scientific reports

Check for updates

OPEN

Deep RNA sequencing of intensive care unit patients with COVID-19

Alger M. Fredericks[1], Maximilian S. Jentzsch[1], William G. Cioffi[1], Maya Cohen[2], William G. Fairbrother[3], Shivam J. Gandhi[3], Elizabeth O. Harrington[3], Gerard J. Nau[4], Jonathan S. Reichner[1], Corey E. Ventetuolo[2], Mitchell M. Levy[2], Alfred Ayala[1] & Sean F. Monaghan[1]✉

COVID-19 has impacted millions of patients across the world. Molecular testing occurring now identifies the presence of the virus at the sampling site: nasopharynx, nares, or oral cavity. RNA sequencing has the potential to establish both the presence of the virus and define the host's response in COVID-19. Single center, prospective study of patients with COVID-19 admitted to the intensive care unit where deep RNA sequencing (> 100 million reads) of peripheral blood with computational biology analysis was done. All patients had positive SARS-CoV-2 PCR. Clinical data was prospectively collected. We enrolled fifteen patients at a single hospital. Patients were critically ill with a mortality of 47% and 67% were on a ventilator. All the patients had the SARS-CoV-2 RNA identified in the blood in addition to RNA from other viruses, bacteria, and archaea. The expression of many immune modulating genes, including PD-L1 and PD-L2, were significantly different in patients who died from COVID-19. Some proteins were influenced by alternative transcription and splicing events, as seen in HLA-C, HLA-E, NRP1 and NRP2. Entropy calculated from alternative RNA splicing and transcription start/end predicted mortality in these patients. Current upper respiratory tract testing for COVID-19 only determines if the virus is present. Deep RNA sequencing with appropriate computational biology may provide important prognostic information and point to therapeutic foci to be precisely targeted in future studies.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing coronavirus disease 2019 (COVID-19) has led to millions of cases worldwide[1]. Current testing is by polymerase chain reaction to detect viral RNA in the nares[2], but provides no insight into the host response. Patients with COVID-19 that require intensive care unit (ICU) care are sick and difficult to manage, thus, there is a need for other diagnostic tests during the hospital stay to assist the clinicians.

Deep RNA sequencing refers to a process of sequencing where (at least) 100 million reads of sequence are generated per sample. Deep sequencing allows for the study of low abundance RNA and biologic processes beyond gene expression. Typically, RNA sequencing data is aligned to the genome of interest, such as aligning to human genes when the sample comes from a human. Reads that do not align to the genome of interest are usually discarded. When the RNA sequencing is performed with this large number of reads, it could be used to identify the presence of specific pathogens in the blood by aligning the reads that would have been discarded to other genomes of interest. In COVID-19, sequencing reads of SARS-CoV-2 may provide insight into the biology of the virus during active illness. In addition, secondary infections could be identified, potentially allowing for better, pathogen-directed antibiotic treatment.

The host response to the virus is responsible for some of the morbidity and mortality observed[3]. Acute respiratory distress syndrome (ARDS) is the most common complication encountered with COVID-19[3]. Our laboratory has shown that there are significant changes in alternative RNA splicing and transcription start and end in ARDS as assessed by deep RNA sequencing[4]. These changes are thought to be due to the physiology of ARDS, e.g., hypoxia and acidosis, which are known to influence splicing. Whether this occurs in patients infected by COVID-19 is not known.

[1]Division of Surgical Research, Department of Surgery, Alpert Medical School of Brown University/Rhode Island Hospital, 593 Eddy Street, Middle House 211, Providence, RI 02903, USA. [2]Division of Pulmonary, Critical Care, and Sleep Medicine, Department of Medicine, Alpert Medical School of Brown University/Rhode Island Hospital, Providence, USA. [3]Brown University, Providence, USA. [4]Division of Infectious Disease, Department of Medicine, Alpert Medical School of Brown University /Rhode Island Hospital, Providence, USA. ✉email: smonaghan@lifespan.org

While RNA sequencing can be used to measure immune modulating gene expression, an alternative approach is the evaluation of global entropy, or disorder in the processing of RNA[5]. In this study, we propose that this entropy metric combined with Principal Component Analysis (PCA) can be leveraged to distinguish COVID-19 patients that develop life-threatening illness from those likely to recover.

Here we examine deep RNA sequencing data from patients in the ICU with COVID-19 to characterize both pathogens and host responses. We evaluate the sequences for the presence of the SARS-CoV-2 virus and other potential infectious agents. The host response to COVID-19 is also characterized. The long-term goal is to combine these measurements to better assist clinical decision-making.

Methods

Study design, population and setting. The study enrolled ICU participants at a single tertiary care hospital evidence of SARS-CoV-2 infection based on positive PCR from the nasopharynx documented during admission. All participants, or their appropriate surrogate, provided informed consent as approved by the Rhode Island Hospital Institutional Review Board (Approval #: 411616) and all methods were carried out in accordance with relevant guidelines and regulations. Blood samples were collected on day 0 of ICU admission. Clinical data including COVID specific therapies was collected prospectively from the electronic medical record and participants were followed until hospital discharge or death. Ordinal scale was collected as previously described[6]; along with sepsis and associated SOFA score[7] and the diagnosis of ARDS[8].

RNA extraction and sequencing. Whole blood was collected in PAXgene tubes (Qiagen, Germantown, MD) and sent to Genewiz (South Plainfield, NJ) for RNA extraction, ribosomal RNA depletion and sequencing. Sequencing was done on Illumina HiSeq machines to provide 150 base pair, paired-end reads. Libraries were made by standard approaches by Genewiz using both the Globin Zero Gold (Illumina) and NEBNext Ultra RNA Library Prep Kit (New England Biolabs)[9]. Libraries were prepared to have three samples per lane. Each lane provided 350 million reads ensuring each sample had > 100 million reads. Raw data was returned on password protected external hard drives to ensure the security of the genomic data.

Computational biology and statistical analysis. All computational analysis was done blinded to the clinical data. The data was assessed for quality control using FastQC[10]. RNA sequencing data was aligned to the human genome utilizing the STAR aligner[11]. Reads that aligned to the human genome were separated and are now referred to as 'mapped' reads. Reads that did not align to the human genome, which are typically discarded during standard RNA sequencing analysis, were kept and identified as 'unmapped' reads. The unmapped reads then aligned to the SARS-CoV-2 genome (NC_045512) and counted per sample using Magic-BLAST[12]. In addition, a coverage map of the SARS-CoV-2 genome was generated using all the subjects to identify the gene expression patterns of the virus in critically ill COVID-19 patients. The unmapped reads were further analyzed with Kraken2[13] using the PlusPFP index[14] to identify other bacterial, fungal, archaeal and viral pathogens.

Reads that aligned to the human genome, the mapped reads, also underwent analysis for gene expression via edgeR and alternative RNA splicing/alternative transcription start/end via Whippet[5]. When comparisons were made between groups (died vs. survived) differential gene expression was set with thresholds of both $p < 0.05$ and ± 1.5 $\log_2$ fold change. p values were measured by digital gene expression (DGE). DGE utilizes the negative binomial distribution as model to compare over dispersion across the dataset. Values are subsequently adjusted using a quantile maximum likelihood estimator to establish a Fischer's exact test with improved performance, separating biological from technical variation. All statistical methods done are embedded in the Whippet software. Alternative splicing was defined as core exon, alternative acceptor splice site, alternative donor splice site, retained intron, alternative first exon and alternative last exon. Alternative transcription start/end events were defined as tandem transcription start site and tandem alternative polyadenylation site. Alternative RNA splicing and alternative transcription start/end events were also compared between groups[5]. Significance was set at great than 2 $\log_2$ fold change as previously described[4]. Genes identified from the analysis of mapped reads were then evaluated by GO enrichment analysis (PANTHER Overrepresentation released 20200728)[15].

Whippet was also used to generate an entropy value for every identified alternative splicing and transcription event of each gene. These entropy values are created without the need for groups used in the gene expression analysis. In order to visualize this data a principal component analysis (PCA) was conducted to reduce the dimensionality of the dataset and to obtain an unsupervised overview of trends in entropy values among the samples. Raw entropy values from all samples were concatenated into one matrix and missing values were replaced with column means. Mortality was then overlaid onto the PCA plot to assess the ability of these raw entropy values to predict this outcome in this sample set. This analysis was done in R[16].

Ethics approval. Institutional Review Board Approval # 411616.

Results

Study population, participant characteristics, and RNA sequencing. Fifteen participants were enrolled and had blood samples drawn on the first day of their ICU stay. Clinical and demographic data is reported in Table 1. The majority of participants were male (73%) and there were a diverse distribution in terms of race (60% not white) and ethnicity (60% Hispanic). The most common co-morbidity was hypertension and the median BMI was almost 30. Forty percent of participants had ARDS at the time the samples was drawn and the patients were distributed across the top of the ordinal scale[6] with a score of 5 as the most common in 53% of the patients. Most participants required a ventilator (67%) and 20% progressed to extracorporeal membrane

	All patients (n = 15)	Alive (n = 8)	Died (n = 7)
Median age (IQR)	66.2 (41.6–72.1)	53.1 (34.5–72.8)	66.2 (50.0–73.3)
Sex			
Female—no. (%)	4 (27)	2 (25)	2 (29)
Male—no. (%)	11 (73)	6 (75)	5 (71)
Race			
White—no. (%)	6 (40)	2 (25)	4 (57)
Black or African American—no. (%)	2 (13)	1 (13)	1 (14)
Other race—no. (%)	7 (47)	5 (62)	2 (29)
Ethnicity			
Hispanic—no. (%)	9 (60)	4 (50)	5 (71)
Not Hispanic—no. (%)	6 (40)	4 (50)	2 (29)
Chronic underlying conditions—no. (%)			
Hypertension	9 (60)	4 (50)	5 (71)
Cardiovascular disease	2 (13)	0 (0)	2 (29)
Congestive heart failure	2 (13)	1 (13)	1 (14)
Diabetes	1 (7)	0 (0)	1 (14)
COPD	1 (7)	0 (0)	1 (14)
Renal failure	1 (7)	0 (0)	1 (14)
Malignancy	1 (7)	1 (13)	0 (0)
Liver disease	0 (0)	0 (0)	0 (0)
Median BMI (IQR)	29. 6 (27.3–31.9)	27.7 (24.8–31.9)	30.3 (29.5–35)
At the time of sample			
ARDS—no. (%)	6 (40)	3 (38)	3 (43)
Vassopressor support—no. (%)	2 (13)	1 (13)	1 (14)
Median SOFA score (IQR)	6 (3–8)	7.5 (4–8)	3 (2–6)
Median APACHE II score (IQR)	18 (13.5–23)	20.5 (17.3–24.5)	12 (9–22)
Ordinal scale 4—no. (%)	3 (20)	3 (38)	0(0)
Ordinal scale 5—no. (%)	8 (53)	1 (13)	7 (100)
Ordinal scale 6—no. (%)	1 (7)	1 (13)	0 (0)
Ordinal scale 7—no. (%)	3 (20)	3 (38)	0 (0)
Hospital course			
Median ventilator days (IQR)	6 (0–30)	7.5 (0–45.75)	2 (0–27)
Median ICU length of stay (IQR)	15 (6.5–35.5)	9 (3.5–36.25)	18 (8–44)
Median hospital length of stay (IQR)	22 (13–49.5)	25 (8.75–71.25)	22 (15–45)
ECMO—no. (%)	3 (20)	2 (25)	1 (14)
Acute renal replacement—no. (%)	4 (27)	2 (25)	2 (29)
Thrombotic event—no. (%)	6 (40)	3 (38)	3 (43)
Death—no. (%)	7 (47)		
Discharge from hospital—no. (%)	8 (53)		
Median D dimer (IQR)	2980 (974–4312)	2657 (885–4433)	3370 (923–7180)
TEG hypercoagulable—no. (%)	6 (40)	3 (38)	3 (43)
Remdesivir—no. (%)	8 (53)	4 (50)	4 (57)
Plasma—no. (%)	2 (13)	1 (13)	1 (14)

Table 1. Demographics.

oxygenation (ECMO); 27% required renal replacement. The median length of hospital stay was 22 days with a mortality rate of 47%.

All samples had sufficient RNA and RNA integrity numbers (RIN)[17] were adequate. The median of sequencing was 125,687,784 reads (95% CI 122,164,763 to 135,800,242) and greater than 90% of those reads were more than thirty bases. After using FastQC[10], all samples had mean quality scores over 30. The reads mapped to the human genome 62–66% of the time (Supplemental Table S1).

Identification of SARS-CoV-2 and other pathogens. Among the fifteen participant samples all participants had SARS-CoV-2 RNA detected. There were a total of 676 reads that align to the SARS-CoV-2 genome with each patient having between 18 and 98 reads. (Fig. 1a) The majority of the reads corresponded to the RNA dependent RNA polymerase and N protein genes (Fig. 1b). RNA from other pathogens including bacteria, viruses and archaea were identified in the blood of all patients (Table 2). Two participants had fungal RNA iden-

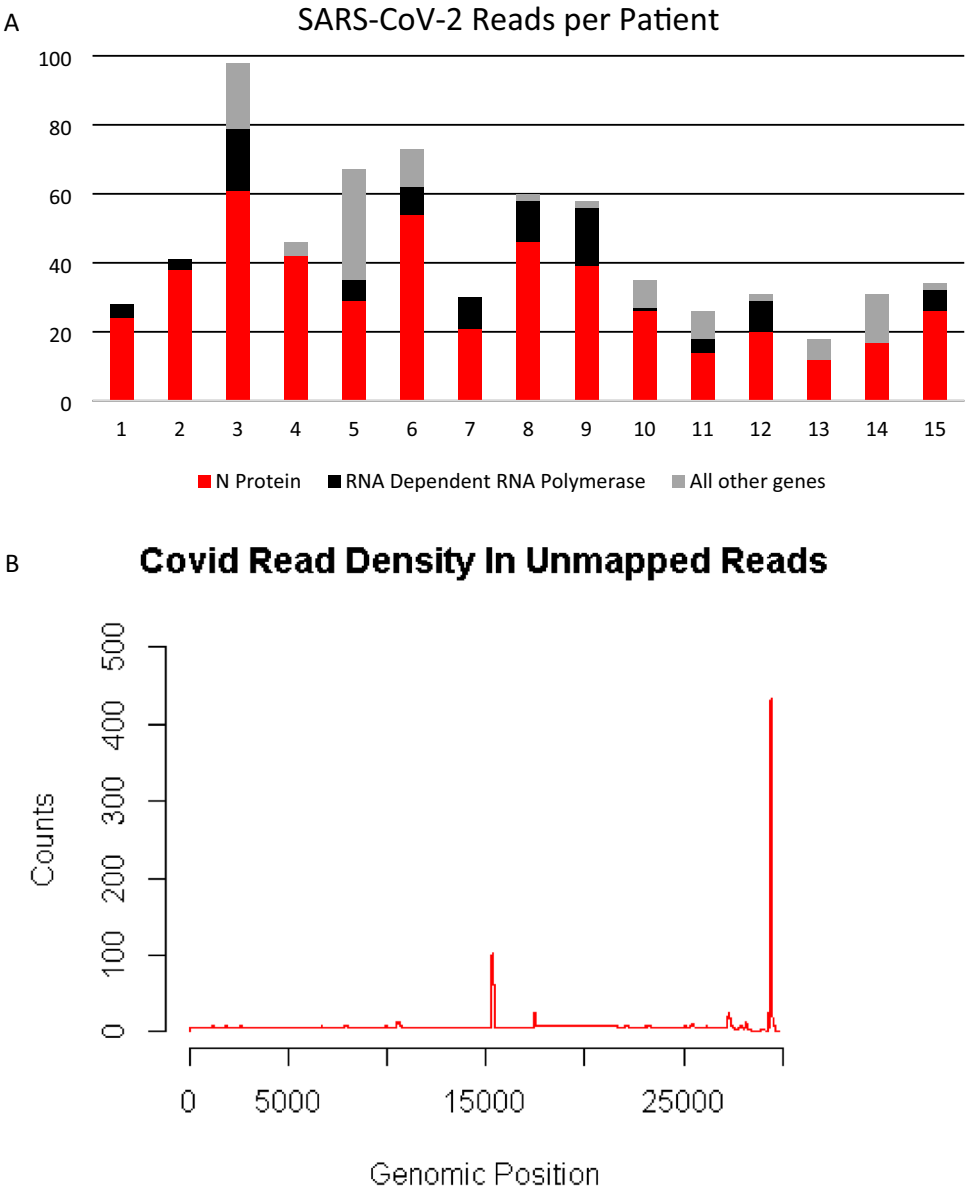

Figure 1. (**A**) Top panel is the number of reads aligning to the SARS-CoV-2 genome from each patient. Most reads aligned to loci encoding the N protein (red bar) or the RNA dependent RNA polymerase (black bar). (**B**) Bottom panel is the location where the cumulative reads from all the patients align to the SARS-CoV-2 genome. Genes encoding the RNA dependent RNA polymerase and the N protein are at positions ~ 15,000 and ~ 29,000, respectively.

tified (not described in Table 2). Despite alignment to a robust database of organisms, each participant still had hundreds of thousands of unclassified reads. (Table 2) The taxonomy classification of "other sequences" (28384) align to elements of cellular organisms (bacterial, archaea, plant), but do not have enough specificity to identify a single species are listed (Table 2). The top ten bacterial reads by count at the species level for each patient is listed in Supplemental Table 2. The top bacterial sequences from all patients were from either *Acinetobacter baumannii* or *Chryseobacterium gallinarum*. In patients who had the most counts of *C. gallinarum, A. baumannii* had significantly reduced counts compared to the counts in other patients (148.1 vs. 50,905.3, $p < 0.05$). Although sequences corresponding to *A. baumannii or C. gallinarum* were found in all patients, none of the patients had positive blood cultures drawn around the time of these samples. No counts of bacteria, virus, archaea, (Table 2) or specific bacteria (Supplemental Table S2) correlated with mortality.

Genomic differences between participants who lived and those who died. Among participants who died there were 86 genes that increased in expression and 207 that decreased in expression (top results in Table 3, full list in Supplemental Table S3, Supplemental Fig. S1). There were 88 significant alternative splicing events occurring in 84 unique genes (Top results Table 3, full list Supplemental Table S4) and 2093 alterna-

	Unclassified	Bacteria	Archaea	Other sequences	Virus
Patient 1	166,039	252,424	160	388,924	317
Patient 2	219,873	213,967	167	325,150	398
Patient 3	216,633	258,383	286	463,595	621
Patient 4	272,019	122,270	230	637,551	374
Patient 5	266,733	119,180	224	552,125	363
Patient 6	217,155	179,383	292	506,456	564
Patient 7	144,690	111,078	153	321,752	302
Patient 8	262,426	194,038	310	557,825	586
Patient 9	294,609	325,558	1055	475,941	768
Patient 10	273,603	272,536	211	373,893	403
Patient 11	222,280	175,230	169	333,625	368
Patient 12	284,819	113,546	223	615,463	271
Patient 13	308,700	103,060	238	469,993	298
Patient 14	235,961	109,451	183	485,807	449
Patient 15	179,668	130,323	188	419,667	353

Table 2. Counts per patient from Kraken2.

Gene expression	Alternative splicing	Alternative transcription
Top 5 increase (gene (fold change)) in patients that died versus lived		
NPC1L1 (8.1)	ITGB2 (3.7)	NCF2 (8.7)
VWA3B (5.7)	DPYD (3.6)	STX3 (8.4)
ABCB11 (5.5)	SMCHD1 (3.4)	YWHAB (8.2)
OR6C4 (5.1)	NF2 (2.9)	MAP2K6 (7.8)
UGT2A3 (4.7)	LRRK2 (2.9)	N4BP1 (7.3)
Gene expression	**Alternative splicing**	**Alternative transcription**
Top 5 decrease (gene (fold change)) in patients that died versus lived		
ABCA4 (−8.4)	VPS13B (−3.8)	STEAP4 (−10.6)
GJB6 (−7.3)	STRN4 (−3.3)	IQGAP1 (−8.7)
CACNA2D (−7.2)	HLA-C (−3.2)	TAP2 (−8.6)
RNF17 (−6.2)	FMR1 (−3.2)	PRRC2B (−8.4)
CTCFL (−6.1)	HLA-E (−3.1)	SNX10 (−8.4)
Gene expression + alternative splicing	**Gene expression + alternative transcription**	**Alternative splicing + alternative transcription**
Significant in both categories in patients that died versus lived		
ABCA13	ABCB11, ADARB2, ADGRD1, ASIC1, ATP2C2, CCL2, CNR1, CNTNAP2, DLG2, EPS8, FAM107A, GBP1, GBP5, HESX1, KCNMA1, MPP2, MRO, MSR1, NRP1, NRP2, PRLR, SLC1A, SLC1A3, STAC, STEAP3, TGFBI, VWA3B	ANKMY1, CCDC32, CXCR2, DUSP18, FAM214B, FGF11, FMR1, IQGAP1, KDM5D, KLC4, LGALS3BP, MBNL1, NF2, SLC24A1, TAF1C, TCF3, ZDHHC24, ZSCAN18
Gene expression	**Alternative splicing**	**Alternative transcription**
Top 5 Gene Ontology (GO) term analysis (pathway (fold change)—contributing genes if less than 10)		
Protein Homotrimerization (19.8)—SLC1A2, ASIC1, HLA-G, SCARA5 Negative regulation of interleukin-10 Production (16.5)—MMP8, CD274, PDCD1LG2, IL23R Protein trimerization (16.5)—SLC1A2, ASIC1, HLA-G, SCARA5 Retinal ganglion cell axon guidance (16.5)—NRP1, NRCAM, EPHB2, POU4F3 Axonal fasciculation (12.9)—NRP1, NRCAM, EPHB2, CNR1	None	RNA Polymerase I Preinitiation Complex Assembly (7.9)—TAF1C, BAZ2A, PIH1D1, SMARCA4, UBTF, SMARCB1 Regulation of RNA splicing (2.5)—34 genes Regulation of mRNA Splicing, via spliceosome (2.3)—23 genes Regulation of mRNA Processing (2.2)—29 genes Purine containing compound biosynthetic process (2.1)—32 genes

Table 3. Gene difference between patients that died versus lived.

tive transcription events occurring in 1769 unique genes (Top results Table 3, full list Supplemental Table S5). ABCA13 was the only gene that had significant expression and alternative splicing events. Twenty-seven genes had significant expression and alternative transcription start/end differences (Table 3). Eighteen genes had significantly different alternative splicing and alternative transcription start/end (Table 3).

The genes that were significant between groups then underwent GO term analysis to assess significant enrichment for a biological process. The top GO terms for gene expression and alternative transcription are listed in

Table 3 (full list Supplemental Tables S6, S7). There were no significant GO terms for the genes impacted by alternative splicing.

RNA entropy as a diagnostic tool. From the over 100 million RNA sequencing reads for each participant, computational analysis via Whippet assigns an entropy value for over 380,000 RNA splicing events and alternative transcription start/end events. Principal component analysis was then applied to these > 380,000 entropy scores for each of the 15 participants and the first two principal components were plotted against each other (Fig. 2). The sample points were then labeled based on their survival status. Survival status was not part of the principal component analysis itself. Participants whose PC2 value was above 0.00 had a mortality rate of 75% (6/8), up from the total group mortality of 46% (7/15) and significantly more than the 14% for those who land below that line (1/7, $p = 0.04$).

Discussion

This project used deep RNA sequencing of whole blood from participants in the ICU with COVID-19 as a novel diagnostic tool. The protocol extracted RNA from the whole blood, as opposed to fractionating the whole blood specimen. Analysis of whole blood increased the breadth of RNA being sequenced, both cell associated and cell-free, and its simplicity for clinical practice. Alternatively, more complicated techniques, such as single cell sequencing may speak more to pathogenesis but adds to the complexity of the protocol and analysis. Despite its isolation from whole blood, the RNA was of high quality (Supplement Table S1). A novel finding using RNA from whole blood from critically ill participants is that only 62–67% of the reads mapped to the human genome. This is less than the 85–97% of reads that typically map to the reference genome[18]. One major drawback is the timing needed for RNA sequencing and analysis. Currently, sequencing machines take ~ 18 h to generate data. The analysis can take additional time and is not yet clinically standardized. As technology advances and speed improves, however, this data will be increasingly accessible in the care of ICU patients.

SARS-CoV-2 RNA was identified in the unmapped reads in all patients (Fig. 1a). This supports that detection of SARS-CoV-2 in the serum has been associated with clinical deterioration[19,20] and RT-PCR identified the SARS-CoV-2 virus in the blood more often in the ICU patients than in the non-ICU patients[21]. The total number of reads in our dataset did not correlate with any outcomes, including mortality, ARDS, or coagulopathy. The low number of total reads, approximately 700 from nearly 2 billion from all the samples, explains the lack of success from other researchers identifying the virus in the blood. In early reports, RT-PCR directed at the N protein gene identified viral RNA in the plasma in 15% of patients[22]. Our data demonstrate the two most abundant genes in blood were the RNA Dependent RNA Polymerase and the N protein (Fig. 1b). With this data we propose these locations (RNA dependent RNA polymerase or N protein) as potential therapeutic or diagnostic targets[23]. In addition, this data shows that deep RNA sequencing of blood versus nasal swab, BAL or single cell RNA sequencing of blood produces drastically different results[24,25].

Other authors have called for robust testing for potential co-infections with SARS-CoV-2[26]. With deep sequencing and computational analysis we have identified the RNA from multiple bacteria, viruses, and archaea in all of the specimens, as well as fungal RNA in two participants. This suggests deep RNA sequencing with computational analysis may be a novel tool for the identification of co-infections. More data is required with comparison to gold standards such as blood culture and pathogen-specific PCR. However, RNA sequencing has the benefit of being able to identify all pathogens with known genomes, including both RNA and DNA based organisms. Moreover, unclassified reads that do not align to any known organism (Table 2) or the other sequences that have cellular organism elements (Table 2) could provide evidence of novel pathogens before a genome is sequenced or the pathogen is cultured.

Critically ill COVID-19 patients provide a difficult clinical dilemma as it pertains to antibiotics. In severely ill patients, clinicians are more likely to prescribe antibiotics despite there not being an identified pathogen[27]. With identification of bacteria known to cause human disease from the RNA sequencing data, appropriate antibiotics could be prescribed to these patients. In this data set, we show that there were significantly more counts of *Acinetobacter baumannii* in a portion of patients (Supplemental Table S2) and this bacterium has been associated with COVID-19[28]. Using a precision medicine approach with these data, patients with significantly elevated levels may potentially be treated with directed antibiotics, in the absence of more time-consuming positive culture data. While there was no difference in survival in participants with versus without identified bacteria in this study, antibiotic use was not standardized or prescribed prospectively based upon our results. In addition, analysis of the unmapped reads aligning to *Acinetobacter baumannii* (averaging over 50,000 among the six with increased reads) could provide insights into genes that are expressed in critical illness and provide novel diagnostic and therapeutic targets.

The immune response to SARS-CoV-2 has been the focus of much research since the pandemic started[29]. The successful use of corticosteroids in the critically ill with COVID-19 emphasizes the importance of the immune system in this disease[30,31]. Because a significant proportion of COVID-19 patients do not respond to corticosteroids, there are still calls for a more precise approach[32]. PD-1 expression is increased in certain cell populations in patients with COVID-19[3,33] but the uses of immune checkpoint inhibitors in cancer patients has been associated with more severe COVID-19[34]. Other authors suggest that immune checkpoint inhibitors may be useful in COVID-19[35]. Our data shows that patients who died had increased expression of PD-L1 and PD-L2 (Supplemental Fig. S1, CD274 and PDCD1Lg1, Table 3). This suggests that immune checkpoint inhibitors targeted against the PD-1 system might be considered in those patients identified to have increased expression of PD-L1 and PD-L2 because of their higher risk of death after ICU admission. The counterintuitive nature of the PD-L1 and PD-L2 expression signifies the complexity of this system and further work is needed. However, a limitation of this study is we are not able to distinguish between harm or benefit from the increase in checkpoint

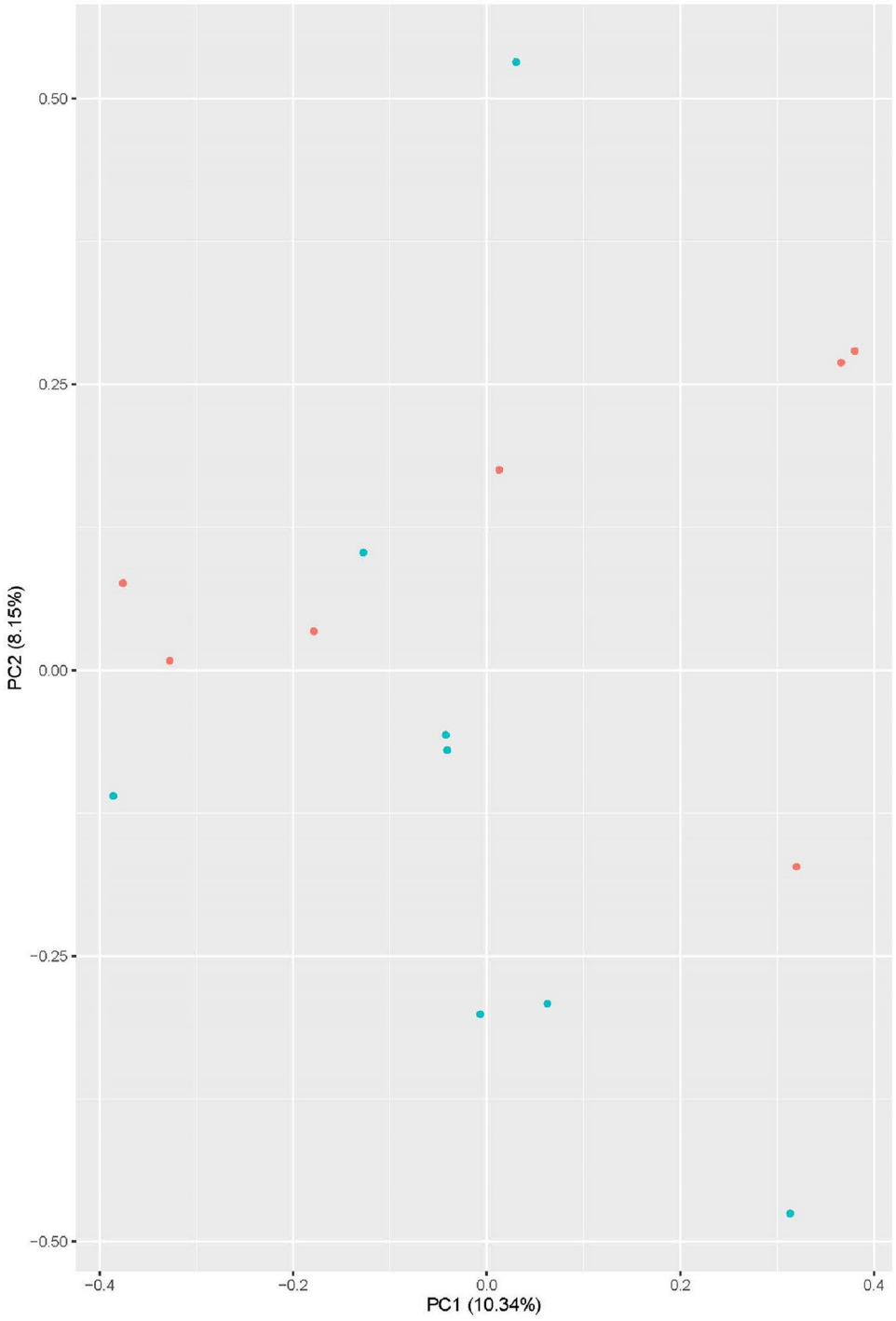

Figure 2. A graph created by the principal component analysis of the > 380,000 entropy values related to alternative RNA splicing and alternative transcription start/end. Patients labeled in red died from COVID-19 and surviving patients are labeled with green dots. Mortality rate above PC2 = 0 is 75% and below is 14% ($p = 0.04$).

proteins. If this could be ascertained by future work, this technology could identify patients who would receive the most benefit. In addition, from the GO term analysis, both PD-1 receptors and other genes impact "Negative Regulation of Interleukin 10" (Table 3) and this could be a better target. Other intriguing results from the GO term analysis include two dealing with protein trimerization, something the spike protein is known to do, and RNA processing, a topic previously commented on in critical illness[4].

Numerous other immune targets are identified from these genomic changes. ITGB2 has increased splicing of the last exon and is an immune cell integrin (Table 3, Table S4). Human leukocyte antigens have been associated

with COVID-19[36,37], however; no one has assessed splicing of these genes. HLA-C and HLA-E are both class 1 heavy chains and have decreased alternative splicing (Table 3, Table S4). Genetic defects of NCF2 are associated with chronic granulomatous disease, an immune suppressed phenotype[38]. In patients who died, NCF2 had increased alternative transcription end (Table 3, Table S5). N4BP1 is a cellular factor that interacts with HIV[39,40] and has increased alternative transcription end in the patients who died (Table 3, Table S5). N4BP1 is induced by interferon and the interferon response has been implicated in COVID-19[41,42]. Our data supports the role for interferons in COVID-19 as patients who died had 2.5-fold increase in expression of interferon 1 alpha (Supplemental Table S3, IFNA1). Clinical features of COVID-19 also correlate with some of the genes identified. OR6C4 is an olfactory gene which we identified has exhibiting a fivefold increased in expression in patients that died (Table 3, Supplemental Table S3). This finding suggests that loss of smell may signify milder disease among patients in the ICU. Thrombotic complications are common in COVID-19 patients (9.5%) and patients admitted to the ICU have a higher incidence of venous thromboembolism[43]. Patients who died have significant decrease in gene expression and multiple changes in alternative transcription end (Table 3, Supplemental Tables S3, S5) of both NRP1 and NRP2. Both these genes are associated with coagulation[44] and the COVID-19 spike protein binds both these receptors[45]. Previous work has shown that there is increased expression in both genes in the lungs of patients with COVID-19 when compared to controls[46]. In our study, the decrease NRP1 and NRP2 were seen in ICU patients who died compared to ICU patients who survived.

Many studies have attempted to utilize clinical data to predict mortality in COVID-19[47,48] and some focus on cytokines[49]. For simplicity all these attempt to identify a few variables to predict mortality. Here we utilize over 380,000 variables with PCA to create a figure that improves mortality prediction based upon where the patient is on the graph (Fig. 2, 75% versus 14%). A limitation to this form of analysis is that the PCA cannot identify a specific gene or event most responsible for outcomes; it uses all 380,000 data points. An additional limitation is the small sample size of this study. This work was done on just fifteen patients, however we plan to add future data from clinically relevant patients to this model to ensure its validity. More advanced analysis of the PCA plots to identify the most important features has been suggested, however the small sample size limits the utility. Other COVID-19 data sets could be used but they are limited by the depth of sequencing. In addition, Table 3 shows splicing events that were significantly different between groups. Although this method does not provide target for intervention, accurate assessment of prognosis using sequencing technology might be valuable to inform end of life care discussions in the ICU. Future work that will be done with this PCA plot based upon RNA splicing entropy will allow for the understanding of which factors influence the outcome the most. However, one advantage in the PCA plot is that it includes all the RNA splicing entropy values in one analysis. It essentially correlates how well the patient is splicing towards the clinical outcome of mortality.

Despite the limitations of this single-center study with a small patient number, we were still able to document that deep RNA sequencing and appropriate computational analysis yields valuable insight into the pathogenesis and host response of COVID-19 in critically ill patients. Novel drug targets were identified from SARS-CoV-2 RNA and the host response, including RNA dependent RNA polymerase, the N protein, and the PD-1 immune checkpoint pathway. The presence of pathogen RNA in the blood suggests co-infection should be reconsidered. Most importantly, PCA of the entropy of > 380,000 events allowed use to group patients into those likely to die versus those likely to live, and this may be helpful in family discussions with critically ill patients. Translating these results to clinical practice will improve the diagnosis, assessment of prognosis, and therapy of COVID-19.

Data availability

See Online Supplement for publication of extensive data.

Code availability

All code used is cited in the text.

Received: 12 August 2021; Accepted: 9 September 2022
Published online: 21 September 2022

References

1. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533 (2020).
2. Sethuraman, N., Jeremiah, S. S. & Ryo, A. Interpreting diagnostic tests for SARS-CoV-2. *JAMA* **323**, 2249 (2020).
3. Bouadma, L. *et al.* Immune alterations in a patient with SARS-CoV-2-related acute respiratory distress syndrome. *J. Clin. Immunol.* **40**, 1–11 (2020).
4. Fredericks, A. M., Wang, L. J., Fairbrother, W. G., Ayala, A. & Monaghan, S. F. Alternative RNA splicing and alternative transcription start/end in acute respiratory distress syndrome. *Intens. Care Med.* **46**, 813 (2020).
5. Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C. H. & Blencowe, B. J. Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. *Mol. Cell* **72**, 187–200 (2018).
6. Beigel, J. H. *et al.* Remdesivir for the treatment of covid-19—Preliminary report. *N. Engl. J. Med.* **383**, 1813 (2020).
7. Singer, M. *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 801–810 (2016).
8. Ferguson, N. D. *et al.* The Berlin definition of ARDS: An expanded rationale, justification, and supplementary material. *Intens. Care Med.* **38**, 1573–1582 (2012).
9. Aho, E. R. *et al.* Displacement of WDR5 from chromatin by a WIN site inhibitor with picomolar affinity. *Cell Rep.* **26**, 2916–2928 (2019).
10. Andrews, S. A quality control tool for high throughput sequence data. FastQC. In *A Quality Control Tool for High Throughput Sequence Data* (FastQC, 2014).
11. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford)* **29**, 15–21 (2013).
12. Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B. & Madden, T. L. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinform.* **20**, 405 (2019).

13. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
14. Editor (eds) Book, City.
15. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
16. Team RC (2018) R: A language and environment for statistical computing. In *Book R: A Language and Environment for Statistical Computing* (2018).
17. Fleige, S. & Pfaffl, M. W. RNA integrity and the effect on the real-time qRT-PCR performance. *Mol. Aspects Med.* **27**, 126–139 (2006).
18. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
19. The COVID-19 Investigation Team. Clinical and virologic characteristics of the first 12 patients with coronavirus disease 2019 (COVID-19) in the United States. *Nat. Med.* **26**, 861 (2020).
20. Chen, W. *et al.* Detectable 2019-nCoV viral RNA in blood is a strong indicator for the further clinical severity. *Emerg. Microbes Infect.* **9**, 469–473 (2020).
21. Fang, Z. *et al.* Comparisons of viral shedding time of SARS-CoV-2 of different samples in ICU and non-ICU patients. *J. Infect.* **81**, 147 (2020).
22. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London)* **395**, 497–506 (2020).
23. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459 (2020).
24. Bost, P. *et al.* Deciphering the state of immune silence in fatal COVID-19 patients. *Nat. Commun.* **12**, 1428 (2021).
25. Finkel, Y. *et al.* The coding capacity of SARS-CoV-2. *Nature* **589**, 125–130 (2021).
26. Lai, C. C., Wang, C. Y. & Hsueh, P. R. Co-infections among patients with COVID-19: The need for combination therapy with non-anti-SARS-CoV-2 agents?. *J. Microbiol. Immunol. Infect.* **53**, 505–512 (2020).
27. Feng, Y. *et al.* COVID-19 with different severity: A multi-center study of clinical features. *Am. J. Respir. Crit. Care Med.* **201**, 1380 (2020).
28. Sharifipour, E. *et al.* Evaluation of bacterial co-infections of the respiratory tract in COVID-19 patients admitted to ICU. *BMC Infect. Dis.* **20**, 646 (2020).
29. Poland, G. A., Ovsyannikova, I. G. & Kennedy, R. B. SARS-CoV-2 immunity: Review and applications to phase 3 vaccine candidates. *Lancet (London)* **396**, 1595 (2020).
30. TRC Group. Dexamethasone in hospitalized patients with Covid-19—preliminary report. *N. Engl. J. Med.* **384**, 693 (2020).
31. Prescott, H. C. & Rice, T. W. Corticosteroids in COVID-19 ARDS: Evidence and hope during the pandemic. *JAMA* **324**, 1292–1295 (2020).
32. Waterer, G. W. & Rello, J. Steroids and COVID-19: We need a precision approach, not one size fits all. *Infect. Dis. Therapy* **9**, 701 (2020).
33. Bellesi, S. *et al.* Increased CD95 (Fas) and PD-1 expression in peripheral blood T lymphocytes in COVID-19 patients. *Br. J. Haematol.* **191**, 207 (2020).
34. Robilotti, E. V. *et al.* Determinants of COVID-19 disease severity in patients with cancer. *Nat. Med.* **26**, 1218–1223 (2020).
35. Vivarelli, S. *et al.* Cancer management during COVID-19 pandemic: Is immune checkpoint inhibitors-based immunotherapy harmful or beneficial?. *Cancers* **12**, 2237 (2020).
36. Lorente, L. *et al.* HLA genetic polymorphisms and prognosis of patients with COVID-19. *Med. Intens.* **45**, 96 (2020).
37. Wang, W., Zhang, W., Zhang, J., He, J. & Zhu, F. Distribution of HLA allele frequencies in 82 Chinese individuals with coronavirus disease-2019 (COVID-19). *Hla* **96**, 194–196 (2020).
38. Roos, D. Chronic granulomatous disease. *Methods Mol. Biol. (Clifton)* **1982**, 531–542 (2019).
39. Nchioua, R., Bosso, M., Kmiec, D. & Kirchhoff, F. Cellular factors targeting HIV-1 transcription and viral RNA transcripts. *Viruses* **12**, 495 (2020).
40. Yamasoba, D. *et al.* N4BP1 restricts HIV-1 and its inactivation by MALT1 promotes viral reactivation. *Nat. Microbiol.* **4**, 1532–1544 (2019).
41. Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science (New York)* **369**, 718–724 (2020).
42. Lei, X. *et al.* Activation and evasion of type I interferon responses by SARS-CoV-2. *Nat. Commun.* **11**, 3810 (2020).
43. Al-Samkari, H. *et al.* COVID and coagulation: bleeding and thrombotic manifestations of SARS-CoV2 infection. *Blood* **136**, 489 (2020).
44. Rossignol, M., Gagnon, M. L. & Klagsbrun, M. Genomic organization of human neuropilin-1 and neuropilin-2 genes: Identification and distribution of splice variants and soluble isoforms. *Genomics* **70**, 211–222 (2000).
45. Daly, J. L. *et al.* Neuropilin-1 is a host factor for SARS-CoV-2 infection. *Science (New York)* **370**, 861 (2020).
46. Ackermann, M. *et al.* Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in Covid-19. *N. Engl. J. Med.* **383**, 120–128 (2020).
47. Tian, W. *et al.* Predictors of mortality in hospitalized COVID-19 patients: A systematic review and meta-analysis. *J. Med. Virol.* **92**, 1875 (2020).
48. Zhang, L. *et al.* D-dimer levels on admission to predict in-hospital mortality in patients with Covid-19. *J. Thromb. Haemost.* **18**, 1324–1329 (2020).
49. 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).

Author contributions

Drs. S.F.M. and A.M.F. had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: S.F.M., A.M.F., M.S.J., S.J.G., G.J.N., M.M.L., A.A. Acquisition, analysis, or interpretation of data: S.F.M., A.M.F., M.S.J., M.C., S.J.G., G.J.N., A.A. Drafting of the manuscript: S.F.M., A.M.F. Critical revision of the manuscript for important intellectual content: W.G.C., W.G.F., E.O.H., G.J.N., J.S.R., C.E.V., M.M.L., A.A. Statistical analysis: S.F.M., A.M.F., M.S.J., S.J.G. Obtained funding: S.F.M., A.M.F., W.G.F., E.O.H., G.J.N., J.S.R., C.E.V., A.A. Administrative, technical, or material support: W.G.C., G.J.N., J.S.R., C.E.V., M.M.L., A.A. Supervision: S.F.M., W.G.C., M.M.L., A.A. All authors reviewed the manuscript.

Funding

This study was supported by funding from the US National Institutes of Health: P20 GM103652 (SFM, WGF, EOH, AA), T32 HL134625 (AMF, EOH), R01 GM 127472 (WGF), P20 GM121344 (GJN), R01 HL147525 (JSR), R01 HL141268 (CEV), R35 GM118097 (AA).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-20139-1.

Correspondence and requests for materials should be addressed to S.F.M.

Reprints and permissions information is available at www.nature.com/reprints.

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