Title

Host methylation predicts SARS-CoV-2 infection and clinical outcome.

Journal

Communications Medicine, 1(1)

ISSN

2730-664X

Authors

Konigsberg, Iain R
Barnes, Bret
Campbell, Monica
et al.

Publication Date

2021

DOI

10.1038/s43856-021-00042-y

Peer reviewed

communications medicine

ARTICLE

https://doi.org/10.1038/s43856-021-00042-y OPEN

Host methylation predicts SARS-CoV-2 infection and clinical outcome

Iain R. Konigsberg[1,7], Bret Barnes[2,7], Monica Campbell[1], Elizabeth Davidson[1], Yingfei Zhen[1], Olivia Pallisard[1], Meher Preethi Boorgula[1], Corey Cox[1], Debmalya Nandy[3], Souvik Seal[3], Kristy Crooks[1], Evan Sticca[1], Genelle F. Harrison[1], Andrew Hopkinson[1], Alexis Vest[1], Cosby G. Arnold[1], Michael G. Kahn[1], David P. Kao[1], Brett R. Peterson[1], Stephen J. Wicks[1], Debashis Ghosh[3], Steve Horvath[4], Wanding Zhou[5], Rasika A. Mathias[1,6], Paul J. Norman[1], Rishi Porecha[2], Ivana V. Yang[1,8], Christopher R. Gignoux[1,8], Andrew A. Monte[1,8], Alem Taye[2,8] & Kathleen C. Barnes[1,8]✉

Abstract

Background Since the onset of the SARS-CoV-2 pandemic, most clinical testing has focused on RT-PCR[1]. Host epigenome manipulation post coronavirus infection[2–4] suggests that DNA methylation signatures may differentiate patients with SARS-CoV-2 infection from uninfected individuals, and help predict COVID-19 disease severity, even at initial presentation.

Methods We customized Illumina's Infinium MethylationEPIC array to enhance immune response detection and profiled peripheral blood samples from 164 COVID-19 patients with longitudinal measurements of disease severity and 296 patient controls.

Results Epigenome-wide association analysis revealed 13,033 genome-wide significant methylation sites for case-vs-control status. Genes and pathways involved in interferon signaling and viral response were significantly enriched among differentially methylated sites. We observe highly significant associations at genes previously reported in genetic association studies (*e.g. IRF7*, *OAS1*). Using machine learning techniques, models built using sparse regression yielded highly predictive findings: cross-validated best fit AUC was 93.6% for case-vs-control status, and 79.1%, 80.8%, and 84.4% for hospitalization, ICU admission, and progression to death, respectively.

Conclusions In summary, the strong COVID-19-specific epigenetic signature in peripheral blood driven by key immune-related pathways related to infection status, disease severity, and clinical deterioration provides insights useful for diagnosis and prognosis of patients with viral infections.

Plain language summary

Viral infections affect the body in many ways, including via changes to the epigenome, the sum of chemical modifications to an individual's collection of genes that affect gene activity. Here, we analyzed the epigenome in blood samples from people with and without COVID-19 to determine whether we could find changes consistent with SARS-CoV-2 infection. Using a combination of statistical and machine learning techniques, we identify markers of SARS-CoV-2 infection as well as of severity and progression of COVID-19 disease. These signals of disease progression were present from the initial blood draw when first walking into the hospital. Together, these approaches demonstrate the potential of measuring the epigenome for monitoring SARS-CoV-2 status and severity.

[1] School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [2] Illumina, Inc., San Diego, CA, USA. [3] Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [4] University of California Los Angeles, Los Angeles, CA, USA. [5] The Children's Hospital of Philadelphia, Philadelphia, PA, USA. [6] Johns Hopkins University, Baltimore, MD, USA. [7] These authors contributed equally: Iain R. Konigsberg, Bret Barnes. [8] These authors jointly supervised this work: Ivana V. Yang, Christopher R. Gignoux, Andrew A. Monte, Alem Taye, Kathleen C. Barnes. ✉email: kathleen.barnes@cuanschutz.edu

Coronaviruses (CoV) comprise a large group of human and animal pathogens, including the novel enveloped RNA betacoronavirus referred to as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[5]. This pathogen is associated with coronavirus disease 2019 (COVID-19) first identified in Wuhan, China in 2019[6] and declared a pandemic on March 11, 2020[7]. Since the onset of the pandemic, multiple tests for diagnosing COVID-19 have been launched, including real-time reverse transcriptase–polymerase chain reaction (RT-PCR), specific antibody detection, and next-generation sequencing assays that query for current or past infections[1]. With the exception of next-generation sequencing, which can discern viral subtypes, most diagnostic tests are viral strain dependent, can carry a high false negative rate, do not discern if the virus is viable and replicating, and do not predict clinical outcomes of infection[1,8,9]. For example, pre-symptomatic patients may test negative[10,11] while patients who have recovered may continue to test positive though they are no longer infectious[12]. Accurate diagnostics are urgently required to control continued communal spread, to better understand host response, and for the development of vaccines and antivirals[13].

Individuals infected with SARS-CoV-2 have a variable course of infection, ranging from asymptomatic to death. Although the fatality rate varies tremendously according to demographic characteristics and co-morbidities[14], the U.S. ranks as one of the countries with the highest COVID-19 mortality rates[15]. Identification of which SARS-CoV-2-infected patients are most likely to develop severe disease would enable clinicians to triage patients via augmented clinical decision support. Having more information on disease severity has recently become critical due to widespread lack of hospital and intensive care unit (ICU) capacity, necessitating difficult decisions about resource triage. To our knowledge, no test can predict COVID-19 clinical course or severity, although work on cytokine abundance ratios after hospitalization has been proposed as a prognostic indicator of severe outcomes[16].

There is considerable evidence that enveloped RNA viruses such as CoV can manipulate the host's epigenome via evolved functions that antagonize and regulate the host innate immune antiviral defense processes[2,3], specifically via DNA methylation. Viral-mediated antagonism of antigen-presentation gene expression in the case of Middle East respiratory syndrome coronavirus (MERS-CoV) was shown to occur via DNA methylation[4]. DNA methylation changes at cytosine-phosphate-guanine (CpG) sites have been increasingly leveraged in the emerging field of clinical epigenetics to characterize unique epigenetic signatures that diagnose disease. To date, considerable success has been demonstrated in developing highly accurate and robust machine learning (ML)-based disease classifiers using DNA methylation patterns to differentiate Mendelian disorders[17], behavior disorders[18], coronary artery disease[19], and some cancers[20–22]. Consequently integration of a methylation-based disease classification can result in relevant improvement in clinical practice[23,24].

With a goal to leverage Illumina's Infinium MethylationEPIC Array to classify differential methylation signatures of SARS-CoV-2-positive (hereafter referred to as SARS-CoV-2+, regardless of additional symptoms) and control peripheral blood DNA samples (either confirmed SARS-CoV-2 negative or samples collected prior to the SARS-CoV-2 pandemic), we conducted this study to determine whether DNA methylation patterns could differentiate SARS-CoV-2-infected patients from non-infected patients from whole blood obtained from patients. Our secondary objective was to determine whether DNA methylation patterns could differentiate patients with SARS-CoV-2 infection who go on to develop severe disease. In this study, we identified a strong COVID-19-specific epigenetic signature in peripheral blood driven by key immune-related pathways related to SARS-CoV-2 infection status, disease severity, and clinical deterioration.

Methods

Source of data. This protocol was reviewed and approved by the Colorado Multiple Institutional Review Board (COMIRB) and the research adheres to the ethical principles of research outlined in the U.S. Federal Policy for the Protection of Human Subjects. SARS-CoV-2+ were defined as those patients who tested positive for SARS-CoV-2 infection via a routine diagnostic RT-PCR assay in the Biobank at the Colorado Center for Personalized Medicine (Thermo Fisher Scientific, Waltham, MA) or in the UCHealth University of Colorado Hospital Clinical Laboratory (Roche Diagnostics, Indianapolis, IN) of a nasopharyngeal swab collected in viral transport media; controls were defined as those who tested negative. Peripheral blood DNA samples were collected in EDTA tubes from patients seen at the UCHealth University of Colorado Hospital and tested for SARS-CoV-2 epigenetic signatures starting on March 1, 2020. Blood specimens were collected from patients consented to the University of Colorado COVID-19 Biorepository (https://research.cuanschutz.edu/university-research/covid-19-clinical-research/covid-19-biobank-specimen-repository) or the University of Colorado Emergency Medicine Specimen Bank (EMSB)[25]. Control subjects included patients from each study who tested negative for SARS-CoV-2 infection during the index visit. Through the University of Colorado COVID-19 Biorepository and the EMSB, patients tested were consented for blood collection and data abstraction from their electronic health record (EHR). Data obtained from EHR abstraction included demographics, past medical history, laboratory testing (including SARS-CoV-2), treatments, vital signs, hospital disposition, and clinical outcomes. In addition, previously collected samples from patients with acute upper respiratory viral infections (SARS-CoV-2 negative/pan-negative for upper respiratory viral infections/positive for non-SARS-CoV-2 upper respiratory viral infections) between February 5, 2018 and January 1, 2020 were obtained through the EMSB as SARS-CoV-2-negative controls. Additional biospecimens included discarded clinical samples from patients not approached for biorepository enrollment through the UCHealth University of Colorado Hospital Clinical Laboratory. Discarded samples were linked to a limited EHR dataset through the Colorado Center for Personalized Medicine's health data warehouse, Health Data Compass, and then deidentified. The limited dataset included age, gender, race, ethnicity, viral test status (SARS-CoV-2 and other upper respiratory viruses), and clinical outcomes. The use of discarded samples and accompanying limited datasets was determined to be exempt from Institutional Review Board approval and the need for informed consent by COMIRB. All samples were frozen at −20 °C after collection prior to processing for methylation analyses.

Customization of the Infinium MethylationEPIC Array. Following a literature review of known epigenetic associations with respiratory viral infections from recent CoV outbreaks, we selected additional content to enrich Illumina's Infinium MethylationEPIC Array[26]. We specifically enriched for known HLA alleles accounting for known genomic variation[27] as well as multiple alternative haplotypes and unpublished reference sequences spanning the major histocompatibility complex genomic region, the natural killer cell immunoreceptor, and other immunogenetic loci (e.g., cytokines, interferon response genes), to enhance the sensitivity of immune response detection. The custom panel targeted 262 genes with 7831 additional probes. While the majority of the additional probes targeted unique sequences

within the genome, a number of probes were intentionally designed to target genomic sequences with a limited degree of repetitiveness. The list of genes and the Illumina IDs for the probes that target these genes are given in Supplementary Data 1.

Methylation array and quality assessment

DNA extraction. Biospecimens were accessioned and tracked via the Colorado Anschutz Research Genetics Organization (CARGO) laboratory information management system (LIMS). Genomic DNA was extracted from SARS-CoV-2+ peripheral blood on the bead-based, automated extraction Maxwell(R) RSC System (Promega) in a biological safety cabinet in compliance with CDC safety guidelines and procedures for handling SARS-CoV-2 biospecimens (biospecimens from SARS-CoV-2+ cases) and from controls on the Autogen FlexSTAR+ using the Autogen's FlexiGene Blood Extraction Kit (Holliston, MA). All DNA samples were quantified using both absorbance (NanoDrop 2000; Thermo Fisher Scientific, Waltham, MA) and fluorescence-based methods (Qubit; Thermo Fisher Scientific, Waltham, MA) using standard dyes selective for double-stranded DNA, minimizing the effects of contaminants that affect the quantitation. DNA quality was assessed using an Agilent TapeStation (Agilent, Santa Clara, CA). Samples were then uploaded to CARGO's LIMS, barcoded, and labeled.

Bisulfite conversion and amplification. Purified DNA samples were processed using the Zymo EZ-96 DNA Methylation bisulfite conversion kits (Zymo, Irvine, CA) as described previously[28]. The product of this process contains cytosine converted to uracil if it was previously unmethylated. The bisulfite-treated DNA was subjected to whole-genome amplification via random hexamer priming and Phi29 DNA polymerase, and the amplification products were then enzymatically fragmented, purified from dNTPs, primers, and enzymes, and applied to the Illumina chip as described elsewhere[29].

Hybridization and single-base extension. The bisulfite-converted amplified DNA products were denatured into single strands and hybridized to the customized Infinium 850K BeadChip (EPIC+; Illumina Inc., San Diego, CA) via allele-specific annealing to either the methylation-specific probe or the non-methylation probe. Hybridization to the chip was followed by single-base extension with labeled di-deoxynucleotides according to Illumina's Infinium protocol at the CARGO laboratory[28].

Fluorescence staining and scanning of chip. The hybridized BeadChips were stained, washed, and scanned to show the intensities of the un-methylated and methylated bead types using Illumina's iScan System.

Data processing and quality control (QC). IDAT files were processed, filtered, and normalized using the SeSAMe R package[30]. Type I probe channel was empirically determined from signal intensities. Probe detection *P* values (representing the ability to differentiate true signal from background fluorescence) were calculated for each color channel using pOOBAH, which leverages the fluorescence of out-of-band (OOB) probes. Normalization was performed using noob, which uses OOB probes to perform a normal-exponential deconvolution of fluorescent intensities[31]. Finally, a common dye bias that results in greater intensities in the red color channel was corrected to ensure that the distribution of intensities in the two color channels were equal. Probes with detection *P* values >0.05 were removed, as well as probes overlapping single-nucleotide polymorphisms with global minor allele frequency >1% in dbSNP, probes with poor mapping, and probes containing non-unique sequence according to Zhou et al.[32]. Beta values were logit-transformed into *M* values for modeling. Probes with >25% missingness were removed. Remaining missing values were then imputed with mean probe *M* value.

Selection of discovery/training and testing cohorts and controls. Case–control analyses were performed using the entire genotyped dataset passing epigenetics QC, with SARS-CoV-2 infection status determined as described above (see Fig. 1 for a summary of the workflow). Analyses were repeated including and excluding controls with other upper respiratory infections validated by clinical respiratory panels. Measurements of disease severity and progression (e.g., hospitalization, ICU admittance, ventilator use) were extracted from chart review within the UCHealth EHR.

Control for batch effect and robustness of the identified epigenetic signatures. To minimize possible batch effects and other sources of variability, samples were split into SARS-CoV-2+ and SARS-CoV-2-negative control sets, randomized within sets to account for unavailable phenotypes, and then distributed across chips. To reduce batch and plating effects a minimum of two SARS-CoV-2+ and two SARS-CoV-2-negative control samples were run on each chip (12 chips per plate, 8 samples each) and positive/negative status was randomized across the chip.

Epigenome-wide association study (EWAS) with COVID-19 disease status. Preprocessing was performed using the GLINT[33] package for association testing and estimating components to adjust for population structure (EPISTRUCTURE[34]) and we used ReFACTor[35] to account for cell-type proportions. We chose ReFACTor to account for cell proportion information in a data-driven fashion. The linear mixed-effects model in GLINT was fit to each probe, testing for differences based on COVID-19 disease status while correcting for age, sex, chip position, 6 ReFACTor components, 1 EPISTRUCTURE component, and a variance component representing individual covariance[36]. Enrichment of top hits in common databases was performed using enrichR[37]. Probes were sorted by adjusted *P* value and the top 800 genes to which differentially methylated probes map were used as input to perform overrepresentation enrichment analysis within Gene Ontology (GO) categories, Kyoto Encyclopedia of Genes and Genomes pathways (KEGG), BioPlanet, and WikiPathways[38–41]. Probes were annotated to CpG island and genic regions using annotatr[42].

Clinical outcome stratification. Clinical data were abstracted via detailed chart review for all EMSB patients. COVID-19 disease severity was determined by an ordered severity score of (1) discharged from emergency department; (2) admitted to inpatient care; (3) progressed to ICU; and (4) death. We also determined a hospital duration variable, where individuals without a measured hospital stay (i.e., discharged from the emergency department) were assigned 0 and individuals who died were removed from the cohort for length of stay analysis to minimize bias associated with timing of decisions to withdraw care.

Construction and validation of a prediction model. Predictive modeling was performed using the Lasso[43] and Elastic Net[44] algorithms for sparse penalized regression modeling available in the *glmnet* software package[45]. For each prediction model, only autosomal methylation probes passing QC were included, to remove potential confounding from sex-linked chromosomes. No demographic, clinical, or cell count variables were included in the

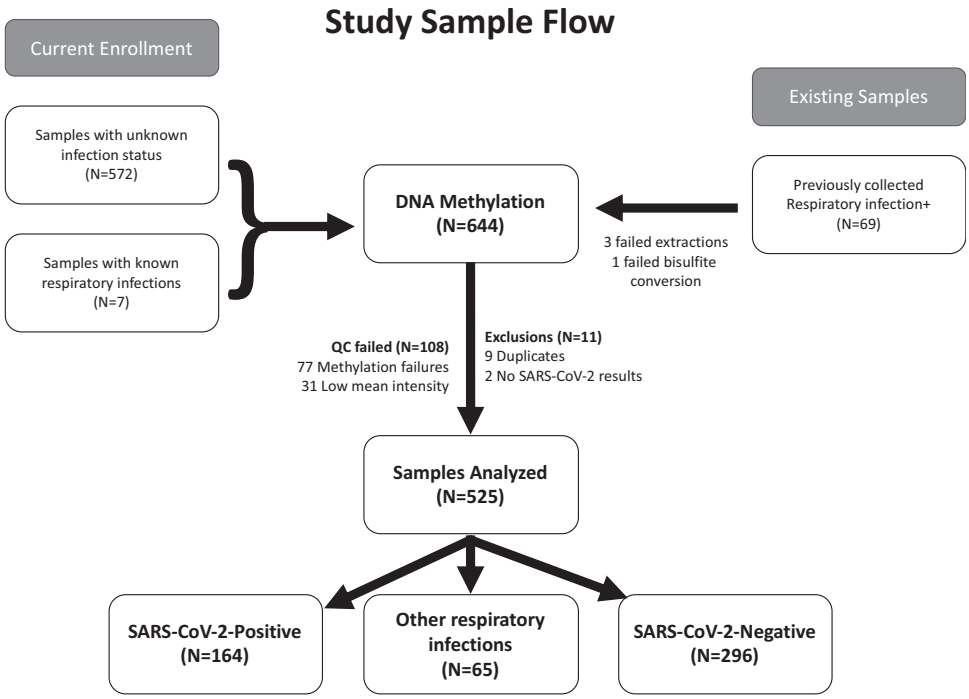

Fig. 1 Flowchart of the study sample collection. Six hundred and forty-eight samples were collected for analysis, of which 644 were processed on MethylationEPIC arrays. Five hundred and twenty-five arrays passed quality control and were included in the final analysis.

predictive models, requiring the algorithm to pick CpG sites with strong enough associations to surpass the level of penalization of the hyperparameters across the entire least angle regression path. For each trait of interest, a separate model was created and best-fitting parameters were chosen after tenfold cross-validation either by maximizing area under the receiver-operator characteristic curve (AUC for dichotomous traits) or minimizing mean-squared error (MSE for quantitative traits). Each was fit across a grid of parameters representing various strengths of penalization and combination of L1 and L2 penalties under the weighted elastic net model. Both the days of hospitalization and case severity were modeled as continuous outcomes. To assess performance for quantitative traits in a manner comparable to dichotomous traits, we swept across potential cutpoints to estimate AUCs for this newly derived dichotomous variable. While case–control status was the primary phenotype of interest, measures of severity were assessed in SARS-CoV-2+ cases only.

To estimate stability of estimation in parameters, we performed 100 iterations of model training and testing. Within each iteration for case–control and severity outcomes, we employed tenfold cross-validation to derive the model and a held-out set of 30% removed from train/test to gauge out-of-sample performance of the best-fitting model. Our train/test and validation splits were created within each stratum to preserve representation across all outcomes and reflect the distribution across the total dataset. For hospitalization duration, the train/test/validation models had instability in convergence and so we reverted to a train/test model using the tenfold cross-validation within the default cv.glmnet() function. We assessed overall performance for the dichotomous COVID+/COVID− case–control status using out-of-sample AUC, the F1 score (a measure of the relationship between precision and recall), the distribution of best-fit λ penalty via cross-validation, and the number of probes chosen in the final model. For the quantitative outcomes, we assessed overall performance using out-of-sample R^2, the slope of the model, and λ number of probes. Finally, these were stratified each across the elastic net weights (α) from 0.01 to 1, representing the proportion of ridge (L2) vs Lasso (L1) penalty to choose a final model. All models included nonzero λ to encourage sparsity (a L2-only model would include prediction from the entire array). Final models described in results were chosen based on best-performing (maximum R^2 or AUC) vs median values for each chosen set of hyperparameters. The final, out-of-sample best-fit prediction for each outcome was considered the "methylation score" used in downstream modeling, characterization of association, and determination of potential confounding with demographic and blood cell proportion characteristics.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Results

Study cohort. We identified 675 patients tested for either SARS-CoV-2 or other acute upper respiratory infections. Of these, 164 were SARS-CoV-2+ by RT-PCR, 58 historical EMSB patients had positive (non-SARS-CoV-2) acute upper respiratory viral RT-PCR tests, 7 had positive (non-SARS-CoV-2) acute upper respiratory viral RT-PCR tests during the pandemic, and 296 were negative for all viral infections and thus served as controls. We excluded 32 samples from the dataset as these were derived from a run with failed hybridization and removed 8 duplicates, resulting in a final cohort of 525 (Fig. 1). Supplementary Table 1 summarizes the demographics and clinical outcomes of patients tested, including proportion of patients with other acute upper respiratory infections. Incidences of non-SARS-CoV-2 respiratory infections are displayed in Supplementary Table 2. The median time from sample collection to hospital admission was 0 days (interquartile range (IQR): 0, 1). In all, 83.4% of samples were collected on the day of admission and only 8.7% were collected >5 days after hospital admission. Samples from patients

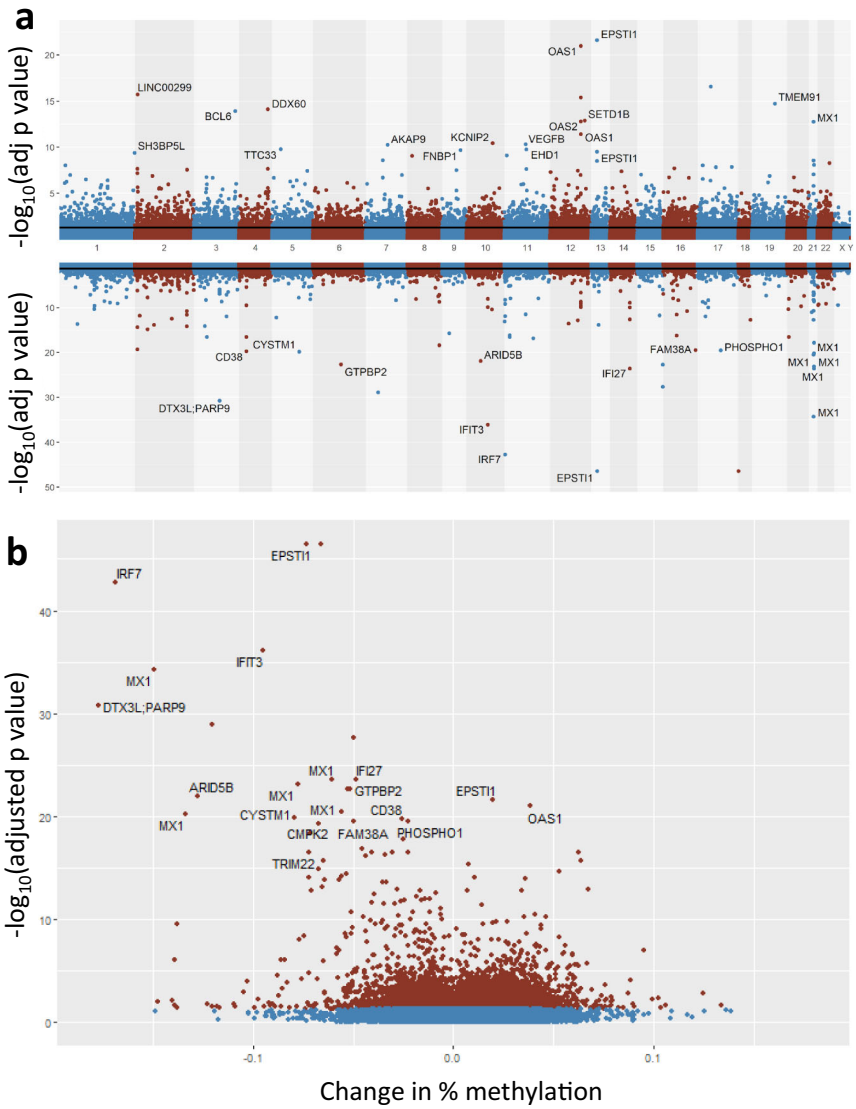

Fig. 2 Differentially methylated CpGs associated with SARS-CoV-2 infection. a Miami plot (top panel) of hypermethylated (top) and hypomethylated (bottom) probes in SARS-CoV-2+ compared to control samples. Significance lines represent FDR-adjusted *P* value <0.05 threshold. **b** Volcano plot of significant (red; FDR-adjusted *P* value <0.05) CpG sites (blue CpG sites have FDR-adjusted *P* value >0.05). Change in percentage methylation on the *x* axis represents the difference in average beta value at a site between cases and controls. Probes for intergenic CpG sites do not have gene annotations. Data used to plot this figure are available as Supplementary Data 5.

who were SARS-CoV-2 positive were drawn with the first blood sample in the emergency department 83% (median blood draw: 0 days, IQR: 0, 1 days) of the time; other samples drawn later in the hospital admission in this group were from patients who developed COVID while admitted to the hospital. Samples from two SARS-CoV-2-positive patients were obtained 6 and 9 days prior to hospital admission. Samples from SARS-CoV-2-negative patients were drawn with the first blood sample in the emergency department 80% (median blood draw: 0 days, IQR: 0, 2 days) of the time and 95% were drawn within 7 days of hospital admission. No samples were obtained on days before hospital admission in the SARS-CoV-2-negative patients.

Disease-specific DNA methylation signature and differentially methylated probes. We first performed an EWAS to identify biological signals associated with COVID-19 disease status. After adjustment for age, sex, array position (batch effect), cell proportions via ReFACTor and ancestry via EPISTRUCTURE components, EWAS of COVID-19 disease status in 164 SARS-CoV-2+ compared to 296 controls yielded 13,033 significant CpGs mapping to 6117 unique genes at false discovery rate (FDR)-adjusted *P* value < 0.05 (Fig. 2 and Supplementary Data 2), with moderate inflation that is typical of EWAS[46] (Supplementary Fig. 1). In total, we observed 35 probes with an unadjusted *P* value $< 10^{-20}$, and 183 with an unadjusted *P* value $< 10^{-10}$. Significant probes overlap 1625 CpG islands and 1001 FANTOM5[47] enhancers (Supplementary Fig. 2). We observed that 52.1% of all significant probes are hypermethylated; however, 78% of the top 100 probes sorted by adjusted *P* value are hypomethylated (Fisher's Exact Test *P* value $= 9.46 \times 10^{-8}$). Custom probes on the EPIC+ chip are enriched in significant EWAS results (*P* value $= 9.94 \times 10^{-7}$, Fisher's Exact Test): specifically, 1.72% of EPIC probes are significant as opposed to 2.51% of custom probes. Principal component analysis of top associations reveals clustering by COVID-19 disease status (Supplementary Fig. 3). Because of concerns that population admixture may confound results, the COVID-19 disease status EWAS was repeated with EHR-defined race and ethnicity as

additional covariates beyond that modeled via EPISTRUCTURE and mixed-effects modeling. This had a minimal effect on results.

Top hypomethylated CpG sites show strong enrichment for interferon and viral response-related pathways including Type I Interferon Signaling Pathway (KEGG, adjusted P value = 7.40×10^{-10}) and Negative Regulation of Viral Genome Replication (GO:BP, adjusted P value = 1.93×10^{-6}; Supplementary Fig. 4a). Hypermethylated CpG sites also show enrichment for relevant biological processes such as Focal Adhesion (GO:CC, adjusted P value = 0.0187; Supplementary Fig. 4b). cg17114584, the third most significant probe with an adjusted P value of 1.78×10^{-43}, shows 16.9% hypomethylation in cases. This CpG is located in exon 6 of the interferon regulatory factor 7 (*IRF7*). *IRF7* encodes a transcription factor that regulates the expression of interferon a and b, as well as interferon-stimulated genes. Other top CpGs are in genes relevant to viral response: *OAS1* (2'-5'-oligoadenylate synthetase 1) is interferon-induced and activates RNase L, which degrades viral (and cellular) RNA (adjusted P value 1.05×10^{-21}, 3.8% methylation change). *MX1* encodes an interferon-induced GTPase that inhibits viral replication. *DTX3L* and *PARP9* form a complex that is involved in interferon-mediated antiviral defenses. This complex has also been shown to promote M1 polarization in macrophages by preventing STAT1 phosphorylation[48]. *IFIT3* encodes another interferon-induced antiviral protein. Overall, we observe strong hypomethylation of interferon- and viral response-related pathways, which is expected as these pathways are activated transcriptionally in SARS-CoV-2+ individuals[49].

Specificity of the COVID-19 disease signature from other respiratory infections. We next compared 164 SARS-CoV-2+ samples to 65 samples with other upper respiratory infections to determine the specificity of the methylation signature to SARS-CoV-2. This analysis yielded 1501 significant CpGs (adjusted P value < 0.05) (Supplementary Data 3), of which 780 (52%) were present in the SARS-CoV-2+ compared to controls analysis (Fig. 3). Comparison of 65 other (non-SARS-CoV-2) upper respiratory infection samples to controls yielded 516 significant CpGs (Supplementary Data 4), of which 116 (22%) were present in the SARS-CoV-2+ compared to controls analysis. Furthermore, examination of the strength of the signal demonstrates that the shared probes in the SARS-CoV-2+ vs control and SARS-CoV-2+ vs other upper respiratory infections analysis have low P values and high effect sizes, whereas this is not the case for probes shared by SARS-CoV-2+ vs control and other upper respiratory infections vs control analyses (Supplementary Fig. 5a). These comparisons suggest high specificity of the COVID-19 disease epigenetic signature. To further investigate this, we examined the significant CpGs from our COVID-19 disease signature compared to control EWAS. We observe the same trend of high correlation of effect sizes (methylation change) in SARS-CoV-2+ compared to control and SARS-CoV-2+ compared to other respiratory infections (Pearson $R = 0.87$; $P < 2.2 \times 10^{-16}$) and very low correlations of effect sizes in SARS-CoV-2+ compared to control and other upper respiratory infections compared to control analyses (Pearson $R = -0.027$; $P = 0.0022$) (Supplementary Fig. 5b). While we do not have sufficient power to examine the specific viruses (other CoV, influenza, etc.), these results strongly point to the specificity of our COVID-19 disease epigenetic signature to detect SARS-CoV-2 infection.

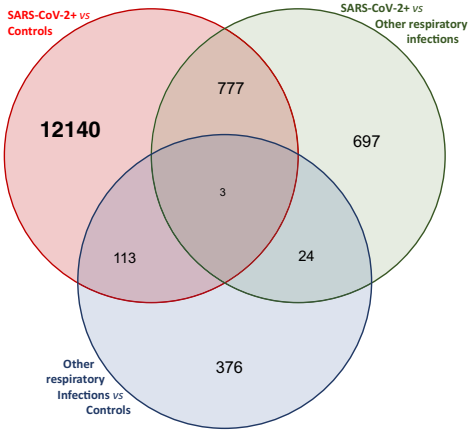

Fig. 3 Overlap of differentially methylated CpGs between disease groups. Venn diagram of overlaps between SARS-CoV-2+–Control EWAS (13,033 significant probes), SARS-CoV-2+–other respiratory infection EWAS (516 significant probes), and other respiratory infection–Control EWAS (1501 significant probes).

Development and validation of a classification model for prediction of disease classes and disease severity. To combine methylation data across the genome into a single predictor, we employed ML models of sparse regression trained via cross-validated *glmnet*[45] as described in "Methods." To determine the sensitivity of our model, 460 subjects (SARS-CoV-2+ vs controls) from the testing cohort were supplied to the classification model, with prediction optimized after the approach defined in "Methods." Only methylation probes were used in feature selection. All models showed relative stability across iterations (Supplementary Fig. 6) and yielded sparse results. Details of each top model are available in Supplementary Table 3. The best-fitting model has a performance of 93.6% in cross-validation for detecting SARS-CoV-2 infection (Fig. 4a, b). Model performance was similar in females and males (93.7 and 93.5%, respectively). In addition, model performance on older individuals and younger individuals (median age = 56 years) was comparable: 94.4 and 92.8%, respectively. Similarly, race/ethnicity information was not significantly correlated with case–control score (all groups $P > 0.05$). When age and race/ethnicity categories were included in a multivariable model along with our prediction score, no additional covariates significantly predicted COVID-19 disease status (all other $P > 0.4$). Similarly, BMI was not associated ($P \sim 0.4$).

To determine the direct association of methylation with clinical outcomes, an additional logistic regression was performed for the subset of individuals with complete blood cell count (CBC) data (341 individuals total). The inclusion of additional blood cell count data did not impact the association between the methylation score and outcome (P value < 2×10^{-16} with or without adjustment), and in the larger CBC model (including total hematocrit, white blood cell count, platelets, neutrophils, lymphocytes, monocytes, eosinophils, and basophils), only hematocrit ($P \sim 0.05$) approached nominal significance. The inclusion of hematocrit moderately improved Akaike information criterion in logistic regression but with limited performance increase in multivariable modeling AUC (93.6 vs 94.1%).

Severity analysis focused on hospital length of stay (median duration: 6 days, IQR 3–11, max 53 days), as well as across the spectrum of severity (34 discharged from emergency room, 84 hospitalized, 35 admitted to ICU, and 11 deaths). The best-fitting model for hospital duration had a cutpoint at 20 days, yielding an AUC of 79.6% with 14 individuals with longer stays vs 135 with shorter stays (or 0 days in hospital) (Fig. 4c). Dichotomizing the best-fit severity measurements yields AUCs of 79.1, 80.8, and 84.4 for hospital admission vs discharge, floor hospital admission vs ICU, and survival vs death, respectively (Fig. 4d).

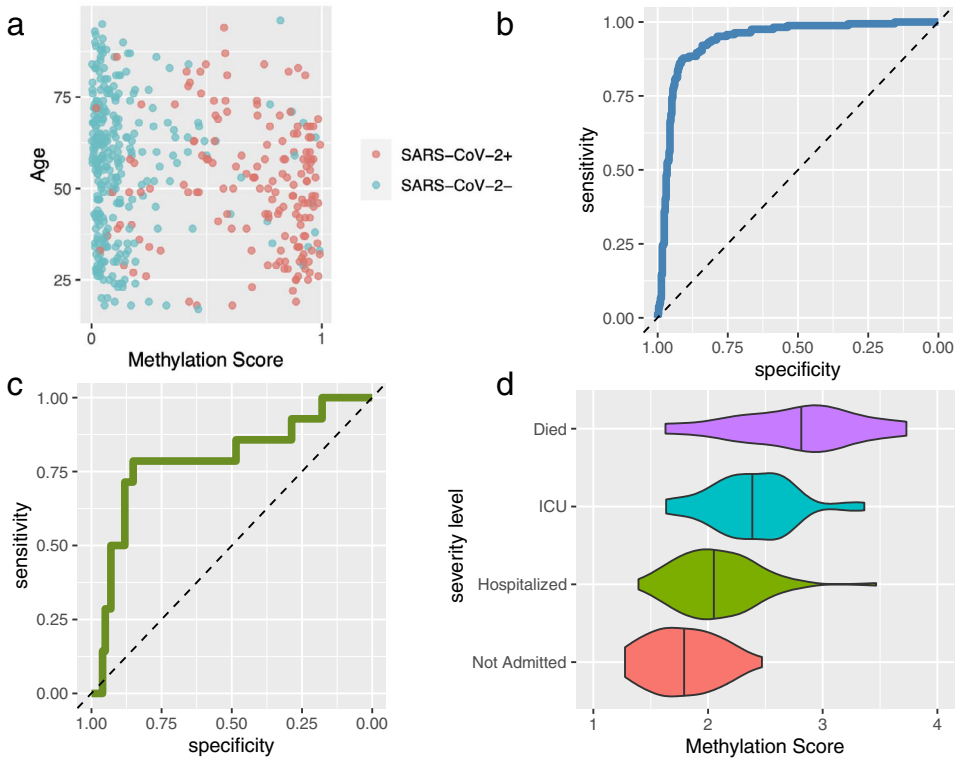

Fig. 4 Performance of SARS-CoV-2 infection status and severity predictive models. a Out-of-sample case–control methylation score for all 460 individuals (164 SARS-CoV-2+, 296 SARS-CoV-2−) compared to case–control status, plotted by biological age. **b** Receiver-operating characteristic (ROC) curve for data in **a**. **c** ROC curve of cross-validated prediction of long hospital duration. **d** Violin and jittered scatter plots of severity methylation scores for each outcome in cases. Data used to plot this figure are available as Supplementary Data 6.

Discussion

Here we report DNA methylation profiling in conjunction with analysis using ML techniques to identify a SARS-CoV-2-specific epigenetic signature in peripheral blood from a large cohort of individuals tested using conventional RT-PCR technology. We also describe the development of a classification algorithm that has high sensitivity and specificity in predicting infection and in-hospital clinical deterioration and that confidently rejects the probability of healthy individuals to be affected by SARS-CoV-2 infection. While any predictive signal invites concern of potential confounding, the methylation signature (derived solely from CpGs, not including any clinical or demographic information) we observe is not driven by confounding either from demographics or typical laboratory measurements (e.g., blood cell counts, BMI). Our findings suggest that measurement of methylation signals that arise during and after SARS-CoV-2 infection may provide clinicians the ability to detect viral infection as well as predict patient clinical course after viral challenge. Unlike sequencing, RT-PCR, and antibody tests, the methylation array is able to predict the severity of SARS-CoV-2 infection and ultimately could provide clinicians with information on how to manage patients infected with SARS-CoV-2.

Our results support the hypothesis that the host epigenome, as measured in peripheral blood, is modified by infection from SARS-CoV-2 and can be used to identify novel biology and it is useful for clinical diagnosis, prognosis, and triage. Despite being a heterogeneous tissue, we relied on peripheral blood as the target tissue because it has proven to be a reliable source for generating epigenetic signatures and disease classifiers in other settings[50–56]. In this study, we observed many methylation changes that are, on average, >10% differentially methylated in the SARS-CoV-2+ group, including *IRF7* and *MX1* interferon-related genes. These are much larger effect sizes than typically observed in EWAS in peripheral blood[57] and similar to the clinical utility of epigenetics observed in cancer[22]. We did not observe confounding by cell proportions, measured by CBC from the EHR, providing strong support for the epigenetic signature of SARS-CoV-2. Although cell-type heterogeneity can be a strong confounder in epigenetic studies[58–60], we did not pursue adjustment for cell proportions beyond adjustment for cell-type proportions using ReFACTor[35] because our primary objective is to develop a COVID-19 disease-specific diagnostic methylation platform, rather than interrogate the underlying pathology.

To validate the customized EPIC methylation platform as a reliable tool for the clinical diagnosis of COVID-19 disease, we performed an EWAS with SARS-CoV-2 infection status. We observed that the epigenetic signature of SARS-CoV-2 infection is enriched for pathways related to host viral response, and specifically for Type I Interferon signaling that is a hallmark of host response to this virus[61]. Our findings of altered DNA methylation in interferon response genes are in concordance with published results of changes in the expression of interferon response genes by SARS-CoV and MERS-CoV viruses through changes in histone modifications[2,3]. One of the most significant probes (adjusted $P = 1.77 \times 10^{-43}$, 16.9% hypomethylation) is located in the gene encoding *IRF7*; loss-of-function variants in 13 genes including *IRF7* were recently found to be associated with life-threatening COVID-19-associated pneumonia[62]. Another interferon-induced gene, *OAS1*, was similarly significant (adjusted P value 1.05×10^{-21}, 3.8% methylation change). In a recent GWAS on critical illness due to SARS-CoV-2, significant associations and replication were observed for variants in the OAS gene cluster, which includes *OAS1*[63], for which variants had previously been associated in candidate gene studies of SARS-

CoV infection[61,64]. Also, in a Mendelian randomization study it was recently shown that increased circulating OAS1 proteins were associated with reduced SARS-CoV-2 susceptibility and disease severity[65]. Collectively, published genomics studies support several of the strongest associations observed in our study.

Previous work also demonstrated that viruses that cause severe disease (e.g., MERS-CoV, H5N1) alter host response by changing methylation landscape of antigen-presenting genes in the HLA region[4]. While we did not observe genome-wide significant signals at classical HLA alleles, we observed six FDR $q < 0.05$ probes in the region, in *HLA-V*, *DOA*, *DQA1*, *DQA2*, and *DRA*, albeit with attenuated significance compared to top CpGs (minimum $q \sim 0.0109$), suggesting that the mechanism of host manipulation by SARS-CoV-2 may be different. However, these results should be interpreted with caution as interrogation of the HLA region is complex; *HLA-V* for example is a pseudogene[66].

As the signatures identified in this study appear to be reactive to the disease, aspects of the disease process are expected to impact these results. Namely, we anticipate these changes to be time-sensitive, as the infection will need to have spread enough to induce methylation changes. Similarly, our case–control variables were defined by RT-PCR, which can carry a high false negative rate depending on the stage of infection and timing of sample collection[9], and may have reduced the classification accuracy. However, we have follow-up EHR information for the patients in this cohort, which minimizes the risk of misclassification bias. We do not expect this potential confounder to affect the measures of severity used in this study as these were determined directly from chart review, but we acknowledge that, for the initial analysis, the numbers of cases may have limited the statistical power and prognostic ability of ML. With additional cases that account for inherent genetic variability within the population, methylation patterns will become more refined and the AUC of these ML models to predict disease severity is likely to increase. While "duration of hospital stay" may not be as immediately actionable as predicting ICU admittance or ventilator use, and it is confounded by pre-existing frailty, social support (or lack of), socioeconomic status, and need for ongoing care once the acute illness has receded, the increased variability in the continuous outcome provides improved signal as observed both in our EWAS and our ML modeling. For this analysis, the 11 individuals who died were removed from duration analyses, as their length of stay would be difficult to compare to those who survived. Although the emerging field of epigenetics has demonstrated actionable classification with much smaller sample sizes in contrast to traditional GWAS in other common disease domains[67], we recognize that additional cases, and in particular understanding the less-severe end of the spectrum (which are likely to be under-reported in data from health systems), will improve our understanding of outcomes across the spectrum of disease severity. We note that, even in our limited sample sizes, the AUCs for ICU admittance still indicate there is signal that can be resolved through future collections. Another limitation of our work is the specificity of the epigenetic signature to SARS-CoV-2 over other respiratory infections. Initial targeted epigenetic analyses demonstrate a trend toward differential methylation, though these findings are limited by low numbers. Currently, we are targeting the collection of biospecimens from patients with respiratory infections other than SARS-CoV-2 for these follow-up studies.

Researchers have previously compared the robustness of DNA methylation profiling vs RNA transcriptome profiling in developing classifiers for different disease states[24,68–70]. One of the advantages of DNA methylation analysis compared to RNA analysis arises from the relative stability of deoxyribonucleic acid over ribonucleic acid[9,71]. The inherent instability of RNA, due to its 2'-OH group and the ubiquitous presence of ribonucleases, requires the use of plasticware, buffers, and processing reagents that are devoid of chemical and enzymatic species that stimulate RNA hydrolysis. Contamination even with a small amount of ribonuclease can degrade RNA samples to the degree where they cannot be analyzed.

The strong signature of viral-driven epigenetic changes may have the ability to detect SARS-CoV-2 infection in patients who never develop symptoms (asymptomatic) and in patients who are not yet symptomatic (pre-symptomatic)[72]. While asymptomatic testing following exposure has increased in recent months, the current testing strategy in the U.S. still predominantly targets symptomatic patients despite estimates that asymptomatic patients represent 40–45% of infected individuals[10,72]. Transmission during the incubation period has been reported, and the viral load of symptomatic and asymptomatic patients is similar[73–76]. The relationship between SARS-CoV-2 viral shedding and risk of transmission is unclear, and the percentage of transmission attributable to asymptomatic or pre-symptomatic infection of SARS-CoV-2 is unknown[77]. We believe that the epigenetics platform may efficiently identify asymptomatic and pre-symptomatic infections, which may, if applied broadly, aid in limiting the spread of SARS-CoV-2.

Due to the widespread occurrence of SARS-CoV-2 and progression to COVID-19 disease, there is the need for scalable testing technologies that can deployed on the national level for surveillance, screening, and prognosis for those infected. The purpose of this study was to identify high-confidence host methylation biomarkers that are able to indicate SARS-CoV-2 infection and predict clinical course of the viral disease in a given patient. This study is a first step toward selecting biomarkers for inclusion on a high-throughput methylation beadchip array specifically for the clinical diagnosis of COVID-19 disease that is also cost-effective given the added value of predicting subsequent clinical outcomes. To that end, we focused on sparse predictive models. Notably, these models are not significantly confounded by demographics or blood cell count information, denoting their specificity to the current infection of the patient, and reducing concern of overfitting to one patient sub-population. These biomarkers can also be used in risk stratification of SARS-CoV-2-infected patients, an unmet need given that none of the existing testing modalities (nucleic acid amplification tests, antigen tests, serology/antibody tests) can achieve this level of specificity. By identifying DNA methylation patterns associated with critical illness, we contend that a methylation test will provide patient-specific treatment targets before critical illness ensues. Preemptive dexamethasone[11,78], anticoagulation[12], or new pharmacologic targets may prevent mortality, guided by these epigenetics patterns. Although our findings must be complemented with further clinical assessment, our model has shown its capacity to leverage methylation quantification as an innovative strategy to generate epigenetic signatures that assess host response to SARS-CoV-2, which is scalable and may have the ability to confirm positive tests in asymptomatic patients and entire communities, and may ultimately differentially diagnose other viruses causing similar symptoms all within in a comprehensive high-throughput manner.

Data availability

The datasets generated during the current study are available in the Gene Expression Omnibus repository (accession GSE167202) and include original .idat array files and the final processed data matrix for DNA methylation analyses. Source data used to generate Figs. 2 and 4 are available as Supplementary Data 5 and 6.

Code availability

Raw array data were processed using SeSAMe 1.7.6 in R 4.0.1. EWAS was carried out using GLINT 1.0.4 on the command line. Machine learning analyses were done using

Glmnet v2.0-18 and Data.table v1.11.4 in R 3.5.1. Plotting and consolidation was done in R 4.1.0 using ggplot2 v2_3.3.3 and Data.table v1.14.0. All packages are available through CRAN and Bioconductor.

Received: 4 March 2021; Accepted: 24 September 2021;
Published online: 26 October 2021

References

1. FDA. Coronavirus disease 2019. (COVID-19) emergency use authorizations for medical devices. https://www.fda.gov/medical-devices/emergency-situations-medical-devices/emergency-use-authorizations (2020).
2. Menachery, V. D. et al. Pathogenic influenza viruses and coronaviruses utilize similar and contrasting approaches to control interferon-stimulated gene responses. *mBio* **5**, e01174–01114 (2014).
3. Marazzi, I. et al. Suppression of the antiviral response by an influenza histone mimic. *Nature* **483**, 428–433 (2012).
4. Menachery, V. D. et al. MERS-CoV and H5N1 influenza virus antagonize antigen presentation by altering the epigenetic landscape. *Proc. Natl Acad. Sci. USA* **115**, E1012–E1021 (2018).
5. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
6. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
7. Roser, M., Ritchie, H., Ortiz-Ospina, E. & Hasell, J. Coronavirus disease (COVID-19) – statistics and research. Our world in data. https://ourworldindata.org/coronavirus (2020).
8. Woloshin, S., Patel, N. & Kesselheim, A. S. False negative tests for SARS-CoV-2 infection - challenges and implications. *N. Engl. J. Med.* **383**, e38 (2020).
9. Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D. & Lessler, J. 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 (2020).
10. Fox-Lewis, S., Muttaiyah, S., Rahnama, F., McAuliffe, G. & Roberts, S. An understanding of discordant SARS-CoV-2 test results: an examination of the data from a central Auckland laboratory. *N. Z. Med. J.* **133**, 81–88 (2020).
11. Aslam, A. et al. SARS CoV-2 surveillance and exposure in the perioperative setting with universal testing and personal protective equipment (PPE) policies. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa1607 (2020).
12. Lan, L. et al. Positive RT-PCR test results in patients recovered from COVID-19. *JAMA* **323**, 1502–1503 (2020).
13. Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A. & Kozlakidis, Z. Considerations for diagnostic COVID-19 tests. *Nat. Rev. Microbiol.* https://doi.org/10.1038/s41579-020-00461-z (2020).
14. Garibaldi, B. T. et al. Patient trajectories among persons hospitalized for COVID-19: a cohort study. *Ann. Intern. Med.* https://doi.org/10.7326/M20-3905 (2020).
15. Bilinski, A. & Emanuel, E. J. COVID-19 and excess all-cause mortality in the US and 18 comparison countries. *JAMA* **324**, 2100–2102 (2020).
16. McElvaney, O. J. et al. A linear prognostic score based on the ratio of interleukin-6 to interleukin-10 predicts outcomes in COVID-19. *EBioMedicine* **61**, 103026 (2020).
17. Aref-Eshghi, E. et al. Evaluation of DNA methylation episignatures for diagnosis and phenotype correlations in 42 Mendelian neurodevelopmental disorders. *Am. J. Hum. Genet.* **106**, 356–370 (2020).
18. Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S. & Ollikainen, M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics* **11**, 1469–1486 (2019).
19. Agha, G. et al. Blood leukocyte DNA methylation predicts risk of future myocardial infarction and coronary heart disease. *Circulation* **140**, 645–657 (2019).
20. Jurmeister, P. et al. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci. Transl. Med.* **11**, eaaw8513 (2019).
21. Benhamida, J. K. et al. Reliable clinical MLH1 promoter hypermethylation assessment using a high-throughput genome-wide methylation array platform. *J. Mol. Diagn.* **22**, 368–375 (2020).
22. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
23. Karimi, S. et al. The central nervous system tumor methylation classifier changes neuro-oncology practice for challenging brain tumor diagnoses and directly impacts patient care. *Clin. Epigenetics* **11**, 185 (2019).
24. Korshunov, A. et al. DNA-methylation profiling discloses significant advantages over NanoString method for molecular classification of medulloblastoma. *Acta Neuropathol.* **134**, 965–967 (2017).
25. Saben, J. L. et al. The emergency medicine specimen bank: an innovative approach to biobanking in acute care. *Acad. Emerg. Med.* **26**, 639–647 (2019).
26. Pidsley, R. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
27. Robinson, J. et al. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–D431 (2015).
28. Boorgula, M. P. et al. Replicated methylation changes associated with eczema herpeticum and allergic response. *Clin. Epigenetics* **11**, 122 (2019).
29. Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* **37**, 549–554 (2005).
30. Zhou, W., Triche, T. J. Jr., Laird, P. W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.* **46**, e123 (2018).
31. Triche, T. J. Jr., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41**, e90 (2013).
32. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
33. Rahmani, E. et al. GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data. *Bioinformatics* **33**, 1870–1872 (2017).
34. Rahmani, E. et al. Genome-wide methylation data mirror ancestry information. *Epigenetics Chromatin* **10**, 1 (2017).
35. Rahmani, E. et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* **13**, 443–445 (2016).
36. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
37. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
38. Harris, M. A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
39. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
40. Huang, R. et al. The NCATS BioPlanet - an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. *Front. Pharmacol.* **10**, 445 (2019).
41. Slenter, D. N. et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**, D661–D667 (2018).
42. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
43. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).
44. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
45. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
46. van Iterson, M., van Zwet, E. W., Consortium, B. & Heijmans, B. T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* **18**, 19 (2017).
47. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
48. Iwata, H. et al. PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. *Nat. Commun.* **7**, 12849 (2016).
49. Lee, J. S. & Shin, E. C. The type I interferon response in COVID-19: implications for treatment. *Nat. Rev. Immunol.* **20**, 585–586 (2020).
50. Jiang, H. et al. DNA methylation markers in the diagnosis and prognosis of common leukemias. *Signal. Transduct. Target. Ther.* **5**, 3 (2020).
51. Ciolfi, A. et al. Frameshift mutations at the C-terminus of HIST1H1E result in a specific DNA hypomethylation signature. *Clin. Epigenetics* **12**, 7 (2020).
52. Wang, L., Ni, S., Du, Z. & Li, X. A six-CpG-based methylation markers for the diagnosis of ovarian cancer in blood. *J. Cell. Biochem.* **121**, 1409–1419 (2020).
53. Imgenberg-Kreuz, J. et al. Shared and unique patterns of DNA methylation in systemic lupus erythematosus and primary Sjogren's syndrome. *Front. Immunol.* **10**, 1686 (2019).
54. Bend, E. G. et al. Gene domain-specific DNA methylation episignatures highlight distinct molecular entities of ADNP syndrome. *Clin. Epigenetics* **11**, 64 (2019).
55. Wang, C., Chen, L., Yang, Y., Zhang, M. & Wong, G. Identification of potential blood biomarkers for Parkinson's disease by gene expression and DNA methylation data integration analysis. *Clin. Epigenetics* **11**, 24 (2019).
56. Panagopoulou, M. et al. Circulating cell-free DNA in breast cancer: size profiling, levels, and methylation patterns lead to prognostic and predictive classifiers. *Oncogene* **38**, 3387–3401 (2019).

57. Breton, C. V. et al. Small-magnitude effect sizes in epigenetic end points are important in children's environmental health studies: The Children's Environmental Health and Disease Prevention Research Center's Epigenetics Working Group. *Environ. Health Perspect.* **125**, 511–526 (2017).
58. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12**, 529–541 (2011).
59. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
60. Michels, K. B. et al. Recommendations for the design and analysis of epigenome-wide association studies. *Nat. Methods* **10**, 949–955 (2013).
61. Hamano, E. et al. Polymorphisms of interferon-inducible genes OAS-1 and MxA associated with SARS in the Vietnamese population. *Biochem. Biophys. Res. Commun.* **329**, 1234–1239 (2005).
62. Zhang, Q. et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, eabd4570 (2020).
63. Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
64. He, J. et al. Association of SARS susceptibility with single nucleic acid polymorphisms of OAS1 and MxA genes: a case-control study. *BMC Infect. Dis.* **6**, 106 (2006).
65. Zhou, Y. et al. Coagulation factors and the incidence of COVID-19 severity: Mendelian randomization analyses and supporting evidence. *Signal. Transduct. Target. Ther.* **6**, 222 (2021).
66. Horton, R. et al. Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899 (2004).
67. Tsai, P. C. & Bell, J. T. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int. J. Epidemiol.* **44**, 1429–1441 (2015).
68. Sadikovic, B., Aref-Eshghi, E., Levy, M. A. & Rodenhiser, D. DNA methylation signatures in mendelian developmental disorders as a diagnostic bridge between genotype and phenotype. *Epigenomics* **11**, 563–575 (2019).
69. Hovestadt, V. et al. Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta Neuropathol.* **125**, 913–916 (2013).
70. Vasudevan, H. N. et al. DNA methylation profiling demonstrates superior diagnostic classification to RNA-sequencing in a case of metastatic meningioma. *Acta Neuropathol. Commun.* **8**, 82 (2020).
71. Voet, D. & Voet, J. *Biochemistry* 4th edn (Wiley & Sons, 2011).
72. Oran, D. P. & Topol, E. J. Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. *Ann. Intern. Med.* **173**, 362–367 (2020).
73. Tan, F. et al. Viral transmission and clinical features in asymptomatic carriers of SARS-CoV-2 in Wuhan, China. *Front. Med.* **7**, 547 (2020).
74. Buitrago-Garcia, D. et al. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: a living systematic review and meta-analysis. *PLoS Med.* **17**, e1003346 (2020).
75. Zou, L. et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N. Engl. J. Med.* **382**, 1177–1179 (2020).
76. Hu, Z. et al. Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. *Sci. China Life Sci.* **63**, 706–711 (2020).
77. Liu, Y. et al. Viral dynamics in mild and severe cases of COVID-19. *Lancet Infect. Dis.* **20**, 656–657 (2020).
78. RC Group et al. Dexamethasone in hospitalized patients with Covid-19 - preliminary report. *N. Engl. J. Med.* **384**, 693–704 (2021).

Acknowledgements

We thank all study subjects at the University of Colorado Anschutz Medical Campus (CU-AMC) for their participation, along with Dr. Thomas Flaig, Dr. Adrie Van Bokhoven, and Dr. Alison Lakin for their leadership and technical support through the University of Colorado COVID-19 Biorepository; Dr. Richard Zane for his support on behalf of UCHealth; and Michelle Edelmann, Andrew Hadd, and Olivia Tintea for technical assistance. From the Illumina, Inc. team, we thank Krishna Bose, Anita Pottekat, Steven Gruber, John Picuri, Jay Kaufman, and Jason Johnson. This work was supported in part by the Health Data Compass Data Warehouse project (https://www.healthdatacompass.org), the Biobank at the Colorado Center for Personalized Medicine, and by grant funding from the University of Colorado Anschutz Medical Campus, Chancellor Discovery Innovation Fund.

Author contributions

K.C.B. and A.T. conceived the study. K.C.B., I.V.Y., C.R.G., R.A.M., P.J.N., A.T., R.P., and B.B. contributed to the design and/or production of the customized EPIC chip. K.C.B., K.C., B.R.P., S.J.W., and A.A.M. contributed to the development of the institutional pipeline for collecting, accessing, and/or testing of biospecimens. O.P., A.H., A.V., C.G.A., M.G.K., D.P.K., and A.A.M. contributed to the clinical informatics pipeline for this study. M.C., E.D., Y.Z., M.B., R.A.M., I.V.Y., and K.C.B. contributed to the design and/or operations of the methylation quantification. I.R.K., G.F.H., M.B., C.C., and I.V.Y. acquired and collated methylation data and/or performed the EWAS analyses. B.B., D.N., S.S., E.S., D.G., S.H., W.Z., and C.R.G. developed and tested the machine learning-based disease classifiers. I.R.K., C.R.G., I.V.Y., A.A.M., A.T., and K.C.B. drafted the manuscript and revised according to co-author suggestions. All authors critically reviewed the manuscript, suggested revisions as needed, and approved the final version.

Competing interests

We declare the following competing interests: B.B., R.P., and A.T. are employees at Illumina, Inc. and B.B., R.P., A.A.M., and A.T. own stock in Illumina, Inc. The rest of the authors declare that they have no relevant conflicts of interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-021-00042-y.

Correspondence and requests for materials should be addressed to Kathleen C. Barnes.

Peer review information *Communications Medicine* thanks Andrew Conway Morris and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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