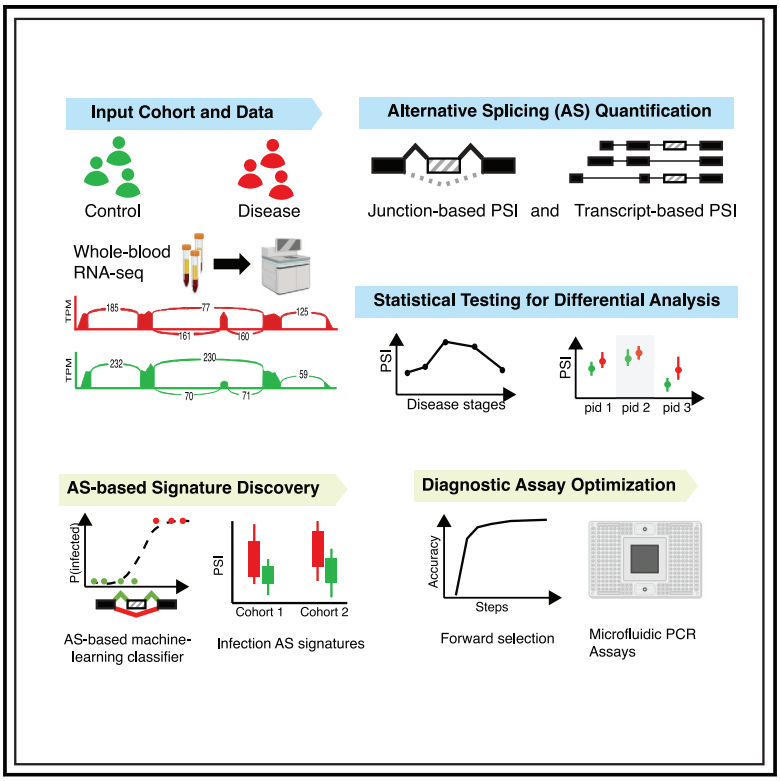
Cell Reports
Methods

Report

Blood RNA alternative splicing events as diagnostic biomarkers for infectious disease

Graphical abstract



Highlights

- We present a computational framework for alternative splicing (AS) diagnostic markers
- Our AS biomarkers outperform gene-expression biomarkers in COVID-19 detection
- Microfluidic PCR diagnostic assay of AS biomarkers achieves greater than 98% accuracy
- We interpret the biological importance of identified AS biomarkers

Authors

Zijun Zhang, Natalie Sauerwald, Antonio Cappuccio, ..., Andrew G. Letizia, Stuart C. Sealfon, Olga G. Troyanskaya

Correspondence

stuart.sealfon@mssm.edu (S.C.S.), ogt@genomics.princeton.edu (O.G.T.)

In brief

Host-based response assays can diagnose infectious disease earlier and more precisely than pathogen-based tests. However, the role of RNA alternative splicing (AS) remains unexplored. Zhang et al. present a computational framework for AS diagnostic biomarkers. Using SARS-CoV-2 as a case study, they demonstrate the improved accuracy of AS biomarkers for COVID-19 diagnosis.

Zhang et al., 2023, Cell Reports Methods *3*, 100395
February 27, 2023
https://doi.org/10.1016/j.crmeth.2023.100395

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Report

Blood RNA alternative splicing events as diagnostic biomarkers for infectious disease

Zijun Zhang,[1,2] Natalie Sauerwald,[1] Antonio Cappuccio,[3] Irene Ramos,[3] Venugopalan D. Nair,[3] German Nudelman,[3] Elena Zaslavsky,[3] Yongchao Ge,[3] Angelo Gaitas,[3] Hui Ren,[4] Joel Brockman,[4] Jennifer Geis,[4] Naveen Ramalingam,[4] David King,[4] Micah T. McClain,[5] Christopher W. Woods,[5] Ricardo Henao,[5] Thomas W. Burke,[5] Ephraim L. Tsalik,[5] Carl W. Goforth,[6] Rhonda A. Lizewski,[7] Stephen E. Lizewski,[7] Dawn L. Weir,[6] Andrew G. Letizia,[6] Stuart C. Sealfon,[3,*] and Olga G. Troyanskaya[1,8,9,10,*]

[1]Center for Computational Biology, Flatiron Institute, New York, NY 10010, USA
[2]Division of Artificial Intelligence in Medicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
[3]Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[4]Fluidigm Corporation, South San Francisco, CA 94080, USA
[5]Center for Applied Genomics and Precision Medicine, Duke University School of Medicine, Durham, NC 27710, USA
[6]Naval Medical Research Center, Silver Spring, MD, USA
[7]Naval Medical Research Unit SIX, Lima, Peru
[8]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[9]Department of Computer Science, Princeton University, Princeton, NJ 08544, USA
[10]Lead contact
*Correspondence: stuart.sealfon@mssm.edu (S.C.S.), ogt@genomics.princeton.edu (O.G.T.)
https://doi.org/10.1016/j.crmeth.2023.100395

MOTIVATION Host-based response assays (HRAs) can often diagnose infectious disease earlier and more precisely than pathogen-based tests. However, the role of RNA alternative splicing (AS) in HRAs remains unexplored, as existing HRAs are restricted to gene expression signatures. We report a computational framework for the identification, optimization, and evaluation of blood AS-based diagnostic assay development for infectious disease. Using SARS-CoV-2 infection as a case study, we demonstrate the improved accuracy of AS biomarkers for COVID-19 diagnosis when compared against six reported transcriptome signatures and when implemented as a microfluidic PCR diagnostic assay.

SUMMARY

Assays detecting blood transcriptome changes are studied for infectious disease diagnosis. Blood-based RNA alternative splicing (AS) events, which have not been well characterized in pathogen infection, have potential normalization and assay platform stability advantages over gene expression for diagnosis. Here, we present a computational framework for developing AS diagnostic biomarkers. Leveraging a large prospective cohort of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection and whole-blood RNA sequencing (RNA-seq) data, we identify a major functional AS program switch upon viral infection. Using an independent cohort, we demonstrate the improved accuracy of AS biomarkers for SARS-CoV-2 diagnosis compared with six reported transcriptome signatures. We then optimize a subset of AS-based biomarkers to develop microfluidic PCR diagnostic assays. This assay achieves nearly perfect test accuracy (61/62 = 98.4%) using a naive principal component classifier, significantly more accurate than a gene expression PCR assay in the same cohort. Therefore, our RNA splicing computational framework enables a promising avenue for host-response diagnosis of infection.

INTRODUCTION

Host-based response assays (HRAs) are being developed for diagnosis of infectious disease.[1,2] Compared with conventional pathogen-based nucleic acid amplification tests (NAATs), HRAs target transcriptional alterations in the host blood instead of pathogenic materials. During the early window of infection, NAATs are known to have relatively high false negative rate,[3] while HRAs are shown to be sensitive as early as 12 h after viral challenge.[4] This characteristic of HRAs has been leveraged to develop early detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections.[5] Similarly, HRAs have been developed where traditional NAATs have failed, such as discriminating bacterial versus viral infections[1,6] and early detection of respiratory viral infection.[7]

However, the role of RNA alternative splicing (AS) in HRAs remains unexplored, as existing HRAs are restricted to gene expression signatures. During AS, an exon (or part of an exon) in the pre-mRNA will be selectively included or excluded from the final mRNA product, and an AS event is quantified by the ratio between the exon-inclusion isoforms and exon-skipping isoforms.[8] As such, AS events have potential normalization and assay platform stability advantages compared with gene expression for diagnostic HRAs. Platform stability is essential for effective biomarker assays in clinics because biomarkers are typically discovered from genome-scale experimental platforms (e.g., Illumina RNA sequencing [RNA-seq]) but are implemented as targeted profiling assays (e.g., microfluidic devices) for portable and scalable clinical use. By measuring the AS event as a ratio, the inclusion level of an exon normalizes to a proportion between 0 and 1; platform-specific bias and technical noises are expected to cancel out in the ratio, as they equivalently affect exon-inclusion and -skipping isoforms.

Using AS events as HRA markers is well grounded biologically. Almost all multi-exon human genes undergo AS, significantly diversifying the human transcriptome and proteome. Within infected cells, viral infection often disrupts host AS; evidence of hijacked host splicing machinery has been identified for Zika virus, human cytomegalovirus (HCMV), and SARS-CoV-2 (Thompson et al.,[9] Banerjee et al.,[10] and Zhou et al.[11]). On the other hand, within human immune cells, a widespread splicing program switch has been observed upon external stimuli, including differential splicing of cell surface markers as well as transcriptional and RNA regulators,[12,13] underlining the functional importance of splicing modulation during immune activation and response.

Here, we report a computational framework for the identification, optimization, and evaluation of blood AS-based diagnostic assay development for infectious disease. Using SARS-CoV-2 infection as a case study, we demonstrated a robust set of AS signatures superior to the gene expression-based markers derived from the same cohort, as well as six previously reported transcriptome signatures. Functional analysis revealed significant enrichment of differential splicing events in immune-specific protein domains and genes. Therefore, we propose a highly performant set of host-based AS-centered biomarkers for SARS-CoV-2 infection detection and demonstrate a promising avenue for design and implementation of host-based diagnostics by leveraging RNA splicing as robust diagnostic biomarkers.

RESULTS

Robust workflow for AS diagnostic biomarker development

To identify robust RNA AS events as diagnostic biomarkers, we developed a computational framework for optimizing AS signatures in flexible experimental designs, including longitudinal and cross-sectional studies. Starting with a cohort of healthy subjects and infected subjects, whole-blood specimens are collected and sequenced by RNA-seq (Figure 1A). We then quantify the AS events from RNA-seq datasets by computing the percent spliced in (PSI) for each AS event using two complementary statistical measurements.[14–16] Because the two measurement approaches rely on related but different empirical information of junction counts and transcript expression levels, respectively (STAR Methods), joint statistical modeling of PSIs measured by different approaches enables us to identify robust, measurement-agnostic AS variations. Our framework leverages a flexible regression framework to model PSI variations with respect to disease status while controlling for confounding covariates (Figure 1B). Therefore, our framework can effectively analyze cross-sectional experimental design as well as more complex experimental designs such as longitudinal studies.

We applied this framework to investigate AS variations during SARS-CoV-2 infection in a large prospective cohort named COVID-19 Health Response for Marines (CHARM).[17] To account for the covariance of multiple RNA-seq data longitudinally sampled from the same subject and to control for confounding variables in the data, we developed linear mixed models in our framework to analyze the dependency between disease progression stages (Figure 1B). Importantly, our framework also enabled us to explicitly control for non-disease-related splicing changes (e.g., due to the subject's sex or military training). After estimating the regression coefficients and testing for statistical significance, we identified over 1,000 significant disease-associated AS events.

Leveraging these AS changes, we built a host-based AS-centered diagnostic assay for SARS-CoV-2 infection. Using the differential AS (DAS) events, we trained a logistic regression classifier that was accurate across cohorts and detected infections undetected by serial PCR tests (Figure 1C). Our AS-based classifier outperformed gene expression-based markers derived from the same cohort as well as six previously reported transcriptome signatures, demonstrating the superior consistency and robustness of AS biomarkers for viral infection. To further optimize a subset of the markers in an AS-based microfluidic biomarker assay, we performed a forward selection to select non-redundant AS signatures (Figure 1D). Our microfluidic assay achieved nearly perfect accuracy in an independent cohort (accuracy = 98.4%). Finally, we characterized SARS-CoV-2-infection-induced AS variations with functional, network, and biomarker analyses, identifying putative upstream splicing factors and temporal dynamics in splicing regulation (Figure 1E). Below, we describe in detail the identification and comprehensive evaluations of AS-based diagnostic biomarkers, followed by a functional analysis of DAS events for COVID-19.

Host-based AS events accurately detect SARS-CoV-2 infection across cohorts

In the CHARM cohort, we performed 1,176 whole-blood RNA-seq from n = 371 US Marine recruits. After a 14-day quarantine, we longitudinally collected whole blood for RNA-seq, immunoglobulin G (IgG) and IgM antibody tests, and PCR testing for SARS-CoV-2 virus. Based on the timeline of SARS-CoV-2 PCR test results, samples were divided into control (more than 1 week before first PCR+ test) and disease stages of pre (within 2 weeks pre-PCR+), first (first-time PCR+), mid (follow-up PCR+), and post (post-infection; see details in Figure 2A).

Using the CHARM cohort of longitudinal SARS-CoV-2 infection, our framework analyzed four distinct types of AS events and identified a substantial number of DAS events in each of the four types of AS events across the disease stages (Figure 2B).

CellPress
OPEN ACCESS

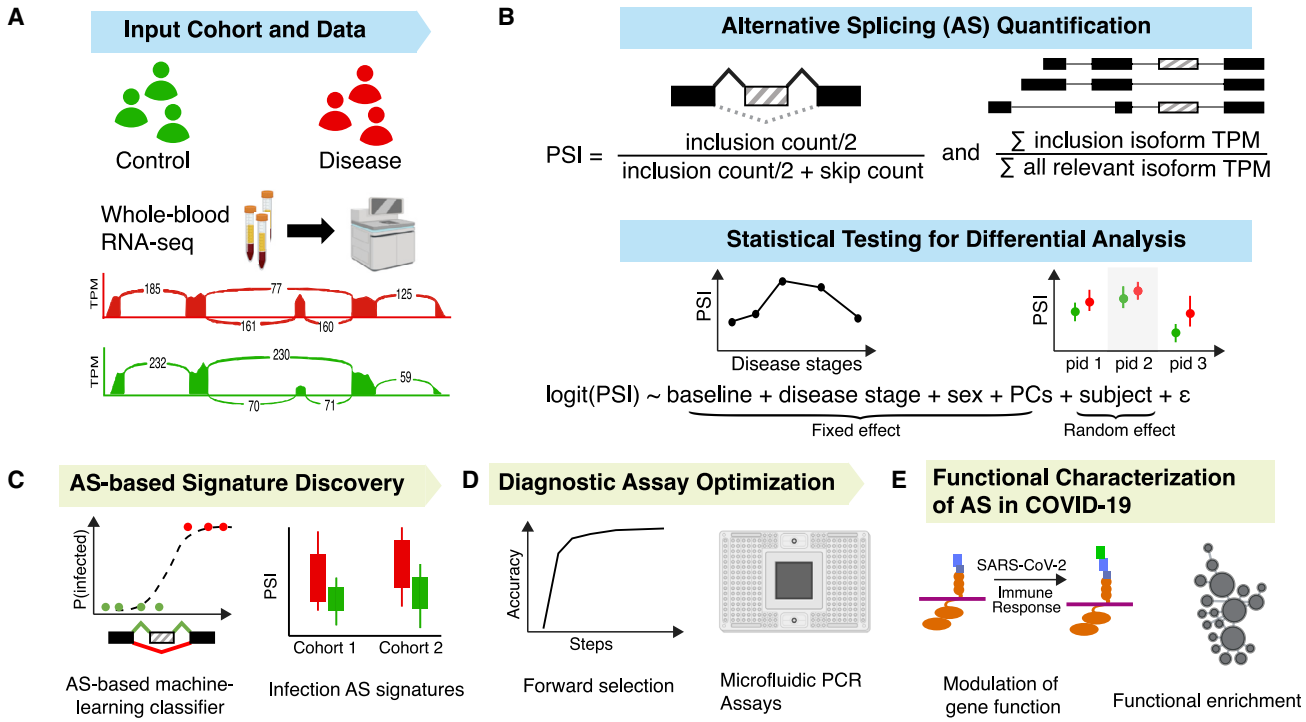


Figure 1. Overview of analysis workflow to develop robust alternative splicing diagnostic biomarkers

(A) Input data for our framework included cohort metadata and whole-blood RNA-seq data to develop alternative splicing (AS)-based diagnostic markers.

(B) These data were then processed in two steps: (1) AS levels were quantified by two statistical approaches, and (2) differential analysis was performed by a regression framework, allowing for cross-sectional and longitudinal experimental designs.

(C) Classification of disease versus controls using AS-based biomarkers. PSI values for AS events were used as features to train a machine-learning classifier, which we evaluated in independent cohorts.

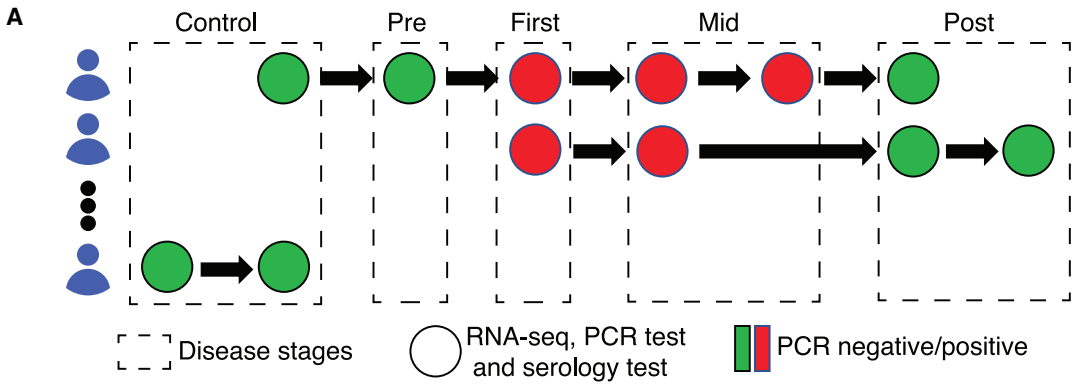
(D) We performed a forward selection to select non-redundant AS signatures. The optimized, small-footprint signature set was then implemented in microfluidic devices for diagnostic testing.

(E) Functional analysis of AS variations on the exon and gene levels. Blue band: data processing; green band: signature development and analysis.

Most splicing program changes occurred during the PCR+ phase of the SARS-CoV-2 infection (first and mid). Qualitatively, we show that DAS events identified at the first versus control comparison can separate infected samples and healthy controls by unsupervised clustering (Figure S1), suggesting an acute, homogeneous immune response mediated by AS regulation to SARS-CoV-2 infection when subjects first turn PCR+.

Therefore, we built logistic regression models using the first PCR+ samples from the CHARM cohort to classify active SARS-CoV-2 infection. The PSI values of DAS events were used as features for a given false discovery rate (FDR) cutoff, and we varied the FDR cutoffs to generate signature sets of different sizes (STAR Methods). We tested our classifiers on an independent cohort processed at the Duke Medical Center (herein referred to as the Duke cohort). The Duke cohort provided a challenging test set as it was focusing on older individuals (average age = 46, SD = 18.6; Table S1) as opposed to the CHARM cohort of healthy, young, asymptomatic/mildly symptomatic Marine recruits. As a control experiment, we applied the same model training procedure to the gene expression values of top differentially expressed genes (DEGs) ranked by their differential test FDRs (STAR Methods). The testing performance peaked at around 300 features for both DAS- and DEG-based classifiers, while including more features afterward induced overfitting for both types of classifiers. Furthermore, when conditioned on the same number of features (n = 100–600 features), DAS classifiers always outperformed DEG classifiers ($p = 0.047$, ANOVA; Figure 3A). Since the set of diagnostic markers needs to be minimized in size while maintaining optimal performance, this suggested the potential of using DAS biomarkers for better diagnostic assay design.

To consolidate our observations, we additionally examined six publicly available gene expression signatures for SARS-CoV-2 infection.[18–23] We denote the best-performing signature sets we derived from the CHARM cohort as CHARM DAS and DEG, respectively. All public signature sets were processed identically to CHARM DEG. Briefly, in order to rigorously compare the quality of signature sets without biases induced by cohort and model differences in previous studies, we re-trained classifiers on the CHARM cohort using different previously reported gene signature sets as features and tested the classifiers on the Duke cohort (STAR Methods). We found that the CHARM DAS signature set performed the best among all signature sets, outperforming all DEG-based signatures by a large margin (Figure 3B). The signature set from McClain et al. was excluded from this comparison because the same Duke cohort was used as the



B

Types of Alternative Splicing Events	Pre (n=48)	First (n=72)	Mid (n=254)	Post (n=334)
Skipped exon (SE)	44	671	1,027	168
Alternative 5' splice site (A5SS)	11	195	261	70
Alternative 3' splice site (A3SS)	17	194	309	111
Retained intron (RI)	21	242	267	225

Figure 2. Longitudinal analysis of differential AS events in SARS-CoV-2 infection

(A) Experimental design for the COVID-19 Health Response for Marines (CHARM) cohort. Whole-blood specimens from subjects were collected longitudinally and sequenced by RNA-seq. Controls were PCR-negative (PCR−) samples with no positive antibody tests. The first sample with a PCR+ test for a given subject was labeled as first. PCR− RNA-seq samples within 2 weeks before first for these subjects were labeled as pre-infection. PCR+ samples after the first sample were labeled as mid, while RNA-seq samples from infected subjects that turned PCR− were labeled as post-infection (post).

(B) Definitions of the four AS events analyzed and the number of statistically significant (FDR < 0.05) events identified for each disease stage.

discovery set in McClain et al. While this signature set performed best among DEG-based signatures (test area under the receiver operating curve [AUROC] = 0.806), it is still subpar to the CHARM DAS signatures. The DAS classifier achieved a testing AUROC of 0.85 on the Duke cohort, demonstrating the superiority of AS signatures compared with previously published gene expression signatures (Figure 3C).

Using PSI of CHARM DAS signature set as features, we assessed the infection likelihood for all CHARM samples by a 10-fold cross-validation prediction (AUROC = 0.922 ± 0.057) for each individual throughout the time course (Figure S2A). Strikingly, for one particular subject with consistently PCR− test results and no symptoms reported (shown in Figure 3D), we observed a strong peak in predicted probability of infection at day 42, with predicted probability gradually decreasing over time. This subject had negative SARS-CoV-2 antibody tests in early time points (T0 and T42), demonstrating no vaccination (as the cohort was enrolled during May–November 2020) nor pre-exposure to SARS-CoV-2. At day 56, this subject had a positive IgG test (7.01) and IgM test (1.85) specific to SARS-CoV-2 viral proteins, suggesting that the PCR tests were false negatives. Together, these results strongly suggest that this subject was an asymptomatic infected subject that consistently tested negative through a series of four PCR tests, and our AS-based classifier was accurate enough to capture this case missed by PCR testing and symptom monitoring. Furthermore, our classifier likely generalized the AS patterns in response to SARS-CoV-2 infection beyond just learning symptoms. We predicted a high infection score for the asymptomatic SARS-CoV-2-infected subject (Figure 3D); conversely, in SARS-CoV-2 PCR-/sero-negative control samples, having symptoms did not significantly increase the predicted infection scores (Figure S2B).

AS-based host-response assay demonstrates across-platform consistency

Encouraged by the superior performance of our CHARM DAS classifier, we further sought to optimize a small set of splicing biomarkers to test on microfluidic devices. Compared with

CellPress
OPEN ACCESS

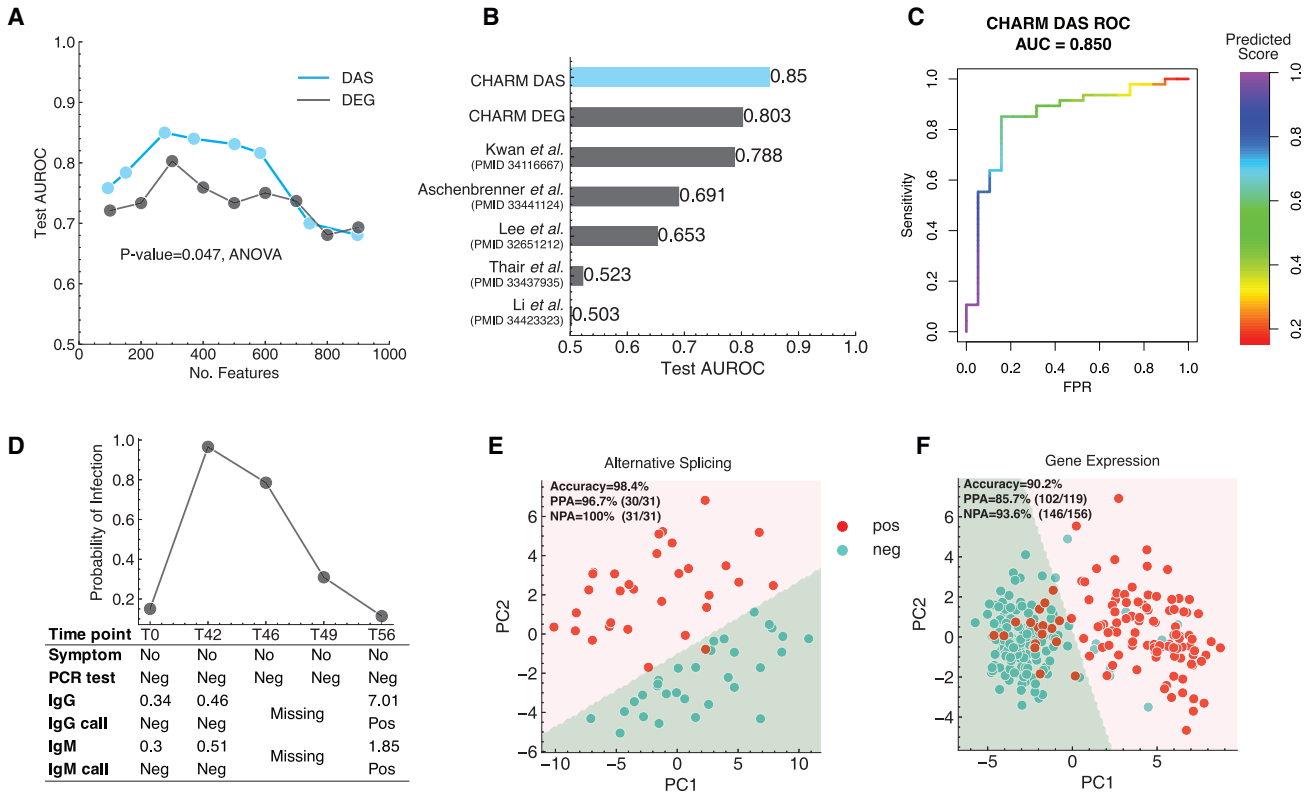


Figure 3. Biomarkers of AS events cross-cohort predictive modeling

(A) Comparison of AS and gene expression classification AUROC on the Duke cohort using varying numbers of features.

(B) Comparison of the best-performing DAS and DEG signatures (CHARM DAS and DEG) against five public transcriptome signatures on the Duke cohort.

(C) AUROC plot for the cross-cohort SARS-CoV-2 infection classifier. Color bar represents the predicted scores.

(D) Cross-validation predictions for control samples identified an infected subject that was mislabeled as healthy control despite serial PCR tests.

(E) PCA of AS biomarkers measured by microfluidic devices in a third independent clinical cohort.

(F) PCA of gene expression biomarkers measured by microfluidic devices generated by Cappuccio et al. Decision boundaries are fitted by a support vector machine with a linear kernel. PPA, positive percent agreement; NPA, negative percent agreement.

The Duke cohort was used for evaluation in (A)–(C); the cross-validation result of the CHARM cohort was used in (D); and the Fluidigm cohort was used for evaluation in (E) and (F).

Illumina RNA-seq, the microfluidic devices are portable as a diagnostic device but are limited in the capacity of measurable probes and present distinct technical biases. We performed a forward selection of biomarkers to select non-redundant biomarkers by optimizing the classifier's performance in both CHARM and Duke cohorts, resulting in a set of 27 AS biomarkers (Figure S2C). The full list of selected biomarkers is in Table S2. When tested on a third independent clinical cohort with balanced COVID-19 versus healthy control (n = 62), who were not included in any previous analyses, the microfluidic device with 27 AS biomarkers achieved nearly perfect linear separation (accuracy = 98.4%, 95% confidence interval [CI]: 90.2%–99.9%) using a simple principal-component analysis, with a 96.7% (95% CI: 81.5%–99.8%) positive percent agreement (PPA) and 100% (95% CI: 86.3%–100%) negative percent agreement (NPA) (Figure 3E).

By contrast, we re-analyzed the microfluidic PCR data based on gene expression markers from a recent report that investigated the same CHARM cohort.[5] Based on a signature set discovered by whole-blood RNA-seq DEG biomarkers (n = 41), this study profiled a subset of n = 275 subjects from the CHARM cohort on the microfluidic PCR platform. Applying the same principal components linear classifier to the DEG principal-component analysis (PCA), we computed its accuracy at 90.2%, with PPA = 85.7% and NPA = 93.6% (Figure 3F). To account for the sample size differences between the two microfluidic assays, we downsampled from the gene expression assay to match the sample size of the AS assay and computed the accuracy over 10,000 downsamples. When sample sizes were matched, the AS biomarker assay was still significantly more accurate than the gene expression assay (p = 0.011; Figure S2D). We further demonstrated that the within-group, between-individual variability of our AS biomarker assay was statistically smaller than both non-PSI-transformed raw probe values and gene expression probes by Cappuccio et al. (Figure S2E). Since this effect is specific to PSI-transformed probe values, it suggests the AS biomarker assay had a smaller technical variance, likely due to its internal normalization property (see methods S1).

Therefore, we propose a highly performant set of host-based AS-centered biomarkers for SARS-CoV-2 infection detection

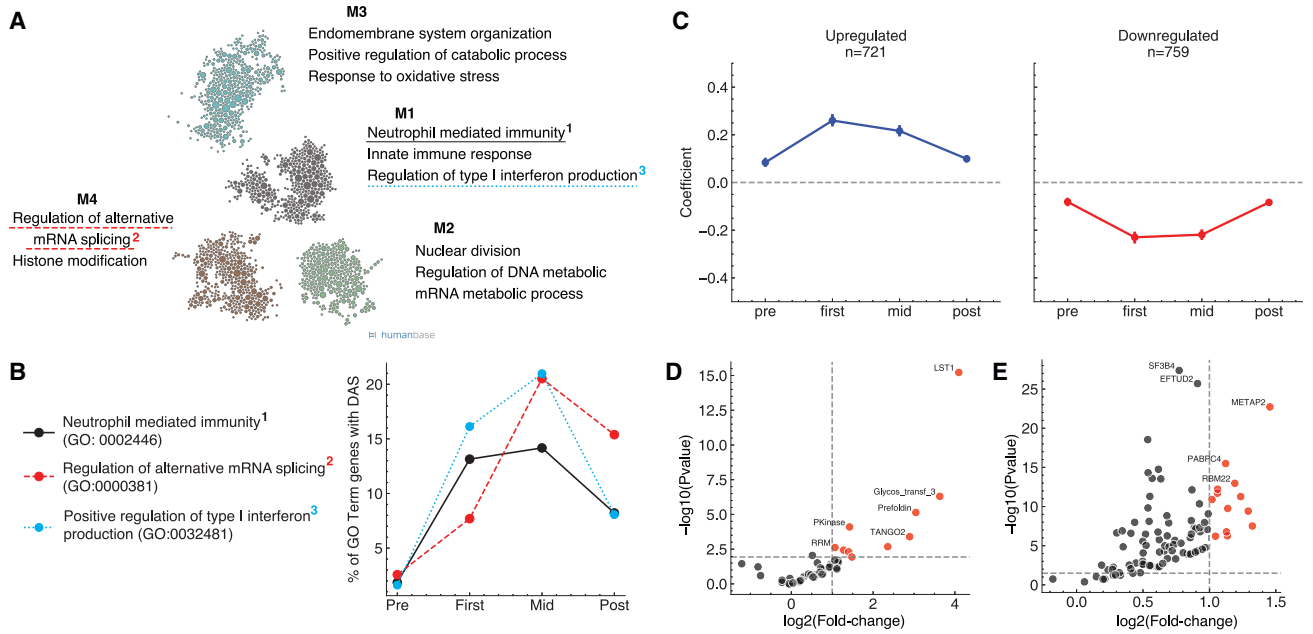


Figure 4. Functional analysis for DAS genes

(A) Whole-blood tissue-specific gene network enrichment of genes undergoing DAS. Representative terms for each module are shown; see Table S5 for the full list.

(B) The percentage of genes undergoing DAS in three representative GO terms along the temporal course of disease progression.

(C) The estimated coefficients represent an unbiased disease-induced AS difference between disease progression stages and healthy controls. Shown here are the average of coefficients for upregulated and downregulated SE events. Horizontal bars are 95% bootstrapping confidence intervals. See Figure S3B for other AS types.

(D and E) Volcano plots for the enrichment of (D) protein domains annotated by Pfam and (E) RBP binding sites identified by ENCODE eCLIP. p values were calculated from Fisher's exact test. Vertical dashed line represents cut off at fold change >2, and horizontal dashed line represents FDR <0.05.

and demonstrate a promising avenue for design and implementation of host-based diagnostic by leveraging RNA splicing as robust diagnostic biomarkers.

Interpreting biological importance of AS biomarkers

We sought to interrogate the biological basis of the superior performance of AS-based diagnostic markers. To understand the functional characteristics of genes undergoing DAS in response to SARS-CoV-2 infection, we employed blood-specific functional networks (https://hb.flatironinstitute.org/).[24] Tissue-specific functional networks capture tissue-specific protein function, interactions, and pathway activities.[25] Louvain community clustering on the whole blood tissue-specific network of the genes undergoing DAS in our data reveals four modules with statistically significant top Gene Ontology (GO) terms (FDR < 0.01; Figure 4A). These include a module focused on innate immune activation and chemokine signaling (M1) and a module on cell division to potentially facilitate proliferation of T cells as part of the adaptive immunity (M2), as well as a module representing phagocytosis and oxidative stress (M3). Finally, we found that a module (M4) of genes involved in splicing regulation appear to themselves be significantly alternatively spliced.

We also examined the temporal dynamics of biological processes affected by DAS (Figure 4B). Specifically, we summarized the significantly enriched Biological Process GO terms into clusters (STAR Methods; Figure S3A) and analyzed the largest three GO clusters for temporal changes in numbers of DAS genes at each stage through pre-infection, infection (first and mid), and post-infection. Genes related to neutrophil-mediated immunity were differentially spliced at a comparable ratio between first and mid; by contrast, regulation of type I interferon production was substantially more differentially spliced at mid compared with first. Interestingly, the regulation of AS also came up as strongly enriched in DAS genes, forming an auto-regulatory loop of splicing factors and target AS exons (Table S3).

Next, we systematically analyzed the temporal dynamics of the DAS events on the exon level (Figure 4C). By meta-analyzing the estimated coefficients from the linear mixed models, we observed the temporal changes of AS differences where the molecular level response was the most dramatic at first, followed by mid. We also identified a set of DAS events with consistent cross-cohort patterns (Table S4; two representative signatures in Figures S2F and S2G). *GALNS* is a lysosomal exohydrolase involved in innate immune response.[26] Increased inclusion of exon 2, which belongs to the sulfatase and phosphodiest protein domain families, suggests enhanced enzymatic activity to facilitate the activation of innate immune pathways. The second event is a retained intron (RI) event in *LST1*, which is thought to be involved in modulating immune responses.[27] Inclusion of the target intron will disrupt the LST1 protein domain and introduce multiple premature stop codons; our analysis shows that this

disruptive AS event is significantly reduced during viral infection. This demonstrates that AS can modulate gene functions independently of the cellular transcriptional control.

For functional characterizations of the DAS events, we focused on the protein domains (Pfam database)[28] and experimentally annotated RNA-binding protein (RBP) binding sites (CLIPdb and ENCODE databases)[29,30] retained in the DAS exons (STAR Methods). For protein domain analysis, LST1 is the most enriched protein family (Figure 4D). Other enriched domains include those that represent post-translational regulation and signaling cascades (glycosyl transfer family a/b and PKinase) and cell migration and cytokine secretion (prefoldin, TANGO2), as well as post-transcriptional regulation (RRM) in response to the viral infection. We also identified a number of RBPs whose binding sites are significantly enriched in the DAS exons compared with all exons, including RBM22, METAP2, and PABPC4 (Figure 4E). PABPC4 is known to be upregulated in activated T cells and might be necessary for the regulation of stability of labile mRNA species in activated T cells.[31,32]

In summary, we systematically characterized the biological basis for the DAS biomarkers during SARS-CoV-2 infection, identifying significant splicing changes involved in immune response both through RBP regulation and protein domain functional modulation.

DISCUSSION

SARS-CoV-2 is a highly contagious virus that has caused a worldwide pandemic. We analyzed the dynamics of host-based response to SARS-CoV-2, focusing on AS. Our analysis leverages a large cohort of US Marine recruits. We found many immune-related genes undergoing DAS. Since AS can alter genes and their protein products, these findings complement the existing gene and protein expression-based findings and open new avenues for understanding the molecular mechanisms of human immune responses against SARS-CoV-2.

The CHARM cohort is very homogeneous in age and health status, providing molecular profiling from many individuals with few confounding factors typical of SARS-CoV-2 human studies. Despite the uniqueness of this cohort, we showed that the AS signatures we found in young adults with mild and asymptomatic infections generalized exceptionally well to two other cohorts, including older individuals. This demonstrates that the AS signatures are robust, viral-infection-induced molecular changes.

The large and homogeneous discovery CHARM cohort in our study may have contributed to the superiority of our SARS-CoV-2 signatures. Compared with previously published signature sets, both CHARM DEG and DAS signatures demonstrated better performance (Figure 3B). Statistically, the uncertainty of identifying the optimal signature set can stem from either limited sample size (data uncertainty) or inexact model estimation (model uncertainty), which both could lead to an underperforming, less-generalizable signature set. While our cohort increases the statistical power in signature discovery, we also note that this cohort cannot be used to systematically evaluate the cross-reactivity of our AS signatures to other viral and bacterial infections. In general, comprehensive evaluation of signature specificity in other diseases remains a key and open challenge for HRAs. Furthermore, previous molecular profilings for infectious disease patient cohorts were largely based on microarrays, which poses an additional challenge for AS analysis. This highlights the importance of cohort recruitment and experiment design in future studies.

In clinical settings, gene and protein expression is usually the primary analysis of interest, partially because of the straightforward functional interpretations. While splicing analysis appears to be underrepresented in such settings, we demonstrate that the analysis of AS is also a powerful tool and can shed new light on the molecular mechanisms that otherwise could be missed by analyses based solely on gene and protein expression.

A host-based diagnostic tool can potentially reduce the spread of infectious diseases by detecting pathogen infections prior to symptom manifest, complementing the conventional pathogen-based diagnosis largely restricted by symptom onset. Leveraging the large longitudinal cohort in this study, we developed a host-based, AS-centered predictive model that can accurately identify SARS-CoV-2-infected samples from serial PCR false negative tests. Our computational framework facilitates further exploration for host-based AS dynamics in other infectious diseases, enabling new avenues for biomarker design and implementation.

Limitations of the study

First, the CHARM cohort is homogeneous in age and health status with few confounding factors, unlike most typical human studies of SARS-CoV-2. To provide as fair a comparison as possible of the gene set signatures rather than the generalizability of the models, we applied the same modeling methodology and used the same training data to evaluate all signatures. Due to the cohort differences, the DEG performances trained in our CHARM cohort may be subpar to the model trained using the previous published cohorts.

Furthermore, we note that the relative performance of signatures could be attributed to many factors, including but not limited to training data quality and sample size, methodology for signature selection, and ensuring biological relevance of the signature cohort similarity. While the results for any particular signature may not result from the difference between DAS and DEG approaches, the failure of any DEG signature to outperform DAS supports the value of the DAS methodology we describe.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Quantification of alternative splicing events
  - Mixed model analysis of longitudinal splicing changes

- Cohort definition and experimental design
- RNA library preparation and sequencing
- Machine learning predictor training and evaluation
- Microfluidic marker selection and analysis
- Characterization of alternatively spliced exons
- Functional and network analysis of alternatively spliced genes

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2023.100395.

ACKNOWLEDGMENTS

We thank the many US Navy corpsmen who assisted in the logistics and sample acquisition and the devoted Marine recruits who volunteered for this study. This study was approved by the Naval Medical Research Center (NMRC) institutional review board (IRB), protocol number NMRC.2020.0006, in compliance with all applicable US federal regulations governing the protection of human subjects. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense or the US government. A.G.L., C.W.G., R.A.L., S.E.L., and D.L.W. are military service members. This work was prepared as part of their official duties. This work was supported/funded by work unit numbers Defense Advanced Research Projects Agency contract number N6600119C4022 (S.C.S.), Defense Health Agency grant 9700130 through the Naval Medical Research Center (A.G.L.), National Institutes of Health grant R01GM071966 (O.G.T.), and Simons Foundation grant 395506 (O.G.T.).

AUTHOR CONTRIBUTIONS

Conceptualization, Z.Z., O.G.T., and S.C.S.; methodology and software, Z.Z.; formal analysis, Z.Z., N.S., and A.C.; investigation, I.R., V.D.N., G.N., E.Z., C.W.G., R.A.L., S.E.L., D.L.W., and A.G.L.; validation, H.R., J.B., J.G., N.R., D.K., M.T.M., C.W.W., R.H., T.W.B., and E.L.T.; resources, S.C.S. and O.G.T.; data curation, Y.G.; funding acquisition, A.G.L., S.C.S., and O.G.T.; project administration, A.G.L., S.C.S., and O.G.T.; supervision, S.C.S. and O.G.T.; writing – original draft, Z.Z., N.S., O.G.T., and S.C.S.; writing – review & editing, Z.Z., N.S., A.C., O.G.T., and S.C.S.

DECLARATION OF INTERESTS

A provisional patent application on this work has been submitted, and the Icahn School of Medicine at Mount Sinai and Princeton University are in discussions about licensing the technology. Z.Z., O.G.T., and S.C.S. are co-inventors of this technology and may benefit from any licensing.

Received: July 27, 2022
Revised: October 31, 2022
Accepted: January 9, 2023
Published: January 12, 2023

REFERENCES

1. Buonsenso, D., Sodero, G., and Valentini, P. (2022). Transcript host-RNA signatures to discriminate bacterial and viral infections in febrile children. Pediatr. Res. *91*, 454–463.
2. Galtung, N., Diehl-Wiesenecker, E., Lehmann, D., Markmann, N., Bergström, W.H., Wacker, J., Liesenfeld, O., Mayhew, M., Buturovic, L., Luethy, R., et al. (2022). Prospective validation of a transcriptomic severity classifier among patients with suspected acute infection and sepsis in the emergency department. Eur. J. Emerg. Med. *29*, 357–365. https://doi.org/10.1097/MEJ.0000000000000931.
3. Kucirka, L.M., Lauer, S.A., Laeyendecker, O., Boon, D., and Lessler, J. (2020). Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time since exposure. Ann. Intern. Med. *173*, 262–267.
4. Huang, Y., Zaas, A.K., Rao, A., Dobigeon, N., Woolf, P.J., Veldman, T., Øien, N.C., McClain, M.T., Varkey, J.B., Nicholson, B., et al. (2011). Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. PLoS Genet. *7*, e1002234.
5. Cappuccio, A., Geis, J., Ge, Y., Nair, V.D., Ramalingam, N., Mao, W., Chikina, M., Letizia, A.G., and Sealfon, S.C. (2022). Earlier detection of SARS-CoV-2 infection by blood RNA signature microfluidics assay. Clin. Transl. Discov. *2*, e47.
6. Tsalik, E.L., Henao, R., Montgomery, J.L., Nawrocki, J.W., Aydin, M., Lydon, E.C., Ko, E.R., Petzold, E., Nicholson, B.P., Cairns, C.B., et al. (2021). Discriminating bacterial and viral infection using a rapid host gene expression test. Crit. Care Med. *49*, 1651–1663.
7. McClain, M.T., Constantine, F.J., Nicholson, B.P., Nichols, M., Burke, T.W., Henao, R., Jones, D.C., Hudson, L.L., Jaggers, L.B., Veldman, T., et al. (2021). A blood-based host gene expression assay for early detection of respiratory viral infection: an index-cluster prospective cohort study. Lancet Infect. Dis. *21*, 396–404.
8. Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The expanding landscape of alternative splicing variation in human populations. Am. J. Hum. Genet. *102*, 11–26.
9. Thompson, M.G., Dittmar, M., Mallory, M.J., Bhat, P., Ferretti, M.B., Fontoura, B.M., Cherry, S., and Lynch, K.W. (2020). Viral-induced alternative splicing of host genes promotes influenza replication. Elife *9*, e55500. https://doi.org/10.7554/eLife.55500.
10. Banerjee, A.K., Blanco, M.R., Bruce, E.A., Honson, D.D., Chen, L.M., Chow, A., Bhat, P., Ollikainen, N., Quinodoz, S.A., Loney, C., et al. (2020). SARS-CoV-2 disrupts splicing, translation, and protein trafficking to suppress host defenses. Cell *183*, 1325–1339.e21.
11. Zhou, C., Liu, S., Song, W., Luo, S., Meng, G., Yang, C., Yang, H., Ma, J., Wang, L., Gao, S., et al. (2018). Characterization of viral RNA splicing using whole-transcriptome datasets from host species. Sci. Rep. *8*, 3273.
12. Martinez, N.M., and Lynch, K.W. (2013). Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. Immunol. Rev. *253*, 216–236.
13. Schaub, A., and Glasmacher, E. (2017). Splicing in immune cells-mechanistic insights and emerging topics. Int. Immunol. *29*, 173–181.
14. Cieślik, M., and Chinnaiyan, A.M. (2017). Cancer transcriptome profiling at the juncture of clinical translation. Nat. Rev. Genet. *19*, 93–109.
15. Zhang, Z., Pan, Z., Ying, Y., Xie, Z., Adhikari, S., Phillips, J., Carstens, R.P., Black, D.L., Wu, Y., and Xing, Y. (2019). Deep-learning augmented RNA-seq analysis of transcript splicing. Nat. Methods *16*, 307–310.
16. Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. *19*, 40.
17. Letizia, A.G., Ramos, I., Obla, A., Goforth, C., Weir, D.L., Ge, Y., Bamman, M.M., Dutta, J., Ellis, E., Estrella, L., et al. (2020). SARS-CoV-2 transmission among marine recruits during quarantine. N. Engl. J. Med. *383*, 2407–2416.
18. Thair, S.A., He, Y.D., Hasin-Brumshtein, Y., Sakaram, S., Pandya, R., Toh, J., Rawling, D., Remmel, M., Coyle, S., Dalekos, G.N., et al. (2021). Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections. iScience *24*, 101947.
19. Lee, J.S., Park, S., Jeong, H.W., Ahn, J.Y., Choi, S.J., Lee, H., Choi, B., Nam, S.K., Sa, M., Kwon, J.-S., et al. (2020). Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. Sci. Immunol. *5*, eabd1554. https://doi.org/10.1126/sciimmunol.abd1554.
20. Li, H.K., Kaforou, M., Rodriguez-Manzano, J., Channon-Wells, S., Moniri, A., Habgood-Coote, D., Gupta, R.K., Mills, E.A., Arancon, D., Lin, J., et al.

(2021). Discovery and validation of a three-gene signature to distinguish COVID-19 and other viral infections in emergency infectious disease presentations: a case-control and observational cohort study. Lancet. Microbe *2*, e594–e603.

21. McClain, M.T., Constantine, F.J., Henao, R., Liu, Y., Tsalik, E.L., Burke, T.W., Steinbrink, J.M., Petzold, E., Nicholson, B.P., Rolfe, R., et al. (2021). Dysregulated transcriptional responses to SARS-CoV-2 in the periphery. Nat. Commun. *12*, 1079.
22. Aschenbrenner, A.C., Mouktaroudi, M., Krämer, B., Oestreich, M., Antonakos, N., Nuesch-Germano, M., Gkizeli, K., Bonaguro, L., Reusch, N., Baßler, K., et al. (2021). Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. Genome Med. *13*, 7.
23. Kwan, P.K.W., Cross, G.B., Naftalin, C.M., Ahidjo, B.A., Mok, C.K., Fanusi, F., Permata Sari, I., Chia, S.C., Kumar, S.K., Alagha, R., et al. (2021). A blood RNA transcriptome signature for COVID-19. BMC Med. Genom. *14*, 155.
24. Wong, A.K., Krishnan, A., and Troyanskaya, O.G. (2018). Giant 2.0: genome-scale integrated analysis of gene networks in tissues. Nucleic Acids Res. *46*, W65–W70.
25. Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576.
26. Rivera-Colón, Y., Schutsky, E.K., Kita, A.Z., and Garman, S.C. (2012). The structure of human GALNS reveals the molecular basis for mucopolysaccharidosis IV A. J. Mol. Biol. *423*, 736–751.
27. Rollinger-Holzinger, I., Eibl, B., Pauly, M., Griesser, U., Hentges, F., Auer, B., Pall, G., Schratzberger, P., Niederwieser, D., Weiss, E.H., and Zwierzina, H. (2000). LST1: a gene with extensive alternative splicing and immunomodulatory function. J. Immunol. *164*, 3169–3176.
28. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, D412–D419.
29. Yang, Y.-C.T., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J., and Lu, Z.J. (2015). CLIPdb: a CLIP-seq database for protein-RNA interactions. BMC Genom. *16*, 51.
30. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. Nature *583*, 711–719.
31. Yang, H., Duckett, C.S., and Lindsten, T. (1995). iPABP, an inducible poly(A)-binding protein detected in activated human T cells. Mol. Cell Biol. *15*, 6770–6776.
32. Turner, M., and Díaz-Muñoz, M.D. (2018). RNA-binding proteins control gene expression and cell fate in the immune system. Nat. Immunol. *19*, 120–129.
33. Sauerwald, N., Zhang, Z., Ramos, I., Nair, V.D., Soares-Schanoski, A., Ge, Y., Mao, W., Alshammary, H., Gonzalez-Reiche, A.S., van de Guchte, A., et al. (2022). Pre-infection antiviral innate immunity contributes to sex differences in SARS-CoV-2 infection. Cell Syst. *13*, 924–931.e4.
34. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
35. Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J.G.R., Storey, R., Swarbreck, D., et al. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol. *7* (*Suppl 1*), S4.1–S4.9.
36. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.
37. Shen, S., Park, J.W., Lu, Z.-X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. USA *111*, E5593–E5601.
38. Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
RNA sequencing	Illumina	N/A
Microfluidic PCR assay	Fluidigm	Biomark HD
PCR testing for SARS-CoV-2	Thermo-Fisher	TaqPath COVID-19 Combo Kit
Deposited data		
RNA-sequencing	Sauerwald et al.[33]	GEO: GSE198449
Software and algorithms		
Github repository for Jupyter notebooks of data analysis	This paper	https://github.com/zj-zhang/CHARM-AlternativeSplicing https://doi.org/10.5281/zenodo.7455070
Github repository for statistical software	This paper	https://github.com/zhanglab-aim/JEMM https://doi.org/10.5281/zenodo.7455172

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Olga G. Troyanskaya (ogt@genomics.princeton.edu).

Materials availability

This study did not generate new materials.

Data and code availability

- The raw sequencing data and subjects metadata used in this study are deposited to NCBI GEO with accession number GSE198449.
- Code implementation, processed results, and reproducible analyses are publicly available at the URL https://github.com/zj-zhang/CHARM-AlternativeSplicing. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Quantification of alternative splicing events

A uniform pipeline was employed to process all RNA-seq fastq files. In particular, STAR (v2.7.4)[34] was used to align the reads to hg38 genome build with Gencode v34 index.[35] To quantify gene expression levels, kallisto (v0.46.0)[36] was used to pseudo-align RNA-seq reads to Gencode v34 transcripts. Throughout this study, Gencode v34 genome annotation was used as the reference gene annotations wherever applicable.

To reduce potential counting bias, two distinct approaches leveraging different aspects of RNA-seq reads to quantify Percent Spliced In (PSI) were employed and combined. Using genome read alignment generated by STAR as input, the junction read counts for alternative splicing events were counted by DARTS/rMATS-turbo.[15,37] Using the transcript quantifications generated by kallisto as input, the ratio between longer and shorter isoforms were computed by SUPPA2.[16] We analyzed four basic types of alternative splicing events, i.e., skipped exons, alternative 5′ splice sites, alternative 3′ splice sites, and retained introns. This approach allowed us to identify measurement-agnostic differential splicing events for applicable event types and overcome the technical bias due to relatively shallow sequencing coverage per RNA-seq dataset (Figure S4).

Mixed model analysis of longitudinal splicing changes

To identify the alternatively spliced exons upon SARS-CoV-2 infection, we employed a linear mixed model regression framework to study the dependency of exon usage on disease stage, sex, and potential confounding factors:

$$logit(\psi_{ijl}) = \mu_i + \alpha Sex_j + \beta Disease_j + P_{ij} + \delta_i 1(\psi_{ijl} \in \Psi_{JCT}) + \sum_k \gamma_k PC_{kj} + \varepsilon_{ij}$$

where ψ_{ijl} is the inclusion level for alternative splicing event i in the RNA-seq sample j measured by approach l, where l is either exon-exon junction counts or isoform ratios. μ_i is the baseline inclusion level for AS event i. Sex_j and $Disease_j$ are sex and annotated disease stage for sample j with regression coefficients α and β, respectively. P_{ij} is the random effect for sample j to account for the covariance among multiple RNA-seq samples coming from the same subject. δ_i quantifies the difference between measurement approaches for AS event i if ψ_{ijl} is measured by counting exon-exon junction reads $\psi_{ijl} \in \Psi_{JCT}$ as compared to isoform ratios, and $1(\cdot)$ is the indicator function.

Finally, to control for potential batch effects, we performed principal component analysis using all PSI levels. Among the first 10 principal components (PC), PCs with no correlation with the biological variables of sex or disease stage (Pearson correlation $p > 0.01$) were considered potential confounders. These were included in the regression model to estimate their coefficients γ_k and control for their potential confounding effects.

The regression fixed-effect coefficients and random-effect variance components were estimated by python implementation statsmodels (v0.11.1). Statistical significance was determined by Wald tests and P-values were multiple-testing corrected by Benjamini-Hochberg False Discovery Rate (FDR). Alternative splicing events with FDR<0.05 for β were considered as disease-stage dependent. Detailed code implementation and reproducible analyses are publicly available at the URL https://github.com/zj-zhang/CHARM-AlternativeSplicing.

To separate true infection-induced AS changes from those induced by military training, we ran a modified version of the above model and identified training-specific AS events to exclude from further analysis. To this end, we ran a modified version of the disease-specific model on all control samples including time since enrollment as a variable:

$$logit(\psi_{ijl}) = \mu_i + \alpha Sex_j + \beta_1 TimePoint_j + \beta_2 Sex_j \times TimePoint_j + P_{ij} + \delta_i 1(\psi_{ijl} \in \Psi_{JCT}) + \sum_k \gamma_k PC_{kj} + \varepsilon_{ij}$$

where $TimePoint_j$ is the collection time of sample j with respect to the initial enrollment of the subject. Alternative splicing events with FDR<0.05 for either the main time effect β_1 or its interaction term with sex β_2 were deemed as subject to military training effects and as such excluded from the downstream analyses.

Cohort definition and experimental design

In the COVID-19 Health Action Response (CHARM) study, whole-blood specimens were collected from n = 371 US Marine recruits in a longitudinal cohort recruited by the study entering basic training at Parris Island, South Carolina from May to November, 2020.[17] 301 out of 371 subjects (81.1%) had repeated measures. Biospecimens were sequenced by Illumina high-throughput sequencing with paired-end reads of 101 bp at an average depth of 25 million. All study participants were tested for SARS-CoV-2 by PCR, had serum drawn to assess antibody status, and were administered a symptom questionnaire as well as demographic information at enrollment, and approximately 7, 14, 28, 42, and 56 days afterward. SARS-CoV-2 qPCR testing was performed in mid-turbinate nares swabs and were performed within 48 h of sample collection at high complexity Clinical Laboratory Improvement Amendments-certified laboratories using the US Food and Drug Administration-authorized Thermo Fisher TaqPath COVID-19 Combo Kit (Thermo Fisher Scientific, Waltham, MA, USA). Lab24Inc (Boca Raton, FL, USA) performed PCR testing from study initiation (May 11, 2020) until Aug 24, 2020, and the Naval Medical Research Center (Silver Spring, MD, USA) from Aug 24, 2020, until the conclusion of the study (Nov 2, 2020).

Depending on the infection status determined by PCR and antibody tests, RNA-seq samples were annotated to five disease stages. RNA-seq samples from initial enrollment with negative PCR tests (PCR-) and negative antibody calls were annotated as healthy controls. The RNA-seq samples collected at the last PCR-test before the subject turned PCR positive (PCR+) were annotated as pre-infection, or Pre. The RNA-seq samples at the first time a subject turned PCR+ were annotated as first-time infection, or First. Following the initial PCR+, all other RNA-seq samples while the subject remained PCR+ were annotated as mid-infection, or Mid. Finally, the RNA-seq samples from the subjects that turned PCR-after the previous infection were annotated as post-infection, or Post. None of the SARS-CoV-2 infected subjects in the CHARM cohort required hospitalization nor any treatment.

An independent test set from a cohort of 47 COVID-19 patients and 19 healthy controls from Duke University Hospital (herein referred to as Duke cohort) was profiled by whole-blood RNA-seq and processed using the same software pipeline as the CHARM study. The Duke cohort consisted COVID-19 patients with distinct characteristics from the CHARM training set (see Table S1). The Duke cohort remained unexposed to the classifier during training and validation. Prediction accuracy was measured using AUROC for the binary classification of infected subjects vs healthy controls in the held-out Duke dataset as an independent evaluation.

RNA library preparation and sequencing

Total RNA from PAXgene preserved blood was extracted using the Agencourt RNAdvance Blood Kit (Beckman Coulter) on a BioMek FXP Laboratory Automation Workstation (Beckman Coulter). Concentration and integrity (RIN) of isolated RNA were determined using Quant-iT RiboGreen RNA Assay Kit (Thermo Fisher) and an RNA Standard Sensitivity Kit (DNF-471, Agilent Technologies, Santa Clara, CA, USA) on a Fragment Analyzer Automated CE system (Agilent Technologies), respectively. Subsequently, cDNA libraries were constructed from total RNA using the Universal Plus mRNA-Seq kit (Tecan Genomics, San Carlos, CA, United States) in a Biomek i7 Automated Workstation (Beckman Coulter). Briefly, mRNA was isolated from purified 300 ng total RNA using oligo-dT beads

and used to synthesize cDNA following the manufacturer's instructions. The transcripts for rRNA and globin were further depleted using the AnyDeplete kit (Tecan Genomics) prior to the amplification of libraries. Library concentration was assessed fluorometrically using the Qubit dsDNA HS Kit (Thermo Fisher), and quality was assessed with the Genomic DNA 50Kb Analysis Kit (DNF-467, Agilent Technologies). Following library preparation, samples were pooled, and preliminary sequencing of cDNA libraries (average read depth of 90,000 reads) was performed using a MiSeq system (Illumina) to confirm library quality and concentration. Deep sequencing was subsequently performed using an S4 flow cell in a NovaSeq sequencing system (Illumina) (average read depth ~30 million pairs of 2 × 100 bp reads) at New York Genome Center.

Machine learning predictor training and evaluation

Logistic regression was employed as the classifier to distinguish infected subjects from the healthy controls. To train the classifier, samples with first-time PCR+ (First) and healthy controls without a pre-infection (Pre) sample were considered, while subjects with both control and pre-infection samples were also held-out to identify potential early immune responses markers. To address the class imbalance, positive samples were up-weighted by their ratios in the training set, while negative samples were kept a default weight of 1. Parameters were trained with 10-fold cross validation on the CHARM samples.

We additionally examined six publicly available gene expression signatures for SARS-CoV-2 infection.[18–23] We denote the best performing signature sets we derived from the CHARM cohort as CHARM DAS and DEG, respectively. To rigorously compare the quality of signature sets and remove other confounding effects (such as differences in discovery cohorts and machine learning algorithms used in previous studies), we re-trained classifiers using the same logistic regression classifier on the same CHARM cohort, tested using the same held-out Duke cohort, while only varying the gene signature sets. All public signature sets were processed identically to CHARM DEG, that is, getting the gene expression matrix based on the signature set as features, training a logistic regression using these features on the CHARM cohort, then testing on Duke cohort. Classification accuracies were measured by area under the receiver-operating curves (AUROC).

Microfluidic marker selection and analysis

To select a smaller set of biomarkers for testing on microfluidic devices, we ranked each AS event by their absolute coefficient values in the classifier trained in the CHARM cohort, and performed a forward selection to select non-redundant signatures. Briefly, we went from the top of the ranked AS event list, and only added one event to the feature set, if adding it to the current feature set improved the discriminative power in the Duke cohort. A set of n = 27 AS biomarkers were selected by this process and transferred to Fluidigm Corp for independent validation on a clinical cohort of n = 31 SARS-CoV-2 infected samples vs n = 31 healthy controls.

We re-analyzed a previously published microfluidic PCR data based on gene expression markers[5] by following the same principal component analysis as the AS assay. The first two principal components for both assays were fitted by a support vector machine with a linear kernel. Based on the linear separation of samples, accuracy, positive percent agreement and negative percent agreement were calculated. To accommodate for the sample size differences between the gene expression and AS assay, we downsampled the gene expression-profiled samples to match the same number of positive and negative samples in the AS assay for 10,000 times. P-value was computed as the frequency that the downsampled gene expression accuracies were equal or greater than the observed accuracy in the AS assay.

Characterization of alternatively spliced exons

Protein domains for each exon were annotated by Pfam and downloaded from UCSC table browser (https://genome.ucsc.edu/). The clan information (v32.0) for each protein domain was accessed from the Pfam ftp server (http://ftp.ebi.ac.uk/pub/databases/Pfam/). Alternative spliced exons were annotated on a per-splice site basis, where each alternative splicing event (across all four types of alternative splicing we analyzed) had four key splice sites that uniquely identified an event. Annotated domains for all four splice sites per event were assigned to that splicing event, and subsequently merged into the clan wherever applicable. Only protein domains with at least 1 observation were analyzed.

RNA-binding proteins binding sites were profiled by eCLIP assays from ENCODE. eCLIP peaks were downloaded from ENCODE data portal (https://www.encodeproject.org/), and IDR peaks were used as a set of high-confidence peaks for the RBP enrichment analysis. Similar to the domain analysis, the alternative splicing events were annotated to RBP binding sites on a per splice-site basis. Peaks that overlapped within ±250 bp to a splice site were annotated to the AS event. Annotated RBP binding sites for all four splice sites per event were assigned to that splicing event.

Fisher exact test was employed to test the enrichment of protein domains and RBP binding sites. Foreground was defined as the significantly differentially spliced events, and background was defined as all tested events. To increase statistical power, we pooled different types of significant DAS events from all disease stages, and analyzed them for the enrichment of various protein domains and RBP binding sites. Subsequently, the ratios of protein domains, or RBP binding sites, were compared between foreground and background and P-values were multiple-testing corrected by FDR.

Functional and network analysis of alternatively spliced genes

Functional networks in HumanBase[24] were used to analyze the blood tissue-specific network modules and functional annotation enrichment for differentially spliced genes across the disease stages. Genes with any significant DAS events were included as

alternatively spliced genes in the functional network analysis. Detected modules for the whole blood-specific network were analyzed for the enriched Gene Ontology (GO) terms. To reduce and cluster similar GO terms, we used ReviGO web server (http://revigo.irb.hr)[38] to summarize GO terms for analysis of the temporal changes.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests and computational analyses are explained in each subsection of the STAR Methods. Quantification of AS level from RNA-seq datasets was described in Section "quantification of alternative splicing events" in STAR Methods. Statistical analysis for calling DAS events was described in Section "mixed model analysis of longitudinal splicing changes" in STAR Methods. Identification and evaluation of AS and transcriptome biomarkers was described in Section "machine learning predictor training and evaluation". Analysis of diagnostic assay based on DAS and DEG signatures was described in "microfluidic marker selection and analysis".