Integrated analysis of COVID-19 multi-omics data for eQTLs reveals genetic mechanisms underlying disease severity

Jeongha Lee,[1]* Eun Young Jeon,[1]* Liyang Yu,[2]* Hye-Yeong Jo,[3] Sang Cheol Kim,[3] Woong-Yang Park,[4,5] Hyun-Young Park,[6] Siming Zhao,[2]** Murim Choi[1]**

[1]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea

[2]Department of Biomedical Data Science, Dartmouth Cancer Center, Dartmouth College, Hanover, NH, USA.

[3]Division of Healthcare and Artificial Intelligence, Department of Precision Medicine, Korea National Institute of Health, Korea Disease Control and Prevention Agency, Cheongju, Republic of Korea

[4]Samsung Genome Institute, Samsung Medical Center, Seoul 06351, Republic of Korea

[5]Geninus Inc, Seoul, Republic of Korea

[6]Korea National Institute of Health, Korea Disease Control and Prevention Agency, Cheongju, Republic of Korea

*. Co-1st authors

**. Co-correspondences

Abstract

The global pandemic caused by the SARS-CoV-2 virus provided an unprecedented opportunity to investigate genetic factors influencing the disease severity of the viral infection. Despite a plethora of recent research on both SARS-CoV-2 and COVID-19, few have taken a systems biology approach to address individual-level variation, especially based on non-European populations. Accordingly, we analyzed multi-omics data generated at three timepoints from 193 Korean COVID-19 patients with mild or severe symptoms, composed of whole genome sequencing, blood-based single-cell RNA-sequencing (2.15M cells), 195 cytokine profiles, and human leukocyte antigen (HLA) allele data. We identified expression quantitative trait loci (eQTLs), disease severity interacting eQTLs (n = 388), and disease progression interacting eQTLs (n = 945) for various cell types. We elucidated a complex regulatory mechanism involving *HLA* genes and their targets, and identified genetic determinants of cytokine levels. Finally, we show how regulation of ieQTLs is established by upstream transcription factors (TFs), illustrating complex regulation of the *IGFBP7* ieQTL by a combined action of two TFs, which is potentially important in conferring differential severity. This study illuminates an efficient molecular interrogation framework that can be applied toward understanding infectious disease progression in individuals of different genotypes.

Introduction

Despite their importance in human biology and diseases, elucidation of the genetic mechanisms that influences individual variation in susceptibility to infection and the severity of infectious diseases has been challenging due to the high contribution of non-genetic factors (*i.e.*, low heritability) and involvement of multiple genetic variants.[1] The recent emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and associated global Coronavirus disease 2019 (COVID-19) pandemic offers an unprecedented opportunity to recruit and study large cohorts across diverse ethnic groups and conduct a genetic evaluation of virus-driven disease susceptibility and severity.

Several factors, including age, sex, and comorbidity are known to contribute to COVID-19 susceptibility and severity.[2-5] To address the genetic contribution of differential severity in addition to these factors, the COVID-19 Host Genetics Initiative (HGI; https://www.covid19hg.org/) determined genome-wide associated variants from more than 100K infected individuals with differing COVID-19 severity.[6,7] Another effort took a rare variant analysis approach and elucidated a list of genes that for which certain alleles confer severe symptoms upon the infection.[8,9] However, neither of these studies examined the complex networks of genomic variants and gene expression that are cell-type-specific, which may eventually affect disease severity and prognosis. Moreover, both studies were mainly conducted on individuals of European ancestries, overlooking the genetics of COVID-19 in other ethnic groups. Indeed, COVID-19 studies using non-

European ancestries have successfully identified novel signals that play roles in the virus infection and disease progression.[10-12]

Analysis of expression quantitative traits (eQTLs) allows for identification of genetic variants that influence gene expression in a tissue,[13-15] which can reveal novel insights into the genetic regulation of gene expression.[16,17] This approach has recently gained popularity due to its ability to explain disease-associated GWAS loci (*i.e.*, colocalization).[18,19] However, previous colocalization efforts for COVID-19-associated loci were performed using public databases based on healthy individuals, rather than patient-originated datasets.[18,20-22] Also, a recent study provided genetic evidence that not all GWAS loci can be tied to an eQTL as GWAS loci and eQTLs are inherently different in many genetic aspects,[23] which may explain the low colocalization rate (20-30%). Moreover, most of the previous eQTL studies lacked clinical implication due to analyses being performed on healthy tissues, or on a combined set of healthy and diseased tissues. To circumvent this limitation, we and others have extended the eQTL approach to identify eQTLs that are functional in a disease-, condition-, or cell-state-specific manner. The resulting response-eQTLs and interaction-eQTLs (ieQTLs) have demonstrated their utility in elucidating novel, individual-specific genetic pathways and drug targets.[14,15,24] Furthermore, our multi-omics approach allows the probing of genetic mechanisms underlying the expression of *HLA* genes and cytokines, which are critical factors in the determination of disease severity.

In this study, we analyzed multi-omics data from 193 Korean COVID-19 patients. The patients were classified by disease severity into the mild and severe groups, and serial sampling and data generation yielded 495 runs of single-cell RNA sequencing (scRNA-

seq) and cytokine levels. Both eQTLs and ieQTLs, collectively termed as COVID-eQTLs, were called and integrated with genetic loci previously associated with COVID-19. Through this, we uncovered a number of molecular mechanisms explaining the association of previously-identified or novel factors with infection status. Our approach further brought to light an ieQTL that regulates *IGFBP7* expression through the complex interaction of two transcription factors (TFs), CEBPG and PATZ1, which may contribute to differential severity of disease. Overall, these variants can be effectively utilized to predict the severity of virus infection.

Results

COVID-eQTLs signify COVID-19 severity

The collection of participants and generation of multi-omics data used in this study were previously described.[25] Briefly, Korean participants that were afflicted with COVID-19 during October 2020 to September 2022 were classified into the “Mild” (*n* = 96) or “Severe” (*n* = 97) groups based on the severity of host response to infection following the WHO guideline (Fig. 1a, Supplementary Table 1 and Supplementary Figure 1).[26] Blood samples from each individual were collected three times at two-week intervals, with the first sampling performed on the day that the infection was confirmed; this resulted in six sample groups: mild timepoints M1, M2, and M3 and severe timepoints S1, S2, and S3. The omics data processed for this study included the following: blood-based scRNA-seq, whole genome sequencing (WGS), human leukocyte antigen (HLA) sequencing (HLA-seq), and blood-based cytokine profiles (Fig. 1a and Supplementary Table 2). Using these data, we sought to understand the contributions of genetic variants in determining disease severity following SARS-CoV-2 infection through gene expression regulation. To do this, we called eQTLs and ieQTLs which interact with COVID-19 severity (sv-ieQTL) or progression time point (tp-ieQTL) (Fig. 1b, Supplementary Tables 3 and 4) for various cell types. We will refer to all eQTLs and ieQTLs identified here as COVID-eQTLs.

Using the scRNA-seq data, 2.15M cells from 495 samples across six points were clustered and annotated, revealing the Severe group to have a slight increase in the proportion of monocytes (1.2-fold) and a decrease in that of CD4+ T cells (0.73-fold)

compared to the Mild group (Fig. 1c; Supplementary Table 5), which was consistent with previous observations.[27,28] We also investigated the expression patterns of interferon-stimulated genes (ISGs) in each group across cell types using interferon gene sets from previous studies.[29-31] Consistent with those studies, ISG expression was more increased in the Mild group across all cell types, and the expression declined over time (Supplementary Figure 2).

Principal component analysis (PCA)-mediated clustering using effect size value of eQTLs from major cell type clusters revealed disease severity (represented by PC2) to be a second determinator following cell type (represented by PC1) (Fig. 1d). Meanwhile, PC3 and PC4 divided the samples by myeloid vs. lymphoid origin (Supplementary Figure 3). While a majority of eQTLs recurrently appeared across all major cell types (39,981/71,265 = 56.1%; Fig. 1e), sv- and tp-ieQTLs showed less overlap, being mostly comprised of unique ieQTLs found in single cell types (number of unique vs. all ieQTLs for sv-ieQTLs: 345/388 = 88.9% and tp-ieQTLs: 791/945 = 83.7%; Fig. 1e). This finding suggests that ieQTL may function in a more cell-type-specific manner compared to eQTLs, which provide more shared regulatory activities across multiple cell types. Fig. 1f-h show example cases of eQTL, sv-ieQTL, and tp-ieQTLs.

Colocalization of COVID-eQTLs with COVID-19 GWAS loci

We compared our COVID-eQTL with previously reported bulk-eQTLs identified from peripheral blood samples (*i.e.*, GTEx blood data and Japan COVID-19 data[10,32]). We observed strong correlation of eQTL effect size (Fig. 2a) for eQTLs, indicating that the majority of variant regulatory effects are shared with other datasets across different

genetic ancestries. One of the advantages of eQTL analysis is that eQTL colocalized with GWAS variants can confer the molecular mechanism underlying the genetic association, through expression changes of nearby genes.[33] Therefore, we sought to determine how many of the loci identified from COVID-19 GWAS were explained by our COVID-eQTL set.[6,7] Overall, we observed a high probability of eQTL colocalization with COVID-19 GWAS signal (PP.H4 ≥ 0.8) in 16 genes, with the greatest number of colocalizations occurring for monocyte eQTLs (Fig. 2b). This level of colocalization aligns with findings from previous COVID-19 GWAS-eQTL colocalization studies in people of European ancestries.[18,20-22] Considering that there is only a marginal difference between cell types in terms of the number of COVID-eQTLs called (Fig. 1e), the enrichment of monocyte eQTLs in GWAS variants associated with COVID-19 severity is notable, providing evidence for the importance of the innate immune response in disease severity.[27,28,34] Colocalized eGenes mostly belonged to the *HLA* locus or were previously colocalized genes (Fig. 2c).[20,21,35] Example colocalized loci in the *ABO* and *TOMM7* loci are shown in Fig. 2d. We further conducted a colocalization analysis using GWAS results from immune-related diseases, which revealed eQTLs in B cells and monocytes to harbor the highest numbers of overlapping loci (Fig. 2e). This result implies involvement of cell type-specific determination of COVID-19 severity through comparison with GWAS loci.

COVID-eQTLs regulate genes of high constraint and expression

Next, we analyzed various genetic features in relation to our COVID-eQTL calls. Plotting the distance of COVID-eQTLs from transcription start sites (TSS) revealed eQTLs to be

more centered around TSS compared to ieQTLs, suggesting that the expression regulation posed by ieQTLs is less promoter-centered and therefore more likely to be context-specific (Fig. 3a). When considering COVID-eQTLs according to whether they are common among examined cell types or distinct to a single cell type, we found cell-type-specific COVID-eQTLs to have larger effect sizes than shared COVID-eQTLs with different strengths (eQTLs: two-tailed *t*-test $P = 4.73 \times 10^{-5}$ ieQTLs: two-tailed *t*-test $P = 0.075$; Fig. 3b left, middle and Supplementary Figures 4 and 5). Similarly, when eQTLs were separated by sharing versus uniqueness across mild and severe disease status (*i.e.*, eQTLs), the latter group showed a larger effect sizes (*i.e.*, sv-ieQTLs) (two-tailed *t*-test $P = 2.26 \times 10^{-135}$; Fig. 3b right). This result implies that regulatory variants which influence cell-type specificity or disease severity through changes in gene expression have a more pronounced influence compared to those found in shared cell-type or disease state. Enrichment analysis revealed eGenes to be enriched in COVID-19 and other immune-related pathways, demonstrating our calling pipeline to have identified a collection of biologically plausible genes (Fig. 3c). Enrichments of COVID-eQTLs in epigenetic and genomic features across all major cell types further revealed that they are preferentially found in or near TSS, enhancers, 5'UTRs, and exons (Fig. 3d and Supplementary Figure 6).

Next, we sought to compare the characteristics of genes that behave as ieGenes, eGenes, differentially expressed genes (DEGs), GWAS associated genes, or none of the above. We selected two pathways from the Reactome Pathway database[36] that showed significant enrichment for ieGene in monocytes in relation to disease severity based on ToppGene[37] analysis, innate immune system ($P_{adj} = 1.94 \times 10^{-2}$) and adaptive

immune system (P_{adj} = 1.94 x 10^{-2}) pathways. Then, we compared the gene constraint scores (o/e constraint score from the gnomAD) and expression levels in monocytes of the genes within each pathway (Fig. 3e). Interestingly, genes whose expression was influenced by ieQTLs (*i.e.*, ieGenes) exhibited the highest constraint scores and expression, followed by eGenes, DEGs, and GWAS associated genes. This result was replicated in different cell types, but not in unrelated pathways (Supplementary Figures 7 and 8). This finding suggests that ieQTLs comprise a preferred method of expression regulation for critical and highly expressed genes; this regulatory effect is context-dependent, and hence avoids ubiquitous large fluctuations in expression.

Genetic regulation pertaining to HLA alleles

HLA genes play a central role in antigen processing and presentation to T cells, and their allele type and variation in expression have been closely associated with susceptibility to numerous infections and immune-related diseases,[38-41] including response to COVID-19 infection.[42-44] Here, colocalization analysis revealed high overlap of COVID-eQTLs and COVID-19 GWAS signals in *HLA* genes, supporting the importance of *HLA* gene regulation (Fig. 2c). Interestingly, COVID-eQTLs from the HLA region in chromosome 6 comprise approximately 14% of all calls, further suggesting delicate tuning of the *HLA* gene expression to be an important factor (Supplementary Figure 9). Therefore, we sought to understand the genetic regulation of *HLA* gene expression and its association with COVID-19 response. We categorized regulation involving *HLA* alleles into two scenarios (Fig. 4a): (1) *cis*-regulation, where expression is regulated by the presence of eSNP or *HLA* allele (*i.e.*, eAllele) in antigen presenting

cells (APCs), and (2) *trans*-regulation, where genes within APC or in T cells are regulated in an intra-cellular or inter-cellular manner by *HLA* genes, in the setting of T cell-APC interaction (Supplementary Table 6).

One notable example demonstrates the regulation of *HLA-B* by a nearby SNP (rs3819294) and its allele (HLA-B*15:01). While rs3819294 acts as a universal eQTL, functioning in both the Mild and Severe groups (Figs. 4b and 4c), HLA-B*15:01 shows regulatory effects only in the Mild group, reducing the expression of the gene (Fig. 4d). This finding provides a molecular mechanism underlying the association of HLA-B*15:01 with asymptomatic COVID-19 response.[44] The allele frequency of HLA-B*15:01 in East Asians does not differ too much from that of Europeans (0.073 in our cohort, 0.028-0.085 in East Asians, and 0.022-0.089 in Europeans; Supplementary Table 7),[45] ruling it out from acting as a confounding factor. To further understand how that asymptomatic effect is established, we examined target genes in APCs and T cells for which expression levels is associated with the presence of HLA-B*15:01. Notably, HLA-B*15:01 carriers in the Mild group showed decreased *SPPL2A* expression in B cells (SPPL2A cleaves CD74 intramembrane, which activates the NF-kB pathway and expression of IL-8, both known to be increased in severe COVID-19[30,46-50]), and increased *ILF3* expression in CD8+ T cells (ILF3 activates interferon type 1 signaling[51]) (Fig. 4e and 4f).

Additional representative examples include eQTLs for *HLA-DQA2* in B cells (Supplementary Figure 10) and *HLA-DPB1* in dendritic cells (DCs), the latter being composed of two independent SNPs, rs7757520 and rs13203715, resolved by conditioning analysis (Figs. 4g and 4h). We applied a linear regression model to

characterize *trans*-regulated genes in CD4+ T cells influenced by *HLA-DPB1* expression in DCs (see Methods). Not only were these sets of genes (83 for the Mild and 111 for the Severe group) enriched in ontologies and pathways relevant to virus infection and host responses, those enrichments were much enhanced in the Severe group, suggesting the differential regulation of *HLA-DPB1* (Fig. 4i) to mediate stronger antiviral response and excessive immune activation. Taken together, our set of COVID-eQTLs allowed elucidation of the complex regulation of *HLA* gene expression and its consequences in T cells with regard to conferring differential severity of COVID-19 (Figs. 4f and 4j). Furthermore, the results demonstrated the complementary roles of eQTL and HLA allele-eQTL analyses in identifying genetic variants and their regulatory effects, highlighting the importance of integrating the two approaches.

Genetic regulation of cytokine expression

Levels of circulating cytokines have critical utility in the response to viral infection, including in COVID-19.[52-54] Accordingly we explored blood cytokine levels for protein QTL (pQTL) calling and compared the results with COVID-eQTLs relating to cellular cytokine RNA levels (Supplementary Table 8). Heritability determination revealed similar values for cytokine proteins (mean = 0.04) and transcripts (in the pseudo-bulk level, mean = 0.04) (Fig. 5a and Supplementary Table 9). Notably, cytokines with high heritability were not enriched with COVID-19-related cytokines or DEPs (Wilcoxon rank-sum test P = 0.93 and 0.79, respectively; Supplementary Figure 11), nor associated with protein multimerization, post-translational cleavage, or protein half-life (data not shown). We present IL6R and MICA, which are important in the inflammatory response

to COVID-19,[55,56] harbored strong pQTLs but non-functional eQTLs (Fig. 5b). Direct comparisons of protein and RNA expression patterns revealed enhanced detection and separation by disease severity for some of the cytokines (Fig. 5c and Supplementary Figure 12). Notably, a region in chromosome 17 encompassing *CCL23* and *CCL18* represents an example of independent regulation by two neighboring pQTLs (Fig. 5d). Conditional analysis revealed that the top SNPs for each protein to behave in a completely independent manner, thereby explaining the fact that these proteins have low correlation of expression despite the encoding genes being close to each other (Supplementary Figure 13).

Identification of upstream factors leading to severity-specific eQTLs

It is critical to understand factors that act upstream of many eQTLs as they may represent master regulators that transmit physiological cues in the cell to affect the expression of many downstream genes. However, upstream factors that regulate eQTLs are less studied compared to the downstream effects posed by eQTLs. To examine such factors, we devised a model where differences in TF activity (*i.e.*, "regulons") pose a change in target gene (*i.e.*, ieGenes) expression through binding to a regulatory region with differential binding affinity (*i.e.*, ieSNP) (Fig. 6a). Our initial pool was composed of 388 sv-ieQTLs, called from every examined cell type, and 376 TFs, as defined by SCENIC,[57] totaling 145,888 regulon-ieQTL pairs. After initial filtering by significant changes in TF-DNA binding ability using FIMO[58], 567 regulon-ieQTL pairs remained (Supplementary Table 10). Then, as a stringent filtering strategy, we applied three criteria to identify regulons significantly correlated with (i) differential activity

between the Mild and Severe groups, (ii) differential level of downstream gene (ieGene) expression, and (iii) ieQTL effect size, under the assumption that a change of regulon activity in a cell causes a change in ieQTL effect size (Figs. 6b and 6c). This strategy yielded 18 regulon-ieQTL pairs (Figs. 6b-d), among which we were especially intrigued by the regulation of *IGFBP7* by TFs CEBPG and PATZ1 in monocytes. Although its implication in COVID-19 remains elusive, IGFBP7 poses multifaceted functions in inflammatory contexts, such as acute kidney injury and cancers.[59,60] Both CEBPG and PATZ1 display significant differential DNA binding ability in relation to the genotype of rs881382 (*i.e.*, a Mild-specific ieSNP in monocytes, Supplementary Figure 14), with the CEBPG motif containing the alternative allele (T) whereas the PATZ1 motif contains the reference allele (G) (Fig. 6e). The region flanking rs881382, located in *IGFBP7* intron 1, is an open chromatin region according to ATAC-seq, DNase-seq, and H3K27ac and PATZ1 ChIP-seq from the ENCODE dataset (Fig. 6f and Supplementary Figure 15). Activity of the CEBPG regulon was higher in the Severe group compared to the Mild group (two-tailed *t*-test $P = 3.56 \times 10^{-15}$), whereas activity of the PATZ1 regulon displayed an opposite pattern (two-tailed *t*-test $P = 3.36 \times 10^{-21}$) (Fig. 6g). Expression of *IGFBP7* showed positive correlation with CEBPG as a transcriptional activator (Pearson's $R = 0.66$), while a negative correlation was observed with PATZ1 as a transcriptional repressor (Pearson's $R = -0.39$) (Fig. 6h).[61] Interestingly, the effect size of the *IGFBP7* eQTL was inversely correlated with CEBPG activity (Pearson's $R = -0.42$) but positively correlated with PATZ1 activity (Pearson's $R = 0.79$) (Fig. 6i), causing eQTL effect size to be higher in the Mild group (*i.e.*, establishment of a Mild-specific ieQTL) (Fig. 6j). Based on this finding, we propose a model where combinatorial

alterations of TF activities in disease status lead to changes in downstream gene expression and thereby constitute a sophisticated gene expression regulatory mechanism (Fig. 6k).

Discussion

Genetic susceptibility of host response to virus infection carries profound implications in both clinical and social systems, and the recent COVID-19 pandemic provided an unprecedented opportunity to systematically address such susceptibility. Indeed, many studies investigated differential genetic and transcriptional profiles in patients with varying disease severity.[6-8,18,20,27,62] However, these GWAS variants explain only a limited portion of phenotypic variation, and the rare pathogenic variants are found in only a relatively small portion of patients.

In this study, we used genome, single-cell transcriptome, and blood cytokine data generated from Korean patients infected with COVID-19 to systemically delineate genetic differences between patients with mild or severe responses by calling COVID-eQTLs, or eQTLs and ieQTLs that drive differential gene expression in relation to disease severity and progression. This approach led us to identify 71,686 eQTLs and 1,211 sv- and tp-ieQTLs across seven main cell types of peripheral blood mononuclear cells (PBMCs).

Our finding emphasize the importance and utility of ieQTLs in regulating genes in the context of viral infection. The ieQTL strategy allows consideration of the inherent heterogeneity in cell states among different samples of the same clinical status and hence discovery of variants that regulate gene expressions in a cell-state-dependent manner. This approach was previously utilized in lung and liver diseases to reveal disease-specific regulatory effects and identify pathways associated with cell state.[14,24] Here, we observed genes regulated by ieQTLs to feature both strong constraint and

higher expression than genes in other classes, such as those associated with GWAS loci, DEGs, or non-implicated genes (Fig. 3b and 3e). This observation suggests that genes under control of eQTLs and ieQTLs may perform more essential functions than GWAS genes in the context of COVID-19 response.

Intriguingly, a number of features that signify disease severity were most prominent in monocytes. Monocytes harbored the largest numbers of eQTLs and ieQTLs (Fig, 1e), eQTLs detected from monocytes demonstrated the highest colocalization with COVID-19-related traits (Fig. 2b), and monocytes exhibited the greatest enrichment of TFs and sv-ieQTL pairs (Fig. 6c). These findings add further evidence of the pivotal role of innate immune cells in cytokine production and the determination of COVID-19 severity.[12,28,30,62]

Scrutinization of *HLA* gene expression regulation and its impact on T cell gene expression allowed us to unravel the molecular mechanism underlying the association of *HLA* genotype with disease outcome. In particular, the allele HLA-B:15:01 is known to be associated with asymptomatic COVID-19 and we demonstrated this allele to reduce expression of *HLA-B* and induce CD8+ T cell activation genes only in the Mild group (Figs. 4b-f).[44] We further present a novel involvement of *in trans* of *HLA-DPB1* expression regulation with expression of virus response and interferon signaling in CD4+ T cells (Figs. 4g-j). These results illuminate the utility of eQTL approach in probing the action of *HLA* genes and their implications in disease severity.

Current understanding of the genetic regulation of cytokine levels is limited. Previous studies have reported high pathogen-dependent variability when cells are stimulated, including by virus.[63-65] Our quantitative comparison of genetic influence on cytokine

protein expression in blood and gene expression in PBMCs revealed genetic factors to have similar ability to predict RNA and cytokine levels.

One advantage of eQTL approach is that it can instantly identify pathways whose activity may be dependent on individuals' genotypes. Another is that it enables the elucidation of immediate upstream factors, such as TFs, that directly regulates many eQTLs. IGFBP7 is a versatile secreted protein with roles in acute kidney injury, cancers, and vascular diseases, which it enacts through regulating fibrosis, inflammation, senescence, and oxidative stress.[60,66] It is expressed in many cell types of PBMCs and implicated as an upstream regulator of genes important in inflammation.[67,68] Expression of *IGFBP7* is considered dynamic, with controversial reports supporting different patterns of expression depending on the disease model[68-70]. Through probing a regulatory network involving the ieQTL and upstream regulators of *IGFBP7* expression, we determined that a complex interaction of the two TFs with different activities, CEBPG and PATZ1, with the allele status of rs881382 contributes to the cell-state-specific regulation of *IGFBP7*. We envision that this approach will be used to accurately understand the behavior of many interesting ieQTLs in the future.

While our approach mainly considered the systematic combination of genome and transcriptomic data, investigating the epigenetic regulation of these genes in COVID-19 patients may provide additional mechanistic insights into differential disease severity. For example, probing of open chromatin regions in hospitalized COVID-19 patients revealed an enrichment of CEBP TFs in monocytes, which aligns with our findings.[71] In addition, more fine stratification of cell states (*e.g.*, gene modules) using a larger cohort

may provide further insights into the regulation of genes whose expression relates to disease severity.

In summary, we combined extensive multimodal data from COVID-19 patients and identified differential regulation of gene expression in relation to disease severity across cell types. Our COVID-eQTL analysis pipeline will serve as a novel framework for elucidating the biological determination of disease severity and as a tool for discovering predictive biomarkers, not only for SARS-CoV-2, but also for other future viral pandemics.

Acknowledgements

We appreciate participation of the COVID-19 patients and clinicians involved in the study. We thank Jungmin Choi, Seunghee Hong, Joon Yong An, Yongjin Yoo, and Sung Eun Hong for for their critical reading of the manuscript. This work was partly supported by the Korean Research Foundation (2021R1A2C3014067 to M.C.), by the Development of Heterogeneous Healthcare Data and Artificial Intelligence (2024-NI-009-00), and by the Geisel School of Medicine at Dartmouth's Center for Quantitative Biology through a grant from the US National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM130454 and R35GM154925 (to S.Z.).

Data availability

All clinical data and biospecimens from this project are available from the National Biobank of Korea (NBK) under prior approval (https://nih.go.kr/biobank). Also, omics dataset from this project are deposited at the Clinical and Omics data Archieve (CODA; https://coda.nih.go.kr) via accession numbers CODA_D21001 and CODA_D23017.

Code availability

Code for the analysis is available at https://github.com/snu-mchoi-lab/COVID-QTL. Most analysis are performed using R. The open-source tools used for data analysis are referred in Methods or Supplementary Methods.

Author information

Contributions

W.-Y.P., M.C. conceived the study. J.L., E.Y.J., M.C. designed the experiments and analyses. S.Z., M.C. acquired funding and provided supervision. J.L., E.Y.J. performed snRNA-seq data analysis, genome data processing, eQTL calling, and downstream analyses. L.Y., S.Z. called interaction-eQTLs and performed downstream anayses. J.L., E.Y.J., L.Y., S.Z., M.C. wrote the paper. H.-Y.J., S.C.K., W.-Y.P., H.-Y.P. contributed to data generation, preprocessing, and administrative tasks.

Corresponding authors

These authors jointly supervised this work: Siming Zhao, Murim Choi

Ethics declarations

The authors declare no conflicts of interest that pertain to this work.

Online Methods

Cohort and data generation

Details of patient recruitment, patient classification, sample collection, clinical and genetic data generation and processing procedures were described previously.[25] The study protocol was approved by the Institutional Review Boards (CNUH 2020-12-002-008, SEOUL 2021- 02-016, SMC-2021-03-160, and 2020-09-03-C-A).

Most patients were infected with the B.1.497, B.1,619, or B.1.620 strains, reflecting the major COVID-19 lineages at the time of the patient recruitment.[25] Classification of patients by COVID-19 severity was performed following the NIH treatment guidelines[26]. Patients classified as “severe” presented COVID-19 symptoms with pneumonia and one of the following: respiratory rate of more than 30 breaths/min, oxygen saturation <93% at room temperature, <300 mmHg of PaO2/FiO2, or rapid progression of infiltration (>50%) observed by chest computed tomography imaging. In this study, the Mild group is composed of asymptomatic (*n* = 11), mild (*n* = 35), and moderate (*n* = 50) cases, and the Severe group of severe (*n* = 57) and critical (*n* = 40) cases, as defined previously (Supplementary Table 1).[25]

Use of the data in this study was approved, and the data downloaded from the Clinical & Omics Data Archive (CODA), Korea National Institute of Health, Republic of Korea (https://coda.nih.go.kr/), under the title of “Establishment of multi-omics data related with COVID19”.

scRNA-seq data processing

PBMCs were acquired to construct scRNA-seq libraries using the 10X Chromium Next GEM Single Cell 5' Library kit. Generation of scRNA-seq matrix was described in a previous study.[25] PBMCs from four participants were pooled together at the equal number of cells. Doublet detection was performed using Scrublet to identify potential doublets in the data.[72] Single-cell RNA-sequencing data from patients were processed using Scanpy v1.6.0[73] and Seurat v5.[74] Cells that expressed <200 genes were excluded from the analysis, and genes expressed in fewer than three cells were removed. Cells with >15% mitochondrial RNA content were likewise excluded from further analysis. Data normalization, variable feature selection, and dimensionality reduction were performed with batch effect correction using Harmony.[75] Clustering was conducted at a resolution of two, and uniform manifold approximation and projection (UMAP) was used for visualization. Cell type annotations were obtained via Azimuth,[76] and the sketched assay was projected onto the full dataset for comprehensive integration.

WGS data processing

To call genetic variants in the participants, WGS reads were aligned to the human genome reference build GRCh38. gVCF files were generated as described previously[25] and subsequent steps for consolidation of gVCFs, joint-calling, and variants recalibration were conducted using the GATK 4.1.6.0 "germline short variant discovery" pipeline.[77] Only variants passing the recalibration step were included. Additional filtering steps to remove low-quality variants included, excluding low-complexity regions,[78] applying allelic balance cutoffs of 0.22 for SNPs and 0.2 for indels, restricting indel sizes to ≤50bp, and including only biallelic variants. Samples were pruned based on sex

discordancy, missing genotype rate >0.05, kinship coefficients >0.177, and population stratification adjustments using PCA (Supplementary Figure 1). Population stratification was evaluated using the 1000 Genomes Project and the Korean Variant Archive (KOVA).[79]

Pseudobulk eQTL and interaction-eQTL calling

For eQTL analysis, samples with both genotype and phenotype data available were included. Preprocessing steps for eQTL input were conducted separately for each combination of condition and cell type. To ensure robust signal detection, seven major cell types were analyzed (B cells, CD4+ T cells, CD8+ T cells, NK cells, dendritic cells, monocytes, and other T cells). For each combination, count matrices were aggregated at the individual level, followed by quantile normalization across samples and inverse normal transformation for gene expression values. Corresponding genotype data for each combination were extracted and filtered based on the following criteria: missingness <5%, minor allele frequency (MAF) > 0.05, Hardy-Weinberg equilibrium $P > 1.0 \times 10^{-5}$, and LD clumping using an R^2 threshold of 0.9, sorting variants by their MAF. Age, sex, the top three genotype PCs and expression PCs generated from PCAforQTL were used as covariates.[80] For the cis-eQTL mapping, TensorQTL was used with 'cis_nominal' mode.[81]

To assess the sharing patterns of eQTLs across cell types and increase eQTL identification power, the Multivariate Adaptive Shrinkage (MASH) method was applied for each condition.[82] In this analysis, the 15,000 tests with the lowest 'pval_nominal' values from TensorQTL were designated as the strong set, and a random subset of

200,000 tests was used. Following MASH procedures, we learned correlation structure among null tests using the random subset and data-driven covariance matrices using the strong subset. We then fit the mashr model to the random subset to estimate prior parameters, and subsequently derived posterior summaries for the strong set and the complete dataset. eQTLs with local false sign rate (LFSR) under 0.05 were used for further analysis.

For the interaction-eQTL (ieQTL) analysis, either severity status (mild or severe) or timepoint (1,2, or 3) was used as the interaction term in a linear mixed model (LMM). The model was defined as follows:

$$Y = G + sex + age + genoPCs + gexPCs + Cond + G \times Cond + \mu_{sample} + \square,$$

where Y refers to gene expression, G to genotype, and sex, age, genoPCs (genotype PCs), and gexPCs (expression PCs) are included as covariates. The term Cond refers to the condition variable (severity or timepoint), and μ_{sample} represents the random effect for samples. For sv-ieQTL analysis, one LMM with severity as the condition variable was constructed (M-S). To capture additional ieQTL signals, pairwise comparison of severity-specific conditions (M1-S1, M2-S2, and M3-S3) was also performed. For tp-ieQTL analysis, pairwise comparisons between timepoints (M1-M2, M1-M3, M2-M3, S1-S2, S1-S3, and S2-S3) were performed using the LMM model. To incorporate cell-type sharing patterns and increase power, MASH analysis was applied similarly as in eQTL calling, and a threshold of LFSR < 0.05 was used to determine statistical significance.

Comparative analyses with other eQTL sets and GWAS

To assess the correlations of our eQTL data with external datasets, we obtained bulk whole blood eQTL data from the GTEx v8 dataset and the Japan Taskforce.[10] A lead eSNP per gene for each cell type and shared across M1, M3, S1, S2, and S3 was extracted based on nominal *P*-values from TensorQTL. Then, we searched for those SNPs in external eQTLs for an exact match comparison and quantified the alignment using Pearson correlation coefficients.

To evaluate colocalization of GWAS signal with our eQTL dataset, the COLOC package was used.[83] The priors for the probabilities of casual associations (p1, p2, and p3) were set to 1×10^{-4}, 1×10^{-4}, and 1×10^{-5}, respectively. COVID-19 GWAS data were obtained from the COVID-19 Host Genetics Initiative (HGI; https://www.covid19hg.org/), while GWAS data for other immune diseases were sourced from the GWAS Catalog (https://www.ebi.ac.uk/gwas/ and Supplementary Table 11). Genomic regions with $P \leq 1.0 \times 10^{-5}$ and extending 250 Kbp upstream and downstream of each locus were extracted. Genes within these regions were considered for analysis if they contained significant eQTL calls, as identified by MASH[82]. COLOC analysis was then conducted on these genes, provided there were at least 20 SNPs shared between the GWAS and eQTL datasets.[84]

Analysis of gene features

Gene Ontology, KEGG pathway, and gene set enrichment analysis (GSEA) analyses were performed for eGenes from each condition and cell type using clusterProfiler.[85] Gene members for Reactome pathways were download from 'https://www.gsea-

msigdb.org/gsea/index.jsp'. Reactome pathway titles and IDs used in this study are as follows:

REACTOME_INNATE_IMMUNE_SYSMTE [R-HSA-168249; M1036]

REACTOME_ADAPTIVE_IMMUNE_SYSTEM [R-HSA-1280218; M1058]

REACTOME_ CELL_CYCLE [R-HSA-1640170; M543]

REACTOME_MUSCLE_CONSTRACTION [R-HSA-397014; M5485]

REACTOME_REPRODUCTION [R-HSA-1474165; M26956]

REACTOME_METABOLISM_OF_LIPIDS [R-HSA-556833; M27451]

To assess the enrichment of eQTLs in epigenomic markers and genomic annotations, we utilized the *fenrich* module from QTLtools.[86] Chromatin state data were obtained from the Roadmap Epigenomics Project using the 15-state model, focusing on the five immune cell types with cell annotations relevant to the project (*i.e.*, B [E032], NK [E046], Mono [E029], CD8+ T [E048, E047], and CD4+ T [E045, E044, E039, E037, E042])[87]. Genomic annotations was performed according to the Comprehensive gene annotation from GENCODE release 26.[88] All isoforms of each gene were collapsed into a single transcript using the *collapse_annotation.py* script from the gtex-pipeline provided by the Broad Institute ('https://github.com/broadinstitute/gtex-pipeline/tree/master/gene_model'). Enrichment tests were performed on lead SNPs using QTLtools with the '--permute 1000' option.

HLA genotyping and calling HLA-allele eQTLs

HLA genotyping was performed as described previously.[25] HLA data were aligned to human genome build GRCh38, and HLA types for each sample were determined based

on the International IMGT/HLA database.[89] Ambiguous HLA calls, entries lacking heterozygous position labels, and missing data were filtered out and the first two fields were used. To maximize the number of samples with complete HLA type information, we focused on the following loci: *HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1,* and *HLA-DRB1*.

For HLA allele-eQTL analysis, we utilized samples with paired HLA allele and scRNA-seq data (Supplementary Table 2). To leverage the HLA type data as genotype input for HLA allele-eQTL analysis, we included only HLA types with MAF ≥ 0.05. A dominant model was applied, treating homozygous calls equivalently to heterozygous calls. Dosage values were assigned as 0 for non-carriers and 1 for carriers. For the expression data, additional filtering was performed on the pseudo-bulk gene expression matrix used for sc-eQTL analysis. Genes were excluded if they were expressed in fewer than three cells, showed no expression variance across individuals, or exhibited a single expression level in over 65% of individuals. Covariates included age, sex and gene expression PCs. TensorQTL was used for HLA allele-eQTL analysis.

To estimate the density of eQTL calls within the *HLA* locus, we tallied the number of eQTLs located within 500 bp upstream and 500 bp downstream of each eSNP. These eQTL density values were then fitted to a negative binominal model, from which the predicted mean and dispersion parameter were extracted. Finally, for each density count, the probability of observing a value greater than the count was calculated using the cumulative distribution function of the negative binomial distribution.

To identify genes in T cells whose expression is associated with *HLA* expression in antigen processing cells (APCs), an individual-level pseudo-bulk gene expression

matrix was used to align putative interacting pairs of T cells and APCs. Afterwards, linear regression was performed with SNP genotype, age, sex, and genotype PCs as covariates. Genes with FDR below 0.05 were utilized for gene enrichment analysis.

Cytokine data processing, pQTL calling, and heritability estimation

Genotype data and quantifications of 195 cytokines data were collected from 395 individuals as described previously (Supplementary Table 2).[25] Briefly, a customized panel based on the Luminex MAGPIX platform (LUMINEX Corporation, USA) was used to measure cytokines in plasma from the participants. For the Mild groups, data were available for 269 individuals at timepoint 1, 104 individuals at timepoint 2, and 187 individuals at timepoint 3. For the Severe group, data were collected from 126 individuals at timepoint 1, 113 individuals at timepoint 2, and 97 individuals at timepoint 3. Raw cytokine levels were first log-transformed and then mapped to genotype data using a linear regression model, with age and sex as covariates. Cytokines with zero values in more than 15% of the samples were excluded. To call protein QTLs (pQTLs), cytokine identifiers were converted to corresponding gene names, and the gene positions were annotated with GRCh38. Due to a high rate of false-positive calls, the region on chromosome 17 from positions 508,668 to 714,839 was excluded from the analysis. A Bonferroni-adjusted *P*-value of < 0.05 was used as the threshold to identify significant pQTLs.

Genomic heritability of cytokine levels was estimated using Genome-wide Complex Trait Analysis (GCTA) version 1.93.2.[90] The genetic relationship matrix (GRM) was constructed for each cytokine using SNPs located within its *cis*-region, including only

the Mild 1 and Severe 1 groups, by applying the --grm option. Log-transformed cytokine levels from all disease severities and timepoints served as the phenotypic variables. Phenotypic data for each cytokine were input separately in phenotype files. Age, sex, and severity were included as covariates for estimating the proportions of phenotypic variance explained by the SNPs. Heritability estimates were obtained using the restricted maximum likelihood method.

Identification of upstream regulators of sv-ieQTLs

To identify putative upstream regulators of sv-ieQTL, firstly SCENIC[57] was used to identify regulon activities across cell types and conditions. Since there are 2.15M cells in our scRNA matrix, we leveraged the Python-based geosketch[91] to subset an informative set of 200K cells, retaining only cells with more than 200 genes, less than 4,000 genes, and less than 15% of transcripts of mitochondrial origin; additionally, only genes with expression detected in more than three cells were used. The sketched matrix was then extracted in loom format and utilized for SCENIC analysis. Subsequently, we proposed several criteria to narrow the set of putative upstream regulators: (1) significant changes in TF DNA binding ability due to the presence of ieSNP, (2) absolute correlation more than 0.3 for regulon activity and ieGene expression across individuals, (3) absolute correlation more than 0.3 for regulon activity and eQTL effect size across the six conditions, and (4) regulon with activity biased to either mild or severe status. Binding motif occurrences were identified using FIMO from the MEME Suite program.[58] Flanking genomic sequences 15 bp upstream and downstream of each ieSNP were used as input, and binding was considered to occur

for ieSNP-TF pairs with *P*-values less than 1 x 10^{-4}. We used only those pairs that displayed a differential binding profile in relation to the ieSNP reference and alternative alleles. Regulon activity bias dependent on severity status was quantified by logistic regression using the *glm* function (R version 4.4.1). Degree of bias was estimataed as the log of the odd ratio, with an absolute value of at least 50 considered to indicate differential regulon activity.

For visualization of epigenetic data in the UCSC genome browser, we downloaded call sets from the ENCODE portal[92] (https://www.encodeproject.org/) with the following identifiers: ENCFF060KZZ (ATAC-seq), ENCFF778SYZ (DNse-seq), ENCFF580XLE (CRE), ENCFF503MRH (H3K27ac ChIP-seq), ENCSR549PVK (PATZ1 ChIP-seq), and ENCSR490LWA (CEBPG ChIP-seq). ATAC-seq, DNase-seq, and H3K27ac ChIP-seq positioned were converted to GRCh38 coordinates with ucsc-liftover.

References

1. Kwok, A.J., Mentzer, A. & Knight, J.C. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nat Rev Genet* **22**, 137-153 (2021).
2. O'Driscoll, M. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**, 140-145 (2021).
3. Takahashi, T. *et al.* Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* **588**, 315-320 (2020).
4. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054-1062 (2020).
5. Docherty, A.B. *et al.* Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* **369**, m1985 (2020).
6. Initiative, C.-H.G. Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472-477 (2021).
7. Initiative, C.-H.G. A second update on mapping the human genetic architecture of COVID-19. *Nature* **621**, E7-E26 (2023).
8. Zhang, Q., Bastard, P., Effort, C.H.G., Cobat, A. & Casanova, J.L. Human genetic and immunological determinants of critical COVID-19 pneumonia. *Nature* **603**, 587-598 (2022).
9. Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**(2020).
10. Wang, Q.S. *et al.* The whole blood transcriptional regulation landscape in 465 COVID-19 infected samples from Japan COVID-19 Task Force. *Nat Commun* **13**, 4830 (2022).
11. Li, Y. *et al.* Genome-wide association study of COVID-19 severity among the Chinese population. *Cell Discov* **7**, 76 (2021).
12. Edahiro, R. *et al.* Single-cell analyses and host genetics highlight the role of innate immune cells in COVID-19 severity. *Nat Genet* **55**, 753-767 (2023).
13. Vosa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300-1310 (2021).
14. Natri, H.M. *et al.* Cell-type-specific and disease-associated expression quantitative trait loci in the human lung. *Nat Genet* **56**, 595-604 (2024).
15. Yoo, T. *et al.* Disease-specific eQTL screening reveals an anti-fibrotic effect of AGXT2 in non-alcoholic fatty liver disease. *J Hepatol* **75**, 514-523 (2021).
16. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
17. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
18. D'Antonio, M. *et al.* SARS-CoV-2 susceptibility and COVID-19 disease severity are associated with genetic variants affecting gene expression in a variety of tissues. *Cell Rep* **37**, 110020 (2021).
19. Baranova, A., Cao, H. & Zhang, F. Unraveling Risk Genes of COVID-19 by Multi-Omics Integrative Analyses. *Front Med (Lausanne)* **8**, 738687 (2021).

681 20. Schmiedel, B.J. *et al.* COVID-19 genetic risk variants are associated with expression of
682 multiple genes in diverse immune cell types. *Nat Commun* **12**, 6760 (2021).
683 21. Aquino, Y. *et al.* Dissecting human population variation in single-cell responses to SARS-
684 CoV-2. *Nature* **621**, 120-128 (2023).
685 22. Ota, M. *et al.* Dynamic landscape of immune cell-specific gene regulation in immune-
686 mediated diseases. *Cell* **184**, 3006-3021 e17 (2021).
687 23. Mostafavi, H., Spence, J.P., Naqvi, S. & Pritchard, J.K. Systematic differences in discovery
688 of genetic effects on gene expression and complex traits. *Nat Genet* **55**, 1866-1875
689 (2023).
690 24. Hong, S.E. *et al.* Identification of genes conferring individual-level variation responsible
691 for metabolic dysfunction-associated steatohepatitis using single-cell eQTL analysis.
692 *bioRxiv*, 2024.09.20.614203 (2024).
693 25. Jo, H.Y. *et al.* Establishment of the large-scale longitudinal multi-omics dataset in COVID-
694 19 patients: data profile and biospecimen. *BMB Rep* **55**, 465-471 (2022).
695 26. COVID-19 Treatment Guidelines Panel. Coronavirus Disease 2019 (COVID-19) Treatment
696 Guidelines. National Institutes of Health. Available at
697 https://www.covid19treatmentguidelines.nih.gov/. Accessed [6/26/2024].
698 27. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-
699 19. *Nat Med* **27**, 904-916 (2021).
700 28. Ren, X. *et al.* COVID-19 immune features revealed by a large-scale single-cell
701 transcriptome atlas. *Cell* **184**, 1895-1913 e19 (2021).
702 29. Wilk, A.J. *et al.* A single-cell atlas of the peripheral immune response in patients with
703 severe COVID-19. *Nat Med* **26**, 1070-1076 (2020).
704 30. Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in
705 severe COVID-19 patients. *Science* **369**, 718-724 (2020).
706 31. Combes, A.J. *et al.* Global absence and targeting of protective immune states in severe
707 COVID-19. *Nature* **591**, 124-130 (2021).
708 32. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human
709 tissues. *Science* **369**, 1318-1330 (2020).
710 33. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to
711 enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).
712 34. Martinez, F.O., Combes, T.W., Orsenigo, F. & Gordon, S. Monocyte activation in systemic
713 Covid-19 infection: Assay and rationale. *EBioMedicine* **59**, 102964 (2020).
714 35. D'Antonio, M. *et al.* SARS-CoV-2 susceptibility and COVID-19 disease severity are
715 associated with genetic variants affecting gene expression in a variety of tissues. *Cell*
716 *Rep* **39**, 110968 (2022).
717 36. Milacic, M. *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* **52**,
718 D672-D678 (2024).
719 37. Chen, J., Bardes, E.E., Aronow, B.J. & Jegga, A.G. ToppGene Suite for gene list
720 enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305-11
721 (2009).
722 38. Trowsdale, J. & Knight, J.C. Major histocompatibility complex genomics and human
723 disease. *Annu Rev Genomics Hum Genet* **14**, 301-23 (2013).

724 39. Johansson, T., Partanen, J. & Saavalainen, P. HLA allele-specific expression: Methods,
725 disease associations, and relevance in hematopoietic stem cell transplantation. *Front*
726 *Immunol* **13**, 1007425 (2022).
727 40. Kulkarni, S. *et al.* Differential microRNA regulation of HLA-C expression and its
728 association with HIV control. *Nature* **472**, 495-8 (2011).
729 41. Krishna, C. *et al.* The influence of HLA genetic variation on plasma protein expression.
730 *Nat Commun* **15**, 6469 (2024).
731 42. Tavasolian, F. *et al.* HLA, Immune Response, and Susceptibility to COVID-19. *Front*
732 *Immunol* **11**, 601886 (2020).
733 43. Lindeboom, R.G.H. *et al.* Human SARS-CoV-2 challenge uncovers local and systemic
734 response dynamics. *Nature* **631**, 189-198 (2024).
735 44. Augusto, D.G. *et al.* A common allele of HLA is associated with asymptomatic SARS-CoV-
736 2 infection. *Nature* **620**, 128-136 (2023).
737 45. Gonzalez-Galarza, F.F. *et al.* Allele frequency net database (AFND) 2020 update: gold-
738 standard data classification, open access genotype data and new query tools. *Nucleic*
739 *Acids Res* **48**, D783-D788 (2020).
740 46. Li, L. *et al.* Interleukin-8 as a Biomarker for Disease Prognosis of Coronavirus Disease-
741 2019 Patients. *Front Immunol* **11**, 602395 (2020).
742 47. Schneppenheim, J. *et al.* The intramembrane protease SPPL2a promotes B cell
743 development and controls endosomal traffic by cleavage of the invariant chain. *J Exp*
744 *Med* **210**, 41-58 (2013).
745 48. Beisner, D.R. *et al.* The intramembrane protease Sppl2a is required for B cell and DC
746 development and survival via cleavage of the invariant chain. *J Exp Med* **210**, 23-30
747 (2013).
748 49. Matza, D., Wolstein, O., Dikstein, R. & Shachar, I. Invariant chain induces B cell
749 maturation by activating a TAF(II)105-NF-kappaB-dependent transcription program. *J*
750 *Biol Chem* **276**, 27203-6 (2001).
751 50. Binsky, I. *et al.* IL-8 secreted in a macrophage migration-inhibitory factor- and CD74-
752 dependent manner regulates B cell chronic lymphocytic leukemia survival. *Proc Natl*
753 *Acad Sci U S A* **104**, 13408-13 (2007).
754 51. Watson, S.F., Bellora, N. & Macias, S. ILF3 contributes to the establishment of the
755 antiviral type I interferon program. *Nucleic Acids Res* **48**, 116-129 (2020).
756 52. Jing, X. *et al.* Association between inflammatory cytokines and anti-SARS-CoV-2
757 antibodies in hospitalized patients with COVID-19. *Immun Ageing* **19**, 12 (2022).
758 53. Calvo-Alvarez, E. *et al.* Multiplex array analysis of circulating cytokines and chemokines
759 in COVID-19 patients during the first wave of the SARS-CoV-2 pandemic in Milan, Italy.
760 *BMC Immunol* **25**, 49 (2024).
761 54. Liu, Q.Q. *et al.* Cytokines and their relationship with the severity and prognosis of
762 coronavirus disease 2019 (COVID-19): a retrospective cohort study. *BMJ Open* **10**,
763 e041471 (2020).
764 55. Farzad, F., Yaghoubi, N., Jabbari-Azad, F., Mahmoudi, M. & Mohammadi, M. Prognostic
765 Value of Serum MICA Levels as a Marker of Severity in COVID-19 Patients. *Immunol*
766 *Invest* **51**, 1856-1866 (2022).

767 56. Group, R.C. Baricitinib in patients admitted to hospital with COVID-19 (RECOVERY): a
768 randomised, controlled, open-label, platform trial and updated meta-analysis. *Lancet*
769 **400**, 359-368 (2022).
770 57. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat*
771 *Methods* **14**, 1083-1086 (2017).
772 58. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif.
773 *Bioinformatics* **27**, 1017-8 (2011).
774 59. Yu, J.T. *et al.* Insulin-like growth factor binding protein 7 promotes acute kidney injury
775 by alleviating poly ADP ribose polymerase 1 degradation. *Kidney Int* **102**, 828-844 (2022).
776 60. Zhang, L. *et al.* IGFBP7 inhibits cell proliferation by suppressing AKT activity and cell
777 cycle progression in thyroid carcinoma. *Cell Biosci* **9**, 44 (2019).
778 61. Abramova, A. *et al.* The transcription factor MAZR preferentially acts as a transcriptional
779 repressor in mast cells and plays a minor role in the regulation of effector functions in
780 response to FcepsilonRI stimulation. *PLoS One* **8**, e77677 (2013).
781 62. Yao, C. *et al.* Cell-type-specific immune dysregulation in severely ill COVID-19 patients.
782 *Cell Rep* **34**, 108943 (2021).
783 63. Brodin, P. *et al.* Variation in the human immune system is largely driven by non-
784 heritable influences. *Cell* **160**, 37-47 (2015).
785 64. Li, Y. *et al.* A Functional Genomics Approach to Understand Variation in Cytokine
786 Production in Humans. *Cell* **167**, 1099-1110 e14 (2016).
787 65. Li, Y. *et al.* Inter-individual variability and genetic influences on cytokine responses to
788 bacteria and fungi. *Nat Med* **22**, 952-60 (2016).
789 66. Wang, X. *et al.* IGFBP7 regulates sepsis-induced epithelial-mesenchymal transition
790 through ERK1/2 signaling. *Acta Biochim Biophys Sin (Shanghai)* **51**, 799-806 (2019).
791 67. Chen, L., Hui, L. & Li, J. The multifaceted role of insulin-like growth factor binding protein
792 7. *Front Cell Dev Biol* **12**, 1420862 (2024).
793 68. Katoh, M. *et al.* Vaccine Therapy for Heart Failure Targeting the Inflammatory Cytokine
794 Igfbp7. *Circulation* **150**, 374-389 (2024).
795 69. Mo, W., Deng, L., Cheng, Y., Ge, S. & Wang, J. IGFBP7 regulates cell proliferation and
796 migration through JAK/STAT pathway in gastric cancer and is regulated by DNA and RNA
797 methylation. *J Cell Mol Med* **28**, e70080 (2024).
798 70. Qiu, L. & Huang, Z. Downregulation of IGFBP7 Alleviates LPS-induced Inflammation and
799 Apoptosis in WI-38 Cells via Enhancing Mitophagy. *Cell Biochem Biophys* (2024).
800 71. Zhang, B. *et al.* Altered and allele-specific open chromatin landscape reveals epigenetic
801 and genetic regulators of innate immunity in COVID-19. *Cell Genom* **3**, 100232 (2023).
802 72. Wolock, S.L., Lopez, R. & Klein, A.M. Scrublet: Computational Identification of Cell
803 Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291 e9 (2019).
804 73. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data
805 analysis. *Genome Biol* **19**, 15 (2018).
806 74. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell
807 analysis. *Nat Biotechnol* **42**, 293-304 (2024).
808 75. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with
809 Harmony. *Nat Methods* **16**, 1289-1296 (2019).

76. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e29 (2021).
77. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
78. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843-51 (2014).
79. Lee, S. *et al.* Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. *Sci Rep* **7**, 4287 (2017).
80. Zhou, H.J., Li, L., Li, Y., Li, W. & Li, J.J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* **23**, 210 (2022).
81. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol* **20**, 228 (2019).
82. Urbut, S.M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187-195 (2019).
83. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
84. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 1273-1300 (2020).
85. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-7 (2012).
86. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8**, 15452 (2017).
87. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
88. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).
89. Robinson, J. *et al.* IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res* **31**, 311-4 (2003).
90. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
91. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. *Cell Syst* **8**, 483-493 e7 (2019).
92. Sloan, C.A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**, D726-32 (2016).

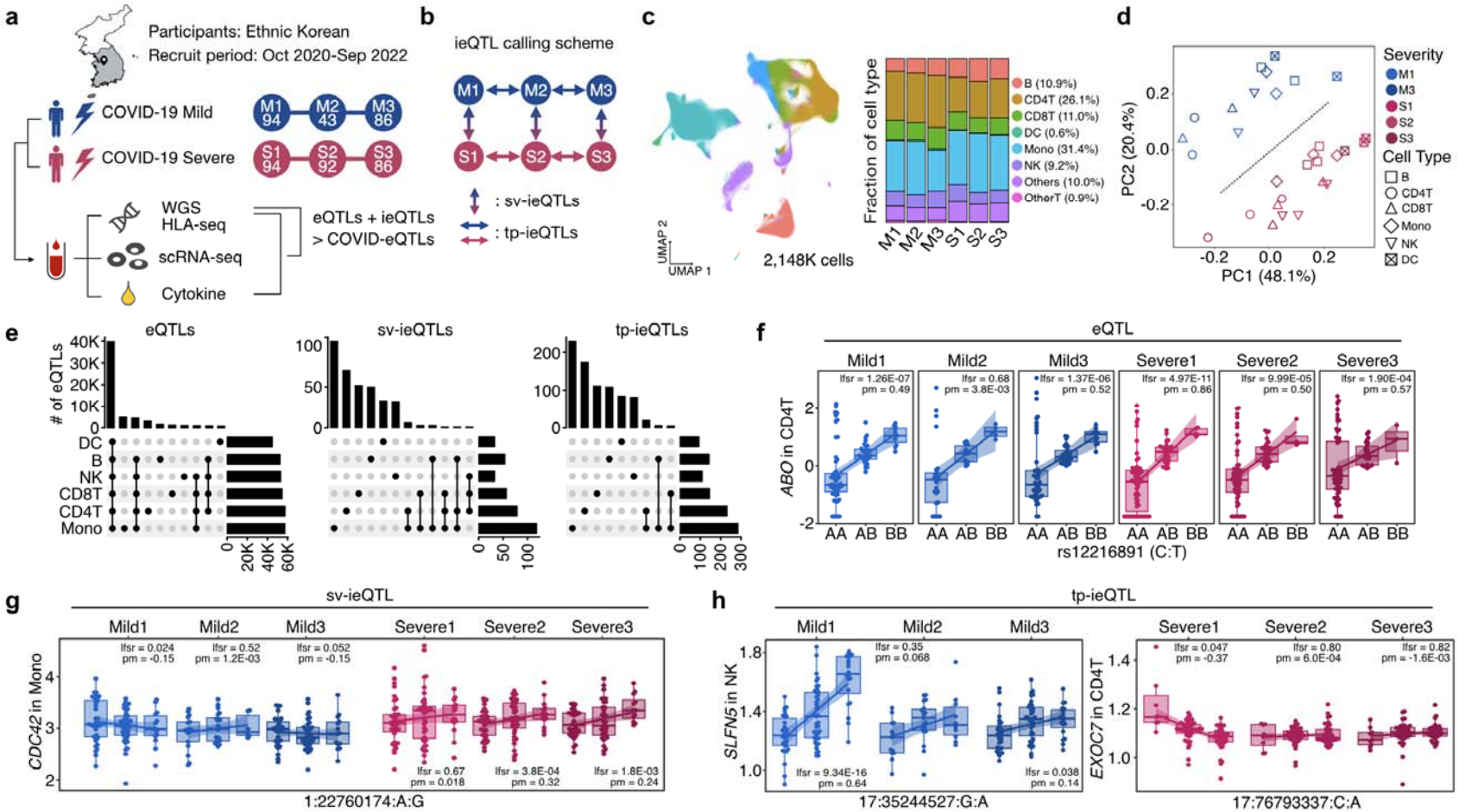

Figure 1. COVID-eQTL study scheme. a, Cohort description and data generation scheme. **b,** ieQTL calling scheme. sv, severity; tp, timepoint. **c,** UMAP (left) and bar plot (right) displaying PBMC scRNA-seq profile of the participants. **d,** PCA clustering of cell types, showing division by cell type and severity. **e,** Counts of eQTLs (left) and ieQTLs across cell types.

f, An example eQTL (rs12216891) that regulates *ABO* gene expression. **g,** An example sv-ieQTL that regulates *CDC42* gene expression only in the Severe condition. **h,** Example tp-ieQTLs that regulate *SLFN5* and *EXOC7* gene expression at different timepoints. lfsr, local false signe rate; pm, posterior mean.

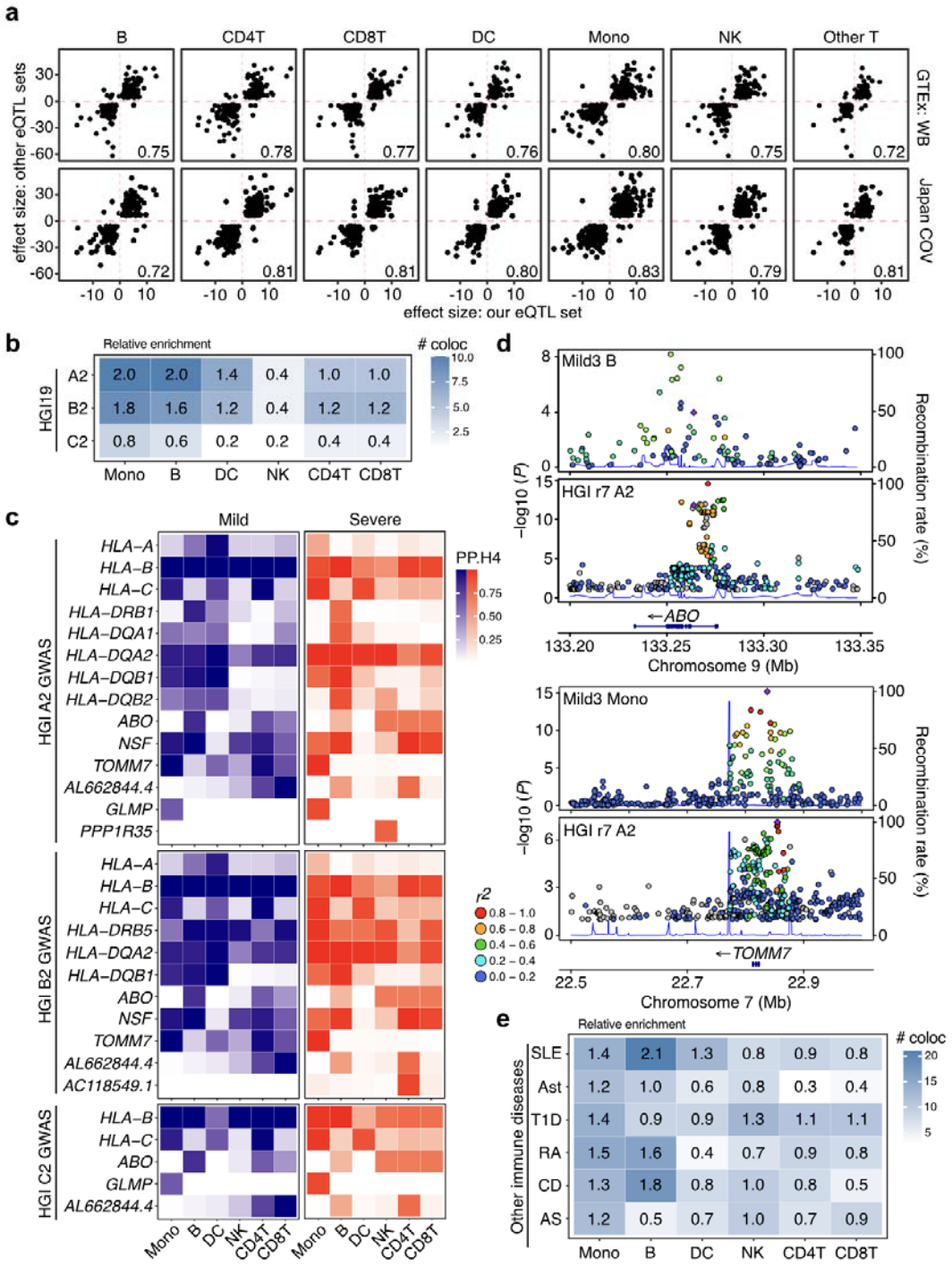

Figure 2. Replication of COVID-eQTLs. a, Scatter plot of β values in the COVID-eQTL set versus the GTEx whole blood (GTEx: WB) and Japan COV datasets. Number in the lower right corner denotes correlation coefficient. **b,** Heatmap depicting enrichment of

colocalization signals for COVID-19 GWAS by cell type. HGI COVID-19 GWAS codes denote as follows: A2, critically ill COVID-19 patients vs. population controls, B2, hospitalized COVID-19 patients vs. non-hospitalized COVID-19 patients, and C2, reported SARS-CoV-2 infected individuals vs. population controls. Values on tiles represent the relative enrichment of colocalized eQTLs for each GWAS phenotype in the given cell types. **c,** Heatmap depicting the result of colocalization of eQTLs with A2, B2 and C2. eGenes of colocalized eQTLs are displayed by cell type with the posterior probability H4 (PP.H4) value. **d,** Regional plots for two example colocalized loci, showing COVID-eQTL (top) and COVID-19 GWAS (bottom) signals for each locus. **e,** Heatmap depicting enrichment of colocalization signals for autoimmune diseases by cell type. SLE, systemic lupus erythematosus; Ast, asthma; T1D, type 1 diabetes; RA, rheumatoid arthritis; CD, Crohn's disease; AS, ankylosing spondylitis.

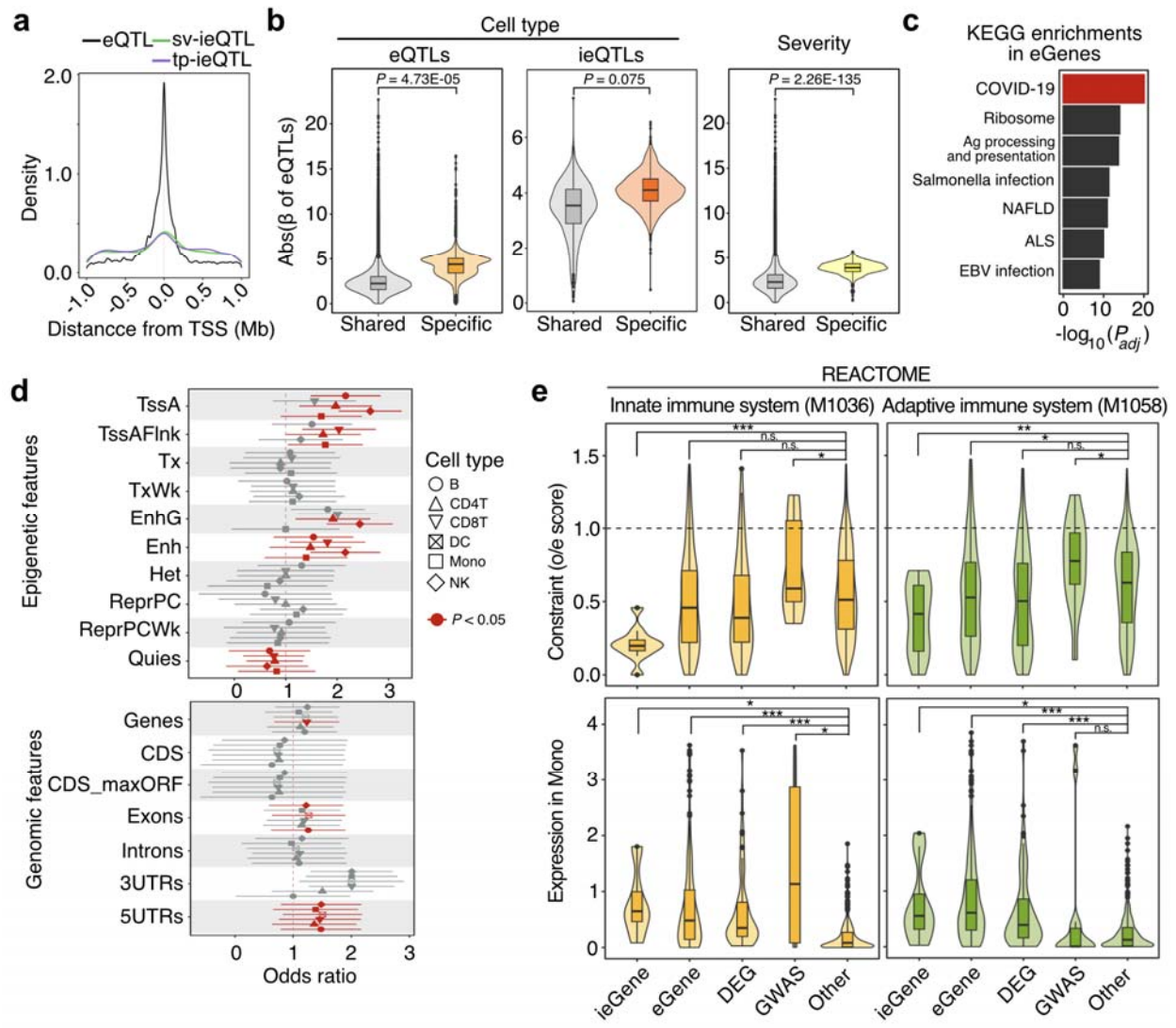

Figure 3. Genetic features of COVID-eQTLs. a, Distribution of COVID-eQTL positions relative to TSS. **b,** Difference in mean of β values (*i.e.*, effect sizes) between shared and specific eQTLs according to cell types (left: eQTLs, middle: ieQTLs) or condition (right). **c,** KEGG pathways enriched among eGenes. Ag, antigen; NAFLD, non-alcoholic fatty liver disease; ALS, amyotrophic lateral sclerosis; EBV, Epstein-Barr virus. **d,** Enrichment of eQTLs in epigenetic and genomic features. Red dots denote significant enrichment ($P < 0.05$). Error bars represent standard error. TssA, active TSS; TssAFlnk, flanking active TSS; Tx, strong transcription; TxWk, weak transcription; EnhG, genetic

enhancer; Enh, enhancer; Het, heterochromatin; ReprPC, repressed polycomb; ReprPCWk, weakly repressed polycomb; Quies, quiescent. **e,** Distribution of gene constraint scores (o/e constraint score) and expression levels in monocytes by genes categorization: ieGenes, eGenes, DEGs, GWAS-associated, or other. Genes are members of the innate immune system (left; Reactome ID: M1036) or adaptive immune system (right; Reactome ID: M1058) pathways.

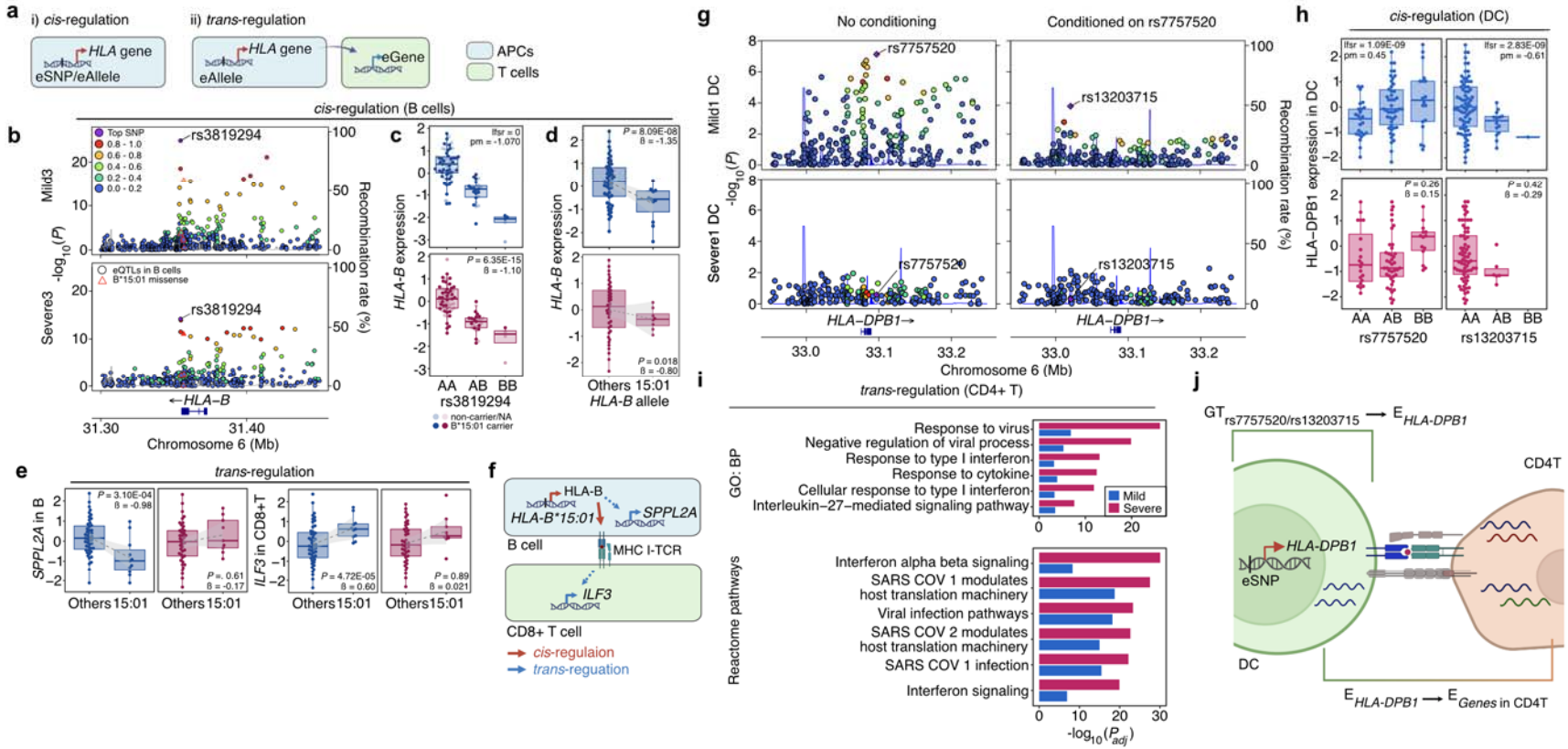

Figure 4. Gene expression regulatory network involving *HLA* gene expression. a, Schematic of the two modes of gene expression regulation. **b,** Regional plots of *HLA-B* locus eQTLs in B cells from Mild 3 (top) and Severe 3 (bottom) timepoints. **c,** *cis*-regulation of *HLA-B* expression by an eSNP (rs3819294). **d,** *cis*-regulation of *HLA-B* expression by the HLA-B:15:01 allele. **e,** Alteration of *SPPL2A* expression by *HLA-B* allele in B cells, and by *ILF3* in CD8+ T cells. **f,** Schematic diagram of the effect of HLA-B*15:01. **g,** Regional distribution of eQTLs near *HLA-DBP1* in Mild 1 and Severe

1 B cells, with (left) or without (right) conditioning on rs7757520. **h,** Expression of *HLA-DPB1* in relation to rs7757520 (left) and rs13203715 (right) genotype. **i,** GO and pathway enrichment analysis of CD4+ T cell genes that correlate with *HLA-DBP1* expression in DCs. **j,** Diagram of a model leading to the identification of a pathway that changes genes in CD4+ T cells due to change in *HLA-DPB1* expression in APCs.

Figure 5. Regulation of blood cytokine production. a, Distribution of heritability values for cytokine protein level (Protein) or single cell-based pseudo-bulk cytokine gene expression level (RNA). Cytokines showing different expression between the Mild and Severe groups are marked as DEPs. **b,** Cytokines (IL6R and MICA) showing strong pQTLs and weak eQTLs. **c,** Scatter plot of TNF and IL6R protein versus RNA expression. **d,** A locus in chromosome 17 showing pQTLs with strong but independent effects on CCL23 and CCL18.

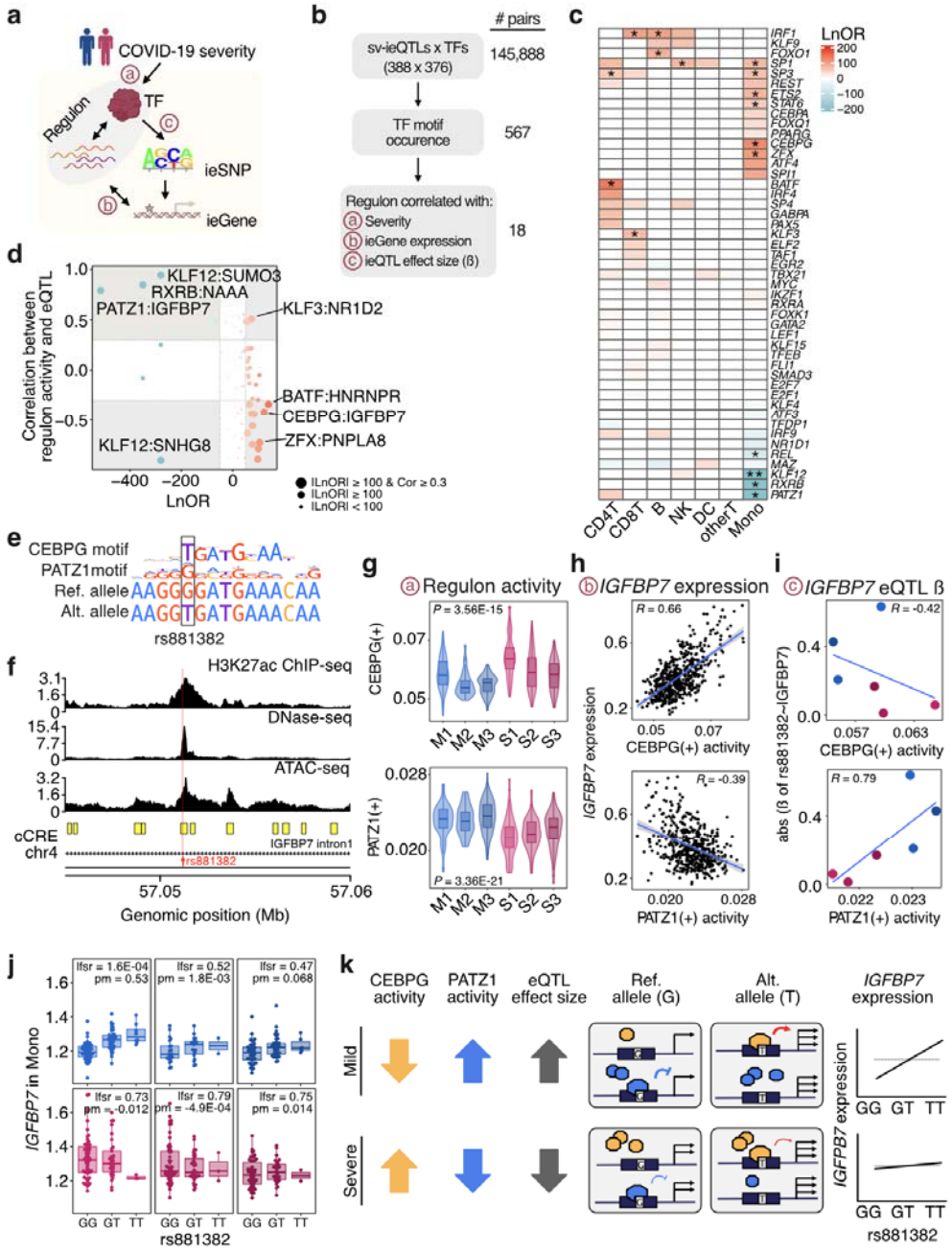

Figure 6. Pathways leading to the regulation of genes that differentiate severe and mild symptoms. a, Schematic representation of the expected upstream mechanism of ieQTLs. **b,** Flow chart displaying a strategy to select strong regulon-ieQTL pairs. **c,** Heatmap showing log odds ratio for the candidate regulon-ieQTL pairs. A positive log odds ratio indicates regulons with greater activity in the Severe group,

while a negative ratio reflects higher activity in the Mild group. **d,** Association of correlation coefficients of regulon activity with ieQTL effect size with log odds ratios (LnOR) for condition-specific regulon activity. Eighteen pairs were highly skewed toward a specific severity group and showed high correlation of regulon activity and eQTL effect size, indicated by grey areas. **e,** Binding motifs of CEBPG and PATZ1 along with the observed sequences flanking rs881382. **f,** Epigenetic tracks of the region flanking rs881382. **g,** CEBPG and PATZ1 regulon activity in monocytes across conditions. **h,** Correlation of regulon activity and *IGFBP7* expression in monocytes. **i,** Correlation of regulon activity and the effect size of the *IGFBP7* eQTL. **j,** Severity-specific regulation of *IGFBP7* expression by the ieQTL under mild conditions. **k,** Proposed model for the mild-diease-specific *IGFBP7* ieQTL.