panel-VB) to be a potential biomarker signature (Panel-VB) to distinguish between viral and bacterial infections.

For Panel-VB, we performed the following tests to evaluate its predictive performance in the datasets containing both viral and bacterial infections such as (a) Discovery Set, (b) Validation Set-1, and (c) Validation Set-3 (an independent validation cohort generated from a South Indian population (BL-VB) containing 16 bacterial and 14 viral samples). ROC analysis of Panel-VB in Discovery showed weighted mean AUROC of 0.97 (95% CI: 0.95–0.98) with a weighted mean sensitivity 0.84 (95% CI: 0.78–0.91) and specificity of 0.95 (95% CI: 0.93–0.97) (Fig. 3a). In case of Validation Set-1, Panel-VB showed weighted mean AUROC of 0.97 (95% CI 0.96–0.99) with a weighted sensitivity 0.93 (95% CI: 0.89–0.96) and specificity of 0.97 (95% CI: 0.95–0.99) (Fig. 3b). Next, we tested the performance of our signature (Panel-VB) in our BL-VB cohort. We found a clear separation of viral from bacterial diseases (AUROC: 1) (Fig. 3c), indicating that the signature performs well for the studied South Indian population as well.

3.4. VB_{10} score formulation

The Panel-VB is clearly seen to be sufficient to separate viral and bacterial infection samples from the predictive performance analysis. Indeed, a clear clustering pattern in the discovery set was observed where all viral datasets were grouped into one category and bacterial

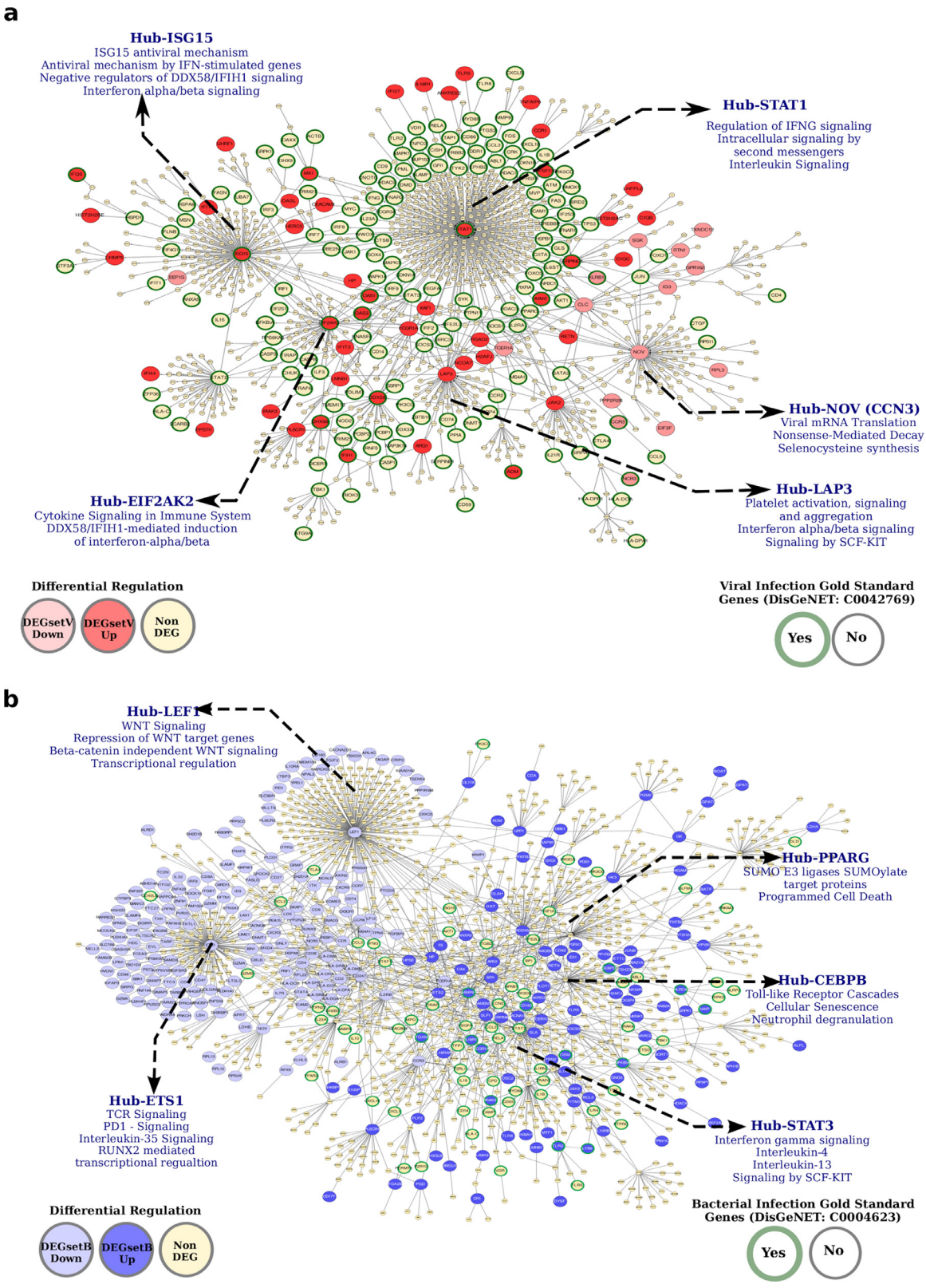


Fig. 2. Networks depicting the 'response cores' in (a) viral and (b) bacterial infections. The networks in each case correspond to the top-ranked perturbations in infection as compared to healthy controls. The viral core consists of 1043 nodes and 1,151 edges, of which 62 belong to DEGsetV (46-up, 15-down, FC $> \pm 1.5$, $q \leq 0.01$) while the bacterial core consists of 1393 nodes, 1845 edges of which 287 belong to DEGsetB (104-up, 183-down, FC $> \pm 1.5$, $q \leq 0.01$). The hubs are labeled by their respective functional categories (from Reactome) obtained through a pathway enrichment analysis of the hub gene and its first neighbors using a hypergeometric test ($q \leq 0.01$).

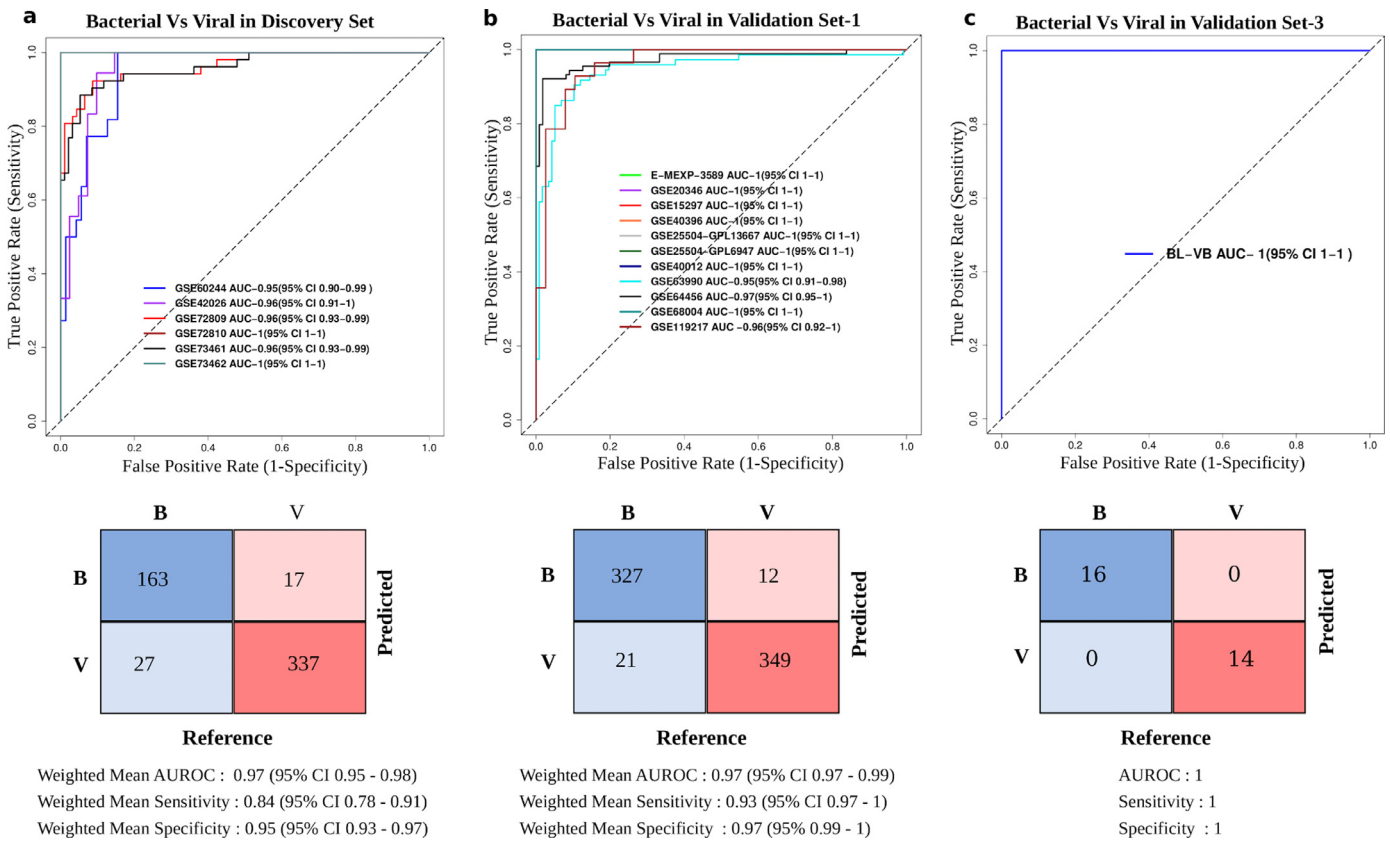


Fig. 3. ROC curves showing the predictive performance of Panel-VB in (a) Discovery Set, (b) Validation Set-1 and (c) Validation Set-3 (BL–VB Cohort). Summary confusion matrix, weighted mean AUROC, weighted mean sensitivity and specificity computed for the respective meta-set is shown in the below panel. AUROC - Area Under the Receiver Operating Characteristics Curve.

into another category (Fig. 4a). As a critical next step towards translation into the clinic, we devised a new score (VB_{10}), which captured the essence of the variation of the gene panel. The expression of the genes in the Panel-VB was combined into a single VB_{10} score for each patient as described in Eq. 4.

$$VB_{10} = [GM(PanelB_{UP}) - GM(PanelV_{UP},\ PanelB_{DOWN})] * \left[\frac{NPanelB_{UP}}{NPanelV_{UP} + NPanelB_{DOWN}}\right] \quad (4)$$

where GM refers to the geometric mean of normalized gene expression values, $PanelB_{UP}$ and $PanelB_{DOWN}$ refer to the upregulated and downregulated Panel-B genes respectively and $PanelV_{UP}$ refers to upregulated Panel-V genes (as compared to healthy controls). $NPanelB_{UP}$, $NPanelV_{UP}$ and $NPanelB_{DOWN}$ indicate the number of genes in the respective set and were used in Eq. (4) to factor in the number of genes considered for computing the score, as per the scaling method described earlier [24]. A stepwise calculation of VB_{10} -score for a representative bacterial and viral sample is shown in Fig. 4b.

3.5. VB_{10} blood test – a diagnostic score to aid clinical decisions

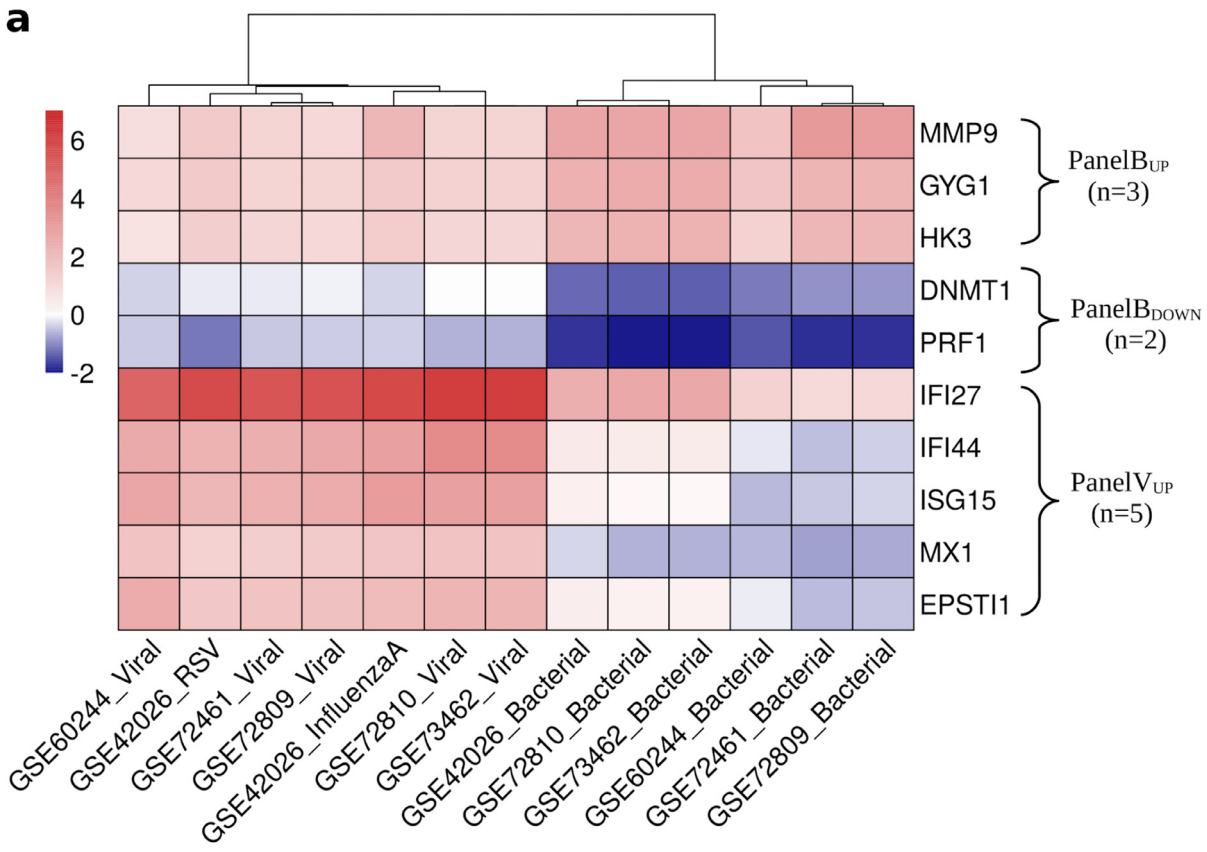
VB_{10}, a standalone score forms the basis for the VB_{10} blood test, as it can be evaluated in individual samples, alleviating the need to compare with healthy controls. The expression of the genes in the Panel-VB was combined into a single VB_{10} score for each patient. The score is devised such that a positive value indicates a bacterial infection whereas a negative value indicates a viral infection (Fig. 5a and b). The global validation of VB_{10} score in the publicly available blood transcriptomes showed a weighted mean AUROC of 0.94 (95% CI: 0.91–0.98), indicating that the score, presented as a single number retains the classification power of the gene signature (Fig. S5a).

Further, in the South Indian Cohort (BL-VB) containing 16 confirmed bacterial and 14 confirmed viral infection samples, VB_{10} scores showed AUC of 1 with sensitivity of 0.94 and specificity of 1 (Fig. 5c). Finally, we have computed probabilities for the VB_{10} score using the 2996 publicly available whole blood transcriptome samples belonging to patients with viral and bacterial infections and provide a measure of confidence to interpret a score report of any given sample (Fig. 5d). Our analysis indicates that a VB_{10} score of >0.5 indicates a bacterial infection with a probability >0.8, whereas a VB_{10} score >1.0 indicates a bacterial infection with a probability >0.9. Similarly, a VB_{10} score of -0.5 or lower indicates a viral infection with a probability of >0.95 whereas a VB_{10} score of -1.0 or lower indicates a viral infection with an even higher probability (of 0.97). This brings out a question of what range of scores are seen in healthy subjects. To address this, we plotted the distribution of VB_{10} scores for the pool of 1,093 healthy controls present in our study datasets. The plot clearly indicates that a majority of the healthy samples show VB_{10} scores ranging from -0.25 to $+0.5$ (Fig. S5b), centered around a median value of 0 indicating them to be of neither viral nor bacterial infections (Fig. 5d).

3.6. Performance of VB_{10}-score in different clinical scenario

Next, we analyze how our score performs in a range of clinical scenarios,

(a) ***Indeterminate infection – samples with unconfirmed diagnosis:*** In a few cases, based on the clinical presentation, the sample can only be labeled as a suspected bacterial or suspected viral, but the diagnosis is often unconfirmed. From the BL-VB cohort, we had 8 samples of this nature and refer to them as the indeterminate infection category. All 8 were culture negative.



b

VB$_{10}$ - Score Computation

Panel-VB	HK3	GYG1	MMP9	DNMT1	PRF1	MX1	IFI27	IFI44	ISG15	EPSTI1
Normalised gene expression values (Bacterial)	12.58	13.11	14.26	6.98	9.01	8.05	4.64	5.43	7.5	6.2
Normalised gene expression values (Viral)	8.77	8.74	6.9	9.85	10.73	13.3	11.91	11.6	13.4	11.8
			PanelB$_{UP}$		**PanelB$_{DOWN}$**				**PanelV$_{UP}$**	

VB$_{10}$ - score calculation = (GM(PanelB$_{UP}$) – GM(PanelB$_{DOWN}$, PanelV$_{UP}$)) x (NpanelB$_{UP}$ /(NPanelB$_{DOWN}$ + NpanelV$_{UP}$)) **(Eq. 4)**

VB$_{10}$ - score = ((HK3 x GYG1 x MMP9)$^{1/3}$ – (DNMT1 x PRF1 x MX1 x IFI27 x IFI44 x ISG15 x EPSTI1)$^{1/7}$) x (3/7)

VB$_{10}$ (Bacterial infection) = ((12.58 x 13.11 x 14.26)$^{1/3}$ – (6.98 x 9.01 x 8.05 x 4.64 x 5.43 x 7.5 x 6.2)$^{1/7}$) x (3/7)

= (13.31 – 6.68) x (0.43) = 6.63 x 0.43 = **2.85**

VB$_{10}$ (Viral infection) = ((8.77 x 8.74 x 6.9)$^{1/3}$ – (9.85 x 10.73 x 13.3 x 11.91 x 11.6 x 13.4 x 11.8)$^{1/7}$) x (3/7)

= (8.09 – 11.74) x (0.43) = **-1.57**

Fig. 4. VB$_{10}$- score formulation. (a) A heatmap showing the differential transcriptome profile of Panel-V and Panel-B genes in the Discovery Set. The figure shows a clear and distinct clustering of known viral and bacterial samples. '*lmfitted*' coefficients of viral and bacterial differential transcriptomes with reference to their matched controls from the respective discovery datasets were used for generating the heatmap. *HK3, GYG1* and *MMP9* constitute PanelB$_{UP}$; *DNMT1* and *PRF1* form PanelB$_{DOWN}$, whereas *IFI27, IFI44, MX1, ISG15* and *EPSTI1* form PanelV$_{UP}$. (b) An illustration showing the stepwise computation of VB$_{10}$- score for a sample bacterial and viral cases.

For these samples, we measured the transcript abundances using the nanostring technology (and subsequently confirmed through qRT-PCR for a subset of these samples) (Table S5). Our VB$_{10}$ score identified 6 of them as clearly bacterial and 2 of them as viral (Fig. 5c; Table S6), which were consistent with subsequent clinical investigations including hemograms, serology tests and response to antibiotic treatment.

(b) ***Recovery-*** We tested if our score is capable of reflecting recovery from infection. From the pool of datasets included in this study, eight datasets (bacterial: GSE42827, GSE72946 &

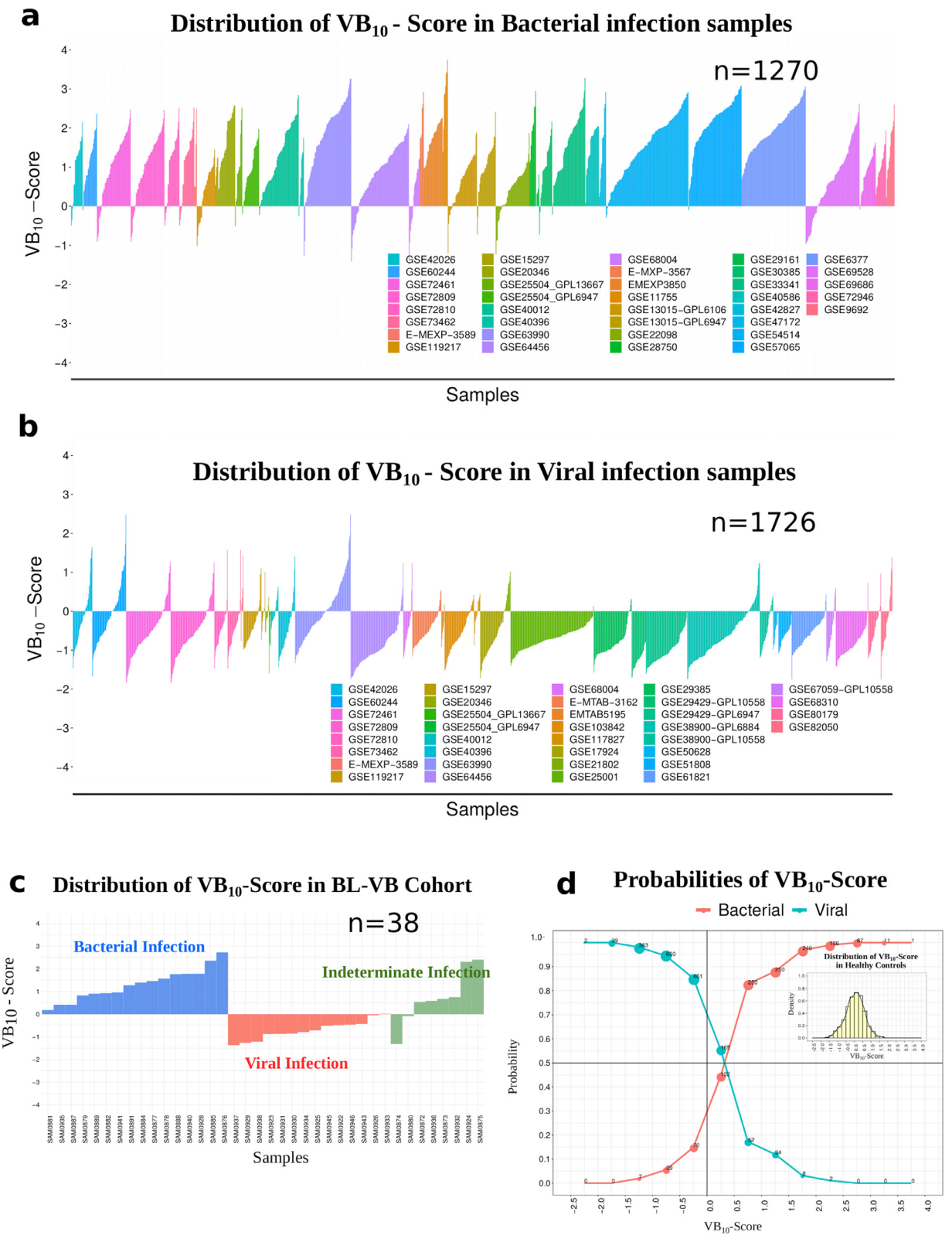


Fig. 5. Evaluation of the VB_{10}-Score. (a) A waterfall plot showing the VB_{10} -scores in 1270 publicly available bacterial infection samples from 37 datasets, with samples from each dataset sorted by their VB_{10} scores and each dataset was indicated by different color (legend in the inset). (b) A similar plot for 1726 publicly available viral infection samples. The 36 datasets are indicated in different colors (legend in the inset) and samples in each are sorted by their VB_{10} scores. (c) A similar plot for VB_{10} – Scores in the BL-VB Cohort (38 samples: Bacterial, Viral, and indeterminate infection category). Color coding is based on the infection category. Sample labels are shown in the x-axis. Those in green represent samples with clinically unconfirmed diagnosis. (d) Joint Probability Density computed from the VB_{10}-Scores of publicly available viral (represented in cyan) and bacterial (red) infection samples. The numbers in the circle correspond to the samples belonging to that bin. Distribution of VB_{10}-Score for the healthy controls is provided in the inset.

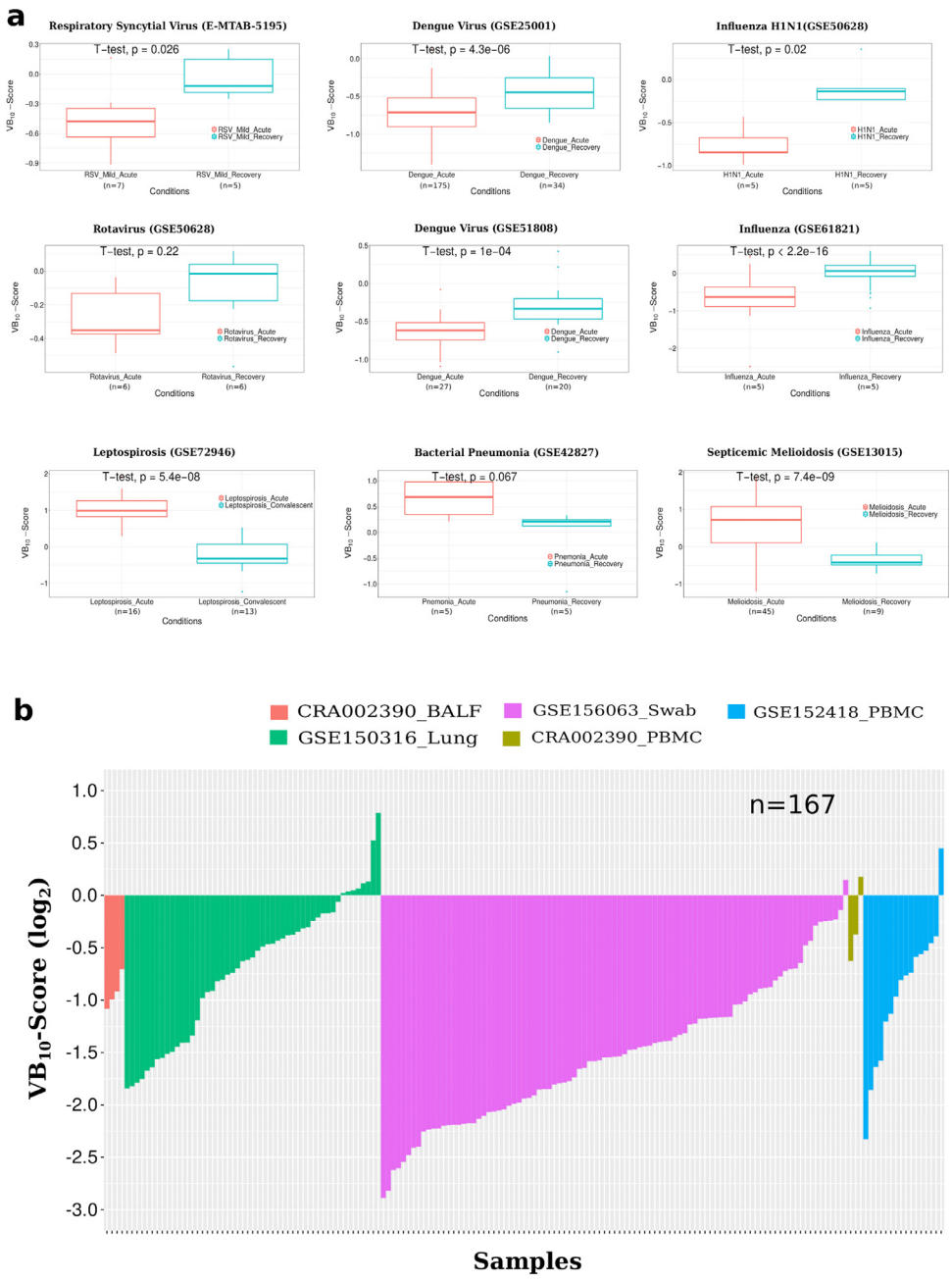


Fig. 6. Performance of VB_{10} -Score in different clinical scenarios. (a) Boxplot showing the VB_{10} -scores in the acute infection and the respective recovery data for the publicly available viral and bacterial infection samples, with the significance computed using the student *t*-test. (b) A waterfall plot showing the VB_{10} -scores in the publicly available COVID-19 samples (n = 167) from four different datasets. Each dataset is represented by different color, corresponding to samples infected with COVID-19. . Peripheral blood mononuclear cell (PBMC) and bronchoalveolar lavage fluid (BALF) patient samples from CRA2002390 dataset are shown in different colors. Samples in each study are sorted by their VB_{10} scores.

GSE13015 and viral: E-MTAB-5195, GSE25001, GSE50628, GSE51808 & GSE61821) contained clinical parameters indicative of recovery. We find that our VB_{10} score in these datasets showed the expected trend in all cases (Fig. 6a), indicating that the score captured the recovery status from the infection.

(c) ***Performance evaluation of VB_{10} in Non-infectious controls*:** Non-infectious controls are the most relevant control group since they represent the population in whom testing would occur. Hence, we evaluated the performance of VB_{10} in discriminating Viral/ Bacterial from non-infectious controls (asthma, COPD, non-infectious sepsis/SIRS). Our results show that VB_{10} significantly differentiates (a) bacterial from pathological matched controls and (b) viral from pathological matched controls with AUC-ROC of 0.83 (95% CI 0.81 – 0.85) and 0.89 (95% CI 0.88 – 0.90), respectively in the validation cohorts (Fig. S6a, 6b, 6c).

(d) ***Performance evaluation of VB_{10} in different age groups* -** Differentiating between bacterial and viral infections among different age groups of patients is often a critical requirement in the clinic. The publicly available datasets that we have analyzed in this study included several neonatal, infant, pediatric and adult samples. We find that our VB_{10} score in the validation datasets (Validation Set-1 and Set-2) show high diagnostic accuracy to distinguish bacterial from viral infections in neonates with AUROC of 0.99 (95% CI 0.95–1), infant with AUROC of 0.95 (95% CI 0.93–0.98), pediatric with AUROC of 0.91 (95% CI 0.88–0.95) and adult with AUROC of 0.96 (95% CI 0.95–0.97) (Fig. S7). This

strongly indicates that the score performs well in all age groups.

(e) ***Disease spectrum*** - We analyzed how our score fares for different bacterial and viral diseases and hence analyzed the disease spectrum covered by the available data. The datasets that we have analyzed, put together were associated with about 12 diseases which includes acute respiratory infections, bronchiolitis, chronic obstructive pulmonary disease, chronic kidney disorder, dengue fever, febrile illness, gastroenteritis, infective endocarditis, leptospirosis, meningitis, pneumonia, and sepsis. The bacterial etiologies included *Staphylococcus, Streptococcus, Chlamydophila, Burkholderia, Leptospira, Neisseria, Acinetobacter, Escherichia coli, Citrobacter, Pseudomonas and Proteus*, while the viral etiologies included Influenza, Respiratory Syncytial Virus, Adenovirus, Human coronavirus, Human metapneumovirus, Human Herpesvirus 6, Enterovirus, Cytomegalovirus, Rhinovirus and Dengue virus. Samples from these, form a part of the data analyzed in Fig. 5a and b. It is clear from the figures that the VB_{10} score shows high performance across different viral and bacterial etiologies in a broad class of disease. In this study, we have excluded atypical bacterial (eg., *Mycobacterium tuberculosis* and salmonella) for two main reasons (i) the immune response elicited by the host towards these pathogens are markedly different from the acute viral and bacterial infections and (ii) there are clear tests available for diagnosing these and therefore, clinically there is no compelling requirement for including these in the general VB_{10} score.

(f) ***COVID-19:*** At present, there is an ongoing pandemic due to SARS-CoV-2 infection (COVID-19) that has been causing a very large number of deaths globally and considerable disruption to normal activities world over [65,66]. We evaluated if our score could be useful in detecting COVID-19 infections using the publicly available patient transcriptome data capturing host response to SARS-CoV-2. Towards this, we considered four publicly available bulk transcriptome datasets (CRA002390, GSE150316, GSE156063 and GSE152418) containing 167 COVID-19 samples from different sample sources [67,68]. Raw counts of the respective datasets were normalized by size factors using DESeq2 package in R [69]. Next, we computed patient-wise VB_{10}-score by taking the fold variation in expression of the genes in our panel-VB. We find that the score clearly indicates a viral infection in almost all cases and with > 0.95 probability (Fig. 6b). This suggests that the VB_{10} score could be tested for differentiating between COVID-19 infections from common bacterial respiratory infections.

3.7. Benchmarking against prior biomarker panels with associated diagnostic scores

Among the various panels that have been reported so far [18,20,22–24,70,71], only two of them contains < 10 genes and have diagnostic scores associated with them. The scores enable testing the biomarkers on individual samples and increase their readiness for implementation in the clinic. We report a rigorous comparison of the performance of our VB_{10} score, the underlying Panel-V, B and VB in 2,996 samples from 56 datasets with the two prior panels and their scores. The first is a seven gene based bacterial/viral metascore (hereafter this gene panel (and score) will be referred to as Sweeney7 (Sweeney7-Score)) that the authors have used for distinguishing viral from bacterial infections in sepsis [24]. The second, the Disease Risk Score (DRS) based on *FAM89A* and *IFI44L* (hereafter this gene panel (and its score) will be referred to as Herberg2 (Herberg2-Score)) [18], which the authors have used for a similar purpose in pediatric febrile illness. 2 genes *IFI27* and *HK3* from the Sweeney7 panel are also a part of our Panel-VB, while there is no overlap with the Herberg2 panel. To test how our Panel-VB fares in comparison to these panels, we computed standard classification metrics of all three signatures for the validation datasets. We found that Panel-VB fared well in terms of accuracy, sensitivity, specificity, and AUC in comparison to the other two signature panels (Table S7). The performance of the sub-panels Panel-V and Panel-B in the Validation Set-1 and Validation Set-2 datasets are clearly better as compared to the corresponding panels from the previous two signatures (Tables. S8, S9). Score level comparison demonstrates VB_{10} score is performed in par with Sweeney7-Score and better than Herberg2-Score in terms of specificity (Data file S10).

As clear from the discussion so far, different computational approaches yield different panels, as their identification is based on different perspectives. This in fact illustrates the need for probing transcriptome datasets with independent approaches. Our network approach uses an unbiased screening of the transcriptome to identify the panels and yet, most of the genes in the Sweeney7 and Herberg2 panels were absent in our final list. We carried out a systematic evaluation at each step of the pipeline to determine the step at which they were eliminated (Table 2). Except for *HK3* and *IFI27* from Sweeney7, all other genes failed to satisfy at least one of the three filters. Besides *IFI27*, other viral markers from both these panels were not present in our viral response core and were not significantly differentially expressed in all the viral diseases. The bacterial markers from these panels, although formed a part of our bacterial response core, failed to show significant differential expression in comparison with healthy controls as well viral vs bacterial comparisons.

Overall, our signature, which was independently derived and different from the first two, shows high accuracy and improved specificity as compared to Sweeney7 and improved in both sensitivity and accuracy as compared to Herberg2.

4. Discussion

Whole blood transcriptomes in different diseases have consistently indicated high promise as diagnostic biomarkers. This holds for the problem being investigated in this work, which is to discriminate bacterial from viral infections, as several studies have described distinct host response patterns to these two disease classes [15–17]. The next logical step is to push towards translation and facilitate their clinical use. Several critical issues must be addressed before a

Table 2
Assessment of genes in prior signatures in the current biomarker discovery pipeline. A cross(X) indicates not meeting the criteria.

	Viral Markers					Bacterial Markers				
	Sweeney7			Herberg2		Sweeney7				Herberg2
Biomarkers	IFI27	JUP	LAX1	IFI44L	Biomarkers	HK3	TNIP1	GPAA1	CTSB	FAM89A
Transcripts common across discovery datasets	√	√	√	√	Transcripts common across discovery datasets	√	√	√	√	√
Transcripts mapped onto hPPiN−V2.0	√	√	√	√	Transcripts mapped onto hPPiN − V2.0	√	√	√	√	√
Viral Response Core	√	X	X	X	Bacterial Response Core	√	√	√	√	X
DEGsetV (V Vs HC)	√	X	X	√	DEGsetB (B Vs HC)	√	X	X	X	X
DEGsetVB (V Vs B)	√	√	√	√	DEGsetVB (V Vs B)	√	X	X	X	X
Panel -V	√	**X**	**X**	**X**	**Panel -B**	√	**X**	**X**	**X**	**X**

biomarker discovery can translate to clinical use, which include (a) establishing the need for a biomarker and defining the context, (b) establishing the ability of the biomarker to achieve acceptable diagnostic accuracy (given the clinical context of interest), (c) demonstrating sufficient generality - in particular a biomarker should show high accuracy in a population where it is intended to be used and (d) making it accessible as a simple readout to the clinician, for it to be a candidate for routine clinical use. Our work meets all these requirements. The need for a biomarker to distinguish between viral and bacterial infections is acute and evident from the growing burden of AMR. The clinical context is clear too as a good biomarker can assist the clinician in deciding whether to prescribe antibiotics and have a far-reaching effect on making therapy more effective and safer. The need is the highest in developing countries like India [7,72].

In this work, we have discovered a 10 gene marker panel and tested its performance for detecting viral and bacterial infections, and discriminating between them with high accuracy, sensitivity, and specificity. Based on the panel, we develop a new diagnostic score and show that our score can correctly detect if the infection in a given sample is due to viral or bacterial etiologies in more than 2,996 cases in all. An ultimate test to assess the clinical utility of the diagnostic score is to measure its ability to guide decision-making in terms of whether or not to prescribe antibiotics. In this study, we do this retrospectively and show that if we were to use our score as a diagnostic test, we would be able to match the diagnosis and the decision made by a clinician in almost all cases. A current limitation is that our score has not been tested for identifying co-infections. To test it in co-infection scenarios, we would require information on the primary infection and the superinfection for each sample. Such information is not available for the datasets that are publicly available, and it was therefore not included in our objectives. However, the individual panels (Panel-V and Panel-B) are likely to be useful in detecting the co-infection status.

Genetic heterogeneity and biological variability are major factors that limit the progression of candidate biomarkers to the clinic. Our method that includes the use of networks to model the host response to infections as an early step, largely addresses these limitations. Network-based biomarker selection methods have been shown to be naturally resistant to batch variation, making them highly effective with high reproducibility [28,73]. Evaluation of our signature on multiple ethnicities and populations, especially including those where it is intended to be used, addresses the problem posed by genetic heterogeneity. Identifying a specific gene panel and studying large meta-datasets from multiple cohorts alleviate the problem of biological variability, which can be due to a multitude of confounding factors. A biomarker must show variations at a level over and above the variations due to these confounders. A single gene as a biomarker is rarely sufficient for catering to a wide cross-section of people or multiple populations as it is unlikely to be a clear DEG in all patients. Instead, the combined effect of a panel of genes has higher promise as a biomarker, since in any given patient, at least some genes in the panel are highly likely to exhibit expected variations.

Finally, focusing on mechanistically relevant genes in the panel reduces the chance of failure in predicting clinical behavior. Our multi-gene biomarker Panel-VB comprises *MX1, EPSTI1, ISG15, IFI27* and *IFI44* as being characteristic of viruses while five others are characteristic of bacterial infections, comprising *GYG1, MMP9, HK3, DNMT1* and *PRF1*. The role of guanosine triphosphate (GTP)-metabolizing (*MX1*), Interferon Alpha Inducible Protein 27 (*IFI27*) and Interferon Induced Protein 44 (*IFI44*) in cellular antiviral response against a wide range of RNA and DNA viruses is well established [57,74]. Epithelial Stromal Interaction 1 (*EPSTI1*), an IL-28A-mediated interferon-inducible gene is known to mediate antiviral activity through RNA-dependent protein kinase (PKR) genes [75]. Glycogenin 1 (*GYG1*), involved in glycogen synthesis, is known to be a part of a neonatal immune-metabolic network associated with bacterial infections [76,77]. Matrix metalloproteinase 9 (*MMP9*), a member of a family of proteolytic enzymes is known to perform multiple roles in the immune response to infection and has been paradoxically linked to the degradation of the extracellular matrix, gelatinases, and collectins, leading to a loss of its innate immune functions including aggregation of bacteria and phagocytosis [78,79]. Hexokinase 3 (*HK3*), that is selectively expressed in hematopoietic cells and subsets of immune cells is an innate immune receptor, acts as an innate sensor during bacterial infection. It recognizes sugars from bacterial peptidoglycans and dissociates it from the mitochondrial outer membrane, triggering the downstream activation of inflammasome [80]. DNA methyltransferase 1 (*DNMT1*) is involved in maintenance and propagation of DNA methylation patterns to the newly synthesized strands. DNA methylation is known to be a transcriptional regulator of the immune system and have a critical role in T cell development, function, and survival [81]. Perforin 1 coded by *PRF1* is essential for secretory granule-dependent cell death, and combat pathogen load in a variety of infections [82].

Overall, we present a new RNA based biomarker signature and a new blood test to distinguish between viral and bacterial infections that can guide a physician in choosing an optimal treatment plan including a decision of whether to prescribe antibiotics. In a clinical setting, we believe this test will help enable the judicious use of antibiotics and reduce the AMR burden.

Contributors

NC: Conceptualization, Funding acquisition, Project administration, Investigation, Supervision, Writing-orginal draft, Writing-review & editing. SR: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing- review & editing. UB: Methodology, Validation. GD and RK: Resources, Validation. CT: Validation. AS, DC, KNB: Methodology, Resources.

Data sharing statement

The study design, protocol and statistical analysis are provided in the main manuscript and the supplementary data files. The access to the data generated and analysed in this study will be provided upon reasonable request to the corresponding author.

Declaration of Competing Interest

NC and SR have obtained a provisional patent for Panel-VB and VB_{10}- score (IN Application No: 202041015738). NC is a co-founder of qBiome Research Pvt Ltd and Healthseq Precision Medicine Pvt Ltd, which have no role in this manuscript. The other authors have no conflicts to disclose.

Acknowledgments

The authors thank Grand Challenges India, Biotechnology Industry Research Assistance Council (BIRAC), Department of Biotechnology, Govt. of India for the funding. KNB thanks Science and Engineering Research Board (SERB), DST for the award of J.C. Bose National Fellowship (No. SB/S2/JCB-025/2016 dated 25.7.15). We also thank the patients and healthy control volunteers who contributed clinical samples for performing this study and the researchers who made their data publicly available. We acknowledge theraCUES, Bangalore, India for performing nanostring. We also thank Dr. Madhulika Mishra, formerly in our laboratory and currently in the European Bioinformatics Institute for critically reading the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2021.103352.

References

[1] Bloom DE, Cadarette D. Infectious disease threats in the twenty-first century: strengthening the global response. Front Immunol 2019;10 [cited 2020 Apr 6] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6447676/.

[2] Kennedy JL, Haberling DL, Huang CC, Lessa FC, Lucero DE, Daskalakis DC, et al. Infectious Disease Hospitalizations: United States, 2001 to 2014. Chest 2019;156(2):255–68.

[3] Zheng J. SARS-CoV-2: an emerging coronavirus that causes a global threat. Int J Biol Sci 2020;16(10):1678–85.

[4] Aabenhus R, Hansen MP, Saust LT, Bjerrum L. Characterisation of antibiotic prescriptions for acute respiratory tract infections in Danish general practice: a retrospective registry based cohort study. NPJ Primary Care Respir Med 2017;27(1):37.

[5] Hecker MT, Aron DC, Patel NP, Lehmann MK, Donskey CJ. Unnecessary use of antimicrobials in hospitalized patients: current patterns of misuse with an emphasis on the antianaerobic spectrum of activity. Arch Intern Med 2003;163(8):972–8.

[6] Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, et al. Antibiotic resistance-the need for global solutions. Lancet Infect Dis 2013;13(12):1057–98.

[7] Laxminarayan R, Chaudhury RR. Antibiotic resistance in india: drivers and opportunities for action. PLoS Med 2016;13(3) [cited 2020 Mar 21]Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4775002/.

[8] Sherwin R, Winters ME, Vilke GM, Wardi G. Does early and appropriate antibiotic administration improve mortality in emergency department patients with severe sepsis or septic shock? J Emerg Med 2017;53(4):588–95.

[9] Daniel P, Rodrigo C, Mckeever TM, Woodhead M, Welham S, Lim WS. Time to first antibiotic and mortality in adults hospitalised with community-acquired pneumonia: a matched-propensity analysis. Thorax 2016;71(6):568–70.

[10] Bloomfield MG, Balm MND, Blackmore TK. Molecular testing for viral and bacterial enteric pathogens: gold standard for viruses, but don't let culture go just yet? Pathology 2015;47(3):227–33.

[11] Riley PA. Principles of microscopy, culture and serology-based diagnostics. Medicine 2017;45(10):639–44.

[12] Korppi M, Heiskanen-Kosma T, Leinonen M. White blood cells, C-reactive protein and erythrocyte sedimentation rate in pneumococcal pneumonia in children. Eur Respir J 1997;10(5):1125–9.

[13] Reinhart K, Bauer M, Riedemann NC, Hartog CS. New approaches to sepsis: molecular diagnostics and biomarkers. Clin Microbiol Rev 2012;25(4):609–34.

[14] Farooq A, Colón-Franco JM. Procalcitonin and its limitations: why a biomarker's best isn't good enough. J Appl Lab Med 2019;3(4):716–9.

[15] Holcomb ZE, Tsalik EL, Woods CW, McClain MT. Host-based peripheral blood gene expression analysis for diagnosis of infectious diseases. J Clin Microbiol 2017;55(2):360–8.

[16] Lydon EC, Ko ER, Tsalik EL. The host response as a tool for infectious disease diagnosis and management. Expert Rev Mol Diagn 2018;18(8):723–38.

[17] Ramilo O, Mejías A. Shifting the paradigm: host gene signatures for diagnosis of infectious diseases. Cell Host Microbe 2009;6(3):199–200.

[18] Herberg JA, Kaforou M, Wright VJ, Shailes H, Eleftherohorinou H, Hoggart CJ, et al. Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. JAMA 2016;316(8):835–45.

[19] Lydon EC, Henao R, Burke TW, Aydin M, Nicholson BP, Glickman SW, et al. Validation of a host response test to distinguish bacterial and viral respiratory infection. EBioMedicine 2019;48:453–61.

[20] Mahajan P, Kuppermann N, Mejias A, Suarez N, Chaussabel D, Casper TC, et al. Association of RNA biosignatures with bacterial infections in febrile infants aged 60 days or younger. JAMA 2016;316(8):846–57.

[21] Mayhew MB, Buturovic L, Luethy R, Midic U, Moore AR, Roque JA, et al. A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. Nat Commun 2020:11. [cited 2020 Oct 5]Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7055276/.

[22] Parnell GP, McLean AS, Booth DR, Armstrong NJ, Nalos M, Huang SJ, et al. A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia. Crit Care 2012;16(4):R157.

[23] Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, et al. Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood 2007;109(5):2066–77.

[24] Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Sci Transl Med 2016;8(346):346ra91.

[25] Tang BM, Shojaei M, Parnell GP, Huang S, Nalos M, Teoh S, et al. A novel immune biomarker IFI27 discriminates between influenza and bacteria in patients with suspected respiratory infection. Eur Respir J 2017;49(6).

[26] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207–10.

[27] Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 2007;35(Database issue):D747–50.

[28] Goh WWB, Wong L. Advancing clinical proteomics via analysis based on biological complexes: a tale of five paradigms. J Proteome Res 2016;15(9):3167–79.

[29] Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods 2015;12(3):179–85.

[30] Geller G, Dvoskin R, Thio CL, Duggal P, Lewis MH, Bailey TC, et al. Genomics and infectious disease: a call to identify the ethical, legal and social implications for public health and clinical practice. Genome Med 2014;6(11):106.

[31] McDermott JE, Wang J, Mitchell H, Webb-Robertson B-J, Hafen R, Ramey J, et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin Med Diagn 2013;7(1):37–51.

[32] Metri R, Mohan A, Nsengimana J, Pozniak J, Molina-Paris C, Newton-Bishop J, et al. Identification of a gene signature for discriminating metastatic from primary melanoma using a molecular interaction network approach. Sci Rep 2017;7(1):17314.

[33] Sambarey A, Devaprasad A, Mohan A, Ahmed A, Nayak S, Swaminathan S, et al. Unbiased identification of blood-based biomarkers for pulmonary tuberculosis by modeling and mining molecular interaction networks. EBioMedicine 2017;15:112–26.

[34] Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics 2010;26(19):2363–7.

[35] Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004;20(3):307–15.

[36] Gharaibeh RZ, Fodor AA, Gibas CJ. Background correction using dinucleotide affinities improves the performance of GCRMA. BMC Bioinform 2008;9 Available from:. doi: 10.1186/1471-2105-9-452.

[37] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43(7):e47.

[38] Smyth GK. limma: linear models for microarray data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinformatics and computational biology solutions using r and bioconductor. New York: Springer; 2005 Available from:. doi: 10.1007/0-387-29362-0_23..

[39] Sambarey A, Devaprasad A, Baloni P, Mishra M, Mohan A, Tyagi P, et al. Meta-analysis of host response networks identifies a common core in tuberculosis. NPJ Syst Biol Appl 2017;3 [Internet].Available from:. doi: 10.1038/s41540-017-0005-4.

[40] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44 Available from:. doi: 10.1093/nar/gkv1070.

[41] Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods 2016;13(12):966–7.

[42] Fazekas D, Koltai M, Türei D, Módos D, Pálfy M, Dúl Z, et al. SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. BMC Syst Biol 2013;7:7.

[43] Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database 2016 [cited 2020 Jun 18]Available from: https://academic.oup.com/database/article/doi/10.1093/database/baw100/2630482.

[44] Liu Z-P, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database 2015:2015. [cited 2020 Aug 27]Available from: https://academic.oup.com/database/article/doi/10.1093/database/bav095/2433227.

[45] Bovolenta LA, Acencio ML, Lemke N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC Genom 2012;13(1):405.

[46] Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, et al. TRRUST: a reference database of human transcriptional regulatory interactions. Sci Rep 2015;5:11432.

[47] Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, et al. TFCat: the curated catalog of mouse and human transcription factors. Genome Biol 2009;10(3):R29.

[48] Ravichandran S, Chandra N. Interrogation of genome-wide networks in biology: comparison of knowledge-based and statistical methods. Int J Adv Eng Sci Appl Math 2019;11(2):119–37.

[49] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13 Available from:. doi: 10.1101/gr.1239303.

[50] Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, et al. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinform 2017;18(1):142.

[51] Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017;45 Available from:. doi: 10.1093/nar/gkw943.

[52] Rivals I, Personnaz L, Taing L, Potier M-C. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 2007;23(4):401–7.

[53] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform 2011;12:77.

[54] Li J, Fine JP. Weighted area under the receiver operating characteristic curve and its application to gene selection. J R Stat Soc Ser C Appl Stat 2010;59(4):673–92.

[55] Warsinske H, Vashisht R, Khatri P. Host-response-based gene signatures for tuberculosis diagnosis: a systematic comparison of 16 signatures. PLoS Med 2019;16(4):e1002786.

[56] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B 1995:57.

[57] Samuel CE. Antiviral actions of interferons. Clin Microbiol Rev 2001;14(4):778–809 table of contents.

[58] Gale M, Tan SL, Katze MG. Translational control of viral gene expression in eukaryotes. Microbiol Mol Biol Rev 2000;64(2):239–80.
[59] Herbert KM, Nag A. A tale of two RNAs during viral infection: how viruses antagonize mRNAs and small non-coding RNAs in the host cell. Viruses. 2016;8(6).
[60] Oliveira-Nascimento L, Massari P, Wetzler LM. The role of TLR2 in infection and immunity. Front Immunol 2012;3:79.
[61] Takeuchi O, Hoshino K, Kawai T, Sanjo H, Takada H, Ogawa T, et al. Differential roles of TLR2 and TLR4 in recognition of gram-negative and gram-positive bacterial cell wall components. Immunity 1999;11(4):443–51.
[62] Davey MS, Lin C-Y, Roberts GW, Heuston S, Brown AC, Chess JA, et al. Human neutrophil clearance of bacterial pathogens triggers anti-microbial $\gamma\delta$ T cell responses in early infection. PLoS Pathog 2011;7(5):e1002040.
[63] Kobayashi SD, Malachowa N, DeLeo FR. Neutrophils and bacterial immune evasion. J Innate Immun 2018;10(5−6):432–41.
[64] Huang X, Venet F, Wang YL, Lepape A, Yuan Z, Chen Y, et al. PD-1 expression by macrophages plays a pathologic role in altering microbial clearance and the innate inflammatory response to sepsis. PNAS 2009;106(15):6303–8.
[65] Roberton T, Carter ED, Chou VB, Stegmuller AR, Jackson BD, Tam Y, et al. Early estimates of the indirect effects of the COVID-19 pandemic on maternal and child mortality in low-income and middle-income countries: a modelling study. Lancet Glob Health 2020;8(7):e901–8.
[66] Nicola M, Alsafi Z, Sohrabi C, Kerwan A, Al-Jabir A, Iosifidis C, et al. The socio-economic implications of the coronavirus pandemic (COVID-19): a review. Int J Surg 2020;78:185–93.
[67] Cavalli E, Petralia MC, Basile MS, Bramanti A, Bramanti P, Nicoletti F, et al. Transcriptomic analysis of COVID-19 lungs and bronchoalveolar lavage fluid samples reveals predominant B cell activation responses to infection. Int J Mol Med 2020;46(4):1266–73.
[68] Xiong Y, Liu Y, Cao L, Wang D, Guo M, Jiang A, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg Microbes Infect 2020;9(1):761–70.
[69] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
[70] 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(2):213–22.
[71] Tsalik EL, Henao R, Nichols M, Burke T, Ko ER, McClain MT, et al. Host gene expression classifiers diagnose acute respiratory illness etiology. Sci Transl Med 2016;8(322):322ra11.
[72] Ganguly NK, Arora NK, Chandy SJ, Fairoze MN, Gill JPS, Gupta U, et al. Rationalizing antibiotic use to limit antibiotic resistance in India. Indian J Med Res 2011;134:281–94.
[73] Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol 2017;35(6):498–507.
[74] Haller O, Kochs G. Human MxA protein: an interferon-induced dynamin-like GTPase with broad antiviral activity. J Interferon Cytokine Res 2011;31(1):79–87.
[75] Meng X, Yang D, Yu R, Zhu H. EPSTI1 is involved in IL-28A-mediated inhibition of HCV infection. Mediators Inflamm 2015;2015:716315.
[76] Adeva-Andany MM, González-Lucán M, Donapetry-García C, Fernández-Fernández C, Ameneiros-Rodríguez E. Glycogen metabolism in humans. BBA Clin 2016;5 [cited 2020 Mar 24]Available from: https://cyberleninka.org/article/n/1341953.
[77] Smith CL, Dickinson P, Forster T, Craigon M, Ross A, Khondoker MR, et al. Identification of a human neonatal immune-metabolic network associated with bacterial infection. Nat Commun 2014;5:4649.
[78] Hong J-S, Greenlee KJ, Pitchumani R, Lee S-H, Song L, Shan M, et al. Dual protective mechanisms of matrix metalloproteinases 2 and 9 in immune defense against Streptococcus pneumoniae. J Immunol 2011;186(11):6427–36.
[79] Renckens R, Roelofs JJTH, Florquin S, de Vos AF, Lijnen HR, van't Veer C, et al. Matrix metalloproteinase-9 deficiency impairs host defense against abdominal sepsis. J Immunol 2006;176(6):3735–41.
[80] Wolf AJ, Reyes CN, Liang W, Becker C, Shimada K, Wheeler ML, et al. Hexokinase is an innate immune receptor for the detection of bacterial peptidoglycan. Cell 2016;166(3):624–36.
[81] Lee PP, Fitzpatrick DR, Beard C, Jessup HK, Lehar S, Makar KW, et al. A critical role for Dnmt1 and DNA methylation in T cell development, function, and survival. Immunity 2001;15(5):763–74.
[82] van den Broek MF, Hengartner H. The role of perforin in infections and tumour surveillance. Exp Physiol 2000;85(6):681–5.