Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions

In less than nine months, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) killed over a million people, including >25,000 in New York City (NYC) alone. The COVID-19 pandemic caused by SARS-CoV-2 highlights clinical needs to detect infection, track strain evolution, and identify biomarkers of disease course. To address these challenges, we designed a fast (30-minute) colorimetric test (LAMP) for SARS-CoV-2 infection from naso/oropharyngeal swabs and a large-scale shotgun metatranscriptomics platform (total-RNA-seq) for host, viral, and microbial profiling. We applied these methods to clinical specimens gathered from 669 patients in New York City during the first two months of the outbreak, yielding a broad molecular portrait of the emerging COVID-19 disease. We find significant enrichment of a NYC-distinctive clade of the virus (20C), as well as host responses in interferon, ACE, hematological, and olfaction pathways. In addition, we use 50,821 patient records to find that renin-angiotensin-aldosterone system inhibitors have a protective effect for severe COVID-19 outcomes, unlike similar drugs. Finally, spatial transcriptomic data from COVID-19 patient autopsy tissues reveal distinct ACE2 expression loci, with macrophage and neutrophil infiltration in the lungs. These findings can inform public health and may help develop and drive SARS-CoV-2 diagnostic, prevention, and treatment strategies.
n March 2020, the World Health Organization declared a pandemic of the coronavirus disease 2019 (COVID-19), an infection caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)\(^1\). Since the start of the pandemic, more than >100 million cases and 2 million deaths have been reported (https://coronavirus.jhu.edu), with an especially high burden of early cases in New York City (NYC). Genomic epidemiology efforts have already played a crucial public health role in confirming community spread of SARS-CoV-2 in the USA\(^2–4\) as well as in China\(^5\) and around the world\(^6–8\). However, standard approaches to viral profiling (i.e., qRT-PCR or targeted methods) fail to provide information on either the host immune response or microbial co-infections, both of which might impact the clinical presentation of COVID-19 and provide directions for therapeutic intervention and management.

Studies of SARS infection have shown that regulation of the innate immune response is associated with the development of adaptive immunity and disease severity. Moreover, a robust inflammatory response is characterized by an upregulation of cytokines including IL-6, IL-10, and MCP-1 in tissues and serum, as well as infiltration of infected tissues by inflammatory cells such as macrophages\(^9\). SARS-CoV-2 infection studies have shown that viral load peaks during the first week of illness, which may account for the high transmissibility of SARS-CoV-2.\(^2\)

Furthermore, antibody profiling by\(^10\) demonstrates both IgM and IgG antibodies began to increase around the 10th day after symptom onset, and most patients had seroconversion within the first 3 weeks of infection, underscoring the need for rapid testing in the acute phase of infection. Moreover, phenotyping of COVID-19 patients has also been shown to identify molecular signatures to distinguish severity of symptoms\(^11,12\), but limited information about the association of the viral sequences and outcome exists, nor viral propagation within human tissues.

To better understand the impact and progression of SARS-CoV-2 infection, we applied a multi-platform and molecular diagnostic approach to samples collected during the outbreak in NYC. Out of concern for the scalability and sensitivity of standard diagnostic assays (qRT-PCR), we developed and validated a rapid reverse transcription loop-mediated isothermal amplification (RT-LAMP) assay to detect SARS-CoV-2 infection from nasopharyngeal swab specimens and oropharyngeal swab lysates. We simultaneously developed a large-scale host and viral profiling platform employing shotgun metatranscriptomics (total RNA-seq). We applied this total RNA-seq platform to 732 clinical samples, including 669 confirmed or suspected COVID-19 cases at New York Presbyterian-Well Cornell Medical Center (NYPH-WCMC). In addition, we report an observational study of renin–angiotensin–aldosterone system inhibitors and severe COVID-19 outcomes using data \((n = 50,821)\) from NYPH-WCMC and New York-Presbyterian Hospital-Columbia University Irving Medical Center (NYU-CUIMC).

Our results define and map the viral genetic features of the NYC outbreak and associate specific host responses and gene expression perturbations with SARS-CoV-2 infection. We also link findings from host transcriptomics to findings from clinical data related to the angiotensin converting enzyme (ACE) pathway using patients at two NYPH sites. Finally, we used spatial transcriptomics to map ACE expression in SARS-CoV-2 infected lungs from COVID-19 patient autopsies. We have made these data available in the database of genotypes and phenotypes (dbGAP) and also in an interactive analytics portal to enable others to explore additional genomic, transcriptomic, and clinical covariate associations (https://covidgenes.weill.cornell.edu/).

### Results

**Validation of LAMP for detection of SARS-CoV-2.** We developed a colorimetric assay to quickly detect and quantify SARS-CoV-2 viral load in patient samples, using a set of LAMP primers and simple single-tube protocol (Fig. 1a, b). To validate the assay, we first evaluated two synthetic RNAs (see “Methods”) whose sequences matched the viral sequences of patients from Wuhan, China and Melbourne, Australia (Supplementary Fig. 1). The reaction output was measured at 0-, 20-, and 30 min intervals (Fig. 1c) before the samples were heated to 95 °C for inactivation. LAMP fluorescence correlated closely with SARS-CoV-2 RNA viral copies (Fig. 1d), with an overlap of the median signal from negative controls at lower levels (0–10 total copies per reaction) of viral RNA \((n = 10)\). The LoD was found to be between 5 and 25 viral total copies for the dual primer, single-tube reaction \((N\) gene and \(E\) gene), and this was then replicated to show a similar LoD on the second synthetic control (Supplementary Fig. 1b).

We next used a set of 201 samples from known or suspected COVID-19 cases that were tested for SARS-CoV-2 by a commercial qRT-PCR assay (Altona) that has been a standard clinical test at NYPH-WCMC since early March. These comprised 69 nasopharyngeal (NP) swab samples that tested positive (qRT-PCR positives, \(Ct < 40\)) and 132 samples that tested negative (qRT-PCR negative, \(Ct ≥ 40\)) (see “Methods”). qRT-PCR positive samples showed a much higher LAMP fluorescence than qRT-PCR negative samples. Analysis of Receiver Operator Characteristic (ROC) curves yielded an overall sensitivity of 95.6% and specificity of 99.2% (Fig. 1e, Supplementary Fig. 2). These results were also confirmed with a capture-based SARS-CoV-2 panel from Twist (Supplementary Fig. 2h, c). Of note, we obtained similar performance on bulk oropharyngeal swab lysate (Supplementary Fig. 3), including increasing reaction sensitivity as a function of viral copy number, but with deteriorating performance \(Ct > 30\). We observed higher LAMP sensitivity at higher viral loads, as determined by qRT-PCR \(Ct\) values (Supplementary Fig. 4).

**Shotgun metatranscriptomics platform for viral and bacterial detection.** To further investigate the biological characteristics of qRT-PCR positive and negative specimens, as well as to compare to RT-LAMP, we developed a shotgun metatranscriptomics platform utilizing total RNA-seq (RNA-sequencing with ribosomal RNA depletion) to profile all RNAs from patient specimens (Fig. 1a). We sequenced 857 RNA-seq libraries (Fig. 2) across 732 specimens from 669 patients treated for influenza-like illness (ILI) at NYPH-WCMC to an average of 63.2M read pairs per sample. This included 215 qRT-PCR positive samples (201 of which were also tested with LAMP), 17 positive (Vero E6 cells), and 33 negative (buffer) controls. To assess the ability of deep shotgun transcriptomics to identify qRT-PCR false negatives, including those that might arise from uncharacterized SARS-CoV-2 variants, we sequenced 517 qRT-PCR negative samples. Among these qRT-PCR negative samples, 311 were tested for the standard clinical respiratory virus panel (BioFire, which included several common cold coronavirus strains and influenza virus).

First, the taxonomic composition of samples was detailed by aligning total RNA-seq reads to the human reference genome GRCh38 and NCBI reference databases (see “Methods”). Kraken2 classification of non-human sequences revealed abundant bacterial and SARS-CoV-2 RNA among qRT-PCR positive samples and infrequent mappings to fungi, archaea, or other viruses (Fig. 2a). Positive controls and clinical samples with medium/high-viral load-consistent qRT-PCR \(Ct\) values were significantly \((p < 1 \times 10^{-16})\) enriched in SARS-CoV-2 genome alignments (median reads per kilobase per million mapped reads, RPKM) relative to
negative controls and qRT-PCR negative clinical samples (Fig. 2a). LAMP fluorescence, qRT-PCR Ct values, and total-RNA Seq RPKM showed consistent estimates of SARS-CoV-2 viral abundance across these three technologies and 201 clinical specimens ($R_{\text{seq}} \text{ vs. } \text{Ct} = -0.84$, $R_{\text{seq}} \text{ vs. } \text{lamp} = 0.82$, $R_{\text{lamp}} \text{ vs. } \text{Ct} = -0.80$) (Supplementary Fig. 4a, b). Analysis of SARS-CoV-2 read coverage across 517 qRT-PCR negative samples revealed 7 (1.3%) samples with more than 0.01% of reads matching SARS-CoV-2. In summary, these results indicate close concordance with SARS-CoV-2 viral load across diverse diagnostic platforms, with rare examples of occult SARS-CoV-2 infection among qRT-PCR negative cases.

Next, an analysis of total RNA-seq sequences showed the presence of commensal species and viruses across both qRT-PCR positives and negative samples (Fig. 2a). We found a statistically significant higher load of other (i.e., non-SARS-CoV-2) respiratory viruses in qRT-PCR negative vs. positive samples (Wilcoxon test, $P = 0.0008$) (Fig. 2b). Oral and airway commensal taxa were compared between the high-viral load group and other patient categories using linear modeling in log-normal space using MaAsLin2 package (see methods) and relative to the negative controls. This includes a correction for the SARS-CoV-2 or Homo sapiens matching reads, which can show variable depth of coverage due to higher SARS-CoV-2 fraction in the high-viral load group. After this correction, 17 species were significantly depleted in COVID-19 patients, including A. xylosoxidans, E. faecalis, and L. bacterium (Supplementary Fig. 5).

Among respiratory viruses discovered across all patients (COVID+ and COVID−), we found frequent Influenza A (23% of viral positive cases), rhinovirus A (16%), and human metapneumovirus (12%). Overall, we found close concordance between these results and the findings of a transcriptomics based viral detection to the results of a standard (BioFire) respiratory pathogen PCR panel performed within 7 days of the NP swab used for RNA sequencing ($N = 356$ patients across 404 tests) (Fig. 2c). The metatranscriptomics platform yielded an Area Under the Receiver Operator Characteristic (AUROC) curve of 0.890 (F-score = 0.725), and an accuracy of 0.994, sensitivity of 0.758, specificity of 0.997, and precision of 0.694 (see methods). These data also enabled an examination of co-infection within COVID+ patients. Among qRT-PCR positive cases, only 7 (3.2%) harbored sequences mappable to other respiratory viruses. Among these, we found human coronaviruses 229E, NL63 and HKU1, influenza A, and human mastadenovirus and metapneumovirus (Fig. 2b). In summary, these results indicate that common respiratory viruses were frequently implicated in many Influenza-like Illness (ILI) cases that were COVID−, and demonstrate low incidence of co-infection of SARS-CoV-2 with other respiratory viruses (3.2%), matching results (3%) from other studies13.

SARS-CoV-2 assemblies from shotgun metatranscriptomes. The abundance of SARS-CoV-2 alignments from total RNA-seq was sufficient to provide >10× coverage of the viral genomes and yield high quality, full-length assemblies for 155 samples (Supplementary Fig. 6). When examining the viral genomes, we identified a total of 1147 instances of 165 unique variants across...
155 assemblies, including 1143 single nucleotide variants (SNVs) and four deletions. We compared the 155 full-length NYPH-WCMC sequences with 46,581 SARS-CoV-2 sequences obtained from a recent GISAID build (downloaded on 6/16/2020) using the Nextstrain pipeline\(^7\) (Fig. 3a, see Methods). We generated full-length SARS-CoV-2 assemblies for 9 of 517 (1.7%) of samples that were found negative by qRT-PCR (Supplementary Fig. 6a). Each of these demonstrated a high RPKM viral load, an abundance of reads evenly covering the SARS-CoV-2 genome, and high (>0.5) variant allele fractions (VAFs) of SNVs commonly
seen in other positive samples and GISAID (Supplementary Fig. 6a). These results indicate that qRT-PCR false negatives can occur, but are rare.

We then used a New York State-focused build of Nextstrain resulting in the incorporation of our 155 genomes with 3541 sequences from GISAID (Fig. 3a). Analyzing worldwide groupings annotated by the Nextstrain database, we found a high proportion (>74%, 116/155) of our assemblies were associated with clade 20C (Fig. 3b), which overall represents 17% of strains in the GISAID build. Our metatranscriptomic assembly-based clade assignments for NYC strains were consistent with those obtained by parallel targeted-sequencing based studies from the same period (Fig. 3b). Investigating the geographical and temporal distribution of non-NYC samples in 20C, we found a global distribution with a likely origin in Western Europe (Fig. 3c). The first noted cases of 20C in GISAID were recovered in France on 2/21/2020, but since then, this clade has represented a minority (<20%) of Western European cases. In contrast, 20C represented a steady majority (>70%) of NYC cases from March to May 2020. Interestingly, 20C was relatively rare (<50%) in other regions of the US in March, but increased through June. The most striking increase was in the southern USA, where 20C began under 30% prevalence in March and rose to above 80% prevalence in May and June. The SNVs defining the 20C clade were non-synonymous variants targeting genes encoding the nonstructural protein 2 (NSP2), the viral replicase, and ORF3a, a poorly characterized SARS-CoV-2 protein with putative roles in inflammation. These observations refine the results of a previous study that associated NYC cases with clade A2a. This clade is defined by the presence of the Spike p.D614G mutations and comprises a superset of 20C.

While the majority of the known variation in SARS-Cov-2 represents SNVs, one hundred distinct deletion variants have been reported in previous analyses. Of the four deletions we detected in our assemblies, three represented the same in-frame deletion in nsp1 (p.141_143del) (Supplementary Fig. 7a). This deletion was well supported by reads in our samples (Supplementary Fig. 7b). While it was not exclusive to genomes from our study, we only detected this deletion in 168 GISAID genomes (0.3%). Interestingly, rather than forming a single monophyletic subclade, genomes carrying this deletion spanned all major SARS-Cov-2 clades (19A: 12, 19B: 4, 20A: 68, 20B: 30, 20c: 53, none: 4) suggesting that this deletion may have evolved many times during SARS-CoV-2 community spread (Supplementary Fig. 7c). In addition, while the dominant (171/172) version of this deletion involved the in-frame deletion of nucleotide positions 686-694, a single genome included an alternative in-frame deletion of positions 684–692, resulting in the same amino-acid deletion.

Variant calling of assemblies and RNA-seq reference alignments showed most (87.5%) variants have VAFs >0.95 and high (>100) numbers of variant-supporting reads. Analysis of VAF posterior probability distributions identified a subset of alleles whose VAFs were confidently (probability > 0.95) above 0.05 but below 0.95 (Fig. 3d, e). We labeled these alleles as "het" (heterogenous) variants and refer to the remaining (high VAF) consensus assembly variants as "homogeneous." Many of these het variants were associated with robust read support despite...
frequently having VAFs below 0.05 and not being incorporated into the consensus assembly sequence. (Supplementary Fig. 6b). These results indicate that SARS-CoV-2 samples frequently harbor minority viral populations, which can be detected and characterized with shotgun metatranscriptomics. The relatively large numbers of these het variants (>10) in a subset of cases, presumably arising during the course of a single infection and then rising to a reasonably high oligoclonal VAFs (0.1–0.2) level, suggests that this intrahost diversification may be rapid and also associated with positive selection.

**Defining the SARS-CoV-2 host transcriptome.** We then leveraged the total RNA-seq profiles to better understand the host transcriptome during SARS-CoV-2 infection. Cell proportion values and iron (heme) regulation during infection. Accordingly, late erythropoiesis, which indicated a connection to hematological COVID-19. Since ACE2 expression can be increased in patients taking ACEIs andARBs, the observed correlation of viral titer with ACE2 expression may be attributed to the pre-infection use of such inhibitors, which is common in older patients and those with comorbidities.

To address this, we analyzed ACEI/ARB use and severe COVID-19 outcomes in an observational cohort of individuals (n = 50,821) suspected of SARS-CoV-2 infection at New York-Presbyterian Hospitals, comprising 23,170 patients from Columbia University Irving Medical Center (NYP-CUIMC) and 27,651 patients from Weill Cornell Medical Center (NYP-WCMC, Table 1). At both sites, we found evidence that ACEI/ARB use is associated with lower rates of intubation (CUIMC Hazard ratio, HR = 0.79, 95% confidence interval, CI: 0.68–0.91, WCMC HR = 0.62, CI: 0.48–0.81) and lower risk of death (CUIMC HR = 0.66, CI: 0.58–0.75, WCMC HR = 0.73, CI: 0.56–0.95) following confirmed SARS-CoV-2 infection (Fig. 5). Each comparison used propensity matching (PSM) and Cox proportional hazards models with covariate adjustments for age, sex, race, ethnicity, drug indications, and relevant comorbidities (see “Methods”; Table 1).

ACEI/ARB have a number of indications in addition to hypertension, so confounding variables remain a challenge. To address this issue, we evaluated the analysis in a cohort of individuals with recent exposure to one of four antihypertensive medication classes: ACEI/ARB, beta blockers (BB), calcium channel blockers (CCB), and thiazide/thiazide-like diuretics (THZ). Evaluating each drug class separately using PSM and covariate adjustment, we found the rate of intubation was consistently lower for ACEI/ARB exposure (CUIMC HR = 0.74, CI: 0.64–0.85, WCMC HR = 0.73, CI: 0.56–0.95) and higher for THZ exposure (CUIMC HR = 1.96, CI: 1.67–2.31, WCMC HR = 1.90, 1.42–2.55). Conversely, we found no statistically significant associations for CCB (CUIMC HR = 0.92, 0.81–1.05, WCMC HR = 1.04, CI: 0.83–1.30), and only weak evidence of positive associations between intubation and BB exposure (CUIMC HR = 1.16, CI: 1.03–1.31, WCMC HR = 1.10, CI: 0.84–1.27). Meanwhile, we found consistently lower rates of death among patients with ACEI/ARB (CUIMC HR = 0.82, CI: 0.71–0.93, WCMC HR = 0.67, CI: 0.52–0.87), BB (CUIMC HR = 0.81, CI: 0.71–0.91, WCMC HR = 0.74, CI: 0.61–0.91), and CCB (CUIMC HR = 0.63, CI: 0.55–0.72, WCMC HR = 0.69, CI: 0.55–0.86) exposures, while THZ exposure had harmful associations (CUIMC HR = 1.03, CI: 0.87–1.22, WCMC HR = 1.43, CI: 1.04–1.96). In summary, among SARS-CoV-2–infected individuals with recent exposure to antihypertensive drugs, ACEI/ARBs were the only class which were associated with lower risk of intubation, while ACEI/ARB, BB, and CCB use was negatively associated with death. (Complete estimates are provided in Supplementary Data 6).

**Spatial expression data.** We next used a spatial transcriptomic technology (GeoMx) to characterize the distribution of a set of >1800 genes (the GeoMx Cancer Transcriptome Atlas probe set, or CTA) across tissues from four COVID-19 patients who died...
and then received an autopsy (Fig. 6); these were compared to three normal lung donors (excess lung material from healthy lung transplant donors). Samples were processed using the Nanostring GeoMx protocols (see methods), including illuminating and UV-cleaving the antibody-linked probes for Areas of Interest (AOIs) within the patients’ lung FFPE sections and alveoli, followed by Next-Generation Sequencing (NGS) profiling on the Illumina NextSeq500. These data showed spatially-restricted expression of ACE2 at the perimeter of the alveoli (Fig. 6b), as well as the presence of the SARS-CoV-2 virus in some of the same sections.

We then used cellular deconvolution methods based on the CTA data across the 107 total AOIs (69 COVID+ and 38 COVID−) to characterize the types of cells present in the lung. The cell proportions showed an increase in macrophages present in the lung tissue (Fig. 6c) compared to normal donors, particularly for patient #3, where 30–52% of the cells found within the AOIs were macrophages, as well as patient #4 (up to 18%). In contrast, normal lung samples showed <1% across all AOIs tested (n = 38 total AOIs). Of note, patient #1 showed up to 3.8% presence of neutrophils in the AOIs profiled, compared to...
Depletion and enrichment of HLA types in COVID patients.

Finally, to probe host genetic diversity in widely expressed genes broadly implicated in immune responses, we also checked RNA data for potential enrichment or depletion of class I HLA type(s) among COVID patients. Using arcasHLA, we tallied frequencies of first-field locus-specific haplogroups (serotype alleles, in HLA parlance) called with default or greater confidence, among SARS-CoV-2 qRT-PCR-positive versus -negative patients who each had at least 100 class I HLA-mapped transcriptome reads. First-pass single-locus Fisher exact tests suggested modest depletion of $\text{HLA-B}^{*}08$ and/or $\text{HLA-B}^{*}18$, and/or enrichment of $\text{HLA-B}^{*}35$, $\text{HLA-B}^{*}39$, and/or $\text{HLA-B}^{*}48$, in COVID patients (Supplementary Fig. 12), but no such disparity in $\text{HLA-A}$ or $\text{HLA-C}$ (Supplementary Fig. 13). To more stringently buffer finite-sample dependence among calls at each locus, and the risk of spurious findings under multiple testing, we permuted qRT-PCR positive and negative labels, preserving sample diplotype calls in 100 randomly seeded runs of 1000 permutations each. Such permutation supported, by consistent extremity of original counts to those under permutation, putative COVID patient-depletion of $\text{HLA-B}^{*}08$ and $\text{HLA-B}^{*}18$ and COVID patient-enrichment of $\text{HLA-B}^{*}39$ and $\text{HLA-B}^{*}48$, but not of $\text{HLA-B}^{*}35$ (Supplementary Table 1 ACEI/ARB characteristics.

| Characteristic | CUIMC No ACEi/ARB | ACEi/ARB | WCMC No ACEi/ARB | ACEi/ARB |
|---------------|-------------------|---------|------------------|---------|
| N (total = 50,821) | 19,299 | 3871 | 25,017 | 2634 |
| Median age (IQR) | 48 (34.1-63.2) | 68.2 (58.9-77.2) | 50 (35-67) | 68 (58-76) |
| Male sex (%) | 7479 (38.8) | 2021 (52.2) | 9357 (37.4) | 1385 (52.6) |
| Race—Asian (%) | 590 (3.1) | 64 (1.7) | 2852 (11.4) | 225 (8.5) |
| Race—Black (%) | 3294 (17.1) | 802 (20.7) | 4031 (16) | 453 (17.2) |
| Race—White (%) | 5942 (30.8) | 1319 (34.1) | 9365 (37.4) | 1064 (40.4) |
| Ethnicity—Hispanic (%) | 5020 (26) | 1534 (39.6) | 3915 (15.6) | 464 (17.6) |
| Asthma (%) | 1816 (9.4) | 715 (18.5) | 1152 (4.6) | 345 (13.1) |
| Chronic kidney disease (%) | 2980 (15.4) | 2131 (55.1) | 3580 (14.3) | 1621 (61.5) |
| Chronic obstructive lung diseases (%) | 846 (4.4) | 545 (14.1) | 609 (2.4) | 283 (10.7) |
| Coronary artery disease (%) | 1253 (6.5) | 1243 (32.1) | 842 (3.4) | 627 (23.8) |
| Diabetes mellitus (%) | 2300 (11.9) | 1999 (51.6) | 1928 (7.7) | 1139 (43.2) |
| Diabetic nephropathy (%) | 266 (1.4) | 286 (7.4) | 217 (0.9) | 236 (9) |
| Diabetic neuropathy (%) | 325 (1.7) | 355 (9.2) | 141 (0.6) | 159 (6) |
| Diabetic retinopathy (%) | 135 (0.7) | 158 (4.1) | 93 (0.4) | 138 (5.2) |
| Diabetic vasculopathy (%) | 123 (0.6) | 193 (5) | 48 (0.2) | 70 (2.7) |
| Heart failure (%) | 1176 (6.1) | 1187 (30.7) | 676 (2.7) | 576 (21.9) |
| Hypertension (%) | 3986 (20.7) | 2928 (75.6) | 3592 (14.4) | 2227 (84.5) |
| Insulin use (%) | 1839 (9.5) | 1642 (42.4) | 910 (3.6) | 729 (27.7) |
| Myocardial infarction (%) | 499 (2.6) | 614 (15.9) | 375 (1.5) | 301 (11.4) |
| Obesity (%) | 3659 (19) | 1610 (41.6) | 3529 (14.1) | 1166 (44.3) |
| Proteinuria (%) | 3063 (15.9) | 1465 (37.8) | 1327 (5.3) | 617 (23.4) |
| HbA1c median (mg/dL) | 5.9 (2315) | 6.7 (1558) | 6.1 (895) | 6.7 (452) |
| Beta blocker (%) | 2208 (11.4) | 2198 (56.8) | 1746 (7) | 1406 (53.4) |
| Calcium channel blocker (%) | 1854 (9.6) | 1902 (49.1) | 1305 (5.2) | 1176 (44.6) |
| Thiazide/thiazide-like diuretic (%) | 664 (3.4) | 1345 (34.7) | 476 (1.9) | 816 (31) |
| Tested for SARS-CoV-2 infection (%) | 17812 (92.3) | 3500 (90.4) | 23578 (94.2) | 2464 (93.5) |
| Confirmed SARS-CoV-2 infection (%) | 4697 (24.3) | 1252 (32.3) | 2878 (11.5) | 479 (18.2) |
| SARS-CoV-2 infection/COVID-19 (%) | 7128 (36.9) | 1730 (44.7) | 4343 (17.4) | 652 (24.8) |
| Intubated/COV+ (%) | 437 (6.1) | 242 (14) | 254 (5.8) | 74 (11.3) |
| Died/COV+ (%) | 504 (7.1) | 261 (15.1) | 226 (5.2) | 69 (10.6) |

Depletion and enrichment of HLA types in COVID patients. Finally, to probe host genetic diversity in widely expressed genes broadly implicated in immune responses, we also checked RNA data for potential enrichment or depletion of class I HLA type(s) among COVID patients. Using arcasHLA, we tallied frequencies of first-field locus-specific haplogroups (serotype alleles, in HLA parlance) called with default or greater confidence, among SARS-CoV-2 qRT-PCR-positive versus -negative patients who each had at least 100 class I HLA-mapped transcriptome reads. First-pass single-locus Fisher exact tests suggested modest depletion of $\text{HLA-B}^{*}08$ and/or $\text{HLA-B}^{*}18$, and/or enrichment of $\text{HLA-B}^{*}35$, $\text{HLA-B}^{*}39$, and/or $\text{HLA-B}^{*}48$, in COVID patients (Supplementary Fig. 12), but no such disparity in $\text{HLA-A}$ or $\text{HLA-$C$}$ (Supplementary Fig. 13). To more stringently buffer finite-sample dependence among calls at each locus, and the risk of spurious findings under multiple testing, we permuted qRT-PCR positive and negative labels, preserving sample diplotype calls in 100 randomly seeded runs of 1000 permutations each. Such permutation supported, by consistent extremity of original counts to those under permutation, putative COVID patient-depletion of $\text{HLA-B}^{*}08$ and $\text{HLA-B}^{*}18$ and COVID patient-enrichment of $\text{HLA-B}^{*}39$ and $\text{HLA-B}^{*}48$, but not of $\text{HLA-B}^{*}35$ (Supplementary Fig. 12B). We caution that these findings, while nominally significant, neglect plausible ancestry-differential COVID-19 infection incidence and/or ascertainment in our NYC cohort (which may strongly confound HLA frequency comparison), as well as intrinsic imprecision of diplotype inference from RNA data, linkage among class I HLA genes (and other major histocompatibility complex loci), and/or other sequence variation distinctive to specific subtypes of surveyed haplogroups.

Discussion

In summary, these data showed that total RNA-sequencing data has great promise as a comprehensive and accurate host and pathogen profiling method. The RNA-seq data enabled a complete genetic map of the viruses, bacteria, host responses, and even HLA subtypes from the sample data source. Of note, a significant subset of our samples, including in nine qRT-PCR negative cases, had sufficient reads to assemble the SARS-CoV-2 genome de novo. Though these likely qRT-PCR false negatives could not be attributed to specific sequence changes (e.g., primer site mutations), their high frequency underscores the limitation of “gold standard” qRT-PCR approaches for SARS-CoV-2 detection. These results also highlight the need for open-source primer design, so these assays can be updated as a more granular picture of strain diversity and evolution is obtained through worldwide sequence efforts. Nevertheless, the rate of SARS-CoV-2 assemblies or reads among qRT-PCR negative cases with ILI symptoms was low, despite a comprehensive search for these sequences among shotgun transcriptomics...
reads. These results indicate that existing standard detection methods are reasonably adapted to patterns of SARS-CoV-2 strain variation.

Notably, our phylogenetic analysis shows that the 20C clade comprises the majority of known NYC samples, including those sequenced outside of this study. Though remaining NYC cases show a wide distribution across all identified clades, the predominance (>80% in NYC) of such a narrowly defined set of sequences within NYC from a minority (<20%) Western European suggests either a founder effect or differential strain fitness. Though the Spike D614G mutation, which defines a superclade of 20C previously known as A2a, has been proposed to increase transmissibility (e.g., via higher viral load), and the low and stable prevalence of 20C in Western Europe since March suggests against differential fitness as an explanation for its dominance in NYC. A more likely explanation involves the early introduction and expansion of one or several 20C cases to NYC in late February, e.g., via the New Rochelle cluster. The gradual increases of 20C in the remainder of the USA post March may similarly reflect migration patterns of New Yorkers to the South and Midwest.

Our results further demonstrate the value of RNA-seq data by revealing distinct, host transcriptional programs that were activated during viral infection of the naso-/oropharynx with SARS-CoV-2. This includes upregulation of specific interferon pathway genes (SHFL, IFI6, IFI27, and IFIT1) that have been previously associated with the innate antiviral host immune response against other positive-strand RNA viruses (e.g., hepatitis C, Dengue virus). These results also provide clinical relevance for recently published results from animal and cellular models of SARS-CoV-2. Our analyses also implicate expression perturbations of the ACE pathway in SARS-CoV-2 host response, including ACE2, and map these patterns from COVID-19 lung samples from autopsies. Patients presenting with COVID-19 frequently harbor comorbidities such as hypertension, diabetes mellitus, and coronary heart diseases, all of which have been associated with increased disease severity. Since these comorbidities are frequently treated with ACEIs and ARBs, one possibility is that these medications may make patients more susceptible to SARS-CoV-2 infection, but this has been limited in prior studies.

Here, we examined the risk of medication use and comorbidities in the context of ACEI or ARB in a large, retrospective

Fig. 5 Host risk to SARS-CoV-2. a Estimates for ACEI/ARB use effect on SARS-CoV-2 infection test result, intubation, and death. Upper panels show estimates from all patients without propensity matching (unmatched, no condition (N.C.)). Lower panels show estimates from propensity-matched patients with recent exposure to antihypertensive (AH) drugs (ACEI, ARB, BB, CCB, and THZ) (Supplementary Data 6). Center points show the estimated coefficient and the error bars show 95% confidence interval. Coefficients represent effect size estimates from logistic regression (infection test outcome) or Cox proportional hazards regression (mechanical respiration and mortality), and can be interpreted as log odds ratios (former outcome) or log hazard ratios (latter outcomes). Colors indicate the variables used for regression on the cohort (drug exposure alone (blue), drug exposure plus age, sex, history of diabetes mellitus (DM), history of hypertension (HTN) (green), drug exposure plus all demographics and risk factors considered (red)). Left panels show estimates from NYP-CUIMC data, right panels show estimates from NYP-WCMC. b Curves for 50,821 patients requiring mechanical respiration (top panels) and survival (bottom panels), as a function of time since confirmed infection and recent ACEI/ARB exposure status. Cohorts created using propensity matching. Left panels give data from NYP-CUIMC, right panels give data from NYP-WCMC. c Comparison of cohorts before and after propensity matching. Each point represents a difference in the mean between exposed and unexposed cohorts, divided by the mean of exposed and unexposed cohorts. Filled circles give these standardized differences after propensity matching, unfilled circles give values before matching.
clinical analysis of 50,821 patient records. These data show that patients with ACEI/ARB exposure had lower risk of severe COVID-19 outcomes, after adjusting for demographics, indications, and other relevant comorbidities. A recent study reported a protective effect of ACEI/ARB exposure on mortality. The associations we found support this finding, and we found similar associations for mechanical respiration requirements among infected patients. Further, our data did not show that other antihypertensive drug classes had any association with mechanical respiration requirements, though BB and CCB were associated with lower risk of death as well. While our data suggest a protective association between ACEI/ARB use and severe COVID-19 from two large New York hospitals (Columbia and Weill Cornell), our results are still preliminary, and caution should be taken in interpreting them. Because of the risk of residual confounding variables, prospective clinical trials would...
be necessary before clinical guidelines should be changed. For example, if some patients are more susceptible because they already express high levels of ACE2, this could help with targeting the ACE pathway in these patients as a prophylactic method. However, if the cells respond to infection with ACE2 expression, and this leads to the cytokine storm seen in patients, then this could be used as a downstream treatment (post-infection), for when ACE2 interacts with TMPRSS2, such as the ongoing trials with camostat mesylate34. Mitigating a fast-spreading pandemic, such as COVID-19, requires scalable diagnostic and screening methods29,30. Such a situation calls for fast, scalable tests readily implemented at home or point-of-care. This one-tube, dual-primer colorimetric SARS-CoV-2 assay has been clinically-validated with qRT-PCR, capture, and total RNA-seq, and thus can help increase testing options; moreover it has recently become available in several commercial forms. Our preliminary results here and elsewhere demonstrate SARS-CoV-2 detection from oral specimens is feasible and even optimal31,32. As such, a LAMP-based approach on such sample types could allow facilities to increase testing capabilities by orders of magnitude. Viral pandemics can have significant, long-lasting detrimental impacts for affected countries, and thus it is crucial to deploy methods that can track and profile cases (e.g., RNA-seq, LAMP, qRT-PCR, capture) and provide a comprehensive view of host and viral biology33. These methods can help mitigate the medical and socioeconomic harm from viral outbreaks, as well as establish protective surveillance networks that can help defend against future pandemics.

Methods

Sample collection and processing. Patient specimens were collected with patient consent at New York-Presbyterian Hospital-Weill Cornell Medical Center (NYPWH-WCMC) and then processed for qRT-PCR. Nasopharyngeal (NP) swab specimens were collected using the BD Universal Viral Transport Media system (Becton, Dickinson and Company, Franklin Lakes, NJ) from symptomatic patients. Samples were collected and processed through the Weill Cornell Medicine Institutional Review Board (IRB) Protocol 19-11021069. All relevant ethical regulations for work with human participants were complied with. Observational cohort analysis (ACEI/ARB) was done through the Columbia University IRB Protocol AAA016.0601. Oropharyngeal samples data from University Hospital of Tuebingen were collected under IRB 243/2020A.

Extraction of Viral RNA and qRT-PCR detection. Total viral RNA was extracted from decontaminated samples using automated nucleic acid extraction on the QIAexpress and the DSP Virus/Pathogen Mini Kit (QIAGEN). One step reverse transcription to cDNA and real-time PCR (RT-PCR) amplification of viral targets, E (envelope) and S (spike) genes and internal control, was performed using the Rotor-Gene Q thermocycler (QIAGEN).

Twist synthetic RNAs. We used two fully synthetic RNAs made by in vitro transcription from Twist Biosciences, which were synthesized in 5 kb pieces with full viral genome coverage of SARS-CoV-2. They were sequence verified to ensure >99.9% viral genome coverage, and come as 1,000,000 copies per µL, 100 µL per tube. The two controls are from Wuhan, China (NM089847.3) and Melbourne, Australia (MT007544.1). Reference sequence designs came from NCBI: https://www.ncbi.nlm.nih.gov/muccore/MT007544 and https://www.ncbi.nlm.nih.gov/muccore/NM089847.3.

Reverse transcriptase, quantitative real-time PCR (RT-PCR). Clinical samples were extracted as described above and then tested with qRT-PCR using primers for the E (envelope) gene, which detects all members of the lineage B of the beta-CoVs, including all SARS, SARS-like, and SARS-related viruses, and a second primer set for the S (spike) gene, which specifically detects the SARS-CoV-2 virus. The reaction also contains an internal control serving as an extraction control and a control for PCR inhibition.

Samples were annotated using qRT-PCR cycle threshold (Ct) value for SARS-CoV-2 primers. Subjects with Ct less than or equal to 18 were assigned “high-viral load” label. Ct between 18 and 24 were assigned “medium viral load” and Ct between 24 and 40 were assigned “low viral load” classes, with anything above Ct of 40 was classified as negative. We also predicted a combined viral load score using Ct, GloMax QuantFluor read-out from LAMP experiments and fraction of SARS-CoV-2 matching NGS reads in a sample. For this score (40-Ct), (LAMP read-out) and (log10(SARS-CoV-2 fraction + 1e-6)) were all normalized between zero and one individually, and summed together using a combination weight of 5 for Ct, 3 for LAMP and 2 for NGS.

LAMP primer sequences. Primers (Supplementary Table 7) were designed using PrimerExplorer (v4.0), as per guidelines in18 to find LAMP-compatible primers for the SARS-CoV-2 reference genome (NCBI). LAMP’s inherent specificity (using 4–6 primers vs. 2 for PCR amplification), in combination with this in silico analysis, suggested little risk of cross-reactivity that might hinder specificity of N-gene primers sensitive for SARS-CoV-2 (Supplementary Data 5). Overall, chosen primers had less than 80% homology with the vast majority of tested pathogen sequences. For any organisms where a primer hit >80% homology, only one of the six primers had significant homology making an amplified product extremely unlikely. Overall, the results of this analysis predict no significant cross-reactivity or microbial interference. We also assessed the potential impact of sequence variation in circulating strains that might lead to poor amplification. In the thousands of sequences deposited in GISAID (Shu and McCauley, 2017), only one site in the priming region was observed to be polymorphic. The polymorphism (T30359C) was only observed in 106 of 6753 (1.5%) sequences with coverage of this region. This variant overlaps the priming site of the LB primer but is not near a 3-prime end and is not anticipated to cause amplification failure. The data from Fig. 1 show the use of a single-tube, dual-primer protocol, wherein both the N2 gene and E-gene primers are present.

The LAMP reaction setup. For each well or Eppendorf tube, we used a set of six primers (above) for Gene N, the M1800 Colorimetric LAMP Master Mix (NEB), water, and 11.5 µL of the sample. The protocol is as follows:

1. Reagents added:
   a. 12.5 µL M1800 LAMP mix (NEB).
   b. 1–5 µL LAMP primers (Gene N or N2/E mix)
   c. 1–11.5 µL of sample.
   d. Remaining volume (to 25 µL) H₂O.

2. Vortex, spin down.

3. Place on Thermocycler at 65 °C for 30 min with lid at 105 °C.

4. Remove tubes, place on ice for 5 s.

5. Visualize over lab bench/ice/paper.

Oropharyngeal lysate LAMP run. Nasopharyngeal and/or oropharyngeal swab samples from 201 patients were collected (IRB 243/2020A Tuebingen) using a dry cotton swab (cliniswab DS, Aptaca Spa (Italy), #2170/SZ/CS). Crude extraction was performed according to pending unpublished European patent [No. 20168 593.0]. In summary, the dry swab was transferred to a 15 mL falcon (Greiner Bio-one, #188,271), filled with 0.5 mL saline solution, and shaken vigorously for 30 min. Afterwards the liquid is transferred to a screw-cap (Sarstedt, #72.692.005). 1µl of the crude extract was added to 12.5 µL 2X NEB LAMP master mix (M1800), 2.5 µL of water, 1 µl 25× primer master mix Gen N, and was incubated at 65 °C for 30–40 min. A sample of a patient who tested positive using an approved qRT-PCR test (sample #1123) was used as an internal control. The read-out was performed visually by color change from pink to yellow or orange.

The RNA isolation was performed with 50 µl of the crude extract on the QIAexpress with the DSP Virus/Pathogen Kit (Qiagen, #937055). The RT-PCR was performed using 5 µl of 85 µl eluate with TIB MolBio Lightmix MODULAR SARS AND WUHAN CoV E-Gene Kit. Analysis was done with the LightCycler(R) 480 II software and calculated CP values were used for statics and graphical analysis. The Standard curve with a synthetic RNA control (Twist Bioscience, #MT007544.1) was generated using the LAMP assay and in parallel with the qRT-PCR. The control RNA was diluted serially tenfold with water and the absolute copy number, ranging from 10^6 to 10^-5 was analyzed. Gel electrophoresis after visual read-out of the LAMP assay was done by loading 5 µl of the LAMP reaction with 5 µl 2× loading dye on a 1.5% agarose (Seakem LE Agarose, Lonza #50004) gel for 40 min. A sample of a patient who tested positive using an approved qRT-PCR test (sample #1123) was used as an internal control. The read-out was performed visually by color change from pink to yellow or orange.

Light intensity and data processing. Completed reactions were analyzed with the Promega GloMax Explorer (Promega GM5300) fluorometer using the QuantFluor ONE dioxane DNA system (Promega E4871). This system recorded fluorometric read-out from each well using an emission filter of 500-550nm, an excitation filter set at
blue 475 nm, and a high sensitivity setting on the Glomax software. Values were then tabulated and compared with controls (positive and negative). The intensity threshold of 2.5× negative control was used as the threshold for positive detection.

**DNAse treatment, rRNA depletion, and RNAseq library construction.** For library preparation, all samples’ total nucleic acid (TNA) were treated with DNAse I (Zymo Research, Catalog # E1010), which cuts both double-stranded and single-stranded DNA. Post-DNAse digested samples were then put into the NEBNext rRNA depletion v2 (Human/Mouse/Rat), Ultra II Directional RNA (10 ng), and Unique Dual-Index Primer Pairs were used following the vendor protocol from New England Bioslabs (except for the first flowcell, see Supplementary Figures). Kits were supplied from a single manufacturer lot. Completed libraries were quantified by Qubit or equivalent and run on a Bioanalyzer or equivalent for size determination. Libraries were pooled and sent to the WCM Genomics Core or HudsonAlpha, which sequenced them using the HiSeq4000 v2 instrument by Qubit (Invitrogen/Scientific), TapeStation 2200 (Agilent), and qRT-PCR using the Kapa Biosystems Illumina library quantification kit.

**Taxonomic classification of sequence data.** All complete genome or chromosome level assemblies from ResSeq database for archaea, bacteria, protozoa, fungi, human and viruses including SARS-CoV and SARS-CoV-2 genomes were downloaded and used for building a classification database for Kraken2 (k = 35, E = 343). To get an approximation for the positive and negative classification rate, the BBMap random-reads script was used to simulate 10 million 150 bp paired-end Illumina reads from the database sequences46. For the negative test all sequences in the database excluding SARS-CoV and SARS-CoV-2 genome were removed from the sequences and the simulated reads were mapped with the Kraken2 database (Supplementary Data 1).

For the positive test, the same process was repeated using only SARS-CoV-2 genome (Supplementary Data 2). Positive results show >99% of SARS-CoV-2 reads uniquely map to either SARS-CoV or SARS-CoV-2, with the remaining 1% are ambiguous, potentially matching multiple taxa (Supplementary Data 1, 2). All sequences were classified using the Kraken2 database. To remove the potential contamination of reads that are homologous across multiple species we used Kraken2 outputs to filter sequences to either human (uniquely matching Homo sapiens and no other taxon in our database), SARS-CoV-2 (either matching SARS- CoV or SARS-CoV-2 but homologous between these two viruses), and remaining reads that may be coming from unclassified, archaeal, bacterial, viral, fungal, protozoan or ambiguously mapping reads to human or SARS-CoV-37).

Using kraken2 classifications common respiratory pathogens were identified in clinical samples. Any SARS-CoV-2 negative sample with >0.01% relative abundance (normalization after the exclusion of any human, SARS-CoV-2 and uncharacterized reads) for presence of viral pathogens were classified as potential unrelated viral infection (Supplementary Fig. 9). These samples were used as controls during specific differential expression comparisons to compare the common effects of viral infections on host cells from SARS-CoV-2 infection.

**Viral genome assembly.** Reads unambiguously mapping to SARS-CoV or SARS-CoV-2 were aligned to the Wuhan-Hu-1 (Genbank accession MN908947.3) reference using the BBMap random-reads script and the default parameters. Variants were called using iVar, and pileup and consensus sequences were generated using samtools38–41. Any sample with >99% coverage above 10x depth for SARS-CoV-2 genome were taken as reliable samples, which resulted in 155 samples (146 positive, 9 negative). 155 clinical samples were compared to the 59,658 SARS-CoV-2 sequences from GISAID (as of June 16, 2020)42. All sequence filtering, alignments, phylogenetic inference, temporal ordering of sequences, and geographic reconstruction of likely transmission events were done using Nextstrain44,45. Nextstrain was configured to generate a New York State-focused build. Sequence identity and coverage metrics were calculated using Mview46. Phylogenetic trees were created using Nextstrain’s augur as described above, and visualized using the ggtree package in R47.

**Viral variant calling and allelic fraction estimation.** Full-length viral consensus sequences were aligned to the Wuhan-Hu-1 reference using bwa mem48 with default settings. Variants were identified by enumerating the coordinates and query/reference subsequence associated with mismatches (SNV) and gaps in the query (deletion) and reference (insertion) using R/Bioconductor (GenomicRanges, Rsamtools, Biostats packages) and gChain [https://github.com/mskclab/gChain] packages. Exhaustive variant calling on read alignments was additionally performed using bcftools mpileup and call, with variant read support (VAF, alternate allele count) enumerated with the R/Bioconductor Rsamtools package.

For each variant, a posteriori distribution of VAF was computed using a beta distribution with shape parameters \(\alpha\) and \(\beta\), where \(\alpha =\) allele counts and \(\beta =\) pseudo count of 0.5. Variants were classified as het (heterogenous) if \(P(VAF > 0.05 \cap VAF < 0.95) > 0.95\). For a given specimen, posterior VAF distributions of \(k\) heterogenous variants were then integrated using a histogram density estimator by summing the posterior VAF density across all variants at each point on a grid of \(k\) x \(k\) for each sample. This (unnormalized) mixture density was visualized alongside the individual VAF densities as an estimate of the probability density of a putative viral subclone.

**Cell deconvolution analysis.** Bulk RNAseq data were deconvolved into cell composition matrices by MUSIC algorithm38, via reference single-cell RNAseq from upper respiratory epithelium obtained from nasal brushes and upper airway and lung cells48.

**Human transcriptome analysis.** The reads that mapped unambiguously to the human reference genome via Kraken2 were used to detect the host transcriptional response to the virus. Reads matching Homo sapiens were trimmed with TrimGalore, aligned with STAR (v2.6.1d) to the human reference build GRC38 and the MUSCLE v33 transcriptome reference, gene expression was quantified using featureCounts, stringTie and salmon using the nf-core RNAseq pipeline50–53. Sample QC was reported using fastqc, RSeQC, qualimap, dupradar, Preseq, and MultiQCC6–66. Samples that had more than 10 million human mapped reads were used for differential expression analysis. Reads, as reported by feature Counts, were normalized using trimmed mean of M-values (TMM) and R for visualization purposes in log-scale61. Limma voom and DESeq2 were used to call differentially expressed with either Positive cases vs. Negative, or viral load (High/ Medium/Low/None excluding any samples with evidence of other viral infections) as reported by each qRT-PCR cycle threshold (Ct) values, or using the inverted normalized Ct value as continuous response for viral levels, where Ct of 15 is 1.0 and Ct of >40 is taken as 062. Genes with BH-adjusted \(p\) value < 0.01 and absolute log2 fold-change greater than 0.58 (at least 50% change in either direction) were taken as significantly differentially regulated53. The same approaches were repeated correcting for potential confounders in our data in two ways. In the first correction clinical metadata (as predicted by MUSIC) was added as a covariate to our model. For the second correction SVA was run on the data and the resulting two surrogate variables were included in a multivariate model54. The complete gene list for all comparisons are given in Supp Data 3. Resulting gene sets were ranked using log2 fold-change values within each comparison and put into GSEA (Supplementary Data 4) to calculate enrichment scores. We also used the MSigDB, MGI Mammalian Phenotypes database and ENCODE transcription factor binding set54–68. Any signature with adjusted \(p\) value < 0.01 and absolute normalized enrichment score (NES) ≥ 1.5 were reported (Supplementary Data 3).

**Cross-reactivity analysis.** Primers were compared with a list of sequences from organisms from the same genetic family as SARS-CoV-2 and other high-priority organisms listed in the United States Food and Drug Administration’s Emergency Use Authorization Template (https://www.fda.gov/media/135900/download). Using the sequence names in the EUA template, the NCBI taxonomic database was queried to find the highest quality representative sequences for more detailed analysis. Primers were compared to this database using Blast 2.8.1 and the following parameters (word size: 7, match score: 2, mismatch score: 5, gap extend cost: 2). Up to 1000 hits with \(e\) value > 0.1 were reported.

**Inclusivity analysis.** Unique, full-length, human-sample sequences were downloaded from the GISAID web interface. These sequences were aligned to NC_045512.2 (Wuhan SARS-CoV-2) using minimap2 -x asm5 and visually inspected using IGV 2.8.0 with allele frequency threshold set to 0.01.

**ACE inhibitor/angiotensin receptor blocker clinical cohort analysis.** We estimated the effects of ACEI/ARB exposure on COVID-19 using an observational case-control analysis of electronic health record data from two NYC hospitals (CUIMC and WCMC). Due to legal and institutional restrictions on data sharing, we performed the analysis at each site separately. We used data from clinical encounters occurring between March 10 and June 15, 2020. Our study included all individuals who received a valid SARS-CoV-2 swab test result at any time in 2020. We used two cohorts for comparison67: all individuals who received a positive SARS-CoV-2 swab test result and67 all individuals from the first cohort who were exposed to any of four anti- hypertensive drug classes (ACEI/ARB, BB, CCB, and THZ). A patient was considered exposed to one of these drug classes if they received and filled a prescription for a drug in that class at any time in 2020.

We considered two severe COVID-19 outcomes: intubation and death following confirmed infection. Individuals with a single positive test result were considered to have been infected and were included in our analysis. Intubation or death following confirmed infection was considered related to COVID-19. As we cannot determine from our data the date of infection, an individual became at risk of intubation and death due to COVID-19 7 days before their first positive test result or their first COVID-19 diagnosis. We determined mortality from a death note filed by a resident or primary provider that records the date and time of death. Intubation was used as an intermediary endpoint and is a proxy for a patient undergoing arrest and outcomes are outlined in a separate file used for patients with SARS-CoV-2 infection to record that this procedure was completed. In addition, we validated outcome data derived from notes against the patient’s medical record using manual review. As a comparison, we also performed a logistic regression using the same approach for the outcome of the infection test result itself using the full cohort of tested individuals.

We gathered data on the following 19 covariates, which were used both for propensity matching and for covariate adjustment: age, sex, race, ethnicity,
hypertension, obesity, chronic kidney disease, coronary artery disease, heart failure, myocardial infarction, diabetes mellitus, diabetic nephropathy, diabetic neuropathy, diabetic retinopathy, diabetic vasculopathy, asthma, chronic obstructive lung diseases, proteinuria, and insulin use. Individuals without demographic information (first four covariates) were excluded from the analysis. The final 15 variables represent disease history and are defined in detail in the supplement.

**Experimental statistical methods.** Using a combination of two cohorts and two severe outcomes, we made four propensity-matched, adjusted estimates of effect for ACEI/ARB. For comparison, we performed the same analysis using three other antihypertensive classes (BB, CCB, THZ). To inspect the degree of confounding due to covariates, we also estimated each effect with and without propensity matching and three levels of covariate adjustment. We used a Cox proportional hazards model for both outcomes. Propensity matching used logistic regression with the above 19 covariates using a 1:5 ratio of controls to cases in each of 100 propensity score bins, selecting none when the bin did not contain both cases and controls.

**Statistical and visualization software.** All electronic health data analyses were performed in Python 3.7 and all models were fit using R 3.6.3. Survival analyses (Cox regressions and survival curves) were performed with the survival package for R, version 3.1-12. Additional statistical analyses, processing, transformation, and visualization of genomic data were completed in R/Bioconductor (“Rsamtools”, “GenomicRanges”, and “Biostrings”) and additional Imlielinski Lab R packages (“gTrack”, “gChain”, “gUids”, and “RseqLib”) available at https://github.com/makdub. Visualization of phylogenies was completed using Auspice and the ‘giraffes’ and ‘ape’ libraries for R.

**The sequential organ failure assessment (SOFA).** The SOFA score is a severity of illness score that sums six separate organ dysfunction subscores, was used to characterize the burden of organ dysfunction. For the central nervous system, kidney, liver, and coagulation organ dysfunction subscores, traditional SOFA methodology was used, based on patient status in the first 24 h following admission to the hospital (can be whatever timeframe that you used). The respiratory SOFA subscore requires the selection of the lowest PaO2/FiO2 of the 24 h period and associates a lower ratio with a more severe dysfunction score, in situations outside of an ICU setting a PaO2 is often not available. We used a common measurement technique to implement a PaO2 from the FiO2 and PO2. The cardiovascular SOFA subscore was updated with additional vasopressors according to a norepinephrine equivalence table from a clinical trial. Specifically, phenylephrine and vasopressin were converted to a norepinephrine equivalence. Missing data for each subscore was treated as normal.

**Spatial gene expression.** RNA from FFPE lung sections was extracted with the RNeasy FFPE Kit (Qiagen). Only the samples with A260/A280 ratios of ≥1.6 were used. At least 100 nanograms of RNA was used for hybridization with the nCounter® PanCancer 10360® Gene Expression Panel spiked with the COVID-19 panel, according to the manufacturer’s instructions and quantified by the nCounter MAX Analysis System (NanoString Technologies, Seattle, WA, USA). Transcripts counts were normalized to ERCC positive controls and housekeeper reference gene expression prior to analysis.

**RNA/NGS Slide Preparation for GeoMx DSP.** For GeoMx DSP slide preparation, we followed GeoMx DSP slide prep user manual (MAN-10087-04). Briefly, tissue slides were baked in a drying oven at 60°C for 1 h and then loaded to Leica Biosystems BOND RX FFPE for deparaffinization and rehydration. After the target retrieval step, tissues were treated with Proteinase K solution to expose RNA targets followed by fixation with 10% BAE. After all tissue pre-treatments were done, tissue slides were unloaded from the Leica Biosystems BOND RX and incubated with RNA capture probe mix (CTA COVID-19 Spike-in panel). The next day, tissues were washed and stained with tissue visualization markers: CD68-647 (Novus Bio, NBP2-34736AF647), CD45-594 at 1:10 (NanoString Technologies), PanCK-532 at 1:20 (BioGenex, 352-600-AF532), CD68-647 at 1:400 (Novus Bio, NBP2-34736AF647), CD3-594 at 1:400 (Abcam, ab196147), CD45-594 at 1:10 (NanoString Technologies), and/or SYTO 13 at 1:10 (Thermo Scientific S7573). Following sequencing of sample libraries, quality control was performed on the fastq files to ensure the sequencing reads met pre-established cutoffs for number of reads and quality using FastQC (version 0.11.8)72 and MultiQC (version 1.8)72. Quality filtering and adapter trimming were performed using BBduk tools (version 38.76, https://sourceforge.net/projects/bbmap/). Remaining reads were aligned to the Ensembl GRCh38 human reference genome assembly (Release 33) using STAR (version 2.7.0f)76 and gene frequencies were counted using feature Counts (version 2.0.0) within the Subread package77. Comparative analysis of DEGs was performed using a generalized linear model (GLM) implemented in the edgeR Bioconductor package (version 3.30.3)78, using a Benjamin–Hochberg corrected p value of <0.01 (Supplementary Data 3). Hierarchical clustering of DEGs was performed in R (version 4.0.0) using the ComplexHeatmap and pheatmap package, and figures were produced using the ggplot2 package. For NP and WB, the top 100 DEGs sorted by p value with a corrected p value of <0.01 and <0.001 were included. For the comparison between hospitalized and outpatients, all the DEGs with a Bonferroni corrected p value of <0.01 were included. Clustering was performed based on Euclidean distance with complete linkage, after exclusion of noncoding genes.
Signaling pathway analyses and heatmaps were generated using Ingenuity Pathway Analysis (IPA) software (QIAGEN)\(^2\). The molecule activity predictor tool of IPA was used to predict gene upregulation or downregulation and pathway activation or inhibition. The enrichment score \( p \) value was used to evaluate the significance of the overlap between predicted and observed genes, while the \( z \)-score was used to assess the match between observed and predicted regulation or downregulation.

Classifiers were developed using scikit-learn (version 1.2.2)\(^3\) in Python. Several different classifier models were evaluated in parallel and the one with optimal performance on the training data was selected. These candidate classifier models included a Linear Support Vector Machine, Linear Discriminant Analysis, and a Deep Neural Network, all within the scikit-learn package. Reduced gene panels were selected\(^4\) and a customized reverse search across the resulting feature set. This search iteratively removed the remaining gene with the lowest significance as measured by its Lasso coefficient, performed classifier training, and reported sensitivity, specificity, and accuracy across the training set. These results were then manually reviewed to balance each of them with a priority placed on specificity and number of genes. Receiver operating characteristic (ROC) curves were generated using ROC package in R\(^5\).

**HLA analysis.** We mapped RNA reads to human reference genome GRCh38 using STAR v2.7.3a, and called likely HLA diplotypes for each sample by arcasHLA (Docker image quay.io/chi/arcas_hla:0.1.0, as pulled from master branch of https://github.com/RabadanLab/arcasHLA on 8 June 2020).

**Hybrid capture-base next-generation sequencing.** Five positive and one negative clinical samples were used to compare LAMP results with RT-qPCR and a hybrid capture-based NGS assay. RNA was isolated and purified in duplicates (250 \( \mu \)l input) using the Direct-zol DNA/RNA MiniPrep kit (Zymo). Viral transport media was used as negative extraction control. Extracted and purified RNA samples were converted to cDNA through random priming using Random Primer 6, Protoscript II First Strand cDNA Synthesis Kit, and NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Kit (New England BioLabs) The cDNA samples were converted to Illumina TruSeq-compatible libraries with Twist’s Library Preparation Kit with Unique Dual Indices (Twist Bioscience). After cDNA generation, samples were pooled in 8-plex reactions using 1875 ng of each barcoded library. Hybridization reactions were performed for 2 h using Twist Fast Hybridization reagents and SARS-CoV-2 Research Panel, a biotin-bound panel that targets libraries containing the SARS-CoV-2 virus. Hybrid capture libraries were spiked with 1% PhiX and sequenced on an Illumina NextSeq 550 platform using a NextSeq500/550 High Output kit (Illumina) set to 150bp single end reads. Extracted RNA was also tested using RT-qPCR (GenArrayisión COVID-19 duplex RT-qPCR) to get CT values, and LAMP. Hybrid capture NGS data were analyzed using the COVID-DX Software (Biotia Inc.), optimized for the SARS-CoV-2 NGS Assay. The software includes removal of low-quality reads, alignment to the SARS-CoV-2 genome (NC_045512.2), and elimination of off-target reads by alignment to other genomes including human (GrCh38) and 26 additional microbial genomes. Presence or absence of SARS-CoV-2 was determined by calculating coverage of NC_045512.2 at 1× depth using a sliding window analysis (window size 1000 and step size 100) and the enrichment score was used to evaluate the significance of the reference genome (NC_045512.2) and its coverage across the viral genome\(^6\).–\(^10\).

**Ethical approval.** Tissue samples were provided by the Weill Cornell Medicine Department of Pathology. The Tissue Procurement Facility operates under IRB approved protocol and follows guidelines set by Health Insurance Portability and Accountability Act (HIPAA). Experiments using samples from human subjects were conducted in accordance with local regulations and with the approval of the IRB at the Weill Cornell Medicine. The autopsy samples were collected under IRB protocols 20-04021814 and 19-11021-069. Autopsy consent is provided on all cases by the next of kin. This consent includes use of tissue for research. Observational cohort analysis (ACEI/ARB) was done through the Columbia University IRB Protocol AAL16001 and the Weill Cornell Medicine IRB protocol 20-04021820. Oropharyngeal samples from University Hospital of Tuebingen were collected under IRB 243/2020A.

**Data availability.** All raw sequence files (FASTQs) and metadata for specimens, including per-run metrics and QC data, have been submitted to the database of Genotypes and Phenotypes dbGAP (accession #38851 and ID phs002258.v1.p1). A total of 155 complete viral sequence reads were assembled from these data. The GenBank accession IDs for the viral genomes are MW493170–MW493863. Source data are provided with this paper. Received: 3 September 2020; Accepted: 25 January 2021; Published online: 12 March 2021

**References**

1. He, F., Deng, Y. & Li, W. Coronavirus disease 2019 (COVID-19): what we know? J. Med. Virol. 92, 719–725 (2020).
2. Fauver, J. R. et al. Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. Cell 181, 990–996 (2020).
3. Gonzalez-Reiche, A. et al. Introductions and early spread of SARS-CoV-2 in the New York City area. MedXAdv https://doi.org/10.1101/2020.04.08.20056929v1 (2020).
4. Zhao, Z., Sokhansanj, B. A. & Rosen, G. Characterizing geographical and temporal dynamics of novel coronavirus SARS-CoV-2 using informative subtype markers. BioRxiv https://doi.org/10.1101/2020.04.07.030759v3 (2020).
5. Lu, J. et al. Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell 181, 997–1003 (2020).
6. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20, P533–P543 (2020).
7. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34, 4121–4123 (2018).
8. Meyers, L. et al. Enterovirus D68 outbreak detection through a syndromic disease epidemiology network. J. Clin. Virol. 124(Mar), 104262 (2020).
9. Li, C. K. & Xiaoong, X. Host immune responses to SARS coronavirus in humans. Mol. Biol. SARS Coronavirus 22, 259–278 (2009).
10. Long, Q. X. et al. Antibody responses to SARS-CoV-2 in patients with COVID-19. Nat. Med. 26, 845–848 (2020).
11. Catanzaro, M. et al. Immune response in COVID-19: addressing a pharmacological challenge by targeting pathways triggered by SARS-CoV-2. Signal Transduct. Target. Ther. 5, 1, 84 (2020).
12. Lucas, C. et al. Longitudinal assays reveal immunological misfiring in severe COVID-19. Nature 584, 463–469 (2020).
13. Lansd, U., Lim, B., Baskaran, V. & Lim, W. S. Co-infections in people with COVID-19: a systematic review and meta-analysis. J. Infect. 81(Aug), 266–275 (2020).
14. Sui, K. L. et al. Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. FASEB J https://doi.org/10.1096/fj.201802418R (2019).
15. Koyama, T., Platt, D. & Parida, L. Variant analysis of SARS-CoV-2 genomes. Bull. World Health Organ 98(Jul), 495–504 (2020).
16. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181, 271–280 (2020).
17. Sungnak, W. et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat. Med. 26, 681–687 (2020).
18. Oudshoorn, D. et al. HERC6 is the main E3 ligase for global ISG15 conjugation in mouse cells. PLoS ONE 7, e29870 (2012).
19. Blanco-Melo, D. et al. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell 181, 1036–1045 (2020).
20. Updegraff, B. L. et al. Transmembrane protease TMPRSS2B112 promotes lung cancer growth by enhancing lactate export and glycolytic metabolism. Cell Rep 25(Nov), 2223–2233.e6 (2018).
21. Ajikia, R. S., Phillips, J. D. & Kushner, J. P. Biosynthesis of heme in mammals. Biochim. Biophys. Acta 1763(Jul), 733–736 (2006).
22. Agata, J. et al. Olmesartan is an angiotensin II receptor blocker with an inhibitory effect on angiotensin-converting enzyme. Hypertens. Res. 29, 865–874 (2006).
23. Fang, L., Karakiulakis, G. & Roth, M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? Lancet. Respir. Med. 8, e21 (2020).
24. Orenbuch, R. et al. arcasHLA: high-resolution HLA typing from RNAseq. Bioinformatics 36, 33–40 (2020).
25. Mackay, M. J. et al. The COVID-19 XPRIZE and the need for scalable, fast, and widespread testing. Nat. Biotechnol. 38, 1021–1024 (2020).
26. Montell, V. et al. Inhibition of SARS-CoV-2 Infections in Engineered Human Tissues Using Clinical-Grade Soluble Human ACE2. Cell. 181, 905–913 (2020).
27. Ferrario, C. M. et al. Effect of angiotensin-converting enzyme inhibition and angiotensin II receptor blockers on cardiac angiotensin-converting enzyme 2. Circulation 111, 2605–2610 (2005).
28. Zhang, P. et al. Association of inpatient use of angiotensin converting enzyme inhibitors and angiotensin II receptor blockers with mortality among patients with hypertension hospitalized with COVID-19. Circ. Res. 126, 1671–1681 (2020).
89. Jenkins, C. & Osburn, B. In silico approach to accelerate the development of mass spectrometry-based proteomics methods for detection of viral proteins: application to COVID-19. bioRxiv https://www.biorxiv.org/content/10.1101/2020.03.08.988348v2 (2020).

90. Konrad, R. et al. Rapid establishment of laboratory diagnostics for the novel coronavirus SARS-CoV-2 in Bavaria, Germany, February 2020. Eurosurveillance 25, 2000173 (2020).

91. Kim, D. et al. The architecture of SARS-CoV-2 transcriptome. Science 368, 489–493 (2020).

92. Narayanan, K., Ramirez, S. I., Lokugamage, K. G. & Makino, S. Coronavirus Pionzio, Melanie Robinson, and Madison Robison. We thank AWS for sponsoring storage from Nanostring who supported the spatial transcriptomics analysis. The authors wish to thank Eileen Dimalanta and Ted Davis for technical discussions. We would like to thank Y.K. Biolabs for providing the reagents for preliminary testing of the LAMP protocols, as well as B.W.L. for their contribution to the peer review of this work. Peer reviewer reports are available.

93. Lee, J. S. et al. Immunophenotyping of COVID-19 and interferon nonstructural protein 1: Common and distinct functions in the regulation of virus replication and IFN-stimulated gene expression. Cell. Immunol. 370, 106–116 (2020).

94. Lauer, S. A. et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann. Intern. Med. 172, 515–524 (2020).

95. Lee, J. S. et al. Interferon-related genes in COVID-19 patients. Gene 773, 146069 (2021).

96. Littman, D. R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science 368, 489–493 (2020).

97. Narayan, K., Ramirez, S. I., Lokugamage, K. G. & Makino, S. Coronavirus nonstructural protein 1: Common and distinct functions in the regulation of host and viral gene expression. Virus Res 202(April), 89–100 (2015).

98. Patel, A. B. & Verma, A. COVID-19 and angiotensin-converting enzyme inhibitors and angiotensin receptor blockers. What is the evidence? JAMA Cardiol. 5, 1769–1779 (2020).

99. Quilty, B., Cliftord, S., Flasche, S. & Eggo, R. M. Effectiveness of airport screening at detecting travellers infected with 2019-nCoV. MedRxiv https://doi.org/10.1101/2020.01.31.20019265v1 (2020).

100. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28(15 September), i333–i339 (2012).

101. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).

102. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).

103. Rodríguez-Santana, J., Burchard, E. G. & Seibold, M. A. Type 2 and interferon inflammation strongly regulate SARS-CoV-2 related gene expression in the airway epithelium. Biorxiv https://www.biorxiv.org/content/10.1101/2020.04.09.034454v1 (2020).

104. Sanantia, J. L. et al. Transmission potential of SARS-CoV-2 in viral shedding observed at the University of Nebraska Medical Center. MedRxiv https://doi.org/10.1101/2020.03.23.20039446v2 (2020).

105. Sajuthi, S. P. et al. Type 2 and interferon inflammation regulate SARS-CoV-2 entry factor expression in the airway epithelium. Nat Commun. 11, 3139 https://doi.org/10.1038/s41467-020-18783-2 (2020).

106. Taiaroa, G. et al. Direct RNA sequencing and early evolution of SARS-CoV-2. BioRxiv https://doi.org/10.1101/2020.03.05.976167v1 (2020).

107. Tang, X. et al. On the origin and continuing evolution of SARS-CoV-2. Natl. Sci. Rev. 7, 1012–1023 (2020).

108. Chen, Y. & Lanjuan, L. SARS-CoV-2: virus dynamics and host response. Lancet Infect. Dis. 20(5), 515–516 (2020).

Acknowledgements
We thank New York-Presbyterian Hospital, the Clinical Translational Science Center (UL1 TR00457), the Joint Clinical Trials Office, the Core Facilities at Weill Cornell Medicine, the Clinical Laboratories at New York-Presbyterian Hospital, the Scientific Computing Unit (SCU), OneCodex, the XSEDE Supercomputing Resources and the GIAD Initiative curators and submitters (Supp Data 7). We also thank New England Biolabs for providing the reagents for preliminary testing of the LAMP protocols, as well as Eileen Dimalanta and Ted Davis for technical discussions. We would like to thank Y.K. from Nanostring who supported the spatial transcriptomics analysis. The authors wish to thank the following members of the HudsonAlpha Discovery team who supported the RNASeq experiments described in the paper: Colleen Cowan, John Mote, Arianna Pionzio, Melanie Robinson, and Madison Robison. We thank AWS for sponsoring storage on DNAnexus. We are grateful for support from Cynthia Polsky, the STARR Foundation (I13-0052) the Vallerie Foundation, the WorldQuant Foundation, The Pershing Square Sohn Cancer Research Alliance, Citadel, the National Institutes of Health (R01MH117406, R25EBO20934, R01AI151059), the Bill and Melinda Gates Foundation (OPP1151094), the NSF (1840275), the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR000457, CTSC), the Intramural Research Program of the National Library of Medicine, NIH, and the Alfred P. Sloan Foundation (G-2015-13964). F.I.S. supported by the National Institute of Allergy and Infectious Diseases (U19AI144297-01). M.Z. supported by T15LM007079. N.P.T and UOG supported by R35GM131905. N.A.L was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number TL1TR002386 and DB, DP and TI were supported by a grant from the BadenWürttemberg Ministry of Science and Art. IH is supported by IR35GM138152-01.

Author contributions
C.E.M. led the study design and coordination and H.R. led the clinical collection and validation work in the NYP CLIA laboratory, as well as M.I. Overall supervision and protocol development and implementation for the Zymo RNAclean and NEB assays (S.L.). D.B. and C.M.Z. performed the LAMP experiments to validate the assay, established a method to quantify LAMP output, and developed a protocol for clinical use of the assay. C.W., D.X., P.R., J.G., L.C., and I.X. assisted with sample preparation and sequencing. D.B., A.I., D.P., and T.I. developed and validated the testing from raw oropharyngeal swabs. C.E.M., D.D., J.F., M.I., A.S., J.R., M.M., E.A., I.H., D.M., T.M., J.T.M., B.W., M.Z., U.G., N.P.T., N.A.L., A.F., C.S.C., F.S., N.L., M.S., D.P., M.Z.V.R., E.S., E.S., C.H., A.G., Y.S., V.S., F., A.F., C.S.C., N.P., and N.T. performed analyses. D.D., C.M.Z., N.A.L., M.S., B.Y., K.R., C.B. coordinated and collected samples. E.A. submitted the IRB application and helped with clinical coordination. E.A., S.L., M.I., M.S., L.F.W., M.L., M.C., H.R., K.R., C.S.C., N.M.P., B.L., N.T., Y.K., J.R., T.H., S.W., M.B., D.M., M.C.R., D.N.S., J.B., H.W., N.O.H, J.R., Y.G., P.A.S., A.S., J.T.M., D.T.M., A.I., E.S., T.R.C., J.S., A.C., P.Y., A.M., S.S., I.H., A.B., T.L., M.G., S.W., C.G., N.P.T., U.O.G., and R.S. helped plan, refine, and interpret analyses. All authors reviewed, edited, and approved the paper.

Competing interests
N.T. and B.W.L. are employees at New England Biolabs. R.E.S. is on the scientific advisory board of Mirromatrix Inc. Biotia employees and advisors include M.C.R., D.N.S., J.B., H.W., C.E.M., and N.O.D. D.B. is a co-founder of Poppy Inc. A.F and C.S.C are employees of DNAnexus. Authors not listed here do not have competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21361-7.

Correspondence and requests for materials should be addressed to R.E.S., N.T., H.R., M.I. or C.E.M.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/. © The Author(s) 2021
Daniel Butler\textsuperscript{1,3,4}, Christopher Mozsary\textsuperscript{1,3,4}, Cem Meydan\textsuperscript{1,2,3,4}, Jonathan Fox\textsuperscript{1,2,3,4}, Joel Rosiene\textsuperscript{4,5,34}, Alon Shaiber\textsuperscript{4,6,34}, David Danko\textsuperscript{1,34}, Ebrahim Afshinnekoo\textsuperscript{1,2,3}, Matthew MacKay\textsuperscript{1}, Fritz J. Sedlacek\textsuperscript{7}, Nikolay A. Ivanov\textsuperscript{1,2,8}, Maria Sierra\textsuperscript{1,2}, Diana Pohle\textsuperscript{9}, Michael Zietz\textsuperscript{10}, Undina Gisladottir\textsuperscript{10}, Vijendra Ramrall\textsuperscript{10,11}, Evan T. Sholle\textsuperscript{12}, Edward J. Schenck\textsuperscript{13}, Craig D. Westover\textsuperscript{1}, Ciaran Hassan\textsuperscript{1}, Krista Ryon\textsuperscript{1}, Benjamin Young\textsuperscript{1}, Chandrima Bhattacharya\textsuperscript{1}, Dianna L. Ng\textsuperscript{14}, Andrea C. Granados\textsuperscript{14,15}, Yale A. Santos\textsuperscript{14,15}, Venice Servellita\textsuperscript{14,15}, Scot Federman\textsuperscript{14,15}, Phyllis Ruggiero\textsuperscript{5}, Arkarachai Fungtammasan\textsuperscript{16}, Chen-Shan Chin\textsuperscript{16}, Nathaniel M. Pearson\textsuperscript{17}, Bradley W. Langhorst\textsuperscript{18}, Nathan A. Tanner\textsuperscript{18}, Youngmi Kim\textsuperscript{19}, Jason W. Reeves\textsuperscript{19}, Tyler D. Hether\textsuperscript{19}, Sarah E. Warren\textsuperscript{19}, Michael Bailey\textsuperscript{19}, Justyna Gawryś\textsuperscript{5}, Dmitry Meleshko\textsuperscript{1,20}, Dong Xu\textsuperscript{21}, Mara Couto-Rodriguez\textsuperscript{22}, Dorottya Nagy-Szakal\textsuperscript{22,23}, Joseph Barrows\textsuperscript{22}, Heather Wells\textsuperscript{22}, Niamh B. O’Hara\textsuperscript{22,23}, Jeffrey A. Rosenfeld\textsuperscript{24,25}, Ying Chen\textsuperscript{24}, Peter A. D. Steel\textsuperscript{26}, Amos J. Shemesh\textsuperscript{26}, Jenny Xiang\textsuperscript{21}, Jean Thierry-Mieg\textsuperscript{27}, Danielle Thierry-Mieg\textsuperscript{27}, Angelika Iftner\textsuperscript{9}, Daniela Bezdan\textsuperscript{9}, Elizabeth Sanchez\textsuperscript{13}, Thomas R. Campion Jr.\textsuperscript{12,28}, John Sipley\textsuperscript{5}, Lin Cong\textsuperscript{5}, Aryrn Crane\textsuperscript{5}, Priya Velu\textsuperscript{5}, Ari M. Melnick\textsuperscript{13}, Sagi Shapira\textsuperscript{10}, Iman Hajirasouliha\textsuperscript{1,2,6}, Alain Borczuk\textsuperscript{13}, Thomas Iftner\textsuperscript{9}, Mirella Salvatore\textsuperscript{13,28}, Massimo Loda\textsuperscript{5}, Lars F. Westblade\textsuperscript{5,13}, Melissa Cushing\textsuperscript{5}, Shixiu Wu\textsuperscript{29,30}, Shawn Levy\textsuperscript{31}, Charles Chiu\textsuperscript{14,15,32}, Robert E. Schwartz\textsuperscript{13,33}, Nicholas Tattonetti\textsuperscript{10,33}, Hanna Rennert\textsuperscript{5}\textsuperscript{35}, Marcin Imielinski\textsuperscript{4,5,6}\textsuperscript{35} & Christopher E. Mason\textsuperscript{1,2,3,33}\textsuperscript{36}

\textsuperscript{1}Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{2}The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{3}WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{4}New York Genome Center, New York, NY, USA. \textsuperscript{5}Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{6}Englander Institute for Precision Medicine and the Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{7}Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. \textsuperscript{8}Clinical & Translational Science Center, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{9}Institute of Medical Virology and Epidemiology of Viral Diseases, University Hospital Tübingen, Tübingen, Germany. \textsuperscript{10}Department of Biomedical Informatics, Department of Systems Biology, Department of Medicine, Institute for Genomic Medicine, Columbia University, New York, NY, USA. \textsuperscript{11}Department of Cellular, Molecular Physiology & Biophysics, Columbia University, New York, NY, USA. \textsuperscript{12}Information Technologies & Services Department, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{13}Reflexion in Biomedical Informatics, Department of Medicine, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{14}Department of Laboratory Medicine, University of California, San Francisco, CA, USA. \textsuperscript{15}UCSF—Abbott Viral Diagnostics and Discovery Center, San Francisco, CA, USA. \textsuperscript{16}DNA nexus, Inc., Mountain View, CA, USA. \textsuperscript{17}Root Deep Insight, Boston, MA, USA. \textsuperscript{18}New England Biolabs, Ipswich, MA, USA. \textsuperscript{19}NanoString Technologies, Seattle, WA, USA. \textsuperscript{20}Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{21}Genomics Resources Core Facility, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{22}Department of Biomedical Information Science, SUNY Downstate Health Sciences University, New York, NY, USA. \textsuperscript{23}Rutgers Cancer Institute of New Jersey, New York, NJ, USA. \textsuperscript{24}Department of Pathology, Robert Wood Johnson Medical School, New York, NJ, USA. \textsuperscript{25}Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{26}National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. \textsuperscript{27}Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{28}Hangzhou Cancer Institute, Hangzhou, China. \textsuperscript{29}Department of Radiation Oncology, Hangzhou Cancer Hospital, Hangzhou, China. \textsuperscript{30}The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. \textsuperscript{31}These authors contributed equally: Daniel Butler, Christopher Mozsary, Cem Meydan, Jonathan Fox, Joel Rosiene, Alon Shaiber, David Danko. \textsuperscript{32}email: res2025@med.cornell.edu; npt2105@cumc.columbia.edu; har2006@med.cornell.edu; mai9037@med.cornell.edu; chm2042@med.cornell.edu