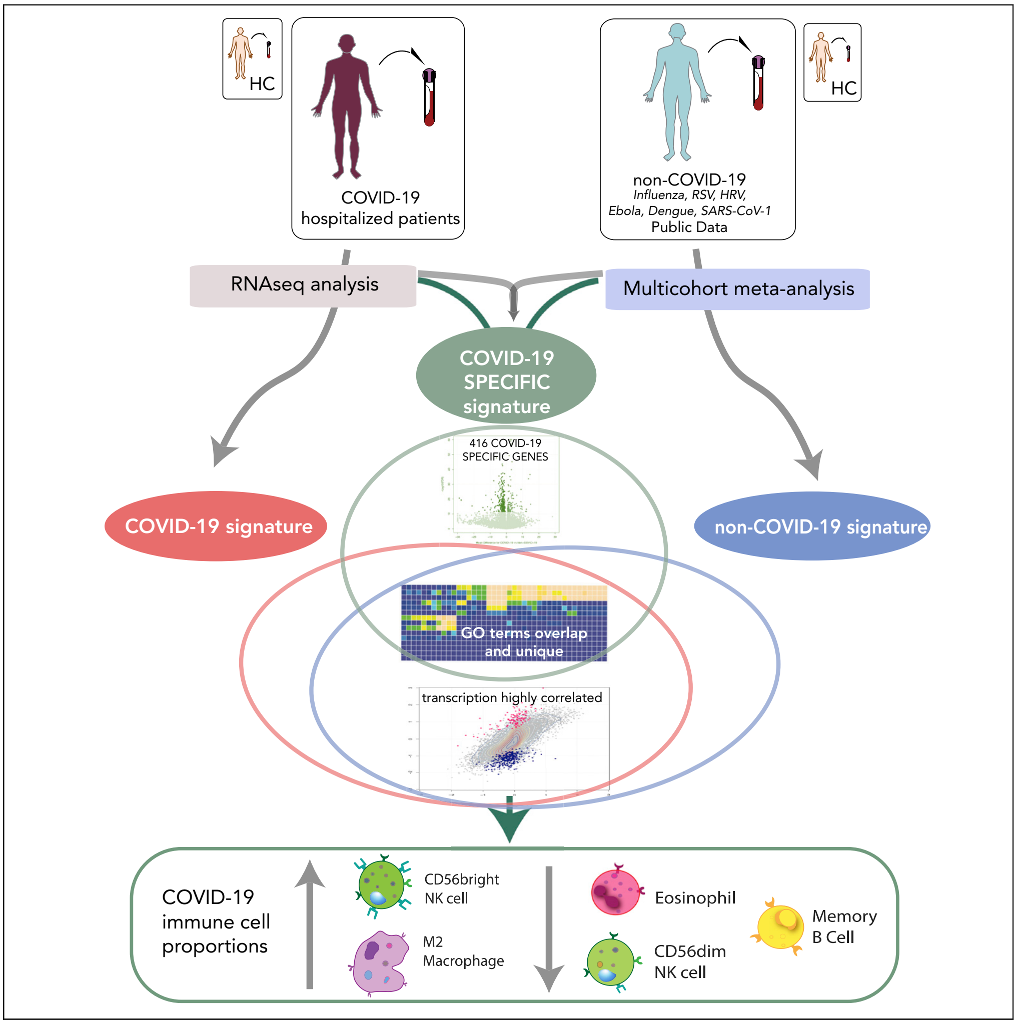
iScience

CellPress
OPEN ACCESS

Article

Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections



Simone A. Thair,
Yudong D. He,
Yehudit Hasin-Brumshtein, ...,
Purvesh Khatri,
Evangelos J. Giamarellos-Bourboulis,
Timothy E. Sweeney

tsweeney@inflammatix.com

Highlights

Whole blood transcriptomics were generated via RNAseq for 62 COVID-19 patients

Curated 23 whole blood transcriptomic studies (1855 samples) of non-COVID-19 viral infections

Discovered 416 COVID-19-specific genes, despite overall correlation with non-COVID-19

Revealed subset of immune cell proportions discordantly shifted in COVID-19 infections

Thair et al., iScience *24*, 101947
January 22, 2021 © 2020 The Authors.
https://doi.org/10.1016/j.isci.2020.101947

iScience

CellPress OPEN ACCESS

Article

Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections

Simone A. Thair,[1,9] Yudong D. He,[1,9] Yehudit Hasin-Brumshtein,[1] Suraj Sakaram,[1] Rushika Pandya,[1] Jiaying Toh,[2,3] David Rawling,[1] Melissa Remmel,[1] Sabrina Coyle,[1] George N. Dalekos,[4] Ioannis Koutsodimitropoulos,[5] Glykeria Vlachogianni,[6] Eleni Gkeka,[7] Eleni Karakike,[8] Georgia Damoraki,[8] Nikolaos Antonakos,[8] Purvesh Khatri,[2,3,10] Evangelos J. Giamarellos-Bourboulis,[8,9,10] and Timothy E. Sweeney[1,9,10,11,*]

Summary

The pandemic 2019 novel coronavirus disease (COVID-19) shares certain clinical characteristics with other acute viral infections. We studied the whole-blood transcriptomic host response to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) using RNAseq from 24 healthy controls and 62 prospectively enrolled patients with COVID-19. We then compared these data to non-COVID-19 viral infections, curated from 23 independent studies profiling 1,855 blood samples covering six viruses (influenza, respiratory syncytial virus (RSV), human rhinovirus (HRV), severe acute respiratory syndrome coronavirus 1 (SARS-CoV-1), Ebola, dengue). We show gene expression changes in COVID-19 versus non-COVID-19 viral infections are highly correlated ($r = 0.74$, $p < 0.001$). However, we also found 416 genes specific to COVID-19. Inspection of top genes revealed dynamic immune evasion and counter host responses specific to COVID-19. Statistical deconvolution of cell proportions maps many cell type proportions concordantly shifting. Discordantly increased in COVID-19 were CD56^{bright} natural killer cells and M2 macrophages. The concordant and discordant responses mapped out here provide a window to explore the pathophysiology of the host response to SARS-CoV-2.

Introduction

A novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has developed into a global pandemic, resulting in more than 47.9 million cases and 1,221871 deaths across 235 countries as we write (WHO, accessed 5 Nov 2020) (Zhou et al., 2020). Contextually, this pandemic has surpassed the severe acute respiratory syndrome coronavirus 1 (SARS-CoV-1) 2003 pandemic by almost 6000-fold in total cases whereby SARS-CoV-1 resulted in 8,098 cases, took 12 months to contain, and had a 9.6% mortality rate (World Health Organization (WHO) accessed 1 Jun 2020). The novel SARS-CoV-2 virus, the causative agent for 2019 novel coronavirus disease (COVID-19), is highly communicable and despite urgent and resource-intensive efforts globally, we have no proven vaccine or efficacious treatment available (Callaway, 2020).

Early in a pandemic, it is imperative to understand what is similar in the host response to the novel virus when compared to other known viruses in order to rapidly rule in or rule out recyclable treatments and/or vaccination strategies. At the same time, it is also critical to understand the differences in this disease in order to search for novel therapeutics. The human immune system has evolved over millions of years to protect the host from microbes (Medzhitov, 2007; Longo et al., 2015). Understanding the overlap, or lack thereof for the most basic immunological features such as the virus's ability to inhibit the interferon response or to infect host cells with an antibody-dependent infection enhancement, can drive medicine rapidly in a life-saving direction (Jaume et al., 2012; Wang et al., 2016; Mesev et al., 2019; Blanco-Melo et al., 2020). In the last decade alone, we have already responded to pandemics of H1N1, chikungunya, Zika, and near-pandemics of two other coronaviruses, SARS-CoV-1 and Middle East respiratory syndrome-related coronavirus (MERS), from which valuable insights can be applied (Morens and Fauci,

[1] Inflammatix, Inc., 863 Mitten Road, 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

[9] These authors contributed equally

[10] Senior authors

[11] Lead contact

*Correspondence: tsweeney@inflammatix.com

https://doi.org/10.1016/j.isci.2020.101947

2020). COVID-19 clearly shares 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 studies (Drakesmith and Prentice, 2008; Blanco-Melo et al., 2020; Catanzaro et al., 2020; Wilson et al., 2020). 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 rate (Tay et al., 2020).

The COVID-19 pandemic has resulted in the halt of normal life across the globe in an attempt to slow the spread of the virus. Computational methods leveraging data generated prior to the pandemic present an advantage to push forward the aforementioned knowledge discovery. Studies comparing COVID-19 to healthy controls (HCs) are useful; however, they do not explain the similarities and differences seen in the COVID-19 syndrome versus other viral infections; hence, we have leveraged our multicohort, conormalization method to execute a head-to-head comparison of COVID-19 to non-COVID-19 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 data sets 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 infections (Sweeney et al., 2015, 2016b, 2018a), of note viral infections (Andres-Terre et al., 2015) but also in autoimmune diseases, vaccination, tuberculosis, cancer, and organ transplant (Li et al., 2012; Khatri et al., 2013; Chen et al., 2014; Andres-Terre et al., 2015; Sweeney et al., 2015, 2016a, 2016b, 2018a, 2018b; Sweeney and Khatri, 2015; Warsinske et al., 2018a, 2018b; Haynes et al., 2020; Mayhew et al., 2020). We have shown in methodological work that this method produces results with the greatest reproducibility in independent cohorts (Sweeney et al., 2017).

In this work, we used RNAseq to profile whole blood samples from 62 patients with COVID-19 prospectively enrolled in Athens, Greece, together with 24 HCs. 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, respiratory syncytial virus (RSV), human rhinovirus (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 and demonstrate that a large portion of the response is in fact similar to previous viral infections. This is immensely valuable as it demonstrates that it is this conserved host response that allows for pandemic preparedness and response. Our implementation of computational methods comparing SARS-CoV-2 to known circulating viruses yields a COVID-19-specific gene signature for differentiating the host response, which warrants further investigation.

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 HCs (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, Table S2A) with absolute Hedges' g effect size (ES) which is the difference between groups as a proportion of variability in the groups (Hedges' g ES) ≥ 1 and false discovery rate (FDR) ≤0.05%), referred to as the "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 pneumonia (Diao et al., 2020; Lagunas-Rangel, 2020; Liu et al., 2020; Qin et al., 2020).

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

Based on our previous results (Andres-Terre et al., 2015), 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

Table 1. Baseline characteristics table for patients with COVID-19

Characteristic	Patients with COVID-19
N	62
Age in years: median [IQR] (n)	61 [52,70] (61)
Gender = male (%)	40 (65)
SOFA (sequential organ failure assessment) score	2 [1,4] (61)
APACHE II (Acute Physiology And Chronic Health Evaluation II)	6.5 [4,9] (56)
Pneumonia severity index	89.5 [65,104.5] (48)
White blood cell (mm3)	6180 [4910,8420] (59)
Neutrophils	75.5 [65.43,84.13] (59)
Lymphocytes	15.69 [10.5,22.55] (59)
Platelets (k/mm3)	195.2 [158.8, 238.8] (58)
Lactate (mmol/L)	1.55 [1.04,2.08] (30)
pO2.FiO2 (mmHg)	255.35 [112.5,310.8] (50)
Creatinine (mg/dL)	0.9 [0.7,1.015] (58)
PCT (procalcitonin) (ng/mL)	0.1 [0.04,0.41] (49)
CRP (C-reactive protein) (mg/L)	78.85 [29.48,175.8] (60)
Days between onset symptoms and sampling	6 [4,8] (53)
Days between intubation and sampling	1 [0.5, 1.5] (23)
Days between hospital admission and intubation	2 [1, 3.5] (23)

All continuous variables are reported as median and interquartile ranges (IQRs) (n).

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 analysis (Hedges and Pigott, 2001) found that even with high inter-study heterogeneity, we had more than 80% statistical power at p value = 0.01 for detecting absolute Hedges' g ES > 0.43 in these data sets (Figure S2). The multi-cohort analysis of 1,324 transcriptome profiles (652 patients with non-COVID-19 viral infections, 672 HCs) from these 14 studies using MetaIntegrator (Haynes et al., 2017) identified 635 differentially expressed genes (314 over-expressed, 321 under-expressed). The area under the curve (AUC) of a receiver operator characteristics (ROC) curve represents the discriminatory ability of the score to correctly identify true positive and/or true negatives. The closer to 1 the value is, the better the performance of the test, for example, a test that can discriminate if a patient or sample is virally infected or healthy. ROC plots for all of the discovery data sets 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, Table S2A). 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 (Figures 2B and 2C).

Validation of host response genes to viral infections in multiple independent data sets

Next, we confirmed that the non-COVID-19 viral signature is conserved across viruses by validating it in several independent data sets. 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 HCs; 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 HCs, 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 HCs with high accuracy (Figure 3B). Third, we tested whether the non-COVID-19 viral signature would also distinguish patients with COVID-19 from HCs. We calculated the non-COVID-19 viral score for each of 62 patients with COVID-19 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).

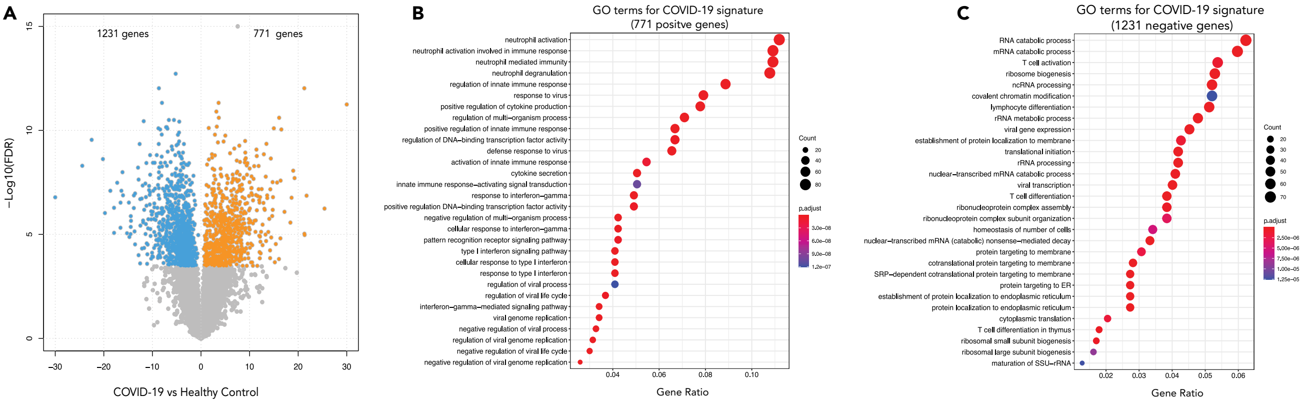


Figure 1. RNA-seq data for patients with COVID-19 versus healthy control and pathway analysis of the COVID-19 signature

(A–C) (A) Significance score [defined as -log10(FDR)] versus mean difference of co-normalized log2-transformed expression data between patients with COVID-19 (n = 62) and healthy controls (n = 24). The chosen cutoff of ES ≥ 1 or ≤ −1 with FDR ≤0.05% yields the 2,002 COVID-19 signatures, including 771 positively regulated genes and 1,231 negatively regulated genes. GO term enrichment analysis of positive (B) and negative (C) gene sets reveals increased neutrophil function enrichment and decreased T-cell-related pathways (gene ratios represent the number of genes in our gene set within that pathway). The gene ratio (x axis) is the ratio of the number of genes in our data enriched in a given gene set (pathway) to the total number of genes in that pathway.

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 HCs across 9,818 genes that were present across all data sets. 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 (Hedges' 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 scatterplot by different colors to highlight their relationships (Figure 4A and Table S2A). 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 (Hedges' g ES ≥ 1, FDR ≤0.05%), as well as 220 genes concordantly under-expressed with (Hedges' g ES ≤ −1, FDR ≤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*) is over-expressed in COVID-19 and under-expressed in non-COVID-19 viral infections and Atlastin GTPase 3 (*ATL3*) is 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 pathophysiology (Hopp et al., 2020). *ATL3* is required for endoplasmic reticulum (ER) membrane junctions and may be linked to viral replication sites (Monel et al., 2019).

Therefore, in order to identify a statistically significant set of genes differentially expressed in patients with COVID-19 compared to those with other viral infections, we employed COCONUT to conormalize the two disease types into a single matrix for comparison of 62 patients with COVID-19 versus 652 patients with non-COVID-19 viral infection. Conormalization with COCONUT allows for pooling of data across data sets while simultaneously removing batch-to-batch technical variance in a bias-free manner (Sweeney et al., 2016b). At Hedges' g | ES| ≥ 1 with FDR ≤0.05%, we found 416 genes we refer to as the "COVID-19-specific gene signature", 114 over-expressed and 302 under-expressed in patients with COVID-19 than in those with non-COVID-19 viral infection (Figures 4B, Tables S2A and S2B). 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) is highlighted in the same scatterplot from panel a (Figure 4C).

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

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 hr of admission	Respiratory viral infection	111	40	71	Adults
GSE40012	GPL6947	Parnell GP	22898401	On admission to intensive care unit (ICU)	H1N1 influenza A	24	18	8	Adults
GSE40396	GPL10558	Hu X	23858444	On hospitalization	Febrile children with viral infection	44	22	22	Infants
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	Children
GSE67059	GPL6947	Heinonen S	26571305	Within 48 hr of admission to emergency department(ED)	HRV +/− symptoms	101	21	80	Infants
EMEXP3589	GPL10332	Almansa R	22852767	Within 24 hr of admission to ICU	Infected chronic obstructive pulmonary disease (COPD) in ICU with viral infections	9	4	5	Adults
GSE82050	GPL21185	Tang BM	28619954	Within 24 hr of admission	Influenza	39	15	24	Adults
GSE68310	GPL10558	Zhai Y	26070066	Within 48 hr 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	Children
GSE111368	GPL10558	Dunning J	29777224	Within 24 hr of admission	Seasonal flu study, acute timepoints	163	130	33	Adults
GSE77087	GPL10558	de Steenhuijsen Piters WA	27135599	Within 24 hr of hospitalization	RSV	59	18	41	Infants
GSE66099	GPL570	Alder MN; Sweeney TE	27635771; 25972003	Admission to ICU	Viral infection	58	47	11	Children
GSE27131	GPL6244	Berdal J	21781987	On hospitalization	Severe flu A	14	7	7	Adult
Total						1324	672	652	

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).

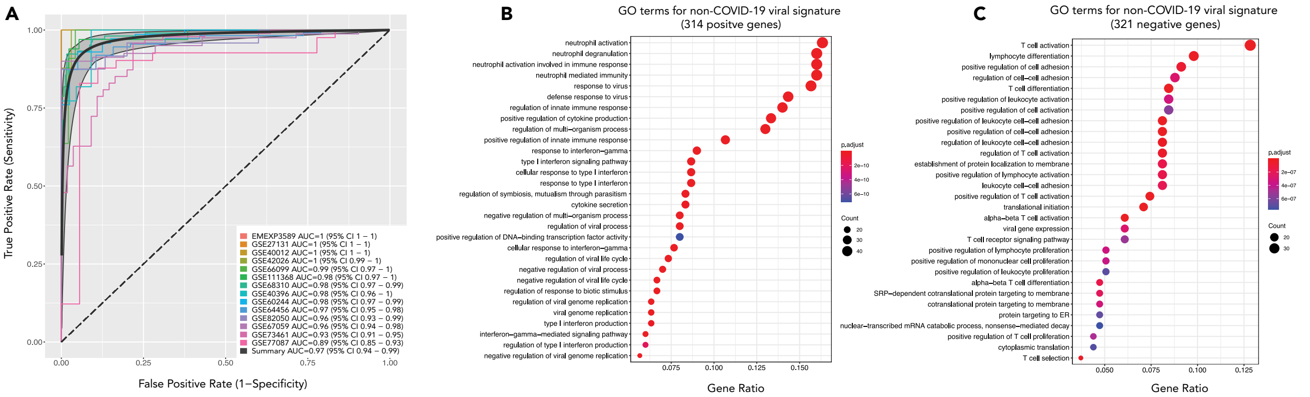


Figure 2. Metaintegration of 14 non-COVID-19 viral disease data sets and pathway analysis of non-COVID-19 signature genes

(A–C) (A) ROC plots of the 635 non-COVID-19 viral signatures discovered using multicohort analysis with a cutoff of ES ≥ 1 or ≤ −1 and FDR ≤0.05% resulting in 314 positively regulated genes and 321 negatively regulated genes then plotted individually for each of the 14 data sets of viral infections (n = 652) and healthy controls (n = 672) identified. The consistent and high AUC values indicate that the signature is representative of all data sets, thereby embracing the heterogeneity which will increase generalizability. GO term enrichment analysis of positive (B) and negative (C) gene sets reveals 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).

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 gene lists from the 9 groups by quadrant in Figure 4A (Figure 5C, Table S2A). 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 (Table S3, Figure S3), 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 ≤10%) and lymphoid cells (CD4+ and CD8+ alpha-beta T cells, B cells) decreased significantly (FDR ≤10%) during viral infection (Figure 6A, Table S4). These results are in line with recent reports demonstrating increased neutrophil and decreased T-cell counts in patients with COVID-19 (Diao et al., 2020; Liu et al., 2020; Qin et al., 2020). 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) (Figures 6B and S4). Indeed, when considering changes within each data set, 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 (Figures 6C, Table S4), 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 H1N1[20]. Cell types that increased in COVID-19 relative to non-COVID-19 were CD56^{bright} natural killer (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

Table 3. Data sets for validation of the non-COVID-19 viral versus healthy signature

Accession	Platform	First author	PMID	Timing of diagnosis	Disease	Total sample number	N healthy controls	N viral	Age
GSE117827	GPL23126	Yu J	30339221	Within 24 hr of hospitalization	HRV	24	6	18	Children
GSE20346	GPL6947	Parnell G	21408152	At peak symptoms	Influenza	37	18	19	Unknown
GSE34205	GPL570	Ioannidis I	22398282	Within 42–72 hr of hospitalization	Influenza/ RSV	101	22	79	Infants
GSE103842	GPL10558	Rodriguez-Fernandez R	29045741	Within 24 hr of hospitalization	RSV	74	12	62	Infants
Total						236	58	178	
GSE5972	GPL4387	Cameron MJ	17537853	Within 24 hr of hospitalization	SARS (CoV1)	64	10	54	Adults
GSE122692	GPL16686	Reynard S	30626757	Within 24 hr of hospitalization	Ebola	45	8	37	Adults
EMTAB3162	GPL570	van de Weg CA	25768297	On admission	Dengue	36	15	21	Adults and children
GSE51808	GPL13158	Kwissa M	24981333	On admission	Dengue	37	9	28	Adults and children
GSE38246	GPL15615	Popper SJ	23285306	Within 24 h of hospitalization	Dengue	113	8	105	Children
Total						295	50	245	

[a]indicates data sets not eligible for COCONUT

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 populations[20].

Discussion

Understanding the pathophysiology of COVID-19 is critical to finding new treatments. Defining the portion of the host response to a novel pandemic virus that is similar to current circulating viral infections is imperative as treatment options are unknown and vaccines non-existent in the early months and thus repurposing drugs that have passed the United States Food and Drug Administration (FDA) safety trials can potentially be informed here. Simultaneously, identifying the biology of the host response that is not similar to circulating viruses may help rank the order with which drugs are repurposed if they do not bolster areas of the immune system succumbing to a virus for which we have no direct immune memory or offer novel targets for new drugs. 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, valuable information under pandemic circumstances, our study highlights some key differences.

The scatterplot of the correlation of the differential expression (relative to HCs) 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 using this method. *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). Viral replication can occur in infected cells due to a hinderance of the function of the immune cells drawn in to kill infected cells; as well, there are reports of SARS-CoV-1 and SARS-CoV-2 directly infecting immune cells themselves (Gu et al., 2005; Hu et al., 2012; Pontelli et al., 2020). As our data are from whole blood RNA, we cannot conclude precisely which of these mechanisms are responsible for the shifts in these genes' expression; however, prior reports suggest that both genes may be involved in viral replication

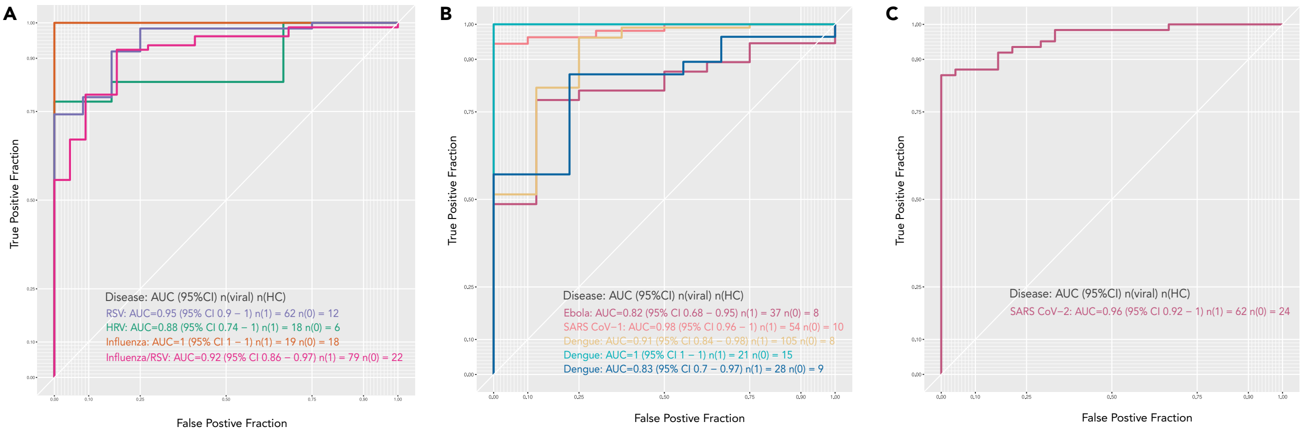


Figure 3. Validation of a global host immune response to viral infections

(A) ROC performance of 635 non-COVID-19 signatures in 4 independent respiratory viral infection data sets including HRV, RSV, picornavirus, and influenza.
(B) ROC performance in 5 additional cohorts of other viral infections to illustrate that this signature is broadly applicable to many viruses [Ebola (GSE122692), SARS CoV-1 (GSE5972), and dengue (GSE38246, EMTAB3162, GSE51808)].
(C) The signature is also tested in the 62 patients with COVID-19 and 24 HCs.

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, 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 proliferation (Koeller et al., 1989). High levels of ferritin are also indicative of macrophage activation syndrome and have been observed in patients with COVID-19 (Ravelli, 2002; Bataille et al., 2020; Dimopoulos et al., 2020; Giamarellos-Bourboulis et al., 2020). *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 destruction (Inoue and Tsai, 2013). Lack of *ATL3* 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 Zika (Monel et al., 2019), although Zika was not studied here.

The power of using COCONUT to combine heterogeneous data sets allowed for a pooled, head-to-head comparison of COVID-19 with non-COVID-19 viral infections, resulting in a 416 gene "COVID-19-specific gene signature" (Table S2B). 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 to those that are heavily annotated and studied, often referred to as the "streetlight effect", so absence of evidence does not denote evidence of absence of coordinated differential response (Haynes et al., 2018a, 2018b; Tomczak et al., 2018). Conversely, this novel combination of genes with these particular effect sizes warrants further investigation as a potential route for novel discoveries (Damelin et al., 2017; Haynes et al., 2018b). Simply reviewing what is known of the immunological function of the top two over- and top two under-expressed genes ranked by Hedges' g ES contextualizes *ACO1* and *ATL3* further with hints of a battle of host versus "novel" pathogen, never encountered by the immune system before. The impact on the function of host immune cells during SARS-CoV-1 and MERS infection is driven by their non-structural proteins and affects the normal production of cytokines compared to that of currently circulating viral infections, such as the repression of interferon proteins/ interferons (IFNs) (Hu et al., 2012; Shah et al., 2020). Recently, Blanco-Melo et al. revealed a dysregulated host response indicative of reduced innate antiviral defenses coupled with excessive cytokine production using cell lines, ferrets, and correlating with two deceased patients with COVID-19 (Blanco-Melo et al., 2020), a phenomenon of novel virus escape mechanisms from host defenses, of which we complement here with even larger numbers of entirely human data.

The most under-expressed gene in the "COVID-19-specific gene signature" is ZC3H13. Knocking this gene down was associated with less RNA methylation N6-methyladenosine (m^6A), an epigenetic modification commonly found in the viral RNA genomes of hepatitis C virus (HCV), Zika, dengue, yellow fever, and West Nile virus (Wen et al., 2018). Depletion of m^6A methyltransferases increase HVC viral particle

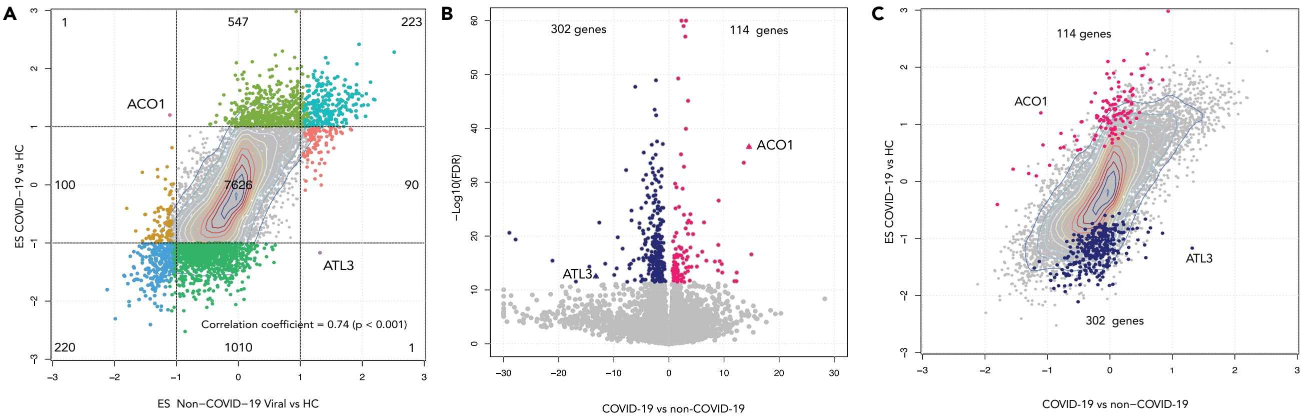


Figure 4. Comparison of COVID-19 signature with non-COVID-19 signature

(A) Scatterplot of effect size for all 9,818 genes commonly present in all data sets between non-COVID-19 vs HC (x axis) and COVID-19 vs HC (y axis). Two thousand two COVID-19 signature genes from Figures 1 and 635 non-COVID-19 signature genes from Figure 2 are overlayed and colored, each of the 9 quadrants have a different color to allow for easy visualization of the overlap of Hedges' g ES from each signature. For example, teal in the top right quadrant are the genes that have an Hedges' g ES ≥ 1 for both the 2,002 COVID-19 signature genes and the 635 non-COVID-19 signature genes. Concordant host response between COVID-19 and other viral infections is reflected by 223 commonly positively (teal, top right) and 220 negatively (blue, bottom left) 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) Using COCONUT conormalized data combined with a head-to-head comparison of COVID-19 and non-COVID-19 viral infections using Hedges' g ES ≥ 1 or ≤ −1 with FDR ≤0.05% yields 416 COVID-19-specific signatures, including 114 positively regulated genes and 302 negatively regulated genes. Significance score [defined as -log10(FDR)] vs mean difference of co-normalized log2-transformed expression data between patients with COVID-19 (n = 62) vs other viral infections (n = 652).

(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 scatterplot in (A).

production (Gokhale et al., 2016), which would imply more SARS-CoV-2 viral replication. *ATL3* as mentioned is also included in the "COVID-19-specific gene signature" and is under-expressed in COVID-19. When *ATL3* was knocked down, there was less Zika replication, implying that the under-expression is a host counteractive protective mechanism. The second most under-expressed gene is *AMIGO1*, a gene for which very little is known; however, recent studies on this family of genes (Kuja-Panula et al., 2003) suggest a cell adhesion function. Cell adhesion molecules are a key component of combatting pathogen infections, without which the host may not mount an appropriate response (Etzioni, 1996). Since the "COVID-19-specific gene signature" is derived from direct comparison of COVID-19 versus non-COVID-19 infections, *ZC3H13*, *AMIGO1*, and *ATL3* under-expressed in COVID-19 equates to higher expression in non-COVID-19 infections. One possible interpretation of this under-expression of *ZC3H13* and *AMIGO1* in COVID-19 that could be investigated in future studies is that this novel virus may be inhibiting their expression to escape the host responses that are otherwise functional for previously circulating viral infections.

If indeed the under-expression of *ATL3* in the "COVID-19-specific gene signature" illustrates the tipping scales between the microbe and host and similar to Zika infections, less of this gene expression results in less viral replication; this would imply a protective mechanism rather that host immune evasion. In fact, coronaviruses bud into the ER-Golgi intermediate compartment and in MERS, the C-terminal domain of the M protein was found to contain a trans-Golgi localization signal (Perrier et al., 2019); thus, the role of ATL3 as a way to control viral protein budding presents an exciting avenue for future work. Further to which, the top two over-expressed genes of the "COVID-19-specific gene signature" are coiled-coil and C2 domain containing 2A (*CC2D2A*) and human homeostatic iron regulator or high FE2+ (HFE). CC2D2A plays a critical role in cilia formation (Veleri et al., 2014). Primary cilia microtubule-based sensory organelles that detect mechanical and chemical stimuli are found in almost all cells in the body (Garcia-Gonzalo and Reiter, 2012). Following T-cell receptor signaling, the ciliary trafficking machinery is used to provide spatial control of immune synapses at the interface with the antigen-presenting cell for signaling (Stephen et al., 2018). HFE is a non-classical major histocompatibility (MHC) protein (HLA-H). Mutations disable the ability of this protein to bind β2-microglobulin, a component of the HLA class I molecule, which normally present peptides derived from cytosolic proteins; stagnating presentation of peptide loaded MHC class I

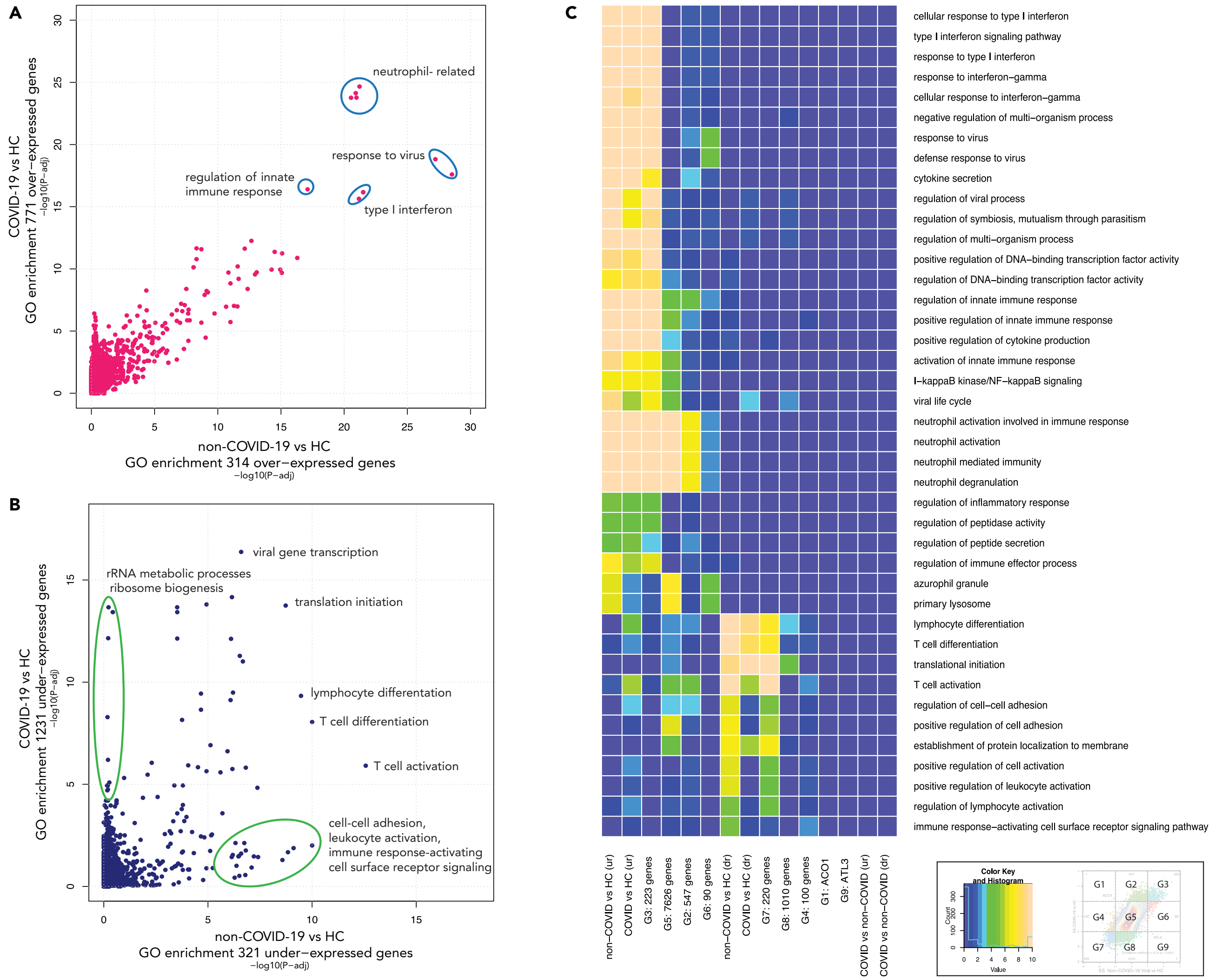


Figure 5. Summary of pathway analysis results

Scatterplots of the significance level from pathway enrichment analysis between COVID-19 and non-COVID-19 viral infections obtained for positive genes in (A) and negative genes in (B), respectively. Significance is defined as -log10(BH-corrected p value) for each pathway. The concordance is seen in results for up-regulated genes between COVID-19 and non-COVID-19, while a degree of discordance is evident in down-regulated genes between COVID-19 and non-COVID-19. (C) Heatmap summary of pathway enrichment analysis for 15 gene sets of interest including COVID-19 vs HC (+) and (−), non-COVID-19 viral vs HC (+) and (−), COVID-19 vs non-COVID-19 viral (+) and (−), as well as gene lists from 9 groups segmented in Figure 4A as labeled in the legend key box. Values between 1 and 10 of -log10(BH-corrected p value) are plotted. ur, up-regulated; dr, down-regulated.

molecules at the cell surface (Hollerer et al., 2017). While largely responsible for presenting ‘‘endogenous’’ peptides, during viral infection, this class of HLA is responsible for loading of viral peptides at the ER and trafficking those to the cell surface (Hollerer et al., 2017). HFE is essential in this function as these peptides are presented to T cells or NK cells. The two genes, therefore, are both involved in an effective immune signaling between virally infected cells and the host. HFE is pleotropic in function, and it binds with the transferrin receptor thus reducing affinity for iron loaded transferrin, resulting in less cytoplasmic iron (Taneri et al., 2020). ACO1 is bifunctional as well, a key modulator of mitochondrial iron metabolism, and it is also an essential enzyme in the Krebs cycle (Wood, 2006). Iron metabolism and ATP production are essential for the function of the cell and the proliferation of immune cells. Here, we observe over-expression of *CC2D2A*, *HFE*, and *ACO1* in COVID-19 infections and lower expression in non-COVID-19 previously circulating infections. We interpret this COVID-19 over-expression of genes not intensely involved in non-COVID-19 infections as avenues for future exploration as possible counteractive measures for the novel immune evasion eluded to by the under-expression of *ZC3H13* and *AMIGO1* described above.

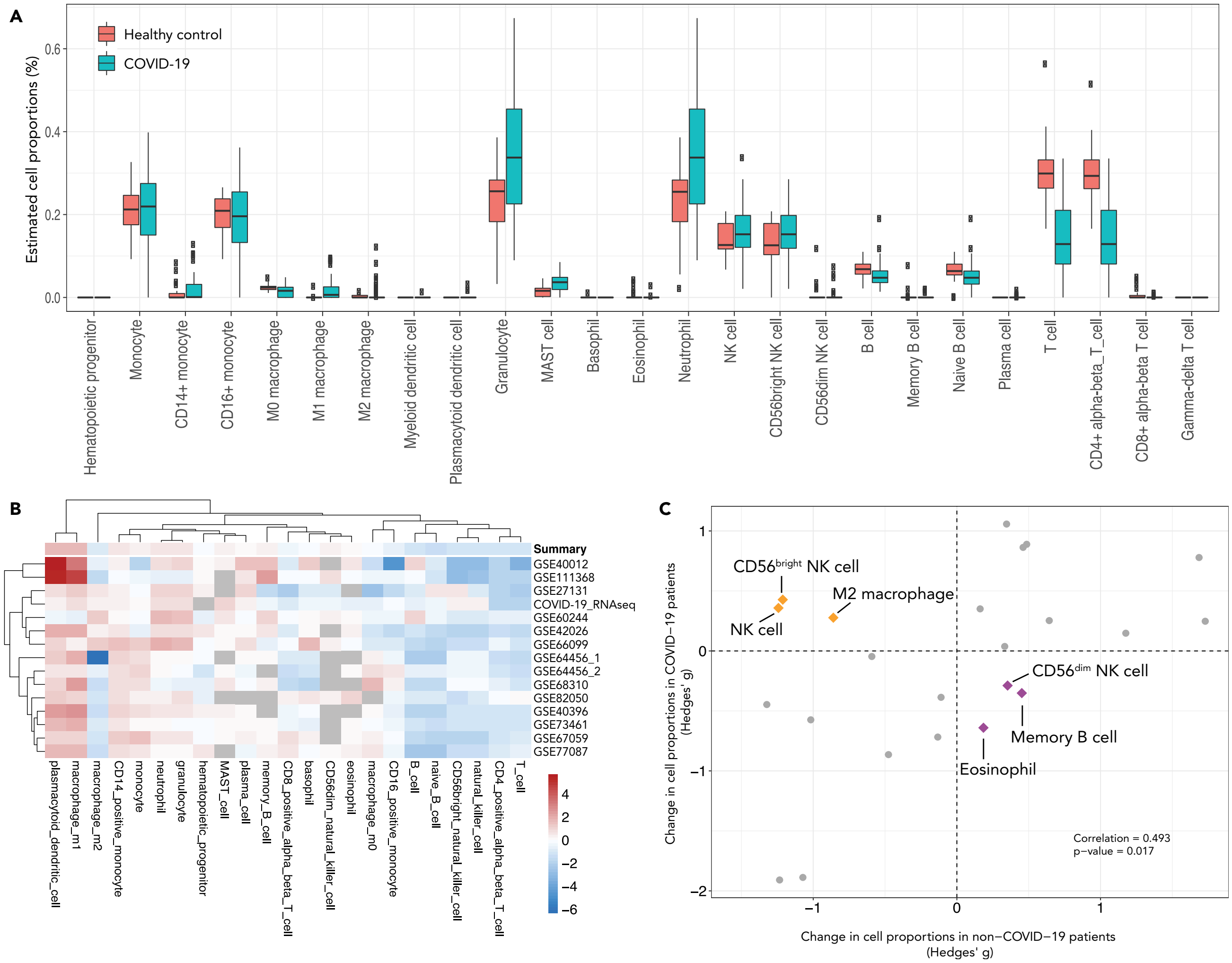


Figure 6. Statistical deconvolution of bulk transcriptome profiles using immunoStates of COVID-19 versus non-COVID-19 viral infections

(A) Changes in cell proportions when comparing patients with COVID-19 to healthy controls. Note the trends of increased neutrophil and decreased T-cell proportions (median and interquartile range [IQR]).

(B) Heatmap of changes in cell proportions of all data sets: 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.

All of these genes and their functions need to be molecularly investigated to determine their true role; here, we use them as an illustration of both novel immune evasion and immune defense systems. These measures and counter measures will likely be somewhat different for each patient as they progress through the disease. We see in this cohort gene expression indicative of a beneficial host response whereby HLA class I molecules present viral peptides to the host response for identification and destruction via over-expression of *HFE* and *CC2D2A*, carefully managed iron metabolism and energy production via HFE and ACO1. However, how much over-expression is needed in order to overcome the SARS-CoV-2 virus is not known, and not surprisingly, there are trials underway for the use of pegylated interferon alpha in patients with COVID-19 (2020). This drug is FDA approved for treatment of viral infections such as HCV (Tan et al., 2004; Nile et al., 2020) and showed promise in combination with ribavirin in patients with MERS (Omrani et al., 2014), as one of its mechanisms of action increases MHC class I function (Nile et al., 2020).

Within this signature, we also find genes commonly studied in cancer (e.g. *TP53*, *AKT*, *VEGF*, and *CYCS*). Interestingly, primary cilia house a number of oncogenic molecules including smoothened, *KRAS*, epidermal growth factor receptor, and platelet-derived growth factor receptor (Jenks et al., 2018), and thus, the role in the immune response to COVID-19 would need further investigation. Of the 416 COVID-19-specific genes, we also observe multiple superfamily members of ATP-binding cassette transporters, which facilitate the interaction of multiple immune cells with various classes of lipids. In macrophages and lymphocytes, this alters the plasticity of the cell, dampening the immune response to viral invasion (Hubler and Kennedy, 2016). As well as *ZC3H13*, this gene set includes many other zinc finger proteins. Zinc (Zn^{2+}) homeostasis in the cell is tightly regulated as viruses need Zn^{2+} for newly synthesized viral proteins (Lazarczyk and Favre, 2008).

In place of GO terms directly derived from our "COVID-19-specific gene signature", Figure 5 illustrates the comparison of COVID-19 versus HC to non-COVID-19 versus HC GO terms. We found many downregulated pathways are discordant when comparing to HCs. 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 translation of host mRNA while allowing viral translation to continue (Huang et al., 2011). An opposite 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, and immune response-activating cell surface receptor signaling, suggesting a less effective immune response in patients with COVID-19. Of particular interest was the observation that while both diseases had enriched GO terms for type-1 interferon signaling pathways, the significance of this enrichment was lower in COVID-19 (Figure 5). The inspection of the 6 genes above mirrors these discordant pathway findings, supporting the concept of novel biology specific to COVID-19 within a largely similar response to other viruses.

Interestingly, the immune cell proportions are mostly consistent across COVID-19 and non-COVID-19 data sets. Our results are in line with several recent studies that found high neutrophil-lymphocyte ratio in patients with COVID-19 (Diao et al., 2020; Lagunas-Rangel, 2020; Liu et al., 2020; Qin et al., 2020). 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 responses (Vivier et al., 2008). 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 patients with COVID-19 and when considering factors most explanatory in those that recovered the cells that were the most dynamic included $CD56^{dim}$ NK cells (Sun et al., 2020).

When comparing COVID-19 to non-COVID-19 viral infections, we see M1 macrophage proportions are similar to those 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 survive (Benoit et al., 2008). However, in general, monocytotropic viruses including SARS-CoV-1 have evolved mechanisms to interfere with effective macrophage polarization (Hu et al., 2012), favoring the M2 population for immune evasion. For example, virus-induced macrophage depletion is executed by viruses that carry pro-antiapoptotic proteins, thus initially reducing the number of M1s to skew population to M2 and avoid attack, and then further suppress the production and action of type I IFNs, stunting the progression of M1 macrophage polarization (Laura C Miller, 2015). This shift we see in the proportion of M2 macrophages in COVID-19 versus non-COVID-19 viral infections indicates that this novel pathogen may be executing these immune evasion techniques with a high degree of success. We see that eosinophils and $CD56^{dim}$ natural killer (NK) cells are lower in COVID-19 versus non-COVID-19 infections, which replicated in a system-level study over time using mass cytometry and Olink assays where both cell types increased in abundance from a low level at the acute phase to a normal level in the recovery phase (Rodriguez et al., 2020). As well, decreased B cell and increased M2 macrophage cells were observed in a study of 3939 patients with COVID-19 from China and pose many avenues for novel therapies (Wang et al., 2020).

In conclusion, we here provide bulk RNAseq profiling of peripheral blood in COVID-19 in comparison to HCs which we derived a signature of 2002 genes for investigation of the biology and potentially pathophysiology of this disease, the "COVID-19 signature genes". We compiled an extensive database of non-COVID-19 viral infections across many platforms, ages, diseases, and locations globally to compare to HCs using metaintegration to derive a set of 635 genes representing the host response to known viral pathogens, the "non-COVID-19 signature genes". We then used COCONUT to conormalize all of the data and directly compare COVID-19 to non-COVID-19 viral infections resulting in a signature of 416 genes, the "COVID-19-specific gene signature". We used all of these analyses to identify both the similarities and differences in the underlying host response. While we found 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. The genes *ACO1* and *ATL3* were identified as an intersect of gene signatures for COVID-19 versus HCs and non-COVID-19 versus HCs, which were further contextualized when considering the top ranking genes of the novel "COVID-19-specific gene signature", suggesting we have illuminated novel biology of the host immune response to a totally novel viral infection, but our findings will need to be replicated in further clinical studies. In summary, COVID-19 gene expression is highly correlated with known viral infection gene expression and has similar shifts in the immune cell proportions known to play a role in viral response but also shows discordant shifts in immune cells that are novel and reflect other recent publications, key information at the onset of a pandemic to leverage our prior and mounting viral infection knowledge. Our computational methods allowed for a head-to-head comparison of COVID-19 and non-COVID-19 viral infection resulting in a novel 416 gene signature, of which many of the genes with the largest Hedges' g ES have well-known immune functions; however, GO terms were not significant suggesting the magnitude and combination of the genes that discriminate the host response to this novel virus can be disseminated to the scientific community at large to investigate whether this novel combination of genes yields any targetable pathophysiology.

Limitations of the study

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 during 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 analyzed 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 Hedges g ES and FDR values for all genes (Table S2A) to enable re-analysis of these genes based on thresholds that others may deem more appropriate. Figure S3 is also provided to show the GO term enrichment results by varying cutoffs.

Resource availability

Lead contact

Timothy E Sweeney, MD, PhD, tsweeney@inflammatix.com, 863 Mitten Rd, Suite 104, Burlingame, CA 94010.

Material availability

This study did not generate any new unique reagents and/or materials.

Data and code availability

The publicly available studies can be accessed on GEO under their respective study IDs. The COVID-19 cohort is deposited in the Gene Expression Omnibus (GEO) database: GSE152641. Results were generated using R packages COCONUT and MetaIntegrator; both methods have been published and are publicly available R packages. The RNAseq pipeline used to process COVID-19 cohort is described in the methods section.

Additional supplemental items including Transparent Methods are available from Mendeley Data: https://doi.org/10.17632/t4twwtvv7r.1.

Methods

All methods can be found in the accompanying Transparent Methods supplemental file.

Supplemental Information

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101947.

Acknowledgments

We are grateful to Ashley Prasse Miller, Mario Esquivel, and Oliver Liesenfeld of the Inflammatix Clinical Affairs team for clinical sample availability and Luciano Brocchieri and Silvia Tomaletti of TB-SEQ for helpful discussion. J.T. is funded by a National Science Scholarship (PhD) from The Agency for Science, Technology and Research (A*STAR), Singapore.

Author contributions

T.E.S., Y.D.H., P.K., and E.J.G.B. designed the study; G.N.D., J.K., G.V., E.G., E.K., G.D., N.A. and E.J.G.B. conducted clinical studies; S.A.T., Y.D.H., Y.H., S.S., R.P., J.T., and P.K. performed bioinformatics analysis; D.R., M.R., and S.C. processed samples; S.A.T., Y.D.H., P.K., and T.E.S. wrote the manuscript; all authors critically revised and approved the manuscript.

Declaration of interests

S.A.T., Y.D.H., Y.H., S.S., R.P., D.R., M.R., S.C., and T.E.S. are employees of and stockholders in Inflammatix, Inc. P.K. reports being a shareholder and a consultant to Inflammatix, Inc. E.J.G.B. reports receiving 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 Horizon 2020 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.

Received: July 10, 2020
Revised: November 11, 2020
Accepted: December 11, 2020
Published: January 22, 2021

References

Andres-Terre, M., McGuire, H.M., Pouliot, Y., Bongen, E., Sweeney, T.E., Tato, C.M., and Khatri, P. (2015). Integrated, multi-cohort analysis identifies conserved Transcriptional signatures across multiple respiratory viruses. Immunity. https://doi.org/10.1016/j.immuni.2015.11.003.

Bataille, S., Pedinielli, N., and Bergougnioux, J.-P. (2020). Could ferritin help the screening for COVID-19 in hemodialysis patients? Kidney Int. *98*, 235.

Benoit, M., Desnues, B., and Mege, J.L. (2008). Macrophage polarization in bacterial infections. J. Immunol. *181*, 3733.

Blanco-Melo, D., Nilsson-Payant, B.E., Liu, W.C., Uhl, S., Hoagland, D., Møller, R., Jordan, T.X., Oishi, K., Panis, M., Sachs, D., et al. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell. https://doi.org/10.1016/j.cell.2020.04.026.

Callaway, E. (2020). The race for Coronavirus vaccines. Nature *580*, 576–577.

Catanzaro, M., Fagiani, F., Racchi, M., Corsini, E., Govoni, S., and Lanni, C. (2020). Immune response in COVID-19: addressing a pharmacological challenge by targeting pathways triggered by SARS-CoV-2. Signal Transduct Target. Ther. *5*, 84.

Chen, R., Khatri, P., Mazur, P.K., Polin, M., Zheng, Y., Vaka, D., Hoang, C.D., Shrager, J., Xu, Y., Vicent, S., et al. (2014). A meta-Analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. Cancer Res. *74*, 2892.

Damelin, M., Bankovich, A., Bernstein, J., Lucas, J., Chen, L., Williams, S., Park, A., Aguilar, J., Ernstoff, E., Charati, M., et al. (2017). A PTK7-targeted antibody-drug conjugate reduces tumor-initiating cells and induces sustained tumor regressions. Sci. Transl Med. *9*, https://doi.org/10.1126/scitranslmed.aag2611.

Diao, B., Wang, C., Tan, Y., Chen, X., Liu, Y., Ning, L., Chen, L., Li, M., Liu, Y., Wang, G., et al. (2020). Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19). Front. Immunol. *11*, https://doi.org/10.3389/fimmu.2020.00827.

Dimopoulos, G., de Mast, Q., Markou, N., Theodorakopoulou, M., Komnos, A., Mouktaroudi, M., Netea, M.G., Spyridopoulos, T., Verheggen, R.J., Hoogerwerf, J., et al. (2020). Favorable anakinra responses IN severe COVID-19 patients with secondary hemophagocytic lymphohistiocytosis. Cell Host Microbe *28*, 117.

Drakesmith, H., and Prentice, A. (2008). Viral infection and iron metabolism. Nat. Rev. Microbiol. *6*, 541.

Etzioni, A. (1996). Adhesion molecules-their role in health and disease. Pediatr. Res. *39*, 191.

Garcia-Gonzalo, F.R., and Reiter, J.F. (2012). Scoring a backstage pass: mechanisms of ciliogenesis and ciliary access. J. Cell Biol. *197*, 697.

Giamarellos-Bourboulis, E.J., Netea, M.G., Rovina, N., Akinosoglou, K., Antoniadou, A., Antonakos, N., Damoraki, G., Gkavogianni, T., Adami, M.E., Katsaounou, P., and Ntaganou, M. (2020). Complex immune dysregulation in COVID-19 patients with severe respiratory failure. Cell Host and Microbe. https://doi.org/10.1016/j.chom.2020.04.009.

Gokhale, N.S., McIntyre, A.B.R., McFadden, M.J., Roder, A.E., Kennedy, E.M., Gandara, J.A., Hopcraft, S.E., Quicke, K.M., Vazquez, C., Willer, J., et al. (2016). N6-Methyladenosine in flaviviridae viral RNA genomes regulates infection. Cell Host Microbe *20*, 654.

Gu, J., Gong, E., Zhang, B., Zheng, J., Gao, Z., Zhong, Y., Zou, W., Zhan, J., Wang, S., Xie, Z., et al. (2005). Multiple organ infection and the pathogenesis of SARS. J. Exp. Med. *202*, 415.

Haynes, W.A., Vallania, F., Liu, C., Bongen, E., Tomczak, A., Andres-Terrè, M., Lofgren, S., Tam, A., Deisseroth, C.A., Li, M.D., et al. (2017). Empowering multi-cohort gene expression analysis to increase reproducibility. In Pacific Symposium on Biocomputing.

Haynes, W.A., Vashisht, R., Vallania, F., Liu, C., Gaskin, G.L., Bongen, E., Lofgren, S., Sweeney, T.E., Utz, P.J., Shah, N.H., and Khatri, P. (2018a). Integrated molecular, clinical, and ontological analysis identifies overlooked disease relationships. bioRxiv. https://doi.org/10.1101/214833.

Haynes, W.A., Haddon, D.J., Diep, V.K., Khatri, A., Bongen, E., Yiu, G., Balboni, I., Bolen, C.R., Mao, R., Utz, P.J., and Khatri, P. (2020). Integrated, multicohort analysis reveals unified signature of systemic lupus erythematosus. JCI Insight. https://doi.org/10.1172/jci.insight.122312.

Haynes, W.A., Tomczak, A., and Khatri, P. (2018b). Gene annotation bias impedes biomedical research. Sci. Rep. *8*, https://doi.org/10.1038/s41598-018-19333-x.

Hedges, L.V., and Pigott, T.D. (2001). The power of statistical tests in meta-analysis. Psychol. Methods *6*, 203–217.

Hollerer, I., Bachmann, A., and Muckenthaler, M.U. (2017). Pathophysiological consequences and benefits of HFE mutations: 20 years of research. Haematologica *102*, 809.

Hopp, M.-T., et al. (2020). Unravelling the debate on heme effects in COVID-19 infections. bioRxiv. https://doi.org/10.1101/2020.06.09.142125.

Hu, W., Yen, Y.T., Singh, S., Kao, C.L., and Wu-Hsieh, B.A. (2012). SARS-CoV regulates immune function-related gene expression in human monocytic cells. Viral Immunol. *25*, 277.

Huang, C., Lokugamage, K.G., Rozovics, J.M., Narayanan, K., Semler, B.L., and Makino, S. (2011). SARS coronavirus nsp1 protein induces template-dependent endonucleolytic cleavage of mRNAs: viral mRNAs are resistant to nsp1-induced RNA cleavage. PLoS Pathog. *7*, e1002433.

Hubler, M.J., and Kennedy, A.J. (2016). Role of lipids in the metabolism and activation of immune cells. J. Nutr. Biochem. *34*, 1.

Inoue, T., and Tsai, B. (2013). How viruses use the endoplasmic reticulum for entry, replication, and assembly. Cold Spring Harb Perspect. Biol. *5*, a013250.

Jaume, M., Yip, M.S., Kam, Y.W., Cheung, C.Y., Kien, F., Roberts, A., Li, P.H., Dutry, I., Escriou, N., Daeron, M., et al. (2012). SARS CoV subunit vaccine: Antibodymediated neutralisation and enhancement. Hong Kong Med. J. *18*, 31.

Jenks, A.D., Vyse, S., Wong, J.P., Kostaras, E., Keller, D., Burgoyne, T., Shoemark, A., Tsalikis, A., de la Roche, M., Michaelis, M., et al. (2018). Primary cilia mediate diverse kinase inhibitor resistance mechanisms in cancer. Cell Rep. https://doi.org/10.1016/j.celrep.2018.05.016.

Khatri, P., Roedder, S., Kimura, N., De Vusser, K., Morgan, A.A., Gong, Y., Fischbein, M.P., Robbins, R.C., Naesens, M., Butte, A.J., and Sarwal, M.M. (2013). A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. J. Exp. Med. *210*, 2205.

Koeller, D.M., Casey, J.L., Hentze, M.W., Gerhardt, E.M., Chan, L.N., Klausner, R.D., and Harford, J.B. (1989). A cytosolic protein binds to structural elements within the iron regulatory region of the transferrin receptor mRNA. Proc. Natl. Acad. Sci. U S A *86*, 3574.

Kuja-Panula, J., Kiiltomäki, M., Yamashiro, T., Rouhiainen, A., and Rauvala, H. (2003). AMIGO, a transmembrane protein implicated in axon tract development, defines a novel protein family with leucine-rich repeats. J. Cell Biol. *160*, 963.

Lagunas-Rangel, F.A. (2020). 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. https://doi.org/10.1002/jmv.25819.

Laura C Miller, Y.S. (2015). Macrophage polarization in virus-host interactions. J. Clin. Cell Immunol. *06*, https://doi.org/10.4172/2155-9899.1000311.

Lazarczyk, M., and Favre, M. (2008). Role of Zn2+ ions in host-virus interactions. J. Virol. *82*, 11486.

Li, L., Khatri, P., Sigdel, T.K., Tran, T., Ying, L., Vitalone, M.J., Chen, A., Hsieh, S., Dai, H., Zhang, M., et al. (2012). A peripheral blood diagnostic test for acute rejection in renal transplantation. Am. J. Transpl. *12*, 2710.

Liu, J., Liu, Y., Xiang, P., Pu, L., Xiong, H., Li, C., Zhang, M., Tan, J., Xu, Y., Song, R., et al. (2020). Neutrophil-to-lymphocyte ratio predicts critical illness patients with 2019 coronavirus disease in the early stage. J. Transl. Med. *18*, 206.

Longo, Dan, Fauci, Anthony, Dennis Kasper, S.H., Jameson, J., and Loscalzo, J. (2015). Harrison's Principles of Internal Medicine: Volumes 1 and 2, Nineteenth Edition (Mcgraw-hill).

Mayhew, M.B., Buturovic, L., Luethy, R., Midic, U., Moore, A.R., Roque, J.A., Shaller, B.D., Asuni, T., Rawling, D., Remmel, M., et al. (2020). A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. Nat. Commun. *11*, https://doi.org/10.1038/s41467-020-14975-w.

Medzhitov, R. (2007). Recognition of microorganisms and activation of the immune response. Nature *449*, 819.

Mesev, E.V., LeDesma, R.A., and Ploss, A. (2019). Decoding type I and III interferon signalling during viral infection. Nat. Microbiol. *4*, 914.

Monel, B., Rajah, M.M., Hafirassou, M.L., Sid Ahmed, S., Burlaud-Gaillard, J., Zhu, P.-P., Nevers, Q., Buchrieser, J., Porrot, F., Meunier, C., et al. (2019). Atlastin endoplasmic reticulum-shaping proteins facilitate Zika virus replication. J. Virol. *93*, https://doi.org/10.1128/jvi.01047-19.

Morens, D.M., and Fauci, A.S. (2020). Emerging pandemic diseases: how we got to COVID-19. Cell *182*, 1077.

Nile, S.H., et al. (2020). COVID-19: Pathogenesis, Cytokine Storm and Therapeutic Potential of Interferons. Cytokine Growth Factor Rev. https://doi.org/10.1016/j.cytogfr.2020.05.002.

Omrani, A.S., Saad, M.M., Baig, K., Bahloul, A., Abdul-Matin, M., Alaidaroos, A.Y., Almakhlafi, G.A., Albarrak, M.M., Memish, Z.A., and Albarrak, A.M. (2014). Ribavirin and interferon alfa-2a for severe Middle East respiratory syndrome coronavirus infection: a retrospective cohort study. Lancet Infect. Dis. *14*, 1090.

Perrier, A., Bonnin, A., Desmarets, L., Danneels, A., Goffard, A., Rouillé, Y., Dubuisson, J., and Belouzard, S. (2019). The C-terminal domain of the MERS coronavirus M protein contains a trans-Golgi network localization signal. J. Biol. Chem. *294*, 14406.

Pontelli, M.C., et al. (2020). Infection of human lymphomononuclear cells by SARS-CoV-2. bioRxiv. https://doi.org/10.1101/2020.07.28.225912.

Qin, C., Zhou, L., Hu, Z., Zhang, S., Yang, S., Tao, Y., Xie, C., Ma, K., Shang, K., Wang, W., and Tian, D.-S. (2020). Dysregulation of immune response in patients with coronavirus 2019 (COVID-19) in wuhan, China. Clin. Infect. Dis. *71*, 762.

Ravelli, A. (2002). Macrophage activation syndrome. Curr. Opin. Rheumatol. *14*, 548.

Rodriguez, L., Pekkarinen, P.T., Lakshmikanth, T., Tan, Z., Consiglio, C.R., Pou, C., Chen, Y., Mugabo, C.H., Nguyen, N.A., Nowlan, K., et al. (2020). Systems-level immunomonitoring from acute to recovery phase of severe COVID-19. Cell Rep. Med. *1*, 100078.

Shah, V.K., Firmal, P., Alam, A., Ganguly, D., and Chattopadhyay, S. (2020). Overview of immune response during SARS-CoV-2 infection: lessons from the past. Front. Immunol. *11*, https://doi.org/10.3389/fimmu.2020.01949.

Stephen, L.A., ElMaghloob, Y., McIlwraith, M.J., Yelland, T., Castro Sanchez, P., Roda-Navarro, P., and Ismail, S. (2018). The ciliary machinery is repurposed for t cell immune synapse trafficking of LCK. Dev. Cell. https://doi.org/10.1016/j.devcel.2018.08.012.

Sun, S., Cai, X., Wang, H., He, G., Lin, Y., Lu, B., Chen, C., Pan, Y., and Hu, X. (2020). Abnormalities of peripheral blood system in patients with COVID-19 in Wenzhou, China. Clin. Chim. Acta *507*, 174.

Sweeney, T.E., and Khatri, P. (2015). Comprehensive validation of the FAIM3:PLAC8 ratio in time-matched public gene expression data. Am. J. Respir. Crit. Care Med. *192*, 1260.

Sweeney, T.E., Shidham, A., Wong, H.R., and Khatri, P. (2015). A comprehensive time-course-based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set. Sci. Transl Med. *7*, 287ra71.

Sweeney, T.E., Braviak, L., Tato, C.M., and Khatri, P. (2016a). Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. Lancet Respir. Med. *4*, 213.

Sweeney, T.E., Wong, H.R., and Khatri, P. (2016b). Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Sci. Transl. Med. *8*, 346ra91.

Sweeney, T.E., Haynes, W.A., Vallania, F., Ioannidis, J.P., and Khatri, P. (2017). Methods to increase reproducibility in differential gene expression via meta-analysis. Nucleic Acids Res. *45*, e1.

Sweeney, T.E., Perumal, T.M., Henao, R., Nichols, M., Howrylak, J.A., Choi, A.M., Bermejo-Martin, J.F., Almansa, R., Tamayo, E., Davenport, E.E., et al. (2018a). A community approach to mortality prediction in sepsis via gene expression analysis. Nat. Commun. *9*, https://doi.org/10.1038/s41467-018-03078-2.

Sweeney, T.E., Azad, T.D., Donato, M., Haynes, W.A., Perumal, T.M., Henao, R., Bermejo-Martin, J.F., Almansa, R., Tamayo, E., Howrylak, J.A., et al. (2018b). Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters. Crit. Care Med. *46*, 915.

Tan, E.L.C., Ooi, E.E., Lin, C.Y., Tan, H.C., Ling, A.E., Lim, B., and Stanton, L.W. (2004). Inhibition of SARS coronavirus infection in vitro with clinically approved antiviral drugs. Emerg. Infect. Dis. https://doi.org/10.3201/eid1004.030458.

Taneri, P.E., et al. (2020). Anemia and iron metabolism in COVID-19: a systematic review and meta-analysis. medRxiv. https://doi.org/10.1101/2020.06.04.20122267.

Tay, M.Z., Poh, C.M., Rénia, L., MacAry, P.A., and Ng, L.F.P. (2020). The trinity of COVID-19: immunity, inflammation and intervention. Nat. Rev. Immunol. *20*, 363.

Tomczak, A., Mortensen, J.M., Winnenburg, R., Liu, C., Alessi, D.T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N.H., et al. (2018). Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Sci. Rep. *8*, https://doi.org/10.1038/s41598-018-23395-2.

Veleri, S., Manjunath, S.H., Fariss, R.N., May-Simera, H., Brooks, M., Foskett, T.A., Gao, C., Longo, T.A., Liu, P., Nagashima, K., et al. (2014). Ciliopathy-associated gene Cc2d2a promotes assembly of subdistal appendages on the mother centriole during cilia biogenesis. Nat. Commun. *5*, 4207.

Vivier, E., Tomasello, E., Baratin, M., Walzer, T., and Ugolini, S. (2008). Functions of natural killer cells. Nat. Immunol. *9*, 503.

Wang, Q., Zhang, L., Kuwahara, K., Li, L., Liu, Z., Li, T., Zhu, H., Liu, J., Xu, Y., Xie, J., et al. (2016). Immunodominant SARS coronavirus epitopes in humans elicited both enhancing and neutralizing effects on infection in non-human primates. ACS Infect. Dis. *2*, 361.

Wang, J., Jiang, M., Chen, X., and Montaner, L.J. (2020). Cytokine storm and leukocyte changes in mild versus severe SARS-CoV-2 infection: review of 3939 COVID-19 patients in China and emerging pathogenesis and therapy concepts. J. Leukoc. Biol. *108*, 17.

Warsinske, H., et al. (2018a). Prospective validation of three-gene whole blood diagnostic for active tuberculosis predicts disease progression and response to treatment. Am. J. Respir. Crit. Care Med.

Warsinske, H.C., Rao, A.M., Moreira, F.M.F., Santos, P.C.P., Liu, A.B., Scott, M., Malherbe, S.T., Ronacher, K., Walzl, G., Winter, J., et al. (2018b). 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 *1*, e183779.

Wen, J., Lv, R., Ma, H., Shen, H., He, C., Wang, J., Jiao, F., Liu, H., Yang, P., Tan, L., et al. (2018). Zc3h13 regulates nuclear RNA m6A methylation and mouse embryonic stem cell self-renewal. Mol. Cell *69*, 1028.

Wilson, J.G., et al. (2020). Cytokine Profile in Plasma of Severe COVID-19 Does Not Differ from ARDS and Sepsis. medRxiv. https://doi.org/10.1101/2020.05.15.20103549.

Wood, E.J. (2006). 'Marks' basic medical biochemistry: a clinical approach (Second Edition). Biochem. Mol. Biol. Edu. https://doi.org/10.1002/bmb.2006.494034052660.

Zhou, P., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. https://doi.org/10.1038/s41586-020-2012-7.

iScience, Volume *24*

Supplemental Information

Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections

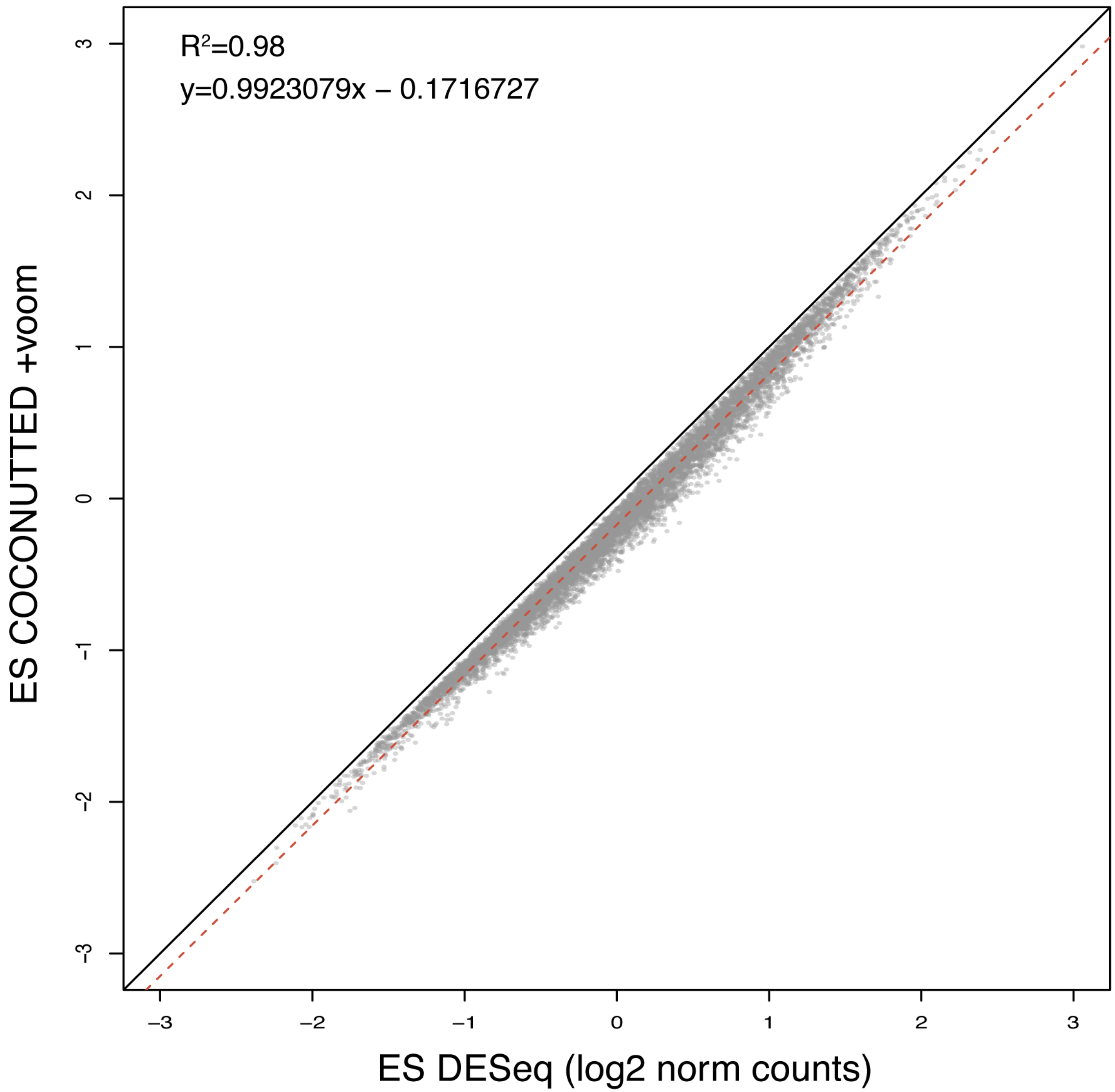
Simone A. Thair, Yudong D. He, Yehudit Hasin-Brumshtein, Suraj Sakaram, Rushika Pandya, Jiaying Toh, David Rawling, Melissa Remmel, Sabrina Coyle, George N. Dalekos, Ioannis Koutsodimitropoulos, Glykeria Vlachogianni, Eleni Gkeka, Eleni Karakike, Georgia Damoraki, Nikolaos Antonakos, Purvesh Khatri, Evangelos J. Giamarellos-Bourboulis, and Timothy E. Sweeney

Supplementary Figure 1. Effect size of DESeq and post COCONUT voom transformed expression data correlate, related to Figure 1.

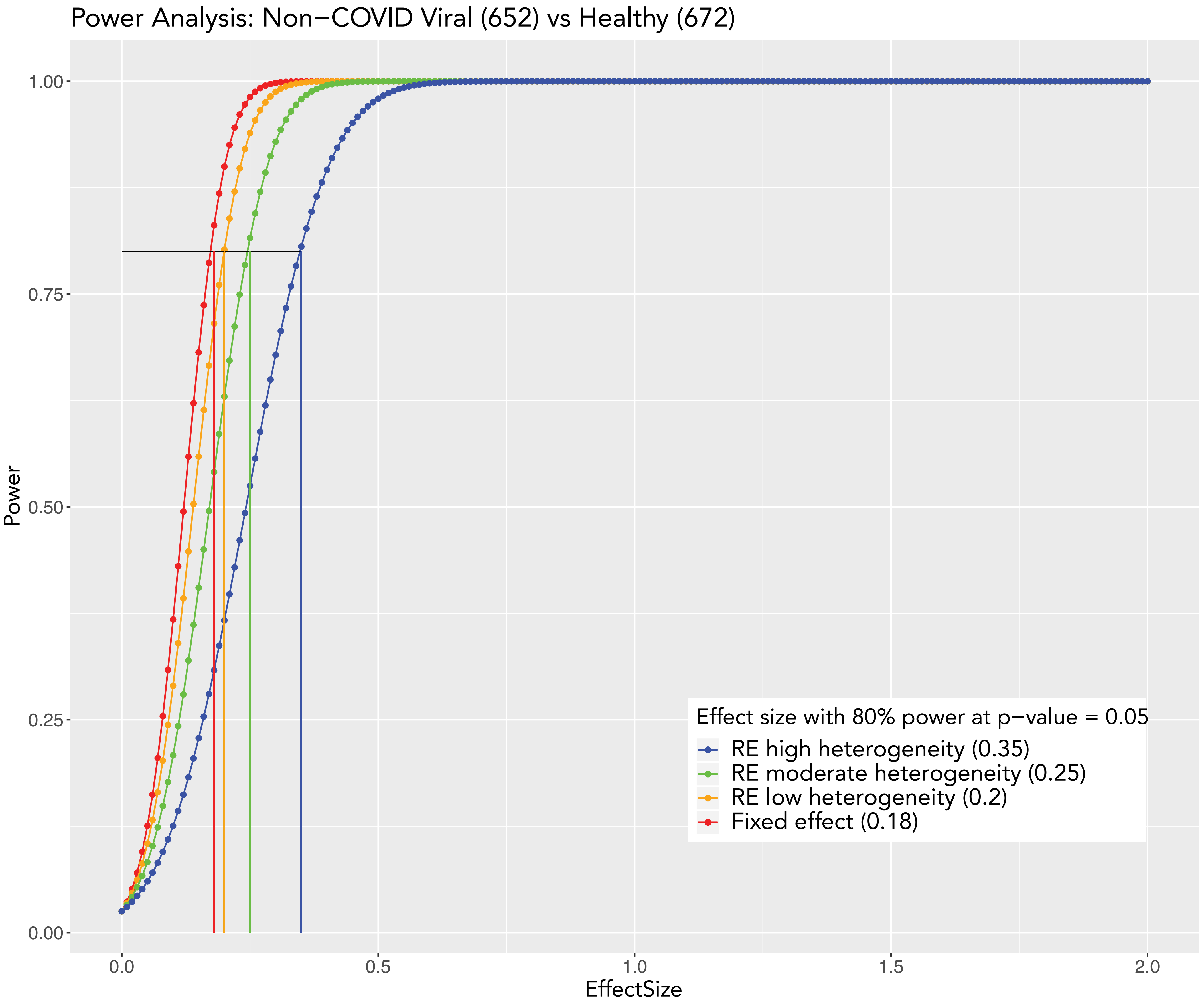
Supplementary Figure 2. Power Analysis. non-COVID-19 (n=652) versus healthy controls (n=672), related to Table 2 and Figure 2.

Supplementary Figure 3. Heatmap of significance score defined as -log10(P-adjusted) from GO term enrichment analyses. Columns contain 45 gene sets including three gene gets (pos, neg, and all) each from COVID-19 vs HC, non-COVID-19 vs HC, or COVID-19 vs non-COVID-19 comparisons with the chosen cutoff and 3 more or less stringent cutoffs, together with 9 gene sets from Figure 4A (see Supplementary Table 4). Rows represent 252 GO terms filtered from a total 8422, in 10 groups by k-means. Related to Figure 4.

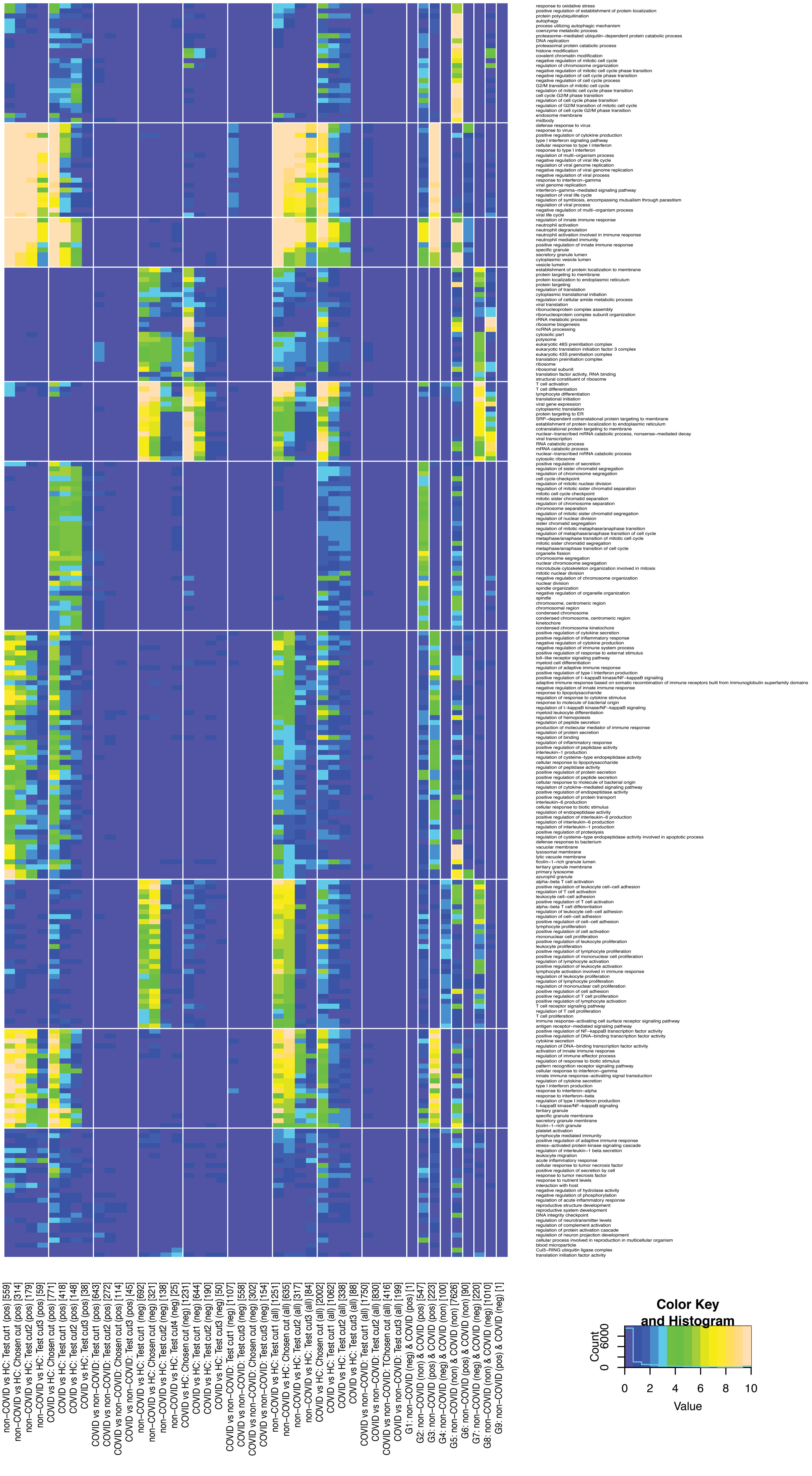
Supplementary Figure 4. Forest plots of cell deconvolution estimates for all studies where estimation was possible (median and interquartile range (IQR)), related to Figure 6.



Supplementary Figure 1. Effect size of DESeq and post COCONUT voom transformed expression data correlate.



Supplementary Figure 2. Power Analysis. non-COVID-19 (n=652) versus healthy controls (n=672).



Supplementary Figure 3. Heatmap of significance score defined as -log10(P-adjusted) from GO term enrichment analyses.

GSE111368
GSE27131
GSE40012
GSE40396
GSE42026
GSE60244
GSE64456
GSE64456
GSE66099
GSE67059GPL6947
GSE68310
GSE73461
GSE77087
GSE82050
Summary
Standardized Mean Difference (log2 scale)
monocyte
−3 −2 −1 0 1 2
CD14_positive_monocyte
−1.0 −0.5 0.0 0.5 1.0 1.5 2.0
CD16_positive_monocyte
−7 −6 −5 −4 −3 −2 −1 0 1
natural_killer_cell
−5 −4 −3 −2 −1 0
CD56bright_natural_killer_cell
−5 −4 −3 −2 −1 0 1
CD56dim_natural_killer_cell
−2.5 −1.5 −0.5 0.0 0.5 1.0
T_cell
−5 −4 −3 −2 −1 0 1
CD4_positive_alpha_beta_T_cell
−3.5 −2.5 −1.5 −0.5 0.5
CD8_positive_alpha_beta_T_cell
−4 −3 −2 −1 0 1
gamma_delta_T_cell
−2.4 −2.0 −1.6 −1.2 −0.8
B_cell
−4 −3 −2 −1 0 1 2
memory_B_cell
−3 −2 −1 0 1 2 3
naive_B_cell
−4 −3 −2 −1 0 1
plasma_cell
−1.5 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
granulocyte
−1.0 0.0 0.5 1.0 1.5 2.0 2.5
MAST_cell
−1.0 −0.5 0.0 0.5 1.0
basophil
−3 −2 −1 0 1 2
eosinophil
−1.0 −0.5 0.0 0.5 1.0 1.5
neutrophil
−1.0 0.0 0.5 1.0 1.5 2.0 2.5 3.0
hematopoietic_progenitor
−2.0 −1.0 −0.5 0.0 0.5 1.0 1.5
macrophage_m0
−5 −4 −3 −2 −1 0 1 2
macrophage_m1
−2 −1 0 1 2 3 4 5
macrophage_m2
−8 −6 −4 −2 0 2
myeloid_dendritic_cell
0.0 0.5 1.0 1.5 2.0
plasmacytoid_dendritic_cell
−1 0 1 2 3 4 5 6 7 8

Supplementary Figure 4. Forest plots of cell deconvolution estimates for all studies where estimation was possible (median and interquartile range (IQR))

SupplementaryTable1_RNAseq_tech_data.xlsx, related to Figure 1.
(excel file, external to this pdf)

SupplementaryTable2_9818genes.xlsx, related to Figure 1a, 2a, 4, 5c
(excel file, external to this pdf)

SupplementaryTable3_GeneSet_Summary.pdf, related to Figure 4.

SupplementaryTable4_Immunostates.pdf, related to Figure 6.

		ALL	POS	NEG	Comment
COVID vs HC comparison					
Chosen cutoff	IESI>= 1.0 & FDR < 0.05%	2,002	771	1,231	***COVID-19 signature***
Test cutoff	IESI>= 1.2 & FDR < 0.05%	1,062	418	644	
Test cutoff	IESI>= 1.5 & FDR < 0.05%	338	148	190	
Test cutoff	IESI>= 1.8 & FDR < 0.05%	88	38	50	
non-COVID vs HC comparison					
Test cutoff	IESI>= 0.8 & FDR < 0.05%	1,251	559	692	
Chosen cutoff	IESI>= 1.0 & FDR < 0.05%	635	314	321	**non-COVID-19 viral signature**
Test cutoff	IESI>= 1.2 & FDR < 0.05%	317	179	138	
Test cutoff	IESI>= 1.5 & FDR < 0.05%	84	59	25	
COVID vs non-COVID comparison					
Test cutoff	IESI>= 0.6 & FDR < 0.05%	1,750	643	1,107	
Test cutoff	IESI>= 0.8 & FDR < 0.05%	830	272	558	
Chosen cutoff	IESI>= 1.0 & FDR < 0.05%	416	114	302	***COVID-19-specific genes***
Test cutoff	IESI>= 1.2 & FDR < 0.05%	199	45	154	

Concordant and discordant between COVID vs HC and non-COVID vs HC

Group	Change in COVID vs HC	Change in COVID vs HC	# Genes	Comment
G1	under-expressed	over-expressed	1	under- in non-COVID and over- in COVID: gene *ACO1*
G2	un-changed	over-expressed	547	over-expressed only in COVID
G3	over-expressed	over-expressed	223	concordantly over-expressed in both
G4	under-expressed	un-changed	100	under-expressed only in non-COVID
G5	un-changed	un-changed	7,626	unchanged in both
G6	over-expressed	un-changed	90	over-expressed only in COVID
G7	under-expressed	under-expressed	220	concordantly under-expressed in both
G8	un-changed	under-expressed	1,010	under-expressed only in COVID
G9	over-expressed	under-expressed	1	over- in non-COVD and under- in COVID: gene *ATL3*

SupplementaryTable3_GeneSet_Summary.pdf, related to Figure 4.

Cell_type	effectSize_VvsHC_discovery	effectSizePval_VvsHC_discovery	effectSizeFDR_VvsHC_discovery	effectSize_COVID_RNAseq	effectSizePval_COVID_RNAseq	effectSizeFDR_COVID_RNAseq
hematopoietic_progenitor	-0.100	4.69E-01	5.10E-01	na	na	na
monocyte	0.333	7.55E-02	8.99E-02	0.037356319	0.87652478	0.87652478
CD14_positive_monocyte	0.643	2.49E-08	1.56E-07	0.252959215	0.294325632	0.350616078
CD16_positive_monocyte	-0.592	2.38E-03	4.97E-03	-0.046220905	0.8475549	0.87652478
macrophage_m0	-0.133	6.62E-01	6.78E-01	-0.717995492	0.003630674	0.010438188
macrophage_m1	1.686	1.46E-06	4.44E-06	0.778568269	0.001690821	0.005555553
macrophage_m2	-0.860	6.23E-03	1.20E-02	0.277036473	0.251080557	0.320825156
myeloid_dendritic_cell	1.176	4.82E-02	6.69E-02	0.147697164	0.539442115	0.590817555
plasmacytoid_dendritic_cell	1.729	1.19E-07	5.97E-07	0.247456957	0.304883546	0.350616078
granulocyte	0.461	3.06E-02	4.77E-02	0.862896132	0.000547814	0.002099952
MAST_cell	0.346	1.22E-06	4.34E-06	1.05765339	3.17E-05	0.000243023
basophil	-0.110	6.78E-01	6.78E-01	-0.386601363	0.110580827	0.195643002
eosinophil	0.185	2.55E-02	4.25E-02	-0.64034035	0.009112743	0.02328812
neutrophil	0.486	2.00E-02	3.57E-02	0.88885828	0.000381853	0.002099952
natural_killer_cell	-1.242	3.54E-07	1.47E-06	0.358153451	0.138925266	0.211711188
CD56bright_natural_killer_cell	-1.212	1.60E-06	4.44E-06	0.427431699	0.078226784	0.149934669
CD56dim_natural_killer_cell	0.353	5.79E-02	7.62E-02	-0.28855208	0.232075462	0.313984449
B_cell	-1.017	1.00E-04	2.28E-04	-0.574249089	0.018865325	0.043390247
memory_B_cell	0.452	6.36E-02	7.96E-02	-0.351391802	0.146437667	0.211711188
naive_B_cell	-1.323	1.04E-15	1.30E-14	-0.446724386	0.065922016	0.137836942
plasma_cell	0.162	1.23E-01	1.40E-01	0.350652985	0.147277348	0.211711188
T_cell	-1.234	9.76E-18	2.44E-16	-1.909044387	1.48E-11	2.48E-10
CD4_positive_alpha_beta_T_cell	-1.072	3.02E-13	2.52E-12	-1.887953367	2.15E-11	2.48E-10
CD8_positive_alpha_beta_T_cell	-0.476	3.59E-02	5.28E-02	-0.864275406	0.000537496	0.002099952
gamma_delta_T_cell	-1.546	2.46E-05	6.15E-05	na	na	na

SupplementaryTable4_Immunostates.pdf, related to Figure 6.

Transparent 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). Patients were enrolled within the first 24 hours of hospital admission using inclusion criteria of identification of a new lower respiratory tract infection due to COVID-19 defined as the presence of new infiltrate in chest X-ray or chest computed tomography indicative of COVID-19 in a patient without any contact with any healthcare facility the last 90 days. 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 hours from MV(Giamarellos-Bourboulis *et al.*, 2020). 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 approval number 30/20 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 Inflammatix 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. Donors were verbally screened to have no inflammation, infection, illness symptoms, (including no fever or antibiotics within 3 days of sampling) or to be immunocompromised. These samples were drawn prior to July 15, 2019, at least 6 months before the first COVID-19 case reported in the US. All samples were tested and negative for HIV, West Nile, Hepatitis B, and Hepatitis C by molecular or antibody-based testing. The age (median and interquartile range (IR) was 36 (29-45.25) and was 70.8% male.

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 FastQC(Andrews S, 2018). The adapter sequence and 3 bases on the 3′ end of the reads was trimmed using cutadapt as a commonly used procedure(Martin, 2011).
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 gtf(Frankish *et al.*, 2019) 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)(Pagès *et al.*, 2017).
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) method (R package v.3.28.0) (Robinson, McCarthy and Smyth, 2009). The voom method (limma R package v.3.41.18) was then used to transform counts into normalized log2-CPM **(Supplementary Figure 1)**(Ritchie *et al.*, 2015).

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. While there is no consensus in the community, we have used this method across a multitude of studies, being that if more than one probe mapped to a specific gene, probes were summarized with a fixed-effects model because a gene within a sample can have only one expression value(Ramasamy *et al.*, 2008).

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. The split between discovery and validation is driven first, by computational technicality whereby the 3 datasets that are not COCONUT conormalized automatically are held out for validation. Second, we held out pandemic and non-respiratory viral infections (eg. Dengue) for test of the signature as a type of "global" viral signature. Third, of the remaining respiratory/ non-pandemic we split as per described best practices(Sweeney *et al.*, 2017) in concert with similar distribution of the types of viruses in discovery and validation.

Integrated with the voom-transformed RNAseq dataset for COVID-19, they were conormalized together using COCONUT (R package v. 1.0.2) (Sweeney, Wong and Khatri, 2016). COCONUT uses COMBAT empiric-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)(Hedges and Olkin, 1985) for each gene was calculated for COVID-19 (62) versus HC (24) two-group comparison test from the COCONUT conormalized output. Hedges' g is the difference between groups as a proportion of variability in the groups and is calculated as:

$$g = \frac{(\bar{X}_1 - \bar{X}_2)}{S_{pooled}}$$

Whereby $\bar{X}_1$ and $\bar{X}_2$ are sample means in two groups. This is divided by the "within-groups" standard deviation which is S_{pooled}

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Where n_1 and n_2 are the sample sizes in the two groups, and S_1 and S_2 are the respective standard deviations. The estimation of effect size for smaller studies is corrected with J.

$$J = 1 - \frac{3}{4df - 1}$$

P-value was calculated using a student's t-test and adjusted using the Benjamini-Hochberg method to obtain the False Discovery Rate (FDR). Hedges' g ES threshold of ≥ 1 or ≤ -1 in combination with FDR threshold of ≤ 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 **(see Detailed Meta-Analysis section below for an expanded description).**

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 multi-cohort analysis with conormalized data as input, we utilized a well-established MetaIntegrator (version 2.1.1)(Haynes *et al.*, 2017). 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 study(DerSimonian and Laird, 2015) **(Supplementary Methods).** We used an ES threshold ≥ 1 or ≤ -1 with FDR ≤ 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:

$$viral\ score = \ zscore\big(GeoMean(pos) - GeoMean(neg)\big)$$

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 or AUROC) 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-Hochberg(Benjamini and Hochberg, 1995) method to obtain the False Discovery Rate (FDR). ES threshold ≥ 1 or ≤ -1 in combination with FDR threshold of ≤ 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 Analysis (GSEA)(Subramanian *et al.*, 2005). 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 hyper-geometric test were adjusted using the Benjamini-Hochberg method(Benjamini and Hochberg, 1995). 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 (Bongen *et al.*, 2018; Vallania *et al.*, 2018). 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 fixed(Vallania *et al.*, 2018). 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 proportions(Bongen *et al.*, 2018; Roy Chowdhury *et al.*, 2018; Vallania *et al.*, 2018; Scott *et al.*, 2019). 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.

DETAILED META-INTEGRATION STATISTICAL METHODS

The use of Hedges' g (Hedges, 1981; Hedges and Olkin, 1985) effect size (ES) stems from a need for the standardized mean difference to transform all effect sizes to a common metric, and thus enables us to include different outcome measures in the same synthesis. To estimate the standardized mean difference (g) can be calculated as

$$g = \frac{(\bar{X}_1 - \bar{X}_2)}{S_{pooled}}$$

Whereby $\bar{X}_1$ and $\bar{X}_2$ are sample means in two groups.
This is divided by the "within- groups" standard deviation which is S_{pooled}

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Where n_1 and n_2 are the sample sizes in the two groups, and S_1 and S_2 are the respective standard deviations.
The estimation of effect size for smaller studies is corrected with J.
By pooling the two estimates of the standard deviation results in a more accurate estimate of their common by including the variance as well as the means.
Hedges (1981) determined that there is a small bias in d (namely Cohen's d) (Cohen, 1988) in small sample sizes, resulting in the addition of a correction factor J

$$J = 1 - \frac{3}{4df - 1}$$

Whereby the df used to estimate S_{pooled} in two independent groups for example would be

$n_1 + n_2 - 2$. J is always less than one and as samples sizes increase, J becomes closer to 1 and thusly impacts smaller sample sizes appropriately without major adjustment in large sample sizes. This is ideal in meta-analysis where the sizes of the studies available often vary.

Thus a gene's Hedges' g effect size represents the difference between groups transformed as a common metric taking into account the proportion of variability in the groups.

This dovetails with the MetaIntegration methods we have developed (MetaIntegrator R package v. 2.1.1)(Haynes *et al.*, 2017), whereby to then pool these effect sizes across datasets, the summary effect size g_s is calculated using a random effects model:

$$g_s = \frac{\sum_i^n W_i g_i}{\sum_i^n W_i}$$

Where n is the number of datasets, W_i is a weight equal to $\frac{1}{(V_i + T^2)}$, where V_i is the variance of that gene within a given dataset I and T^2 is the inter-dataset variation estimated using the DerSimonian- Liard method(DerSimonian and Laird, 2015), determined to be optimal for our methods but may be further investigated by each individual research group (Sweeney *et al.*, 2017).

Standard error for the summary effect size is derived with

$$SE_{g_s} = \sqrt{\frac{1}{\sum_i^n W_i}}$$

From which a p-value is calculated and corrected for using the Benjamini-Hochberg(Benjamini and Hochberg, 1995) false discovery rate (FDR) correction for multiple hypotheses. Fisher's method is used for combining p-values across the studies as the log sum of

$$F_{up} = -2\sum_{i}^{n} \log{(p_i)}$$

For up-regulated genes and again for down regulated genes, these are again corrected with the Benjamini- Hochberg method.

Genes can then be filtered based on ES, ES FDR, Fisher's FDR and if desired, MetaIntegrator allows for the inclusion of genes that meet these criteria in a leave one dataset out analyses and can further be queried for Cochrane's Q value for evaluating heterogeneity of effect size estimates between studies:

$$Q = \sum_{i}^{n} W_i\,(g_i - \ g_i)^2$$

Allowing for the user to increase the number of genes to pursue perhaps in the case of biological interest or restrict the number of genes when searching for predictive signatures in a disease type (Haynes *et al.,* 2017; Sweeney *et al.,* 2017).

Andrews S (2018) 'FastQC A Quality control tool for high throughput sequence data', *Babraham Bioinfo*.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B (Methodological)*. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Bongen, E. *et al.* (2018) 'KLRD1-expressing natural killer cells predict influenza susceptibility', *Genome Medicine*. doi: 10.1186/s13073-018-0554-1.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Science (2nd Edition), Statistical Power Anaylsis for the Behavioral Sciences*.

DerSimonian, R. and Laird, N. (2015) 'Meta-analysis in clinical trials revisited', *Contemporary Clinical Trials*. doi: 10.1016/j.cct.2015.09.002.

Frankish, A. *et al.* (2019) 'GENCODE reference annotation for the human and mouse genomes', *Nucleic Acids Research*. doi: 10.1093/nar/gky955.

Giamarellos-Bourboulis, E. J. *et al.* (2020) 'Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure', *Cell Host and Microbe*. doi: 10.1016/j.chom.2020.04.009.

Haynes, W. A. *et al.* (2017) 'Empowering multi-cohort gene expression analysis to increase reproducibility', in *Pacific Symposium on Biocomputing*. doi: 10.1142/9789813207813_0015.

Hedges, L. V. (1981) 'Distribution Theory for Glass's Estimator of Effect size and Related Estimators', *Journal of Educational Statistics*. doi: 10.3102/10769986006002107.

Hedges, L. V and Olkin, I. (1985) *Statistical Methodblogy in Meta-Analysis., Statistical Methodblogy in Meta-Analysis*.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*. doi: 10.14806/ej.17.1.200.

Pagès, H. *et al.* (2017) 'Package "AnnotationDbi"', *Bioconductor Package Maintainer*.

Ramasamy, A. *et al.* (2008) 'Key issues in conducting a meta-analysis of gene expression microarray datasets', *PLoS Medicine*. doi: 10.1371/journal.pmed.0050184.

Ritchie, M. E. *et al.* (2015) 'Limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*. doi: 10.1093/nar/gkv007.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2009) 'edgeR: A Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*. doi: 10.1093/bioinformatics/btp616.

Roy Chowdhury, R. *et al.* (2018) 'A multi-cohort study of the immune factors associated with M. tuberculosis infection outcomes', *Nature*. doi: 10.1038/s41586-018-0439-x.

Scott, M. K. D. *et al.* (2019) 'Increased monocyte count as a cellular biomarker for poor outcomes in fibrotic diseases: a retrospective, multicentre cohort study', *The Lancet Respiratory Medicine*. doi: 10.1016/S2213-2600(18)30508-3.

Subramanian, A. *et al.* (2005) 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0506580102.

Sweeney, T. E. *et al.* (2017) 'Methods to increase reproducibility in differential gene expression via meta-analysis', *Nucleic Acids Research*. doi: 10.1093/nar/gkw797.

Sweeney, T. E., Wong, H. R. and Khatri, P. (2016) 'Robust classification of bacterial and viral infections via integrated host gene expression diagnostics', *Science Translational Medicine*. doi: 10.1126/scitranslmed.aaf7165.

Vallania, F. *et al.* (2018) 'Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases', *Nature Communications*. doi: 10.1038/s41467-018-07242-6.