Cell Systems

Article

Multiomic analysis reveals cell-type-specific molecular determinants of COVID-19 severity

Highlights

- Machine learning combines GWAS with single-cell omics to discover COVID-19 risk genes
- The discovered severe COVID-19 risk genes account for 77% of the observed heritability
- Genetic risk for severe COVID-19 is focused within NK cells and T cells
- Mendelian randomization and single-cell multiomics highlight CD56^{bright} NK cells

Authors

Sai Zhang, Johnathan Cooper-Knock, Annika K. Weimer, ..., Mark M. Davis, Philip S. Tsao, Michael P. Snyder

Correspondence

ptsao@stanford.edu (P.S.T.), mpsnyder@stanford.edu (M.P.S.)

In brief

Zhang et al. apply a machine learning method that integrates single-cell multiomics with GWAS summary statistics for gene discovery. Application to severe COVID-19 identifies >1,000 risk genes, which account for 77% of the observed heritability. Genetic risk is focused within NK cells, CD56^{bright} cytokine-producing NK cells in particular, highlighting the dysfunction of these cells as a determinant of severe disease.

Zhang et al., 2022, Cell Systems *13*, 598–614
August 17, 2022
https://doi.org/10.1016/j.cels.2022.05.007

CellPress

CellPress
OPEN ACCESS

Cell Systems

Article

Multiomic analysis reveals cell-type-specific molecular determinants of COVID-19 severity

Sai Zhang,[1,2,3,4,18] Johnathan Cooper-Knock,[5,18] Annika K. Weimer,[1,3] Minyi Shi,[1,3] Lina Kozhaya,[6] Derya Unutmaz,[6] Calum Harvey,[5] Thomas H. Julian,[5] Simone Furini,[7] Elisa Frullanti,[7,8] Francesca Fava,[7,8,9] Alessandra Renieri,[7,8,9] Peng Gao,[1,3] Xiaotao Shen,[1,3] Ilia Sarah Timpanaro,[10] Kevin P. Kenna,[10] J. Kenneth Baillie,[11,12,13] Mark M. Davis,[14,15,16] Philip S. Tsao,[2,4,17,*] and Michael P. Snyder[1,3,4,19,*]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[2]VA Palo Alto Epidemiology Research and Information Center for Genomics, VA Palo Alto Health Care System, Palo Alto, CA 94304, USA
[3]Center for Genomics and Personalized Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[4]Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA 94305, USA
[5]Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield S10 2HQ, UK
[6]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA
[7]Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy
[8]Medical Genetics, Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy
[9]Genetica Medica, Azienda Ospedaliero-Universitaria Senese, 53100 Siena, Italy
[10]Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, 3584 CX Utrecht, the Netherlands
[11]Roslin Institute, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK
[12]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK
[13]Intensive Care Unit, Royal Infirmary of Edinburgh, Edinburgh EH16 4SA, UK
[14]Institute for Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA 94305, USA
[15]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA
[16]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305, USA
[17]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[18]These authors contributed equally
[19]Lead contact
*Correspondence: ptsao@stanford.edu (P.S.T.), mpsnyder@stanford.edu (M.P.S.)
https://doi.org/10.1016/j.cels.2022.05.007

SUMMARY

The determinants of severe COVID-19 in healthy adults are poorly understood, which limits the opportunity for early intervention. We present a multiomic analysis using machine learning to characterize the genomic basis of COVID-19 severity. We use single-cell multiome profiling of human lungs to link genetic signals to cell-type-specific functions. We discover >1,000 risk genes across 19 cell types, which account for 77% of the SNP-based heritability for severe disease. Genetic risk is particularly focused within natural killer (NK) cells and T cells, placing the dysfunction of these cells upstream of severe disease. Mendelian randomization and single-cell profiling of human NK cells support the role of NK cells and further localize genetic risk to CD56^{bright} NK cells, which are key cytokine producers during the innate immune response. Rare variant analysis confirms the enrichment of severe-disease-associated genetic variation within NK-cell risk genes. Our study provides insights into the pathogenesis of severe COVID-19 with potential therapeutic targets.

INTRODUCTION

Infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) giving rise to coronavirus disease 2019 (COVID-19) has caused a global pandemic with almost unprecedented morbidity and mortality (Dong et al., 2020). Severity of COVID-19 is markedly variable ranging from an asymptomatic infection to fatal multiorgan failure. Severity correlates with age and co-morbidities (Shang et al., 2020) but not exclusively (Li et al., 2020). Host genetics is thought to be an essential determinant of severity (The COVID-19 Host Genetics Initiative, 2020), but this is poorly understood. Improved tools to identify individuals at risk of severe COVID-19 could facilitate life-saving precision medicine.

Profiles of the immune response associated with severe COVID-19 have been largely observational; they have produced conflicting conclusions and struggled to assign causality. Studies have variously linked severity to CD8 T cells (Mathew et al., 2020), CD19 B cells (Sosa-Hernández et al., 2020), eosinophils (Lucas et al., 2020), and myeloid cells (Arunachalam et al.,

2020), and even specific subtypes of immune cells such as adaptive natural killer (NK) cells (Maucourant et al., 2020). Single-cell omics profiling has demonstrated the differential function of various immune cell types in severe disease (Delorey et al., 2021a; Liao et al., 2020; Melms et al., 2021; Ren et al., 2021; Stephenson et al., 2021; Zhang et al., 2020a), but this work has focused on the transcriptome profiled during acute COVID-19, which is likely to be modified by the ongoing infection and is not necessarily an upstream determinant of outcome. We have focused on host genetics so as to circumvent this problem because genetic variation is largely fixed at conception.

There have been several efforts to address the genetic basis of COVID-19 severity (Severe COVID-19 GWAS Group et al., 2020; Shelton et al., 2021), including large-scale genome-wide association studies (GWASs) (COVID-19 Host Genetics Initiative, 2021; Pairo-Castineira et al., 2021) and rare variant approaches (Benetti et al., 2020; Kosmicki et al., 2021; Novelli et al., 2020; Wang et al., 2020b). However, the discovery power of those studies is limited, and the biological interpretation of those identified loci has been difficult, partially because of the confounding effects of patient age and comorbidities (Huang et al., 2020). To avoid a similar problem, we have integrated genetic signals with single-cell multiome profiling of human lungs so as to map cell-type-specific functions.

A primary cause of morbidity and mortality in COVID-19 is respiratory disease and specifically, a hyperinflammatory response within the lung that occurs in an age-independent manner (Brodin, 2021). This is the basis of a number of interventions based on immunosuppression (Mehta et al., 2020), which have repurposed treatments used for other diseases, particularly autoimmune diseases. Efficacy and side-effect profiles are likely to be improved by a COVID-19-specific immunomodulatory approach.

To understand the genomic basis of COVID-19 severity and gain insights into its molecular mechanisms, we sought to combine the genetic variation associated with severe COVID-19 together with single-cell-resolution functional profiling of human lungs. Using RefMap, a machine learning algorithm, we recently developed for genetic and epigenetic integration (Zhang et al., 2022), we identified over 1,000 genes associated with critical illness across 19 cell types, which account for 77% of SNP-based heritability for severe COVID-19; this represents a 5-fold increase over traditional approaches (COVID-19 Host Genetics Initiative, 2021). Analysis of single-cell transcriptomic profiling of respiratory tissues revealed downregulation of our risk genes in corresponding cell types in severe COVID-19 patients. Network analysis identified multiple protein-protein interaction (PPI) modules enriched with risk genes, unveiling additional cell-type-specific mechanisms underlying severe COVID-19. Heritability analysis and Mendalian randomization (MR) confirmed an important role for NK cells, specifically CD56^{bright} NK cells, in driving severe disease that extends previous literature (Maucourant et al., 2020) and adds causal inference. Rare variant analysis provided orthogonal evidence to further support the association of NK-cell risk genes with severe disease. Altogether, our study unravels a genomic landscape of COVID-19 severity and provides a better understanding of the disease pathogenesis, with potential for new prevention strategies and therapeutic targets.

RESULTS

Mapping cell-type-specific risk genes for severe COVID-19

We used the RefMap machine learning model (Zhang et al., 2022; STAR Methods) to identify the genomic regions and genes associated with severe COVID-19. Briefly, RefMap is a Bayesian network that combines genetic signals (e.g., SNP *Z* scores) with functional genomic profiling (e.g., ATAC-seq and ChIP-seq) to map risk regions associated with complex diseases. With RefMap, we can search the genome for functional regions within which disease-associated genetic variation is significantly shifted from the null distribution. This reduces the size of the search space and increases the statistical power. Rather than testing SNPs one by one, RefMap models the genetic architecture (i.e., all of the SNPs) of diseases using a unified probabilistic model that captures more complex genetic structures and avoids a multiple testing correction that would limit the statistical power. The power of the RefMap model for gene discovery and recovery of missing heritability has been demonstrated in our recent work (Zhang et al., 2022).

Here, to achieve cell-type-specific resolution within multicellular tissue, we modified RefMap and integrated single-cell multiome profiling of human lungs with COVID-19 GWAS data (Figure 1). In particular, we obtained summary statistics (COVID-19 Host Genetics Initiative [COVID19-hg], European [EUR], Release 5, phenotype definition A2; 5,101 cases versus 1,383,241 population controls) from the largest GWAS study of COVID-19 (COVID-19 Host Genetics Initiative, 2021), where age, sex, and 20 first principal components were included in the analysis as covariates. Severe COVID-19 was defined by the requirement for respiratory support or death attributed to COVID-19. Human lung single-cell multiome profiling, including single nucleus RNA sequencing (snRNA-seq) and single nucleus assay for transposase-accessible chromatin using sequencing (snATAC-seq), was retrieved from a recent study of healthy individuals (Wang et al., 2020a). Nineteen cell types were identified in both snATAC-seq and snRNA-seq profiles, including epithelial (alveolar type 1 [AT1], alveolar type 2 [AT2], club, ciliated, basal, and pulmonary neuroendocrine [PNEC]), mesenchymal (myofibroblast, pericyte, matrix fibroblast 1 [matrix fib. 1], and matrix fibroblast 2 [matrix fib. 2]), endothelial (arterial, lymphatic, capillary 1 [cap1], and capillary 2 [cap2]), and hematopoietic (macrophage, B cell, T cell, NK cell, and enucleated erythrocyte) cell types. We adopted these 19 cell types as the reference set within lung tissue throughout our study. Based on snATAC-seq peaks called in one or more of the 19 cell types, we used RefMap to identify disease-associated genomic regions from the COVID-19 GWAS summary data, resulting in 6,662 1-kb regions that passed the 5% significance threshold (referred to as RefMap COVID-19 regions hereafter; STAR Methods). By examining the intersection of identified regions with open chromatin in individual cell types based on corresponding snATAC-seq peaks, we derived cell-type-specific RefMap regions (mean per cell type = 1,733, standard deviation [SD] = 624; Figure 2A).

Next, we sought to map the target genes of RefMap COVID-19 regions in a cell-type-specific manner (Figure 1). In particular, we identified the regulatory targets that are expressed in the corresponding cell type for individual RefMap regions (STAR

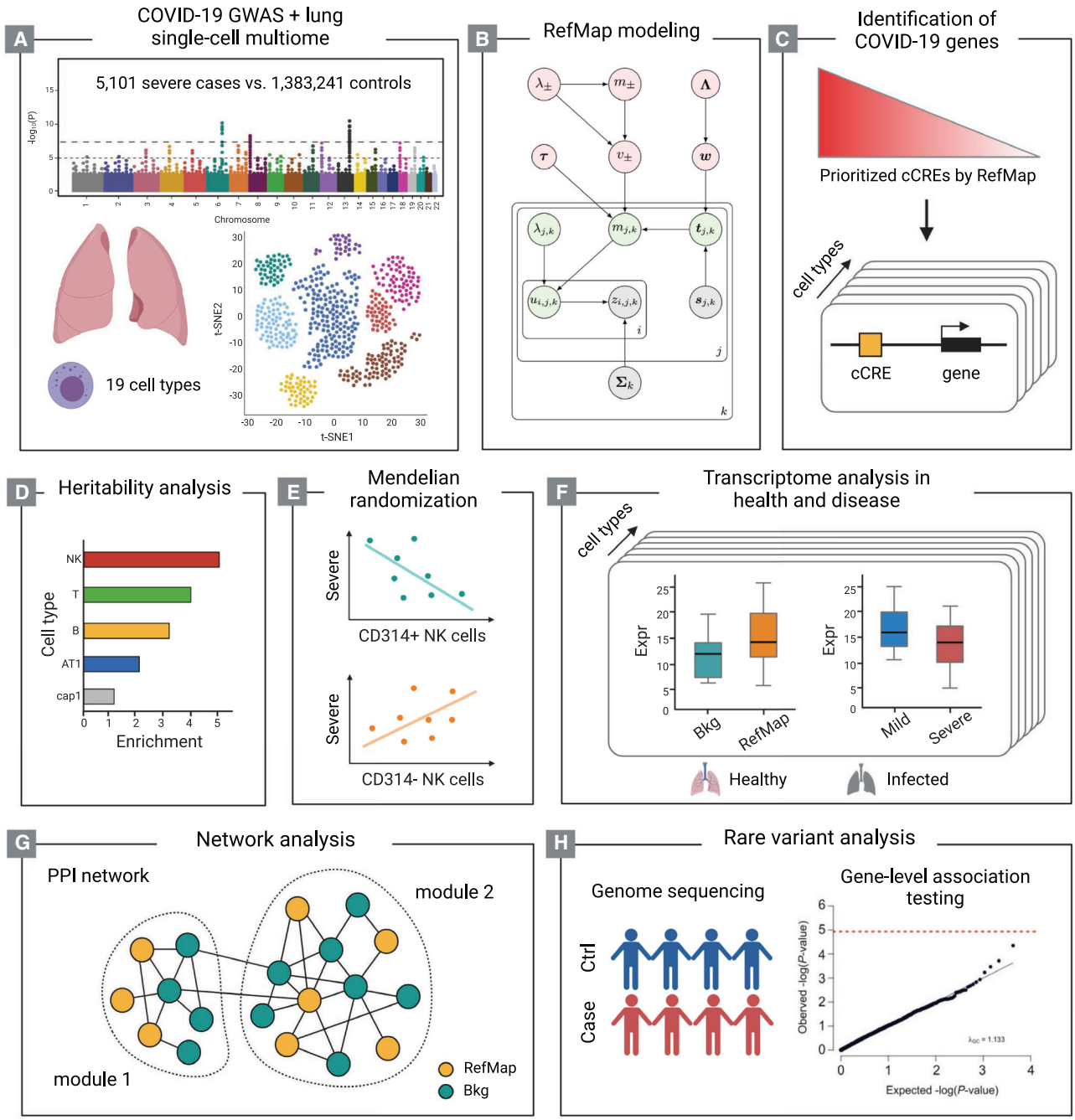


Figure 1. Schematic of the study design

(A-H) The COVID-19 GWAS and human lung single-cell multiome (A) are integrated by the RefMap model shown in (B), where gray nodes represent observations, green nodes are local hidden variables, and pink nodes indicate global hidden variables (STAR Methods). Cell-type-specific risk genes are mapped using single-cell multiome profiling (C). Heritability analysis (D), Mendelian randomization (E), transcriptome analysis (F), and network analysis (G) together characterize the functional importance of RefMap genes, particularly for NK cells, in severe COVID-19. Rare variant analysis (H) orthogonally supports the role of NK cells in severe disease. cCRE, candidate *cis*-regulatory element.

See also Table S1.

Methods). In total, we discovered 1,370 genes (referred to RefMap COVID-19 genes hereafter; mean per cell type = 280 and SD = 80; Figure 2B; Table S1) associated with the severe disease. Interestingly, hematopoietic cells have the largest number of unique RefMap regions and genes among all major cell types (Figure 2C); for example, there is a significant enrichment of unique RefMap regions observed for hematopoietic cells versus epithelial cells (p = 5.2e−3, odds ratio [OR] = 1.15, Fisher's exact

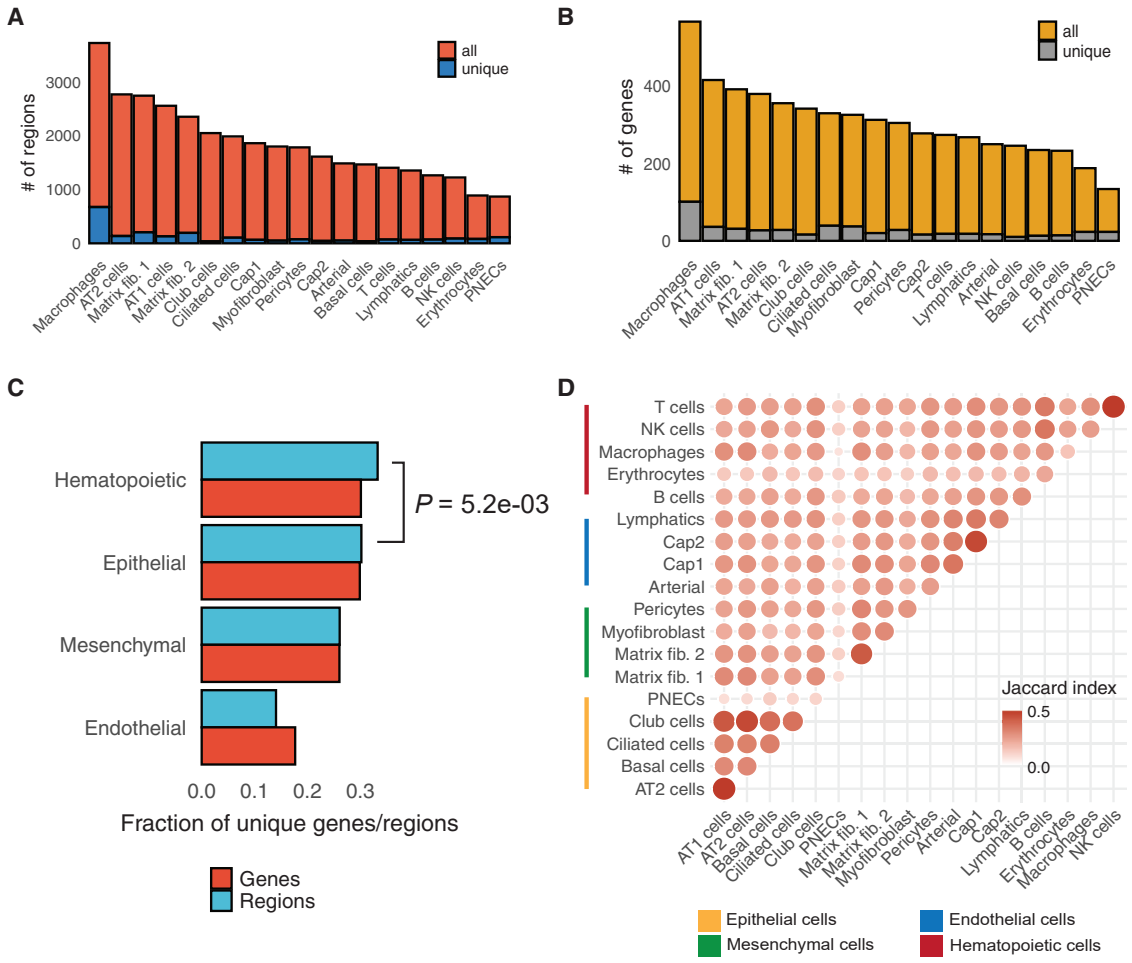


Figure 2. RefMap identifies cell-type-specific risk genes associated with severe COVID-19

(A) Total number and number of unique genomic regions containing genetic variation associated with severe COVID-19 for 19 different cell types.

(B) Total number and number of unique genes implicated by genetic variation associated with severe COVID-19 for 19 different cell types.

(C) Fraction of unique genomic regions and genes associated with severe COVID-19 for major cell types; statistical comparison of enrichment was determined by Fisher's exact test.

(D) Similarity between different cell types quantified by the overlap of RefMap genes. Gene set overlap was calculated by the Jaccard similarity index.

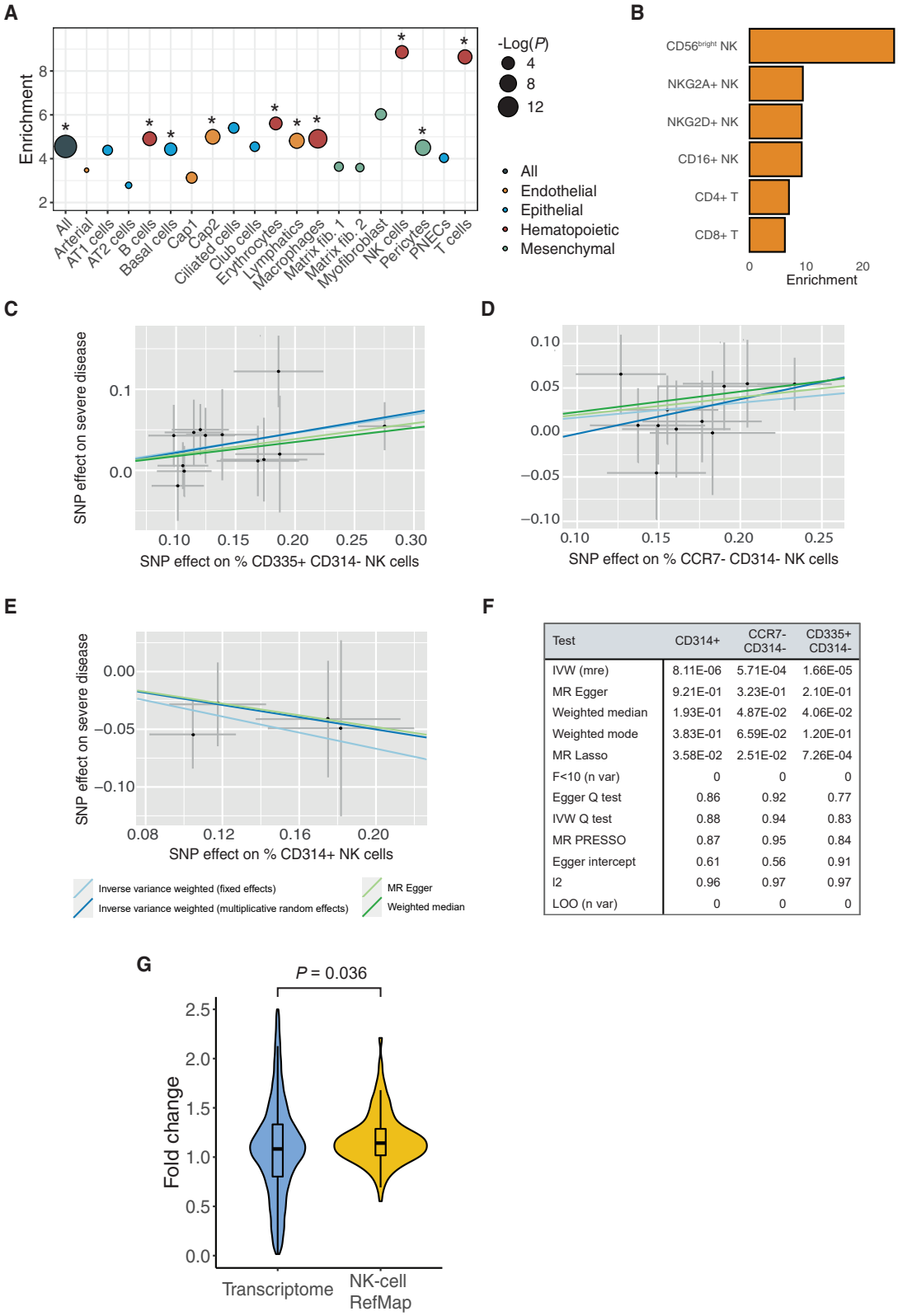
See also Table S1.

test; Figure 2C). To profile the cell-cell interactions underlying severe COVID-19 from a genetic perspective, we constructed a cell correlation matrix based on the overlap of RefMap genes between cell types (Figure 2D). We discovered that the correlation is strongest between functionally related cells, demonstrating that the RefMap signal is consistent with known biology (Wang et al., 2020a).

To replicate our findings, we obtained available summary statistics of SNPs associated with severe COVID-19 ($p < 1e{-}4$, n = 5,779) from a GWAS of an entirely independent cohort (the 23andMe cohort, 15,434 COVID-19-positive cases, and 1,035,598 population controls) (Shelton et al., 2021). The total union of RefMap regions is significantly enriched with SNPs associated with multiple COVID-19 phenotypes defined in this new dataset (mean p = 1.24e−3, Fisher's exact test; Table S2; STAR Methods). Specifically, the most significant enrichment is with SNPs associated with COVID-19 requiring respiratory support (mean p = 3.83e−4, mean OR = 9.52, mean standard error [SE] = 42.0, Fisher's exact test; Table S2). As further confirmation, we obtained GWAS summary statistics for a whole-genome sequencing (WGS) cohort of severe COVID-19 (the GenOMICC cohort; STAR Methods) (Kousathanas et al., 2021), including 7,491 severe COVID-19 patients and 48,400 population controls. Although this dataset overlaps the COVID19-hg cohort (EUR, Release 5, phenotype A2), 78% of the severe COVID-19 cases and 83% of the controls are distinct. Again, the total union of RefMap regions is significantly enriched with SNPs associated with severe COVID-19 in this dataset (mean p = 3.43e−4, mean OR = 10.29, mean SE = 48.0, Fisher's exact test; Table S2).

RefMap COVID-19 genes highlight known disease biology

The RefMap COVID-19 gene list contains known driver genes for severe COVID-19. For instance, a recent study mapped the



Test	CD314+	CCR7- CD314-	CD335+ CD314-
IVW (mre)	8.11E-06	5.71E-04	1.66E-05
MR Egger	9.21E-01	3.23E-01	2.10E-01
Weighted median	1.93E-01	4.87E-02	4.06E-02
Weighted mode	3.83E-01	6.59E-02	1.20E-01
MR Lasso	3.58E-02	2.51E-02	7.26E-04
F<10 (n var)	0	0	0
Egger Q test	0.86	0.92	0.77
IVW Q test	0.88	0.94	0.83
MR PRESSO	0.87	0.95	0.84
Egger intercept	0.61	0.56	0.91
I2	0.96	0.97	0.97
LOO (n var)	0	0	0

Figure 3. Severe-COVID-19-associated common variants are linked to multiple cell types

(A) Heritability enrichment estimated by LDSC for different cell types. Enrichment was calculated as the proportion of total SNP-based heritability adjusted for SNP number. The Benjamini-Hochberg (BH) procedure was used to calculate FDRs throughout the study. *: FDR < 0.1.

(legend continued on next page)

3p21.31 risk locus of severe COVID-19 (Pairo-Castineira et al., 2021) to altered expression of *LZTFL1* in lung epithelial cells and in ciliated epithelial cells (Downes et al., 2021). Extensive functional follow-up analysis demonstrated that the 3p21.31 risk genotype is associated with increased expression of *LZTFL1* by ciliated epithelial cells and inhibition of epithelial-mesenchymal transition (EMT). EMT is a component of the innate immune response that is thought to be beneficial in the context of SARS-CoV-2 infection via repair of damaged tissues and reduced expression of ACE2 and TMPRSS2, which may limit intracellular viral uptake (Downes et al., 2021). Consistent with this prior evidence, *LZTFL1* is also identified as a RefMap gene and is specifically associated with ciliated epithelial cells (Table S1).

In addition, excessive activity of ADAM9, a metalloprotease, has been associated with severe COVID-19 (Carapito et al., 2021). ADAM9 promotes the conversion of monocytes and macrophages to multinucleated giant cells as part of the immune response (Chou et al., 2020). Notably, *ADAM9* is identified as a RefMap gene expressed by cell types associated with the extracellular matrix (ECM), including matrix fibroblasts, myofibroblasts, and pericytes (Table S1). ADAM proteins, including ADAM9, are typically localized to the ECM (Chou et al., 2020), and hence, our finding is consistent with known biology.

Heritability enrichment within COVID-19 risk genes

Linkage disequilibrium score regression (LDSC) (Bulik-Sullivan et al., 2015) has been used to measure the total SNP-based heritability (h^2) of severe COVID-19 (COVID-19 Host Genetics Initiative, 2021). Here, we examined the partitioning of SNP-based heritability for severe COVID-19 within RefMap genes (STAR Methods). We discovered that the heritability of severe COVID-19 (COVID19-hg, EUR, Release 5, phenotype definition A2) is significantly enriched within RefMap genes (proportion of SNP-based heritability = 77%, OR = 4.6, SE = 0.78, p = 1.55e−7; Figure 3A; Table S3). We also compared RefMap with several widely used methods for genetic association study (STAR Methods), including naive GWAS (COVID-19 Host Genetics Initiative, 2021), MAGMA (Delorey et al., 2021b), PAINTOR (Kichaev et al., 2014), and Pascal (Lamparter et al., 2016), and observed superior performance of RefMap in terms of the proportion of recovered heritability (up to 5-fold increase; Table S3). The proportion of SNP-based heritability for hospitalized COVID-19 (COVID19-hg, EUR, Release 5, phenotype definition B2) within RefMap genes is 62% and for COVID-19 independent of severity (COVID19-hg, EUR, Release 5, phenotype definition C2) it is 52% (Table S3). In both cases, the improvement in captured heritability based on RefMap compared with traditional methods is up to 3-fold. Consistent with the design of our model, the recovered heritability is highest for severe COVID-19. As further confirmation, we calculated recovered heritability based on the GenOMICC GWAS. RefMap genes are highly enriched with heritability for severe COVID-19 in this replication cohort (proportion of SNP-based heritability = 50%, OR = 3.0, SE = 0.48, p = 5.12e−6; Table S3).

Prioritizing cell types by heritability partitioning

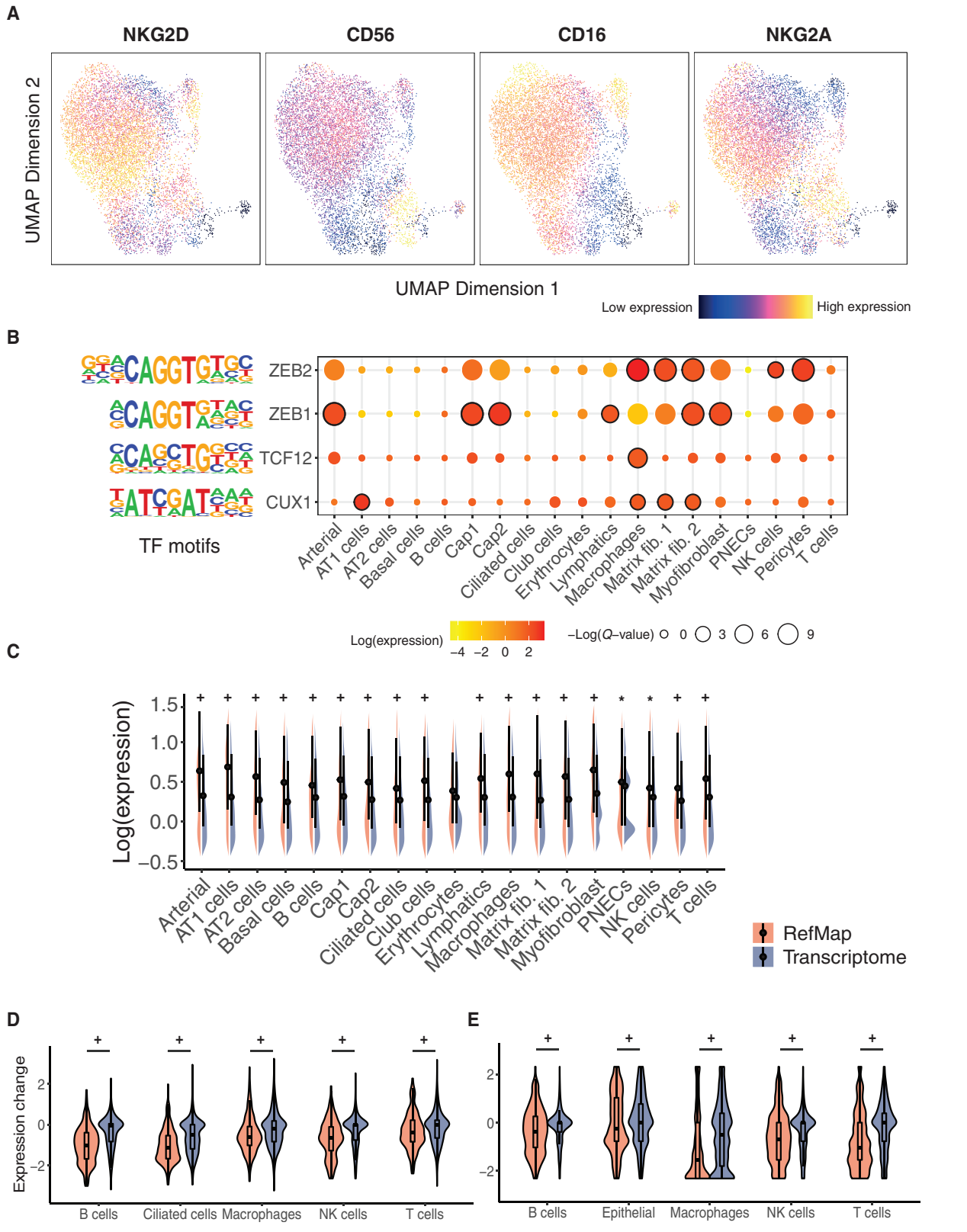
Next, we used cell-type-specific RefMap COVID-19 genes to determine which cell types are relatively more important in the development of severe COVID-19. Specifically, we calculated the partitioned heritability per cell type within the severe COVID-19 GWAS (A2) and also within GWAS for hospitalized versus non-hospitalized COVID-19 (B2) and COVID-19 versus population (C2) (STAR Methods). For severe COVID-19, of all 19 cell types tested, NK-cell and T cell genes are the most enriched with SNP-based heritability (NK cells: OR = 8.87, SE = 3.68, p = 0.016; T cells: OR = 8.64, SE = 3.28, p = 0.005; Figure 3A; Table S3). The same is also true for hospitalized COVID-19 (NK cells: OR = 10.57, SE = 4.95, p = 0.039; T cells: OR = 8.67, SE = 4.13, p = 0.041), but this enrichment is not statistically significant for COVID-19 irrespective of severity (NK cells: OR = 5.74, SE = 3.09, p = 0.077; T cells: OR = 4.16, SE = 2.56, p = 0.18; Table S3). In the GenOMICC GWAS, we obtained a similar result, whereby NK cells and T cells are the most enriched with SNP-based heritability for severe COVID-19 (NK cells: OR = 5.13, SE = 1.68, p = 0.01; T cells: OR = 4.28, SE = 1.32, p = 0.01; Table S3) compared with other cell types.

We extended our analysis of T cells by dividing them into $CD4^+$ and $CD8^+$ subtypes based on fluorescence-activated cell sorting (FACS) purified single cell RNA sequencing (scRNA-seq) profiling of the human lung (Travaglini et al., 2020). We first examined the relative expression of RefMap T cell genes within $CD4^+$ and $CD8^+$ T cells but discovered no significant difference between these two cell types (p = 0.34, two-tailed Wilcoxon rank-sum test). Furthermore, when we split RefMap T cell genes into $CD4^+$ and $CD8^+$ specific genes based on relative overexpression (fold change [FC] > 1.5 between subtypes; STAR Methods), we observed no significant difference in the heritability enrichment for severe COVID-19 between these two gene sets ($CD4^+$: OR = 6.95, $CD8^+$: OR = 6.23; Figure 3B; Table S3), suggesting that both subtypes may contribute to the immune response leading to severe disease. We observed similar results ($CD4^+$: OR = 5.40, $CD8^+$: OR = 4.66; Table S3) using the GenOMICC GWAS.

Dissecting NK-cell subtypes using MR and single-cell multiome profiling

NK cells have diverse biological functions, and therefore, we sought to understand which of these functions is a determinant

(B) Heritability enrichment for different subsets of T cells and NK cells.
(C–E) Significant Mendelian randomization results for three exposures linked to severe COVID-19, including blood counts of CD335+ CD314− (C), CCR7− CD314− (D), and CD314+ (E), NK cells; points indicate effect size (β) and standard errors for each SNP-outcome relationship.
(F) Sensitivity analyses and robust tests for MR analyses (STAR Methods).
(G) Comparative gene expression analysis of RefMap NK-cell genes in NKG2D+ and NKG2D− NK cells. Fold change was calculated as the ratio of gene expression levels in NKG2D+ NK cells to NKG2D− NK cells. The transcriptome was defined by all the expressed genes (with at least one unique molecular identifier [UMI]) in NK cells. Violin plots show the distributions of fold change values within each group, and boxplots indicate the median, interquartile range (IQR), Q1 − 1.5 × IQR, and Q3 + 1.5 × IQR. Distributions were compared by one-tailed Wilcoxon rank-sum test.
See also Figures S1 and S2 and Table S3.



(legend on next page)

of COVID-19 severity. Two-sample MR facilitates identification of a causal relationship between an exposure and an outcome (Smith, 2010). We examined whether NK-cell populations measured in the blood are causally related to severe COVID-19. In total, 26 GWAS measures of NK-cell subtypes were identified based on a previous study in which flow cytometry was used to quantify both counts of immune cell subsets and immune-cell-surface protein expression (Roederer et al., 2015; STAR Methods). After harmonizing exposure and outcome genetic instruments, we excluded tests with less than five SNPs that are likely to be underpowered (STAR Methods). With MR, three exposures were shown to be causally related to severe COVID-19 after correcting for multiple testing ($p < 2e{-}3$, multiplicative random effects [MREs] inverse-variance weighted [IVW] estimate). All three exposures relate to the proportion of NK cells expressing NKG2D/CD314 on the cell surface (Roederer et al., 2015), which is the major receptor responsible for NK-cell activation (Raulet, 2003). A higher proportion of NKG2D/CD314− cells is linked to severe COVID-19 (A2) (Figures 3C and 3D), but a higher proportion of NKG2D/CD314+ cells is protective (Figure 3E). Checks for genetic pleiotropy (MR Egger intercept not significantly different from zero, $p > 0.05$; Figure 3F) and instrument heterogeneity ($p > 0.05$, Cochran's Q test, and $I^2_{GX} > 0.95$; Figure 3F) were satisfactory. Moreover, robust measures are significant for all three exposures (Figure 3F). Furthermore, MR revealed that the proportion of NKG2D/CD314− cells is also causally associated with hospitalized COVID-19 (B2) and with COVID-19 independent of severity (C2) (Figure S1), but in each case, the effect size is reduced compared with severe COVID-19 (A2). To validate these findings, we repeated the MR analysis of NKG2D/CD314 cell count in the GenOMICC GWAS; again, a higher proportion of NKG2D/CD314− cells is linked to severe COVID-19 ($p = 0.035$; Figure S2).

Inspired by our MR analysis, we tested if the expression of RefMap NK-cell genes reflects a functional difference between NKG2D/CD314+ and NKG2D/CD314− cells. We examined gene expression using scRNA-seq data from healthy lungs (Travaglini et al., 2020); we discovered that RefMap NK-cell genes are expressed at a higher level within NKG2D/CD314+ cells compared with NKG2D/CD314− cells ($p = 0.036$, one-tailed Wilcoxon rank-sum test; Figure 3G). We conclude from these analyses that NK-cell activation via NKG2D/CD314 receptors may be protective against severe COVID-19. We note that the NKG2D/CD314 signaling pathway is not specific with respect to diverse NK-cell functions.

To achieve a finer characterization of NK-cell subtypes, we isolated NK cells from an uninfected human donor and performed single-cell multiome profiling (STAR Methods). NKG2D/CD314+ NK cells and various functional subtypes including $CD56^{bright}$, CD16+, and NKG2A+ NK cells were identified based on relative expression and open chromatin over the gene body and promoter (Figure 4A; STAR Methods). As expected, most NK cells are NKG2D/CD314+ (Figure 4A), and moreover, RefMap NK-cell genes that are relatively overexpressed in NKG2D/CD314+ compared with NKG2D/CD314− cells ($FC > 1.5$; STAR Methods) are not significantly enriched with heritability of severe COVID-19 compared with the total set of RefMap NK-cell genes (OR = 9.2 for NKG2D/CD314+ NK-cell genes versus OR = 8.9 for all NK-cell genes; Figure 3B; Table S3). The major functional subdivision of NK cells is between $CD56^{bright}$ NK cells, which are responsible for cytokine production and immunomodulation, and $CD56^{dim}$ NK cells, which are directly cytotoxic (Michel et al., 2016). A prominent mechanism for activation of cytotoxic NK cells is via CD16 ligand crosslinking, and therefore, it is expected that $CD56^{bright}$ NK cells are largely non-overlapping with CD16+ NK cells (Romee et al., 2013) as we observed (Figure 4A). Interestingly, RefMap NK-cell genes are relatively overexpressed in $CD56^{bright}$ cells compared with $CD56^{dim}$ cells ($FC = 1.3$, $p = 0.02$, Student's t test). Moreover, RefMap NK-cell genes overexpressed in $CD56^{bright}$ NK cells ($FC > 1.5$) are highly enriched with heritability for severe COVID-19 compared with the total set of RefMap NK-cell genes (OR = 24.5 for $CD56^{bright}$ NK-cell genes versus OR = 8.9 for all NK-cell genes; Figure 3B; Table S3). This result was replicated in the GenOMICC GWAS (OR = 16.5 for $CD56^{bright}$ NK-cell genes versus OR = 5.1 for all NK-cell genes; Table S3). This difference in heritability enrichment is statistically significant when considering the two datasets together ($FC = 2.93$, $p = 0.045$, Student's t test). To further develop this observation, we used MR to test whether NK-cell surface expression of CD56 (Roederer et al., 2015) is causally linked to severe COVID-19; the test was relatively underpowered at the chosen significance threshold (STAR Methods), but higher expression of CD56 was associated with lower risk of severe COVID-19 (MRE IVW $p = 0.01$). We conclude that severe COVID-19 may be associated with the function of $CD56^{bright}$ NK cells; both our MR analysis and transcriptome study suggest that reduced function of these

Figure 4. Transcriptomic signature of RefMap COVID-19 genes in different cell types

(A) UMAP of iterative latent semantic indexing (LSI) for combined gene expression and open chromatin over the gene body and promoter. Cells are colored by relative expression of NKG2D, CD56, CD16, and NKG2A, respectively. Expression is quantified as $\log_2$(normalized gene counts + 1); yellow/orange cells have relatively high expression of each marker.

(B) Enriched TF motifs of RefMap COVID-19 regions across 19 cell types. Relative enrichment per cell type is indicated by circle size and significant enrichment (HOMER, Q value < 0.1) is annotated with black circle; TF expression is indicated by color according to log(normalized gene counts + 1). Only highly expressed (expression level in the top 95% in corresponding cell types) TFs were considered.

(C) Gene expression analysis of RefMap genes across different cell types in healthy lungs. The transcriptome was defined as the total set of expressed genes for each cell type. Violin plots show the distributions of log expression levels within each group, and point plots indicate the median and IQR.

(D and E) Comparative gene expression analysis of cell-type-specific RefMap genes in severe COVID-19 patients versus moderately affected patients based on scRNA-seq datasets from Ren et al. (D) and Liao et al. (E), respectively. The *Z* score of Wilcoxon rank-sum test was used to indicate the change of gene expression between severe and moderate patient groups, where a positive value means higher gene expression in severe patients. Violin plots show the distribution of gene expression changes within each group, and boxplots indicate the median, IQR, Q1 − 1.5 × IQR, and Q3 + 1.5 × IQR. *: $0.01 \leq FDR < 0.1$. +: $FDR < 0.01$. Distributions were compared by one-tailed Wilcoxon rank-sum test.

See also Figure S3.

cells may be an upstream cause of SARS-CoV-2 infection. The observed heritability enrichment of severe COVID-19 associated with CD56^{bright} NK cells is significantly larger than that of any other profiled cell type including subtypes of T cells (Figure 3B). This has functional implications because CD56^{bright} cells are responsible for production of IFN-γ and other immunomodulatory cytokines important for the innate immune response (He et al., 2004). Our interpretation is that deficiency of the NK-cell-derived IFN-γ defense may precipitate uncontrolled viral replication.

Unlike CD56^{bright} NK cells, RefMap NK-cell genes overexpressed in CD16+ NK cells (FC > 1.5) are not significantly enriched with heritability of severe COVID-19 compared with NK-cell genes overall (OR = 9.4 for CD16+ NK-cell genes versus OR = 8.9 for all NK-cell genes; Figure 3B; Table S3). We considered one additional functional subtype of NK cells: NKG2A+ NK cells. A preponderance NKG2A+ NK cells has been associated with reduced diversity and potency of both the cytotoxic and chemokine producing arms of the NK-cell response linked to genetic variation within the HLA-B gene (Horowitz et al., 2016). NK-cell genes overexpressed in NKG2A+ NK cells are not significantly enriched with heritability for severe COVID-19 compared with NK-cell genes overall (OR = 9.7 for NKG2A+ NK-cell genes versus OR = 8.9 for all NK-cell genes; Figure 3B; Table S3). We conclude that this may not be an important distinction driving severe COVID-19. This is consistent with fine-mapping studies of the HLA locus that have not revealed genetic variation significantly linked to severe COVID-19 (Degenhardt et al., 2021), and *in vitro* studies showing that HLA blockade of NK cells does not alter their capacity for control of SARS-CoV-2 replication (Witkowski et al., 2021).

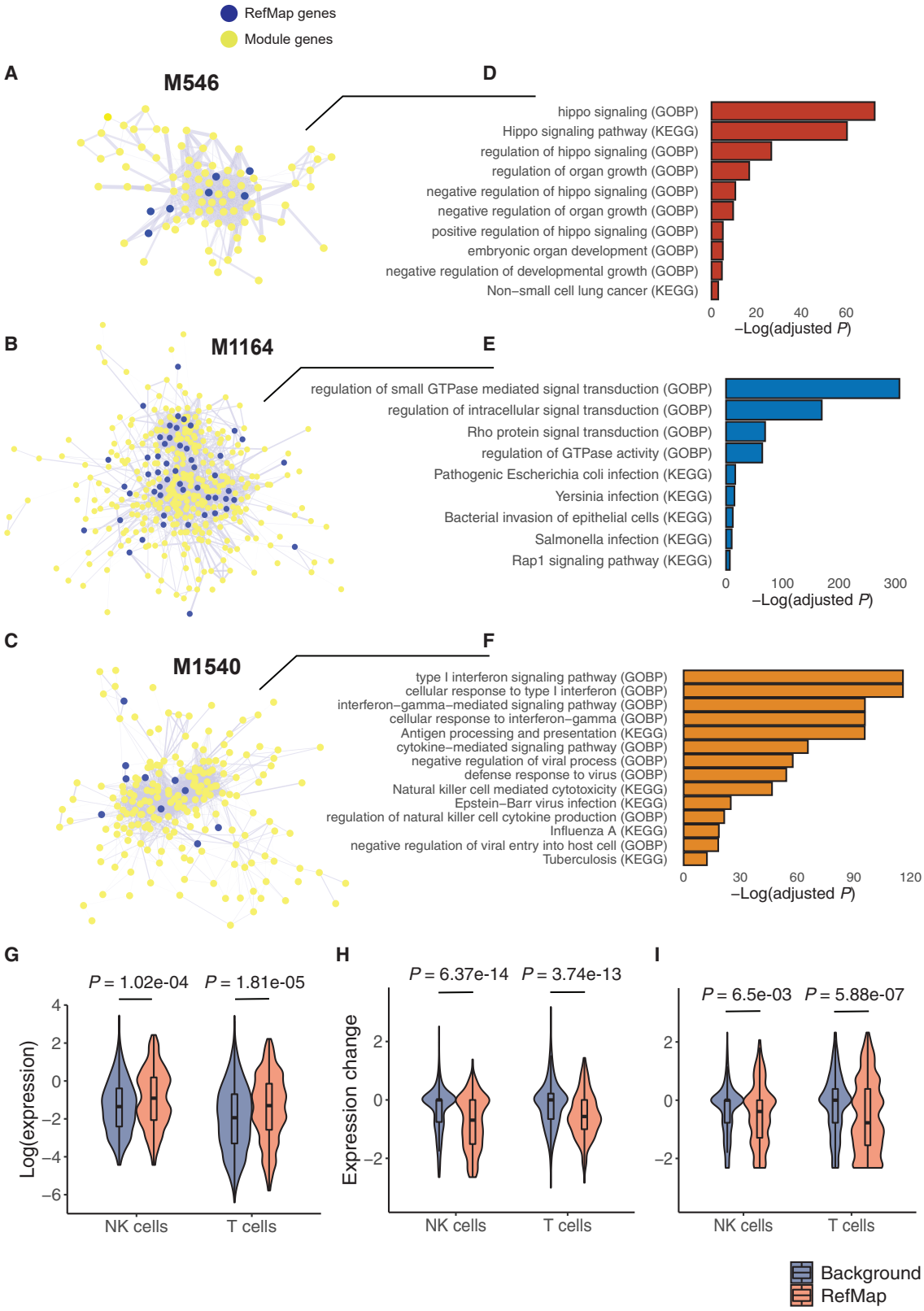
Functional profiling of severe COVID-19 risk genes

We sought to systematically characterize the functional roles of RefMap COVID-19 regions and genes across the 19 cell types in both healthy and diseased contexts. We first performed motif enrichment analysis (Heinz et al., 2010) (STAR Methods) for RefMap regions, which identified four transcription factors (TFs), including CUX1, TCF12, ZEB1, and ZEB2, whose binding motifs are enriched in at least one of 19 cell types (HOMER, Q value < 0.1 and expression percentile >95; Figure 4B). Interestingly, although T cell and NK-cell RefMap gene lists were equally enriched with heritability for severe COVID-19 (Figure 3A), only NK-cell risk regions were enriched with a TF binding motif: ZEB2. ZEB2 is an essential driver of NK-cell maturation (van Helden et al., 2015), and immature NK cells from ZEB2-null mice perform deficient immunosurveillance *in vivo* even when *in vitro* functions are maintained (Bi and Wang, 2020). We conclude that deficient NK-cell function leading to severe COVID-19 may be a result of failed maturation.

Next, we performed functional enrichment analysis for RefMap genes using Enrichr (Kuleshov et al., 2016; Tables S4 and S5). We observed that RefMap NK-cell genes are enriched with pathways and gene ontology (GO) terms related to intra- and inter-cellular signaling important for NK-cell activation, including "Phospholipase D signaling pathway" (Balboa et al., 1992), "Antigen processing and presentation," "regulation of small-GTPase-mediated signal transduction" (GO:0051056) (Watzl and Long, 2010), and "regulation of intracellular signal transduction" (GO:1902531) (adjusted p < 0.1; Tables S4 and S5). This is consistent with the hypothesis that COVID-19 severity is determined by failed activation of NK cells. Other cell-type-specific RefMap gene lists are also enriched with relevant biological pathways. For example, AT2-cell genes are linked to pathways associated with viral infection such as "human papillomavirus infection" and "viral carcinogenesis" (adjusted p < 0.1; Table S5), which is consistent with the established role of AT2 cells as the initial site of SARS-CoV-2 entry into host cells (Hoffmann et al., 2020). T cell genes are enriched with "IL-17-signaling pathway" (adjusted p = 0.021; Table S5), which is interesting in light of previous literature highlighting the production of IL-17 by T cells from COVID-19 patients as a potential therapeutic target (De Biasi et al., 2020).

We investigated the baseline expression pattern of RefMap genes in healthy lungs. In particular, we calculated mean expression levels of genes in different cell types based on lung snRNA-seq data from Wang et al., 2020a), and then compared the expression of RefMap genes with the total set of expressed genes in each cell type. Interestingly, although the gene expression level was not an input to the RefMap model, RefMap genes are expressed at a higher level compared with expressed genes in all 19 cell types, including immune and epithelial cells (false discovery rate [FDR] < 0.1, one-tailed Wilcoxon rank-sum test; Figure 4C) but with the exception of pericytes (FDR = 0.11, *Z* score = 1.25); notably, pericytes may be downstream in the pathogenesis of COVID-19 because they are protected by an endothelial barrier (He et al., 2020). This supports the functional significance of RefMap genes across multiple cell types in healthy human lungs. As a negative control, we performed a similar expression comparison between non-developmental genes and all expressed genes in lungs, which yielded no significant difference (Figure S3; STAR Methods).

Finally, given that RefMap COVID-19 genes were identified via their associated regulatory elements (i.e., snATAC-seq peaks), we examined whether there is any expression change for RefMap genes in the context of SARS-CoV-2 infection. We obtained scRNA-seq data from the respiratory system for a large COVID-19 cohort (Ren et al., 2021), including 12 bronchoalveolar lavage fluid (BALF) samples, 22 sputum samples, and 1 sample of pleural fluid mononuclear cells (PFMCs) from 27 severely and 8 mildly affected patients. Severity was classified based on the World Health Organization (WHO) guidelines (https://www.who.int/publications/i/item/WHO-2019-nCoV-clinical-2021-1). For individual cell types, we compared the expression level of RefMap genes in severe patients versus moderately affected patients (STAR Methods). Compared with the background transcriptome, we observed that RefMap genes are expressed at a lower level in corresponding cell types from severe patients compared with moderate patients (FDR < 0.01, one-tailed Wilcoxon rank-sum test; Figure 4D), supporting the functional significance of RepMap genes in severe COVID-19. As a replication experiment, we carried out a similar analysis based on an independent COVID-19 scRNA-seq dataset (Liao et al., 2020), including 9 BALF samples from 6 severe patients and 3 moderate patients (STAR Methods). The lower expression of RefMap genes in severe patients is consistent across multiple cell types (FDR < 0.01, one-tailed Wilcoxon rank-sum test; Figure 4E). Altogether, these transcriptome-based orthogonal analyses are



(legend on next page)

consistent with the hypothesis that identified cell-type-specific RefMap genes are functionally linked to COVID-19 severity.

Identification of gene modules associated with severe COVID-19

To delineate cell-type-specific molecular mechanisms underlying severe COVID-19, we next mapped RefMap COVID-19 genes to the global PPI network and then inspected functional enrichment of COVID-19-associated network modules. In particular, we extracted high-confidence (combined score >700) PPIs from STRING v11.0 (Szklarczyk et al., 2019), which include 17,161 proteins and 839,522 protein interactions. To eliminate the bias of hub genes (Krishnan et al., 2016), we performed the random walk with restart algorithm over the raw PPI network to construct a smoothed network based on edges with weights in the top 5% (STAR Methods). Next, this smoothed PPI network was decomposed into non-overlapping subnetworks using the Leiden algorithm (Traag et al., 2019). This process yielded 1,681 different modules (Table S6) in which genes within modules are densely connected but sparsely connected with genes in other modules.

Six modules including M62 (n = 370; mesenchymal cells), M148 (n = 364; mesenchymal cells), M546 (n = 90; epithelial cells), M750 (n = 281; endothelial and mesenchymal cells), M1164 (n = 396; endothelial, epithelial, hematopoietic, and mesenchymal cells), and M1540 (n = 226; hematopoietic cells) are significantly enriched with at least one cell-type-specific RefMap gene list (FDR <0.1, hypergeometric test; Figures 5A–5C; Table S6). In particular, RefMap genes specific to ciliated epithelial cells are enriched in module M546 (FDR < 0.1, hypergeometric test; Figure 5A; Table S6), which is enriched with biological functions including "hippo signaling" (GO:0035329) (adjusted *p<0.1*; Figure 5D). Hippo signaling is involved in the EMT response (Lei et al., 2008), and hence, this result is consistent literature linking the RefMap risk gene *LZTFL1* to severe COVID-19 via increased expression within ciliated epithelial cells and reduced EMT. RefMap genes expressed by several epithelial cell types including AT1, AT2, basal, and ciliated cells are enriched within module M1164 (FDR < 0.1, hypergeometric test; Figure 5B; Table S6), which is linked to infection (e.g., "Bacterial invasion of epithelial cells") and intracellular signaling (e.g., "regulation of small-GTPase-mediated signal transduction" (GO:0051056) and "Rho protein signal transduction" (GO:0007266)) (adjusted p < 0.1; Figure 5E). Moreover, NK-cell and T cell RefMap genes are specifically enriched within module M1540 (FDR < 0.1, hypergeometric test; Figure 5C; Table S6), which is linked to interferon signaling (e.g., "interferon-gamma-mediated signaling pathway" (GO:0060333) and "positive regulation of NK-cell cytokine production" (GO:0002729)) (adjusted p < 0.1; Figure 5F). Interestingly, M1540 is also enriched with gene expression linked to CD56+ NK cells (adjusted p < 0.1, Human Gene Atlas).

To further characterize the function of identified modules, we investigated the expression patterns of module genes based on scRNA-seq profiling of healthy and diseased tissues. In particular, genes in module M1540 are relatively overexpressed in NK and T cells of healthy lungs (Wang et al., 2020a) compared with the background transcriptome (p = 1.02e−4 and p = 1.81e−5, respectively, one-tailed Wilcoxon rank-sum test; Figure 5G). In contrast, in respiratory samples infected with SARS-CoV-2, we observed a downregulation of M1540 genes in NK and T cells in patients suffering severe disease (Ren et al., 2021) (p = 6.37e−14 and p = 3.74e−13, respectively, one-tailed Wilcoxon rank-sum test; Figure 5H). This observation was replicated in another cohort (Liao et al., 2020) (p = 6.5e−3 and p = 5.88e−7 for NK and T cells, respectively, one-tailed Wilcoxon rank-sum test; Figure 5I). These results together are consistent with our previous findings, suggesting that interferon signaling by NK cells can prevent severe COVID-19.

Rare variant association analysis of COVID-19 risk genes

To verify the set of RefMap genes, we utilized a rare variant analysis that is orthogonal to our methods described to this point, which rely on common variant analysis. Rare variants were used to construct gene-level mutation burden in a large meta-analysis, including 4,964 severe COVID-19 patients and 570,461 population controls (STAR Methods). Rare variants were identified by minor allele frequency (MAF) < 1% (STAR Methods). Only loss-of-function (LoF) mutations were considered, including nonsense mutations, frameshift mutations, and splice-site mutations (STAR Methods). A random-effect (Dersimonian-Laird) meta-analysis was performed to combine results from multiple independent cohorts. Of NK-cell RefMap genes, 225 are present within this dataset where 18 are significantly enriched (p < 0.05, REGENIE; Mbatchou et al., 2021) with rare variants associated with severe COVID-19. This enrichment is statistically significant (FDR < 0.1, permutation test; Figures 6A and 6B; STAR Methods). These 18 genes are significantly enriched with biological functions including "negative regulation of cell adhesion" (GO:0007162) (adjusted p = 2.1e−4, OR = 82.5), which is consistent with a role in recruitment of NK cells to an area of infection. One of these 18 genes is *APOBEC3G* (Figure 6B), which is a cytidine deaminase implicated in the immune defense against *Coronaviridae* (Milewska et al., 2018; Wang and Wang, 2009) and in enhancement of NK-cell antiviral function

Figure 5. PPI modules enriched with COVID-19 genes and their functional characterization

(A–C) Three PPI network modules, including M546 (A), M1164 (B), and M1540 (C), are significantly enriched with ciliated-cell gene, epithelial-cell genes, and immune-cell genes, respectively. Blue nodes represent RefMap COVID-19 genes and yellow nodes indicate other genes within each module. Edge thickness is proportional to STRING confidence score (>400).

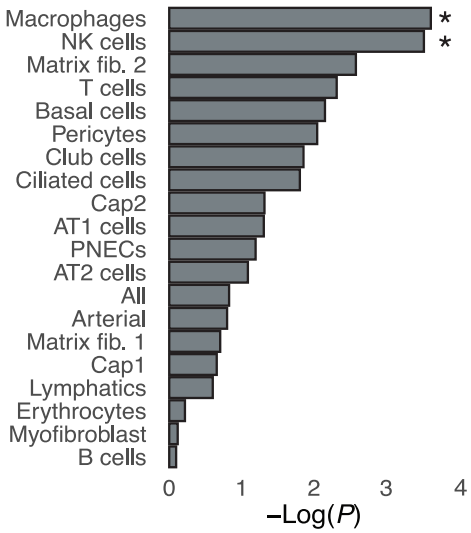
(D–F) Gene functions that are significantly enriched (Fisher's exact test, adjusted p < 0.1) in modules M546 (D), M1164 (E), and M1540 (F). GOBP, GO biological process.

(G) Gene expression analysis of module genes in NK and T cells. The transcriptome was defined as the total set of expressed genes in NK and T cells, respectively.

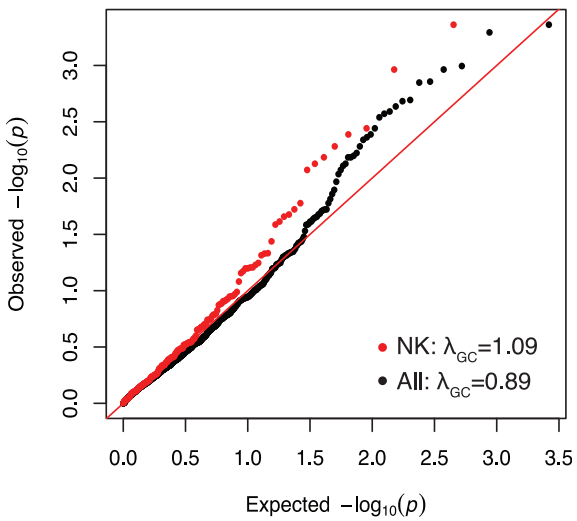
(H and I) Comparative gene expression analysis of module genes in severe COVID-19 patients versus moderate patients based on scRNA-seq datasets from Ren et al. (H) and Liao et al. (I), respectively. The *Z* score of Wilcoxon rank-sum test was used to indicate the change of gene expression between severe and moderate patient groups. Violin plots show the distribution of gene expression changes within each group, and boxplots indicate the median, IQR, Q1 − 1.5 × IQR, and Q3 + 1.5 × IQR. Distributions were compared by one-tailed Wilcoxon rank-sum test.

See also Table S6.

A



C



B

Gene	Effect	SE	*P*-value	Cases	Controls	MAC
APOBEC3G	1.142597	0.3980405	0.004098	4166	568193	837
CMTM7	4.3951113	1.6693336	0.008467	1562	530113	70
FNDC3B	3.3909267	1.5220001	0.02588	1685	530607	52
GRAP2	1.8907092	0.7901131	0.01671	3719	552392	174
HDAC4	2.296866	1.1535357	0.04646	1763	549913	52
HIVEP3	1.8490742	0.5255951	0.0004347	2309	560204	339
IL17RA	1.4887629	0.634454	0.01895	3905	566000	254
NF2	1.0959715	0.5513719	0.04684	1749	558569	731
NOTCH1	2.1166519	0.9240593	0.02199	1999	545559	152
PITPNB	3.3541076	1.4894145	0.02432	1749	558569	74
PLXND1	1.667027	0.5102646	0.001087	3779	553498	307
PTBP2	3.0119456	1.4398726	0.03646	1638	517166	25
RBM6	-2.7524372	1.1931974	0.02107	3069	509201	93
RHOA	3.6892732	1.3212193	0.005233	1263	400410	13
SLC20A2	2.9348089	1.0968631	0.007459	1800	547954	77
TBC1D2B	1.7838432	0.904058	0.04848	3748	564773	182
TNFRSF1B	2.5571908	0.8787731	0.003615	3232	559779	63
ZNF638	1.5041852	0.5531376	0.006541	3957	566509	808

Figure 6. Rare variant analysis supports the association of NK cells with severe COVID-19

(A) Enrichment analysis of cell-type-specific RefMap COVID-19 genes based on rare variant associations. *: Q value < 0.1, permutation test.

(B) Overlapping genes between the RefMap NK-cell gene set and the rare-variant-associated gene set ($p < 0.05$, REGENIE). MAC, minor allele count.

(C) Q-Q plot of rare variant association test for RefMap NK-cell genes and all RefMap genes, including expected p values based on the null hypothesis and observed p values by REGENIE.

See also Figure S4 and Table S7.

(Norman et al., 2011). Moreover, the mean p value and median p value of NK-cell RefMap genes in the rare variant test are significantly lower than expected by chance (p = 0.02 and p = 0.03, respectively, permutation test; STAR Methods). The enrichment of rare variants associated with severe COVID-19 within NK-cell RefMap genes is greater than for RefMap genes overall (λ_{GC} = 1.09 for NK-cell genes versus λ_{GC} = 0.89 for all RefMap genes; Figure 6C; STAR Methods) or any other cell-type-specific RefMap gene list, again, consistent with a role of NK cells in severe COVID-19.

Similar enrichment with rare variants associated with severe COVID-19 was observed for macrophage RefMap genes, which supports a role for this cell type in susceptibility to COVID-19; this idea has been proposed previously by others (Delorey et al., 2021a). Of 450 macrophage RefMap genes, 32 are significantly enriched (p < 0.05, REGENIE) with rare variants associated with severe COVID-19 (FDR<0.1, permutation test; Figure 6A; Table S7; STAR Methods). These 32 genes are enriched with biological functions including "response to chemokine" (GO:1990868) (adjusted p = 0.02, OR = 111), which is consistent with a role in immune cell recruitment. Overall, our data suggest that rare LoF genetic variation associated with severe COVID-19 selectively impairs migration of immune cells to the site of viral entry, in accordance with previous literature (Hussman, 2020).

DISCUSSION

The COVID-19 pandemic is a global health crisis (Dong et al., 2020). Vaccination efforts have led to early successes (Shilo et al., 2021), but the prospect of evolving variants capable of immune-escape (Darby and Hiscox, 2021) highlights the importance of efforts to better understand the COVID-19 pathogenesis and to develop effective treatments. Host genetic determinants of disease severity have been investigated (Severe Covid-19 GWAS Group et al., 2020; Benetti et al., 2020; COVID-19 Host Genetics Initiative, 2021; Kosmicki et al., 2021; Novelli et al., 2020; Pairo-Castineira et al., 2021; Shelton et al., 2021; Wang et al., 2020b), but the findings and functional interpretations so far have been limited (Huang et al., 2020). In contrast, studies of the immune response accompanying severe COVID-19 (Arunachalam et al., 2020; Lucas et al., 2020; Mathew et al., 2020; Sosa-Hernández et al., 2020) have struggled to establish causality leading to a diverse array of candidates and little consensus. Our contributions are an integrated analysis of common and rare host genetic variation causally linked to severe COVID-19, together with biological interpretations via single-cell omics profiling of lung tissue and identification of >1,000 risk genes explaining the majority of SNP-based heritability of severe COVID-19.

Our RefMap analysis, which integrates GWAS summary statistics with epigenetic profiles (Zhang et al., 2022), uncovers a landscape of cellular dysfunction within lung tissue leading to severe COVID-19. Our findings are consistent with previous literature, but our focus on host genetics allows us to make conclusions about upstream causation, which has been missing from previous studies. In particular, we highlight the failure of cytokine production by CD56^{bright} NK cells. Our MR analyses revealed that genetic predisposition to lower counts of mature NK cells is associated with increased risk of severe COVID-19 and NK-cell-specific RefMap risk genes are enriched for binding motifs for ZEB2, a TF involved in NK-cell maturation. Our rare variant analysis revealed that LoF variants, which impair NK-cell function and potentially NK-cell recruitment to a site of infection, increase risk of severe COVID-19. Interestingly, we showed that LoF variants within *APOBEC3G* increase the risk of severe COVID-19, which is consistent with previous literature linking this protein to the NK-mediated immune response to *Coronaviridae* (Milewska et al., 2018; Norman et al., 2011; Wang and Wang, 2009).

Previous work has implicated both the IFN-γ response and the importance of CD56^{bright} NK cell function in development of severe COVID-19. Our contribution is that we have arrived at this result through study of host genetics, which are fixed at conception and so necessarily upstream of SARS-CoV-2 infection. This also helps to explain an apparent contradiction between our results that predict a failure of the IFN-γ response in severe COVID-19 and studies that have positively correlated serum IFN-γ with COVID-19 mortality (Gadotti et al., 2020). In a purely observational study, it is impossible to distinguish cause and effect, and individuals with mild disease will likely have a lower viral load and therefore an attenuated IFN-γ response. The key comparison is the capacity of NK cells to produce a suitable immune response to SARS-CoV-2 at the earliest stages of infection or even before infection has actually occurred. Given that experimentally introducing viruses to healthy individuals is not feasible, we believe that our genetics-based approach is an optimal method to determine which events are upstream and truly causal in the development of severe COVID-19. Other observational studies are entirely consistent with our own observations; for example, NK-cell counts at the time of admission predict the rate of decline in viral load (Witkowski et al., 2021), which is exactly what we would predict from our MR results. The same study observed a relationship between low blood counts of CD56^{bright} NK cells in the first week after COVID-19 symptom onset and increased likelihood of severe disease, which directly mirrors our own conclusions regarding this NK-cell subtype. Another study of immune cell profiles during active infection observed a relative contraction of NKG2D+ and CD56^{bright} NK-cell counts in patients with poor COVID-19 outcomes (Varchetta et al., 2021). These studies are of course important, but our work now shows that these findings are dependent on host factors and not determined by viral properties. Ultimately, this could enable prediction of COVID-19 risk in uninfected individuals. In the cancer field, NK-cell stimulation has been postulated as a therapeutic strategy (Hu et al., 2019). We propose that this strategy could protect at-risk individuals in future waves of COVID-19.

We highlight the importance of T cells and NK cells in particular. Other immune cell types, such as B cells, carry less genetic risk for severe COVID-19. This is consistent with previous literature demonstrating robust immune responses to infection with SARS-CoV-2, even in patients with dramatic B cell depletion due to hematological malignancy (Bange et al., 2021).

Our study has several limitations. Our genetic discovery data are largely focused on European ancestry, which may limit widespread applicability. In part, this was because we did not have access to individual-level data, and therefore, we necessarily performed out-sample LD estimation, which can lead to false

positive results (Benner et al., 2017). To mitigate this, both our transcriptome analyses and rare variant burden testing encompassed several population groups. We also analyzed the gene intolerance to LoF variants using probability of loss of function intolerance (pLI) and observed/expected ratio (o/e) scores (Karczewski et al., 2020; Lek et al., 2016), which measure the excess of LoF variants for individual genes compared with the genome-wide expected baseline. The pLI score and o/e scores are calculated based on ExAC and gnomAD exome cohorts, respectively, which involve multiple populations. Interestingly, we observed that our RefMap NK-cell genes have significantly larger pLI scores and smaller o/e scores compared with the total set of protein-coding genes (pLI: $p = 3.16e{-}16$, Wilcoxon rank-sum test; o/e: $p = 1.33e{-}16$, Wilcoxon rank-sum test; Figure S5), suggesting that the RefMap NK-cell genes are intolerant to LoF mutations, regardless of populations and demonstrating their functional significance. The validation of our genes in orthogonal, and independent data from diverse ancestries indicate the population transposability of our findings. In addition, in the absence of tissue-specific QTL data, we mapped risk genes from regulatory regions using the "closest-gene" method, which has been widely used in the absence of 3D genome profiling (e.g., Hi-C) (Buniello et al., 2019; Mountjoy et al., 2020). Indeed, in the context of mapping disease risk loci, assigning the *cis*-regulatory elements (CREs) to their closest neighbors yields comparable results compared with more advanced methods (Nasser et al., 2021). Using decaying weights (e.g., exponentially) based on the distance from CREs to the transcription start sites (TSSs) in predicting gene expression also gives competitive results (Zhou et al., 2018). These previous studies indicate that the closest-gene method is a reasonable method when lacking other orthogonal molecular profiling. Next, our analysis did not have sufficient resolution to detect all potentially relevant cellular subtypes, and some rare cell types, such as ionocytes (Montoro et al., 2018), went completely undetected in the profiling data we used.

Finally, we have not performed any *in vitro* experimental follow-up studies; we expect that cultured NK cells derived from individuals carrying an at-risk genotype would demonstrate reduced expression of RefMap COVID-19 genes and reduced ability to control viral replication in co-culture with an appropriate cell type, such as Calu3 cells. We note that use of this system by others confirmed our prediction that NKG2D signaling is necessary for NK cell control of viral replication (Witkowski et al., 2021). However, there are limitations of such *in vitro* systems that fail to capture all the interactions between the numerous cell types present *in vivo*. A recent study revealed that manipulating the activation of NK cells in a mouse model resulted in a significantly higher viral burden (Wang et al., 2021), but there are limitations to animal models that do not necessarily translate to human disease. We suggest that our genetic profile could be used to guide a prospective study and even a protective intervention trial in human patients.

In conclusion, we have uncovered a genetic architecture of severe COVID-19 integrated with single cell-resolution biological functions. Both common and rare variant analyses have highlighted NK-cell activation as a key determinant of disease severity. Our framework can be applied to decipher the genetic and biological basis of other complex diseases.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - The RefMap model
  - Mapping cell-type-specific genes from RefMap regions
  - Validation of RefMap results in 23andMe and GenOMICC datasets
  - Heritability analysis
  - Implementation details of MAGMA, PAINTOR and Pascal
  - Mendelian randomization
  - Single-cell multiome profiling of NK cells
  - Deriving cell-type-specific genes for T-cell and NK-cell subtypes
  - Motif enrichment analysis
  - Transcriptome analysis
  - Network analysis
  - Rare-variant burden testing

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2022.05.007.

ACKNOWLEDGMENTS

We acknowledge the Stanford Genetics Bioinformatics Service Center (GBSC) for providing computational infrastructure for this study. This study was supported by the National Institutes of Health (1S10OD023452-01 to GBSC; CEGS 5P50HG00773504, 1P50HL083800, 1R01HL101388, 1R01-HL122939, S10OD025212, P30DK116074, and UM1HG009442 to M.P.S.) and the Wellcome Trust (216596/Z/19/Z to J.C.-K.). Figure 1 was created with BioRender.com. We thank the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/) for releasing the summary statistics data of GWAS and rare variant association study. We also thank GenOMICC (https://genomicc.org/) for sharing the GWAS summary statistics data to us.

AUTHOR CONTRIBUTIONS

S.Z., J.C.-K., P.S.T., and M.P.S. conceived and designed the study. S.Z., J.C.-K., A.K.W., M.S., L.K., D.U., C.H., T.H.J., J.K.B., P.S.T., and M.P.S. were responsible for data acquisition. S.Z., J.C.-K., C.H., T.H.J., and J.K.B. were responsible for data analysis. S.Z., J.C.-K., A.K.W., C.H., T.H.J., S.F., E.F., F.F., A.R., P.G., X.S., I.S.T., K.P.K., J.K.B., M.M.D., P.S.T., and M.P.S. were responsible for the interpretation of the findings. S.Z., J.C.-K., P.S.T., and M.P.S. drafted the manuscript with assistance from all authors. All authors meet the four ICMJE authorship criteria and were responsible for revising the manuscript, approving the final version for publication, and for accuracy and integrity of the work.

DECLARATION OF INTERESTS

M.P.S. is a co-founder and member of the scientific advisory board of Personalis, Qbio, January, SensOmics, Protos, Mirvie, NiMo, Onza, and Oralome. He

is also on the scientific advisory board of Danaher, Genapsys, and Jupiter. No other authors have competing interests.

Received: December 8, 2021
Revised: April 2, 2022
Accepted: May 18, 2022
Published: June 14, 2022

REFERENCES

Arunachalam, P.S., Wimmers, F., Mok, C.K.P., Perera, R.A.P.M., Scott, M., Hagan, T., Sigal, N., Feng, Y., Bristow, L., Tak-Yin Tsang, O., et al. (2020). Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science *369*, 1210–1220.

Balboa, M.A., Balsinde, J., Aramburu, J., Mollinedo, F., and López-Botet, M. (1992). Phospholipase D activation in human natural killer cells through the Kp43 and CD16 surface antigens takes place by different mechanisms. Involvement of the phospholipase D pathway in tumor necrosis factor alpha synthesis. J. Exp. Med. *176*, 9–17.

Bange, E.M., Han, N.A., Wileyto, P., Kim, J.Y., Gouma, S., Robinson, J., Greenplate, A.R., Hwee, M.A., Porterfield, F., Owoyemi, O., et al. (2021). CD8+ T cells contribute to survival in patients with COVID-19 and hematologic cancer. Nat. Med. *27*, 1280–1289.

Benetti, E., Giliberti, A., Emiliozzi, A., Valentino, F., Bergantini, L., Fallerini, C., Anedda, F., Amitrano, S., Conticini, E., Tita, R., et al. (2020). Clinical and molecular characterization of COVID-19 hospitalized patients. PLoS One *15*, e0242534.

Benner, C., Havulinna, A.S., Järvelin, M.-R., Salomaa, V., Ripatti, S., and Pirinen, M. (2017). Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. Am. J. Hum. Genet. *101*, 539–551.

Bi, J., and Wang, X. (2020). Molecular regulation of NK cell maturation. Front. Immunol. *11*, 1945.

Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational inference: a review for statisticians. J. Am. Stat. Assoc. *112*, 859–877.

Bowden, J., Del Greco, F., Minelli, C., Smith, G.D., Sheehan, N.A., and Thompson, J.R. (2016). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. Int. J. Epidemiol. *45*, 1961–1974.

Bowden, J., Hemani, G., and Davey Smith, G.D. (2018). Invited commentary: detecting individual and global horizontal pleiotropy in Mendelian randomization—a job for the humble heterogeneity statistic? Am. J. Epidemiol. *187*, 2681–2685.

Brodin, P. (2021). Immune determinants of COVID-19 disease presentation and severity. Nat. Med. *27*, 28–33.

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.

Burgess, S., and Thompson, S.G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. Eur. J. Epidemiol. *32*, 377–389.

Burgess, S., and Thompson, S.G.; CRP CHD Genetics Collaboration (2011). Avoiding bias from weak instruments in Mendelian randomization studies. Int. J. Epidemiol. *40*, 755–764.

Burgess, S., Davey Smith, G., Davies, N.M., Dudbridge, F., Gill, D., Glymour, M.M., Hartwig, F.P., Holmes, M.V., Minelli, C., Relton, C.L., et al. (2019). Guidelines for performing Mendelian randomization investigations. Wellcome Open Res. *4*, 186.

Carapito, R., Li, R., Helms, J., Carapito, C., Gujja, S., Rolli, V., Guimaraes, R., Malagon-Lopez, J., Spinnhirny, P., Lederle, A., et al. (2021). Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort. Sci. Transl. Med. *14*, eabj7521.

Cheng, S.H., and Higham, N.J. (1998). A modified Cholesky algorithm based on a symmetric indefinite factorization. SIAM J. Matrix Anal. Appl. *19*, 1097–1110.

Choi, K.W., Chen, C.-Y., Stein, M.B., Klimentidis, Y.C., Wang, M.-J., Koenen, K.C., and Smoller, J.W.; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2019). Assessment of bidirectional relationships between physical activity and depression among adults: a 2-sample Mendelian randomization study. JAMA Psychiatry *76*, 399–408.

Chou, C.-W., Huang, Y.-K., Kuo, T.-T., Liu, J.-P., and Sher, Y.-P. (2020). An overview of ADAM9: structure, activation, and regulation in human diseases. Int. J. Mol. Sci. *21*, 7790.

COVID-19 Host Genetics Initiative (2021). Mapping the human genetic architecture of COVID-19. Nature *600*, 472–477.

Darby, A.C., and Hiscox, J.A. (2021). Covid-19: variants and vaccination. BMJ *372*, n771.

De Biasi, S., Meschiari, M., Gibellini, L., Bellinazzi, C., Borella, R., Fidanza, L., Gozzi, L., Iannone, A., Lo Tartaro, D., Mattioli, M., et al. (2020). Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. Nat. Commun. *11*, 3434.

de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. *11*, e1004219.

Degenhardt, F., Ellinghaus, D., Juzenas, S., Lerga-Jaso, J., Wendorff, M., Maya-Miles, D., Uellendahl-Werth, F., ElAbd, H., Arora, J., Özer, O., et al. (2021). New susceptibility loci for severe COVID-19 by detailed GWAS analysis in European populations. Preprint at medRxiv. 10.1101/2021.07.21.21260624.

Delorey, T.M., Ziegler, C.G.K., Heimberg, G., Normand, R., Yang, Y., Segerstolpe, Å., Abbondanza, D., Fleming, S.J., Subramanian, A., Montoro, D.T., et al. (2021a). COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature *595*, 107–113.

Delorey, T.M., Ziegler, C.G.K., Heimberg, G., Normand, R., Yang, Y., Segerstolpe, A., Abbondanza, D., Fleming, S.J., Subramanian, A., Montoro, D.T., et al. (2021b). A single-cell and spatial atlas of autopsy tissues reveals pathology and cellular targets of SARS-CoV-2. Preprint at bioRxiv. 10.1101/2021.02.25.430130.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. *20*, 533–534.

Downes, D.J., Cross, A.R., Hua, P., Roberts, N., Schwessinger, R., Cutler, A.J., Munis, A.M., Brown, J., Mielczarek, O., de Andrea, C.E., et al. (2021). Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. Nat. Genet. *53*, 1606–1615.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

Gadotti, A.C., de Castro Deus, M., Telles, J.P., Wind, R., Goes, M., Garcia Charello Ossoski, R., de Padua, A.M., de Noronha, L., Moreno-Amaral, A., Baena, C.P., et al. (2020). IFN-γ is an independent risk factor associated with mortality in patients with moderate and severe COVID-19 infection. Virus Res. *289*, 198171.

Gauthier, L., Morel, A., Anceriz, N., Rossi, B., Blanchard-Alvarez, A., Grondin, G., Trichard, S., Cesari, C., Sapet, M., Bosco, F., et al. (2019). Multifunctional natural killer cell engagers targeting NKp46 trigger protective tumor immunity. Cell *177*, 1701–1713.e16.

Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. *53*, 403–411.

Hartwig, F.P., Davies, N.M., Hemani, G., and Davey Smith, G. (2016). Two-sample Mendelian randomization: avoiding the downsides of a powerful,

widely applicable but potentially fallible technique. Int. J. Epidemiol. *45*, 1717–1726.

Harva, M., and Kabán, A. (2007). Variational learning for rectified factor analysis. Signal Process. *87*, 509–527.

He, L., Mäe, M.A., Muhl, L., Sun, Y., Pietilä, R., Nahar, K., Liébanas, E.V., Fagerlund, M.J., Oldner, A., Liu, J., et al. (2020). Pericyte-specific vascular expression of SARS-CoV-2 receptor ACE2—implications for microvascular inflammation and hypercoagulopathy in COVID-19. Preprint at bioRxiv. 10.1101/2020.05.11.088500.

He, X.-S., Draghi, M., Mahmood, K., Holmes, T.H., Kemble, G.W., Dekker, C.L., Arvin, A.M., Parham, P., and Greenberg, H.B. (2004). T cell–dependent production of IFN-γ by NK cells in response to influenza A virus. J. Clin. Invest. *114*, 1812–1819.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell *181*, 271–280.e8.

Horowitz, A., Djaoud, Z., Nemat-Gorgani, N., Blokhuis, J., Hilton, H.G., Béziat, V., Malmberg, K.-J., Norman, P.J., Guethlein, L.A., and Parham, P. (2016). Class I HLA haplotypes form two schools that educate NK cells in different ways. Sci. Immunol. *1*, eaag1672.

Hu, W., Wang, G., Huang, D., Sui, M., and Xu, Y. (2019). Cancer immunotherapy based on natural killer cells: current progress and new opportunities. Front. Immunol. *10*, 1205.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet *395*, 497–506.

Hussman, J.P. (2020). Cellular and molecular pathways of COVID-19 and potential points of therapeutic intervention. Front. Pharmacol. *11*, 1169.

Julian, T.H., Glascow, N., Barry, A.D.F., Moll, T., Harvey, C., Klimentidis, Y.C., Newell, M., Zhang, S., Snyder, M.P., Cooper-Knock, J., et al. (2021). Physical exercise is a risk factor for amyotrophic lateral sclerosis: convergent evidence from Mendelian randomisation, transcriptomics and risk genotypes. EBioMedicine *68*, 103397.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.

Kosmicki, J.A., Horowitz, J.E., Banerjee, N., Lanche, R., Marcketta, A., Maxwell, E., Bai, X., Sun, D., Backman, J.D., Sharma, D., et al. (2021). A catalog of associations between rare coding variants and COVID-19 outcomes. Preprint at medRxiv. 10.1101/2020.10.28.20221804.

Kousathanas, A., Pairo-Castineira, E., Rawlik, K., Stuckey, A., Odhams, C.A., Walker, S., Russell, C.D., Malinauskas, T., Millar, J., Elliott, K.S., et al. (2021). Whole genome sequencing identifies multiple loci for critical illness caused by COVID-19. Preprint at medRxiv. 10.1101/2021.09.02.21262965.

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. *19*, 1454–1462.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97.

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput. Biol. *12*, e1004714.

Lei, Q.-Y., Zhang, H., Zhao, B., Zha, Z.-Y., Bai, F., Pei, X.-H., Zhao, S., Xiong, Y., and Guan, K.-L. (2008). TAZ promotes cell proliferation and epithelial-mesenchymal transition and is inhibited by the hippo pathway. Mol. Cell. Biol. *28*, 2426–2436.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, X., Marmar, T., Xu, Q., Tu, J., Yin, Y., Tao, Q., Chen, H., Shen, T., and Xu, D. (2020). Predictive indicators of severe COVID-19 independent of comorbidities and advanced age: a nested case- control study. Epidemiol. Infect. *148*, e255.

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat. Med. *26*, 842–844.

Lucas, C., Wong, P., Klein, J., Castro, T.B.R., Silva, J., Sundaram, M., Ellingson, M.K., Mao, T., Oh, J.E., Israelow, B., et al. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature *584*, 463–469.

Machiela, M.J., and Chanock, S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics *31*, 3555–3557.

Mathew, D., Giles, J.R., Baxter, A.E., Oldridge, D.A., Greenplate, A.R., Wu, J.E., Alanio, C., Kuri-Cervantes, L., Pampena, M.B., D'Andrea, K., et al. (2020). Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science *369*, eabc8511.

Maucourant, C., Filipovic, I., Ponzetta, A., Aleman, S., Cornillet, M., Hertwig, L., Strunz, B., Lentini, A., Reinius, B., Brownlie, D., et al. (2020). Natural killer cell immunotypes related to COVID-19 disease severity. Sci. Immunol. *5*, eabd6832.

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. *53*, 1097–1103.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl variant effect predictor. Genome Biol. *17*, 122.

Mehta, P., McAuley, D.F., Brown, M., Sanchez, E., Tattersall, R.S., and Manson, J.J.; HLH Across Speciality Collaboration, UK (2020). COVID-19: consider cytokine storm syndromes and immunosuppression. Lancet *395*, 1033–1034.

Melms, J.C., Biermann, J., Huang, H., Wang, Y., Nair, A., Tagore, S., Katsyv, I., Rendeiro, A.F., Amin, A.D., Schapiro, D., et al. (2021). A molecular single-cell lung atlas of lethal COVID-19. Nature *595*, 114–119.

Michel, T., Poli, A., Cuapio, A., Briquemont, B., Iserentant, G., Ollert, M., and Zimmer, J. (2016). Human CD56bright NK cells: an update. J. Immunol. *196*, 2923–2931.

Milewska, A., Kindler, E., Vkovski, P., Zeglen, S., Ochman, M., Thiel, V., Rajfur, Z., and Pyrc, K. (2018). APOBEC3-mediated restriction of RNA virus replication. Sci. Rep. *8*, 5960.

Montoro, D.T., Haber, A.L., Biton, M., Vinarsky, V., Lin, B., Birket, S.E., Yuan, F., Chen, S., Leung, H.M., Villoria, J., et al. (2018). A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature *560*, 319–324.

Mountjoy, E., Schmidt, E.M., Carmona, M., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Schwartzentruber, J., Karim, M.A., et al. (2020). Open Targets Genetics: an open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Preprint at bioRxiv. 10.1101/2020.09.16.299271.

Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature *593*, 238–243.

Norman, J.M., Mashiba, M., McNamara, L.A., Onafuwa-Nuga, A., Chiari-Fort, E., Shen, W., and Collins, K.L. (2011). The antiviral factor APOBEC3G

enhances the recognition of HIV-infected primary T cells by natural killer cells. Nat. Immunol. *12*, 975–983.

Novelli, A., Biancolella, M., Borgiani, P., Cocciadiferro, D., Colona, V.L., D'Apice, M.R., Rogliani, P., Zaffina, S., Leonardis, F., Campana, A., et al. (2020). Analysis of ACE2 genetic variants by direct exome sequencing in 99 SARS-CoV-2 positive patients. Hum. Genomics *14*, 29.

Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A.D., Rawlik, K., Pasko, D., Walker, S., Parkinson, N., Fourman, M.H., Russell, C.D., et al. (2021). Genetic mechanisms of critical illness in COVID-19. Nature *591*, 92–98.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raulet, D.H. (2003). Roles of the NKG2D immunoreceptor and its ligands. Nat. Rev. Immunol. *3*, 781–790.

Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., et al. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. Cell *184*, 1895–1913.e19.

Renoux, V.M., Zriwil, A., Peitzsch, C., Michaëlsson, J., Friberg, D., Soneji, S., and Sitnicka, E. (2015). Identification of a human natural killer cell lineage-restricted progenitor in fetal and adult tissues. Immunity *43*, 394–407.

Roederer, M., Quaye, L., Mangino, M., Beddall, M.H., Mahnke, Y., Chattopadhyay, P., Tosi, I., Napolitano, L., Terranova Barberio, M., Menni, C., et al. (2015). The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. Cell *161*, 387–403.

Romee, R., Foley, B., Lenvik, T., Wang, Y., Zhang, B., Ankarlo, D., Luo, X., Cooley, S., Verneris, M., Walcheck, B., et al. (2013). NK cell CD16 surface expression and function is regulated by a disintegrin and metalloprotease-17 (ADAM17). Blood *121*, 3599–3608.

Severe Covid-19 GWAS Group, Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J., Prati, D., Baselli, G., et al. (2020). Genomewide association study of severe Covid-19 with respiratory failure. N. Engl. J. Med. *383*, 1522–1534.

Shang, Y., Liu, T., Wei, Y., Li, J., Shao, L., Liu, M., Zhang, Y., Zhao, Z., Xu, H., Peng, Z., et al. (2020). Scoring systems for predicting mortality for severe patients with COVID-19. EClinicalMedicine *24*, 100426.

Shelton, J.F., The 23andMe COVID-19 Team, Shastri, A.J., Ye, C., Weldon, C.H., Filshtein-Sonmez, T., Coker, D., Symons, A., Esparza-Gordillo, J., Aslibekyan, S., et al. (2021). Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. Nat. Genet. *53*, 801–808.

Shilo, S., Rossman, H., and Segal, E. (2021). Signals of hope: gauging the impact of a rapid national vaccination campaign. Nat. Rev. Immunol. *21*, 198–199.

Smith, G.D. (2010). Mendelian randomization for strengthening causal inference in observational studies: application to gene × environment interactions. Perspect. Psychol. Sci. *5*, 527–545.

Sosa-Hernández, V.A., Torres-Ruíz, J., Cervantes-Díaz, R., Romero-Ramírez, S., Páez-Franco, J.C., Meza-Sánchez, D.E., Juárez-Vega, G., Pérez-Fragoso, A., Ortiz-Navarrete, V., Ponce-de-León, A., et al. (2020). B cell subsets as severity-associated signatures in COVID-19 patients. Front. Immunol. *11*, 611004.

Stephenson, E., Reynolds, G., Botting, R.A., Calero-Nieto, F.J., Morgan, M., Tuong, Z.K., Bach, K., Sungnak, W., Worlock, K.B., Yoshida, M., et al. (2021). The cellular immune response to COVID-19 deciphered by single cell multi-omics across three UK centres. Preprint at medRxiv. 10.1101/2021.01.13.21249725.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

The COVID-19 Host Genetics Initiative (2020). The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur. J. Hum. Genet. *28*, 715–718.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619–625.

Turro, E., Astle, W.J., Megy, K., Graf, S., Greene, D., Shamardina, O., Allen, H.L., Sanchis-Juan, A., Frontini, M., Thys, C., et al. (2021). Whole-genome sequencing of patients with rare diseases in a national health system. Nature *583*, 96–102.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27.

van Helden, M.J., Goossens, S., Daussy, C., Mathieu, A.-L., Faure, F., Marçais, A., Vandamme, N., Farla, N., Mayol, K., Viel, S., et al. (2015). Terminal NK cell maturation is controlled by concerted actions of T-bet and Zeb2 and is essential for melanoma rejection. J. Exp. Med. *212*, 2015–2025.

Varchetta, S., Mele, D., Oliviero, B., Mantovani, S., Ludovisi, S., Cerino, A., Bruno, R., Castelli, A., Mosconi, M., Vecchia, M., et al. (2021). Unique immunological profile in patients with COVID-19. Cell. Mol. Immunol. *18*, 604–612.

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat. Genet. *50*, 693–698.

Wang, A., Chiou, J., Poirion, O.B., Buchanan, J., Valdez, M.J., Verheyden, J.M., Hou, X., Kudtarkar, P., Narendra, S., Newsome, J.M., et al. (2020a). Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. Elife *9*, e62522.

Wang, E.Y., Mao, T., Klein, J., Dai, Y., Huck, J.D., Jaycox, J.R., Liu, F., Zhou, T., Israelow, B., Wong, P., et al. (2021). Diverse Functional Autoantibodies in Patients with COVID-19 (Nature Publishing).

Wang, F., Huang, S., Gao, R., Zhou, Y., Lai, C., Li, Z., Xian, W., Qian, X., Li, Z., Huang, Y., et al. (2020b). Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility. Cell Discov. *6*, 83.

Wang, S., Cho, H., Zhai, C., Berger, B., and Peng, J. (2015). Exploiting ontology graph for predicting sparsely annotated gene function. Bioinformatics *31*, i357–i364.

Wang, S.-M., and Wang, C.-T. (2009). APOBEC3G cytidine deaminase association with coronavirus nucleocapsid protein. Virology *388*, 112–120.

Watzl, C., and Long, E.O. (2010). Signal transduction during activation and inhibition of natural killer cells. Curr. Protoc. Immunol. *Chapter 11*, Unit 11.9B.

Witkowski, M., Tizian, C., Ferreira-Gomes, M., Niemeyer, D., Jones, T.C., Heinrich, F., Frischbutter, S., Angermair, S., Hohnstein, T., Mattiola, I., et al. (2021). Untimely TGFβ responses in COVID-19 limit antiviral functions of NK cells. Nature *600*, 295–301.

Wootton, R.E., Lawn, R.B., Millard, L.A.C., Davies, N.M., Taylor, A.E., Munafò, M.R., Timpson, N.J., Davis, O.S.P., Davey Smith, G., and Haworth, C.M.A. (2018). Evaluation of the causal effects between subjective wellbeing and cardiometabolic health: Mendelian randomisation study. BMJ *362*, k3788.

Zhang, J.-Y., Wang, X.-M., Xing, X., Xu, Z., Zhang, C., Song, J.-W., Fan, X., Xia, P., Fu, J.-L., Wang, S.-Y., et al. (2020a). Single-cell landscape of immunological responses in patients with COVID-19. Nat. Immunol. *21*, 1107–1118.

Zhang, S., Cooper-Knock, J., Weimer, A.K., Shi, M., Moll, T., Marshall, J.N.G., Harvey, C., Nezhad, H.G., Franklin, J., Souza, C.D.S., et al. (2022). Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. Neuron *110* (6), 992–1008.e11.

Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based *ab initio* prediction of variant effects on expression and disease risk. Nat. Genet. *50*, 1171–1179.

STAR★METHODS

KEY RESOURCES TABLE

Reagent or Resource	Source	Identifier
Critical Commercial Assays		
NK Cell isolation kit	Miltenyi Biotec	#130-092-657
Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Bundle	10X Genomics	PN-1000283
Experimental Models: Cell Lines		
Homo sapiens male adult (33 years) CD14-negative NK primary cell nuclear fraction	The Jackson Laboratory	N/A
Software and Algorithms		
RefMap	https://github.com/szhang1112/refmap	DOI: 10.5281/zenodo.5774249
ArchR	https://github.com/GreenleafLab/ArchR	N/A
CellRanger ARC	10X Genomics	N/A
SKAT-O	https://cran.r-project.org/web/packages/SKAT/index.html	N/A
R v4.0.1	https://cran.r-project.org/mirrors.html	N/A
snpStats	https://www.bioconductor.org/packages/release/bioc/html/snpStats.html	N/A
VariantAnnotation	https://www.bioconductor.org/packages/release/bioc/html/VariantAnnotation.html	N/A
VAutils	https://github.com/oyhel/vautils/	N/A
PLINK V1.90	http://zzz.bwh.harvard.edu/plink/download.shtml	N/A
PRISM 7	GraphPad	N/A
IGV v2.4.16	https://software.broadinstitute.org/software/igv/	N/A
MATLAB R2018b	MathWorks	N/A
MAGMA v1.08	https://ctg.cncr.nl/software/magma	N/A
Pascal	https://www2.unil.ch/cbg/index.php?title=Pascal	N/A
PAINTOR v3.0	https://github.com/gkichaev/PAINTOR_V3.0	N/A
LD Score Regression	https://github.com/bulik/ldsc	N/A
HOMER v4.11	http://homer.ucsd.edu/homer/	N/A
Data		
NK single-cell multiome data	encodeproject.org	https://www.encodeproject.org/experiments/ENCSR710NDM/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Micheal P. Snyder (mpsnyder@stanford.edu).

Materials availability

All unique/stable reagents generated in this study are available from the lead contact without restriction.

Data and code availability

- ***Source data statement***: This paper analyzes existing, publicly available published datasets with the exception of single-cell multiome data of NK cells which have been deposited at encodeproject.org and are publicly available as of the date of publication. The accession number is listed in the key resources table.
- ***Code statement***: All original code has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

The RefMap model

The general idea of RefMap is to model the genetic associations within individual genomic regions whose function is informed by epigenetic features. In particular, in RefMap allele *Z*-scores were calculated as *Z=b/*se, where *b* and *se* are effect size and standard error, respectively, as reported by the COVID-19 GWAS (COVID-19 Host Genetics Initiative, 2021) (COVID-19 Host Genetics Initiative, Release 5, phenotype definition A2, EUR only) where the sample age, sex, and ancestry information were included as covariates. Given *Z*-scores and lung snATAC-seq peaks, we aim to identify functional genomic regions in which the *Z*-score distribution is significantly shifted from the null distribution, informing the disease association. Suppose we have K 1Mb linkage disequilibrium (LD) blocks, where each LD block contains J_k (k=1,..., K) 1kb regions and each region harbors $I_{j,k}$ (j=1,..., J_k, $I_{j,k}$>0) SNPs, the *Z*-scores follow a multivariate normal distribution, i.e.,

$$\boldsymbol{z}_k|\boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{\Sigma}_k\boldsymbol{u}_k, \boldsymbol{\Sigma}_k),\ k = 1, \cdots, K, \quad \text{(Equation 1)}$$

where the *Z*-score of the i-th SNP in the j-th region of the k-th block is denoted as $z_{i,j,k}$ (i=1,..., $I_{j,k}$) and $\boldsymbol{u}_k$ are the effect sizes that can be expressed as

$$\boldsymbol{u}_k = \left[u^T_{1:I_{1,k},1,k}, \cdots, u^T_{1:I_{j,k}j,k}, \cdots u^T_{1:I_{J_k,k},J_k,k}\right]^T. \quad \text{(Equation 2)}$$

$\boldsymbol{\Sigma}_k \in \mathbb{R}^{I_k \times I_k}$ in Equation 1 represents the in-sample LD matrix comprising of the pairwise Pearson correlation coefficients between pairs of SNPs within the k-th block, and I_k is the total number of SNPs given by $I_k = \Sigma_{j=1}^{J_k} I_{j,k}$. Here, since we have no access to the individual-level data, we used EUR samples from the 1000 Genomes Project (Phase 3) to estimate $\boldsymbol{\Sigma}_k$, which yields the out-sample LD matrix. A modified Cholesky algorithm (Cheng and Higham, 1998) was used to get a symmetric positive definite (SPD) approximation of the LD matrix.

Further, we assume $u_{i,j,k}$ (i=1,..., $I_{j,k}$) are independent and identically distributed (i.i.d.), following a normal distribution given by

$$u_{i,j,k}|m_{j,k}, \lambda_{j,k} \sim N\left(m_{j,k}, \lambda_{\mathrm{j,k}}^{-1}\right), i = 1, \cdots, I_{j,k}, \quad \text{(Equation 3)}$$

in which the precision $\lambda_{j,k}$ follows a Gamma distribution, i.e.,

$$\lambda_{j,k} \sim \mathrm{Gamma}(a_0, b_0). \quad \text{(Equation 4)}$$

Furthermore, to characterize the shift of the expectation in Equation 3 from the null distribution, we model $m_{j,k}$ by a three-component Gaussian mixture model (GMM) given by

$$m_{j,k}|t_{j,k}, v_{-1}, v_{+1}, \tau_0, \tau_{-1} \sim \underbrace{\mathcal{N}\left(-v_{-1}, \tau_{-1}^{-1}\right)^{t^{(-1)}_{j,k}}}_{\text{negative}} \underbrace{\mathcal{N}\left(0, \tau_0^{-1}\right)^{t^{(0)}_{j,k}}}_{\text{zero}} \underbrace{\mathcal{N}\left(v_{+1}, \tau_{+1}^{-1}\right)^{t^{(+1)}_{j,k}}}_{\text{positive}}, \quad \text{(Equation 5)}$$

in which the precisions follow

$$\tau_{-1}, \tau_0, \tau_{+1} \sim \mathrm{Gamma}(a_0, b_0), \quad \text{(Equation 6)}$$

and v_{-1} and v_{+1} are non-negative variables measuring the absolute values of effect size shifts for the negative and positive components, respectively.

To impose non-negativity over v_{-1} and v_{+1}, we adopt the rectification nonlinearity technique proposed previously (Harva and Kabán, 2007). In detail, we assume v_{-1} and v_{+1} follow

$$v_{-1}|m_{-1}, \lambda_{-1} \sim \Re^N(m_{-1}, \lambda_{-1}), \quad \text{(Equation 7)}$$

$$v_{+1}|m_{+1}, \lambda_{-1} \sim \Re^N(m_{+1}, \lambda_{+1}), \quad \text{(Equation 8)}$$

in which the rectified Gaussian distribution is defined via a dumb variable. We then define v_{-1} and v_{+1} by

$$v_{-1} = \max(r_{-1}, 0), \quad \text{(Equation 9)}$$

$$v_{+1} = \max(r_{+1}, 0), \quad \text{(Equation 10)}$$

which guarantees that v_{-1} and v_{+1} are non-negative. The dump variable r_{-1} and r_{+1} follow the Gaussian distributions given by

$$r_{-1}|m_{-1}, \lambda_{-1} \sim \mathcal{N}\left(m_{-1}, \lambda_{-1}^{-1}\right), \quad \text{(Equation 11)}$$

$$r_{+1}|m_{+1},\lambda_{+1} \sim \mathcal{N}\left(m_{+1},\lambda_{+1}^{-1}\right), \quad \text{(Equation 12)}$$

where $m_{\pm}$ and $\lambda_{\pm}$ follow the Gaussian-Gamma distributions, i.e.,

$$m_{-1},\lambda_{-1} \sim \mathcal{N}\left(\mu_0,(\beta_0\lambda_{-1})^{-1}\right)\text{Gamma}(a_0,b_0), \quad \text{(Equation 13)}$$

$$m_{+1},\lambda_{+1} \sim \mathcal{N}\left(\mu_0,(\beta_0\lambda_{+1})^{-1}\right)\text{Gamma}(a_0,b_0). \quad \text{(Equation 14)}$$

Note that an advantage of the rectification nonlinearity is it is tractable in the variational inference framework (Harva and Kabán, 2007).

The indicator variables in Equation 5 denote whether that region is disease-associated or not. Indeed, we define the region to be disease-associated if $t_{j,k}^{(-1)} = 1$ or $t_{j,k}^{(+1)} = 1$, and to be non-associated otherwise. To simplify the analysis, we put a symmetry over $t_{j,k}^{(-1)}$ and $t_{j,k}^{(+1)}$, and define the distribution by

$$p(t_{j,k}|\pi_{j,k}) = (0.5\pi_{j,k})^{t_{j,k}^{(-1)}}(1-\pi_{j,k})^{t_{j,k}^{(0)}}(0.5\pi_{j,k})^{t_{j,k}^{(+1)}}, j = 1,\cdots,j_k, k = 1,\cdots,K. \quad \text{(Equation 15)}$$

Furthermore, to incorporate the functional information into the modeling, we define the probability parameter $\pi_{j,k}$ in Equation 15 as

$$\pi_{j,k} = \sigma\left(\boldsymbol{w}^T\boldsymbol{s}_{j,k}\right), \quad \text{(Equation 16)}$$

where $\sigma(.)$ is the sigmoid function, $\boldsymbol{s}_{j,k}$ is the vector of epigenetic features for the j-th region in the k-th LD block, and the weight vector $\boldsymbol{w}$ follows a multivariate normal distribution, i.e.,

$$\boldsymbol{w}|\boldsymbol{\Lambda} \sim \mathcal{N}\left(0,\boldsymbol{\Lambda}^{-1}\right), \quad \text{(Equation 17)}$$

and $\boldsymbol{\Lambda}$ follows

$$\boldsymbol{\Lambda} \sim \mathcal{W}(\boldsymbol{W}_0,\nu_0). \quad \text{(Equation 18)}$$

In this study, the epigenetic feature $\boldsymbol{s}_{j,k}$ was calculated as the overlapping ratios of that region with the snATAC-seq peaks detected in any of the cell types in healthy human lungs. All priors in RefMap are defined based on the conjugacy rule to make the variational inference tractable.

Based on the model defined in Equations 1–18, we are interested in calculating the posterior probability $p(\boldsymbol{T}|\boldsymbol{Z},\boldsymbol{S})$, where the mean-field variational inference (MFVI) and local variational method (Blei et al., 2017) were adopted to solve the intractability. Here, we used variational inference, an approximate inference framework, because of its superiority in convergence rate compared to sampling methods. More technical details, including a coordinate ascent-based inference algorithm, can be found in Zhang et al. (2022). In this study, we ran the MFVI algorithm per chromosome to further accelerate the computation. The Q^+- and Q^--scores were defined as $q\left(t^{(+1)}=1\right)$ and $q\left(t^{(-1)}=1\right)$, respectively, and we also defined the Q-score as $Q=Q^+ + Q^-$. RefMap regions were identified by Q^+- or Q^--score >0.95.

Mapping cell-type-specific genes from RefMap regions

For each cell type within lung tissue, we defined cell-type-specific RefMap regions as the overlap between RefMap regions and the total set of snATAC-seq peaks detected in that cell type. Cell-type-specific RefMap genes were then identified if the extended gene body (i.e., the region up to 10kb either side of the annotated gene body) overlapped with any of the cell-type-specific regions (Turro et al., 2021). To get the final gene lists, non-expressed genes in corresponding cell types were excluded from RefMap genes in the downstream analysis. In addition, we note that there are non-adult samples (~30 weeks gestation and ~3 years) sequenced in the single cell profiling data (Wang et al., 2020a). To remove the bias towards lung development, we first calculated the fold change of gene expression levels between the adult sample (~30 years) and non-adult ones, and defined non-developmental genes (nDG) as those with FC>1.5. Only RefMap genes that were identified as expressed and non-developmental in each cell type were kept for downstream analysis.

Validation of RefMap results in 23andMe and GenOMICC datasets

We calculated the overlap of total RefMap regions and of cell-type-specific RefMap regions with genomic regions shown to contain COVID-19-associated SNPs (p < 1e-04) based on the GWAS of an independent cohort recruited by 23andMe (Shelton et al., 2021). To determine whether the observed overlap is statistically significant, we examined the average overlap with ten sets of control regions of equivalent length to RefMap regions. Control regions were +/-1Mb-5Mb distant from the RefMap regions (Quinlan and Hall, 2010). The same procedure was performed for the GenOMICC dataset (Kousathanas et al., 2021).

SNP association with COVID-19 in the 23andMe and GenOMICC datasets was calculated as previously described (Kousathanas et al., 2021; Pairo-Castineira et al., 2021).

Heritability analysis

We used LD score regression (LDSC) (Bulik-Sullivan et al., 2015) to calculate the SNP-based heritability based on different GWAS and different gene sets in this study. Partitioned heritability was calculated as previously described (Finucane et al., 2015). Briefly, for all gene lists, we examined the proportion of total SNP-based heritability carried by SNPs +/-100kb from the transcription start site (TSS) of each gene in the list. Enrichment was calculated by comparing the ratio of partitioned heritability to the quantity of genetic material.

Implementation details of MAGMA, PAINTOR and Pascal

MAGMA (v1.08) (de Leeuw et al., 2015) and Pascal (Lamparter et al., 2016) were applied using default settings. Input consisted of summary statistics for all SNPs genome-wide as measured in the COVID-19 GWAS (COVID-19 Host Genetics Initiative, 2021). We employed PAINTOR (v3.0) following the guidance provided in Kichaev et al. (2014) and https://github.com/gkichaev/PAINTOR_V3.0/. The genome was annotated based on the snATAC-seq peaks detected in human lungs (Wang et al., 2020a). We ran the algorithm in the MCMC mode. All other parameters in PAINTOR were left to be default. In all cases, we estimated the LD structure using EUR samples from the 1000 Genomes Project phase 3.

Mendelian randomization

In total, 46 GWAS measures of NK cell subtypes were identified from the IEU Open GWAS Project, including "met-b-124", "met-b-245", "met-b-242", "met-b-237", "met-b-258", "met-b-246", "met-b-249", "met-b-140", "met-b-240", "met-b-123", "met-b-250", "met-b-239", "met-b-120", "met-b-154", "met-b-247", "met-b-251", "met-b-238", "met-b-243", "met-b-244", "met-b-153", "met-b-248", "met-b-152", "met-b-122", "met-b-121", "met-b-252", and "met-b-241" (Roederer et al., 2015). Exposure SNPs or instrumental variables (IVs) are chosen based on an arbitrary *P*-value cutoff (Choi et al., 2019; Wootton et al., 2018). A cutoff that is too low will lose informative instruments, but a cutoff that is too high could introduce non-informative instruments. We chose to set the cutoff at 5e-06 in line with our previous work (Julian et al., 2021). We employed a series of sensitivity analyses to ensure that our analysis was not confounded by invalid IVs. Identified SNPs were clumped for independence using PLINK clumping in the TwoSampleMR tool (Purcell et al., 2007). A stringent cutoff of $R^2 \leq 0.001$ and a window of 10,000kb were used for clumping within a European reference panel. Where SNPs were in LD, those with the lowest *P*-value were retained. SNPs that were not present in the reference panel were excluded. Where an exposure SNP was unavailable in the outcome dataset, a proxy with a high degree of LD ($R^2 \geq 0.9$) was identified in LDlink within a European reference population (Machiela and Chanock, 2015). Where a proxy was identified to be present in both datasets, the target SNP was replaced with the proxy in both exposure and outcome datasets in order to avoid phasing issues (Hartwig et al., 2016). Where a SNP was not present in both datasets and no SNP was available in sufficient LD, the SNP was excluded from the analysis. The effects of SNPs on outcomes and exposures were harmonized in order to ensure that the beta values were signed with respect to the same alleles. For palindromic alleles, those with minor allele frequency (MAF) > 0.42 were omitted from the analysis in order to reduce the risk of errors due to strand issues (Hartwig et al., 2016).

The MR measure with the greatest power is the inverse-variance weighted (IVW) method, but this is contingent upon the exposure IV assumptions being satisfied (Burgess and Thompson, 2017). With the inclusion of a large number of SNPs within the exposure IV, it is possible that not all variants included are valid instruments and therefore, in the event of a significant result, it is necessary to include a range of robust methods which provide valid results under various violations of MR principles at the expense of power (Burgess et al., 2019). Robust methods applied in this study include MR-Egger, MR-PRESSO, weighted median, weighted mode, and MR-Lasso.

With respect to the IVW analysis, a fixed-effects (FE) model is indicated in the case of homogeneous data, whilst a multiplicative random effects (MRE) model is more suitable for heterogeneous data. Burgess et al. recommended that an MRE model be implemented when using GWAS summary data to account for heterogeneity in variant-specific causal estimates (Burgess et al., 2019). In the interest of transparency, we calculated both results but present the MRE in the text.

MR analyses should include evaluation of exposure IV strength. In order to achieve this, we provided the *F*-statistic, MR-Egger intercept, MR-PRESSO global test, Cochran's Q test, and I^2 for our data. The *F*-statistic is a measure of instrument strength with >10 indicating a sufficiently strong instrument (Burgess et al., 2011). We provided *F*-statistics for individual exposure SNPs and the instrument as a whole. Cochran's Q test is an indicator of heterogeneity in the exposure dataset and serves as a useful indicator that horizontal pleiotropy is present as well as directing decisions to implement FE or MRE IVW approaches (Bowden et al., 2018). The MR-Egger intercept test determines whether there is directional horizontal pleiotropy. The MR-PRESSO global test determines if there are statistically significant outliers within the exposure-outcome analysis (Verbanck et al., 2018). I^2 was calculated as a measure of heterogeneity between variant specific causal estimates, with a low I^2 indicating that Egger is more likely to be biased towards the null (Bowden et al., 2016). Finally, we performed a leave-one-out analysis using the method of best fit for each exposure SNP within the IV in order to determine if any single variants were exerting a disproportionate effect upon the results of our analysis (Burgess et al., 2019). The same MR analysis was conducted for COVID19-hg and GenOMICC GWAS independently.

Single-cell multiome profiling of NK cells

To obtain the NK single-cell multiome data, human PBMCs were isolated from a 33 year old male and sorted including CD14 exclusion (Renoux et al., 2015) to isolate NK cells (as utilized in Gauthier et al. (2019). NK cells were isolated by a series of monoclonal

antibodies conjugated to magnetic beads (Miltenyi Biotech). The final step was sorting (AutoMACS pro separator, Miltenyi Biotech) to achieve CD14 exclusion (monocytes). A detailed nuclei isolation protocol can be found at https://www.encodeproject.org/experiments/ENCSR710NDM/.

After nuclei isolation the 10x multiome protocol (Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Bundle, PN-1000283) was followed according to manufacturer's instructions (10X Genomics, USA) and can be accessed at https://assets.ctfassets.net/an68im79xiti/1MrGvRY2vJF0Fc1lcn3BPW/d24d1099ee82656f5c230cd08bdaec8b/CG000338_ChromiumNextGEM_Multiome_ATAC_GEX_User_Guide_RevD.pdf or https://www.encodeproject.org/experiments/ENCSR710NDM/.

Libraries were sequenced on a NovaSeq6000 instrument. The scATAC-seq libraries were sequences as follows: Read 1N, 50 cycles; i7 Index, 8 cycles; i5 Index, 24 Cycles; Read 2N, 49 cycles. The snRNA-seq libraries were sequenced as follows: Read 1, 28 cycles; i7 Index, 10 cycles; i5 Index, 10 Cycles; Read 2, 90 cycles.

12,834 cells were profiled; 5,046 ATAC fragments and 858 expressed genes were identified per cell. TSS enrichment score for the scATAC-seq data was 11. Raw data was analyzed using CellRanger ARC (10X Genomics).

Deriving cell-type-specific genes for T-cell and NK-cell subtypes

We derived $CD4^+$ and $CD8^+$ T-cell genes from RefMap T-cell genes based on their relative expression (Travaglini et al., 2020). In particular, we defined $CD4^+$ T-cell genes as those genes higher expressed in $CD4^+$ T cells compared to $CD8^+$ T cells (FC>1.5). The $CD8^+$ T-cells genes were defined similarly. Identified gene sets were then analyzed in downstream analysis such as heritability partitioning as described above (see section "heritability analysis").

NK-cell multiome data was analyzed using ArchR (Granja et al., 2021). NK-cell subtypes were identified based on relative expression and open chromatin over the gene body and promoter, respectively. In downstream analysis doublets (a single droplet containing multiple cells) were excluded. Missing data was imputed using MAGIC (Markov affinity-based graph imputation of cells) (van Dijk et al., 2018). Dimensionality reduction was performed by iterative latent semantic indexing (LSI) and data was visualized using uniform manifold approximation and projection (UMAP). RefMap NK-cell genes were assigned to NK-cell subtypes based on relative expression within NK multiome data. For each subtype, a suitable control group provided a comparison to identify genes overexpressed (FC>1.5) in the subgroup of interest: $CD56^{bright}$ NK cells were compared to $CD56^{null}$ NK cells; NKG2D+, CD16+ and NKG2A+ NK cells were compared to NKG2D-, CD16- and NKG2A- NK cells respectively. Whereas the latter three groups were identified based on presence/absence of expression, $CD56^{bright/dim}$ NK cells were identified by first excluding cells with no CD56 expression, and then dividing the remaining cells based on the overall median expression of CD56. The method applied here is analogous to that used to assign RefMap T-cell genes to $CD4^+$/$CD8^+$ T-cells.

Motif enrichment analysis

Motif enrichment analysis was performed using HOMER v4.11 (Heinz et al., 2010). Specifically, to focus on RefMap regions and remove enrichment bias from the general chromatin accessibility, we used the original snATAC-seq peaks as the background in the analysis. Other parameters were left to be default.

Transcriptome analysis

Four single-cell RNA-seq datasets were used in the transcriptome analyses, including human healthy lungs (Travaglini et al., 2020; Wang et al., 2020a) and COVID-19 patients (Liao et al., 2020; Ren et al., 2021). Data after quality control (QC) was acquired for each study. Only samples from the respiratory system were considered in the analyses. For the healthy lung data, only expressed genes were considered in the analysis. For the disease samples, we removed the overlap of severe patients between the two cohorts (Liao et al., 2020; Ren et al., 2021). In the comparative expression analysis of severe versus moderate patients, to stabilize the analysis we estimated the change of gene expression levels using the Z-score estimated from Wilcoxon rank-sum test, wherein a positive Z-score indicates a higher expression level in severe patients and a negative value suggests the lower expression. The Benjamini-Hochberg (BH) procedure was used for multiple testing correction throughout the study.

Network analysis

We first downloaded the human PPIs from STRING v11, including 19,567 proteins and 11,759,455 protein interactions. To eliminate the bias caused by hub proteins, we first carried out the random walk with restart algorithm (Wang et al., 2015) over the PPI network, wherein the restart probability was set to 0.5, resulting in a smoothed network after retaining the top 5% predicted edges. To decompose the network into different subnetworks/modules, we performed the Leiden algorithm (Traag et al., 2019), a community detection algorithm that searches for densely connected modules by optimizing the modularity. After the algorithm converged, we obtained 1,681 modules with an average size of 9.98 nodes (SD=53.35; Table S6).

Rare-variant burden testing

Rare-variant burden testing was performed to determine whether any genes were differentially enriched with rare variants between severe COVID-19 patients and population controls. MAF were checked against gnomAD/ESP, as well as a pooled common variant list obtained from 19 of the 21 participating cohorts; variants with MAF >1% were removed. We utilized an additive model in which a score of 0 was assigned if there were no deleterious variants identified in a particular gene; 1 if at least one deleterious variant, but all heterozygous; and 2 if at least one homozygous deleterious variant. All burden tests were performed using REGENIE v2 (Mbatchou

et al., 2021) with firth correction. LoF variants were identified by HIGH impact in the Ensembl annotations from VEP (McLaren et al., 2016). Meta-analysis was performed in two steps: first an inverse-variance weighted fixed-effect meta-analysis was performed to obtain summary statistics of the same ancestry. Then a random-effect (Dersimonian-Laird) meta-analysis was applied across the resulting mono-ancestry summary statistics. To avoid inflation of test statistics we removed all genes for which the number of controls was <5,000 or the MAC<10. A QQ-plot confirmed that there was no significant genomic inflation (λ_{GC}=0.825; Figure S4). Significant enrichment of rare LoF variants associated with severe COVID-19 for each cell-type-specific RefMap gene set was determined by comparing with 10,000 gene sets of the same size randomly selected from genes which passed all stages of QC.