https://doi.org/10.1038/s41588-021-00955-3

nature genetics

Identification of *LZTFL1* as a candidate effector gene at a COVID-19 risk locus

Damien J. Downes[1], Amy R. Cross[2,13], Peng Hua[1,13], Nigel Roberts[1], Ron Schwessinger[1,3], Antony J. Cutler[4,12], Altar M. Munis[5], Jill Brown[1], Olga Mielczarek[4], Carlos E. de Andrea[6], Ignacio Melero[7], COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium*, Deborah R. Gill[5], Stephen C. Hyde[5], Julian C. Knight[4,8,9], John A. Todd[4], Stephen N. Sansom[10], Fadi Issa[2,11], James O. J. Davies[1,11] ✉ and Jim R. Hughes[1,3] ✉

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) disease (COVID-19) pandemic has caused millions of deaths worldwide. Genome-wide association studies identified the 3p21.31 region as conferring a twofold increased risk of respiratory failure. Here, using a combined multiomics and machine learning approach, we identify the gain-of-function risk A allele of an SNP, rs17713054G>A, as a probable causative variant. We show with chromosome conformation capture and gene-expression analysis that the rs17713054-affected enhancer upregulates the interacting gene, leucine zipper transcription factor like 1 (*LZTFL1*). Selective spatial transcriptomic analysis of lung biopsies from patients with COVID-19 shows the presence of signals associated with epithelial–mesenchymal transition (EMT), a viral response pathway that is regulated by *LZTFL1*. We conclude that pulmonary epithelial cells undergoing EMT, rather than immune cells, are likely responsible for the 3p21.31-associated risk. Since the 3p21.31 effect is conferred by a gain-of-function, *LZTFL1* may represent a therapeutic target.

The COVID-19 pandemic is estimated to have caused over 4.6 million deaths so far[1,2]. The predominant cause of mortality is pneumonia and severe acute respiratory distress syndrome[3]. However, COVID-19 can cause multiple organ failure through cytokine release, microvascular and macrovascular thrombosis, endothelial damage, acute kidney injury and myocarditis[4–6]. Genome-wide association studies (GWAS) are important for identifying candidate genes and pathways that predispose to complex diseases[7]; genetically validated drug targets are more likely to lead to approved drugs[8]. Two large GWAS were carried out to determine whether common variants drive susceptibility to severe COVID-19 (refs. [9,10]). Both studies identified a region of chromosome 3p21.31 as having the strongest association, while a third study also identified this locus as conferring susceptibility to infection[11]. The 3p21.31 risk haplotype, which arises from Neanderthal DNA[12] and is currently unexplained with regards to the causal variant(s), causal gene(s) and specific role in COVID-19, confers a twofold increased risk of respiratory failure from COVID-19 (refs. [9,10]) and an over twofold increased risk of mortality for individuals under 60 (ref. [13]). Additionally, the risk variants at this locus are carried by >60% of individuals with South Asian ancestry (SAS), compared to 15% of European ancestry (EUR) groups, partially explaining the ongoing higher death rate in this population in the UK[14,15].

Identifying the causal gene(s) and mechanism(s) behind GWAS hits poses several challenges. First, a causative variant is usually in linkage disequilibrium (LD) with many other variants and these can take different forms (SNPs, insertions, deletions and structural polymorphisms). Second, the genetic signals are completely cell type-agnostic, which makes it challenging to identify appropriate experimental models for further investigation. Third, there are multiple mechanisms by which variants can have an effect. Alteration of the protein-coding sequence or RNA splicing, both of which are relatively straightforward to disentangle, account for fewer than 20% of associations in polygenic disease[16]. The remaining variants and their target gene(s) can be very difficult to decode. Many are thought to lie within *cis*-regulatory elements[17], such as enhancers, which are short DNA sequences that often control tissue- and developmental stage-specific gene expression. Deciphering the variants that affect enhancers is challenging because many enhancers are only active in specific cell types or at specific times; enhancers are often distant in the linear DNA sequence (often 10^4–10^6 base pairs (bp)) from the genes they control and the effects of sequence changes are not straightforward to predict.

We developed a comprehensive platform for decoding the effects of sequence variation identified by GWAS[16] (Extended Data Fig. 1a). This combines computational and wet lab approaches to delineate

[1]Department of Medicine, Medical Research Council Molecular Haematology Unit, Medical Research Council Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [2]Nuffield Department of Surgical Sciences, Transplantation Research and Immunology Group,University of Oxford, Oxford, UK. [3]Department of Medicine, Medical Research Council Weatherall Institute of Molecular Medicine Centre for Computational Biology, University of Oxford, Oxford, UK. [4]Nuffield Department of Medicine, Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [5]Department of Medicine, Gene Medicine Group, Nuffield Division of Clinical Laboratory Sciences, Radcliffe University of Oxford, Oxford, UK. [6]Department of Pathology, Clínica Universidad de Navarra, Pamplona, Spain. [7]Division of Immunology and Immunotherapy, Centre for Applied Medical Research, University of Navarra, Pamplona, Spain. [8]Chinese Academy of Medical Science Oxford Institute, University of Oxford, Oxford, UK. [9]National Institute for Health Research Oxford Biomedical Research Centre, Oxford, UK. [10]Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK. [11]Oxford University Hospitals National Health Service Foundation Trust, Oxford, UK. [12]Present address: Immunology Research Unit, GlaxoSmithKline, Stevenage, UK. [13]These authors contributed equally: Amy R. Cross, Peng Hua. *A list of members and their affiliations appears in the Supplementary information. ✉e-mail: james.davies@imm.ox.ac.uk; jim.hughes@imm.ox.ac.uk

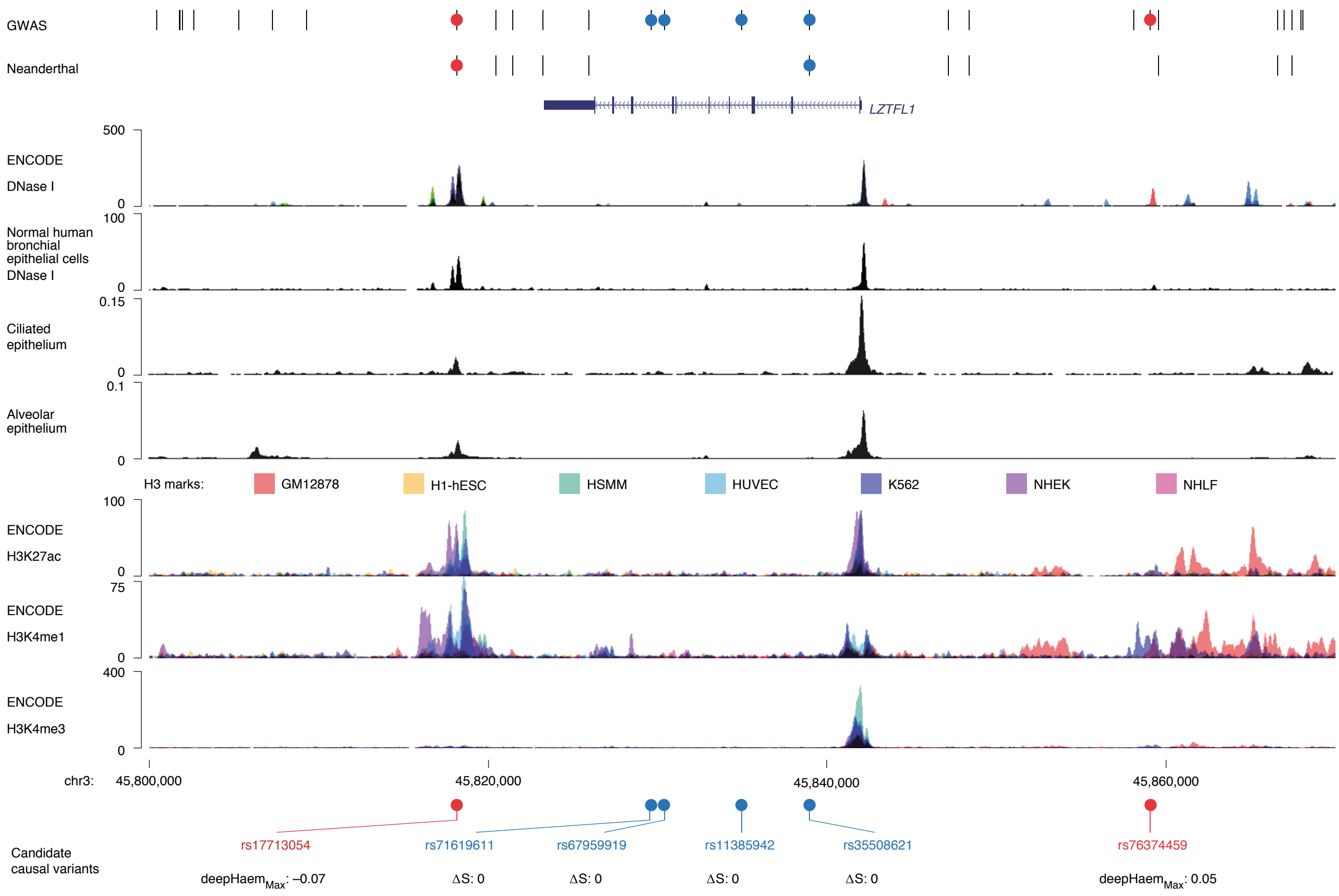


Fig. 1 | Identification of a potentially causative COVID-19 risk variant. COVID-19 risk variants from GWAS were assessed for multiple mechanisms. All genome-wide-significant variants and linked variants are shown (GWAS) as are variants present in the Vindija Neanderthal[12] risk haplotype. The circles indicate variants assessed for splicing changes (blue circles, SpliceAI[18]: ΔS score (0–1, where 1 is the most damaging)), and presence in *cis*-regulatory elements using open chromatin in 95 ENCODE overlaid DNase I datasets (red circles), normal human bronchial epithelial cells and scATAC-seq from fetal ciliated and alveolar epithelia[34]. Histone H3 modification tracks show the presence of marks associated with active transcription (H3K27ac) at enhancers (H3K4me1) and promoters (H3K4me3). Variants in open chromatin are given deepHaem damage scores (0–1) with sign indicating increased (−) or decreased (+) accessibility. The region shown is chr3:45,800,000–45,870,000, hg38. HSMM, human skeletal muscle myoblast; NHEK, normal human epidermal keratinocyte.

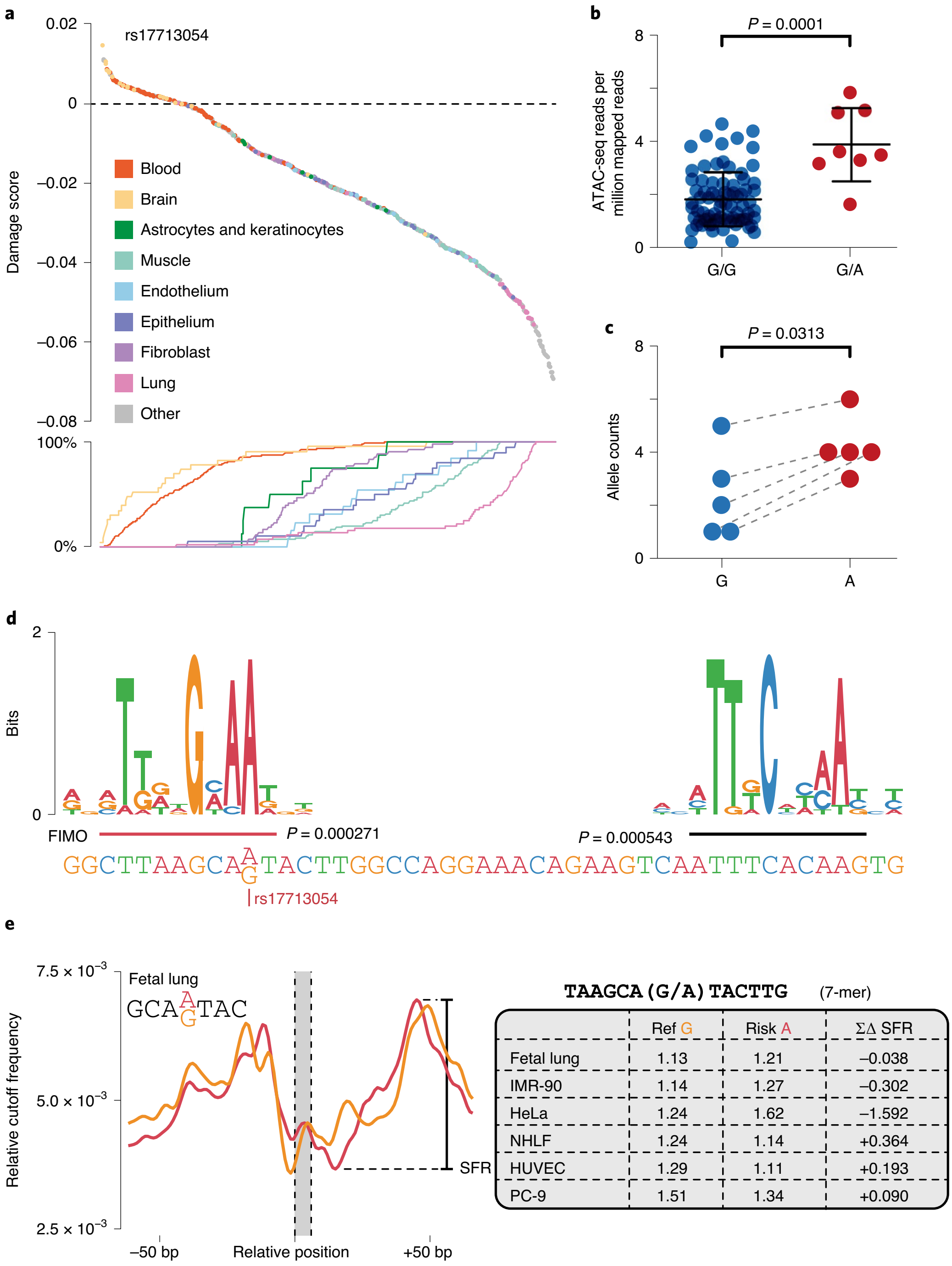
the identity of causative variants, the cell types involved and effector genes. Initially, GWAS-identified haplotypes were screened for potential protein-coding sequence variants. Variants altering splice sites were then assessed using a combination of machine learning[18] and RNA sequencing (RNA-seq) analysis. Conventional genomic approaches were then combined with machine learning[19] to define whether variants were found within, and affected, *cis*-regulatory sequences from a panel of disease-relevant cell types; this allows for the identification of the key cell type(s) and the determination of the likely causative variant. Subsequently, chromosome conformation capture (3C) analysis[20–22] was used to identify the gene promoters, which physically contacted the candidate enhancer sequence in the relevant cell type(s); these data were integrated with gene-expression analyses. Finally, genome editing was used to validate the regulatory effects of prioritized variants.

In this study, we applied this approach to identify rs17713054 as a probable causative variant and *LZTFL1* as a candidate effector gene in pulmonary epithelial cells as contributing to the strong COVID-19 association at the 3p21.31 locus, with EMT identified as a relevant infection response pathway.

Results

The rs17713054 risk allele generates a CCAAT/enhancer binding protein beta motif. The 3p21.31 region contains variants associated with the autoimmune diseases type 1 diabetes[23] and multiple sclerosis[24], although the lead and tag variants identified in these studies are not in high LD with those associated with COVID-19 severity (Extended Data Fig. 1b). There are 28 candidate risk variants in LD with the original genome-wide significant SNPs[9] at 3p21.31 ($r^2 > 0.8$, EUR; Extended Data Fig. 1c). None of these variants affect coding sequences. One SNP, rs35624553, is in the 3′-UTR of the gene *LZTFL1* (Fig. 1) but this is not a conserved microRNA (miRNA) binding site[25] and neither miR*dSNP*[26] nor MicroSNiPer[27] predict that the variant alters miRNA binding. Four other variants are within *LZTFL1* introns, including the lead SNP rs11385942 (ref. [9]). None of these are predicted to alter messenger RNA splicing of *LZTFL1*, either by machine learning with SpliceAI[18] or splicing quantitative trait locus (sQTL)-based approaches[28], and the nearest exon junction to these variants is approximately 500 bp (Fig. 1). Therefore, a *cis*-regulatory mechanism is the most likely explanation for this haplotype.

We first examined open chromatin from 24 diverse immune cell populations[29] (including T, B, natural killer and dendritic cells) in resting and stimulated states but did not identify any of the 28 severe COVID-19-associated variants at 3p21.31 in open chromatin (Extended Data Fig. 1d), making it unlikely that a *cis*-regulatory mechanism in these immune cell types is responsible. By considering open chromatin data from 95 diverse cell types,



	Ref G	Risk A	ΣΔ SFR
Fetal lung	1.13	1.21	–0.038
IMR-90	1.14	1.27	–0.302
HeLa	1.24	1.62	–1.592
NHLF	1.24	1.14	+0.364
HUVEC	1.29	1.11	+0.193
PC-9	1.51	1.34	+0.090

Fig. 2 | rs17713054 creates a CEBPB motif. a, Ranked deepHaem chromatin accessibility damage scores for the risk A allele of rs17713054 in 694 cell types including primary cells. The line plot shows the cumulative percentage of samples for each tissue, an indication that lung tissue is enriched in the highly ranked damaging variants. **b**, Quantification of ATAC-seq reads in the rs17713054 enhancer (chr3:45,817,661–45,818,660, hg38) from aortic endothelium[37]. The bars show the mean and 1 s.d. Two-tailed Mann–Whitney *U*-test, comparing accessibility of the two genotypes; G/G *n* = 78 and G/A *n* = 8 independent experiments. **c**, ATAC-seq reads over rs17713054 alleles in heterozygous individuals; the gray lines denote paired counts from a single replicate. One-sided Wilcoxon matched-pairs signed-rank test, testing the higher accessibility of the A allele, *n* = 5. Three replicates were excluded due to low coverage. **d**, CEBPB DNA binding motif over the sequence around the rs17713054 risk A and non-risk G alleles. The *P* values for the motifs were determined using FIMO with reference and variant sequence for the entire enhancer and JASPAR motif MA0466.1. The motif over rs17713054 was only identified in the sequence with the A allele. **e**, Sasquatch DNase I hypersensitivity profile and shoulder-footprint ratio (SFR) scores for rs17713054 risk and non-risk (ref G) alleles using DNase I datasets for a subset of cells with open chromatin at this site. Larger SFR scores indicate a deeper footprint associated with greater likelihood of being bound by a transcription factor. ΔSFR scores are generated by subtracting the risk A SFR from the ref G SFR; negative values show an increased footprint depth in the risk allele.

we identified 2 SNPs, rs17713054 and rs76374459, which are found in open chromatin[30] (Fig. 1 and Extended Data Fig. 2). Machine learning approaches have proven accurate at predicting allele-specific changes in transcription factor binding and chromatin accessibility[31,32], including de novo gain-of-function changes[33]. We previously developed a machine learning model, deepHaem[19], which uses 694

DNase I hypersensitivity and assay for transposase-accessible chromatin using sequencing (ATAC-seq) datasets to predict changes to active regulatory elements. Importantly, deepHaem predicted that the 26 variants not found in open chromatin have no strong gain-of-function effect in any cell type (Extended Data Fig. 3).

Of the two variants in open chromatin, rs76374459 is unlikely to be causative. It is not contained within the Vindija Neanderthal risk haplotype[12] and is not in tight LD with the 3p21.31 lead SNPs from either of two GWAS[9,10] (rs11385942, $r^2=0.737/0.058$, EUR/SAS; rs73064425, $r^2=0.747/0.058$, EUR/SAS). In addition, it is in an erythroid-specific enhancer, a cell type not strongly implicated in SARS-CoV-2 infection; it is not predicted by machine learning to cause damaging effects (Fig. 1 and Extended Data Figs. 2 and 4). In contrast, rs17713054 is likely to be a causative SNP since it is in tight LD with both lead SNPs (rs11385942, $r^2=1.0/1.0$, EUR/SAS; rs73064425, $r^2=0.986/0.995$, EUR/SAS), is located in open chromatin in numerous COVID-19-relevant cell types, including epithelial and endothelial cells (Fig. 1 and Extended Data Fig. 2), where it is marked by epigenetic modifications associated with active enhancers (histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac)). Inspection of single-cell ATAC-seq (scATAC-seq) from healthy lung[34,35] showed that this enhancer is present in several lung epithelial cell types, including the ciliated epithelium and club cells that line the respiratory tract, and in type 1 and type 2 pneumocytes, which form the alveoli (Fig. 1 and Extended Data Fig. 5). Interestingly, deepHaem predicted that the rs17713054 risk allele, which is the minor allele A (minor allele frequency (MAF): 0.0817 EUR, 0.377 SAS[36]), acts as a gain-of-function mechanism by augmenting an existing enhancer, resulting in increased chromatin accessibility in both epithelial and endothelial cells and particularly in primary lung tissue (Fig. 2a). Analysis of ATAC-seq for human aortic endothelial cells from 48 individuals[37] showed that the rs17713054-containing enhancer was significantly more accessible in heterozygous A/G donors than homozygous G/G donors (Fig. 2b); in heterozygous samples, more reads originated the risk A allele than the non-risk G allele (Fig. 2c).

Sequence analysis showed that the risk allele generates a second CCAAT/enhancer binding protein beta (CEBPB) motif[38] in the enhancer (Fig. 2d). The biological relevance of this new motif is supported by strong expression of CEBPB in lung tissue[28] and chromatin immunoprecipitation followed by sequencing (ChIP-seq) of CEBPB in HeLa, A549 alveolar basal epithelial adenocarcinoma and IMR-90 lung fibroblast cells[39]––which are homozygous G/G non-risk––showing weak binding at the enhancer (Extended Data Fig. 6a–d). Furthermore, deepHaem predicted that rs17713054-A would lead to increased CEBPB binding in IMR-90 and A549 cells (Extended Data Fig. 6e). An orthogonal DNase I hypersensitivity footprinting-based approach, Sasquatch[40], uses genome-wide, cell type-specific motif footprints to predict how sequence-specific changes alter transcription factor binding. This found that motifs containing either allele have strong DNase I footprints. When comparing motifs with the risk A allele with the non-risk G allele, risk A motifs showed a weak gain in accessibility in fetal lung and IMR-90 lung fibroblast cells (Fig. 2e), corroborating a gain-of-function mechanism.

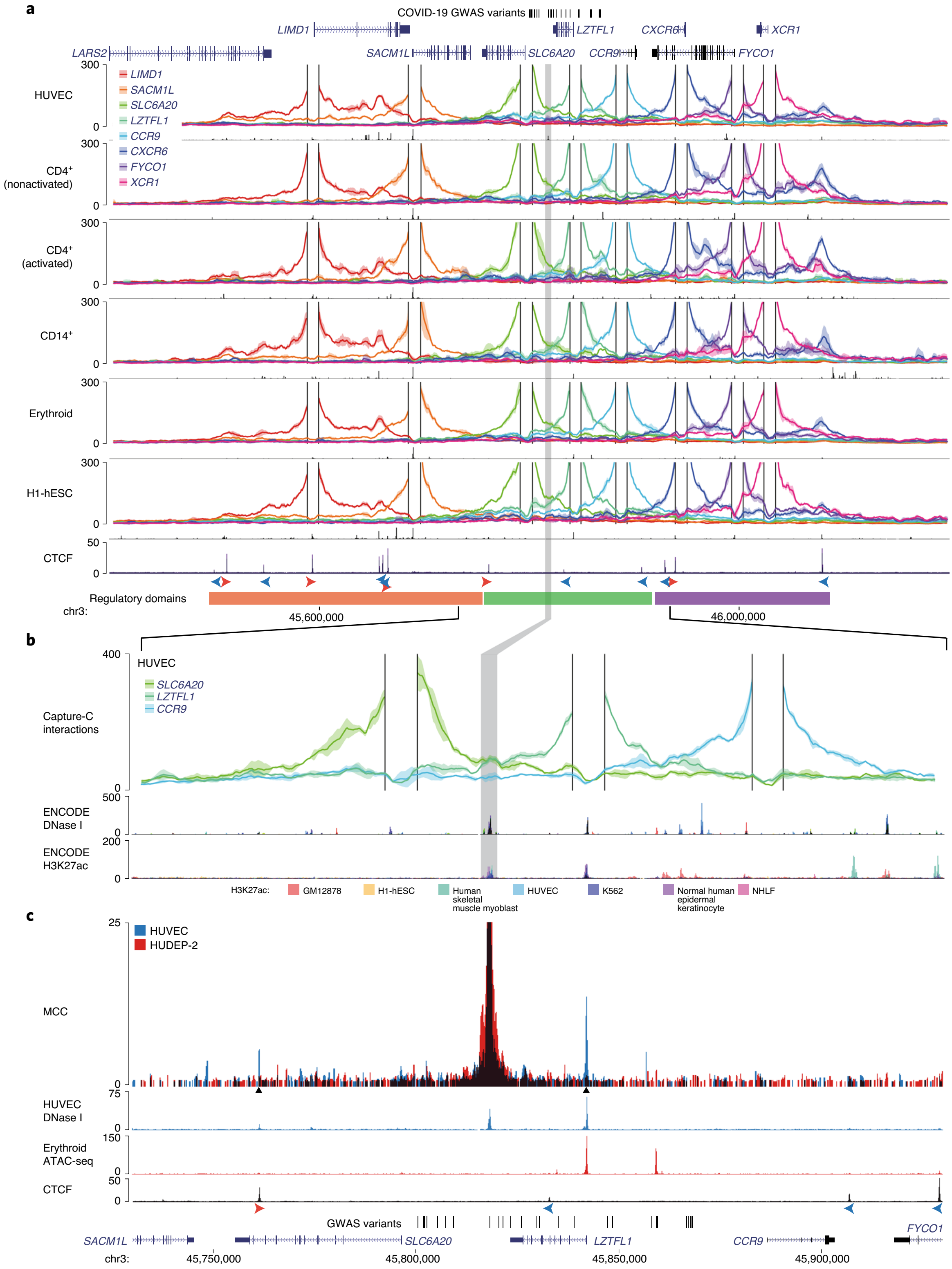
The rs1773054 enhancer interacts with the *LZTFL1* promoter. The 3p21.31 locus is gene-dense and contains several candidates that could potentially be involved in COVID-19 pathogenesis. These include three chemokine receptors: *CCR9* (which encodes a lymphocyte-expressed C-C chemokine receptor[41]); *CXCR6* (which is associated with sarcoidosis and is a coreceptor for HIV[42,43]); and *XCR1* (which encodes a X-C chemokine receptor). Transcriptome-wide association study (TWAS) analysis also identified *CCR2*, *CCR3* and *FYCO1*, which lie up to 500 kilobases (kb) away, as candidate effector genes for the 3p21.31 COVID-19 association[10]. In addition, there are the two nearest genes that are less well studied: *SLC6A20* (the SIT1 imino acid transporter associated with glycinuria[44]) and *LZTFL1* (ref. [45]), the homozygous loss of which causes the classical ciliopathy Bardet–Biedl syndrome[46,47].

To identify candidate target genes of the rs17713054 enhancer we performed NuTi Capture-C[20,21] from the promoters of genes in surrounding regulatory domains (Methods) in primary human umbilical vein endothelial cells (HUVECs) where the rs17713054 enhancer is accessible, as well as resting and stimulated primary CD4$^+$ T cells, primary CD14$^+$ monocytes, CD71$^+$ CD235$^+$ erythroid cells and H1 human embryonic stem cells (H1-hESCs), where the enhancer is not accessible. In all cell types tested, all 28 COVID-19-associated variants fell within a domain of interaction that contained only the promoters of *LZTFL1*, *SLC6A20* and *CCR9*, and is delimited by convergent CTCF boundary motifs (Fig. 3a). Within this domain, the promoters of both *LZTFL1* and *SLC6A20* interacted more strongly with the rs17713054 enhancer than *CCR9* (Fig. 3b). Reciprocal Capture-C from the rs17713054 enhancer also showed that its interactions were primarily constrained to the same domain (Extended Data Fig. 7a). Notably, inside this domain, several tissue-specific enhancers could be seen for immune, erythroid and endothelial cell types, altering the interaction profile of the ubiquitously accessible *LZTFL1* promoter and indicating dynamic regulation (Supplementary Fig. 1).

We went on to perform Micro Capture-C (MCC), a 3C method that provides higher resolution data than conventional approaches[22], from the rs17713054 enhancer in endothelial cells. MCC in HUVECs delineated significant tissue-specific interaction with the *LZTFL1* promoter and the nearest upstream boundary CTCF site but no other significant peaks of interactions with any of the other gene promoters in the region (Fig. 3c and Extended Data Fig. 7a). Importantly, we did not find a peak of interaction with *SLC6A20*, probably because ENCODE datasets show that *SLC6A20* carries Polycomb repression marks in endothelial (HUVEC) and normal human lung fibroblast (NHLF) cells (Extended Data Fig. 7b). Additionally, the *LZTFL1* promoter was more consistently accessible in cells where rs17713054 was also accessible (Extended Data Fig. 7c,d). Therefore, *LZTFL1* is the most likely direct regulatory target of the rs17713054-containing epithelial–endothelial–fibroblast enhancer.

Fig. 3 | The interaction landscape of the severe COVID-19 risk locus. a, DpnII Capture-C-derived mean interaction count ($n=3$ for all except CD14$^+$, $n=2$) and 1 s.d. (shading) for gene promoters in HUVECs, resting and activated T cells (CD4$^+$ nonactivated/activated), monocytes (CD14$^+$), CD71$^+$ CD235$^+$ erythroid cells and H1-hESCs. The enhancer containing rs17713054 is highlighted by a gray box. ATAC-seq/DNase I for each cell type is shown underneath in black. The CTCF track shows binding of the CCAAT-binding factor that acts as a boundary with forward and reverse motif orientation shown with arrowheads (red and blue, respectively). Three broad regulatory domains were identified as regions with overlapping interactions (region: chr3:45,400,000–46,200,000, hg38). Per-fragment interactions were smoothed using 400-bp bins and an 8-kb window. **b**, The rs17713054 regulatory domain in endothelial cells (HUVECs). Overlaid DNase I shows accessible sites in 95 cell types and H3K27ac shows active elements (region: chr3:45,730,000–45,930,000, hg38). Per-fragment interactions were smoothed using 250-bp bins and a 5-kb window. The solid line shows the mean interaction count ($n=3$ independent samples) with 1 s.d. (shading). **c**, MCC of the rs17713054 enhancer in endothelial (HUVECs, blue) and erythroid (HUDEP-2, red) cells with tissue-specific open chromatin tracks ($n=3$). Peak analysis of MCC using LanceOtron to compare the HUVEC and HUDEP-2 profiles identified two significantly enriched peaks in HUVEC cells (black triangles, $P\leq1\times10^{-999}$) that correspond to the *LZTFL1* promoter and upstream CTCF site.

rs17713054 A is associated with higher gene expression in the lung. Disease biology, deepHaem, TWAS analysis[10] and a phenome-wide association study[11] identified lung tissue and function as key for the 3p21.31 COVID-19 association. Analysis of whole-lung RNA-seq[28] showed that *LZTFL1* is strongly expressed in the lung (Fig. 4a) and single-cell RNA-seq (scRNA-seq)[48] showed that *LZTFL1* is present throughout the respiratory epithelium but predominantly expressed in ciliated cells (Fig. 4b,c). Of the other candidate genes identified in this study and elsewhere[10,49,50] (*SLC6A20*, *CCR2*, *CCR3*, *CCR9*, *CXCR6* and *FYCO1*), only *SLC6A20* and *FYCO1* were consistently expressed in both lung bulk and scRNA-seq datasets, although *CCR2* and *CXCR6* were found in bulk RNA-seq. *FYCO1* was found in most cell types and *SLC6A20* was restricted to goblet cells and alveolar type 2 pneumocytes (Fig. 4 and Extended Data Fig. 8). Analysis using the Genotype-Tissue Expression[28] (GTEx) portal for expression quantitative trait loci (eQTLs) showed that the



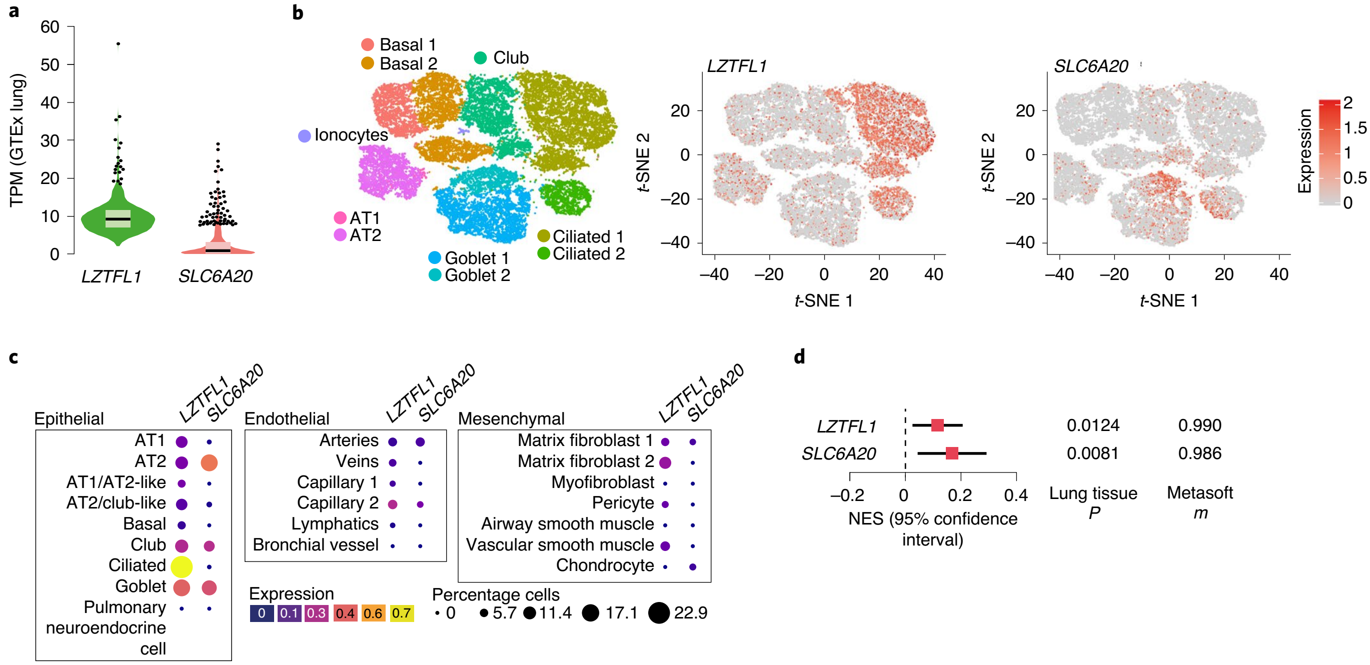


Fig. 4 | Pulmonary expression analysis of *LZTFL1* and *SLC6A20*. **a**, GTEx whole-lung RNA-seq expression profiles for *LZTFL1* and *SLC6A20* as transcripts per million (TPM). For the violin plots, minima and maxima are the top and bottom of the violin, the black lines show the means, the ends of the pale regions denote the first and third quartiles and the black dots denote outliers ($n=578$ independent samples). **b**, 10x Genomics Chromium droplet scRNA-seq from the upper and lower airways and lung parenchyma[48] from healthy volunteers or deceased transplant donors with ten epithelial populations (left). scRNA-seq expression profiles for *LZTFL1* (middle) and *SLC6A20* (right). **c**, Chromium single-nucleus RNA-seq[35] from non-diseased adult lung ($n=3$) with 22 epithelial, endothelial and mesenchymal populations, including AT1 and AT2 pneumocytes. **d**, GTEx eQTL analysis the rs17713054 risk A allele in the lung ($n=515$ independent samples). The normalized effect size (NES) is the slope of the linear regression comparing the alternate (A) allele to the reference (G) allele. NES are calculated in a normalized space where magnitude has no direct biological interpretation. The lines show the 95% confidence interval, with significance values for single-tissue (two-sided *P* value without multiple test correction) and multi-tissue (PP/*m* value) analyses.

rs17713054 A risk allele was associated with higher levels of expression in the lung of *LZTFL1* and *SLC6A20* but not the other genes (Fig. 4d and Extended Data Fig. 8). Colocalization analysis[51] showed that these GWAS and eQTL associations are more likely as a result of a single variant (posterior probability (PP) = 0.2657) than two distinct variants (PP = 0.0566).

CRISPR–Cas9 genome editing[52] allows the possibility to test the role of the rs17713054 enhancer in the regulation of *LZTFL1* and *SLC6A20*. Since the enhancer shows accessibility in epithelial, endothelial and mesenchymal cells (Extended Data Fig. 9a), we used CRISPR–Cas9 ribonucleoprotein (RNP) editing to delete either a 108- or 191-bp region at high efficiency (>70%) from H441 distal lung epithelial cells, adult blood outgrowth endothelial cells, HUVECs and IMR-90 lung fibroblast cells (Extended Data Fig. 9b–d and Supplementary Fig. 2). Using real-time quantitative PCR (qPCR) we detected no effect on *LZTFL1* expression after enhancer deletion (Extended Data Fig. 9e), which is consistent with a study that CRISPR interference in the 16HBE14o- bronchial epithelial cell line had no effect on nearby gene expression[50]. Since *SLC6A20* is Polycomb-repressed in fibroblasts and endothelial cells, it was undetectable by qPCR with reverse transcription (RT–qPCR). To understand the unexpected result, we generated H3K27ac ChIP-seq in all four cell types (Extended Data Fig. 9f,g). The rs17713054 enhancer lacked strong H3K27ac and was probably inactive, explaining the lack of effect seen by deletion. Therefore, a suitable cell model for testing the effects of rs17713054, particularly in the lung epithelium, is not currently available.

Epithelial dysfunction in the COVID-19 lung. Given that the rs17713054 enhancer is present and *LZTFL1* is expressed in lung epithelial cells, the respiratory epithelium is of particular interest for understanding the association at 3p21.31. EMT, a developmental pathway that allows terminally differentiated epithelial cells to dedifferentiate and acquire mesenchymal identity, plays a key role in the innate immune response, is a consequence of lung inflammation and is involved in both the development and resolution of pneumonitis[53–56]. SARS-CoV-2 is known to induce EMT in both lung carcinoma cell lines and in the respiratory tract[57,58] and LZTFL1 is known to regulate EMT through Wnt/β-catenin, hedgehog and transforming growth factor-β (TGF-β) signaling[59,60]. In the context of malignancy, increased levels of LZTFL1 inhibits EMT, whereas decreased LZTFL1 promotes EMT[45,59,60].

Defining EMT in complex tissues is challenging due to its diverse and dynamic nature but can be achieved through a combined assessment of cellular reorganization, an abundance of fibroblasts (which are a product of EMT), presence of EMT-promoting signaling pathways and coexpression of epithelial and mesenchymal markers[61]. Consistent with the work by others[62,63], we saw widespread epithelial dysfunction and diffuse alveolar damage with reorganization indicative of EMT evident in postmortem biopsies of three patients with COVID-19. Dysfunction in ciliated airways included denudation, hyperplasia and squamous metaplasia (Fig. 5a). Features of diffuse alveolar damage included pneumocyte hyperplasia, hyaline membrane deposition, immune inflammation, fine and focal fibrosis and squamous metaplasia (Fig. 5b). Between the areas of interstitial expansion and fibrotic foci, there was an accumulation of fibroblasts, which is generally absent from healthy lung tissue.

We previously generated selective spatial transcriptomics from 46 areas of postmortem biopsies from patients with critical COVID-19 covering a spectrum of alveolar injury[64]. To explore the expression

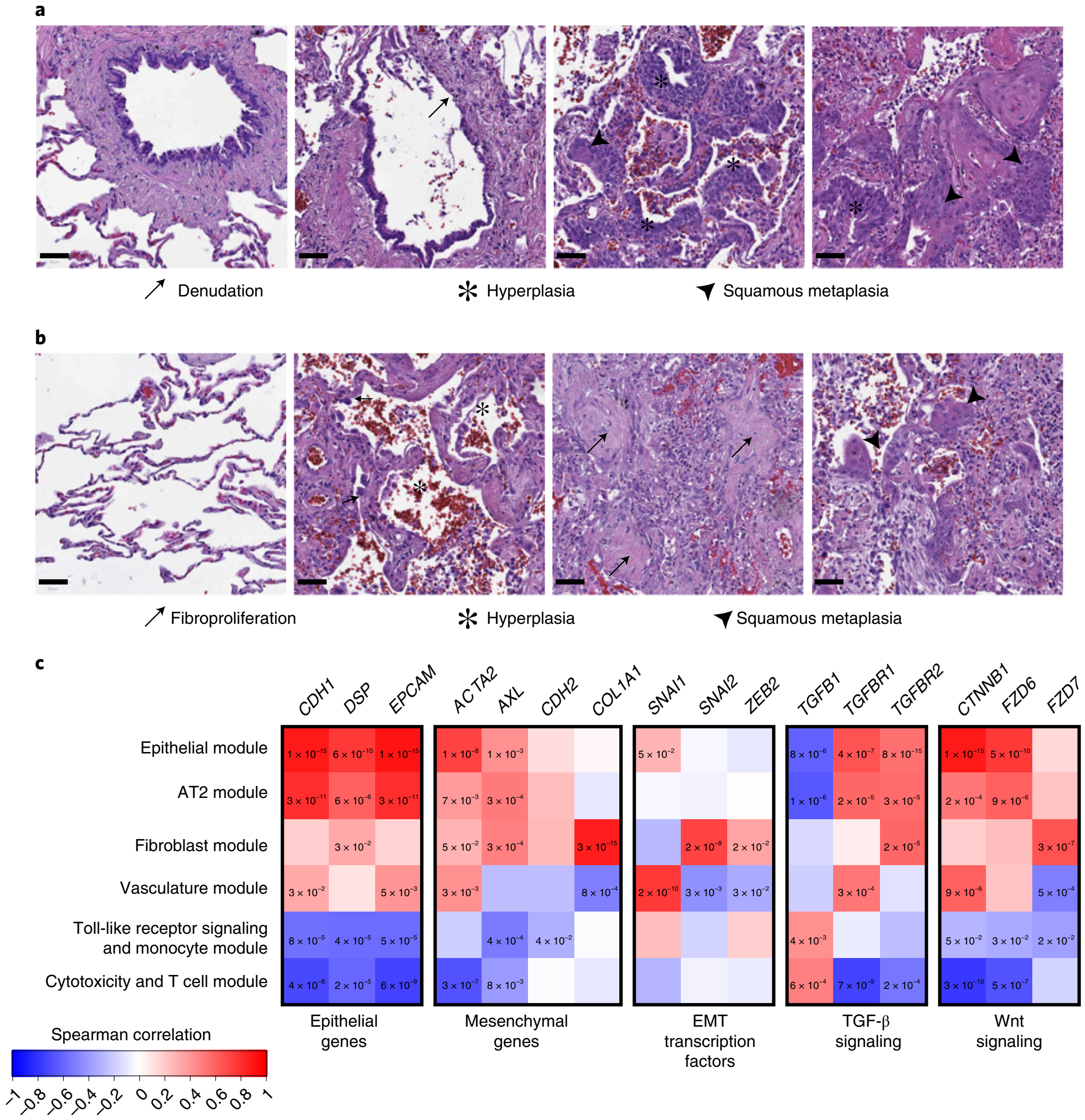


Fig. 5 | The lungs of patients with COVID-19 show signals of EMT. a,b, H&E-stained biopsies of the ciliated respiratory epithelium on bronchiole (**a**) and of alveolar space (**b**) in healthy lung (left) and the lung of patients with COVID-19 (middle and right). The samples of patients with COVID-19 are representative images from the staining of biopsies from three individuals and show loss of ciliated cell-lined bronchioles (denudation) and loss of alveolar monolayers populated by AT1 pneumocytes with few AT2 pneumocytes, with alveolar wall expansion and fine interstitial fibrosis. Scale bars, 50 µM. **c**, Spearman correlation of gene-expression profiles for EMT-related genes with the eigengenes of cell type modules identified by WGCNA analysis from spatially resolved expression data from the lung of patients with COVID-19. *P* values were identified by two-sided Hmisc analysis (without multiple test correction); values for significant correlations ($P < 0.05$) are shown and all correlation and *P* values are in the Source data.

profiles of EMT-relevant genes we used both a cell deconvolution approach[65], to estimate cell abundance through gene transcripts, and a weighted gene correlation network analysis[66] (WGCNA), to identify modules of coregulated gene-expression patterns that were assigned to cell types or biological processes. As expected, epithelial marker genes (*CDH1*, *EPCAM*) were naturally associated with alveolar type (AT) 1 and AT2 pneumocytes, as well as both of the epithelial and AT2 pneumocyte WGCNA modules (Fig. 5c and Extended Data Fig. 10). However, AT1 was also positively associated with the hallmark EMT gene *ACTA2* (actin alpha 2, smooth muscle; Hmisc rcorr asymptomatic $P = 0.0014$), as were both the AT2 and epithelial modules ($P = 0.0069$ and $P = 9.59 \times 10^{-9}$, respectively). These two modules were also positively associated with a second mesenchymal EMT marker gene, the receptor tyrosine kinase encoding *AXL* ($P = 0.0002$ and $P = 0.0031$). We next investigated EMT-associated transcription factors, finding *SNAI1* (snail family transcriptional repressor 1) positively associated with the epithelial module ($P = 0.0491$) and AT1 cells ($P = 0.0432$), while fibroblasts were associated with *SNAI2* ($P = 1.08 \times 10^{-6}$) and the fibroblast module was associated with both *SNAI2* ($P = 1.54 \times 10^{-8}$) and *ZEB2* (zinc finger E-box binding homeobox 2; $P = 0.0144$). Finally, we investigated the Wnt/β-catenin and TGF-β pathways, finding that both pneumocyte subtypes (AT1, AT2) and both epithelial modules were associated with TGF-β signaling receptor genes (*TGFBR1* and *TGFBR2*) and Wnt signaling genes that encode β-catenin and frizzled receptors (*CTNNB1* and *FZD6*). By contrast, neither CD8$^+$ T cells nor the cytotoxicity and T cell module expressed epithelial or mesenchymal genes but they expressed *TGFB1* ($P = 0.0029$ and $P = 0.0005$, respectively).

The colocalized expression of mesenchymal genes with epithelial cells, along with the expression of EMT transcription factors and associated signaling pathways is indicative of the EMT process, highlighting the relevance of this cellular reorganization pathway in COVID-19. Therefore, the modulation of EMT by LZTFL1 may be of relevance to the pathological outcome of COVID-19 infection.

Discussion

We applied a machine learning and molecular biology platform for decoding GWAS hits and identified a relatively unstudied gene, *LZTFL1*, as a candidate causal gene potentially responsible for the twofold increased risk of respiratory failure from COVID-19 associated with 3p21.31. The risk allele of the SNP, rs17713054 A, leads to increased transcription through augmentation of an epithelial–endothelial–fibroblast enhancer, facilitated by the addition of a second CEBPB binding motif.

MCC identified *LZTFL1* as the only gene to specifically interact with the rs17713054 enhancer. However, it is possible *LZTFL1* may not be the sole causal gene at 3p21.31. Two TWAS identified 11 candidate genes at this locus[10,49], including *LZTFL1* and *SLC6A20*, but only these two genes have strong 3C contacts with the rs17713054 enhancer and lung eQTLs. TWAS cannot differentiate between direct and indirect regulation[67]. The absence of a 3C interaction with COVID-19 severity-associated variants suggests that there may be an indirect effect for other genes, with the caveat that it is possible that a direct effect may occur in an untested cell type. While the ultrahigh resolution MCC approach only identified physical contacts between *LZTFL1* and rs17713054, traditional 3C found both *CCR9* and *SLC6A20* to be in the same regulatory domain. *CCR9* is not expressed in the lung and rs17713054 is not in an active enhancer in immune cells, where *CCR9* is expressed. Both *LZTFL1* and *SLC6A20* have higher expression in the presence of the rs17713054 risk allele; it is plausible that in cells where *SLC6A20* is not Polycomb-repressed (for example, goblet cells and AT2 pneumocytes), it also directly interacts with the rs17713054 enhancer and would thus be affected by the risk allele.

The biological relevance of *SLC6A20* to COVID-19 is unclear. It is primarily expressed in the kidneys and gastrointestinal tract and its associated Mendelian disease causes renal calculi due to failure of reuptake of glycine in the nephron[44]. Nevertheless, its function as an imino acid transporter is modulated by levels of angiotensin-converting enzyme 2 (ref. [68]) (ACE2), which is a cell receptor for SARS-CoV-2 (ref. [69]). Conversely, *LZTFL1* is widely expressed in pulmonary epithelial cells, including ciliated epithelial cells, which have been identified as one of the main cellular targets for SARS-CoV-2 infection[70]. Furthermore, homozygous loss of *LZTFL1* causes a classical ciliopathy––Bardet–Biedl syndrome[46,47]. The association of 3p21.31 variants with susceptibility to SARS-CoV-2 infection, as well as disease severity, highlights the importance of the respiratory epithelium for this locus[11]. *LZTFL1* encodes a cytosolic leucine zipper protein, which associates with the epithelial marker E-cadherin and is involved in the trafficking of numerous signaling molecules[45,71–74]. We note that upregulation of *LZTFL1* in the context of malignancy inhibits EMT[45,59,60], a pathway known to be part of both wound healing and immune responses[53–56].

Examination of postmortem COVID-19 lung biopsies demonstrated widespread epithelial dysfunction with EMT signatures[62,63]. Consistently, scRNA-seq showed a reduction in the total numbers of epithelial cells after infection[75], with a lower epithelial composition correlating with a more rapid progression from symptom onset to death[76]. The samples analyzed in this study showed few areas of healthy tissue and it is possible that inflammation or neutrophil extracellular traps, rather than direct viral infection, was driving this epithelial dysfunction[58] and that LZTFL1 acts earlier in disease progression, contributing to poor structural resolution of inflammation. Expression profiling of nasal epithelia from patients with COVID-19 detected EMT signals in the upper respiratory tract[57]. Similarly, SARS-CoV-2 infection of both a reconstructed human bronchial epithelium model and Syrian hamster induced dedifferentiation of airway ciliated cells[77], highlighting the relevance of this pathway and cell type. As such, an effect of the 3p21.31 locus in the early epithelial response may contribute to susceptibility to SARS-CoV-2 infection[11]. Although both influenza and SARS-CoV-2 have been shown to induce EMT[57,78], its role in viral infection is not entirely clear. While chronic EMT leads to fibrosis and severe inflammation, acute EMT may be a beneficial response. In the context of viral infection, EMT leads to a reduction of two of the cell receptors of SARS-CoV-2: ACE2 and transmembrane protease serine 2 (TMPRSS2) (refs. [57,79]). A reduction in these cell surface markers as a result of EMT could reduce viral load by decreasing infection efficiency and preventing severe disease. Conversely, EMT allows for epithelial cells to proliferate, repair damaged tissue and replace lost cells, which may be required to overcome severe disease.

For the 3p21.31 COVID-19 risk locus, higher risk is associated with increased expression of *LZTFL1*, a known EMT inhibitor. Higher levels of LZTFL1 may delay the positive effects of an acute EMT response, blocking a reduction in ACE2 and TMPRSS2 levels and/or through slowing EMT-driven tissue repair. Further investigation of the potential role of LZTFL1 and EMT in pulmonary pathogenesis is needed. Our findings suggest that a gain-of-function variant in an inducible enhancer, causing increased expression of *LZTFL1*, may be associated with a worse outcome. This raises the possibility that *LZTFL1* could be a potential therapeutic target for the treatment or prevention of COVID-19.

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/s41588-021-00955-3.

Received: 23 October 2020; Accepted: 22 September 2021;
Published online: 4 November 2021

References

1. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
2. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
3. Marini, J. J., Hotchkiss, J. R. & Broccard, A. F. Bench-to-bedside review: microvascular and airspace linkage in ventilator-induced lung injury. *Crit. Care* **7**, 435–444 (2003).
4. Levi, M., Thachil, J., Iba, T. & Levy, J. H. Coagulation abnormalities and thrombosis in patients with COVID-19. *Lancet Haematol.* **7**, e438–e440 (2020).
5. Varga, Z. et al. Endothelial cell infection and endotheliitis in COVID-19. *Lancet* **395**, 1417–1418 (2020).
6. Ackermann, M. et al. Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in COVID-19. *N. Engl. J. Med.* **383**, 120–128 (2020).
7. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
8. King, E. A., Wade Davis, J. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
9. Ellinghaus, D. et al. Genomewide association study of severe COVID-19 with respiratory failure. *N. Engl. J. Med.* **383**, 1522–1534 (2020).
10. Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
11. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature*, https://doi.org/10.1038/s41586-021-03767-x (2021).
12. Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).

13. Nakanishi, T. et al. Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. *J. Clin. Invest.*, https://doi.org/10.1172/JCI152386 (2021).
14. Nafilyan, V. et al. Ethnic differences in COVID-19 mortality during the first two waves of the Coronavirus Pandemic: a nationwide cohort study of 29 million adults in England. *Eur. J. Epidemiol.* **36**, 605–617 (2021).
15. Intensive Care National Audit & Research Centre. *COVID-19 in critical care: England, Wales and Northern Ireland* (2021).
16. Downes, D. J. et al. An integrated platform to systematically identify causal variants and genes for polygenic human traits. Preprint at *bioRxiv* https://doi.org/10.1101/813618 (2019).
17. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
18. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
19. Schwessinger, R. et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods* **17**, 1118–1124 (2020).
20. Davies, J. O. J. et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).
21. Downes, D. J. et al. High-resolution targeted 3C interrogation of *cis*-regulatory element organisation at genome-wide scale. *Nat. Commun.* **12**, 531 (2021).
22. Hua, P. et al. Defining genome architecture at base-pair resolution. *Nature* **595**, 125–129 (2021).
23. Robertson, C. C. et al. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nat. Genet.* **53**, 962–971 (2021).
24. Patsopoulos, N. A. et al. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).
25. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
26. Bruno, A. E. et al. miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3′UTRs of human genes. *BMC Genomics* **13**, 44 (2012).
27. Barenboim, M., Zoltick, B. J., Guo, Y. & Weinberger, D. R. MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. *Hum. Mutat.* **31**, 1223–1232 (2010).
28. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
29. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).
30. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
31. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
32. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
33. Bozhilov, Y. K. et al. A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker. *Nat. Commun.* **12**, 3806 (2021).
34. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
35. Wang, A. et al. Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. *eLife* **9**, e62522 (2020).
36. Phan, L. et al. *ALFA: Allele Frequency Aggregator* (National Center for Biotechnology Information, U.S. National Library of Medicine, 2020); www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/
37. Stolze, L. K. et al. Systems genetics in human endothelial cells identifies non-coding variants modifying enhancers, expression, and complex disease traits. *Am. J. Hum. Genet.* **106**, 748–763 (2020).
38. Hendricks-Taylor, L. R. et al. The CCAAT/enhancer binding protein (C/EBPα) gene (*CEBPA*) maps to human chromosome 19q13.1 and the related nuclear factor NF-IL6 (C/EBPβ) gene (*CEBPB*) maps to human chromosome 20q13.1. *Genomics* **14**, 12–17 (1992).
39. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
40. Schwessinger, R. et al. Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome Res.* **27**, 1730–1742 (2017).
41. Uehara, S., Grinberg, A., Farber, J. M. & Love, P. E. A role for CCR9 in T lymphocyte development and migration. *J. Immunol.* **168**, 2811–2819 (2002).
42. Liao, F. et al. STRL33, a novel chemokine receptor-like protein, functions as a fusion cofactor for both macrophage-tropic and T cell line-tropic HIV-1. *J. Exp. Med.* **185**, 2015–2023 (1997).
43. Agostini, C. et al. Role for CXCR6 and its ligand CXCL16 in the pathogenesis of T-cell alveolitis in sarcoidosis. *Am. J. Respir. Crit. Care Med.* **172**, 1290–1298 (2005).
44. Bröer, S. et al. Iminoglycinuria and hyperglycinuria are discrete human phenotypes resulting from complex mutations in proline and glycine transporters. *J. Clin. Invest.* **118**, 3881–3892 (2008).
45. Wei, Q. et al. Tumor-suppressive functions of leucine zipper transcription factor-like 1. *Cancer Res.* **70**, 2942–2950 (2010).
46. Zaghloul, N. A. & Katsanis, N. Mechanistic insights into Bardet-Biedl syndrome, a model ciliopathy. *J. Clin. Invest.* **119**, 428–437 (2009).
47. Marion, V. et al. Exome sequencing identifies mutations in *LZTFL1*, a BBSome and smoothened trafficking regulator, in a family with Bardet–Biedl syndrome with situs inversus and insertional polydactyly. *J. Med. Genet.* **49**, 317–321 (2012).
48. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
49. Pathak, G. A. et al. Integrative genomic analyses identify susceptibility genes underlying COVID-19 hospitalization. *Nat. Commun.* **12**, 4569 (2021).
50. Yao, Y. et al. Genome and epigenome editing identify *CCR9* and *SLC6A20* as target genes at the 3p21.31 locus associated with severe COVID-19. *Signal Transduct. Target. Ther.* **6**, 85 (2021).
51. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
52. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
53. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 69–84 (2019).
54. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).
55. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial–mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **15**, 178–196 (2014).
56. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial–mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).
57. Stewart, C. A. et al. Lung cancer models reveal severe acute respiratory syndrome coronavirus 2–induced epithelial-to-mesenchymal transition contributes to coronavirus disease 2019 pathophysiology. *J. Thorac. Oncol.* https://doi.org/10.1016/j.jtho.2021.07.002 (2021).
58. Pandolfi, L. et al. Neutrophil extracellular traps induce the epithelial–mesenchymal transition: implications in post-COVID-19 fibrosis. *Front. Immunol.* **12**, 663303 (2021).
59. Wei, Q. et al. LZTFL1 suppresses lung tumorigenesis by maintaining differentiation of lung epithelial cells. *Oncogene* **35**, 2655–2663 (2016).
60. Wang, L. et al. LZTFL1 suppresses gastric cancer cell migration and invasion through regulating nuclear translocation of β-catenin. *J. Cancer Res. Clin. Oncol.* **140**, 1997–2008 (2014).
61. Yang, J. et al. Guidelines and definitions for research on epithelial–mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **21**, 341–352 (2020).
62. He, J. et al. Single-cell analysis reveals bronchoalveolar epithelial dysfunction in COVID-19 patients. *Protein Cell* **11**, 680–687 (2020).
63. Borczuk, A. C. et al. COVID-19 pulmonary pathology: a multi-institutional autopsy cohort from Italy and New York City. *Mod. Pathol.* **33**, 2156–2168 (2020).
64. Cross, A. R. et al. Spatial transcriptomic characterization of COVID-19 pneumonitis identifies immune pathways related to tissue injury. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.21.449178 (2021).
65. Danaher, P. et al. Advances in mixed cell deconvolution enable quantification of cell types in spatially-resolved gene expression data. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.04.235168 (2020).
66. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
67. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
68. Singer, D. et al. Defective intestinal amino acid absorption in Ace2 null mice. *Am. J. Physiol. Gastrointest. Liver Physiol.* **303**, 686–695 (2012).
69. Vuille-dit-Bille, R. N. et al. Human intestine luminal ACE2 and amino acid transporter expression increased by ACE-inhibitors. *Amino Acids* **47**, 693–705 (2015).
70. Ravindra, N. G. et al. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. *PLoS Biol.* **19**, e3001143 (2021).
71. Promchan, K. & Natarajan, V. Leucine zipper transcription factor-like 1 binds adaptor protein complex-1 and 2 and participates in trafficking of transferrin receptor 1. *PLoS ONE* **15**, e0226298 (2020).

72. Starks, R. D. et al. Regulation of insulin receptor trafficking by Bardet Biedl syndrome proteins. *PLoS Genet.* **11**, e1005311 (2015).
73. Wei, Q. et al. Lztfl1/BBS17 controls energy homeostasis by regulating the leptin signaling in the hypothalamic neurons. *J. Mol. Cell Biol.* **10**, 402–410 (2018).
74. Seo, S. et al. A novel protein LZTFL1 regulates ciliary trafficking of the BBSome and Smoothened. *PLoS Genet.* **7**, e1002358 (2011).
75. Melms, J. C. et al. A molecular single-cell lung atlas of lethal COVID-19. *Nature* **595**, 114–119 (2021).
76. Delorey, T. M. et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595**, 107–113 (2021).
77. Robinot, R. et al. SARS-CoV-2 infection induces the dedifferentiation of multiciliated cells and impairs mucociliary clearance. *Nat. Commun.* **12**, 4354 (2021).
78. Ruan, T. et al. H1N1 influenza virus cross-activates Gli1 to disrupt the intercellular junctions of alveolar epithelial cells. *Cell Rep.* **31**, 107801 (2020).
79. 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.e8 (2020).

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

Methods

Human research ethics compliance. All samples and information were collected with written and signed informed consent. For erythroid cells, peripheral blood was obtained with approval from the North West Research Ethics Committee of the NHS National Research Ethics Services (03/08/097). Blood samples for CD4$^+$ cells were obtained from donors recruited from the Cambridge BioResource. The study was approved by the East of England––Cambridgeshire and Hertfordshire Research Ethics Committee (05/Q0106/20). CD14$^+$ samples were isolated from healthy donors with approval from the Oxfordshire Research Ethics Committee COREC (06/Q1605/55). Patient samples were acquired and analyzed with approval from the ethics committee of the University of Navarra (15/05/2020) and the Medical Sciences Interdivisional Research Ethics Committee of the University of Oxford (approval no. R76045/RE001). Hematopoietic stem and progenitor cells from healthy donors were stored in accordance with the Human Tissue Authority (license no. 12433).

Cell isolation, culture and stimulation. The H1-hESC (https://scicrunch.org/resolver/CVCL_9771) WA01 WiCell cell line (research resource identifier (RRID):CVCL_9771) was grown on Matrigel-coated (Corning) plates in mTeSR1 medium (STEMCELL Technologies). Cells were collected as a single-cell suspension using Accutase (Merck Millipore); fixation was carried out in mTeSR1 medium. Primary neonatal HUVECs (catalog no. CC-2517, Lonza; catalog no. C0035C, Gibco; catalog no. C-12200, PromoCell) were expanded in endothelial cell growth medium (Sigma-Aldrich) up to five passages according to the manufacturer's protocol. For passaging, HUVECs were grown to 60% confluence, washed with Hanks' Balanced Salt Solution at room temperature and subcultured after light trypsinization using trypsin-EDTA (Sigma-Aldrich) at room temperature with trypsin inhibitor (Sigma-Aldrich) added on rounding of the cells to achieve gentle release from the flask. HUVECs were fixed in Roswell Park Memorial Institute (RPMI) 1640 supplemented with 10% FCS. For erythroid cells, CD34$^+$ hematopoietic stem and progenitor cells were isolated from the peripheral blood of 2 healthy males and 1 healthy female and differentiated ex vivo for 13 d as described previously[82]. CD4$^+$ T cells were enriched from whole blood (93–99% pure, RosetteSep Human CD4$^+$ T Cell Enrichment Cocktail; STEMCELL Technologies) and were plated at 250,000 cells per well in U-96 well plates (Greiner) and cultured in medium alone or stimulated with anti-CD3/CD28 T-activator beads (Dynabeads; Thermo Fisher Scientific) at a ratio of 0.3 beads per cell for 4 h at 37 °C in X-VIVO 15 (Lonza), 1% AB serum (Lonza) and penicillin-streptomycin (Thermo Fisher Scientific). Nonactivated or activated CD4$^+$ T cells were pooled after 4 h of culture and fixed in growth medium. For CD14$^+$ cells, peripheral blood mononuclear cells (PBMCs) were obtained by Ficoll-Paque (GE Healthcare) density centrifugation of whole blood collected into EDTA tubes (BD Vacutainer system) or leukocyte cones (NHS Blood and Transplant). Monocyte isolation was carried out by positive selection using magnetic-activated cell sorting with CD14$^+$ beads (Miltenyi Biotec) according to the manufacturer's instructions. IMR-90 (https://scicrunch.org/resolver/CVCL_0347) lung fibroblasts (CCL-186, RRID:CVCL_0347; ATCC) were cultured in Eagle's minimal essential medium supplemented with 10% FCS, 1 mM of sodium pyruvate (Gibco), 1× MEM nonessential amino acids (Gibco) and penicillin-streptomycin (100 U ml^{-1} each). Cells were subcultured every 3 d after light trypsinization using 0.05% trypsin-EDTA (Gibco). Blood outgrowth endothelial cells (BOECs) were isolated as described previously[83]. Briefly 20–40 ml of fresh blood was diluted 1:1 with PBS, layered over Histopaque-1077 (Sigma-Aldrich) and centrifuged for 15 min at 500 *g*, brake off. PBMCs were washed with PBS then resuspended in EGM-2 BulletKit growth medium (Lonza) supplemented with 10% heat-inactivated FCS. Cells were cultured for 21–28 d in collagen-coated flasks until BOEC colonies formed. BOEC colonies were passaged by light trypsinization. BOEC cells were passaged twice before any experimentation to ensure endothelial cell purity, which was also confirmed by FACS and immunofluorescence. BOEC cells were fixed in growth medium. NCI-H441 (https://scicrunch.org/resolver/CVCL_1561) cells (HTB-174, RRID:CVCL_1561; ATCC) were grown in RPMI 1640 medium (Gibco) supplemented with 10% non-heat-inactivated FCS (Sigma-Aldrich) and 1% penicillin-streptomycin (Gibco); cells were given fresh medium every 2 d and passaged by light trypsinization twice weekly. Human umbilical derived erythroid progenitor line 2 cells[84] (HUDEP-2 (https://scicrunch.org/resolver/CVCL_VI06), RRID:CVCL_VI06) were provided by RIKEN and were maintained at 0.7–1.5 × 10^6 cells ml^{-1} in HUDEP expansion medium (serum-free expansion medium, 50 ng ml^{-1} of stem cell factor, 3 IU ml^{-1} of erythropoietin, 10 µM of dexamethasone, 1% L-glutamine, 1% penicillin-streptomycin) and changed into fresh medium containing 2× doxycycline every 2 d.

Variant effect sequence predictions. Linkage analysis was determined using the LDlink web tool v.5.1 (LDproxy, LDpair; https://ldlink.nci.nih.gov/). Candidate variants either achieved genome-wide significance in the first COVID-19 GWAS[9] or were in tight linkage ($r^2 > 0.8$) with lead variants from the first two large COVID-19 GWAS[9,10]. The deepHaem convolutional neural network[19] was trained with 4,384 ENCODE peaks calls (694 open chromatin DNase I/ATAC-seq, 1,750 transcription factor ChIP-seq and 1,940 histone modification ChIP-seq) and is available via GitHub (model 4; https://github.com/rschwess/deepHaem). Identification of CEBPB motifs was performed by Find Individual Motif Occurrences (FIMO)[85] analysis of reference and variant containing the enhancer sequence (chr3:45,817,661–45,818,660, hg38) with the JASPAR[86] motif MA0466.1. Sasquatch[40] was run using the default Workflow 3 settings (v1.0, 7-mer, propensity-based (erythroid), exhaustive) on the web interface (https://sasquatch.molbiol.ox.ac.uk/cgi-bin/foot.cgi). Masked SpliceAI[18] predictions for each variant were extracted from the coding genome scan for substitutions, 1-base insertions and 1–4 base deletions (https://github.com/Illumina/SpliceAI). Conserved miRNA binding sites were identified using TargetScan[25] (v7.2, http://www.targetscan.org/vert_71/). SNP predictions were identified using the miR*dSNP*[26] database (v11.03, http://mirdsnp.ccr.buffalo.edu/browse-genes.php) and the MicroSNiPer[27] web tool (release 19, http://vm24141.virt.gwdg.de/services/microsniper/index.php), using 6-mer, 7-mer, 8-mer and 9-mer settings.

Colocalization analysis. Harmonized summary statistics for severe COVID-19 (ref. [9]) were downloaded from the GWAS Catalog[87] (GCST90000256). Summary statistics for all lung eQTL–variant pairs (V8) in individuals with European-American ancestry were downloaded from the GTEx portal[28]. Coloc[51] v.5.0.1 analysis of variants within 200 kb of the predicted causal variant (rs17713054) was implemented in R. Inputs of GWAS size ($n = 3,795$), GWAS case frequency (0.419), eQTL study size ($n = 515$) and association β, s.e.m., MAFs and *z*-scores were used in a sensitivity analysis[88] that showed a prior probability of colocalization (p_{12}) of 1×10^{-5} tested approximately equal prior probability of both H_3 (two distinct causal variants for the GWAS and eQTL trait) and H_4 (a single causal variant).

3C. Gene promoters were selected for Capture-C using 10-kb resolution Hi-C data on the 3D Genome Browser[89] (http://3dgenome.fsm.northwestern.edu/index.html) from a range of cell types to identify putative regulatory domains and interactions with rs17713054. Capture-C was performed as described previously with either the NG or NuTi method[20,21,90]. Briefly 5–20 million cells were fixed with 2% formaldehyde and 3C libraries were generated using the high-resolution DpnII enzyme. Targeted enrichment was performed using SeqCap reagents (Roche) and 100-mer biotinylated oligonucleotides (Supplementary Table 2) at the optimal titrated concentration[21]. Libraries were sequenced using 75 bp paired-end reads on an Illumina NextSeq Platform to generate over 250,000 reads per viewpoint per sample. For MCC[22], aliquots of 1–2 × 10^7 cells were fixed for 10 min with 2% formaldehyde in 10 ml of growth medium. Formaldehyde was quenched with 125 mM of glycine and cells were pelleted (5 min, 500 *g*, 4 °C) and washed with PBS. Cells were resuspended in 1 ml of PBS and permeabilized with 0.005% digitonin. Cells were pelleted and resuspended in 800 µl of reduced calcium content micrococcal nuclease buffer (10 mM of Tris-HCl, pH 7.5, 1 mM of $CaCl_2$). Chromatin was digested for 1 h at 37 °C inside intact, permeabilized cells in three separate reactions using 5–120 Kunitz units of micrococcal nuclease (New England Biolabs). Digestion was quenched by with 5 mM of EGTA (Sigma-Aldrich). Cells were pelleted and washed with PBS before end-repair and phosphorylation; cells were resuspended in 400 µl of DNA ligase buffer (Thermo Fisher Scientific) supplemented with 400 µM of each of deoxyATP, deoxyCTP, deoxyGTP and deoxyTTP and 5 mM of EGTA, 200 U ml^{-1} of T4 Polynucleotide Kinase (New England Biolabs) and 100 U ml^{-1} DNA Polymerase I, Large (Klenow) Fragment (New England Biolabs) for 2 h at 37 °C. To ligate DNA fragments, T4 DNA ligase (Thermo Fisher Scientific) was added at 300 U ml^{-1} and the reaction was incubated at room temperature for 8 h. Chromatin was de-crosslinked with proteinase K at 65 °C for over 4 h and DNA was extracted using either phenol chloroform with RNase treatment (Roche) and ethanol precipitation or using the DNeasy Blood and Tissue Kit (QIAGEN). MCC libraries were sonicated to 200-bp fragments and indexed using NEBNext Ultra II indexing reagents (New England Biolabs) with the following modifications: 2 µg of DNA was indexed; 5 µl of adapter was used; bead cleanups were performed with 1.5 volumes of AMPure XP beads; and Herculase II PCR reagents (Agilent) were used for the indexing PCR. Target enrichment was performed using double capture with 120-bp biotinylated oligonucleotides (Supplementary Table 3) with SeqCap Reagents (Roche). Enriched libraries were sequenced on the NextSeq platform using 150-bp paired-end reads to generate approximately 1 M reads per viewpoint.

3C data analysis. NuTi Capture-C data were mapped to the hg38 using CCseqBasicS[91] (v5, https://github.com/Hughes-Genome-Group/CCseqBasicS) using Bowtie 2. Briefly, CCseqBasic5 (ref. [92]) trims adapter sequences, flashes read pairs, digests fragments in silico and uses map reads before identifying sequences as either capture and reporter. Replicates were compared using CaptureCompare[93] (v1, https://github.com/Hughes-Genome-Group/CaptureCompare), which normalizes *cis* reporter counts per 100,000 *cis* reporters, generates per-fragment mean counts for each cell type and then bins reporter counts in equally sized regions to generate a windowed profile. For MCC, adapters were removed using TrimGalore[94] v.0.3.1, then fragments were reconstructed with FLASH[95] v.1.2.11 into single sequences using the central area of overlapping reads. Fragments were mapped to the oligonucleotide DNA sequence ±350 bp using BLAT[96] v.35 to identify ligation junctions, allowing splitting of reads into new paired FASTQ files using MCCsplitter.pl v1 and subsequent mapping to hg38 with Bowtie 2 (ref. [97]) v.2.3.5. PCR duplicates were removed from the alignment files with MCCanalyser.pl v1 using both sonicated ends and ligation junction with a wobble of ±2 bp. MCCsplitter.pl and MCCanalyser.pl are available for academic use through the Oxford University Innovation software store (https://process.innovation.ox.ac.uk/software/p/16529a/micro-capture-c-academic/1). MCC tissue-specific peaks for rs17713054 were called using LanceOtron[98] on the web tool 'Find and Score Peaks with Inputs' (v2, https://lanceotron.molbiol.ox.ac.uk) using the HUDEP-2 MCC profile as an input track.

Genome editing. For the deletion of the rs17713054 enhancer, cells were transfected with 5 μg of Alt-R S.p. Cas9 nuclease V3 RNP (Integrated DNA Technologies) and 0.1 nmol each of two guide RNAs (Supplementary Table 4). All transfections were carried out with $1–2 \times 10^5$ cells in 20-μl reactions using a 4D-Nucleofector (Lonza); IMR-90 fibroblast cells were electroporated using Amaxa Cell Line Nucleofector Kit V reagents (Lonza) with program CM-120. HUVECs and BOECs were electroporated using Amaxa P5 Primary Cell 4D-Nucleofector X Kit S reagents (Lonza) with program CA-167 and H441 epithelial cells were electroporated using P3 Primary Cell 4D-Nucleofector X Kit S reagents (Lonza) with program EL-10. Cells were cultured for 24 h in 2 ml of antibiotic-free growth medium in a single well of a 6-well plate before expansion in fully supplemented media. Bulk DNA was extracted using the DNeasy Blood and Tissue Kit and the edited region (chr3:45,817,769–45,818,459; hg38) was amplified using the Platinum PCR SuperMix (Invitrogen) with 5′-GGAAAGAACACGCATAAACCATA-3′ (forward primer) and 5′-CTCATCCCACAGTGAACTAAGAA-3′ (reverse primer). Editing efficiency was determined using a D1000 TapeStation and Sanger sequencing with the forward primer and Synthego ICE analysis (https://ice.synthego.com/#/).

RT–qPCR. For expression analysis, cells were grown to >80% confluence in a single well of a 6-well plate. Cells were lysed by adding 1 ml of TRI Reagent (Sigma-Aldrich), snap-frozen and stored at −80 °C for less than 6 months. RNA was separated by adding 100 μl of 1-bromo-3-chloropropane, centrifuged in a Phase Lock Gel Heavy tube (5Prime) for 5 min at 10,000 *g* and precipitated in an equal volume of isopropanol (500 μl) with 1 μl of GlycoBlue (Thermo Fisher Scientific). DNA was removed using the DNA-free DNA Removal Kit (Invitrogen) and complementary DNA (cDNA) was generated using 1 μg of total RNA with SuperScript III First-Strand Synthesis SuperMix reagents (Thermo Fisher Scientific). qPCR was performed using a 1:10 dilution of cDNA, TaqMan Universal PCR Master Mix II without UNG (Thermo Fisher Scientific) and TaqMan Gene Expression Assays (Thermo Fisher Scientific) for *LZTFL1* (Hs00947898_m1), *SLC6A20* (Hs00610960_m1) and *RPL18* (Hs00965812_g1) with FAM dye label. *LZTFL1* expression was normalized to *RPL18* and relative expression calculated by normalizing to the mean expression of *LZTFL1* in RNP-treated cells from samples of the same cell type processed in the same batch.

ChIP-seq. For ChIP-seq, single-cell suspensions of 10^6 cells ml^{-1} in growth medium were generated after light trypsin treatment. Cells were fixed by adding 1% formaldehyde for 10 min at room temperature, which was quenched by adding glycine at a final concentration of 125 mM. Fixed cells were washed with PBS and snap-frozen. Cell lysis and immunoprecipitation was carried out using the ChIP Assay Kit (Merck Millipore) on 5×10^6 cells in 2 ml of dilution buffer incubated overnight at 4 °C with 1 μl of rabbit polyclonal anti-H3K27ac (1:2,000 dilution; catalog no. ab4729, 0.3 μg, Abcam). DNA was isolated by phenol/chloroform isoamyl alcohol extraction and ethanol precipitation then indexed using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs). Libraries were sequenced using 39-bp paired-end reads on a NextSeq platform. Reads were mapped to hg38 using Bowtie 2 (ref. [97]), PCR duplicates filtered using SAMtools[99] and BigWig files generated with deepTools[100] v2.2.2.

FACS analysis. For FACS, approximately 10^5 cells were resuspended in 100 μl of staining buffer (PBS with 10% FCS) and incubated with 1 μl each of allophycocyanin-conjugated mouse anti-CD14 (1:100 dilution, 2 ng, clone M5E2, catalog no. 301807; BioLegend), phycoerythrin-conjugated mouse anti-CD309/VEGFR2 (1:100 dilution, 2 ng, clone 7D4-6, catalog no. 359903; BioLegend), fluorescein isothiocyanate (FITC)-conjugated mouse anti-CD31/PECAM (1:100 dilution, 2 ng, clone WM59, catalog no. 303103; BioLegend) and PE/Cyanine7-conjugated mouse anti-CD34 (1:100 dilution, 0.5 ng, clone 561, catalog no. 343616; BioLegend) for 20 min at 4 °C. Cell were diluted with 90 μl of staining buffer with 1:5,000 Hoechst 33258 (Thermo Fisher Scientific) and analyzed on an Attune NxT Flow Cytometer. Voltages and compensation were set using single-stain samples with UltraComp eBeads (Thermo Fisher Scientific) for antibodies and cells for Hoechst. Negative and positive populations were established using fluorescence minus one controls. Mononuclear cells were gated using forward scatter (FSC) and side scatter; single cells were gated using FSC-area and FSC-height and live cells were selected using a Hoechst-negative gate in FlowJo v.10.7.

ATAC-seq. ATAC-seq was performed as published elsewhere[101,102] with 7.5×10^4 cells per technical replicate and 2–4 technical replicates per samples. After spinning at 500 *g* for 15 min, cells were resuspended in lysis buffer (10 mM of Tris-HCl, pH 7.5, 10 mM of NaCl, 3 mM of $\mathrm{MgCl_2}$, 0.1% IGEPAL CA-630), centrifuged and nuclei washed with PBS. Nuclei were pelleted, PBS was discarded and nuclei were resuspended in tagmentation buffer (25 μl of 2× tagmentation DNA buffer, 2.5 μl of Tn5 Transposase (Illumina) and 22.5 μl of water) then incubated at 37 °C for 30 min. After transposition DNA was extracted using the MinElute PCR Purification Kit (QIAGEN), half the DNA was amplified for sequencing using the NEBNext High-Fidelity 2× PCR Master Mix (New England Biolabs) and further purified with the QIAquick PCR Purification Kit (QIAGEN). Libraries were sequenced using 39-bp paired-end reads on a NextSeq platform. Reads were mapped to hg38 using Bowtie 2 in NGseqBasic[102] v20.

Immunofluorescence staining and microscopy. Cells were grown for 24–48 h on sterilized coverslips under standard growth conditions and fixed in 4% vol/vol paraformaldehyde in 0.25 M of HEPES for 15 min, followed by permeabilization in 0.2% vol/vol Triton X-100 in PBS for 10 min. After blocking with 10% vol/vol FCS in PBS, von Willebrand's factor was detected using mouse anti-von Willebrand's factor 1:100 (clone F8/86, catalog no. MA5-14029; Invitrogen) and goat anti-mouse Alexa Fluor 488 1:500 (catalog no. A32723; Thermo Fisher Scientific). DNA was stained with 1 μg ml^{-1} of 4,6-diamidino-2-phenylindole (DAPI) in PBS; after washing, coverslips were mounted in VECTASHIELD (Vector Laboratories). Widefield fluorescence imaging was performed on a DeltaVision Elite system (Applied Precision) using a Universal Plan Fluorite 40× 1.30 numerical aperture oil immersion objective (Olympus), a CoolSnap HQ2 charge-coupled device camera (Photometrics) and DAPI (excitation 390/18, emission 435/40) and FITC (excitation 475/28, emission 525/45) filters; 12-bit image stacks were acquired with a z-step of 200 nm giving a voxel size of 161.3 × 161.3 × 200 nm. All images were acquired using the same exposure settings. Using Fiji[103] v2.1.0, three-dimensional images were flattened by maximum intensity projection and displayed at the same minimum/maximum intensity settings. Images were cropped for publication in Adobe Photoshop v.22.4.1.

Patients tissue analyses. Healthy lung samples were sourced from patients with chronic obstructive pulmonary disease during lung tumor resection, with a sample of normal lung acquired away from the tumor. The medical records of patients with COVID-19 were reviewed retrospectively[104] and 3 were selected for in-depth analysis based on their clinical manifestation of acute respiratory distress syndrome, typical COVID-19 histology (with a 4–5 score on the Brescia-COVID Respiratory Severity Scale) and a lung-restricted (absence in heart, liver and kidney biopsies) presence of SARS-CoV-2. Postmortem lung tissues were obtained through open biopsy shortly after death and processed as described previously[104]. Briefly, tissues were immediately fixed in neutral-buffered formalin for <24 h and then paraffin-embedded. Sections (5 μm each) were cut from wedge biopsies (mean size = 1.78 cm^2, s.d. = 0.55 cm^2) for hematoxylin and eosin (H&E) analysis. Sections were analyzed by NanoString GeoMx Digital Spatial Profiling with normalization and downstream analysis by WGCNA[66] and cell deconvolution[65] as described previously[64]. For deconvolution with SpatialDecon in R v.1.0.0, cell profiles were obtained from the Human Cell Atlas healthy lung and scRNA-seq-appended with neutrophil data[105] using the R 'Lung_plus_neut' dataset. Seven relevant cell types were selected for expression analysis from a total of 26 cell types. WGCNA was performed using the WGCNA R package v.1.70-3 and generated 17 biologically assignable modules of which 6 were selected for further analysis. Spearman correlation and unadjusted *P* value generation was performed with the Hmisc R package v.4.5-0 and visualized with corrplot v.0.84.

Public dataset analysis. Unless stated, ENCODE datasets were accessed using the UCSC Genome Browser[106,107], which was also used to generate track figures. ENCODE DNase I BigWig files (hg38) were downloaded from the ENCODE portal (https://www.encodeproject.org/) and analyzed with deepTools[100] (multiBigwigSummary; https://deeptools.readthedocs.io/en/develop/content/tools/multiBigwigSummary.html). Capture-C was analyzed using the CaptureCompendium suite v1[91] mapping to hg38 with Bowtie 2 (ref. [97]) and using default settings. ATAC-seq and H3K27ac ChIP-seq data from erythroid progenitors, immune cells[29,80,81] and aortic endothelium[37] were downloaded from the Gene Expression Omnibus (GEO) (accession nos. GSE74912, GSE115684, GSE118189, GSE139377) and analyzed using NGseqBasic[102] with default settings for Bowtie 2 (ref. [97]). Aortic endothelial samples were genotyped by counting two or more reads from either allele in the combined ATAC-seq and ChIP-seq data. For allelic skew analysis, aortic endothelium ATAC-seq from heterozygous individuals was mapped with Bowtie 2 (ref. [97]) and processed using WASP v0.3.4[108] to correct for reference genome mapping bias. Three replicates with fewer than four remaining reads were excluded from the analysis. Mature erythroid chromatin modification and CTCF data (GSE125926) were previously reported by our group[16], CTCF motifs were identified using the MEME Suite[85] tools (v5.3.0, meme--dna--nmotifs 1--w 19--mod zoops--maxsize 1102788; fimo--thresh 1e-4--motif 1). scRNA-seq data[35,48] were sourced from online portals (Lung Cell Atlas https://asthma.cellgeni.sanger.ac.uk/, Gene Expression Profiling https://www.lungepigenome.org/gene-expression/) on 9 October 2020 and 19 May 2021, respectively. scATAC-seq data[34,35] were sourced from online portals (descartes https://descartes.brotmanbaty.org/bbi/human-chromatin-during-development/dataset/lung, Lung Genome Browser https://www.lungepigenome.org/) on 19 May 2021. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by the National Cancer Institute, National Human Genome Research Institute, National Heart, Lung, and Blood Institute, National Institute on Drug Abuse, National Institute of Mental Health and National Institute of Neurological Disorders and Stroke. The multi-tissue eQTL and expression level data were obtained from the GTEx Portal V8 on the 14 October 2020 (https://gtexportal.org/home/snp/rs17713054).

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

Data availability

Capture-C, Micro Capture-C, ATAC-seq and ChIP-seq data generated for this study (Fig. 3, Extended Data Figs. 7 and 9 and Supplementary Figs. 1 and 2) are available from the GEO under accession nos. GSE159867 and GSE175791). Processed Capture-C data can be visualized on the UCSC Genome Browser (http://datashare.molbiol.ox.ac.uk//datashare/project/fgenomics/publications/Downes_2021_Covid_GWAS/hub.txt) or on the CaptureSee website (https://capturesee.molbiol.ox.ac.uk/projects/capture_compare/3718). Numerical values for Figs. 2a–c and 5d and Extended Data Figs. 2–4, 6, 7, 9 and 10 are available in the Source data. Expression data (Fig. 3 and Extended Data Figs. 6 and 8) was from publicly available sources: GTEx Portal (https://gtexportal.org); Lung Cell Atlas (https://asthma.cellgeni.sanger.ac.uk/); and Lung Genome Browser (https://www.lungepigenome.org/). Publicly available open chromatin data (ATAC-seq/DNase-seq), transcription factor binding data (ChIP-seq) and epigenetic modification (ChIP-seq) data (Figs. 1 and 2, Extended Data Figs. 1, 2, 4–7 and 9 and Supplementary Figs. 1 and 2) were sourced from the ENCODE portal (https://www.encodeproject.org/), the GEO (accession nos. GSE74912, GSE115684, GSE118189, GSE125926), the UCSC Genome Browser (https://genome.ucsc.edu), descartes Human Chromatin Accessibility during Development atlas (https://descartes.brotmanbaty.org/bbi/human-chromatin-during-development/); and the Lung Genome Browser. Masked splicing prediction effects were downloaded from the SpliceAI database (https://github.com/Illumina/SpliceAI). The CEBPB motif (MA0466.1) was downloaded from the JASPAR database (http://jaspar.genereg.net). Conserved miRNA sites were identified on miR*dSNP* (http://mirdsnp.ccr.buffalo.edu/browse-genes.php). Source data are provided with this paper.

Code availability

All custom analysis code and links to software are available on GitHub (https://github.com/Hughes-Genome-Group/Downes_2021_LZTFL1_Covid.git). MCCsplitter.pl and MCCanalyser.pl are only available for academic use through the Oxford University Innovation software store (https://process.innovation.ox.ac.uk/software/p/16529a/micro-capture-c-academic/1).

References

80. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).

81. Ludwig, L. S. et al. Transcriptional states and chromatin accessibility underlying human erythropoiesis. *Cell Rep.* **27**, 3228–3240.e7 (2019).

82. Scott, C. et al. Recapitulation of erythropoiesis in congenital dyserythropoietic anaemia type I (CDA-I) identifies defects in differentiation and nucleolar abnormalities. *Haematologica*, https://doi.org/10.3324/haematol.2020.260158 (2020).

83. Martin-Ramirez, J., Hofman, M., van den Biggelaar, M., Hebbel, R. P. & Voorberg, J. Establishment of outgrowth endothelial cells from peripheral blood. *Nat. Protoc.* **7**, 1709–1715 (2012).

84. Kurita, R. et al. Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. *PLoS ONE* **8**, e59890 (2013).

85. Bailey, T. L. et al. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).

86. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

87. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

88. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, e1008720 (2020).

89. Wang, Y. et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).

90. Downes, D. J. & Hughes, J. R. Chromosome conformation capture with nuclear titrated Capture-C (NuTi Capture-C). *Protoc. Exch.* https://doi.org/10.21203/rs.3.pex-1244/v1 (2020).

91. Telenius, J. M. et al. CaptureCompendium: a comprehensive toolkit for 3C analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2020.02.17.952572 (2020).

92. Telenius, J. M., Davies, J. O. J. & Hughes, J. R. *Hughes-Genome-Group/CCseqBasicS: Release for DOI* https://zenodo.org/record/4196777#.YWQkYBDMKWY (2020).

93. Downes, D. J. et al. *CaptureCompare* https://zenodo.org/record/4194345#.YWQk_hDMKWY (2020).

94. Krueger, F. Trim Galore https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2015).

95. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

96. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

97. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

98. Hentges, L. D., Sergeant, M. J., Downes, D. J., Hughes, J. R. & Taylor, S. LanceOtron: a deep learning peak caller for ATAC-seq, ChIP–seq, and DNase-seq. Preprint at *bioRxiv* https://doi.org/10.1101/2021.01.25.428108 (2021).

99. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

100. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

101. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

102. Telenius, J. M. & Hughes, J. R. NGseqBasic––a single-command UNIX tool for ATAC-seq, DNaseI-seq, Cut-and-Run, and ChIP–seq data mapping, high-resolution visualisation, and quality control. Preprint at *bioRxiv* https://doi.org/10.1101/393413 (2018).

103. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

104. Recalde-Zamacona, B. et al. Histopathological findings in fatal COVID-19 severe acute respiratory syndrome: preliminary experience from a series of 10 Spanish patients. *Thorax* **75**, 1116–1118 (2020).

105. Desai, N. et al. Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nat. Commun.* **11**, 6319 (2020).

106. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

107. Rosenbloom, K. R. et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).

108. van de Geijn, B., Mcvicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).

Acknowledgements

We thank P. Hirschfeld, M. Gosden, S. Carpenter, C. Harrold, L. Nussbaum, Y. Bozhilov, A. King, M. Badat, M. Salmon, L. Hentges, A. Brown, G. Scozzafava, A. Lledó Lara, T. Rostron, J. Torres, C. Eijsbouts, V. Iotchkova and M. Sergeant for their help with cell culture, experiments and data analysis, R. Beagrie for critical reading of the manuscript and D. Higgs for nurturing this research. This GWAS approach was developed as part of the Wellcome Investigation of Genome Wide Association Mechanisms Consortium funded by a Wellcome Strategic Award (no. 106130/Z/14/Z to J.R.H.). This work was also supported by Medical Research Council (MRC) Core Funding (no. MC_UU_00016/14 to J.R.H.). D.J.D. received funding from the Oxford University Medical Science Internal Fund: Pump Priming (no. 0006152). R.S. was supported by a Wellcome Doctoral Programme (no. 203728/Z/16/Z). S.N.S. received Kennedy Trust for Rheumatology Research Core support (no. KENN171803). F.I. received support from the Wellcome (no. 211122/Z/18/Z). J.O.J.D. is funded by an MRC Clinician Scientist Award (no. MR/R008108) and received Wellcome support (no. 098931/Z/12/Z). J.C.K. is a Wellcome Investigator (no. WT204969/Z/16/Z) and is supported by the National Institute for Health Research Oxford Biomedical Research Centre and the Chinese Academy of Medical Sciences Innovation Fund for Medical Science (no. 2018-I2M-2-002). A.M.M., D.R.G. and S.C.H. received support from the Wellcome (no. 110579/Z/15/Z). J.A.T. is supported by a Strategic Award from the Wellcome (no. 107212/Z/15/Z), Juvenile Diabetes Research Foundation (nos. 5-SRA-2015-130-A-N and 4-SRA-2017-473-A-N) and Wellcome Core Award to the Wellcome Centre for Human Genetics (no. 203141/Z/16/Z). C.E.d.A. and I.M. were supported by the Banco Bilbao Vizcaya Foundation 'Ayudas a Equipos de Investigación Científica SARS-CoV-2 y COVID-19'. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

D.J.D., A.J.C., O.M., N.R. and A.M.M. isolated, cultured and fixed the cells and processed 3C material. D.J.D., P.H., A.R.C., J.B., C.E.d.A., I.M., F.I. and J.O.J.D. designed and performed the experiments. D.J.D., P.H., A.R.C., R.S., J.B. and S.N.S. analyzed the data.

C.E.d.A., I.M., D.R.G., S.C.H., J.C.K., J.A.T., F.I. and J.R.H. acquired the funding and oversaw the project. D.J.D., J.O.J.D. and J.R.H. conceived the work, generated the figures and wrote the manuscript.

Competing interests

J.R.H. and J.O.J.D. are founders and shareholders of Nucleome Therapeutics; J.R.H., J.O.J.D., D.J.D. and R.S. are paid consultants for Nucleome Therapeutics. J.R.H. and J.O.J.D. hold patents for Capture-C (nos. WO2017068379A1, EP3365464B1 and US10934578B2) and have a patent application for MCC. J.A.T. is a member of the GSK Human Genetics Advisory Board. The other authors declare no competing interests.

Additional information

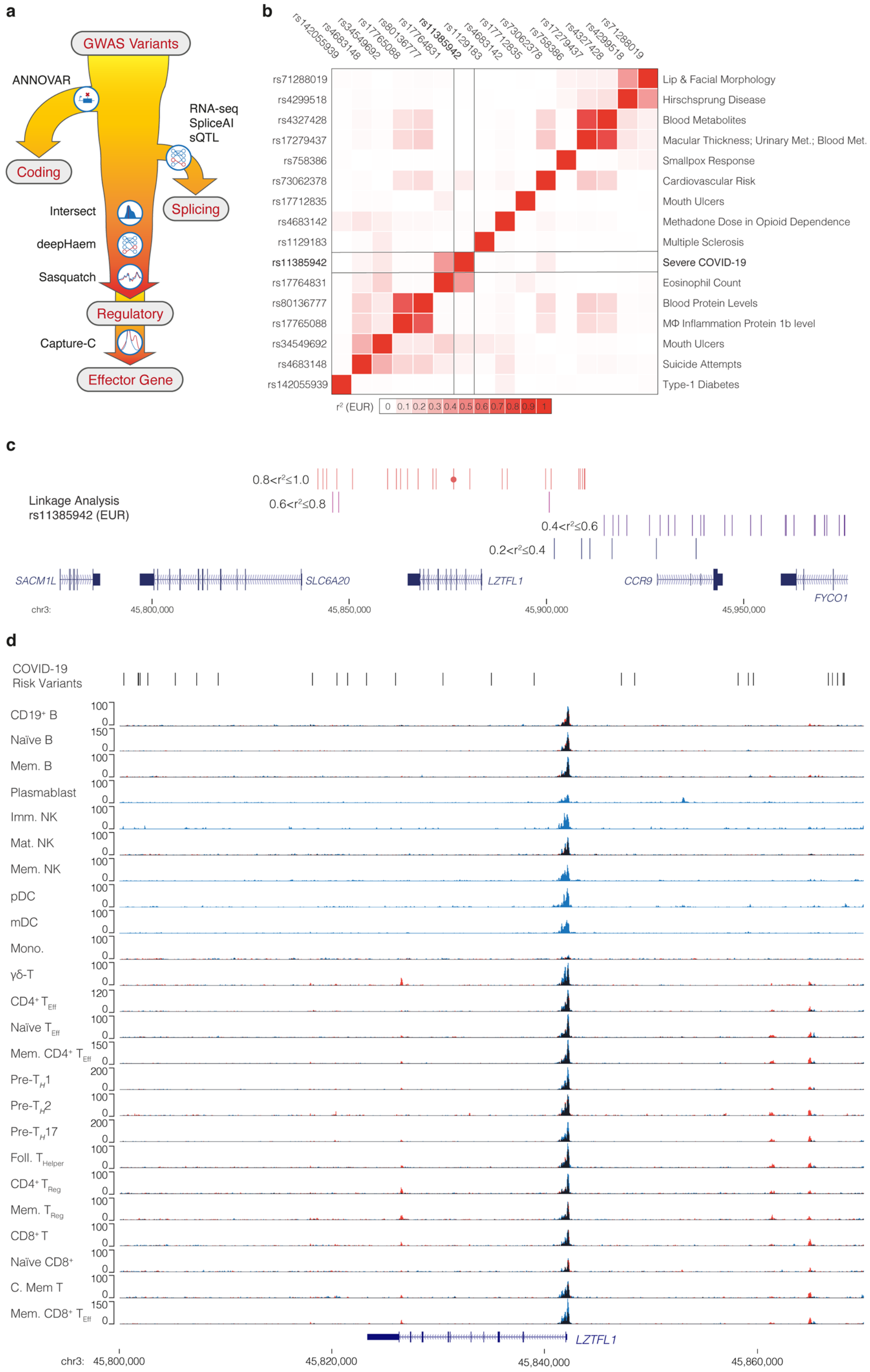
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00955-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00955-3.

Correspondence and requests for materials should be addressed to James O. J. Davies or Jim R. Hughes.

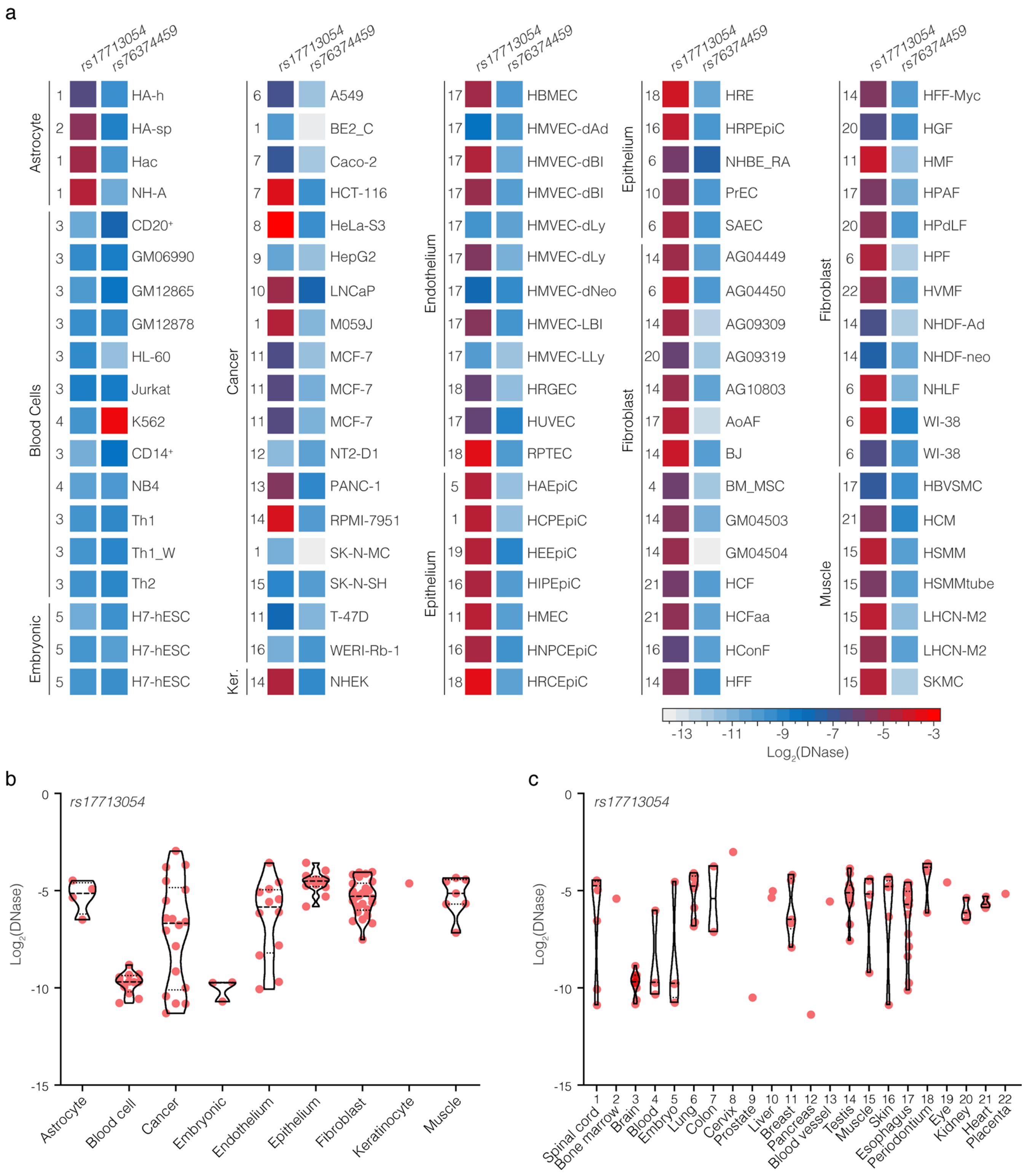
Peer review information *Nature Genetics* thanks Luis Barreiro and the other anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

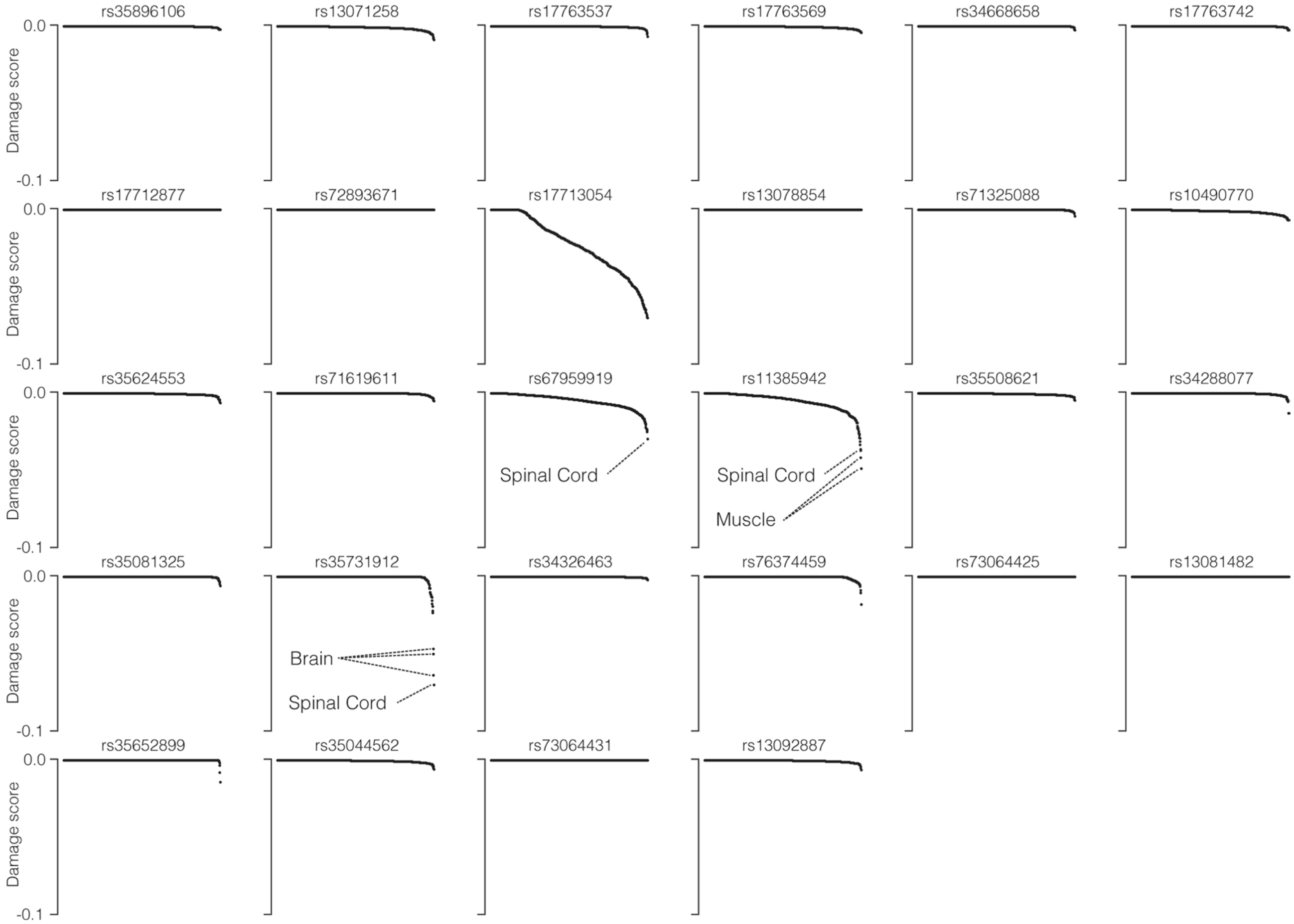


Extended Data Fig. 1 | See next page for caption.

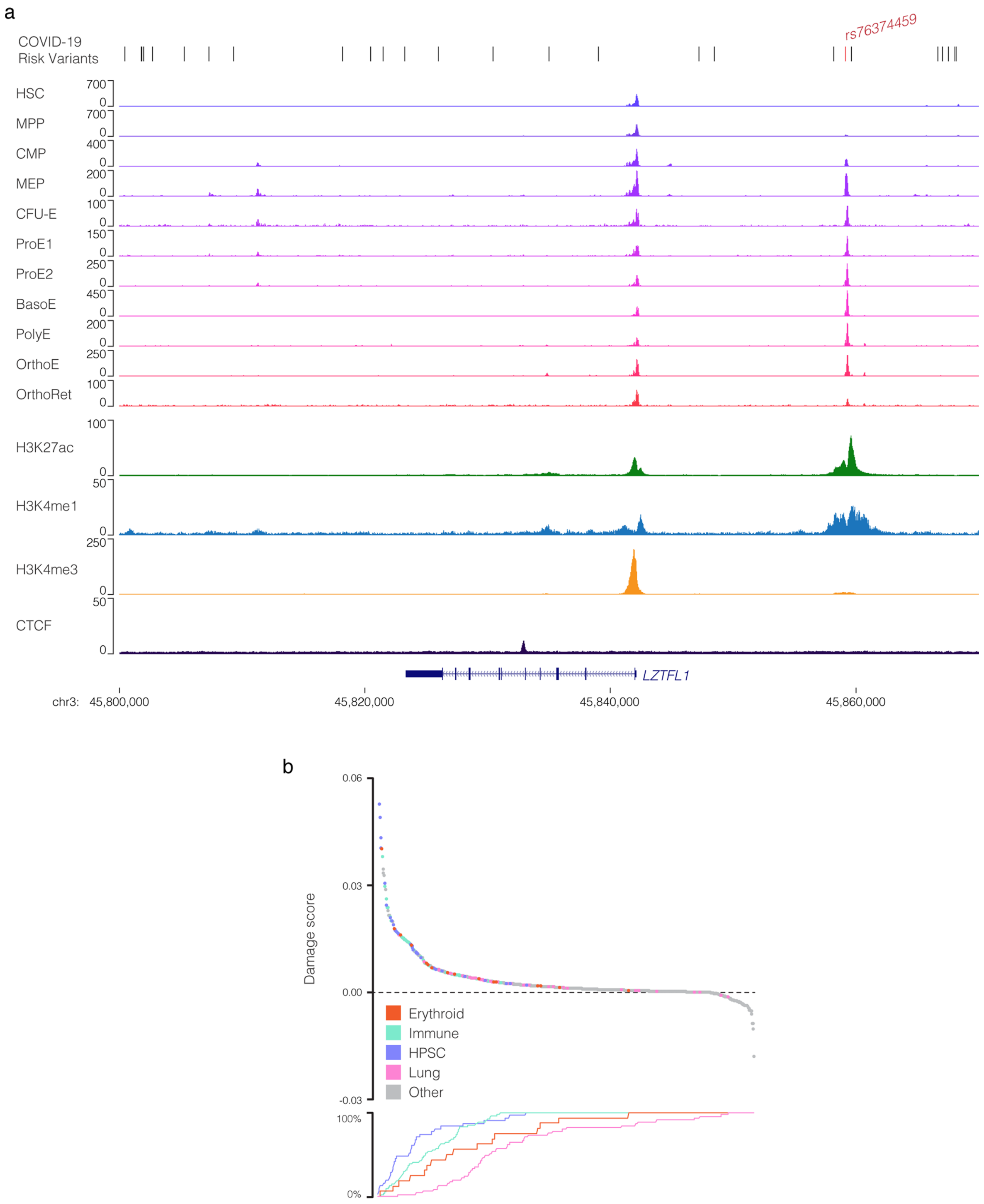
Extended Data Fig. 1 | 3p21.31 severe COVID-19 locus SNPs are not in immune regulatory elements. a, To decode GWAS variants either all genome wide significant variants and/or variants in linkage disequilibrium with sentinel variants are assessed for protein coding changes with ANNOVAR. Remaining variants are then assessed for changes in splicing of expressed genes using the SpliceAI machine learning approach[18] or splicing quantitative trait loci (sQTL). Variants are then intersected with open chromatin with a panel of disease relevant cell types to asses *cis*-regulatory element altering potential. This potential is assessed for effects on open chromatin with deepHaem[19] or transcription factor binding with both deepHaem and Sasquatch[40]. Finally, variants in enhancers are linked to target effector genes using high resolution chromosome conformation capture with NG/NuTi Capture-C[20,21] or Micro Capture-C[22]. **b**, Heatmap of linkage disequilibrium (European; EUR) between a severe COVID-19 lead SNP (rs11385942) with lead SNPs for other GWAS traits identified in the region (chr3:45,710,500-45,954-500, hg38). **c**, Linkage analysis for a 3p21.31 severe COVID-19 lead SNP (rs11385942 - circle) showing variants within 100 kb and $r^2 > 0.2$. No variants with $r^2 > 0.6$ were seen beyond this range. **d**, Overlaid tracks of ATAC-seq from sorted populations of resting (blue) and stimulated (red) immune cells[29]. Overlapping signal appears black. Abbreviations: Memory (Mem.), Immature (Imm.), Mature (Mat.), Natural Killer cells (NK), Plasmacytoid Dendritic cells (pDC), Myeloid Dendritic cells (mDC), Monocytes (Mono.), Effector (Eff.), Helper (H.), Regulatory (Reg.), and Central (C.). Region: chr3:45,800,000-45,870,000, hg38.



Extended Data Fig. 2 | DNase I accessibility over COVID-19 SNPs. a. DNase I signal in each of 95 ENCODE datasets for rs17713054 (chr3:45,817,661-45,818,660, hg38) and rs7634459 (chr3:45,859,001-45,859,500, hg38) which were found in open chromatin. Datasets are grouped according to cell-type, numbers indicate tissue of origin (see panel c). Violin plots of ENCODE DNase I accessibility over rs17713054 grouped by cell type **(b)** and tissue of origin **(c)**. Each sample is shown as a red dot, dashed lines show mean, dotted lines show quartiles.

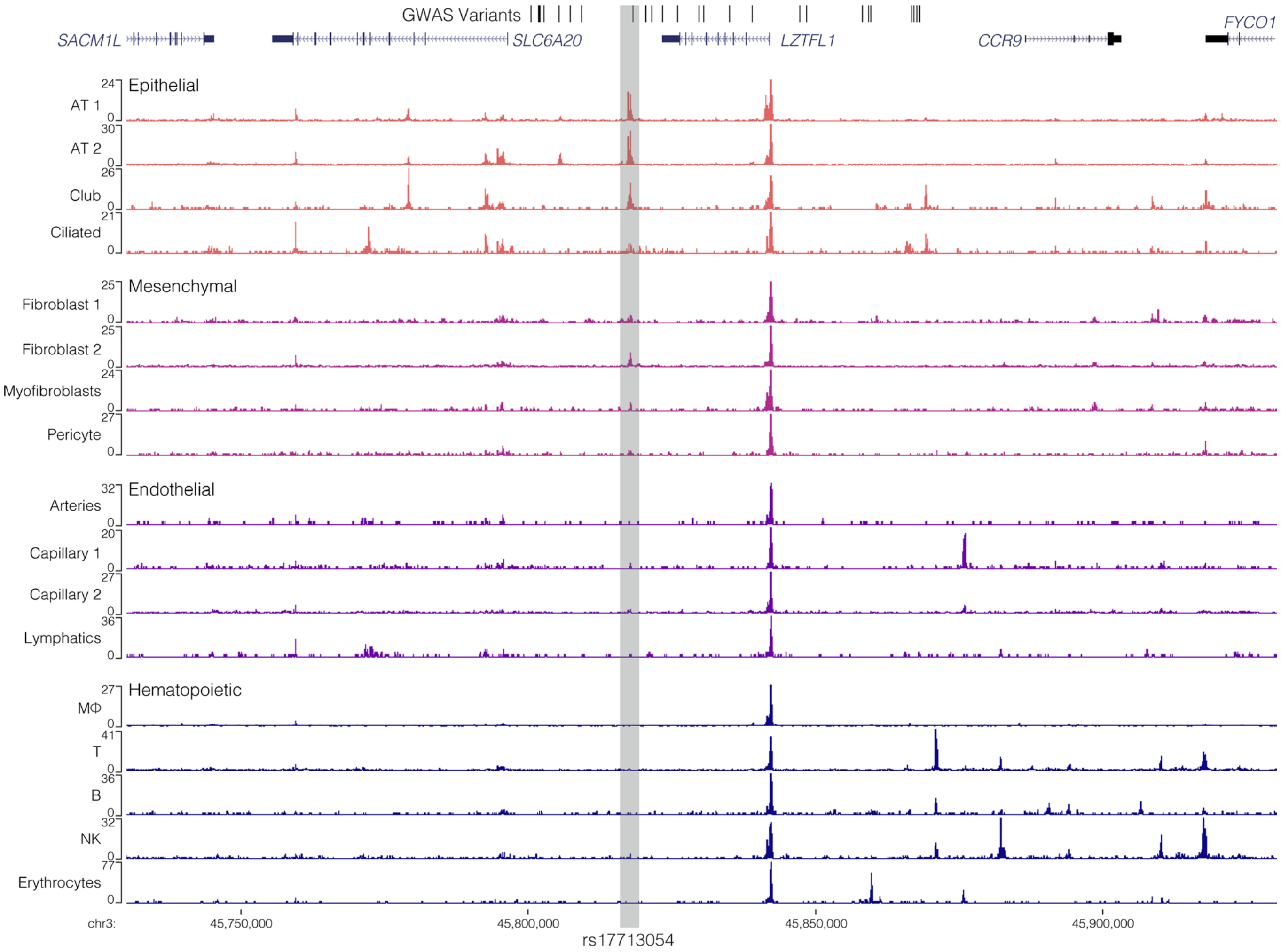


Extended Data Fig. 3 | deepHaem prediction of de novo open chromatin elements. deepHaem[19] negative damage scores, which predict gain-of-accessibility, for the 28 candidate COVID-19 severity variants in 694 cell-types. Positive scores (loss-of-function) were adjusted to zero. In general, variants generating *de novo* regulatory elements[33] have scores lower than -0.1, which was not true for any variant in any cell type.

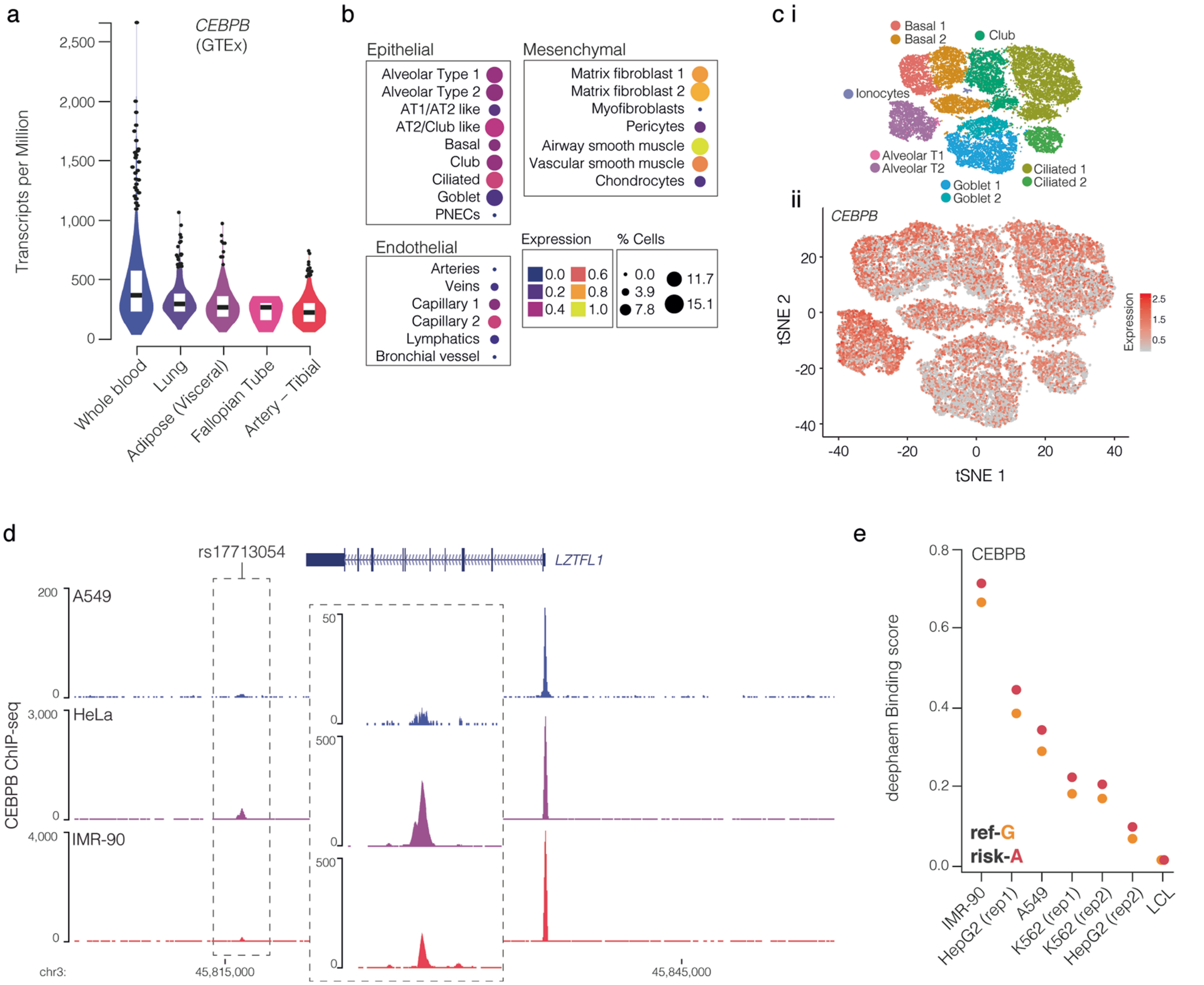


Extended Data Fig. 4 | See next page for caption.

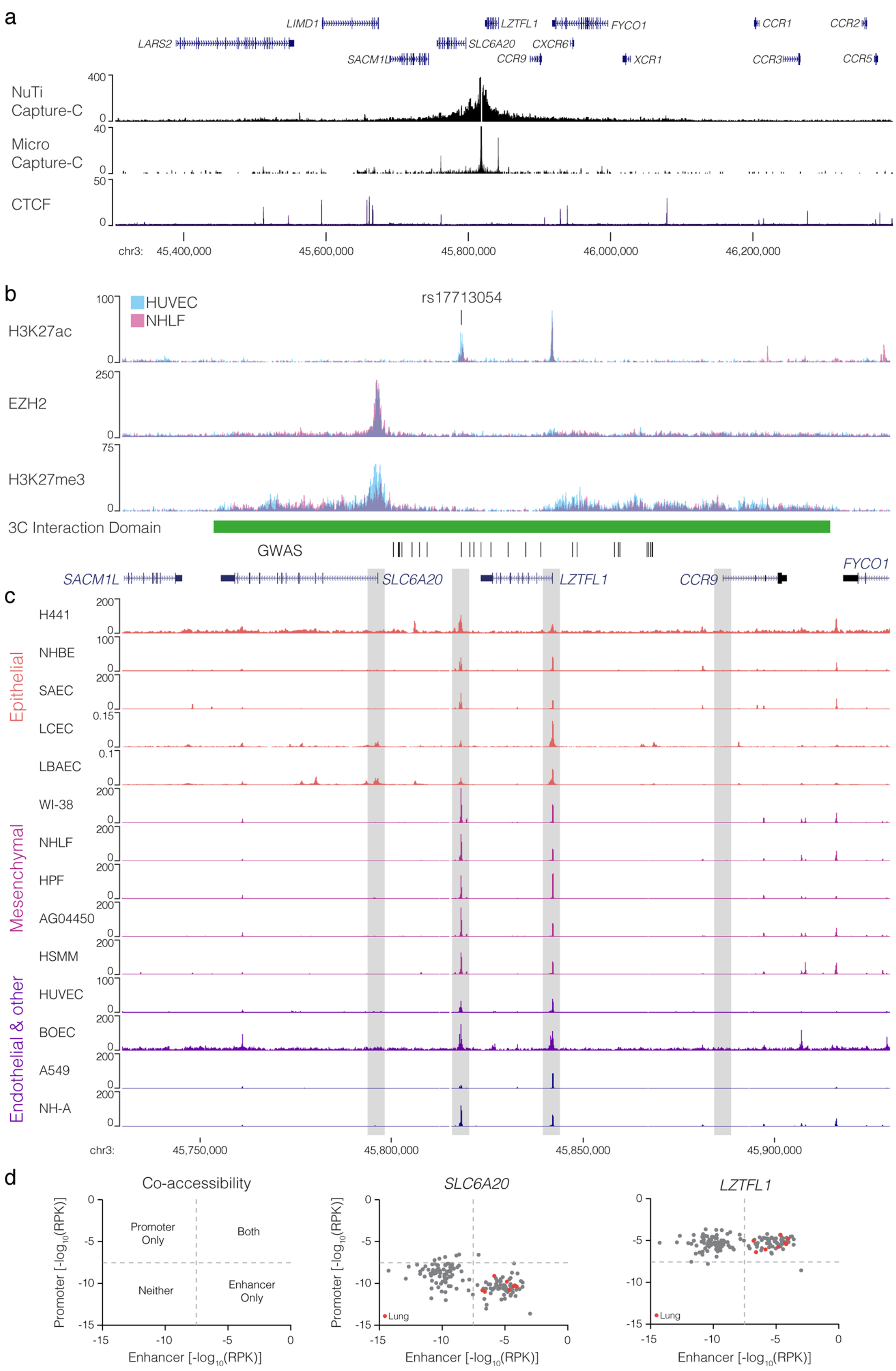
Extended Data Fig. 4 | rs76374459 is likely benign in an erythroid enhancer. ATAC-seq from progenitor[80] and differentiating erythroid cells[81]. Haematopoietic Stem Cells (HSC), Multi-Potent Progenitors (MPP), Common Myeloid Progenitors (CMP), Myeloid-Erythroid Progenitors (MEP) from bone marrow or peripheral blood and erythroid Colony Forming Units (CFU-E), Pro-erythroblasts (ProE1, ProE2), Basophilic Erythroblasts (BasoE), Polychromatic Erythroblasts (PolyE), Orthochromatic Erythroblasts (OrthoE) and Orthochromatic/Reticulocytes (OrthoRet). ChIP-seq tracks from CD71$^+$ CD23$^+$ mature erythroid cells[16] show presence of marks associated with active transcription (H3K27ac), enhancers (H3K4me1), promoters (H3K4me3) and boundaries (CTCF). **b**, deepHaem damage score for the risk-C allele versus non-risk-G allele of rs76374459 associated with severe COVID-19 in 694 cell-types. rs763774458 is found in open chromatin through-out erythropoiesis. A positive score predicts loss of accessibility, a negative score predicts increased accessibility.



Extended Data Fig. 5 | Single nucleus ATAC-seq in adult lungs. Chromium single nucleus ATAC-seq from non-diseased adult lung[35] (n = 3) with 17 epithelial, endothelial, mesenchymal and hematopoietic populations, including Alveolar Type (AT) 1 and 2 Pneumocytes, Macrophage (MΦ) and Natural Killer (NK) cells. The rs17713054 containing element is highlighted in grey.

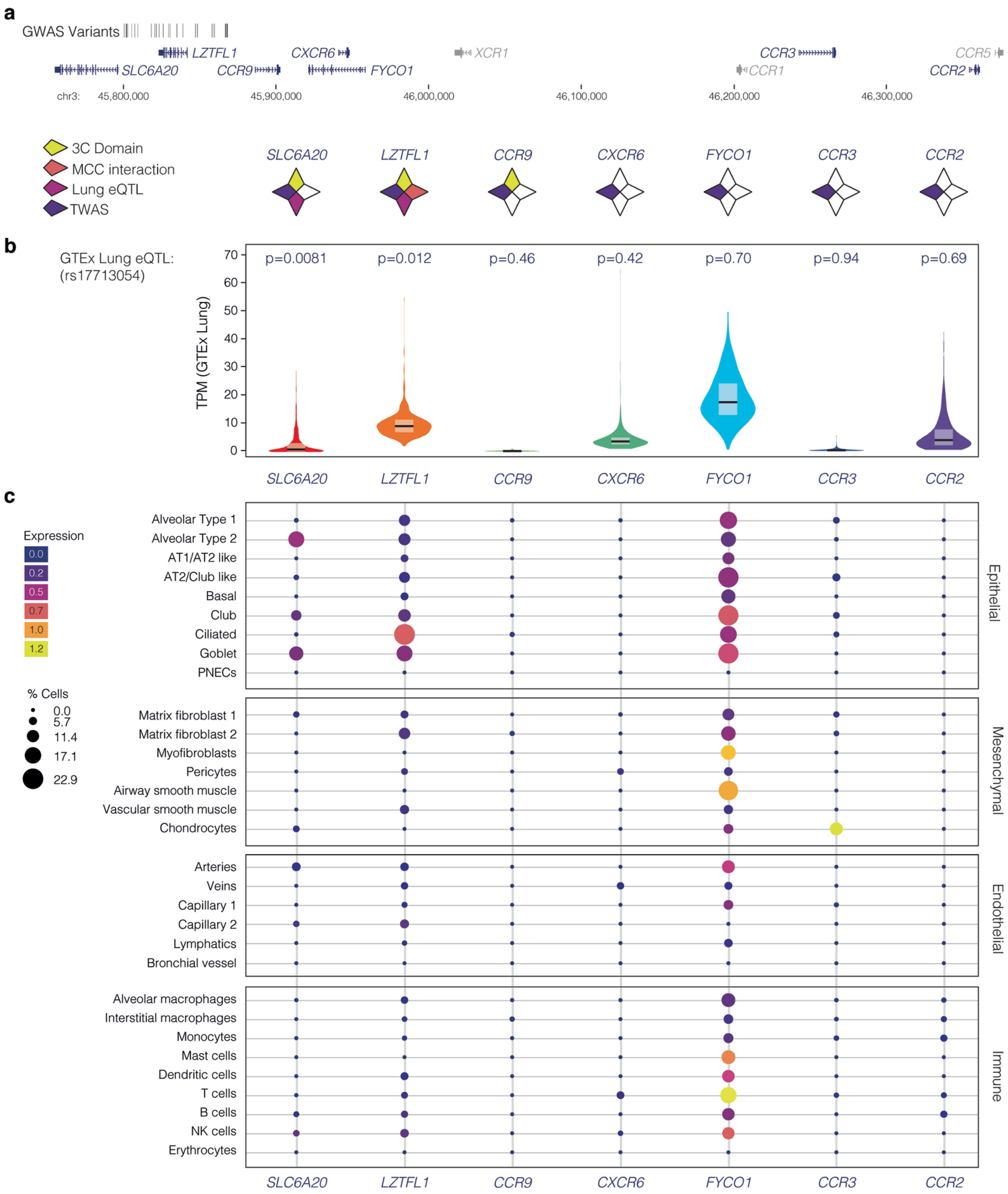


Extended Data Fig. 6 | Pulmonary expression and binding analysis of CEBPB. **a**, GTEx top five expressed tissues for CEBPB. For violin plots, minima and maxima are the top and bottom of the violin, black lines show means, ends of the pale regions denote first and third quartiles, and black dots denote outliers. Data from independent samples for Whole blood (n = 755), Lung (n = 578), Adipose (n = 541), Fallopian Tube (n = 9), Artery (n = 663). **b**, Chromium single nucleus RNA-seq from non-diseased adult lung[35] (n = 3 independent samples) with 22 epithelial, endothelial and mesenchymal populations, including Alveolar Type (AT) 1 and 2 Pneumocytes and Pulmonary Neuroendocrine cells (PNECs). **c**, 10x Genomics Chromium droplet single-cell RNA sequencing (scRNA-seq) from upper and lower airways and lung parenchyma[34] from healthy volunteers or deceased transplant donors with ten epithelial populations **(i)** with expression profiles for *CEBPB* **(ii)**. **d**, ENCODE ChIP-seq for CEBPB in A549 alveolar basal epithelial adenocarcinoma cells, HeLa cells, and IMR-90 lung fibroblast cells with inset region (chr3:45,805,000-45,855,000; hg38) showing the rs17713054 containing enhancer. **e**, DeepHeam ChIP-seq binding prediction score for CEBPB in lung fibroblast (IMR-90), alveolar basal epithelial adenocarcinoma (A549), the erythroleukaemia line (K562), human endothelial kidney cells (HEK293), and the GM12878 lymphoblastoid cell line (LCL) predicts increased binding to the risk-A allele.

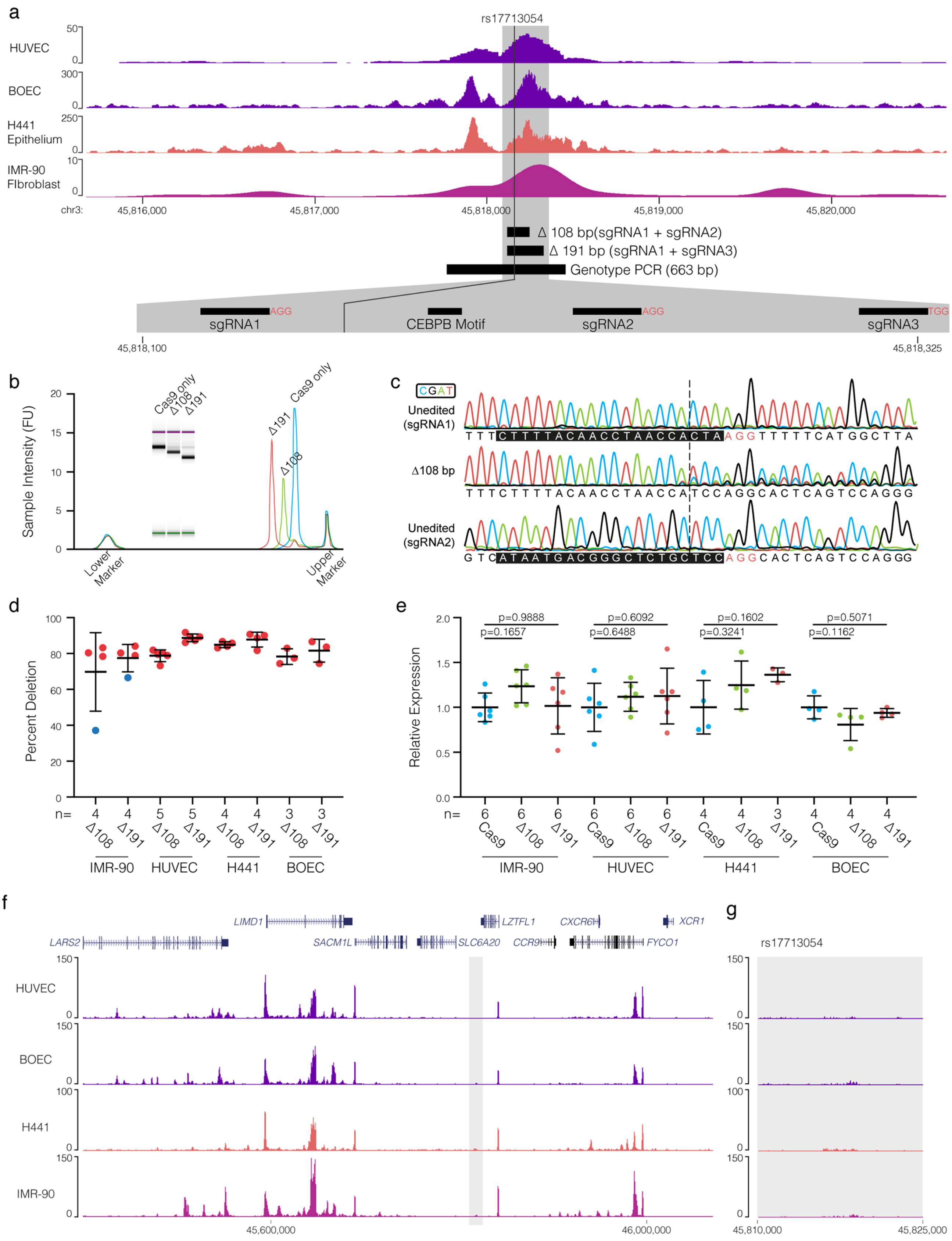


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | LZTFL1 is a direct target of rs17713054. a, NuTi Capture-C and Micro Capture-C from the rs17713054 enhancer in Endothelial cells (HUVEC) shows specific interaction with only the promoter of *LZTFL1* and an upstream CTCF site. CTCF track shows binding of the CCCTC-binding factor which acts as a boundary. **b**, ENCODE ChIP-seq for the active chromatin mark (H3K27ac), the repressive chromatin mark (H3K27me3) and EZH2, a member of the Polycomb Repressive Complex 2, in endothelial (HUVEC) and normal human lung fibroblast (NHLF) cells. Green bar denotes the 3C regulatory domain as identified by 3 C analysis. **c**, ENCODE DNase I seq tracks from a range of cell types and tissues, including airway epithelium and bronchial epithelium, where the rs17713054 enhancer is active. In these cell types the *LZTFL1* promoter is DNase I accessible, but neither the *CCR9* promoter nor the *SLC6A20* promoter are. Region shown is chr3: 45,730,000-45,930,000 (hg38). **d**, Paired accessibility analysis of read counts per kilobase (RPK) over the *LZTFL1* and *SLC6A20* promoters and the rs17713054 enhancer in 156 ENCODE, immune and erythroid open chromatin datasets. Only the *LZTFL1* promoter is widely accessible in the same cells as the affected enhancer.

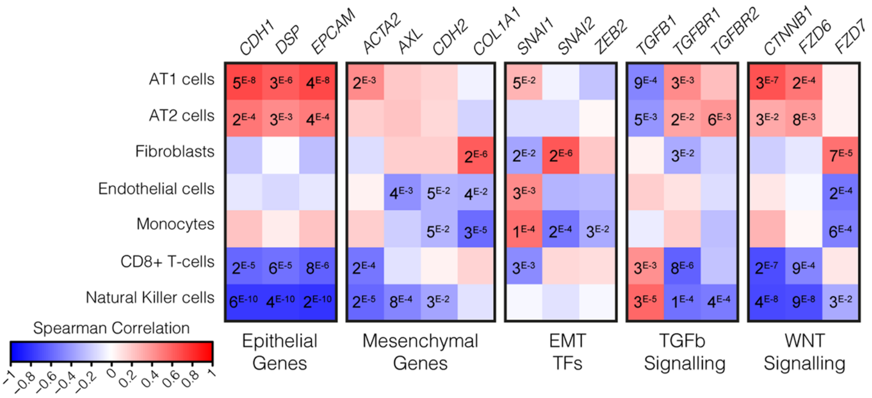


Extended Data Fig. 8 | Expression and eQTL analysis of 3p21.31 candidate lung effector genes. **a**, Genomic position of genes identified as 3p21.31 candidate causal genes with method of identification, including two TWAS[10,49]. **b**, GTEx whole lung RNA-seq expression profiles for candidate causal genes as transcripts per million (TPM) with rs17713054 eQTL two-sided *P* value for lung. For violin plots, minima and maxima are the top and bottom of the violin, black lines show means, ends of the pale regions denote first and third quartiles, and black dots denote outliers. n = 578 independent samples. **c**, Chromium single nucleus RNA-seq[35] from non-diseased adult lung (n = 3), including Alveolar Type 1 (AT1) and Type 2 (AT2) Pneumocytes and Pulmonary Neuroendocrine cells (PNECs).



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

Extended Data Fig. 9 | CRISPR-Cas9 deletion of the rs17713054 enhancer. a, ENCODE DNase I-seq in HUVEC and IMR-90 cells and ATAC-seq in Blood Outgrowth Endothelial Cells (BOECs) and H441 epithelial cells showing the rs17713054 containing enhancer with schematic of generated deletions and short guide RNA (sgRNA) binding sites. **b**, Example D1000 trace of genotyping PCR product amplified from cells transfected with Cas9 protein only, Cas9 protein with sgRNA1+2 (Δ108), or Cas9 protein with sgRNA1+3 (Δ191). **c**, Example Sanger sequencing trace following ICE analysis over the sgRNA1 and sgRNA2 binding sites in unedited cells, and the double strand break repair site in cells containing the 108 bp deletions. sgRNA sequence shown by black boxes, protospacer adjacent motif sites shown with red letters. **d**, Calculated deletion efficiency for each sgRNA pair and cell type. Transfections failing to achieve >70% deletion (blue circles) were excluded from expression analyses. n shown are for independent transfections **e**, Expression of *LZTFL1* normalized to *RPS18* and expressed as relative to the mean expression in Cas9 only treated cells for each cell type. Corrected *P* values from an ordinary one-way ANOVA with Dunnett's multiple comparisons test. n shown are for independent samples from at least 3 independent transfections. For d,e bars show mean and one standard deviation. **f**, ChIP-seq for the active transcription marker (H3K27ac) was performed in umbilical vein endothelial cells (HUVECs), blood outgrowth endothelial cells (BOECs), H441 lung epithelial cells, and IMR-90 lung fibroblast cells. The rs17713054 enhancer (grey box, **g**) lacks strong modification under standard growth conditions in these cells.



Extended Data Fig. 10 | COVID-19 patient lungs show signals of EMT. Spearman correlation of gene expression profiles for EMT-related genes with the cell-types identified by deconvolution. AT1: Alveolar Type 1 pneumocytes, AT2: Alveolar Type 2 pneumocytes. *P* values were identified by two-sided Hmisc analysis (without multiple test correction), values for significant correlations are shown and all correlation and *P* values are in Source Data.

nature research

Corresponding author(s): James Davies
Jim Hughes

Last updated by author(s): Aug 31, 2021

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

Commercially available software was used for data collection:
Sequencing: Illumina NextSeq Platform using the NextSeq System Suite (v2)
FACS: Attune NxT software (v3.0)
RT-qPCR: StepOnePlus™ Real-Time PCR System

Data analysis

- Chromosome conformation capture data was processed using CaptureCompendium which is comprised of CCseqBasicS (v5, github.com/Hughes-Genome-Group/CCseqBasicS) and captureCompare (v1, github.com/Hughes-Genome-Group/CaptureCompare).
- Downloaded ENCODE bigwigs were processed with deepTools (v2.2.2)
- CTCF and CEBPB motifs were identified using the MEME-Suite tools (v5.3.0)
- ATAC-seq and ChIP fastq data was processed with NGseqBasic (v20; https://www.biorxiv.org/content/early/2018/08/16/393413) which uses FASTQC (v0.11.4), bowtie2 (v2.4.2), UCSCtools (v3.8.5), and Samtools (v0.1.19), and also with deepTools (v2.2.2)
- Linkage analysis was determined using the LDlink webtool (v5.1, LDproxy, LDpair; https://ldlink.nci.nih.gov/)
- Immunofluorescence images were processed in Fiji (v2.1.0, https://imagej.net/Fiji) and Adobe Photoshop (v22.4.1)
- MCC was processed with TrimGalore (v0..3.1), FLASH (v1.2.11), BLAT (v35), bowtie2 (v2.3.5), MCCsplitter.pl (v1) and MCCanalyser.pl (v1). MCCsplitter.pl and MCCanalyser.pl are available for academic use through the Oxford University Innovation software store https://processinnovation.ox.ac.uk/software/p/16529a/micro-capture-c-academic/1.
- MCC peaks were called using LanceOtron (v2, https://lanceotron.molbiol.ox.ac.uk; https://www.biorxiv.org/content/10.1101/2021.01.25.428108v2)
- Statistical analysis was performed in Prism (v8.0e; https://www.graphpad.com/scientific-software/prism/)
- Variant damage was predicted using sasquatch (http://apps-dev.molbiol.ox.ac.uk/sasquatch/cgi-bin/foot.cgi) and deepHaem (https://github.com/rschwess/deepHaem)
- FACS was analysed in FlowJo (v10.7)
- Editing efficiency was determined using D1000 tapestation software (https://www.agilent.com/en/product/automated-electrophoresis/tapestation-systems/tapestation-software/tapestation-software-379381) and Synthego Ice analysis (v2.0, https://ice.synthego.com/#/)

- miRNA predictions used TargetScan (v7.2 http://www.targetscan.org/vert_72/), miRdSNP (v11.03; http://mirdsnp.ccr.buffalo.edu/) and MicroSNiPer (Release 19, http://vm24141.virt.gwdg.de/services/microsniper/)
- Digital spatial profiling was analysed with R packages: SpatialDecon (v1.0.0), (v1.70-3) Hmisc (v4.5-0) and corrplot (v0.84).
- Colocalisation analysis was performed with the R coloc package (v5.1.0)
- Allelic bias mapping was performed with WASP (v0.3.4)

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

Capture-C, Micro Capture-C, ATAC-seq and ChIP-seq data generated for this study (Fig 3, Extended Data Figs. 7, 9 and Supplementary Figs. 1, 2) are available from the Gene Expression Omnibus (GSE159867, GSE175791). Processed Capture-C data can be visualised on UCSC (http://datashare.molbiol.ox.ac.uk//datashare/project/fgenomics/publications/Downes_2021_Covid_GWAS/hub.txt) or on the CaptureSee website (https://capturesee.molbiol.ox.ac.uk/projects/capture_compare/3718). Numerical values for Figs. 2a-c, and 5d, and Extended Data Figs. 2, 3, 4, 6, 7, 9, 10 are available in Source Data. Expression data (Fig. 3, Extended Data Figs. 6,8) was from publicly available sources: GTEx (https://gtexportal.org), the Lung Cell Atlas (https://asthma.cellgeni.sanger.ac.uk/) and the Lung Epigenome (https://www.lungepigenome.org/). Publicly available open chromatin data (ATAC-seq/DNase-seq), transcription factor binding data (ChIP-seq), and epigenetic modifications (ChIP-seq) data, (Figs. 1,2, Extended Data Figs. 1, 2, 4-7, 9) were sourced either from the ENCODE portal (https://www.encodeproject.org/), the Gene Expression Omnibus (GSE74912, GSE115684, GSE118189, GSE125926) or the UCSC genome browser (https://genome.ucsc.edu), descartes human developmental accessibility atlas https://descartes.brotmanbaty.org/bbi/human-chromatin-during-development/) and the Lung Epigenome (https://www.lungepigenome.org/). Masked splicing prediction effects were downloaded from the SpliceAI database (https://github.com/Illumina/SpliceAI). CEBPB motif (MA0466.1) was downloaded from the JASPER database (http://jaspar.genereg.net). Conserved miRNA sites were identified on miRdSNP (http://mirdsnp.ccr.buffalo.edu/browse-genes.php).

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

Life sciences study design

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

Sample size

- For Capture-C, a sample size of three independent replicates was chosen per cell type to enable potential statistical analysis with DESeq2 consistent with previous studies.
- For analysis of accessibility, sample size was determined by published data and genotype of sampled individuals but was sufficient to allow implementation of a non-parametric analysis.
- For qPCR expression analyses, at least 3 replicates of each condition were used to allow one-way ANOVA analysis in three categories/groups.
- For in situ transcriptomics, 13-17 regions were selected from each individual to give multiple samples of severe and mild diffuse alveolar damage and ensure balance across the three individuals.

Data exclusions

- For editing experiments bulk populations with <70% deletion efficiency were excluded from expression analyses.
- For digital spatial profiling, sequencing quality per area was examined and a single under-sequenced area with zero deduplicated reads was excluded. Additionally, outlier probes were excluded (geomean probe in all AOI/geomean of all probes for a transcript = <0.1; or probe failure in the Grubbs outlier test in over 20% AOIs).
- Allele counts from three samples were excluded from ATAC-seq bias analysis as they did not meet a predetermined filter of 4 or more reads.
- No other data was excluded.

Replication

- For each cell type, three independent replicates of Capture-C were performed and examined to visually inspect for high reproducibility, which was observed. To generate independent replicates for primary blood cell types (CD4, CD14, Erythroid) independent donors were used. For HUVEC primary cells, three independent samples were acquired by sourcing from three companies (Lonza, Gibco, PromoCell). For the H1-hESC, H441 and IMR-90 cell lines, samples were grown independently.
- CRISPR/Cas9 experiments were performed with between 2 and 6 independent transfections per combination of guide RNAs per cell-type and results were replicated in all experiments.
- For qPCR expression analysis results were successfully replicated across multiple experiments.
- For IF, staining was performed twice with both replicates producing consistent results.
- For H&E staining one biopsy was taken per individual (total of 3 individuals), all 3 samples were stained and showed consistent, which were also consistent with results described in publications cited in this paper.
- No attempts were made to replicate in situ transcriptomic results as the data was generated from unique tissue samples from patients.
- ChIP-seq for H3K27ac was not repeated but results were visually compared with similar public datasets and successfully reproduced known sites of activation (Supp Fig 2).

Randomization

- Random grouping of samples was not relevant to this study as comparison was between distinct cell-types, tissues, and genotypes.
- For CRISPR-Cas9, cells were sub-divided from a single pool and randomly assigned for treatment with each condition, cells were cultured in identical conditions and media, and experiments performed simultaneously.

Blinding

Blinding was not relevant to this study. Groupings were predetermined by inherent cell-type, tissue-type, or genotype and therefore not subject to bias. Data generation and processing followed an identical experimental and analytical pipeline for all samples to avoid potential bias. Biological material for comparison was processed simultaneously. Exclusion criteria were quantitative and established prior to analysis and not changed. For H&E comparison, patient samples are so distinct from healthy controls that blinding is not worthwhile.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used

ChIP:
Rabbit polyclonal anti-H3K27ac (AbCam, ab4729, 0.3 µg, lot: GB3205523-I, 1:2,000)

FACS:
APC conjugated mouse monoclonal anti-CD14 1:100 (2 ng, Clone: M5E2, BioLegend Cat: 301807, Lot: B266608, 1:100)
PE conjugated mouse monoclonal anti-CD309/VEGFR2 1:100 (2 ng, Clone: 7D4-6, BioLegend Cat: 359903, Lot: B245460, 1:100)
FITC conjugated mouse monoclonal anti-CD31/PECAM 1:100 (2 ng, Clone: WM59, BioLegend Cat: 303103, Lot: B287895, 1:100)
PE/Cy7 conjugated mouse monoclonal anti-CD34 1:100 (0.5 ng, Clone: 561, BioLegend Cat: 343616, Lot: B257238, 1:100)

IF:
Mouse monoclonal anti-VWF 1:100 (Clone F8/86, MA5-14029, Invitrogen, Lot: WB3184744C)
Alexa 488 conjugated goat polyclonal anti-mouse 1:500 (A-11029, ThermoFisher Scientific, Lot: 2120125)

Validation

ChIP:
- All AbCam antibodies are validated by Western blot analysis prior to sale. From https://www.abcam.com/histone-h3-acetyl-k27-antibody-chip-grade-ab4729.html?productWallTab=ShowAll
"All batches of ab4729 are tested in Peptide Array against peptides to different Histone H3 modifications. Six dilutions of each peptide are printed on to the Peptide Array in triplicate and results are averaged before being plotted on to a graph. Results show strong binding to Histone H3 - acetyl K27 peptide (ab24404), indicating that this antibody specifically recognises the Histone H3 - acetyl K27 modification."
- Following sequencing, tracks were compared with equivalent H3K27ac datasets to ensure a comparable profile.

FACS:
Each batch of Biolegend antibodies is tested against cells know to be positive and negative for target proteins. From https://www.biolegend.com/en-us/quality/quality-control :
- Specificity testing of 1-3 target cell types with either single- or multi-color analysis (including positive and negative cell types).
- Once specificity is confirmed, each new lot must perform with similar intensity to the in-date reference lot. Brightness (MFI) is evaluated from both positive and negative populations.
- Each lot product is validated by QC testing with a series of titration dilutions.

IF:
Antibody specificity was confirmed by the manufacturers using Western Blot analysis. From Invitrogen (VWF Antibody (MA5-14029): "This Antibody was verified by Relative expression to ensure that the antibody binds to the antigen stated. Antibody specificity was demonstrated by detection of differential basal expression of the target across cell models owing to their inherent genetic constitution. Expression of VWF was observed specifically in HUVEC cells using VWF Mouse Monoclonal Antibody (Product # MA5-14029) in western blot."
- In addition to manufacturer testing, two VWF non-expressing cell lines were stained to demonstrate specificity (Supp. Fig. 2).
- Coverslips and cells without primary antibody were stained with only the secondary antibody (Goat anti-mouse antibodies) to ensure specificity.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	H1-hESC cells came from WiCell. NCI-H441 cells were sourced from ATCC via LGC standards. IMR-90 cells were sourced from ATCC via LGC standards. HUDEP-2 were provided by RIKEN
Authentication	Cells were not specifically authenticated. However, ATAC-seq and H3K27ac (which show highly cell type specific profiles) were consistent with ENCODE datasets for matching or similar cell types. HUVEC, BOEC and HUDEP cells were analyzed with FACS to ensure expression of specific Markers. HUVEC and BOEC cells were further analyzed by IF to ensure expression of the endothelial marker VWF. IMR-90 and NCI-H441 were not specifically authenticated beyond cellular morphology.
Mycoplasma contamination	Cell lines were routinely tested for mycoplasma and only used if found negative. [LookOut Mycoplasma qPCR Detection Kit (MP0040A, Sigma-Aldrich)]
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Whole blood was drawn from a range of healthy males and females (25-60 years old) with no genotyping information.
Recruitment	Healthy donors were recruited from the local community by self-nomination and selected based on general good health (no self-reported chronic conditions) and availability to regularly donate blood. All donors were of mixed European ancestry, however this is unlikely to affect results. No other potential biases are expected.
Ethics oversight	North West Research Ethics Committee of NHS National Research Ethics Services, UK (03/08/097) East of England – Cambridgeshire and Hertfordshire Research Ethics Committee (05/Q0106/20) Oxfordshire Research Ethics Committee COREC (06/Q1605/55). Human Tissue Authority (License 12433) Ethics Committee of University of Navarra, Spain (15/05/2020) Medical Sciences Interdivisional Research Ethics Committee of the University of Oxford (Approval R76045/RE001)

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

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	Raw reads and processed files are available from the Gene Expression Omnibus (GSE175791)
Files in database submission	BOEC_H3K27ac.hg38.bw ChIP_BOEC_Don053_H3K27ac_rep1_R1.fastq.gz ChIP_BOEC_Don053_H3K27ac_rep1_R2.fastq.gz ChIP_H441_H3K27ac_rep1_R1.fastq.gz ChIP_H441_H3K27ac_rep1_R2.fastq.gz ChIP_HUVEC_H3K27ac_rep1_R1.fastq.gz ChIP_HUVEC_H3K27ac_rep1_R2.fastq.gz ChIP_IMR90_H3K27ac_rep2_R1.fastq.gz ChIP_IMR90_H3K27ac_rep2_R2.fastq.gz H441_H3K27ac.hg38.bw HUVEC_H3K27ac.hg38.bw IMR90_H3K27ac.hg38.bw BOEC_Input.hg38.bw ChIP_BOEC_Don053_Input_rep1_R1.fastq.gz ChIP_BOEC_Don053_Input_rep1_R2.fastq.gz ChIP_H441_Input_rep1_R1.fastq.gz ChIP_H441_Input_rep1_R2.fastq.gz ChIP_HUVEC_Input_rep1_R1.fastq.gz ChIP_HUVEC_Input_rep1_R2.fastq.gz ChIP_IMR90_Input_rep2_R1.fastq.gz ChIP_IMR90_Input_rep2_R2.fastq.gz H441_Input.hg38.bw

	HUVEC_Input.hg38.bw IMR90_Input.hg38.bw
Genome browser session (e.g. UCSC)	Datahub for loading into UCSC http://datashare.molbiol.ox.ac.uk//datashare/project/fgenomics/publications/Downes_2021_Covid_GWAS/hub.txt

Methodology

Replicates	One replicate was performed per IP per cell type with a 10% input (no IP control)
Sequencing depth	10% Input samples: BOEC: 32,574,141 total read pairs; 23,210,386 mapped read pairs; 1,277,826 filtered PCR duplicates. IMR-90: 15,535,096 total read pairs; 8,785,802 mapped read pairs; 337,344 filtered PCR duplicates. H441: 24,375,724 total read pairs; 15,246,962 mapped read pairs; 415,664 filtered PCR duplicates. HUVEC: 30,388,466 total read pairs; 21,027,888 mapped read pairs; 544,441 filtered PCR duplicates. H3K27ac samples: BOEC: 30,590,762 total read pairs; 27,985,799 mapped read pairs; 3,673,873 filtered PCR duplicates. IMR-90: 19,282,519 total read pairs; 16,474,295 mapped read pairs; 755,665 filtered PCR duplicates. H441: 25,311,763 total read pairs; 23,116,406 mapped read pairs; 1,279,269 filtered PCR duplicates. HUVEC: 25,475,462 total read pairs; 23,250,184 mapped read pairs; 1,160,345 filtered PCR duplicates.
Antibodies	1:2,000 Polyclonal rabbit anti-H3K27ac (AbCam, ab4729, 0.3 µg, lot: GB3205523-I)
Peak calling parameters	No peak calling of ChIP-seq was performed
Data quality	ChIP quality was performed prior to sequencing by qPCR to ensure at least 10-fold enrichment at an enhancer over an intergenic control region. Following sequencing ChIP profiles were confirmed with matching or equivalent cell type datasets
Software	Sequencing data was collected on an Illumina NextSeq Platform using the NextSeq System Suite (v2) and mapped using bowtie2

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	For FACS, ~105 cells were resuspended in 100 µL staining buffer (PBS with 10% FBS) and incubated with 1 µL each APC conjugated mouse anti-CD14 (2 ng, Clone: M5E2, BioLegend Cat: 301807, Lot: B266608), PE conjugated mouse anti-CD309/VEGFR2 (2 ng, Clone: 7D4-6, BioLegend Cat: 359903, Lot: B245460), FITC conjugated mouse anti-CD31/PECAM (2 ng, Clone: WM59, BioLegend Cat: 303103, Lot: B287895), and PE/Cy7 conjugated mouse anti-CD34 (0.5 ng, Clone: 561, BioLegend Cat: 343616, Lot: B257238) for 20 min at 4ºC. Cell were diluted with 90 µL staining buffer with 1:5,000 Hoechst 33258 (ThermoFisher)
Instrument	Attune NxT Flow Cytometer
Software	Data collection: Attune NxT software (v3.0) Data analysis: FlowJo v10.7
Cell population abundance	Simple staining analysis of bulk monoculture rather than sorting of mixed populations was performed.
Gating strategy	Voltages and compensation were set using single stain samples with UltraComp eBeads (ThermoFisher) for antibodies and cells for Hoescst. Negative and positive populations were established using Fluorescence Minus One Controls. Mononuclear cells were gated using forward scatter (FSC) and side scatter, single cells gated using FSC-area and FSC-height, and live cells selected using a Hoechst negative gate.

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