

















































OPEN

Single-cell multi-omics analysis of the immune response in COVID-19

Emily Stephenson^{1,43}, Gary Reynolds^{1,43}, Rachel A. Botting^{1,43}, Fernando J. Calero-Nieto ^{2,43}, Michael D. Morgan ^{3,4,43}, Zewen Kelvin Tuong ^{5,6,43}, Karsten Bach^{3,4,43}, Waradon Sunngak ^{6,43}, Kaylee B. Worlock⁷, Masahiro Yoshida ⁷, Natsuhiko Kumasaka ⁶, Katarzyna Kania⁴, Justin Engelbert¹, Bayanne Olabi¹, Jarmila Stremenova Spegarova ⁸, Nicola K. Wilson ², Nicole Mende ², Laura Jardine¹, Louis C. S. Gardner ¹, Issac Goh ¹, Dave Horsfall ¹, Jim McGrath¹, Simone Webb ¹, Michael W. Mather ¹, Rik G. H. Lindeboom ⁶, Emma Dann ⁶, Ni Huang⁶, Krzysztof Polanski ⁶, Elena Prigmore⁶, Florian Gothe^{8,24}, Jonathan Scott⁸, Rebecca P. Payne ⁸, Kenneth F. Baker ^{8,9}, Aidan T. Hanrath^{8,10}, Ina C. D. Schim van der Loeff ⁸, Andrew S. Barr ¹⁰, Amada Sanchez-Gonzalez¹⁰, Laura Bergamaschi^{11,12}, Federica Mescia ^{11,12}, Josephine L. Barnes ⁷, Eliz Kilich¹³, Angus de Wilton¹³, Anita Saigal¹⁴, Aarash Saleh¹⁴, Sam M. Janes ^{7,13}, Claire M. Smith ¹⁵, Nusayhah Gopee^{1,16}, Caroline Wilson^{1,17}, Paul Coupland ⁴, Jonathan M. Coxhead¹, Vladimir Yu Kiselev⁶, Stijn van Dongen⁶, Jaume Bacardit ¹⁸, Hamish W. King ^{6,19}, Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration*, Anthony J. Rostron ^{8,20}, A. John Simpson ⁸, Sophie Hambleton ⁸, Elisa Laurenti ², Paul A. Lyons ^{11,12}, Kerstin B. Meyer ^{6,44}, Marko Z. Nikolic ^{7,13,44}, Christopher J. A. Duncan^{8,10,44}, Kenneth G. C. Smith ^{11,12,44}, Sarah A. Teichmann ^{6,21,44} , Menna R. Clatworthy ^{5,6,12,22,23,44} , John C. Marioni ^{3,4,6,44} , Berthold Göttgens ^{2,44}  and Muzlifah Haniffa ^{1,6,9,16,44} 

Analysis of human blood immune cells provides insights into the coordinated response to viral infections such as severe acute respiratory syndrome coronavirus 2, which causes coronavirus disease 2019 (COVID-19). We performed single-cell transcriptome, surface proteome and T and B lymphocyte antigen receptor analyses of over 780,000 peripheral blood mononuclear cells from a cross-sectional cohort of 130 patients with varying severities of COVID-19. We identified expansion of nonclassical monocytes expressing complement transcripts (*CD16⁺C1QA/B/C⁺*) that sequester platelets and were predicted to replenish the alveolar macrophage pool in COVID-19. Early, uncommitted CD34⁺ hematopoietic stem/progenitor cells were primed toward megakaryopoiesis, accompanied by expanded megakaryocyte-committed progenitors and increased platelet activation. Clonally expanded CD8⁺ T cells and an increased ratio of CD8⁺ effector T cells to effector memory T cells characterized severe disease, while circulating follicular helper T cells accompanied mild disease. We observed a relative loss of IgA2 in symptomatic disease despite an overall expansion of plasmablasts and plasma cells. Our study highlights the coordinated immune response that contributes to COVID-19 pathogenesis and reveals discrete cellular components that can be targeted for therapy.

The COVID-19 global pandemic has caused >120 million infections and 2.6 million deaths (as of 17 March 2021)^{1,2}. Symptoms vary in severity and include acute respiratory distress syndrome, thrombosis and organ failure³. COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a single-stranded RNA betacoronavirus that enters host cells through receptors such as angiotensin-converting enzyme 2 (ACE2) and neuropilin (NRP1), which are expressed widely, including in nasal epithelium^{4–6}.

Several studies have highlighted a complex network of peripheral blood immune responses in COVID-19 infection^{7,8}. A reduction in

T cells with disease severity and reduced interferon (IFN)- γ production by lymphocytes have been reported⁹. However, an expansion of highly cytotoxic effector T cell subsets in moderate to severe disease^{10,11} and higher expression of exhaustion markers programmed cell death protein 1 and Tim-3 on CD8⁺ T cells have been described in patients receiving intensive care therapy⁵. In severe cases, classical monocytes have been shown to display a type 1 IFN inflammatory signature⁶; however, low levels of IFN α coupled with a reduction in plasmacytoid dendritic cells (DCs) have been reported in patients with critical disease¹². Emergency and dysregulated myelopoiesis, and expanded activated megakaryocytes have also

A full list of author affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper.

been reported^{13–15}. Proliferating plasmablasts and extrafollicular B cell activation are present in critically ill patients, but despite high levels of SARS-CoV-2-specific antibodies and antibody-secreting cells, many of these patients do not recover^{8,15}. To better understand the coordinated systemic immune response in individuals with asymptomatic and symptomatic COVID-19, we performed combined single-cell transcriptome, cell-surface protein and lymphocyte antigen receptor repertoire analysis of peripheral blood of a cross-sectional patient cohort and integrated results across three UK medical centers.

Results

Altered cellular profiles across COVID-19 severities. We generated single-cell combined transcriptome and surface proteome data from peripheral blood mononuclear cells (PBMCs) from individuals with asymptomatic, mild, moderate, severe and critical¹⁶ COVID-19 across three UK centers in Newcastle, Cambridge and London (Fig. 1a, Extended Data Fig. 1a and Supplementary Table 1). Controls included healthy volunteers, individuals with non-COVID-19 severe respiratory illness and healthy volunteers administered with intravenous lipopolysaccharide (IV-LPS) as a surrogate for an acute systemic inflammatory response (Fig. 1a and Supplementary Table 2). We sequenced 1,141,860 cells from 143 samples. Following computational doublet removal, 781,123 cells passed quality control (QC; minimum of 200 genes and <10% mitochondrial reads per cell; Extended Data Fig. 1b). Data were integrated using Harmony¹⁷ with good mixing of cells by the kBET statistic calculated for each cluster across samples (rejection rate improved from 0.62 to 0.36 following integration, $P < 2.1 \times 10^{-8}$ by Wilcoxon paired signed-rank test; Extended Data Fig. 1c,d).

Following Leiden clustering, cells were manually annotated based on the RNA expression of known marker genes supported by surface protein expression of markers employed in flow cytometry to discriminate subpopulations (Extended Data Fig. 1e). We defined 18 cell subsets (Fig. 1b), with an additional 27 cell states identified following subclustering (Figs. 1b, 2a, 3a,b and 4a,b and Supplementary Table 3). Our annotation was further validated using Azimuth, whereby more than 50% of cells were mapped and matched to a unique cluster in 32/33 of the clusters defined in the Azimuth PBMC dataset (Methods; proliferating CD8⁺ T cells mapped across two clusters). Clusters unique to our data included proliferating monocytes, innate lymphoid cell subpopulations and isotype-specific plasma cells (Extended Data Fig. 1f).

We observed a relative expansion of proliferating lymphocytes, proliferating monocytes, platelets and mobilized hematopoietic stem and progenitor cells (HSPCs) with worsening disease. Plasmablasts and B cells were expanded in severe and critical disease (Fig. 1c and Extended Data Fig. 2a). These changes matched trends in clinical blood cell counts (Extended Data Fig. 2b and Supplementary Table 4). To assess the broader impacts of patient characteristics and clinical metadata on the altered proportion of cell types/states, we used a Poisson linear mixed model (Methods), which predicted the COVID-19 swab result (Bonferroni-corrected logistic regression (BF-corrected LR), $P = 1.1 \times 10^{-3}$; Methods), disease severity at blood sampling (BF-corrected LR, $P = 8.9 \times 10^{-8}$) and center (contributed by increased red blood cells (RBCs) and reduced monocytes in the Cambridge patient cohort; (BF-corrected LR, $P = 2.0 \times 10^{-142}$) as the main contributing factors among seven different clinical/technical factors (Extended Data Fig. 2c). PBMC composition varied depending on symptom duration, with increased relative frequency of plasmacytoid dendritic cells (pDCs), natural killer (NK) cells, CD14⁺ and CD16⁺ monocytes (false discovery rate (FDR), 10%) and decreased relative frequencies of B cells, regulatory T (T_{reg}) cells, RBCs, platelets and CD4⁺ T cells, with a longer symptomatic interval before hospital admission (Extended Data Fig. 2d). These changes may be due to a subset

of individuals in the critically ill category who reported a longer time since symptom onset, consistent with a protracted course of infection in critical disease (Extended Data Fig. 2e,f). However, concordant changes in immune cell composition were observed when excluding patients with either the longest symptom durations (>24 d) or critical disease (Extended Data Fig. 2g), indicating that disease severity changes were not driven by symptom duration. Cell abundance results were also in agreement when performing a leave-one-out analysis (Extended Data Fig. 2h).

We observed expression of type I/III interferon response genes in monocytes, DCs and HSPCs across the spectrum of COVID-19 severity, but not in individuals challenged with IV-LPS, in keeping with the importance of type I and III interferons in innate immune responses to viral infection (Fig. 1d). Type I/III interferon response-related genes were recently implicated in genome-wide association studies (GWAS) for COVID-19 susceptibility^{18,19}. *IFNAR2* was both upregulated in individuals with COVID-19 compared to healthy controls in most circulating cell types and highly expressed by plasmablasts, monocytes and DCs (Extended Data Fig. 2i).

Multiplexed analysis of 45 proteins in serum showed two contrasting profiles between mild/moderate and severe/critical patients. CCL4, CXCL10, interleukin (IL)-7 and IL-1 α were associated with severe and critical disease, suggesting an augmented drive for monocyte and NK lymphocyte recruitment, as well as support for T cell activity/pathology (Extended Data Fig. 2j and Supplementary Table 5).

We used Cydar²⁰ to characterize the immune landscape changes with disease severity based on surface protein expression by dividing cells into phenotypic hyperspheres. We quantified the number of cells from each severity group within the hyperspheres, identifying 608 hyperspheres that differed significantly in abundance with increasing severity (spatial FDR < 0.05; Fig. 1e). Differentially abundant hyperspheres were present in all major immune compartments. Notably, we found an increase in B cells (CD19⁺/CD20⁺), plasma cells (CD38⁺) and HSPCs (CD34⁺), as well as remodeling of the myeloid compartment¹³ (Fig. 1e).

Mononuclear phagocytes and HSPC changes. Transcriptome and surface proteome analysis of blood mononuclear phagocytes (MPs) identified known DC subsets (pDC, ASDC (*AXL*⁺*SIGLEC6*⁺ DC), DC1, DC2 and DC3) and several monocyte states (Fig. 2a,b). Three CD14⁺ monocyte states were present (proliferating, classical CD14⁺ and activated CD83⁺) in addition to two CD16⁺ monocyte states (nonclassical CD16⁺ and C1QA/B/C⁺CD16⁺; Fig. 2a,b). Proliferating monocytes and DCs expressing *MKI67* and *TOP2A* were increased with disease severity (Fig. 2a,b). In contrast, the frequencies of DC2 and DC3 were reduced. Proliferating monocytes, previously reported by flow cytometry analysis of blood from patients with COVID-19¹⁴, transcriptionally resembled CD14⁺ monocytes and was the only population to change significantly with symptom duration. (Fig. 2a,b and Extended Data Fig. 3a). Proliferating DCs resembled the DC2 subset (Fig. 2a,b). Rare C1QA/B/C-expressing CD16⁺ monocytes were the only source of C1 complement components (Fig. 2b and Extended Data Fig. 3b).

We previously demonstrated egress of blood DCs and monocytes to the alveolar space with rapid acquisition of a lung molecular profile following human inhalational LPS challenge²¹. To investigate the relationship between circulating and lung alveolar MPs in COVID-19, we compared the transcriptome profile of blood DCs and monocytes with their bronchoalveolar lavage (BAL) counterparts using recently published data (*GSE145926*)²² (Extended Data Fig. 3d). Partition-based graph abstraction (PAGA) suggested transcriptional similarity between circulating CD14⁺ monocytes and BAL macrophages in health, aligning with recent data demonstrating that BAL macrophages can arise from circulating CD14⁺ monocytes

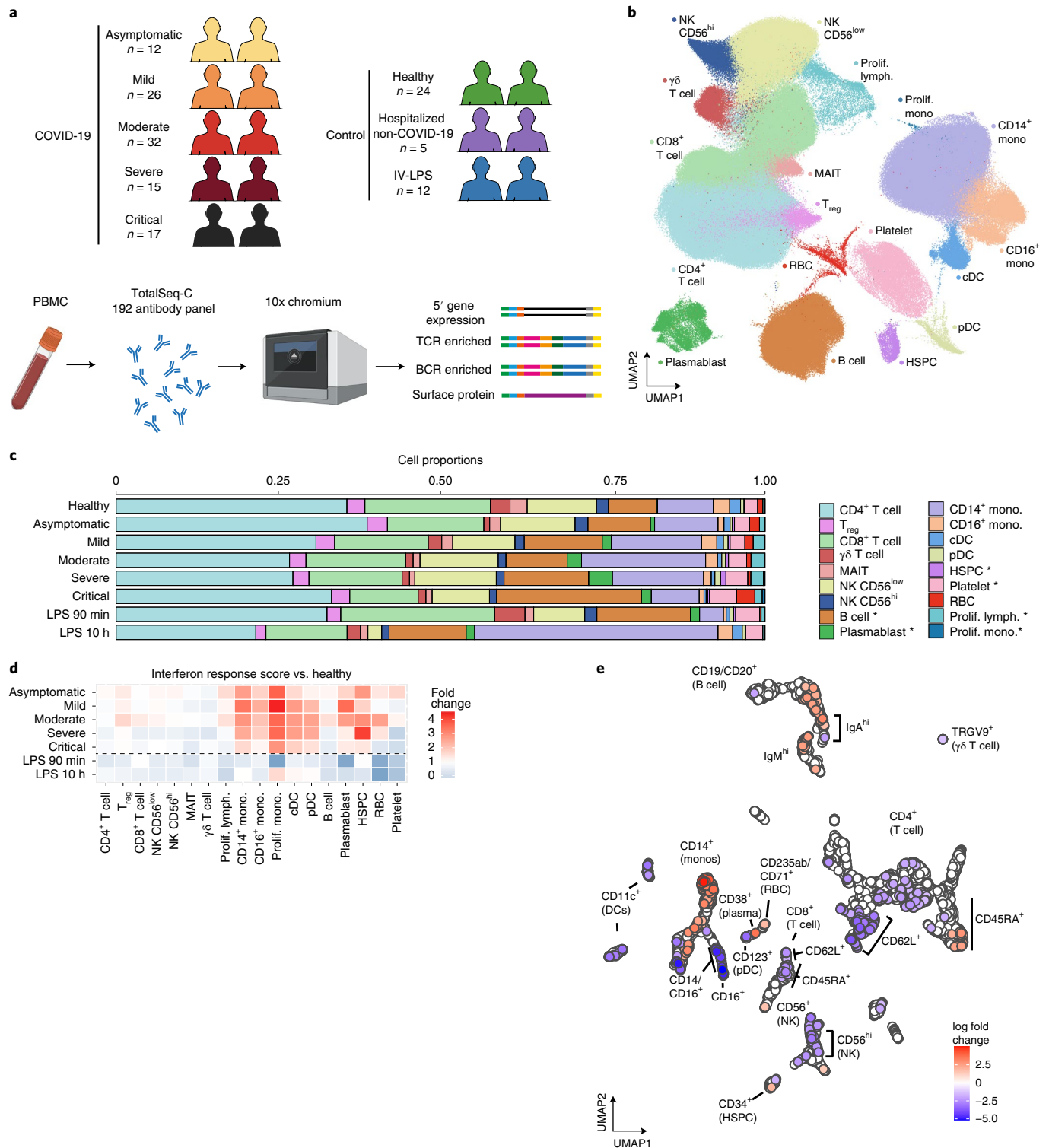


Fig. 1 | Single-cell multi-omics analysis of PBMCs from individuals with COVID-19 and controls. a, Overview of the participants included and the samples and data collected. Figure was created using [BioRender.com](https://www.biorender.com). **b**, UMAP visualization of all 781,123 cells after QC. Leiden clusters based on 5' gene expression shown and colored by cell type. Lymph, lymphocyte; mono, monocyte; prolif, proliferating. **c**, Bar plot of the proportion of cell types shown in **b**, separated by condition and COVID-19 severity status. Hypothesis testing was performed using quasi-likelihood *F*-test comparing healthy controls to individuals with COVID-19 for linear trends across disease severity groups (healthy > asymptomatic > mild > moderate > severe > critical). Differentially abundant cell types were determined using a 10% FDR and are marked with an asterisk. **d**, Enrichment of interferon response of each cell state separated by severity. IFN response was calculated using a published gene list ([GO:0034340](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GO:0034340)). **e**, UMAP computed using batch-corrected mean staining intensities of 188 antibodies for 4,241 hyperspheres. Each hypersphere represents an area in the 188-dimensional space and is colored by significant (spatial FDR < 0.05) severity-associated changes in abundance of cells within that space.

(Fig. 2c)²³. In COVID-19, there was greater transcriptional similarity between BAL macrophages and circulating C1QA/B/C⁺CD16⁺ monocytes (Fig. 2c), suggesting a differential origin of alveolar macrophages in healthy donors and individuals with COVID-19. Both BAL macrophages and C1QA/B/C⁺CD16⁺ monocytes express *FCGR3A* and *C1QA/B/C* and type I interferon response genes (Fig. 2a). Myeloid hyperinflammation causing lung and peripheral tissue damage via secretion of inflammatory cytokines such as IL-6 and tumor necrosis factor (TNF) in COVID-19 has been reported and in our analysis were primarily expressed by tissue rather than blood MPs (Fig. 2c). Genes differentially expressed by blood monocytes identified *S100A8*, previously reported in COVID-19 as contributing to the cytokine storm in severe infection²⁴. BAL macrophages expressed leukocyte-recruiting chemokines including *CCL2*, *CCL4*, *CCL7* and *CCL8* (Fig. 2d).

Tissue DCs respond to local inflammation and pathogen challenge by migrating to the draining lymph node to activate naïve T cells. BAL macrophages contain a population of mature, migratory DCs that express *CCR7* and *LAMP3* but downregulate DC-specific markers, such as *CD1C* and *CLEC9A* (Extended Data Fig. 3c). These migratory DCs express *IL10* in health, but *TNF* and the common IL-12 and IL-23 subunit *IL12B* in COVID-19, suggesting altered capacity for T cell polarization (Fig. 2e). In peripheral blood, C1QA/B/C⁺CD16⁺ monocytes expressed the highest amount of type I IFN response genes compared to all myeloid cells (Fig. 2f and Supplementary Tables 6 and 7). We detected minimal TNF-mediated or IL-6-mediated JAK–STAT signaling activation in circulating monocytes and DCs, but this was upregulated by COVID-19 BAL MPs (Fig. 2f, Supplementary Tables 6 and 7).

Coagulation abnormalities and monocyte–platelet aggregates have been previously reported in COVID-19 (refs. ^{25,26}), leading us to investigate predicted receptor–ligand interactions between monocytes and platelets using the CellPhoneDB repository. The expression levels of *SIRPA:CD47*, *FPR1:ANXA1*, *FPR2:APP* between monocytes:platelets were highest in the C1QA/B/C⁺CD16⁺ monocytes (Fig. 2g). Using protein data, we identified ICAM1 interactions on platelets with CD11a/b/c/CD18 primarily on C1QA/B/C⁺CD16⁺ monocytes and CD16⁺ monocytes (Extended Data Fig. 3d), accompanied by increased expression of surface proteins indicative of platelet activation (Fig. 2h).

Our large dataset of 781,123 PBMCs allowed us to interrogate 3,297 CD34⁺ HSPCs. Leiden clustering and uniform manifold and projection (UMAP) visualization showed a cloud-like representation, consistent with a stem/progenitor cell landscape previously described for bone marrow HSPCs²⁷ (Fig. 2i and Extended Data

Fig. 3e). Absence of CD38 mRNA and protein expression marks the most immature cells within the CD34⁺ compartment, while expression of markers such as *GATA1*, *MPO* and *PF4* characterizes distinct erythroid, myeloid and megakaryocytic progenitor populations, respectively (Fig. 2i). Accordingly, we were able to annotate six transcriptional clusters as CD34⁺CD38⁻ HSPCs, CD34⁺CD38⁺ early progenitor HSPCs and CD34⁺CD38⁺ erythroid, megakaryocytic and myeloid progenitors, as well as a small population distinguished by the expression of cell cycle (S phase) genes (Fig. 2i). Megakaryocyte progenitors were absent in healthy and asymptomatic individuals but made up ~5% of CD34⁺ cells in all symptomatic individuals (Fig. 2j). As peripheral blood is not a site for hematopoiesis²⁸, this finding likely reflects COVID-19-mediated perturbation of normal homeostatic functioning of the bone marrow HSPC compartment.

In light of our earlier observations of platelet activation and enhanced C1QA/B/C⁺CD16⁺ monocyte–platelet interactions (Fig. 2g,h), the appearance of CD34⁺ megakaryocyte progenitors was of particular interest, as it suggested a rebalancing of the stem/progenitor cell compartment. To further explore this hypothesis, we generated differential gene expression signatures between the megakaryocyte, myeloid and erythroid progenitor clusters to interrogate potential early activation or priming in the most immature HSPC clusters (Extended Data Fig. 3f). We observed enrichment of the megakaryocyte progenitor signature in the CD38⁺ HSPC populations in moderate COVID-19 compared to the healthy condition (Fig. 2k and Supplementary Table 8), but no enrichment of erythroid or myeloid signatures in either CD38⁻ or CD38⁺ HSPCs (Extended Data Fig. 3g and Supplementary Table 8). Our earlier observation of increased platelet activation within the context of normal platelet counts (Fig. 2h and Extended Data Fig. 2b) is thus consistent with a model whereby a rebalancing of the HSPC compartment toward megakaryopoiesis may be compensating for peripheral platelet consumption in COVID-19.

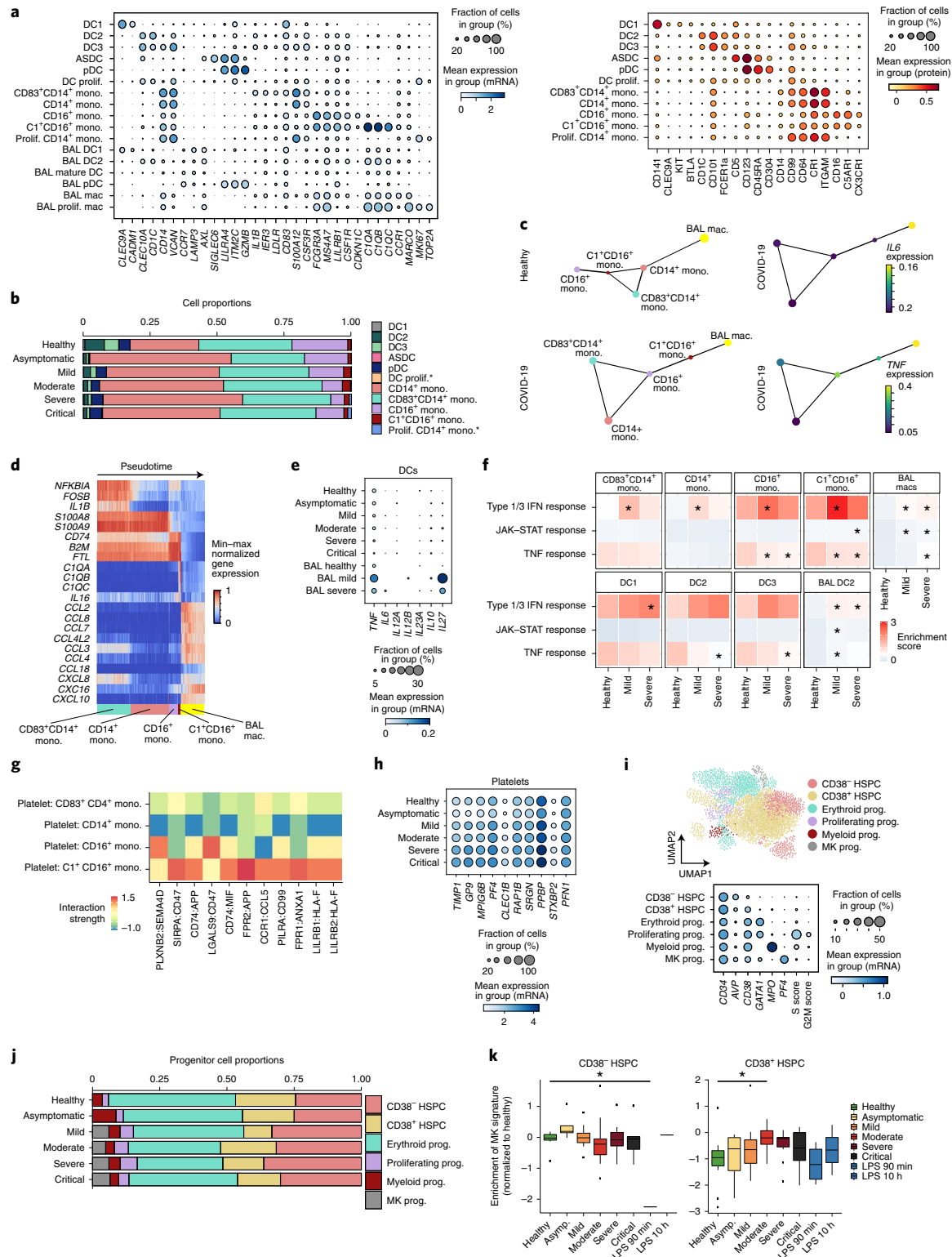
T lymphocytes and T cell receptor changes. Fine-resolution clustering of mRNA profiles revealed 11 initial clusters of CD4⁺ T cells, CD8⁺ T cells and innate-like T cells including $\gamma\delta$ T cells, NK T cells and mucosal-associated invariant T (MAIT) cells (Fig. 3a and Extended Data Fig. 4). Annotations were refined further using RNA expression of effector cytokines and surface protein expression (Fig. 3a–c).

Cellular composition of the T cell compartment varied between healthy and infected groups (Fig. 3d). Based on their relative proportions and differential abundance testing (FDR 10%), we found activated CD4⁺ T cells expressing *IL22*, circulating follicular helper

Fig. 2 | Expansion of complement-expressing nonclassical monocytes and megakaryocyte-primed progenitor cells and increased platelet activation with COVID-19 disease severity. **a**, Dot plots of gene (left) and surface protein (right) expression for myeloid populations. **b**, Bar plot of the proportion of myeloid populations from the Newcastle and London sites. Hypothesis testing was performed using a quasi-likelihood *F*-test comparing healthy controls to individuals with COVID-19. Differentially abundant cell types were determined using a 10% FDR and are marked with an asterisk. **c**, PAGA graph representing connectivity between clusters defined in **a** for healthy (top left) and COVID-19 (bottom left) monocytes and BAL macrophages (mac). Expression of *IL6* (top right) and *TNF* (bottom right) in each cluster along the predicted path for COVID-19 monocytes. **d**, Expression of differentially expressed cytokines between CD83⁺CD14⁺ monocytes and BAL macrophages shown by cells ordered by pseudotime calculated for cells from **c**. **e**, Dot plot of gene expression of DC-derived T cell polarizing cytokines in peripheral blood DC2 cells and mature BAL DCs. **f**, Heat map displaying gene-set enrichment scores for type I/3 IFN response, TNF response and JAK–STAT signatures in the myeloid populations. Asterisks indicate significance compared to healthy controls. Absolute values and other comparisons are provided in Supplementary Table 7. **g**, Heat map of predicted ligand–receptor interactions between platelets and monocyte subsets, using RNA data. **h**, Dot plot of significant differentially expressed genes between samples from healthy donors and individuals with COVID-19 filtered for platelet activation markers. **i**, UMAP representation of HSPCs (top) and dot plot of gene expression markers used to annotate clusters (bottom). MK, megakaryocyte; prog, progenitor. **j**, Bar chart depicting the proportion of progenitors. **k**, Box plots displaying the enrichment of a megakaryocyte progenitor signature in CD34⁺CD38⁺ HSPCs (right) and CD34⁺CD38⁻ (left), averaged per donor scores. Comparisons were made by an analysis of variance (ANOVA) with pairwise comparisons using Tukey's test. Asterisks above bars indicate significance and are colored by the severity for which they were compared to. Absolute values are provided in Supplementary Table 8. Boxes denote the interquartile range (IQR), and the median is shown as horizontal bars. Whiskers extend to 1.5 times the IQR, and outliers are shown as individual points (*P* values: CD38-negative cells in healthy versus LPS group (90 min), 0.3×10^{-3} ; CD38-positive cells in healthy versus moderate group, 0.7×10^{-3}).

T_{TH} cells, type 1 helper T (T_{H1}) cells, CD8⁺ effector memory T (T_{EM}) cells and MAIT cells relatively enriched in individuals with asymptomatic and mild infection, with NKT, proliferating CD8⁺ and CD4⁺, and CD8⁺ terminal effector T (T_{TE}) cells enriched in individuals with more severe infection (Fig. 3e and Extended Data Fig. 5a,b). Treating disease severity as an ordinal variable (Methods), multiple cell populations displayed nonlinear differences across disease severity (proliferating CD4⁺ and CD8⁺ T cells, CD8⁺ T_{TE}, CD4⁺

T_{H1}, CD4⁺ T_{H17}, CD4⁺ central memory T (T_{CM}) and IL-22⁺CD4⁺ T cells), illustrating the complex compositional changes to peripheral T cells that occur with COVID-19 (Fig. 3e and Extended Data Fig. 5b). IL-22-expressing CD4⁺ T cells seen in asymptomatic and mild disease could be associated with tissue-protective responses that may restrict immunopathology (Fig. 3e) as previously shown for IL-22 in influenza A virus infection²⁹ and lower viral load in COVID-19 patients' lungs³⁰. Proliferating CD4⁺ and CD8⁺ T cells



coexpressed exhaustion marker genes *LAG3* and *TOX* (Extended Data Fig. 5c), in keeping with previous studies of patients with severe COVID-19 (ref. 5). In contrast to disease severity, CD4⁺ T_{H1}, CD4⁺ T_{H2}, CD4⁺IL-22⁺ and CD4⁺ T_{CM} cells were enriched among individuals with longer symptom duration, while effector populations with a cytotoxic phenotype (CD8⁺ T_{TE}, CD8⁺ T_{EM}, MAIT and NK T cells) were enriched in individuals with shorter symptom duration (Extended Data Fig. 5d).

Differential gene expression analysis across disease severity (FDR 1%) and gene-set enrichment analysis (GSEA) identified pathways associated with inflammation and T cell activation across multiple subsets, including IL-2–STAT5 signaling, mTORC1 signaling, inflammatory response, IFN γ response, and IL-6–JAK–STAT3 signaling (Extended Data Fig. 5e). The increased activation and cytotoxic phenotype in T cells from individuals with COVID-19 was functionally validated by flow cytometry analysis of PBMCs stimulated ex vivo with SARS-CoV-2 peptide showing upregulation of CD137 and CD107 α (Extended Data Fig. 5f).

T cell receptor (TCR) clonality analysis showed that effector CD8⁺ T cells were the most clonally expanded (odds ratio (OR) (95% confidence interval (CI)) 1.81 (1.58–2.10)), $P=2.49\times 10^{-15}$ and their relative proportion increased with disease severity (Fig. 3f,g and Extended Data Fig. 5g,h and Supplementary Tables 6 and 7). Conversely, the relative proportion of clonally expanded effector memory CD8⁺ T_{EM} cells decreased in individuals with more severe disease (OR (95% CI) 0.87 (0.72–1.04), $P=0.26$; Fig. 3f,g and Supplementary Tables 9 and 10). These clonal alterations were primarily driven by severity rather than differences in symptom duration for more severely ill patients, as CD8⁺ T_{EM} clones expanded in individuals who had longer symptom duration, in line with a more developed infection trajectory in these individuals (per day, OR (95% CI) 1.02 (1.01–1.03), $P=2.66\times 10^{-10}$). The ratio of effector CD8⁺ T cells to CD8⁺ T_{EM} cells (TE:EM ratio) correlated with disease severity (linear model β 2.97, $P=2.92\times 10^{-18}$; Fig. 3g and Supplementary Tables 9 and 10), suggesting that CD8⁺ T cell differentiation outcome may contribute to both antiviral protection and immunopathology, as previously reported in animal models³¹, although bystander T cell activation cannot be excluded.

B lymphocytes and B cell receptor changes. Re-clustering of B cells and plasma cells identified nine clusters that were annotated according to canonical marker expression (Fig. 4a,b), and previously published transcriptional signatures (Extended Data Fig. 6a). This included immature, naïve, switched and non-switched memory B cells, and a cluster of cells that enriched for markers previously described in exhausted memory B cells^{32,33} (Fig. 4a,b and Extended Data Fig. 6a). We also found a large population of plasmablasts with negative expression of CD19 and CD20, with high expression of the proliferation marker *MKI67*, consistent with previous

reports on severe SARS-CoV-2 infection^{15,24}, as well as IgM⁺, IgG⁺ and IgA⁺ plasma cells (Fig. 4a,b). In individuals with symptomatic COVID-19, there was a significant expansion of plasmablasts and plasma cells (Fig. 4c and Extended Data Fig. 6b). The magnitude of this expansion increased from mild to moderate disease but was attenuated in severe to critical disease. This observation persisted even after accounting for days from symptom onset (Extended Data Fig. 6b). IgA⁺ cells were decreased in individuals with symptomatic COVID-19 due to a significant decrease of the IgA2 subclass (Fig. 4d and Extended Data Fig. 6b,c), suggestive of the maintenance of an effective mucosal humoral response in asymptomatic individuals. In parallel, we observed the greatest expansion of circulating follicular helper T (cT_{FH}) cells in asymptomatic individuals and a strong positive correlation between cT_{FH} cells and plasma cells in individuals with asymptomatic/mild disease that was lost with increasing disease severity (Figs. 3e and 4e and Extended Data Fig. 5a,b). Together, this suggests the presence of coordinated T cell and B cell responses in asymptomatic and mild disease, generating effective antiviral humoral immunity that becomes uncoupled in severe and critical disease. This is consistent with previous findings relating to the requirement of T_{FH} cells for optimal antibody responses and high-quality neutralizing antibodies in viral infection³⁴.

GSEA analysis identified interferon alpha response and interferon gamma response pathway genes enriched in all B cell subsets in individuals with COVID-19, and this was more marked in those with asymptomatic or mild disease, and attenuated in severe and critical disease (Fig. 4f and Extended Data Fig. 6d). The magnitude of type I interferon transcriptional response in B cells mirrored serum IFN α levels, which were highest in individuals with mild disease (Extended Data Fig. 2j), suggesting that the low expression of IFN response genes in B cells in severe or critical disease does not reflect an inability of B cells to respond to IFN α , but rather attenuation of IFN α . This may be because the initial antiviral response has waned in patients with severe or critical disease or because these individuals fail to sustain adequate IFN α production by myeloid cells and pDCs following symptom onset as previously reported⁷. Longitudinal sampling would be required to distinguish these two possibilities.

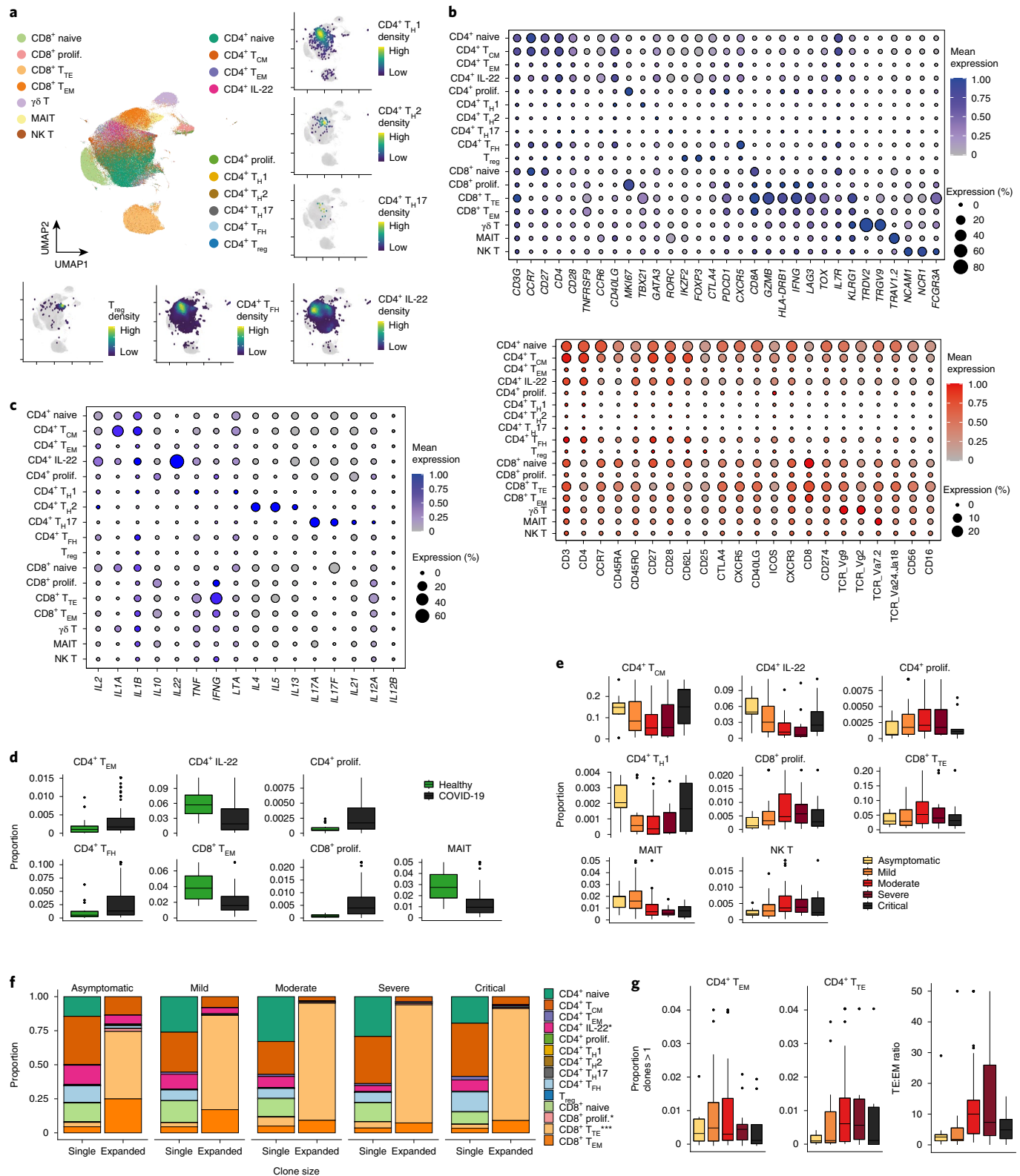
In asymptomatic individuals, TNF signaling via nuclear factor kappa B (NF- κ B) pathway genes was enriched in immature, naïve and switched memory B cells, but decreased in immature B cells and plasma cells in critical and severe disease (Fig. 4f and Extended Data Fig. 6d). Assessment of the leading-edge genes in this pathway demonstrated their markedly higher expression in all B cell and plasmablast/cell subsets in asymptomatic individuals with COVID-19 compared with those with symptomatic disease (Fig. 4g and Extended Data Fig. 6e). TNF was barely detectable in COVID-19 serum samples and highest in individuals with moderate disease (Extended Data Fig. 2j), suggesting that another

Fig. 3 | Compositional and clonal analyses of T lymphocytes illustrate the expansion of effector subsets. **a**, UMAP visualization of 309,617 T cells based on gene expression shown and colored by cell type. Insets show the two-dimensional kernel density estimates of select T cell types in UMAP space. **b**, Dot plots of gene (top) and surface protein (bottom) expression for populations shown in **a**. **c**, Dot plots of gene expression of cytokine genes for populations shown in **a**. **d**, Box plots of cell type proportions that are differentially abundant between healthy donors and individuals with COVID-19. Boxes denote the IQR, and the median is shown as horizontal bars. Whiskers extend to 1.5 times the IQR and outliers are shown as individual points ($n=24$ healthy, $n=86$ COVID-19 biologically independent samples). **e**, Box plots of the proportion of cell types shown in **a**. Only cell types showing trends of changes by severity status are shown. Boxes denote IQR with median shown as horizontal bars. Whiskers extend to 1.5 times the IQR, and outliers are shown as individual points ($n=9$ asymptomatic, $n=23$ mild, $n=30$ moderate, $n=13$ severe, $n=10$ critical biologically independent samples). **f**, Bar plots show the frequency of clonal T cells. Expanded clones denote clonotypes observed more than once. Asterisks indicate significance after multiple-testing correction (logistic regression using two-sided t -test with Benjamini–Hochberg FDR correction; CD4⁺ T_{CM} adjusted $P=0.119$, CD4⁺ T_{EM} adjusted $P=0.472$, CD4⁺IL-22⁺ adjusted $P=0.01$, CD4⁺ prolif. adjusted $P=0.993$, CD4⁺ T_{H1} adjusted $P=0.993$, CD4⁺ T_{FH} adjusted $P=0.109$, T_{reg} adjusted $P=0.993$, CD8⁺ prolif. adjusted $P=0.016$, CD8⁺ T_{TE} adjusted $P=2.49\times 10^{-15}$, CD8⁺ T_{EM} adjusted $P=0.259$). **g**, Box plots of the proportion of clonally expanded CD8⁺ T_{EM} cells (left), effector CD8⁺ T cells (middle) and the ratio of effector CD8⁺ T cells to CD8⁺ T_{EM} cells (right). Boxes denote the IQR, and the median is shown as horizontal bars. Whiskers extend to 1.5 times the IQR, and outliers are shown as individual points. Legend is as in **e**.

cytokine, for example IL-6, or stimulus may be responsible for NF- κ B activation in asymptomatic individuals with COVID-19.

Hypoxia pathway genes were enriched in immature and naïve B cells only in asymptomatic individuals (Fig. 4f and Extended Data Fig. 6d). Since these individuals are unlikely to be hypoxic (given their lack of symptoms), we postulated that this signature

may reflect another hypoxia inducible factor-activating stimulus, which includes B cell receptor (BCR) cross-linking³⁵. We assessed the expression of genes associated with BCR activation, such as *CD79A* and *CD79B*, and downstream kinases such as *BTK* in B cell subsets. Overall, BCR activation-associated genes were most highly expressed in B cells in healthy controls, followed by asymptomatic



individuals with COVID-19, with lower expression observed in all symptomatic COVID-19 groups (Fig. 4g and Extended Data Fig. 6e). BCR activation threshold is also modulated by immune tyrosine inhibitory motif-containing receptors that recruit phosphatases, increasing the activation threshold of B cells³⁶. BCR inhibitory gene expression was limited, but *CD22* was detectable across B cell subsets in asymptomatic COVID-19, while *FCGR2B*, *CD72* and *PTPN6* expression was evident in B cells in severe COVID-19 (Fig. 4g and Extended Data Fig. 6e). Together, this analysis suggests that B cells in asymptomatic individuals with COVID-19 and those with mild disease have a more pronounced response to interferons, increased NF- κ B activation and a higher expression of genes associated with BCR activation signaling, suggesting a potential for greater BCR activation. Longitudinal analysis of patient samples will be required to establish if this is due to avid responses early in disease that prevent progression to a more severe phenotype or the immune response in early disease.

Following activation, B cells differentiate into antibody-producing plasma cells, accompanied by a progressive increase in oxidative metabolism^{37,38}. We observed differences in metabolic gene pathway expression in plasmablasts and plasma cells between disease severity categories, with enrichment of oxidative phosphorylation pathway genes in all disease groups, and a relative increase in glycolysis pathway genes in asymptomatic patient plasmablasts when compared to symptomatic disease groups (Fig. 4f and Extended Data Fig. 6e).

We next assessed BCR clonality using ‘dandelion’, a single-cell BCR-sequencing analysis package (Methods), and found substantially more clonal expansion in symptomatic individuals with COVID-19 (Fig. 4h and Extended Data Fig. 7). Expanded clonotypes were found across all major cell types with larger clonotypes present primarily in plasmablast/plasma cell clusters (Extended Data Fig. 8a,b). Within the expanded clonotypes, there was some evidence of class switching within symptomatic COVID-19 groups but not in the asymptomatic/healthy individuals (Extended Data Fig. 8c). Unlike other large-scale single-cell RNA-sequencing (scRNA-seq) studies with BCR profiling^{15,24}, there was no obvious bias of immunoglobulin heavy-chain variable (*IGHV*) gene usage in general (Extended Data Fig. 9a). Disaggregating the *IGHV* gene usage data to individual gender groups showed that only *IGHV1-46* was significantly increased in women with critical COVID-19 relative to healthy controls (Extended Data Fig. 9a). Some related BCRs were present in different individuals, with more incidence of V and J gene usage and related amino acid sequences of heavy-chain and light-chain CDR3s observed in individuals with severe or critical disease, and in individuals from one center (Newcastle; Fig. 4i), which could arise due to local variants of the virus driving expansion

of specific B cell clones. We note that none of these related BCRs were found to be expanded in the individuals, which was expected as only a relatively small number of B cells per individual were sampled. It would have been unlikely to find exactly matching heavy-chain and light-chain sequences across different individuals (even when allowing for somatic hypermutation variation), given the expected low coverage that arises from a small number of single cells (relative to bulk BCR sequencing). Finally, we observed disproportionate distribution in clonotype size, whether considering expanded or all clonotypes, and increased BCR mutations between men and women with COVID-19, with greater levels of both in women compared with men (Fig. 4j and Extended Data Fig. 9b). These differences in clonal expansion of B cells are consistent with a role in previous reports of worse outcomes in COVID-19 in men^{39,40}.

We summarize the immunological cellular and molecular profiles observed in our study distinguishing features between asymptomatic/mild disease from severe/critical disease (Fig. 5). Future longitudinal studies may enable us to distinguish if the distinct responses in asymptomatic and milder disease prevent progression to severe phenotypes.

Discussion

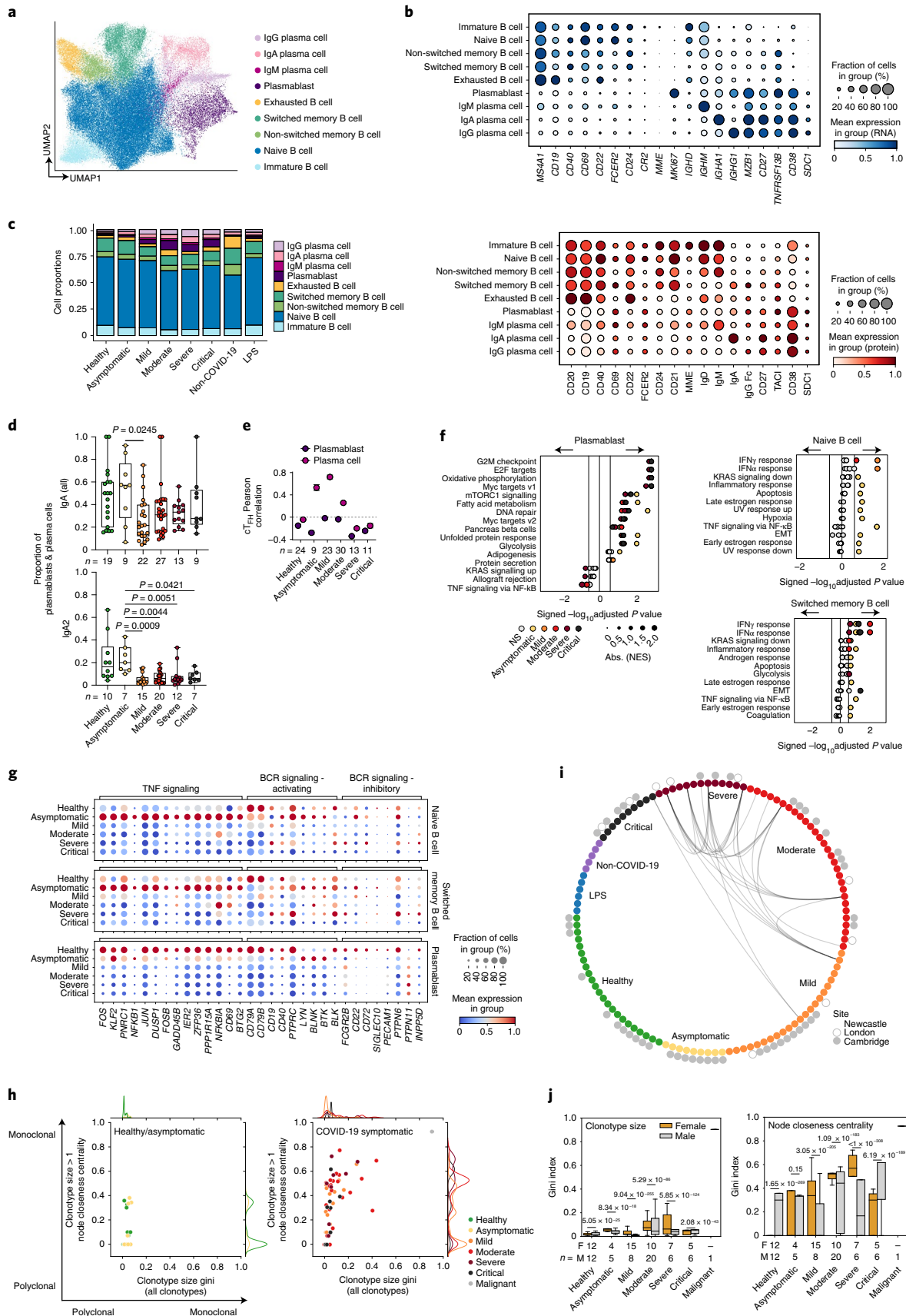
Our cross-sectional multi-omics PBMC survey revealed several new insights into COVID-19 pathogenesis. Firstly, peripheral blood monocytes and DCs exhibit an interferon response to infection. We identified CD1QA/B/C⁺CD16⁺ monocytes, coexpressing receptors and ligands for interactions with platelets, that are predicted to replenish alveolar macrophages in COVID-19. Secondly, altered hematopoiesis is evident in the peripheral circulation with megakaryocyte-primed gene expression in the earliest CD34⁺CD38⁺HSPCs, and expanded megakaryocyte progenitors in the response to COVID-19. We reveal a balance in protective versus immunopathogenic adaptive immune responses in COVID-19 patients. Previous studies have reported expanded proliferative CD4⁺ and CD8⁺ T cells with disease severity²⁴, but a reduction in $\gamma\delta$ T cells^{24,41}, consistent with our study. In addition, we observed enrichment of T_{H1} cells in asymptomatic donors, consistent with previously reported IFN γ and IL-2 antigen-specific T cells in asymptomatic individuals⁴². We report expansion of CD8⁺ effector T cells, which likely include antigen-specific short-lived effector cells that could lead to uncontrolled inflammation and immunopathology, expanding on previous reports^{43–46}.

The expansion of plasmablasts and plasma cells is less evident in critical and severe disease than in moderate and mild disease, in contrast to previous studies that reported the diminished plasmablast expansion in convalescent stages and not within active disease⁴⁵. This response is paralleled by the T_{FH} profile in individuals

Fig. 4 | Single-cell analysis of B lymphocytes and BCR repertoire reveal plasmablast expansion and clonality differences between genders. **a**, UMAP visualization of 74,019 cells in the B cell lineage identified from gene expression data. **b**, Dot plots of gene (top) and surface protein (bottom) expression for populations shown in **a**. **c**, Bar plot of the mean proportion of cell types shown in **a**. **d**, Proportion of total IgA and IgA2 in plasmablast and plasma cells based on BCR data. Kruskal-Wallis test with Benjamini-Hochberg correction. **e**, Coordinated changes between T_{FH} and B cells assessed by differential correlation analysis (empirical $P \leq 0.1$). Shown is the Pearson correlation (\pm bootstrap s.e.m.) between T_{FH} proportions and plasmablast or plasma cell (combined); only significant trends are shown. **f**, GSEA of MSigDB hallmark signatures in naive B cells, switched memory B cells and plasmablasts for asymptomatic/symptomatic COVID-19 versus healthy groups. Size of circles indicate (absolute) normalized enrichment score (NES). GSEA (permutation) nominal $P < 0.05$ and FDR < 0.25 denoted by non-gray colored dots. EMT, epithelial-mesenchymal transition; UV, ultraviolet. **g**, Dot plots of genes related to TNF signaling and BCR signaling in naive B cells, switched memory B cells and plasmablasts. Size of circles indicates the percentage of cells expressing the gene, and color gradient corresponds to increasing mean expression value. **h**, Scatterplot of clonotype size by node closeness centrality gini indices with marginal histograms indicating the distribution. Each dot represents an individual. **i**, BCR overlap incidence plot. Nodes correspond to individual donors colored by (inner ring) severity and (outer ring) site from which samples were collected. Edges indicate if at least one cell from each individual displayed an identical combination of heavy and light-chain V and J gene usage with CDR3 similarity allowance ($\geq 85\%$). **j**, Clonotype size (left) and node closeness centrality gini indices (right) separated by gender. Mann-Whitney U test with Benjamini-Hochberg correction between the gender groups within each severity status. Color of adjusted P values indicates the gender group with the higher mean value. The box portion of the box plots extends from the 25th to 75th percentiles, whiskers extend from the smallest to largest values, and the middle line corresponds to the median. NS, not significant.

with COVID-19 and is consistent with postmortem observations showing a lack of germinal centers in lymph nodes and spleen in individuals with fatal COVID-19 and a decrease in T_{FH} cells⁴⁴.

Our data revealed a significant decrease in IgA2 in symptomatic COVID-19 compared to asymptomatic donors, suggesting that maintenance of a robust mucosal humoral immune response



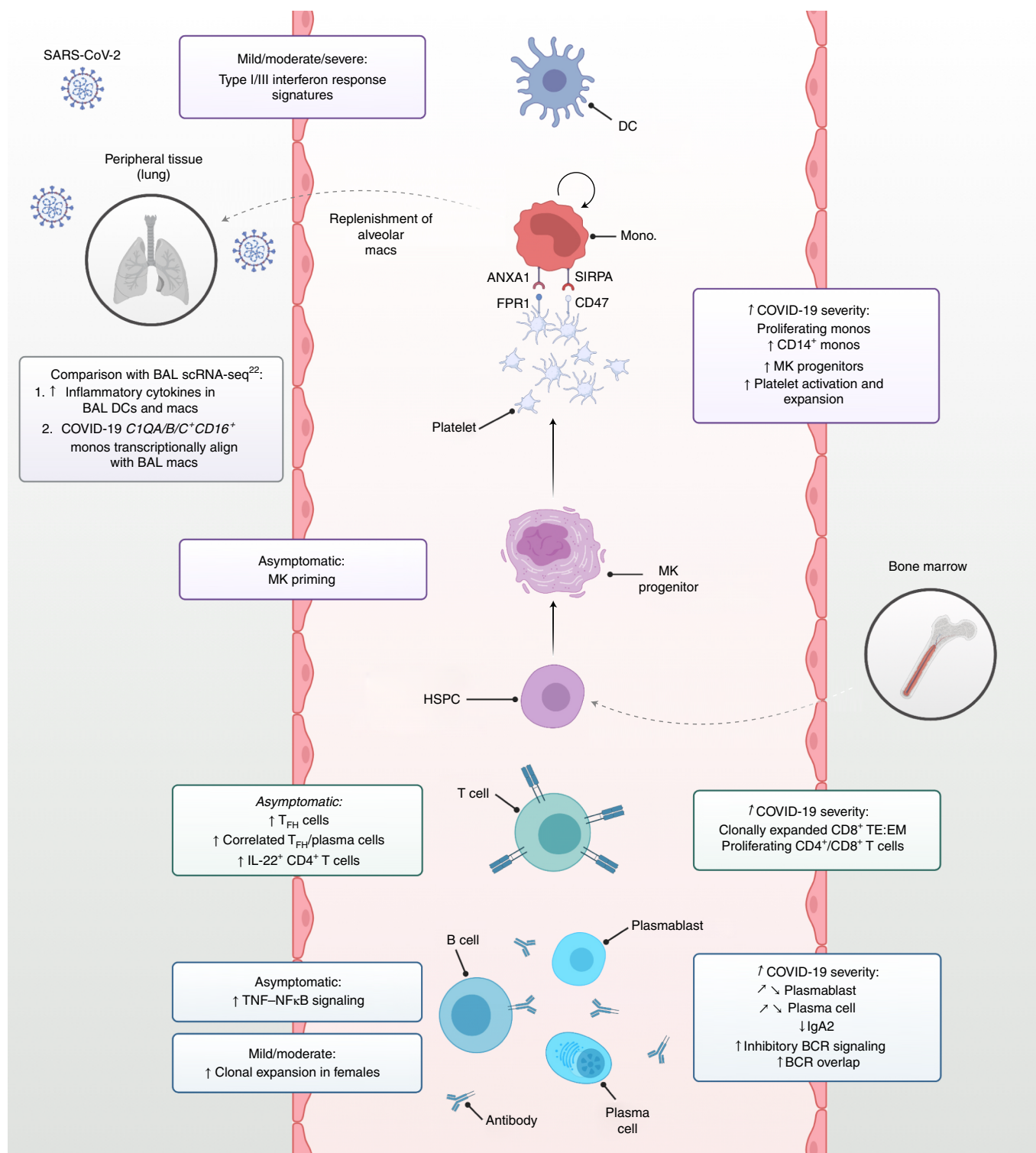


Fig. 5 | Integrated framework of the peripheral immune response in COVID-19. Schematic illustration of study highlights. Created with BioRender.com.

may influence the fate of individuals infected with SARS-CoV-2. We observed a diminished IFN α response in the B cell compartments of individuals with critical and severe disease, further emphasizing a crucial role of these responses in outcomes, as previously reported in patients with COVID-19 who had type I IFN antibodies¹⁷. Our data also suggest differential BCR clonality and mutation frequencies between gender groups, which may contribute to the differing clinical outcomes observed between men and women

with COVID-19⁴⁸. Our study provides a valuable resource, exploitable for translational studies, and a template for future integrative meta-analysis of single-cell multi-omics datasets from individuals with COVID-19 worldwide.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-021-01329-2>.

Received: 15 January 2021; Accepted: 23 March 2021;
Published online: 20 April 2021

References

- Bedford, J. et al. COVID-19: towards controlling of a pandemic. *Lancet* **395**, 1015–1018 (2020).
- John Hopkins University. COVID-19 map. Johns Hopkins Coronavirus Resource Center. <https://coronavirus.jhu.edu/map.html> (2020).
- Chen, N. et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395**, 507–513 (2020).
- Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280 (2020).
- Diao, B. et al. Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19). *Front. Immunol.* **11**, 827 (2020).
- Lee, J. S. et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci. Immunol.* **5**, eabd1554 (2020).
- Arunachalam, P. S. et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* **369**, 1210–1220 (2020).
- Woodruff, M. C. et al. Extrafollicular B cell responses correlate with neutralizing antibodies and morbidity in COVID-19. *Nat. Immunol.* **21**, 1506–1516 (2020).
- Chen, G. et al. Clinical and immunological features of severe and moderate coronavirus disease 2019. *J. Clin. Invest.* **130**, 2620–2629 (2020).
- Zhang, J.-Y. et al. Single-cell landscape of immunological responses in patients with COVID-19. *Nat. Immunol.* **21**, 1107–1118 (2020).
- Liu, C. et al. Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell* <https://doi.org/10.1016/j.cell.2021.02.018> (2021).
- Hadjadi, J. et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718–724 (2020).
- Schulte-Schrepping, J. et al. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* **182**, 1419–1440 (2020).
- Mann, E. R. et al. Longitudinal immune profiling reveals key myeloid signatures associated with COVID-19. *Sci. Immunol.* **5**, eabd6197 (2020).
- Bernardes, J. P. et al. Longitudinal multi-omics analyses identify responses of megakaryocytes, erythroid cells and plasmablasts as hallmarks of severe COVID-19. *Immunity* **53**, 1296–1314 (2020).
- World Health Organization. Clinical management of COVID-19. <https://www.who.int/publications/i/item/clinical-management-of-covid-19> (2021).
- Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
- Severe COVID-19 GWAS Group et al. Genome-wide association study of severe COVID-19 with respiratory failure. *N. Engl. J. Med.* **383**, 1522–1534 (2020).
- Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in COVID-19. *Nature* <https://doi.org/10.1038/s41586-020-03065-y> (2020).
- Lun, A. T. L., Richard, A. C. & Marioni, J. C. Testing for differential abundance in mass cytometry data. *Nat. Methods* **14**, 707–709 (2017).
- Jardine, L. et al. Lipopolysaccharide inhalation recruits monocytes and dendritic cell subsets to the alveolar airspace. *Nat. Commun.* **10**, 1999 (2019).
- Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
- Evren, E. et al. Distinct developmental pathways from blood monocytes generate human lung macrophage diversity. *Immunity* <https://doi.org/10.1016/j.immuni.2020.12.003> (2020).
- Ren, X. et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* <https://doi.org/10.1016/j.cell.2021.01.053> (2021).
- Levi, M., Thachil, J., Iba, T. & Levy, J. H. Coagulation abnormalities and thrombosis in patients with COVID-19. *Lancet Haematol.* **7**, e438–e440 (2020).
- Hottz, E. D. et al. Platelet activation and platelet-monocyte aggregate formation trigger tissue factor expression in patients with severe COVID-19. *Blood* **136**, 1330–1341 (2020).
- Velten, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).
- Mende, N. et al. Quantitative and molecular differences distinguish adult human medullary and extramedullary haematopoietic stem and progenitor cell landscapes. Preprint at *BioRxiv* <https://doi.org/10.1101/2020.01.26.919753> (2020).
- Ivanov, S. et al. Interleukin-22 reduces lung inflammation during influenza A virus infection and protects against secondary bacterial infection. *J. Virol.* **87**, 6911–6924 (2013).
- Desai, N. et al. Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nat. Commun.* **11**, 6319 (2020).
- Joshi, N. S. et al. Inflammation directs memory precursor and short-lived effector CD8⁺ T cell fates via the graded expression of T-bet transcription factor. *Immunity* **27**, 281–295 (2007).
- Monaco, G. et al. RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep.* **26**, 1627–1640 (2019).
- Adlowitz, D. G. et al. Expansion of activated peripheral blood memory B cells in rheumatoid arthritis, impact of B cell depletion therapy, and biomarkers of response. *PLoS ONE* **10**, e0128269 (2015).
- Crotty, S. T. T follicular helper cell biology: a decade of discovery and diseases. *Immunity* **50**, 1132–1148 (2019).
- Meng, X. et al. Hypoxia-inducible factor-1 α is a critical transcription factor for IL-10-producing B cells in autoimmune disease. *Nat. Commun.* **9**, 251 (2018).
- Pritchard, N. R. & Smith, K. G. C. B cell inhibitory receptors and autoimmunity. *Immunology* **108**, 263–273 (2003).
- Doughty, C. A. et al. Antigen receptor-mediated changes in glucose metabolism in B lymphocytes: role of phosphatidylinositol 3-kinase signaling in the glycolytic control of growth. *Blood* **107**, 4458–4465 (2006).
- Boothby, M. & Rickert, R. C. Metabolic regulation of the immune humoral response. *Immunity* **46**, 743–755 (2017).
- Guan, W.-J. et al. Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).
- Onder, G., Rezza, G. & Brusaferro, S. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *J. Am. Med. Assoc.* **323**, 1775–1776 (2020).
- Parrot, T. et al. MAIT cell activation and dynamics associated with COVID-19 disease severity. *Sci. Immunol.* **5**, eabe1670 (2020).
- Le Bert, N. et al. Highly functional virus-specific cellular immune response in asymptomatic SARS-CoV-2 infection. *J. Exp. Med.* **218**, e20202617 (2021).
- Rydzynski Moderbacher, C. et al. Antigen-specific adaptive immunity to SARS-CoV-2 in acute COVID-19 and associations with age and disease severity. *Cell* **183**, 996–1012 (2020).
- Kaneko, N. et al. Loss of Bcl-6-expressing T follicular helper cells and germinal centers in COVID-19. *Cell* **183**, 143–157 (2020).
- Maloy, K. J. et al. CD4⁺ T cell subsets during virus infection. Protective capacity depends on effector cytokine secretion and on migratory capability. *J. Exp. Med.* **191**, 2159–2170 (2000).
- Sahin, U. et al. COVID-19 vaccine BNT162b1 elicits human antibody and T_H1 T cell responses. *Nature* **586**, 594–599 (2020).
- Bastard, P. et al. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370**, eabd4585 (2020).
- Takahashi, T. et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* **588**, 315–320 (2020).

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

¹Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK. ²Wellcome - MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. ³European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, UK. ⁴Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. ⁵Molecular Immunity Unit, Department of Medicine, University of Cambridge, Cambridge, UK. ⁶Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. ⁷UCL Respiratory, Division of Medicine, University College London, London, UK. ⁸Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK. ⁹NIHR Newcastle Biomedical Research Centre, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. ¹⁰Department of Infection and Tropical Medicine, Newcastle upon Tyne Hospitals NHS Foundation, Newcastle upon Tyne, UK. ¹¹Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, Cambridge, UK. ¹²Department of Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. ¹³University College London Hospitals NHS Foundation Trust, London, UK. ¹⁴Royal Free Hospital NHS Foundation Trust, London, UK. ¹⁵UCL Great Ormond Street Institute of Child Health, London, UK. ¹⁶Department of Dermatology, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. ¹⁷The Innovation Lab Integrated COVID Hub North East, Newcastle Upon Tyne, UK. ¹⁸School of Computing, Newcastle University, Newcastle Upon Tyne, UK. ¹⁹Centre for Immunobiology, Blizard Institute, Queen Mary University of London, London, UK. ²⁰Integrated Critical Care Unit, Sunderland Royal Hospital, South Tyneside and Sunderland NHS Foundation Trust, Sunderland, UK. ²¹Theory of Condensed Matter Group, Cavendish Laboratory/Department of Physics, University of Cambridge, Cambridge, UK. ²²Cambridge Institute for Therapeutic Immunology and Infectious Disease, Cambridge Biomedical Campus, Cambridge, UK. ²³NIHR Cambridge Biomedical Research Centre, Cambridge, UK. ²⁴Present address: Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, Ludwig-Maximilians-Universität Munich, Munich, Germany. ⁴³These authors contributed equally: Emily Stephenson, Gary Reynolds, Rachel A. Botting, Fernando J. Calero-Nieto, Michael D. Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak. ⁴⁴These authors jointly supervised this work: Kerstin B. Meyer, Marko Z. Nikolic, Christopher J. A. Duncan, Ken Smith, Sarah A. Teichmann, Menna R. Clatworthy, John C. Marioni, Berthold Göttgens, Muzlifah Haniffa. ✉e-mail: st9@sanger.ac.uk; mrc38@cam.ac.uk; marioni@ebi.ac.uk; bg200@cam.ac.uk; m.a.haniffa@ncl.ac.uk

Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration

Stephen Baker^{11,12}, John R. Bradley^{12,25}, Gordon Dougan^{11,12}, Ian G. Goodfellow²⁶, Ravindra K. Gupta^{11,12}, Christoph Hess^{11,12,27,28}, Nathalie Kingston^{25,29}, Paul J. Lehner^{11,12}, Paul A. Lyons^{11,12}, Nicholas J. Matheson^{11,12}, Willem H. Owehand²⁹, Caroline Saunders³⁰, Kenneth G. C. Smith^{11,12}, Charlotte Summers^{12,31,32,33}, James E. D. Thaventhiran^{11,12,34}, Mark Toshner^{12,31,32}, Michael P. Weekes¹¹, Ashlea Bucke³⁰, Jo Calder³⁰, Laura Canna³⁰, Jason Domingo³⁰, Anne Elmer³⁰, Stewart Fuller³⁰, Julie Harris³⁵, Sarah Hewitt³⁰, Jane Kennet³⁰, Sherly Jose³⁰, Jenny Kourampa³⁰, Anne Meadows³⁰, Criona O'Brien³⁵, Jane Price³⁰, Cherry Publico³⁰, Rebecca Rastall³⁰, Carla Ribeiro³⁰, Jane Rowlands³⁰, Valentina Ruffolo³⁰, Hugo Tordesillas³⁰, Ben Bullman¹¹, Benjamin J. Dunmore¹², Stuart Fawke³⁶, Stefan Gräf^{12,25,29}, Josh Hodgson¹², Christopher Huang¹², Kelvin Hunter^{11,12}, Emma Jones³⁷, Ekaterina Legchenko¹², Cecilia Matara¹², Jennifer Martin¹², Ciara O'Donnell¹², Linda Pointon¹², Nicole Pond^{11,12}, Joy Shih¹², Rachel Sutcliffe¹², Tobias Tilly¹², Carmen Treacy¹², Zhen Tong¹², Jennifer Wood¹², Marta Wylot³⁸, Laura Bergamaschi^{11,12}, Ariana Betancourt^{11,12}, Georgie Bower^{11,12}, Aloka De Sa¹², Madeline Epping^{11,12}, Oisín Huhn³⁴, Sarah Jackson¹², Isobel Jarvis¹², Jimmy Marsden¹², Francesca Nice³⁹, Georgina Okecha¹², Omar Omarjee¹², Marianne Perera¹², Nathan Richoz¹², Rahul Sharma¹², Lori Turner^{11,12}, Eckart M. D. D. De Bie¹², Katherine Bunclark¹², Masa Josipovic³¹, Michael Mackay¹², Federica Mescia^{11,12}, Alice Michael³², Sabrina Rossi⁴⁰, Mayurun Selvan¹², Sarah Spencer⁴¹, Cissy Yong⁴⁰, Ali Ansaripour³², Lucy Mwaura³², Caroline Patterson³², Gary Polwarth³², Petra Polgarova³³, Giovanni di Stefano³³, John Allison^{25,29}, Helen Butcher^{25,42}, Daniela Caputo^{25,42}, Debbie Clapham-Riley^{25,42}, Eleanor Dewhurst^{25,42}, Anita Furlong^{25,42}, Barbara Graves^{25,42}, Jennifer Gray^{25,42}, Tasmin Ivers^{25,42}, Mary Kasanicki^{25,33}, Emma Le Gresley^{25,42}, Rachel Linger^{25,42}, Sarah Meloy^{25,42}, Francesca Muldoon^{25,42}, Nigel Ovington^{25,29}, Sofia Papadia^{25,42}, Isabel Phelan^{25,42}, Hannah Stark^{25,42}, Kathleen E. Stirrups^{25,29}, Paul Townsend^{25,29}, Neil Walker^{25,29} and Jennifer Webster^{25,42}

²⁵NIHR BioResource, Cambridge University Hospitals NHS Foundation, Cambridge Biomedical Campus, Cambridge, UK. ²⁶Division of Virology, Department of Pathology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. ²⁷Department of Biomedicine, University and University Hospital Basel, Basel, Switzerland. ²⁸Botnar Research Centre for Child Health (BRCC), University Basel & ETH Zurich, Basel, Switzerland. ²⁹Department of

Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. ³⁰Cambridge Clinical Research Centre, NIHR Clinical Research Facility, Cambridge University Hospitals NHS Foundation Trust, Addenbrooke's Hospital, Cambridge, UK. ³¹Metabolic Research Laboratories, Wellcome Trust-Medical Research Council Institute of Metabolic Science, University of Cambridge, Cambridge, UK. ³²Royal Papworth Hospital NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, UK. ³³Addenbrooke's Hospital, Cambridge, UK. ³⁴Department of Obstetrics & Gynaecology, The Rosie Maternity Hospital, Robinson Way, Cambridge, UK. ³⁵Department of Paediatrics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. ³⁶Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. ³⁷Department of Veterinary Medicine, Cambridge, UK. ³⁸Department of Biochemistry, University of Cambridge, Cambridge, UK. ³⁹Cancer Molecular Diagnostics Laboratory, Department of Oncology, University of Cambridge, Cambridge, UK. ⁴⁰Department of Surgery, Addenbrooke's Hospital, Cambridge, UK. ⁴¹Department of Clinical Biochemistry and Immunology, Addenbrooke's Hospital, Cambridge, UK. ⁴²Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

Methods

Ethics and sample collection. *Newcastle.* Participants were recruited and consented under the Newcastle Biobank (Research Ethics Committee (REC) no. 17/NE/0361; Integrated Research Application System (IRAS) no. 233551) study and ethical governance. For the COVID-19-positive samples and healthy controls, peripheral blood was collected in EDTA tubes and serum separator tubes and processed within 4 h of collection.

For the IV-LPS control samples, ethical approval was granted by a REC (17/YH/0021). Healthy volunteers provided informed, written consent and passed the inclusion criteria as set out in the study protocol (Supplementary File 1.) LPS was obtained from Clinical Center Reference Endotoxin (94332B1, National Institutes of Health) and injected intravenously as a bolus dose of 2 ng kg⁻¹. Blood samples were taken before IV-LPS administration (baseline), at 90 min, and at 10 h after challenge. Venous blood was drawn from an 18-gauge venous cannula and was collected into EDTA and serum separator tubes. Only samples from 90 min and 10 h were analyzed in this study.

Cambridge. Study participants were recruited between 31 March 2020 and 20 July 2020 from patients attending Addenbrooke's Hospital with suspected COVID-19 or a confirmed diagnosis of COVID-19 by nucleic acid amplification test (including point-of-care testing), patients admitted to Royal Papworth Hospital NHS Foundation Trust or Cambridge and Peterborough Foundation Trust with a confirmed diagnosis of COVID-19, together with health-care workers identified through staff screening with a positive PCR test for SARS-CoV-2. Controls were recruited among hospital staff attending Addenbrooke's serology screening program, and selected to cover the whole age spectrum of COVID-19-positive study participants, across gender groups. Only controls with negative serology results (45 of 47) were subsequently included in the study. Recruitment of inpatients at Addenbrooke's Hospital and health-care workers was undertaken by the National Institute for Health Research (NIHR) Cambridge Clinical Research Facility outreach team and the NIHR BioResource research nurse team. Ethical approval was obtained from the East of England Cambridge Central Research Ethics Committee (NIHR BioResource, REC no. 17/EE/0025; 'Genetic variation and altered leukocyte function in health and disease (GANDALF)', REC no. 08/H0308/176). Informed consent was obtained from all participants. Each participant provided 27 ml of peripheral venous blood collected into a 9-ml sodium citrate tube.

London. Participants aged 18 years and older were recruited from two large hospital sites in London, United Kingdom, namely University College London (UCL) Hospitals NHS Foundation Trust and Royal Free London NHS Foundation Trust during the height of the pandemic in the United Kingdom (April to July 2020). All participants provided informed consent. Ethical approval was obtained through the Living Airway Biobank, administered through UCL Great Ormond Street Institute of Child Health (REC no. 19/NW/0171, IRAS project no. 261511), as well as by the local R&D departments at both hospitals. At daily virtual COVID-19 coordination meetings, suitable participants were chosen from a list of newly diagnosed and admitted patients within the preceding 24 h (based on a positive nasopharyngeal swab for SARS-CoV-2). Participants with typical clinical and radiological COVID-19 features but with a negative screening test for SARS-CoV-2 were excluded. Other exclusion criteria included active hematological malignancy or cancer, known immunodeficiencies, sepsis from any cause and a blood transfusion within 4 weeks. Maximal severity of COVID-19 was determined retrospectively by identifying the presence of symptoms, the need for oxygen supplementation and the level of respiratory support. Peripheral blood sampling was performed before inclusion to any pharmacological interventional trials.

Samples were collected and transferred to a category level 3 facility at UCL and processed within 2 h of sample collection. Peripheral blood was centrifuged after adding Ficoll-Paque Plus, and PBMCs, serum and neutrophils were separated, collected and frozen for later processing.

Clinical status assignment. Clinical metadata were collected at the point-of-care sample collection, including current oxygen requirements and location. This was used to assign disease severity status. Participants based on a ward and not requiring oxygen were defined as having 'mild' disease. Participants outside of an intensive care unit (ICU) environment requiring oxygen were defined as having 'moderate' disease. All patients in ICU and/or requiring noninvasive ventilation were defined as having 'severe' disease. Participants requiring intubation and ventilation were defined as having 'critical' disease. There were no patients in ICU that did not require supplemental oxygen.

PBMC isolation and dead cell removal. *Newcastle.* PBMCs were isolated from blood samples using Lymphoprep (StemCell Technologies) density gradient centrifugation according to the manufacturer's instructions. Single-cell suspensions were then washed with Dulbecco's PBS (Sigma) and frozen in aliquots containing 5–10 million cells in 90% (vol/vol) heat-inactivated FCS (Gibco) and 10% (vol/vol) DMSO (Sigma-Aldrich). On the day of the experiment, the cells were thawed for 1 min, transferred to wash buffer (PBS supplemented with 2% (vol/vol) FCS and 2 mM EDTA) and centrifuged at 500g for 5 min. Resuspended cells were

passed through a 30- μ m filter and counted before live-cell magnetic-activated cell sorting (MACS) enrichment with the dead cell removal kit (Miltenyi Biotec), per the manufacturer's instructions. Cell pellets were resuspended in microbeads and incubated at room temperature for 15 min. Each stained sample was passed through an LS column and rinsed with binding buffer (all from Miltenyi Biotec) before centrifugation. Cell pellets were resuspended in wash buffer and counted for antibody staining by cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq).

Cambridge. PBMCs were isolated using Leucosep tubes (Greiner Bio-One) with Histopaque 1077 (Sigma) by centrifugation at 800g for 15 min at room temperature. PBMCs at the interface were collected, rinsed twice with autoMACS running buffer (Miltenyi Biotec) and cryopreserved in FBS with 10% DMSO. All samples were processed within 4 h of collection. Purified PBMCs were thawed at 37°C, transferred to a 50-ml tube, and ten volumes of prewarmed thawing medium (IMDM; Gibco, 12440-053) with 50% (vol/vol) FCS (not heat inactivated; PAN-Biotech, P40-37500) and 0.1 mg ml⁻¹ DNase I (Worthington, LS002139) were added slowly and dropwise, followed by centrifugation at 500g for 5 min. The pellet was resuspended in 1 ml of FACS buffer (PBS; Sigma, D8537) with 3% (vol/vol) heat-inactivated FCS, and the viability of each sample was assessed by counting in an improved Neubauer chamber using Trypan blue. Pools of four samples were generated by combining 0.5 million live cells per individual (2 million live cells in total). The pools were washed twice in FACS buffer (10 ml and 2 ml, respectively) followed by centrifugation for 5 min at 500g. The pellet was then resuspended in 35 μ l of FACS buffer and the viability of each pool was assessed.

London. Peripheral whole blood was collected in EDTA tubes and processed fresh via Ficoll-Paque Plus separation (GE Healthcare, 17144002). The blood was first diluted with 5 ml 2 mM EDTA-PBS (Invitrogen, 1555785-038), before 10–20 ml of diluted blood was carefully layered onto 15 ml of Ficoll in a 50-ml falcon tube. If the sample volume was less than 5 ml, blood was diluted with an equal volume of EDTA-PBS and layered onto 3 ml Ficoll. The sample was centrifuged at 800g for 20 min at room temperature. The plasma layer was carefully removed and the PBMC layer collected using a sterile Pasteur pipette. The PBMC layer was washed with three volumes of EDTA-PBS by centrifugation at 500g for 10 min. The pellet was suspended in EDTA-PBS and centrifuged again at 300g for 5 min. The PBMC pellet was collected and the cell number and viability assessed using Trypan blue. Cell freezing medium (90% FBS and 10% DMSO) was added dropwise to PBMCs slowly on ice and the mixture cryopreserved at –80°C until further full-sample processing.

TotalSeq-C antibody staining and 10x Chromium loading. *Newcastle.*

Approximately 200,000 cells from each donor were stained with Human TruStain FcX Fc Blocking Reagent (BioLegend, 422302) for 10 min at room temperature. The cells were then stained with the custom panel TotalSeq-C (BioLegend, 99813; Supplementary Table 1) for 30 min at 4°C. Cells were then washed twice with PBS supplemented with 2% (vol/vol) FCS and 2 mM EDTA (Sigma) before resuspending in PBS and counting. Approximately 20,000–30,000 cells per sample were loaded onto the 10x Chromium controller using Chromium NextGEM Single Cell V(D)J Reagent kits v1.1 with Feature Barcoding technology for Cell-Surface Protein (10x Genomics) according to the manufacturer's protocol.

Cambridge. Half a million viable cells were resuspended in 25 μ l of FACS buffer and incubated with 2.5 μ l of Human TruStain FcX Fc blocking reagent (BioLegend, 422302) for 10 min at 4°C. The TotalSeq-C antibody cocktail (BioLegend 99813; Supplementary Table 1) was centrifuged at 14,000g at 4°C for 1 min, resuspended in 52 μ l of FACS buffer, incubated at room temperature for 5 min and centrifuged at 14,000g at 4°C for 10 min. Around 25 μ l of solution was subsequently added to each sample pool and incubated for 30 min at 4°C in the dark. Pools were washed three times with 27 volumes (1.4 ml) of FACS buffer, followed by centrifugation at 500g for 5 min. The pellet was resuspended in 62.5 μ l of 1 \times PBS + 0.04% BSA (Ambion, AM2616), filtered through a 40- μ m cell strainer (Flowmi, H13680-0040), and viable cells of each sample pool were counted in an improved Neubauer chamber using Trypan blue. Around 50,000 live cells (up to a maximum of 60,000 total cells) for each pool were processed using Single Cell V(D)J 5' version 1.1 (1000020) together with Single Cell 5' Feature Barcode library kit (1000080), Single Cell V(D)J Enrichment Kit, human B cells (1000016) and Single Cell V(D)J Enrichment Kit, human T cells (10x Genomics, 1000005) according to the manufacturer's protocols.

London. Frozen PBMC samples were thawed quickly in a water bath at 37°C. Warm RPMI 1640 medium (20–30 ml) containing 10% FBS was added slowly to the cells before centrifuging at 300g for 5 min. The pellet was then washed with 5 ml RPMI 1640-FBS and centrifuged again (300g for 5 min). The PBMC pellet was collected and cell number and viability determined using Trypan blue. PBMCs from four different donors were then pooled together at equal numbers (1.25 \times 10⁵ PBMCs from each donor) to make up 5.0 \times 10⁵ cells in total. The remaining cells were used for DNA extraction (Qiagen, 69504). The pooled PBMCs were stained with TotalSeq-C antibodies (BioLegend, 99814) according to manufacturer's

instructions. After incubating with half a vial of TotalSeq-C for 30 min at 4 °C, PBMCs were washed three times by centrifugation at 500g for 5 min at 4 °C. PBMCs were counted again and processed immediately for 10x 5' single-cell capture (Chromium Next GEM Single Cell V(D)J Reagent Kit v1.1 with Feature Barcoding technology for Cell-Surface Protein-Rev D protocol). Two lanes of 25,000 cells were loaded per pool on a 10x chip.

Library preparation and sequencing. *Newcastle and London.* Gene expression, TCR-enriched and BCR-enriched libraries were prepared for each sample according to the manufacturer's protocol (10x Genomics). Cell-surface protein libraries were subjected to double the manufacturer's recommended primer concentration and seven to eight amplification cycles during the sample index PCR to reduce the likelihood of daisy chains forming. Libraries were pooled per participant using a ratio of 6:2:1:1 for gene expression, cell-surface protein, TCR-enriched and BCR-enriched libraries, respectively. All libraries were sequenced using a NovaSeq 6000 (Illumina) to achieve a minimum of 50,000 paired-end reads per cell for gene expression and 20,000 paired-end reads per cell for cell-surface protein, TCR-enriched and BCR-enriched libraries.

Cambridge. The samples were subjected to 12 cycles of cDNA amplification and 8 cycles for the cell-surface protein library construction. Following this, the libraries were processed according to the manufacturer's protocol. Libraries were pooled per sample using a ratio of 9:2.4:1:0.6 for gene expression, cell-surface protein expression, and TCR and BCR enrichment libraries, respectively. Samples were sequenced using a NovaSeq 6000 (Illumina), using S1 flow cells.

Alignment and quantification. Droplet libraries were processed using Cell Ranger v4.0. Reads were aligned to the GRCh38 human genome concatenated to the SARS-CoV-2 genome (NCBI SARS-CoV-2 isolate Wuhan-Hu-1) using STAR⁴⁹ (v2.5.1b) and unique molecular identifiers (UMIs) deduplicated. CITE-seq UMIs were counted for GEX and ADT libraries simultaneously to generate feature-X droplet UMI count matrices.

Doublet identification. *Newcastle.* Scrublet (v0.2.1) was applied to each sample to generate a doublet score. These formed a bimodal distribution so the tool's automatic threshold was applied.

Cambridge. Non-empty droplets were called within each multiplexed pool of donors using the emptyDrops function implemented in the Bioconductor package DropletUtils (v1.10.3), using a UMI threshold of 100 and FDR of 1%. The probability of being a doublet was estimated for each cell per sample (that is, one 10x lane) using the 'doubletCells' function in 'scrn' based on highly variable genes (HVGs). Next, we used 'cluster_walktrap' on the shared nearest neighbor graph that was computed on HVGs to form highly resolved clusters per sample. Per-sample clusters with either a median doublet score greater than the median + 2.5× median absolute deviation (MAD) or clusters containing more than the median + 2.5× MAD genotype doublets were tagged as doublets. This was followed by a second round of highly resolved clustering across the whole dataset, in which again cells belonging to clusters with a high proportion (>60%) of cells previously labeled as doublets were also defined as doublets.

London. For pooled donor CITE-seq samples, the donor ID of each cell was determined by genotype-based demultiplexing using soupcorell (v2)⁵⁰. Soupcorell analyses were performed with 'skip_remap' enabled and a set of known donor genotypes given under the 'common_variants' parameter. The donor ID of each soupcorell genotype cluster was annotated by comparing each soupcorell genotype to the set of known genotypes. Droplets that contained more than one genotype according to soupcorell were flagged as 'ground-truth' doublets for heterotypic doublet identification. Ground-truth doublets were used by DoubletFinder (v2.0.3)⁵¹ to empirically determine an optimal 'pK' value for doublet detection. DoubletFinder analysis was performed on each sample separately using ten principal components (PCs), a 'pN' value of 0.25, and the 'nExp' parameter estimated from the fraction of ground-truth doublets and the number of pooled donors.

CITE-seq background signal removal. Background antibody-specific and nonspecific staining was subtracted from ADT counts in each data set from the 3 data acquisition sites separately. Antibody-derived tag (ADT) counts for each protein were first normalized using counts per million and log transformed, with a pseudocount of +1. To estimate the background signal for each protein, a two-component Gaussian mixture model, implemented in the R package function 'mclust' (v5.4.7), was fit across the droplets with a total UMI count of >10 and <100 from each experimental sample separately. The mean of the first Gaussian mixture model component for each protein was then subtracted from the log counts per million from the QC-passed droplets in the respective experimental sample.

Quality control, normalization, embedding and clustering. Combined raw data from the three centers was filtered to remove cells that expressed fewer

than 200 genes and >10% mitochondrial reads. Data were normalized (scanpy: normalize_total), log + 1 corrected (scanpy: log1p) and HVGs identified using the Seurat vst algorithm (scanpy: highly_variable_genes). Harmony was used to adjust PCs by sample ID and used to generate the neighborhood graph and embedded using UMAP. Clustering was performed using the Leiden algorithm with an initial resolution of 3. For initial clustering, differentially expressed genes were calculated using the Wilcoxon rank-sum test.

Cluster differential abundance testing. Numbers of cells of each cell subtype were quantified in each participant and control sample (donors) to compute a matrix of cell type × donor counts. Cell type abundance counts were modeled as a function of either disease severity (as an ordinal variable: healthy < asymptomatic < mild < moderate < severe < critical) or days from symptom onset, adjusting for age, gender, batch and days from onset, in a negative binomial generalized linear model (NB GLM), implemented in the Bioconductor package edgeR. Counts were normalized in the model using the (log) of the total numbers of all cells captured for each donor. Hypothesis testing was performed using a quasi-likelihood *F*-test for either a linear or a quadratic trend across disease severity groups (asymptomatic > mild > moderate > severe > critical), or comparing healthy controls to SARS-CoV-2-infected donors (healthy versus all asymptomatic, mild, moderate, severe and critical groups). Differentially abundant cell types were determined using a 10% FDR. Due to compositional differences across sites, when analyzing differential abundance of myeloid populations (Fig. 2), only samples from Newcastle and London were included.

Relative importance of metadata on cell type composition. The number of cells for each sample (*N* = 110 samples in total with complete metadata) and cell type (18 different cell types in total) combination was modeled with a generalized linear mixed model with a Poisson outcome. The five clinical factors (COVID-19 swab result, age, gender, disease severity at day 0 and days from onset) and the two technical factors (patient and sequencing center) were fitted as random effects to overcome the collinearity among the factors. The effect of each clinical/technical factor on cell type composition was estimated by the interaction term with the cell type. The likelihood ratio test was performed to assess the statistical significance of each factor on cell type abundance by removing one interaction term from the full model at a time. The number of factors was used to adjust multiple testing with the Bonferroni approach. The 'glmer' function in the lme4 package implemented on R was used to fit the model. The standard error of variance parameter for each factor was estimated using the 'numDeriv' package.

Cydar analysis. We utilized cydar to identify changes in cell composition across the different severity groups based on the protein data alone. First, the background-corrected protein counts from the three different sites were integrated using the 'fastMNN' method (*k* = 20, *d* = 50, *cos.norm* = TRUE) in batchelor (v1.6.2). The batch-corrected counts for 188 proteins (four rat/mouse antibody isotypes were removed) were then used to construct hyperspheres using the 'countCells' function (downsample = 7) with the tolerance parameter chosen so that each hypersphere had at least 20 cells, estimated using the 'neighborDistances' function. To assess whether the abundance of cells in each hypersphere were associated with disease status, hypersphere counts were analyzed using the quasi-likelihood method in edgeR (v3.32.1). After filtering out hyperspheres with an average count per sample below five, we fitted a mean-dependent trend to the NB dispersion estimates. The trended dispersion for each hypersphere was used to fit an NB GLM using the log-transformed total number of cells as the offset for each sample and blocking for gender, age and batch. The quasi-likelihood *F*-test was used to compute *P* values for each hypersphere, which were corrected for multiple testing using the spatial FDR method in cydar.

Comparisons of PBMC annotation using the Azimuth tool. The final annotation of PBMCs was compared to a published PBMC annotation using the Azimuth tool (<http://azimuth.satijalab.org/app/azimuth/>). Because of size restrictions of 100,000 cells, our data were subsampled to 10% of the total cells. After running the algorithm, results with a prediction score < 0.5 were removed (5.8% of total removed). For each cluster in the COVID-19 PBMC data, the percentage of cells mapped to each cluster in the Azimuth annotation was calculated.

Interferon, TNF and JAK-STAT response scoring. A list of genes related to response to type I interferons was obtained from the GSEA Molecular Signatures Database (MSigDB; GO:0034340). Enrichment of the interferon score was measured using the 'tl.score_genes' tool in 'scanpy', which subtracts the average expression of all genes in the dataset from the average expression of the genes in this list. The scores were averaged across clusters and clinical status and expressed as a fold change over the interferon score in the equivalent healthy cluster.

kBET analysis. The kBET⁵² algorithm (<https://github.com/theislab/kBET/>) was run for each cluster (Fig. 1) using the UMAP coordinates generated from Harmony-adjusted PCs and the sample number as the batch factor. The same procedure was then performed with the same annotation but using the UMAP coordinates generated from non-Harmony-adjusted PCs. The resultant rejection

rates were averaged across clusters and compared using a Wilcoxon paired signed-rank test.

Bronchoalveolar lavage data analysis. scRNA-seq data from BAL was obtained from the Gene Expression Omnibus (accession no. GSE145926)²². Raw data were analyzed using the same pipeline as for PBMC data, specifically using the same QC cutoffs (minimum of 200 genes and <10% mitochondrial reads per cell), and data were batch corrected using Harmony by donor ID. To gain greater resolution of MPs, the DCs and macrophages were analyzed with further rounds of subclustering to identify DC1, DC2 and mature DC subsets.

PAGA analysis of blood monocytes and BAL macrophages. Annotated raw expression datasets of BAL macrophages and COVID-19 PBMCs were merged and data log normalized and scaled as for the original datasets. The top 3,000 HVGs were chosen using the Seurat 'vst' method and used for downstream analysis. PCs were batch corrected by donor and used to build a neighborhood graph. The PAGA tool in scanpy (tl.paga) was used to generate the abstracted graph between clusters.

CellphoneDB. CellphoneDB³³ was used to assess putative interactions between monocytes (CD14_mono, CD83_CD14_mono, C1_CD16_mono, CD16_mono and Prolif_mono) and platelets. The tool was run for 100 iterations, and an expression threshold of 0.25 (limiting the analysis to genes expressed by 25% of cells). For downstream analysis, we focused on interactions between platelets and any monocyte subset.

Platelet activation. Differentially expressed genes in platelets between healthy control and COVID-19 samples were filtered for those predicted to be involved in platelet activation (https://www.gsea-msigdb.org/gsea/msigdb/cards/REACTOME_PLATELET_ACTIVATION_SIGNALING_AND_AGGREGATION).

HSPC commitment scoring. HSPCs were subsetted from the data and Leiden clusters generated using the same pipeline and parameters as for the whole PBMC dataset. Differentially expressed genes between the HSPC clusters that showed evidence of lineage commitment (megakaryocyte, erythroid and myeloid) were calculated using the FindAllMarkers tool in Seurat (with thresholds of genes expressed by 25% of cells and with a log fold change of 0.25), and genes with an adjusted *P*-value cutoff of 0.05 were used to generate gene signatures for each. Enrichment scores of these signatures in the CD38-negative and CD38-positive HSPC clusters were calculated using the tl.score_genes tool in scanpy (v1.6.0). The average expression of these enrichment scores in the CD38-negative and CD38-positive HSPC clusters was averaged by donor then compared across clinical states. Differences between groups were assessed using ANOVA with pairwise comparisons using Tukey's test.

Multiplex cytokine analysis. Serum was obtained from peripheral blood in red-topped serum Vacutainers (BD, 367815) and allowed to clot for at least 30 min before centrifugation (800g for 10 min) to separate the serum. After collection, serum was frozen at -80°C and thawed on ice on the day of experiment. The assay was carried out using the Cytokine/Chemokine/Growth Factor 45-Plex Human ProcartaPlex Panel 1 kit (Invitrogen, EPX450-12171-901), utilizing the Luminex xMAP technology and according to the manufacturer's protocol. Each sample was run in duplicate. The values of each analyte were detected using the MAGPIX system and analyzed using the ProcartaPlex Analyst v1.0 Software (Thermo Fisher Scientific).

Restimulation of PBMCs with SARS-Cov-2 peptide S. Purified PMBCs were thawed at 37°C, transferred into a 15-ml tube with 10 ml prewarmed complete culture media RPMI 1640 medium (Sigma-Aldrich, R0883) supplemented with 10% (vol/vol) FCS (Gibco, 10270-106) and 1% (vol/vol) penicillin-streptomycin (100 U ml⁻¹ and 100 µg ml⁻¹, respectively; Sigma-Aldrich, P0781) and 1% (vol/vol) L-glutamine (2 mM; Sigma-Aldrich, G7513), referred to as RPMI 10, followed by centrifugation at 500g for 5 min. Cell pellet was resuspended in 500 µl RPMI 10 with added DNase (1 µg ml⁻¹; Merck, 10104159001), divided into five wells of a round-bottom 96-well plate and left to rest at 37°C for 1 h. Cells were stimulated with SARS-CoV-2 PepTivator peptide S for pan-HLA (2 µg ml⁻¹; Miltenyi Biotec, 136-126-700) and PMA/ionomycin as a control (2 µl ml⁻¹ Cell Activation cocktail; BioLegend, 423301), and incubated at 37°C for 2 h. Negative controls were left untreated. Brefeldin A (2 µg ml⁻¹; GolgiPlug, BD Bioscience, 555029) and anti-CD107a-BB700 (1:50 dilution; clone H4A3; BD Bioscience, 566558) was added for an additional 4 h into all conditions. Cells were stained for detection of activation-induced markers and intracellular cytokines 6 h after stimulation and subjected to flow cytometry.

Flow cytometry of stimulated cells. PBMCs stimulated for 6 h with the SARS-CoV-2 peptide were washed with PBS, and cell-surface stained for 1 h at room temperature: anti-CD14-FITC (1:50 dilution; clone M5E2; BD Biosciences, 555397), anti-CD19-FITC (1:50 dilution; clone 4G7; BD Biosciences, 345776), anti-CD137-PE-Dazzle594 (1:50 dilution; clone 4B4-1; BioLegend, 309826), anti-CCR7-PE-Cy7 (1:50 dilution; clone G043H7; BioLegend,

353226), anti-CD45RO-APC-H7 (1:50 dilution; clone UCHL1; BD Biosciences, 561137), anti-CD28-BV480 (1:50 dilution; clone CD28.2; BD Biosciences, 566110), anti-CD4-BV785 (1:100 dilution; clone SK3; BioLegend, 344642), anti-CD3-BUV395 (1:50 dilution; clone UCHT1; BD Biosciences, 563546), anti-CD8-BUV496 (1:100 dilution; clone RPA-T8; BD Biosciences, 564804), anti-CD25-BUV737 (1:100 dilution; clone 2A3; BD Biosciences, 612806) and viability dye Zombie Yellow (1:200 dilution; BioLegend, 423104). Cells were washed with PBS 2% (vol/vol) FCS, fixed with 4% (wt/vol) paraformaldehyde (Thermo Fisher Scientific, 28908) and kept at 4°C overnight. Subsequently, cells were washed with PBS, permeabilized with Perm/Wash buffer (BD Biosciences, 554723) according to the manufacturer's instructions, and stained with intracellular antibodies for 1 h on ice: anti-IL10-PE (1:10 dilution; clone JES3-19F1; BD Biosciences, 559330), anti-IFN-APC (1:25 dilution; Miltenyi Biotec, 130-090-762), anti-TNF-AF700 (1:50 dilution; clone MAb11; BioLegend, 502928), anti-IL-2-BV421 (1:100 dilution; clone 5344.111; BD Biosciences, 562914) and anti-CD154-BV605 (1:50 dilution; clone 24-31; BioLegend, 310826). Cells were washed, transferred to flow cytometry 5-ml tubes, and acquired on a Symphony A5 flow cytometer (BD Biosciences). Data were analyzed by FlowJo v10 (BD Biosciences).

GSEA analysis. Preranked gene-set analysis on MSigDB v7.2 Hallmark gene sets⁵⁴ was performed using preranked gene lists with the fgsea⁵⁵ package in R. Genes were preranked according to signed $-\log_{10} P$ values for all preranked gene-set analysis procedures. For B cells, generation of the rank gene list was performed using a Wilcoxon rank-sum test (via 'tl.rank_genes_groups' in scanpy) with each day 0 COVID-19 statuses (asymptomatic to symptomatic critical) as the 'tests' versus day 0 healthy samples as the 'reference/controls'.

T cell clustering, annotation and visualization. Droplets labeled as T cells (CD4, CD8, T_{reg}, MAIT and γδ T) were subset from those in Fig. 1b and reclustered using a set of HVGs calculated within each batch, the union of which was used to estimate the first 50 PCs across cells using the 'irlba' R package (v2.3.3). Batch effects were removed across the first 30 PCs using the fastMNN⁵⁶ implementation in the Bioconductor package batchelor (v1.6.2; *k* = 50). A *k*-nearest-neighbor graph (*k* = 20) was computed across these 30 batch-integrated PCs using the 'buildKNNGraph' function implemented in the Bioconductor package scan (v1.18.3), which was then used to group cells into connected communities using Louvain⁵⁷ clustering implemented in the R package 'igraph' (v1.2.6). Clusters that displayed mixed profiles of T and other lymphoid lineages, that is, CD19, CD20 and immunoglobulin genes, were classed as doublets and removed from downstream analyses. Clusters indicative of NK cells (CD3-CD56⁺) were subsequently annotated as such and removed from T cell analyses. Remaining clusters were annotated using a combination of canonical protein and mRNA markers for major αβ T cells (CD4, CD8, CCR7, CD45RA, CD45RO, CD62L, CD27, CD38, CD44, CXCR5, CD40LG, CCR7, FOXP3 and IKZF2), γδ T cells (Vγ9, Vγ2, TRGV9 and TRDV2) and invariant T cells, MAIT (Vα24-Jα18 and TRAV1.2) and NK T (CD3, CD16, CD56, NCAM1, NCR1 and FCGR3A) cells. Polarized CD4⁺ T cell annotations were refined using the combination of transcription factor genes and expressed cytokines for the respective T_H cell types: T_H1 (IFNG, TBX21 and TNFA), T_H2 (GATA3, IL4 and IL5) and T_H17 (RORC, IL17A, IL17F and IL21). Where clusters appeared heterogeneous in their expression of T cell lineage markers, single-cell annotations were refined based on the coexpression of specific marker gene and protein pairs. Dot plots to visualize marker protein and mRNA expression across clusters were generated using the R package 'ggplot2' (v3.3.3). UMAP³⁸ was used to project all single T cells into a two-dimensional space (*k* = 31) using the first 30 batch-integrated PCs as input with the R package 'umap' R v4.0.3 and Bioconductor v3.12 were used for all analyses.

T cell differential gene expression analysis. Differential gene expression analysis was performed across COVID-19 disease severity groups, ordered from healthy > asymptomatic > mild > moderate > severe > critical. Donor pseudo-bulk samples were first created by aggregating gene counts for each annotated T cell type, within each donor, where there were at least 20 cells of that type. Genes with fewer than three counts in any given pseudo-bulk sample, or fewer than five counts in total across donor pseudo-bulk samples, were removed before analysis. Differential gene expression testing was performed using an NB GLM implemented in the Bioconductor package edgeR^{59,60} (v3.32.1). Statistically significant differentially expressed genes were defined with FDR < 0.1. Functional annotation enrichment was performed using the Bioconductor package enrichR⁶¹ (v3.0). Upregulated and downregulated differentially expressed genes in each T cell type were used as input, testing separately against the MSigDB Hallmark 2020 and Transcription Factor Protein-Protein Interactions gene sets. Significant enrichments were defined with 1% FDR.

T cell receptor analysis. Single-cell TCRs were computed from the TCR-sequencing data using Cell Ranger v4.0.0. The unfiltered outputs of reconstructed TCR contigs across all three sites (Newcastle, Cambridge and London) were combined before filtering using: (1) full-length CDR3, (2) each cell droplet barcode matched a TCR droplet barcode and (3) productive CDR3

spanning V + J genes. Chain-specific TCR clones were defined for each observed α -chain and β -chain by first concatenating the V, J and identical CDR3 nucleotide sequences. For each single T cell, these chains were then combined to form a single clonotype, removing cells that contained: (1) greater than two β -chains and greater than two α -chains and (2) a single α -chain or a single β -chain only. T cells with exactly two β -chains and one α -chain, or those with exactly two α -chains and one β -chain were retained. TCR clonotypes were counted within each donor sample, and expanded clones were defined where more than one cell was assigned to the TCR clonotype.

The proportion of expanded clones as a function of a linear trend across disease severity groups was modeled using logistic regression, adjusted for age, gender and batch. A separate model was run for each T cell subtype that contained at least five cells assigned to the expanded TCR clonotypes. Linear trend *P* values were corrected for multiple testing using the Benjamini–Hochberg procedure⁶².

The TE:EM ratio was calculated within each donor, using the number of observed expanded clonotypes. The TE:EM ratio change across COVID-19 severity was tested using a robust linear model implemented in the R package ‘robustbase’ (v0.93-7), regressing the TE:EM ratio on either disease severity as an ordered linear variable (asymptomatic > mild > moderate > severe > critical) or symptom duration, adjusted for age, gender and batch. Statistical significance was defined based on the linear trend across disease severity ($P \leq 0.01$). An equivalent analysis was performed, restricted to participants with a shorter symptom duration (≤ 24 d).

Differential correlation analysis. Changes in the correlations between PBMC cell types were computed using a differential correlation analysis, implemented in the R package DCARS⁶³ (v0.3.5). Cell type proportions were computed by normalizing the counts of each cell type within each donor by the total number of cells captured for that donor sample. Donor samples were ranked according to their disease severity (healthy > asymptomatic > mild > moderate > severe > critical). Differential correlation analysis was then performed between CD4.T_{H1} versus all B cell types. Statistically significant differentially correlated cell types were defined with empirical *P* value ≤ 0.1 , estimated from 10,000 permutations.

BCR V(D)J analysis. Single-cell V(D)J data from the 5' Chromium 10x kit were initially processed with Cell Ranger v2 pipeline (4.0.0). BCR contigs contained in ‘filtered_contigs.fasta’ and ‘filtered_contig_annotations.csv’ from all three sites were then preprocessed using ‘immcantation’ inspired preprocessing pipeline⁶⁴ implemented in the dandelion Python package; dandelion is a new single-cell BCR-seq analysis package for 10x Chromium 5' data. All steps outlined below were performed using dandelion v0.0.27.post2 and are available at <https://github.com/clatworthy/dandelion/>.

BCR preprocessing. Individual BCR contigs were reannotated with ‘igblastn’ v1.1.15 using the IMGT reference database (downloaded on 30 June 2020)⁶⁵ by calling changeo’s ‘AssignGenes.py’ script, and reannotated contigs in ‘blast’ format were parsed into the Adaptive Immune Receptor Repertoire standards 1.3 format with changeo’s ‘MakeDB.py’ script. Amino acid sequence alignment information not present in the output from blast format was retrieved from reannotation with igblastn in AIRR format. Heavy-chain V-gene alleles were corrected for individual genotypes with TIGGER⁶⁶ (v1.0.0) using a modified ‘tigger-genotype.R’ script from ‘immcantation’ suite. Germline sequences were reconstructed based on the genotype corrected V-gene assignments using changeo’s (v1.0.1) ‘CreateGermlines.py’ script; contigs which failed germline sequence reconstruction were removed from further analysis. Constant genes were reannotated using blastn (v2.10.0+) with CH1 regions of constant gene sequences from IMGT followed by pairwise alignment against curated sequences to correct assignment errors due to insufficient length of constant regions.

BCR filtering. Contigs assigned to cells that passed QC on the transcriptome data were retained for further QC assessment, which included the following checks: (1) contigs with mismatched locus and V, J and constant gene assignments were removed from the analysis; (2) cell barcodes with multiple heavy-chain contigs were flagged for filtering (exceptions to this were (a) when the multiple heavy-chain contigs were assessed to have identical V(D)J sequences but assigned as different contigs belonging to the same cell by Cell Ranger v2 pipeline, (b) when there was a clear dominance (assessed by difference in UMI count) by a particular contig, and (c) if and when there was presence of one IgM and one IgD contig assigned to a single cell barcode; in the first two cases, the contig with the highest UMI count was retained); (3) cell barcodes with multiple light-chain contigs were flagged for filtering; and (4) in situations where cell barcodes were matched with only light-chain contigs, the contigs were dropped from the V(D)J data but the transcriptome barcode was retained.

B cell clone/clonotype definition. BCRs were grouped into clones/clonotypes based on the following sequential criteria that apply to both heavy-chain and light-chain contigs: (1) identical V and J gene usage, (2) identical junctional CDR3 amino acid length, and (3) at least 85% amino acid sequence similarity at the CDR3 junction (based on hamming distance). Light-chain pairing was performed using the same criteria within each heavy-chain clone. Only samples collected at day 0 of

the study were analyzed from this step onwards and clones/clonotypes were called across the entire dataset; the sample from one of the donors who was subsequently found to have a B cell malignancy was separated from the analysis and processed independently.

B cell clone/clonotype network. Single-cell BCR networks were constructed using adjacency matrices computed from pairwise Levenshtein distance of the full amino acid sequence alignment for BCR(s) contained in every pair of cells within each disease severity cohort. Construction of the Levenshtein distance matrices were performed separately for heavy-chain and light-chain contigs, and the sum of the total edit distance across all layers/matrices was used as the final adjacency matrix. To construct the BCR neighborhood graph, a minimum-spanning tree was constructed on the adjacency matrix for each clone/clonotype, creating a simple graph with edges indicating the shortest edit distance between a B cell and its nearest neighbor. Cells with identical BCRs, that is, cells with a total pairwise edit distance of zero, were then connected to the graph to recover edges trimmed off during the minimum-spanning-tree construction step. Fruchterman–Reingold graph layout was generated using a modified method to prevent singletons from flying out to infinity in ‘networkx’ (v2.5). Visualization of the resulting single-cell BCR network was achieved via transfer of the graph to relevant ‘anndata’ slots, allowing for access to plotting tools in scanpy.

The use of the BCR network properties for computing gini indices was inspired from bulk BCR-sequencing network analysis methods where distributions of clone sizes and vertex sizes (sum of identical BCR reads) in BCR clone networks were used to infer the relationships between BCR clonality, somatic hypermutation and diversity⁶⁷. However, there are challenges with native implementation of this approach for single-cell data. Firstly, to enable calculation of network-based clone/cluster and vertex/node size distribution, BCR networks needed to be reduced such that nodes/cells with identical BCRs had to be merged and counted; this required the reconstruction of BCR networks per sample and discarding single-cell-level information. Furthermore, the process of node contraction and counting of merging events requires substantial computation time and resource. Secondly, this approach is dependent on sufficient coverage of the BCR repertoire, as the BCRs from the number of cells sampled (after QC) may not necessarily recapitulate the entire repertoire, which may underrepresent or overrepresent merged counts for gini index calculation. We propose the use of node closeness centrality computed on each expanded clone (clone size > 1) as an alternative metric to emulate the statistics to adapt to the single-cell nature of the data; closeness centrality defines how close and central each node is with respect to other nodes in the graph; therefore, cells with identical BCRs will have high closeness centrality scores, due to the way the BCR network is constructed in dandelion. Thus, we can quickly calculate if cells across clones, and/or samples overall, in the entire graph display proportionately/disproportionately high or low closeness centrality scores. One caveat to the current implementation is that it is only meaningful if there are clonotypes with at least two cells, as scores will only be computed for non-singleton components of the graph. Gini indices are computed using ‘skbio.diversity.alpha.gini_index’ (scikit-bio v0.5.6) with the ‘trapezoids’ method after clone definition and network generation. Summary visualization was performed using plotting tools in ‘seaborn’ (v0.11.0).

Definition of BCR convergence across participants. BCR overlap was determined by collapsing sharing incidence of V and J gene usage and CDR3 amino acid sequences, in both heavy and light chains, between individuals into a binarized format (1 or 0). The information is turned into an adjacency matrix where an edge is created between two individuals if there is at least one clonotype (at least one cell from each individual displays an identical combination of heavy-chain and light-chain V and J gene usage with allowance for somatic hypermutation at the CDR3 junctional region) that is similar between the two individuals. Visualization was achieved using the ‘CircosPlot’ function from ‘nxviz’ package (v0.6.2).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The dataset from our study can be explored interactively through the web portal <https://covid19cellatlas.org/>. The data object, as a h5ad file, can also be downloaded from <https://covid19cellatlas.org/>. The processed data are available to download from Array Express under accession number E-MTAB-10026. Source data are provided with this paper.

Code availability

All data analysis scripts are available on <https://github.com/scCOVID-19/COVIDPBMC/>.

References

49. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

50. Heaton, H. et al. Souporecell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
51. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA-sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 (2019).
52. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
53. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).
54. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
55. Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene-set enrichment analysis. Preprint at *BioRxiv* <https://doi.org/10.1101/060012> (2019).
56. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
57. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
58. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at <https://arxiv.org/abs/1802.03426> (2018).
59. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
60. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
61. Kuleshov, M. V. et al. Enrichr: a comprehensive gene-set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
62. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).
63. Ghazanfar, S., Strbenac, D., Ormerod, J. T., Yang, J. Y. H. & Patrick, E. DCARS: differential correlation across ranked samples. *Bioinformatics* **35**, 823–829 (2019).
64. Gupta, N. T. et al. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).
65. Lefranc, M.-P. IMGT, the International ImMunoGeneTics Information System. *Novartis Found. Symp.* **254**, 126–142 <http://imgt.cines.fr> (2008).
66. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl Acad. Sci. USA* **112**, E862–E870 (2015).
67. Bashford-Rogers, R. J. M. et al. Network properties derived from deep sequencing of human B cell receptor repertoires delineate B cell populations. *Genome Res* **23**, 1874–1884 (2013).
68. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The human cell atlas bone marrow single-cell interactive web portal. *Exp. Hematol.* **68**, 51–61 (2018).

Acknowledgements

This publication is part of the Human Cell Atlas (<https://www.humancellatlas.org/publications/>). We acknowledge assistance from P. Vegh, J. Fletcher and D. Dixon to the IV-LPS study and funding from the Wellcome Human Cell Atlas Strategic Science Support (WT211276/Z/18/Z). We acknowledge B. Yeung and K. 'Kit' Nazor from BioLegend for helpful discussions in the optimization of the CITE-seq protocol. We acknowledge B. Stewart for his helpful contributions toward curation of the final dataset. We acknowledge R. Queen and R. Hussain from Newcastle University Genomics Core Facility for technical assistance. We acknowledge M. Glanville and the Newcastle University Infectious Diseases Facility for biosafety level 3 laboratory access and assistance. We acknowledge M. Prete for assistance with online data hosting and interactivity. We thank NIHR BioResource volunteers for their participation, and gratefully acknowledge NIHR BioResource centers, NHS Trusts and staff for their contribution. We thank the NIHR, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Programme. A.J.S. is an NIHR senior investigator; the views expressed in this article are those of the author(s) and not

necessarily those of the NHS, the NIHR or the Department of Health and Social Care. M.H. is funded by Wellcome (WT107931/Z/15/Z), The Lister Institute for Preventive Medicine and Newcastle NIHR Biomedical Research Centre. B.G. is funded by Wellcome (206328/Z/17/Z), MRC (MR/S036113/1), and the Aging Biology Foundation. S.A.T. is funded by Wellcome (WT206194), ERC Consolidator and EU MRG-Grammar awards. Z.K.T. and M.R.C. are supported by a Medical Research Council Human Cell Atlas Research Grant (MR/S035842/1). J.C.M. is supported by core funding from Cancer Research UK (C9545/A29580) and from the European Molecular Biology Laboratory. M.R.C. is supported by a Versus Arthritis Cure Challenge Research Grant (21777), and an NIHR Research Professorship (RP-2017-08-ST2-002). S.H. is funded by Wellcome (207556_Z_17_Z). H.W.K. is supported by a Sir Henry Wellcome Postdoctoral Fellowship (213555/Z/18/Z). K.B.M. acknowledges funding from the Chan Zuckerberg Initiative (2017-174169), Wellcome (WT211276/Z/18/Z) and Sanger Core (WT206194). M.Z.N. acknowledges funding from a UKRI Innovation/Rutherford Fund Fellowship allocated by the MRC and the UK Regenerative Medicine Platform (MR/5005579/1). M.Z.N. and K.B.M. have been funded by the Rosetrees Trust (M944). K.B.W. is funded by University College London, Birkbeck MRC Doctoral Training Programme. J.L.B. acknowledges funding from the MRC and the UK Regenerative Medicine Platform (MR/5005579/1). M.Y. is funded by The Jikei University School of Medicine. A.T.H. is funded by the NIHR (ACF-2018-01-004) and the BMA Foundation. K.F.B. is funded by an NIHR Clinical Lectureship (CL-2017-01-004). M.W.M. is funded by Action Medical Research (GN2779). N.M. is supported by a DFG Research Fellowship (ME 5209/1-1). C.J.A.D. is funded by Wellcome (211153/Z/18/Z) and acknowledges support from the Barbour Foundation. S.W. is funded by a Barbour Foundation PhD studentship. E.L. is supported by a Sir Henry Dale fellowship from Wellcome/Royal Society (107630/Z/15/Z), BBSRC (BB/P002293/1), and core support grants from Wellcome and MRC to the Wellcome-MRC Cambridge Stem Cell Institute (203151/Z/16/Z). C.M.S. is currently supported by grants from BBSRC (BB/V006738/1), GOSH Children's charity (COVID_CSsmith_017), Wellcome Trust (212516/Z/18/Z) and NIHR GOSH Biomedical research centre. This work was partly funded by UKRI/NIHR through the UK Coronavirus Immunology Consortium.

Author contributions

M.H., B.G., S.T., J.M. and M.R.C. conceived and directed this study. A.J.S. and A.J.R. conceived the IV-LPS study. M.Z.N. and K.M. conceived and led the UCL study. C.J.A.D., K.S., P.L., M.Z.N., F.M., E.K., A.d.W., A. Saigal, A. Saleh, S.M.J., A.T.H., K.F.B., I.C.D.S.v.d.L., L.C.S.G., A.S.B., A.S.-G. and L.B. organized sample acquisition, recruited participants, collected samples and clinical metadata. E.S., R.A.B., F.G., J.S., R.P.P., K.B.W., M.Y. and J.L.B. isolated PBMCs. R.A.B., E.S., K.B.W., M.Y., J.L.B., N.M., F.J.C.-N. performed 10x and CITE-seq analysis. E.S., J.E. and K.K. prepared sequencing libraries. E.P., J.C. and P.C. conducted the sequencing. G.R., Z.K.T., K.B., M.D.M., W.S., N.K., E.S., S.v.D., V.K., N.H., R.L., K.P. and E.D. analyzed the data. J.M.G., D.H., V.Y.K. and S.v.D. facilitated online data hosting. M.H., B.G., S.T., J.M., M.R.C., M.Z.N., K.M., G.R., Z.K.T., K.B., M.D.M., W.S., N.K., R.A.B., E.S., L.J., S.W. and J.S.S. interpreted the data. J.S.S. performed flow cytometry. E.S. performed multiplex cytokine analysis. M.H., B.G., S.T., M.R.C., J.M., E.S., G.R., L.J., B.O., M.D.M., K.B., N.K., Z.K.T. and W.S. wrote the manuscript. N.M., L.C.S.G., S.W., K.F.B., C.W., J.M.C., H.W.K., S.H., E.L., M.Z.N., K.M., K.B.M. and K.S. edited the manuscript. Z.K.T. and H.W.K. developed software (dandelion).

Competing interests

S.A.T. has received remunerations for consulting and Scientific Advisory Board work from Genentech, Biogen, Roche and GlaxoSmithKline, as well as Foresite Labs over the past 3 years. All other authors declare no competing interests.

Additional information

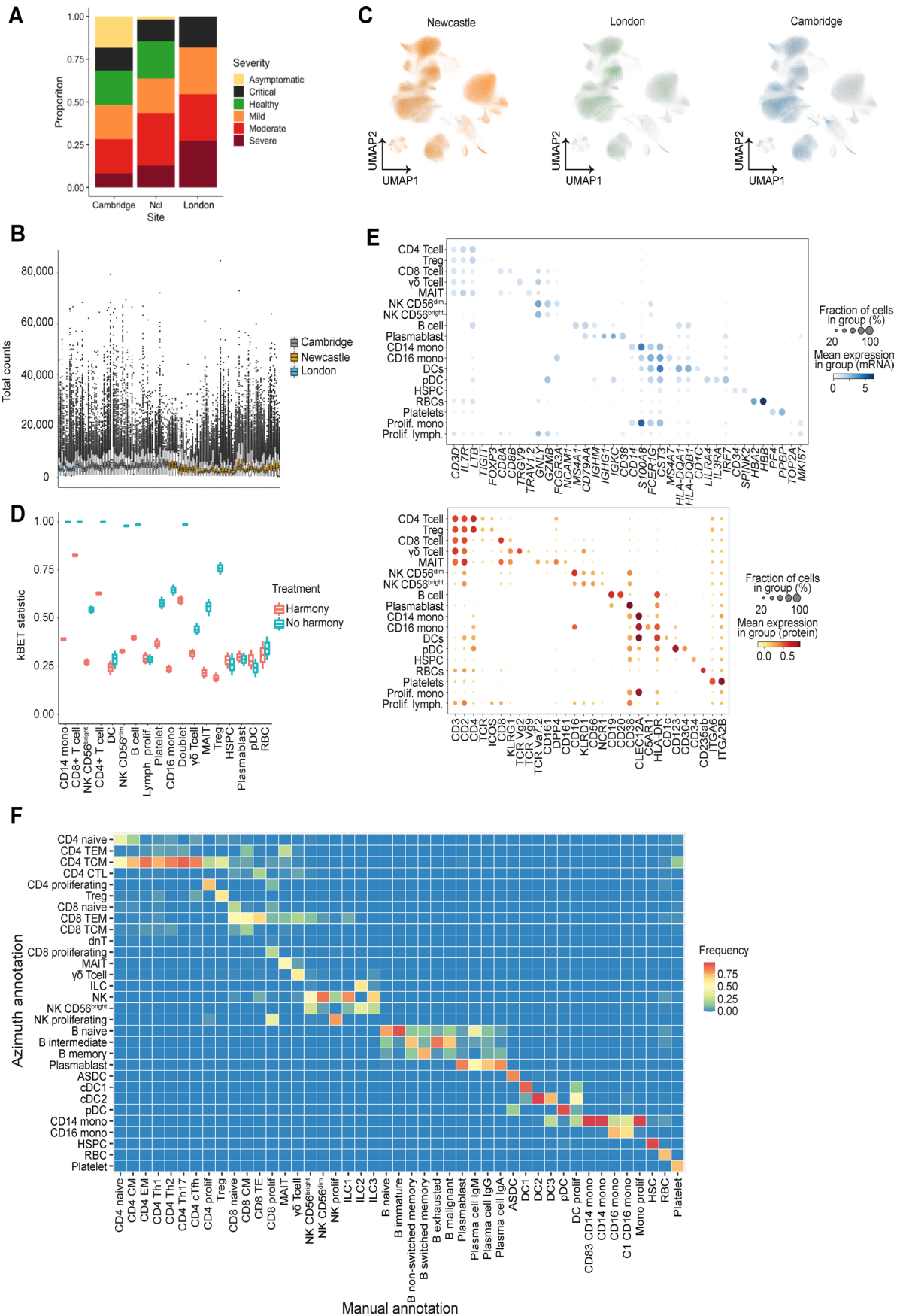
Extended data is available for this paper at <https://doi.org/10.1038/s41591-021-01329-2>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-021-01329-2>.

Correspondence and requests for materials should be addressed to S.A.T., M.R.C., J.C.M., B.G. or M.H.

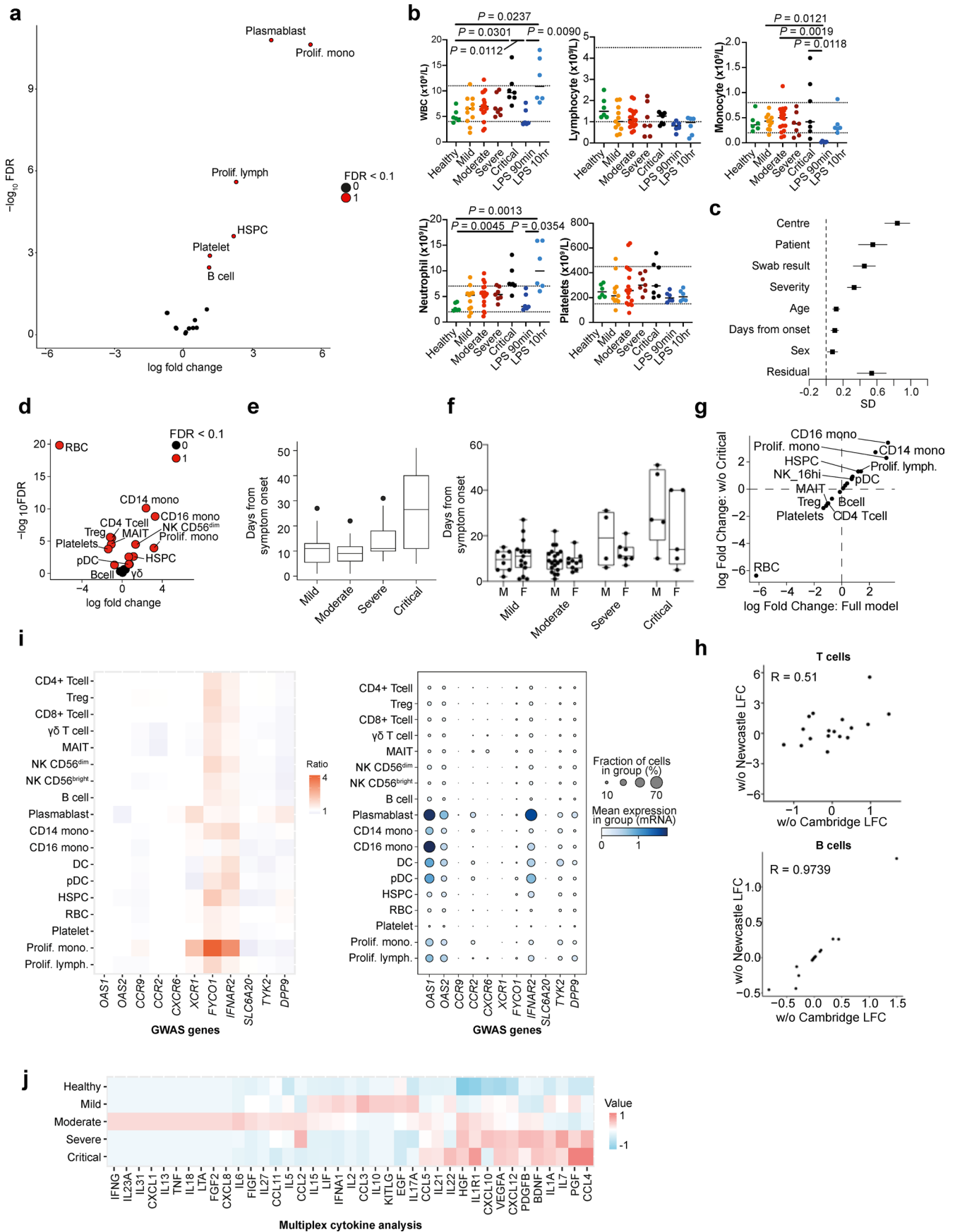
Peer review information *Nature Medicine* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Saheli Sadanand was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

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



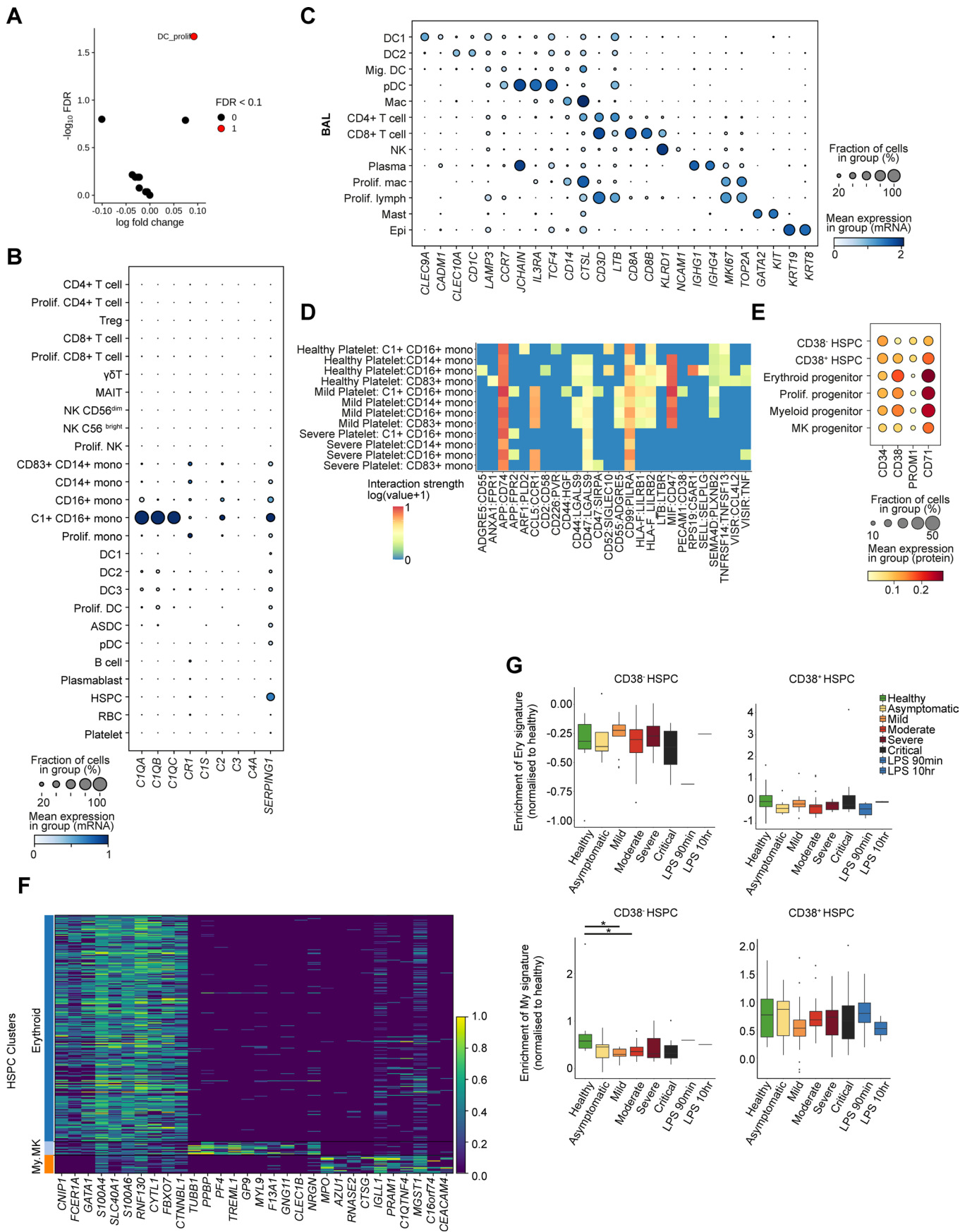
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Single-cell analysis quality control and cell type definition for COVID PBMC single cell analysis. **a**, Bar chart showing the composition of sample severities across the three sites (n = 23 mild, n = 30 moderate, n = 11, severe, n = 10 critical biologically independent samples). **b**, Scatter plot displaying the total number of gene counts per sample from each site. **c**, UMAP visualisations from Fig. 1b coloured by site. **d**, Boxplot of kBET results calculated both before and after batch correction with Harmony for each cluster in Fig. 1b kBET statistic calculation using patient ID as the batch factor (n = 130 biologically independent samples, n = 627,172 cells in 1 experiment). Hinges indicate to 25th and 75th percentile and whiskers to lowest and highest value in 1.5*interquartile range. **e**, Dot plots of 5' gene expression (top; blue) and surface protein (bottom; red) expression for populations shown in Fig. 1a where the colour is scaled by mean expression and the dot size is proportional to the percent of the population expressing the gene/protein, respectively. **f**, Tile plot showing percentage concordance between COVID-19 PBMC annotation (x-axis) and Azimuth annotation (y-axis) (<https://satijalab.org/azimuth/>).



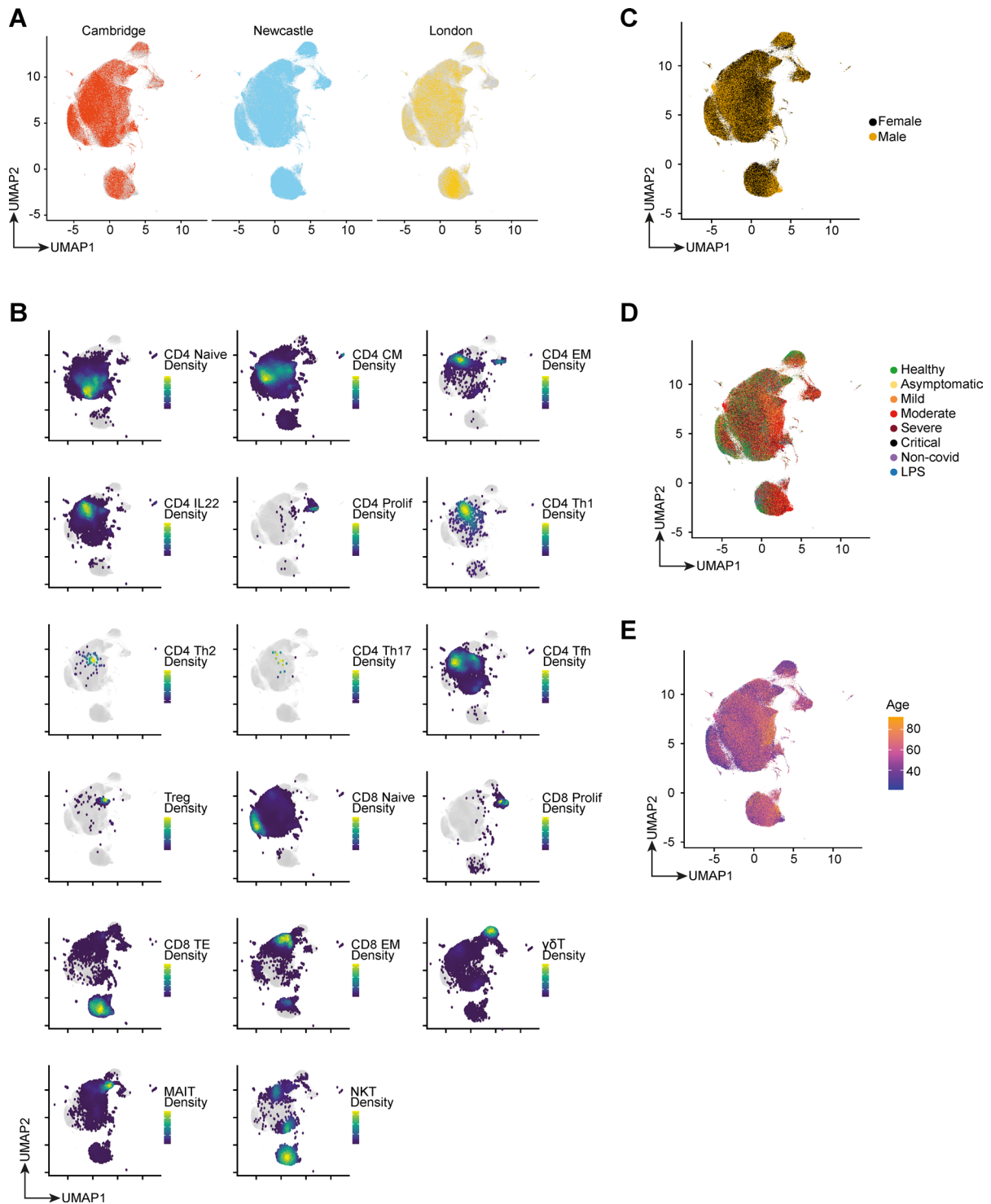
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Differential abundance analysis and expression of GWAS hits related to cytokines, chemokines and growth factors. **a**, Volcano plots showing differential abundance testing performed using quasi-likelihood F-test comparing healthy controls to cases for linear trends across disease severity groups. Differentially abundant cell types were determined using a 10% false discovery rate (FDR). **b**, Scatter plots showing blood counts for Newcastle data grouped by severity status (n = 6 healthy and LPS samples; n = 11 mild, n = 17 moderate, n = 7 severe, and n = 7 critical biologically independent samples). Dotted lines mark the normal ranges. Significance determined using Kruskal-Wallis with Dunn's post hoc corrected for multiple comparison. WBC, whole blood count. **c**, Forest plot showing the standard deviation of each clinical/technical factor estimated by the Poisson generalised linear mixed model. Error bars show the standard error estimated from the Fisher information matrix (n = 130 biologically independent samples) SD, standard deviation. **d**, Volcano plots showing differential abundance testing according to time since symptom onset. Differentially abundant (FDR 10%) points are shown in red. **e**, Box plots displaying the duration of COVID-19 symptoms from the onset by severity (n = 23 mild, n = 30 moderate, n = 11, severe, n = 10 critical biologically independent samples). Boxes denote IQR with median shown as horizontal bars. Whiskers extend to 1.5x the IQR; outliers are shown as individual points. **f**, Box plots displaying the duration of COVID-19 symptoms from the onset split by severity and sex (n = 23 mild, n = 30 moderate, n = 11, severe, n = 10 critical biologically independent samples). Boxes denote IQR with median shown as horizontal bars. Whiskers extend to 1.5x the IQR; outliers are shown as individual points. **g**, Correlated log fold-changes of cell type abundance changes as a function of symptom duration with (x-axis) and without critically ill patients (y-axis). **h**, Differential abundance testing with a leave-one-out analysis for the T cells (top) of B cells (bottom) (FDR10%). **i**, Heat map displaying fold change over healthy (left) and dot plot of gene expression where the colour is scaled by mean expression and the dot size is proportional to the percent of the population expressing the gene (right) for genes associated with COVID-19 identified in recent GWAS studies^{18,19} for the cell populations in Fig. 1b. **j**, Heat map displaying normalised values of cytokine, chemokine and growth factors in serum of patients with COVID-19.

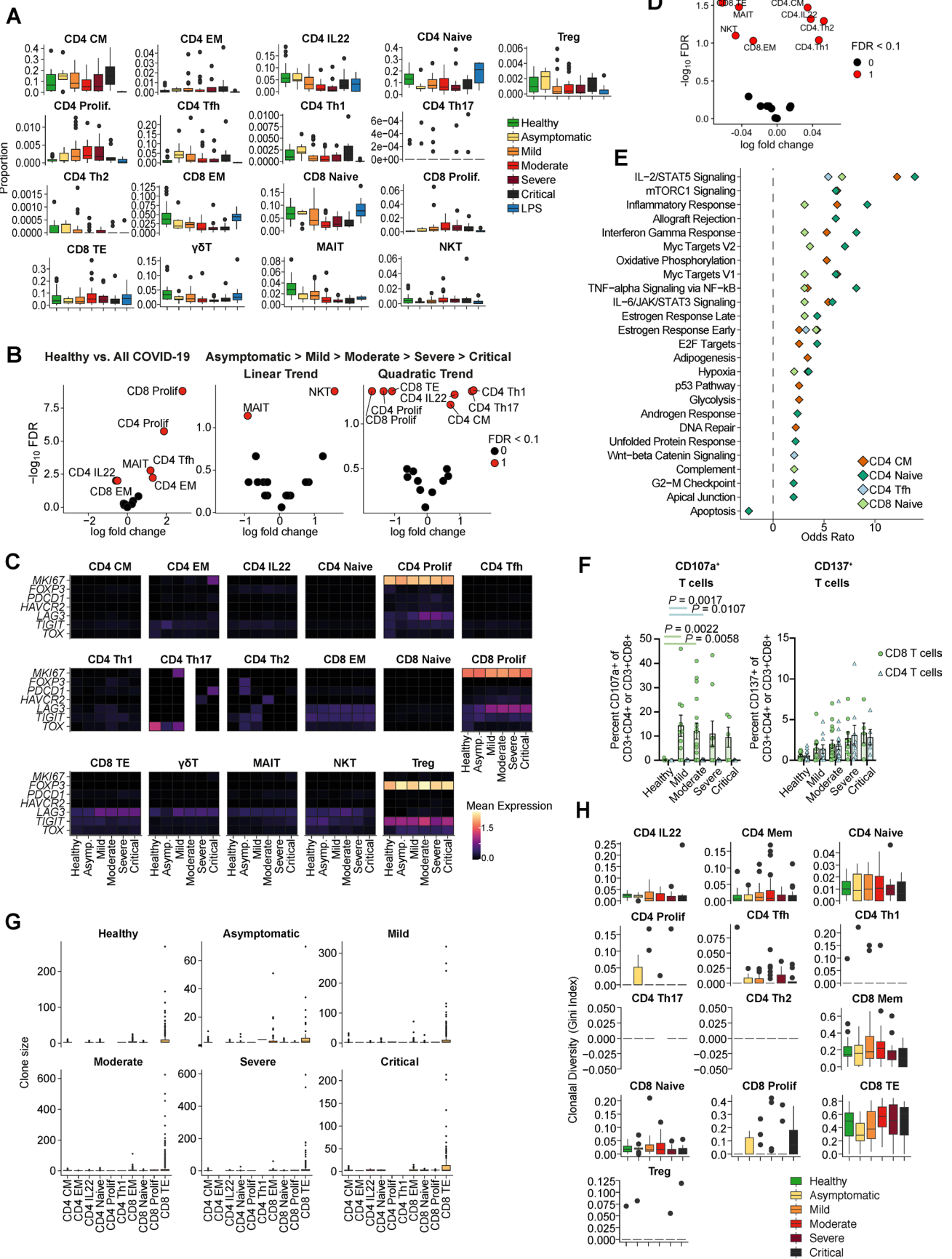


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Myeloid comparisons with bronchial alveolar lavage dataset and receptor-ligand interaction analysis between megakaryocyte, myeloid and progenitor cells. **a**, Volcano plot showing differential abundance testing according to time since symptom onset for the myeloid populations. Differentially abundant (FDR 10%) points are shown in red. **b**, Dot plots of gene expression of C1 complement components for cells in Fig. 1b where the colour is scaled by mean expression and the dot size is proportional to the percent of the population expressing the gene. **c**, Dot plots of gene expression of a recently published BAL dataset (accession number GSE145926²³) for genes in Fig. 2a where the colour is scaled by mean expression and the dot size is proportional to the percent of the population expressing the gene. **d**, Heat map of CellPhoneDB predicted ligand:receptor interactions between platelets and monocyte subsets, based on the protein data. **e**, Dot plots of expression of protein markers used to annotate clusters in Fig. 2i. MK, Megakaryocyte. **f**, Heatmap of differentially expressed genes between megakaryocyte, myeloid and erythroid progenitor clusters. MK, megakaryocyte; My, myeloid. **g**, Box plots displaying enrichment of an erythroid signature (top) and a myeloid signature (bottom) found in CD34⁻CD38⁻ (left) and CD34⁺CD38⁺ HSPCs (right), separated by severity. Asterisks above bars indicate significance and are coloured by the severity for which they were compared to. Absolute values are provided in Supplementary Tables 8 and 9. (n = 120 biologically independent samples, n = 3297 cells in 1 experiment). Boxes denote IQR with median shown as horizontal bars. Whiskers extend to 1.5x the IQR; outliers are shown as individual points (n = 24 healthy, n = 86 COVID-19 biologically independent samples) (p-values: Myeloid signature in CD38 negative HSPCs, Healthy vs. Mild: 0.8×10^{-3} , Healthy vs. Moderate 0.02).



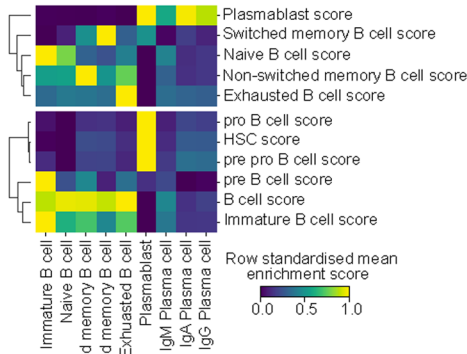
Extended Data Fig. 4 | UMAP embedding of T lymphoid clusters showing cell type estimation and covariates. a, UMAP visualisation of 309,617 T cells separated by sources of donors. **b**, UMAP visualisation showing 2-dimensional kernel density estimates of each T cell type in UMAP space. **c-e**, UMAP visualisation of T cells coloured by gender (**c**), disease severity status (**d**) and age (**e**).



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Differential abundance testing, gene set enrichment analysis and clonal diversity analysis of T lymphoid compartment. a, Box plots showing the proportion of cell types shown in Fig. 3a, $n=108$ biologically independent samples. Boxes denote IQR with median shown as horizontal bars. Whiskers extend to 1.5x the IQR; outliers are shown as individual points. **b,** Volcano plots showing results of differential abundance testing. Abundance counts were modelled either comparing healthy vs. COVID-19, or as a function of severity. Hypothesis testing was performed using quasi-likelihood F-test comparing healthy controls to cases, or for either a linear or quadratic trend across disease severity. Differentially abundant cell types were determined using a 10% false discovery rate (FDR). **c,** Heatmaps showing mean expression levels across T cell subsets for suppressive (*FOXP3*), proliferating (*MKI67*) and exhaustion markers (*PDCD1*, *HAVCR2*, *LAG3*, *TIGIT*, *TOX*). Columns denote the mean log-normalised expression within each severity category and healthy controls. **d,** Volcano plot showing differential abundance testing according to time since symptom onset for the T cell populations. Differentially abundant (FDR 10%) points are shown in red. **e,** Gene set enrichment (Methods) in each T cell type based on differential gene expression (DGE) analysis was performed across COVID-19 disease severity groups, ordered from healthy > asymptomatic > mild > moderate > severe > critical. Statistically significant DE genes were defined with FDR < 1%. Significant enrichments were defined with 10% FDR. **f,** Bar plots showing percent (mean \pm SEM) of CD3⁺CD4⁺ (blue) and CD3⁺CD8⁺ (green) T cells expressing CD107a (left) and CD137 (right) in response to SARS-CoV-2 S peptide stimulation ($n=12$ healthy, $n=10$ mild, $n=17$ moderate, $n=9$ severe, and $n=5$ critical biologically independent samples). Significance determined using Kruskal-Wallis with Dunn's post-hoc corrected for multiple comparison. **g,** Box plots showing clone size distribution for each T cell subset ($n=9$ asymptomatic, $n=19$ mild, $n=29$ moderate, $n=12$ severe, $n=10$ critical biologically independent samples). Boxes denote IQR with median shown as horizontal bars. Whiskers extend to 1.5x the IQR; outliers are shown as individual points. **h,** Box plots showing clonal diversity for each T cell subset ($n=22$ healthy, $n=12$ asymptomatic, $n=22$ mild, $n=31$ moderate, $n=14$ severe, $n=13$ critical biologically independent samples). Boxes denote IQR with median shown as horizontal bars. Whiskers extend to 1.5x the IQR; outliers are shown as individual points.

A



B

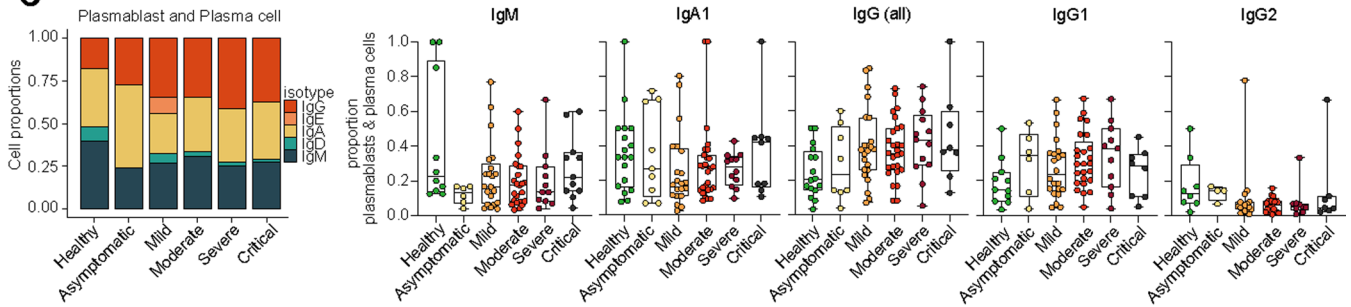
Kruskal-Wallis Test

	Plasmablast			Plasma cell			B_exhausted	B_memory		B_naive	B_immature
	IgM	IgA	IgG	IgM	IgA	IgG		switched	non-switched		
Healthy vs. Asymptomatic	0.6411	0.2874	0.8789	0.1813	0.8110	0.5073	0.9828	0.7875	0.2066		
Healthy vs. Mild	0.0017	0.1589	0.0183	0.0810	0.7494	0.0084	0.6988	0.7875	0.1111		
Healthy vs. Moderate	<0.0001	0.0693	0.0026	0.0058	0.0585	0.0104	0.9828	0.1524	0.0023		
Healthy vs. Severe	0.0001	0.0322	0.0026	0.0008	0.8468	0.1391	0.9082	0.1524	0.0190		
Healthy vs. Critical	0.0017	0.2545	0.0563	0.1421	0.1180	0.0151	0.9082	0.6336	0.0675		
Healthy vs. Non_covid	0.8696	0.6207	0.8789	0.4733	0.0741	0.9568	0.2338	0.1524	0.3094		
Healthy vs. LPS	0.7344	0.9891	0.8789	0.6552	0.7656	0.8126	0.9082	0.7875	0.6402		
Asymptomatic vs. LPS	0.9319	0.3069	0.8789	0.4758	0.7494	0.8410	0.9828	0.9659	0.2066		
Mild vs. LPS	0.2109	0.2545	0.0563	0.4758	0.7053	0.2155	0.9828	0.7679	0.1516		
Moderate vs. LPS	0.0528	0.1662	0.0183	0.2374	0.1070	0.2518	0.9082	0.1524	0.0206		
Severe vs. LPS	0.0528	0.0693	0.0158	0.0810	0.7494	0.5073	0.9828	0.1854	0.0374		
Critical vs. LPS	0.1167	0.3069	0.0960	0.4758	0.1647	0.2155	0.9828	0.6336	0.0952		
Non_covid vs. LPS	0.7344	0.6207	0.8789	0.6552	0.0835	0.8126	0.2338	0.1524	0.2585		

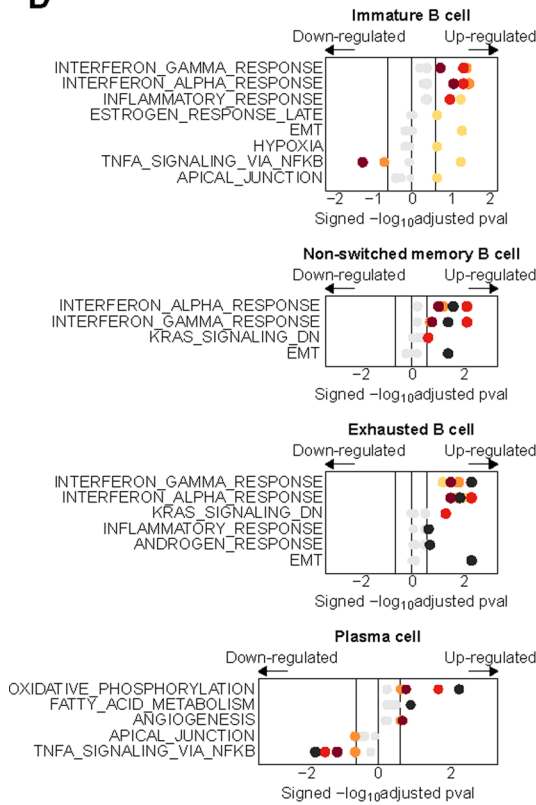
Differential abundance test (asymptomatic > mild > moderate > severe > critical)

CellType	Linear changes					Quadratic changes				
	logFC	logCPM	F	PValue	FDR	logFC	logCPM	F	PValue	FDR
B_exhausted	-0.022	15.476	2.586	0.111	0.167	0.506	15.476	2.137	0.147	0.223
B_immature	-0.010	16.091	0.729	0.396	0.396	-0.036	16.091	0.016	0.900	0.900
B_naive	0.005	19.179	1.274	0.262	0.337	-0.158	19.179	1.884	0.173	0.223
B_non-switched_memory	0.016	15.539	2.738	0.102	0.167	-0.258	15.539	0.955	0.331	0.373
B_switched_memory	0.020	16.571	4.427	0.038	0.115	-0.457	16.571	3.324	0.072	0.161
Plasma_cell_IgA	-0.036	15.441	6.413	0.013	0.059	0.500	15.441	2.006	0.160	0.223
Plasma_cell_IgG	-0.016	15.622	1.044	0.310	0.348	0.803	15.622	4.254	0.042	0.127
Plasma_cell_IgM	-0.028	14.211	3.425	0.068	0.152	1.577	14.211	14.748	0.002	0.002
Plasmablast	-0.051	15.670	6.870	0.010	0.059	1.536	15.670	10.682	0.002	0.007

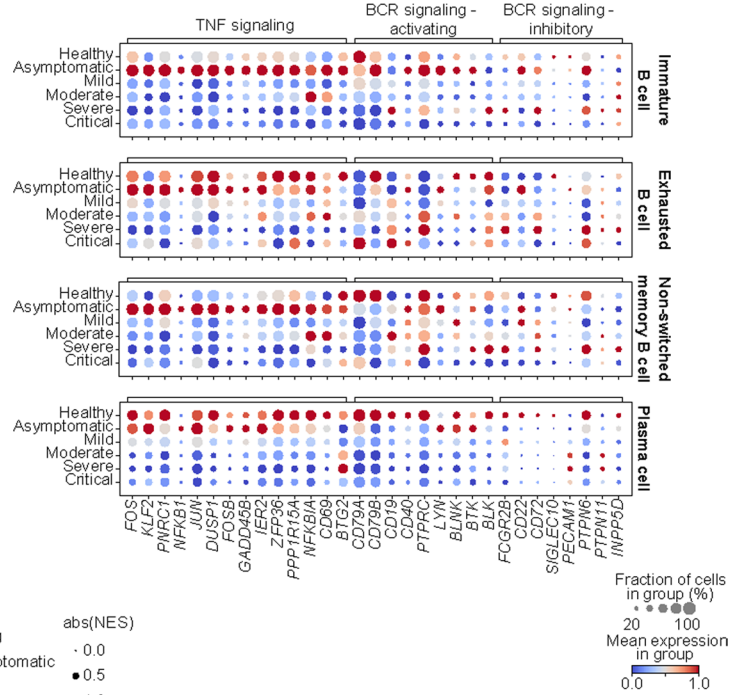
C



D

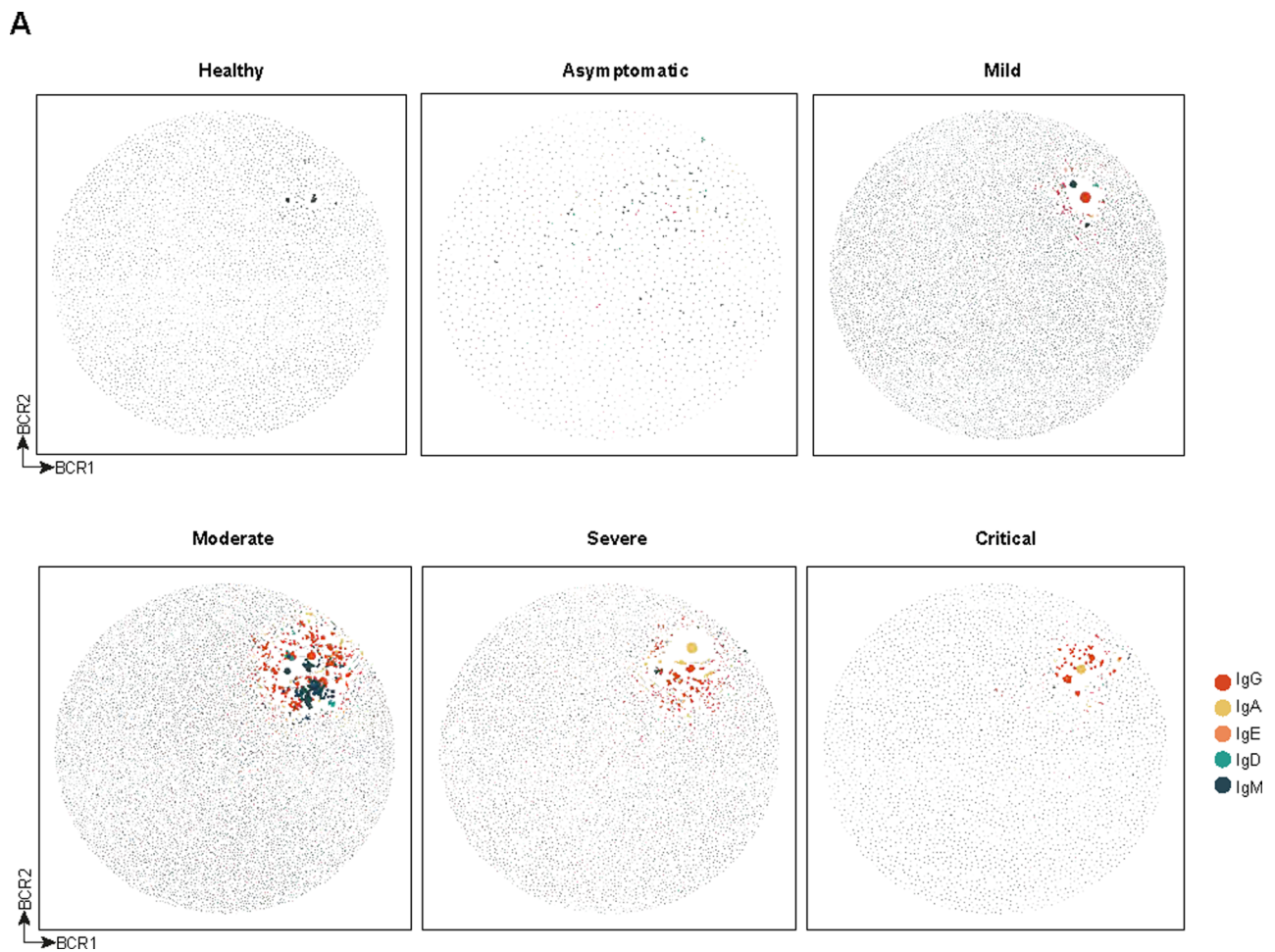


E

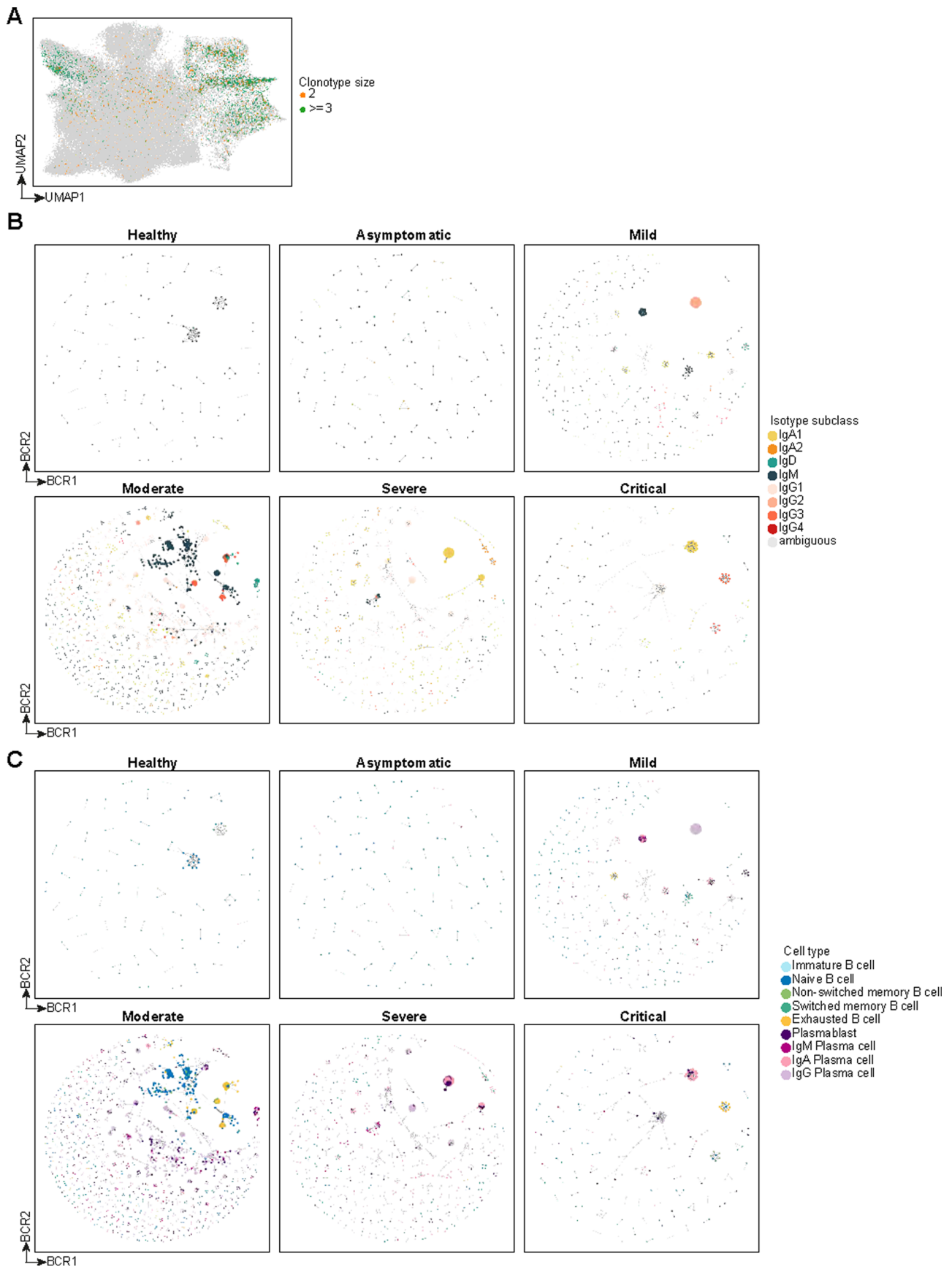


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Single-cell analysis of B lymphocytes proportion and gene set differences across COVID-19 severity. **a**, Heatmap of mean gene set enrichment scores of (top) adult peripheral blood B cell signatures³² and (bottom) bone marrow B cell signatures⁶⁸. Row enrichment value is scaled from 0-1 and colour gradient corresponds to increasing mean enrichment score. **b**, (Top) Kruskal-Wallis test results with Benjamini-Hochberg false discovery correction for cell type proportion differences in plasmablast/plasma cells between severities. Significance is denoted by Benjamini-Hochberg corrected $P < 0.05$ (red text). (Bottom) Cell type abundance counts were modelled as a function of severity. Hypothesis testing was performed using quasi-likelihood F-test comparing asymptomatic to symptomatic covid, for either a linear or quadratic trend across severities. Differentially abundant cell types were determined using a 10% false discovery rate (FDR). **c**, (Left) Bar plot showing the mean proportion of plasmablast/plasma cells expressing IgA, IgD, IgE, IgG or IgM, based on V(D)J information. (Right) Proportion of IgM, IgA1, total IgG, IgG1, and IgG2 subclass in plasmablast and plasma cells. Kruskal-Wallis test results with Benjamini-Hochberg false discovery correction for cell type proportion differences in plasmablast/plasma cells between severities showed no significant differences. $n = 10, 6, 21, 24, 11, 11$ (IgM), $n = 18, 9, 21, 27, 12, 9$ (IgA1), $n = 16, 8, 23, 27, 13, 8$ (IgG (all)), $n = 11, 6, 22, 25, 11, 8$ (IgG1) and $n = 8, 4, 16, 23, 9, 7$ (IgG2) biologically independent patients/samples for healthy, asymptomatic, mild, moderate, severe and critical respectively. **d**, GSEA of pathways in B cell subsets for asymptomatic/symptomatic COVID versus healthy. Size of circles indicate (absolute) normalised enrichment score (NES). Pathways were considered statistically significant if GSEA (permutation) nominal $P < 0.05$ and FDR < 0.25 (denoted by non-grey coloured dots; boundary lines in the middle marks FDR = 0.25). EMT, Epithelial-mesenchymal transition. **e**, Dot plots of TNF signalling molecules, activating and inhibitory BCR signaling molecules (5' gene expression data) in immature B cells, non-switched memory B cells, 'exhausted' B cells and plasma cells. Size of circles indicate percent of cells expressing the gene and colour gradient corresponds to increasing mean expression value.



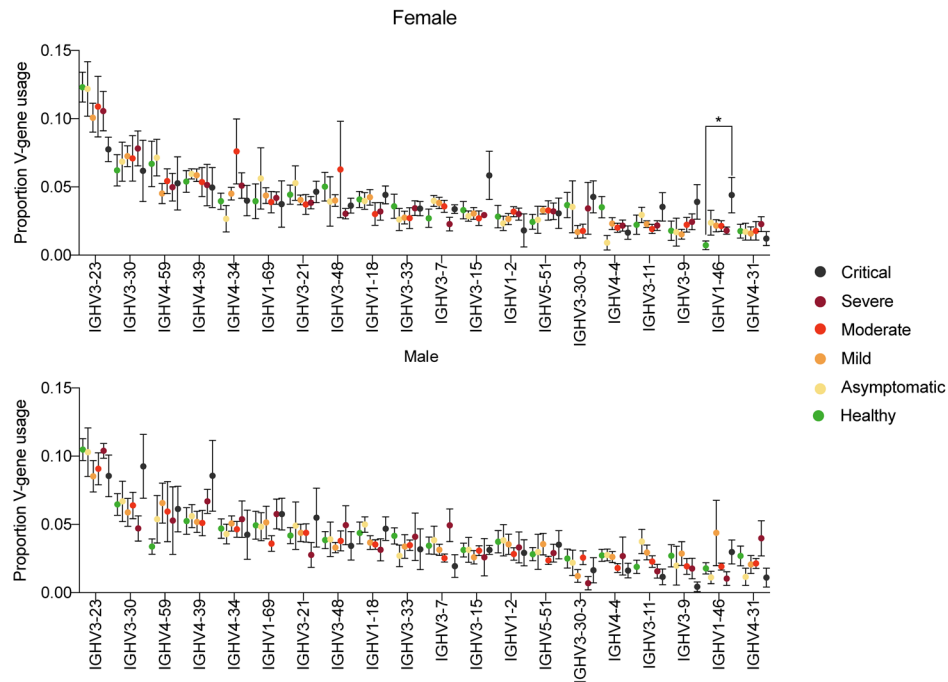
Extended Data Fig. 7 | Single-cell BCR networks across COVID-19 severity. Single-cell BCR network plots for each severity status coloured by heavy chain isotype class (IgM, IgD, IgA, IgE, or IgG). Each circle/node corresponds to a single B cell with a corresponding set of BCR(s). Each clonotype is presented as a minimally connected graph with edge widths scaled to $1/d+1$ for edge weight d where d corresponds to the total (Levenshtein) edit distance of BCRs between two cells. Size of nodes is scaled according to increasing node closeness centrality scores that is nodes that are highly central to a clonotype network will be larger.



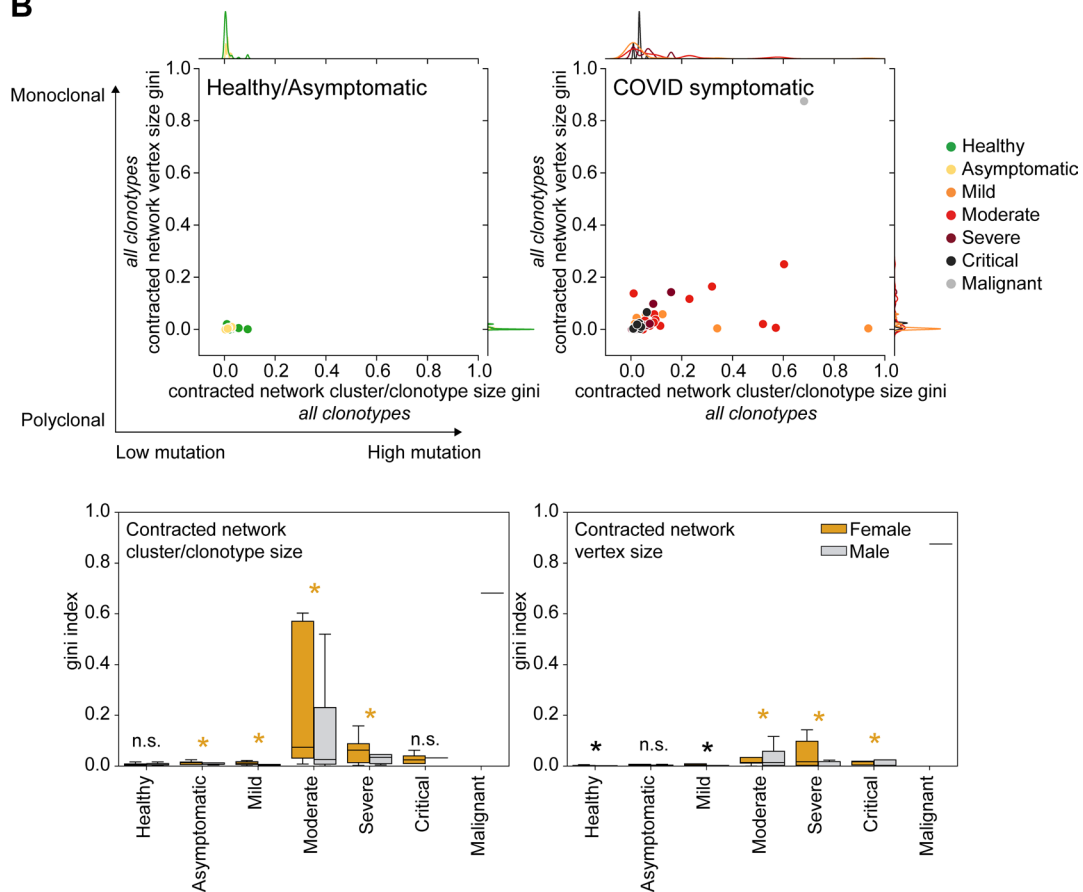
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Single-cell BCR clonotype expansion analysis across COVID-19 severity. **a**, UMAP visualisation of B cell lineage and coloured by clonotype size in the V(D)J data. Only expanded clonotypes are coloured (clonotype size > 2). **b**, Single-cell BCR network plots for each severity status coloured by assigned cell type. **c**, Single-cell BCR network plots for each severity status coloured by heavy chain isotype subclass (IgM, IgD, IgA1, IgA2, IgE, IgG1, IgG2, IgG3 or IgG4). Each circle/node corresponds to a single B cell with a corresponding set of BCR(s). Each clonotype is presented as a minimally connected graph with edge widths scaled to $1/d + 1$ for edge weight d where d corresponds to the total (Levenshtein) edit distance of BCRs between two cells. Size of nodes is scaled according to increasing node closeness centrality scores that is nodes that are highly central to a clonotype network will be larger.

A



B



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Single-cell BCR analysis between genders across COVID-19 severity. a. Summary plot for BCR IGHV gene usage split by gender (top: female; bottom: male). Each data point is presented as the mean gene usage proportion \pm standard error of the mean for samples within each sample group. Kruskal-Wallis test with Benjamini-Hochberg $*P < 0.05$ (adjusted $p = 0.035$) $n = 12, 5, 8, 20, 6$ and 6 biologically independent male patients/samples and $n = 12, 4, 15, 10, 7$ and 5 biologically independent female patients/samples for healthy, asymptomatic, mild, moderate, severe and critical respectively. **b.** (Top) Scatter plot of clonotype/cluster size by vertex size gini indices computed from contracted BCR networks (identical nodes are merged and counted). Each dot represents the gini indices of an individual coloured by severity status. Marginal histograms indicate the distribution of samples in a given severity status along the axes. (Bottom, left) Cluster/clonotype size (contracted network) gini indices separated by gender. (Bottom, right) Vertex size (contracted network) gini indices separated by gender. The boxes extend from the 25th to 75th percentiles. The whiskers go down to the smallest value and up to the largest. The line in the middle of the boxes is plotted at the median. Statistical tests were performed with non-parametric Mann-Whitney U test between the gender groups within each severity status and were considered statistically significant if Benjamini-Hochberg corrected $P < 0.05$ (denoted by *; n.s. denotes not significant). The Benjamini-Hochberg adjusted p-values are as follows: cluster/clonotype size (contracted network) gini indices comparisons: 3.42×10^{-1} , 1.91×10^{-25} , 7.71×10^{-272} , 1.31×10^{-169} , 3.84×10^{-113} and 2.98×10^{-1} ; vertex size (contracted network) gini indices comparisons: 1.12×10^{-57} , 2.17×10^{-1} , $< 1 \times 10^{-308}$, 3.64×10^{-49} , 2.69×10^{-53} and 2.37×10^{-04} - for healthy, asymptomatic, mild, moderate, severe and critical respectively. Colour of asterisks indicates which gender group displays a higher mean gini index (yellow: female; grey: male). $n = 12, 5, 8, 20, 6$ and 6 biologically independent male patients/samples and $n = 12, 4, 15, 10, 7$ and 5 biologically independent female patients/samples for healthy, asymptomatic, mild, moderate, severe and critical respectively. $n = 1$ for malignant (male).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used for data collection.

Data analysis

All data analysis scripts are available on <https://github.com/scCOVID-19/COVIDPBMC>

Alignment and quantification

Droplet libraries were processed using Cellranger v4.0. Reads were aligned to the GRCh38 human genome concatenated to the SARS-Cov-2 genome (NCBI SARS-CoV-2 isolate Wuhan-Hu-1) using STAR50 (v2.5.1b) and unique molecular identifiers (UMIs) deduplicated. CITE-seq UMIs were counted for GEX and ADT libraries simultaneously to generate feature X droplet UMI count matrices.

Doublet identification

Newcastle:

Scrublet (v0.2.1) was applied to each sample to generate a doublet score. These formed a bimodal distribution so the tool's automatic threshold was applied.

Cambridge:

Non-empty droplets were called within each multiplexed pool of donors using the emptyDrops function implemented in the Bioconductor package DropletUtils (v1.10.3), using a UMI threshold of 100 and FDR of 1%. The probability of being a doublet was estimated for each cell per sample (that is one 10x lane) using the "doubletCells" function in scran based on highly variable genes (HVGs). Next, we used "cluster_walktrap" on the SNN-Graph that was computed on HVGs to form highly resolved clusters per sample. Per-sample clusters with either a median doublet score greater than the median + 2.5 x MAD or clusters containing more than the median + 2.5 MAD genotype doublets were tagged as doublets. This was followed by a second round of highly-resolved clustering across the whole data set, in which again cells belonging to clusters with a high proportion (> 60%) of cells previously labelled as doublets were also defined as doublets.

London:

For pooled donor CITE-seq samples, the donor ID of each cell was determined by genotype-based demultiplexing using souporcell version 251. Souporcell analyses were performed with 'skip_remap' enabled and a set of known donor genotypes given under the 'common_variants' parameter. The donor ID of each souporcell genotype cluster was annotated by comparing each souporcell genotype to the set of known genotypes. Droplets that contained more than one genotype according to souporcell were flagged as 'ground-truth' doublets for heterotypic doublet identification. Ground-truth doublets were used by DoubletFinder 2.0.352 to empirically determine an optimal 'pK' value for doublet detection. DoubletFinder analysis was performed on each sample separately using 10 principal components, a 'pN' value of 0.25, and the 'nExp' parameter estimated from the fraction of ground-truth doublets and the number of pooled donors.

CITE-seq background signal removal

Background antibody- and non-specific staining was subtracted from ADT counts in each data set from the 3 data acquisition sites separately. ADT counts for each protein were first normalised using counts per million (CPM) and log transformed, with a +1 pseudocount. To estimate the background signal for each protein, a 2-component gaussian mixture model (GMM), implemented in the mclust (v5.4.7) R package function Mclust, was fit across the droplets with a total UMI count > 10 and < 100 from each experimental sample separately. The mean of the first GMM component for each protein was then subtracted from the log CPM from the QC-passed droplets in the respective experimental sample.

Quality control, normalisation, embedding and clustering

Combined raw data from the three centres was filtered to remove those that expressed fewer than 200 genes and >10% mitochondrial reads. Data was normalised (scanpy: normalize_total), log+1 corrected (scanpy: log1p) and highly variable genes identified using the Seurat vst algorithm (scanpy: highly_variable_genes). Harmony was used to adjust principal components by sample ID and used to generate the neighbourhood graph and embedded using UMAP. Clustering was performed using the Leiden algorithm with an initial resolution of 3. For initial clustering, differentially expressed genes were calculated using Wilcoxon rank-sum test.

Relative importance of metadata on cell type composition

The number of cells for each sample (N=110 samples in total with complete metadata) and cell type (18 different cell types in total) combination was modelled with a generalised linear mixed model with a Poisson outcome (see Supplementary File 2 for detail). The 5 clinical factors (COVID-19 swab result, age, sex, disease severity at day 0 and days from onset) and the 2 technical factors (patient and sequencing centre) were fitted as random effects to overcome the collinearity among the factors. The effect of each clinical/technical factor on cell type composition was estimated by the interaction term with the cell type. The likelihood ratio test was performed to assess the statistical significance of each factor on cell type abundance by removing one interaction term from the full model at a time. The number of factors was used to adjust multiple testing with the Bonferroni approach. The 'glmer' function in the lme4 package implemented on R was used to fit the model. The standard error of variance parameter for each factor was estimated using the numDeriv package.

Cydar Analysis

We utilized cydar to identify changes in cell composition across the different severity groups based on the protein data alone. First, the background-corrected protein counts from the three different sites were integrated using the 'fastMNN' method ($k = 20$, $d = 50$, $\text{cos.norm} = \text{TRUE}$) in batchelor (v1.6.2). The batch-corrected counts for 188 proteins (4 rat/mouse antibody isotypes were removed) were then used to construct hyperspheres using the 'countCells' function (downsample = 7) with the tolerance parameter chosen so that each hypersphere has at least 20 cells which was estimated using the 'neighborDistances' function. To assess whether the abundance of cells in each hypersphere are associated with disease status, hypersphere counts were analyzed using the quasi-likelihood (QL) method in edgeR (v3.32.1). After filtering out hyperspheres with an average count per sample below 5 we fitted a mean-dependent trend to the NB dispersion estimates. The trended dispersion for each hypersphere was used to fit a NB GLM using the log-transformed total number of cells as the offset for each sample and blocking for sex, age and batch. The QL F-test was used to compute P values for each hypersphere which were corrected for multiple testing using the spatial FDR method in cydar.

Comparisons of PBMC annotation using the Azimuth tool

The final annotation of PBMCs was compared to a published PBMC annotation using the Azimuth tool (<http://azimuth.satijalab.org/app/azimuth>). Because of size restrictions of 100,000 cells, our data was subsampled to 10% of the total cells. After running the algorithm, results with a prediction score < 0.5 were removed (5.8% of total removed). For each cluster in the COVID-19 PBMC data, the percentage of cells mapped to each cluster in the Azimuth annotation was calculated.

Interferon, TNF and JAK-STAT response scoring

A list of genes related to response to type I interferons was obtained from the GSEA Molecular Signatures Database (MSigDB) (GO: 0034340). Enrichment of the interferon score was measured using the tl.score_genes tool in scanpy which subtracts the average expression of all genes in the dataset from the average expression of the genes in this list. The scores were averaged across clusters and clinical status and expressed as a fold-change over the interferon score in the equivalent healthy cluster.

kBET analysis

The kBET53 algorithm (<https://github.com/theislab/kBET>) was run for each cluster defined in Fig. 1 using the Uniform manifold and projection (UMAP) coordinates generated from Harmony-adjusted principal components, and the sample number as the batch factor. The same procedure was then performed using the same annotation but using the UMAP coordinates generated from non-Harmony-adjusted principal components. The resultant rejection rates were averaged across clusters and compared using a Wilcoxon paired signed rank test.

Bronchoalveolar lavage data analysis

ScRNAseq data from BAL was obtained from GEO (accession number GSE14592622). Raw data was analysed using the same pipeline as PBMC data, specifically using the same quality control cut-offs (min of 200 genes and <10% mitochondrial reads/cell) and batch-corrected using Harmony by donor ID. To gain greater resolution of mononuclear phagocytes the DC and macrophages were analysed with further rounds of sub-clustering to identify DC1, DC2 and mature DC.

PAGA analysis of blood monocytes and BAL macrophages

Annotated raw expression datasets of BAL macrophages and COVID-19 PBMCs were merged and data log-normalised and scaled as for the original datasets. The top 3000 highly variable genes were chosen using the Seurat "vst" method and used for downstream analysis. Principal components were batch corrected by donor and used to build a neighborhood graph. The PAGA tool in scanpy (tl.paga) was used to generate the abstracted graph between clusters.

CellphoneDB

CellphoneDB54 was used to assess putative interactions between monocytes (CD14_mono, CD83_CD14_mono, C1_CD16_mono, CD16_mono, Prolif_mono) and platelets. The tool was run for 100 iterations and an expression threshold of 0.25 (limiting the analysis to

genes expressed by 25% of cells). For downstream analysis we focused on interactions between platelets and any monocyte subset.

Platelet activation

Differentially expressed genes in platelets between healthy controls and COVID-19 were filtered for those predicted to be involved in platelet activation (https://www.gsea-msigdb.org/gsea/msigdb/cards/REACTOME_PLATELET_ACTIVATION_SIGNALING_AND_AGGREGATION)

HPSC commitment scoring

HPSCs were subsetted from the data and Leiden clusters generated using the same pipeline and parameters as for the whole PBMC dataset. Differentially expressed genes between the HSPC clusters that showed evidence of lineage commitment (MK, Erythroid and Myeloid) were calculated using FindAllMarkers tool in Seurat (with thresholds of genes expressed by 25% of cells and with a log fold-change of 0.25) and genes with an adjusted p-value cut-off of 0.05 were used to generate gene signatures for each. Enrichment of these signatures in the CD38 negative and CD38 positive HSPC clusters were calculated using the `tl.score_genes` in scanpy (v1.6.0). The average expression of these enrichment scores in the CD38 negative and CD38 positive HSPC clusters was averaged by donor then compared across clinical states. Differences between groups were assessed using ANOVA with pairwise comparisons using Tukey's test.

Multiplex cytokine analysis

The values of each analyte were detected using the MAGPIX® system and analysed using the ProcartaPlex Analyst version 1.0 Software (ThermoFisher Scientific).

Flow Cytometry of stimulated cells

Data were analysed by FlowJo V10 (BD Biosciences).

GSEA analysis

Pre-ranked gene set analysis (prGSEA) on MSigDB v7.2 Hallmark genesets55 was performed using pre-ranked gene lists with `fgsea56` in R. Genes were pre-ranked according to signed $-\log_{10}$ P-values for all prGSEA procedures. For B cells, generation of rank gene list was performed using Wilcoxon rank sum test (via `tl.rank_genes_groups` in scanpy) with each Day 0 COVID statuses (asymptomatic to symptomatic critical) as the "tests" versus Day 0 Healthy samples as "reference/control".

T cell clustering, annotation and visualisation

Droplets labelled as T cells ("CD4", "CD8", "Treg", "MAIT", "gdT") were subset from those in Fig. 1B and re-clustered using a set of HVGs calculated within each batch, the union of which were used to estimate the first 50 principal components across cells using the `irlba` R package (v2.3.3). Batch effects were removed across the first 30 PCs using the `fastMNN57` implementation in the Bioconductor package `batchelor` (v1.6.2) ($k=50$). A k-nearest neighbour graph ($k=20$) was computed across these 30 batch-integrated PCs using the `buildKNNGraph` function implemented in the Bioconductor package `scrn` (v1.18.3), which was then used to group cells into connected communities using Louvain58 clustering implemented in the R package `igraph` (v1.2.6). Clusters that displayed mixed profiles of T and other lymphoid lineages, i.e. CD19, CD20 and immunoglobulin genes, were classed as doublets and removed from down-stream analyses. Clusters indicative of NK cells (CD3-CD56+) were subsequently annotated as such and removed from T cell analyses. Remaining clusters were annotated using a combination of canonical protein & mRNA (italicised) markers for major $\alpha\beta$ T cells (CD4, CD8, CCR7, CD45RA, CD45RO, CD62L, CD27, CD38, CD44, CXCR5, CD40LG, CCR7, FOXP3, IKZF2), $\gamma\delta$ T cells (V γ 9, V γ 2, TRGV9, TRDV2) and invariant T cells; MAIT (V α 24-J α 18, TRAV1.2), NKT (CD3, CD16, CD56, NCAM1, NCR1, FCGR3A). Polarized CD4+ T cell annotations were refined using the combination of transcription factor genes and expressed cytokines for the respective helper T cell types: Th1 (IFNG, TBX21, TNFA), Th2 (GATA3, IL4, IL5), Th17 (RORC, IL17A, IL17F, IL21). Where clusters appeared heterogeneous in their expression of T cell lineage markers, single cell annotations were refined based on the co-expression of specific marker gene and protein pairs. Dot plots to visualise marker protein and mRNA expression across clusters were generated using the R package `ggplot2` (v3.3.3). UMAP59 was used to project all single T cells into a 2D space ($k=31$) using the first 30 batch-integrated PCs as input using the R package `umap`. R version 4.0.3 and Bioconductor version 3.12 were used for all analyses.

T cell differential gene expression analysis

Differential gene expression (DGE) analysis was performed across COVID-19 disease severity groups, ordered from healthy > asymptomatic > mild > moderate > severe > critical. Donor pseudo-bulk samples were first created by aggregating gene counts for each annotated T cell type, within each donor, where there were at least 20 cells of that type. Genes with fewer than 3 counts in any given pseudo-bulk, or fewer than 5 counts in total across donor pseudo-bulk samples, were removed prior to analysis. DGE testing was performed using a negative binomial generalized linear model (NB GLM) implemented in the Bioconductor package `edgeR60,61` (v3.32.1). Statistically significant DE genes were defined with FDR < 0.1. Functional annotation enrichment was performed using the Bioconductor package `enrichR62`(v3.0). Up- and down-regulated DE genes in each T cell type were used as input, testing separately against the MSigDB Hallmark 2020 and Transcription Factor Protein-Protein Interactions gene sets. Significant enrichments were defined with 1% FDR.

T cell receptor analysis

Single-cell TCRs were computed from the TCR-seq data using Cellranger v4.0.0. The unfiltered output of reconstructed TCR contigs across all 3 sites (Newcastle, Cambridge, London) were combined prior to filtering using: 1) full length CDR3, 2) droplet barcode matched a T cell droplet, 3) productive CDR3 spanning V+J genes. Chain-specific TCR clones were defined for each observed α and β chain by first concatenating the V, J and identical CDR3 nucleotide sequences. For each single T cell, these chains were then combined to form a single clonotype, removing cells that contained: 1) > 2 β chains and > 2 α chains, 2) a single α or a single β chain only. T cells with exactly 2 β chains and 1 α chain, or those with exactly 2 α chains and 1 β chain were retained. TCR clonotypes were counted within each donor sample, and expanded clones were defined where > 1 cell was assigned to the TCR clonotype.

The proportion of expanded clones as a function of a linear trend across disease severity groups was modelled using logistic regression, adjusted for age, gender and batch. A separate model was run for each T cell subtype which contained at least 5 cells assigned to the expanded TCR clonotypes. Linear trend p-values were corrected for multiple testing using the Benjamini & Hochberg procedure63.

The TE:EM ratio was calculated within each donor, using the number of observed expanded clonotypes. The TE:EM ratio change across COVID-19 severity was tested using a robust linear model implemented in the R package `robustbase` (v0.93-7), regressing TE:EM ratio on either disease severity as an ordered linear variable (asymptomatic > mild > moderate > severe > critical) or symptom duration, adjusted for age, gender and batch. Statistical significance was defined based on the linear trend across disease severity ($p \leq 0.01$). An equivalent analysis was performed, restricted to patients with a shorter symptom duration (≤ 24 days).

Differential correlation analysis

Changes in the correlations between PBMC cell types were computed using a differential correlation analysis, implemented in the R package DCARS64 (v0.3.5). Cell type proportions were computed by normalizing the counts of each cell type within each donor by the total number of cells captured for that donor sample. Donor samples were ranked according to their disease severity (healthy > asymptomatic > mild > moderate > severe > critical). Differential correlation analysis was then performed between CD4.Tfh vs all B cell types. Statistically significant differentially correlated cell types were defined with empirical p-value ≤ 0.1 , estimated from 10,000 permutations.

BCR V(D)J analysis

Single-cell V(D)J data from the 5' Chromium 10x kit were initially processed with cellranger-*vdj* (4.0.0). BCR contigs contained in `filtered_contigs.fasta` and `filtered_contig_annotations.csv` from all three sites were then pre-processed using `imccantation` inspired preprocessing pipeline⁶⁵ implemented in the `dandelion` python package; `dandelion` is a novel single cell BCR-seq analysis package for 10x Chromium 5' data. All steps outlined below are performed using `dandelion v0.0.27.post2` and is available at <https://github.com/clatworthylab/dandelion>.

BCR preprocessing

Individual BCR contigs were re-annotated with `igblastn v1.1.15` using the IMGT reference database (date downloaded: 30-June-2020)⁶⁶ by calling `changeo's AssignGenes.py` script and re-annotated contigs in blast format were parsed into the Adaptive Immune Receptor Repertoire (AIRR) standards 1.3 format with `changeo's MakeDB.py` script. Amino acid sequence alignment information not present in the output from blast format were retrieved from re-annotation with `igblastn` in airr format. Heavy chain V-gene alleles were corrected for individual genotypes with `TigGER67 (v1.0.0)` using a modified `tigger-genotype.R` script from `imccantation` suite. Germline sequences were reconstructed based on the genotype corrected V-gene assignments using `changeo's (v1.0.1) CreateGermlines.py` script; contigs which fail germline sequence reconstruction were removed from further analysis. Constant genes were re-annotated using `blastn (v2.10.0+)` with CH1 regions of constant gene sequences from IMGT followed by pairwise alignment against curated sequences to correct assignment errors due to insufficient length of constant regions.

BCR filtering

Contigs assigned to cells that passed quality control on the transcriptome data were retained for further quality control assessment, which includes checks for: i) contigs with mismatched locus, V-, J- and constant gene assignments were removed from the analysis; ii) cell barcodes with multiple heavy chain contigs were flagged for filtering. Exceptions to this would be when a) the multiple heavy chain contigs were assessed to have identical V(D)J sequences but assigned as different contigs belonging to the same cell by `cellranger-vdj`, b) when there is a clear dominance (assessed by difference in UMI count) by a particular contig, and c) if and when there is presence of one IgM and one IgD contig assigned to a single cell barcode. In the first two cases, the contig with the highest UMI count is retained; iii) cell barcodes with multiple light chain contigs were flagged for filtering; iv) in situations where cell barcodes are matched with only light chain contigs, the contigs would be dropped from the V(D)J data but transcriptome barcode will be retained.

B cell clone/clonotype definition

BCRs were grouped into clones/clonotypes based on the following sequential criterion that applies to both heavy chain and light chain contigs – i) identical V- and J- gene usage, ii) identical junctional CDR3 amino acid length, and iii) at least 85% amino acid sequence similarity at the CDR3 junction (based on hamming distance). Light chain pairing is performed using the same criterion within each heavy chain clone. Only samples collected at day 0 of the study were analyzed from this step onwards and clones/clonotypes were called across the entire dataset; the sample from one of the donors who was subsequently found to have a B cell malignancy was separated from the analysis and processed independently.

B cell clone/clonotype network

Single-cell BCR networks were constructed using adjacency matrices computed from pairwise Levenshtein distance of the full amino acid sequence alignment for BCR(s) contained in every pair of cells within each disease severity cohort. Construction of the Levenshtein distance matrices were performed separately for heavy chain and light chain contigs and the sum of the total edit distance across all layers/matrices was used as the final adjacency matrix. To construct the BCR neighborhood graph, a minimum spanning tree was constructed on the adjacency matrix for each clone/clonotype, creating a simple graph with edges indicating the shortest edit distance between a B cell and its nearest neighbor. Cells with identical BCRs i.e. cells with a total pairwise edit distance of zero are then connected to the graph to recover edges trimmed off during the minimum spanning tree construction step. Fruchterman-reingold graph layout was generated using a modified method to prevent singletons from flying out to infinity in `networkx (v2.5)`. Visualisation of the resulting single-cell BCR network is achieved via transferring the graph to relevant `anndata` slots, allowing for access to plotting tools in `scanpy`.

The use of the BCR network properties for computing gini indices was inspired from bulk BCR-seq network analysis methods where distribution of clone sizes and vertex sizes (sum of identical BCR reads) in BCR clone networks were used to infer the relationships between BCR clonality, somatic hypermutation and diversity⁶⁸. However, there are challenges with native implementation of this approach for single-cell data. Firstly, to enable calculation of network-based clone/cluster and vertex/node size distribution, BCR networks needed to be reduced such that nodes/cells with identical BCRs had to be merged and counted; this required the re-construction of BCR networks per sample and discarding single-cell level information. Furthermore, the process of node contraction and counting of merging events requires significant computation time and resource. Secondly, this approach is dependent on sufficient coverage of the BCR repertoire, as the BCRs from the number of cells sampled (post-QC) may not necessarily recapitulate the entire repertoire, which may under- or over-represent merged counts for gini index calculation. We propose the use of node closeness centrality computed on each expanded clone (clone size > 1) as an alternative metric to emulate the statistics to adapt to the single-cell nature of the data; closeness centrality defines how close and central each node is with respect to other nodes in the graph and therefore cells with identical BCRs will have high closeness centrality scores, due to the way the BCR network is constructed in `dandelion`. Thus, we can quickly calculate if cells across clones, and/or samples overall, in the entire graph display proportionately/disproportionately high or low closeness centrality scores. One caveat to the current implementation is that it is only meaningful if there are clonotypes with at least two cells as scores will only be computed for non-singleton components of the graph. Gini indices are computed using `skbio.diversity.alpha.gini_index (scikit-bio v0.5.6)` with the `trapezoids` method after clone definition and network generation. Summary visualisation was performed using plotting tools in `seaborn (v0.11.0)`.

Definition of BCR convergence across patients

BCR overlap was determined by collapsing sharing incidence of V- and J- gene usage and CDR3 amino acid sequences, in both heavy and light chains, between individuals into a binarized format (1 or 0). The information is turned into an adjacency matrix where an edge is created between two individuals if there is at least one clonotype (at least 1 cell from each individual displays an identical combination of heavy and light chain V- and J- gene usage with allowance for somatic hypermutation at the CDR3 junctional region) that is similar between the two

individuals. Visualisation is achieved using the CircosPlot function from nxviz package (v0.6.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The dataset from our study can be explored interactively through a web portal: <https://covid19cellatlas.org>. The data object, as a h5ad file, can also be downloaded from <https://covid19cellatlas.org>. The processed data is available to download from Array Express using accession number E-MTAB-10026.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The sample size was not predetermined and all samples that were available were processed. Data from samples across the three sites were pooled to increase sample size for statistical power.

Data exclusions

Doublets:

Any cells labeled as doublets (as defined below) were excluded from the analysis.

Newcastle:

Scrublet (v0.2.1) was applied to each sample to generate a doublet score. These formed a bimodal distribution so the tool's automatic threshold was applied.

Cambridge:

Non-empty droplets were called within each multiplexed pool of donors using the emptyDrops function implemented in the Bioconductor package DropletUtils (v1.10.3), using a UMI threshold of 100 and FDR of 1%. The probability of being a doublet was estimated for each cell per sample (that is one 10x lane) using the "doubletCells" function in scran based on highly variable genes (HVGs). Next, we used "cluster_walktrap" on the SNN-Graph that was computed on HVGs to form highly resolved clusters per sample. Per-sample clusters with either a median doublet score greater than the median + 2.5 x MAD or clusters containing more than the median + 2.5 MAD genotype doublets were tagged as doublets. This was followed by a second round of highly-resolved clustering across the whole data set, in which again cells belonging to clusters with a high proportion (> 60%) of cells previously labelled as doublets were also defined as doublets.

London:

For pooled donor CITE-seq samples, the donor ID of each cell was determined by genotype-based demultiplexing using souporcell version 251. Souporcell analyses were performed with 'skip_remap' enabled and a set of known donor genotypes given under the 'common_variants' parameter. The donor ID of each souporcell genotype cluster was annotated by comparing each souporcell genotype to the set of known genotypes. Droplets that contained more than one genotype according to souporcell were flagged as 'ground-truth' doublets for heterotypic doublet identification. Ground-truth doublets were used by DoubletFinder 2.0.352 to empirically determine an optimal 'pK' value for doublet detection. DoubletFinder analysis was performed on each sample separately using 10 principal components, a 'pN' value of 0.25, and the 'nExp' parameter estimated from the fraction of ground-truth doublets and the number of pooled donors.

Low quality cells:

Combined raw data from the three centres was filtered to remove those that expressed fewer than 200 genes and >10% mitochondrial reads.

Myeloid cells:

Due to compositional differences across sites, when analysing differential abundance of myeloid populations (Figure 2), only samples from Newcastle and London were included.

Replication

Similar results were obtained using orthogonal methodological approaches including experiments performed across three laboratories. Due to sample availability, we were unable to include technical replicates.

Randomization

Samples were not allocated into experimental groups.

Blinding

Blinding was not relevant as it is a single arm study.

Reporting for specific materials, systems and methods

Materials & experimental systems

Methods

- n/a | Involved in the study
- Antibodies
- Eukaryotic cell lines
- Palaeontology and archaeology
- Animals and other organisms
- Human research participants
- Clinical data
- Dual use research of concern

- n/a | Involved in the study
- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Antibodies

Antibodies used

anti-CD107a-BB700 antibody (1:50, clone H4A3, BD Bioscience, 566558), anti-CD14-FITC (1:50, clone M5E2, BD Biosciences, 555397), anti-CD19-FITC (1:50, clone 4G7, BD Biosciences, 345776), anti-CD137-Pe-Dazzle594 (1:50, clone 4B4-1, Biolegend, 309826), anti-CCR7-PE-Cy7 (1:50, clone G043H7, Biolegend, 353226), anti-CD45RO-APC-H7 (1:50, clone UCHL1, BD Biosciences, 561137), anti-CD28-BV480 (1:50, clone CD28.2, BD Biosciences, 566110), anti-CD4-BV785 (1:100, clone SK3, Biolegend, 344642), anti-CD3-BUV395 (1:50, clone UCHT1, BD Biosciences, 563546), anti-CD8-BUV496 (1:100, clone RPA-T8, BD Biosciences, 564804), anti-CD25-BUV737 (1:100, clone 2A3, BD Biosciences, 612806) and viability dye Zombie Yellow (1:200, Biolegend, 423104) anti-IL10-PE (1:10, clone JES3-19F1, BD Biosciences, 559330), anti-IFN-APC (1:25, Miltenyi Biotec, 130-090-762), anti-TNF-AF700 (1:50, clone MAb11, Biolegend, 502928), anti-IL2-BV421 (1:100, clone 5344.111, BD Biosciences, 562914), anti-CD154-BV605 (1:50, clone 24-31, Biolegend, 310826). TotalSeq™-C Human Universal Cocktail, V1.0, Biolegend.

Validation

The following antibodies have been validated to use as a 1:10 dilution.
anti-IL10-PE (clone JES3-19F1, BD Biosciences, 559330)

The following antibodies have been validated to use as a 1:25 dilution.
anti-IFN-APC (1:25, Miltenyi Biotec, 130-090-762)

The following antibodies have been validated to use as a 1:50 dilution.
anti-CD107a-BB700 antibody (clone H4A3, BD Bioscience, 566558)
anti-CD14-FITC (clone M5E2, BD Biosciences, 555397)
anti-CD19-FITC (clone 4G7, BD Biosciences, 345776)
anti-CD137-Pe-Dazzle594 (clone 4B4-1, Biolegend, 309826)
anti-CCR7-PE-Cy7 (clone G043H7, Biolegend, 353226)
anti-CD45RO-APC-H7 (clone UCHL1, BD Biosciences, 561137)
anti-CD28-BV480 (clone CD28.2, BD Biosciences, 566110)
anti-CD3-BUV395 (clone UCHT1, BD Biosciences, 563546)
anti-TNF-AF700 (clone MAb11, Biolegend, 502928)
anti-CD154-BV605 (clone 24-31, Biolegend, 310826)

The following antibodies have been validated to use as a 1:100 dilution.
anti-CD4-BV785 (clone SK3, Biolegend, 344642)
anti-CD8-BUV496 (clone RPA-T8, BD Biosciences, 564804)
anti-CD25-BUV737 (clone 2A3, BD Biosciences, 612806)
anti-IL2-BV421 (1:100, clone 5344.111, BD Biosciences, 562914)

We validated the following antibody to use 1 vial per 1 million cells.
TotalSeq™-C Human Universal Cocktail, V1.0, Biolegend

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Population characteristics are outlined in Supplementary Table 2.
We collected samples from patients aged 21-86 and both male and female.
Other characteristics were not used as covariates in the study.

Recruitment

Newcastle:
Patients were recruited and consented under the Newcastle Biobank (REC 17/NE/0361, IRAS 233551) study and ethical governance.

Cambridge:
Study participants were recruited between 31/3/2020 and 20/7/2020 from patients attending Addenbrooke's Hospital with a suspected or nucleic acid amplification test (NAAT) confirmed diagnosis of COVID-19 (including point of care testing (Collier et al., 2020; Mlcochova et al., 2020)), patients admitted to Royal Papworth Hospital NHS Foundation Trust or Cambridge and Peterborough Foundation Trust with a confirmed diagnosis of COVID-19, together with Health Care Workers identified through staff screening as PCR positive for SARS-CoV-2 (Rivett et al., 2020). Controls were recruited among hospital staff

attending Addenbrooke's serology screening programme, and selected to cover the whole age spectrum of COVID-19 positive study participants, across both genders. Only controls with negative serology results (45 out of 47) were subsequently included in the study. Recruitment of inpatients at Addenbrooke's Hospital and Health Care Workers was undertaken by the NIHR Cambridge Clinical Research Facility outreach team and the NIHR BioResource research nurse team. Ethical approval was obtained from the East of England – Cambridge Central Research Ethics Committee ("NIHR BioResource" REC ref 17/EE/0025, and "Genetic variation AND Altered Leukocyte Function in health and disease - GANDALF" REC ref 08/H0308/176). All participants provided informed consent.

London:

Subjects 18 years and older were recruited from two large hospital sites in London, United Kingdom, namely University College London Hospitals NHS Foundation Trust and Royal Free London NHS Foundation Trust during the height of the pandemic in the United Kingdom (April to July 2020) and gave informed consent.

Ethics oversight

Newcastle Biobank (REC 17/NE/0361, IRAS 233551)
 IV-LPS: REC (17/YH/0021)
 ("NIHR BioResource" REC ref 17/EE/0025, and "Genetic variation AND Altered Leukocyte Function in health and disease - GANDALF" REC ref 08/H0308/176)
 UCL Great Ormond Street Institute of Child Health (REC reference: 19/NW/0171, IRAS project ID 261511)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

PBMCs were isolated from blood samples using Lymphoprep (StemCell Technologies) density gradient centrifugation as per manufacturer's instructions. Single cell suspensions were then washed with Dulbecco's phosphate buffered saline (PBS) (Sigma) and frozen in 5-10 million cell aliquots in 90% (v/v) heat inactivated fetal calf serum (FCS) (Gibco) 10% (v/v) DMSO (Sigma Aldrich). Purified PMBC were thawed at 37°C, transferred into a 15 mL tube with 10 mL pre-warmed complete culture media RPMI-1640 medium (Sigma Aldrich, R0883) supplemented with 10% (v/v) FCS (Gibco, 10270-106), 1% (v/v) Penicillin/Streptomycin (100 U/mL and 100 µg/mL respectively; Sigma Aldrich, P0781) and 1% (v/v) L-Glutamine (2 mM; Sigma Aldrich, G7513), referred as RPMI10, followed by centrifugation at 500 g for 5 min. Cell pellet was resuspended in 500 µL RPMI10 with added DNase (1 µg/mL, Merck, 10104159001), divided into 5 wells of round bottom 96-well plate and left to rest at 37°C for an hour. Cells were stimulated with SARS-CoV-2 PepTivator peptide S for pan-HLA (2 µg/mL, Miltenyi Biotec, 136-126-700) and PMA/Ionomycin as a control (2 µL/mL, Cell Activation cocktail, Biolegend, 423301), and incubated at 37°C for 2 h. Negative controls were left untreated. Brefeldin A (2 µg/mL, GolgiPlug, BD Bioscience, 555029) and anti-CD107a-BB700 antibody (1:50, clone H4A3, BD Bioscience, 566558) was added for additional 4 h into all conditions. Cells were stained for detection of activation induced markers and intracellular cytokines 6 h after stimulation and subjected to flow cytometry. PBMC stimulated for 6 h with the SARS-Cov-2 peptide were washed with PBS, and cell surface stained for 1 h at room temperature: anti-CD14-FITC (1:50, clone M5E2, BD Biosciences, 555397), anti-CD19-FITC (1:50, clone 4G7, BD Biosciences, 345776), anti-CD137-Pe-Dazzle594 (1:50, clone 4B4-1, Biolegend, 309826), anti-CCR7-PE-Cy7 (1:50, clone G043H7, Biolegend, 353226), anti-CD45RO-APC-H7 (1:50, clone UCHL1, BD Biosciences, 561137), anti-CD28-BV480 (1:50, clone CD28.2, BD Biosciences, 566110), anti-CD4-BV785 (1:100, clone SK3, Biolegend, 344642), anti-CD3-BUV395 (1:50, clone UCHT1, BD Biosciences, 563546), anti-CD8-BUV496 (1:100, clone RPA-T8, BD Biosciences, 564804), anti-CD25-BUV737 (1:100, clone 2A3, BD Biosciences, 612806) and viability dye Zombie Yellow (1:200, Biolegend, 423104). Cells were washed with PBS 2% (v/v) FCS, fixed with 4% (w/v) paraformaldehyde (ThermoFisher Scientific, 28908) and kept at 4°C overnight. Subsequently, cells were washed with PBS, permeabilized with Perm/Wash buffer (BD Biosciences, 554723) according manufacturer's instruction, and stained with intracellular antibodies for 1 h on ice: anti-IL10-PE (1:10, clone JES3-19F1, BD Biosciences, 559330), anti-IFN-APC (1:25, Miltenyi Biotec, 130-090-762), anti-TNF-AF700 (1:50, clone MAb11, Biolegend, 502928), anti-IL2-BV421 (1:100, clone 5344.111, BD Biosciences, 562914), anti-CD154-BV605 (1:50, clone 24-31, Biolegend, 310826). Cells were washed, transferred to flow cytometry 5 mL tubes.

Instrument

Symphony A5 flow cytometer (BD Biosciences)

Software

FlowJo V10 (BD Biosciences)

Cell population abundance

No sorting was performed

Gating strategy

Cells were gated using the following strategy. Live, single > Lineage, including CD3, CD4 and CD8 > Non "dump" including CD14 and CD19 > CD107a or CD137

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.