Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics

NHLBI LungMap Consortium; Human Cell Atlas Lung Biological Network; Muus, Christoph; Luecken, Malte D; Eraslan, Gökcen; Sikkema, Lisa; Waghray, Avinash; Heimberg, Graham; Kobayashi, Yoshihiko; Vaishnav, Eeshit Dhaval

Published in:
Nature Medicine

DOI:
10.1038/s41591-020-01227-z

ANALYSIS
https://doi.org/10.1038/s41591-020-01227-z

nature medicine

Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics

Angiotensin-converting enzyme 2 (ACE2) and accessory proteases (TMPRSS2 and CTSL) are needed for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cellular entry, and their expression may shed light on viral tropism and impact across the body. We assessed the cell-type-specific expression of *ACE2*, *TMPRSS2* and *CTSL* across 107 single-cell RNA-sequencing studies from different tissues. *ACE2*, *TMPRSS2* and *CTSL* are coexpressed in specific subsets of respiratory epithelial cells in the nasal passages, airways and alveoli, and in cells from other organs associated with coronavirus disease 2019 (COVID-19) transmission or pathology. We performed a meta-analysis of 31 lung single-cell RNA-sequencing studies with 1,320,896 cells from 377 nasal, airway and lung parenchyma samples from 228 individuals. This revealed cell-type-specific associations of age, sex and smoking with expression levels of *ACE2*, *TMPRSS2* and *CTSL*. Expression of entry factors increased with age and in males, including in airway secretory cells and alveolar type 2 cells. Expression programs shared by *ACE2*$^+$*TMPRSS2*$^+$ cells in nasal, lung and gut tissues included genes that may mediate viral entry, key immune functions and epithelial–macrophage cross-talk, such as genes involved in the interleukin-6, interleukin-1, tumor necrosis factor and complement pathways. Cell-type-specific expression patterns may contribute to the pathogenesis of COVID-19, and our work highlights putative molecular pathways for therapeutic intervention.

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, can manifest with pathologies in multiple systems, including the lungs and airways, gastrointestinal tract, kidney, liver and heart, and multi-organ failure[1–3]. SARS-CoV-2 RNA has been found in nasal and throat secretions, saliva and stool specimens[4].

Virion infection of host cells is initiated by the viral spike (S) protein binding to ACE2. *ACE2* expression has been correlated with increased viral load in human cell lines[5,6] and in mice[7]. Viral infection further requires proteolytic cleavage of the S protein, and TMPRSS2 or cathepsin L, encoded by the *CTSL* gene, can provide this role for cellular entry[8].

There is substantial variation in the clinical consequences of infection across individuals, from asymptomatic illness to death. Disease severity and mortality rise with age[9,10], with a slightly higher incidence and mortality in men[2]. Children are significantly less likely to develop severe acute disease[11]. Smoking may be associated with more severe disease[12]. Finally, adults with preexisting cardiovascular disease may have higher rates of disease acuity and death[2].

Identifying specific cell types that can be infected by SARS-CoV-2 and relating SARS-CoV-2 entry factors to key covariates like age or sex could inform our understanding of COVID-19 tropism and heterogeneity in disease outcomes. The Human Cell Atlas (HCA) community has generated single-cell atlases of diverse tissues in healthy individuals, which can now be leveraged to enable such studies. Early analyses of HCA data revealed that some of the cells of the nasal passages, airways, lung parenchyma and gut express *ACE2* and *TMPRSS2* (refs. [13,14]), most notably nasal goblet cells and multiciliated cells[13] in the airways and AT2 cells in the distal lung[13,15,16], and identified *ACE2* and *TMPRSS2* expression in colonic enterocytes[13,17].

Here, we chart the cell-type-specific expression patterns of *ACE2* and accessory proteases by integrated analysis of 116 single-cell and single-nucleus RNA-sequencing (scRNA-seq and snRNA-seq) studies, including 31 studies of the lung and airways, and 85 studies of other diverse tissues. With the lung and airway studies, to our knowledge, we performed the first single-cell meta-analysis of atlas datasets associating cell-type-specific changes in expression level with age, sex and smoking status. We identify cross-tissue and tissue-specific gene programs enriched in immune-associated genes in *ACE2*$^+$*TMPRSS2*$^+$ cells and highlight other proteases that are significantly coexpressed with *ACE2* and could play a role in infection.

Results

Double-positive *ACE2*$^+$*TMPRSS2*$^+$ cells across the lung, airways and other organs associated with COVID-19. We enumerated the proportion of double-positive *ACE2*$^+$*TMPRSS2*$^+$ cells and *ACE2*$^+$*CTSL*$^+$ cells across 92 human scRNA-seq or snRNA-seq studies, including 7 of the lung and airways (Fig. 1, Methods and Supplementary Table 1 and 2). We surveyed published datasets, assigning cells to five broad categories (Fig. 1a,b, Extended Data Figs. 1 and 2 and Supplementary Table 1), and analyzed more finely annotated published and unpublished datasets (Methods, Fig. 1c,d and Supplementary Tables 1 and 3).

ACE2$^+$*TMPRSS2*$^+$ epithelial cells were most prevalent (in order) within the ileum, liver, lung, nasal mucosa, bladder, testis, prostate and kidney (Fig. 1a). Consistent with previous reports[18], double-positive *ACE2*$^+$*TMPRSS2*$^+$ cells in the nose and airways were largely secretory goblet and multiciliated cells, and double-positive cells in the distal lung were largely alveolar type 2 (AT2) cells (Fig. 1c and Extended Data Fig. 3a). *ACE2* and *TMPRSS2* expression in secretory and AT2 cells is also supported by single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) from the primary carina and subpleural parenchyma of one adult individual, respectively, as well as secretory and multiciliated cells, and to a lesser extent some basal and tuft cells (Supplementary Fig. 1a–d; $n=3$ samples per location, $n=1$ patient; Methods). In a larger aggregation of lung and nasal datasets (Methods), we observed *ACE2*$^+$*TMPRSS2*$^+$ cells in various lung epithelial cells in pediatric samples (Extended Data Fig. 3b,c), also supported by single-cell chromatin accessibility by transposome hypersensitive sites sequencing (scTHS-seq)[19] (Extended Data Fig. 4 and Methods). Significant double-positive *ACE2*$^+$*TMPRSS2*$^+$ cells in other tissues included enterocytes, pancreatic ductal cells, prostate luminal

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

epithelial cells, brain oligodendrocytes, kidney proximal tubular cells and principal cells of the collecting duct, inhibitory enteric neurons, heart fibroblasts/pericytes, and fibroblasts and pericytes in multiple tissues (Fig. 1a–c). Notably, some of the cell types in which there were double-positive cells (including brain oligodendrocytes, multiciliated cells of the upper respiratory tract and sustentacular cells in olfactory epithelium) are cell types that also express *MYRF* (albeit not always significant triple expressors; Supplementary Fig. 2). MYRF is a transcription factor that induces expression of myelin basic protein and myelin oligodendrocyte glycoprotein[20]. Autoimmune reactions against these proteins are known to potentially induce neurological symptoms (Discussion).

ACE2$^+$*CTSL*$^+$ coexpressing cells were enriched among AT1 and AT2 cells, enterocytes, ventricular cardiomyocytes and heart macrophages, as well as fibroblasts and pericytes in multiple tissues, including the placenta, heart, lung, kidney and enteric nervous system (ENS; Fig. 1d). We did not observe substantial *ACE2* mRNA expression in scRNA-seq profiles in the bone marrow or cord blood (Fig. 1a,b), although there was *ACE2* expression in alveolar and heart macrophages (Extended Data Fig. 5). Notably, in human placenta[21–23], *ACE2* was expressed (1.4%) in maternal decidual/stromal cells, maternal pericytes and fetal extravillous trophoblasts, cytotrophoblasts and syncytiotrophoblasts in both first-trimester and term placenta (Fig. 1d). While there was little expression of *TMPRSS2* (0.2%), *CTSL* was expressed in most cells (56%), and there were *ACE2*$^+$*CTSL*$^+$ double-positive cells (1.3%).

Cell-type-specific expression of additional proteases that may be relevant to infection. SARS-CoV-2 infects cells in the absence of *TMPRSS2* (ref. [8]), so additional proteases likely play roles in proteolytic cleavage of viral proteins for entry and egress. To predict such proteases, we tested the coexpression of *ACE2* with each of 625 annotated human protease genes[24] in a declined donor transplant dataset ('regev/rajagopal'; Supplementary Table 1). *TMPRSS2* was significantly coexpressed in multiple lung epithelial cell types (Fig. 2a and Supplementary Tables 4 and 5), as were multiple members of the proprotein convertase subtilisin kexin (*PCSK*) family (Fig. 2a,b), including *FURIN*, *PCSK2*, *PCSK5*, *PCSK6* and *PCSK7* in AT2 cells. Proprotein convertases have known roles in coronavirus S-protein priming. We obtained similar results in an independent dataset from 40 samples (Extended Data Fig. 6a,b, Supplementary Table 1 and datasets 'kropski', 'lafyatis/rojas', 'misharin_new', 'nawijn/teichmann', 'northwestern_misharin_ 2018reyfman' and 'sanger_meyer_2019madissoon'; collectively referred to as 'aggregated lung'). As previously reported[25], the SARS-CoV-2 S protein has a polybasic motif in the S1/S2 region (Extended Data Fig. 6c) that corresponds to cleavage motifs of PCSK family proteases (Extended Data Fig. 6d)[25] and an additional site at the S2′ position (Extended Data Fig. 6e)[26].

FURIN, *PCSK5* and *PCSK7* were coexpressed with *ACE2* across multiple lung cell types (Fig. 2c and Extended Data Fig. 6f). *PCSK1* and *PCSK2* were mostly restricted to neuroendocrine cells[27]; *PCSK2* was also detected in some AT2 cells (Fig. 2d and Extended Data Fig. 6g). In AT2 cells, proximal multiciliated cells and basal cells, dual expression of PCSK proteases with *ACE2* was at fractions comparable to or higher than that of *ACE2*$^+$*TMPRSS2*$^+$ cells (Fig. 2e and Extended Data Fig. 6h). Coexpression was significant across other tissues (Extended Data Fig. 6i,j), including liver, ileum, kidney and nasal airways.

Because different host proteases may contribute to different stages of the viral life cycle[26], we examined the prevalence of *ACE2*$^+$*TMPRSS2*$^+$*PCSK*$^+$ triple-positive cells in the lung. *ACE2*$^+$*TMPRSS2*$^+$*PCSK7*$^+$ were the main triple-positive cells in multiciliated (0.75%) and secretory (0.72%) cells of proximal airways, and *ACE2*$^+$*TMPRSS2*$^+$*FURIN*$^+$ triple-positive cells were the most common within AT2 cells (0.36%; Extended Data Fig. 6k). Among all known human proteases (Fig. 2f and Supplementary Fig. 3), cathepsins (*CTSB*, *CTSC*, *CTSD*, *CTSL* and *CTSS*), proteasome subunits (*PSMB2, PSMB4* and *PSMB5*) and complement proteases (*C1R*, *C2* and *CFI*) were the most commonly coexpressed with *ACE2* in lung epithelial cell types.

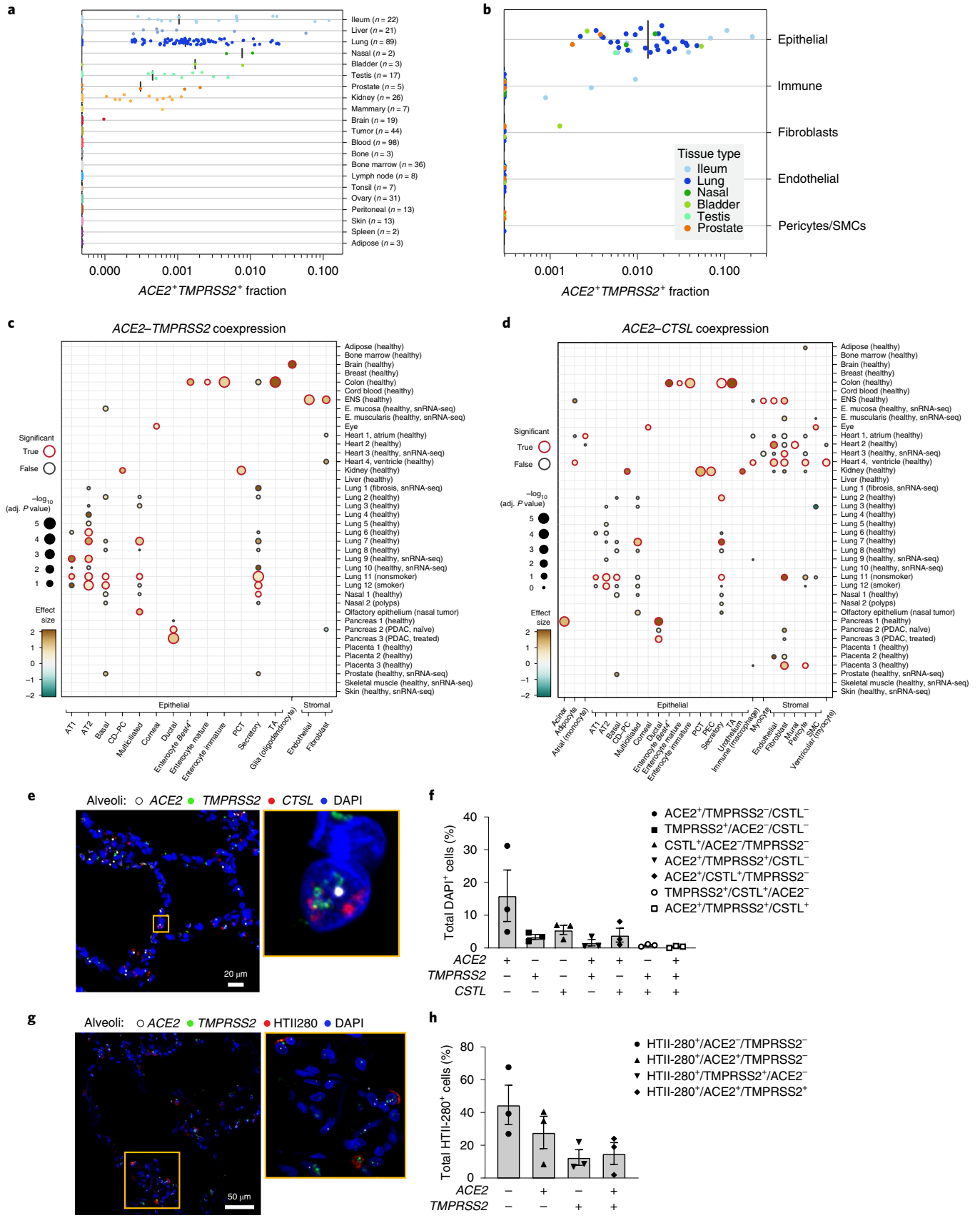
Orthogonal validation of *ACE2*, *TMPRSS2* and *CTSL* expression in the lungs. As *ACE2* expression was quite low, we next validated some of these patterns by fluorescence in situ hybridization and immunofluorescence in tissue sections of airways and alveoli from three healthy donor lungs that were rejected for lung transplantation. *ACE2*, *CTSL* and *TMPRSS2* were coexpressed by fluorescence in situ hybridization in alveolar cells, albeit at low levels (Fig. 1e,f). Co-staining with cell-type-specific markers showed *ACE2* expression and *TMPRSS2* expression in some HTII-280$^+$ AT2 cells (Fig. 1g,h); we confirmed the latter by TMPRSS2 protein immunostaining (Extended Data Fig. 7d). TMPRSS2 protein was expressed at low levels in some AT1 cells (identified by AGER; Extended Data Fig. 7d). Some non-epithelial cells also expressed these three genes. We further validated *ACE2* expression by bulk mRNA-seq of sorted AT2 cells (Extended Data Fig. 7e). Immunohistochemistry with antibodies used previously to block cellular viral entry specifically labeled adult pro-SFTPC$^+$ AT2 cells (Extended Data Fig. 7c, Supplementary Table 6 and Methods).

Previous studies revealed that ACE2 is highly enriched in nasal and intestinal mucous cells[13,14]. While mucous cells are relatively rare in healthy surface airway epithelium, they are abundant in submucosal glands (SMGs). Analysis by scRNA-seq of microdissected SMGs from healthy donors showed enrichment of *ACE2*, *TMPRSS2* and *CTSL* in mucous cells (Extended Data Fig. 7f). In situ analysis confirmed the presence of *ACE2* transcripts in acinar epithelial cells of the SMGs (Extended Data Fig. 7g) and cells expressing *ACE2* in the large airway epithelium (Extended Data Fig. 7).

Association of *ACE2*, *TMPRSS2* and *CTSL* expression in lung and airway cells with age, sex and smoking. We next asked how the expression of *ACE2*, *TMPRSS2* and *CTSL* in specific cell subsets relates to three key covariates associated with more severe disease: age (older individuals), sex (males) and smoking[28]. As no single

Fig. 1 | A cross-tissue survey of *ACE2*$^+$*TMPRSS2*$^+$ cells shows enrichment in cells at reported sites of disease transmission or pathogenesis. **a**,**b**, Double-positive cells were more prevalent in epithelial organs and cells. **a**, Proportion of *ACE2*$^+$*TMPRSS2*$^+$ cells per dataset (dots) from 21 tissues and organs (rows). **b**, Proportion of *ACE2*$^+$*TMPRSS2*$^+$ cells within cell clusters (dots) annotated by broad cell-type categories (rows) within each of the top seven enriched datasets. SMCs, smooth muscle cells. **c**,**d**, Significant coexpression of *ACE2*$^+$*TMPRSS2*$^+$ or *ACE2*$^+$*CTSL*$^+$ highlights cells from tissues implicated in transmission or pathogenesis. Significance of coexpression (dot size; $-\log_{10}$ adjusted (adj.) *P* value), by two-sided Wald test (Methods); red border: false discovery rate (FDR) < 0.1) of *ACE2*$^+$*TMPRSS2*$^+$ (**c**) or *ACE2*$^+$*CTSL*$^+$ (**d**) and effect size (dot color, color bar) for finely annotated cell classes (columns) from diverse tissues (rows). Only tissues and cells in at least one significant coexpression relationship are shown (Methods). PDAC, pancreatic ductal adenocarcinoma; CD–PC, collecting duct principal cell; PEC, parietal epithelial cell; PCT, proximal convoluted tubule; TA, transit amplifying. **e**–**h**, In situ validation of double-positive cells in the lung, airways and SMGs ($n = 3$ donors per experiment, images of three randomly chosen areas per donor). Proximity ligation in situ hybridization (PLISH) and immunostaining (**e** and **g**) and quantification (error bars: standard errors; **f** and **h**) in human adult lung alveoli for *ACE2* (white), *TMPRSS2* (green) and *CTSL* (**e**) (red; total of 1,487 DAPI-positive cells examined for quantification (**f**)) and *ACE2* (white), *TMPRSS2* (green) and HTII-280 (**g**) (red; total of 482 HTII-280-positive cells examined for quantification (**h**)).

dataset to date was sufficiently large, we aggregated samples across 31 scRNA-seq and snRNA-seq studies (Supplementary Table 2; 14 published[16,18,29–38]; 17 not yet published[39,40] at the time of writing). This analysis spanned 1,320,896 cells from 228 individuals without known lung disease or from histologically normal-appearing lung adjacent to the site of disease, across 377 nasal, lung and



airway samples from brushes, scrapings, biopsies, bronchoalveolar lavages, resections or entire lungs that could not be used for transplant or postmortem examinations (Fig. 3a). From unpublished data, we only obtained single-cell expression counts for the three genes (preprocessed by each data generator), total unique molecular identifier (UMI) counts per cell, cell identity annotations (which we harmonized to three resolution levels across studies; Fig. 3a,b, Supplementary Table 2, Extended Data Fig. 8 and Methods), and age, sex and smoking status (when ascertained). We modeled the association between the expression counts of each gene and age, sex and smoking status using a generalized linear model, accounting for technical variation arising from dataset-related factors and covariate interactions (Methods). We fitted this model within each cell type to non-fetal lung data of donors for whom smoking history was known (985,420 cells, 286 samples, 164 donors, 21 datasets) and fitted a model without smoking status covariates to the full non-fetal lung data (1,096,604 cells, 309 samples, 185 donors, 24 datasets).

For simplicity, we treated each cell as an independent observation. This implicitly combines variability in both donors and cells, and, because cells from the same donor are not truly independent observations, can result in inflated *P* values, especially when there are few donors for a particular cell type. To address this, account for covariate interactions and ensure robustness, we (1) used a simple noise model (Poisson) to reduce overfitting of donor variability; (2) confirmed that effect directions of significant associations were consistent in a pseudo-bulk analysis (modeling only donor variation; Methods and Supplementary Data 1–4); (3) confirmed summarized age, sex and smoking associations with a model including interaction terms (Methods and Supplementary Data 1–4); and (4) separated significant associations that passed all above confirmations into 'robust trends' and 'indications' depending on their robustness to holding out individual datasets (Methods and Supplementary Data 1–4). We focused on trends or indications in cell types where *ACE2* and *TMPRSS2* were coexpressed (Fig. 3c): airway epithelial cells (basal, multiciliated and secretory cells), AT1 and AT2 cells and SMG secretory cells.

We found robust trends of *ACE2* expression with age, sex and smoking status in these cell types (Fig. 3d, Extended Data Fig. 9 and Supplementary Figs. 4–6; nonsmoking model results in Supplementary Figs. 7–10): *ACE2* expression increased with age in AT2 cells, and was elevated in males in airway secretory cells and AT1 and AT2 cells. *ACE2* levels were higher in past or current smokers in basal and submucosal secretory cells, and lower in AT2 cells (Fig. 3d). Analysis of bulk RNA-seq data from bronchial brushings[41] indicated an upregulation of both *ACE2* and *TMPRSS2* in current smokers compared with former smokers (Extended Data Fig. 10). Furthermore, we found indications of increased *ACE2* expression with age and in males in multiciliated cells, but those relied on inclusion of the dataset with the most cells and samples ('regev/rajagopal'; Extended Data Fig. 9 and Methods). All above trends and indications for sex and age were validated in a simplified model without smoking status on the full non-fetal lung dataset (Supplementary Fig. 7, Supplementary Data 5–8 and Methods).

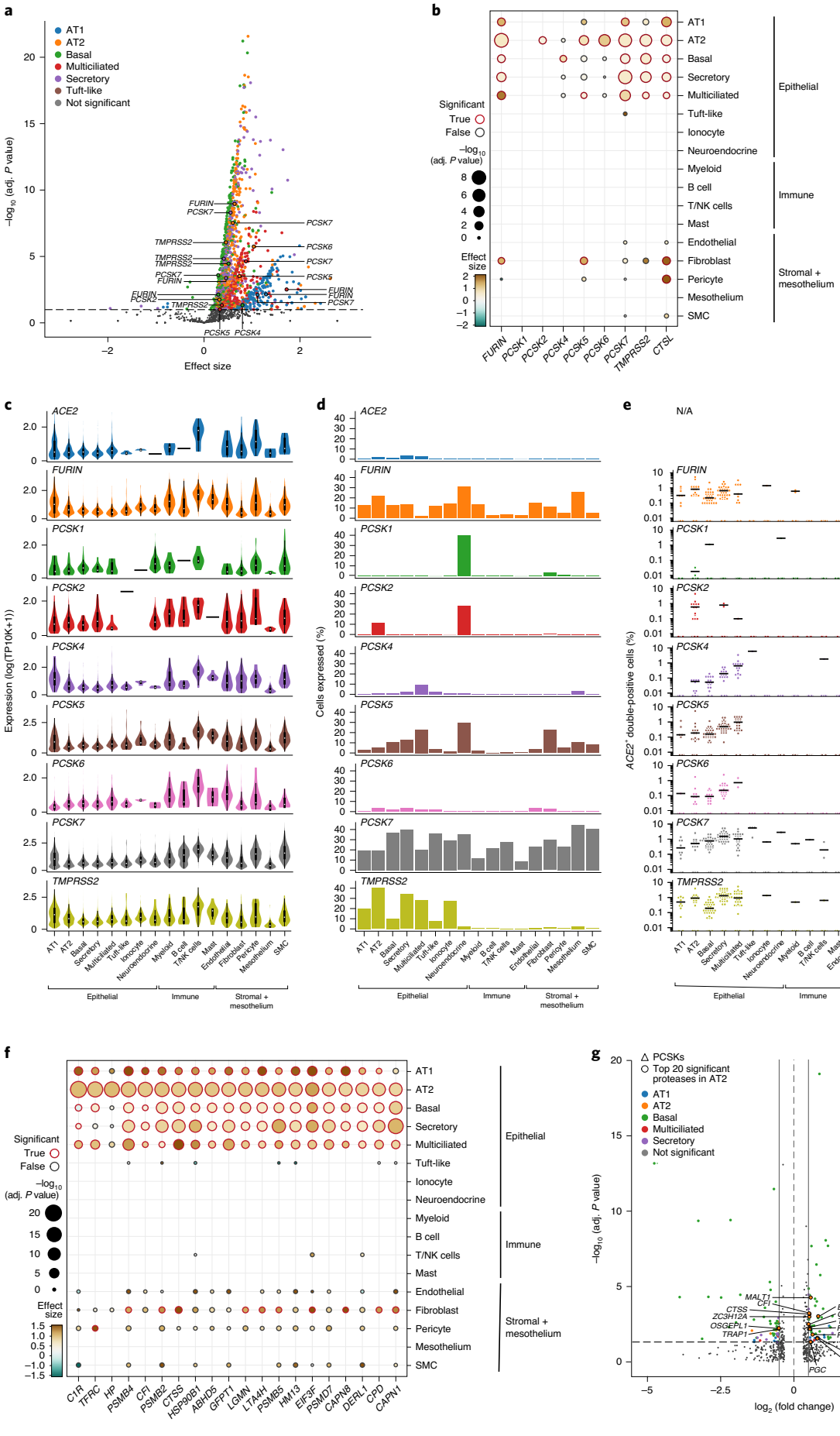
Examining joint trends of *ACE2* and the protease genes within the same cell type, we found robust trends of *ACE2* and *TMPRSS2* coexpression increasing with age in AT2 cells, in males in AT1 cells, and an indication of the two genes being elevated in males in multiciliated cells (*ACE2* indication dependent on the 'regev/rajagopal' dataset; Fig. 3d and Extended Data Fig. 9). *ACE2* and *CTSL* showed robust trends of joint upregulation in males in AT2 cells, and in smokers in submucosal secretory cells. Indications of joint upregulation of these genes were found in males in AT1 cells, and in smokers in basal cells (Fig. 3d, Extended Data Fig. 9 and Methods). All joint trends for age and sex covariates were confirmed on the full non-fetal lung data using the simple model without smoking covariates (Supplementary Fig. 7).

An immune gene program in *ACE2*$^+$*TMPRSS2*$^+$ cells in airway, lung and gut. Our previous analyses revealed immune signaling genes that covary with *ACE2* and *TMPRSS2* in airway and lung cells[13,14]. To explore these in a broader context, we identified tissue and cell programs related to double-positive *ACE2*$^+$*TMPRSS2*$^+$ cells in the nasal epithelium, lung and gut (Supplementary Tables 7–10). Tissue programs are shared across double-positive cells from different cell types in one tissue; cell programs distinguish double-positive cells from the rest of the cells of the same type (Methods).

Tissue programs were enriched in pathways related to viral infection and immune response, including phagosome structure, antigen processing and presentation, and apoptosis (Fig. 4a,b, Supplementary Fig. 11a,b (for selected genes) and Supplementary Tables 7–10). These included *CEACAM5* (lung, nasal and gut programs) and *CEACAM6* (ref. [42]; lung), surface attachment factors for coronavirus S protein; *SLPI* (lung and nasal)[43]; *PIGR* (lung and gut; may promote antibody-dependent enhancement via IgA[44]); and *CXCL17* (lung and nasal)[45]. Tissue programs also had genes associated with cholesterol and lipid metabolic pathways and endocytosis (*DHCR24*, *LCN2* and *FASN*), major histocompatibility complex I and II pathways[46], preparation against cellular injury (interferons; extracellular RNase: *PLAC8* and *TXNIP*), complement (*C3* and *C4BPA*), immune modulation (*BTG1*) and tight junctions (*DST*, *CLDN3* and *CLDN4*).

Cell programs (Fig. 4c,d, Supplementary Fig. 12a–c and Supplementary Tables 7–10) were enriched in many of the same genes and pathways (for example, *CEACAM5*, *CXCL17* and *SLPI*), and further captured unique functions, including tumor necrosis factor (TNF) signaling in lung secretory cells (for example, *RIPK3*; ref. [47]), lysosomal functions in lung secretory and multiciliated cells[48], the immunoproteasome (AT1 cells; Fig. 4c), cytokines, chemokines and their receptors (nasal goblet cells: *CSF3*, *CXCL1*, *CXCL3*, *IL19* and *CCL20*; AT1 cells: *IL1R1*) and genes that encode surfactant

Fig. 2 | *ACE2*–protease coexpression and SARS-CoV-2 S-protein cleavage sites suggest a possible role for additional proteases in infection. a, Multiple proteases were coexpressed with *ACE2* in human lung scRNA-seq data. Scatterplot of significance ($-\log_{10}$ adj. *P* value), by two-sided Wald test (Methods) and effect size of coexpression of each protease gene (dot) with *ACE* within each indicated epithelial cell type (color). Dashed line: significance threshold. *TMPRSS2* and PCSK proteases that were significantly coexpressed with *ACE2* are marked. **b**, *ACE2*–protease coexpression with PCSKs, *TMPRSS2* and *CTSL* across lung cell types. Significance (dot size; $-\log_{10}$ (adj. *P* value), by two-sided Wald test. (Methods)) and effect size (color) for coexpression of *ACE2* with selected proteases (columns) across cell types (rows). NK, natural killer. **c**,**d**, Multiple proteases were expressed across lung cell types. **c**, Distribution of non-zero expression for *AC2*, *PCSK* and *TMPRSS2* across lung cell types. White dot: median non-zero expression. **d**, Proportion of cells expressing *ACE2*, *PCSK* or *TMPRSS2* across lung cell types, ordered by compartment. **e**, *ACE2*$^+$*PCSK*$^+$ double-positive cells across lung cell types. Fraction of different *ACE2*$^+$*PCSK*$^+$ or *ACE2*$^+$*TMPRSS2*$^+$ double-positive cells across lung cell types. Dots: different samples; line: median of non-zero fractions. **f**, *ACE2*–protease coexpression analysis for the 20 most significant human proteases in AT2 cells. Significance (dot size; $-\log_{10}$ (adj. *P* value), by two-sided Wald test (Methods)) and effect size (color) for coexpression of *ACE2* with different proteases (columns) across cell types (rows). **g**, Additional protease expression in *ACE2*$^+$*TMPRSS2*$^+$ double-positive cells. Significance ($-\log_{10}$ adj. *P* value, by two-sided Wald test (Methods)) and fold change of differential expression for each human protease between *ACE2*$^+$*TMPRSS2*$^+$ double-positive versus double-negative cells within each indicated epithelial cell type (color). Significantly differentially expressed proteases within AT2 cells and PCSK across all epithelial cell types are highlighted.


a
AT1
AT2
Basal
Multiciliated
Secretory
Tuft-like
Not significant
−log10 (adj. P value)
Effect size
FURIN
PCSK7
TMPRSS2
PCSK6
PCSK5
PCSK4
PCSK2
PCSK1
b
Significant
True
False
−log10 (adj. P value)
Effect size
AT1
AT2
Basal
Secretory
Multiciliated
Tuft-like
Ionocyte
Neuroendocrine
Myeloid
B cell
T/NK cells
Mast
Endothelial
Fibroblast
Pericyte
Mesothelium
SMC
Epithelial
Immune
Stromal + mesothelium
FURIN
PCSK1
PCSK2
PCSK4
PCSK5
PCSK6
PCSK7
TMPRSS2
CTSL
c
ACE2
FURIN
PCSK1
PCSK2
PCSK4
PCSK5
PCSK6
PCSK7
TMPRSS2
Expression (log(TP10K+1))
Epithelial
Immune
Stromal + mesothelium
d
Cells expressed (%)
e
N/A
ACE2+ double-positive cells (%)
f
Significant
True
False
−log10 (adj. P value)
Effect size
C1R
TFRC
HP
PSMB4
CFI
PSMB2
CTSS
HSP90B1
ABHD5
GFPT1
LGMN
LTA4H
PSMB5
HM13
EIF3F
PSMD7
CAPN8
DERL1
CPD
CAPN1
g
PCSKs
Top 20 significant proteases in AT2
AT1
AT2
Basal
Multiciliated
Secretory
Not significant
−log10 (adj. P value)
log2 (fold change)
MALT1
CFI
CTSS
ZC3H12A
OSGEPL1
TRAP1
BACE1
GGT6
TMPRSS13
PCSK6
PLAT
LTF
MMP7
PGC

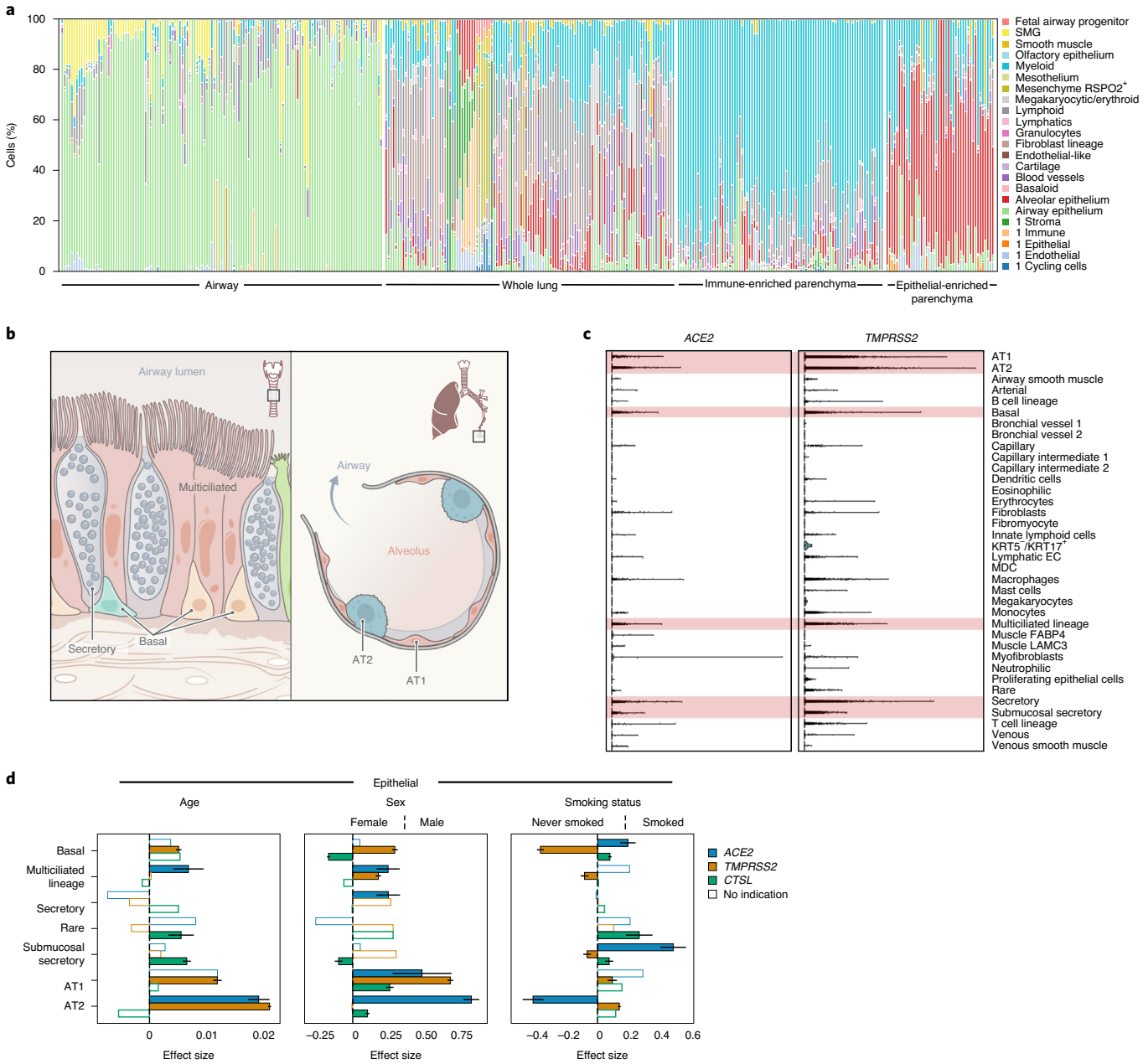


Fig. 3 | *ACE2*, *TMPRSS2* and *CTSL* expression increases with age and in men, and shows cell-type-specific associations with smoking. a, Samples in the aggregated lung and airway dataset partitioned to several classes by their cell composition. Percentage of cells by level 2 cell annotations (annotations with a preceding '1' indicate coarse annotations of cells that had no annotation at level 2) across samples. The 377 samples were ordered by sample composition clusters (Methods). **b**, Schematic of key lung and airway epithelial cell types. **c**, Distribution of normalized *ACE2* and *TMPRSS2* expression across level 3 lung cell types in 1,031,254 cells from 228 donors. Red shading indicates the main cell types that expressed both *ACE2* and *TMPRSS2*. **d**, Age, sex and smoking status associations with expression of *ACE2* (blue), *TMPRSS2* (orange) and *CTSL* (green) in level 3 epithelial cells. The effect size of the association is given as a log fold change (sex and smoking status) or the slope of log expression per year with age. As the age effect size is given per year, it is not directly comparable to the sex and smoking status effect sizes. Positive effect sizes indicate increases with age, in males, and in smokers. Colored bars: associations with an FDR-corrected *P* value < 0.05 (one-sided Wald test on regression model coefficients), consistent effect direction in pseudo-bulk analysis, and consistent results using the model with interaction terms (Methods). White bars: associations that did not pass all of the three above-mentioned evaluation criteria. Error bars: standard errors around coefficient estimates. Error bars are only shown for colored bars (indications or robust trends). Number of cells and donors, respectively, for each cell type: basal: 155,877 and 105; multiciliated lineage: 37,530 and 157; secretory: 22,306 and 140; rare: 2,676 and 71; submucosal secretory: 33,661 and 45; AT1: 29,973 and 101; AT2: 155,512 and 104. EC: endothelial cell; MDC: monocyte-derived cell.

proteins (AT2 cells: *SFTPA* and *SFTPA2*). Cell programs from multiple tissues (Fig. 4c,d) included genes related to TNF signaling, raising the possibility that anti-TNF therapy may impact the expression of *ACE2* and/or *TMPRSS2*. Some of the genes encode proteins that are targets of known drugs[49] (for example, in lung secretory cells: *C3*, *HDAC9*, *IL23A*, *PIK3CA*, *RAMP1* and *SLC7A11*), and other gene products have been shown to interact with SARS-CoV-2 proteins[50], for example, GDF15 (ref. [51]), a central regulator of inflammation[52], and yet others may be related to COVID-19 pathological features, including *MUC1* (ref. [53]; in tissue and specific cell programs), *IL6ST*

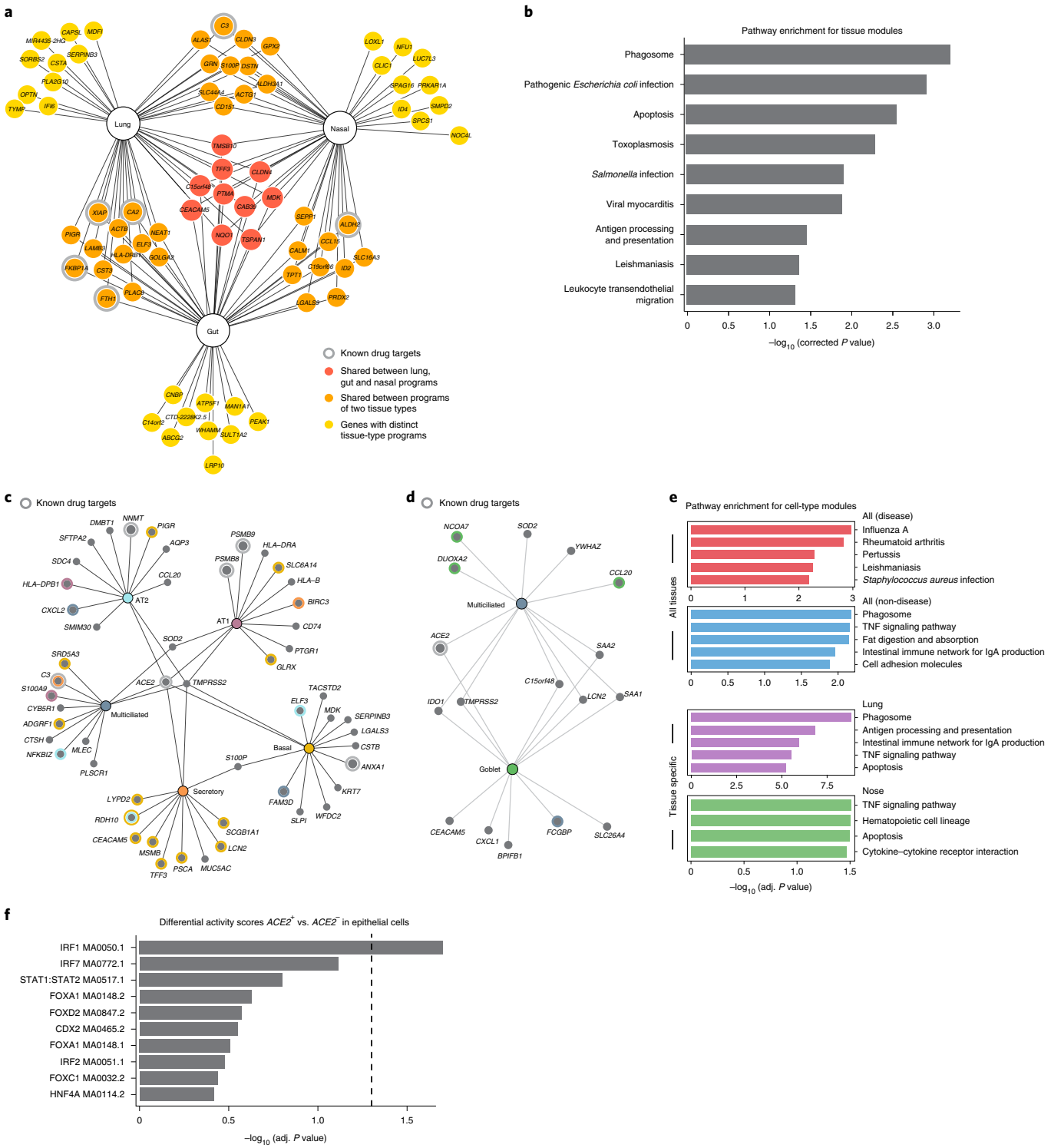


Fig. 4 | Tissue- and cell-type-specific gene modules in *ACE2*$^+$*TMPRSS2*$^+$ cells highlight immune and inflammatory features. a,**b**, Tissue programs of *ACE2*$^+$ *TMPRSS2*$^+$ cells in lung, gut and nasal samples. **a**, Selected tissue program genes. Node: gene; edge: program membership. Genes were selected heuristically for visualization (Methods). **b**, Enrichment was tested using a hypergeometric test exactly as performed by gprofiler in scanpy.queries.enrich ($-\log_{10}$ adj. *P* value) of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway gene sets in the full tissue programs. **c**–**e**, Cell programs of *ACE2*$^+$*TMPRSS2*$^+$ cells. **c**,**d**, Top 12 genes from each cell program recovered for different lung (**c**) or nasal (**d**) epithelial cell types (nodes; colors). Colored concentric circles: overlap with a gene in the top 250 significant genes in other cell types. *ACE2* and *TMPRSS2* were included even if not among the top 12 genes. **e**, Enrichment ($-\log_{10}$ adj. *P* value) of KEGG disease and non-disease pathway gene sets in either highly significant genes across all tissues (top) or in specific tissues (lung and nose; bottom). **f**, Motif activity in immune transcription factors in *ACE2*$^+$ cells. Significance ($-\log_{10}$ adj. *P* value) of the top ten differential 'motif activity scores' (Methods) between epithelial *ACE2*$^+$ cells or *ACE2*$^-$ cells. Epithelial cells are: AT1, AT2, secretory, ciliated, ionocytes and neuroendocrine cells (highlighted in the gray shaded area in Supplementary Fig. 1a). $n=2$ locations: primary carina and lung lobes; $n=3$ samples per location; $n=1$ patient. Motifs were extracted from the JASPAR2020 database, and the motif code is shown in each row. Dashed line: threshold for significance (adj. *P* value of 0.05). *P* values were calculated by logistic regression and likelihood ratio test, adjusted through Bonferroni correction (Methods).

(lung tissue and gut enterocyte programs) and *IL6* (AT2 program; Supplementary Fig. 12d). Other cell types, such as heart pericytes, were enriched for cells coexpressing *ACE2* with *IL6R* or *IL6ST* (Supplementary Fig. 13). The immune-like programs of *ACE2*$^+$ epithelial cells were also reflected in the regulatory features of the *ACE2* locus by scATAC-seq (Fig. 4f). Cell–cell interaction analysis[54] (Methods) predicted interactions (Supplementary Table 11) between AT2 cells (overall or *ACE2*$^+$*TMPRSS2*$^+$) and myeloid cells through oncostatin, complement, interleukin (IL)-1 receptor and colony-stimulating factor signaling.

Conserved expression patterns in mouse models. Preclinical studies of SARS-CoV-2 infection and treatment require model systems that approximate human physiology. Transgenic mouse models that express human ACE2 (hereafter, hACE2) have been identified as a valuable resource to evaluate diverse therapeutics for COVID-19 (ref. [55]). We thus asked whether expression patterns of SARS-CoV-2 entry factors were similar in human and mouse model cell types of interest.

Ace2$^+$*Tmprss2*$^+$ and *Ace2*$^+$*Ctsl*$^+$ double-positive cells were present primarily in club and multiciliated cells in the airway epithelia of healthy mice[56] (Fig. 5a), consistent with human airways (Extended Data Fig. 3a), and increased from 2 to 4 months of age(Fig. 5a,b). Moreover, the expression patterns observed in scRNA-seq data of whole lungs from mice exposed daily to cigarette smoke for 2 months (Fig. 5c–k and Methods) are consistent with our observations in human airway epithelial cells (Fig. 3d and Extended Data Fig. 9a). Upon smoke exposure, there was a significant increase in the number *Ace2*$^+$ cells and *Ace2* expression in airway secretory cell numbers, but not AT2 cells (Fig. 5f–i). There was also agreement in expression patterns between the human placenta and mouse placenta development (Figs. 1c,d and 5l and Supplementary Fig. 14).

Discussion

To the best of our knowledge, this study represents the first single-cell meta-analysis. Our meta-analysis provided the required power to uncover age, sex and smoking associations at single-cell resolution. The contrasting smoking associations of *ACE2* across epithelial cell types show the importance of single-cell resolution, as downregulation in AT2 cells would have been otherwise masked by increases in airway epithelial signal in bulk RNA-seq[57]. Although we have aggregated over 200 donors in our dataset, effects such as race, ethnicity, genetic ancestry, cumulative smoking or healthy tissue with a distal disease site may still confound the associations we have obtained.

Our models included tested covariates, technical covariates and interaction terms, allowing us to uncover complex associations (for example, sex and smoking associations are typically stronger for younger individuals; Supplementary Fig. 5). Modeling the smoking status of a donor was important to reduce background variation and account for the unbalanced distribution of covariates. Fitting this model required aggregating many datasets, harmonized by a consistent cell-type annotation. However, the annotation remains coarse in some cases, where cell labels still aggregate over considerable diversity, and can be further refined in the future. As the HCA grows and further datasets become available, our model could be extended to allow nonlinear associations with the tested covariates. Such associations may uncover, for example, distinct effects in the particularly affected geriatric population. While there is a trend of an increased proportion of *ACE2*$^+$*TMPRSS2*$^+$ cells with age (Extended Data Fig. 3b,c), this cannot be modeled reliably given the compositional diversity (Fig. 3a and Supplementary Fig. 15), potential confounders and limited sample sizes. Further metadata can help to address this.

Our findings in human and mouse models are consistent with respect to smoking and age associations. In line with our human data, we find an increase in *Ace2* expression in maturing mice (2–4 months). Others have reported lower expression of entry factors in aged mice (24 months), showing potential limitations of mice as a model system[58].

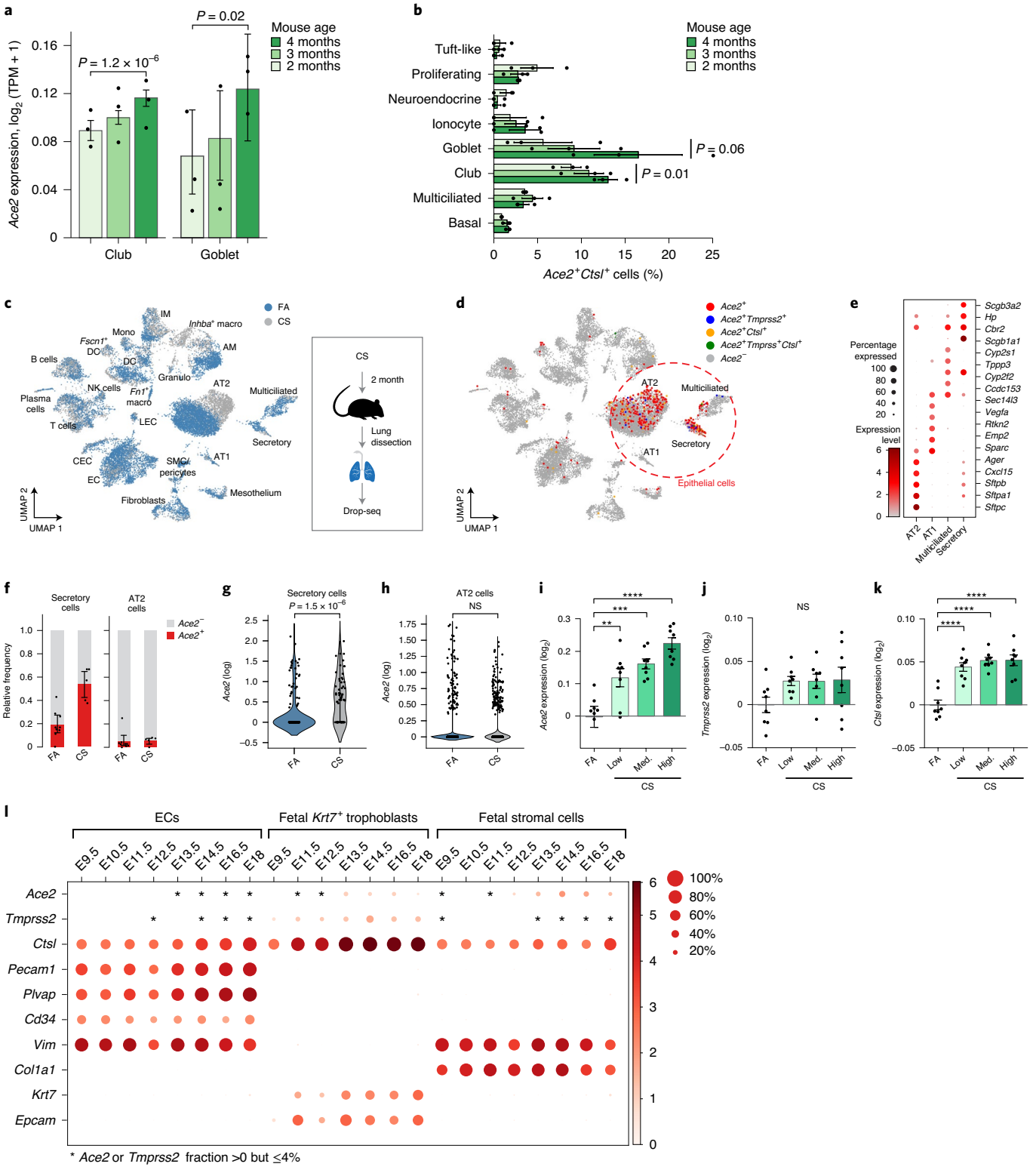
Our comprehensive cross-tissue analysis expands on our[13,14,16,59] and others[60–62] earlier efforts, identifying cell subsets across tissues that may be implicated in transmission or pathogenesis. For example, double-positive cells in the SMGs may be a reservoir for viruses that escape from expulsion associated with severe cough in the airway luminal surface. Another intriguing hypothesis is that neurological symptoms[63–65] and Guillain–Barré syndrome[66] may arise as an autoimmune response to myelin antigens expressed by infected *ACE2*$^+$*TMPRSS2*$^+$ and *ACE2*$^+$ cells that express myelin-producing genes (Supplementary Fig. 2 and Supplementary Table 7).

ACE2 and *TMPRSS2* expression in lung, nasal and gut epithelial cells is associated with programs involving key immunological genes and genes related to viral infection. Expression of *IL6*, *IL6R* and *IL6ST* in lung epithelial cells raises the hypothesis that infection may trigger uncontrolled cytokine expression, as IL-6 levels were reported to increase with COVID-19 severity[67]. The prediction of TNF, complement and IL-1 pathways may suggest a benefit for therapies that target these axes. The accessibility of binding sites for the transcription factors STAT and IRF in scATAC-seq data is consistent with interferon regulation of *ACE2* expression in epithelial cells[14] and with high activity of STAT1, STAT2, IRF1, IRF2,

Fig. 5 | *Ace2*, *Tmprss2* and *Ctsl* expression in mouse in similar cell types, and follows similar patterns with age and smoking. a, Gradual increase in *Ace2* expression by airway epithelial cell type with age. Mean expression of *Ace2* in different airway epithelial cells of mice of three consecutive ages. Shown are replicate mice (dots; $n=3$ for each age), mean (bar) and error bars (s.e.m.). The effect of mouse age was tested using a two-sided Wald test (*P* values). TPM, transcripts per million. **b**, Increase in proportion of *Ace2*$^+$*Ctsl*$^+$ goblet and club cells with age. Percentage of *Ace2*$^+$*Ctsl*$^+$ cells in different airway epithelial cell types of mice of three consecutive ages. The effect of mouse age was tested using a Wald test (*P* values). **c**–**k**, Increase in *Ace2* expression in secretory cells with smoking. Mice were exposed daily to cigarette smoke (CS) or filtered air (FA) as control for 2 months after which cells from whole-lung suspensions were analyzed by scRNA-seq (Drop-seq). AM, alveolar macrophages; IM, interstitial macrophages; DC, dendritic cells; LEC, lymphatic endothelial cells; CEC, capillary endothelial cells; EC, endothelial cells; Mono, monocytes. **c**,**d**, Uniform manifold approximation and projection analysis of scRNA-seq profiles (dots) colored by experimental group (**c**) or by *Ace2*$^+$ cells and indicated double-positive cells (**d**). AT1 and AT2 cells and airway epithelial secretory and ciliated cells are marked by the red dashed line. Macro, macrophage; mono, monocyte. **e**, Marker genes of AT1, AT2, multicilitated and secretory cell clusters. **f**, The relative frequency of *Ace2*$^+$ cells is increased by smoking in airway secretory cells but not AT2 cells. Relative proportion of *Ace2*$^+$ and *Ace2*$^-$ cells in smoke-exposed and control mice of different cell types (FA: $n=9$ mice; CS: $n=5$ mice; error bars represent 95% confidence intervals). **g**,**h**, Expression of *Ace2* was increased in airway secretory cells (FA: 187 cells; CS: 62 cells), but not in AT2 cells (FA: 3,808; CS: 1,882). Distribution of *Ace2* expression in secretory (**g**) and AT2 (**h**) cells from control and smoke-exposed mice (*P* value derived from a Wilcoxon rank-sum test; NS, not significant). **i**–**k**, Reanalysis of published bulk mRNA-seq[69] of lungs exposed to different daily doses of CS show increased expression of *Ace2* (**i**), *Tmprss2* (**j**) and *Ctsl* (**k**) after 5 months of chronic exposure; $n=8$ mice per condition. Bars show the mean, and error bars show the standard error ($^{**}P=0.0046$, $^{***}P=0.0002$ and $^{****}P<0.0001$; one-way ANOVA with Dunnett's multiple comparisons test, compared to FA group.) **l**, Expression in placenta. Mean expression (color) and proportion of expressing cells (dot size) of *Ace2*, *Tmprss2* and *Ctsl* along with marker genes (Supplementary Fig. 14) in single- and double-positive cells from embryonic day (E) 9.5 to E18 of mouse placenta development.

IRF5, IRF7, IRF8 and IRF9 in macrophage states, which increased in patients with severe COVID-19 (ref. [68]). Future lines of inquiry could include investigating the impact of lysosomal genes in lung secretory and multiciliated cells on viral infection and of *RIPK3* expression in airway cells on necroptosis.

Finally, the expression of other potential accessory proteases may help pursue therapeutic hypotheses related to disruption of viral processing via protease inhibition. *FURIN*, *PCSK5* and *PCSK7* are more broadly expressed than *TMPRSS2* across lung cell types (Fig. 2d) and across tissues (Extended Data Fig. 6i). Viral proteins may physically interact with PCSK6 (ref. [50]), which is significantly coexpressed with *ACE2* in AT2 cells (Fig. 2b and Extended Data Fig. 6b). Because PCSK proteases are localized in different membrane compartments[27], they might process SARS-CoV-2 S proteins at different viral stages. Altogether, this could provide SARS-CoV-2 with immense flexibility in entry and egress.



Our meta-analysis provides a detailed molecular and cellular map to aid in our understanding of SARS-CoV-2 transmission, pathogenesis and clinical associations. We herein demonstrated how this can be done despite restrictions on data sharing. As the HCA progresses, we envision such meta-analyses in the context of other diseases, for example, by combining large healthy reference atlases with both epidemiological and genetic risk factors. In parallel, as new atlases are generated from COVID-19 tissues and models, their integration will further advance our understanding of this disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-01227-z.

Received: 16 April 2020; Accepted: 23 December 2020;
Published online: 2 March 2021

References

1. Wang, D. et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020).
2. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
3. Chen, N. et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395**, 507–513 (2020).
4. Wang, W. et al. Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA* https://doi.org/10.1001/jama.2020.3786 (2020).
5. Jia, H. P. et al. ACE2 receptor expression and severe acute respiratory syndrome coronavirus infection depend on differentiation of human airway epithelia. *J. Virol.* **79**, 14614–14621 (2005).
6. Hou, Y. J. et al. SARS-CoV-2 reverse genetics reveals a variable infection gradient in the respiratory tract. *Cell* **182**, 429–446 (2020).
7. McCray, P. B. Jr et al. Lethal infection of K18-*hACE2* mice infected with severe acute respiratory syndrome coronavirus. *J. Virol.* **81**, 813–821 (2007).
8. Walls, A. C. et al. Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292 (2020).
9. Perez-Saez, J. et al. Serology-informed estimates of SARS-CoV-2 infection fatality risk in Geneva, Switzerland. *Lancet Infect. Dis*. https://doi.org/10.1016/S1473-3099(20)30584-3 (2020).
10. Zhou, F. et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
11. Ludvigsson, J. F. Systematic review of COVID-19 in children shows milder cases and a better prognosis than adults. *Acta Paediatr.* **109**, 1088–1095 (2020).
12. Guo, F. R. Smoking links to the severity of COVID-19: an update of a meta-analysis. *J. Med. Virol.* **92**, 2304–2305 (2020).
13. Sungnak, W. et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* **26**, 681–687 (2020).
14. Ziegler, C. G. K. et al. SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. *Cell* **181**, 1016–1035 (2020).
15. Qi, F., Qian, S., Zhang, S. & Zhang, Z. Single-cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. *Biochem. Biophys. Res. Commun.* **526**, 135–140 (2020).
16. Lukassen, S. et al. SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells. *EMBO J*. https://doi.org/10.15252/embj.20105114 (2020).
17. Zhang, H. et al. Specific ACE2 expression in small intestinal enterocytes may cause gastrointestinal symptoms and injury after 2019-nCoV infection. *Int. J. Infect. Dis.* **96**, 19–24 (2020).
18. Ordovas-Montanes, J. et al. Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature* **560**, 649–654 (2018).
19. Sos, B. C. et al. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing assay. *Genome Biol.* **17**, 20 (2016).
20. Emery, B. et al. Myelin gene regulatory factor is a critical transcriptional regulator required for CNS myelination. *Cell* **138**, 172–185 (2009).
21. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
22. Suryawanshi, H. et al. A single-cell survey of the human first-trimester placenta and decidua. *Sci. Adv.* **4**, eaau4788 (2018).
23. Tsang, J. C. H. et al. Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. *Proc. Natl Acad. Sci. USA* **114**, E7786–E7795 (2017).
24. Pérez-Silva, J. G., Español, Y., Velasco, G. & Quesada, V. The Degradome database: expanding roles of mammalian proteases in life and disease. *Nucleic Acids Res.* **44**, D351–D355 (2016).
25. Coutard, B. et al. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* **176**, 104742 (2020).
26. Millet, J. K. & Whittaker, G. R. Physiological and molecular triggers for SARS-CoV membrane fusion and entry into host cells. *Virology* **517**, 3–8 (2018).
27. Seidah, N. G. & Prat, A. The biology and therapeutic targeting of the proprotein convertases. *Nat. Rev. Drug Discov.* **11**, 367–383 (2012).
28. Cai, H. Sex difference and smoking predisposition in patients with COVID-19. *Lancet Respir. Med.* **8**, e20 (2020).
29. Goldfarbmuren, K. C. et al. Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nat. Commun.* **11**, 2485 (2020).
30. Duclos, G. E. et al. Characterizing smoking-induced transcriptional heterogeneity in the human bronchial epithelium at single-cell resolution. *Sci. Adv.* **5**, eaaw3413 (2019).
31. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
32. Reyfman, P. A. et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517–1536 (2019).
33. Madissoon, E. et al. scRNA-seq assessment of the human lung, spleen and esophagus tissue stability after cold preservation. *Genome Biol.* **21**, 1 (2019).
34. Miller, A. J. et al. In vitro and in vivo development of the human airway at single-cell resolution. *Dev. Cell* **53**, 117–128 (2020).
35. Adams, T. S. et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* **6**, eaba1983 (2020).
36. Habermann, A. C. et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.* **6**, eaba1972 (2020).
37. Deprez, M. et al. A single-cell atlas of the human healthy airways. *Am. J. Respir. Crit. Care Med*. https://doi.org/10.1164/rccm.201911-2199OC (2020).
38. Morse, C. et al. Proliferating *SPP1/MERTK*-expressing macrophages in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **54**, 1802441 (2019).
39. Travaglini, K. J., Nabhan, A. N., Penland, L. & Sinha, R. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
40. Mayr, C. H. et al. Integrated single-cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids. *SSRN* https://doi.org/10.2139/ssrn.3538700 (2020).
41. Beane, J. E. et al. Molecular subtyping reveals immune alterations associated with progression of bronchial premalignant lesions. *Nat. Commun.* **10**, 1856 (2019).
42. Chan, C.-M. et al. Carcinoembryonic antigen-related cell adhesion molecule 5 is an important surface attachment factor that facilitates entry of middle east respiratory syndrome coronavirus. *J.Virol.* **90**, 9114–9127 (2016).
43. Wahl, S. M. et al. Secretory leukocyte protease inhibitor in mucosal fluids inhibits HIV-1. *Oral Dis.* **3**, S64–S69 (1997).
44. Turula, H. & Wobus, C. The role of the polymeric immunoglobulin receptor and secretory immunoglobulins during mucosal infection and immunity. *Viruses* **10**, 237 (2018).
45. Burkhardt, A. M. et al. CXCL17 is a mucosal chemokine elevated in idiopathic pulmonary fibrosis that exhibits broad antimicrobial activity. *J. Immunol.* **188**, 6399–6406 (2012).
46. Debbabi, H. et al. Primary type II alveolar epithelial cells present microbial antigens to antigen-specific $CD4^+$ T cells. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **289**, L274–L279 (2005).
47. Yue, Y. et al. SARS-Coronavirus open reading frame-3a drives multimodal necrotic cell death. *Cell Death Dis.* **9**, 904 (2018).
48. Burkard, C. et al. Coronavirus cell entry occurs through the endo-/lysosomal pathway in a proteolysis-dependent manner. *PLoS Pathog.* **10**, e1004502 (2014).
49. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
50. Gordon, D. E. et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
51. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
52. Luan, H. H. et al. GDF15 is an inflammation-induced central mediator of tissue tolerance. *Cell* **178**, 1231–1244 (2019).

53. Dhar, P. & McAuley, J. The role of the cell surface mucin MUC1 as a barrier to infection and regulator of inflammation. *Front. Cell. Infect. Microbiol.* **9**, 117 (2019).
54. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
55. Bao, L. et al. The pathogenicity of SARS-CoV-2 in hACE2 transgenic mice. *Nature* **583**, 830–833 (2020).
56. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
57. Smith, J. C. et al. Cigarette smoke exposure and inflammatory signaling increase the expression of the SARS-CoV-2 receptor ACE2 in the respiratory tract. *Devel. Cell* **53**, 514–529.e3 (2020).
58. Booeshaghi, A. S. & Pachter, L. Decrease in ACE2 mRNA expression in aged mouse lung. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.02.021451 (2020).
59. Tucker Nathan, R. et al. Myocyte-specific upregulation of ACE2 in cardiovascular disease. *Circulation* **142**, 708–710 (2020).
60. Hamming, I. et al. Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J. Pathol.* **203**, 631–637 (2004).
61. Zhao, Y. et al. Single-cell RNA expression profiling of ACE2, the receptor of SARS-CoV-2. *Am. J. Respir. Crit. Care Med.* **202**, 756–759 (2020).
62. Venkatakrishnan, A. J. et al. Knowledge synthesis of 100 million biomedical documents augments the deep expression profiling of coronavirus receptors. *eLife* **9**, e58040 (2020).
63. Mao, L. et al. Neurological manifestations of hospitalized patients with COVID-19 in Wuhan, China: a retrospective case series study. *JAMA Neurol.* **77**, 683–690 (2020).
64. Poyiadji, N. et al. COVID-19-associated acute hemorrhagic necrotizing encephalopathy: CT and MRI features. *Radiology* **296**, E119–E120 (2020).
65. Helms, J., Kremer, S. & Meziani, F. More on neurologic features in severe SARS-CoV-2 infection. *N. Engl. J. Med.* **382**, e110 (2020).
66. Toscano, G. et al. Guillain–Barré syndrome associated with SARS-CoV-2. *N. Engl. J. Med.* **382**, 2574–2576 (2020).
67. Del Valle, D. M. et al. An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nat. Med.* https://doi.org/10.1038/s41591-020-1051-9 (2020).
68. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
69. McInnes, L. et al. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

Christoph Muus[1,2,193] ✉, Malte D. Luecken[3,193] ✉, Gökcen Eraslan[1,193], Lisa Sikkema[3,193], Avinash Waghray[4,5,6,193], Graham Heimberg[1,193], Yoshihiko Kobayashi[7,193], Eeshit Dhaval Vaishnav[1,8,193], Ayshwarya Subramanian[1,193], Christopher Smillie[1,193], Karthik A. Jagadeesh[1,193], Elizabeth Thu Duong[9,193], Evgenij Fiskin[1,193], Elena Torlai Triglia[1,193], Meshal Ansari[10,11,193], Peiwen Cai[12,193], Brian Lin[5,6,13,193], Justin Buchanan[14,15,193], Sijia Chen[16,193], Jian Shu[17,18,193], Adam L. Haber[1,19,193], Hattie Chung[1,193], Daniel T. Montoro[1,193], Taylor Adams[20], Hananeh Aliee[11], Samuel J. Allon[17,21,22], Zaneta Andrusivova[23], Ilias Angelidis[10], Orr Ashenberg[1], Kevin Bassler[24], Christophe Bécavin[25], Inbal Benhar[1], Joseph Bergenstråhle[23], Ludvig Bergenstråhle[23], Liam Bolt[26], Emelie Braun[27], Linh T. Bui[28], Steven Callori[29,30], Mark Chaffin[31], Evgeny Chichelnitskiy[32,33], Joshua Chiou[34], Thomas M. Conlon[10], Michael S. Cuoco[1], Anna S. E. Cuomo[35], Marie Deprez[25], Grant Duclos[36], Denise Fine[37], David S. Fischer[38,39], Shila Ghazanfar[40], Astrid Gillich[41], Bruno Giotti[42], Joshua Gould[1], Minzhe Guo[43], Austin J. Gutierrez[28], Arun C. Habermann[44], Tyler Harvey[1], Peng He[26], Xiaomeng Hou[45,46], Lijuan Hu[27], Yan Hu[47], Alok Jaiswal[1], Lu Ji[48], Peiyong Jiang[48], Theodoros S. Kapellos[49], Christin S. Kuo[50], Ludvig Larsson[51], Michael A. Leney-Greene[1], Kyungtae Lim[52], Monika Litviňuková[53,54], Leif S. Ludwig[1,55], Soeren Lukassen[56,57], Wendy Luo[1], Henrike Maatz[54], Elo Madissoon[58,59], Lira Mamanova[26], Kasidet Manakongtreecheep[17,60,61], Sylvie Leroy[62,63], Christoph H. Mayr[64], Ian M. Mbano[65,66], Alexi M. McAdams[67], Ahmad N. Nabhan[41], Sarah K. Nyquist[17,22,68], Lolita Penland[41], Olivier B. Poirion[45,46], Sergio Poli[20], CanCan Qi[69,70], Rachel Queen[71], Daniel Reichart[72,73], Ivan Rosas[20], Jonas C. Schupp[74], Conor V. Shea[75], Xingyi Shi[75,76], Rahul Sinha[77], Rene V. Sit[41], Kamil Slowikowski[17,60,61], Michal Slyper[1], Neal P. Smith[78], Alex Sountoulidis[79], Maximilian Strunz[80], Travis B. Sullivan[81], Dawei Sun[52], Carlos Talavera-López[82], Peng Tan[1], Jessica Tantivit[17,60,61], Kyle J. Travaglini[41], Nathan R. Tucker[31,83], Katherine A. Vernon[17,84], Marc H. Wadsworth[17,22,85], Julia Waldman[1], Xiuting Wang[12], Ke Xu[75], Wenjun Yan[86,87], William Zhao[12], Carly G. K. Ziegler[17,22,88], The NHLBI LungMap Consortium* and The Human Cell Atlas Lung Biological Network* ✉

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [3]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany. [4]Center for Regenerative Medicine, Massachusetts General Hospital, Boston, MA, USA. [5]Departments of Internal Medicine and Pediatrics, Pulmonary and Critical Care Unit, Massachusetts General Hospital, Boston, MA, USA. [6]Harvard Stem Cell Institute, Cambridge, MA, USA. [7]Department of Cell Biology, Duke University Medical School, Durham, NC, USA. [8]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [9]Division of Respiratory Medicine, Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA. [10]Comprehensive Pneumology Center (CPC)/ Institute of Lung Biology and Disease (ILBD), Helmholtz Zentrum München, Member of the German Center for Lung Research (DZL), Munich, Germany.

[11]Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany. [12]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [13]Center for Regenerative Medicine, Massachusetts General Hospital, Boston, MA, USA. [14]Center for Epigenomics, University of California San Diego School of Medicine, La Jolla, CA, USA. [15]Department of Cellular and Molecular Medicine, University of California San Diego School of Medicine, La Jolla, CA, USA. [16]Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [17]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [18]Whitehead Institute for Biomedical Research, Cambridge, MA, USA. [19]Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [20]Pulmonary, Critical Care and Sleep Medicine, Yale University School of Medicine, New Haven, CT, USA. [21]Institute for Medical Engineering and Science & Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. [22]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. [23]SciLifeLab, Department of Gene Technology, KTH Royal Institute of Technology, Stockolm, Sweden. [24]Department for Genomics & Immunoregulation, LIMES-Institute, University of Bonn, Bonn, Germany. [25]Université Côte d'Azur, CNRS, IPMC, Sophia-Antipolis, France. [26]Wellcome Sanger Institute, Hinxton, UK. [27]Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Stockholm, Sweden. [28]Translational Genomics Research Institute, Phoenix, AZ, USA. [29]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [30]Bioinformatics Program, Boston University School of Medicine, Boston, MA, USA. [31]Precision Cardiology Laboratory, The Broad Institute, Cambridge, MA, USA. [32]Institute of Transplant Immunology, Hannover Medical School, MHH, Hannover, Germany. [33]German Center for Infectious Diseases (DZIF), Braunschweig, Germany. [34]Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, CA, USA. [35]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. [36]Boston University School of Medicine, Boston, MA, USA. [37]Boston University Medical Center, Boston, MA, USA. [38]Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany. [39]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. [40]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [41]Department of Biochemistry and Wall Center for Pulmonary Vascular Disease, Stanford, CA, USA. [42]Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [43]Divisions of Pulmonary Biology; Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [44]Division of Allergy, Pulmonary and Critical Care Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [45]Center for Epigenomics, University of California-San Diego School of Medicine, La Jolla, CA, USA. [46]Department of Cellular and Molecular Medicine, University of California-San Diego School of Medicine, La Jolla, CA, USA. [47]Division of Pulmonary Sciences and Critical Care Medicine, School of Medicine, University of Colorado, Aurora, CO, USA. [48]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, SAR, China. [49]Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany. [50]Division of Pulmonary Medicine, Department of Pediatrics, Stanford University, Stanford, CA, USA. [51]SciLifeLab, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden. [52]Gurdon Institute, University of Cambridge, Cambridge, UK. [53]Cellular Genetics Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [54]Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [55]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [56]Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany. [57]Berlin Institute of Health (BIH), Center for Digital Health, Berlin, Germany. [58]European Molecular Biology Laboratory—European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. [59]Wellcome Sanger Institute, Cellular Genetics Programme Wellcome Genome Campus, Hinxton, UK. [60]Center for Cancer Research, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [61]Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, MA, USA. [62]Pulmonology Department, Université Côte d'Azur, CHU Nice, Nice, France. [63]CNRS, Institut de Pharmacologie Moléculaire et Cellulaire, Sophia-Antipolis, France. [64]Helmholtz Zentrum München, Institute of Lung Biology and Disease, Group Systems Medicine of Chronic Lung Disease, Member of the German Center for Lung Research (DZL), Munich, Germany. [65]Africa Health Research Institute, Durban, South Africa. [66]School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of Kwazulu Natal, Durban, South Africa. [67]Department of Ophthalmology, Harvard Medical School and Massachusetts Eye and Ear, Boston, MA, USA. [68]Computational and Systems Biology, CSAIL, Institute for Medical Engineering and Science & Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. [69]Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children's Hospital, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [70]GRIAC Research Institute, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [71]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, International Centre for Life, Newcastle upon Tyne, UK. [72]Department of Genetics, Harvard Medical School, Boston, MA, USA. [73]Department of Cardiology, University Heart & Vascular Center, University of Hamburg, Hamburg, Germany. [74]Section of Pulmonary, Critical Care, and Sleep Medicine, Yale University School of Medicine, New Haven, CT, USA. [75]Division of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [76]Bioinformatics Program, Boston University, Boston, MA, USA. [77]Institute for Stem Cell Biology and Regenerative Medicine, Stanford Medicine, Stanford, CA, USA. [78]Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Boston, MA, USA. [79]Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Stockholm, Sweden. [80]Comprehensive Pneumology Center (CPC) and Institute of Lung Biology and Disease (ILBD), Helmholtz Zentrum München, Member of the German Center for Lung Research (DZL), Munich, Germany. [81]Lahey Hospital & Medical Center, Burlington, MA, USA. [82]Cellular Genetics Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [83]Masonic Medical Research Institute, Utica, NY, USA. [84]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [85]Institute for Medical Engineering and Science, Department of Chemistry & Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [86]Center for Brain Science, Harvard University, Cambridge, MA, USA. [87]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [88]Harvard–MIT Health Sciences and Technology, Institute for Medical Engineering and Science, Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [193]These authors contributed equally: Christoph Muus, Malte D. Luecken, Gökcen Eraslan, Lisa Sikkema, Avinash Waghray, Graham Heimberg, Yoshihiko Kobayashi, Eeshit Dhaval Vaishnav, Ayshwarya Subramanian, Christopher Smillie, Karthik A. Jagadeesh, Elizabeth Thu Duong, Evgenij Fiskin, Elena Torlai Triglia, Meshal Ansari, Peiwen Cai, Brian Lin, Justin Buchanan, Sijia Chen, Jian Shu, Adam L. Haber, Hattie Chung, Daniel T. Montoro. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: muus@broadinstitute.org; malte.luecken@helmholtz-muenchen.de; hca@humancellatlas.org

The NHLBI LungMap Consortium

Gail H. Deutsch[89], Jennifer Dutra[90,91], Kyle J. Gaulton[46], Jeanne Holden-Wiltse[90,91], Heidie L. Huyck[92], Thomas J. Mariani[93,94], Ravi S. Misra[93], Cory Poole[93], Sebastian Preissl[45,46], Gloria S. Pryhuber[93], Lisa Rogers[93], Xin Sun[95,96], Allen Wang[45,46], Jeffrey A. Whitsett[97] and Yan Xu[98]

[89]Department of Pathology, Seattle Children's Hospital, University of Washington, Seattle, WA, USA. [90]University of Rochester Biocomputational Center, Research Data Integration & Analytics Group, University of Rochester Medical Center, Rochester, NY, USA. [91]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA. [92]Division of Neonatology, Department of Pediatrics, University of Rochester Medical Center, Rochester, Rochester, NY, USA. [93]Division of Neonatology, Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA. [94]Program in Pediatric Molecular and Personalized Medicine, Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA. [95]Department of Pediatrics, University of California-San Diego School of Medicine, La Jolla, CA, USA. [96]Department of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. [97]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [98]Divisions of Pulmonary Biology and Biomedical Informatics, Perinatal Institute, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, USA.

The Human Cell Atlas Lung Biological Network

Jehan Alladina[99], Nicholas E. Banovich[28], Pascal Barbry[25], Jennifer E. Beane[75], Roby P. Bhattacharyya[100,101], Katharine E. Black[99], Alvis Brazma[102], Joshua D. Campbell[75], Josalyn L. Cho[103,104], Joseph Collin[105], Christian Conrad[57,106], Kitty de Jong[107], Tushar Desai[108], Diane Z. Ding[75], Oliver Eickelberg[109], Roland Eils[57,106,110], Patrick T. Ellinor[31,111], Alen Faiz[112], Christine S. Falk[32], Michael Farzan[113], Andrew Gellman[114], Gad Getz[17,115,116], Ian A. Glass[117], Anna Greka[118], Muzlifah Haniffa[119,120,121], Lida P. Hariri[122], Mark W. Hennon[107], Peter Horvath[123,124], Norbert Hübner[54,125,126,127], Deborah T. Hung[72,128,129], Heidie L. Huyck[93], William J. Janssen[130], Dejan Juric[131], Naftali Kaminski[20], Melanie Koenigshoff[47,132], Gerard H. Koppelman[133], Mark A. Krasnow[41], Jonathan A. Kropski[44,134,135], Malte Kuhnemund[136], Robert Lafyatis[137], Majlinda Lako[71], Eric S. Lander[8,138,139], Haeock Lee[140], Marc E. Lenburg[75], Charles-Hugo Marquette[141], Ross J. Metzger[142], Sten Linnarsson[27], Gang Liu[75], Yuk Ming Dennis Lo[48], Joakim Lundeberg[51], John C. Marioni[35,143,144], Sarah A. Mazzilli[75], Benjamin D. Medoff[99], Kerstin B. Meyer[119], Zhichao Miao[119], Alexander V. Misharin[145], Martijn C. Nawijn[146], Marko Z. Nikolić[147], Michela Noseda[148,149], Jose Ordovas-Montanes[6,17,150,151], Gavin Y. Oudit[152,153], Dana Pe'er[154], Joseph E. Powell[155,156], Stephen R. Quake[157,158], Jayaraj Rajagopal[4,6], Purushothama Rao Tata[159], Emma L. Rawlins[160], Aviv Regev[1,161], Mary E. Reid[107], Paul A. Reyfman[145], Kimberly M. Rieger-Christ[81], Mauricio Rojas[162], Orit Rozenblatt-Rosen[163], Kourosh Saeb-Parsy[164], Christos Samakovlis[165,166], Joshua R. Sanes[86,87], Herbert B. Schiller[10], Joachim L. Schultze[24,167], Roland F. Schwarz[168], Ayellet V. Segre[17,169,170], Max A. Seibold[171], Christine E. Seidman[72,172,173], Jon G. Seidman[72], Alex K. Shalek[17,174,175], Douglas P. Shepherd[176], Rahul Sinha[76], Jason R. Spence[177,178,179], Avrum Spira[75,180], Xin Sun[96], Erik Sundström[181], Sarah A. Teichmann[53,182], Fabian J. Theis[183], Alexander M. Tsankov[42], Ludovic Vallier[184,185], Maarten van den Berge[186], Tave A. Van Zyl[67], Alexandra-Chloé Villani[17,60,61], Astrid Weins[187], Ramnik J. Xavier[188], Ali Önder Yildirim[10,189], Laure-Emmanuelle Zaragosi[25], Darin Zerti[71,190], Hongbo Zhang[191], Kun Zhang[192] and Xiaohui Zhang[75]

[99]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [100]Infectious Disease and Microbiome Program, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [101]Infectious Diseases Division, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [102]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK. [103]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [104]Center for Immunology and Inflammatory Diseases, Division of Rheumatology, Allergy and Immunology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [105]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, International Centre for Life, Bioscience West Building, Newcastle upon Tyne, UK. [106]Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany. [107]Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA. [108]Department of Medicine and Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA. [109]Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA. [110]Health Data Science Unit, Heidelberg University Hospital and BioQuant, Heidelberg, Germany. [111]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [112]Respiratory Bioinformatics and Molecular Biology, University of Technology Sydney, Sydney, NSW, Australia. [113]Department of Immunology and Microbiology, The Scripps Research Institute, Jupiter, FL, USA. [114]Department of Statistics, Columbia University, New York, NY, USA. [115]Department of Pathology, Harvard Medical School, Boston, MA, USA. [116]Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [117]Department of Pediatrics, Genetic Medicine, University of Washington, Seattle, WA, USA. [118]Brigham and Women's Hospital, Harvard Medical School, and Broad Institute of MIT and Harvard, Cambridge, MA, USA. [119]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [120]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK.

[121]Department of Dermatology and NIHR Newcastle Biomedical Research Centre, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [122]Division of Pulmonary and Critical Care Medicine and Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [123]Synthetic and Systems Biology Unit, Hungarian Academy of Sciences, Biological Research Center (BRC), Szeged, Hungary. [124]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [125]DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, Berlin, Germany. [126]Berlin Institute of Health (BIH), Berlin, Germany. [127]Charité–Universitätsmedizin, Berlin, Germany. [128]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA. [129]Infectious Disease and Microbiome Program and Core Faculty Member, Broad Institute of MIT & Harvard, Cambridge, MA, USA. [130]Division of Pulmonary, Critical Care and Sleep Medicine, National Jewish Health, Division of Pulmonary Medicine and Critical Care Sciences, University of Colorado Denver, Denver, CO, USA. [131]Department of Medicine, Harvard Medical School and Massachusetts General Hospital Cancer Center, Boston, MA, USA. [132]Lung Repair and Regeneration Unit, Helmholtz-Zentrum Munich, Ludwig-Maximilians-University, University Hospital Grosshadern, Member of the German Center of Lung Research (DZL), Munich, Germany. [133]Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children's Hospital, University of Groningen, University Medical Center Groningen (UMCG), Groningen Research Institute for Asthma and COPD, Groningen, the Netherlands. [134]Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN, USA. [135]Department of Veterans Affairs Medical Center, Nashville, TN, USA. [136]Cartana AB, Stockholm, Sweden. [137]Division of Rheumatology, Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA, USA. [138]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [139]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [140]Department of Biomedicine and Health Sciences, The Catholic University of Korea, Seoul, Korea. [141]FHU OncoAge, CNRS, Inserm, IRCAN team 3, Pulmonology Department, Université Côte d'Azur, CHU de Nice, Nice, France. [142]Department of Biochemistry and Wall Center for Pulmonary Vascular Disease, Stanford University, Stanford, CA, USA. [143]Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. [144]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [145]Division of Pulmonary and Critical Care Medicine, Northwestern University, Chicago, IL, USA. [146]Department of Pathology and Medical Biology, University of Groningen, GRIAC Research Institute, University Medical Center Groningen, Groningen, the Netherlands. [147]UCL Respiratory, Division of Medicine, University College London, London, UK. [148]National Heart and Lung Institute, Imperial College London, London, UK. [149]British Heart Foundation Centre for Research Excellence and Centre for Regenerative Medicine, Imperial College London, London, UK. [150]Division of Gastroenterology Boston Children's Hospital, Boston, MA, USA. [151]Program in Immunology, Harvard Medical School, Boston, MA, USA. [152]Division of Cardiology, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada. [153]Mazankowski Alberta Heart Institute, Edmonton, Alberta, Canada. [154]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [155]Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [156]UNSW Cellular Genomics Futures Institute, University of New South Wales, Sydney, New South Wales, Australia. [157]Departments of Bioengineering and Applied Physics, Stanford University, Stanford, CA, USA. [158]Chan Zuckerberg Biohub, San Francisco, CA, USA. [159]Department of Cell Biology, Regeneration Next Initiative, Duke University School of Medicine, Durham, NC, USA. [160]Wellcome Trust/CRUK Gurdon Institute and Department Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK. [161]Department of Biology, Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. [162]Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [163]Klarman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [164]Department of Surgery, University of Cambridge and NIHR Cambridge Biomedical Research Centre, Cambridge, UK. [165]SciLifeLab, Department of Molecular Biosciences, Stockholm University, Stockholm, Sweden. [166]Cardiopulmonary Institute, Justus Liebig University, Giessen, Germany. [167]PRECISE Platform for Single Cell Genomics & Epigenomics, Germany Center for Neurodegenerative Diseases and University of Bonn, Bonn, Germany. [168]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany. [169]Harvard Medical School, Boston, MA, USA. [170]Ocular Genomics Institute, Department of Ophthalmology, Massachusetts Eye and Ear, Boston, MA, USA. [171]Department of Pediatrics; Center for Genes, Environment, and Health; National Jewish Health, Denver, CO, USA. [172]Cardiovascular Division, Brigham & Women's Hospital, Boston, MA, USA. [173]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. [174]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA. [175]Institute for Medical Engineering and Science (IMES), Koch Institute for Integrative Cancer Research, and Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. [176]Center for Biological Physics and Department of Physics, Arizona State University, Tempe, AZ, USA. [177]Department of Internal Medicine, Gastroenterology, University of Michigan Medical School, Ann Arbor, MI, USA. [178]Department of Cell and Developmental Biology, University of Michigan Medical School, Ann Arbor, MI, USA. [179]Department of Biomedical Engineering, University of Michigan College of Engineering, Ann Arbor, MI, USA. [180]Johnson & Johnson Innovation, Cambridge, MA, USA. [181]Division of Neurogeriatrics, Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Stockholm, Sweden. [182]Department of Physics/Cavendish Laboratory, University of Cambridge, Cambridge, UK. [183]Institute of Computational Biology, Helmholtz Zentrum München and Departments of Mathematics and Life Sciences, Technical University Munich, Munich, Germany. [184]Wellcome and MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, Cambridge, UK. [185]Department of Surgery, Cambridge Biomedical Campus, Cambridge, UK. [186]Department of Pulmonary Diseases and Tuberculosis, University of Groningen, GRIAC Research Institute, University Medical Center Groningen, Groningen, the Netherlands. [187]Department of Pathology, Brigham and Women's Hospital, and Harvard Medical School, Boston, MA, USA. [188]Broad Institute, Department of Molecular Biology and Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [189]Koc University Research Center for Translational Medicine (KUTTAM), Istanbul, Turkey. [190]Microscopy Centre and Department of Applied Clinical Sciences and Biotechnology, University of L'Aquila, L'Aquila, Italy. [191]Key Laboratory for Stem Cells and Tissue Engineering, Ministry of Education and Department of Histology and Embryology of Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China. [192]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA.

Methods

Patient samples. Sample collection underwent Institutional Review Board (IRB) review and approval at the institutions where the samples were originally collected. 'Adipose_healthy_manton_unpublished' was collected under IRB no. 2007P002165/1 (ORSP-3877). Tissue samples from breast, esophagus muscularis, esophagus mucosa, heart, lung, prostate, skeletal muscle and skin, referred to as 'tissue_healthy_regev_snRNA-seq_unpublished', were collected under ORSP-3635. Samples referred to as 'eye_sanes_unpublished' were collected under Dana-Farber/Harvard Cancer Center protocol no. 13-416 and Massachusetts Eye and Ear protocol no. 18-034H. Samples referred to as 'kidney_healthy_greka_unpublished' were collected under Massachusetts General Hospital IRB no. 2011P002692. Samples referred to as 'liver_healthy_manton_unpublished' were collected under IRB no. 02-240/ORSP-1702, as well as ORSP-2630 under ORSP-2169. Lung samples from smokers and non-smokers (41 samples from ten patients, 2–6 locations each) with the suffix 'regev/rajagopal_unpublished' were collected under Massachusetts General Hospital IRB no. 2012P001079/ORSP-3900 under ORSP-3490. Healthy and fibrotic lung samples with the suffix 'xavier_snRNA-seq_unpublished' were collected under Massachusetts General Hospital IRB no. 2003P000555 (CG-5242 under ORSP-3490), and Medoff no. 2015P000319 (CG-5145 under ORSP-3490). Pancreatic ductal adenocarcinoma samples were collected under C. Fernandez-del Castillo, 2003P001289 (CG-4692) under ORSP-3490 at Massachusetts General Hospital. Samples in the dataset 'barbry' were derived from a study that was approved by the Comité de Protection des Personnes Sud Est IV (approval no. 17/081), and informed written consent was obtained from all participants involved. All experiments were performed during 8 months, in accordance with relevant guidelines and French and European regulations. No deviations were made from our approved protocol named 3Asc (An Atlas of Airways at a single-cell level: NCT03437122). Lungs with chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis in the 'kaminski' dataset were obtained from patients undergoing transplant, while healthy lungs were obtained from rejected donor lung organs that underwent lung transplantation at the Brigham and Women's Hospital or donor organs provided by the National Disease Research Interchange. Patient tissues relating to the dataset 'krasnow' were obtained under a protocol approved by Stanford University's Human Subjects Research Compliance Office (IRB no. 15166), and informed consent was obtained from each patient before surgery. The study protocol was approved by the Partners Healthcare IRB (protocol no. 2011P002419). Samples in the dataset 'kropski_banovich' were collected under Vanderbilt IRB nos. 060165 and 171657, and Western IRB no. 20181836 (ethics approval no. 2018/769-31). 'Meyer_b' samples were collected by the Cambridge Biorepository for Translational Medicine, under research ethics committee (REC) approval no. 15/EE/0152. Samples in the dataset 'linnarsson' are covered by 2018/769-31, approved by the Swedish Ethical Review Authority. Samples in the 'misharin' dataset were collected under STU00056197, STU00201137 and STU00202458 and approved by the Northwestern University IRB. Samples in the 'rawlins' dataset were obtained from terminations of pregnancy from Cambridge University Hospitals NHS Foundation Trust under permission from NHS Research Ethical Committee (96/085) and the Joint MRC/Wellcome Trust Human Developmental Biology Resource (grant no. R/R006237/1; www.hdbr.org; REC approval nos. 18/LO/0822 and 18/NE/0290). The studies relating to datasets 'schultze' and 'schultze_falk' were approved by the ethics committees of the University of Bonn and University Hospital Bonn (local ethics vote no. 076/16) and the Medizinische Hochschule Hannover (local ethics vote no. 7414/2017). Fifteen human tracheal airway epithelia in the 'schultze' dataset were isolated from de-identified donors whose lungs were not suitable for transplantation. Lung specimens were obtained from the International Institute for the Advancement of Medicine and the Donor Alliance of Colorado. The National Jewish Health IRB approved the research under protocol nos. HS-3209 and HS-2240. Samples in the 'xu/whitsett' dataset were provided through the federal United Network of Organ Sharing via the National Disease Research Interchange and International Institute for Advancement of Medicine and entered into the National Heart, Lung and Blood Institute (NHLBI) LungMAP Biorepository for Investigations of Diseases of the Lung at the University of Rochester Medical Center, overseen by the IRB as RSRB00047606 (Supplementary Tables 1 and 2).

Integrated analysis of published datasets. Publicly available (Supplementary Table 1) single-cell RNA-seq datasets were downloaded from the Gene Expression Omnibus (GEO). We searched the GEO for datasets that met the following criteria: (1) provided unnormalized count data; (2) were generated using the 10x Genomics Chromium platform; and (3) profiled human samples. These samples spanned a wide range of tissues, including primary tissues, cultured cell lines and chemically or genetically perturbed samples. Applying these filters increases standardization of sample as the vast majority were prepared using the same 10x Chromium instrument and Cell Ranger pipelines.

Datasets comprise one or more samples (individual gene expression matrices), which often correspond to individual experiments or patient samples. In total, this yielded 2,333,199 cells from 469 samples from 64 distinct datasets (Supplementary Table 1). To allow comparison across samples and datasets, we mapped genes using a common dictionary of gene symbols and excluded unrecognized symbols. If a gene from an aggregated master list was not found in a sample, the expression was considered to be zero for every cell in that sample.

After all datasets were collected, we quantified the percentage of cells with >0 UMIs for both *ACE2* and *TMPRSS2* or *ACE2* and *CTSL*. For further analyses with broad cell classes, we only used datasets with more than 15 double-positive cells yielding 252,871 cells from 40 samples.

For integration across datasets, we used two levels of annotations. When possible, every sample was annotated with its tissue of origin based on the available metadata from the GEO. We excluded any sample for which tissue was not specified. For the smaller subset of 252,871 cells, we manually annotated cell clusters with broad cell-type classes using marker genes. These clusters were generated using the harmony-pytorch Python implementation (v0.1.1; https://github.com/lilab-bcb/harmony-pytorch/) of the Harmony scRNA-seq integration method[70] for batch correction and leiden clustering from the Scanpy package (v1.4.5). Clusters without clear markers distinguishing types were excluded from further analysis.

Data were processed using Scanpy. Individual datasets were log normalized (UMIs/10,000 + 1) by column sum and the log1p function ($\ln(10{,}000 \times g_{ij} + 1)$, where a gene's expression profile, g, is the result of the UMI count for each gene, i, for cell j, normalized by the sum of all UMI counts for cell j. This data normalization step was only used for generating the clusters and cell-type annotations.

All other statistical tests for the integrated analysis were performed on the cell's binary classification as double positive or not. For example, for a cell to be considered *ACE2*$^+$, it has >0 *ACE2* transcripts. Double-positive cells have >0 transcripts for both genes of interest. We used Fisher's exact test to determine the statistical dependence between the expression of *ACE2* and *TMPRSS2* or *CTSL* and corrected for multiple testing using the Benjamin–Hochberg method over all tests for each gene pair.

Bronchial brushings from current and former smokers. Bronchial brushings were obtained from high-risk individuals undergoing lung cancer screening at ~1-year intervals by white light and autofluorescence bronchoscopy and computed tomography ($n = 137$ brushings from $n = 50$ patients; GSE109743) and profiled via RNA-seq as described previously[41]. Differential expression analysis of entry factors in former and current smokers was performed via voom-limma[51] using the model:

$$Y_i \sim \text{smoking} + \text{batch} + \text{TIN} + (1|\text{patient}),$$

where smoking denotes the encoded smoking status ('current' or 'former'), batch refers to the experimental batch effect derived from the sequencing run, TIN represents the RNA integrity score, and (1|patient) is a random intercept per patient. Multiple-testing correction was performed via Benjamini–Hochberg to obtain an FDR-corrected P value.

Integrated coexpression analysis of high-resolution cell annotations across tissues. We compiled a compendium of published and unpublished datasets consisting of 2,433,890 cells from 21 tissues and/or organs including adipose, bone marrow, brain, breast, colon, cord blood, ENS, esophagus mucosa, esophagus muscularis, anterior eye, heart, kidney, liver, lung, nasal, olfactory epithelium, pancreas, placenta, prostate, skeletal muscle and skin. After the harmonization of cell-type annotations, *ACE2*-*TMPRSS2* and *ACE2*-*CTSL* expression were assessed using a logistic mixed-effect model:

$$Y_i \sim ACE2 + (1|\text{sample\_id}) \quad (1)$$

where Y_i was the binarized expression level of either *TMPRSS2* or *CTSL*, and covariates were binarized *ACE2* expression in cell i and a sample-level random intercept.

Models were fit separately for each cell type in each dataset. To avoid spurious associations in cell types with very few *ACE2*$^+$ cells and due to very low expression of *ACE2*, we subsampled *ACE2*$^-$ cells to the number of *ACE2*$^+$ cells within each cell type and discarded cell types containing fewer than five cells expressing either *ACE2* or the other gene being tested after the subsampling procedure. The significance of the association between *ACE2* and *TMPRSS2*/*CTSL* was controlled for 10% FDR using the statsmodels Python package (v0.11.1)[71]. Data processing was performed using Scanpy (v1.4.6)[72], and logistic models were fit using lme4 R package (v1.1.21)[73].

Single-cell ATAC-sequencing analysis. *Library generation and sequencing.* We performed single-cell ATAC-seq from primary carina and subpleural parenchyma of one individual ($n = 3$ samples per location). Libraries were generated using the 10x Chromium Controller and the Chromium Single Cell ATAC Library & Gel Bead Kit (1000111) according to the manufacturer's instructions (CG000169-Rev C; CG000168-Rev B) with unpublished modifications relating to cell handling and processing. Briefly, human lung-derived primary cells were processed in 1.5 ml DNA LoBind tubes (Eppendorf), washed in PBS via centrifugation at 400g for 5 min at 4 °C and lysed for 3 min on ice before washing via centrifugation at 500g for 5 min at 4 °C. The supernatant was discarded and lysed cells were diluted in 1× diluted nuclei buffer (10x Genomics) before counting using trypan blue and a

Countess II FL Automated Cell Counter to validate lysis. If large cell clumps were observed, a 40-µm Flowmi cell strainer was used before the tagmentation reaction, followed by Gel Bead-In-Emulsion generation and linear PCR as described in the protocol. After breaking the emulsion, the barcoded tagmented DNA was purified and further amplified to enable sample indexing and enrichment of scATAC-seq libraries. The final libraries were quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

All libraries were sequenced using NextSeq High Output Cartridge kits and a NextSeq 500 sequencer (Illumina), and 10x scATAC-seq libraries were characterized by paired-end sequencing (2×72 cycles).

Initial data processing and quality control. Fastq files were demultiplexed using 10x Genomics Cell Ranger ATAC mkfastq (v1.1.0). We obtained peak-barcode matrices by aligning reads to GRCh38 (CR v1.2.0 pre-built reference) using Cell Ranger ATAC count. Peak-barcode matrices from six channels were normalized per sequencing depth and pooled using the Cell Ranger ATAC command 'aggr'.

The aggregated, depth-normalized, filtered dataset was analyzed with Signac (v0.1.6; https://github.com/timoast/signac/), a Seurat[74] extension developed for the analysis of scATAC-seq data. All the analyses in Signac were run with a random number generator seed set as 1234. Cells that appeared as outliers in quality-control metrics (peak_region_fragments ≤ 750 or peak_region_fragments ≥ 20,000 or blacklist_ratio ≥ 0.025 or nucleosome_signal ≥ 10 or TSS.enrichment ≤ 2) were excluded from the analysis.

Normalization and dimensionality reduction. The aggregated dataset was processed with latent semantic indexing[75], that is, datasets were normalized using term frequency–inverse document frequency. Next, singular value decomposition, ran on all binary features, was used to embed cells in low-dimensional space. UMAP[69] was then applied for visualization, using the first 30 dimensions of the singular value decomposition space.

Gene activity matrix and differential motif activity analysis. A gene activity matrix was calculated as the chromatin accessibility associated with each gene locus (extended to include 2 kb upstream of the transcription start site, as described in the vignette 'analyzing PBMC scATAC-seq' (March 13, 2020; https://satijalab.org/signac/articles/pbmc_vignette.html), using as gene annotation the genes.gtf file provided together with Cell Ranger's ATAC GRCh38-1.2.0 reference genome. For the motif analysis, we note that because epithelial cells with an accessible *ACE2* locus tend to have a higher number of fragments in peaks than cells with inaccessible *ACE2* (Supplementary Fig. 1e), consistent also with higher UMIs in scRNA-seq, some of the cells with inaccessible *ACE2* could be false negatives, thus reducing our power.

Clusters were annotated using label transfer from matching scRNA samples or by literature/expert search of marker 'active' (that is, accessible) genes. Differential motif activity analysis was performed using Signac's implementation of ChromVAR[76], with motif position frequency matrices from JASPAR2020 (ref. [77]; http://jaspar.genereg.net/) selecting transcription factor motifs from human (species = 9606), broadly following the vignette 'motif analysis with Signac' (https://satijalab.org/signac/articles/motif_vignette.html). Cells were identified as positive for *ACE2* and/or *TMPRSS2* (that is, with the loci accessible) if at least one fragment was overlapping with the gene locus or within 2 kb upstream. Differential activity scores between epithelial cells positive for *ACE2* (with the above-mentioned definition of 'positive') and nonexpressing *ACE2* was performed with the FindMarkers function of Seurat (v3.1.1), using the test function set to 'LR' (that is, logistic regression) and the number of counts per peak as the latent variable. The function constructs a logistic regression model predicting group membership based on each motif score individually and compares this to a null model with a likelihood ratio test. An adjusted *P* value is the result of Bonferroni correction.

Immunohistochemistry and proximity ligation in situ hybridization. PLISH was performed as described previously[77]. Briefly, frozen human trachea and distal lung sections were fixed with 4% paraformaldehyde for 20 min, treated with protease (20 µg ml^{-1} proteinase K for lung or pepsin for trachea for 9 min) at 37 °C, and dehydrated with ethanol. The sections were incubated with gene-specific oligonucleotides (Supplementary Table 6) in hybridization buffer (1 M sodium trichloroacetate, 50 mM Tris (pH 7.4), 5 mM EDTA and 0.2 mg ml^{-1} heparin) for 2 h at 37 °C. Common bridge and circle probes were added to the section and incubated for 1 h followed by T4 ligase reaction for 2 h. Rolling circle amplification was performed by using phi29 polymerase (30221, Lucigen) for 12 h at 37 °C. Fluorophore-conjugated detection probe was applied and incubated for 30 min at 37 °C. For the combination of PLISH and immunostaining, sections were incubated with primary antibody for HTII-280 (Terrace Biotech, TB-27AHT2-280), pro-SFTPC (Millipore, ab3786) or ACTA2 (Sigma, F3777) for 1 h at room temperature. Sections were incubated with goat anti-mouse IgM secondary antibody (Thermo Fisher Scientific, A21044) or donkey anti-rabbit IgG secondary antibody (Thermo Fisher Scientific, A32795) for 45 min at room temperature, and then sections were mounted in medium containing DAPI. We imaged three representative areas per patient for three patients in total for the images and quantification shown in Fig. 1 and one representative area for a single patient for Extended Data Fig. 7a,c,d,g. Images were captured using an Olympus confocal microscope FV3000 with Olympus FLUOVIEW FV31S-SW (v2.1.1.98) using a ×20 or ×60 objective.

Bulk mRNA sequencing of sorted cell populations from human lung. Human lung tissue was received from New England Donor Services under the Massachusetts General Hospital approved institutional review board protocol. Tissue was used from three individual patients with no history of lung disease or smoking. Primary bronchus and a piece of right lung lobe were manually dissected out. Single cells were dissociated in HBSS media (Sigma, 55021C) containing collagenase (225 units ml^{-1}), dispase (2.5 units ml^{-1}), elastase (2 units ml^{-1}), pronase (1 mg ml^{-1}), DNAse (1 unit ml^{-1}) and Y-27632 (5 µM) inhibitor. Cell suspension was treated with ACK lysis buffer (Thermo Fisher Scientific A1049201) for 2 min on ice to remove red blood cells. Large airway basal cells were isolated from primary bronchus tissue while small airway basal cells and alveolar type 2 (AT2) cells were isolated from lung lobe tissue. Basal cells from both tissue suspensions were isolated using the anti-human CD271 MicroBeads kit (Miltenyl Biotech, 130-099-023) following the manufacturer's protocol. AT2 cells were isolated with anti-HT2-280 antibody (Terrace Biotech, TB-27AHT2-280) and anti-mouse IgM MicroBeads kit (Miltenyl Biotech, 130-047-301) following the manufacturer's protocol. For mRNA-seq sample preparation, total RNA was isolated using Trizol reagent (Thermo Fisher Scientific, 15596026) following the manufacturer's instructions. Quality and quantity of total RNA was assayed using Nanodrop and an Agilent Bioanalyzer. mRNA-seq libraries were prepared using the TrueSeq protocol from Illumina. For mRNA-seq analysis, the raw fastq files were mapped to the human genome using Tophat (with bowtie2) (version 2.1.1). The output files were processed through Cufflinks (version 2.2.1.3)) and Cuffdiff (version 2.2.1.6) to conduct differential gene expression analysis.

Transposome hypersensitive sites sequencing on human pediatric samples. THS-seq was performed as previously reported[19] on human pediatric samples (full gestation, with no known lung disease) collected at day 1 of life, and again at 14 months, 3 years and 9 years ($n = 1$ at each time point).

Integrated analysis for associating *ACE2*, *TMPRSS2 and CTSL* expression with age, sex and smoking status in nasal, airway and lung cells. To assess the association of age, sex, and smoking status with the expression of *ACE2*, *TMPRSS2* and *CTSL*, we aggregated 31 scRNA-seq datasets of healthy human nasal and lung cells, as well as fetal samples containing the expression counts of only the three genes. Aggregation of these datasets was enabled by harmonizing the cell-type labels of individual datasets and dataset concatenation within Scanpy[71] (v1.4.5.1). We harmonized annotations manually on the basis of provided cell-type labels together with data contributors using a preliminary ontology generated on the basis of five published datasets[31–33,36,38] with three levels of annotations. Level 1 has the lowest resolution and distinguishes epithelial from stromal/mesenchymal, endothelial and immune cells. Level 2 breaks up each of the level 1 categories in the coarsest available further observed annotations. Level 3 in turn splits up the observed level 2 annotations where finer annotations were available (Supplementary Table 2; consent to publish was obtained from all contributors). To compare AT2 cells and their possible fetal progenitors, we mapped progenitor cells labeled 'AT2-like' and 'SpC$^+$ progenitors' to the AT2 label. We further harmonized metadata by collapsing the smoking covariate into 'has smoked' and 'has never smoked' and by taking the mean age where only age ranges were given. This resulted in a dataset of 1,320,896 cells and three genes in 377 samples from 228 donors (the cell by three-gene count matrix with annotations is available on the Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP1257)). We divided the data into fetal (136,450 cells, 41 samples and 34 donors), adult nasal (57,548 cells, 20 samples and 18 donors) and adult lung (1,126,898 cells, 316 samples and 187 donors) datasets based on the metadata provided.

To get an overview of sample diversity, we clustered the samples using the proportion of cells in level 2 cell types as features. Clustering was performed using louvain clustering (resolution of 0.3; louvain package v0.6.1) on a *k*-nearest-neighbor graph ($k = 15$) computed on Euclidean distances over the top five principal components of the cell-type proportion data within Scanpy. This produced four clusters. Sample cluster labels were assigned based on cell-type compositions and metadata for anatomical location that were obtained from the published datasets and via input from the data generators.

Within non-fetal datasets, we modeled the association of age, sex and smoking status with gene expression for *ACE2*, *TMPRSS2* and *CTSL* within each cell type using a generalized linear model with the log total counts per cell as offset and Poisson noise as implemented in Statsmodels[71] (v0.11.1) and using a Wald test from Diffxpy (www.github.com/theislab/diffxpy/; v0.7.3, batchglm v0.7.4). Specifically, we fit the model:

$$Y_{ij} \sim \text{age} + \text{sex} + \text{age} : \text{sex} + \text{smoking} + \text{sex} : \text{smoking} + \text{age} : \text{smoking} + \text{dataset}, \quad (2)$$

which models effects of age, sex and smoking while accounting for potential interactions between covariates and the uneven distribution of covariates across the dataset. Here, Y_{ij} denotes the raw count expression of gene i in cell j; age, sex and smoking denote the modeled covariates; and 'age:sex', 'sex:smoking' and 'age:smoking' represent the interaction terms between these covariates. The interaction terms model whether there is a difference in the smoking effect in men and women, and likewise whether the age effect is different for smokers and non-smokers. We included the 'dataset' term to model the technical variation (for example, sampling and processing differences) between the diverse datasets, and the log total count per cell was used as an offset. Here, the total counts were scaled to have a mean of 1 across all cells before the log was taken. Due to the inclusion of interaction terms, the complex interaction model (2) fits the overall effects of age (k_{age}), sex (k_{sex}) and smoking ($k_{smoking}$) as linear functions of the other two covariates, given by the equations:

$$k_{age}(\text{sex}, \text{smoking}) = \beta_{age} + \text{sex}\ \beta_{age:sex} + \text{smoking}\ \beta_{age:smoking},$$

$$k_{sex}(\text{age}, \text{smoking}) = \beta_{sex} + \text{age}\ \beta_{age:sex} + \text{smoking}\ \beta_{sex:smoking},$$

$$k_{smoking}(\text{age}, \text{sex}) = \beta_{smoking} + \text{age}\ \beta_{age:smoking} + \text{sex}\ \beta_{sex:smoking},$$

Here, β_{age} and $\beta_{age:sex}$ represent the model coefficients for age and the interaction of age and sex in model 2, respectively, and age denotes the age covariate. Sex and smoking covariates were converted into a one-hot encoded format such that sex = 0 denoted females and smoking = 0 denoted non-smokers. As linear dependencies on covariates can be summarized by showing two values per covariate, we displayed effect sizes for the overall age, sex and smoking associations by computing k_{age}, k_{sex} and $k_{smoking}$ for sex $\in \{0,1\}$, smoking $\in \{0,1\}$ and age $\in \{31,62\}$ (the first and third quartiles of the age distribution). Standard errors for these effects were computed with the variance–covariance matrix Σ using $SE = \sqrt{C^T \Sigma C}$, where SE is the standard error and C is the vector of covariate values used to compute the respective overall effect (for example, k_{age}). P values were obtained using a Wald test, and correction for multiple testing was performed over all tests on the same cell-type data using the Benjamini–Hochberg method. To fit this model, we pruned the data to contain only datasets that had at least two donors and for which smoking status metadata were provided. This resulted in a dataset of 985,420 cells and 286 samples from 164 donors for adult lung data. Only 15 donors remained for adult nasal data after this filtering, which we deemed too few to obtain robust results. To obtain cell-type-specific associations, the above model was fit within each cell type for all cell types with at least 1,000 cells.

While cells from different donors are not truly independent observations, model 2 treats them as such and thus models cellular and donor variation jointly. As donor variation tends to be larger than single-cell variation, when most cells come from few donors (either there are few donors, or few donors contribute most of the cells), this can lead to inflation of P values. To counteract this effect, we verified that significant associations were consistent when modeling only donor variation via pseudo-bulk analysis (Supplementary Data 1–4). Furthermore, we tested whether effects were dependent on few donors by holding out datasets.

Pseudo-bulk data were generated by computing the mean for each gene expression value and the number of UMIs (nUMI) covariate for cells in the same cell type and donor. After filtering as described above, model 2 was fit to the data (Supplementary Data 1–4). In contrast to the single-cell model, pseudo-bulk analysis underestimates certainty in modeled effects as uncertainty in the pseudo-bulk means are not taken into account when estimating background variance. Thus, we used only effect directions from pseudo-bulk analysis to validate single-cell associations. In further analysis, we regarded only those associations as confirmed by pseudo-bulk analysis, where the FDR-corrected P value in the single-cell model was below 0.05, and the sign of the estimated effect was consistent in both the single-cell and the pseudo-bulk analysis.

We further separated significant associations into robust trends and indications depending on the holdout analysis. A significant association was regarded as a robust trend if the effect direction is consistent when holding out any dataset when fitting the model (without considering the P value). In the case that holding out one dataset caused the maximum-likelihood estimate of the coefficient to be reversed, we denoted this as the effect no longer being present, which characterized the association as an indication. Two dataset holdouts led to indications in our analysis: the largest declined donor transplant dataset ('regev-rajagopal'; Supplementary Table 2; most cells and most samples; indication in *ACE2* multiciliated lineage age and sex associations, and *CTSL* AT1 sex association) and a declined donor tracheal epithelium dataset ('seibold'; Supplementary Table 2; most donors in the smoking analysis; *CTSL* basal smoking association).

At least four values for each covariate are required to describe a single association in model 2 (for example, male nonsmoker, female nonsmoker, male smoker and female smoker for the k_{age} effect). To summarize these effects and present a single association per covariate, we also fit the simplified model:

$$Y_{ij} \sim \text{age} + \text{sex} + \text{smoking} + \text{dataset} \quad (3)$$

As in model 2, the logarithmized, scaled total counts per cell were used as an offset, data were filtered as described, and multiple-testing correction was performed via Benjamini–Hochberg. To increase the robustness of our reported associations, we again performed pseudo-bulk and holdout analysis. Additionally, to still account for covariate interactions, we discarded associations where the complex model 2 and the simplified model 3 results were inconsistent. Here, consistency was defined by two criteria: at least one model 2 indication or robust trend in the same direction as the model 3 effect, and no model 2 indication or robust trend in the opposite direction to the model 3 effect.

As metadata on smoking status were only available for a subset of the data, we also fitted a reduced version of models 2 and 3 without the smoking covariate on a larger dataset to confirm sex and age associations (Supplementary Data 5–8). The nonsmoking model was fit on 1,096,604 cells in 309 samples from 185 donors of adult lung data. Again, log total count (scaled) was used as an offset, pseudo-bulk and holdout analysis was performed, and associations from the simple model were tested for consistency with the complex model.

Normalizing *ACE2*$^+$*TMPRSS2*$^+$ double-positive fractions of human lung samples. Proportions of *ACE2*$^+$*TMPRSS2*$^+$ cells (Extended Data Fig. 3a and Supplementary Fig. 15) were normalized to account for differences in total UMI counts. Normalization was completed per donor, for each cell type by calculating $\frac{X_{i,j}}{N_{i,j}} \times 10,000$, where $X_{i,j}$ is the double-positive fraction of cell type i in donor j, and $N_{i,j}$ represents the median total UMI count of cells of type i in donor j.

Identification of gene programs using feature importance for a random forest trained to classify *ACE2*$^+$*TMPRSS2*$^+$ versus *ACE2*$^-$*TMPRSS2*$^-$ cells. To infer tissue programs, we trained a random forest classifier to discriminate between double-positive and double-negative cells (excluding *ACE2* and *TMPRSS2*; a 75:25 class-balanced test:train split), generalizing across multiple cell types in one tissue, and ranked genes according to their importance scores in the classifier. To infer cell programs, we performed differential expression analysis between double-positive and double-negative cells within each cell subset.

Importantly, these methods do not assume that *ACE2*$^+$*TMPRSS2*$^+$ cells form a distinct subset within each cell type. Rather, our goal is to leverage the variation among single cells within a single type to identify gene programs that are co-regulated with *ACE2* and *TMPRSS2* within each expressing cell subset.

For each of the lung, nasal and gut datasets, we labeled the cells with non-zero counts for both *ACE2* and *TMPRSS2* as double-positive cells, and the cells with zero counts for both *ACE2* and *TMPRSS2* as double-negative cells. Within each tissue, we identified cell types with greater than ten double-positive cells, and for each of these cell types, we selected the genes with increased expression (log fold change > 0) in double-positive cells compared to double-negative cells (to focus on important 'positive' features). We trained a classifier with a 75:25 train:test split to classify the double-positive cells from double-negative cells within each of these cell types using the 'sklearn' (v0.21.3)[78] 'RandomForestClassifier' function with the following parameters: 'n_estimators' set to 100, the 'criterion as gini', and the 'class_weight' parameter set to 'balanced_subsample'. We first trained individual classifiers separately for each of the cell types and pooled genes with positive feature importance values (using the 'feature_importance'[79] field in the trained RandomForestClassifier object) to train a final double-positive cell versus double-negative classifier across each tissue. We used the top 500 genes, as ranked by their feature importance scores, to define the signature for the gene expression program of double-positive cells for the tissue. This procedure was carried out in lung, nasal and gut datasets, yielding tissue-specific signatures for gene expression programs of double-positive cells from each tissue.

For visualization purposes only, we generated network diagrams using the 'networkx' (v2.2) tool with the ForceAtlas2 graph layout algorithm[80]. We scored genes that appeared in signatures for multiple tissues by their aggregated feature importance (using a plotting heuristic method that used the sum of importance ranks for genes in individual tissues and by assigning a large valued rank (10,000) to a gene that did not appear in a particular tissue) and selected the top ten genes that were shared by each pair of tissues or shared by all tissues along with additional genes that included the ones unique to each tissue's signature to plot in the network visualization. The GO terms enriched in the gene expression programs shared by double-positive cells across tissues were found using g:Profiler (v1.0.0)[81] using the 'scanpy.queries.enrich' tool.

This analysis was performed in two ways: on the original data, as well as after accounting for differences in distribution of the nUMI per cell between double-positive cells and double-negative cells. This was performed by binning the nUMI distribution in the double-positive cells for each tissue into 100 bins and then randomly sampling from the nUMI distribution for the double-negative cells in each bin to match the distribution of the double-positive cells in that bin. The nUMI distributions before and after the matching procedure are shown in Supplementary Fig. 11b.

Identification of gene programs enriched in double-negative cells versus double-negative cells using regression. In parallel, we used a regression framework to recover gene modules enriched in double-positive versus

double-negative cells (Fig. 4c,d and Supplementary Fig. 12a,b) in the nasal, lung and gut datasets. We first restricted our analysis to cell subsets derived from at least two donor individuals that each contained a mixture of double-negative and double-positive cells (nawijn nasal: multiciliated; goblet; regev/rajagopal lung: AT1, AT2, basal, multiciliated, and secretory; aggregated lung: AT2, multiciliated and secretory; regev/xavier colon: *BEST4*$^+$ enterocytes, cycling TA (transit amplifying), enterocytes, immature enterocytes 2 and TA-2). For each of these cell subsets, we then used MAST (v1.8.2)[82] to fit the following regression model to every gene with cells as observations:

$$Y_i \sim X + (1|S),$$

where Y_i is the expression level of gene i in cells, measured in units of $\log_2$(transcripts per 10,000 reads (TP10K) + 1), X is the binary coexpression state of each cell (that is, double-positive versus double-negative cells), and S is the donor that each cell was isolated from. To control for donor-specific effects (that is, batch effects), we used a mixed model with a random intercept that varies for each donor. To fit this model, we subsampled cells from double-positive and double-negative groups to ensure that both the donor distribution and the cell complexity (that is, the number of genes per cell) were evenly matched between the two groups, as follows. First, for each subset, we restricted our analysis to donors containing at least two double-negative and two double-positive cells. Using these samples, we partitioned the cells into ten equally sized bins based on cell complexity and subsampled double-negative cells from each bin to match the cell complexity distribution of the double-positive cells. Finally, we fit the mixed model (above), controlling for both donor and cell complexity.

To build gene modules for double-positive cells, we prioritized genes by requiring that they be expressed in at least 10% of double-positive cells, and to have a model coefficient greater than 0 with an FDR-adjusted P value of less than 0.05 (for the combined coefficient in the hurdle model). After this filtering step, genes were ranked by their model coefficient (that is, estimated effect size). The top 12 genes were selected for network visualization within each cell type (Fig. 4c,d and Supplementary Fig. 12a,b). In three cases (gut cycling TA, TA-2 and *BEST4*$^+$ cells), *RP11* antisense genes were flagged and excluded from visualizations. To visualize overlap across each network, we indicated whether each gene was among the top 250 genes from each of the other cell types. Putative drug targets were identified by querying the DrugBank database[49]. Gene-set enrichment analysis was performed using the R package Enrichr (v1.0)[83], selecting the top 25 genes from each cell type for the pan-tissue analysis ('all' category; Fig. 4e) and the top 50 genes from each cell type for the tissue-specific analyses ('nose' and 'lung' categories; Fig. 4e). We note a few limitations that may influence our results, including nonuniform sampling across donors, variation in cell compositions across regions (for example, distal lung versus carina) and additional cellular heterogeneity that the current level of broad subset annotation may not have captured.

Cell–cell interaction analysis. CellphoneDB[54] (v.2.0.0) was run with default parameters on the ten human lung samples of the regev/rajagopal dataset (41 samples from 10 patients, 2–6 locations each), analyzing the cells from each dissected region separately. For each sample (patient/location combination) and for each cell type, we distinguished double-positive cells (*ACE2* > 0 and *TMPRSS2* > 0) from all others. Only interactions highlighted as significant (that is, present in the 'significant means' output $P < 0.05$) from CellphoneDB were considered. AT2 cells and myeloid cells were present in lung lobe samples from all ten patients, whereas samples from five patients contained both *ACE2*$^+$*TMPRSS2*$^+$ double-positive AT2 cells and myeloid cells.

Coexpression patterns of additional proteases and *IL6*/*IL6R*/*IL6ST*. *ACE2*–protease coexpression (Fig. 2 and Extended Data Fig. 5) and *ACE2*–*IL6*/*IL6R*/*IL6ST* coexpression (Supplementary Fig. 13) were tested via the logistic mixed-effects model described above (model 1).

Mouse smoke exposure experiments. For these experiments, 8- to 10-week-old pathogen-free female wild-type C57BL/6 mice were obtained from Charles River and housed in rooms maintained at constant temperature and humidity with a 12-h light cycle. Animals were allowed food and water ad libitum. All animal experiments were approved by the ethics committee for animal welfare of the local government for the administrative region of Upper Bavaria (Regierungspräsidium Oberbayern) and were conducted under strict governmental and international guidelines in accordance with EU Directive 2010/63/EU. The female C57BL/6 mice ($n = 5$) were whole-body exposed to 100% mainstream cigarette smoke at a particle concentration of 500 mg/m^3, generated from 3R4F research cigarettes (filter removed; Tobacco Research Institute, University of Kentucky), for 50 min twice daily, 5 d per week for 2 months to mimic human smoking habits[84]. Control mice ($n = 3$) were exposed to filtered air, but exposed to the same stress as mice exposed to cigarette smoke.

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

Data availability

Availability of published datasets is summarized in Supplementary Tables 1 and 2. Interactive visualization and download of select (as indicated in Supplementary Tables 1 and 2) human gene expression data can be accessed on the Single Cell Portal at http://broad.io/hcacovid19. The scATAC-Seq data is available on Terra and github (see below) and the scTHS-Seq data is available on GEO (GSE154027). Mouse placenta data can be accessed on the Single Cell Portal at https://singlecell.broadinstitute.org/single_cell/study/SCP1292.

Code availability

Data and an interactive analysis examining the coexpression of genes across datasets can be accessed via the open-source data platform Terra at https://app.terra.bio/#workspaces/kco-incubator/COVID-19_cross_tissue_analysis/. All analysis scripts can further be accessed at https://github.com/theislab/Covid_meta_analysis/.

References

70. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
71. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* 57–61 (Austin, 2010).
72. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
73. West, B. T., Welch, K. B. & Galecki, A. T. *Linear Mixed Models: a Practical Guide Using Statistical Software* 2nd edn. (CRC Press, 2014).
74. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
75. Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
76. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
77. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
78. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
79. Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and Regression Trees* (CRC press, 1984).
80. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* **9**, e98679 (2014).
81. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists. *Nucleic Acids Res.* **47**, W191–W198 (2019).
82. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
83. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server. *Nucleic Acids Res.* **44**, W90–W97 (2016).
84. Jia, J. et al. Cholesterol metabolism promotes B cell positioning during immune pathogenesis of chronic obstructive pulmonary disease. *EMBO Mol. Med.* **10**, e8349 (2018).

Acknowledgements

We thank all donors, patients and their families for their contributions to the studies that are part of our integrated analysis. We thank L. Gaffney and A. Hupalowska for help with figure preparation, C. de Boer for critical reading of the manuscript and E. Spiegel from the statistical consulting core facility at the Institute of Computational Biology, Helmholtz Center Munich, for advice on statistical modeling. N.E.B. is supported by the National Institutes of Health (NIH)/NHLBI (R01HL145372) and the Department of Defense (W81XWH1910416). J.C. is supported by grants from the Medical Research Council (MRC; MR/S035826/1) and the European Research Council (ERC; 614620). R.E. and C.C. are supported by the European Commission (ESPACE/HEuropean Union Horizon 2020 Research and Innovation Program, 874710). T.D. is supported by HubMap consortium and Stanford Child Health Research Institute (Woods Family Faculty Scholarship). O.E. is supported by the Chan Zuckerberg Initiative (CZI) Seed Network and the NIH (1R01HL146519). C.S.F. is supported by DFG, SFB 738 project B3 (DFG FA-483/1-1). I.A.G. and the University of Washington Laboratory of Developmental Biology were supported by NIH award no. 5R24HD000836 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. A.G. is supported by a CZI Seed Network grant. P.H. acknowledges support from the LENDULET-BIOMAG grant (2018-342) and the CZI (CZF2019-002448). N.H. acknowledges support from a British Heart Foundation (BHF)/German Centre for Cardiovascular Research (DZHK) grant, ERC Advanced Grant under the Horizon 2020 Program and the Federal Ministry of Education and Research of Germany in the framework of CaRNAtion. W.J.J. received funding from the NIH (R35HL140039 and R01HL130938). N.K. received funding from NIH grants R01HL127349 and U01HL145567 and an unrestricted grant from Three Lakes Foundation. M.K. received

funding from NIH grant R01HL141380. G.H.K. and M.K. received funding from Horizon2020 HCA 'discovAIR' project (no. 874656). M.A.K. received funding from Howard Hughes Medical Institute, CZI and Wall Center for Pulmonary Vascular Disease. J.A.K. received funding from NIH grants R01HL145372 (J.A.K./N.E.B.) and K08HL130595 (J.A.K.) and the Doris Duke Charitable Foundation (J.A.K.). M.L. received funding from the ERC (614620). H.L. acknowledges funding from the National Research Foundation of Korea. S.A.M., J.C., A.S., M.E.L. and J.B. acknowledge support from a Stand Up to Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Cancer Reserach Grant (SU2C-AACR-DT23-17 to S. M. Dubinett and A.S.). Stand Up to Cancer is a division of the Entertainment Industry Foundation. S.A.M., J.C., M.E.L. and J.B. acknowledge funding from Sponsored Research Agreements with Janssen Pharmaceuticals. J.B. and J.C. acknowledge funding from the Department of Defense (W81XWH1410234). S. Leroy acknowledges funding from Horizon 2020 under grant no. 874656 (discovAIR). S. Linnarsson acknowledges funding from the Knut and Alice Wallenberg Foundation (2015.0041 and 2018.0172), the Erling-Persson Family Foundation (Human Developmental Cell Atlas) and the Swedish Foundation for Strategic Research (SB16-0065 and RIF14-0057). J.L. acknowledges funding from Horizon2020 under grant no. 874656 (discovAIR), the Knut and Alice Wallenberg Foundation (2018.0172) and the Erling-Persson Family Foundation (HDCA). B.D.M. is supported by NIH grant R01 HL133153. K.B.M. acknowledges funding from CZI grant 2017-174169 (5022), Wellcome Trust grants 206194/Z/17/Z and 211276/Z/18/Z, MRC grant MR/S035907/1 and Horizon2020 grant no. 874656 (discovAIR). A.V.M. acknowledges funding from NIH grants HL135124, AG049665 and AI135964 and grant number CZF2019-002438 from the CZI Foundation awarded to the HCA Lung Seed Network. M.C.N. acknowledges funding from from grant number CZF2019-002438 from the CZI Foundation awarded to the HCA Lung Seed Network, GSK, Netherlands Lung Foundation project nos. 5.1.14.020 and 4.1.18.226 and Horizon2020 under grant no. 874656 (discovAIR). M.Z.N. acknowledges funding from Rutherford Fund Fellowship allocated by the MRC and the UK Regenerative Medicine Platform (MR/5005579/1); Rosetrees Trust (grant no. M899). M.N. acknowledges funding from a BHF/DZHK grant and the BHF (PG/16/47/32156), CZI RFA CZF2019-002431e for Research Excellence and the Centre for Regenerative Medicine, Imperial College London. J.O.-M. acknowledges funding from the Richard and Susan Smith Family Foundation. G.Y.O. acknowledges support from the Canada Research Chair, the Canadian Institute of Health Research and the Heart and Stroke Foundation. D.P. acknowledges funding from the Alan and Sandra Gerry Metastasis and Tumor Ecosystems Center. S.R.Q. acknowledges funding from the CZI Biohub. J.R. acknowledges funding from LungMAP and CZI Seed Network. P.R.T. acknowledges funding from R01HL146557 from NHLBI/NIH and CZI-HCA Seed projects. E.L.R. acknowledges funding from the MRC (MR/S035907/1 and MR/P009581/1), Wellcome Trust (109146/Z/15/Z), Core support from the Wellcome Trust (203144/Z/16/Z) and Cancer Research UK (C6946/A24843). A.R. and O.R.-R. were supported by the Howard Hughes Medical Institute, the Klarman Cell Observatory, the Manton Foundation and the CZI. P.A.R. acknowledges funding from the NIH (K08HL146943), a Parker B. Francis Fellowship and an ATS Foundation/Boehringer Ingelheim Pharmaceuticals Research Fellowship in idiopathic pulmonary fibrosis. M.R. acknowledges funding from 1 U01 HL14555-01. K.S.-P. acknowledges funding from NIHR Cambridge Biomedical Research Centre. C.S. acknowledges funding from the Swedish research Council, Swedish Cancer Society, CPI and Horizon2020 under grant no. 874656 (discovAIR). H.S. was supported by grant number CZF2019-002438 from the CZI Foundation awarded to the HCA Lung Seed Network, the German Center for Lung Research and Helmholtz Association, and Horizon2020 under grant no. 874656 (discovAIR). Work by J.S. was supported by J.L.S. funded in part by Boehringer Ingelheim, by the German Research Foundation (DFG; EXC2151/1, ImmunoSensation2-the immune sensory system; project nos. 390873048, 329123747 and 347286815) and by the HGF grant sparse2big. C.E.S. was supported by the Howard Hughes Medical Institute and the NIH (NHLBI; 2R01HL080494). J.G.S. was supported by the NIH (NHLBI; 2R01HL080494). A.K.S. was supported by the Beckman Young Investigator Program, a Sloan Fellowship in Chemistry, the NIH (5U24AI118672) and the Bill and Melinda Gates Foundation. D.P.S. was supported by the CZI Seed Network grant. J.R.S. is supported by the NHLBI (R01HL119215), by the NIAID Novel Alternative Model Systems for Enteric Diseases consortium (U19AI116482) and by grant no. CZF2019-002440 from the CZI DAF, an advised fund of Silicon Valley Community Foundation. F.J.T. was supported by grant no. CZF2019-002438 from the CZI Foundation awarded to the HCA Lung Seed Network, the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI (grant no. ZT-I-PF-5-01), Horizon2020 under grant no. 874656 (discovAIR) and the German Center for Lung Research. A.M.T. was supported by CZI Lung Atlas and National Science Foundation award no. IOS-2028295. L.V. was supported by the ERC advanced grant New-Chol, the Cambridge University Hospitals NIHR Biomedical Research Centre and the core support grant from the Wellcome Trust and MRC of the Wellcome–MRC Cambridge Stem Cell Institute. M.V.D.B. was supported by the Ministry of Economic Affairs and Climate Policy by means of the PPP. R.J.X. was supported by DK 043351, DK114784, AI142784 and DK117263. L.E.Z. was supported by the Agence Nationale de la Recherche (UCAJEDI, ANR-15-IDEX-01; SAHARRA, ANR-19-CE14-0027; France Génomique, ANR-10-INBS-09-03), Fondation pour la Recherche Médicale (DEQ20180339158), CZI (Silicon Valley Foundation, 2017-175159-5022) and Conseil Départemental des Alpes Maritimes (2016-294DGADSH-CV and 2019-390DGADSH-CV). D.Z. was supported by the MRC (MR/S035826/1) and ERC (614620). H.Z. is supported by the National Key R&D Program (2019YFA0801703) and the National Natural Science Foundation of China (31871370). This study was supported by NHLBI Molecular Atlas of Lung Development Program Human Tissue Core grants U01HL122700 and HL148861. J.W., G.H.D. and Y.X. acknowledge support from the NIH, U01 HL148856 LungMap Phase II–building a multidimensional map of developing human lung. X.S. and, A.W. acknowledge support from the NIH, 1U01 HL148867-01.

Author contributions

The principal investigators listed in The HCA Lung Biological Network authors provided resources. Sample collection was performed by A.W., B.L., D.T.M., S.L., I.R., J.C.S., M. Slyper, K.A.V. and the HCA Lung Biological Network. C.M., A.W., Y.K., E.T.D., B.L., D.T.M., T.S.A., S.J.A., Z.A., I.A., K.B., J.B., L. Bergenstråhle, L. Bolt, E.B., L.T.B., S.C., E.C., T.M.C., M.S.C., S.G., A.G., X.H., L.H., Y.H., T.S.K., M.L., L.S.L., W.L., H.M., E.M., L.M., K.M., I.M.M., A.M.M., A.N.N., S.K.N., L.P., C.Q., D.R., R.V.S., M. Slyper, N.P.S., M. Strunz, D.S., J.T., K.J.T., M.H.W., J.W., W.Y. and C.G.K.Z. performed experiments. C.M., M.D.L., G.E., L.S., A.W., Y.K., G.H., E.D.V., A.S., C.S.S., K.A.J., E.T.D., E.F., E.T.T., M.A., P.C., B.L., J.B., S.C., J.S., A.L.H., H.C. and D.T.M. analyzed and interpreted data with input from T.S.A., H.A., S.J.A., O.A., C.B., I.B., M.C., J.C., T.M.C., A.S.C., M.D., G.D., D.S.F., S.G., A.G., B.G., J.G., M.G., A.C.H., T.H., P.H., A.J., L.J., P.J., T.S.K., L.L., M.A.L., S.L., H.M., E.M., C.-H.M., I.M.M., A.M.M., L.P., A.N.N., O.B.P., S.P., S.K.N., R.Q., D.R., C.V.S., X.S., R.S., K.S., T.B.S., C.T., P.T., K.J.T., N.R.T., M.H.W., X.W., W.Y., W.Z. and C.G.Z. A.R., C.M., M.D.L., G.E., L.S., A.W., G.H., E.D.V., A.S., C.S.S., K.A.J., E.F., E.T.T., B.L., S.C., J.S., A.L.H. and D.T.M. wrote the manuscript with input from all authors and guidance from A.R., J.R., F.J.T. and M.C.N.

Competing interests

N.K. was a consultant to Biogen Idec, Boehringer Ingelheim, Third Rock, Pliant, Samumed, NuMedii, Indaloo, Theravance, LifeMax, Three Lake Partners and Optikira and received nonfinancial support from MiRagen. All of these were outside the work reported herein. J.L. is a scientific consultant for 10x Genomics. A.R. is a cofounder and equity holder of Celsius Therapeutics, an equity holder in Immunitas and a SAB member of Thermo Fisher Scientific, Syros Pharmaceuticals, Asimov and Neogene Therapeutics. O.R.-R. and A.R. are coinventors on patent applications filed by the Broad Institute to inventions relating to single-cell genomics applications, such as in PCT/US2018/060860 and US Provisional Application no. 62/745,259. A.K.S. received compensation for consulting and has SAB membership from Honeycomb Biotechnologies, Cellarity, Cogen Therapeutics, Orche Bio and Dahlia Biosciences. S.A.T. was a consultant at Genentech, Biogen and Roche in the last 3 years. F.J.T. reports receiving consulting fees from Roche Diagnostics and ownership interest in Cellarity. L.V. is founder of Definigen and Bilitech, two biotech companies using human pluripotent stem cells and organoid cultures for disease modeling and cell-based therapy. J.A.K. has received advisory board fees from Boehringer Ingelheim, and has research contracts with Genentech. E.S.L. serves on the Board of Directors for Codiak BioSciences and serves on the Scientific Advisory Board of F-Prime Capital Partners and Third Rock Ventures; he is also affiliated with several nonprofit organizations including serving on the Board of Directors of the Innocence Project, Count Me In and Biden Cancer Initiative, and the Board of Trustees for the Parker Institute for Cancer Immunotherapy. He has served and continues to serve on various federal advisory committees. J.L. is a scientific consultant for 10x Genomics. J.B., J.C., M.E.R. and S.A.M. are funded in part by a sponsored research agreement from Janssen Pharmaceuticals. A.S. is an employee of Johnson & Johnson. R.J.X. is a cofounder of Celsius Therapeutics and Jnana Therapeutics, and a consultant at Novartis. All other authors declare no competing interests.

Additional information

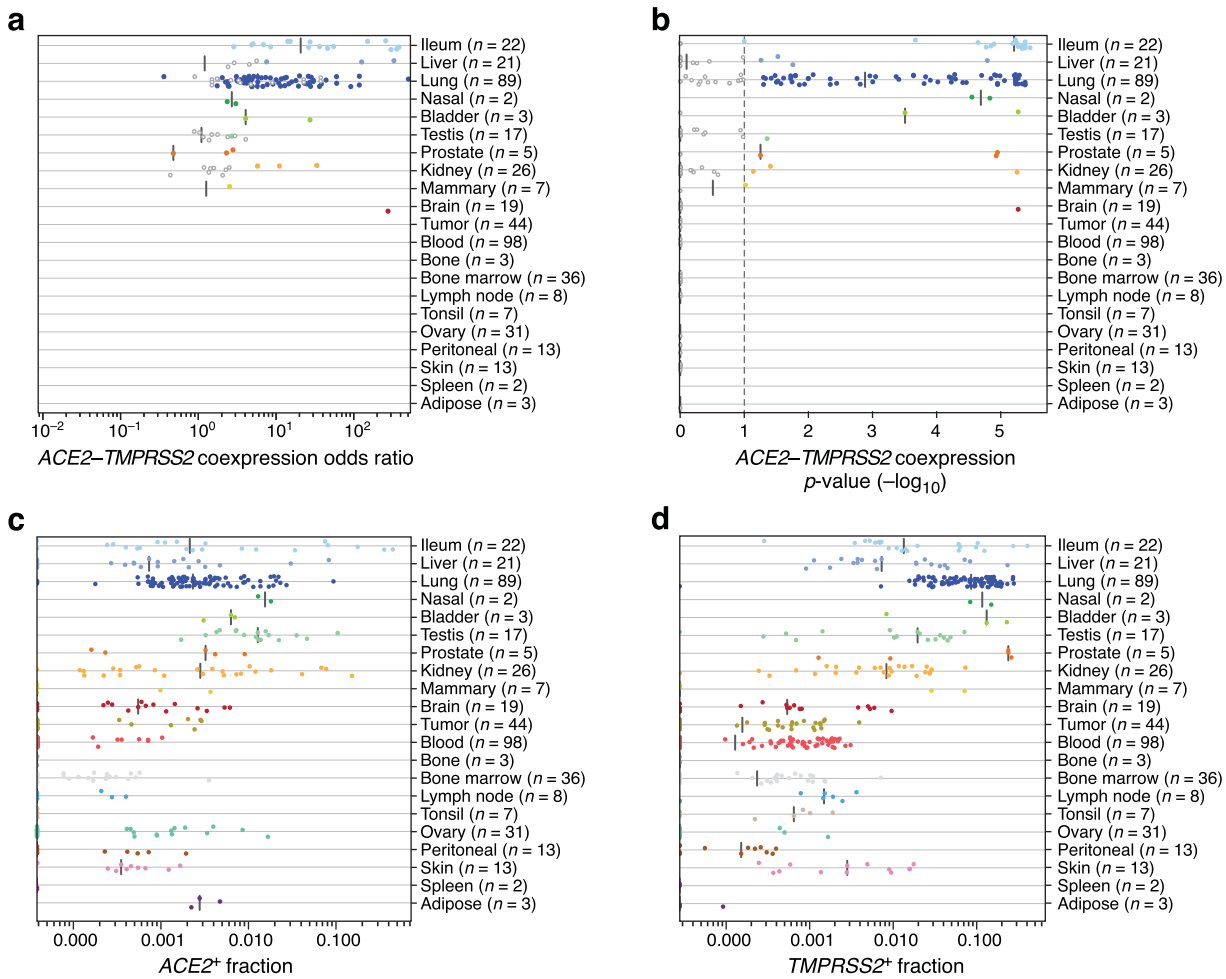
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-01227-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-020-01227-z.

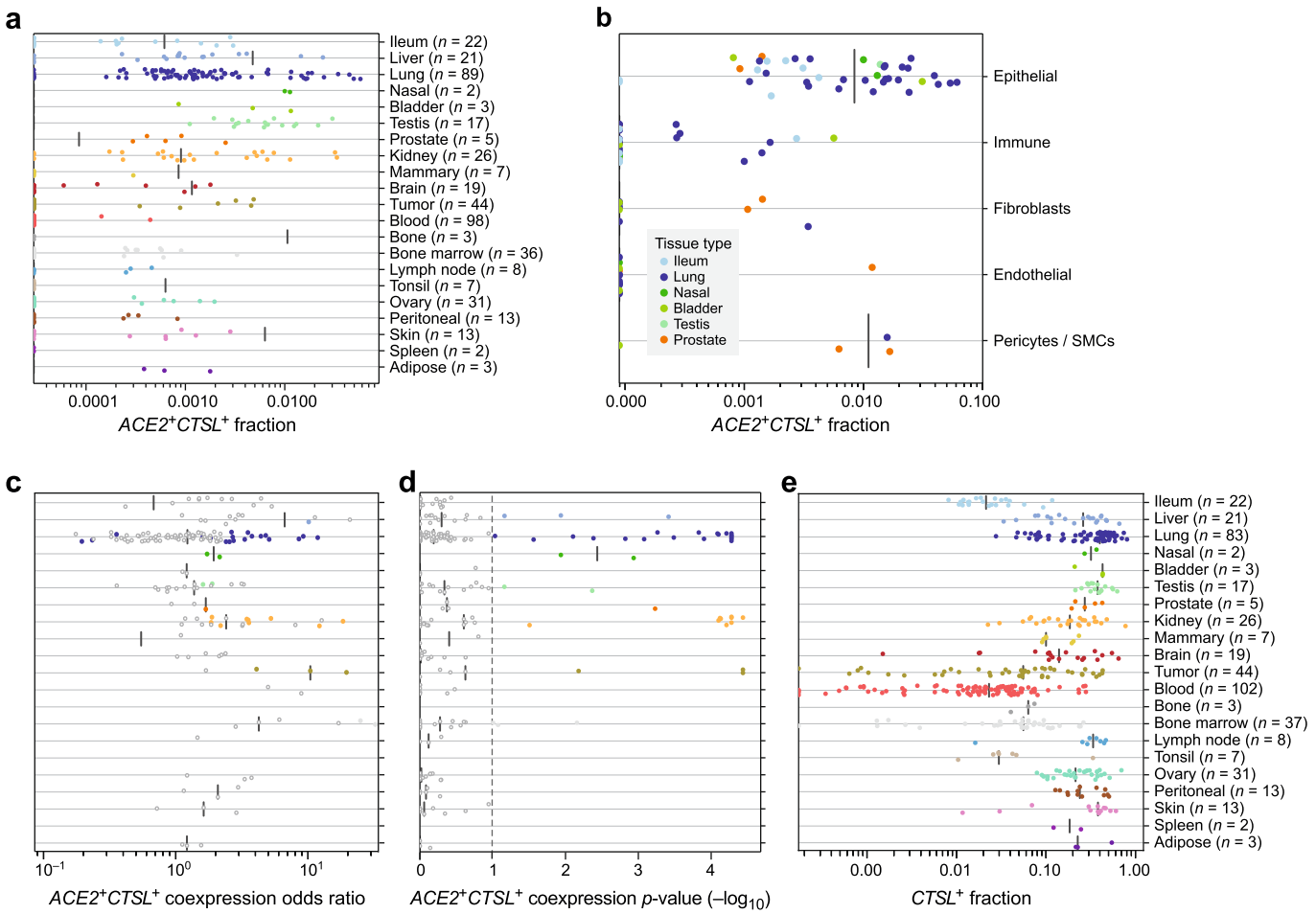
Correspondence and requests for materials should be addressed to C.M. or M.D.L.

Peer review information Saheli Sadanand was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

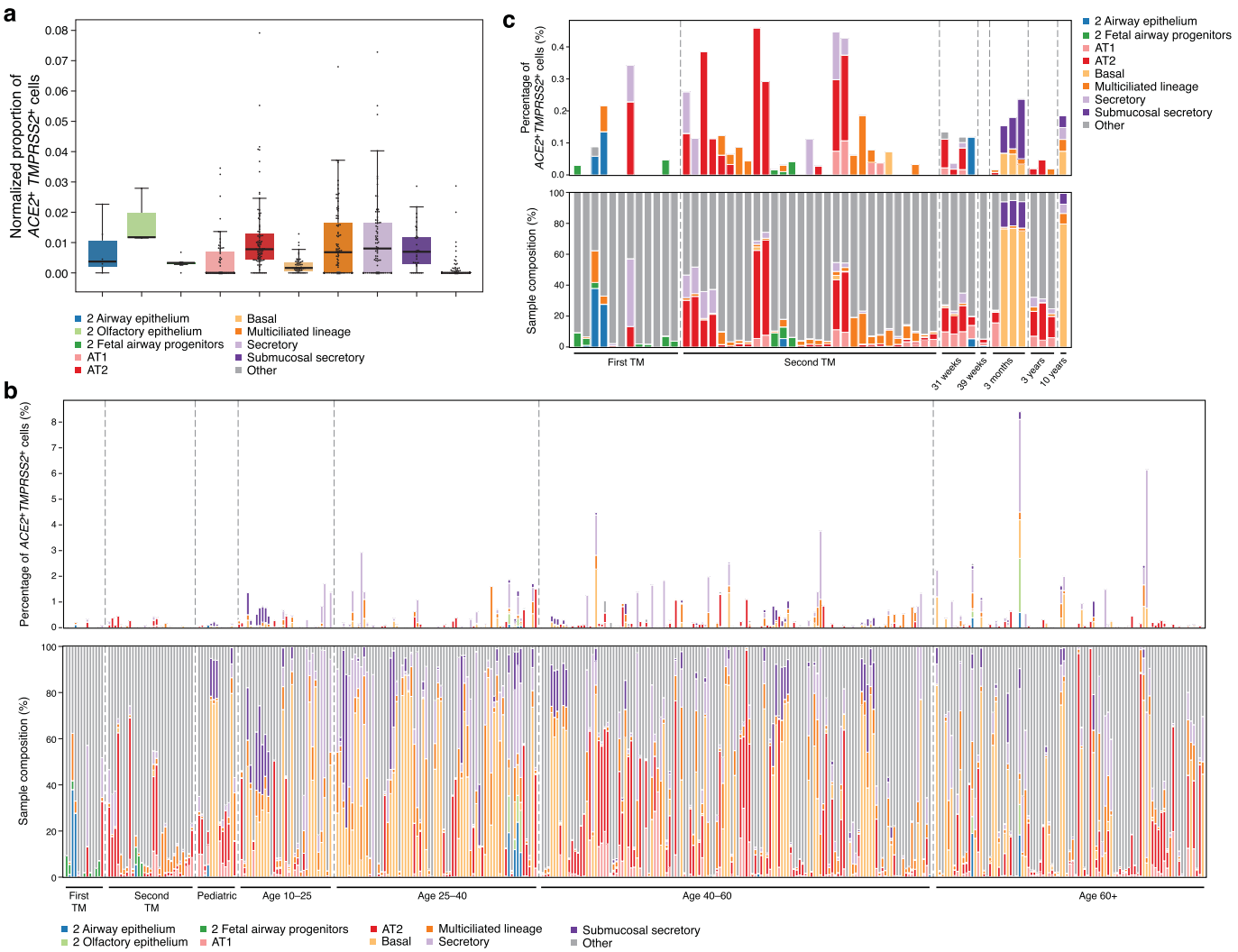
Reprints and permissions information is available at www.nature.com/reprints.



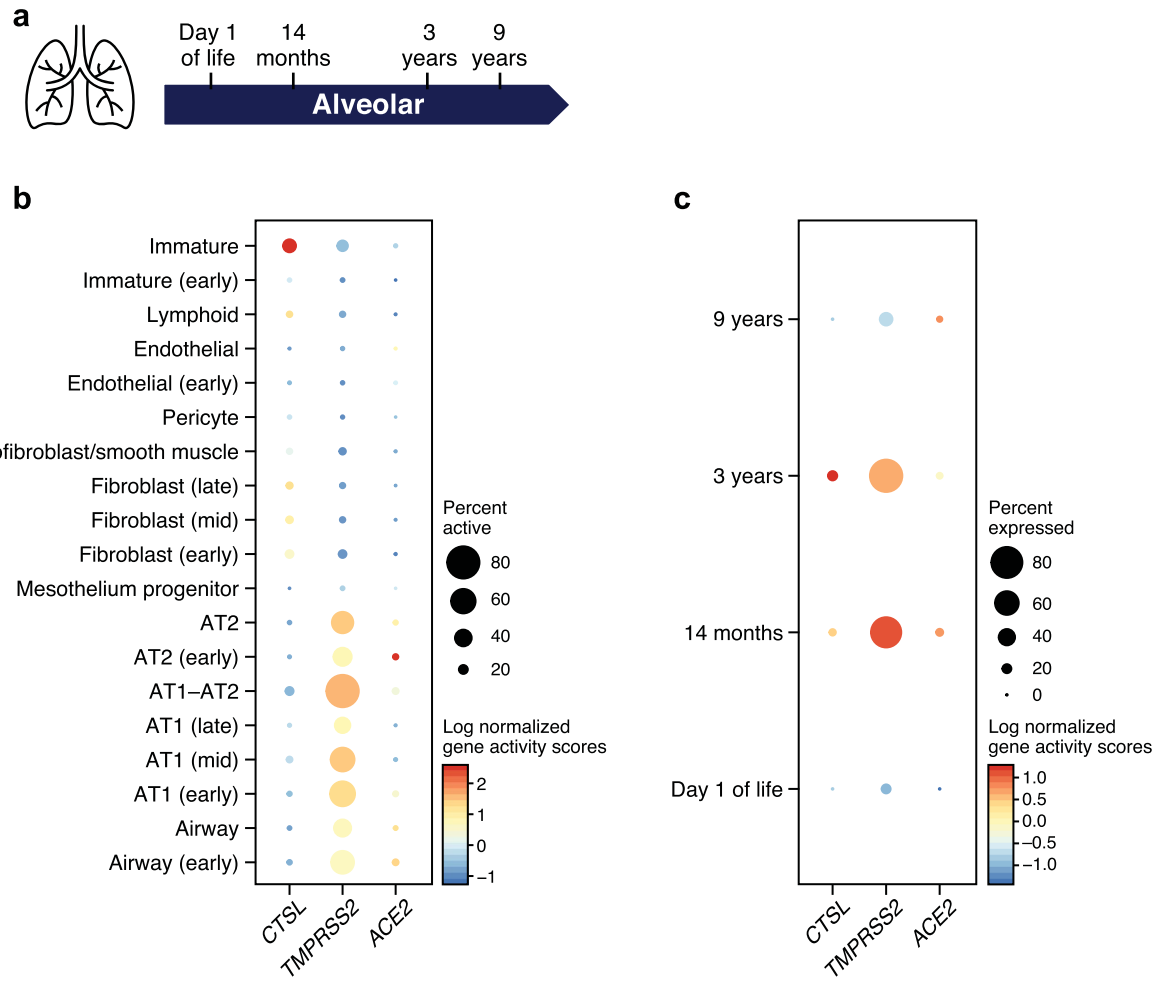
Extended Data Fig. 1 | A cross-tissue survey of ACE2⁺TMPRSS2⁺ cells in published single-cell datasets. a, Odds ratio (x axis) of *ACE2*⁺*TMPRSS2*⁺ coexpression in single-cell datasets (dots) from different tissues (y axis). **b**, Significance (-$\log_{10}$(*p*-value) using two-sided Fisher's exact test, x axis) of coexpression of *ACE2*⁺*TMPRSS2*⁺ in single-cell datasets (dots) from different tissues (y axis). **c,d**, Proportion (x axis) of *ACE2*⁺ cells per dataset (**c**) and *TMPRSS2*⁺ cells per dataset (**d**) across different tissues (y axis).



Extended Data Fig. 2 | A cross-tissue survey of ACE2⁺CTSL⁺ cells in published single-cell datasets. a, Proportion (x axis) of *ACE2*⁺*CTSL*⁺ cells per dataset (dots) across different tissues (y axis). **b**, Proportion (x axis) of *ACE2*⁺*CTSL*⁺ cells within clusters annotated by broad cell-type categories (dots) in each of the top 7 enriched datasets (y axis; color legend, inset). **c**, Odds ratio (x axis) of *ACE2*⁺*CTSL*⁺ coexpression in single-cell datasets (dots) from different tissues (y axis). **d**, Significance (-$\log_{10}$(*p*-value) using two-sided Fisher's exact test, x axis) of coexpression of *ACE2* and *CTSL* in single-cell datasets (dots) from different tissues (y axis). **e,** Proportion (x axis) of *CTSL*⁺ cells per dataset across different tissues (y axis).



Extended Data Fig. 3 | Cellular composition and fraction of ACE2⁺TMPRSS2⁺ cells across the aggregated lung dataset. a, Boxplot of normalized donor fractions of *ACE2⁺TMPRSS2⁺* (double positive - DP) cells per cell type. The box indicates the median and first and third quartile, whiskers extend to points within 1.5 times the interquartile range. For each cell type, only donors that have at least 100 cells of the cell type were included. Cell types with at least 10 *ACE2⁺TMPRSS2⁺* cells in the entire dataset were labeled, the remaining cell types were grouped under 'Other'. Cell type labels preceded by a '2' consist of cells that had no annotation available at level 3 and therefore kept their level 2 annotation. Cells with only level 1 annotations were grouped under 'Other'. (2_Airway epithelium: n=6, 2_Olfactory epithelium: n=3, 2_fetal airway progenitors: n=5, AT1: n=60, AT2: n=92, Basal: n=56, Multiciliated lineage: n=88, Secretory: n=79, Submucosal Secretory: n=35, Other: n=180 donors.). **b**, Percentage of *ACE2⁺TMPRSS2⁺* cells across 377 samples and with sample composition. Top: Percentage *ACE2⁺TMPRSS2⁺* cells in each sample, categorized by level 3 annotations. Bottom: Sample compositions. Samples are ordered by age, with 31-week pre-term births and 39-week full-term births both set to age 0. **c**, Zoom in on fetal and pediatric samples of plot (**b**). Samples are ordered and labeled by age. Fetal samples are partitioned into first and second trimester (TM) and pediatric samples are divided into 31-week pre-term births, 39-week full term births, 3 month, 3 year, and 10 year old children. AT1, 2: alveolar type 1, 2. AT2 progenitor cells were grouped under AT2.

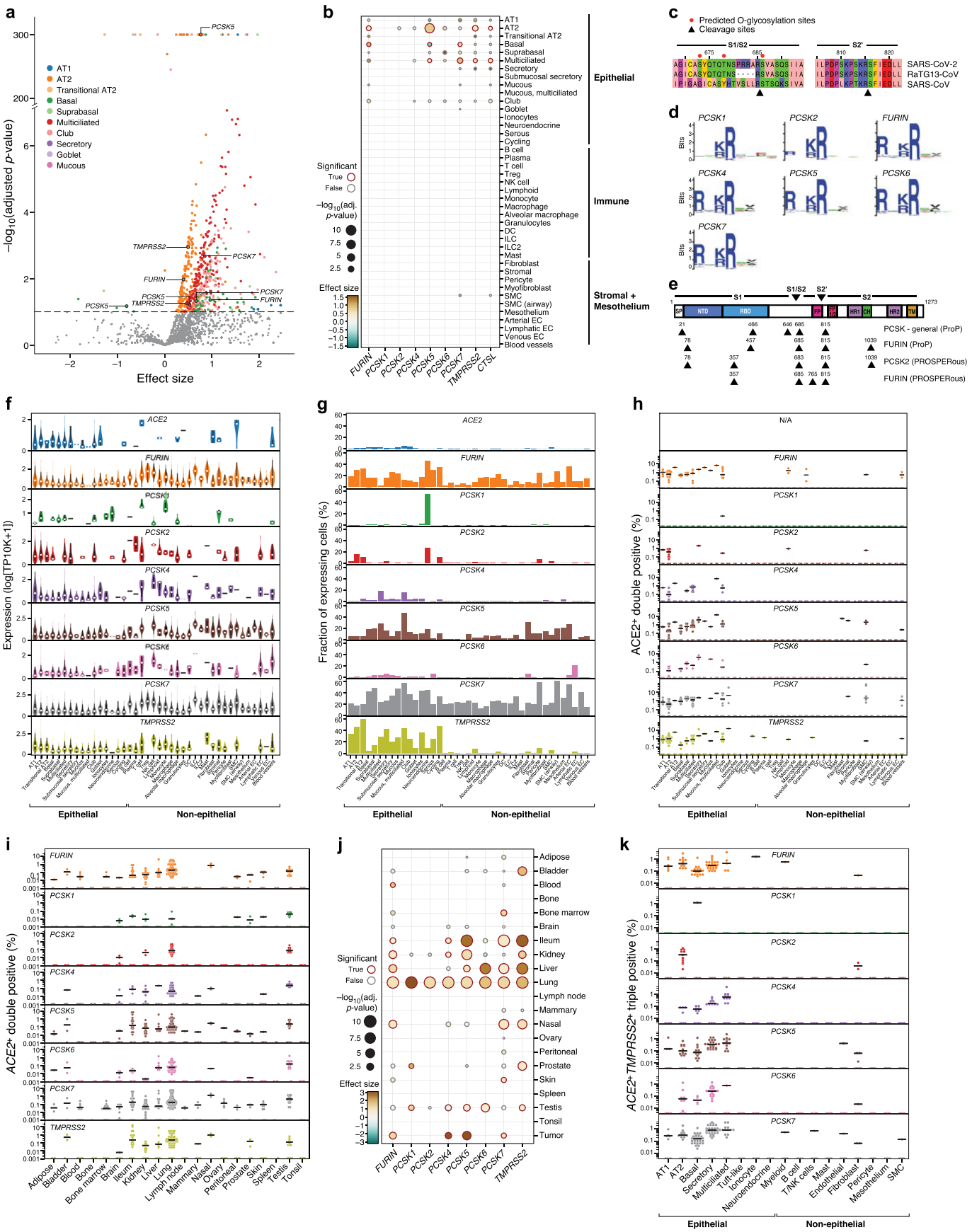


d scTHS-Seq accessibility in all cell types by age: *ACE2*, *TMPRSS2* and *CTSL* (human)

Time point	Day 1	14 mos.	3 yr	9 yr
ACE2$^+$*CTSL*$^+$	81	72	113	4
ACE2$^+$*TMPRSS2*$^+$	349	243	384	54
Total cells	11,444	9077	10,510	4232

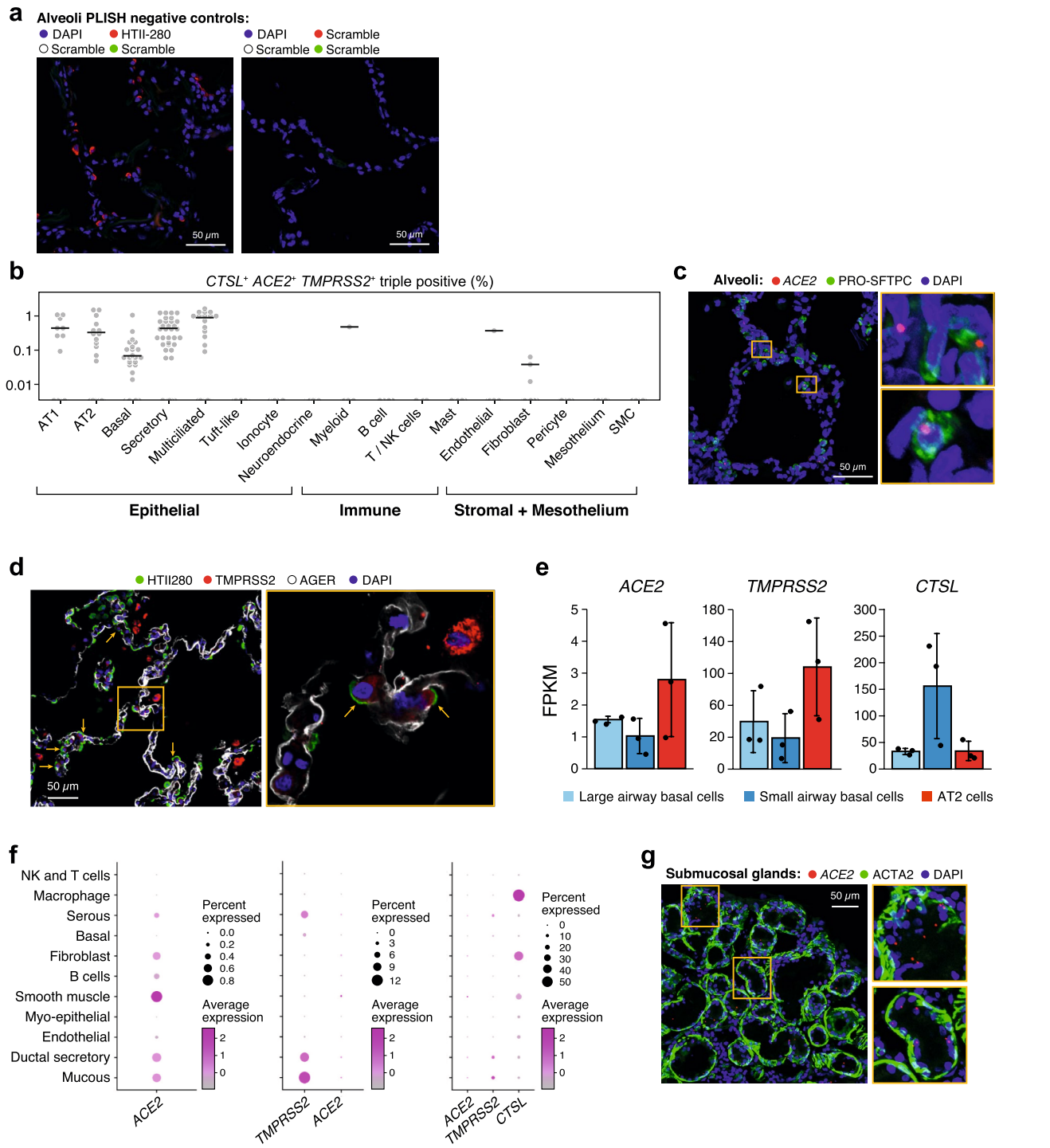
Extended Data Fig. 4 | Chromatin accessibility at the ACE2, TMPRSS and CTSL loci across lung cells in early life. a, Schematic: single-cell chromatin accessibility by transposome hypersensitive sites sequencing (THS-Seq) from human pediatric samples (full gestation, no known lung disease) collected at day 1 of life, 14 months, 3 years, and 9 years (n=1 at each time point). **b**, Accessibility (dot color log normalized gene activity scores), and % of cells with accessible loci (dot size) for the *ACE2, TMPRSS*, and *CTSL* loci (columns) across different cell types (rows) in scTHS-Seq with all time points aggregated. **c**, Accessibility (dot color log normalized gene activity scores), and % of cells with accessible loci (dot size) of *ACE2, TMPRSS* and *CTSL* in AT1–AT2 cells in scTHS-Seq at day 1 of life, 14 months, 3 years, and 9 years (rows). **d**, Number of *ACE2*$^+$*CTSL*$^+$ and *ACE2*$^+$*TMPRSS2*$^+$ cells per time point.

Extended Data Fig. 5 | ACE2 expression across tissues and cell types. Shown are fractions of *ACE2* expressing cells (dot size) and mean *ACE2* expression level in expressing cells (dot color) across datasets (rows) and cell types (columns).



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

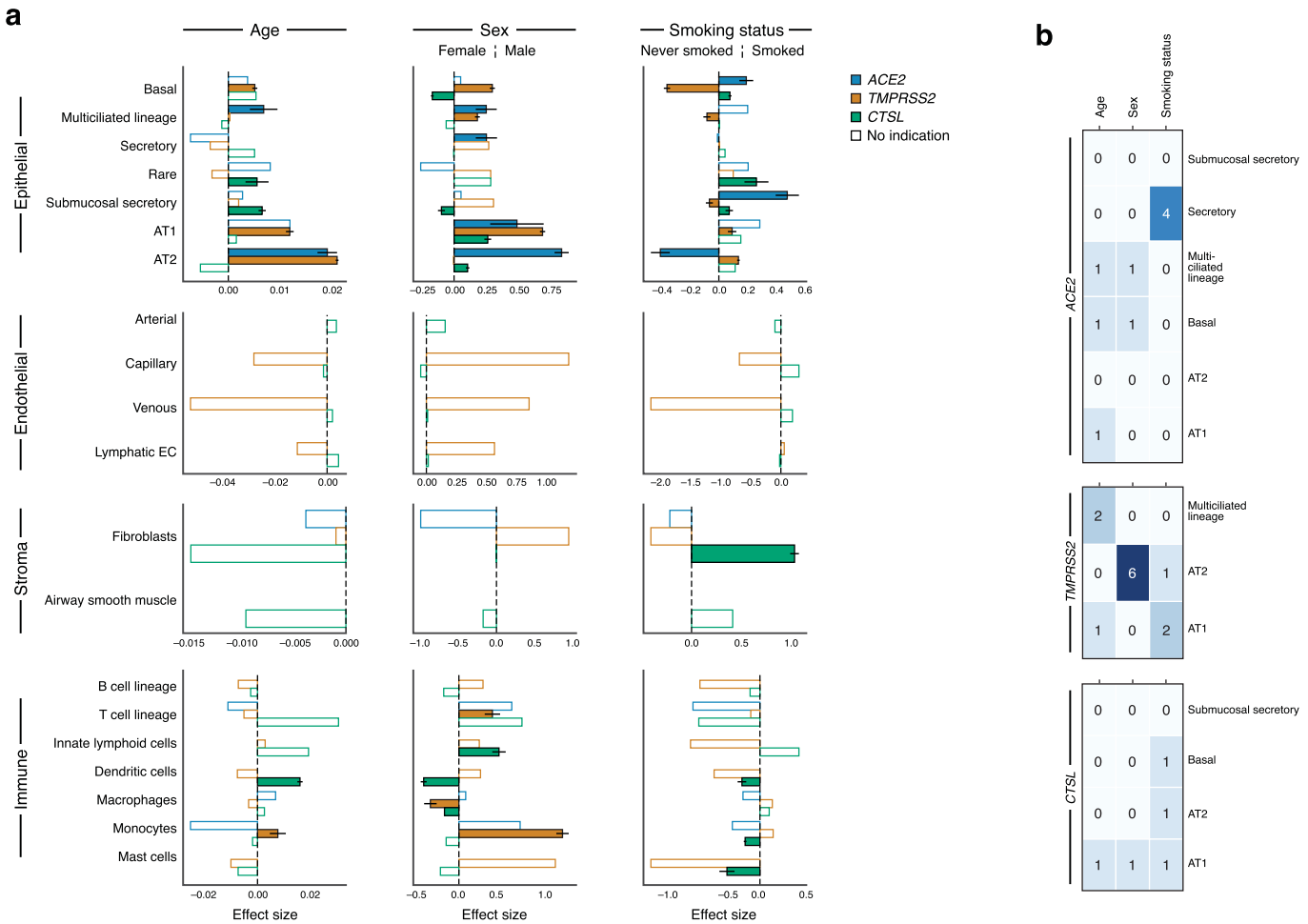
Extended Data Fig. 6 | Additional analyses to identify other proteases that may have a role in infection. a, Multiple proteases are coexpressed with *ACE2* in another human lung scRNA-seq ('aggregated lung'). Scatter plot of significance (y axis, -log10(adjusted p value) by two-sided Wald test. (Methods)) and effect size (x axis) of coexpression of each protease gene (dot) with *ACE2* within each indicated epithelial cell type (color). Dashed line: significance threshold. *TMPRSS2* and PCSKs that significantly coexpressed with *ACE2* are marked. **b**, *ACE2*-protease coexpression with PCSKs, *TMPRSS2* and *CTSL* across lung cell types ('aggregated lung'). Significance (dot size, -log10(adjusted p value) by two-sided Wald test. (Methods)) and effect size (color) for coexpression of *ACE2* with selected proteases (columns) across cell types (rows). **c-d,** Predicted cleavage sites in the SARS-CoV-2 S-protein S1/S2 region. (**c**) Multiple amino acid sequence alignment of SARS-CoV-2 S-protein S1/S2 region with orthologous sequences from other betacoronaviruses (top) and polybasic cleavage sites of other human pathogenic viruses (bottom). **d**, Sequence logo plot showing cleavage site preference derived from MEROPS database for PCSK1, PCSK2, FURIN, PCSK4, PCSK5, PCSK6 and PCSK7. **e**, Protease cleavage sites (triangles) predicted by ProP and PROSPERous in the SARS-CoV-2 spike protein. Top: Full-length SARS-CoV-2 S-protein sequence schematic with predicted functional protein domains and motifs. Numbers: amino acid residues after which cleavage occurs; SP: signal peptide; NTD: N-terminal domain; RBD: Receptor-binding domain; FP: Fusion peptide; FP1/2: Fusion peptide 1/2; HR1: Heptad repeat 1; CH: connecting helix; HR2: Heptad repeat 2; TM: Transmembrane domain. **f,g**, Multiple proteases are expressed across lung cell types ('aggregated lung'). **f**, Distribution of non-zero expression (y axis) for *ACE2*, PCSKs and *TMPRSS2* across lung cell types (x axis). White dot: median non-zero expression. **g**, Proportion of cells (y axis) expressing *ACE2*, *PCSK* family or *TMPRSS2* across lung cell types (x axis), ordered by compartment. **h**, $ACE2^{+}PCSK^{+}$ double positive cells across lung cell types. Fraction (y axis) of different $ACE2^{+}PCSK^{+}$ or $ACE2^{+}TMPRSS2^{+}$ double positive cells across lung cell types, ordered by compartment (x axis). Dots: different samples, line: median of non-zero fractions. **i,j**, $ACE2^{+}PCSK^{+}$ coexpression across human tissues (collection of published scRNA seq datasets). **i**, Percent (y axis) of different $ACE2^{+}PCSK^{+}$ or $ACE2^{+}TMPRSS2^{+}$ double positive cells across human tissues (x axis). Dots: different single-cell datasets, line: median of non-zero fractions. **j**, *ACE2* coexpression with PCSKs or *TMPRSS2* across human tissues. Significance (dot size, -log10(adjusted p value) by two-sided Wald test. (Methods)) and effect size (dot color) of coexpression. **k**, Fraction of $ACE2^{+}TMPRSS2^{+}PCSK^{+}$ cells across lung cell types ('Regev/Rajagopal dataset'). Dots: samples, line: median of non-zero fractions.



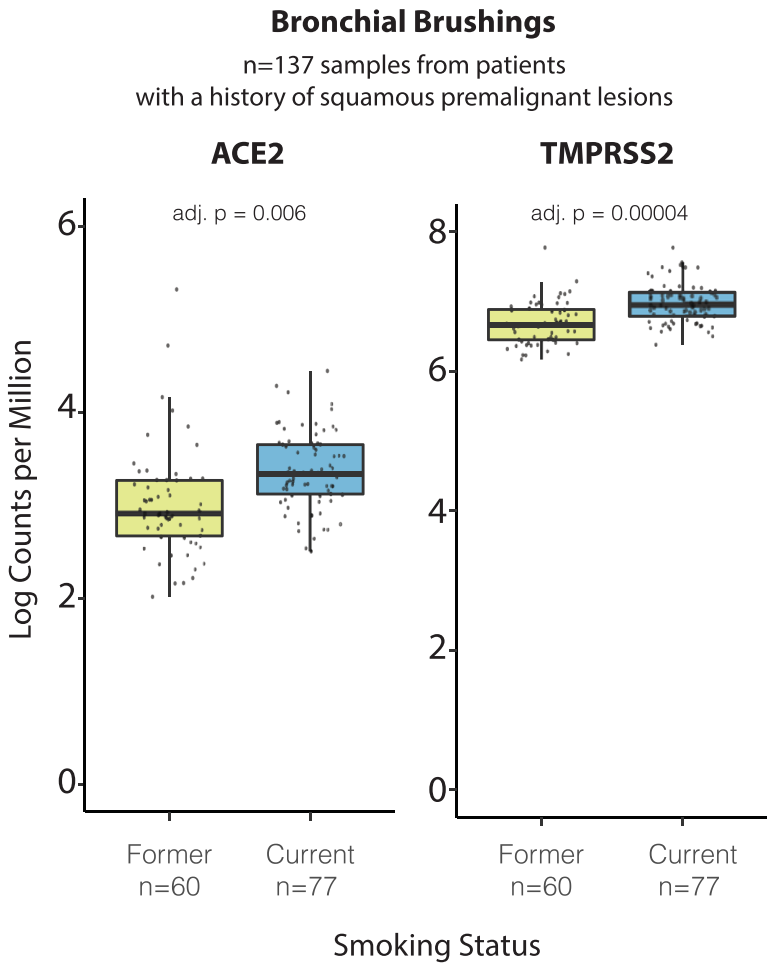
Extended Data Fig. 7 | ACE2, TMPRSS2, CTSL Immunofluorescence and RNA profiling. a, Negative control of PLISH in human lung alveoli. Left shows scrambled probe detection in three indicated colors. Right shows HTII-280 antibody staining (red) with 2 color scramble probe detection. DAPI (blue) indicates nuclei. **b,** Frequency of *ACE2, CTLS* and *TMPRSS2* triple positive cells in each sample (n = 60) (dots) in the Regev/Rajagopal dataset. **c**, PLISH and immunostaining in human adult lung alveoli for *ACE2* (red), PRO-SFTPC (green), DAPI (blue). **d,** Immunostaining in human adult lung alveoli. HTII-280 (green), TMPRSS2 (red) and AGER (white). Blue shows DAPI in nuclei. **e**, Mean expression (y axis, FPKM, from bulk RNA-seq, error bars: standard errors) of *ACE2, CTSL, TMPRSS2* in sorted cells from 3 different human explant donors using the following markers: large and small airway basal cells (NGFR+), AT2 cells (HT-II 280+) and alveolar organoids (HT-II 280+). **f**, Expression in the submucosal gland. Mean expression (color) and proportion of expressing cells (dot size) of *ACE2*, *TMPRSS2* and *CTSL* in key cell types (rows), from scRNA-seq of human large airway submucosal glands. **g**, PLISH and immunostaining in human large airway submucosal glands. *ACE2* (red), *ACTA2* (green) and DAPI (blue). We imaged one representative area for a single patient for a,c,d,g (Methods).

Level 1	Level 2	Level 3
Epithelial	Airway epithelium	Basal Multiciliated Secretory Rare (PNEC, Tuft-like, Ionocyte) proliferat. Epithelial cells KRT5-/KRT17+
	Submucosal gland	Submucosal secretory Acinar
	Alveolar epithelium	AT1 AT2
	Olfactory epithelium	Olfactory epithelium Basaloid
Endothelial	Blood vessels	Arterial Capillary Venous Bronchial Vessel 1 Bronchial Vessel 2 Capillary Intermediate 1 Capillary Intermediate 2
	Lymphatic	Lymphatic EC
Stroma	Fibroblast	Fibroblast Myofibroblast
	Smooth muscle	Airway smooth muscle Venous smooth muscle Fibromyocyte
	Mesothelium	Mesothelium
Immune	Lymphoid	B cell lineage T cell lineage Innate lymphoid
	Myeloid	Dendritic Macrophage Monocyte Mast
	Granulocytes	Basophilic Neutrophilic Eosinophilic
	Megakariocytic	Megakaryocytes Erythrocytes
Cycling cells	Cycling	

Extended Data Fig. 8 | Consensus ontology for lung dataset meta-analysis. An overview of the three-level lung cell ontology used for cell annotation harmonization. For consistent analysis across datasets we mapped annotations to this ontology. PNEC: pulmonary neuroendocrine cells. AT1, 2: alveolar type 1, 2. EC: endothelial cells.



Extended Data Fig. 9 | Age, sex, and smoking status associations with expression of ACE2, TMPRSS2, and CTSL across level 3 cell type annotations modeled without interaction terms. a, Age, sex, and smoking assocations with expression of *ACE2* (blue), *TMPRSS2* (yellow), and *CTSL* (green) modeled without interaction terms on 985,420 cells from 164 donors. Level 3 cell types are shown on the y-axes, and are subdivided by level 1 cell type annotations (top to bottom: epithelial, endothelial, stromal and immune cells). The effect size (x axis) is given as a log fold change (sex, smoking status) or the slope of log expression per year (age). Positive effect sizes indicate increases with age, in males, and in smokers. As the age effect size is given per year, it is not directly comparable to the sex and smoking status effect sizes. Colored bars: associations with an FDR-corrected p-value<0.05 (one-sided Wald test on regression model coefficients), consistent effect direction in pseudo-bulk analysis, and consistent results using the model with interaction terms (Methods). White bars: associations that do not pass all of the three above-mentioned evaluation criteria. Error bars: standard errors around coefficient estimates. Error bars are only shown for colored bars (indications or robust trends) to limit figure size. Only cell types with at least 1000 cells across donors are included. Number of cells and donors per cell type: Basal: 155877, 105, Multiciliated lineage: 37530, 157, Secretory: 22306, 140, Rare: 2676, 71, Submucosal secretory: 33661, 45, AT1: 29973, 101, AT2: 155512, 104, Arterial: 3497, 37, Capillary: 15745, 34, Venous: 7173, 33, Lymphatic EC: 5055, 76, Fibroblasts: 9112, 51, Airway smooth muscle: 1077, 13, B cell lineage: 11761, 90, T cell lineage: 52139, 97, Innate lymphoid cells: 29836, 56, Dendritic cells: 9017, 90, Macrophages: 156964, 89, Monocytes: 42703, 96, Mast cells: 13581 cells, 88 donors. **b**, Robustness of associations to holding out a dataset. The values show the number of held-out datasets that result in loss of association between a given covariate (rows) and *ACE2, TMPRSS2*, or *CTSL* expression in a given cell type (columns). Robust trends are determined by significant effects that are robust to holding out any dataset (0 values). From left to right: results for *ACE2, TMPRSS2*, and *CTSL*. AT1, 2: alveolar type 1, 2. EC: endothelial cell.



Extended Data Fig. 10 | ACE2 and TMPRSS2 are up-regulated in bronchial brushings from current versus former smokers. Boxplots of log counts per million normalized gene expression for *ACE2* and *TMPRSS2* are plotted across current (red, n=70 samples) versus former (green, n=60 samples) smokers. Both genes are significantly up-regulated in current versus former/never (*ACE2*, FDR=0.006; and *TMPRSS2*, FDR=0.00004) based on a linear model using voom-transformed data that included genomic smoking status, batch, and RNA quality (TIN) as covariates and patient as a random effect. Multiple testing correction was performed via Benjamini-Hochberg to obtain an FDR-corrected p-value. (Methods).

nature research

Corresponding author(s): Muus, Luecken, Regev, HCA Lung Biological Network

Last updated by author(s): Dec 3, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	R packages: lme4 (1.1.21), Signac (0.1.6), JASPAR (2020), EnrichR (1.0), Harmony (0.1.1), Seurat (3.1.1), chromVAR (1.6.0), TFBSTools (1.25.1), MAST (1.8.2), gprofiler (1.0.0), GenomeInfoDb (1.22.0), python packages: harmony-pytorch (version 0.1.1), scanpy (versions 1.4.5, 1.4.5.1, 1.4.6), statsmodels (version 0.11.1), sklearn (version 0.21.3), ForceAtlas2 (version 0.3.5), louvain (version 0.6.1), networkx (version 2.2), diffxpy (version 0.7.3), batchglm (version 0.7.4) 10X Genomics CellRanger ATAC mkfastq (version 1.1.0), CellphoneDB (version 2.0.0), Star (Version 20201), Cufflinks (Version 2.2.1.3), Cuffdiff (Version 2.2.1.6)

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

Data and an interactive analysis examining the co-expression of genes across datasets can be accessed via the open-source data platform, Terra at https://app.terra.bio/#workspaces/kco-incubator/COVID-19_cross_tissue_analysis.

Interactive visualization and download of gene expression data can be accessed on the Single Cell Portal at http://broad.io/hcacovid19

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Meta-analysis of single-cell atlases, see main text and supplementary tables for included studies.
Data exclusions	Meta-analysis of single-cell atlases, see main text and supplementary tables for included studies.
Replication	N/A (the data was not specifically collected for this meta-analysis)
Randomization	N/A (the data was not specifically collected for this meta-analysis)
Blinding	N/A (the data was not specifically collected for this meta-analysis)

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

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

Antibodies

Antibodies used

Primary antibodies
ACTA2 Mouse IgG2a, FITC-conjugated (Sigma, F3777, Clone: 1A4, lot: 038M4865V, 1:500)
AGER Goat IgG (R&D systems, AF1145, lot: HCE0719011, 1:200)
HTII-280 Mouse IgM (Terrace Biotech, TB-27AHT2-280, lot: not available, 1:100)
Pro-SFTPC Rabbit IgG (Sigma, ab3786, lot: 3267937, 1:500)
TMPRSS2 Rabbit IgG (Abcam, ab109131, Clone: EPR3862, lot: GR3248440-1, 1:200)

Secondary antibodies
Alexa Fluor 488 Donkey anti-Mouse IgM (Thermo fisher scientific, A21042, lot: 2160416, 1:400)
Alexa Fluor 488 Donkey anti-Rabbit IgG (Thermo fisher scientific, A32795, lot: 1981155, 1:400)
Alexa Fluor 594 Donkey anti-Rabbit IgG (Thermo fisher scientific, A21207, lot: 1987293, 1:400)
Alexa Fluor 594 Goat anti-Mouse IgM (Thermo fisher scientific, A21044, lot: 1806144, 1:400)
Alexa Fluor 647 Donkey anti-Goat IgG (Thermo fisher scientific, A21447, lot:2175459, 1:400)

Validation

Validation statements and relevant citations of the listed antibodies are available in the manufacturer's websites:

Primary antibodies:
ACTA2 Mouse IgG2a, FITC-conjugated(https://www.sigmaaldrich.com/catalog/product/sigma/f3777?lang=en&region=US)
AGER Goat IgG (https://www.rndsystems.com/products/human-mouse-rat-rage-ager-antibody_af1145)
HTII-280 Mouse IgM (https://www.terracebiotech.com/product-page/anti-ht2-280-1ml)
Pro-SFTPC Rabbit IgG (http://www.emdmillipore.com/US/en/product/Anti-Prosurfactant-Protein-C-proSP-C-Antibody,MM_NF-AB3786?ReferrerURL=https%3A%2F%2Fwww.google.com%2F&bd=1)
TMPRSS2 Rabbit IgG (https://www.abcam.com/tmprss2-antibody-epr3862-ab109131.html)

Secondary antibodies:
Alexa Fluor 488 Donkey anti-Mouse IgM (https://www.thermofisher.com/antibody/product/Goat-anti-Mouse-IgM-Heavy-chain-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21042)
Alexa Fluor 488 Donkey anti-Rabbit IgG (https://www.thermofisher.com/antibody/product/Donkey-anti-Rabbit-IgG-H-L-Highly-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A32795)
Alexa Fluor 594 Donkey anti-Rabbit IgG (https://www.thermofisher.com/antibody/product/Donkey-anti-Rabbit-IgG-H-L-Highly-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21207)
Alexa Fluor 594 Goat anti-Mouse IgM (https://www.thermofisher.com/antibody/product/Goat-anti-Mouse-IgM-Heavy-chain-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21044)
Alexa Fluor 647 Donkey anti-Goat IgG (https://www.thermofisher.com/antibody/product/Donkey-anti-Goat-IgG-H-L-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21447)

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Wild-type Mus musculus, C57BL/6J and Cast/EiJ, both Females and Males were ordered from the Jackson Lab. Mice were housed under standard barrier conditions at the Whitehead Institute for Biomedical Research. Experimental mice were obtained by crossing Cast/EiJ mice with C57BL/6J mice. For the smoke exposure experiments, 8 to 10 week old pathogen-free female wild-type C57BL/6 mice were obtained from Charles River (Sulzfeld, Germany).
Wild animals	none
Field-collected samples	none
Ethics oversight	Mice were housed under standard barrier conditions at the Whitehead Institute for Biomedical Research. All experiments performed in this study were in accordance with the relevant animal husbandry standards of the Committee on Animal Care. All mouse exposure experiments were approved by the ethics committee for animal welfare of the local government for the administrative region of Upper Bavaria (Regierungspräsidium Oberbayern) and were conducted under strict governmental and international guidelines in accordance with EU Directive 2010/63/EU.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Provided in supplementary table 2 for lung single-cell datasets, N/A or not specifically collected for the remaining studies.
Recruitment	N/A (the data was not specifically collected for this meta-analysis)
Ethics oversight	Sample collection underwent IRB review and approval at the institutions where the samples were originally collected. "Adipose_Healthy_Manton_unpublished" was collected under IRB 2007P002165/1(ORSP-3877). Tissue samples from breast, esophagus muscularis, esophagus mucosa, heart, lung, prostate, skeletal muscle and skin referred to as "Tissue_Healthy_Regev_snRNA-seq_unpublished" were collected under ORSP-3635. Samples referred to as "Eye_Sanes_unpublished" were collected under Dana Farber / Harvard Cancer Center Protocol Number 13-416 and Massachusetts Eye and Ear Protocol Number 18-034H. Samples referred to as "Kidney_Healthy_Greka_unpublished" were collected under Massachusetts General Hospital IRB number 2011P002692. Samples referred to as "Liver_Healthy_Manton_unpublished" were collected under IRB 02-240; ORSP 1702 as well as and ORSP-2630 under ORSP-2169. Lung samples from smokers and non-smokers (41 samples, 10 patients, 2-6 locations each) with suffix "Regev/ Rajagopal_unpublished" were collected under Massachusetts General Hospital IRB 2012P001079 / (ORSP-3900) under ORSP-3490. Healthy and fibrotic lung samples with suffix "Xavier_snRNA-seq_unpublished" were collected under Massachusetts General Hospital IRB number 2003P000555 (CG-5242) under ORSP-3490, Medoff, 2015P000319 (CG-5145) under ORSP-3490. Pancreas PDAC samples were collected under Fernandez-del Castillo, 2003P001289 (CG-4692) under ORSP-3490 Massachusetts General Hospital IRB number Fernandez-del Castillo, 2003P001289 (CG-4692) under ORSP-3490. Samples in the dataset "Barbry" were derived from a study that was approved by the Comité de Protection des Personnes Sud Est IV (approval number: 17/081) and informed written consent was obtained from all participants involved. All experiments were performed during 8 months, in accordance with relevant guidelines and French and European regulations. No deviations were made from our approved protocol named 3Asc (An Atlas of Airways at a single cell level - ClinicalTrials.gov identifier: NCT03437122). IPF and COPD lungs in the "Kaminski" dataset were obtained from patients undergoing transplant while healthy lungs were from rejected donor lung organs that underwent lung transplantation at the Brigham and Women's Hospital or donor organs provided by the National Disease Research Interchange (NDRI). Patient tissues relating to the dataset "Krasnow" were obtained under a protocol

approved by Stanford University's Human Subjects Research Compliance Office (IRB 15166) and informed consent was obtained from each patient prior to surgery. The study protocol was approved by the Partners Healthcare Institutional Board Review (IRB Protocol # 2011P002419). Samples in the dataset "Kropski_Banovich" were collected under Vanderbilt IRB # 060165, 171657, and Western IRB#20181836. Ethics approval number 2018/769-31. "Meyer_b" were collected under CBTM (Cambridge Biorepository for Translational Medicine), research ethics approval number: UK NHS REC approval reference number 15/EE/0152. Samples in the dataset "Linnarsson" are covered by (2018/769-31) approved by the Swedish Ethical Review Authority. Samples in the "Misharin" dataset were collected under (STU00056197, STU00201137, and STU00202458) approved by the Northwestern University Institutional Review Board. Samples in the "Rawlins" dataset were obtained from terminations of pregnancy from Cambridge University Hospitals NHS Foundation Trust under permission from NHS Research Ethical Committee (96/085) and the Joint MRC/Wellcome Trust Human Developmental Biology Resource (grant R/R006237/1, www.hdbr.org, HDBR London: REC approval 18/LO/0822; HDBR Newcastle: REC approval 18/NE/0290). The studies relating to datasets "Schultze" and "Schultze_Falk" were approved by the ethics committees of the University of Bonn and University hospital Bonn (local ethics vote 076/16) and the Medizinische Hochschule Hannover (local ethics vote 7414/2017). Fifteen human tracheal airway epithelia in the "Schultze" dataset were isolated from de-identified donors whose lungs were not suitable for transplantation. Lung specimens were obtained from the International Institute for the Advancement of Medicine (Edison, NJ) and the Donor Alliance of Colorado. The National Jewish Health Institutional Review Board (IRB) approved the research under IRB protocols HS-3209 and HS-2240. Samples in the "Xu/Whitsett" dataset were provided through the federal United Network of Organ Sharing via the National Disease Research Interchange (NDRI) and International Institute for Advancement of Medicine (IIAM) and entered into the NHLBI LungMAP Biorepository for Investigations of Diseases of the Lung (BRINDL) at the University of Rochester Medical Center, overseen by the IRB as RSRB00047606. (Supplementary Table 1, 2)

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