Published in final edited form as:
Nat Genet. 2021 November 01; 53(11): 1606–1615. doi:10.1038/s41588-021-00955-3.

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

Damien J. Downes[1], **Amy R. Cross**[#2], **Peng Hua**[#1], **Nigel Roberts**[1], **Ron Schwessinger**[1,3], **Antony J. Cutler**[4,5], **Altar M. Munis**[6], **Jill Brown**[1], **Olga Mielczarek**[4], **Carlos E. de Andrea**[7], **Ignacio Melero**[8], **COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium**
A full list of consortium members can be found in Supplementary Table 1. Authors who are COMBAT consortium members:, Damien J. Downes[1], Ron Schwessinger[1,2], Julian C. Knight[3,4,5], John A. Todd[3], Stephen N. Sansom[6], Fadi Issa[7,8], Jim R. Hughes[1,2]
, **Deborah R. Gill**[6], **Stephen C. Hyde**[6], **Julian C. Knight**[4,9,10], **John A. Todd**[4], **Stephen N. Sansom**[11], **Fadi Issa**[2,12], **James O.J. Davies**[1,12,†], **Jim R. Hughes**[1,3,†]

[1]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[2]Transplantation Research and Immunology Group, Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

[3]MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[4]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[6]Gene Medicine Group, Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[7]Department of Pathology, Cífnica Universidad de Navarra, Pamplona, Spain

[8]Division of Immunology and Immunotherapy, Centre for Applied Medical Research (CIMA), University of Navarra, Pamplona, Spain

[9]Chinese Academy of Medical Science (CAMS) Oxford Institute (COI), University of Oxford, Oxford, UK

[†]Corresponding authors: jim.hughes@imm.ox.ac.uk, james.davies@imm.ox.ac.uk.
[5]Present address: Immunology Research Unit, GSK, Stevenage, UK

Author Contributions Statement

D.J.D., A.J.C., O.M., N.R., and A.M.M. isolated, cultured, and fixed 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 experiments. D.J.D., P.H., A.R.C., R.S., J.B., and S.N.S. analyzed 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 funding and oversaw the project. D.J.D., J.O.J.D. and J.R.H. conceived the work, generated figures, and wrote the manuscript.

Competing Interests Statement

J.R.H. and J.O.J.D. are founders and shareholders of, and 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 (WO2017068379A1, EP3365464B1, US10934578B2) and have a patent application for MCC. J.A.T. is member of the GSK Human Genetics Advisory Board. These authors declare no other financial or non-financial interests. The remaining authors declare no competing interests.

[10]NIHR Oxford Biomedical Research Centre, Oxford, UK

[11]Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK

[12]Oxford University Hospitals National Health Service (NHS) Foundation Trust, Oxford, UK

[1]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[2]MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[3]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[4]Chinese Academy of Medical Science (CAMS) Oxford Institute (COI), University of Oxford, Oxford, UK

[5]NIHR Oxford Biomedical Research Centre, Oxford, UK

[6]Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK

[7]Transplantation Research and Immunology Group, Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

[8]Oxford University Hospitals National Health Service (NHS) Foundation Trust, Oxford, UK

[#] These authors contributed equally to this work.

Abstract

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) disease (COVID-19) pandemic has caused millions of deaths worldwide. Genome-wide association studies (GWAS) identified the 3p21.31 region as conferring a two-fold increased risk of respiratory failure. Here, using a combined multiomics and machine-learning approach, we identify the gain-of-function risk A allele of a single-nucleotide polymorphism (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 COVID-19 patient lung biopsies 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 to be responsible for the 3p21.31 associated risk. As the 3p21.31 effect is conferred by a gain-of-function, *LZTFL1* may provide a therapeutic target.

Keywords

COVID-19; Coronavirus; gene regulation; chromosome conformation capture; machine learning; respiratory epithelium

Introduction

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]. GWAS are important for identifying candidate genes and pathways that predispose to complex diseases[7] and genetically validated drug targets are more likely to lead to approved drugs[8]. Two large GWAS have been carried out to determine whether common variants drive susceptibility to severe COVID-19[9,10]. Both studies identified a region of chromosome 3p21.31 as having the strongest association, whilst 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 two-fold increased risk of respiratory failure from COVID-19[9,10], and an over two-fold increased risk of mortality for under 60 year-olds[13]. Additionally, the risk variants at this locus are carried by >60% of individuals with South Asian ancestries (SAS), compared to 15% of white European ancestry (EUR) groups, partially explaining the ongoing higher death rate in this population in the United Kingdom[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). Secondly, the genetic signals are completely cell-type agnostic, which makes it challenging to identify appropriate experimental models for further investigation. Thirdly, 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 bp) from the genes they control, and the effects of sequence changes are not straightforward to predict.

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

Here, we have 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 CEBPB motif

The 3p21.31 region contains variants associated with the autoimmune diseases type-1 diabetes[23] and multiple sclerosis[24], though 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 miRNA binding site[25] and neither miRdSNP[26] nor MicroSNiPer[27] predict the variant alters miRNA binding. Four other variants are within *LZTFL1* introns, including the lead SNP rs11385942[9]. None of these are predicted to alter mRNA splicing of *LZTFL1*, either by machine learning with SpliceAI[18] or splicing quantitative trait loci (sQTL) based approaches[28], and the nearest exon junction to these variants is ~500 bp (Fig. 1). Therefore, a *cis*-regulatory mechanism seems to be 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 we identified two SNPs, rs17713054 and rs76374459, which lie in open chromatin[30] (Fig. 1, 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 for *de novo* gain-of-function changes[33]. We have previously developed a machine-learning model, deepHaem[19], which uses 694 DNase I hypersensitivity and ATAC-seq datasets to predict changes to active regulatory elements. Importantly, deepHaem predicted that the 26 variants not 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, and it is not predicted by machine learning to cause damaging effects (Fig. 1, Extended Data Figs. 2,4).

In contrast, rs17713054 is likely to be a causative SNP as 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, 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 from healthy lung[34,35] shows this enhancer is present in several lung epithelial cell types, including the ciliated epithelia and club cells, which line the respiratory tract, and in Type I and Type II pneumocytes, which form the alveoli (Fig. 1, Extended Data Fig. 5). Interestingly, deepHaem predicts the rs17713054 risk allele, which is the minor allele A (MAF: 0.0817 EUR, 0.377 SAS[36]), acts as a gain-of-function 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 (HAECs) 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), and, in heterozygous samples, more reads originated the risk A allele than the non-risk G allele (Fig. 2c).

Sequence analysis shows 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 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 – shows weak binding at the enhancer (Extended Data Fig. 6a-d). Furthermore, deepHaem predicts rs17713054-A leads 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.

rs1773054 enhancer interacts with *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 co-receptor for HIV[42,43]) and *XCR1* (which encodes a X-C chemokine receptor). Transcriptome-wide association study (TWAS) analysis has also identified *CCR2, CCR3* and *FYCO1*, which lie up to 500 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* (Leucine zipper transcription factor-like 1[45]), 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 (HUVEC) in which the rs17713054 enhancer is accessible, as well as resting and stimulated primary $CD4^+$ T-cells, primary $CD14^+$ monocytes, $CD71^+$ $CD235^+$ erythroid cells, and 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 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, Extended Data Fig. 7a). Importantly, we did not find a peak of interaction with *SLC6A20*, likely because ENCODE datasets show that *SLC6A20* carries Polycomb repression marks in endothelial (HUVEC) and fibroblast (NHLF) cells (Extended Data Fig. 7b). Additionally, the *LZTFL1* promoter is more consistently accessible in cells where rs17713054 is also accessible (Extended Data Fig. 7c,d). Therefore, *LZTFL1* seems to be the most likely direct regulatory target of the rs17713054-containing epithelial-endothelial-fibroblast enhancer.

rs17713054-A associates with higher gene expression in lung

Disease biology, deepHaem, TWAS analysis[10] and a Phenome-wide association study[11] (PheWAS) 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[48] shows that *LZTFL1* is present throughout the respiratory epithelium, but predominantly expressed in ciliated cells (Fig. 4b,c). Of the other candidate genes identified here and elsewhere[10,49,50] (*SLC6A20, CCR2, CCR3, CCR9, CXCR6*, and *FYCO1*), only *SLC6A20* and *FYCO1* are consistently expressed in both lung bulk and single-cell RNA-seq datasets, though *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 II pneumocytes (Fig. 4, Extended Data Fig. 8). Analysis using the Genotype-Tissue Expression[28] (GTEx) portal for expression quantitative trait loci (eQTLs) showed that the rs17713054-A risk-allele is associated with higher levels of expression in the lung of *LZTFL1* and *SLC6A20* but not the other genes (Fig. 4d, Extended Data 8). Colocalization analysis[51] shows that these GWAS and eQTL associations are more likely as a result of a single variant (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 regulation of *LZTFL1* and *SLC6A20*. As the enhancer shows accessibility in epithelial, endothelial, and mesenchymal cells (Extended Data Fig. 9a) we used CRISPR-Cas9 ribonucleoprotein editing to delete either a 108-bp or a 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, Supplementary Fig. 2). Using real-time qPCR we detected no-effect on *LZTFL1* expression following enhancer deletion (Extended Data Fig. 9e) – consistent with a report that CRISPRi in the 16HBE14o- bronchial epithelial cell line had no effect on nearby gene expression[50]. As *SLC6A20* is Polycomb repressed in fibroblasts and endothelial cells it was undetectable by 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 is likely 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 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 de-differentiate and acquire mesenchymal identity, plays a key role in the innate immune response and is a consequence of lung inflammation, 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 TGF-β signalling[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 products of EMT), presence of EMT promoting signaling pathways and co-expression of epithelial and mesenchymal markers[61]. Consistent with work by others[62,63], we saw widespread epithelial dysfunction and diffuse alveolar damage (DAD) with reorganization indicative of EMT evident in post-mortem biopsies of three COVID-19 patients. Dysfunction in ciliated airways included denudation, hyperplasia, and squamous metaplasia (Fig. 5a). Features of DAD 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, generally absent from healthy lung tissue.

We previously generated selective spatial transcriptomics from 46 areas of post-mortem biopsies from critical COVID-19 patients covering a spectrum of alveolar injury[64]. To explore expression 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 co-regulated 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 I (AT1) and Type II (AT2) pneumocytes, as well as both of the Epithelial and Type II pneumocyte WGCNA modules (Fig 5c, Extended Data Fig. 10). However, AT1, was also positively associated with the hallmark EMT gene *ACTA2* (α-smooth muscle actin; Hmisc rcorr asymptomatic $P = 0.0014$), as were both the AT2 and Epithelial modules ($P = 0.0069$ and 9.59×10^{-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 associated with *SNAI2* ($P = 1.08 \times 10^{-6}$) and the fibroblast module 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 both pneumocyte subtypes (AT1, AT2) and both epithelial modules associated with TGF-β signaling receptor genes (*TGFBR1* and *TGFBR2*) and WNT signaling genes which encode β-catenin and frizzled receptors (*CTNNB1, FZD6* and *FZD7*). By contrast, neither CD8$^+$ T-cells nor the cytotoxicity and T-cell module expressed epithelial or mesenchymal genes, but they did express *TGFB1* ($P = 0.0029$ and $P = 0.0005$ respectively). The colocalized expression of mesenchymal genes with epithelial cells, along with 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. The modulation of EMT by LZTFL1 may therefore be of relevance to the pathological outcome of COVID-19 infection.

Discussion

We have 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 two-fold 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 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 TWASs 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. TWASs are unable to 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 the remainder of the genes; with the caveat that it is possible that a direct effect may occur in an untested cell type. Whilst the ultra-high 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 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, and it is plausible that in cells where *SLC6A20* is not Polycomb repressed (e.g. Goblet and Alveolar Type II 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[68] (ACE2); which is a cell receptor for SARS-CoV-2[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 post-mortem COVID-19 lung biopsies demonstrates widespread epithelial dysfunction with EMT signatures[62,63]. Consistently, single-cell RNA-seq shows a reduction in total numbers of epithelial cells following infection[75], with a lower epithelial composition correlating with a more rapid progression from symptom onset to death[76]. The samples analyzed here 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 COVID-19 patients has 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 de-differentiation 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 serine protease 2 (TMPRSS2)[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 treatment or prevention of COVID-19.

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 North West Research Ethics Committee of NHS National Research Ethics Services, UK (03/08/097). Blood samples for CD4$^+$ cells were obtained from donors recruited from the Cambridge BioResource. The study was approved by 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, Spain (15/05/2020) and the Medical Sciences Interdivisional Research Ethics Committee of the University of Oxford (Approval R76045/RE001). Patient samples and, hematopoietic stem and progenitor cells from healthy donors were stored in accordance with Human Tissue Authority (License 12433).

Cell isolation, culture and stimulation

Human ESC line H1 (H1-hESC [https://scicrunch.org/resolver/CVCL_97711; WA01 WiCell, RRID:CVCL_9771) was grown on Matrigel (Corning) coated plates in mTeSR1 medium (StemCell technologies). Cells were harvested as a single-cell suspension using Accutase (EDM Millipore); fixation were carried out in mTeSR1 medium. Primary neonatal Human Umbilical Vein Endothelial Cells (HUVEC; Lonza; CC-2517, Gibco; C0035C, PromoCell; C-12200) were expanded in endothelial cell growth medium (Sigma) up to five passages following the manufacturer's protocol. For passaging, HUVECs were grown to 60% confluence, washed with HBSS at room temperature and sub-cultured following light trypsination using Trypsin-EDTA (Sigma) at room temperature with trypsin inhibitor (Sigma) added upon rounding of the cells to achieve gentle release from the flask. HUVECs were fixed in RPMI supplemented with 10% FBS. For erythroid cells, CD34$^+$ hematopoietic stem and progenitor cells were isolated from peripheral blood of two healthy males and one healthy female and differentiated *ex vivo* for 13 days as previously described[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, Life Technologies) at a ratio of 0.3 beads per cell for 4 hours at 37°C in X-VIVO-15 (Lonza), 1% AB serum (Lonza) and penicillin/ streptomycin (Life Technologies). Non-activated or activated CD4$^+$ T-cells were pooled after 4 hours of culture and fixed in growth media. For CD14$^+$ cells, peripheral blood mononuclear cells 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 Transport). Monocyte isolation was carried out by positive selection using magnetic-activated sorting with CD14$^+$ beads (MACS, Miltenyi) according to the manufacturer's instructions. IMR-90 [https://scicrunch.org/resolver/CVCL_03471 lung fibroblasts (ATCC; CCL-186, RRID:CVCL_0347) were cultured in Eagle's Minimal Essential Medium supplemented with 10% FBS, 1 mM sodium pyruvate (Gibco), 1× MEM Non-Essential Amino Acids (Gibco) and Penicillin-Streptomycin (100 U ml^{-1} each) Cell were sub-cultured every 3 days following light trypsination using 0.05% Trypsin-EDTA (Gibco). Blood outgrowth endothelial cells (BOECs) were isolated as previously described[83]. Briefly 20-40 ml of fresh blood was diluted 1:1 with PBS, layered over Histopaque-1077 (Sigma) and centrifuged for 15 minutes at 500 ×g, brake off. Peripheral blood mononuclear cells were washed with PBS then resuspended in EGM-2 Bullet kit growth media (Lonza) supplemented with 10% heat inactivated FBS. Cells were cultured for 21-28 days in collagen coated flasks until BOEC colonies formed. BOEC colonies were passaged by light trypsinisation. 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_15611 cells (ATCC; HTB-174, RRID:CVCL_1561) were grown in RPMI 1640 medium (Gibco) supplemented with 10% non-heat inactivated FBS (Sigma) and 1% Penicillin-Streptomycin (Gibco), cells were given fresh media every 2 days and passaged by light trypsination twice weekly. Human Umbilical Derived Erythroid Progenitor line 2 cells[84] (HUDEP-2 [https://scicrunch.org/resolver/CVCL_VI061, RRID:CVCL_VI06) were provided by RIKEN and were maintained at 0.7-1.5 × 10^6 cells ml^{-1} in HUDEP expansion medium (SFEM, 50 ng ml^{-1} SCF, 3 IU ml^{-1} EPO, 10 μM DEX, 1% L-Glu, 1% Penstrep) and changed into fresh medium containing 2× doxycycline every 2 days.

Variant effect sequence predictions

Linkage analysis was determined using the LDlink webtool (v5.1, 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 FIMO[85] analysis of reference and variant containing enhancer sequence (chr3:45,817,661-45,818,660, hg38) with the JASPAR[86] motif MA0466.1. Sasquatch[40] was run using default Workflow 3 settings (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] (http://www.targetscan.org/vert_71/). SNP predictions were identified using the miRdSNP[26] database (http://mirdsnp.ccr.buffalo.edu/browse-genes.php) and the MicroSNiPer[27] webtool (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[9] were downloaded from the GWAS Catalog[87] (GCST90000256). Summary statistics for all lung eQTL-variant pairs (V8) in

European-Americans (EUR) were downloaded from the GTEx portal[28]. Coloc[51] (v5.0.1) 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), as well as association β, standard errors, minor allele frequencies, and z-scores were used in a sensitivity analysis[88] which showed a prior probability of co-localization (p12) 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).

Chromosome conformation capture (3C)

Gene promoters were selected for Capture-C using 10-kb resolution Hi-C data on the 3D Genome Browser[89] [url: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 previously described with either the NG or NuTi method[20,21,90]. Briefly 5-20 million cells were fixed with 2% formaldehyde, 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 Micro Capture-C[22] (MCC), aliquots of 1-2 × 10^7 cells were fixed for 10 minutes with 2% formaldehyde in 10 ml of growth medium. Formaldehyde was quenched with 125 mM glycine, and cells were pelleted (5 minutes, 500 ×g, 4°C) and washed with PBS. Cells were resuspended in 1 ml PBS and permeabilized with 0.005% Digitonin. Cells were pelleted and resuspended in 800 *μl* of reduced calcium content micrococcal nuclease buffer (10 mM Tris-HCl pH 7.5, 1 mM $CaCl_2$). Chromatin was digested for 1 hour at 37°C inside intact, permeabilised cells in three separate reactions using 5-120 Kunitz U of micrococcal nuclease (NEB). Digestion was quenched by with 5 mM ethylene glycol-bis(2-aminoethylether)-N,N,N',N'-tetraacetic acid (EGTA; Sigma). Cells were pelleted and washed with PBS, prior to end-repair and phosphorylation; cells were resuspended in 400 μl DNA ligase buffer (Thermo Scientific) supplemented with 400 μM of each of dATP, dCTP, dGTP and dTTP and 5 mM EGTA, 200 U ml^{-1} T4 Polynucleotide U Kinase PNK (NEB) and 100 U ml^{-1} DNA Polymerase I Large Klenow Fragment (NEB) for 2 hours at 37°C. To ligate DNA fragments, T4 DNA ligase (Thermo Scientific) was added at 300 U ml^{-1} and the reaction was incubated at room temperature for 8 hours. Chromatin was de-crosslinked with Proteinase K at 65°C for over 4 hours and DNA 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 (NEB) with the following modifications; 2 μg of DNA was indexed, 5 μl of adaptor was used, bead clean-ups were performed with 1.5 volumes of Ampure XP beads, and with Herculase II PCR reagents (Agilent) 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 ~1 M reads per viewpoint.

3C data analysis

NuTi Capture-C data were mapped to hg38 using CCseqBasicS[91] (github.com/Hughes-Genome-Group/CCseqBasicS) using bowtie2. Briefly, CCseqBasic5[92] trims adaptor sequences, flashes read pairs, *in silico* digests fragments and uses maps reads before identifying sequences as either capture and reporter. Replicates were compared using the CaptureCompare[93] (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, then bins reporter counts in equal sized regions for generating a windowed profile. For MCC, adapters were removed using TrimGalore[94] (v0.3.1) then fragments reconstructed with FLASH[95] (v1.2.11) into single sequences using the central area of overlapping reads. Fragments were mapped to the oligonucleotide DNA sequence ±350 bp using BLAT[96] (v35) to identify ligation junctions, allowing splitting of reads into new paired fastq files using MCCsplitter.pl and subsequent mapping to hg38 with bowtie2[97] (v2.3.5). PCR duplicates were removed from alignment files with MCCanalyser.pl 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 webtool "Find and Score Peaks with Inputs" (https://lanceotron.molbiol.ox.ac.uk) using the HUDEP2 MCC profile as an input track.

Genome editing

For deletion of the rs17713054 enhancer, cells were transfected with 5 μg Alt-R SpCas9 nuclease V3 ribonucleoprotein (IDT) and 0.1 nmol each of two guide RNAs (Supplementary Table 4). All transfections were carried out with 1-2 × 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 hours in 2 ml antibiotic-free growth media in a single well of a 6-well plate before expansion in fully supplemented media. Bulk DNA was extracted using DNeasy Blood and Tissue Kit (Qiagen), and the edited region (chr3:45,817,769-45,818,459; hg38) amplified using 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/#/].

Quantitative reverse transcription PCR

For expression analysis, cells were grown to >80% confluence in a single well of a 6-well plate. Cells were lysed by addition of 1 ml TRI Reagent (Sigma), snap frozen and stored at -80°C for less than 6 months. RNA was separated by addition of 100 μl of 1-bromo-3-chloropropane, centrifugation in a Phase Lock Gel-Heavy tube (5Prime) for 5 minutes at 10,000 ×g, and precipitation in an equal volume of isopropanol (500 µl) with 1 µl of GlycoBlue (Thermo Fisher). DNA was removed using DNA-free DNA Removal Kit (Invitrogen) and cDNA generated using 1 µg of total RNA with SuperScript III First-Strand Synthesis SuperMix reagents (Thermo Fisher). Quantitative PCR was performed using a 1:10 dilution of cDNA, TaqMan Universal PCR Master Mix II without UNG (Thermo Fisher) and TaqMan Gene Expression Assays (Thermo Fisher) for *LZTFL1* (Hs00947898_m1), *SLC6A20* (Hs00610960_m1) and *RPL18* (Hs00965812_g1) with FAM. *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.

Chromatin Immunoprecipitation (ChIP)

For ChIP, single-cell suspensions of 10^6 cells ml^{-1} in growth media were generated after light trypsin treatment. Cells were fixed by addition of 1% formaldehyde for 10 minutes at room temperature, which was quenched by addition of 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 ChIP Assay Kit (Merk Millipore) on 5 × 10^6 cells in 2 ml of dilution buffer incubated overnight at 4°C with 1 µl polyclonal rabbit anti-H3K27ac (AbCam, ab4729, 0.3 µg, lot: GB3205523-I; 1:2,000). DNA was isolated by phenol- chloroform isoamylalcohol extraction and ethanol precipitation then indexed using NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB). Libraries were sequenced using 39-bp paired-end reads on a NextSeq Platform. Reads were mapped to hg38 using bowtie2[97], PCR duplicates filtered using samtools[99], and bigwigs generated with deeptools[100].

Fluorescence-activated cell sorting (FACS) analysis

For FACS, ~10^5 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, 1:100), PE conjugated mouse anti-CD309/VEGFR2 (2 ng, Clone: 7D4-6, BioLegend Cat: 359903, Lot: B245460, 1:100), FITC conjugated mouse anti-CD31/PECAM (2 ng, Clone: WM59, BioLegend Cat: 303103, Lot: B287895, 1:100), and PE/Cy7 conjugated mouse anti-CD34 (0.5 ng, Clone: 561, BioLegend Cat: 343616, Lot: B257238, 1:100) for 20 minutes at 4°C. Cell were diluted with 90 μl staining buffer with 1:5,000 Hoechst 33258 (Thermo Fisher) and analyzed on an Attune NxT Flow Cytometer. Voltages and compensation were set using single stain samples with UltraComp eBeads (Thermo Fisher) 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 gated using FSC-area and FSC-height, and live cells selected using a Hoechst negative gate in FlowJo (v10.7).

Assay for Transposase-Accessible Chromatin

ATAC-seq was performed as published[101,102] with 7.5 × 10^4 cells per technical replicate, and two to four technical replicates per samples. After spinning at 500 ×g for 15 minutes, cells were resuspended in lysis buffer (10 mM Trish-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630), centrifuged and nuclei washed with PBS. Nuclei were pelleted, PBS discarded and resuspended in tagmentation buffer (25 μl 2× TD buffer, 2.5 µl Tn5 Transposase [Illumina], and 22.5 µl water) then incubated at 37°C for 30 minutes. After the transposition DNA was extracted using MinElute PCR purification kit (Qiagen), half the DNA was amplified for sequencing using NEBNext High-Fidelity 2× PCR Master Mix (NEB) and further purified with QIAquick PCR purification kit (Qiagen). Libraries were sequenced using 39-bp paired-end reads on a NextSeq Platform. Reads were mapped to hg38 using bowtie2 in NGseqBasic[102].

Immunofluorescence staining and microscopy

Cells were grown for 24-48 hours on sterilized coverslips under standard growth conditions and fixed in 4% vol/vol para-formaldehyde (PFA) in 0.25 M HEPES for 15 minutes, followed by permeabilization in 0.2% vol/vol Triton X-100 in PBS for 10 minutes. Following blocking with 10% vol/vol fetal calf serum in PBS, von Willebrand's factor (VWF) was detected using mouse anti-VWF 1:100 (Clone F8/86, MA5-14029, Invitrogen) and goat anti-mouse Alexa Fluor 488 1:500 (A32723, Thermo Fisher Scientific). DNA was stained with 1 µg ml^{-1} 4',6-Diamidino-2-phenylindole dihydrochloride (DAPI) in PBS and after washing coverslips were mounted in Vectashield (Vector Laboratories). Widefield fluorescence imaging was performed on a DeltaVision Elite system (Applied Precision) using a UPLFLN 40× 1.30NA oil immersion objective (Olympus), a CoolSnap HQ2 CCD camera (Photometrics), 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 nm × 161.3 nm × 200 nm. All images were acquired using the same exposure settings. Using Fiji[103], 3D images were flattened by maximum intensity projection and displayed at the same minimum/maximum intensity settings. Images were cropped for publication in Adobe Photoshop (v22.4.1).

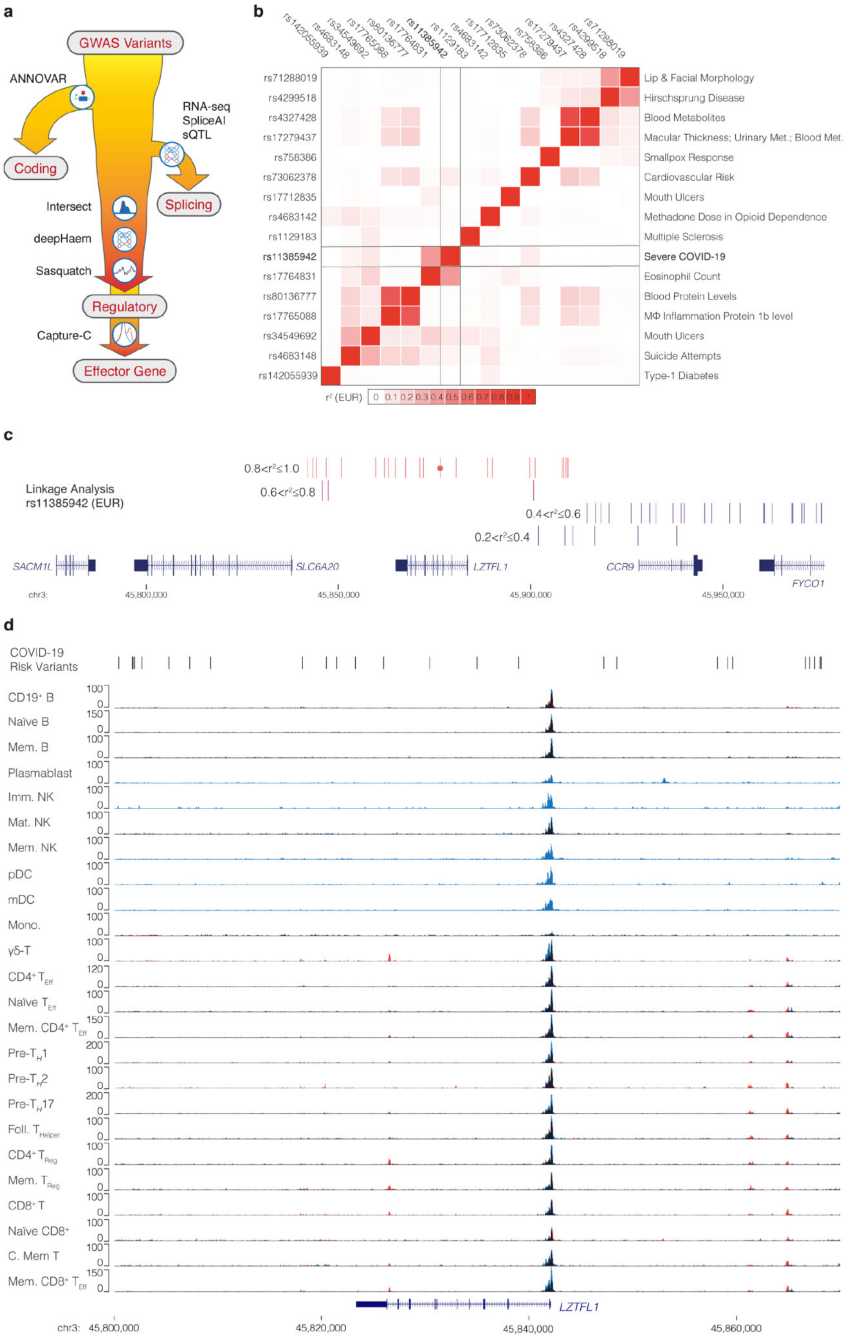
Patients tissue analyses

Healthy lung samples were sourced from chronic obstructive pulmonary disease (COPD) patients during lung tumor resection, with a sample of normal lung acquired away from the tumor. Medical records of COVID-19 patients were retrospectively reviewed[104] and three were selected for in-depth analysis based on: their clinical manifestation of acute respiratory distress syndrome (ARDS), 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. Post-mortem lung tissues were obtained through open biopsy shortly after death and processed as fully described previosly[104]. In brief, tissues were immediately fixed in neutral buffered formalin for <24 hours and then paraffin embedded. Sections (5 µm each) were cut from Wedge biopsies (mean size 1.78 cm^2, standard deviation 0.55 cm^2) for Hematoxylin and Eosin (H&E) analysis. Sections were analyzed by NanoString GeoMx Digital Spatial Profiling (DSP) with normalization and downstream analysis by weighted gene correlation network analysis[66] (WGCNA) and cell deconvolution[65] as previously described[64]. For deconvolution with SpatialDecon in R (v1.0.0), cell profiles were obtained from the Human Cell Atlas healthy lung single-cell RNA-seq appended with neutrophil data[105] using the R package dataset "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 (v1.70-3) and generated 17 biologically assignable modules, of which six were selected for further analysis. Spearman correlation and non-adjusted *P* value generation was performed with the Hmisc R package (v4.5-0) and visualized with corrplot (v0.84).

Public Data Set analysis

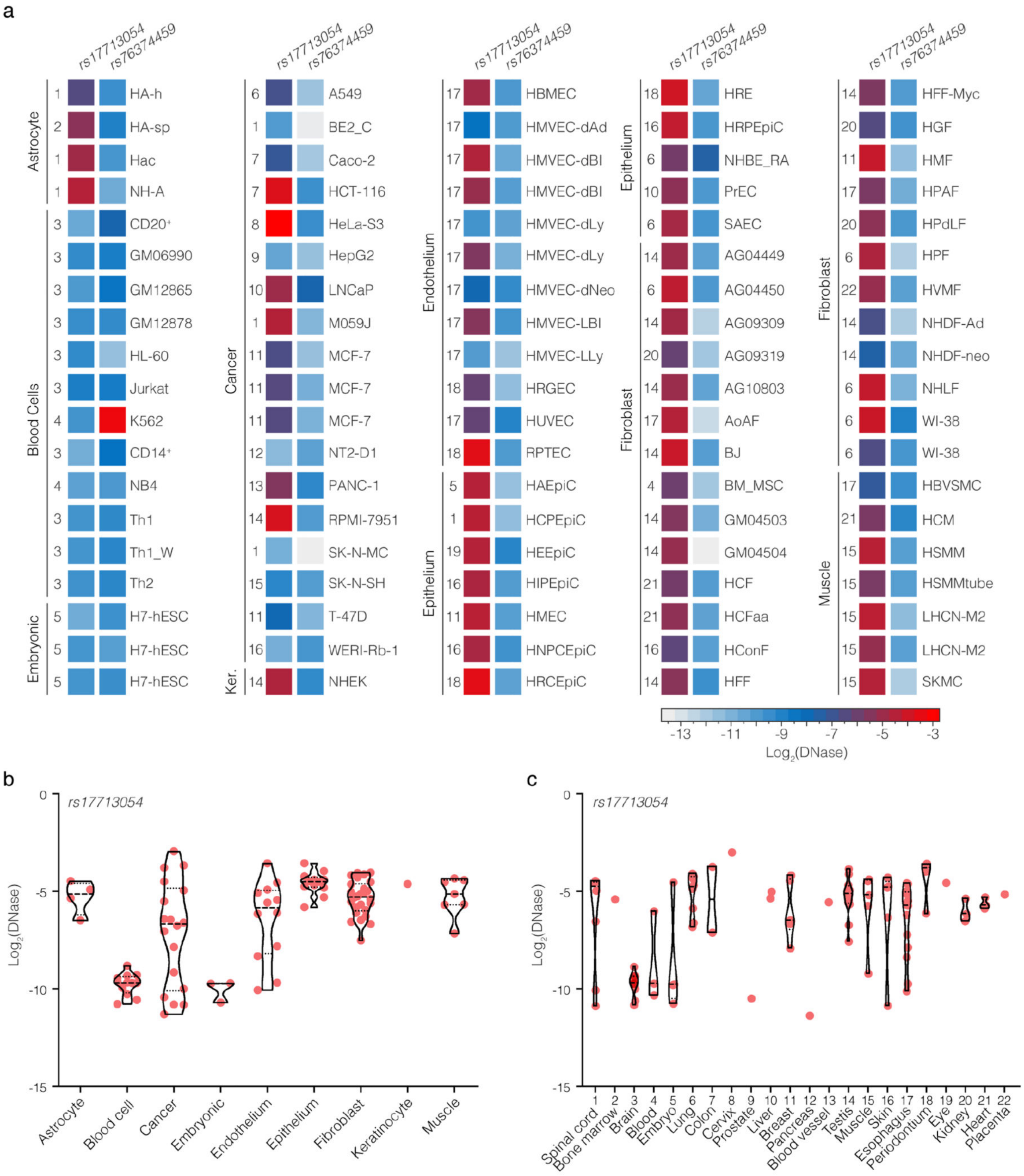
Unless stated, ENCODE datasets were accessed using the UCSC genome browser[106,107], which was also used to generate track figures. ENCODE DNase I bigwigs (hg38) were downloaded from ENCODE portal (https://www.encodeproject.org/) and analyzed with deeptools[100] (multiBigwigSummary). Capture-C was analyzed using the CaptureCompendium suite[91] mapping to hg38 with bowtie2[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 (GSE74912, GSE115684, GSE118189, GSE139377) and analyzed using NGseqBasic[102] with default settings for bowtie2[97]. Aortic endothelial samples were genotyped by counting two or more reads from either allele in combined ATAC-seq and ChIP-seq data. For allelic skew analysis, aortic endothelium ATAC-seq from heterozygous individuals was mapped with bowtie2[97] and processed using WASP[108] to correct for reference genome mapping bias. Three replicates with fewer than four remaining reads were excluded from analysis. Mature erythroid chromatin modification and CTCF data (GSE125926) have been previously reported by our group[16], CTCF motifs were identified using MEME-SUITE[85] tools, (meme --dna --nmotifs 1 --w 19 --mod zoops --maxsize 1102788; fimo --thresh 1e-4 --motif 1). Single-cell RNA-seq data[35,48] were sourced from online portal (https://asthma.cellgeni.sanger.ac.uk/, https://www.lungepigenome.org/gene-expression/) on 9th Oct. 2020 and 19th May 2021 respectively. Single-cell ATAC-seq data[34,35] were sourced from online portals (https://descartes.brotmanbaty.org/bbi/human-chromatin-during-development/dataset/lung, https://www.lungepigenome.org/) on 19th May 2021. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The multi-tissue eQTL and expression level data were obtained from the GTEx Portal V8 on the 14th Oct. 2020 (https://gtexportal.org/home/snp/rs17713054).

Extended Data



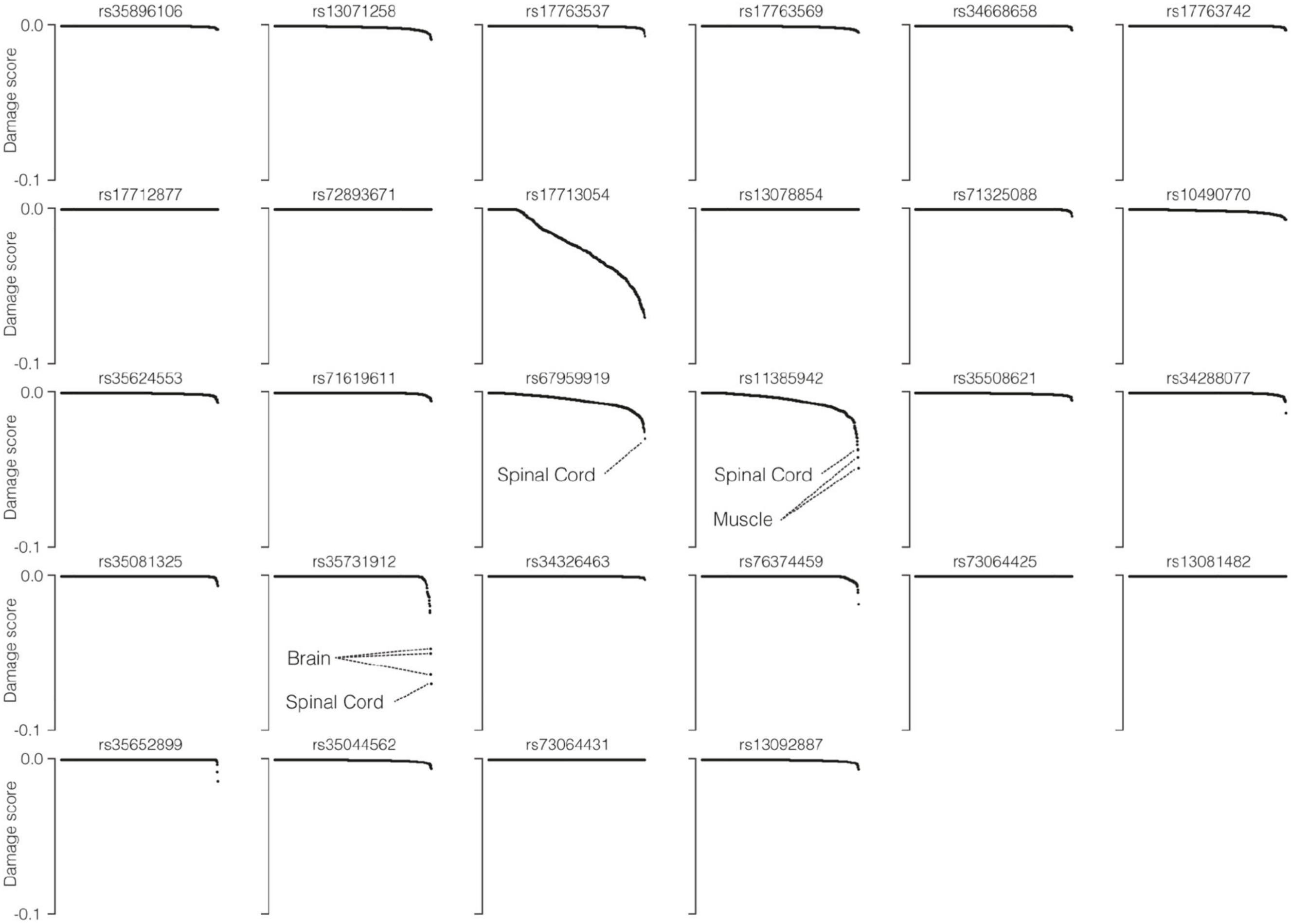
Extended Data Figure 1. 3p21.31 severe COVID-19 locus SNPs are not in immune regulatory elements.

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.



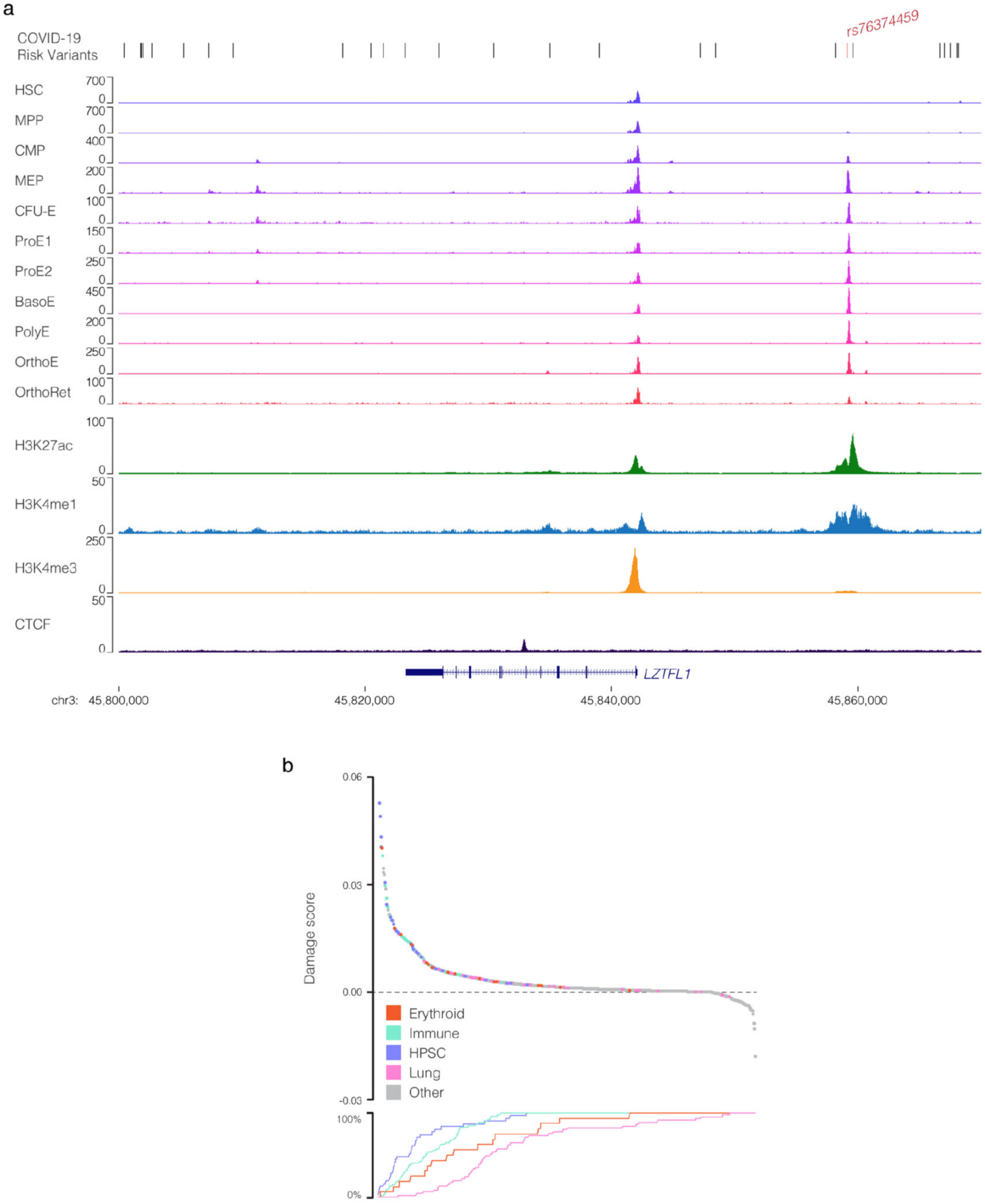
Extended Data Figure 2. DNase I accessibility over COVID-19 SNPs.

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.



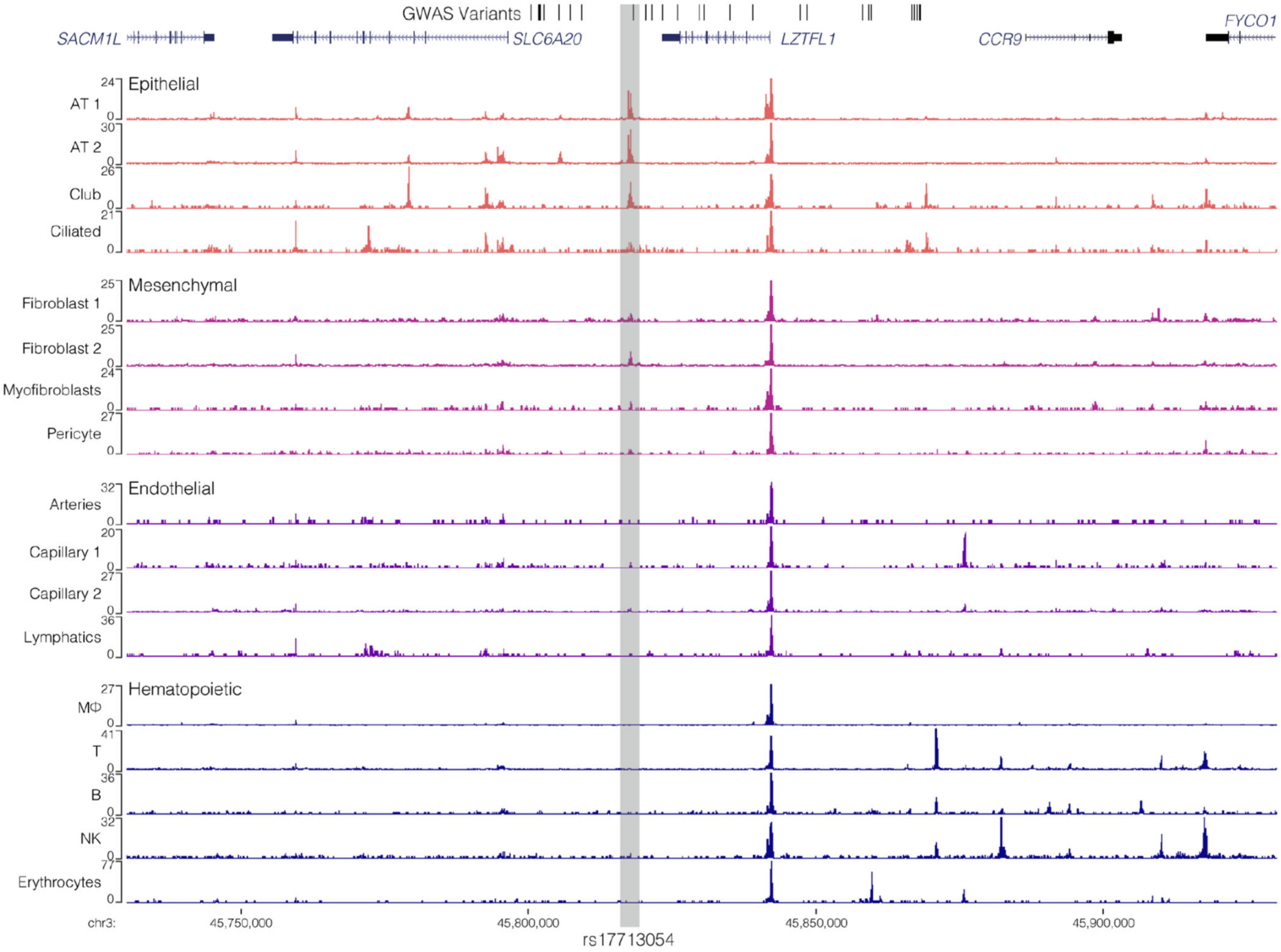
Extended Data Figure 3. deepHaem prediction of *de novo* open chromatin elements.

deepHaem[19] negative damage score, 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.



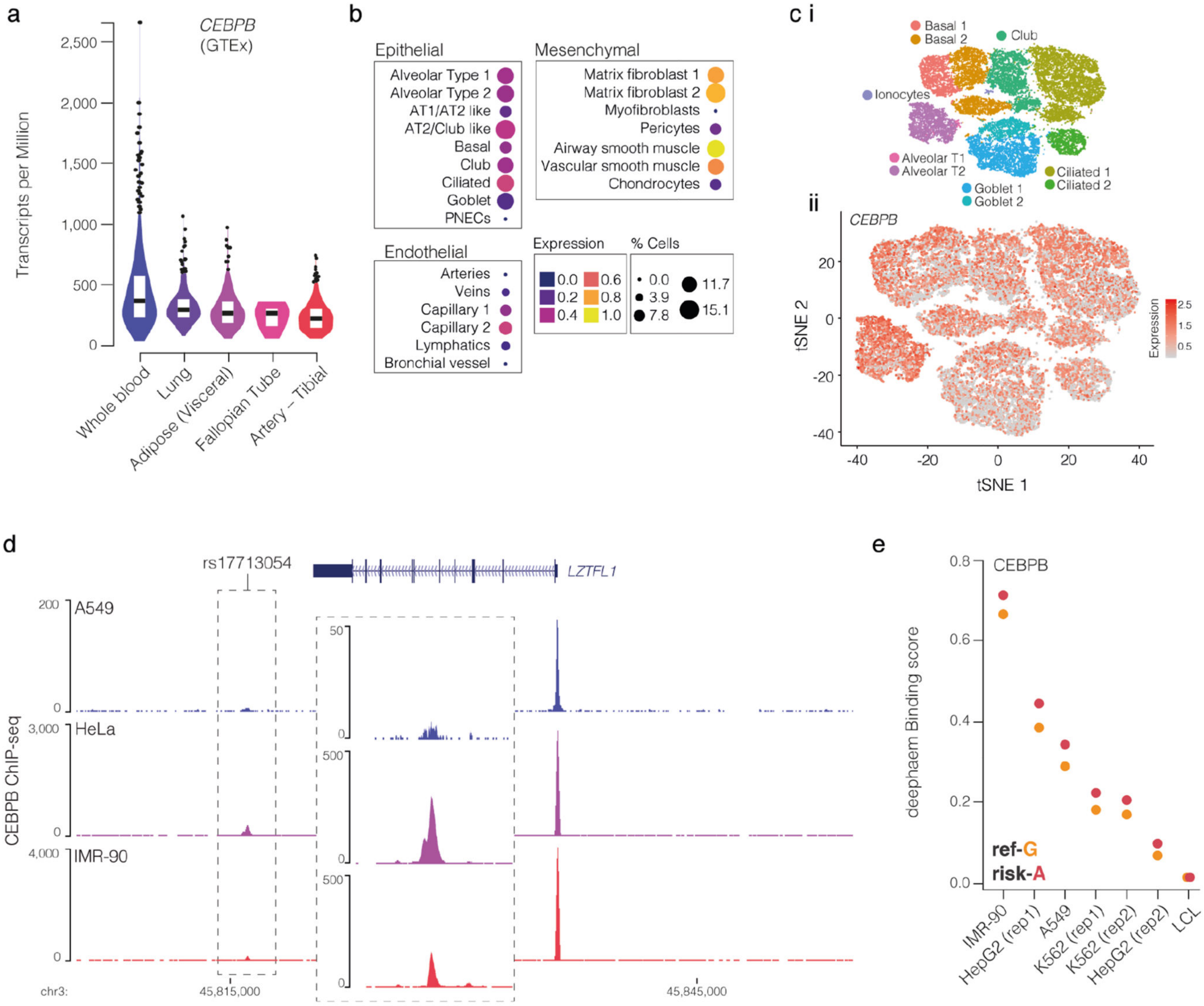
Extended Data Figure 4. rs76374459 is likely benign in an erythroid enhancer.

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 Figure 5. Single nucleus ATAC-seq in adult lung.

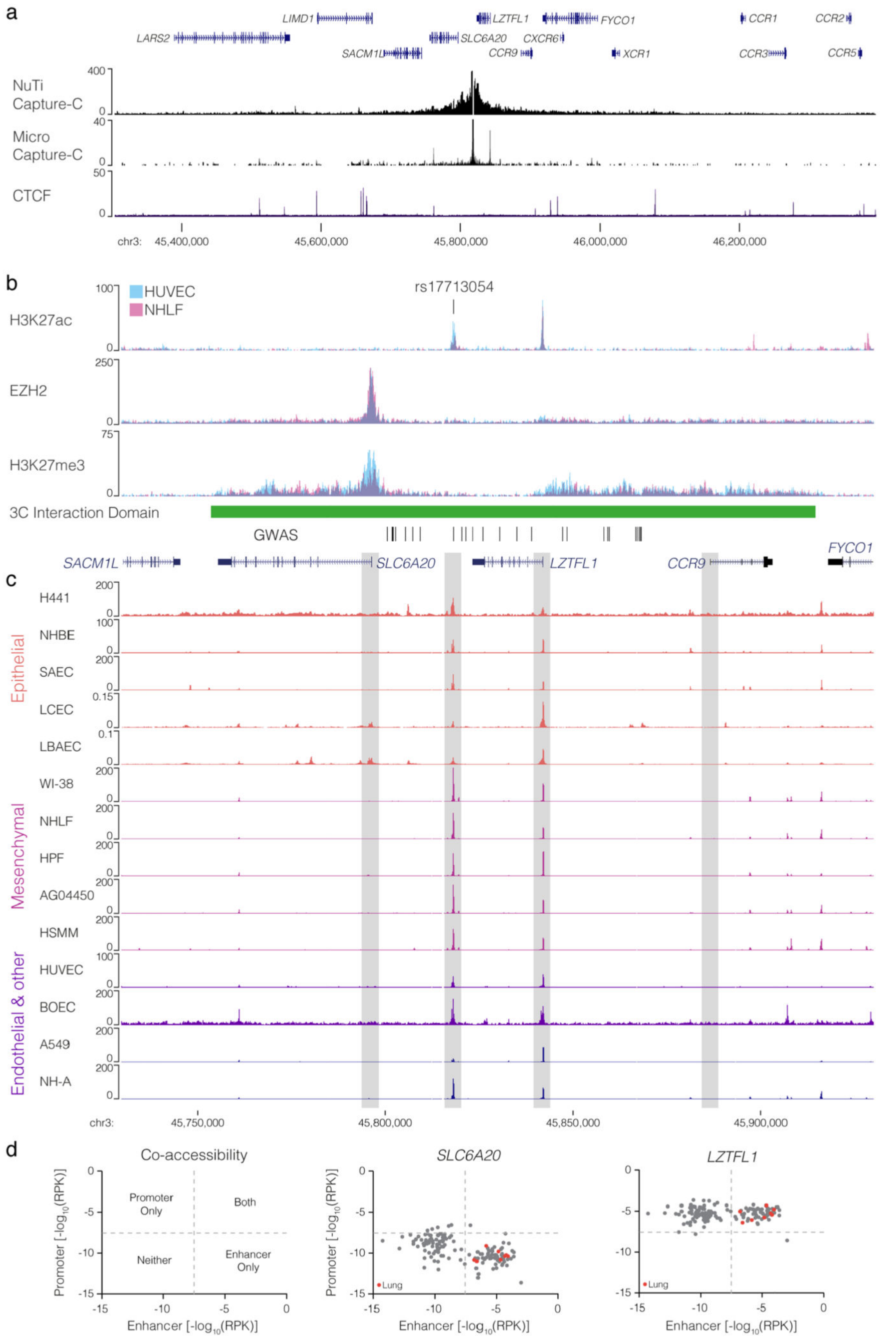
Chromium single nucleus ATAC-seq from non-diseased adult lung[35] (n=3) with 17 epithelial, endothelial, mesenchymal and hematopoietic populations, including Alveolar Type (AT) 1 and 2 Pneumocytes, Macrophage (MΦ) and Natural Killer (NK) cells. The rs17713054 containing element is highlighted in grey.



Extended Data Figure 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,**

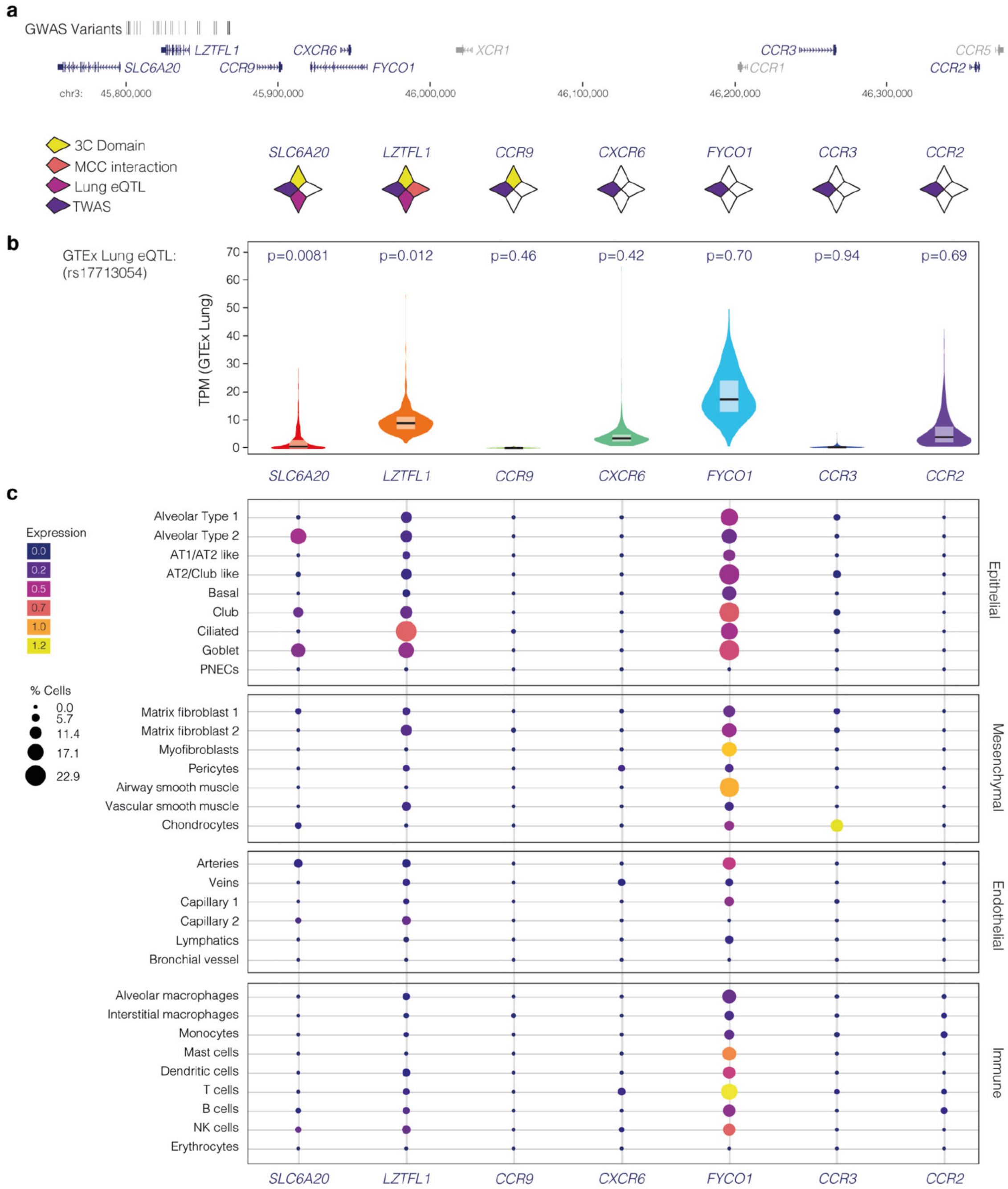
DeepHeam ChIP-seq binding prediction score for CEBPB in lung fibroblast (IMR-90), alveolar basal epithelial adenocarcinoma (A549), the erythroleukaemia line (K562), human endothelial kidney cells (HEK293), and the GM12878 lymphoblastoid cell line (LCL) predicts increased binding to the risk-A allele.



Extended Data Figure 7. *LZTFL1* is a most likely 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

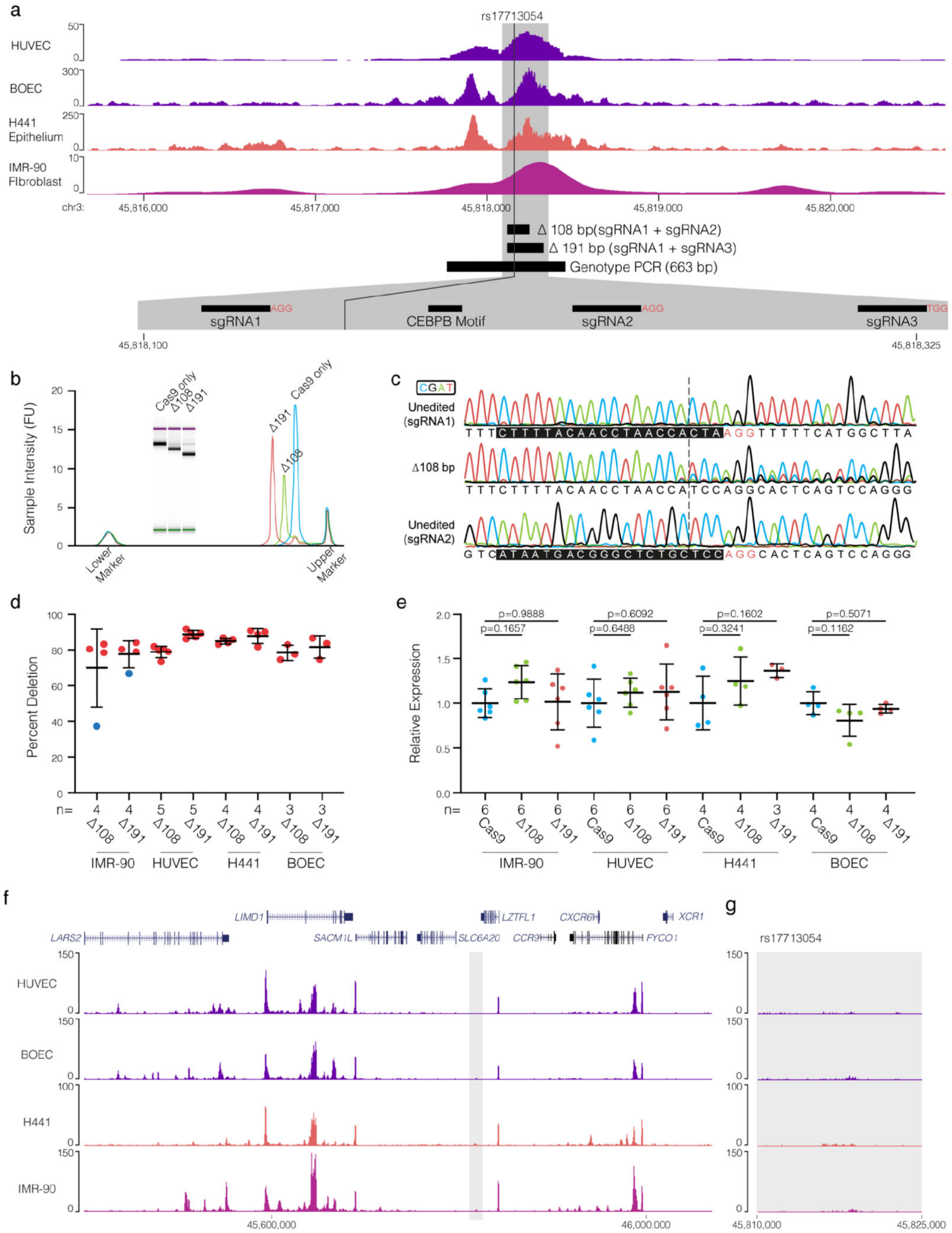
CTCF site (triangles). CTCF track shows binding of the CCAAT-binding factor which acts as a boundary. **b,** ENCODE ChIP-seq for the active chromatin mark (H3K27ac), the repressive chromatin mark (H3K27me3) and EZH2, a member of the Polycomb Repressive Complex 2, in endothelial (HUVEC) and normal human lung fibroblast (NHLF) cells. Green bar denotes the 3C regulatory domain as identified by 3C 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 Figure 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 TWASs[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

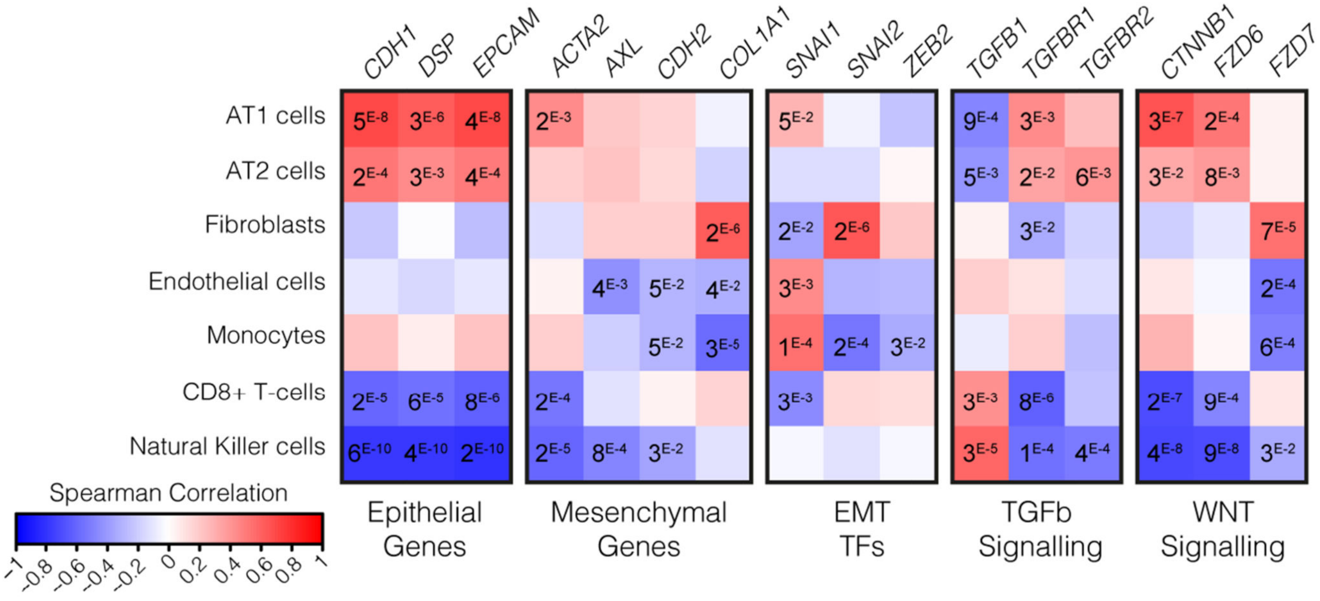
RNA-seq[35] from non-diseased adult lung (n=3), including Alveolar Type 1 (AT) and Type 2 (AT2) Pneumocytes and Pulmonary Neuroendocrine cells (PNECs).



Extended Data Figure 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, PAM 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 Figure 10. COVID-19 patient lung shows signals of EMT.

Spearman correlation of gene expression profiles for EMT-related genes with the cell-types identified by deconvolution. AT1: Alveolar Type1 pneumocytes, AT2: Alveolar Type2 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Priscila Hirschfeld, Matt Gosden, Steph Carpenter, Caz Harrold, Lea Nussbaum, Yavor Bozhilov, Andrew King, Mohsin Badat, Minnie Salmon, Lance Hentges, Andrew Brown, Guiseppe Scozzafava, Alicia Lledó Lara, Tim Rostron, Jason Torres, Chris Eijsbouts, Valentina Iotchkova, and Martin Sergeant for their help with cell culture, experiments, and data analysis, Rob Beagrie for critical reading of the manuscript, and Doug Higgs for nurturing this research.

This GWAS approach was developed as part of the WIGWAM Consortium (Wellcome Investigation of Genome Wide Association Mechanisms) funded by a Wellcome Strategic Award (106130/Z/14/Z, J.R.H.). This work was also supported by Medical Research Council (MRC) Core Funding (MC_UU_00016/14, J.R.H.). D.J.D. received funding from the Oxford University Medical Science Internal Fund: Pump Priming (0006152). R.S. was supported by a Wellcome Doctoral Programme (203728/Z/16/Z). S.N.S. received Kennedy Trust for Rheumatology Research Core support (KENN171803). F.I. received support from Wellcome (211122/Z/18/Z). J.O.J.D. is funded by an MRC Clinician Scientist Award (MR/R008108) and received Wellcome support (098931/Z/12/Z). J.C.K. is a Wellcome Investigator (WT204969/Z/16/Z) and supported by NIHR Oxford Biomedical Research Centre and the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Science (CIFMS), China (2018-I2M-2-002). A.M.M., D.R.G., and S.C.H. received support from Wellcome (110579/Z/15/Z). J.A.T. is supported by a Strategic Award from Wellcome (107212/Z/15/Z), JDRF (5-SRA-2015-130-A-N and 4-SRA-2017-473-A-N) and the Wellcome Core Award to the Wellcome Centre for Human Genetics (203141/Z/16/Z). C.E.A. and I.M. were supported by Banco Bilbao Vizcaya (BBVA) 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.

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 visualized 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, Supplementary Figs. 1,2) were sourced from the ENCODE portal (https://www.encodeproject.org/), the Gene Expression Omnibus (GSE74912, GSE115684, GSE118189, GSE125926), 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 JASPAR database (http://jaspar.genereg.net). Conserved miRNA sites were identified on miRdSNP (http://mirdsnp.ccr.buffalo.edu/browse-genes.php).

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. Note, 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

1. Zhu N, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020; 382 :727–733. [PubMed: 31978945]
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020; 20 :533–534. [PubMed: 32087114]
3. Marini JJ, Hotchkiss JR, Broccard AF. Bench-to-bedside review: Microvascular and airspace linkage in ventilator-induced lung injury. J Am Med Assoc. 2020; 323 2330
4. Levi M, Thachil J, Iba T, Levy JH. Coagulation abnormalities and thrombosis in patients with COVID- 19. Lancet Haematol. 2020; 7 :e438–e440. [PubMed: 32407672]
5. Varga Z, et al. Endothelial cell infection and endotheliitis in COVID-19. Lancet. 2020; 395 :1417–1418. [PubMed: 32325026]
6. Ackermann M, et al. Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in COVID-19. N Engl J Med. 2020; 383 :120–128. [PubMed: 32437596]
7. Visscher PM, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017; 101 :5–22. [PubMed: 28686856]
8. King EA, Wade Davis J, Degner JF. 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. 2019; 15 :1–20.
9. Ellinghaus D, et al. Genomewide Association Study of Severe COVID-19 with Respiratory Failure. N Engl J Med. 2020; doi: 10.1056/NEJMoa2020283
10. Pairo-Castineira E, et al. Genetic mechanisms of critical illness in Covid-19. Nature. 2020; doi: 10.1038/s41586-020-03065-y
11. Initiative, T. C.-19 H. G. Mapping the human genetic architecture of COVID-19. Nature. 2021; doi: 10.1101/2021.03.10.21252820
12. Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature. 2020; doi: 10.1038/s41586-020-2818-3
13. Nakanishi T, et al. Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. medRxiv. 2021 2021.03.07.21252875
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. 2021; doi: 10.1007/s10654-021-00765-1
15. ICNARC. ICNARC report on COVID-19 in critical care: England, Wales and Northern Ireland. 2021.
16. Downes DJ, et al. An integrated platform to systematically identify causal variants and genes for polygenic human traits. bioRxiv. 2019; 813618 doi: 10.1101/813618
17. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci. 2009; 106 :9362–9367. [PubMed: 19474294]
18. Jaganathan K, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019; 176 :535–548. e24 [PubMed: 30661751]
19. Schwessinger R, et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nat Methods. 2020; doi: 10.1038/s41592-020-0960-3
20. Davies JOJ, et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat Methods. 2016; 13 :74–80. [PubMed: 26595209]
21. Downes DJ, et al. High-resolution targeted 3C interrogation of cis-regulatory element organisation at genome-wide scale. Nat Commun. 2020
22. Hua P, et al. Defining genome architecture at base-pair resolution. Nature. 2021; doi: 10.1038/s41586-021-03639-4

23. Robertson CC, et al. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells,genes and drug targets for type 1 diabetes. Nat Genet. 2021; doi: 10.1038/s41588-021-00880-5

24. Patsopoulos NA, et al. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science (80-). 2019; 365

25. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015; 4 :1–38.

26. Bruno AE, et al. miRdSNP: A database of disease-associated SNPs and microRNA target sites on 3’UTRs of human genes. BMC Genomics. 2012; 13

27. Barenboim M, Zoltick BJ, Guo Y, Weinberger DR. MicroSNiPer: A web tool for prediction of SNP effects on putative microRNA targets. Hum Mutat. 2010; 31 :1223–1232. [PubMed: 20809528]

28. Genomics H. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020; 369 :1318–1330. [PubMed: 32913098]

29. Calderon D, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat Genet. 2019; 51 :1494–1505. [PubMed: 31570894]

30. Thurman RE, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489 :75–82. [PubMed: 22955617]

31. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016; :990–999. DOI: 10.1101/gr.200535.115 [PubMed: 27197224]

32. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015; 12

33. Bozhilov YK, et al. A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker. Nat Commun. 2021

34. Domcke S, et al. A human cell atlas of fetal chromatin accessibility. Science (80-). 2020; 370

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. 2020; 9 :1–28.

36. Phan, L, , et al. National Center for Biotechnology Information. National Library of Medicine; U.S: 2020. Available at: www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/

37. Stolze LK, et al. Systems Genetics in Human Endothelial Cells Identifies Non-coding Variants Modifying Enhancers, Expression, and Complex Disease Traits. Am J Hum Genet. 2020; 106 :748–763. [PubMed: 32442411]

38. Hendricks-Taylor LR, 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. 1992; 14 :12–17. [PubMed: 1427819]

39. ENCODE. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489 :57–74. [PubMed: 22955616]

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. 2017; 27 :1730–1742. [PubMed: 28904015]

41. Uehara S, Grinberg A, Farber JM, Love PE. A Role for CCR9 in T Lymphocyte Development and Migration. J Immunol. 2002; 168 :2811–2819. [PubMed: 11884450]

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. 1997; 185 :2015–2023. [PubMed: 9166430]

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. 2005; 172 :1290–1298. [PubMed: 16100013]

44. Broer S, et al. Iminoglycinuria and hyperglycinuria are discrete human phenotypes resulting from complex mutations in proline and glycine transporters. J Clin Invest. 2008; 118 :3881–3892. [PubMed: 19033659]

45. Wei Q, et al. Tumor-suppressive functions of leucine zipper transcription factor-like 1. Cancer Res. 2010; 70 :2942–2950. [PubMed: 20233871]

46. Zaghloul N, Katsanis N. Mechanistic insights into Bardet-Biedl syndrome, a model ciliopathy. J Clin Invest. 2009; 119 :428–437. [PubMed: 19252258]

47. Marion V, et al. Exome sequencing identifies mutations in LZTFL1, a bbsome and smoothened trafficking regulator, in a family with bardetebiedl syndrome with situs inversus and insertional polydactyly. J Med Genet. 2012; 49 :317–321. [PubMed: 22510444]

48. Vieira Braga FA, et al. A cellular census of human lungs identifies novel cell states in health and in asthma. Nat Med. 2019; 25 :1153–1163. [PubMed: 31209336]

49. Pathak GA, et al. Integrative genomic analyses identify susceptibility genes underlying COVID-19 hospitalization. Nat Commun. 2021; :1–11. DOI: 10.1038/s41467-021-24824-z [PubMed: 33397941]

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. 2021; 6 :2020–2022.

51. Giambartolomei C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 2014; 10

52. Mali P, et al. RNA-Guided Human Genome Engineering via Cas9. Science (80-). 2013; 339 :823–826.

53. Dongre A, Weinberg RA. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. Nat Rev Mol Cell Biol. 2019; 20 :69–84. [PubMed: 30459476]

54. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. J Clin Invest. 2009; 119 :1420–1428. [PubMed: 19487818]

55. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. Nat Rev Mol Cell Biol. 2014; 15 :178–196. [PubMed: 24556840]

56. Thiery JP, Acloque H, Huang RYJ, Nieto MA. Epithelial-Mesenchymal Transitions in Development and Disease. Cell. 2009; 139 :871–890. [PubMed: 19945376]

57. Stewart CA, et al. Lung cancer models reveal SARS-CoV-2-induced EMT contributes to COVID-19 pathophysiology. J Thorac Oncol. 2021

58. Pandolfi L, et al. Neutrophil extracellular traps induce the epithelial-mesenchymal transition: implications in post-COVID-19 fibrosis. Front Immunol. 2021; 12

59. Wei Q, et al. LZTFL1 suppresses lung tumorigenesis by maintaining differentiation of lung epithelial cells. Oncogene. 2016; 35 :2655–2663. [PubMed: 26364604]

60. Wang L, et al. LZTFL1 suppresses gastric cancer cell migration and invasion through regulating nuclear translocation of P-catenin. J Cancer Res Clin Oncol. 2014; 140 :1997–2008. [PubMed: 25005785]

61. Yang J, et al. Guidelines and definitions for research on epithelial-mesenchymal transition. Nat Rev Mol Cell Biol. 2020; 21 :341–352. [PubMed: 32300252]

62. He J, et al. Single-cell analysis reveals bronchoalveolar epithelial dysfunction in COVID-19 patients. Protein Cell. 2020; 11 :680–687. [PubMed: 32671793]

63. Borczuk AC, et al. COVID-19 pulmonary pathology: a multi-institutional autopsy cohort from Italy and New York City. Mod Pathol. 2020; 33 :2156–2168. [PubMed: 32879413]

64. Cross AR, et al. Spatial transcriptomic characterization of COVID-19 pneumonitis identifies immune pathways related to tissue injury. bioRxiv. 2021; 2021.06.21.449178 doi: 10.1101/2021.06.21.449178

65. Danaher P, et al. Advances in mixed cell deconvolution enable quantification of cell types in spatially- resolved gene expression data. bioRxiv. 2020; 2020.08.04.235168 doi: 10.1101/2020.08.04.235168

66. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9

67. Gusev A, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016; 48 :245–252. [PubMed: 26854917]

68. Singer D, et al. Defective intestinal amino acid absorption in Ace2 null mice. Am J Physiol - Gastrointest Liver Physiol. 2012; 303 :686–695.

69. Vuille-dit-Bille RN, et al. Human intestine luminal ACE2 and amino acid transporter expression increased by ACE-inhibitors. Amino Acids. 2015; 47 :693–705. [PubMed: 25534429]

70. Ravindra NG, 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. 2021; 19 :1–24.

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. 2020; 15 :1–25.

72. Starks RD, et al. Regulation of Insulin Receptor Trafficking by Bardet Biedl Syndrome Proteins. PLoS Genet. 2015; 11 :1–16.

73. Wei Q, et al. Lztfl1/BBS17 controls energy homeostasis by regulating the leptin signaling in the hypothalamic neurons. J Mol Cell Biol. 2018; 10 :402–410. [PubMed: 30423168]

74. Seo S, et al. A novel protein LZTFL1 regulates ciliary trafficking of the BBSome and smoothened. PLoS Genet. 2011; 7

75. Melms JC, et al. A molecular single-cell lung atlas of lethal COVID-19. Nature. 2021; doi: 10.1038/s41586-021-03569-1

76. Delorey TM, et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature. 2021; doi: 10.1038/s41586-021-03570-8

77. Robinot R, et al. SARS-CoV-2 infection induces the dedifferentiation of multiciliated cells and impairs mucociliary clearance. Nat Commun. 2021; :1–16. DOI: 10.1038/s41467-021-24521-x [PubMed: 33397941]

78. Ruan T, et al. H1N1 Influenza Virus Cross-Activates Gli1 to Disrupt the Intercellular Junctions of Alveolar Epithelial Cells. Cell Rep. 31 2020; 107801 [PubMed: 32610119]

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. 2020; 181 :271–280. e8 [PubMed: 32142651]

80. Corces MR, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016; 48 :1193–1203. [PubMed: 27526324]

81. Ludwig LS, et al. Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis. CellReports. 27 2019; :3228–3240. e7

82. Scott C, et al. Recapitulation of erythropoiesis in congenital dyserythropoietic anaemia type I (CDA-I) identifies defects in differentiation and nucleolar abnormalities. Haematologica. 2020; :1–27. DOI: 10.1101/744367

83. Martin-Ramirez J, Hofman M, Van Den Biggelaar M, Hebbel RP, Voorberg J. Establishment of outgrowth endothelial cells from peripheral blood. Nat Protoc. 2012; 7 :1709–1715. [PubMed: 22918388]

84. Kurita R, et al. Establishment of Immortalized Human Erythroid Progenitor Cell Lines Able to Produce Enucleated Red Blood Cells. PLoS One. 2013; 8

85. Bailey TL, et al. MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. 2009; 37 :202–208.

86. Fornes O, et al. JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. Nucleic Acids Res. 2020; 48 :D87–D92. [PubMed: 31701148]

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. 2019; 47 :D1005–D1012. [PubMed: 30445434]

88. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. 2020; 16 :1–20.

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. 2018; :1–12. DOI: 10.1101/112268 [PubMed: 29301551]

90. Downes DJ, Hughes JR. Chromosome Conformation Capture with Nuclear Titrated Capture-C (NuTi Capture-C). Protoc Exch. 2020 :1–20.

91. Telenius JM, et al. CaptureCompendium: a comprehensive toolkit for 3C analysis. bioRxiv. 2020; :1–18. DOI: 10.1101/2020.02.17.952572

92. Telenius, JM, Davies, JOJ, Hughes, JR. CCseqBasic. GitHub. 2020.

93. Downes DJ, et al. CaptureCompare. GitHub. 2020; doi: 10.5281/zenodo.4194345

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

95. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011; 27 :2957–2963. [PubMed: 21903629]

96. Kent WJ. BLAT---The BLAST-Like Alignment Tool. Genome Res. 2002; 12 :656–664. [PubMed: 11932250]

97. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9 :357–359. [PubMed: 22388286]

98. Hentges LD, Sergeant MJ, Downes DJ, Hughes JR, Taylor S. LanceOtron: a deep learning peak caller for ATAC-seq, ChIP-seq, and DNase-seq. bioRxiv. 2021; 2021.01.25.428108 doi: 10.1101/2021.01.25.428108

99. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25 :2078–2079. [PubMed: 19505943]

100. Ramirez F, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016; 44 :W160–W165. [PubMed: 27079975]

101. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013; 10 :1213–8. [PubMed: 24097267]

102. Telenius JM, Hughes JR. 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. bioRxiv. 2018; 393413 doi: 10.1101/393413

103. Schindelin J, et al. Fiji: An open-source platform for biological-image analysis. Nat Methods. 2012; 9 :676–682. [PubMed: 22743772]

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. 2020; 75 :1116–1118. [PubMed: 32839288]

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

106. Kent WJ, et al. The Human Genome Browser at UCSC. Genome Res. 2002; 12 :996–1006. [PubMed: 12045153]

107. Rosenbloom KR, et al. ENCODE Data in the UCSC Genome Browser: year 5 update. 2013; 41 :56–63.

108. van de Geijn B, Mcvicker G, Gilad Y, Pritchard JK. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015; 12 :1061–1063. [PubMed: 26366987]

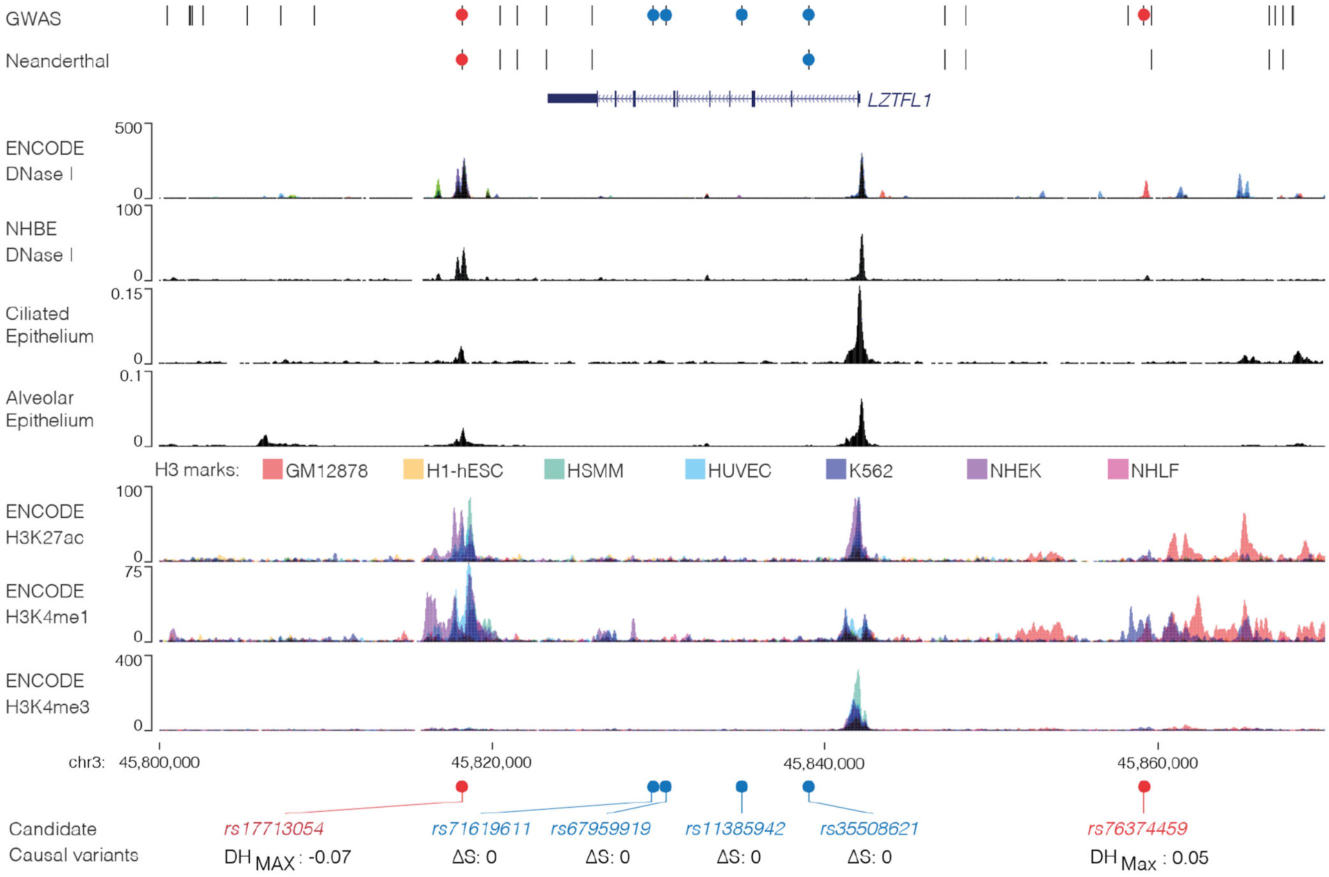
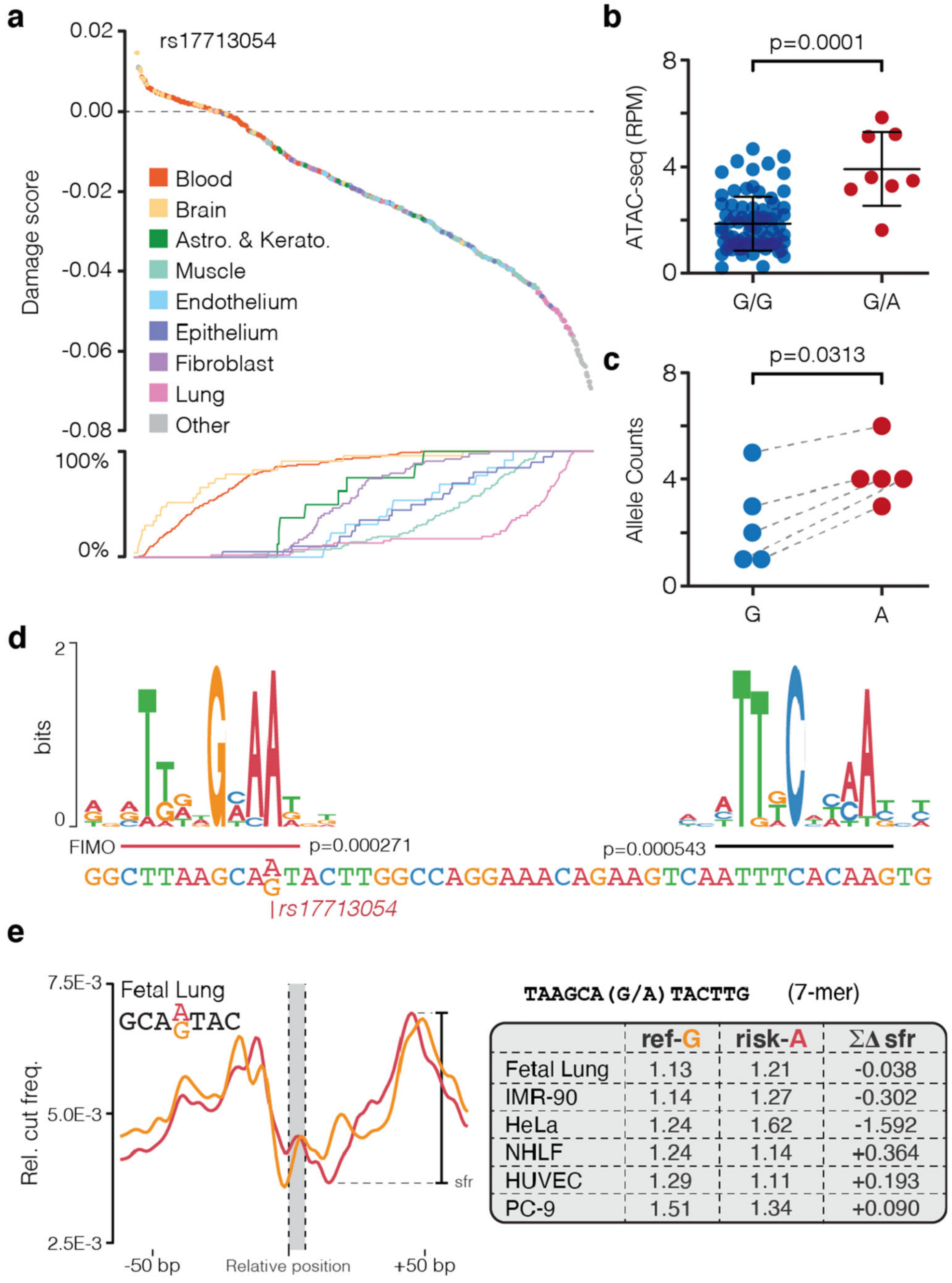


Figure 1. Identification of a potentially causative COVID-19 risk variant.

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. Circles indicate variants assessed for splicing changes (blue circles, SpliceAI[18]: ΔS score [0-1, where 1 is 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 (NHBE), and single-cell ATAC-seq from fetal ciliated epithelium and alveolar epithelium[34]. Histone H3 modification tracks show presence of marks associated with active transcription (H3K27ac) at enhancers (H3K4me1) and promoters (H3K4me3). Variants in open chromatin are given deepHaem damage scores (DH, 0-1) with sign indicating increased (-) or decreased (+) accessibility. Region shown is chr3:45,800,000-45,870,000, hg38.



TAAGCA(G/A)TACTTG (7-mer)

	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

Figure 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. Line plot shows cumulative percentage of samples for each tissue, 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. Bars show mean and one standard deviation. Two-tailed Mann-Whitney rank sum test, testing different 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, grey lines denote paired counts from a single replicate. One-sided Wilcoxon matched-pairs signed rank test, testing higher accessibility of the A allele, n = 5. Three replicates were excluded due to low coverage. **d,** CEBPB DNA binding motif over sequence around the rs17713054 risk-A and non-risk-G alleles. *P* values for 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 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 risk-A sfr from ref-G sfr, negative values show an increased footprint depth in the risk allele.

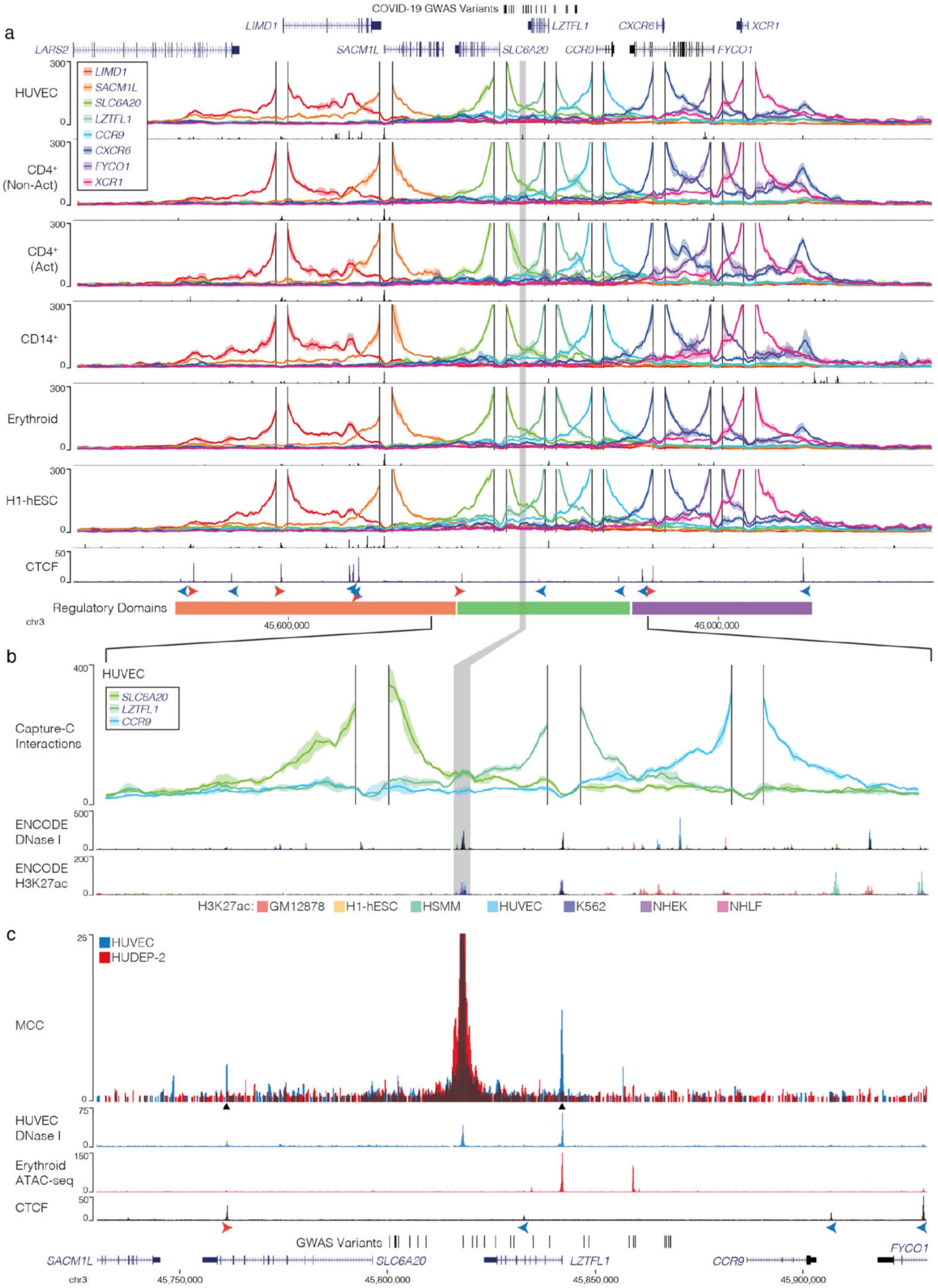


Figure 3. The interaction landscape of the severe COVID-19 risk locus.

a, *DpnII* Capture-C derived mean interaction count (n = 3 for all except CD14$^+$: n = 2) and one standard deviation (shading) for gene promoters in human vein endothelial cells (HUVEC), resting and activated T-Cells (CD4$^+$ Non-Act/Act), monocytes (CD14$^+$), CD235$^+$ CD71$^+$ erythroid cells and human embryonic stem cells (H1-hESCs). The enhancer containing rs17713054 is highlighted by a grey box. ATAC-seq/DNase I for each cell-type is shown underneath in black. CTCF track shows binding of the CCAAT-binding factor which 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 (HUVEC). 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. Solid line shows mean interaction count (n = 3 independent samples) with one standard deviation (shading). **c,** Micro Capture-C (MCC) of the rs17713054 enhancer in endothelial (HUVEC, blue) and erythroid (HUDEP-2, red) cells with tissue specific open chromatin tracks (n = 3). Peak analysis of MCC using LanceOtron to compare HUVEC and HUDEP-2 profiles identified two significantly enriched peaks in HUVEC cells (black triangles, $P \leq 1 \times 10^{-999}$) which correspond to the *LZTFL1* promoter and the upstream CTCF site.

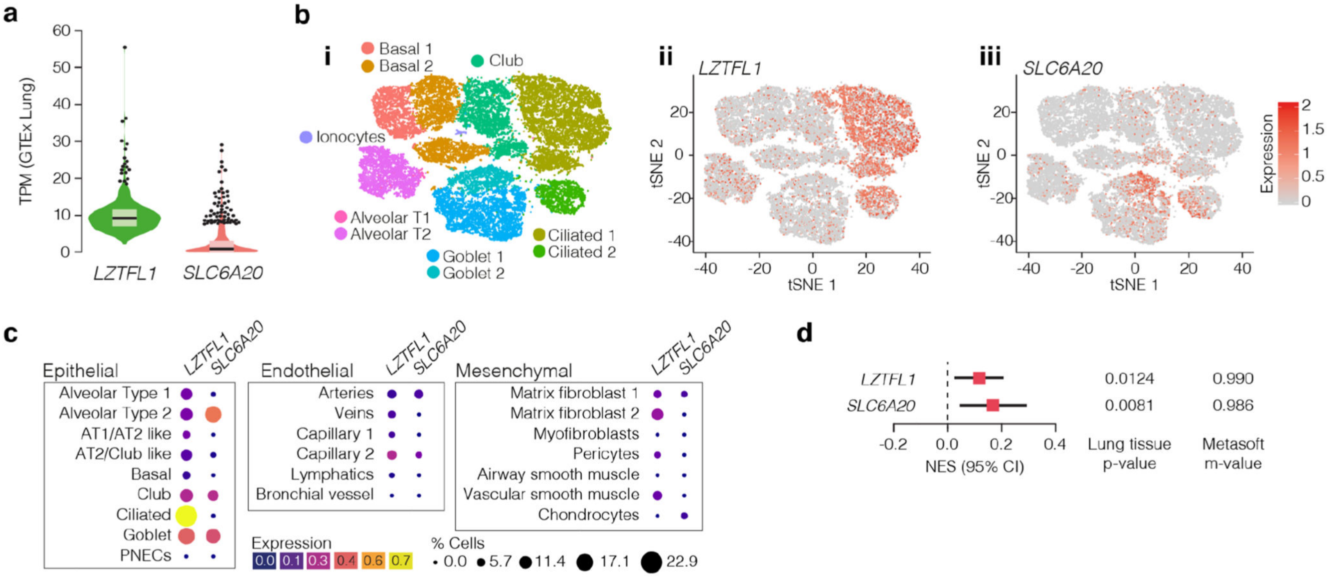


Figure 4. Pulmonary expression analysis of *LZTFL1* and *SLC6A20*.

a, GTEx whole-lung RNA-seq expression profiles for *LZTFL1* and *SLC6A20* as transcripts per million (TPM). For 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). **b,** 10x Genomics Chromium droplet single-cell RNA sequencing (scRNA-seq) from upper and lower airways and lung parenchyma[48] from healthy volunteers or deceased transplant donors with ten epithelial populations (**i**). scRNA-seq expression profiles for *LZTFL1* **(ii)** and *SLC6A20* **(iii)**. **c,** Chromium single-nucleus RNA-seq[35] from non-diseased adult lung (n = 3) with 22 epithelial, endothelial and mesenchymal populations, including alveolar type 1 (AT) and type 2 (AT2) pneumocytes and pulmonary neuroendocrine cells (PNECs). **d**, GTEx eQTL analysis the rs17713054 risk-A allele in lung (n = 515 independent samples). 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. Lines show the 95% confidence interval, with significance values for single tissue (two-sided *P* value without multiple test correction) and multi-tissue (posterior probability/m-value) analyses.

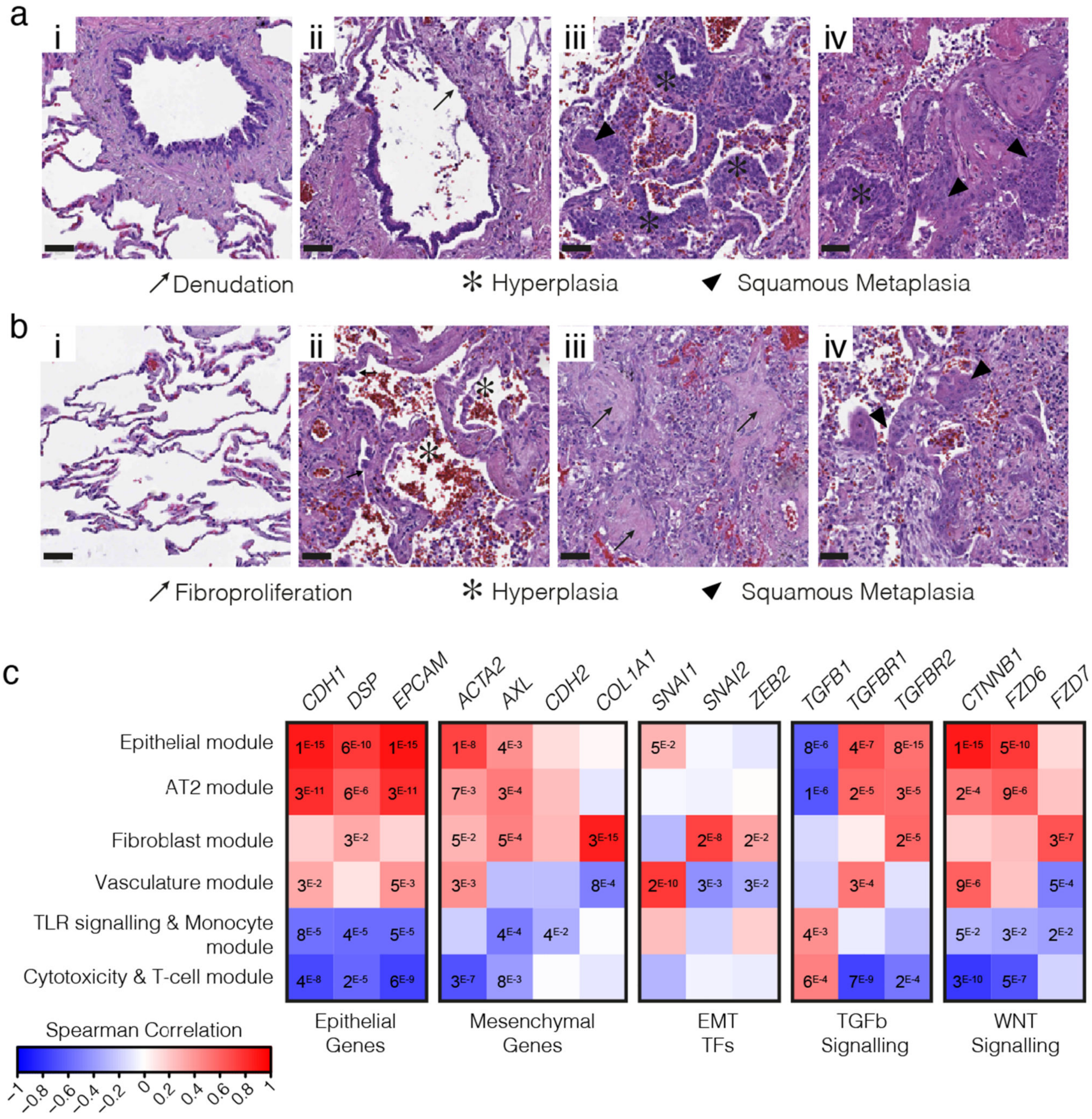


Figure 5. COVID-19 patient lungs show signals of EMT.

Hematoxylin and eosin (H&E) stained biopsies of the ciliated respiratory epithelium on bronchiole (**a**) and of alveolar space (**b**) in healthy lung (**i**) and COVID-19 patient lung (**ii-iv**). COVID-19 patient samples are representative images from staining of biopsies from 3 individuals and show loss of ciliated cell lined bronchioles (denudation) and loss of alveolar monolayers populated by alveolar type I pneumocytes with few type II pneumocytes, with alveolar wall expansion and fine interstitial fibrosis. Scale bars show 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 COVID-19 patient lung. *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 Source Data.