Journal of Infection 87 (2023) 538–550

Contents lists available at ScienceDirect

Journal of Infection

journal homepage: www.elsevier.com/locate/jinf

# Analysis of blood and nasal epithelial transcriptomes to identify mechanisms associated with control of SARS-CoV-2 viral load in the upper respiratory tract

Mahdi Moradi Marjaneh [a,b,c,*], Joseph D. Challenger [d], Antonio Salas [e,f,g], Alberto Gómez-Carballa [e,f,g], Abilash Sivananthan [a], Irene Rivero-Calle [f,g,h], Gema Barbeito-Castiñeiras [i], Cher Y. Foo [j], Yue Wu [k], Felicity Liew [l], Heather R. Jackson [a,b], Dominic Habgood-Coote [a,b], Giselle D'Souza [a,b], Samuel J. Nichols [a,b], Victoria J. Wright [a,b], Michael Levin [a,b], Myrsini Kaforou [a,b], Ryan S. Thwaites [l], Lucy C. Okell [d], Federico Martinón-Torres [f,g,h], Aubrey J. Cunnington [a,b,**], on behalf of the PERFORM Consortium [1], GEN-COVID Study Group (http://gencovid.eu)

[a] Section of Paediatric Infectious Disease, Department of Infectious Disease, Imperial College London, London, UK
[b] Centre for Paediatrics and Child Health, Imperial College London, London, UK
[c] Section of Virology, Department of Infectious Diseases, Imperial College London, London, UK
[d] Medical Research Council Centre for Global Infections Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK
[e] Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain
[f] Genetics, Vaccines and Infections Research Group (GENVIP), Instituto de Investigación Sanitaria de Santiago, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[g] Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBER-ES), Madrid, Spain
[h] Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[i] Servicio de Microbiología y Parasitología, Complejo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[j] School of Medicine, Imperial College London, London, UK
[k] Department of Surgery and Cancer, Imperial College London, St. Mary's Hospital, London, UK
[l] National Heart and Lung Institute, Imperial College London, London, UK

ARTICLE INFO

Accepted 16 October 2023
Available online 18 October 2023

*Keywords:*
COVID-19
SARS-CoV-2
Upper respiratory tract

SUMMARY

*Objectives:* The amount of SARS-CoV-2 detected in the upper respiratory tract (URT viral load) is a key driver of transmission of infection. Current evidence suggests that mechanisms constraining URT viral load are different from those controlling lower respiratory tract viral load and disease severity. Understanding such mechanisms may help to develop treatments and vaccine strategies to reduce transmission. Combining mathematical modelling of URT viral load dynamics with transcriptome analyses we aimed to identify mechanisms controlling URT viral load.

*Methods:* COVID-19 patients were recruited in Spain during the first wave of the pandemic. RNA sequencing of peripheral blood and targeted NanoString *n*Counter transcriptome analysis of nasal epithelium were

*Abbreviations:* AIPL1, Aryl Hydrocarbon Receptor Interacting Protein Like 1; C7orf33, Chromosome 7 Open Reading Frame 33; CCL4, C-C Motif Chemokine Ligand 4; COL4A4, Collagen Type IV Alpha 4 Chain; COVID-19, coronavirus disease; CpG, Cytosine-phosphate-Guanine; Ct value, cycle threshold value; CYP2F1, Cytochrome P450 Family 2 Subfamily F Member 1; FEZ1, Fasciculation and Elongation Protein Zeta 1; GALNT17, Polypeptide N-Acetylgalactosaminyltransferase 17; GNLY, Granulysin; IFNA14, Interferon Alpha 14; IFNA2, Interferon Alpha-2; IFNAR1 and IFNAR2, Interferon Alpha and Beta Receptor Subunits 1 and 2; IFNL3, Interferon Lambda 3; IFNLR, Interferon Lambda Receptor; IL-22, Interleukin-22; IL22RA2, Interleukin 22 Receptor Subunit Alpha 2; IL4I1, Interleukin 4-induced gene-1; IPA, Ingenuity Pathway Analysis; ISGs, interferon-stimulated genes; JAK/STAT, Janus Kinase/Signal Transducer and Activator of Transcription; LRT, lower respiratory tract; NK cells, natural killer cells; PCA, principal component analysis; PTK2, Protein Tyrosine Kinase 2; RBP3, Retinol Binding Protein 3; RNA-Seq, RNA sequencing; RORC, RAR Related Orphan Receptor C; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; STING1, Stimulator of Interferon Response CGAMP Interactor 1; Th22, T helper cells type 22; TLR9, Toll Like Receptor 9; TUBB1, Tubulin Beta 1 Class VI; UGT1A1, UDP Glucuronosyltransferase Family 1 Member A1; URT, upper respiratory tract; VL, viral load; WGCNA, weighted correlation network analysis

* Correspondence to: VC1, Medical School Building, St Mary's Campus Norfolk Place, London W2 1PG, UK.
** Correspondence to: 244, Medical School Building, St Mary's Campus Norfolk Place, London W2 1PG, UK.
*E-mail addresses:* m.moradi@imperial.ac.uk (M. Moradi Marjaneh), a.cunnington@imperial.ac.uk (A.J. Cunnington).
[1] Personalised Risk Assessment in Febrile Illness to Optimise Real-Life Management (PERFORM), London, UK.

https://doi.org/10.1016/j.jinf.2023.10.009

Viral load
Mathematical modelling
Transcriptome
Gene network analysis

performed and gene expression analysed in relation to paired URT viral load samples collected within 15 days of symptom onset. Proportions of major immune cells in blood were estimated from transcriptional data using computational differential estimation. Weighted correlation network analysis (adjusted for cell proportions) and fixed transcriptional repertoire analysis were used to identify associations with URT viral load, quantified as standard deviations (z-scores) from an expected trajectory over time.

*Results:* Eighty-two subjects (50% female, median age 54 years (range 3–73)) with COVID-19 were recruited. Paired URT viral load samples were available for 16 blood transcriptome samples, and 17 respiratory epithelial transcriptome samples. Natural Killer (NK) cells were the only blood cell type significantly correlated with URT viral load z-scores (r = −0.62, $P$ = 0.010). Twenty-four blood gene expression modules were significantly correlated with URT viral load z-score, the most significant being a module of genes connected around *IFNA14* (Interferon Alpha-14) expression (r = −0.60, $P$ = 1e-10). In fixed repertoire analysis, prostanoid-related gene expression was significantly associated with higher viral load. In nasal epithelium, only *GNLY* (granulysin) gene expression showed significant negative correlation with viral load.

*Conclusions:* Correlations between the transcriptional host response and inter-individual variations in SARS-CoV-2 URT viral load, revealed many molecular mechanisms plausibly favouring or constraining viral replication. Existing evidence corroborates many of these mechanisms, including likely roles for NK cells, granulysin, prostanoids and interferon alpha-14. Inhibition of prostanoid production and administration of interferon alpha-14 may be attractive transmission-blocking interventions.



## Background

The advent of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) leading to the coronavirus disease 2019 (COVID-19) has placed an enormous burden on affected individuals, healthcare systems, and economies worldwide. SARS-CoV-2 is highly transmissible and causes a wide range of severity from asymptomatic infection to severe disease and death. The amount of SARS-CoV-2 detected in the upper respiratory tract of infected individuals (URT viral load) is a key driver of transmission of infection.[1] High URT viral loads can increase household and non-household transmissions by up to nearly 60% and 40%, respectively.[2] Interestingly, URT viral load does not necessarily correlate with the severity of illness, nor is it determined by established risk factors for poor outcome such as age and sex.[3,4] This suggests that the host immune mechanisms involved in constraining the virus in the URT are different from those determining the severity of illness, although such mechanisms have not been fully elucidated. In contrast, high and persistent SARS-CoV-2 shedding in the lower respiratory tract (LRT) is associated with severe disease,[5] indicating differences in the mechanisms underlying control and pathogenesis of SARS-CoV-2 in the URT and LRT. Understanding the mechanisms controlling the viral load in the URT could illuminate new strategies to prevent transmission from infected individuals and might also enable control of the localised infection before it progresses to the LRT, triggering more serious illness.

URT viral load is highly dynamic. It changes over the course of illness due to dynamic interactions with the host immune response; culturable virus and viral RNA levels peak around the time of symptom onset and then gradually decrease to low or undetectable levels over the following 10 days (culturable virus) to two weeks (viral RNA).[3,6–8] Moreover, the kinetics of viral load vary between individuals, presumably determined by variation in immune responses.[3,8] The host response constraining viral load includes both an immediate innate component and a later adaptive response.[3,9,10] With limited in vivo data, researchers have attempted to mathematically model and explain the viral-host interaction and host immune responses to better understand the dynamics of SARS-CoV-2 viral load. We have recently developed a within-host model that has been successful in interpreting URT RNA viral load kinetics in a wide range of data including 2172 serial measurements from 605 subjects, collected from 17 different studies.[3]

Despite the dynamic interaction between the virus and host immune system during SARS-CoV-2 infection and the diversity in such interaction observed between individuals, the immune response involves conserved elements which can be reflected in host transcriptomes.[11] While gene expression is a dynamic process, and a single transcriptomic experiment usually captures only a “snapshot” in time, using robust transcriptional analyses we can pinpoint key biological mechanisms underlying the immune response. The host transcriptomic response in human infection is often studied in peripheral blood leukocytes. This is because peripheral blood leukocytes mount cell-intrinsic responses to pathogens but also mount transcriptional responses to signals arising from the organs through which they circulate. Evaluating the host transcriptome in the context of the dynamics of host-pathogen interaction can be a powerful approach to elucidate mechanisms responsible for the control of pathogen load.[12]

Many previous studies have used transcriptomics to investigate the pathogenesis of severe COVID-19,[13,14] but less attention has focused on the mechanisms controlling the dynamics of URT SARS-CoV-2 viral load. Here we sought to fill this gap in knowledge by combining our previously derived model of the trajectory of URT SARS-CoV-2 viral load[3] with analysis of the transcriptomic host response. By quantifying the deviation of measured viral load from a population average viral load trajectory, at a given time after the onset of symptoms, and correlating this with peripheral blood and nasal epithelial transcriptomes, we identified mechanisms which may restrict or promote viral replication. The mechanistic correlates of URT viral load identified herein may be important to develop new therapeutic and vaccine strategies to block transmission of SARS-CoV-2.

## Results

### *Participants*

Subjects were recruited as part of our previously published studies of the blood[15] and nasal[16] transcriptomes in COVID-19. Transcriptome data was available for 82 COVID-19 patients (50% female, median age 54 years (range 3–73 years)) recruited during the “first wave” (February to May 2020) of COVID-19 in Spain, before vaccination and natural infection became determinants of the immune response to SARS-CoV-2. Whole blood RNA sequencing (RNA-Seq) and nasal epithelium *n*Counter NanoString gene expression assay data were generated (see Methods) for 68 and 24 subjects, respectively, with 10 subjects being included in both analyses (Fig. 1). Clinical characteristics of all subjects are provided in Supplementary Table 1.

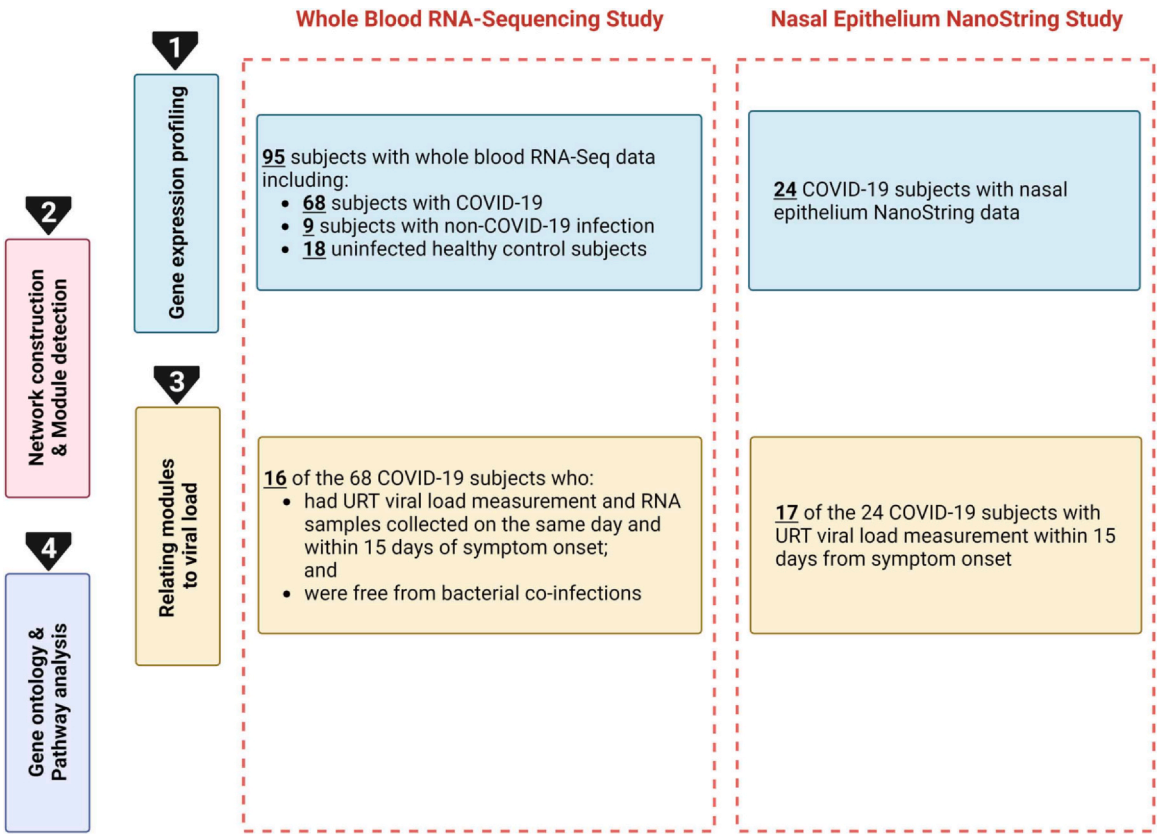


**Fig. 1.** Subject selection for weighted correlation network analysis. The flow chart provides a breakdown of the number of subjects analysed and the selection criteria applied at each analysis step.

The whole blood transcriptome profiles were used to construct gene co-expression networks and detect clusters of interconnected genes (Fig. 1, see below). For gene module discovery, to optimise the generalisability of modules, we included all 68 COVID-19 subjects with RNA-Seq data (regardless of whether they had co-infections), and an additional 18 uninfected healthy control subjects and 9 subjects with non-COVID-19 infections (4 bacterial and 5 viral) (Supplementary Table 1), all sequenced in the same batch. However, of the COVID-19 cases who were free from suspected or proven bacterial co-infections, only 16 had URT viral load measurement and RNA samples collected on the same day and within 15 days of symptom onset (after which any detected viral RNA is exceedingly unlikely to represent culturable virus[17]). Only these 16 subjects were included in analyses correlating URT viral load with the whole blood transcriptome (Table 1, Fig. 2A). Subjects were mostly female (57.1%), with ages ranging from 3 to 78 years (median = 55 years) (Fig. 2B and C). The disease severity was mild (n = 3; 19%), moderate (n = 7; 44%), and severe (n = 6; 37%).

We also performed a *n*Counter NanoString gene expression analysis on nasal epithelium samples from 24 COVID-19 patients, including 17 with URT viral load measurement on the same day as nasal epithelium sample collection and within 15 days from symptom onset (Fig. 1, Table 2). The subjects' ages ranged from 16 to 80 years (median = 47 years), and most had mild disease (n = 9; 53%) or severe disease (n = 6; 35.3%).

*Conversion of viral load measurements to z-scores using viral load regression model*

We recently developed a regression model fitted to viral load measurements within the first 15 days of symptoms across 16 datasets, capturing the viral load variation during the course of infection between different individuals.[3] Here, to quantify whether individual subjects in the current study had higher or lower than average viral load measurements relative to their duration of illness, we used the previously published regression model to calculate a z-score for each viral load measurement representing the deviation of that measurement from the mean viral load trajectory (regression line) at a given time from onset of symptoms (Table 1 and 2, Fig. 2D and E; see Methods).

Viral load z-scores calculated from the data in this study were not associated with the severity of illness (Fig. 2F). In our previous large-scale analysis of COVID-19 subjects,[3] we showed that age and sex did not significantly influence URT viral load dynamics and that URT viral load dynamics did not affect the severity of illness. Therefore, in the present study, we did not adjust the viral load z-scores for these variables.

*Exploring molecular correlates of SARS-CoV-2 viral load using whole blood transcriptomics*

We aimed to identify groups of genes for which expression correlated with viral load z-score, providing insights into the mechanisms controlling viral load. We first performed a gene signature-based deconvolution,[18] as in our previous studies.[12,19] Interestingly, the computed proportion estimate of natural killer (NK) cell population was negatively correlated with viral load z-score (r = −0.62 and $P = 0.010$, Fig. 2G and Supplementary Figure 1). There was insufficient evidence to conclude a significant linear relationship between other leukocyte populations (B-cells, monocytes, neutrophils, $CD4^+$ T-cells, and $CD8^+$ T-cells) and viral load z-score. Gene expression counts were then adjusted for leukocyte mixture to remove the confounding effect of differences in blood leukocyte proportions

**Table 1**
Samples used to correlate URT viral load and whole blood transcriptome.

| Days of illness[a] | Age (year) | Sex | Severity | Average Ct value[b] | Calculated viral load (log10 (viral load/ml)) | Viral load z-score[c] |
|---|---|---|---|---|---|---|
| 3 | 15 | Female | Mild | 27.72 | 5.49 | -0.32 |
| 4 | 55 | Male | Moderate | 37.08 | 2.52 | -2.89 |
| 5 | 70 | Male | Severe | 31.53 | 4.34 | -0.99 |
| 8 | 73 | Female | Severe | 35.87 | 5.60 | 0.80 |
| 8 | 43 | Male | Moderate | 27.41 | 2.91 | -1.71 |
| 9 | 55 | Male | Severe | 35.43 | 3.04 | -1.39 |
| 9* | 53 | Male | Moderate | 32.36 | 3.83 | -0.65 |
| 9 | 36 | Female | Mild | 33.1 | 3.98 | -0.51 |
| 10 | 3 | Female | Mild | 21.03 | 3.23 | -1.00 |
| 10 | 53 | Female | Severe | 34.39 | 7.45 | 2.94 |
| 11 | 72 | Female | Moderate | 29.97 | 4.77 | 0.64 |
| 11 | 60 | Male | Severe | 33.24 | 3.73 | -0.33 |
| 12 | 41 | Female | Moderate | 36.64 | 2.90 | -0.91 |
| 13* | 53 | Male | Severe | 28.62 | 7.93 | 4 |
| 13 | 71 | Male | Moderate | 19.57 | 5.54 | 1.76 |
| 15 | 71 | Female | Moderate | 37.17 | 2.48 | -0.69 |

* These two samples are from the same subject.
[a] How many days after symptom onset the viral load was measured.
[b] Cycle threshold value.
[c] Relative to the days of illness at the time of viral load measurement (see Fig. 2D).

between individuals. To make the best use of the relatively small sample size of the selected 16 samples, we performed dimensionality reduction using weighted correlation network analysis (WGCNA).[20] First, we clustered the RNA-Seq profiles (n = 96) and removed an outlier (Supplementary Figure 2). Then, a gene co-expression network was constructed, and modules (clusters of highly co-expressed genes) were detected using the remaining 95 whole blood RNA-Seq profiles. Next, we correlated the first principal

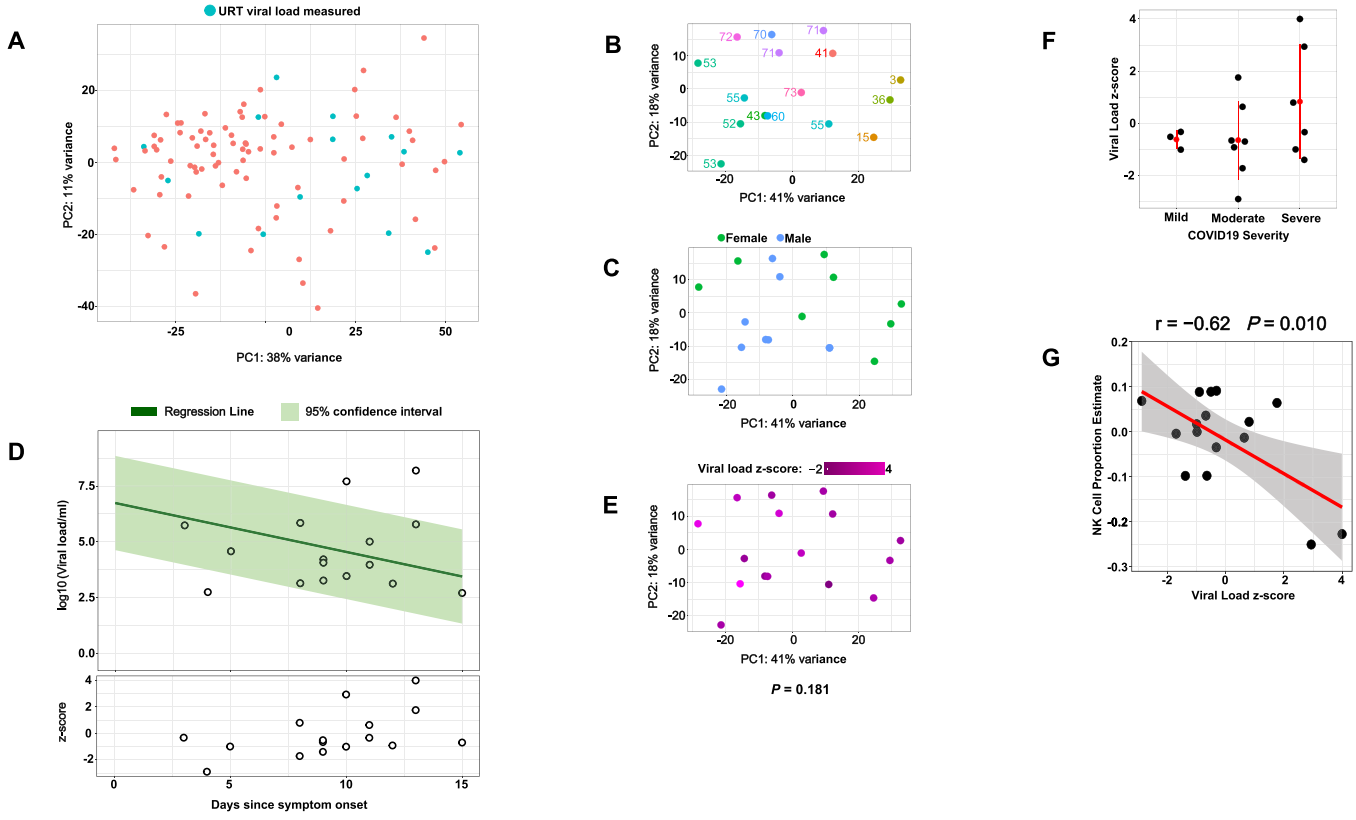


**Fig. 2.** Overview of peripheral blood gene expression and viral load in subjects with COVID-19. A) PCA (principal component analysis) plot of peripheral blood gene expression determined by RNA-Seq. Samples with paired URT viral load measurements are coloured as blue. B) and C) PCA plots represent samples with paired RNA-Seq and viral load data coloured by age and sex, respectively. D) Calculation of viral load z-scores. In the upper panel, the viral load data of the present study (black circles) are plotted against the time since symptom onset. The green line indicates a linear regression model fitted to the viral load data from 16 different datasets previously studied. The shaded green area represents the 95% confidence interval for the regression model. As shown in the lower panel, for each data point, a z-score is calculated as the distance of the data point from the mean trajectory (green line). E) PCA plot of samples with paired data coloured based on viral load z-score. F) Viral load z-score is compared between groups of different COVID-19 severity. Red dots and whiskers represent mean and 1 standard deviation. G) The correlation between computed proportion of NK cells within blood populations and URT viral load is plotted. Each dot represents a sample with paired viral load and blood transcriptome data. The grey shading represents the 95% confidence interval around the line of best fit (red).

**Table 2**
Subjects used to correlate URT viral load with nasal epithelium NanoString profiles.

| Days of illness | Age (year) | Sex | Severity | Average Ct value | Calculated viral load (log10 (viral load/ml)) | Viral load z-score |
|---|---|---|---|---|---|---|
| 0 | 29 | Female | Mild | 27.53 | 5.94 | -0.74 |
| 3 | 16 | Female | Mild | 27.72 | 5.74 | -0.32 |
| 4 | 80 | Female | Severe | 25.36 | 6.46 | 0.56 |
| 5 | 70 | Male | Severe | 31.53 | 4.57 | -0.99 |
| 6 | 43 | Male | Mild | 38.16 | 2.39 | -2.8 |
| 7 | 47 | Male | Mild | 29.9 | 5.01 | -0.17 |
| 8 | 73 | Female | Severe | 27.41 | 5.85 | 0.8 |
| 8 | 34 | Female | Mild | 28.93 | 5.33 | 0.33 |
| 8 | 34 | Male | Mild | 35.06 | 3.41 | -1.45 |
| 9 | 55 | Male | Severe | 35.43 | 3.26 | -1.39 |
| 9 | 38 | Male | Mild | 36.67 | 2.45 | -2.14 |
| 10 | 73 | Male | Severe | 28.86 | 5.32 | 0.72 |
| 10 | 48 | Female | Mild | 28.78 | 5.48 | 0.87 |
| 11 | 26 | Female | Mild | 36.92 | 2.54 | -1.65 |
| 12 | 30 | Male | Moderate | 27.97 | 5.61 | 1.4 |
| 12 | 62 | Male | Severe | 32.49 | 4.29 | 0.18 |
| 13 | 69 | Female | Moderate | 25.3 | 6.36 | 2.3 |

component of each module (module eigengene) to viral load z-scores in the group of 16 samples with paired data, reasoning that inducible mechanisms which restrict viral load would be enriched amongst the most strongly correlated modules. Twenty-four modules were significantly correlated with viral load ($P < 0.01$; Fig. 3, Table 3, and Supplementary Table 2). To aid interpretation, we represented each module by its hub gene (gene with the highest connectivity within the module). Fourteen modules were positively correlated with viral load z-score and 10 were negatively correlated (Fig. 3A). *IFNA14* (Interferon Alpha-14) and *AIPL1* (Aryl Hydrocarbon Receptor Interacting Protein Like 1) modules showed the strongest negative correlation with viral load (r = −0.60 with $P$ = 1e-10 and r = −0.60 with $P$ = 2e-10, respectively). The largest positive correlation was observed for the *AC011455.2* module (r = 0.60, $P$ = 2e-10).

We selected the top 6 significant modules for further analysis: *IFNA14*, *AIPL1*, *AC011455.2*, *C7orf33* (Chromosome 7 Open Reading Frame 33), *IFNL3* (Interferon Lambda 3), and *GALNT17* (Polypeptide N-Acetylgalactosaminyltransferase 17). *AC011455.2*, *C7orf33*, and *GALNT17* modules were positively correlated with viral load z-score and positioned very close to each other in the hierarchical clustering (Fig. 3B). Therefore, we merged their gene sets to form a metamodule (*AC011455.2*/*C7orf33*/*GALNT17*; total gene count = 114) for further data analysis, assuming that the higher gene count would increase power to detect biologically relevant changes. The Fig. 3C heatmap provides an alternative representation of the correlation between the modules and their association with viral load z-score. We used Qiagen's Ingenuity Pathway Analysis (IPA) for biological understanding of the modules.[21] Correlation coefficients between module genes and viral load z-score were used to infer the activity

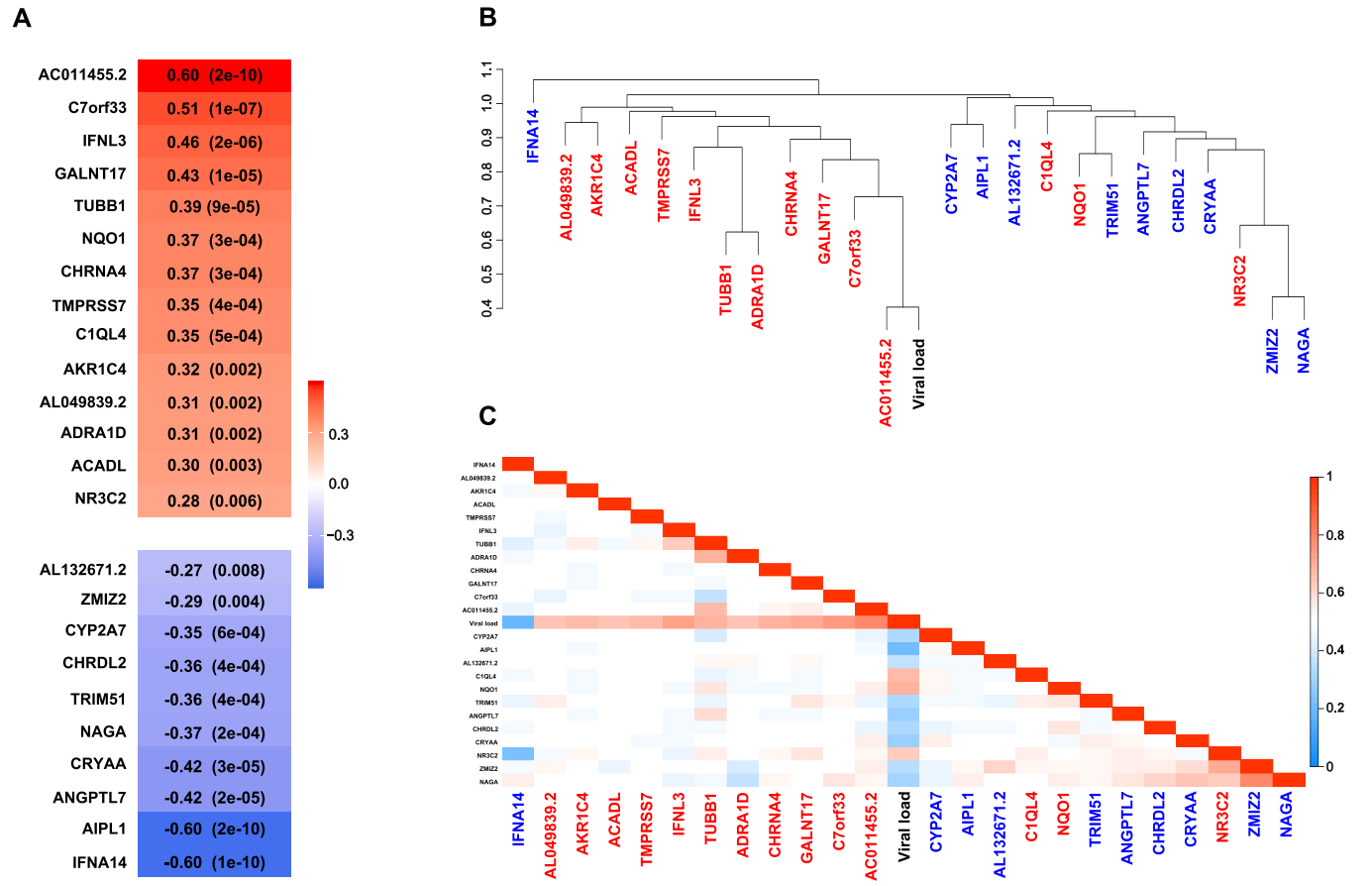


**Fig. 3.** Peripheral blood gene expression modules correlated with viral load z-score. A) For each module, the hub gene, Pearson correlation with viral load z-score and corresponding p-value (in parentheses) are displayed. The Pearson correlation scale is depicted on the right. The modules are ranked based on the correlation between the modules and viral load z-score. B and C) Module network and relationship with viral load z-score. The hierarchical clustering dendrogram of the module eigengenes (B) was generated using all genes in the modules and shows the dissimilarity of eigengenes with the distance measure being one minus correlation. Modules coloured in red and blue are, respectively, positively and negatively correlated with viral load. The heatmap (C) represents module eigengene adjacency calculated as (1 + correlation)/2.

**Table 3**
Modules significantly correlated with viral load z-score. Each module is represented by its hub gene. Complete lists of module genes and their information are provided in Supplementary Table 2.

| Module | Correlation | *P* | Gene count[a] | Genes with highest contribution to the module[b] |
|---|---|---|---|---|
| IFNA14 | -0.60 | 1e-10 | 45 | *IFNA14, ADAD1, IL22, LIN28A, TMEM270* |
| AIPL1 | -0.60 | 2e-10 | 43 | *AIPL1, CFAP100, KCNK3, MOGAT2, OR5D3P, SLC13A2, SPEM1* |
| AC011455.2 | 0.60 | 2e-10 | 41 | *AC011455.2, FRMD1, TTC6, KCNK18, KIF12, SNAI2* |
| C7orf33 | 0.51 | 1e-07 | 36 | *C7orf33, DGKB, GK2, GHRH, PDZD9, SEPTIN14, SMIM40, TTLL2* |
| IFNL3 | 0.46 | 2e-06 | 50 | *IFNL3, STRA8* |
| GALNT17 | 0.43 | 1e-05 | 37 | *GALNT17, GCSAML-AS1, IP6K3, NPY, OR6B2, SIM2, SSU72P2, SSX4* |
| ANGPTL7 | -0.42 | 2e-05 | 43 | *ANGPTL7, CALML3, GFY, HNRNPCL2, SLC6A1, TPTE, UGT1A3* |
| CRYAA | -0.42 | 3e-05 | 57 | *CRYAA*, AC104581.2*, ACSM4*, FABP6*, GBX2*, PASD1*, PCDHA12*, PDYN*, SSMEM1** |
| TUBB1 | 0.39 | 9e-05 | 349 | *TUBB1, NRGN, SELP, MPIG6B, SPARC, GP9, PTGS1, GP6, CTTN, ABLIM3, ARHGAP6, CMTM5, GP1BB, TSPAN9, ITGB5, GUCY1B1, TREML1, PGRMC1, MYLK, ITGB3, ITGA2B, PTCRA* |
| NAGA | -0.37 | 2e-04 | 638 | *NAGA, RASSF4, ZNF385A* |
| NQO1 | 0.37 | 3e-04 | 89 | *ABCG5*, AC003688.1*, AL121899.2*, CCL26*, DRD1*, SERPINB13*, SP5** |
| CHRNA4 | 0.37 | 3e-04 | 51 | *CHRNA4, CPN2, EPYC, GGTLC2, HDGFL1, MARCOL, MUC21, OR1D2, OR4F15, RNASE11, SCEL, SLC35F4, XAGE2* |
| TRIM51 | -0.36 | 4e-04 | 32 | *TRIM51, AL583836.1, EPHA5, KRTAP10–7, LUZP2, TCF23* |
| CHRDL2 | -0.36 | 4e-04 | 31 | *CHRDL2, EGFLAM, OR51L1, PPFIA2, SKOR2, SLCO1B1, TM4SF4* |
| TMPRSS7 | 0.35 | 4e-04 | 48 | *AC097636.1, APOF, BPIFB2, FOXR1, MMP10, MMP12, FLG* |
| C1QL4 | 0.35 | 5e-04 | 42 | *C1QL4, DNMT3L, KRTAP10–8, OR14J1, OVOL3, PMIS2, PPP1R14D, TFAP2D, UBTFL1, H2AC18** |
| CYP2A7 | -0.35 | 6e-04 | 87 | *CYP2A7, DCX, NGB, NR0B1, SLC17A6, SPHKAP, SPRR2D, CEACAM18, OOSP4A* |
| AKR1C4 | 0.32 | 0.002 | 37 | *AKR1C4, RGS21, CCDC63, DSG1, GLRA1, IL20, TMEM174* |
| AL049839.2 | 0.31 | 0.002 | 38 | *AL049839.2, CCDC190, FAM236C, FAM236D, VSNL1, MAGEB1, NRAP, RHCG* |
| ADRA1D | 0.31 | 0.002 | 54 | *ADRA1D, DIO3, EPHA6, H2BC1, IL31, PDE6C, SPATA31D4, TCEAL5, UNCX* |
| ACADL | 0.30 | 0.003 | 172 | *ACADL, AC008770.4, ADCY8, APOA4, ASZ1, BTBD16, CASQ2, CCK, CNTN1, COMP, CYP3A7, CYP4A11, DAZ1, DCAF12L1, DSG4, DUX4, DYDC2, FOXG1, FOXR2, FSHR, GRIA1, GRIA2, GRM6, IFNA13, KIF2B, KLK3, KRTAP2–2, MISP, NPY2R, NRK, NTF3, NTSR2, NXPH1, NXPH2, OR13J1, OR14A2, OR14C36, OR1C1, OR51G2, OR5L1, OR5W2, OR7A10, OR7E24, OTX2, PAX1, PAX7, PHOX2B, PRAMEF7, PRDM9, PRM3, PSG7, RPTN, SCYGR2, SERPINA5, SLC19A3, SLC2A7, ST8SIA3, SULT2A1, TBPL2, TRIML1, TSPY10, TSPY2, TSPY3, TSPY8, UGT1A4, ZIC1, ZSCAN5C* |
| ZMIZ2 | -0.29 | 0.004 | 711 | N/A[c] |
| NR3C2 | 0.28 | 0.006 | 208 | N/A[c] |
| AL132671.2 | -0.27 | 0.008 | 97 | *AL132671.2, C2orf72, CCDC166, FGF23, MBL2, MUC3A, MYOG, SPINK6, STRA6* |

[a] The number of genes involved in each module.
[b] Genes with the absolute value of module membership (the correlation between the module eigengene and gene expression values) > 0.9. Genes with negative module membership are marked with an asterisk.
[c] No genes had module membership higher than 0.9 or lower than −0.9.

pattern (activation or inhibition) of the biological processes involved such as enriched pathways and upstream regulators.

Fig. 4A illustrates top enriched canonical pathways for the *IFNA14*, *AIPL1, AC011455.2/C7orf33/GALNT17,* and *IFNL3* modules (details provided in Supplementary Table 3). Of these, 3 pathways showed an enrichment *P*-value of below 0.001 including 'pathogen-induced cytokine storm signalling' (*P* = 5e-4) and 'IL-22 (Interleukin-22) signalling' (*P* = 8e-4) enriched in the *IFNA14* module, and 'melatonin degradation' (*P* = 2e-4) enriched in the *IFNL3* module. The pathogen-induced cytokine storm signalling pathway encompasses the highest number of genes from the tested module (*IFNA14*, *IL22*, *CCL4* (C-C Motif Chemokine Ligand 4), *CD70*, and *COL4A4* (Collagen Type IV Alpha 4 Chain) from the IFNA14 module) compared to the other enriched pathways identified. *IFNA14* and *IL22*, two main members of the *IFNA14* module (module membership = 0.98 and *P* = 5e-64 for both genes), are key components of the pathway (Fig. 4B). Interestingly in our dataset both genes were negatively correlated with viral load z-score (r = −0.59 and *P* = 3e-10 for both genes), whereas *CCL4*, *CD70*, and *COL4A4* were positively correlated with viral load z-score (Fig. 4B). The IL-22 signalling pathway involves two genes, *IL22* and *IL22RA2* (Interleukin 22 Receptor Subunit Alpha 2), with a high contribution to the *IFNA14* module (*IL22RA2* module membership = 0.75 with *P* = 1e-18) both negatively correlated with viral load z-score (for *IL22RA2*, r = −0.63 and *P* = 7e-12; Fig. 4B). The superpathway of melatonin degradation includes three genes (*CYP2F1* (Cytochrome P450 Family 2 Subfamily F Member 1), *IL4I1* (Interleukin 4-induced gene-1), and *UGT1A1* (UDP Glucuronosyltransferase Family 1 Member A1) from the IFNL3 module.

*STING1* (Stimulator of Interferon Response CGAMP Interactor 1), *RBP3* (Retinol Binding Protein 3), and *RORC* (RAR Related Orphan Receptor C) were predicted to be the most significant upstream regulators of *IFNA14* module genes (*P* = 6e-5, 3e-4, and 4e-4, respectively) (Fig. 4C and D). They interact with *IL22*, *IL22RA2*, *CCL4*, *CD70*, and *FEZ1* (Fasciculation and Elongation Protein Zeta 1), members of the *IFNA14* module which are mapped to the pathogen-induced cytokine storm signalling and IL-22 signalling pathways (Supplementary Tables 4 and 5). None of the identified pathways and regulators showed reliable evidence of activation (absolute value of IPA activation z-score > 2) and therefore we were not able to infer an overall directionality (activation or inhibition) with respect to viral load.

We also applied the *BloodGen3Module* tool[22] to identify gene modules associated with viral load. Unlike WGCNA which detects modules from the analysed gene expression dataset, *BloodGen3Module* uses fixed functionally pre-annotated modules characterising different biological responses of distinct blood cell types. We used RNA-Seq data without adjustment for leukocyte-mixture and evaluated differential expression of these modules between samples with positive and negative viral load z-scores (n = 5 and n = 11, respectively). We identified an aggregate of five modules showing high 'module response' and higher module expression in subjects with positive viral load z-score (aggregate module A34; Fig. 5A and Supplementary Table 6). A module response is defined as the percentage of genes for a given module showing significant differential expression between the groups. From the module aggregate, the Prostanoids module showed the highest module response (97%). Interestingly, we observed a significant overlap between the A34 aggregate module and *TUBB1* (Tubulin Beta 1 Class VI) module which was found to be significantly positively correlated with viral load z-score by WGCNA (r = 0.39 with *P* = 9e-05; Fig. 3).

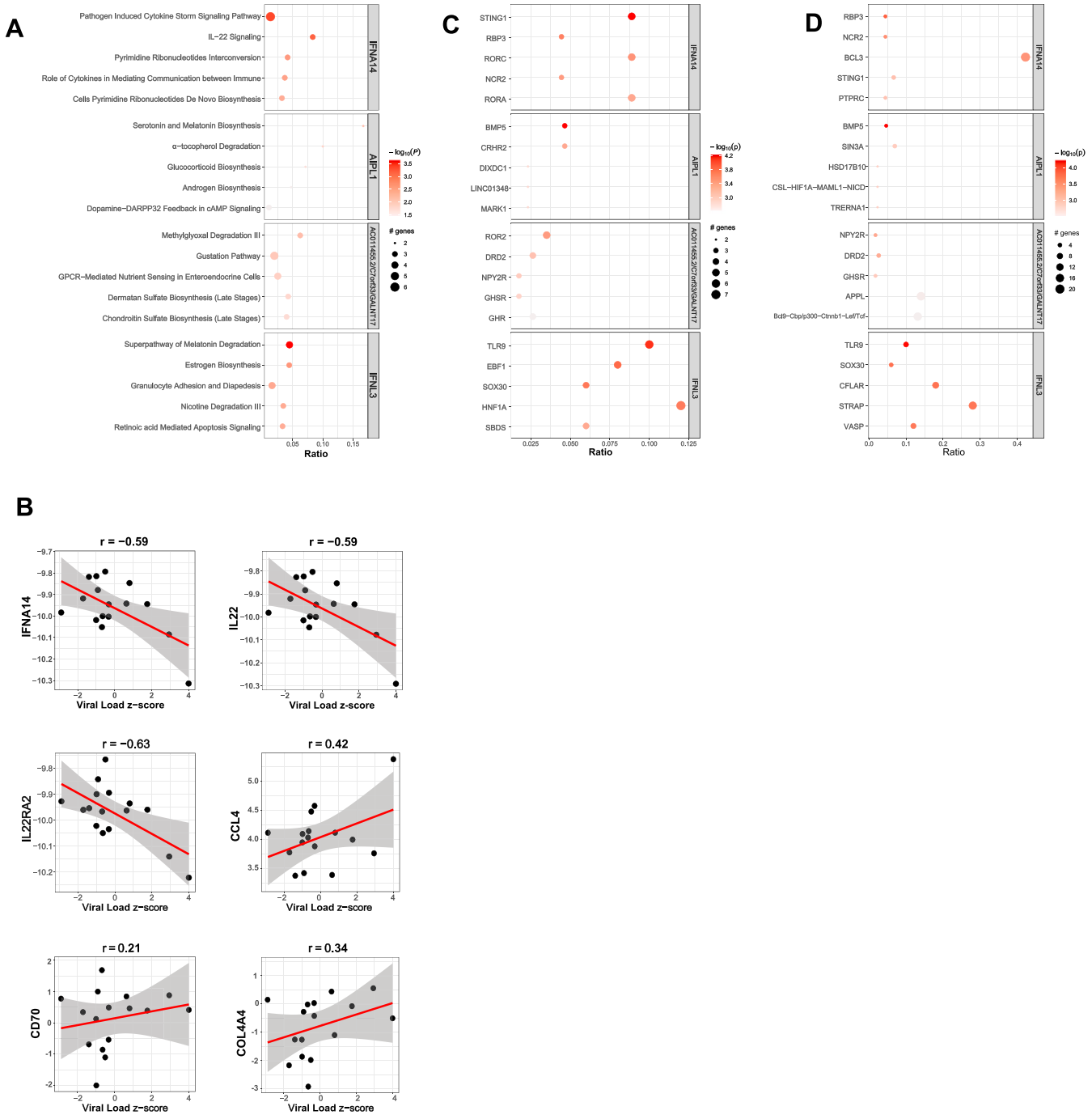


**Fig. 4.** Ingenuity pathway analyses of peripheral blood gene expression modules most strongly correlated with viral load. A) For each module, the top 5 significant pathways are illustrated in descending order of statistical significance as indicated by colour. For each pathway, the size of the corresponding circle represents the number of module genes that map to the pathway. The x-axis shows the ratio of the number of genes common between the corresponding module and pathway divided by the total number of genes that map to the same pathway. B) The correlations between the expression of the main *IFNA14* module genes, adjusted for cell mixture, and URT viral load are plotted. Each dot represents a sample with paired viral load and blood transcriptome data. The grey shading represents the 95% confidence interval around the line of best fit (red). C) and D) For each module, the 5 most significant upstream (C) and master regulators (D) are illustrated in descending order of statistical significance as indicated by colour. For each regulator, the size of the corresponding circle represents the number of module genes downstream to the regulator. The x-axis shows the ratio of the number of module genes downstream to the corresponding regulator divided by the total number of module genes.

Seventy genes including *TUBB1* were common between the A34 and *TUBB1* modules (Fig. 5B and Supplementary Tables 6) while A34 did not overlap with any other WGCNA module correlated with viral load z-score. Among the A34 modules, the Prostanoids module showed the highest overlap with the *TUBB1* module (from 36 genes involved in the Prostanoids module, 30 were also included in the *TUBB1* module).

*Exploring molecular correlates of SARS-CoV-2 viral load using NanoString assay of nasal epithelium*

We analysed RNA isolated from nasal epithelium samples of 24 COVID-19 patients using a NanoString panel of 579 genes involved in core pathways and processes of human immune responses (Supplementary Table 7). Seventeen subjects also had paired URT

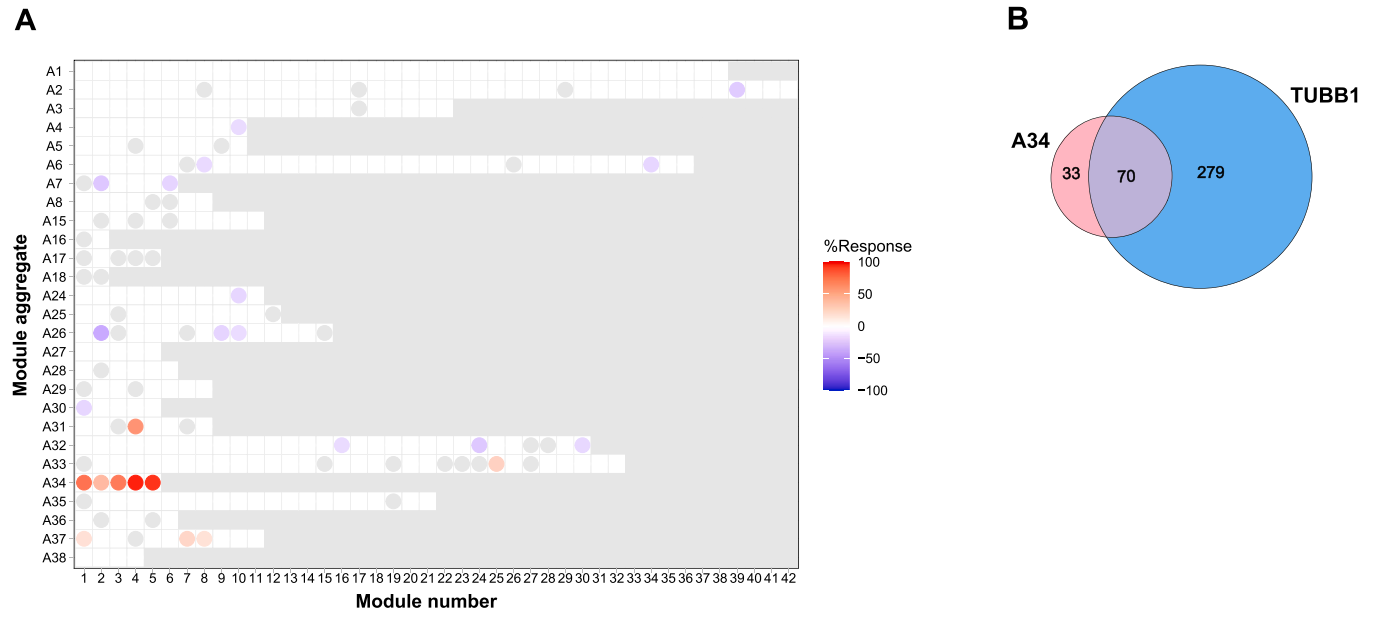


**Fig. 5.** Pre-annotated blood gene expression modules associated with viral load. A) Module fingerprint grid plot. The differential expression of the modules is compared between two groups with positive and negative viral load z-score using t-test with fold change and p-value cut-off of 0.5 and 0.05, respectively. Each module is allocated a fixed position (block) on the grid, where each row represents a 'module aggregate.' The grouping of these modules into separate aggregates was determined by similarities in abundance levels across the 16 reference datasets used by *BloodGen3Module*, which correspond to different immune states. Red and blue spots represent modules with increased and decreased abundance in the positive vs negative viral load z-score group, respectively. The gradient represents 'module response' which is the percentage of genes for a given module showing significant change in abundance between the two groups. Only modules with at least 15% response have been shown. B) Overlap of genes between A34 and *TUBB1* module.

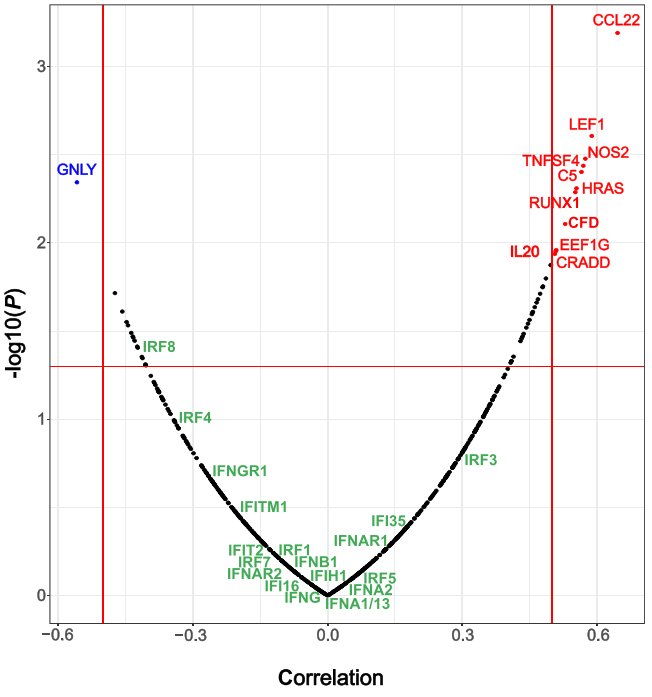


**Fig. 6.** Correlation between nasal epithelium transcriptome and viral load z-score. The volcano plot illustrates correlation coefficients and corresponding p-values. Each dot represents a gene included in the NanoString panel. Genes strongly correlated with viral load z-score (absolute correlation coefficient > 0.5 and P < 0.05) are coloured in red (positive correlation) and blue (negative correlation). Interferon pathway genes are depicted in green.

viral load measurement within 15 days from symptom onset and had no evidence of bacterial co-infection. Using WGCNA we identified seven gene co-expression networks which we refer to as "pseudo-modules" since they were detected using a relatively low number of genes included in the NanoString panel (Supplementary Table 8). Only one pseudo-module was correlated with viral load z-score at significance threshold of 0.05 (*PTK2* (Protein Tyrosine Kinase 2) module; r = 0.485 and *P* = 0.016). Additionally, correlation analysis between individual gene expression and viral load z-score detected significant correlation (absolute correlation coefficient > 0.5 and *P* < 0.05; Fig. 6) in 12 individual genes, 11 of which were positively correlated and one (*GNLY* (Granulysin)) was negatively correlated. In order to gain a deeper understanding of genes that exhibit a positive correlation with viral load, we analysed a list of 88 genes showing moderate to strong positive correlations (r ≥ 0.3, Supplementary Table 9) using IPA. 'NOD1/2 signalling' (*P* = 1e-24) and 'pathogen-induced cytokine storm signalling' (*P* = 9e-23) were found to be the most enriched pathways.

## Discussion

Understanding mechanisms controlling SARS-CoV-2 viral load in the URT can provide valuable leads towards treatment and vaccine strategies aimed at reducing viral transmission.[8] Such strategies have recently been highlighted as potential "game changers" as societies adapt to living with COVID-19.[23] Current evidence suggests that the mechanisms controlling URT viral load may be different from those controlling LRT viral load and disease severity.[3–5] However, our current knowledge concerning control of URT viral load is far from complete. To unravel the biological complexity underlying the control of SARS-CoV-2 viral load, we sought to identify correlates of the variation in viral load which occurs in naturally infected individuals. Of note, we used samples from the first wave of infection in Europe, prior to vaccination, infection-induced immunity, and circulation of important variants of SARS-CoV-2. We included a wide age range and did not adjust for age or sex, in order to maximise heterogeneity and variation in control of viral load and thus detection of the mechanistic correlates. To account for the dynamic nature of URT viral load, which rapidly increases to a peak just before symptom onset and then declines more slowly, we quantified viral

loads by their standardised deviation (z-scores) from a previously derived average trajectory.[3] We correlated viral load z-scores with paired peripheral blood and nasal epithelium transcriptomes.

After excluding individuals with proven or suspected bacterial co-infection and URT SARS-CoV-2 viral load samples taken more than 15 days after symptom onset, relatively small numbers of subjects for blood and nasal transcriptome analysis (n = 16 and n = 17, respectively) remained. This prompted the use of gene modules rather than individual genes for our primary analysis. This would reduce the complexity of large gene networks into relevant modules and increase the statistical power to detect those correlating with viral load. An individual module compromises genes that are more densely connected than expected by chance and often involved in the same biological functions.[24] We applied two different methods to detect gene clusters, *WGCNA* and *BloodGen3Module*. The first identifies modules directly from the gene expression data and the later uses pre-annotated modules.

The peripheral blood module most significantly associated with URT viral load, and least related to other modules, had *IFNA14* as its hub gene. *IFNA14* encodes the type I interferon, interferon α14. Interferons are glycoprotein cytokines made and released by various cell types including host lymphocytes and are known to be key effectors in antiviral responses. However, their pattern of expression and function during SARS-CoV-2 infection is controversial. While some studies suggest protective effects of interferons in severe COVID-19,[25–27] others indicate poor clinical outcomes in those with increased production of interferons.[28–31] There is relatively limited data available on the relationship between interferon expression and variation in human SARS-CoV-2 URT viral load dynamics. Sposito et al. evaluated nasopharyngeal swabs of COVID-19 patients and showed that the expression of type I and III interferons was significantly associated with viral load in patients under 70 years old.[31] However, those aged over 70 years showed no association and/or showed a significantly lower correlation coefficient. This evaluation did not include *IFNA14*. Also, it appears that their viral load measurements were not adjusted for the time between sample collection and symptom onset. In another study of nasopharyngeal swabs taken from COVID-19 patients, Ziegler et al. showed an expansion of ciliated cells responsive to interferon in “high” viral load COVID-19 samples.[32] *IFNA14* has been shown to activate a potent antiviral response via binding to IFNAR1 and IFNAR2 (Interferon Alpha and Beta Receptor Subunits 1 and 2) receptors.[33,34] This triggers the activation of JAK/STAT (Janus Kinase/Signal Transducer and Activator of Transcription) signalling complexes which subsequently induces the expression of ISGs (interferon-stimulated genes) that inhibit virus infection.[35] Cheemarla et al. have shown that under specific circumstances, ISG-mediated defences can significantly inhibit the replication of SARS-CoV-2.[36] The strong negative correlation between the *IFNA14* module as well as *IFNA14* as an individual gene and viral load in our data suggests that IFNA14 signalling could play a key role in controlling SARS-CoV-2 viral load, i.e. increased expression of *IFNA14* restricts viral replication. Schuhenn et al. recently showed that IFNA14 is one of the most potent interferon alpha subtypes inhibiting SARS-CoV-2 replication and can cause a significant reduction of SARS-CoV-2 viral titre by up to $10^5$-fold.[37] Furthermore, unpublished data suggest that, compared to IFNA2 (Interferon Alpha-2), which was used to treat COVID-19 patients in an uncontrolled exploratory study in China,[38] IFNA14 is more efficient at preventing the infection while less detrimental to the immune system.[39] Not only is *IFNA14* important as an individual gene, but it also represents a network of highly connected genes in our data, the *IFNA14* module, which showed a high enrichment of two canonical pathways ‘pathogen-induced cytokine storm signalling’ and ‘IL-22 signalling’. Although SARS-CoV-2 can trigger a ‘cytokine storm’,[40,41] the changes in expression of genes in this pathway were not consistently associated with activation or inhibition of the pathway, agreeing with previous findings that ability to control of URT viral load is dissociated from severity of illness.[3,4] IL-22 is a cytokine released by several immune cells such as Th22 (T helper cells type 22) and plays an important role at mucosal barriers, orchestrating the interaction between the epithelial cell layer and local immune system in response to infections.[42] IL-22 stimulates the IL-22 receptor complex on epithelial cells resulting in downstream activation of JAK-STAT signalling pathway which induces multiple antiviral responses and therefore can be protective during SARS-CoV-2 infection.[42–44] Elevated levels of IL-22 in the plasma have been implicated as a hallmark of severe COVID-19.[29] Taken together there is compelling evidence that the genes in the *IFNA14* module act to reduce URT viral load, and add to the evidence that interferon α14 should be considered as a candidate treatment to reduce viral load in the URT and decrease transmissibility of SARS-CoV-2.

The *AIPL1* module was the second top module negatively corelated with viral load z-score. Unlike *IFNA14*, *AIPL1* is not known to be involved with the pathogenesis of COVID-19, and the enriched pathways for this module contained relatively few genes. Nevertheless, enrichment of the ‘α-tocopherol degradation’ pathway suggests a potential role of α-tocopherol (also known as Vitamin E) in the control of viral load. α-tocopherol is an antioxidant which may enhance the function of innate and adaptive immune cells, for example increasing NK cell activity and the phagocytic capacity of leukocytes, which could bolster the immune response to reduce pathogen load as observed in influenza.[45,46] Emerging evidence suggests that water soluble derivatives of α-tocopherol have potent antiviral properties especially when they are used synergistically with remdesivir to inhibit SARS-CoV-2 RNA-dependent RNA polymerase.[47]

We also identified modules positively correlated with viral load, possibly indicating that these modules are induced in response to increasing amounts of virus or that expression of these genes favours an increase in viral replication. The most significant of these modules was the *IFNL3* module. The hub gene, *IFNL3*, encodes a type III interferon which is a cytokine activated in response to mucosal viral infections and signals through the heterodimeric IFNLR (Interferon Lambda Receptor) that is expressed distinctly in the URT epithelial cells. This stimulates the activation of several transcription factors which upregulate ISGs. Type III interferon signalling pathway is considered slower and induces a weaker ISG response than type 1 interferons.[48,49] The most significant upstream regulator of the *IFNL3* module is TLR9 (Toll Like Receptor 9), which may be stimulated by unmethylated CpG (Cytosine-phosphate-Guanine) sequences during SARS-CoV-2 infection[50] and result in the observed upregulation of *IFNL3* module genes.

Using *BloodGen3Module*, we identified a cluster of blood pre-annotated transcriptional modules (A34) positively correlated with viral load. This was particularly interesting as these modules showed a significant overlap with the *TUBB1* module found to be significantly positively correlated with viral load by WGCNA. From the A34 modules, the Prostanoids module showed the highest module response and overlap with the *TUBB1* module. Prostanoids are a subclass of eicosanoids and regulate the inflammatory response.[51] The observed association between the prostanoids module expression and viral load z-score suggests that high levels of prostanoids may suppress processes which constrain viral load and therefore promote high viral load levels. This is supported by a recent study showing that abrogation of eicosanoid signalling reduces viral load and rescues mice from fatal SARS-CoV-2 infection.[52] It is also intriguing that the top *BloodGen3Module* modules (modules 1, 2, 4 and 5 from aggregate module A34; Supplementary Tables 6) are connected to platelets. TUBB1 protein is specifically expressed in platelets and megakaryocytes and involved in proplatelet synthesis and platelet release.[53] Thus, these findings may indicate an interplay between platelets and high viral load.[54]

In addition to peripheral blood samples, we studied samples taken from the primary infection site, the nasal epithelium. Both blood and nasal transcriptomes can reflect the host immune response to the infection. In a respiratory infection, epithelial cells are directly infected, and peripheral blood leukocytes also respond to signals arising from the site of infection.[55] Rajagopala et al. showed that the initial immune response in the nasal mucosa to SARS-CoV-2 infection is viral load dependent and high viral load enhances interferon signalling in the upper respiratory mucosa.[56] In our study, the difference in the transcriptomic analysis approach we used for the peripheral blood and nasal samples (RNA-Seq and NanoString assay, respectively) made it difficult to compare the results obtained from the two tissue types. The NanoString assay analysed a relatively small number of genes (579 genes involved in immune response) and therefore the data did not yield reliable module level results. However, individual genes correlated with URT viral load z-score were identified that may be of interest. *GNLY* was the only negatively correlated gene representing a likely role in the control of viral load. It is produced by a variety of killer cells such as cytotoxic T lymphocytes and NK cells, and it has both cytolytic and proinflammatory activity.[57] Indeed, the expression of *GNLY* in lymphocytes has been reported to be associated with recovery in COVID-19 suggesting it may play a major role in clearance of infected cells and termination of infection.[58] In agreement with the correlation between *GNLY* and viral load, we also showed that cell proportion estimates of NK cells in peripheral blood were negatively correlated with viral load z-score, highlighting the importance of this cell population in constraining the virus. For example, Witkowski et al. showed COVID-19 patients with normal NK cell numbers demonstrated a more rapid decline of viral load compared to those with low NK cell numbers.[59] Finally, the link between the blood and nasal transcriptome results can be also established through the presence of interferon-mediated processes. An example of this is the overlap we observed between the enriched pathways in the *IFNA14* module, and the genes positively correlated with viral load in the nasal samples, especially in the context of pathogen-induced cytokine storm signalling.

Our study was limited by the relatively small sample size which may have reduced the statistical power and resulted in missed opportunities to capture some biological signals. Additionally, we cannot establish from this data whether the molecular mechanisms identified are cause or consequence of the viral load, although there are plausible mechanisms which suggest causal roles in some cases. The relevance of these mechanisms for emerging variants of SARS-CoV-2, in individuals with vaccination-induced or naturally-acquired immunity, will require further investigation.

Our study also has several methodological limitations. We do not have any way to validate consistency of sampling of nasal epithelial transcriptomes, meaning there may be variation in sample composition. We deduced viral load from the amount of viral RNA, rather than the count of virions or culturable virus plaque forming units. We inferred the variation in viral load (z-score) using a mathematical model based on data from studies employing different designs and participant characteristics. Lastly, the method we employed for blood cell deconvolution does not include minor cell populations such as eosinophils, which may still be important in COVID-19.[60]

## Conclusions

To our knowledge, this is the most comprehensive study focusing on identifying molecular correlates of the SARS-CoV-2 viral load control in the URT. We identified numerous molecular processes which may contribute to the control of URT viral load. These candidate mechanisms can be the focus of further functional studies and may lead to new strategies to prevent COVID-19 and reduce SARS-CoV-2 transmission.

## Methods

### *Study design and participants*

We studied 82 COVID-19 patients recruited through the GEN-COVID study (www.gencovid.eu), a multi-centre and prospective cohort designed to evaluate the effect of genetic factors on SARS-CoV-2 infection. Subjects were recruited at Hospital Clínico Universitario de Santiago de Compostela (Galicia, Spain) between March 2020 and May 2020, during the first wave of infections in Spain, before significant levels of infection- and vaccine-induced immunity in the community. COVID-19 was defined according to the Spanish national guidelines (https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos.htm).
The severity of the disease was defined as mild, moderate, and severe based on WHO scoring for COVID-19 patients and as described previously.[15,61] Symptom onset data were acquired by either using patient questionnaires or, in cases where this was not feasible (the most severe cases), by obtaining this information from a family member. We also included 18 uninfected controls and 9 subjects with non-COVID-19 infections recruited through the PERFORM Consortium.

### *Sample collection*

Blood samples and nasal epithelium specimens were collected at the same time at hospital for moderate and severe COVID-19 subjects and at home for subjects with mild disease. Whole blood was collected into PAXgene blood RNA tubes (PreAnalytiX) and nasal epithelium samples were collected in Oragene CP-190 kit (DNA Genotek). Nasal swabs were taken by trained healthcare staff following a consistent standard operating procedure. For a given sample, the flocked swab was inserted into the nasal passage until a slight resistance was met. The flocked swab was then rotated slightly while brushing the mucosa for 5 s to ensure maximum absorbency and cell recovery. Samples were processed as described previously.[15,16]

One COVID-19 subject contributed two paired sets of samples (viral load and blood RNA-Seq; Table 1) collected 3 days apart. We included both as they showed a noticeable difference in viral load z-score and hence were informative.

### *RNA isolation*

Total RNA was isolated from blood and nasal epithelium samples using PAXgene blood miRNA extraction and RNeasy microkit, respectively, according to the manufacturer's protocols (Qiagen). RNA amount and integrity were assessed using TapeStation 4200 (Agilent). RNA quality was checked based on DV200 metric to ensure that sufficient percentage (over 50%) of RNA fragments were greater than 200 nucleotides in length and also to estimate the optimal sample input for the *nCounter* NanoString analysis.

### *Viral load measurements*

#### *Viral load quantification*

Nasopharyngeal samples were collected in Universal Transport Medium (UTM) tubes supplied by COPAN (www.copangroup.com) and assessed for the presence and viral load of SARS-CoV-2. We detected viral particles using a multiplex real-time PCR with the Allplex™SARS-CoV-2 Assay (Seegene). Viral load values (viral copies per ml) were computed from the Ct values as described previously.[3]

#### *Calculation of viral load z-scores*

A regression model of the average trajectory of viral load over time and quantification of variation between individuals, using data

from 16 datasets, was reported previously.[3] Viral load values from the present study were compared to this previously derived regression line to assess whether a particular viral load measurement, sampled a certain number of days after symptom onset, was higher or lower than average. A 'z-score' was calculated for each data point by calculating its deviation from the mean trajectory and dividing by the standard deviation of the variation in viral load around the mean trajectory (Fig. 2D).

*RNA sequencing*

Paired-end sequencing was performed at The Wellcome Centre for Human Genetics in Oxford, UK as described previously.[15] Sequencing was carried out using Novaseq6000 platform providing 150 bp paired end reads.

*RNA-Seq upstream analyses*

Adapter trimming and quality control of sequencing reads were performed with Trimmomatic version 0.36 and FastQC version 0.11.7, respectively.[62,63] The reads were then mapped against hg38 reference genome using STAR version 2.7.1a.[64] RSEM version 1.3.1 was used for transcript quantification.[65] Next, we performed a gene signature-based deconvolution using CellCODE as in our previous work and adjusted gene expression for leukocyte (B cells, monocytes, neutrophils, NK cells, $CD4^+$ T cells, and $CD8^+$ T cells) mixture.[12,18,19]

*NanoString experiment*

*NanoString nCounter assay*

We analysed immunological gene expression profiles of nasal epithelium using the *SPRINT nCounter* system (NanoString Technologies) with the Human Immunology V2 Panel (579 genes covering the core pathways and processes of the immune response, and 15 internal reference genes for data normalisation). The detail of the assay is described previously.[16]

*Differential gene expression analysis*

The gene expression counts adjusted for leukocyte mixture were correlated with viral load z-scores using *glmQLFit* and *glmQLFTest* functions in edgeR package.[66]

*Weighted correlation network analysis*

Gene counts were normalised using variance stabilising transformation (*VST*) function of DESeq2 R package[67] and adjusted for leukocyte mixture using *removeBatchEffect* function of limma R package.[68] We used WGCNA version 1.71 R package for weighted correlation network analysis.[20]

*Module repertoire analysis*

We applied *BloodGen3Module* version 1.4.0 R package[22] to the normalised gene expression counts unadjusted for cell-mixture from 16 samples with paired viral load, collected in the absence of bacterial co-infection. The package encompasses 382 functionally annotated blood transcriptional modules which have been grouped into 38 "aggregates" (A1-A38). The differential expression of the modules was compared between two groups with positive and negative viral load z-scores ($n = 5$ and $n = 11$, respectively) using t-test with fold change and p-value cut-off of 0.5 and 0.05, respectively. For each module, we computed 'module response' as the percentage of genes for the module showing significant differential expression between the two groups.

*Ingenuity pathway analysis*

We employed Qiagen's Ingenuity Pathway Analysis (IPA) to gain biological insights into the modules, including the identification of enriched pathways and potential regulators. For example, the regulators were predicted based on their expected causal effects on module genes, with these expectations drawn from the literature compiled within the IPA database. The analysis involved a thorough examination of the known targets associated with each regulator in our dataset. The inferred directions of changes in the module genes, as determined by their correlation coefficients with viral load z-score, were compared to the expectations derived from the literature, and the activity patterns (activation or inhibition) of the regulators were issued accordingly.

*Further statistical analysis*

The normality of distributions was assessed using the Shapiro-Wilk normality test. Pearson correlation was used to analyse the degree of association between two continuous variables. An independent-samples t-test and one-way ANOVA with Tukey's post hoc test were used to compare continuous variables between two and multiple groups, respectively.

## Funding

This work was supported by UKRI (MRC) and the DHSC (NIHR) (Grant Ref: MR/V027409/1). This study also received support from the European Union's Horizon 2020 program under GA No 668303 (PERFORM); Instituto de Salud Carlos III ([ISCIII] TRINEO: PI22/00162; DIAVIR: DTS19/00049; Resvi-Omics: PI19/01039 [AS]; ReSVinext: PI16/01569 [F.M.-T.]; Enterogen: PI19/01090 [F.M.-T.]; OMI-COVI-VAC (PI22/00406 [F.M.-T.] cofinanciados FEDER), GAIN: Grupos con Potencial de Crecimiento (IN607B 2020/08, [A.S.]); ACIS: BI-BACVIR (PRIS-3, [A.S.]), and CovidPhy (SA 304C, [A.S.]); and consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CB21/06/00103; F.M.-T.); GEN-COVID (IN845D 2020/23, F.M.-T.) and Grupos de Referencia Competitiva (IIN607A2021/05, F.M.-T.). The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

M.M.M. is supported in part by the NIHR Biomedical Research Centre of Imperial College NHS Trust.

J.D.C. and L.C.O. acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union.

F.L. is supported by an MRC clinical training fellowship [award MR/W000970/1]. F.L. and R.S.T. are supported by the UK Coronavirus Immunology Consortium (UKCIC).

H.R.J. received support from the Wellcome Trust (4-year PhD programme, grant number 215214/Z/19/Z).

M.K. acknowledges support from the Wellcome Trust and the Medical Research Foundation Grants (206508/Z/17/Z and MRF-160-0008-ELP-KAFO-C0801).

L.C.O. declares grant funding from Merck Group on an unrelated project.

## Ethics approval and consent to participate

The GEN-COVID and PERFORM (Personalised Risk assessment in Febrile illness to Optimise Real-life Management across the European Union; perform2020.org/) studies were conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of Galicia (CEIC, ref 2020/178, 18/03/2020) and St Mary's Research Ethics Committee (16/LO/1684, 25/02/2013), respectively. A written informed consent was obtained for each participant. If this was not done at the time of sampling, a retrospective consent was sought at the earliest appropriate opportunity.

Here, we have replaced sample and subject IDs with identifiers that cannot reveal the identity of the study subjects.

## Consent for publication

Not applicable.

## CRediT authorship contribution statement

Conceptualization: A.J.C, J.D.C., L.C.O., R.S.T. Data curation: M.M.M., I.R.-C., A.G.-C., D.H.-C., H.R.J., C.Y.F., Y.W., F.L. Formal analysis: M.M.M., J.D.C, A.S., A.G.-C., M.K. Funding acquisition: A.J.C, J.D.C, L.C.O., R.S.T., M.L., M.K., F.M.-T., I.R.-C., A.G.-C., A.S. Investigation: I.R.-C., A.G.-C., V.J.W., S.N., G.D'S., D.H.-C. Methodology: M.M.M., J.D.C., A.G.-C., A.S., H.R.J., V.J.W., D.H.-C. Project administration: A.J.C, A.S., A.G.-C., F.M.-T. Resources: A.J.C., F.M.-T., I.R.-C., A.G.-C., M.K. Software: M.M.M, H.R.J., D.H.-C. Supervision: A.J.C, L.C.O., R.S.T., M.K., V.J.W., M.L., F.M.-T. Validation: M.M.M. Visualisation: M.M.M. Writing – original draft: M.M.M., A.J.C. Writing – review & editing: ALL.

## Data availability

Raw RNA-Seq data and corresponding metadata are available on ArrayExpress under the accession E-MTAB-12791. NanoString nCounter data and corresponding metadata are available at https://github.com/MahdiMoradiMarjaneh/COVID19_viral_load. Codes used in the analyses can be accessed on the same GitHub repository.

## Declaration of Competing Interest

The authors declare that they have no competing interests.

## Acknowledgements

The members and affiliations of the PERFORM consortium and the GEN-COVID (www.gencovid.eu) study group are listed in the Supplementary text file.

Figure 1 and the visual abstract figure were created with "BioRender.com".

For the purpose of open access, the authors have applied a 'Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jinf.2023.10.009.

## References

1. Cornelissen L, Andre E. *Understanding the drivers of transmission of SARS-CoV-2.* Lancet Infect Dis 2021;**21**(5):580–1.
2. Marc A, Kerioui M, Blanquart F, Bertrand J, Mitja O, Corbacho-Monne M, et al. *Quantifying the relationship between SARS-CoV-2 viral load and infectiousness.* Elife 2021;**10**:e69302.
3. Challenger JD, Foo CY, Wu Y, Yan AWC, Marjaneh MM, Liew F, et al. *Modelling upper respiratory viral load dynamics of SARS-CoV-2.* BMC Med 2022;**20**(1):25.
4. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, et al. *Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C mortality score.* BMJ 2020;**370**:m3339.
5. Chen PZ, Bobrovitz N, Premji ZA, Koopmans M, Fisman DN, Gu FX. *SARS-CoV-2 shedding dynamics across the respiratory tract, sex, and disease severity for adult and pediatric COVID-19.* Elife 2021;**10**:e70458.
6. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. *SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis.* Lancet Microbe 2021;**2**(1):e13–22.
7. Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, et al. *Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility.* N Engl J Med 2020;**382**(22):2081–90.
8. Puhach O, Meyer B, Eckerle I. *SARS-CoV-2 viral load and shedding kinetics.* Nat Rev Microbiol 2023;**21**(3):147–61.
9. Diamond MS, Kanneganti TD. *Innate immunity: the first line of defense against SARS-CoV-2.* Nat Immunol 2022;**23**(2):165–76.
10. Blanco-Melo D, Nilsson-Payant BE, Liu WC, Uhl S, Hoagland D, Moller R, et al. *Imbalanced host response to SARS-CoV-2 drives development of COVID-19.* Cell 2020;**181**(5):1036–1045.e9.
11. McClain MT, Constantine FJ, Henao R, Liu Y, Tsalik EL, Burke TW, et al. *Dysregulated transcriptional responses to SARS-CoV-2 in the periphery.* Nat Commun 2021;**12**(1):1079.
12. Georgiadou A, Lee HJ, Walther M, van Beek AE, Fitriani F, Wouters D, et al. *Modelling pathogen load dynamics to elucidate mechanistic determinants of host-Plasmodium falciparum interactions.* Nat Microbiol 2019;**4**(9):1592–602.
13. Nain Z, Barman SK, Sheam MM, Syed SB, Samad A, Quinn JMW, et al. *Transcriptomic studies revealed pathophysiological impact of COVID-19 to predominant health conditions.* Brief Bioinform 2021;**22**(6):bbab197.
14. Islam A, Khan MA, Ahmed R, Hossain MS, Kabir SMT, Islam MS, et al. *Transcriptome of nasopharyngeal samples from COVID-19 patients and a comparative analysis with other SARS-CoV-2 infection models reveal disparate host responses against SARS-CoV-2.* J Transl Med 2021;**19**(1):32.
15. Jackson H, Rivero Calle I, Broderick C, Habgood-Coote D, D'Souza G, Nichols S, et al. *Characterisation of the blood RNA host response underpinning severity in COVID-19 patients.* Sci Rep 2022;**12**(1):12216.
16. Gomez-Carballa A, Rivero-Calle I, Pardo-Seco J, Gomez-Rial J, Rivero-Velasco C, Rodriguez-Nunez N, et al. *A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity.* Environ Res 2022;**210**:112890.
17. Singanayagam A, Patel M, Charlett A, Lopez Bernal J, Saliba V, Ellis J, et al. *Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020.* Eur Surveill 2020;**25**(32):2001483.
18. Chikina M, Zaslavsky E, Sealfon SC. *CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations.* Bioinformatics 2015;**31**(10):1584–91.
19. Lee HJ, Georgiadou A, Walther M, Nwakanma D, Stewart LB, Levin M, et al. *Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria.* Sci Transl Med 2018;**10**(447):eaar3619.
20. Langfelder P, Horvath S. *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinforma 2008;**9**:559.
21. Kramer A, Green J, Pollard Jr. J, Tugendreich S. *Causal analysis approaches in ingenuity pathway analysis.* Bioinformatics 2014;**30**(4):523–30.
22. Altman MC, Rinchai D, Baldwin N, Toufiq M, Whalen E, Garand M, et al. *Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data.* Nat Commun 2021;**12**(1):4385.
23. Kozlov M. *Could a nose spray a day keep COVID away?* Nature 2022. https://doi.org/10.1038/d41586-022-03341-z
24. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. *From molecular to modular cell biology.* Nature 1999;**402**(Suppl.6761):C47–52.
25. Combes AJ, Courau T, Kuhn NF, Hu KH, Ray A, Chen WS, et al. *Global absence and targeting of protective immune states in severe COVID-19.* Nature 2021;**591**(7848):124–30.
26. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. *Genetic mechanisms of critical illness in COVID-19.* Nature 2021;**591**(7848):92–8.
27. Wang EY, Mao T, Klein J, Dai Y, Huck JD, Jaycox JR, et al. *Diverse functional autoantibodies in patients with COVID-19.* Nature 2021;**595**(7866):283–8.
28. Broggi A, Ghosh S, Sposito B, Spreafico R, Balzarini F, Lo Cascio A, et al. *Type III interferons disrupt the lung epithelial barrier upon viral recognition.* Science 2020;**369**(6504):706–12.
29. Lucas C, Wong P, Klein J, Castro TBR, Silva J, Sundaram M, et al. *Longitudinal analyses reveal immunological misfiring in severe COVID-19.* Nature 2020;**584**(7821):463–9.
30. Major J, Crotta S, Llorian M, McCabe TM, Gad HH, Priestnall SL, et al. *Type I and III interferons disrupt lung epithelial repair during recovery from viral infection.* Science 2020;**369**(6504):712–7.
31. Sposito B, Broggi A, Pandolfi L, Crotta S, Clementi N, Ferrarese R, et al. *The interferon landscape along the respiratory tract impacts the severity of COVID-19.* Cell 2021;**184**(19):4953–4968.e16.
32. Ziegler CGK, Miao VN, Owings AH, Navia AW, Tang Y, Bromley JD, et al. *Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19.* Cell 2021;**184**(18):4713–4733.e22.
33. Shim JM, Kim J, Tenson T, Min JY, Kainov DE. *Influenza virus infection, interferon response, viral counter-response, and apoptosis.* Viruses 2017;**9**(8):223.
34. Sun J, Wang J, Yuan X, Wu X, Sui T, Wu A, et al. *Regulation of Early Host Immune Responses Shapes the Pathogenicity of Avian Influenza A Virus.* Front Microbiol 2019;**10**:2007.
35. Hoffmann HH, Schneider WM, Rice CM. *Interferons and viruses: an evolutionary arms race of molecular interactions.* Trends Immunol 2015;**36**(3):124–38.
36. Cheemarla NR, Watkins TA, Mihaylova VT, Wang B, Zhao D, Wang G, et al. *Dynamic innate immune response determines susceptibility to SARS-CoV-2 infection and early replication kinetics.* J Exp Med 2021;**218**(8):e20210583.
37. Schuhenn J, Meister TL, Todt D, Bracht T, Schork K, Billaud JN, et al. *Differential interferon-alpha subtype induced immune signatures are associated with suppression of SARS-CoV-2 infection.* Proc Natl Acad Sci U S A 2022;**119**(8):e2111600119.

38. Zhou Q, Chen V, Shannon CP, Wei XS, Xiang X, Wang X, et al. *Interferon-alpha2b Treatment for COVID-19.* Front Immunol 2020;**11**:1061.
39. Ball K. Assessing the efficacy of non-canonical IFNA-subtypes as inhibitors of SARS-CoV-2 replication: Medical Research Scotland; 2020. Available from: ⟨https://www.medicalresearchscotland.org.uk/awards/assessing-the-efficacy-of-non-canonical-ifna-subtypes-as-inhibitors-of-sars-cov-2-replication/⟩.
40. Ragab D, Salah Eldin H, Taeimah M, Khattab R, Salem R. *The COVID-19 Cytokine Storm; what we know so far.* Front Immunol 2020;**11**:1446.
41. Ye Q, Wang B, Mao J. *The pathogenesis and treatment of the `Cytokine Storm' in COVID-19.* J Infect 2020;**80**(6):607–13.
42. Ahn D, Prince A. *Participation of the IL-10RB related cytokines, IL-22 and IFN-lambda in defense of the airway mucosal barrier.* Front Cell Infect Microbiol 2020;**10**:300.
43. Albayrak N, Orte Cano C, Karimi S, Dogahe D, Van Praet A, Godefroid A, et al. *Distinct expression patterns of interleukin-22 receptor 1 on blood hematopoietic cells in SARS-CoV-2 infection.* Front Immunol 2022;**13**:769839.
44. Klooster JPT, Bol-Schoenmakers M, van Summeren K, van Vliet ALW, de Haan CAM, van Kuppeveld FJM, et al. *Enterocytes, fibroblasts and myeloid cells synergize in anti-bacterial and anti-viral pathways with IL22 as the central cytokine.* Commun Biol 2021;**4**(1):631.
45. Han SN, Meydani SN. *Impact of vitamin E on immune function and its clinical implications.* Expert Rev Clin Immunol 2006;**2**(4):561–7.
46. Lewis ED, Meydani SN, Wu D. *Regulatory role of vitamin E in the immune system and inflammation.* IUBMB Life 2019;**71**(4):487–94.
47. Pacl HT, Tipper JL, Sevalkar RR, Crouse A, Crowder C, UAB Precision Medicine Institute, et al. *Water-soluble tocopherol derivatives inhibit SARS-CoV-2 RNA-dependent RNA polymerase.* bioRxiv 2021. https://doi.org/10.1101/2021.07.13.449251
48. Pervolaraki K, Rastgou Talemi S, Albrecht D, Bormann F, Bamford C, Mendoza JL, et al. *Differential induction of interferon stimulated genes between type I and type III interferons is independent of interferon receptor abundance.* PLoS Pathog 2018;**14**(11):e1007420.
49. Peterson ST, Kennedy EA, Brigleb PH, Taylor GM, Urbanek K, Bricker TL, et al. *Disruption of type III interferon (IFN) genes Ifnl2 and Ifnl3 recapitulates loss of the type III IFN receptor in the mucosal antiviral response.* J Virol 2019;**93**(22). https://doi.org/10.1128/jvi.01073-19
50. Bezemer GFG, Garssen J. *TLR9 and COVID-19: a multidisciplinary theory of a multifaceted therapeutic target.* Front Pharm 2020;**11**:601685.
51. Schmid T, Brune B. *Prostanoids and resolution of inflammation - beyond the lipid-mediator class switch.* Front Immunol 2021;**12**:714042.
52. Wong LR, Zheng J, Wilhelmsen K, Li K, Ortiz ME, Schnicker NJ, et al. *Eicosanoid signalling blockade protects middle-aged mice from severe COVID-19.* Nature 2022;**605**(7908):146–51.
53. Patel SR, Richardson JL, Schulze H, Kahle E, Galjart N, Drabek K, et al. *Differential roles of microtubule assembly and sliding in proplatelet formation by megakaryocytes.* Blood 2005;**106**(13):4076–85.
54. Ercan H, Schrottmaier WC, Pirabe A, Schmuckenschlager A, Pereyra D, Santol J, et al. *Platelet phenotype analysis of COVID-19 patients reveals progressive changes in the activation of integrin alphaIIbbeta3, F13A1, the SARS-CoV-2 target EIF4A1 and annexin A5.* Front Cardiovasc Med 2021;**8**:779073.
55. Yu J, Peterson DR, Baran AM, Bhattacharya S, Wylie TN, Falsey AR, et al. *Host gene expression in nose and blood for the diagnosis of viral respiratory infection.* J Infect Dis 2019;**219**(7):1151–61.
56. Rajagopala SV, Strickland BA, Pakala SB, Kimura KS, Shilts MH, Rosas-Salazar C, et al. *Mucosal gene expression in response to SARS-CoV-2 is associated with viral load.* J Virol 2023;**97**(2):e0147822.
57. Dotiwala F, Lieberman J. *Granulysin: killer lymphocyte safeguard against microbes.* Curr Opin Immunol 2019;**60**:19–29.
58. Zhang JY, Wang XM, Xing X, Xu Z, Zhang C, Song JW, et al. *Single-cell landscape of immunological responses in patients with COVID-19.* Nat Immunol 2020;**21**(9):1107–18.
59. Witkowski M, Tizian C, Ferreira-Gomes M, Niemeyer D, Jones TC, Heinrich F, et al. *Untimely TGFbeta responses in COVID-19 limit antiviral functions of NK cells.* Nature 2021;**600**(7888):295–301.
60. Eijmael M, Janssens N, le Cessie S, van Dooren Y, Koster T, Karim F. *Coronavirus disease 2019 and peripheral blood eosinophil counts: a retrospective study.* Infection 2021;**49**(6):1325–9.
61. WHO. Clinical management of COVID-19: interim guidance, 27 May 2020; 2020.
62. Bolger AM, Lohse M, Usadel B. *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics 2014;**30**(15):2114–20.
63. Andrews S. FastQC: a quality control tool for high throughput sequence data; 2010. Available from: ⟨http://www.bioinformatics.babraham.ac.uk/projects/fastqc/⟩.
64. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics 2013;**29**(1):15–21.
65. Li B, Dewey CN. *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinform 2011;**12**:323.
66. Robinson MD, McCarthy DJ, Smyth GK. *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics 2010;**26**(1):139–40.
67. Love MI, Huber W, Anders S. *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol 2014;**15**(12):550.
68. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res 2015;**43**(7):e47.