

Transcriptomic Similarities and Differences in Host Response between SARS-CoV-2 and Other Viral Infections

Simone A. Thair^{1,*}, Yudong D. He^{1,*}, Yehudit Hasin-Brumshtein¹, Suraj Sakaram¹, Rushika Pandya¹, Jiaying Toh^{2,3}, David Rawling¹, Melissa Remmel¹, Sabrina Coyle¹, George N. Dalekos⁴, Ioannis Koutsodimitropoulos⁵, Glykeria Vlachogianni⁶, Eleni Gkeka⁷, Eleni Karakike⁸, Georgia Damoraki⁸, Nikolaos Antonakos⁸, Purvesh Khatri^{2,3,+}, Evangelos J Giamarellos-Bourboulis^{8,+}, Timothy E Sweeney^{1,+}

Affiliations

1. Inflammatrix, Inc., 863 Mitten Rd, Suite 104, Burlingame, CA 94010, USA
2. Institute for Immunity, Transplantation and Infection, School of Medicine, Stanford University, Palo Alto, CA 94305, USA
3. Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA 94305, USA
4. Department of Internal Medicine, University of Thessaly, Larissa General Hospital, Greece
5. Intensive Care Unit, Latseion General Hospital of Elefsis, Greece
6. Intensive Care Unit, Aghios Dimitrios Thessaloniki General Hospital, Greece
7. Intensive Care Unit, AHEPA Thessaloniki General Hospital, Greece
8. 4th Department of Internal Medicine, National and Kapodistrian University of Athens, Medical School, 124 62 Athens, Greece

* Co-first authors

+ Co-senior authors

Corresponding author:

Timothy E Sweeney, MD, PhD

tsweeney@inflammatrix.com

863 Mitten Rd, Suite 104

Burlingame, CA 94010

Abstract

COVID-19 is a pandemic that shares certain clinical characteristics with other acute viral infections. Here, we studied the whole-blood transcriptomic host response to SARS-CoV-2 and compared it with other viral infections to understand similarities and differences in host response. Using RNAseq we profiled peripheral blood from 24 healthy controls and 62 prospectively enrolled patients with community-acquired lower respiratory tract infection by SARS-Cov-2 within the first 24 hours of hospital admission. We also compiled and curated 23 independent studies that profiled 1,855 blood samples from patients with one of six viruses (influenza, RSV, HRV, ebola, Dengue, and SARS-CoV-1). We show gene expression changes in peripheral blood in patients with COVID-19 versus healthy controls are highly correlated with changes in response to other viral infections ($r=0.74$, $p<0.001$). However, two genes, *ACO1* and *ATL3*, show significantly opposite changes between conditions. Pathway analysis in patients with COVID-19 or other viral infections versus healthy controls identified similar pathways including neutrophil activation, innate immune response, immune response to viral infection, and cytokine production for over-expressed genes. Conversely, for under-expressed genes, pathways indicated repression of lymphocyte differentiation and T cell activation. When comparing transcriptome profiles of patients with COVID-19 directly with those with other viral infections, we found 114 and 302 genes were over- or under-expressed, respectively, during COVID-19. Pathways analysis did not identify any significant pathways in these genes, suggesting novel responses to further study. Statistical deconvolution using immunoStates found that M1 macrophages, plasmacytoid dendritic cells, CD14⁺ monocytes, CD4⁺ T cells, and total B cells showed change consistently in the same direction across all viral infections including COVID-19. Those that increased in COVID-19 but decreased in non-COVID-19 viral infections were CD56^{bright} NK cells, M2 macrophages, and total NK cells. The concordant and discordant responses mapped out here provide a window to explore the pathophysiology of COVID-19 versus other viral infections and show clear differences in signaling pathways and cellularity as part of the host response to SARS-CoV-2.

Introduction

A novel coronavirus, SARS-CoV-2, has developed into a global pandemic, resulting in more than 7.6 million cases worldwide with over 427,000 deaths as we write (WHO accessed 14Jun2020)¹. Contextually this pandemic is likely to surpass the SARS-CoV-1 2003 pandemic by 1000-fold whereby SARS-CoV-1 resulted in 8,098 cases, took 12 months to contain and had a 9.6% mortality rate (WHO accessed 1Jun2020). The novel SARS-CoV-2 virus, the causative agent for COVID-19 disease, is highly communicable and despite urgent and resource-intensive efforts globally, we have no vaccine or efficacious treatment in sight².

COVID-19 clearly shares some immunological features with other viral responses, such as interferon activation, simultaneous repression of immune cells, and changes in metabolism including glucose and iron regulation as shown by cytokine and cytometry studies³⁻⁵. However, while many acute viral infections can lead to critical illness and death, COVID-19 appears both quantitatively and qualitatively to differ when compared to other acute viral infections. Notable features of COVID-19 include high rates of acute respiratory distress requiring mechanical ventilation; clinical coagulopathy; features of a cytokine storm and/or viral sepsis, and a high case fatality rate⁶. Thus, while studies comparing COVID-19 to healthy controls (HC) are useful, they do not explain the similarities and differences seen in the COVID-19 syndrome vs other viral infections.

Our approach involves a multi-cohort analysis of transcriptomic host response data to investigate host inflammation. The core discovery method leverages biological, clinical, and technical heterogeneity across datasets to identify generalizable disease biomarkers. We have repeatedly demonstrated that host response can be a generalizable sensitive and specific diagnostic and prognostic marker for presence, type, and severity of infections⁷⁻⁹, but also in autoimmune diseases, vaccination, TB, cancer and organ transplant^{7,8,17-20,9-16}. We have shown in methodological work that this method produces results with the greatest reproducibility in independent cohorts²¹.

In this work, we used RNAseq to profile whole blood samples from 62 COVID-19 patients prospectively enrolled in Athens, Greece, together with 24 healthy controls. We simultaneously compiled a database of clinical viral infections from 23 studies of > 1,800 samples to represent the conserved immune response to a broad range of viral infections including influenza, RSV, HRV, SARS-CoV-1, ebola, and dengue. We here report on the results of a comparison of host responses to SARS-CoV-2 and other viruses. We mapped out their similarities and differences at the gene level, pathway level, and cell proportion level, as a first step to gain a better understanding of this novel pandemic virus.

Methods

SAMPLE ACQUISITION AND PROCESSING

COVID-19 samples from Hellenic Sepsis Study Cohort

A total of 76 adult patients with SARS-CoV-2 pneumonia were prospectively enrolled from April 1st to May 4th by department participating in the Hellenic Sepsis Study Group (www.sepsis.gr) using the inclusion criteria already described elsewhere²². Lower respiratory tract infection was defined as the presence of infiltrates in chest X-ray or chest computed tomography compatible with COVID-19. SARS-Cov-2 was detected by positive molecular testing of respiratory secretions. For patients who required mechanical ventilation (MV), blood sampling was performed within the first 24 h from MV. Exclusion criteria were infection by the human immunodeficiency virus, neutropenia, and any previous intake of immunosuppressive medication (corticosteroids, anti-cytokine biologicals, and biological response modifiers). The studies were conducted under the 30/20 approval by the National Ethics Committee of Greece. Written informed consent was provided by patients or by first-degree relatives in cases where patients were unable to consent.

Whole blood was drawn in PAXgene tubes at enrollment along with other standard laboratory parameters. Data collection included demographic information, clinical scores (SOFA, APACHE II), laboratory results, length of stay and clinical outcomes. Patients were followed up daily for 30 days; outcomes were defined as severe respiratory failure (PaO₂/FiO₂ ratio less than 150 requiring MV) or death. PAXgene Blood RNA samples were shipped to Inflammatrix for processing.

Healthy control sample sourcing

Blood RNA tubes were prospectively collected from healthy controls (HC) through a commercial vendor (BioIVT) under IRB approval (Western IRB #2016165) using informed consent. Patients were non-febrile and verbally screened to confirm that no signs or symptoms of infection were present within 3 days prior to sample collection and that they were not currently undergoing antibiotic treatment, nor had not taken antibiotics within 3 days prior to sample collection. Furthermore, all samples were negative for HIV, West Nile, Hepatitis B, and Hepatitis C by molecular or antibody-based testing.

RNA extraction protocol

Prior to processing, samples in PAXgene Blood RNA tubes from 76 COVID-19 patients and 24 healthy controls were removed from -80C to thaw at room temperature for two hours. The samples were then inverted several times to achieve homogeneity, after which 3 mL aliquots were removed for processing. RNA was extracted from these samples using a modified version of the RNeasy Mini Kit (QIAGEN) protocol executed on the a QIAcube automated workstation. PAXgene samples comprise of whole blood in PAXgene stabilizing solution. The sample is diluted with PBS, then centrifuged at 3,000 x g to pellet precipitated nucleic acids. Pellets were washed with molecular biology grade water and again pelleted via centrifugation at 3,000 x g. Pelleted material is resuspended in Buffer RLT (QIAGEN). Using the automated QIAcube, samples are then subjected to treatment by Proteinase K and gDNA elimination via columns (QIAGEN). Flow-through was mixed with isopropanol and passed over a MinElute (QIAGEN) spin column. The column was washed with 80% ethanol and purified nucleic acid was eluted in RNase-free water. Purified RNA was heat denatured at 55° C for 5 minutes, then snap-cooled on ice. RNA was quantitated using a Qubit fluorimeter with the Quant-iT RNA Assay kit (Thermo-Fisher). Samples with an RNA integrity number (RIN) below 7 (BioAnalyzer, Agilent) did not proceed to sequencing, resulting in 62 COVID-19 samples and 24 HC samples for sequencing.

RNAseq library preparation

Total RNA samples were depleted of globin RNA using the GLOBINclear kit (Invitrogen) following the procedure described by the manufacturer. Globin-depleted RNA was quantified using the Qubit RNA High Sensitivity kit (Life Technologies) and 10ng of globin-depleted RNA was then used for rRNA depletion and RNAseq library preparation using the SMARTer Stranded Total RNAseq kit v2 Pico Input Mammalian (Takara Bio) following the manufacturer's protocol. RNAseq libraries were then quantified using the Qubit dsDNA High Sensitivity kit (Life Technologies) and their quality and size evaluated by a Fragment Analyzer High Sensitivity Small Fragment kit (Agilent Technologies).

RNA sequencing

A total of 86 RNAseq libraries generated above were pooled and sequenced on an Illumina NovaSeq6000 Sequencing System (Illumina) in a paired-end fashion (2 x 100 cycles). 41 M to 124 M paired-end reads were obtained for each sample obtained for each sample. Fastq files were used as input for RNAseq data processing. Library prep and sequencing were performed at TB-SEQ (Palo Alto, CA).

DATA PROCESSING AND ANALYSIS

RNAseq data processing

Trimming: Quality control (QC) assessment of the reads was done using FastQC²³. The adapter sequence and 3 bases on the 3' end of the reads was trimmed using cutadapt as a commonly used procedure²⁴.

Alignment: Trimmed reads were mapped to a reference genome index generated based on the human genome, GRCh38, and a transcriptome reference, GENCODE v32 primary assembly gtf²⁵ with the sjdbOverhang option set to 100 (default), using STAR aligner (v2.7.3a).

Quantitation: Mapped reads were quantified as per Ensembl transcript ID as defined in GENCODE v32 annotation. Reads were summed across Ensembl transcript IDs mapping to Entrez gene IDs in order to compare them with other viral data assayed by microarrays (AnnotationDbi from Bioconductor)²⁶.

Data Quality: Various QC metrics prior to and post trimming were examined to assess data quality as a standard procedure for RNAseq data. Additionally, the distributions of raw and trimmed counts were assessed and Principal Component Analysis (PCA) with various cutoffs was performed for QC. All 86 samples passed standard QC metrics and the resulting counts matrix (12,142 Entrez genes by 86 samples) was used in subsequent data integration steps (**Supplementary Table 1**).

Normalization and voom transformation of RNAseq counts

Low-expressed genes were filtered using the following cutoff: max counts per million (CPM) less than 5 across all 86 samples. Normalization factors were obtained using edgeR's Trimmed Mean of M values (TMM) method²⁷. The voom method was then used to transform counts into normalized log₂-CPM (**Supplementary Figure 1**)²⁸. Data is available at Gene Expression Omnibus (GEO) repository (GSE152641).

Non-COVID-19 viral dataset selection

Transcriptomic data of clinical respiratory infections caused by viruses other than SARS-CoV-2 were surveyed from Gene Expression Omnibus (GEO) and ArrayExpress for inclusion to define a conserved host response signature for non-COVID-19 viral infection. We identified 23 such independent datasets that profiled a total of 1,855 peripheral blood samples (PBMCs or whole blood) from patients (infants, children, or adults) with one of six viral infections (influenza, RSV, HRV, ebola, dengue, SARS-CoV-1, but not SARS-CoV-2). Collectively the 23 datasets comprised of 780 samples from healthy controls and 1,075 from patients with a viral infection represent biological, clinical, and technical heterogeneity observed in the real-world patient population with viral infections.

Non-COVID-19 viral dataset processing

Raw microarray data for each dataset was renormalized (when available) using standardized methods. Affymetrix arrays were renormalized using the robust multichip average (RMA) method. Illumina, Agilent, GE, and other commercial arrays were renormalized via normal-exponential background correction followed by quantile normalization. Data were log₂-transformed. Probe to gene (Entrez ID) summarization was performed within each study using the mean signal intensity for probes mapping to a single gene.

COCONUT conormalization of all data sets

Of the 23 non-COVID-19 viral infections datasets, 20 datasets with a total of 879 viral infected patients and 754 HCs met the criteria for conormalization: 1) the dataset must have HCs, and 2) the dataset was obtained on a single-channel microarray platform. Integrated with the voom-transformed RNAseq dataset for COVID-19, they were conormalized together using COCONUT as previously described⁸. COCONUT uses COMBAT empirical-Bayes conormalization on healthy controls to derive correction factors for diseased patients. The technique integrates datasets such that (i) no bias is introduced to the diseased samples, (ii) there is no change to the distribution of a gene within a study, and (iii) each gene shares the same distribution across healthy controls between studies after normalization. This COCONUT conormalized expression data comprising of a total of 941 (COVID-19 and non-COVID-19) viral patients and 778 HCs across 9,818 genes common across 11 platforms were used as input data to perform the following multicohort and integrated analyses.

COVID-19 versus healthy control comparison

Hedges' *g* effect size (ES)²⁹ for each gene was calculated for COVID-19 (62) versus HC (24) two-group comparison test from the COCONUT conormalized output. P-value was calculated using a student's t-test and adjusted using the Benjamini-Hochberg method to obtain the False Discovery Rate (FDR). ES threshold of ≥ 1 or ≤ -1 in combination with FDR threshold of $\leq 0.05\%$ was used to identify genes whose expressions are over- or under-expressed in COVID-19 infected patients than in the mean value of HCs.

Non-COVID-19 viral versus healthy controls comparison

14 datasets composed of 1,324 whole blood and PBMC samples were chosen for the discovery cohort, of which 652 were from respiratory viral infected patients (viral) and 672 samples were from HCs patients. As a multicohort analysis with conormalized data as input, we utilized a well-established MetaIntegrator (version 2.1.1) as described previously³⁰. Briefly, Hedges' *g* ES was computed for each gene within a study between viral and HC. ESs for genes across studies was summarized using the DerSimonian & Laird random-effects model, where each ES is weighted by the inverse of the variance in that study³¹. We used an ES threshold ≥ 1 or ≤ -1 with FDR $\leq 0.05\%$ to identify signature genes (**Supplementary Table 2**).

Validation of non-COVID-19 viral infection signature

The signature genes identified based on 14 discovery datasets were evaluated for prediction of viral infections from HC with a score calculated for each sample using the following formula:

$$\text{viral score} = zscore(\text{GeoMean}(\text{pos}) - \text{GeoMean}(\text{neg}))$$

The score is a rescaled difference between geometric means of positive (over-expressed) genes and negative (under-expressed) genes. Receiver-operating characteristics (ROC) plots are generated for held out validation datasets and the Area Under the ROC (AUC) is used as a performance metric. For validation of the non-COVID-19 viral signature, we compiled 9 datasets comprised of 6 held out from the COCONUT expression data, plus 3 normalized as per platform requirements without COCONUT (**Table 3**). We then tested this signature first using 4 datasets comprising of 178 respiratory viral infection samples and 58 HCs (236 total) (**Table 3**). We then further validated this signature in 5 datasets of other viral etiology (245 viral and 50 HC, 295 total) (**Table 3**).

COVID-19 versus non-COVID-19 viral Comparison

Hedges' g ES was calculated for each gene in a COVID-19 (62) and non-COVID-19 viral (652) two-group comparison test from the COCONUT conormalized expression data. P-value was calculated using a Welch's t-test assuming unequal variance and sample sizes and adjusted using the Benjamini-Hochberg³² method to obtain the False Discovery Rate (FDR). ES threshold ≥ 1 or ≤ -1 in combination with FDR threshold of $\leq 0.05\%$ was used to identify signature genes.

PATHWAY AND IMMUNOSTATES ANALYSIS

Pathway Analysis

Each over- or under-expressed gene set from comparisons between COVID-19 vs HC, non-COVID-19 viral infection vs HC, and COVID-19 vs non-COVID-19 viral infection was subjected to a pathway analysis with Gene Set Enrichment Analysis³³. We tested significance of over-representation of genes in each of the pathways reflected in Gene Ontology (GO) including biological process (BP), molecular function (MF), and cellular compartment (CC). The human transcriptome reference is used as background and the p-values from the hypergeometric test were adjusted using the Benjamini-Hochberg method³². Top-ranked pathways common between COVID-19 and non-COVID-19, and specific separately to COVID-19 or non-COVID-19 viral infections were selected.

ImmunoStates Analysis

A statistical deconvolution method was used to estimate the percentage of 25 immune cell types in the peripheral blood transcriptome data^{34,35}. Statistical deconvolution estimates the percentage of various cell types present in a blood transcriptome profile. It uses a set of pre-defined genes that represent cell types of interest, called a basis matrix, and a variant of linear regression to make estimates. Previously, it was demonstrated that different methods produce highly correlated estimates of cellular proportions once basis matrix is fixed³⁴. Here, immunoStates (MetaIntegrator) was used as a basis matrix because it has been shown to reduce the effect of the biological and technical heterogeneity in transcriptome data on statistical deconvolution and identify robust changes in immune cell proportions³⁴⁻³⁷. The 14 non-COVID-19 viral discovery datasets and the COVID-19 dataset were deconvolved separately, then change in proportion of a given cell type between healthy controls and the infected patients of each dataset was estimated.

Results

Differential expression analysis of transcriptome profiles of patients with COVID-19

We prospectively enrolled and sequenced RNAseq from whole blood from 62 patients with COVID-19 and 24 healthy controls (**Table 1**). Differential expression analysis of 86 peripheral blood samples identified 2,002 differentially expressed genes (771 over-expressed, 1,231 under-expressed; **Figure 1a**, **Supplementary Table 2**) with absolute ES ≥ 1 and FDR $\leq 0.05\%$, referred to as COVID-19 signature. We performed pathway enrichment analysis of the COVID-19 signature using Gene Ontology (GO) terms. The 30 most significant pathways for 771 over-expressed genes included neutrophil activation, innate immune response, immune response to viral infection, type-I interferon signaling and cytokine production (**Figure 1b**), and for 1,231 under-expressed genes include lymphocyte differentiation and T cell activation and regulation (**Figure 1c**). These results suggest that in response to SARS-CoV-2 infection T cells are suppressed whereas neutrophils are activated as a hallmark of its overwhelming host response represented in the transcriptomic changes. High neutrophil-to-lymphocyte ratios have been observed as a marker of severity in sepsis, cancer, and pneumonia³⁸⁻⁴¹.

Identification of host response genes to viral infections through multi-cohort analysis

Based on our previous results¹⁴, we hypothesized that there is a conserved immune response to respiratory viral infections irrespective of age and genetic background of a patient or a virus. We identified 23 studies of acute viral infection, and from these selected 14 as our discovery set for a non-COVID-19 viral signature (**Table 2**), and 9 were held out for validation. Statistical power analysis⁴² found that even with high inter-study heterogeneity, we had more than 80% statistical power at p-value = 0.01 for detecting absolute ES > 0.43 in these datasets (**Supplementary Figure 2**). The multi-cohort analysis of 1,324 transcriptome profiles (652 non-COVID-19 viral patients, 672 healthy controls) from these 14 studies using MetaIntegrator³⁰ identified 635 differentially expressed genes (314 over-expressed, 321 under-expressed). ROC plots for all of the discovery datasets using this signature illustrate the high sensitivity and specificity this gene list possesses, indicating genes that are highly discriminatory and hence likely to represent this conserved signature (**Figure 2a**, **Supplementary Table 2**). We refer to these 635 genes in short as the non-COVID-19 viral signature. Similar to the COVID-19 signature, GO analysis of over- and under-expressed genes in the non-COVID-19 viral signature identified a similar set of pathways highlighted by neutrophil and T cell activation, respectively (**Figure 2b, 2c**).

Validation of host response genes to viral infections in multiple independent datasets

Next, we confirmed that the non-COVID-19 viral signature is conserved across viruses by validating it in several independent datasets. We calculate the non-COVID-19 viral score for a sample as the difference in geometric means of over-expressed and under-expressed genes. In four independent studies consisting of 236 samples (178 viral infections, 58 healthy controls; **Table 3**), the score accurately distinguished patients with a respiratory viral infection (influenza, HRV, or RSV) from HCs (**Figure 3a**).

Second, we investigated whether the non-COVID-19 viral signature is observed in other severe viral infections including ebola, dengue, and SARS-CoV-1 in five independent studies (50 HC, 54 SARS-CoV-1, 37 ebola, 154 dengue). In each study, the non-COVID-19 viral score also distinguished patients with a viral infection from healthy controls with high accuracy (**Figure 3b**).

Third, we tested whether the non-COVID-19 viral signature would also distinguish patients with COVID-19 from healthy controls. We calculated the non-COVID-19 viral score for each of 62 COVID-19 patients together with 24 HCs using the conormalized expression data. We found that non-COVID-19 viral score separated patients with COVID-19 from HCs with an AUC of 0.96 (**Figure 3c**), similar to SARS-CoV-1 (AUC=0.98).

Comparison of COVID-19 profile with non-COVID-19 viral infection profile

Next, we investigated similarities and differences in host response to SARS-CoV-2 and other respiratory viruses by comparing change in expression with respect to healthy controls across 9,818 genes that were present across all datasets. When considering the entire transcriptome, there was high correlation ($r = 0.74$, $p < 0.001$) between change in expression in response to SARS-CoV-2 or other respiratory viruses (ES from COVID-19 vs HC comparison is plotted against ES from non-COVID-19 vs HC comparison in **Figure 4a**). We visualized 2,002 COVID-19 signature genes and 635 non-COVID-19 signature genes in the same ES scatter plot by different colors to highlight their relationships (**Figure 4a, Supplementary Table 2**). We observe that 7,626 genes uncolored in the middle (gray, with higher density in the center shown by contours) out of 9,818 profiled (77.7%) are not in the signature genes in either COVID-19 or non-COVID-19 viral infections. Given the high correlation ($r = 0.74$), it is not surprising that 223 genes are concordantly over-expressed ($ES \geq 1$, $FDR \leq 0.05\%$) as well as 220 genes concordantly under-expressed with ($ES \leq -1$, $FDR \leq 0.05\%$). Of the remaining genes from the non-COVID-19 signature, there are 90 genes over-expressed and 100 genes under-expressed in non-COVID-19, however these had ES between -1 and 1 in the distribution of the COVID-19 ESs. As well, of the remaining genes from the COVID-19 signature, there are 547 genes over-expressed and 1,010 genes under-expressed in COVID-19 that had ES between -1 and 1 in the distribution of the non-COVID-19 ESs. We only found two genes that were completely discordant, thus completely oppositely regulated in COVID-19 and non-COVID-19 viral infections: Aconitase1 (*ACO1*) over-expressed in COVID-19 and under-expressed in non-COVID-19 viral infections and Atlantin GTPase 3 (*ATL3*) over-expressed in non-COVID-19 viral infections and under-expressed in COVID-19. Interestingly, *ACO1* is involved in iron metabolism, and heme appears to be interlinked with COVID-19 pathophysiology⁴³. *ATL3* is required for endoplasmic reticulum (ER) membrane junctions and may be linked to viral replication sites⁴⁴.

Therefore, in order to identify a statistically significant set of genes differentially expressed in COVID-19 patients compared to those with other viral infections, we employed COCONUT to conormalize the two disease types into a single matrix for comparison of 62 COVID-19 patients versus 652 non-COVID-19 viral infection patients. Using COCONUT allows for comparison across datasets with heterogeneity while simultaneously creating a way to calculate an FDR for the gene effect size when compared “head-to-head” or “disease to disease” directly, rather than looking for correlated and anti-correlated genes for which ES and FDR are calculated separately. At $|ES| \geq 1$ with $FDR \leq 0.05\%$, we found 416 genes as COVID-19-specific genes, 114 over-expressed and 302 under-expressed in patients with COVID-19 than in those with non-COVID-19 viral infection (**Figure 4b**). To illustrate the gain in identification of genes to investigate and re-iterate the value in this statistical method, this set of genes from (b) are highlighted in the same scatter plot from panel a (**Figure 4c**).

Unlike the COVID-19 and non-COVID-19 viral signatures, the pathway analysis of this gene set did not identify any statistically significant GO terms, potentially indicating novel pathophysiology unique to COVID-19. This combination of genes may include those less well annotated within pathways and thus less likely to result in statistically significance assignment to a pathway. Nonetheless, top-ranked but statistically insignificant GO terms include muscle contraction, regulation of epithelial cell proliferation, and biological processes involved in lung and respiratory development for 114 positive genes, as well as pathways related to T cell homeostasis and T cell differentiation for 302 negative genes. The significance of these pathways in connection with clinical manifestation needs to be investigated further.

Similarities and differences in pathways between COVID-19 and non-COVID-19 viral infection

We expanded our comparison of significant pathways in response to SARS-CoV-2 versus non-COVID-19 viruses by including all pathways instead of only 30 most significant pathways. We found pathways for over-expressed

genes are highly concordant between patients with COVID-19 and non-COVID-19 viral infections (**Figure 5a**), pathways for under-expressed genes are discordant (**Figure 5b**).

To amalgamate these findings we performed hierarchical clustering of all pathway analysis results of all gene sets of interest including *three signature sets*: 1) COVID-19 vs HC (771 over- and 1,231 under-expressed), 2) non-COVID-19 viral vs HC (314 over- and 321 under-expressed), and 3) COVID-19 vs non-COVID-19 viral (114 over- and 302 under-expressed) as well as the *8 gene groups* based on concordance between signatures (**Figure 5c, Supplementary Table 2**). To check the dependency of GO term enrichment results on the cutoffs for selecting signature genes, we tested three additional cutoffs (less or more stringent than the chosen one) each for COVID-19 vs HC, non-COVID-19 vs HC, or COVID-19 vs non-COVID-19 comparison. The results for over-expressed, under-expressed, and all genes from each cutoff together with the 9 gene sets from **Figure 4a** show a merging and comprehensive picture of pathway analysis results (**Supplementary Table 3, Supplementary Figure 4**) allowing one to focus on pathways of interest, either commonly significant across gene sets or uniquely significant in a gene set or a combination of genes of interest.

Similarities and differences in changes in immune cell proportions between COVID-19 and non-COVID-19 viral infection

We estimated proportions of 25 immune cell types in bulk gene expression in blood samples from patients with COVID-19 or non-COVID-19 viral infections using immunoStates. In patients with COVID-19, we found immune cells from myeloid-lineage (M1 macrophages, neutrophils, and MAST cells) increased significantly ($FDR \leq 10\%$), and lymphoid cells (CD4+ and CD8+ alpha-beta T cells, B cells) decreased significantly ($FDR \leq 10\%$) during viral infection (**Figure 6a, Supplementary Table 3**). These results are in line with recent reports demonstrating increased neutrophil and decreased T cell counts in COVID-19 patients³⁹⁻⁴¹. In patients with non-COVID-19 viral infections, we observed significant increase in proportion for myeloid cells (M1 macrophages, CD14+ monocytes, MAST cells), and significant decrease in proportion for lymphoid cells (CD4+ and CD8+ T cells, gamma-delta T cells, B cells) (**Figure 6b, Supplementary Figure 3**). Indeed, when considering changes within each dataset, M1 macrophages, plasmacytoid dendritic cells, CD14+ monocytes, CD4+ T cells, and total T cells showed change consistently in the same direction across all viral infections including COVID-19 (**Figure 6b**).

We observed an overall correlation of 0.493 ($p=0.017$) for change in cellular proportions in patients with COVID-19 compared to non-COVID-19 viral infections (**Figure 6c, Supplementary Table 3**), where all but 6 cell types changed in the same direction, though not all changes were statistically significant. We again observed increased neutrophil and decreased T cell counts in COVID-19 which is in line with a recent study that compared COVID-19 to the 2009 H1N1²⁰. Cell types that increased in COVID-19 relative to non-COVID-19 were CD56^{bright} NK cells, M2 macrophages, and total NK cells. Those that decreased in non-COVID-19 relative to COVID-19 were CD56^{dim} NK cells, memory B cells, and eosinophils. Although change in memory B cells was not statistically significant, the direction of change is expected as patients with non-COVID-19 infection are highly likely to have memory to those viruses, whereas SARS-CoV-2 is a novel coronavirus with no pre-existing memory in the population. Similar findings are reported when the absolute cell counts were measured by flow cytometry in smaller patient populations²⁰.

Discussion

Understanding the pathophysiology of COVID-19 is critical to finding new treatments. Here we take a host response transcriptomics approach using peripheral blood transcriptomics of the immune response to COVID-19 (n=62) compared to 652 non-COVID-19 viral infections spanning 6 viruses. While the vast majority of the host immune response appears to be similar between COVID-19 and other viruses, our study highlights some key differences.

The scatter plot of the correlation of the differential expression of non-COVID-19 viral infections versus COVID-19 infections illustrates this large proportion of concordance and seemingly small amount of discordance (Figure 4). We found only two genes, *ACO1* and *ATL3*, that were expressed in opposite directions. *ACO1* was over-expressed in COVID-19 versus HC and under-expressed in non-COVID-19 viral infections versus HC, whereas *ATL3* entirely oppositely regulated (Figure 4). Prior reports suggest that both genes may be involved in viral replication and immune evasion. *ACO1* is an iron-sulfur protein that regulates ferritin and transferrin. When cellular iron levels are low, the protein binds to iron-responsive elements (IREs), which represses translation of ferritin (a protein that stores iron), and simultaneously stabilizes the normally rapidly degraded transferrin receptor mRNA allowing for translation of the receptor and more cellular uptake of iron, which is required for proliferation⁴⁵. High levels of ferritin are also indicative of macrophage activation syndrome (MAS) and have been observed in COVID-19 patients^{22,46–48}. *ATL3* is a member of the integral membrane GTPases. Proper formation of ER tubules is affected by mutations in this gene. Viruses are known to target host organelles to enter a host cell and avoid destruction⁴⁹. Lack of ATL results in delayed cargo exit and coat assembly for budding from the ER which is necessary for export of cytokines and chemokines in response to infection; *ATL3* has been linked directly to viral replication in Zika⁴⁴, although Zika was not studied here.

The power of using COCONUT to combine heterogeneous datasets allowed for a pooled, head-to-head comparison of COVID-19 with non-COVID-19 viral infections. Interestingly, the differentially expressed genes in this analysis were not enriched for any GO terms. However, there is bias in the annotation of gene ontologies, so absence of evidence does not denote evidence of absence of coordinated differential response^{50,51}. Indeed, **Figure 5** illustrates the comparison of COVID-19 to non-COVID-19 GO terms. We found many downregulated pathways are discordant when comparing to healthy controls. Within these, a cluster of pathways that are high in COVID-19 and low in non-COVID-19 viral infections involve ribosome related processes. In SARS-CoV-1 infections it was determined that viral nsp1 disrupts ribosomal function⁵². The inverse cluster of pathways that are high in non-COVID-19 viral infections and low in COVID-19 positively regulate cell-cell adhesion, cell activation, leukocyte activation, immune response-activating cell surface receptor signaling, perhaps suggesting a more dysregulated immune response. While the host response to SARS-CoV-2 in essence is highly similar to other viral infections, it does clearly have some molecular differences. Of particular interest was the observation that while both diseases had increased type-1 interferon signaling pathways, the magnitude of this pathway response was lower in the COVID-19 (**Figure 5**).

Interestingly, the consistency in the change in the immune cell proportions are mostly consistent across COVID-19 and non-COVID-19 datasets. Our results are in line with several recent studies that found high neutrophil-lymphocyte-ratio (NLR) in COVID-19 patients^{38–41}. Expansion of CD56^{bright} NK cells is common in many viral infections, as part of recognizing and killing virally infected cells while orchestrating adaptive immune responses⁵³. Comparing patients with COVID-19 to HCs shows an increase in NK cells (**Figure 6a**), largely driven by the CD56^{bright} population. When compared to non-COVID-19 viral infections the increase in NK cell (via CD56^{bright} NK cell) proportion remains high in the COVID-19 infections. This phenomenon was also directly observed using mass spectrometry to measure cell abundance over time in COVID-19 patients, and when

considering factors most explanatory in those that recovered the cells that were the most dynamic included CD56^{dim} NK cells⁵⁴.

When comparing COVID-19 to non-COVID-19 viral infections, we see M1 macrophage proportions are similar to that of other viral diseases, but the elevated M2 response is discordant. M1 macrophages are pro-inflammatory and kill invaders, whereas M2 macrophages are considered anti-inflammatory and reparative. A large body of work in bacterial sepsis found that individuals with high M1 profiles had increased mortality whereas those with a more evenly balanced M1/M2 were more likely to survive⁵⁵. However, in general, monocytotropic viruses including SARS-CoV-1 have evolved mechanisms to interfere with effective macrophage polarization, favoring the M2 population for immune evasion. For example, virus-induced macrophage depletion is executed by viruses that carry pro-apoptotic proteins, thus initially reducing the number of M1s to skew population to M2 and avoid attack, then further suppress the production and action of type I IFNs, stunting the progression of M1 macrophage polarization⁵⁶. This shift we see in the proportion of M2 macrophages in COVID-19 versus non-COVID-19 viral infections indicate that this novel pathogen may be executing these immune evasion techniques with a high degree of success.

Our study has some limitations due to the design of using public data for non-COVID-19 comparison. First, due to the limited nature of clinical studies in a pandemic, we had just 62 patients with COVID-19 compared to >650 with other viral infections, creating class imbalance in their comparison. Second, we did not investigate effects of severity on host response as this was mostly unavailable. It is possible that differences in severity between this COVID-19 cohort and the other viral cohorts was a confounder in our analysis. Third, we analysed differential expression at single pre-set significance and effect size thresholds. Choosing different thresholds (e.g., thresholds based on 80% statistical power in each analysis) would have identified different sets of differentially expressed genes. We provide ES and FDR values for all genes (**Supplementary Table 2**) to enable re-analysis of these genes based on thresholds that others may deem more appropriate. **Supplementary Figure 4** is also provided to show the GO term enrichment results by varying cut-offs.

Conclusions

We here provide bulk RNAseq profiling of peripheral blood in COVID-19 in comparison to healthy controls which we derived a signature of 2002 genes for investigation of the biology and potentially pathophysiology of this disease. We compiled an extensive database of non-COVID-19 viral infections across many platforms, ages, diseases and locations globally to compare to healthy controls using meta-integration to derive a set of 635 genes representing the host response to known viral pathogens. We then used COCONUT to connormalize all of the data and directly compare COVID-19 to non-COVID-19 viral infections resulting in a signature of 416 genes. We used all of these analyses to identify both the similarities and differences in the underlying host response. While we identified that a large proportion of the host response is similar to that of other infections, we also identified key differences in individual genes, pathways, and cellularity that are suggestive of the clinical differences observed in COVID-19. Of particular interest are the potential roles of ACO1 and ATL3 in describing the differential host response in COVID-19 complemented by the 416 genes that may identify novel biology and further the understanding of ACO1 and ATL3, but our findings will need to be replicated in further clinical studies.

Acknowledgements

We are grateful to Ashley Prasse Miller, Mario Esquivel, and Oliver Liesenfeld of Inflammatrix Clinical Affairs team for clinical sample availability, and Luciano Brocchieri and Silvia Tomaletti of TB-SEQ for helpful discussion.

Competing interests

SAT, YDH, YH, SS, RP, DR, MR, SC, and TES are employees of, and stockholders in, Inflammatrix, Inc. PK is a shareholder and a consultant to Inflammatrix, Inc. EJGB has received honoraria from AbbVie USA, Abbott CH, InflaRx GmbH, MSD Greece, XBiotech Inc. and Angelini Italy; independent educational grants from AbbVie, Abbott, Astellas Pharma Europe, AxisShield, bioMérieux Inc, InflaRx GmbH, and XBiotech Inc; and funding from the FrameWork 7 program HemoSpec (granted to the National and Kapodistrian University of Athens), the Horizon2020 Marie-Curie Project European Sepsis Academy (granted to the National and Kapodistrian University of Athens), and the Horizon 2020 European Grant ImmunoSep (granted to the Hellenic Institute for the Study of Sepsis). The other authors declare no competing interests.

Author contributions

TES, YDH, PK, and EJGB designed the study; GND, JK, GV, EG, EK, GD, NA and EJGB conducted clinical studies; SAT, YDH, YH, SS, RP, JT, PK performed bioinformatics analysis; DR, MR, and SC processed samples; SAT, YDH, PK, and TES wrote the manuscript; all authors critically revised and approved the manuscript.

Data availability

The public cohorts are available under their respective study IDs. The COVID-19 cohort is deposited in GEO for public reuse (GSE152641).

References

1. Zhou P, Yang X Lou, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020. doi:10.1038/s41586-020-2012-7
2. Callaway E. The race for Coronavirus vaccines. *Nature*. 2020.
3. Wilson JG, Simpson LJ, Ferreira A-M, et al. Cytokine profile in plasma of severe COVID-19 does not differ from ARDS and sepsis. *medRxiv*. January 2020:2020.05.15.20103549. doi:10.1101/2020.05.15.20103549
4. Catanzaro M, Fagiani F, Racchi M, Corsini E, Govoni S, Lanni C. Immune response in COVID-19: addressing a pharmacological challenge by targeting pathways triggered by SARS-CoV-2. *Signal Transduct Target Ther*. 2020;5(1):84. doi:10.1038/s41392-020-0191-1
5. Drakesmith H, Prentice A. Viral infection and iron metabolism. *Nat Rev Microbiol*. 2008. doi:10.1038/nrmicro1930
6. Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol*. 2020. doi:10.1038/s41577-020-0311-8
7. Sweeney TE, Shidham A, Wong HR, Khatri P. A comprehensive time-course-based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set. *Sci Transl Med*. 2015. doi:10.1126/scitranslmed.aaa5993
8. Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Sci Transl Med*. 2016. doi:10.1126/scitranslmed.aaf7165
9. Sweeney TE, Perumal TM, Henao R, et al. A community approach to mortality prediction in sepsis via gene expression analysis. *Nat Commun*. 2018. doi:10.1038/s41467-018-03078-2
10. Li L, Khatri P, Sigdel TK, et al. A peripheral blood diagnostic test for acute rejection in renal transplantation. *Am J Transplant*. 2012. doi:10.1111/j.1600-6143.2012.04253.x
11. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: A multicohort analysis. *Lancet Respir Med*. 2016. doi:10.1016/S2213-2600(16)00048-5
12. Sweeney TE, Khatri P. Comprehensive validation of the FAIM3: PLAC8 ratio in time-matched public gene expression data. *Am J Respir Crit Care Med*. 2015. doi:10.1164/rccm.201507-1321LE
13. Haynes WA, Haddon DJ, Diep VK, et al. Integrated, multicohort analysis reveals unified signature of systemic lupus erythematosus. *JCI Insight*. 2020. doi:10.1172/jci.insight.122312
14. Andres-Terre M, McGuire HM, Pouliot Y, et al. Integrated, Multi-cohort Analysis Identifies Conserved Transcriptional Signatures across Multiple Respiratory Viruses. *Immunity*. 2015. doi:10.1016/j.immuni.2015.11.003
15. Mayhew MB, Buturovic L, Luethy R, et al. A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. *Nat Commun*. 2020. doi:10.1038/s41467-020-14975-w
16. Khatri P, Roedder S, Kimura N, et al. A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. *J Exp Med*. 2013. doi:10.1084/jem.20122709
17. Warsinske HC, Rao AM, Moreira FMF, et al. Assessment of Validity of a Blood-Based 3-Gene Signature Score for Progression and Diagnosis of Tuberculosis, Disease Severity, and Treatment Response. *JAMA Netw open*. 2018. doi:10.1001/jamanetworkopen.2018.3779
18. Sweeney TE, Azad TD, Donato M, et al. Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters. *Crit Care Med*. 2018. doi:10.1097/CCM.0000000000003084
19. Chen R, Khatri P, Mazur PK, et al. A meta-Analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. *Cancer Res*. 2014. doi:10.1158/0008-5472.CAN-13-2775
20. Warsinske H, Croda J, Liu AB, et al. Prospective validation of three-gene whole blood diagnostic for active tuberculosis predicts disease progression and response to treatment. *Am J Respir Crit Care Med*. 2018.

21. Sweeney TE, Haynes WA, Vallania F, Ioannidis JP, Khatri P. Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Res.* 2017. doi:10.1093/nar/gkw797
22. Giamarellos-Bourboulis EJ, Netea MG, Rovina N, et al. Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure. *Cell Host Microbe.* 2020. doi:10.1016/j.chom.2020.04.009
23. Andrews S. FastQC A Quality control tool for high throughput sequence data. *Babraham Bioinfo.* 2018.
24. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* 2011. doi:10.14806/ej.17.1.200
25. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019. doi:10.1093/nar/gky955
26. Pagès H, Carlson M, Falcon S, Maintainer NL. Package 'AnnotationDbi.' *Bioconductor Packag Maint.* 2017.
27. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2009. doi:10.1093/bioinformatics/btp616
28. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015. doi:10.1093/nar/gkv007
29. Hedges L V, Olkin I. *Statistical Methodology in Meta-Analysis.*; 1985.
30. Haynes WA, Vallania F, Liu C, et al. Empowering multi-cohort gene expression analysis to increase reproducibility. In: *Pacific Symposium on Biocomputing.* ; 2017. doi:10.1142/9789813207813_0015
31. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. *Contemp Clin Trials.* 2015. doi:10.1016/j.cct.2015.09.002
32. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B.* 1995. doi:10.1111/j.2517-6161.1995.tb02031.x
33. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005. doi:10.1073/pnas.0506580102
34. Vallania F, Tam A, Lofgren S, et al. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nat Commun.* 2018. doi:10.1038/s41467-018-07242-6
35. Bongen E, Vallania F, Utz PJ, Khatri P. KLRD1-expressing natural killer cells predict influenza susceptibility. *Genome Med.* 2018. doi:10.1186/s13073-018-0554-1
36. Roy Chowdhury R, Vallania F, Yang Q, et al. A multi-cohort study of the immune factors associated with M. tuberculosis infection outcomes. *Nature.* 2018. doi:10.1038/s41586-018-0439-x
37. Scott MKD, Quinn K, Li Q, et al. Increased monocyte count as a cellular biomarker for poor outcomes in fibrotic diseases: a retrospective, multicentre cohort study. *Lancet Respir Med.* 2019. doi:10.1016/S2213-2600(18)30508-3
38. Lagunas-Rangel FA. Neutrophil-to-lymphocyte ratio and lymphocyte-to-C-reactive protein ratio in patients with severe coronavirus disease 2019 (COVID-19): A meta-analysis. *J Med Virol.* 2020. doi:10.1002/jmv.25819
39. Liu J, Liu Y, Xiang P, et al. Neutrophil-to-lymphocyte ratio predicts critical illness patients with 2019 coronavirus disease in the early stage. *J Transl Med.* 2020;18(1):206. doi:10.1186/s12967-020-02374-0
40. Diao B, Wang C, Tan Y, et al. Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19). *Front Immunol.* 2020. doi:10.3389/fimmu.2020.00827
41. Qin C, Zhou L, Hu Z, et al. Dysregulation of immune response in patients with COVID-19 in Wuhan, China. *Clin Infect Dis.* 2020. doi:10.1093/cid/ciaa248
42. Hedges L V, Pigott TD. The power of statistical tests in meta-analysis. *Psychol Methods.* 2001;6(3):203-217.
43. Hopp M-T, Domingo-Fernández D, Gadiya Y, et al. Unravelling the debate on heme effects in COVID-19 infections. *bioRxiv.* January 2020:2020.06.09.142125. doi:10.1101/2020.06.09.142125
44. Monel B, Rajah MM, Hafirassou ML, et al. Atlantin Endoplasmic Reticulum-Shaping Proteins Facilitate

- Zika Virus Replication. *J Virol*. 2019. doi:10.1128/jvi.01047-19
45. Koeller DM, Casey JL, Hentze MW, et al. A cytosolic protein binds to structural elements within the iron regulatory region of the transferrin receptor mRNA. *Proc Natl Acad Sci U S A*. 1989. doi:10.1073/pnas.86.10.3574
 46. Dimopoulos G, de Mast Q, Markou N, et al. FAVORABLE ANAKINRA RESPONSES IN SEVERE COVID-19 PATIENTS WITH SECONDARY HEMOPHAGOCYTIC LYMPHOHISTIOCYTOSIS. *Cell Host Microbe*. 2020. doi:10.1016/j.chom.2020.05.007
 47. Bataille S, Pardinielli N, Bergougnioux J-P. Could ferritin help the screening for COVID-19 in hemodialysis patients? *Kidney Int*. 2020. doi:10.1016/j.kint.2020.04.017
 48. Ravelli A. Macrophage activation syndrome. *Curr Opin Rheumatol*. 2002. doi:10.1097/00002281-200209000-00012
 49. Inoue T, Tsai B. How viruses use the endoplasmic reticulum for entry, replication, and assembly. *Cold Spring Harb Perspect Biol*. 2013. doi:10.1101/cshperspect.a013250
 50. Haynes W, Vashisht R, Vallania F, et al. Integrated molecular, clinical, and ontological analysis identifies overlooked disease relationships. *bioRxiv*. 2018. doi:10.1101/214833
 51. Tomczak A, Mortensen JM, Winnenburger R, et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci Rep*. 2018. doi:10.1038/s41598-018-23395-2
 52. Huang C, Lokugamage KG, Rozovics JM, Narayanan K, Semler BL, Makino S. SARS coronavirus nsp1 protein induces template-dependent endonucleolytic cleavage of mRNAs: Viral mRNAs are resistant to nsp1-induced RNA cleavage. *PLoS Pathog*. 2011. doi:10.1371/journal.ppat.1002433
 53. Vivier E, Tomasello E, Baratin M, Walzer T, Ugolini S. Functions of natural killer cells. *Nat Immunol*. 2008. doi:10.1038/ni1582
 54. Sun S, Cai X, Wang H, et al. Abnormalities of peripheral blood system in patients with COVID-19 in Wenzhou, China. *Clin Chim Acta*. 2020. doi:10.1016/j.cca.2020.04.024
 55. Benoit M, Desnues B, Mege J-L. Macrophage Polarization in Bacterial Infections. *J Immunol*. 2008. doi:10.4049/jimmunol.181.6.3733
 56. Laura C Miller YS. Macrophage Polarization in Virus-Host Interactions. *J Clin Cell Immunol*. 2015. doi:10.4172/2155-9899.1000311

	COVID-19 Patients
n	62
Age in years: median [IQR] (n)	61 [52,70] (61)
Gender = Male (%)	40 (65)
SOFA score	2 [1,4] (61)
APACHE II	6.5 [4,9] (56)
Pneumonia severity index	89.5 [65,104.5] (48)
White blood cell (mm ³)	6180 [4910,8420] (59)
Neutrophils	75.5 [65.43,84.13] (59)
Lymphocytes	15.69 [10.5,22.55] (59)
Platelets (k/ mm ³)	195.2 [158.8, 238.8] (58)
Lactate (mmol/l)	1.55 [1.04,2.08] (30)
pO ₂ .FiO ₂ (mmHg)	255.35 [112.5,310.8] (50)
Creatinine (mg/dl)	0.9 [0.7,1.015] (58)
PCT (ng/ml)	0.1 [0.04,0.41] (49)
CRP (mg/l)	78.85 [29.48,175.8] (60)
Days btwn onset symptoms and sampling	6 [4,8] (53)
Days btwn intubation and sampling	1 [0.5, 1.5] (23)
Days btwn hospital admission and intubation	2 [1, 3.5] (23)

Table 1. Baseline characteristics table for COVID-19 patients. All continuous variables are reported as median and interquartile ranges [IQR] (n).

Accession	Platform	First Author	PMID	Timing of Diagnosis	Disease	Total Sample number	N Healthy Controls	N Viral	Age
GSE60244	GPL10558	Suarez NM	25637350	Within 24 h of admission	Respiratory viral infection	111	40	71	Adults
GSE40012	GPL6947	Parnell GP	22898401	On admission to ICU	H1N1 influenza A	24	18	8	Adults
GSE40396	GPL10558	Hu X	23858444	On hospitalization	Febrile children with viral infection	44	22	22	Pediatrics
GSE64456	GPL10558	Mahajan P	27552618	On hospitalization	Febrile children with viral infection	130	19	111	Infants
GSE42026	GPL6947	Herberg JA	23901082	On hospitalization	H1N1, RSV	74	33	41	Pediatrics
GSE67059	GPL6947	Heinonen S	26571305	Within 48h of admission/ ED	HRV +/- symptoms	101	21	80	Pediatrics
EMEXP3589	GPL10332	Almansa R	22852767	Within 24h of admission to ICU	Infected COPD in ICU with viral infections	9	4	5	Adults
GSE82050	GPL21185	Tang BM	28619954	Within 24 h of admission	Influenza	39	15	24	Adults
GSE68310	GPL10558	Zhai Y	26070066	Within 48 hours of acute respiratory infection onset	Influenza and other respiratory viral infections	347	243	104	Adults
GSE73461	GPL10558	Wright VJ	30083721	On presentation of symptoms	Viral infection	149	55	94	Pediatrics
GSE111368	GPL10558	Dunning J	29777224	Within 24 h of admission	seasonal flu study acute timepoints	163	130	33	Adults
GSE77087	GPL10558	de Steenhuijsen Pijters WA	27135599	Within 24h of hospitalization	RSV	59	18	41	Pediatrics
GSE66099	GPL570	Alder MN; Sweeney TE	27635771; 25972003	Admission to ICU	Viral infection	58	47	11	Pediatrics
GSE27131	GPL6244	Berdal J	21781987	On hospitalization	Severe Flu A	14	7	7	Adult
TOTAL						1324	672	652	

Table 2. 14 datasets used for discovery of the non-COVID-19 viral immune response.

Validation Datasets (Respiratory Viral)

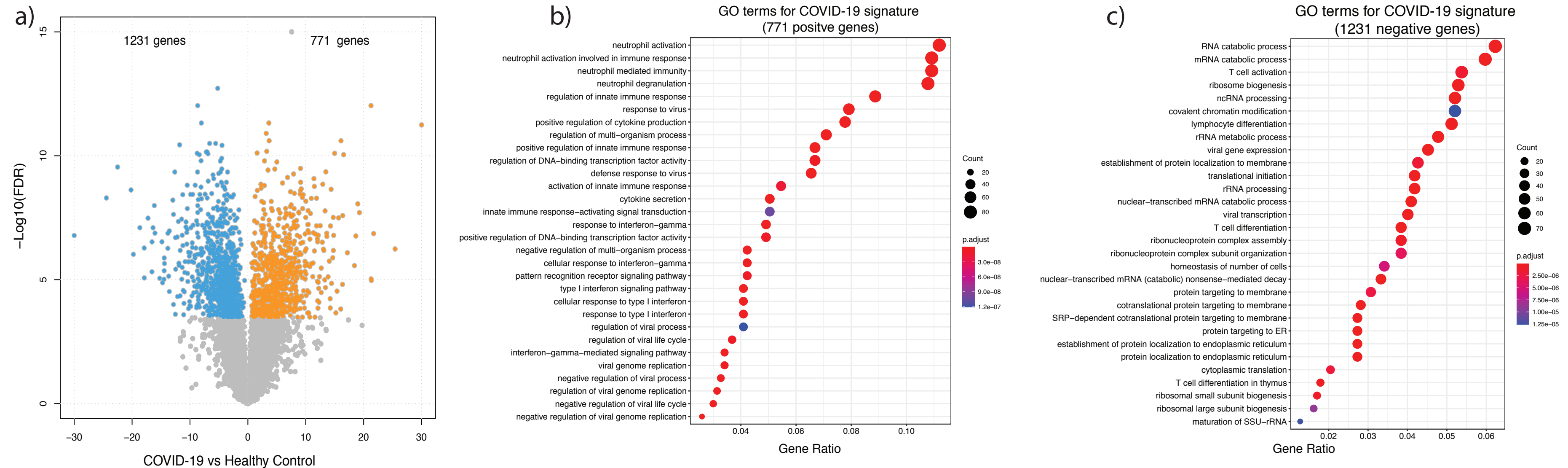
Accession	Platform	First Author	PMID	Timing of Diagnosis	Disease	Total Sample number	N Healthy Controls	N Viral	Age
GSE117827	GPL23126	Yu J	30339221	Within 24h of hospitalization	HRV	24	6	18	Pediatric
GSE20346	GPL6947	Parnell G	21408152	At peak symptoms	Influenza	37	18	19	Unknown
GSE34205	GPL570	Ioannidis I	22398282	Within 42-72h of hospitalization	Influenza/ RSV	101	22	79	Pediatric
GSE103842	GPL10558	Rodriguez-Fernandez R	29045741	Within 24h of hospitalization	RSV	74	12	62	Pediatric
TOTAL						236	58	178	

Validation Datasets (Other Viral)

Accession	Platform	First Author	PMID	Timing of Diagnosis	Disease	Total Sample number	N Healthy Controls	N Viral	Age
GSE5972	GPL4387	Cameron MJ	17537853	Within 24h of hospitalization	SARS (CoV1) *	64	10	54	Adults
GSE122692	GPL16686	Reynard S	30626757	Within 24h of hospitalization	Ebola *	45	8	37	Adults
EMTAB3162	GPL570	van de Weg CA	25768297	On admission	Dengue	36	15	21	Adults and Pediatric
GSE51808	GPL13158	Kwissa M	24981333	On admission	Dengue	37	9	28	Adults and Pediatric
GSE38246	GPL15615	Popper SJ	23285306	Within 24h of hospitalization	Dengue *	113	8	105	Pediatric
TOTAL						295	50	245	

Table 3. Datasets for validation of the non-COVID-19 viral vs healthy signature.

* indicates datasets not eligible for COCONUT



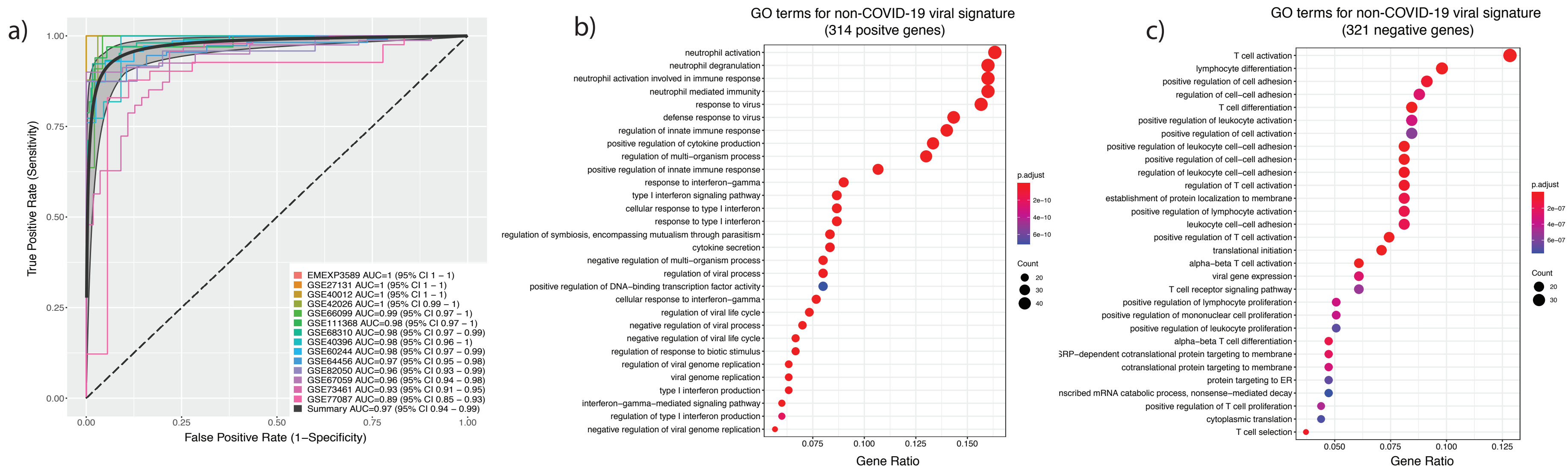


Figure 2. MetaIntegration of 14 non-COVID-19 viral disease datasets. (a) Multicohort analysis of 14 datasets of viral infections (n = 652) and healthy controls (n = 672) identified 635 non-COVID-19 viral signature, including 314 positively regulated genes and 321 negatively regulated genes at the chosen cutoff of $ES \geq 1$ or ≤ -1 with $FDR \leq 0.05\%$. GO term enrichment analysis of positive (b) and negative (c) gene sets reveal increased neutrophil function enrichment and decreased T cell related pathways, similar to those in Figure 1 (Gene Ratios represent the number of genes in our gene set within that pathway).

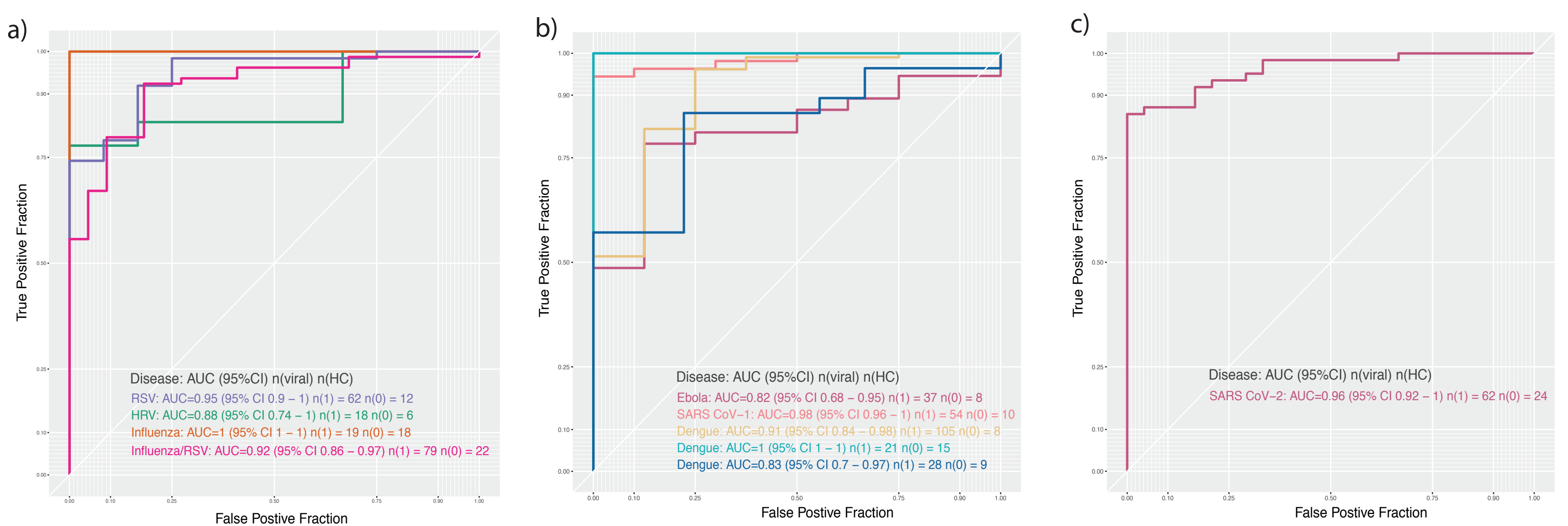


Figure 3. Validation of a global host immune response to viral infections. (a) ROC performance of 635 non-COVID-19 signature in 4 independent datasets including HRV, RSV, picornavirus and influenza. (b) ROC performance in 5 additional cohorts of disease not included in discovery [Ebola (GSE122692), SARS CoV-1 (GSE5972) and Dengue (GSE38246, EMTAB3162, GSE51808)]. (c) The signature is also tested in the 62 COVID-19 patients and 24 HCs.

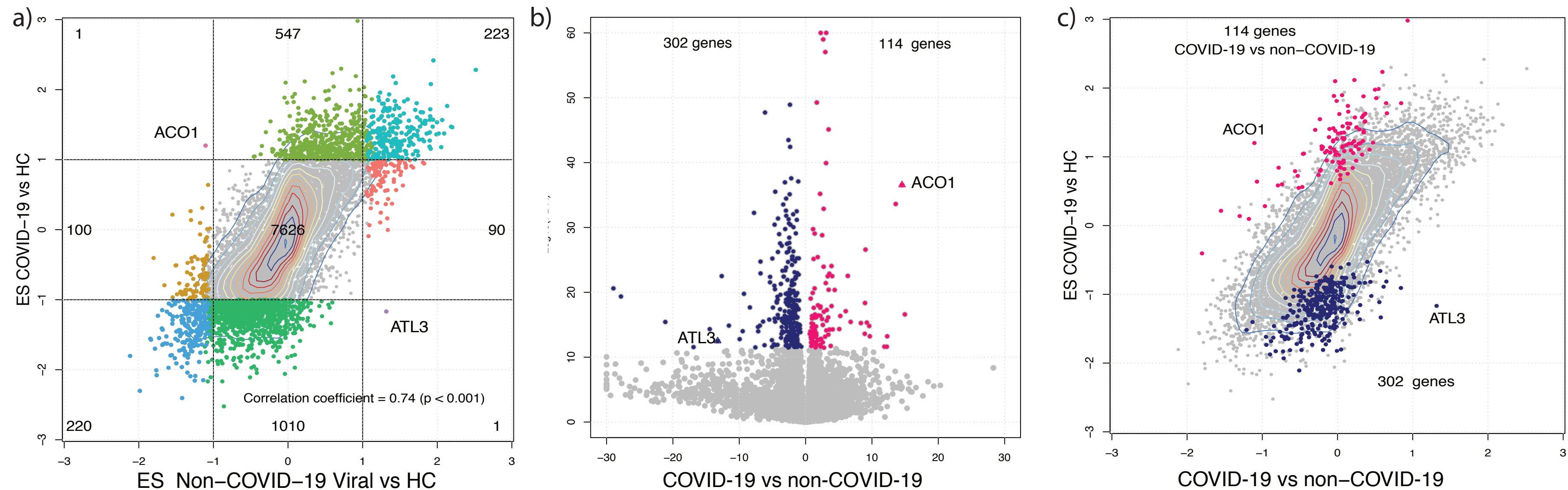
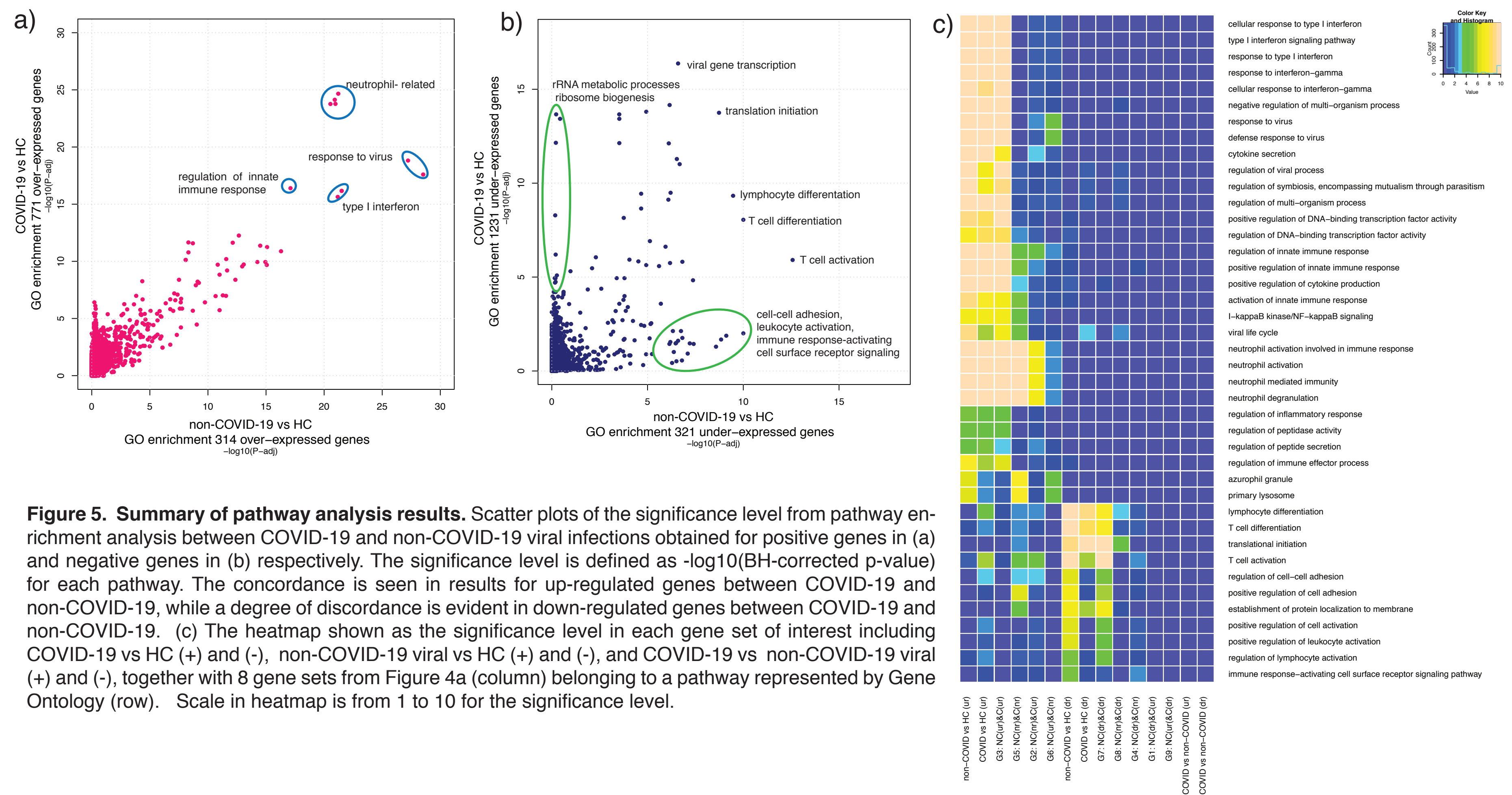
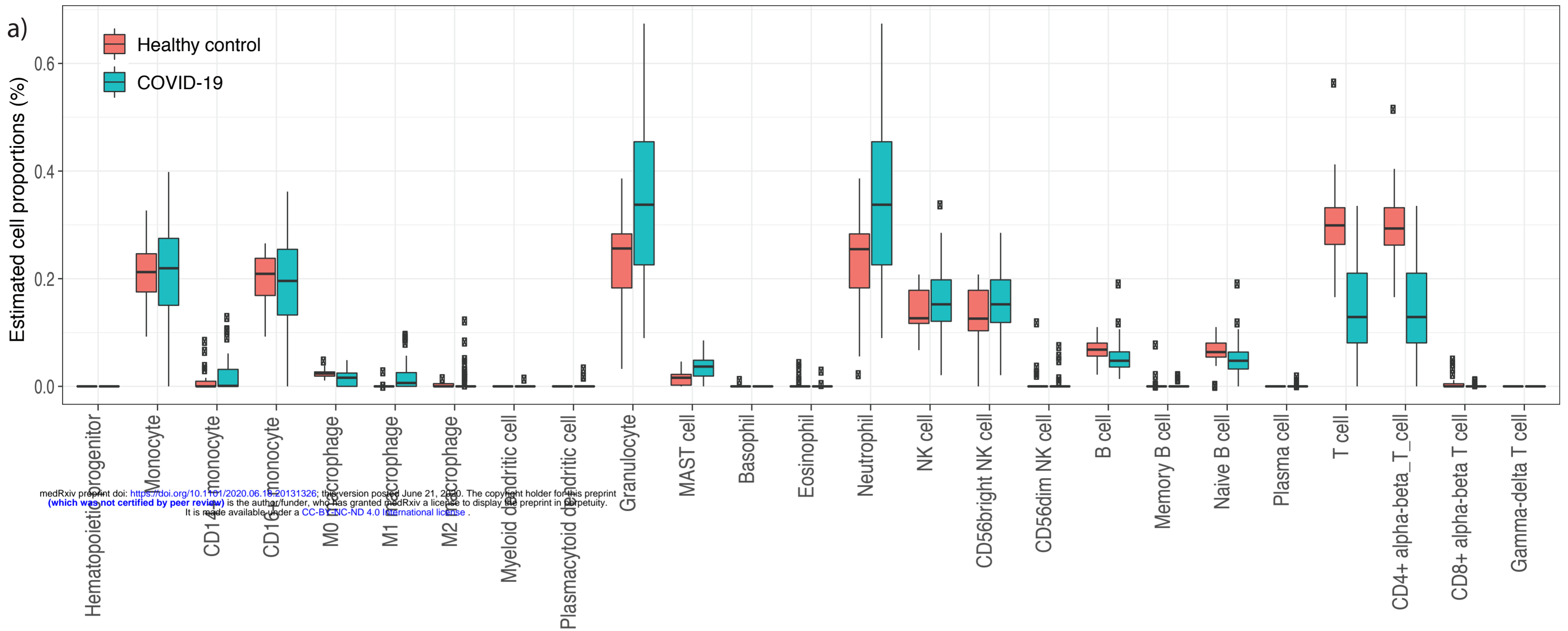


Figure 4. Comparison of COVID-19 signature with non-COVID-19 signature. (a) Scatter plot of effect size for all 9,818 genes commonly present in all datasets between non-COVID-19 vs HC (x-axis) and COVID-19 vs HC (y-axis). 2,002 COVID-19 signature genes from Figure 1 and 635 non-COVID-19 signature genes from Figure 2 are colored distinguishably in 9 quadrats. Concordant host response between COVID-19 and other viral infections is reflected by 223 commonly positively and 220 negatively regulated genes in both. Discordant response is only seen in ACO1 whose expression is positively regulated in COVID-19 but negatively regulated in non-COVID, and in ATL3 whose expression is negatively regulated in COVID-19 but positively regulated in non-COVID-19. (b) The head-to-head comparison between COVID-19 and other viral infections was made possible by using co-normalized data by COCONUT. Significance score [defined as $-\log_{10}(\text{FDR})$] vs mean difference of co-normalized log₂-transformed expression data between COVID-19 patients (n = 62) vs other viral infections (n = 652). The chosen cutoff of $\text{ES} \geq 1$ or ≤ -1 with $\text{FDR} \leq 0.05\%$ yields 416 COVID-19 specific signature, including 114 positively regulated genes and 302 negatively regulated genes. (c) To illustrate the overlap of (a) and (b), the 416 COVID-19 specific signature genes from head-to-head comparison in (b) are shown in the same scatter plot in (a).





medRxiv preprint doi: <https://doi.org/10.1101/2020.06.18.20131326>; this version posted June 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

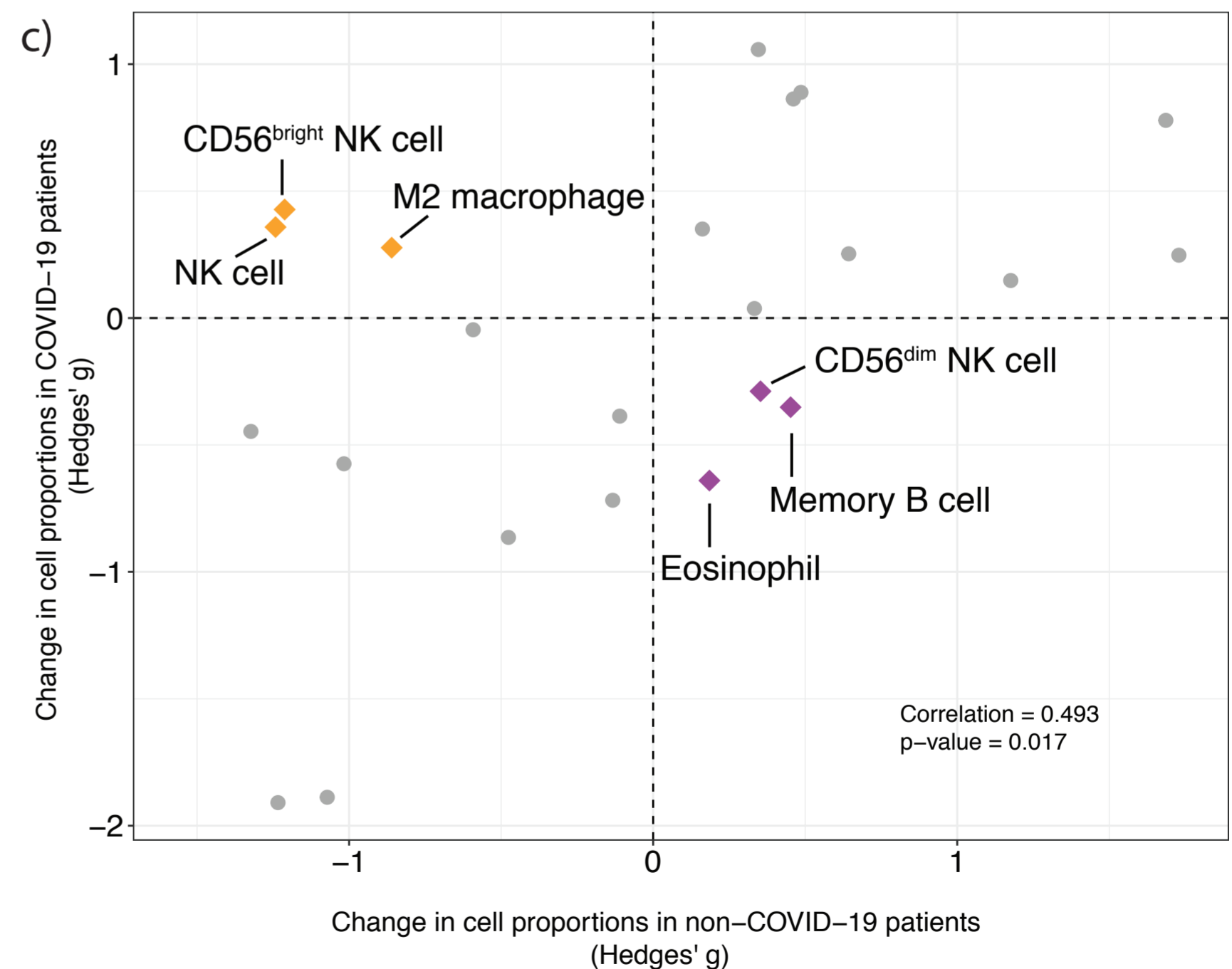
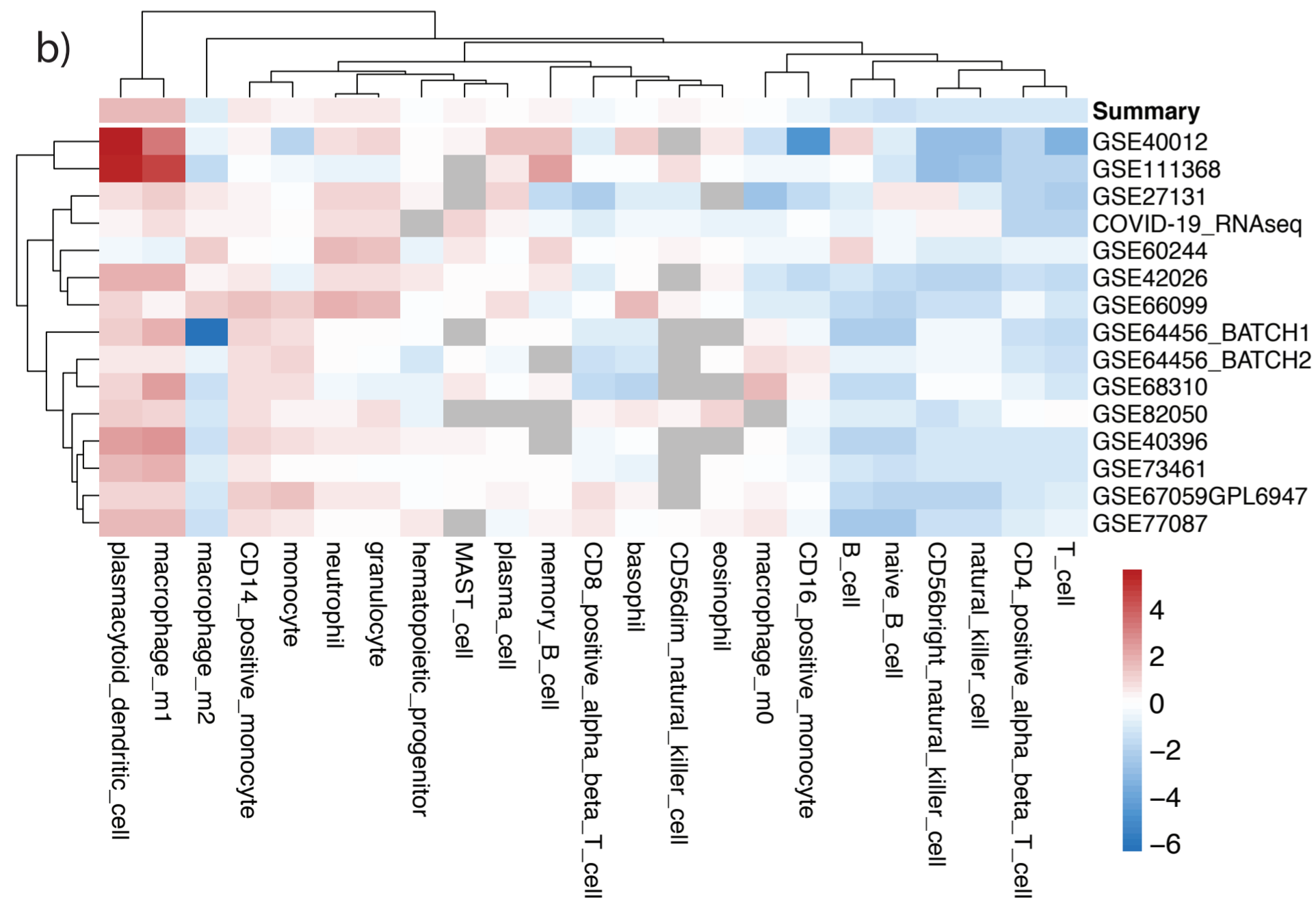


Figure 6. Statistical deconvolution of bulk transcriptome profiles using immunoStates of COVID-19 versus non-COVID-19 viral infections. a) Changes in cell proportions when comparing COVID-19 patients to healthy controls. Note the trends of increased neutrophil and decreased T cell proportions. b) Heatmap of changes in cell proportions of all datasets: non-COVID-19 and COVID-19 c) Concordant and discordant changes in cellular proportions comparing COVID-19 to non-COVID-19 viral infections. Cell types that increased in COVID-19 (hence decreased in non-COVID-19) were CD56^{bright} NK cells, M2 macrophages, and total NK cells. Those that decreased in non-COVID-19 but increased in COVID-19 were CD56^{dim} NK cells, memory B cells, and eosinophils.