nature genetics

Article

https://doi.org/10.1038/s41588-024-01896-3

Statistically and functionally fine-mapped blood eQTLs and pQTLs from 1,405 humans reveal distinct regulation patterns and disease relevance

Received: 21 July 2023

Accepted: 6 August 2024

Published online: 24 September 2024

Check for updates

Qingbo S. Wang[1,2] ✉, Takanori Hasegawa[3], Ho Namkoong[4] ✉, Ryunosuke Saiki[5], Ryuya Edahiro[2,6], Kyuto Sonehara[1,2], Hiromu Tanaka[7], Shuhei Azekawa[7], Shotaro Chubachi[7], Yugo Takahashi[8], Saori Sakaue[2,9,10,11], Shinichi Namba[2], Kenichi Yamamoto[2,12,13], Yuichi Shiraishi[14], Kenichi Chiba[14], Hiroko Tanaka[3], Hideki Makishima[5], Yasuhito Nannya[5], Zicong Zhang[15], Rika Tsujikawa[15], Ryuji Koike[16], Tomomi Takano[17], Makoto Ishii[18], Akinori Kimura[19], Fumitaka Inoue[15], Takanori Kanai[20], Koichi Fukunaga[7], Seishi Ogawa[5,15], Seiya Imoto[21], Satoru Miyano[3], Yukinori Okada[1,2,22,23,24] ✉ & Japan COVID-19 Task Force*

Studying the genetic regulation of protein expression (through protein quantitative trait loci (pQTLs)) offers a deeper understanding of regulatory variants uncharacterized by mRNA expression regulation (expression QTLs (eQTLs)) studies. Here we report *cis*-eQTL and *cis*-pQTL statistical fine-mapping from 1,405 genotyped samples with blood mRNA and 2,932 plasma samples of protein expression, as part of the Japan COVID-19 Task Force (JCTF). Fine-mapped eQTLs ($n = 3,464$) were enriched for 932 variants validated with a massively parallel reporter assay. Fine-mapped pQTLs ($n = 582$) were enriched for missense variations on structured and extracellular domains, although the possibility of epitope-binding artifacts remains. *Trans*-eQTL and *trans*-pQTL analysis highlighted associations of class I HLA allele variation with *KIR* genes. We contrast the multi-tissue origin of plasma protein with blood mRNA, contributing to the limited colocalization level, distinct regulatory mechanisms and trait relevance of eQTLs and pQTLs. We report a negative correlation between *ABO* mRNA and protein expression because of linkage disequilibrium between distinct nearby eQTLs and pQTLs.

Studies of genetic regulation of mRNA expression (expression quantitative trait locus (eQTL) studies) is highly informative in interpreting associations between genetic variation and human diseases[1,2]. However, mRNA expression is a limited proxy of protein expression, which affects human phenotypes in a more direct manner[3–5].

Analysis of genetic regulation of protein expressions (protein QTL (pQTL) studies) is gaining popularity[6–11], owing to the development of high-throughput affinity-based assays that enable one-shot measurements of thousands of proteins in large-scale biobank cohorts. Sun et al.[7] performed a pQTL study in 3,301 samples on 3,622 plasma proteins measured using SOMAscan and identified nearly 2,000 pQTLs. Zhang et al.[8] performed large-scale pQTL fine-mapping in European and African populations and discussed druggability. The UK Biobank (UKB) Pharma Proteomics Project (PPP) has performed pQTL mapping in more than 50,000 samples[12–14], uncovering common and rare variant contributions to

A full list of affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper.
✉e-mail: qingbow@m.u-tokyo.ac.jp; hounamugun@keio.jp; yuki-okada@m.u-tokyo.ac.jp

variation in protein expression that do not necessarily involve the effect of eQTLs in major tissues.

Complementing such studies, having another layer of diversity by studying East Asian (EAS) populations would be promising because multi-cohort genetic and proteomic studies have been effective in identifying drug targets[15–18]. In addition, instead of using an external eQTL dataset to examine colocalization between eQTLs and pQTLs, building on earlier studies with multimodal measurements of mRNA and protein expression in the same sample set[19] in this era of large-scale proteomics would be effective in identifying disease-causing variants in multiple layers as it minimizes the potential loss of discovery power due to differences in technical and clinical backgrounds.

In this study, we systematically characterized shared versus mRNA expression or protein-specific QTLs using a collection of 1,405 genotyped samples, whole-blood mRNA and 2,932 plasma protein expression data (measured using Olink Explore 3072) in the Japan COVID-19 Task Force (JCTF[20,21]), including 998 samples and 2,211 genes with both mRNA and protein measurements (Extended Data Fig. 1). Using statistical fine-mapping, we describe the distinctive features for each class of QTLs, show the relevance of protein-specific QTLs to complex traits and provide examples where protein-specific QTLs explain variant–disease associations. We then discuss the dynamics of mRNA and protein expression as a function of disease (COVID-19) severity. Finally, we characterize that eQTLs and pQTLs in complex linkage disequilibrium (LD) in the *ABO* locus might possibly create a negative correlation between mRNA and protein expression.

Results

An expanded catalog of fine-mapped whole-blood eQTLs

We identified 3,464 putative causal (posterior inclusion probability (PIP) > 0.9) *cis*-eQTLs from statistical fine-mapping of 165,410,953 variant–gene pairs in 1,019 samples. The power gain compared to the previous version of our fine-mapping call from 465 samples[20] is represented by a 2.96-fold increase in the number of putative causal eQTLs (from 1,169 to 3,464; Extended Data Fig. 2a); the largely maintained quality of fine-mapping was validated by functional enrichment and consistency with external datasets (Extended Data Fig. 2b, and Supplementary Figs. 1 and 2). The list of splice QTLs (sQTLs) was also updated (Supplementary Fig. 3).

For further validation of statistically fine-mapped eQTLs, we performed a massively parallel reporter assay (MPRA[22,23]) targeting over 10,000 variants with nontrivial causal evidence in the previous call (PIP ≥ 0.1). The fraction of MPRA hits (Supplementary Data 1 and Methods) increased along with the PIPs. For example, 13.9% of variants with a PIP of 1 presented expression modifier effects in either or both K562 and HepG2 at a false discovery rate (FDR) of less than 0.1 (2.2-fold enrichment compared to 6.3% at a PIP < 0.1; Fisher's exact test $P = 2.5 \times 10^{-6}$; Extended Data Fig. 2c). In addition, we observed increased concordance between the direction of the eQTL effects and the effects in the MPRA, along with the confidence in each metric. For example, when focusing on 60 variants with a PIP greater than 0.99 and an MPRA hit FDR lower than 0.01 (that is, tier 1) in either tissue, an effect direction concordance of around 80% was observed in both cell types (82.5% in K562 and 79.4% in HepG2 cells; Extended Data Fig. 2d). Taken together, these results provide orthogonal support for our fine-mapping results.

A fine-mapped pQTL resource from 1,384 EAS samples

Using the expression of 2,932 plasma proteins measured with the Olink Explore 3072 assay for 1,384 samples, we performed genome-wide *cis*-pQTL calling and fine-mapping (Supplementary Fig. 4 and Methods). More than 40% of the measured proteins ($n = 1{,}191$ proteins, 40.6%) had at least one variant with $P < 5 \times 10^{-8}$ (Fig. 1a). Statistical fine-mapping identified $n = 582$ putative causal (PIP ≥ 0.9) pQTLs.

To validate our pQTL calling, we compared our result with multiple external studies. First, we compared our results with the recent large-scale pQTL mapping results from the UKB PPP study[12] and observed a high correlation in effect size estimates (Fig. 1b). The correlation increased along with the causal evidence (PIP) in our dataset; it was the highest when the population background matched (that is, when compared to $n = 262$ EAS samples in the UKB; Pearson $r = 0.97$ when PIP = 1; first row in Fig. 1b). The second highest correlation was observed in European (EUR) samples in the UKB ($n > 40{,}000$ samples; Pearson $r = 0.93$ when PIP = 1; second row in Fig. 1b), followed by African (AFR) samples in the UKB ($n = 931$; Pearson $r = 0.85$ when PIP = 1). The fraction of the lead variants in the UKB PPP dataset also increased along with the PIP (as an approximation of causality in the absence of fine-mapping data in the UKB PPP; Supplementary Fig. 5). Such concordance between two datasets, both measured using the Olink Explore 3072 assay, validate our pQTL fine-mapping. Lower consistency was observed when comparing with other datasets based on different technological platforms[8,24–26] (Fig. 1c), as discussed in detail in Supplementary Figs. 6 and 7 and in Supplementary Note.

Functional characterization of fine-mapped pQTLs

To characterize fine-mapped pQTLs, we tested the enrichment of major functional annotations. For noncoding annotations, the 5′ and 3′ untranslated regions (UTRs) were significantly enriched for putative causal pQTLs (521.5-fold and 167.6-fold; Fisher's exact test $P = 8.8 \times 10^{-77}$ and 5.3×10^{-86}, respectively); for coding annotations, missense and predicted loss of function (pLoF) showed enrichments an order of magnitude higher (2109.2-fold and 8046.9-fold; Fisher's exact test $P < 10^{-100}$ for both), which is consistent with our understanding that pQTL effects could be driven by both transcriptional and posttranscriptional regulations (Fig. 2a).

Focusing on missense pQTLs, we tested the enrichment of protein structure and localization-related annotations (Fig. 2b). Our analysis highlighted two factors: (1) putative causal pQTLs were enriched for protein domains that form an alpha-helix or beta-sheet structure (2.48-fold and 1.96-fold; Fisher's exact test $P = 7.6 \times 10^{-6}$ and 1.3×10^{-3}, respectively) and (2) they were enriched for extracellular domains (or depleted for cytoplasmic domains; 1.43-fold and 0.23-fold; Fisher's exact test $P = 4.5 \times 10^{-3}$ and 1.8×10^{-3}, respectively). These observations suggest that missense mutations could show pQTL effects by affecting protein structure and stability, especially in an extracellular environment, as discussed in ref. 9, who reported an enrichment of secreted protein-related features in genes harboring pQTLs (Supplementary Fig. 7).

On the other hand, as protein expression in our study was measured with an affinity-based assay using antibodies specific to each protein, the observed pQTL effects might reflect differences in binding affinity introduced by genomic variations, rather than differences in physiological protein abundance. To evaluate such potential 'epitope effects', we investigated the replication rate of fine-mapped pQTLs stratified according to missense and other functional annotations (ARIC[8] instead of the UKB PPP was used because fine-mapping data were only available for ARIC). Missense variations were the most enriched functional annotation, even when focusing on possibly causal pQTLs replicated in the ARIC study (PIP > 0.1 in both studies; Fig. 2c and Supplementary Fig. 8). On the other hand, replicated possibly causal missense pQTLs often presented opposite effect directions (Fig. 2d; 13 of 57 variants). A possible model derived from these observations is that a missense variant increases affinity to an antibody in one platform while decreasing affinity to the probe in another platform; thus, the 'true' effect size and direction in the physiological context is uncertain, warranting future large-scale experimental validation of pQTLs as done in the MPRA for eQTLs.

A subset of causal pQTLs in noncoding regions possibly presents pQTL effects caused by transcriptional regulation in related tissues. To characterize such pQTL variants as colocalizing[27] with eQTL variants, we compared our fine-mapping results with the eQTL fine-mapping results from the Genotype-Tissue Expression (GTEx) project v.8 (ref. 2) in 49 major tissues (Supplementary Fig. 9a,b and Methods). The liver presented the highest enrichment, followed by whole blood and

https://doi.org/10.1038/s41588-024-01896-3

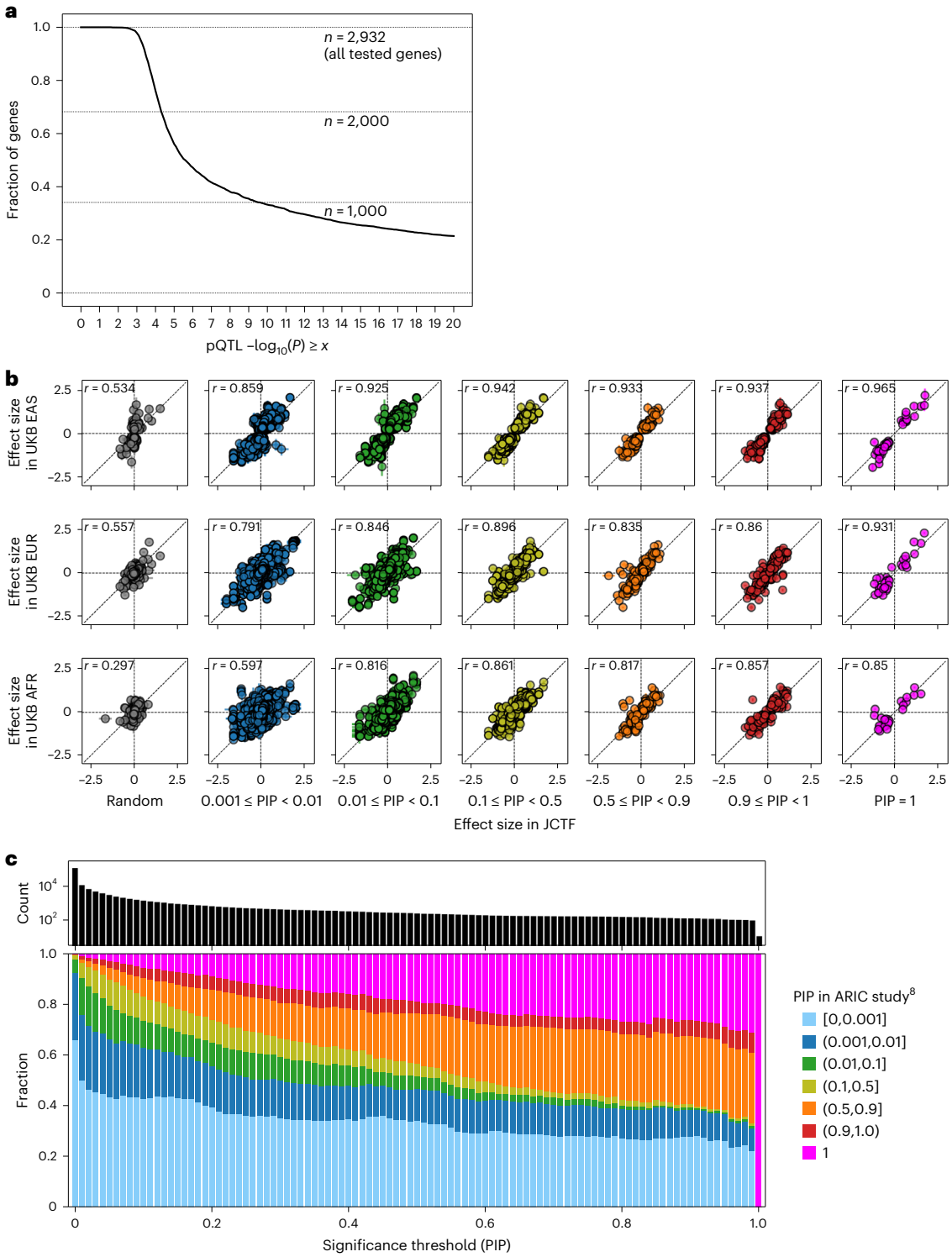


Fig. 1 | A fine-mapped pQTL resource from 1,384 EAS samples. a, Fraction of genes passing the minimum *P* threshold. **b**, Concordance of variant effect size in an external large pQTL dataset (the UKB PPP), stratified according to the population in the UKB PPP (row) and the PIP of our pQTL fine-mapping (column). For simplicity, only variants with $P < 0.05$ in the UKB PPP are included. **c**, Concordance of PIP with an external pQTL fine-mapping dataset that used the SOMAscan platform (ARIC)[8].

spleen, which is consistent with our understanding that a large proportion of plasma proteins are secreted from the liver[28], as well as hematopoietic organs (2.44-fold, 2.15-fold and 1.92-fold, respectively; $P = 2 \times 10^{-3}$ for the spleen; Fig. 2e). The testis and most brain-related tissues were significantly depleted for colocalization (for example, 0.33-fold, $P = 1 \times 10^{-3}$ for the testis), which is consistent with the previous literature[29].

Contrasting eQTLs and pQTLs highlights distinct functional features

Although colocalization between eQTLs and pQTLs has been investigated previously[6,7,9,12,19,30,31], when eQTL and pQTL data are from different cohorts, the correlation between two association statistics can be low because of LD structure differences even when the underlying causal

https://doi.org/10.1038/s41588-024-01896-3

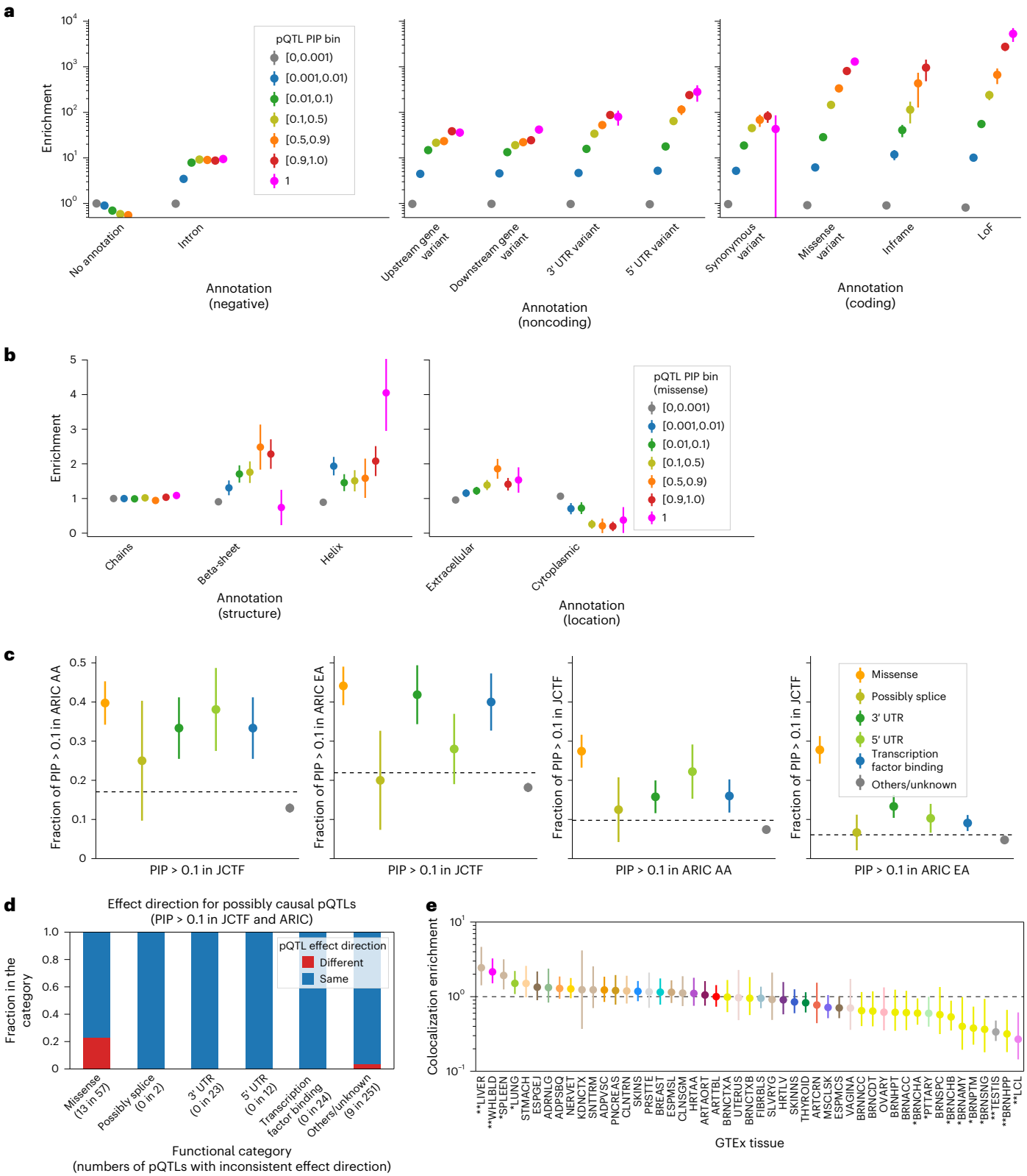


Fig. 2 | Functional characterization of fine-mapped pQTLs. a, Enrichment of major functional annotations along with pQTL PIPs. The missing dot in the category corresponds to $n = 0$. **b**, Enrichment of protein-level structure and localization annotations along with pQTL PIPs for missense variants. **c**, Concordance between the fine-mapped pQTL PIPs in our dataset and the ARIC dataset, stratified according to major functional annotations. **d**, Enrichment of missense variation in possibly causal (PIP > 0.1) pQTLs, with inconsistent effect direction between datasets. **e**, Enrichment of colocalization score between pQTLs and 49 fine-mapped eQTLs in the GTEx. $^{*}P < 0.05$, $^{**}P < 0.05/49$. **a**,**b**, $n = 27{,}354{,}135, 164{,}350, 32{,}868, 4{,}257, 562, 508$ and 81 variant–gene pairs are included in each category. **c**, $n = 1{,}008, 1{,}304, 1{,}759$ and 4,705 variant–gene pairs are included in each panel. **e**, All 27,556,761 variant–gene pairs are included. Those missing in the GTEx were omitted from the downsampling process (Methods).

variants are the same. Calling eQTLs and pQTLs from the same samples is better suited to colocalization analysis as LD contributes equally to both QTL calls. Thus, we used $n = 998$ samples and 2,211 genes where both RNA and protein expression profiles existed, fine-mapped the eQTLs and pQTLs within the 998 samples once more and compared their effect sizes (Fig. 3a). While the correlation was relatively high for putative causal eQTLs and pQTLs (Fig. 3a, bottom right; $r = 0.80$ when both eQTL and pQTL PIPs are greater than 0.9), there were many fine-mapped eQTLs and pQTLs that are probably causal in either mRNA or protein expression in the blood but not both (for example, top-right and bottom-left in Fig. 3a).

Further, we created mRNA-adjusted protein expression and protein-adjusted mRNA expression profiles, and performed adjusted mRNA or protein expression QTL[19] calling followed by statistical fine-mapping (Supplementary Fig. 10). Using a catalog of fine-mapped eQTLs, pQTLs, protein-adjusted eQTLs and mRNA-adjusted pQTLs, we classified putative causal variant–gene pairs into three classes: (1) 'colocalizing QTLs', confidently acting on both mRNA and protein expression; (2) 'mRNA-specific QTLs'; and (3) 'protein-specific QTLs' (Methods). Nearly half of colocalizing QTLs are probably explained by splice variation (17 of 42 (40.5%) had an sQTL PIP > 0.9, in contrast to 20.2% for mRNA-specific QTLs and 6.4% for protein-specific QTLs; Fisher's exact test $P = 5.3 \times 10^{-3}$ and 5.3×10^{-8}; Fig. 3b). When focusing on noncoding annotations, mRNA-specific QTLs were enriched for transcription factor binding sites and marginally for 5′ UTR variants (1.96-fold and 1.91-fold compared to protein-specific QTLs; $P = 1.2 \times 10^{-2}$ and 5.3×10^{-2} testing the difference of the odds ratio). The enrichment level of 3′ UTR variants for mRNA and protein-specific QTLs was comparable (67.8-fold and 86.0-fold, $P > 0.05$), supporting the notion that RNA sequencing (RNA-seq)-based methods for testing the effects of 3′ UTR variants[32,33], which often involve post-transcriptional modifications, are harder than those of 5′ UTR variants. For coding annotations, missense variants were strongly enriched for protein-specific QTLs as expected (8.95-fold compared to mRNA-specific QTLs, $P = 1.1 \times 10^{-15}$; Fig. 3c and Supplementary Fig. 11a).

Buffering of blood mRNA expression variation in constrained genes

Turning to a gene-centric view, we classified 557 genes carrying variants belonging to one (or, rarely, more than one) of the three QTL classes defined above (Supplementary Data 2 and Fig. 3d). We then examined their mRNA expression level across tissues in the GTEx, as well as their constraint from the Genome Aggregation Database (gnomAD) (LoF observed/expected upper bound fraction (LOEUF)[34]; Fig. 3e and Supplementary Fig. 11b). Genes carrying mRNA-specific QTLs had a higher number of expressed tissues in the GTEx (expressed in 46 versus 41 or 34 tissues on average, t-test $P = 6.3 \times 10^{-4}$ and 1.4×10^{-3}), and were more highly constrained compared to those carrying colocalization QTLs or protein-specific QTLs (mean LOEUF = 0.90 versus 1.29 or 1.02, t-test $P = 1.7 \times 10^{-5}$ and 4.0×10^{-3}). Our observation suggests that, in highly constrained and ubiquitously expressed genes, transcriptional regulation in the blood is more likely to be buffered in plasma, presumably resulting in maintained functionality.

Limited colocalization between eQTLs and pQTLs

We next shifted our focus to the intersection of eQTL and pQTL effects (that is, colocalization). In our QTL classification, even with a colocalization posterior probability (CLPP) threshold of 0.1 (Methods), the number of genes carrying possibly colocalizing QTLs were far fewer than those carrying mRNA or protein-specific QTLs (68 versus 260 or 158). Using an alternative method (Mendelian randomization; Supplementary Fig. 12 and Methods) also suggested limited overlap; we explored possible reasons for such a low colocalization level.

First, we tested colocalization with a larger blood eQTL dataset (eQTLgen) to evaluate the potential power loss attributed to a low sample size. Even with a lenient assumption that a significant P value in eQTLgen is a sign of colocalization (in reality, many are noncausal, tagged variants), the colocalization level remained low (only less than 25% were resolved; Supplementary Fig. 13a–d). Second, we compared our colocalization detection method with another state-of-the-art method (SuSiE-coloc). While more colocalization could be declared depending on the threshold setting, the overall fraction of genes with colocalizing evidence remained low (<15% at a PP.H4 = 0.1 threshold; Supplementary Fig. 13e,f).

Last is the contribution of tissues other than blood. We stratified the gene sets into distinct classes based on the mRNA expression level in whole blood relative to other tissues in the GTEx and observed that genes with a higher percentage of mRNA expression had a higher correlation of blood mRNA and plasma protein expression (average Pearson $r = 0.26$ for genes with more than 50% expression in the blood versus 0.021 for genes with less than 1% expression, t-test $P = 1.2 \times 10^{-8}$, where percentage expression was defined as the transcripts per million (TPM) in whole blood divided by the sum of TPM across 54 tissues in the GTEx; Fig. 3f). Such differences in the mRNA expression level in whole blood presented major differences in colocalization level (Fig. 3g). Only 5% of lowly (<1%) mRNA-expressed genes had colocalization evidence at a 0.1 CLPP threshold, while over 20% did for highly (>5%) mRNA-expressed genes. Considering additional biological properties, such as active secretion of proteins into the bloodstream (Supplementary Fig. 13g,h) or testing colocalization with fine-mapped variants in the GTEx across 49 tissues also increased the colocalization signal (from approximately 5% to over 10%; Supplementary Figs. 9c and 13i). Nevertheless, the fraction of genes with colocalization evidence remained low, even with the most lenient threshold (<33% at an ultra-lenient 0.001 CLPP threshold across GTEx tissues).

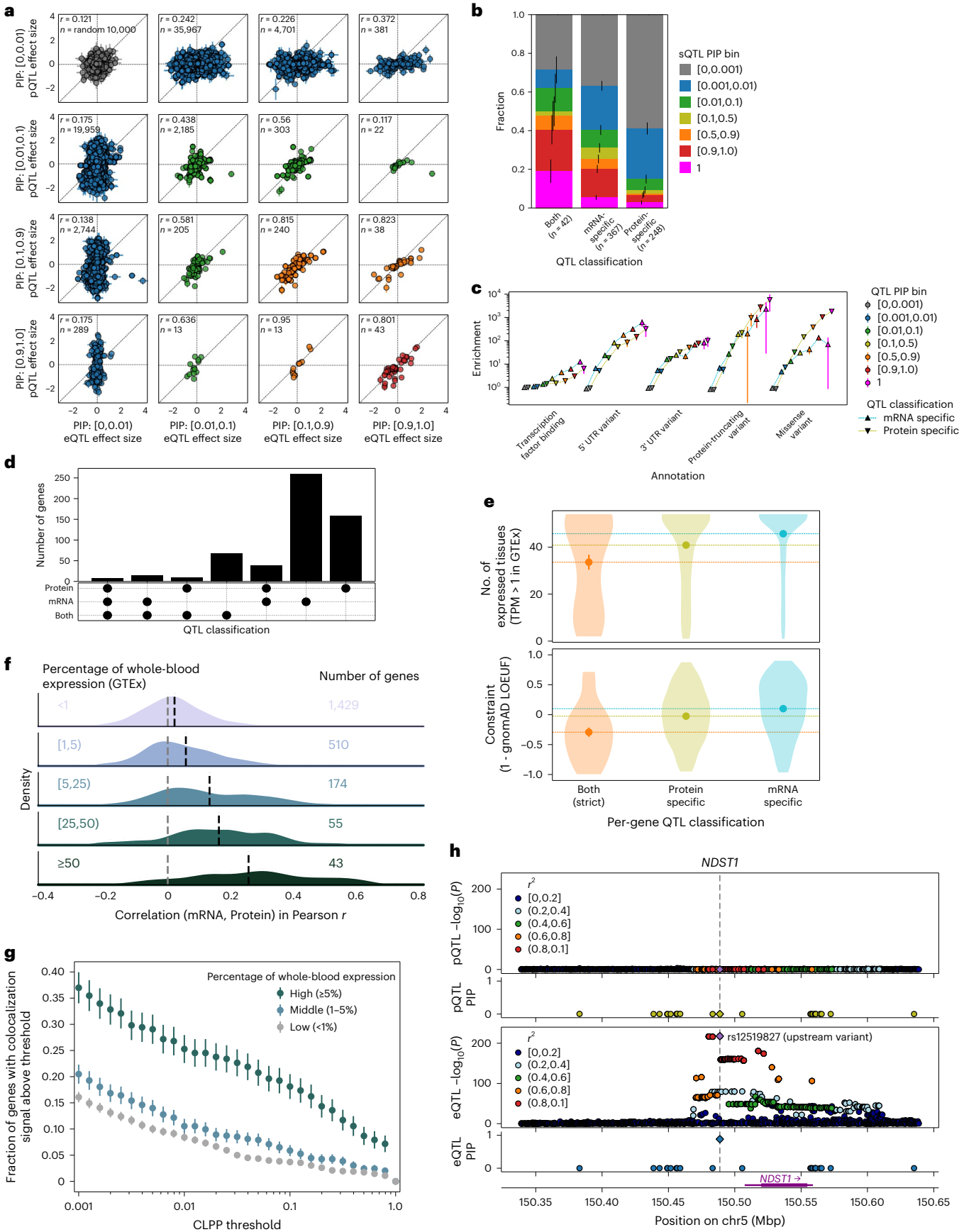
Thus, our observations are in line with previous reports[19] that only a sizable fraction of plasma pQTLs are mediated through eQTL effects in major tissues.

Examples of limited colocalization between eQTLs and pQTLs

As an example of blood mRNA-specific regulation on a constrained gene, we highlight rs12519827 (chr5:150488763:G:A), an intronic variant on the *NDST1* gene (Fig. 3h). The variant has a strong negative effect on blood mRNA expression ($P < 1 \times 10^{-200}$, PIP = 1, $\beta = -0.68$) but little or no effect on the protein expression level in plasma ($P = 0.41$, PIP = 0). *NDST1* is a nonsecreted protein, highly constrained (LoF observed/expected = 3/41.1 = 0.07 in gnomAD) and mouse homozygous knockout lethal[35] with known association with severe intellectual developmental disorder[36]. Associations with complex traits did not reach genome-wide significance in the Biobank Japan (BBJ)[37,38] ($P = 2.6 \times 10^{-7}$ for basophil count and $P > 1 \times 10^{-4}$ for others). Notably, the mRNA expression of *NDST1* in whole blood was the lowest among all tissues in the GTEx (median TPM = 8.9); the eQTL effect of rs12519827 is highly specific to

Fig. 3 | Characterization of mRNA-specific and protein-specific or colocalizing QTLs and genes. a, Correlation between eQTL and pQTL effect sizes in our dataset, stratified according to the PIPs from the fine-mapping of each QTL. **b**, Fraction of fine-mapped sQTLs for different QTL categories. **c**, Enrichment of major functional annotations for mRNA-specific or protein-specific QTL PIP bins. **d**, Number of genes harboring mRNA-specific or protein-specific or colocalizing QTLs. **e**, Distribution of constraint score (LOEUF) and number of mRNA-expressing (TPM > 1) tissues in the GTEx for different QTL categories. **f**, Ridge plot showing the distribution of correlation between mRNA and protein expression in 998 samples, stratified according to the percentage of whole-blood expression in the GTEx. **g**, Fraction with colocalization evidence according to varying thresholds, stratified according to the percentage of whole-blood expression in the GTEx. Each stratum contains 1,429, 510 and 272 (174 + 55 + 43) genes, as visible in **f**. **h**, Locus zoom around the *NDST1* gene, as an example of blood mRNA-specific regulation on a constrained gene.

https://doi.org/10.1038/s41588-024-01896-3



whole blood ($P = 1.3 \times 10^{-97}$ in whole blood but more than 5×10^{-8} in all other tissues), suggesting that the whole-blood-specific eQTL effect has only a negligible effect and is buffered by relatively high and stable mRNA expression in other tissues relevant to plasma protein expression. As another example of a gene with limited colocalization evidence, we highlight the unresolved *ALDH2* locus[39–42] in Supplementary Note, Supplementary Fig. 14 and Supplementary Table 1.

pQTLs may be more relevant to complex traits than eQTLs

The observation that variation in mRNA expression in constrained genes is often buffered suggests possible differences in contribution to complex traits. We used large-scale fine-mapping results for 79 traits from the BBJ to quantify the level of trait-causal variant colocalization (Fig. 4a, Supplementary Fig. 15 and Supplementary Data 3). Protein-specific QTLs had significantly higher enrichment compared to mRNA-specific QTLs (7.6-fold versus 5.6-fold, $P = 2.7 \times 10^{-3}$, ratio of means test for PIP > 0.01 combined), suggesting that alteration in protein expression is more relevant to complex traits. We replicated these observations using the fine-mapping results for 96 traits in the UKB[38,43], stratified according to trait categories, validating that protein-specific QTL enrichment is not simply driven by a specific biological trait (protein-specific QTL enrichment was greater; $P < 0.05$ except for psychological traits; Fig. 4b). Enrichment is not likely to be driven by simple measurement traits such as blood cell count or lipid levels as enrichment persisted in several major disease and complex traits, although not universal to all traits as the opposite enrichment pattern was observed for body mass index (Supplementary Fig. 15a). Stratification according to the percentage of mRNA expression of the genes suggested that the power gain of protein-specific QTLs is mainly from genes that are relatively lowly mRNA-expressed in the blood (Fig. 4c and Supplementary Fig. 15d), highlighting the value of the multi-tissue origin of plasma proteins. Comparison with nonspecific QTLs is more nuanced, as described in detail in the Supplementary Note.

As specific examples, we highlighted three scenarios where plasma pQTL fine-mapping is beneficial in addition to eQTL fine-mapping in whole blood for complex trait-causal variant interpretation: (1) a 3′ UTR variant possibly involved in posttranscriptional regulation (rs884205 on the *TNFRSF11A* gene; Fig. 4d); (2) a missense variant probably affecting protein stability (rs429358, the well-known constituent of the APOE-ε4 allele[44]; Fig. 4e and Supplementary Fig. 16); and (3) mRNA expression regulation taking place in non-blood tissue (rs13395911 on *EFHD1* and rs28372783 on *CDH1*; Supplementary Fig. 15h,i), with detailed interpretation in Supplementary Note.

Characterizing the *trans*-pQTL effects

The regulatory effects of genetic variants on distal genes (that is, *trans*-regulation) are much lower than those on a nearby gene (that is, *cis*-regulation). Thus, identifying genome-wide *trans*-regulatory variants requires a sample size much larger than that of a canonical *cis*-QTL analysis[45–48]. In this study, we used two different approaches to maximize the prior and to control for the multiple test burden (Supplementary Note).

First, we focused on the lead *cis*-pQTL (that is, the variant with the lowest *cis*-pQTL *P* value) for each gene and tested their genome-wide *trans*-effects. We observed a distribution shift in the test statistics for the lead *cis*-variant in *trans*-pQTL effects compared to random variants (Fig. 5a), validating the strategy. Notably, we replicated the observation of a *trans*-pQTL 'hotspot' in the *ABO* gene, regulating the expressions of 26 genes genome-wide (at a Bonferroni significance; $P < 1.1 \times 10^{-8}$; Fig. 5b).

Second, we focused on variation in the classical alleles of the HLA[49] genes (and *MICA*, an HLA-like gene located within the major histocompatibility complex (MHC); Fig. 5c) and examined the genome-wide *trans*-eQTL and *trans*-pQTL effect. This identified significant associations ($P < 1 \times 10^{-5}$) for 42 genes (Supplementary Data 4), including those between class 1 HLA alleles (especially, *HLA-C*) and killer immunoglobulin-like receptor (*KIR*) genes (*KIRDL2* at the mRNA level and KIRDL3 at the protein level; $P = 7.2 \times 10^{-9}$ and $P = 1.2 \times 10^{-25}$), providing a genetic basis for the well-known molecular interaction between KIR and class 1 HLA at the protein–protein level[50–52].

For HLA, comparing *trans*-mRNA-regulated and protein-regulated genes, there was zero overlap at the threshold of $P < 1 \times 10^{-5}$, potentially suggesting that plasma pQTL analysis better captures the dynamics of immunological responses across cells and organs led by the variation in HLA alleles, whereas the blood transcriptome is more narrowly focused on peripheral immune cell responses.

Buffering of context-specific (severe COVID-19) eQTL effects

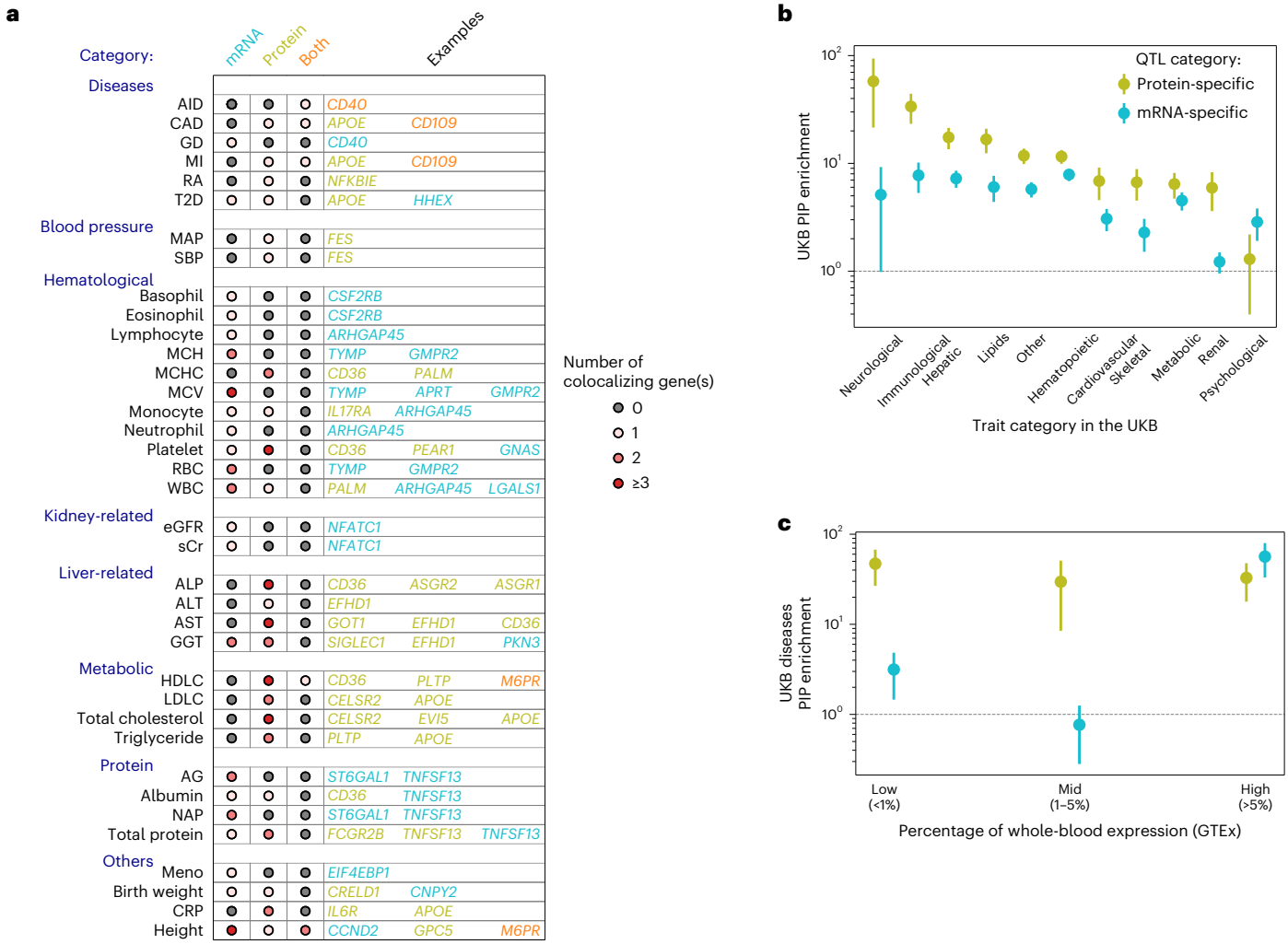
Our study cohort was ascertained for COVID-19-positive cases, classified into four levels of severity. We used this phenotype to understand the dynamics of mRNA and protein expression in response to infectious diseases.

First, we observed an increase in the correlation between mRNA and protein expression in patients with severe COVID-19 (Fig. 6a and Supplementary Fig. 17), presumably because of the natural immune response triggering an increase in relevant genes at both the blood mRNA and plasma protein level. As an example, interleukin-1 receptor-like 1 (IL1RL1), a well-known chemokine, had high correlation specifically in the severe and most severe states (Fig. 6b; $r = 0.50$ in the most severe disease and 0.08 in the asymptomatic state). In addition, the shared biological entities of blood mRNA and plasma protein dynamics were visible at the module level (Supplementary Fig. 18).

Then, we investigated the genetic regulation landscape in several different infectious disease statuses[53]. Specifically, we performed COVID-19 severity interaction QTL (iQTL) analysis as described previously in[20] and compared the interaction landscape at the mRNA and protein levels (Supplementary Figs. 19 and 20 and Supplementary Data 5). The interaction between disease status and pQTL effects was less significant compared to eQTL effects (Fig. 6c–f). We attribute this to buffering from high and stable expression of mRNAs in other tissues, such as lymphocytes (Fig. 6c), the dynamics of solid components missing in plasma and the limitation of detection (LOD) specific to the quantification of affinity-based protein expression (Supplementary Fig. 21), with further details described in Supplementary Note.

Fig. 4 | Complex trait colocalization. a, Number of colocalizing genes for each of the major BBJ trait–QTL pairs, along with gene examples. The color coding of the genes corresponds to the QTL classifications. **b**, Enrichment of complex trait PIPs from the UKB according to trait category, for variants with putative mRNA-specific or protein-specific causal QTL effects. **c**, Enrichment of complex trait PIPs from the UKB, stratified according to the percentage of whole-blood expression in the GTEx. **d**,**e**, Specific examples where pQTLs colocalize with complex trait-causal variants (*TNFRSF11A* (**d**) and *APOE* (**e**)). **b**, $n = 3{,}206$ protein-specific and $n = 5{,}308$ mRNA expression-specific QTLs passing the PIP > 0.1 threshold were included. **c**, Those QTLs were further divided into $n = 3{,}181$, 1,344 and 799 mRNA-specific QTLs and $n = 1{,}908$, 723 and 585 protein-specific QTLs in each whole-blood mRNA expression level stratum (from low to high), after removing the ones where GTEx data were missing. AID, autoimmune disease; AG, albumin to globulin ratio; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; CAD, coronary artery disease; CRP, C-reactive protein; eGFR, estimated glomerular filtration rate; GD, Graves' disease; GGT, γ-glutamyltransferase; HDLC, high-density lipoprotein cholesterol; LDLC, low-density lipoprotein cholesterol; MAP, mean arterial pressure; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; Meno, age at menopause; MI, myocardial infarction; NAP, nucleosome assembly protein; RA, rheumatoid arthritis; RBC, red blood cell; SBP, systolic blood pressure; sCr, serum creatinine response; T2D, type 2 diabetes; WBC, white blood cell.

https://doi.org/10.1038/s41588-024-01896-3



https://doi.org/10.1038/s41588-024-01896-3

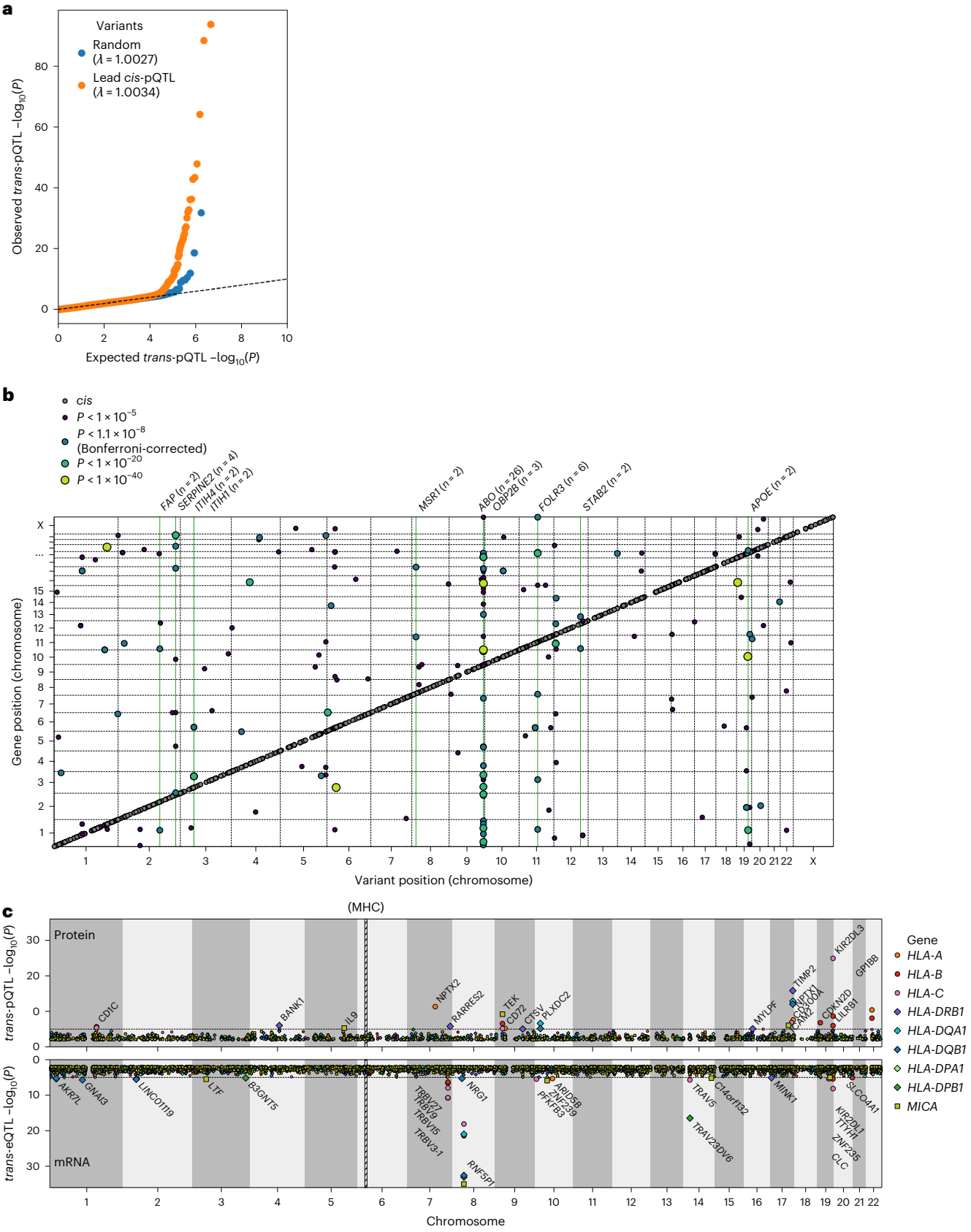


Fig. 5 | The *trans*-QTL analysis. a, Q–Q plot comparing the *P* distribution when testing *trans*-pQTL effects for random variants (blue) or lead *cis*-pQTL variants (orange). **b**, Overview of the *trans*-pQTL effects. **c**, Miami plot for genome-wide *trans*-eQTL and *trans*-pQTL effects of variation in *HLA* genes. Genes passing the suggestive $P < 1 \times 10^{-5}$ threshold have been enlarged and annotated (in bold if the gene had both mRNA and protein measurements). The largest association in eQTL effect for *RNF5P1* was probably driven by a sequencing error[61]. In the plot, we omitted chromosome X, which contained only one significant pQTL association (*CFP*; $P = 9 \times 10^{-5}$), for visual simplicity.

https://doi.org/10.1038/s41588-024-01896-3

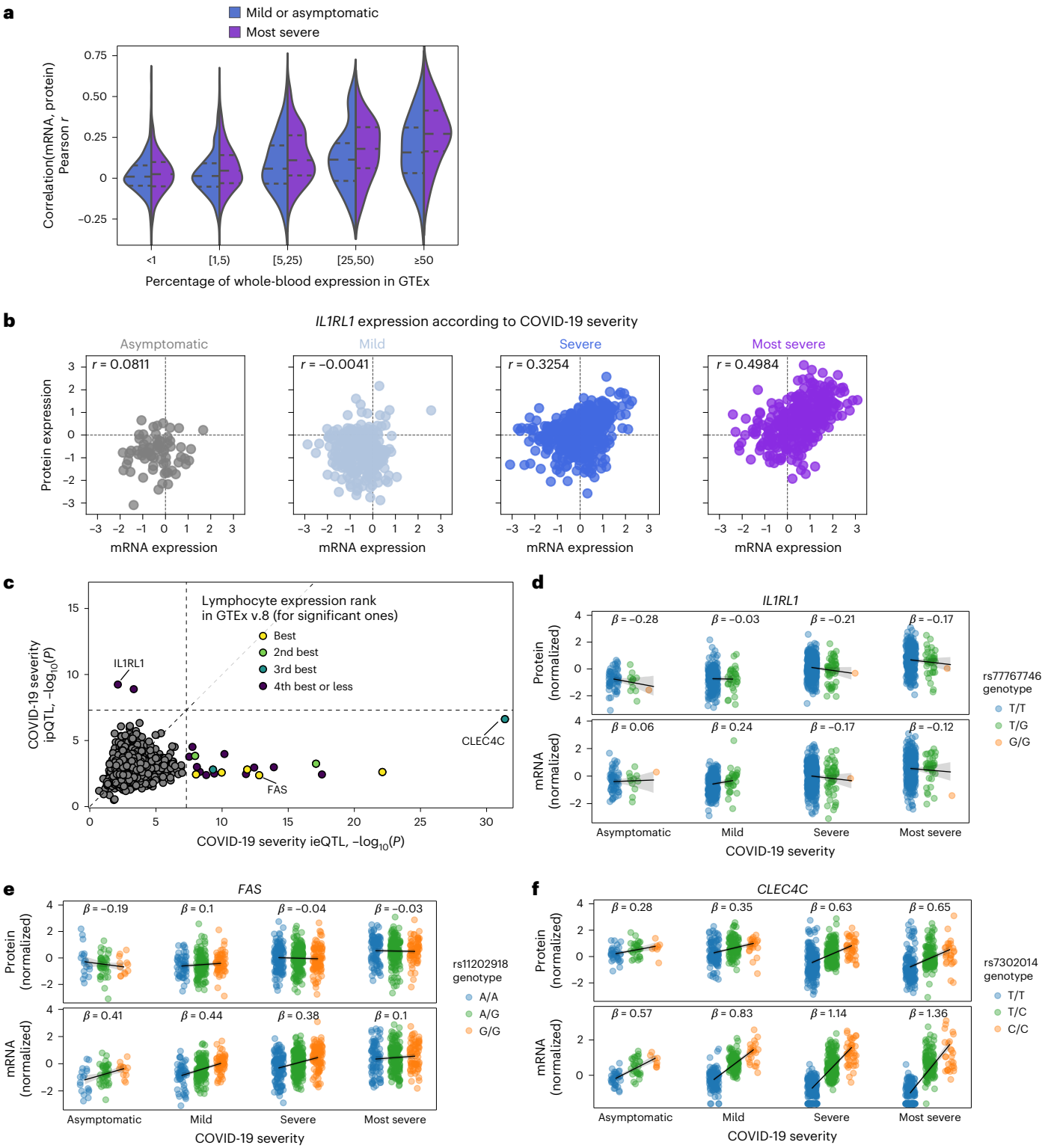


Fig. 6 | COVID-19 severity interaction eQTLs (ieQTLs) and interaction pQTLs (ipQTLs). a, Increase of correlation between mRNA and protein expression in severe COVID-19 across a ranging fraction of whole-blood expression in the GTEx. The boxes inside the violin plots show the 25%, 50% and 75% quantiles. **b**, An example of increased correlation between mRNA and protein expression along with COVID-19 severity (*IL1RL1* gene). **c**, Scatter plot comparing the significance ($-\log_{10}(P)$) of ieQTLs (*x* axis) and ipQTLs (*y* axis), colored according to the rank of expression in lymphocytes in the GTEx, when significant. **d**, Scatter plot showing the effect of rs77767746 on *IL1RL1* mRNA and protein expression. **e**, Scatter plot showing the effect of rs11202918 on *FAS* mRNA and protein expression. **f**, Scatter plot showing the effect of rs11055602 on *CLEC4C* mRNA and protein expression.

https://doi.org/10.1038/s41588-024-01896-3

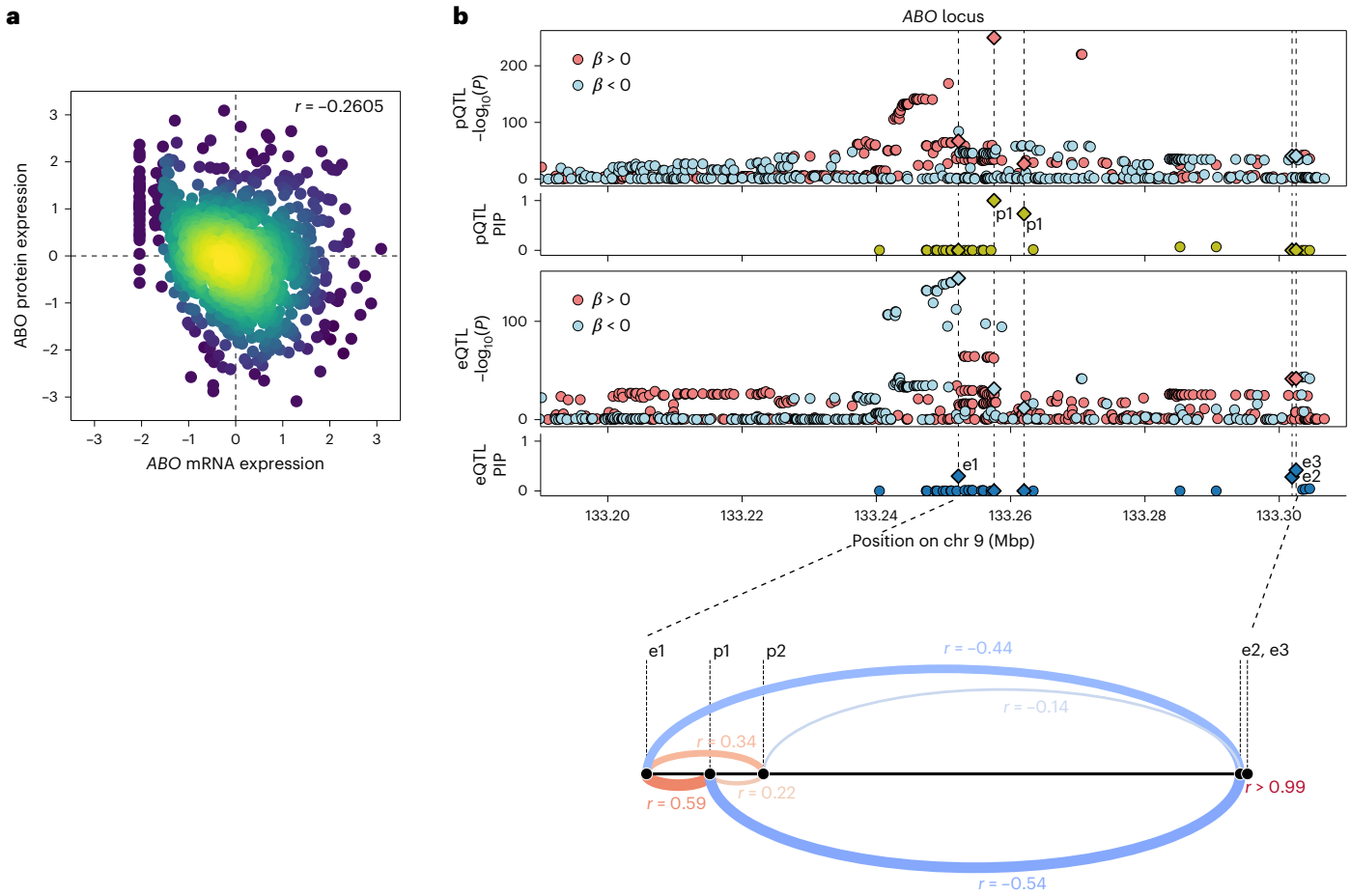


Fig. 7 | Negative correlation between whole-blood mRNA and plasma protein expression in *ABO*. **a**, Density scatter plot showing negatively correlated mRNA and protein expression for the *ABO* gene. **b**, Locus zoom around the *ABO* locus containing two putative causal eQTLs and three putative causal pQTLs, and visualization of LD between those five causal variants. **c**, Normalized protein (*x* axis) and mRNA (*y* axis) expression for the rs8176719 genotype, with and without controlling for LD with two nearby putative causal eQTLs. **d**, Normalized protein (*x* axis) and mRNA (*y* axis) expression for the rs656105 genotype, with and without controlling for LD with another nearby putative causal pQTL.

Interpreting a negative correlation between mRNA and protein expression

Consistent with previous studies[54,55], the correlation between mRNA and protein expression was limited in our dataset (Fig. 3f and Supplementary Fig. 17a,b; mean Pearson $r = 0.047$); a large fraction of genes had a negative correlation between mRNA and protein expression (866 of 2,211 (39.1%)). While multiple mechanisms could contribute to such a negative correlation[56], we highlight *ABO* ($r = -0.26$; Fig. 7a), where multiple distinct causal variants on the same gene acting at different stages of the central dogma in the opposite direction (eQTLs and pQTLs;

Fig. 7b) might be a potential cause of the negative correlation. The *ABO* gene has partially overlapping eQTL and pQTL peaks with a different effect direction; statistical fine-mapping identified five distinct putative *cis*-causal variants (three separate eQTLs plus two pQTLs, hereafter called e1, e2, e3, p1 and p2, with further details shown in Supplementary Table 2). Although p1 had a significant positive effect on protein expression and a negative effect on mRNA expression (Fig. 7c), controlling the LD with e1 and e2 (and e3, which is in near-perfect LD with e2) resolved the negative effect on mRNA expression. Similarly, although e2 had a significant positive effect on protein expression and a negative effect on mRNA expression (Fig. 7d), controlling the LD with p1 nearly resolved the negative effect on protein expression. Functional annotations of these variants further supported the distinct causal mechanisms. p1 (chr9:133257521:T:TC, rs8176719) is a frameshift variant on exon 6, thereby acting during translation, whereas e1 (chr9:133252214:G:A, rs9411476) and e2 (chr9:133301911:T:C, rs656105) had medium-to-high sQTL evidence (sQTL PIP = 0.97 and 0.11, respectively, although not canonical splice sites), that is, in the earlier stage of the central dogma and more subject to buffering at the multi-tissue level. For example, a stable supply of the specific mRNA isoform in tissues other than blood may dominate the observed eQTL and sQTL effect on blood mRNA and possibly mask the effect of the variant at the plasma protein level. These observations highlight a complex regulation landscape of *ABO* mRNA and protein expression by multiple entangled causal variants that form a negative correlation between mRNA and protein expression in the blood.

Discussion

In this study, we presented a collection of whole-blood mRNA and plasma protein expression data from 1,405 genotyped Japanese samples. Our eQTL analysis presented an expanded catalog of 3,464 fine-mapped putative causal eQTLs at single-variant resolution, including 932 validated by an MPRA, thereby allowing functional prioritization of putative causal mRNA regulatory variants even in the case of tight LD. Our pQTL analysis of 2,932 proteins presented a catalog of 582 fine-mapped putative causal pQTLs, allowing us to carry out detailed functional characterization, ranging from mediation by eQTL effects in several tissues (for example, liver or spleen) to protein structure disruption.

Combining mRNA and protein expression for 998 samples, we compared mRNA-specific or protein-specific versus shared regulatory effects to highlight distinct characteristics, such as enrichment of sQTLs for shared regulatory variants and higher constraint for genes specifically regulated by mRNA. We reported a limited level of colocalization between causal eQTLs and pQTLs, attributed to fundamental differences in mRNA versus plasma protein in their origin and biological property. Furthermore, we reported a higher proportion of trait-causal and disease-causal variant colocalization for protein-specific QTLs compared to mRNA-specific QTLs, especially in genes with a low mRNA expression fraction in the blood, as well as clear differences in the *trans*-regulatory landscape of class I HLA variations and their connection to KIR families, all highlighting the value of plasma protein expression studies on top of blood mRNA expression studies.

We also showed that the interaction of QTL effects with COVID-19 severity was milder for protein expression compared to mRNA expression, probably because of active expression of mRNA in lymphocytes regardless of mRNA regulation in the blood, as well as the limited dynamic range of affinity-based protein measurement (LOD). Finally, taking the *ABO* locus as an example, we showed that negative correlation could arise due to the LD of distinct nearby causal eQTLs and pQTLs, warranting the need to consider the combination of variants[57] acting on different layers of regulatory mechanisms.

The limitations of our study include: (1) the existence of measurement-specific effects by altering epitope-binding sites (epitope effects). Although our analysis using multiple cohorts spanning multiple technological platforms suggested a certain level of consistency of pQTL architecture, it also suggested that epitope effects have a role in inconsistent effect size estimations, especially in missense variants; (2) the existence of measurements below the LOD, which probably decreased the power of pQTL calling, especially when the LOD fraction correlated with the biological properties of the sample (for example, COVID-19 severity); and (3) the fact that our pQTL analysis was restricted to approximately 3,000 genes where measurements were available.

With refined understanding of epitope effects, identification of larger number of putative causal pQTLs by increasing the study size and the integration of more detailed machine learning-derived variant features, such as effects on folding[58,59], we envision a future where functional priors specific for pQTLs similar to our previous work on eQTLs[60] could be leveraged to better understand protein regulation. Finally, although we presented an example where nearby eQTLs and pQTLs with an opposite effect direction exist in LD, a systematic, genome-wide evaluation of such QTL 'entanglement' is yet to be performed. Digesting regulatory mechanisms in light of multiple regulatory layers, including splicing, isoform level expression and the dynamics of mRNA and protein turnover, would be important for several aspects, such as estimation of disease heritability mediated at different layers of the central dogma or developments of drugs targeting mRNA or protein expression specifically.

Our study, leveraging a rich dataset of 1,405 EAS individuals with genomics, transcriptomics and proteomics measurements, serves as an important step forward for deciphering the complex multiomics landscape of human genetic variation in health and disease. All the eQTL and pQTL summary statistics, and the MPRA results, are publicly available via the Japan Omics Browser v.0.1 (https://japan-omics.jp/).

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01896-3.

References

1. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
2. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
3. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**, 535–550 (2016).
4. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644 (2020).
5. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
6. Pietzner, M. et al. Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
7. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
8. Zhang, J. et al. Plasma proteome analyses in individuals of European and African ancestry identify *cis*-pQTLs and models for proteome-wide association studies. *Nat. Genet.* **54**, 593–602 (2022).
9. Koprulu, M. et al. Proteogenomic links to human metabolic diseases. *Nat. Metab.* **5**, 516–528 (2023).
10. Brown, A. A. et al. Genetic analysis of blood molecular phenotypes reveals common properties in the regulatory networks affecting complex traits. *Nat. Commun.* **14**, 5062 (2023).
11. Zhao, J. H. et al. Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets. *Nat. Immunol.* **24**, 1540–1551 (2023).

12. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
13. Dhindsa, R. S. et al. Rare variant associations with plasma protein levels in the UK Biobank. *Nature* **622**, 339–347 (2023).
14. Eldjarn, G. H. et al. Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348–358 (2023).
15. Xu, F. et al. Genome-wide genotype-serum proteome mapping provides insights into the cross-ancestry differences in cardiometabolic disease susceptibility. *Nat. Commun.* **14**, 896 (2023).
16. Zhao, H. et al. Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. *Cell Genom.* **2**, 100195 (2022).
17. Namba, S., Konuma, T., Wu, K.-H., Zhou, W. & Okada, Y. A practical guideline of genomics-driven drug discovery in the era of global biobank meta-analysis. *Cell Genom.* **2**, 100190 (2022).
18. Folkersen, L. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).
19. Battle, A. et al. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
20. 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).
21. Namkoong, H. et al. *DOCK2* is involved in the host genetics and biology of severe COVID-19. *Nature* **609**, 754–760 (2022).
22. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
23. Gordon, M. G. et al. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412 (2020).
24. Pietzner, M. et al. Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.* **12**, 6822 (2021).
25. Dammer, E. B. et al. Multi-platform proteomic analysis of Alzheimer's disease cerebrospinal fluid and plasma reveals network biomarkers associated with proteostasis and the matrisome. *Alzheimers Res. Ther.* **14**, 174 (2022).
26. Katz, D. H. et al. Proteomic profiling platforms head to head: leveraging genetics and clinical traits to compare aptamer- and antibody-based methods. *Sci. Adv.* **8**, eabm5164 (2022).
27. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
28. Schreiber, G. The synthesis and secretion of plasma proteins in the liver. *Pathology* **10**, 394 (1978).
29. Jiang, L. et al. A quantitative proteome map of the human body. *Cell* **183**, 269–283 (2020).
30. He, B., Shi, J., Wang, X., Jiang, H. & Zhu, H.-J. Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. *BMC Biol.* **18**, 97 (2020).
31. Toikumo, S., Xu, H., Gelernter, J., Kember, R. L. & Kranzler, H. R. Integrating human brain proteomic data with genome-wide association study findings identifies novel brain proteins in substance use traits. *Neuropsychopharmacology* **47**, 2292–2299 (2022).
32. Mayr, C. What are 3′ UTRs doing? *Cold Spring Harb. Perspect. Biol.* **11**, a034728 (2019).
33. Griesemer, D. et al. Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247–5260 (2021).
34. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
35. Ringvall, M. et al. Defective heparan sulfate biosynthesis and neonatal lethality in mice lacking *N*-deacetylase/*N*-sulfotransferase-1. *J. Biol. Chem.* **275**, 25926–25930 (2000).
36. Reuter, M. S. et al. NDST1 missense mutations in autosomal recessive intellectual disability. *Am. J. Med. Genet. A* **164**, 2753–2763 (2014).
37. Sakaue, S. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
38. Kanai, M. et al. Insights from complex trait fine-mapping across diverse populations. Preprint at *medRxiv* https://doi.org/10.1101/2021.09.03.21262975v1 (2021).
39. Okada, Y. et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).
40. Matoba, N. et al. GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. *Nat. Hum. Behav.* **4**, 308–316 (2020).
41. Tomofuji, Y. et al. Prokaryotic and viral genomes recovered from 787 Japanese gut metagenomes revealed microbial features linked to diets, populations, and diseases. *Cell Genom.* **2**, 100219 (2022).
42. Sakaue, S. et al. Functional variants in *ADH1B* and *ALDH2* are non-additively associated with all-cause mortality in Japanese population. *Eur. J. Hum. Genet.* **28**, 378–382 (2020).
43. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
44. Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C. & Bu, G. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nat. Rev. Neurol.* **15**, 501–518 (2019).
45. Võsa, 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).
46. Fauman, E. B. & Hyde, C. An optimal variant to gene distance window derived from an empirical definition of *cis* and *trans* protein QTLs. *BMC Bioinformatics* **23**, 169 (2022).
47. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
48. Ferkingstad, E. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).
49. Hirata, J. et al. Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).
50. Rajagopalan, S. & Long, E. O. Understanding how combinations of HLA and KIR genes influence disease. *J. Exp. Med.* **201**, 1025–1029 (2005).
51. Moradi, S. et al. Structural plasticity of KIR2DL2 and KIR2DL3 enables altered docking geometries atop HLA-C. *Nat. Commun.* **12**, 2173 (2021).
52. Sakaue, S. et al. Decoding the diversity of killer immunoglobulin-like receptors by deep sequencing and a high-resolution imputation method. *Cell Genom.* **2**, 100101 (2022).
53. Kanai, M. et al. A second update on mapping the human genetic architecture of COVID-19. *Nature* **621**, E7–E26 (2023).
54. Franks, A., Airoldi, E. & Slavov, N. Post-transcriptional regulation across human tissues. *PLoS Comput. Biol.* **13**, e1005535 (2017).
55. Gry, M. et al. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 365 (2009).
56. Takemon, Y. et al. Proteomic and transcriptomic profiling reveal different aspects of aging in the kidney. *eLife* **10**, e62585 (2021).
57. Wang, Q. et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.* **11**, 2539 (2020).

58. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
59. Pak, M. A. et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLoS ONE* **18**, e0282689 (2023).
60. Wang, Q. S. et al. Leveraging supervised learning for functionally informed fine-mapping of *cis*-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
61. Sharon, E. et al. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.* **48**, 995–1002 (2016).

[1]Department of Genome Informatics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan. [2]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. [3]M&D Data Science Center, Tokyo Medical and Dental University, Tokyo, Japan. [4]Department of Infectious Diseases, Keio University School of Medicine, Tokyo, Japan. [5]Department of Pathology and Tumor Biology, Kyoto University, Kyoto, Japan. [6]Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine, Suita, Japan. [7]Division of Pulmonary Medicine, Department of Medicine, Keio University School of Medicine, Tokyo, Japan. [8]Faculty of Medicine, Osaka University, Suita, Japan. [9]Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [10]Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [11]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [12]Laboratory of Children's Health and Genetics, Division of Health Science, Osaka University Graduate School of Medicine, Suita, Japan. [13]Department of Pediatrics, Osaka University Graduate School of Medicine, Suita, Japan. [14]Division of Genome Analysis Platform Development, National Cancer Center Research Institute, Tokyo, Japan. [15]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan. [16]Health Science Research and Development Center (HeRD), Tokyo Medical and Dental University, Tokyo, Japan. [17]Laboratory of Veterinary Infectious Disease, Department of Veterinary Medicine, Kitasato University, Tokyo, Japan. [18]Department of Respiratory Medicine, Nagoya University Graduate School of Medicine, Nagoya, Japan. [19]Institute of Research, Tokyo Medical and Dental University, Tokyo, Japan. [20]Division of Gastroenterology and Hepatology, Department of Medicine, Keio University School of Medicine, Tokyo, Japan. [21]Division of Health Medical Intelligence, Human Genome Center, the Institute of Medical Science, University of Tokyo, Tokyo, Japan. [22]Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [23]Department of Immunopathology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan. [24]Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, Suita, Japan.
✉e-mail: qingbow@m.u-tokyo.ac.jp; hounamugun@keio.jp; yuki-okada@m.u-tokyo.ac.jp

Japan COVID-19 Task Force

Qingbo S. Wang[1,2], Takanori Hasegawa[3], Ho Namkoong[4], Ryunosuke Saiki[5], Ryuya Edahiro[2,6], Kyuto Sonehara[1,2], Hiromu Tanaka[7], Shuhei Azekawa[7], Shotaro Chubachi[7], Shinichi Namba[2], Kenichi Yamamoto[2,12,13], Yasuhito Nannya[5], Ryuji Koike[16], Tomomi Takano[17], Makoto Ishii[18], Akinori Kimura[19], Takanori Kanai[20], Koichi Fukunaga[7], Seishi Ogawa[5,15], Seiya Imoto[21], Satoru Miyano[3] & Yukinori Okada[1,2,22,23,24]

Methods

Ethics

We complied with all relevant ethical regulations. This study was approved by the ethics committees of the Keio University School of Medicine, the Osaka University Graduate School of Medicine and affiliated institutes. Written informed consent was obtained from all participants.

The JCTF

The study participants were recruited through the Japan COVID-19 Task Force (JCTF), which is described in detail in [21]. In this study, 1,405 genotyped samples from the JCTF cohort presenting varying levels of the COVID-19 phenotype at the time of recruitment and passing stringent quality control (QC) steps (as described in detail in the next sections) were analyzed. COVID-19 severity was categorized according to four levels: most severe for patients in the intensive care unit or requiring intubation and ventilation ($n = 501$); severe for others requiring oxygen support ($n = 494$); mild for other symptomatic patients (for example, those with shortness of breath; $n = 332$); and asymptomatic for those without COVID-19-related symptoms ($n = 78$). In 1,405 genotyped samples, mRNA expression in the whole blood was measured with RNA-seq in 1,019 samples, protein expression in the plasma was measured in 1,384 samples and 998 samples were at the intersection.

Genotyping, RNA-seq and protein expression measurement

Genotyping was performed using an Infinium Asian Screening Array (Illumina). Stringent sample and variant-level QC filters were applied (for example, sample call rate greater than 0.98, variant call rate greater than 0.99), resulting in $n = 1,405$ samples and a total of $n = 13,355,923$ variants ($n = 502,364$ genotyped and $n = 12,853,559$ imputed). For imputation, we extended our in-house and population-specific imputation reference panel; the extended, Japanese-specific reference panel included $n = 4,561$ whole-genome sequenced (WGS) data from multiple studies (for example, $n = 1,939$ from the BBJ study[39] and $n = 141$ WGS data from [62]), and was higher in number and population specificity compared to the imputation panel used in our previous transcriptomics study of the JCTF[20], which consisted of WGS data from 1,037 Japanese samples[63] plus those from the 1000 Genomes Project. We expected potential bias because the disease-ascertained nature of the WGS data in the reference panel was relatively low, as we observed an improvement in the eQTL fine-mapping results when compared to using the previous reference panel (Supplementary Fig. 1), suggesting a major benefit of a larger sample size and population specificity (Supplementary Note).

We lifted over the hg19 genotypes to hg38 using the Genome Analysis Toolkit (GATK) LiftoverVcf tool, and filtered out those without unique mapping. For the downstream QTL calling and fine-mapping analyses, we applied additional filtering based on the minor allele count (MAC) and imputation quality, while using a relatively lenient threshold (MAC > 2 and imputation $R^2 > 0.6$). The loose threshold setting was based on recent simulation-based observations[38] that fine-mapping benefits more from including low-frequency variants even with a relatively limited quality, and was validated with functional enrichment analyses (although warning us that low-frequency variants could be slightly enriched for false positives; Supplementary Fig. 2i–l and Supplementary Note).

RNA-seq was performed using the NovaSeq 6000 platform (Illumina) with paired end reads (read length of 100 bp), using the S4 Reagent Kit (200 cycles). Plasma protein expression was measured using the Olink Explore 3072 platform. The QC steps are described in the next sections.

eQTL and sQTL fine-mapping

We followed our previous pipeline for eQTL calling in principle, which is based on the GTEx pipeline and described in detail in [20]. To quantify mRNA expression, RNA-seq data were first aligned to the hg38 human reference genome using STAR v.2.5.3a. Transcripts were quantified using RSEM v.1.3.0. The following criteria were applied for sample QC: $0.5 \times 10^8 <$ number of mapped reads $< 3 \times 10^8$, mapping rate > 0.97, intergenic rate <0.05, rRNA rate <0.05, base mismatch rate <0.005 and intersample correlation deviation measure > −15 (the threshold was slightly different from the GTEx pipeline or our previous pipeline to fit the observed data distribution; Supplementary Fig. 22). mRNA expression was then trimmed mean of M component (TMM)-normalized while low-expression data were filtered out as in the GTEx pipeline. An eQTL call was performed using fastQTL, including 60 PEER factors, sex and five genotype principal components as covariates. Potential variations originating from technical factors, such as RNA integrity number, were thought to be captured by the PEER factors (Supplementary Fig. 23). Performing eQTL mapping for each disease severity stratum and combining the test statistics using a fixed effects model yielded generally consistent results (Supplementary Fig. 24).

Fine-mapping of eQTLs and other QTLs was performed based on our previous pipeline (that is, using FINEMAP v.1.3.1 and susieR v.0.11.43 with default parameters, where the inputs were the summary statistics and in-sample covariate-adjusted LD matrix), with two minor changes. First, we changed the number of single effects in the SuSiE model from ten to five, to be consistent with the default parameter in FINEMAP. This change in number led to slightly conservative but largely consistent results, as previously observed in ref. 20. Second, we performed fine-mapping of all genes regardless of their minimum P value (whereas in a previous version[20], we restricted this to genes with a minimum $P < 5 \times 10^{-8}$), as we had a larger sample size and higher power. This could result in a small but limited number of false positives (described in detail in ref. 64). Throughout the article, we report PIP as the minimum of the output from FINEMAP and SuSiE, and define putative causal QTLs as PIP greater than 0.9.

sQTL calling and fine-mapping were also performed mainly based on the previous pipeline (that is, the splicing level was quantified using LeafCutter v.0.2.7 and 15 PEER factors were included) with the changes described above and the following additional modifications to reduce possible false positive calls due to noisy annotation-free splice variant quantification: (1) we defined the *cis*-window to be ±0.1 Mb from the center of the intron cluster; and (2) we omitted several additional filters from the GTEx pipeline (WASP and z-score-based filtering). Instead, we applied an additional stringent filtering step after nominal sQTL calling, which filtered out intron clusters with a minimum $P > 5 \times 10^{-8}$, analogous to our eQTL fine-mapping in a smaller sample[20].

pQTL fine-mapping

The Olink Explore 3072 platform quantifies the expression of each protein in a normalized scale (normalized protein expression (NPX)). As measuring proteins was separated into three batches for logistical reasons, we bridge-normalized the NPX values using the OlinkAnalyze R package[65], using 16 intersecting samples as bridging samples. The distribution of COVID-19 severity was similar between batches (Supplementary Fig. 4). To be consistent with the eQTL fine-mapping pipeline and other major pQTL studies using Olink data, we further inverse-normal transformed the bridge-normalized NPX matrix. Samples with QC warning flags were removed. We did not apply additional sample QC after confirming that there were no outlier and major batch effects in the first two principal component spaces (Supplementary Fig. 4) after bridge normalization. Although we observed a high percentage of measurements below the LOD for a subset of genes, we followed the guideline on the Olink website (see 'How is LOD estimated for Olink Explore 3072/384' and 'Explore HT and what is recommended downstream usage?' at https://olink.com/faq) and did not explicitly remove or adjust those entries. We instead evaluated the potential effect of samples below the LOD post hoc and found no major bias introduced by the inclusion of samples below the LOD (Supplementary Fig. 21 and Supplementary Note).

The gene names from the Olink platform were converted into canonical gene names and Ensembl gene IDs based on gencode v.30. For the assays corresponding to multiple gene names separated by an underbar '_' (for example, CGB3_CGB5_CGB8), we empirically split the entry into multiple entries with an identical value (in the case above, we interpreted that three proteins, CGB3, CGB5 and CGB8, had the same measurement value). Small numbers of genes where any gene name alias in GeneCards (https://www.genecards.org/) did not match the gene name in gencode v.30 were excluded from the analysis. When a gene name was mapped to multiple Ensembl IDs, we exploded the matrix and included all the Ensembl IDs separately. When multiple measurements (either multiple gene names or assays) were mapped to an Ensembl ID, we collapsed the measurements by taking the mean. In all such cases, the differences between measurements were minimal, limiting the bias introduced by this step (Supplementary Fig. 4).

After creating the protein expression matrix as described above, QTL calling and fine-mapping were performed according to the same steps as in eQTL fine-mapping (that is, sex, five genotype principal components and 60 PEER factors aimed at capturing sample-to-sample technical variation, each recalculated within the genotype and expression matrix), except that the step corresponding to filtering the genes based on TPM and the number of samples with nonzero TPM did not exist. Thus, 2,932 proteins were included in QTL calling. We acknowledge the slight difference in the number of unique proteins measured compared to a recent large-scale analysis using the same Olink Explore 3072 platform, such as the UKB PPP, because of minor differences in the data processing step, as described above. The biological properties of these 2,932 genes compared to all coding genes; for example, enrichment in inflammatory functions, as noted on their website (https://olink.com/products/olink-explore-3072-384) are summarized in Supplementary Fig. 25.

pQTL replication

The UKB PPP data were downloaded from the Synapse portal (https://www.synapse.org/#!Synapse:syn51365301). We matched each canonical gene name as in gencode v.30 and the variant ID in hg19 to perform the comparison. For the effect size concordance analysis, pQTLs with $P > 0.05$ in the UKB PPP dataset were not included for simplicity. The ARIC[8] data (the full summary statistics and PIPs) were kindly shared by the authors (http://nilanjanchatterjeelab.org/pwas/). We matched the Ensembl gene IDs based on gencode v.30 and the variant IDs in hg38, removed any unmatched entries and performed the comparison. For the EPIC-Norfolk study, we downloaded the supplementary data from ref. 9 and matched the Ensembl gene IDs and the variant IDs (lifted to hg38). For the mass spectrometry-derived pQTL data, we downloaded the supplementary data from ref. 15 (Supplementary Data 3, summary of identified pQTLs) and matched them on the variant ID (lifted to hg38). As the mass spectrometry data did not directly nominate the affected genes, we matched the variant ID alone and assumed that the gene with the most significant *cis*-pQTL effect in our dataset was the affected gene to make the comparison. Different prefiltering strategies could contribute to the differences in replication rate, deeming our comparison of the replication rate between cohorts as semiquantitative.

mRNA-specific and protein-specific eQTL fine-mapping

To investigate the mRNA-specific or protein-specific expression of QTLs, we focused on 998 samples and 2,211 genes with both QC-passed mRNA and protein expression measurements. Normalization was reperformed within this reduced sample-by-gene matrix, for mRNA and protein expression separately. For each normalized mRNA (or protein) expression datum, linear regression using protein (or mRNA) expression of the same gene as the only variable was performed to calculate the regression coefficient. The mRNA (or protein) expression data where the linear effect of protein (or mRNA) were regressed out using the model described above were used as the mRNA-specific or protein-specific expression matrix. QTL calling and fine-mapping were performed in the same manner as described by others (that is, preparing the same set of covariates in the QTL calling and with the same parameter setting in the fine-mapping algorithm).

Targeted *trans*-eQTL and *trans*-pQTL calling

We first focused on the common lead pQTLs (variants with the lowest pQTL *P* value for each gene, minor allele frequency > 0.01) and tested their genome-wide pQTL effects (Bonferroni-corrected $P = 0.05/4{,}569{,}247$, where 4,569,247 is 1,569 common lead variants times 2,932 genes and minus 31,061 *cis*-variant–gene pairs within a 5-Mb distance to the TSS) using tensorQTL, including the same set of covariates as in *cis*-pQTL mapping. Both the lead pQTLs and corresponding random control variants were filtered to minor allele frequency >0.01 for consistency.

When focusing on the genetic variation in classical *HLA* genes to test genome-wide *trans*-eQTL and *trans*-pQTL effects, DEEP*HLA[66] was used to impute HLA alleles for the 998 samples with both mRNA and protein measurements. We focused on 144 four-digit alleles passing QC (imputation $R^2 > 0.7$ and minor allele frequency > 0.01) mainly on classical *HLA* genes (*HLA*-A, *HLA-B*, *HLA-C*, *DRB1*, *DQA1*, *DQB1*, *DPA1*, *DPB1* and *MICA*). tensorQTL was used to test the *trans*-effect (defined by >5 Mb distance between the center of the MHC region and the gene TSS) of each of the four-digit alleles on genome-wide mRNA and protein expression. For each of the *HLA* genes, the minimum *P* value over all the four-digit alleles were displayed in a Miami plot; those passing a suggestive $P < 1 \times 10^{-5}$ threshold were annotated. The MHC region (defined as chr6:25726063–33400644) was not included in the *cis*-QTL fine-mapping analyses because of their high complexity leading to lower mapping quality (that is, while we controlled any biases due to a difference in LD structure by focusing on the same set of 998 samples, we left the fine-mapping of causal variants within *HLA*, whether in *cis* or *trans*, on mRNA or protein expression, as future work).

ieQTL and ipQTL calling

COVID-19 severity ieQTL calling (ieQTL calling) was performed according to the pipeline described previously in ref. 20. Briefly, we used COVID-19 severity classified into four levels as the interaction term, kept all the other terms such as principal components and PEER factors, and used tensorQTL to obtain the *P* value from the likelihood ratio test. ipQTL calling was performed in a similar fashion. Results were largely consistent when excluding age from the covariates (Supplementary Fig. 20); they were far from identical when testing for the interaction with age instead of COVID-19 phenotype, suggesting that the inclusion of age was not introducing bias and that the COVID-19 interaction effects were not simply driven by shifts in age distribution (Supplementary Note). Samples with both RNA-seq and protein measurements ($n = 998$) were used for ieQTL and ipQTL calling.

MPRA

Massively parallel reporter assay (MPRA) is a high-throughput method that allows quantification of variant effects by measuring the transcriptional activity of many reporters inserted with different sequence elements. Our MPRA library contained 24,000 oligonucleotides, allowing us to test nearly 12,000 variants across the genome (that is, (24,000 − no. of controls)/2). While the library design is described in a separate manuscript in preparation, using K562 and HepG2 cells, we systematically tested variants with PIP > 0.1 in the previous eQTL fine-mapping from the JCTF[20], PIP > 0.1 in the previous functionally informed fine-mapping of the GTEx whole-blood eQTLs[60], and a small number of additional variants with possible phenotypic effects, while removing variants with an insertion and deletion length greater than 70.

The MPRA experiments were performed according to the steps in ref. 23. Briefly, the MPRA library was synthesized by Agilent, amplified using PCR, adding random barcodes and cloned into the pLS-SceI vector (no. 137725, Addgene). Sequence–barcode associations in the plasmid library were determined by sequencing using the NextSeq

Mid Output 300 cycle kit. The plasmid library was packaged with lentivirus and infected into 2.8 million HepG2 or 10 million K562 cells at a multiplicity of infection of 50 and 10, respectively. For each cell line, three independent infections were performed to obtain three biological replicates. After 3 days, genomic DNA and total RNA were extracted using an AllPrep DNA/RNA Mini Kit (no. 80204, QIAGEN). Total RNA was reverse-transcribed to generate complementary DNA, using Superscript IV Reverse Transcriptase (Thermo Fisher Scientific). Integrated and transcribed barcodes were amplified with specific primers while incorporating the 16-bp unique molecular identifier and Illumina sequencing adapters. Barcodes were then sequenced using the NextSeq High Output 75 cycle kit.

MPRAflow[23] was used to associate the barcodes to individual 200-bp sequences and to count the number of DNA and RNA reads (Supplementary Fig. 26). After filtering noisy barcodes with fewer than five DNA counts in a replicate, as well as noisy variants with fewer than five total barcode–replicate pairs in reference or alternative alleles, we aggregated all barcode counts for each sequence across replicates and defined the allelic effect of each variant as $\log_2$((total RNA count/total DNA count in alt)/(total RNA count/total DNA count in ref)). We then performed permutation-based significance tests to obtain the *P* value for each variant and used the Benjamini–Hochberg method to estimate the FDR. Variants passing an FDR threshold of 0.01 were defined as tier 1 EMVars; other variants passing an FDR threshold of 0.1 were defined as tier 2 EMVars. As described in detail in the Supplementary Note, our analysis was based on a simplified model where RNA count follows a Poisson distribution with $\lambda = \alpha \times$ (DNA count), where the translation rate α is a function of the 200-bp sequence alone.

As shown in Extended Data Fig. 2c,d, we did not observe a major difference in enrichment level between K562 and HepG2 cells, even though K562 cells are physiologically more relevant to blood, suggesting that the effect of culturability, transduction efficiency and other methodological details could differ depending on cell type.

Defining shared and specific QTLs

To investigate shared and distinct QTL effects on mRNA and protein expression, we calculated the product of the eQTL and pQTL PIP (recalculated using the identical set of $n = 998$ samples for each of the intersecting $n = 2{,}211$ genes) to define the CLPP. As eQTL and pQTL calling from identical samples were nonindependent, we deemed this CLPP assignment as conservative, and thus took different thresholds to declare (single-variant-level) colocalization: CLPP ≥ 0.9 for 'colocalizing QTLs' and CLPP ≥ 0.1 for 'possibly colocalizing QTLs'. We deem how to set the threshold to declare colocalization as an open problem, although as analyzed in the Results section, using different thresholds did not qualitatively change the results.

When performing coloc, we turned the *P* values into Bayes factors assuming a single causal variant per gene and calculated the PP.H1 to PP.H4 with the default prior probability setting described in their manuscript[67]. When performing SuSiE-coloc, we calculated the PP.H1 to PP.H4 for each combination of pure mRNA and protein QTL credible sets in a gene and let the maximum PP.H4 be the colocalization probability for a gene. (We did not relax the purity filter for credible sets, which could result in slightly conservative results compared to their original implementation.) For the main results, we did not use canonical colocalization[67] because we observed more than one causal eQTL or pQTL effect in each. SuSiE-coloc[68] was also not chosen as our main method because we aimed to also use the results from FINEMAP. We took the minimum PIPs from SuSiE and FINEMAP, resulting in a slightly more conservative CLPP estimation than using one of the two (Supplementary Fig. 13i).

mRNA-specific QTLs were defined as those with a protein-adjusted eQTL PIP ≥ 0.9 and CLPP < 0.01 (CLPP as described above); protein-specific QTLs were defined as mRNA-adjusted pQTL PIP ≥ 0.9 and CLPP < 0.01.

Functional enrichment

Variant Effect Predictor (VEP) v.108 was used to annotate the variant–gene pairs with several functional annotations. Instead of using the canonical 'most severe consequence', which could be tagged by the effect on other genes, we parsed the annotation at the unit of the variant–gene pairs to be more specific. When multiple annotations were included, we did not filter to the most severe ones canonically; thus, each variant–gene pair could be positive for more than one annotation. Protein-specific annotations were obtained as BED files from UniProt through the UCSC Genome Browser. We used pybedtools to check whether a variant intersected with each specific protein annotation.

Enrichment of a category *C* specific functional annotation given a bin *B* PIP was defined as the likelihood ratio compared to a random draw (that is, $p(v \in C \mid v \in B)/p(v \in C)$). The error bar denotes the s.e.m. of the numerator (that is, we assumed that the denominator contained a large number of variant–gene pairs and that the error could be trivial). Plots are often displayed in log-scale for visualization purposes. The colocalization enrichment score between our pQTL calling and eQTL calling in GTEx v.8 across 49 tissues (Fig. 2e) was calculated as follows: (1) We obtained variant–gene–tissue triads with PIP > 0.001 in GTEx v.8 (where the PIP was calculated with a uniform prior and the minimum of SuSiE and FINEMAP was taken, as described in ref. 60) and filtered out missense variants. We annotated the pQTL PIP in JCTF as a function of variant–gene while filling with zero when either the PIP was lower than 0.001 or missing in the JCTF. (2) Then, for each tissue *T*, we calculated the average of the CLPP, defined as the product of the eQTL and pQTL PIPs. Let this be U_T. (3) We randomly selected the same number of variant–gene–tissue triads from tissues other than *T* and calculated the average CLPP. We repeated the process (that is, sampling the same number of variant–gene pairs excluding the tissue of interest and calculating the average CLPP) for 1,000 times (let these be $V_{T,1}$, $V_{T,2}$,…,$V_{T,1000}$); (4) our point estimate of the enrichment score is the mean of U_T divided by V_T and the error bar is the 2.5–97.5% quantile. Although the estimation could be deflated as it does not distinguish between variants missing versus having a PIP below the 0.001 threshold, and the power was dependent on the sample size for each GTEx tissue, we expected to qualitatively characterize the different level of colocalization between plasma proteome and each tissue.

Complex trait and disease analysis

BBJ and UKB fine-mapping data, described in detail in ref. 38, were downloaded from the National Bioscience Database Center (NBDC) Human Database (accession no. hum0197) and the Finucane lab (https://www.finucanelab.org/data). PIP enrichment was defined for each variant–gene–trait triad as the sum($\mathrm{PIP}_{\mathrm{molQTL}} \times \mathrm{PIP}_{\mathrm{trait}}$) (those with $\mathrm{PIP}_{\mathrm{trait}} < 0.001$ were not included in this calculation; this had a minimal effect on the quantification); then, the maximum over genes was taken to obtain the per-variant scores. When the analysis was not trait-specific, or according to the trait category, the maximum over the traits in a category was taken. When visualizing the number of colocalizing genes per trait (Fig. 4a), we used a custom threshold of $0.1 \le \min(\mathrm{PIP}_{\mathrm{molQTL}}, \mathrm{PIP}_{\mathrm{trait}})$, instead of using the CLPP, to allow comparison between shared versus specific QTLs (that is, for shared QTLs, the threshold was $0.1 \le \min(\mathrm{PIP}_{\mathrm{eQTL}}, \mathrm{PIP}_{\mathrm{pQTL}}, \mathrm{PIP}_{\mathrm{trait}})$).

Negative expression correlation analysis

For the analysis of mRNA and protein QTL entanglement at the *ABO* locus, we selected five variants corresponding to the top PIP variants from two eQTL credible sets and two pQTL credible sets from SuSiE that were in LD, removing the ones with limited Bayes factors or far away from the variant of highest interest (=p1). When comparing the QTL effect with and without controlling for the others, we restricted the analysis to $n = 998$ samples with both mRNA and protein measurements.

Statistical analysis
All the statistical tests were two-sided. No adjustment was made for the reported *P* values unless it was clearly stated as 'adjusted *P* value'. The error bars denote the s.e.m. unless noted otherwise. When showing enrichment, the enrichment error bar denotes the standard error of the numerator divided by the denominator.

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

Data availability
The summary statistics of the QTL analyses and the RNA-seq expression matrix are available at the NBDC Human Database (accession no. hum0343). The QTL summary statistics are also available at https://japan-omics.jp/. Individual genotype data are available at the European Genome-phenome Archive (accession no. EGAS00001006284). Publicly available datasets used are: BBJ and UKB fine-mapping; NBDC Human Database (accession no. hum0197) and https://www.finucanelab.org/data; the expression modifier score (https://www.finucanelab.org/data); the GTEx *cis*-eQTL data (https://gtexportal.org/home/datasets); the hg38 reference genome (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/); protein-specific annotations from UniProt, obtained through the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTables); protein QTL data from the ARIC study (http://nilanjanchatterjeelab.org/pwas/); and protein QTL data from the UKB PPP study (https://www.synapse.org/#!Synapse:syn51365301).

Code availability
The code used in this study is available at https://github.com/QingboWang/japan_covid_taskforce_multi_omics and has been deposited via Zenodo at https://doi.org/10.5281/zenodo.11169201 (ref. 69). The software and tools used for data analysis and visualization are: DEEP*HLA v.1.0.0 (https://zenodo.org/record/4478902)[70]; fastQTL v.2.165 (http://fastqtl.sourceforge.net); FINEMAP v.1.3.1 (http://www.christianbenner.com/); GATK v.4.1.9.0 LiftoverVcf (https://gatk.broadinstitute.org/); the GTEx pipeline (https://github.com/broadinstitute/gtex-pipeline); LeafCutter v.0.2.7 (https://davidaknowles.github.io/leafcutter/index.html); matplotlib v.3.3.4 (https://matplotlib.org); MPRAflow v.2.3.5 (https://mpraflow.readthedocs.io/en/latest/index.html); NumPy v.1.20.1 (https://numpy.org); OlinkAnalyze v.3.4.1 (https://cran.r-project.org/web/packages/OlinkAnalyze/index.html); pandas v.1.1.4 (https://pandas.pydata.org); pybedtools v.0.9.0 (https://daler.github.io/pybedtools/); PyWGCNA v.1.20.3 (https://github.com/mortazavilab/PyWGCNA); RSEM v.1.3.0 (https://deweylab.github.io/RSEM/); scikit-learn v.0.24.1 (https://scikit-learn.github.io/stable); SciPy v.1.6.2 (https://scipy.org/); seaborn v.0.11.1 (https://seaborn.pydata.org); STAR v.2.5.3a and v.2.6.0 (https://github.com/alexdobin/STAR); susieR v.0.11.43 (https://github.com/stephenslab/susieR); tensorQTL v.1.0.5 (https://github.com/broadinstitute/tensorqtl); TwoSampleMR v.0.5.7 (https://mrcieu.github.io/TwoSampleMR/articles/introduction.html); and VEP v.108 (https://asia.ensembl.org/Homo_sapiens/Tools/VEP/).

References

62. Sonehara, K. et al. Genetic architecture of microRNA expression and its link to complex diseases in the Japanese population. *Hum. Mol. Genet.* **31**, 1806–1820 (2022).
63. Akiyama, M. et al. Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **10**, 4393 (2019).
64. Wang, Q. S. et al. Estimating gene-level false discovery probability improves eQTL statistical fine-mapping precision. *NAR Genom. Bioinform.* **5**, lqad090 (2023).
65. Nevola, K. et al. OlinkAnalyze: Facilitate analysis of proteomic data from Olink. R version 3.4.1 https://cran.r-project.org/web/packages/OlinkAnalyze/index.html (2023).
66. Naito, T. et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nat. Commun.* **12**, 1639 (2021).
67. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
68. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).
69. Wang, Q. S. QingboWang/japan_covid_taskforce_multi_omics: v1.0 (v1.0). *Zenodo* https://doi.org/10.5281/zenodo.11169202 (2024).
70. tatsuhikonaito/DEEP-HLA: First release of DEEP*HLA (v.1.0.0). *Zenodo* https://zenodo.org/records/4478902 (2021).

Acknowledgements

We thank all the participants involved in this study, and all the members of the JCTF for their support. We thank J. Kitano, the e-Parcel Corporation and the Ascend Corporation for supporting the JCTF. This study was supported by the Japan Agency for Medical Research and Development (AMED) (nos. JP23kk0305022, JP22ek0410075, JP23km0405211, JP23km0405217, JP23ek0109594, JP23ek0410113, JP223fa627002, JP223fa627010, JP233fa627011, JP23zf0127008, JP23tm0524002, JP22fk0108510, JP21fk0108553, JP21fk0108431, JP20fk0108415 and JP20fk0108452), JST CREST (no. JPMJCR20H2), JST FOREST (no. JPMJFR225Y), JST PRESTO (no. JPMJPR21R7), JST Moonshot R&D (nos. JPMJMS2021 and JPMJMS2024), MHLW (no. 20CA2054), JSPS KAKENHI (nos. 22H00476 and 23K14233), the Nakajima Foundation, the Uehara Memorial Foundation, the Takeda Science Foundation, the Mitsubishi Foundation and the Bioinformatics Initiative of the Osaka University Graduate School of Medicine, the Institute for Open and Transdisciplinary Research Initiatives, the Center for Infectious Disease Education and Research and the Center for Advanced Modality and DDS, Osaka University. The super-computing resource was provided by the Human Genome Center (University of Tokyo).

Author contributions
Q.S.W. and Y.O. designed the study. Q.S.W., Y.T., H.N., T.H., S.I., S.M. and Y.O. analyzed the data. Q.S.W. wrote the manuscript. Y.O. reviewed and edited the manuscript. H.N. and Y.O. supervised the work. All authors and the JCTF contributed to the generation of the primary data incorporated in the study, provided inputs and approved the manuscript.

Competing interests
Q.S.W. is an employee of Calico Life Sciences. The other authors declare no competing interests.

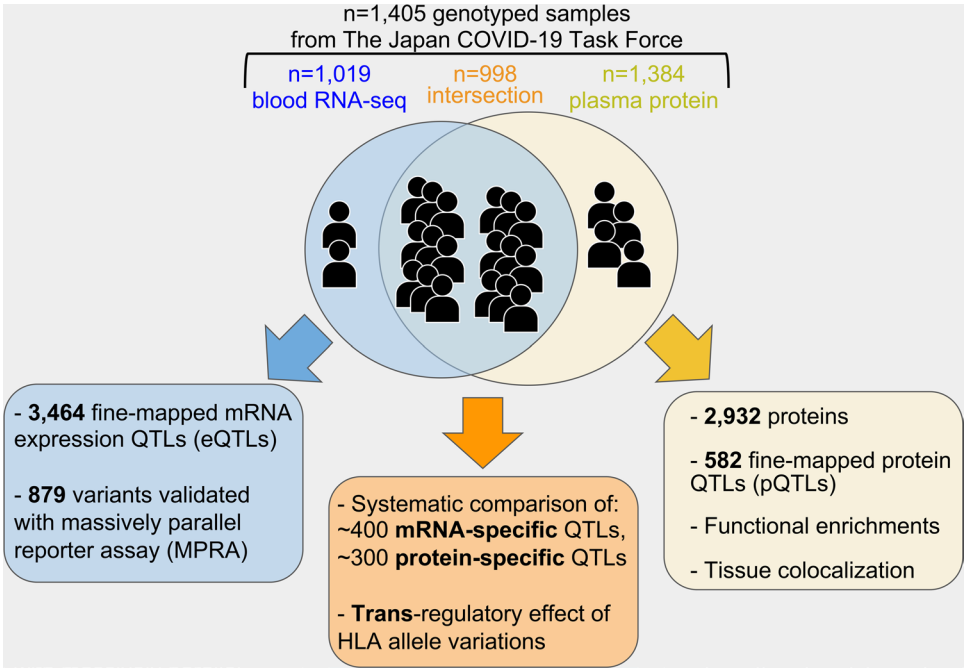
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01896-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01896-3.

Correspondence and requests for materials should be addressed to Qingbo S. Wang, Ho Namkoong or Yukinori Okada.

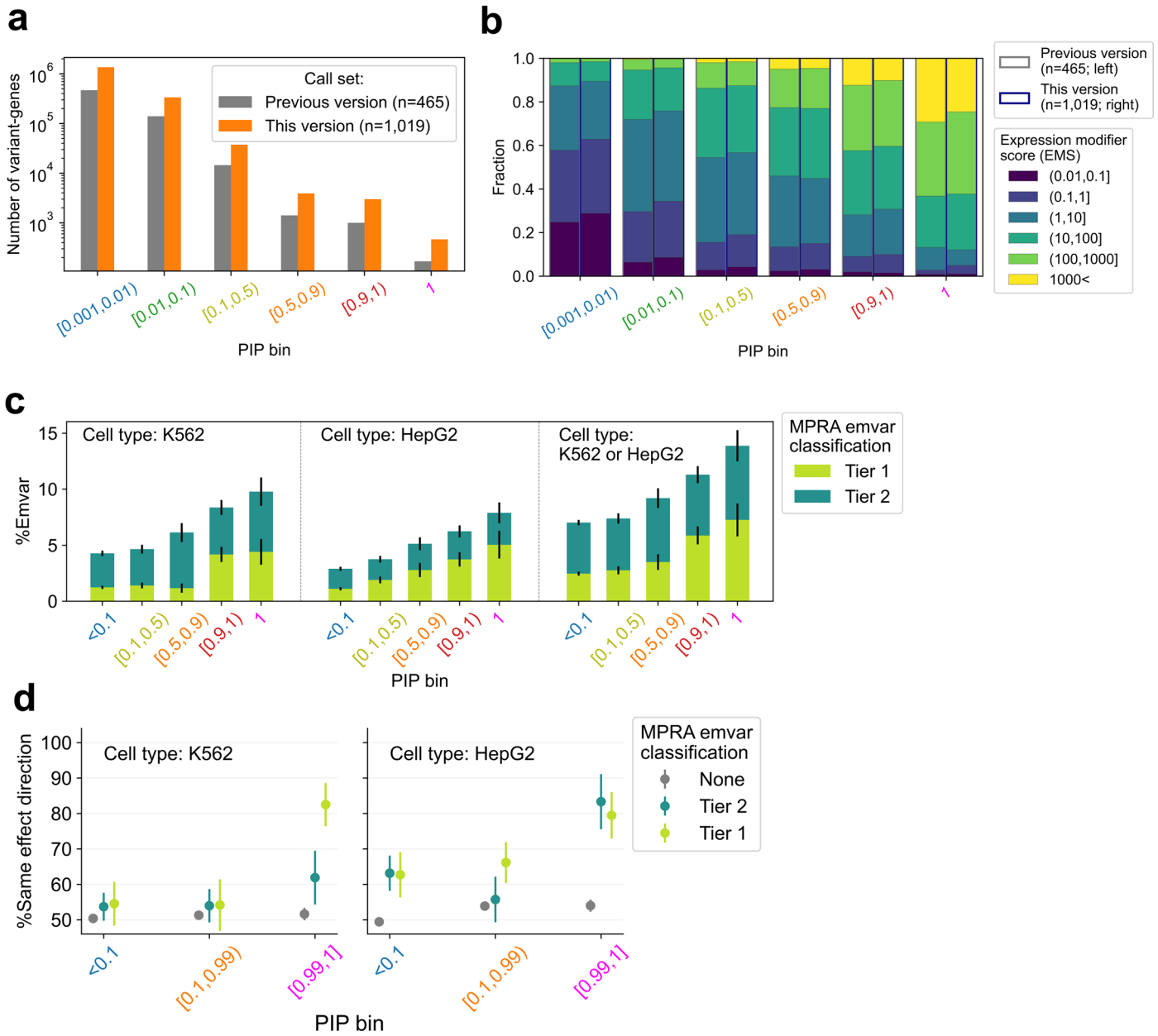
Peer review information *Nature Genetics* thanks Anders Malarstig, Clint Miller and Maik Pietzner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-024-01896-3



Extended Data Fig. 1 | Overview of the study. We performed mRNA expression QTL (eQTL) fine-mapping from 1,019 RNA-sequenced samples, pQTL fine-mapping from 1,384 protein measured samples, as well as mRNA or protein specific QTL fine-mapping from 998 samples with both measures, all genotyped and processed in a single platform as part of the Japan COVID-19 Task Force[20]. Massive parallel reporter assay (MPRA) was performed for validation of a subset of fine-mapped eQTLs.

https://doi.org/10.1038/s41588-024-01896-3



Extended Data Fig. 2 | eQTL fine-mapping expanded. a. Comparison of the numbers of eQTLs in our dataset compared to the previous release. **b.** Functional score (the expression modifier score = EMS) enrichment in eQTLs along with the posterior inclusion probability (PIP). **c.** Percentage of expression modifying variants (emvars) experimentally validated in massive parallel reporter assay (MPRA). Tier 1 corresponds to FDR < 0.01 and tier 2 to FDR < 0.1. n in each bin = 7,418, 2,060, 685, 885 and 317 variants. **d.** Percentage of agreement between the direction of variant effects in eQTL or MPRA study. n in each bin = 7,418, 2,992 and 955 variants.

nature portfolio

Corresponding author(s): Qingbo Wang, Ho Namkoong, Yukinori Okada

Last updated by author(s): 06/20/2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection.
Data analysis	The set of softwares and tools used for the analysis as well as data visualization are listed as below; DEEP*HLA v1.0.0 (https://zenodo.org/record/4478902) fastQTL v2.165 (http://fastqtl.sourceforge.net) FINEMAP v1.3.1 (http://www.christianbenner.com/) GATK v4.1.9.0 LiftoverVcf (https://gatk.broadinstitute.org/) GTEx pipeline (https://github.com/broadinstitute/gtex-pipeline) LeafCutter v0.2.7 (https://davidaknowles.github.io/leafcutter/index.html) matplotlib v3.3.4 (https://matplotlib.org) MPRAflow v2.3.5 (https://mpraflow.readthedocs.io/en/latest/index.html) numpy v1.20.1 (https://numpy.org) OlinkAnalyze v3.4.1 (https://cran.r-project.org/web/packages/OlinkAnalyze/index.html) pandas v1.1.4 (https://pandas.pydata.org) pybedtools v0.9.0 (https://daler.github.io/pybedtools/) pyWGCNA v1.20.3 (https://github.com/mortazavilab/PyWGCNA) RSEM v1.3.0 (https://deweylab.github.io/RSEM/) scikit-learn v0.24.1 (https://scikit-learn.github.io/stable) scipy v1.6.2 (http://scikit-learn.github.io/stable) seaborn v0.11.1 (https://seaborn.pydata.org) STAR v2.5.3a and v2.6.0 (https://github.com/alexdobin/STAR)

susieR v0.11.43 (https://github.com/stephenslab/susieR)
tensorQTL v1.0.5 (https://github.com/broadinstitute/tensorqtl)
TwoSampleMR v0.5.7 (https://mrcieu.github.io/TwoSampleMR/articles/introduction.html)
Variant Effect Predictor (VEP) version 108 web interface (https://asia.ensembl.org/Homo_sapiens/Tools/VEP/)

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The summary statistics of QTL analyses, as well as the RNA-seq expression matrix are available at the National Bioscience Database Center (NBDC) Human Database (accession code: hum0343; https://humandbs.biosciencedbc.jp/en/hum0343). The QTL summary statistics are also visible at the browser https://japan-omics.jp/. The individual genotype data is available at European Genome-Phenome Archive (EGA) (accession code: EGAS00001006284; https://ega-archive.org/studies/EGAS00001006284).

The list of publicly available datasets used are listed below:
Biobank Japan (BBJ) and UK Biobank (UKB) fine-mapping:
NBDC Human Database (accession code: hum0197; https://humandbs.biosciencedbc.jp/en/hum0197) and https://www.finucanelab.org/data
hg38 reference genome: https://hgdownload.soe.ucsc.edu/goldenPath/hg38/
The expression modifier score (EMS): https://www.finucanelab.org/data
Genotype-Tissue Expression (GTEx) cis-eQTL data: https://gtexportal.org/home/datasets
Protein specific annotations from Uniprot, obtained through the UCSC genome browser; https://genome.ucsc.edu/cgi-bin/hgTables
Protein-QTL data from the ARIC study; http://nilanjanchatterjeelab.org/pwas/
Protein-QTL data from the UKB-PPP study; https://www.synapse.org/#!Synapse:syn51365301

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No selection based on sexual orientation was performed.
Reporting on race, ethnicity, or other socially relevant groupings	We did not pre-select study samples based on any social groupings.
Population characteristics	Study participants are of East Asian ancestry (age mean = 59.9 yrs old, sd = 17.1), tested positive for PCR test results.
Recruitment	We enrolled participants diagnosed as COVID-19 positive by physicians using the clinical manifestation and PCR test results at one of the >100 the affiliated hospitals participating to Japan COVID-19 Task Force. Any subjects with obtained informed consent were included without further biases. Due to the nature of COVID-19 susceptibility, the number of male participants were larger than that of females (997 / 1,405 = 70.1% male).
Ethics oversight	This study was approved by the ethical committees of Keio University School of Medicine, Osaka University Graduate School of Medicine, and affiliated institutes. Informed consent was obtained from all participants.

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	The study participants were recruited through Japan COVID-19 Task Force. Whole blood-RNA-sequencing and/or plasma protein expression assay was performed for a subset of the genotyped samples (n=1405) and analyzed in this study. Although no sample size was predetermined due to the unpredictable nature of the COVID-19 outbreak, the current sample size is deemed reasonable, as it is larger than that of major

	bulk RNA-seq studies such as GTEx (n = 670 for whole blood) and the size of east Asian samples in a major pQTL study (UKB-PPP, east Asian n=262).
Data exclusions	Stringent sample and variant level quality control (QC) filters were applied (e.g. call rate >0.97). The distribution of the quality are available at Supplementary Figures
Replication	Although we did not attempt to replicate our results by constructing another dataset of the same nature due to its uniqueness, we replicated our main findings by comparing with existing databases such as GTEx and UK Biobank.
Randomization	We did not need to use randomization in this study because this is a genotype-gene expression association study. All the samples with available accessibility to genotype and RNA/protein expression data passing quality control threshold were included in the analysis.
Blinding	We did not apply blinding of the samples because this is a genotype-gene expression association study and no intervention was conducted in our study.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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