It is made available under a CC-BY-NC-ND 4.0 International license .

A gene expression-based diagnostic classifier for identification of severe COVID-19 and 1

multisystem inflammatory syndrome in children (MIS-C) 2

- Alicia Sotomayor-Gonzalez^{1#}, Conor J. Loy^{2#}, Jenny Nguyen^{1#}, Venice Servellita^{1#}, Sanchita 3
- Bhattacharva⁴, Joan Lenz², Meagan Williams⁵, Will Suslovic⁵, Alexandre P, Cheng², Andrew 4
- Bliss², Prachi Saldhi¹, Jessica Streithorst¹, Hee Jae Huh⁶, Abiodun Foresythe², Miriam 5
- Osequera¹, Katrina de la Cruz¹, Noah Brazer¹, Nathan Wood³, Charlotte Hsieh³, Burak Bahar⁵, 6
- Amelia Gliwa¹, Kushmita Bhakta⁶, Maria A. Perez⁶, Evan J. Anderson^{6,7}, Ann Chahroudi⁶, 7
- Meghan Delaney⁵, Atul J. Butte⁴, Roberta DeBiasi⁵, Christina A. Rostad⁶, Iwijn De Vlaminck^{2*}, 8
- Charles Chiu^{1,8*}. 9
- 10
- 11
- Affiliations: ¹Department of Laboratory Medicine, University of California, San Francisco, San Francisco, 12
- CA, USA 13
- ²Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA 14
- ³UCSF Benioff Children's Hospital, Oakland, CA, USA 15
- ⁴Bakar Computational Health Sciences Institute, University of California, San Francisco, San 16
- Francisco, CA, USA 17
- ⁵Children's National Hospital, Washington D.C., USA 18
- ⁶Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan 19
- University School of Medicine, Seoul, South Korea 20
- ⁶Department of Pediatrics and Center for Childhood Infections at Vaccines, Emory University 21
- 22 School of Medicine and Children's Healthcare of Atlanta, Atlanta, GA, USA-
- 23 ⁷Department of Medicine, Emory University School of Medicine, Atlanta, GA, US
- ⁸Department of Medicine, Division of Infectious Diseases, University of California, San 24
- Francisco, San Francisco, CA, USA 25
- 26
- [#]These authors contributed equally 27
- *Co-corresponding authors: charles.chiu@ucsf.edu, id93@cornell.edu 28
- 29

It is made available under a CC-BY-NC-ND 4.0 International license .

31 **KEY WORDS:** multisystem inflammatory syndrome in children (MIS-C), SARS-CoV-2,

coronavirus disease 2019, cell-free RNA, whole blood RNA, RNA sequencing, RNA-Seq,

33 clinical severity, nucleic acid sequencing, host response, disease biomarkers, classifier

- ³⁴ models, machine learning, random forest, reverse transcription polymerase chain reaction.
- 35

36 **ABSTRACT**

MIS-C is a severe hyperinflammatory condition with involvement of multiple organs that 37 occurs in children who had COVID-19 infection. Accurate diagnostic tests are needed to guide 38 management and appropriate treatment and to inform clinical trials of experimental drugs and 39 vaccines, yet the diagnosis of MIS-C is highly challenging due to overlapping clinical features 40 with other acute syndromes in hospitalized patients. Here we developed a gene expression-41 based classifier for MIS-C by RNA-Seq transcriptome profiling and machine learning based 42 analyses of 195 whole blood RNA and 76 plasma cell-free RNA samples from 191 subjects, 43 including 95 MIS-C patients, 66 COVID-19 infected patients with moderately severe to severe 44 disease, and 30 uninfected controls. We divided the group into a training set (70%) and test 45 set (30%). After selection of the top 300 differentially expressed genes in the training set, we 46 47 simultaneously trained 13 classification models to distinguish patients with MIS-C and COVID-19 from controls using five-fold cross-validation and grid search hyperparameter tuning. The 48 final optimal classifier models had 100% diagnostic accuracy for MIS-C (versus non-MIS-C) 49 50 and 85% accuracy for severe COVID-19 (versus mild/asymptomatic COVID-19). Orthogonal validation of a random subset of 11 genes from the final models using quantitative RT-PCR 51 52 confirmed the differential expression and ability to discriminate MIS-C and COVID-19 from

It is made available under a CC-BY-NC-ND 4.0 International license .

controls. These results underscore the utility of a gene expression classifier for diagnosis of
 MIS-C and severe COVID-19 as specific and objective biomarkers for these conditions.

55

56 **INTRODUCTION**

Multisystem inflammatory syndrome in children (MIS-C) is a severe post-infectious 57 complication of SARS-CoV-2 infection in pediatric patients that is characterized by severe 58 disease with systemic hyperinflammation and multi-organ involvement¹. On average, 59 symptoms of MIS-C first present two to four weeks after acute COVID-19 illness and can 60 involve a constellation of respiratory, gastrointestinal, cardiac, renal, dermatologic, and 61 neurological symptoms that in 60% of cases result in hospitalization and possible ICU stay with 62 invasive mechanical ventilation required due to inadequate oxygenation². Diagnosis of MIS-C 63 is challenging due to overlapping clinical features with other hyperinflammatory illnesses, such 64 as Kawasaki Disease (KD) and Toxic Shock Syndrome (TSS)^{3,4}, and the lack of objective 65 biomarker based diagnostic tests hinders accurate diagnosis and effective management and 66 treatment for this condition⁵. As of August 2022, the CDC has reported 8,798 cases of MIS-C 67 and 71 deaths attributed to MIS-C in children⁶, defined as individuals under 21 years of age. 68 Transcriptome analysis by RNA sequencing (RNA-Seq) has been shown to be useful in 69 the diagnosis of rare genetic diseases⁷ and infections such as Lyme disease⁸, influenza⁹, and 70 COVID-19¹⁰. In a previous study¹¹, analysis of plasma cell-free RNA from patients with MIS-C 71 72 or severe COVID-19 yielded distinct signatures of cell injury and death between these two disease states, including the involvement of unexpected pathways such as endothelial and 73 neuronal Schwann cell signaling. These signatures were different from those from whole blood 74 RNA profiling¹⁰⁻¹³, which showed upregulation of pro-inflammatory signaling pathways in 75

It is made available under a CC-BY-NC-ND 4.0 International license .

76 COVID-19 but downregulation of T cell-associated pathways in MIS-C. Given the findings of distinct signaling pathways¹¹, we hypothesized that whole blood and plasma would be 77 promising analytes for the development of diagnostic assays for severe COVID-19 and MIS-C. 78 79 Here we trained machine learning algorithms to identify panels of differentially expressed genes that can distinguish MIS-C or severe COVID-19 from uninfected controls 80 (donors or patients with other inflammatory diseases). We obtained performance accuracies 81 for the gene panels of 85-100% and confirmed the differential expression of a subset of genes 82 by gRT-PCR. Our results lay the groundwork for the development and clinical validation of 83 multiplexed RNA gene expression-based assays for MIS-C and severe COVID-19. 84

85

RESULTS 86

LTS Transcriptome profiling using RNA-Seq was performed on 195 whole blood RNA 87 (wbRNA) samples and 76 plasma cell-free RNA (cfRNA) samples from 191 subjects, including 88 95 MIS-C patients, 66 COVID-19 infected patients, and 30 uninfected controls. A mean of 31 89 million reads were generated per whole blood sample and a mean of 8.6 million reads were 90 generated per cell-free sample. No batch effect based on collection center was observed. 91

We divided the samples into two sets, a training set (70%) and a test set (30%), with 92 consideration to sample origin and severity of disease (Fig. 1A). Next, we performed feature 93 selection on the training set. We removed genes with low counts and near zero variance. We 94 selected the top 300 relevant features selected using differential expression/abundance 95 analysis in DESeg 2^{14} (Benjamini-Hochberg adjusted p-value < 0.05, ranked by fold change). 96

97

It is made available under a CC-BY-NC-ND 4.0 International license .



It is made available under a CC-BY-NC-ND 4.0 International license .

108	that optimize classifier performance, we used Receiver Operating Characteristic (ROC)
109	analysis. Next, we applied the trained models to the test set samples, with the classification
110	score threshold determined from the training step, and we quantified the performance of each
111	model for both wbRNA and cfRNA biomarkers (Fig. 2A).
112	We observed high classification performance comparing samples from patients with
113	MIS-C and COVID-19 for both wbRNA and cfRNA using 9 machine learning algorithms (Fig.
114	2A , test and train area under the curve (AUC) > 0.95). The generalized linear models with
115	Ridge and LASSO feature selection performed the best for both wbRNA and cfRNA (Fig. 2B-
116	C , wbRNA: accuracy=0.95, sensitivity=1, specificity=0.83; cfRNA: accuracy=0.93,
117	sensitivity=0.96, specificity=0.89). We also observed high classification performance
118	comparing samples from patients with MIS-C and good health for both wbRNA and cfRNA and
119	for most models, as would be expected. Surprisingly, we observed lower classification
120	performance comparing samples from patients with COVID-19 and good health, particularly in
121	the cell-free RNA samples (Fig. 2A). We attribute this to overfitting of the training model, due
122	to a combination of a small sample size for this comparison and the heterogeneity of the
123	affected population. For the severe versus mild/asymptomatic COVID-19 comparison using
124	wbRNA, the best performing algorithms included Random Forest (RF) Extra Trees, Naïve
125	Bayes (NB), and Classification and Regression Trees (RPART), with all models ddyielding an
126	accuracy of >85% (Fig. 3A).
127	Next, we incorporated clinical metadata into our modeling using the Random Forest
128	Extra Trees algorithm (Fig. 4A-B). We observed high classification performance for both
129	wbRNA and cfRNA in differentiating between samples from patients with MIS-C and COVID-

130 19. Incorporating the clinical metadata increased the performance of the cfRNA model, but did

It is made available under a CC-BY-NC-ND 4.0 International license .

131	not impact the performance of the wbRNA model (Fig. 4C). Both models were performing well
132	to begin with, and we believe that a larger data set would be needed to better measure the
133	difference incorporating clinical metadata has on classification performance.
134	
135	
136	
137	
138	ils
139	for detaile
140	int DOLID.
141	anuscript
142	see mai

It is made available under a CC-BY-NC-ND 4.0 International license .





144 Figure 2. Comparison of machine learning classification algorithms using RNA-seq

145 data. (A) Receiver operator curve (ROC) area under the curve (AUC) metrics of 13 machine

It is made available under a CC-BY-NC-ND 4.0 International license .

learning algorithms and logistic regression for training and test sets in cfRNA or wbRNA across
 comparisons. (B) Train and test performance of a generalized linear model machine learning
 algorithm using Ridge feature selection in distinguishing MIS-C and Moderate/Severe COVID-

149 19 in cfRNA and **(C)** wbRNA.



150



152 COVID-19 severity. (A) Accuracy, sensitivity, and specificity of 13 machine learning

algorithms and logistic regression for training and test sets in wbRNA. (B) Receiver operator

It is made available under a CC-BY-NC-ND 4.0 International license .

curve (ROC) area under the curve (AUC) plot of wbRNA test set using the RF Extra Trees 154

algorithm. (C) Violin plot of classifier scores from the RF Extra Trees algorithm on the train and 155

test sets. (D) Confusion matrix of RF Extra Trees algorithm performance on the test set. 156

157



159

160 Figure 4. A composite model incorporating RNA-seg and clinical metadata in classifying

MIS-C from Moderate/Severe COVID-19 (A) Train and test performance of a Random Forest 161

It is made available under a CC-BY-NC-ND 4.0 International license .

Extra Trees classification algorithm in distinguishing MIS-C and Moderate/Severe COVID-19 in cfRNA and **(B)** wbRNA utilizing RNA-seq and clinical metadata. **(C)** Comparison of ROC AUC metrics between Random Forest Extra Trees classification algorithms utilizing RNA-seq data with and without the addition of clinical metadata in distinguishing MIS-C and Moderate/Severe COVID-19 in cfRNA and wbRNA.

167

Finally, we evaluated the performance of gene expression from a subset of top differentially 168 expressed genes (DEGs) in whole blood samples from MIS-C, severe COVID-19 (excluding 169 MIS-C cases), and uninfected controls. A subset of 12 differentially expressed genes 170 (FCER1A, ADAMTS2, CD177, LMOD1, KCNA5, OLAH, OTOF, TIFAB, B3GALT2, ITGA7, 171 PLA267, and IFI27) were randomly selected and tested. The results, expressed in cycle 172 threshold (C) values, were concordant with the relative differences in expression levels and 173 direction of gene expression (upregulation or downregulation), as reported in the previous 174 study¹¹. Differences din expression between MIS-C or severe COVID-19 and uninfected 175 176 controls were statistically significant for four genes (ADAMTS2, CD177, OLAH, and TIFAB), whereas the differences between MIS-C and severe COVID-19 were significant for three 177 genes (ADAMTS2, CD177, and OLAH). 178

179

It is made available under a CC-BY-NC-ND 4.0 International license .

181 Figure 5. Confirmation of differential gene expression by quantitative RT-PCR (qRT-

- **PCR).** Twelve genes predicted to distinguish MIS-C and severe COVID-19 from uninfected
- 183 controls were tested by qRT-PCR with 3 sample replicates and 2 assay replicates per gene.
- 184 Genes highlighted in red text are downregulated in MIS-C and COVID-19 compared to
- controls. The symbols and error bars denote the mean and standard deviation, respectively,
- 186 for the 3 sample replicates.



Differentially Expressed Gene

187

188 DISCUSSION

Here we developed classifier models for severe manifestations of COVID-19, including MIS-C and moderate-to-severe, non-MIS-C COVID-19, consisting mostly of pneumonia cases, that can result in hospitalization and adverse clinical outcomes such as ICU admission, endorgan failure, and death¹⁵. Transcriptome profiling by RNA-Seq was performed on 195 wbRNA and 76 plasma cfRNA samples from 191 subjects, and sequencing data from 70% of samples assigned to a training set were used to generate models with 100% accuracy in discriminating MIS-C from non-MIS-C, and >85% accuracy in severe COVID-19 versus mild/asymptomatic It is made available under a CC-BY-NC-ND 4.0 International license .

COVID-19. We subsequently confirmed differential gene expression among MIS-C, severe
 COVID-19, and uninfected controls by orthogonal qRT-PCR of 11 genes taken from the final
 model. These findings underscore the potential clinical utility of gene expression-based
 classification in the development and validation of diagnostic assays for MIS-C and severe
 COVID-19.

A high degree of overlap in symptomatology and clinical presentation between severe 201 manifestations of COVID-19 and other acute illnesses in hospitalized patients has been 202 reported^{3,4}. Acute syndromes that can mimic MIS-C include Kawasaki disease³, toxic shock 203 syndrome³, bacterial or viral sepsis⁴, and even non-infectious conditions such as congestive 204 heart failure, whether or not directly related to MIS-C¹⁶. In contrast, acute illnesses that can 205 mimic severe COVID-19 include many infections¹⁷⁻¹⁹, including those caused by respiratory 206 207 viruses (e.g., influenza virus, parainfluenza virus, adenovirus, etc.), bacterial pneumonia, including tuberculosis, malaria, and chronic obstructive pulmonary disease exacerbation. 208 However, whereas molecular or antigen tests for SARS-CoV-2 from nasal swabs can readily 209 210 diagnose COVID-19, and diagnostic tests for many other illnesses in the differential diagnosis are available, specific biomarkers and tests for MIS-C are lacking to date. Such diagnostic 211 tests for MIS-C would be useful to inform accurate and timely management of patients with 212 inflammatory diseases that have clinically overlapping presentations. They may also be used 213 as a "companion diagnostic" to clinical trials of drugs and/or vaccines by providing an objective 214 measure of the response to and effectiveness of an intervention 20 . 215

216 Similarly, there is an urgent need for diagnostic tests that can establish whether a 217 patient has severe COVID-19. This is especially important as the widespread availability of 218 effective vaccines to prevent severe complications of COVID-19 has led to a sharp decline in

It is made available under a CC-BY-NC-ND 4.0 International license .

the number of cases in the United States and worldwide as of September 2022²¹, which may decrease clinical vigilance for patients at high risk of life-threatening complications or death from COVID-19. Timely diagnosis can enable patients to promptly receive antiviral therapies such as ritonavir-boosted nirmatrelvir(Paxlovid)²², the effectiveness of which wanes in patients with delayed diagnosis and more severe disease, thereby decreasing lengths of stay in the hospital and reducing utilization of limited health care resources.

Our study employed orthogonal confirmation of transcriptome profiling results by 225 multiplex gRT-PCR²³. This approach not only supports the accuracy of the gene expression-226 based models, but also highlights how these assays may be introduced into the clinical setting 227 soon. Clinical multiplex gRT-PCR syndromic panels are now widely available for diagnosis of 228 multiple infectious diseases²⁴⁻²⁶, including neurological infections, acute respiratory illness, and 229 diarrheal disease, or gastroenteritis. These panels have the advantage of moderate to high-230 throughput, batch testing capability, and low cost, none of which is the case with next-231 generation sequencing based platforms. Thus, diagnostic assays based on a condensed panel 232 233 of 30-50 genes may be more conducive to clinical laboratory workflows than those based on next-generation sequencing. 234

Our study has some limitations. First, we had a very limited number of samples from patients with "MIS-C-like" illnesses – Kawasaki disease, toxic shock syndrome, and/or acute bacterial sepsis^{3,4}. Comparisons between MIS-C and these aforementioned diseases is probably more useful than comparisons between MIS-C and COVID-19 or donor controls. Second, without longitudinal samples, we were unable to ascertain the prognostic utility of classifier models in predicting clinical outcomes, whether patients will clinically deteriorate and develop more severe disease over time. Third, although we used a fully independent test set in

It is made available under a CC-BY-NC-ND 4.0 International license .

these analyses, the divergence in assay performance between the training and test set data

suggests that the models may be slightly overfit; additional sample sizes are likely needed to

address this problem.

- 245
- 246

247 **METHODS**

248 Ethics Statement. The University of California, San Francisco (UCSF) Institutional Review

Board (IRB) (#21-33403), San Francisco, CA; Emory University IRB (STUDY00000723),

Atlanta, GA; Childrens National Medical Center (CNMC) IRB (Pro00010632), Washington, DC;

and Cornell University IRB for Human Participants (2012010003), New York, NY each

252 approved the protocols for this study. All samples and patient information were de-identified for

analysis and sharing with collaboration institutions. At Emory University the IRB approved

254 protocol was a prospective enrollment study under which parents provided consent and

children assent as appropriate for age. At CNMC and UCSF, the IRB protocols were no

subject contact sample biobanking protocols under which content was not obtained and data

257 was extracted from medical charts.

258

Sample Acquisition UCSF. Samples were acquired from UCSF as previously described¹¹.
Briefly, hospitalized pediatric patients were identified as having COVID-19 by testing positive
with SARS-CoV-2 real-time PCT (RT-PCR). Whole blood samples were collected in EDTA
lavender top tubes and diluted 1:1: in DNA/RNA shield (Zymo Research). Remaining blood
was centrifuged at 2500 rpm for 15 min and the available plasma was retained. All samples
were stored in a -80°C freezer until used.

It is made available under a CC-BY-NC-ND 4.0 International license .

2	c	г.
2	ь	Э

266	Sample Acquisition Emory and Children's Healthcare of Atlanta. Samples were acquired
267	from Emory and Children's Healthcare of Atlanta as previously described ¹¹ . Briefly, pediatric
268	patients were classified as having COVID-19 via SARS-CoV-2 RT-PCR and as having MIS-C if
269	they met the CDC case definition. Controls were healthy outpatients with no known history of
270	COVID-19 who volunteered for specimen collection. Whole blood was collected in EDTA
271	lavender top tubes and aliquoted for plasma extraction via centrifugation at 2500 rpm for
272	15min. Samples were stored in a -80 $^\circ$ C freezer and shipped on dry ice to either UCSF or
273	Cornell for analysis.
274	int DOI TO
275	Sample Acquisition Children's National. Samples were acquired from Children's National as
276	previously described ¹¹ . Briefly, pediatric patients were classified as having MIS-C if they met
277	the CDC case definition. Whole blood samples were collected and centrifuged at 1300g for 5
278	minutes at room temperature. Plasma was aliquoted into a cryovial and frozen at -80°C. A
279	DMSO-based cryopreservative (Cryostor CS10) was added in a 1:1 ratio to the cell pellet and
280	then frozen at -80°C in a controlled rate freezing container (i.e., Mr. Frosty) and then
281	transferred to liquid nitrogen within 1 week.
282	Clinical Data. Patients were stratified as previously described ¹¹ . For the purposes of this
283	study, patients were classified as having MIS-C by multidisciplinary teams which adjudicated
284	whether a patient met the CDC case definition of MIS-C. COVID-19 was defined as any patient
285	with PCR-confirmed SARS-CoV-2 infection within the preceding 14 days who did not also
286	meet the MIS-C case definition. Clinical data was abstracted from the medical record and
287	inputted into a shared REDCap database housed at UCSF.

It is made available under a CC-BY-NC-ND 4.0 International license .

288

289 **Severity.** Patients were assigned a severity using the following criteria:

309	Sample processing and sequencing. Samples were processed as described previously ¹¹ .
308	
307	from COVID-19 or MIS-C. These patients were almost always admitted to the ICU.
306	diabetic ketoacidosis, hemodynamic instability requiring vasopressors) and/or dying
305	failure requiring dialysis, coagulation abnormalities resulting in bleeding or stroke,
304	membrane oxygenation (ECMO) and/or evidence of end-organ failure (acute renal
303	pressure (BIPAP), intubation with mechanical ventilation, or extracorporeal
302	high-flow oxygen requirement (high-flow nasal cannula, bilevel positive airway
301	• Severe: The patient must have been hospitalized for COVID-19 or MIS-C with either
300	abnormalities, etc.) and/or MIS-C.
299	extremities, diarchea, encephalopathy, renal insufficiency, mild coagulation
298	neonatal fever, dehydration, new diagnosis diabetes, acute appendicitis, necrosis of
297	disease and/or any systemic/non-respiratory symptoms attributed to COVID-19 (e.g
296	<u>Moderate</u> : The patient must have been hospitalized due to COVID-19 respiratory
295	throat, cough, rhinorrhea, loss of sense of smell or taste from COVID-19 only.
294	COVID-19) or if hospitalized, only upper respiratory symptoms, including fever, sore
293	<u>Mild</u> : This included all outpatient cases (who did not require hospitalization for
292	hospitalized for another cause or not hospitalized.
291	nasopharyngeal RT- PCR but no symptoms of COVID-19, regardless of whether
290	<u>Asymptomatic</u> : This included patients with evidence of SARS-CoV-2 infection by

Briefly, samples were received on dry ice, RNA was extracted, libraries prepared, and

It is made available under a CC-BY-NC-ND 4.0 International license .

311 sequenced on a NextSeq or NovaSeq Illumina sequencer. Sequencing data was processed

using a custom bioinformatics pipeline which included quality filtering and trimming, alignment

to the human GRCh38 reference genome, and counting of gene features.

314

Gene expression analyses. Whole blood samples from three different categories (MIS-C, 315 severe COVID-19, and healthy controls) were extracted as previously described and eluted in 316 200 ul. RT-PCR was performed using 5 ul of TagMan Fast Advanced Master Mix (Applied 317 Biosystems 4369514), 1 ul of probe (predesigned TagMan Probes, Thermo Fisher), 6 ul of 318 nuclease-free water, and 8 ul of extracted material. All reactions were performed in a 319 QuantStudio Real-Time PCR (Thermo Fisher) following this thermal-cycling conditions: 320 incubation at 50°C for 2 minutes, enzyme activation 95°C for 20 seconds, and 40 cycles of 321 322 denature step at 95°C for 3 seconds and anneal/extend 60°C for 30 seconds. Results were analyzed using the QuantStudio software. 323

324

325 **Machine learning.** Machine learning and model training was done using R (v4.1.1) with packages Caret (v6.0.90), tidyverse (v1.3.1), pROC (v1.18.0), PRROC (v1.3.1), DESeq2 326 327 (v1.34.0), and data.table (v1.14.2). Sample metadata and count matrices were loaded as dataframes, and split 70/30 into a training set and a test set. Hospital of origin and severity of 328 disease were considered while splitting the data to minimize differences in the training and test 329 330 sets. Relevant features for model training were selected by filtering and differential expression/abundance analysis. First, genes with low counts (sum counts per million across 331 samples < 15) and near zero variance (R caret package nearZeroVar function) were removed. 332 333 Next, the top 300 genes were selected using differential expression/abundance analysis using

It is made available under a CC-BY-NC-ND 4.0 International license .

DESeq2¹⁴ as ranked by absolute log2 fold change (adjusted p.value < 0.05, basemean > 5). 334 Machine learning algorithms were trained using the subset meta data using 5-fold cross 335 validation and grid search hyperparameter tuning. Next, class predictions were predicted for 336 337 the training set. Accuracy, sensitivity, specificity, and ROC AUC were used to measure test performance. The classification models used are generalized linear models with Ridge and 338 LASSO feature selection (GLMNETRidge and GLMNETLasso), support vector machines with 339 linear and radial basis function kernels (SVMLin and SVMRAD), random forest (RF), random 340 forest ExtraTrees (EXTRATREES), neural networks (NNET), linear discriminant analysis 341 (LDA), nearest shrunken centroids (PAM), C5.0 (C5), k-nearest neighbors (KNN), naive bayes 342 (NB), CART (RPART), and logistic regression (GLM). 343 Anuscript

344

Acknowledgements. We would like to acknowledge staff members at the UCSF Clinical 345 Laboratories and the UCSF Clinical Microbiology Laboratories for their help in identifying and 346 retrieving patient whole blood samples. We thank the Cornell Genomics Center and the UCSF 347 348 Center for Advanced Technology for helping with sequencing libraries. At Emory, we thank Christopher Choi, Caroline Ciric, Khalel De Castro, Theda Gibson, Hui-Mien Hsiao, Wensheng 349 Li, Austin Lu, Lisa Macoy, Kathy Stephens, Madeline Taylor, Ashley Tippett and the Children's 350 351 Healthcare of Atlanta Research Laboratory for their contributions to specimen and data collection. We thank the patients and their families for contributing their blood to further our 352 understanding of pediatric COVID-19 and MIS-C. 353

- 354
- 355

Author Contributions 356

It is made available under a CC-BY-NC-ND 4.0 International license .

- 357 A.S.-G., C.J.L., V.S., R.D., C.A.R., I.D.V., and C.Y.C. conceived and designed the study.
- C.J.L., A.S.-G., V.S., J.L., A.P.C., and A.B. performed sequencing experiments. A.S.-G., J.N., 358
- J.L., M.E.W., P.S., N.B., J.S., W.S., C.H., N.W., A.F., A.G. K.B., M.A.P., L.H., E.J.A., A.C., 359
- 360 M.D., R.D., C.A.R., and C.Y.C. identified and collected patient samples and clinical metadata.
- C.J.L., V.S., B.B., M.D., R.D., C.A.R., I.D.V., and C.Y.C. analyzed sequencing data and built 361
- models. M.D., A.B., R.D., C.A.R., I.D.V., and C.Y.C. supervised the study. A.S.-G., C.J.L., 362
- V.S., S.B., I.D.V., and C.Y.C. wrote the manuscript and prepared the figures. All authors read 363
- and edited the manuscript and agree to its contents. 364
- 365

366

Conflicts of Interest I.D.V. is a member of the Scientific Advisory Board of Karius Inc., Kanvas Biosciences and 367 GenDX. C.Y.C. is a founder and a member of the Scientific Advisory Board of Delve Bio. I.D.V. 368 and A.P.C. are listed as an inventor on submitted patents pertaining to cell-free DNA (US 369 370 patent applications 63/237,367, 63/056,249, 63/015,095, 16/500,929) and receive consulting 371 fees from Eurofins Viracor. C.A.R. received funding to conduct clinical research unrelated to this manuscript from BioFire Inc., GSK, MedImmune, Micron, Merck, Novavax, PaxVax, 372 Regeneron, Pfizer, and Sanofi-Pasteur. She is co-inventor of patented RSV vaccine 373 374 technology (International PCT Application No. PCT/US2016/058976, filed 12/28/2016 by Emory University), which has been licensed to Meissa Vaccines, Inc. with royalties received. 375 376 Her institution has received funding from NIH to conduct clinical trials of Moderna and Janssen COVID-19 vaccines. E.J.A has consulted for Pfizer, Sanofi Pasteur, GSK, Janssen, and 377 Medscape, and his institution receives funds to conduct clinical research unrelated to this 378 379 manuscript from MedImmune, Regeneron, PaxVax, Pfizer, GSK, Merck, Sanofi-Pasteur,

It is made available under a CC-BY-NC-ND 4.0 International license .

380 Janssen, and Micron. He also serves on a safety monitoring board for Kentucky BioProcessing, Inc. and Sanofi Pasteur. He serves on a data adjudication board for WCG and 381 ACI Clinical. His institution has also received funding from NIH to conduct clinical trials of 382 383 Moderna and Janssen COVID-19 vaccines. A.B. is a co-founder and consultant to Personalis and NuMedii; consultant to Mango Tree Corporation, and in the recent past, Samsung, 10x 384 Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory 385 panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, 386 AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in 387 Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google), 388 Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty 389 Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, 390 391 Nuna Health, Assay Depot, and Vet24seven, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from 392 Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, 393 Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic 394 institutions, medical or disease specific foundations and associations, and health systems. 395 A.B. receives royalty payments through Stanford University, for several patents and other 396 disclosures licensed to NuMedii and Personalis. Research from A.B. has been funded by NIH, 397 Peraton (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert 398 399 Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent 400 past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor Office 401 402 of Planning and Research, California Institute for Regenerative Medicine, LOreal, and

It is made available under a CC-BY-NC-ND 4.0 International license .

- Progenity. The authors have declared that none of these companies or competing interests 403
- had any role in this work or manuscript. 404
- 405

406 Data and Code Availability

- All code will be made available on Github. Processed sequencing data will be deposited in the 407
- National Institutes of Health (NIH) and National Center for Biotechnology Information (NCBI) 408
- . nenotypes (dbGAP). nenotypes (dbGAP). A details see manuscript DOI for details Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) repositories under 409
- restricted access via Database for Genotypes and Phenotypes (dbGAP). 410
- 411
- 412

It is made available under a CC-BY-NC-ND 4.0 International license .

413 **REFERENCES**

- Abrams, J. Y. *et al.* Multisystem Inflammatory Syndrome in Children Associated with
 Severe Acute Respiratory Syndrome Coronavirus 2: A Systematic Review. *J Pediatr* **226**, 45-54 e41 (2020). <u>https://doi.org:10.1016/j.jpeds.2020.08.003</u>
- 417 2 Rey-Jurado, E. *et al.* Deep immunophenotyping reveals biomarkers of MIS-C in a Latin
 418 American cohort. *J Allergy Clin Immunol* (2022).
 419 https://doi.org:10.1016/j.jaci.2022.09.006
- Godfred-Cato, S. *et al.* Distinguishing Multisystem Inflammatory Syndrome in Children
 From COVID-19, Kawasaki Disease and Toxic Shock Syndrome. *Pediatr Infect Dis J* 422
 41, 315-323 (2022). https://doi.org:10.1097/INF.00000000003449
- 4 Kara, Y. *et al.* A Common Problem During the Pandemic Period; Multisystem
 Inflammatory Syndrome in Children or Gram-negative Sepsis? *Pediatr Infect Dis J* 41,
 e29-e30 (2022). <u>https://doi.org:10.1097/INF.0000000003345</u>
- Zambrano, L. D. *et al.* Investigating Health Disparities Associated With Multisystem
 Inflammatory Syndrome in Children After SARS-CoV-2 Infection. *Pediatr Infect Dis J* (2022). https://doi.org:10.1097/INF.00000000003689
- 6 CDC. Health Department-Reported Cases of Multisystem Inflammatory Syndrome in
 Children (MIS-C) in the United States, <<u>https://covid.cdc.gov/covid-data-tracker/#mis-</u>
 national-surveillance> (2022).
- Montgomery, S. B., Bernstein, J. A. & Wheeler, M. T. Toward transcriptomics as a
 primary tool for rare disease investigation. *Cold Spring Harb Mol Case Stud* 8 (2022).
 https://doi.org:10.1101/mcs.a006198
- 8 Servellita, V. *et al.* A diagnostic classifier for gene expression-based identification of
 early Lyme disease. *Commun Med (Lond)* 2, 92 (2022). <u>https://doi.org:10.1038/s43856-</u>
 022-00127-2
- 438 9 Zaas, A. K. *et al.* Gene expression signatures diagnose influenza and other
 439 symptomatic respiratory viral infections in humans. *Cell Host Microbe* 6, 207-217
 440 (2009). <u>https://doi.org:10.1016/j.chom.2009.07.006</u>
- Ng, D. L. *et al.* A diagnostic host response biosignature for COVID-19 from RNA
 profiling of nasal swabs and blood. *Sci Adv* 7 (2021).
 https://doi.org:10.1126/sciadv.abe5984
- Loy, C. J. *et al.* Nucleic acid biomarkers of immune response and cell and tissue
 damage in children with COVID-19 and MIS-C. *medRxiv*, 2022.2006.2021.22276250
 (2022). <u>https://doi.org:10.1101/2022.06.21.22276250</u>
- Beckmann, N. D. *et al.* Downregulation of exhausted cytotoxic T cells in gene
 expression networks of multisystem inflammatory syndrome in children. *Nat Commun*449 **12**, 4854 (2021). <u>https://doi.org:10.1038/s41467-021-24981-1</u>
- Butler, D. *et al.* Shotgun transcriptome, spatial omics, and isothermal profiling of SARS CoV-2 infection reveals unique host responses, viral diversification, and drug
 interactions. *Nat Commun* 12, 1660 (2021). https://doi.org:10.1038/s41467-021-21361-7

- Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion
 for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014).
 <u>https://doi.org:10.1186/s13059-014-0550-8</u>
- Feldstein, L. R. *et al.* Characteristics and Outcomes of US Children and Adolescents
 With Multisystem Inflammatory Syndrome in Children (MIS-C) Compared With Severe
 Acute COVID-19. *JAMA* 325, 1074-1087 (2021).
 https://doi.org:10.1001/jama.2021.2091
- Belhadjer, Z. *et al.* Acute Heart Failure in Multisystem Inflammatory Syndrome in
 Children in the Context of Global SARS-CoV-2 Pandemic. *Circulation* 142, 429-436
 (2020). https://doi.org:10.1161/CIRCULATIONAHA.120.048360
- Coleman, J. J., Manavi, K., Marson, E. J., Botkai, A. H. & Sapey, E. COVID-19: to be or not to be; that is the diagnostic question. *Postgrad Med J* 96, 392-398 (2020).
 <u>https://doi.org:10.1136/postgradmedj-2020-137979</u>
- Fistera, D. *et al.* What about the others: differential diagnosis of COVID-19 in a German emergency department. *BMC Infect Dis* 21, 969 (2021). <u>https://doi.org:10.1186/s12879-021-06663-x</u>
- Stoyanov, G., Dzhenkov, D. & Petkova, L. Non-COVID-19 viral respiratory tract infection
 as causes of death amid the pandemic: a report of two autopsy cases and tips for safe
 practice. *Folia Med (Plovdiv)* 63, 608-612 (2021).
 https://doi.org:10.3897/tolmed.63.e56037
- Dailey, P. J., Elbeik, T. & Holodniy, M. Companion and complementary diagnostics for
 infectious diseases. *Expert Rev Mol Diagn* 20, 619-636 (2020).
 https://doi.org:10.1080/14737159.2020.1724784
- 476 21 Mullen, J. L. *et al. outbreak.info*, <<u>https://outbreak.info/</u>> (2020).
- 477 22 Arbel, R. *et al.* Nirmatrelvir Use and Severe Covid-19 Outcomes during the Omicron
 478 Surge. *N Engl J Med* 387, 790-798 (2022). <u>https://doi.org:10.1056/NEJMoa2204919</u>
- 479 23 Coenye, T. Do results obtained with RNA-sequencing require independent verification?
 480 *Biofilm* 3, 100043 (2021). <u>https://doi.org:10.1016/j.bioflm.2021.100043</u>
- Chang, L. J. *et al.* Accuracy and comparison of two rapid multiplex PCR tests for
 gastroenteritis pathogens: a systematic review and meta-analysis. *BMJ Open Gastroenterol* 8 (2021). <u>https://doi.org:10.1136/bmjgast-2020-000553</u>
- Popowitch, E. B., Kaplan, S., Wu, Z., Tang, Y. W. & Miller, M. B. Comparative
 Performance of the Luminex NxTAG Respiratory Pathogen Panel, GenMark eSensor
 Respiratory Viral Panel, and BioFire FilmArray Respiratory Panel. *Microbiol Spectr* 10,
 e0124822 (2022). <u>https://doi.org:10.1128/spectrum.01248-22</u>
- Tansarli, G. S. & Chapin, K. C. Diagnostic test accuracy of the BioFire(R) FilmArray(R)
 meningitis/encephalitis panel: a systematic review and meta-analysis. *Clin Microbiol Infect* 26, 281-290 (2020). <u>https://doi.org:10.1016/j.cmi.2019.11.016</u>