



RAPID COMMUNICATION

Identification of gene classifier and modules associated with SARS-CoV-2 ARI in nasal swab



During a sudden viral pandemic, where the pathogenesis and nature of the infection are unclear, early and accurate identification of the patients infected with the pandemic-causing pathogen is crucial to prevent excess mortality. During the coronavirus disease 2019 (COVID-19) pandemic, early diagnosis of severe acute respiratory syndrome virus-2 (SARS-CoV-2) is limited by the incidence of the symptoms that are shared with acute respiratory illnesses (ARI) caused by other pathogens, including bacteria and viruses. Although pathogen-specific diagnosis methods like reverse transcription-quantitative polymerase chain reaction (RT-PCR) remain gold-standard for SARS-CoV-2 diagnosis, they also have several disadvantages, such as false negatives due to low or fluctuating viral loads and long processing time with the expertise needed for handling. The host immune response to different pathogen infections can be utilized for pathogen-specific diagnosis. This emerging approach has been shown to differentiate between bacterial and viral infections accurately. Identifying SARS-CoV-2 specific host transcriptome gene classifiers that can differentiate ARI caused by SARS-CoV-2 and other pathogens can aid in the early diagnosis of COVID-19 patients.

In this study, we developed a machine learning-based gene classifier and identified modules associated with SARS-CoV-2 infection. The expressions of the genes present in the classifier are validated in an independent dataset comprising patients infected with SARS-CoV-2 variants of concern (VOCs) (Fig. 1A). Three eligible RNA-seq datasets, GSE188678, GSE179277, and GSE156063, were selected from the gene expression omnibus (GEO) database (Table S1 and Fig. S1).¹ Three individual datasets were merged into a single dataset with 772 nasal swab ARI samples comprising 383 non-viral ARI (NVARI), 139 other respiratory viruses associated ARI (ORVARI), and 250 SARS-CoV-2 associated

ARI (SC2ARI) samples. This dataset was utilized to build a machine learning-based ARI gene classifier using the DaMiRseq pipeline.² The direct integration of three different datasets resulted in significant batch effects identified through the principal component analysis (PCA) plot (Fig. 1B). The batch-effect correction was performed to remove the batch effects. The adjusted counts were visualized through the PCA plot (Fig. 1C), and it was observed that batch effects were diminished and three groups were segregated, making the data suitable for further downstream analysis. Variance stabilizing transformation (vst) normalization and data filtering removed the genes with less expression and inconsistent samples. DaMiRseq pipeline identified 105 small, robust discriminatory features (genes) through backward elimination with partial least-square regression (Fig. S2 and Table S2; Supplementary methods). 16 genes were identified to be eligible for constructing the gene classifier based on the sorted list using the DaMir.Fselect function. The expression of these 16 genes across all the samples was visualized through a cluster plot (Fig. 1D). Out of the 16 genes, *IFI27*, *IFI6*, *IFI44*, *OAS2*, *OAS1*, *IDO1*, *SERPING1*, *USP18*, *C1QB*, *IL1B*, and *CCL5* were the 10 well-known genes to be interferon-inducible and play a role in interferon signaling pathway and the role of genes such as *SEC14L3* and *WFDC6* is unclear. The expression of these interferon-related genes increases substantially in SC2ARI compared to all other ARI conditions. The functional relevance of these genes was tabulated in (Table S3). The classifier model performance was evaluated using stringent 100 iterations and found ensemble approaches of SVM and RF, SVM alone, and RF alone have an accuracy of 83.7%, 84.6%, and 84.2%, respectively. The accuracy across 100 iterations was visualized through a violin plot (Fig. 1E).

Gene co-expression analysis was performed using the CEMiTool R package with default parameters to construct the co-expression network and modules co-expressed

Peer review under responsibility of Chongqing Medical University.

<https://doi.org/10.1016/j.gendis.2023.03.004>

2352-3042/© 2023 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

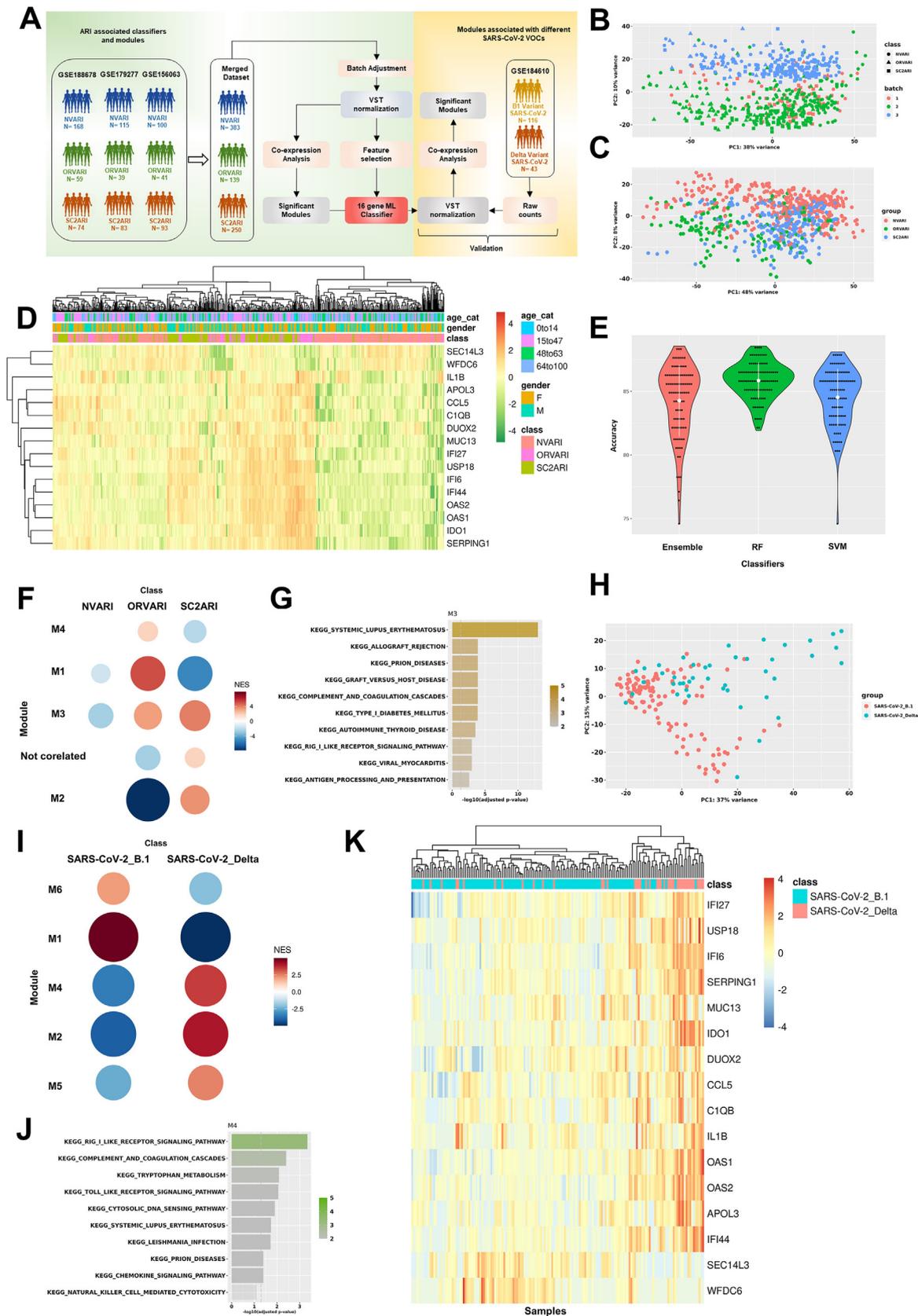


Figure 1 Construction of 16-gene ARI classifier and gene co-expression modules associated with ARI that also dysregulate between SARS-CoV-2 delta and previous SAR-CoV-2 VOCs. (A) Schematic representation of the overall approach implemented in the study. ARI-associated nasal swab classifier and gene modules were identified through machine learning and gene co-expression analysis, respectively. The significant gene modules associated with SARS-CoV-2 B.1 variants and highly contagious delta strain

across each ARI group.³ A total of 4 significant gene modules were identified with a soft threshold beta value of 6 (Fig. 1F; Fig. S3A, 4A, 4C and Table S4). Module 3 consists of 97 genes enriched for the interferon pathway, and its activity is very high in SC2ARI, indicating enhanced antiviral response in SC2ARI (Fig. 1G). Interestingly, most of the genes in the 16-gene ARI classifier belong to this module. Module 4 comprises 42 genes that do not significantly enrich any pathways, and the activity of this module is high in ORVARI and low in SC2ARI. Gene interaction data was used to construct the network identifying hub genes in each module. Module 3 is of high importance because of the higher activity in SC2ARI patients (Fig. S3B), and most of these genes were also found to discriminate between the three types of ARIs significantly.

Among the VOCs, the SARS-CoV-2 delta variant (B.1.167) undoubtedly caused a severe pandemic wave, causing an excessive death rate. Thus, we investigated the GSE184610 dataset to identify the significant modules and genes associated with the delta variant compared with other SARS-CoV-2 B.1 variants.⁴ The raw counts were vst-normalized and visualized through the PCA plot that shows the segregation of both groups (Fig. 1H). The vst-normalized data was used as input for CEMiTool with default parameters that identified 6 significant modules associated with each group (Fig. 1I; Fig. S5A and Table S5). Out of the 6 modules identified, module 4 is of high interest because it had 84 genes that are enriched for interferon-associated genes whose activity was high in the SARS-CoV-2 delta-infected patients. ORA analysis identified the RIG-I signaling pathway to be significantly enriched (Fig. 1J). Gene interaction network analysis identified interferon pathway-related genes such as *MX1*, *IFIT1*, and *IFIT2* as hub genes (Fig. S5B). Another module, M1 enriched for the metabolism-associated pathway, was also identified to be significant in ARI classification (Fig. S6). Module 2 enriched for cytokine and chemokine signaling, while Module 5 enriched for leukocyte trans-

endothelial migration and B-cell receptor pathway (Fig. S7).

Since the module enrichment for the interferon-associated pathway was significantly identified in both ARI and variants datasets, we cross-validated the expression of the 16 genes used to build the ARI classifier in the SARS-CoV-2 delta and B.1 variants of infected patients. Interferon-associated genes were up-regulated in SARS-CoV-2 delta variant infected patients, whereas the expression of *SEC14L3* and *WFDC6* were down-regulated in SARS-CoV-2 delta variant patients in comparison with patients infected with SARS-CoV-2 B.1 variants (Fig. 1K). Module 3 of the ARI analysis and Module 4 of the VOC analysis share many common genes. Similarly, other genes and the associated modules were tabulated (Table S6). The observation of higher induction of interferon-associated genes in SC2ARI compared to other ARIs, and SARS-CoV-2 delta variants compared to the previous B1 variants, reveals the modulation of innate immune pathways employed by evolving variants of SARS-CoV-2.

Our observation is in accordance with the study that identified enhanced interferon-dependent gene expression and hyper-inflammation in the mouse lungs intranasally challenged with the SARS-CoV-2 delta variant compared to other SARS-CoV-2 VOC alpha variants.⁵ Although SARS-CoV-2 VOCs were known to be resistant to interferon induction in multiple studies, our observation of enhanced interferon-dependent gene expression in SARS-CoV-2 delta infection leads to an open question regarding the induction of interferon responses by delta variant and how they use or overcome this effect to establish infection.

Host response-based diagnosis can be a complementary tool to existing methods to diagnose the SARS-CoV-2 infection early. Although host response-based diagnosis is promising, clinical validation of the classifiers in a large cohort of samples with metadata information only warrants the real-world accuracy to be compared with gold standard methods such as RT-PCR.

were identified through the gene co-expression analysis. A comparison of ARI classifier gene expression in the SARS-CoV-2 variant dataset was performed. (B) The raw counts of nasal swab transcriptome datasets, GSE188678, GSE179277, and GSE156063 were merged into a single dataset of 772 samples and the batch effect due to the merging of three independent datasets was visualized through PCA plot. (C) After performing batch effect correction, the PCA plots were used to visualize the batch-adjusted counts. (D) Batch-adjusted counts were subjected to feature selection through the DaMiRseq pipeline that identified 105 genes and 16 genes were selected to construct the gene classifier and visualized through cluster plot. (E) The accuracy of the constructed classifier across 100 iterations was visualized through the violin plot. (F) The batch-adjusted, vst-normalized counts were used as input for the gene co-expression analysis through the CEMiTool package with default parameters. Four significantly co-expressed gene modules and the gene set enrichment analysis of each identified module were visualized through the dot plot. (G) Module 3 enriched for interferon associated pathway and over-representation analysis (ORA) of Module 3 was performed and visualized through a bar plot. (H) GSE184610 dataset derived from the nasal swab of SARS-CoV-2 delta variants and other SARS-CoV-2 B.1 variants was re-analyzed and vst-normalized counts were visualized through PCA plot that showed the segregation of both the groups. (I) Gene co-expression analysis was performed through the CEMiTool package with default parameters. Six significantly co-expressed gene modules and the gene set enrichment analysis of each identified module were visualized through the dot plot. (J) Over-representation analysis of Module M4 identified interferon-related pathways as observed with the Module 3 ARI analysis. (K) The expression level of the 16 genes identified as the ARI classifier was plotted as a heatmap in the SARS-CoV-2 delta variant and other SARS-CoV-2 B1 variants infected patients.

Author contributions

Pandikannan Krishnamoorthy: conceptualization, data curation, formal analysis, validation, writing - original draft, and writing - review & editing; Athira S Raj: data curation, writing - original draft, and writing - review & editing; Himanshu Kumar: conceptualization, writing - original draft, writing - review & editing, project administration, and supervision.

Conflict of interests

All authors declare no conflict of interests. All authors agreed to manuscript publication.

Funding

This work was supported by the Department of Biotechnology India (DBT)-(BT/PR44561/MED/15/217/2021), the Science and Engineering Research Board, India (SERB)-(CRG/2020/001417), and IISER Bhopal.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.gendis.2023.03.004>.

References

1. Mick E, Kamm J, Pisco AO, et al. Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses. *Nat Commun.* 2020; 11(1):5854.
2. Chiesa M, Colombo GI, Piacentini L. DaMiRseq-an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification. *Bioinformatics.* 2018;34(8):1416–1418.
3. Russo PST, Ferreira GR, Cardozo LE, et al. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinf.* 2018;19(1):56.
4. Grimaldi A, Panariello F, Annunziata P, et al. Improved SARS-CoV-2 sequencing surveillance allows the identification of new variants and signatures in infected patients. *Genome Med.* 2022; 14(1):90.
5. Lee KS, Wong TY, Russ BP, et al. SARS-CoV-2 Delta variant induces enhanced pathology and inflammatory responses in K18-hACE2 mice. *PLoS One.* 2022;17(8):e0273430.

Pandikannan Krishnamoorthy ^a, Athira S. Raj ^a,
Himanshu Kumar ^{a,b,*}

^aDepartment of Biological Sciences, Laboratory of Immunology and Infectious Disease Biology, Indian Institute of Science Education and Research (IISER) Bhopal, Bhopal, MP 462066, India

^bLaboratory of Host Defense, WPI Immunology, Frontier Research Centre, Osaka University, Osaka 5650871, Japan

*Corresponding author. Department of Biological Sciences, Laboratory of Immunology and Infectious Disease Biology, Indian Institute of Science Education and Research (IISER) Bhopal, AB-3, Room No. 220, Bhopal By-pass Road, Bhauri, Bhopal, MP 462066, India.

E-mail address: hkumar@iiserb.ac.in (H. Kumar)

14 November 2022
Available online 5 April 2023