nature genetics

Technical Report

https://doi.org/10.1038/s41588-023-01421-y

Mapping interindividual dynamics of innate immune response at single-cell resolution

Received: 28 August 2021

Accepted: 27 April 2023

Published online: 12 June 2023

Check for updates

Natsuhiko Kumasaka[1,2], Raghd Rostom[1,3], Ni Huang[1], Krzysztof Polanski[1], Kerstin B. Meyer[1], Sharad Patel[1], Rachel Boyd[1], Celine Gomez[1], Sam N. Barnett[1], Nikolaos I. Panousis[1], Jeremy Schwartzentruber[1,4], Maya Ghoussaini[1,4], Paul A. Lyons[5,6], Fernando J. Calero-Nieto[7], Berthold Göttgens[7], Josephine L. Barnes[8], Kaylee B. Worlock[8], Masahiro Yoshida[8], Marko Z. Nikolić[8,9], Emily Stephenson[1,10], Gary Reynolds[10], Muzlifah Haniffa[1,10,11,12], John C. Marioni[1,3,13], Oliver Stegle[1,3,14,15], Tzachi Hagai[16,18] ✉ & Sarah A. Teichmann[1,17,18] ✉

Common genetic variants across individuals modulate the cellular response to pathogens and are implicated in diverse immune pathologies, yet how they dynamically alter the response upon infection is not well understood. Here, we triggered antiviral responses in human fibroblasts from 68 healthy donors, and profiled tens of thousands of cells using single-cell RNA-sequencing. We developed GASPACHO (GAuSsian Processes for Association mapping leveraging Cell HeterOgeneity), a statistical approach designed to identify nonlinear dynamic genetic effects across transcriptional trajectories of cells. This approach identified 1,275 expression quantitative trait loci (local false discovery rate 10%) that manifested during the responses, many of which were colocalized with susceptibility loci identified by genome-wide association studies of infectious and autoimmune diseases, including the *OAS1* splicing quantitative trait locus in a COVID-19 susceptibility locus. In summary, our analytical approach provides a unique framework for delineation of the genetic variants that shape a wide spectrum of transcriptional responses at single-cell resolution.

The innate immune response is a cell-autonomous program that induces an antiviral state in infected and nearby cells and alerts the immune system of the invading pathogen[1]. Dysregulation of this response can affect a wide range of inflammatory and autoimmune diseases and determine the outcome of infection[2–6]. Common genetic variants have been shown to modulate transcriptional responses to various viral and bacterial stimuli, and to contribute to disease onset and progression[7–11]. Most past gene-expression-focused studies of this program are based on bulk RNA-sequencing (RNA-seq) technologies, which do not fully elucidate the continuous dynamics of transcriptional changes during the innate immune response. Single-cell genomic technologies are powerful approaches to study cell heterogeneity and transcriptional variability across cells[12]. Furthermore, by utilizing single-cell RNA-seq (scRNA-seq) profiling of tissues composed of several cell lineages, previous studies have successfully performed genetic association mapping of cell-type-specific expression[13–19].

We here use full-length scRNA-seq of dermal fibroblasts from different human individuals, challenged with immune stimuli. Based on the pseudo-temporal reconstruction of these data, we map the transcriptional variation of the innate immune response at single-cell resolution. This provides the foundation for superimposing human genetic variation onto the transcriptional dynamics of this response.

A full list of affiliations appears at the end of the paper. ✉e-mail: tzachiha@tauex.tau.ac.il; st9@sanger.ac.uk

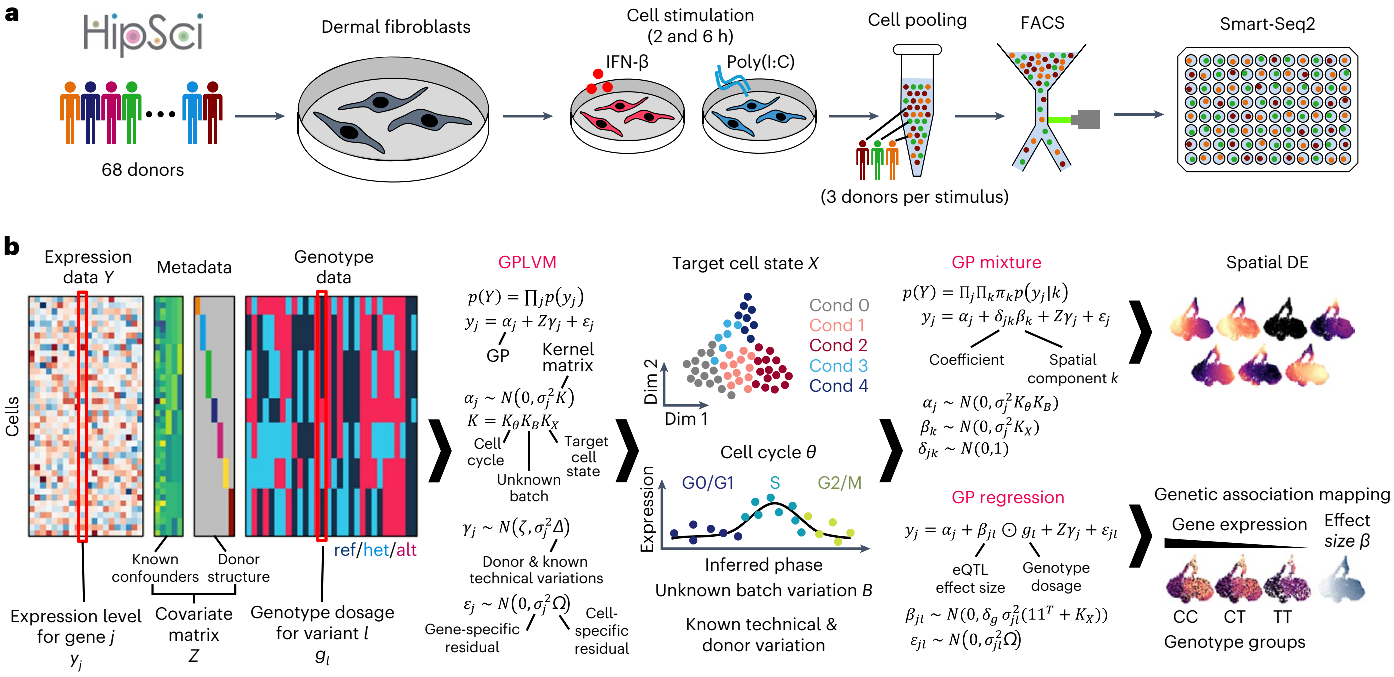

a
HipSci
68 donors
Dermal fibroblasts
Cell stimulation (2 and 6 h)
IFN-β
Poly(I:C)
Cell pooling
(3 donors per stimulus)
FACS
Smart-Seq2
b
Expression data Y
Metadata
Genotype data
Cells
ref/het/alt
Known confounders
Donor structure
Expression level for gene j y_j
Covariate matrix Z
Genotype dosage for variant l g_l
GPLVM
$p(Y) = \prod_j p(y_j)$
$y_j = \alpha_j + Z\gamma_j + \varepsilon_j$
GP
Kernel matrix
$\alpha_j \sim N(0, \sigma_j^2 K)$
$K = K_\theta K_B K_X$
Cell cycle
Unknown batch
Target cell state
$\gamma_j \sim N(\zeta, \sigma_j^2 \Delta)$
Donor & known technical variations
$\varepsilon_j \sim N(0, \sigma_j^2 \Omega)$
Gene-specific residual
Cell-specific residual
Target cell state X
Cond 0
Cond 1
Cond 2
Cond 3
Cond 4
Dim 2
Dim 1
Cell cycle θ
G0/G1
S
G2/M
Expression
Inferred phase
Unknown batch variation B
Known technical & donor variation
GP mixture
$p(Y) = \Pi_j \Pi_k \pi_k p(y_j|k)$
$y_j = \alpha_j + \delta_{jk}\beta_k + Z\gamma_j + \varepsilon_j$
Coefficient
Spatial component k
$\alpha_j \sim N(0, \sigma_j^2 K_\theta K_B)$
$\beta_k \sim N(0, \sigma_j^2 K_X)$
$\delta_{jk} \sim N(0,1)$
Spatial DE
GP regression
$y_j = \alpha_j + \beta_{jl} \odot g_l + Z\gamma_j + \varepsilon_{jl}$
eQTL effect size
Genotype dosage
$\beta_{jl} \sim N(0, \delta_g \sigma_{jl}^2 (11^T + K_X))$
$\varepsilon_{jl} \sim N(0, \sigma_{jl}^2 \Omega)$
Genetic association mapping
Gene expression
Effect size β
CC
CT
TT
Genotype groups

Fig. 1 | Schematics of experiment and statistical analysis. a, Experimental design using the in vitro fibroblast system. **b**, GASPACHO framework. Expression data and relevant metadata (known confounding factors) as well as donor (cell line) structure are used to construct a GPLVM to extract the target cell state, while dissecting cell cycle effect and other known and unknown technical variability including donor–donor variation. The result of the GPLVM is then utilized for the subsequent analyses of spatial DE analysis using a GP mixture model and the genetic association mapping using a GP regression model (Methods). Cond, experimental condition; Dim, dimension; ref/het/alt, reference homozygote/heterozygote/alternative homozygote.

To this end, we develop a statistical approach based on a Gaussian process (GP) latent variable model (GPLVM)[20,21] called GASPACHO (GAuSsian Processes for Association mapping leveraging Cell HeterOgeneity). This allows us to identify expression quantitative trait loci (eQTLs) that manifest at different stages of the response to stimuli.

We find more than a thousand eQTLs, hundreds of which are colocalized with known risk loci of diverse autoimmune and infectious diseases. We perform fine-mapping of the *OAS1* locus, associated with COVID-19, to reveal the imbalanced expression of *OAS1* and *OAS3* genes during the antiviral innate immune response. We further integrate these data with eQTLs from a COVID-19 patient cohort dataset of peripheral blood mononuclear cell (PBMC) scRNA-seq[22], as well as with scRNA-seq data of infected nasal epithelial cells from 33 patients with COVID-19 (ref. 23).

Overall, our study illustrates how coupling single-cell transcriptomics with a cutting-edge statistical approach can identify dynamic effects of human trait-associated genetic variants in different contexts of activation of antiviral innate immunity and, in general, in diverse cellular dynamic processes.

Results

Cell stimulation to study antiviral responses in fibroblasts

To study the innate immune expression program that is triggered upon viral infection, we exposed primary dermal fibroblasts from 68 donors from the Human Induced Pluripotent Stem Cell Initiative (HipSci)[24] to two stimulants: (1) Poly(I:C), a synthetic double-stranded RNA (dsRNA) that is rapidly recognized by viral sensors and elicits primary antiviral and inflammatory responses; and (2) interferon-β (IFN-β), a cytokine that upregulates a secondary wave of response in both infected and bystander cells, and shifts the cells into an antiviral mode, where hundreds of interferon-stimulated genes (ISGs) are upregulated to contain the infection.

We collected cells exposed to each of the two stimuli after 2 and 6 h of stimulation (Fig. 1a). Following this, scRNA-seq profiling was performed using a plate-based full-length transcript approach (Methods). After quality control, 22,188 high-quality cells were obtained across 128 plates, with each plate containing cells from three donors (Fig. 1a). The donor identity for each cell was inferred from scRNA-seq read data using known genotypes made available by HipSci (Extended Data Fig. 1a and Methods). Preliminary analysis showed that our data display high cell-to-cell variability in gene expression both within and across donors, as observed in previous studies by us and others[25–27]. In fact, our data were confounded by various technical and biological factors, including library preparation in different batches, and cell cycle effects (Extended Data Fig. 1b). The complex nature of these data, along with their confounders, motivated us to develop an approach that reveals the genetic and physiologically relevant variation, while computationally masking confounding factors.

Uncovering cell-state dynamics using GPs

Single-cell transcriptomics (as compared with bulk) enables us to uncover hidden states of complex biological processes, while also requiring regression of technical effects and biological variation that is not of interest (for example, proliferation). We developed GASPACHO, which utilizes a GPLVM to uncover the dynamic cell states of interest, while adjusting periodic cell cycle variation and both known and unknown technical variations (such as, donor and Smart-Seq2 plate variations) simultaneously (Fig. 1b and Methods). The use of a GPLVM allows us to capture smooth and continuous nonlinear trends in gene expression along the latent variables, for which other methods such as the standard linear principal component analysis will not work well.

Although there are other models that utilize a GPLVM to study single-cell dynamics[28,29], the unique aspect of our GPLVM approach is that it explicitly takes account of the donor variation as well as other known confounding effects (such as technical batches) as additional random effect terms (Fig. 1b and Methods). These confounders are known to inflate the type I error in downstream analyses, such as in differential expression[30] (DE), leading to false discovery of differentially

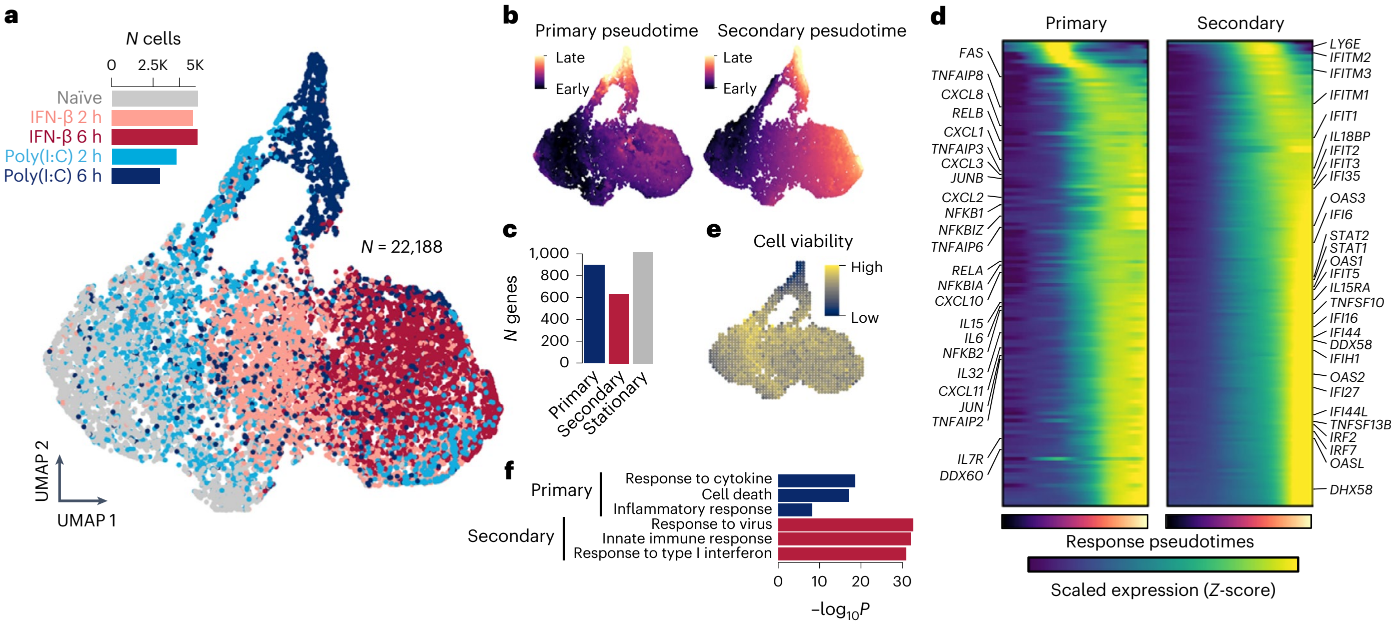


Fig. 2 | Innate immunity captured by GPLVM and GP mixture model. **a**, UMAP of latent variables capturing innate immune variation between cells. Cells are colored by the five experimental time points (gray, naïve state; pink, IFN-β 2 h; brown, IFN-β 6 h; blue, Poly(I:C) 2 h; navy, Poly(I:C) 6 h). **b**, Estimated pseudotime for primary and secondary responses using the GP mixture model. UMAP coordinates are identical to **a**. **c**, Barplot shows the numbers of response and stationary genes. **d**, Heatmaps show dynamic gene expression changes along primary or secondary response pseudotime. The pseudotime color scale corresponds to **b**. The expression color (navy to yellow) shows the magnitude of scaled expression for each gene (Z-score). **e**, UMAP shows a predicted Achilles cell viability using CEVIChE (Methods). UMAP coordinates are identical to **a**. **f**, Barplot shows the enrichment of gene ontology terms for primary and secondary genes. *P* values were computed using Fisher's one-tailed test with Bonferroni multiple testing correction implemented in gprofiler2 on R. K, thousand.

expressed genes. As detailed below, the model output not only enabled us to look at the architecture of the antiviral response in the cell-state space, but also provided a rigorous statistical framework of (1) spatial DE analysis and (2) genetic association mapping using genotype data obtained from the donor of origin for each cell.

Specifically, the gene expression variation in the target cell-state space was inferred by a GP mixture model in which an additional GP component is introduced into the model to capture hidden spatial DE patterns[31] of gene expression in the latent space (Fig. 1b and Methods). The genetic association mapping was also carried out by using a GP regression model in which the effect size of a quantitative trait locus (QTL) was modeled as a GP in the target cell-state space. Here, the additional GP was multiplied by the genotype dosage (the number of alternative alleles for each donor) to capture the gene–environment interaction[32] (Fig. 1b and Methods). Importantly, the eQTL effect is obtained at single-cell resolution, and the model does not require aggregation of single-cell data into pseudo-bulk data, which is a common eQTL mapping strategy. Thus, we can study the effect of genetic variants without losing the continuum of transcriptional dynamics and its spectrum across individual cells. We performed a comprehensive analysis to assess our method in terms of both sensitivity and specificity with another single-cell-based eQTL mapping approach, CellRegMap[33], using simulation-based datasets (section 3 of the Supplementary Notes). We have implemented the software in R, which is available from github (https://github.com/natsuhiko/GASPACHO).

Primary and secondary responses of innate immunity

We first applied the GPLVM to adjust for the cell cycle and unknown batch effects in our data (Extended Data Fig. 2a,b) and successfully extracted the innate immune state embedded in the data (Fig. 2a). We also confirmed that the extracted immune state was independent of cell cycle or the unknown batch variations (Extended Data Fig. 2c). We observed two major cell trajectories: one for response to IFN-β from the naïve state (*x* axis) and the other for response to Poly(I:C) (*y* axis).

We then applied the GP mixture model which revealed two independent innate immune responses, the primary response by virus infection and the secondary response for bystander cells due to IFN-β secretion by the infected cells or direct IFN-β stimulation (Fig. 2b and Extended Data Fig. 3a). Those responses were highly overlapping on the Uniform Manifold Approximation and Projection (UMAP), suggesting those two processes are independently and simultaneously happening in each cell. In total, the GP mixture model discovered 903 and 636 genes upregulated during the primary and secondary responses, respectively (hereafter referred to as primary response genes and secondary response genes), while 1,020 genes were expressed uniformly across all cells in different experimental conditions (referred to as stationary genes) (Fig. 2c and Extended Data Fig. 3b). Many cytokine and chemokine genes were upregulated along the primary response, while ISGs were upregulated along the secondary response (Fig. 2d). Interestingly, the primary response was also correlated with the predicted cell viability by CEVIChE (CEll VIability Calculator from gene Expression) (Fig. 2e and Methods). The gene ontology enrichment analysis for the primary and secondary genes clearly demonstrated that primary response genes are enriched for cell death and inflammatory response, while secondary response genes are enriched for type I interferon response (Fig. 2f and Extended Data Fig. 3c).

Dynamic genetic effect on innate immune response

We then mapped eQTLs along innate immune responses using the GP regression model to assess genetic association in single-cell resolution (Methods). We discovered 1,275 eQTL genes (local false discovery rate (FDR) 10%) among 10,748 genes expressed, at least, in 10% of total cells. We mapped eQTLs in the HLA region where 91 genes were tested, and 25 genes were eQTLs (local FDR 10%). Because the number of eQTLs was high for the modest sample size, we examined whether there was an inflation of test statistics for eQTL variants with lower minor allele frequencies (MAFs) and confirmed that there was no inflation in the Bayes factors nor in the number of discovered eQTLs in lower MAF bins

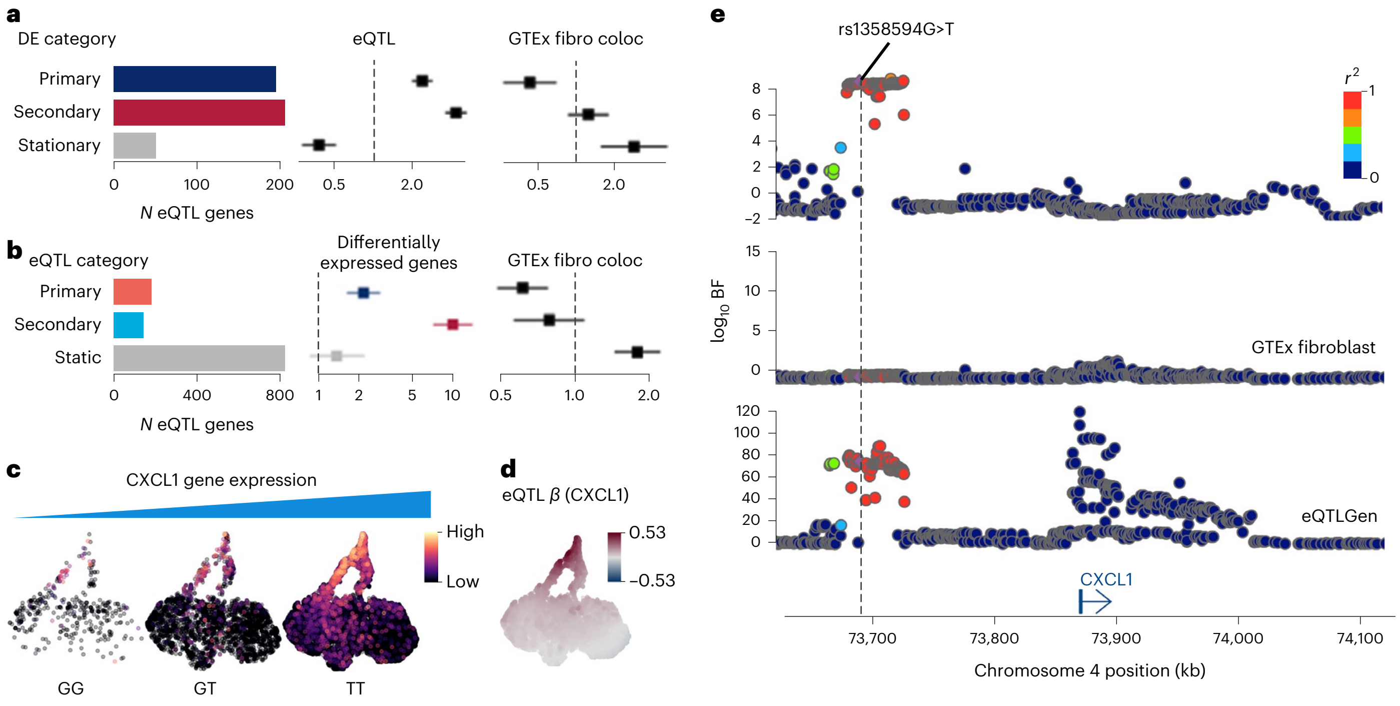


Fig. 3 | Characteristics of response eQTLs mapped using GP regression. **a**, Barplot showing the numbers of eQTLs (local FDR < 10%) that are primary and secondary response genes or stationary genes. Forest plots showing the enrichment of discovered eQTLs for the DE categories and GTEx fibroblast eQTL colocalizations. The error bars in the forest plots show 95% confidence intervals (standard errors) of odds ratios using $N = 10,748$ genes as independent samples (Methods). **b**, Barplot showing the numbers of static and dynamic eQTLs (primary and secondary response eQTLs). Forest plots showing the enrichment of response eQTLs for differentially expressed genes (primary, secondary response and static eQTLs for primary and secondary response genes, and for stationary genes, respectively) and the enrichment of response eQTLs of GTEx fibroblast eQTL colocalization. The error bars in the forest plots show 95% confidence intervals (standard errors) of odds ratios using $N = 10,748$ genes as independent samples (Methods). **c**, UMAPs showing *CXCL1* expression levels stratified by different genotype groups at rs1358594G>T. UMAP coordinates are identical to Fig. 2a. **d**, UMAP shows the distribution of eQTL effect size (β) at rs1358594. The alternative allele (T) is assessed. UMAP coordinates are identical to Fig. 2a. **e**, Locus zoom plot of *CXCL1* eQTL association Bayes factors around the *CXCL1* gene (top, in-house data; middle, GTEx fibroblast eQTL; bottom, eQTLGen blood eQTL). BF, Bayes factor; fibro coloc, colocalisation with GTEx fibroblast eQTLs.

(Extended Data Fig. 4a,b). We note that the eQTL genes discovered are strongly enriched in highly expressed genes (Extended Data Fig. 4c), but this is not the case for differentially expressed genes (for example, stationary genes were depleted in highly expressed genes; Extended Data Fig. 4d).

We found that 15% and 16% of our eQTL genes are primary and secondary response genes, respectively (Fig. 3a and Extended Data Fig. 4e). These genes are strongly enriched within the discovered eQTLs (Fig. 3b and Extended Data Fig. 4f). Because eQTL genes tend to be highly expressed (Extended Data Fig. 4c), we adjusted enrichments for the average expression level (Methods) and confirmed that the result is broadly the same (Extended Data Fig. 4f). We also found that primary response genes are depleted in colocalization between our eQTLs and fibroblast eQTLs from the Genotype-Tissue Expression (GTEx) Project, while the stationary genes are enriched (odds ratio of 2.8) (Fig. 3a, Extended Data Fig. 4g,h and Methods), suggesting our eQTLs are highly context-specific. We note that the odds ratio observed suggests a notable agreement between the two datasets (GTEx and our data), which is meaningful given the experimental and technical differences between the two studies which can strongly affect gene expression and subsequent eQTL discovery.

Next, we classified our eQTLs into static and dynamic eQTLs. We then further classified the dynamic eQTLs into distinct spatial patterns of their effect sizes utilizing a similar GP mixture model that segregated response genes (Methods). We discovered 830 (65%) (posterior probability > 0.9) of our eQTLs were static eQTLs whose genetic effect is ubiquitous in any cellular state (Fig. 3b and Extended Data Fig. 4l,j). Furthermore, 184 (14%) and 141 (11%) of our eQTLs were primary and secondary response eQTLs, respectively, whose patterns resembled the spatial DE patterns of primary and secondary response genes (Extended Data Fig. 4i). We also found that the primary/secondary response eQTLs were enriched with primary/secondary response genes, respectively, while the static eQTLs were enriched with other types of differentially expressed genes (Fig. 3b and Extended Data Fig. 4k). The fact that the static eQTLs are significantly enriched with GTEx fibroblast eQTLs, while the primary response eQTLs are depleted, further supports the notion that our eQTLs are specific to the innate immune response and not detectable in a naïve cellular state. We note here that, as was shown in our simulation study (section 3 of the Supplementary Notes), the GP mixture model is robust against the lower detection limit of gene expression quantified by the sequencing technology. Therefore, the misclassification rate in the real data analysis is likely to be minimal.

As an example, the *CXCL1* gene is a known primary innate immune response gene and a primary response eQTL (as expected from its known functions as an important chemokine in this response). It is mostly expressed in later time points of Poly(I:C)-stimulated cells (Fig. 3c), and its expression level is higher for the alternative allele T at rs1358594 compared with the reference allele G (Fig. 3d). This eQTL signal was discovered more than 100 kilobases (kb) downstream of the gene's transcription start site (TSS) and only detected upon cell stimulation by Poly(I:C), but not in the naïve condition, as also shown in the GTEx naïve fibroblast eQTL data (Fig. 3e), though the gene is highly expressed in the GTEx data (median transcripts per million (TPM) = 111). We note, however, that this eQTL was discovered in eQTLGen data with tens of thousands of blood samples (Fig. 3e). This might suggest that

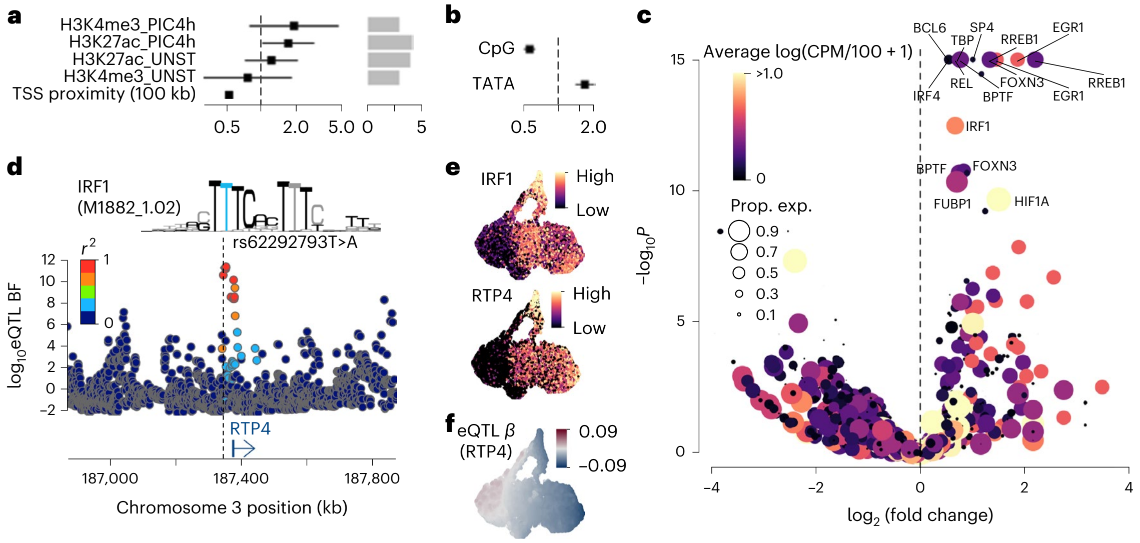


Fig. 4 | Fine-mapping eQTLs with epigenetic data. a, Enrichment eQTLs for regulatory regions characterized by ChIP–seq for the two histone modifications (H3K27ac and H3K4me3) under two different conditions (UNST, unstimulated; PIC4h, Poly (I:C) 4-h stimulation). The error bars in the forest plot show 95% confidence intervals (standard errors) of odds ratios using $N = 10{,}748$ genes as independent samples (Methods). **b**, eQTL enrichment for genes with TATA-box or CGI between TSS and 100 bp upstream. The error bars in the forest plots show 95% confidence intervals (standard errors) of odds ratios using $N = 10{,}748$ genes as independent samples (Methods). **c**, Enrichment of lead eQTL variants for various TF motifs. The color of each point shows the average expression across all cells and the point size shows the frequency of cells with CPM > 0 for each TF gene in our fibroblast data. **d**, Locus zoom plot showing the *RTP4* eQTL association Bayes factors. The lead eQTL variant rs62292793T>A disrupts the putative IRF1 binding motif (M1882_1.02; CIS-BP v.1.02) upstream of the *RTP4* gene. **e**, UMAPs showing expression levels of *IRF1* and *RTP4*. UMAP coordinates are identical to Fig. 2a. **f**, UMAP showing eQTL effect size of *RTP4*. UMAP coordinates are identical to Fig. 2a. CPM, counts per million; Prop. exp., proportion of cells the gene expressed.

the eQTL signal is present in unstimulated conditions in immune cells (such as in blood).

To compare our eQTLs with other datasets, we repeated the colocalization analysis between our stimulated fibroblast eQTLs and all primary GTEx tissues[34] as well as with immune cells (monocytes and induced pluiripotent stem cell (iPSC)-derived macrophages) stimulated using different conditions based on previous studies[7,9]. We found that the naïve GTEx fibroblast showed the highest prior probability of colocalization (Extended Data Fig. 5 and Methods).

Fine-mapping eQTLs with epigenetic data

We have also performed fine-mapping using epigenetic data (histone modification chromatin immunoprecipitation followed by sequencing (ChIP–seq) of active promoters and enhancers) originating in dsRNA-stimulation of human dermal fibroblasts[27] (Methods). We identified that more than 10% of the putative causal eQTL variants discovered by the hierarchical model (Methods) are located in the *cis*-regulatory regions characterized by the ChIP–seq data (Fig. 4a). Those variants are especially enriched for promoter peaks characterized by H3K4me3 antibody in Poly(I:C)-stimulated cells (Fig. 4a). Importantly, our eQTLs were also strongly enriched around the TSS, and the number of eQTLs was reduced by 46% every 100 kb further away from the TSS (Fig. 4a).

We next tested whether promoter architecture affects the variability between individuals. It was previously shown by us and others[27,35] that genes containing TATA boxes in their promoters tend to vary more in transcription between species and conditions and between individual cells responding to an immune stimulus, whereas promoters containing CpG islands (CGIs) tend to vary less and be transcriptionally more homogenous. We observe that genes with TATA-containing promoters are 1.4-times more highly enriched with eQTLs in comparison with genes with CGI-containing promoters (Fig. 4b).

Lastly, using the eQTL variants fine-mapped based on ChIP–seq annotations, we examined which transcription factor (TF) motifs were disrupted by the lead eQTL variants (Methods). We found interferon regulatory factors 1 and 4 (IRF1 and IRF4) as well as REL and ATF4 were significantly enriched (Fig. 4c). An example of putative TF binding disruption was discovered in *RTP4* eQTL (Fig. 4d), where the alternative allele of a promoter-flanking eQTL variant (rs62292793T>A) may disrupt an IRF1 motif that significantly reduces putative TF binding affinity, which subsequently downregulates the *RTP4* expression (Fig. 4e,f). Furthermore, the TATA motif (TBP) is also found to be disrupted by eQTL variants (Fig. 4c), further suggesting the importance of TATA regulation in modulating the response and its variability among individuals, as previously suggested in the mammalian immune response[27,35] and in other systems[36].

Colocalization with autoimmune and infectious disease genome-wide association studies

One of the purposes of eQTL mapping is to uncover the target genes and related cell states at each genetic locus implicated by genome-wide association studies (GWASs) of common complex traits. Here, we tested colocalization of our eQTLs with risk loci from 701 GWASs (each with five or more genome-wide significant loci), of which 112 were broadly immune-related, including autoimmune and chronic inflammatory diseases such as Crohn's disease and infectious diseases such as COVID-19 (Methods). We discovered 7,934 unique gene–trait combinations with the posterior probability of a single shared causal variant between an eQTL and a GWAS locus greater than 0.5. The combinations consisted of 643 different GWAS traits and 988 unique genes. We observed an excess of colocalized eQTLs for immune-related traits over nonimmune traits (Fig. 5a, $P = 2.8 \times 10^{-6}$, and Methods), likely reflecting the known involvement of innate immunity in each of the disease pathologies.

We discovered 48 primary and secondary response genes that were specifically colocalized with 51 autoimmune and infectious disease loci, some of which were colocalized with multiple traits (Fig. 5b). For example, we detected an eQTL for the *ETV7* gene, which encodes a TF in the ETS family and plays a key role in hematopoiesis[37]. The eQTL was colocalized with rheumatoid arthritis (posterior probability = 0.98)

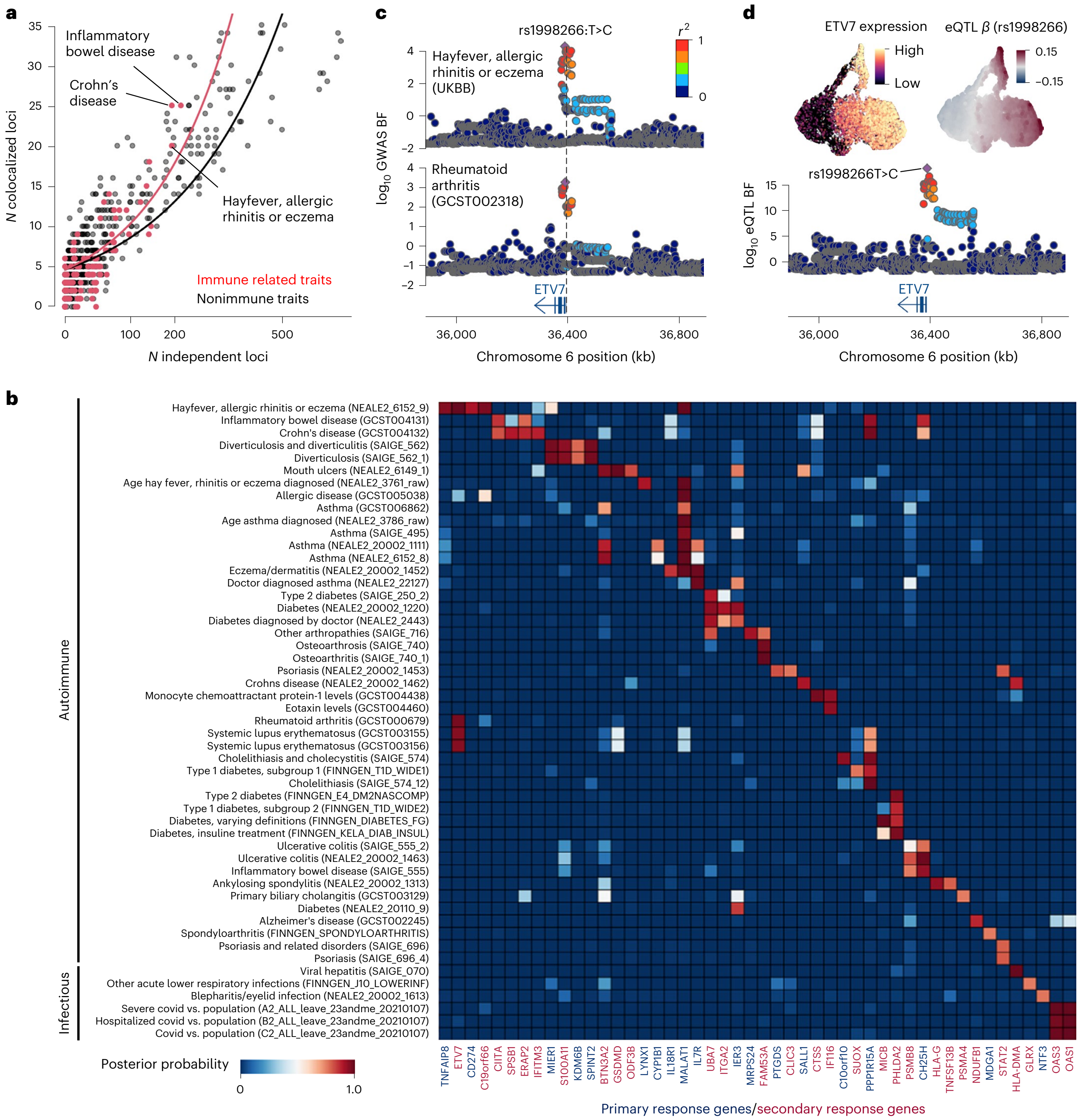


Fig. 5 | eQTL and disease GWAS colocalization. a, Scatter plot shows the number of colocalized eQTLs with posterior probability greater than 0.5 (*y* axis) against the number of independent loci where GWAS and eQTL do not share the putative causal variant (Methods). **b**, Heatmap shows the posterior probability of colocalization between eQTLs and GWAS loci. Only primary (colored by navy) and secondary (colored by red) response genes were shown. **c**, Locus zoom plots show the association of rheumatoid arthritis with hay fever, allergic rhinitis or eczema around the *ETV7* gene. Points are colored by the linkage disequilibrium index (r^2) with the GWAS index variant rs1008266T>C. **d**, UMAPs show the scaled *ETV7* expression and the eQTL effect size (*β*) at the lead eQTL variant rs1998266T>C. Locus zoom plot shows the eQTL association for *ETV7*. UMAP coordinates are identical to Fig. 2a. UKBB, UK BioBank.

and hay fever, allergic rhinitis or eczema (posterior probability = 0.99) (Fig. 5c). The gene is an ISG and the expression is upregulated during secondary response (Fig. 5d). The lead eQTL variant (rs1998266T>C) is shared with the GWAS traits, whose alternative allele C upregulates gene expression in stimulated conditions and also increases the risks of those GWAS traits (Fig. 5d). The alternative allele C also modifies the binding motif of the TF ATF6 putatively bound at the promoter region of *ETV7*, thereby potentially increasing the expression level (Extended Data Fig. 6a–c).

Lastly, we compared our colocalization results with those provided by Open Targets for 2,331 GWAS traits and 16 different eQTL studies (Data availability). We discovered that 6,327 gene–trait pairs

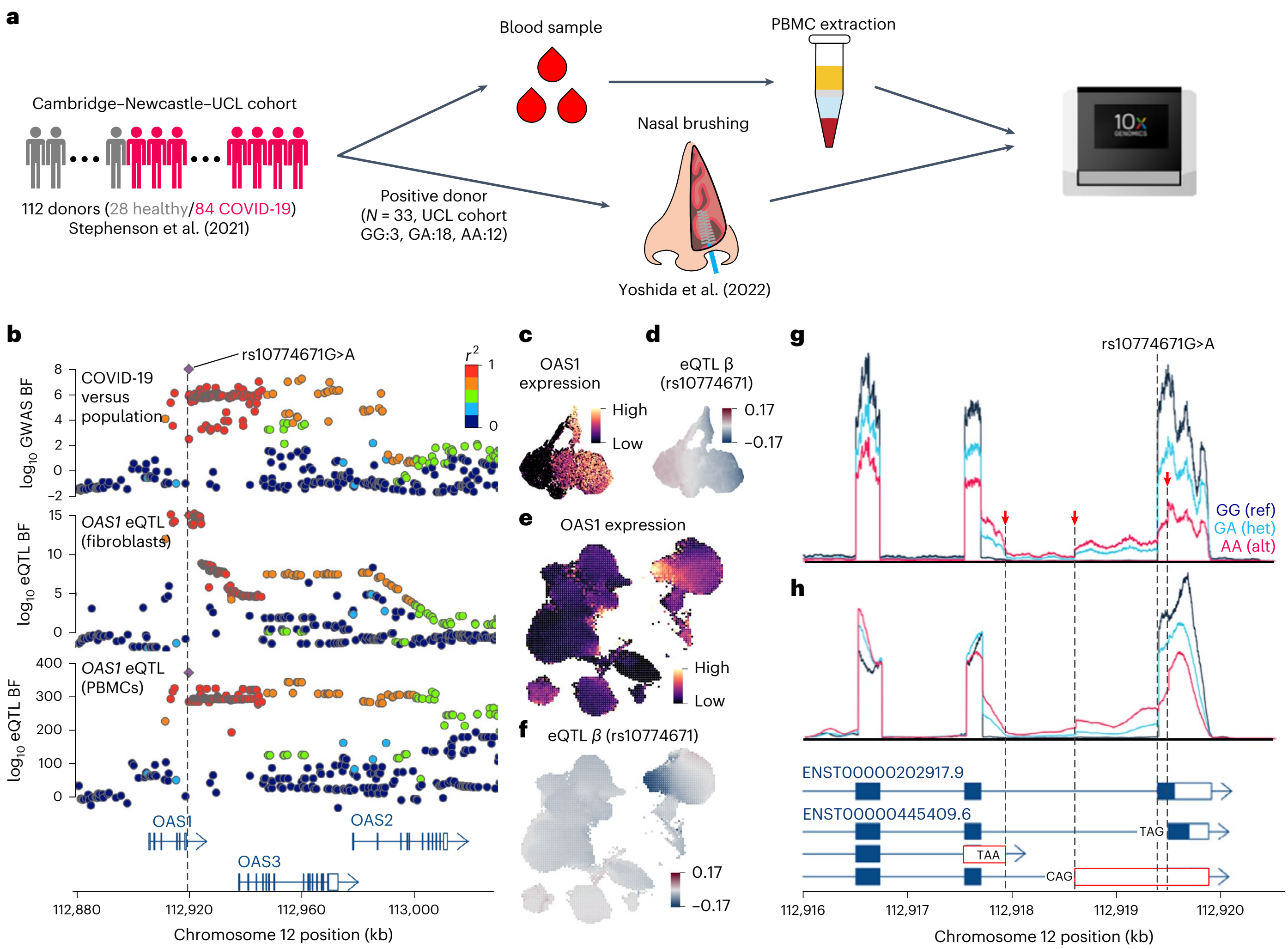


Fig. 6 | Fine-mapping *OAS1* locus using three different model systems. **a**, Schematic of in vivo system, COVID-19 study of PBMCs and nasal brushings to confirm the splicing QTL association in the *OAS1* gene. **b**, Locus zoom plots show the COVID-19 GWAS (COVID-19 versus population) association Bayes factors as well as eQTL associations of the *OAS1* gene in fibroblasts and PBMCs. **c**, UMAP shows the expression levels of the *OAS1* gene in fibroblasts. UMAP coordinates are identical to Fig. 2a. **d**, UMAP shows the eQTL effect size of the *OAS1* gene at rs10774671G>A. UMAP coordinates are identical to Fig. 2a. **e**, UMAP shows *OAS1* expression level in PBMCs. UMAP coordinates are identical to Extended Data Fig. 7a. **f**, UMAP shows the eQTL effect size of the *OAS1* gene at rs10774671G>A in PBMCs. UMAP coordinates are identical to Extended Data Fig. 7a. **g**, Sequencing coverage depth around the splicing variant rs10774671G>A, which creates three different isoforms, two of which are not annotated in Ensembl 90. scRNA-seq reads in fibroblasts were aggregated and stratified by the three different genotype groups (GG, reference homozygote; GA, heterozygote; AA, alternative homozygote). **h**, The 10x RNA-seq coverage depth of epithelial cells around *OAS1* 3′ end in nasal brushing samples taken from 33 COVID-19-positive adult patients was stratified and aggregated by the three genotype groups of rs10774671G>A as demonstrated in **g**. UCL, University College London.

were unique and only found in our colocalization results (posterior probability of colocalization > 0.5), whereas 1,607 gene–trait pairs are shared in both datasets (posterior probability of colocalization > 0.5), of which 615 also overlap with the GTEx fibroblast colocalization analysis (Extended Data Fig. 6d).

Fine-mapping *OAS1* eQTL associated with SARS-CoV-2 infection

In conjunction with the fibroblast system, we used two additional in vivo systems (Fig. 6a) to further fine-map the 12q24.13 (*OAS1*) locus that was reported in a GWAS of SARS-CoV-2 positive-infected individuals against population controls[38] (index SNP: rs10774671G>A). The locus is colocalized with the *OAS1* eQTL in fibroblasts with a posterior probability of 1.0 (Figs. 5b and 6b). *OAS1* is a secondary response gene and is highly expressed upon IFN-β (at 2 h and later) and Poly(I:C) stimulation (at 6 h) (Fig. 6c). The alternative allele A of rs10774671 downregulates the expression level (Fig. 6d).

We investigated our recently published PBMC scRNA-seq data[22] obtained from 112 donors, including 84 COVID-19-positive individuals, and profiled using the CITE-seq approach[39], as an independent in vivo validation of *OAS1* eQTL colocalization with GWAS locus for COVID-19 susceptibility (Methods). There are 18 major blood cell types annotated in this dataset (Extended Data Fig. 7a), of which myeloid cells and certain T cell subtypes show higher expression of the secondary response genes discovered in our fibroblast data (Extended Data Fig. 7b and Methods). As expected, *OAS1* is highly expressed as a secondary response gene in PBMCs (Fig. 6e). In addition, we confirmed that the *OAS1* gene is also a strong eQTL in PBMCs and colocalizes well with the COVID-19 GWAS locus with the posterior probability of 0.99 (Fig. 6b). The GWAS index variant rs10774671G>A is the lead eQTL variant in PBMCs whose alternative allele A is strongly negatively correlated with *OAS1* expression. This is especially clear in $CD16^+$ monocytes, among other immune cell types (Fig. 6f and Extended Data Fig. 7a).

The index SNP rs10774671 is known to be a splicing QTL[40] that disrupts the splicing motif right next to the last exon of the *OAS1* gene (Fig. 6g). In our fibroblast data, this variant also increased the intron expression between the last two exons and created three different isoforms (Fig. 6g), all of which are known to cause impaired OAS1 protein expression[40]. These alternative isoforms are also observed in nasal epithelial cells from 33 brushing samples from a recent COVID-19 cohort[23]. We confirmed that the alternative allele of the lead eQTL variant in stimulated fibroblasts is associated with lower expression of the *OAS1* gene in the epithelial cells and the occurrence of alternative isoforms, as expected (Fig. 6h). This suggests that the in vitro stimulation fibroblast system is an appropriate model to study physiologically relevant eQTLs in the context of infection, such as COVID-19.

In addition to the *OAS1* eQTL, *OAS3* eQTL in fibroblasts was also colocalized with the COVID-19 GWAS locus (posterior probability = 0.99) (Fig. 5b and Extended Data Fig. 7c). Because *OAS1* and *OAS3* are both ISGs, the expression patterns of *OAS1* and *OAS3* along the innate immune response trajectory are expected to be very similar, as indeed observed (Fig. 6c and Extended Data Fig. 7d). However, the direction of the eQTL effect was opposite for the two genes: *OAS1* gene expression is downregulated by the alternative allele of the COVID-19 GWAS index SNP rs10774671G>A, whereas the expression level of *OAS3* gene is upregulated by the alternative allele (Fig. 6d and Extended Data Fig. 7e,f).

Discussion

In this work we developed GASPACHO, a statistical framework that fulfils two tasks: firstly, to infer the trajectory of gene expression over a dynamic process, and secondly, to model nonlinear dynamic genetic effects in every individual cell. Using GASPACHO, we integrated scRNA-seq data from fibroblasts from 68 donors triggered by innate immune stimuli and obtained a low-dimensional gene expression space representing the response dynamics across stimulated cells. This approach provides us with a unique map of interindividual transcriptional variation at single-cell resolution, which was often linked with noncoding regulatory regions (such as TF binding sites), previously profiled during fibroblast stimulation[27]. This approach discovered 1,275 eQTL loci, of which 988 were colocalized with one or more GWAS loci of autoimmune and infectious diseases, including COVID-19 at the *OAS1* locus. We also found 1,607 colocalizations shared with the Open Targets colocalizations database, of which 615 also overlap with the GTEx fibroblast colocalization analysis. We note here that some of the colocalizations in the Open Targets database are likely to be missed because Open Targets has only tested colocalizations against genome-wide significant loci in GWAS traits ($P < 5 \times 10^{-8}$), while we tested colocalizations for all GWAS loci (regardless of their P values) overlapping with 1-megabase (Mb) *cis*-windows of genes tested for eQTL mapping.

Previous studies used scRNA-seq profiling of tissues composed of several cell types from dozens to hundreds of donors, and performed genetic association mapping of cell-type-specific expression by using a pseudo-bulk approach[13–19]. GASPACHO, as well as cellRegMap[33], allows mapping of dynamic genetic effects of gene expression in individual cells. These two approaches are particularly suitable when considering a continuous genetic effect along cellular states rather than several discrete cell populations or states. GASPACHO and CellRegMap both incorporate context-specific donor (donor by context interaction) effects to adjust for dynamic genetic effects, for a better statistical calibration (because dynamic cellular states, such as immune responses, often vary between donors due to environmental and trans genetic effects). However, CellRegMap and GASPACHO differ in their model assumptions on genetic effects: CellRegMap assumes linearity on context-specific genetic (genotype by context interaction) effects, while GASPACHO assumes that those effects are nonlinear, suggesting GASPACHO is more flexible, yet computationally intensive (section 3 of the Supplementary Notes). Therefore, further studies are required to implement faster GP regression in modern computational environments, such as GPU. Lastly, the GPLVM implemented in GASPACHO is currently applicable for rapid analysis of dozens of thousands of cells. An accelerated version of GPLVM will be needed in the future for scaling, and a cutting-edge Bayesian inference technique, such as the stochastic variational inference implemented in GPy, should be able to achieve this goal.

The innate immune response is a genetic program that is elicited by most cells invaded by pathogens; however, the response varies between infected cells in terms of magnitude, the specific set of regulated genes and their cellular fate. This variability is observed both between cells originating from different lineages and between individual cells from the same homogenous cellular population[25–27]. Furthermore, genetic variation has been shown to significantly modulate the innate immune response in many previous studies and its dynamics are thought to be nonlinear[4,6–9,35]. GASPACHO is thus particularly useful to study how genetic effects are associated with different stages and cellular trajectories during this response, as demonstrated, for example, in our analysis of primary and secondary response genes. Furthermore, we also observe eQTLs that appear only during stimulation, as previously suggested using bulk RNA-seq[4,6–9,35], but without the need to partition samples into discretized conditions.

Our in vitro immune stimulation of dermal fibroblasts can be used as a model system to study genetic effects in innate immune responses in primary cells. Using this analysis, we detected 6,327 colocalizations between eQTLs and various autoimmune and infectious disease GWAS loci. While fibroblasts are not the primary cellular target of SARS-CoV-2 infection (which mostly targets epithelial cells), we detected a colocalization between *OAS1* eQTL and COVID-19 GWAS locus, which we then also found in PBMCs from patients with COVID-19. This colocalization was also previously found using bulk RNA-seq[41]. Our findings suggest an association between a particular risk variant (rs10774671) and COVID-19 infection and severity, and that this risk allele may generate alternative isoforms of the *OAS1* gene in nonclassical monocytes in peripheral blood. We further found that these alternative isoforms are expressed in nasal epithelial cells from a set of patients with COVID-19 carrying the alternative allele. Since the alternative allele is also a risk allele in COVID-19 GWAS, this implies that these *OAS1* RNA splicing isoforms may be associated with impaired OAS1 protein expression and viral clearance in host cells, as previously suggested in other viral diseases[40,42]. Interestingly, we also observe a colocalization in this locus between *OAS3* eQTL and COVID-19 GWAS locus; however, in this case, the alternative allele of rs10774671 is linked to an increase in the *OAS3* gene expression level. Further studies are needed to mechanistically determine the impact of OAS3 expression on SARS-CoV-2 infection.

In summary, our study demonstrates how coupling single-cell transcriptomics with a statistical approach can identify dynamic nonlinear effects of genetic variants across cellular contexts.

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/s41588-023-01421-y.

References

1. Kumar, H., Kawai, T. & Akira, S. Pathogen recognition by the innate immune system. *Int. Rev. Immunol.* https://doi.org/10.3109/08830185.2010.529976 (2011).
2. Fumagalli, M. et al. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* https://doi.org/10.1371/journal.pgen.1002355 (2011).

3. Rice, G. I. et al. Gain-of-function mutations in IFIH1 cause a spectrum of human disease phenotypes associated with upregulated type I interferon signaling. *Nat. Genet.* **46**, 503–509 (2014).
4. Brinkworth, J. F. & Barreiro, L. B. The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. *Curr. Opin. Immunol.* **31**, 66–78 (2014).
5. Gorman, J. A. et al. The A946T variant of the RNA sensor IFIH1 mediates an interferon program that limits viral infection but increases the risk for autoimmunity. *Nat. Immunol.* **18**, 744–752 (2017).
6. Quintana-Murci, L. Human immunology through the lens of evolutionary genetics. *Cell* **177**, 184–199 (2019).
7. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
8. Lee, M. N. et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
9. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
10. Ye, C. J. et al. Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. *Genome Res.* **28**, 1812–1825 (2018).
11. Rotival, M. et al. Population variation in miRNAs and isomiRs and their impact on human immunity to infection. *Genome Biol.* **21**, 187 (2020).
12. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
13. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0089-9 (2018).
14. Cuomo, A. S. E. et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).
15. Jerber, J. et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021).
16. Cuomo, A. S. E., Alvari, G., Azodi, C. B., McCarthy, D. J. & Bonder, M. J. Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biol.* **22**, 1–30 (2021).
17. Randolph, H. E. et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science* **374**, 1127–1133 (2021).
18. Yazar, S. et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
19. Perez, R. K. et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
20. Lawrence, N. D. in *Advances in Neural Information Processing Systems 16* (eds Thrun, S. et al.) 329–336 (MIT Press, 2004).
21. Titsias, M. & Lawrence, N. D. Bayesian Gaussian process latent variable model. *Proc. Mach. Learn. Res.*, **9**, 844–851 (2010).
22. Stephenson, E. et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* https://doi.org/10.1038/s41591-021-01329-2 (2021).
23. Yoshida, M. et al. Local and systemic responses to SARS-CoV-2 infection in children and adults. *Nature* **602**, 321–327 (2022).
24. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
25. Shalek, A. K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).
26. Russell, A. B., Trapnell, C. & Bloom, J. D. Extreme heterogeneity of influenza virus infection in single cells. *eLife* **7**, e32303 (2018).
27. Hagai, T. et al. Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197–202 (2018).
28. Reid, J. E. & Wernisch, L. Pseudotime estimation: deconfounding single cell time series. *Bioinformatics* **32**, 2973–2980 (2016).
29. Ahmed, S., Rattray, M. & Boukouvalas, A. GrandPrix: scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics* **35**, 47–54 (2019).
30. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
31. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
32. Moore, R. et al. A linear mixed-model approach to study multivariate gene-environment interactions. *Nat. Genet.* **51**, 180–186 (2019).
33. Cuomo, A. S. E. et al. CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. *Mol. Syst. Biol.* **18**, e10663 (2022).
34. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
35. Schroder, K. et al. Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proc. Natl Acad. Sci. USA* **109**, E944–E953 (2012).
36. Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J. & Hartl, D. L. Genetic properties influencing the evolvability of gene expression. *Science* **317**, 118–121 (2007).
37. Carella, C. et al. The ETS factor TEL2 is a hematopoietic oncoprotein. *Blood* **107**, 1124–1132 (2006).
38. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* https://doi.org/10.1038/s41586-021-03767-x (2021).
39. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
40. Li, H. et al. Identification of a Sjögren's syndrome susceptibility locus at OAS1 that influences isoform switching, protein expression, and responsiveness to type I interferons. *PLoS Genet.* **13**, e1006820 (2017).
41. Baranova, A., Cao, H. & Zhang, F. Unraveling risk genes of COVID-19 by multi-omics integrative analyses. *Front. Med.* **8**, 738687 (2021).
42. Lim, J. K. et al. Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. *PLoS Pathog.* **5**, e1000321 (2009).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [2]Medical Support Center of Japan Environment and Children's Study (JECS), National Center for Child Health and Development, Tokyo, Japan. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. [4]Open Targets, Wellcome Genome Campus, Hinxton, UK. [5]Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge, UK. [6]Department of Medicine, University of Cambridge, Cambridge, UK. [7]Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [8]UCL Respiratory, Division of Medicine, University College London, London, UK. [9]University College London Hospitals NHS Foundation Trust, London, UK. [10]Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK. [11]NIHR Newcastle Biomedical Research Centre, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [12]Department of Dermatology, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [13]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [14]Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany. [15]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [16]Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [17]Theory of Condensed Matter Group, Cavendish Laboratory/Department of Physics, University of Cambridge, Cambridge, UK. [18]These authors contributed equally: Tzachi Hagai, Sarah A. Teichmann. ✉e-mail: tzachiha@tauex.tau.ac.il; st9@sanger.ac.uk

Methods

Ethical compliance

This project was approved by the Wellcome Sanger Institute Animal Welfare and Ethical Review Body and complied with all relevant ethical regulations regarding animal research and human studies. Human cells were obtained from HipSci[24], where they were collected from volunteers recruited from the National Institute for Health and Care Research (NIHR) Cambridge BioResource (written consent was given). Human skin profiling was performed in accordance with protocols approved by the Newcastle Research Ethics Committee (REC approval 08/H0906/95+5). Patients with a confirmed diagnosis of COVID-19 were recruited from Addenbrooke's and Royal Papworth hospitals under ethical approval obtained from the East of England Cambridge Central Research Ethics Committee (NIHR BioResource, REC no. 17/EE/0025). Informed consent was obtained for all participants.

Dermal fibroblast cell culture and stimulation

Primary dermal fibroblast cells from HipSci were used (http://www.hipsci.org/). The cells were derived from healthy individuals spanning a range of ages (from 30 to 79 and 57.2 on average) and both sexes (40 female and 28 male). Following a similar protocol used in our previous work[27], cells were cultured in DMEM (high glucose, pyruvate, Life Technologies), with 10% FBS, GlutaMAX and 1% penicillin-streptomycin. In each experimental batch, we cultured in parallel cells from three different individuals. Cells were split the day before the experiment into separate wells and on the day of experiment were stimulated with either dsRNA (0.5 μg ml^{-1} high-molecular-weight rhodamine-labeled Poly(I:C) (Invivogen, tlrl-pic), transfected with 1 μl ml^{-1} lipofectamine 2000 (Thermo Fisher, 11668027), for 2 or 6 h) or 1,000 U ml^{-1} human recombinant IFN-β (11410-2, PBL), for 2 or 6 h, or left untreated. In this manner, for each individual, we obtained five separate conditions.

After the relevant period of time, cells were detached by trypsinization and resuspended in PBS. Samples from the three individuals with the same treatment were then mixed (for example, 'unstimulated' cells from the three donors would be pooled together). The primary aim of this mixing step was to reduce downstream experimental variability between the three donors, while simultaneously streamlining the collection stage. In this manner, we obtained plates for each of the five conditions, with each having a mixture of all three individuals.

Sorting and single-cell library preparation

Cells were sorted on a Becton Dickinson Influx into 96-well plates containing 2 μl per well of lysis buffer, as described in the Smart-Seq2 protocol[43], or in our previous work[27]. Importantly, each 96-well plate contained cells from the same condition of all three individuals used for each experimental batch. Single cells were sorted individually (using FSC-W versus FSC-H), and apoptotic cells were excluded using DAPI. Rhodamine-positive cells were selected in the Poly(I:C) treatments. Cells from each three-plex cell pool were sorted across four plates. Reverse transcription and complementary DNA amplification were performed according to the Smart-Seq2 protocol (Picelli et al., 2014), and library preparation was performed using an Illumina Nextera kit. Samples were sequenced using paired-end 75-bp reads on an Illumina HiSeq 2500 machine. For library preparation, cells were loaded into 384-well plates. We note that cells in one of the four stimulated conditions were assigned in a 384-well plate in conjunction with cells in naïve condition (rows C, H and M).

Smart-Seq2 data preprocessing and quality control

All sequence data were aligned to human genome assembly GRCh38 using STAR (v.2.5.3a; https://github.com/alexdobin/STAR/releases) and ENSEMBL human gene assembly 90 as the reference transcriptome. We performed adapter trimming of Tn5 transposon and PCR primer sequences using skewer (v.0.1.127; https://github.com/relipmoc/skewer) before alignment. Following alignment, we used featureCounts (v.1.5.3; http://subread.sourceforge.net/) to count fragments for each annotated gene. In total, we observed 58,394 cells, of which 22,188 cells passed the quality control criteria: the minimum number of sequenced fragments (>10,000 autosomal fragments), the minimum number of expressed genes (>500 autosomal genes), mitochondrial fragment percentage (<20%) and the library complexity (percentage of autosomal fragment counts for the top 100 highly expressed genes <30%). We also performed demuxlet[44] (v.0.1.0; https://github.com/statgen/demuxlet) to identify the genetic origin of each cell as well as to remove doublets using the genotype data from HipSci.

Genotype data

We obtained the SNP genotype data from HipSci[24] (Data availability). We also genotyped 112 COVID-19 PBMC samples using the Affymetrix Axiom UK Biobank array (Data availability). We converted the genome coordinates from hg19 to GRCh38 using CrossMap (v.0.5.2; http://crossmap.sourceforge.net/). We then performed the whole-genome imputation using Beagle (v.5.1; https://faculty.washington.edu/browning/beagle/beagle.html) with the reference panel from the 1000 Genomes Project (Data availability).

Cell viability prediction

The cell viability was predicted by the web-based tool CEVIChE (https://saezlab.shinyapps.io/ceviche/). Because the tool is designed for bulk RNA-seq data, we aggregated gene expression levels for neighboring cells based on the UMAP in Fig. 2a. We constructed 30 × 30 equispaced grids and took geometric means of logCPM (log of counts per million) values within each grid.

GPLVM

The GASPACHO framework incorporated a GPLVM as a core model to estimate the latent variables and model parameters subsequently used in the spatial DE analysis and eQTL mapping. We assumed that the gene expression vector $y_j = (y_{ij}; i = 1, \dots, N)^T$ for the gene j across N cells is independently drawn from

$$y_j \sim N(\alpha_j + Z\gamma_j, \sigma_j^2 \Omega)$$

$$\alpha_j \sim N(0, \sigma_j^2 K_\theta K_B K_X)$$

$$\gamma_j \sim N(\zeta, \sigma_j^2 \Delta)$$

where α_j is a baseline GP governed by three different kernel matrices, periodic kernel matrix K_θ for the cell cycle state (θ) and two other squared exponential kernel matrices K_B and K_X for unknown batch effects (B) and the target cell state (X), respectively. Here, Z is a design matrix for the known covariates, such as donor and sequencing plates (Fig. 1b), and γ_j is a random effect to adjust the known confounding effects whose mean and variance were defined by ζ and the diagonal matrix Δ shared across all genes $j = 1, \dots, J$. The residual expression was determined by the gene-specific residual variance σ_j^2 and the cell-specific residual variance $\Omega = \mathrm{diag}(\omega_i; i = 1, \dots, N)$. The variance of the GP and random effect for gene j was properly scaled by the gene-specific residual variance σ_j^2.

The model parameters $\{\Delta, \Omega, \Sigma, \zeta\}$ and the latent variables $\{\theta, B, X\}$ were inferred by maximizing the marginal likelihood

$$L(\theta, B, X, \Delta, \Omega, \Sigma, \zeta) = \prod_{j=1}^{J} \int p(y_j|\alpha_j, \gamma_j) p(\alpha_j) p(\gamma_j) \mathrm{d}\alpha_j \mathrm{d}\gamma_j,$$

where $\Sigma = \mathrm{diag}(\sigma_j^2; j = 1, \dots, J)$. We used the L-BFGS algorithm with the analytic gradient of the likelihood function with respect to the parameters and the latent variables. In reality, the kernel matrices are not tractable for large N; we computed the Titsias bound using the sparse GP[21] to approximate the above likelihood (see section 1.3 of the Supplementary Notes for more details).

GP mixture model for gene classification

We employed a GP mixture model to perform the DE analysis in the target cell-state space defined by X which was estimated by the GPLVM. Specifically, we introduced one extra GP β_k for the kth differentially expressed gene group ($k = 1,\ldots,K$) to which a gene j belongs:

$$y_j \sim N(\alpha_j + \delta_{jk}\beta_k + Z\gamma_j, {\sigma_j}^2 \Omega)$$

$$\alpha_j \sim N(0, {\sigma_j}^2 K_\theta K_B)$$

$$\beta_k \sim N(0, {\sigma_j}^2 K_X)$$

$$\gamma_j \sim N(\zeta, {\sigma_j}^2 \Delta)$$

$$\delta_{jk} \sim N(0, 1)$$

Here, the effect size of the GP was properly scaled by a coefficient δ_{jk} to allow the GP to be both positively and negatively correlated with the gene expression. The model parameters $\{\Delta, \Omega, \Sigma, \zeta\}$ and the latent variables $\{\theta, B, X\}$ were replaced by the estimators of the GPLVM. Then, we maximized the likelihood of a finite mixture of GPs:

$$L(\pi_1, \ldots, \pi_K, \beta_1, \ldots, \beta_K) = \prod_{j=1}^{J} \sum_{k=1}^{K} \int \pi_k p(y_j|\alpha_j, \beta_k, \gamma_j, \delta_{jk}) p(\alpha_j) p(\beta_j) p(\gamma_j) p(\delta_{jk}) \mathrm{d}\alpha_j \mathrm{d}\gamma_j \mathrm{d}\delta_{jk}$$

with respect to π_k and β_k for $k = 1, \ldots, K$. Note that the number of total mixture components K is fixed in the current implementation and $K = 3$ was used in the fibroblast data. We used the sparse approximation to make the likelihood tractable (see section 1.4 of the Supplementary Notes for more details). Note that this model can be readily extended to classify dynamic eQTL effect sizes into finite spatial patterns (see section 1.4.1 of the Supplementary Notes).

For the pseudotime analysis, we computed the posterior mean $E[\beta_k|y_1,\ldots,y_J]$ for the kth component, which provided the underlying cellular states regarding the primary and secondary innate immune responses.

GP regression for association mapping

We employed a GP regression model to map eQTLs in the target cell-state space defined by X which was estimated by the GPLVM. Specifically, we introduced one extra GP β_{jl} for the gene j multiplied by the lth genetic variant $g_l = (g_{l1},\ldots,g_{lN})^T$ whose ith element g_{li} is alternative allele dosages for the individual i as a gene–environment interaction:

$$y_j \sim N(\alpha_j + \beta_{jl} \odot g_l + Z\gamma_j, {\sigma_{jl}}^2 \Omega)$$

$$\alpha_j \sim N(0, {\sigma_{jl}}^2 K_\theta K_B K_X)$$

$$\beta_{jl} \sim N(0, {\delta_g}^2 {\sigma_{jl}}^2 (11^T + K_X))$$

$$\gamma_j \sim N(\zeta, {\sigma_{jl}}^2 \Delta)$$

Here the eQTL effect size was properly scaled by a coefficient δ_g to allow for controlling of the genetic contribution on the expression level. The model parameters $\{\Delta, \Omega, \Sigma, \zeta\}$ and the latent variables $\{\theta, B, X\}$ were replaced by the estimated values obtained by the GPLVM. The Bayes factor of genetic association can be obtained by:

$$\mathrm{BF} = \frac{\int p(y_j|\alpha_j, \beta_{jl}, \gamma_j) p(\alpha_j) p(\beta_{jl}) p(\gamma_j) \mathrm{d}\alpha_j \mathrm{d}\beta_{jl} \mathrm{d}\gamma_j}{\int p(y_j|\alpha_j, \beta_{jl} = 0, \gamma_j) p(\alpha_j) p(\gamma_j) \mathrm{d}\alpha_j \mathrm{d}\gamma_j}$$

where we set $\delta_g = 0.1$ (see section 1.5 of the Supplementary Notes for more details).

As is implemented in CellRegMap, our model can also be extended to take the context-specific donor (context-by-donor interaction) effect into account. Here, the gene expression model can be written as:

$$y_j \sim N\left(\alpha_j + \beta_{jl} \odot g_l + Z\gamma_j + \sum_{i=1}^{N_\mathrm{d}} f_{ij} \odot z_i, {\sigma_j}^2 \Omega\right)$$

$$f_{ij} \sim N(0, {\delta_\mathrm{dxc}}^2 {\sigma_j}^2 K_X);\ i = 1, \ldots, N_d,$$

where f_{ij} denotes an additional GP for the individual i, z_i denotes the indicator vector to specify which cells belong to the individual i and N_d denotes the number of donors in the data. The additional variance parameter δ_dxc for the context-by-donor interaction effect is estimated under the null model using all genes (see section 1.5.1 of the Supplementary Notes for more details). All the real data analyses using the fibroblast data in this manuscript were based on the Bayes factors with this context-specific donor effect.

The eQTL effect size was estimated using the posterior distribution $p(\beta_{jl}|y_j) \propto p(y_j|\beta_{jl})p(\beta_{jl})$ and the posterior mean $E[\beta_{jl}|y_j]$ was computed for each variant l and used for the visualization on a UMAP (see section 1.5.2 of the Supplementary Notes for more details).

Hierarchical model for eQTL mapping and enrichment analysis

We tested genetic variants whose MAF is greater than 0.05 in a 1-Mb *cis*-regulatory window centered at each gene TSS. To control the FDR in a Bayesian framework, we used the hierarchical model[45] to obtain the posterior probability that a gene is an eQTL as well as the posterior probability that a variant is an eQTL variant within the *cis* window. The model allows incorporating various genomic annotations in the gene-level and variant-level as demonstrated previously[45]. We used the ChIP–seq peak annotations obtained by Hagai et al.[27] in conjunction with TSS proximity to estimate the contribution of epigenetic information to the eQTL variant discovery (see section 2.1 of the Supplementary Notes for more details). Note that we only consider genes expressed in at least 10% of the cells, resulting in a tested dataset of 10,748 genes. We did not introduce the gene-level prior probability to weight highly expressed genes for the eQTL discovery.

eQTL enrichment in differentially expressed genes and other annotations

The enrichment analysis was carried out based on the posterior probability Z_j that the gene j is an eQTL obtained from the hierarchical model. We then computed a 2 × 2 table using a corresponding binary annotation X_j (if the gene j belongs to some annotation, for example, a TATA-box, then $X_j = 1$, and otherwise $X_j = 0$) or alternatively the posterior probability $X_j \in [0,1]$ that the gene j is a differentially expressed gene (one of multiple differentially expressed gene categories defined above), such that

$$T_{kl} = \sum_{j=1}^{J} (1 - X_j)^{(1-k)} (1 - Z_j)^{(1-l)} {X_j}^k {Z_j}^l$$

for $k, l = 0, 1$. From the 2-by-2 table T, we computed the log odds ratio $r = \log(T_{00}T_{11}/(T_{01}T_{10}))$ and its standard error $\mathrm{Var}(r) = (1/T_{00} + 1/T_{01} + 1/T_{10} + 1/T_{11})$ to perform hypothesis testing. The confidence interval of the log odds ratio was given by $r \pm 1.96\sqrt{\mathrm{Var}(r)}$. We also computed the P value from the X^2 statistic $\chi^2 = r^2/\mathrm{Var}(r)$.

If the occurrence of eQTLs and an annotation X were confounded by a factor C (such as expression level for a gene), we split genes into 100 quantile bins according to the confounding factor C to compute the log odds ratio and its standard error for each bin as demonstrated above, and then we combined them using the inverse variance method to derive the meta statistic for an adjusted enrichment statistic.

eQTL sharing with GTEx tissues

We used the pairwise hierarchical model[45,46] to jointly map eQTLs in two different cell types (a similar approach to the pairwise fGWAS model[47]). We employed the association Bayes factor at each variant for each gene to compute the regional Bayes factors in a *cis* region of 1 Mb centered at the TSS under the following five different hypotheses:

H_0: a gene is not an eQTL in cell/tissue types 1 and 2.

H_1: a gene is an eQTL in cell/tissue type 1, but not in cell/tissue type 2.

H_2: a gene is an eQTL in cell/tissue type 2, but not in cell/tissue type 1.

H_3: a gene is an eQTL in cell/tissue types 1 and 2 with two independent putative causal variants.

H_4: a gene is an eQTL in cell/tissue types 1 and 2 with the shared putative causal variant.

Those regional Bayes factors were used in a hierarchical model to estimate prior probabilities that eQTLs are shared between two cell types (see section 2.2 of the Supplementary Notes for more details).

The hypothesis testing of eQTL enrichment in different DE categories, which are also colocalized with other eQTLs (such as GTEx fibroblasts), was performed by computing the pseudo counts:

$$T_{kl} = \sum_{j=1}^{J} (1 - X_j)^{(1-k)} \left(Z_j^{(1)} + Z_j^{(3)}\right)^{(1-l)} X_j^{k} \left(Z_j^{(4)}\right)^{l}$$

for $k, l = 0, 1$, where X_j denotes the posterior probability that the gene *j* is differentially expressed, $Z_j^{(1)}$ denotes the probability that the gene *j* is an eQTL in our data and not in a GTEx tissue, $Z_j^{(3)}$ denotes the probability that the gene *j* is an eQTL in our data and a GTEx tissue and not sharing the putative causal variant, and $Z_j^{(4)}$ denotes the probability that the gene *j* is an eQTL both in our data and a GTEx tissue and colocalized. The odds ratio and its standard error were computed as described in the section 'eQTL enrichment in differentially expressed genes and other annotations'.

Annotating TATA and CpG genes

To look for TATA-box motifs in gene promoters, we used TATA-motifs from CIS-BP (Data availability). We used the CpG annotation (Data availability) from the UCSC Genome browser to search for genes whose promoters overlap with a CGI. In both cases, we used the region 100 bp upstream from the TSS as the promoter region and referred to these genes as TATA genes and CpG genes, respectively.

eQTL variant enrichment at TF motifs

The hierarchical model provided the posterior probability that each variant *l* in the *cis*-regulatory region for the gene *j* is the eQTL Z_{jl}, so that $\sum_{l=1}^{L_j} Z_{jl} = 1$ where L_j is the number of variants in the *cis* window. We first selected the lead eQTL variant according to the posterior probability for each gene *j*. We then used the position weight matrices of TF motifs in CIS-BP (Data availability) to call motifs overlapping with lead eQTL variants as described elsewhere[45].

To perform the hypothesis testing that a TF motif is significantly overlapping with eQTL variants, we set $Z_j = \max_{l=1,\ldots,L_j} \{Z_{jl}\}$ and X_j to be the binary variable whose value is $X_j = 1$ if the lead eQTL variant *l* is overlapping with a TF motif; otherwise, $X_j = 0$. We then computed the 2 × 2 table to perform the enrichment analysis as described in the section 'eQTL enrichment in differentially expressed genes and other annotations'.

GWAS summary statistics

GWAS summary statistics were obtained from Open Targets which collected and harmonized summary statistics from the GWAS Catalog, FinnGen and UK Biobank (in total, 4,744 traits) (Data availability). We also downloaded summary statistics of four different COVID-19-related traits for all samples excluding 23andMe ('Very severe respiratory confirmed covid vs. population', 'Hospitalized covid vs. not hospitalized covid', 'Hospitalized covid vs. population' and 'Covid vs. population' in release 5) from the COVID-19 Host Genetics Initiative (Data availability). We selected 701 GWAS traits out of 4,748 traits with the criterion of five or more genome-wide significant loci, of which 112 were broadly immune-related, including autoimmune and chronic inflammatory diseases as well as infectious diseases.

Colocalization with GWAS traits

We used the same pairwise hierarchical model as in the section 'eQTL sharing with GTEx tissues' to perform the GWAS colocalization analysis, where the prior probabilities of the pairwise hierarchical model were fixed as $\{\Pi_1, \Pi_2, \Psi_{12}\} = \{0.2, 0.05, 0.01\}$ so that we can compare different studies with different statistical power to detect GWAS associations due to varying sample sizes. Here, Π_1 denotes the prior probability that a gene is an eQTL, Π_2 denotes that a genomic region for the corresponding gene (a 1-Mb window centered at TSS) is a significant GWAS locus and Ψ_{12} is a prior probability that the eQTL and the GWAS locus are colocalized. To fit the model, we converted the effect sizes and standard errors of each GWAS trait into Bayes factors using Wakefield's approximation[48]. See section 2.3 of the Supplementary Notes for more details.

PBMC data analysis and eQTL mapping

We used human PBMC scRNA-seq data[22] from 112 donors, including 84 COVID-19-positive individuals, profiled with the CITE-seq approach from 10x Genomics. We reduced the full GASPACHO approach to accommodate the PBMC single-cell data of over 700,000 cells in a reasonable time scale. The kernel functions used in the model were restricted to the linear kernel without the cyclic kernel for the cell cycle effect. The latent factors were estimated with the covariates of the number of genes expressed, the number of mapped reads, the sequencing center, sex, age, COVID-19 status, COVID-19 severity, patient ID and the first three genotype principal components. The latent factors were then used to define the two GPs

$$\alpha_j \sim N(0, \sigma_j^2 XX^T)$$

$$\beta_{jl} \sim N(0, \delta_g^2 \sigma_j^2 (11^T + XX^T))$$

for the intercept and the eQTL effect size of variant *l* for gene *j*.

OAS1 locus analysis using COVID-19 nasal brushing samples

To fine-map the *OAS1* locus in cells in vivo infected with COVID-19, we used human single-cell data of 33 nasal brushing samples from patients with COVID-19 from a recent work[23], profiled using CITE-seq. We used the aligned bam files to quantify allele-specific expression at rs10774671 using RASQUAL allele-specific expression caller[49]. The genotypes were assigned by fitting a binomial distribution on the allele-specific expression with probability parameters $p = \{0.01, 0.5, 0.99\}$ for reference homozygote, heterozygote and alternative homozygote, respectively. We inferred that there are three reference homozygotes, 18 heterozygotes and 12 alternative homozygotes at rs10774671 in this dataset.

Reporting summary

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

Data availability

All Smart-Seq2 cram files of our fibroblast data are available from the European Nucleotide Archive (Accession ID: PRJEB20147). The genotype data of fibroblast samples are under managed access and available through the HipSci portal (https://www.hipsci.org/data). The lines used in this study have the identifiers: HPSI0114pf-eipl, HPSI0114pf-fikt, HPSI0114pf-joxm, HPSI0114pf-lexy, HPSI0114pf-rozh, HPSI0114pf-vabj, HPSI0114pf-vass, HPSI0114pf-zoxy, HPSI0115pf-gifk, HPSI0115pf-melw, HPSI0115pf-zihe, HPSI0214pf-feec, HPSI0214pf-heja, HPSI0214pf-pelm, HPSI0215pf-deyz, HPSI0215pf-fawm, HPSI0215pf-hipn, HPSI0215pf-oilg, HPSI0314pf-bubh, HPSI0314pf-cuhk, HPSI0314pf-qonc, HPSI0314pf-wigw, HPSI0314pf-xugn, HPSI0414pf-ceik, HPSI0414pf-gesg, HPSI0414pf-naju, HPSI0414pf-oaqd, HPSI0514pf-fiaj, HPSI0514pf-kuco, HPSI0514pf-puie, HPSI0514pf-rutc, HPSI0514pf-sohd, HPSI0514pf-vuna, HPSI0614pf-ciwj, HPSI0614pf-miaj, HPSI0614pf-oicx, HPSI0714pf-pipw, HPSI0913pf-diku, HPSI0913pf-eika, HPSI0913pf-lise, HPSI0914pf-euts, HPSI0914pf-kajh, HPSI0914pf-laey, HPSI1013pf-garx, HPSI1013pf-jogf, HPSI1013pf-pamv, HPSI1013pf-sebz, HPSI1013pf-wopl, HPSI1013pf-wuye, HPSI1014pf-qayj, HPSI1014pf-sehl, HPSI1014pf-tixi, HPSI1014pf-toss, HPSI1014pf-tuju, HPSI1014pf-vils, HPSI1113pf-bima, HPSI1113pf-dons, HPSI1113pf-eofe, HPSI1113pf-ieki, HPSI1113pf-oaaz, HPSI1113pf-qolg, HPSI1113pf-wahn, HPSI1113pf-wetu, HPSI1114pf-ualf, HPSI1213pf-hehd, HPSI1213pf-nusw, HPSI1213pf-tolg and HPSI1213pf-xuja. The genotype data of COVID-19 PBMC samples around OAS1 gene are available at Zenodo[50]. The genome-wide genotype data are available upon request under the managed access of the NIHR BioResource's Data Access Committee. The annotation of the CpG site was downloaded from the UCSC website (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/cpgIslandExt.txt.gz). The position weight matrices (PWMs) of transcription factor motifs were obtained from CIS-BP (http://cisbp.ccbr.utoronto.ca/bulk.php). The PWMs used to find TATA-box in the gene promotor have the following identifiers: M1641_1.02, M2191_1.02, M4011_1.02, M4266_1.02, M6502_1.02, M1642_1.02, M4010_1.02, M4014_1.02 and M4708_1.02. Open Targets GWAS summary statistics are available from the GWAS Catalog (https://www.ebi.ac.uk/gwas/), FINNGEN (https://www.finngen.fi/en/access_results) and UK Biobank (https://www.nealelab.is/uk-biobank). COVID-19 GWAS summary statistics (release 5) are available from the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/results/r5/). The Open Targets colocalization data are obtained from the website (https://ftp.ebi.ac.uk/pub/databases/opentargets/genetics/210608/). The eQTL summary statistics of GTEx 48 tissues as well as immune cells (iPSC-derived macrophages and monocytes) under different stimulation conditions were obtained from the eQTL catalog (http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/). The 1000 Genomes Project VCF data (version: shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased) were obtained from: http://hgdownload.soe.ucsc.edu/gbdb/hg38/1000Genomes/. The summary statistics of fibroblast eQTLs are available from Zenodo (https://doi.org/10.5281/zenodo.7680146).

Code availability

GASPACHO (v.1.0.0), example data and script code to reproduce the results are available at Zenodo[51].

References

43. Picelli, S. et al. Full-Length RNA-Seq from single cells using smart-seq2. *Nat. Proto.* **9**, 171–181 (2014).
44. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
45. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
46. Young, A. M. H. et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet.* **53**, 861–868 (2021).
47. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
48. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
49. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).
50. Kumasaka, N. COVID-19 PBMC sample information and the VCF file of variants around OAS1 gene. *Zenodo* https://doi.org/10.5281/zenodo.7866287 (2023).
51. Kumasaka, N. natsuhiko/GASPACHO-v1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.7866223 (2023).

Acknowledgements

N.K., R.R., K.B.M., T.H. and S.A.T. were supported by Wellcome Sanger core funding (grant no. WT206194) and the Human Induced Pluripotent Stem Cell Initiative. T.H. was supported by the Israel Science Foundation (grant no. 435/20) and by the Chan Zuckerberg Initiative (Single-Cell Analysis of Inflammation, grant no. DAF2020-217532). R.R. was supported by the BBSRC Doctoral Training Programme. F.J.C.-N. and B.G. were supported by Wellcome (grant no. 206328/Z/17/Z) and MRC (grant no. MR/S036113/1). P.A.L. was supported by the Evelyn Trust (grant no. 20/75) and the UKRI/NIHR through the UK Coronavirus Immunology Consortium. K.B.W. acknowledges funding from University College London, Birkbeck MRC Doctoral Training Programme. M.Z.N. acknowledges funding from an MRC Clinician Scientist Fellowship (grant no. MR/W00111X/1) and the Rutherford Fund Fellowship allocated by the MRC and from the UK Regenerative Medicine Platform 2 (grant no. MR/5005579/1). M.Z.N. and K.B.M. have been funded by the Rosetrees Trust (grant no. M944) and from Action Medical Research (grant no. GN2911). We thank M. D. Morgan for careful reading of the manuscript and data sharing.

Author contributions

R.R., O.S., T.H. and S.A.T. designed the experiments of fibroblast stimulation and single-cell sequencing. R.R., S.P., R.B., C.G., S.N.B. and T.H. performed the experiments. N.K. and S.A.T. developed the analytical approach. N.K., R.R., N.H. and T.H. analyzed the data. N.I.P. provided the HipSci genotype data. J.S. and M.G. provided the Open Targets GWAS summary statistics. K.B.M., P.A.L., F.J.C.-N., B.G., J.L.B., K.B.W., M.Y., M.Z.N., E.S. and M.H. performed genotyping of the PBMC samples. G.R., M.H., and J.M. provided the processed single-cell RNA-seq data of the PBMC samples. M.Y., K.B.W., M.Z.N. and K.B.M. provided the data from the COVID-19 nasal brushing sample. N.K., K.P. and N.H. analyzed the COVID-19 nasal brushing data. N.K., T.H. and S.A.T. wrote the manuscript.

Competing interests

S.A.T. has received remunerations for consulting and Scientific Advisory Board work from TransitionBio, GlaxoSmithKline, Qiagen and Foresite Labs over the past 3 years and is an equity holder of TransitionBio. All other authors declare no competing interests.

Additional information

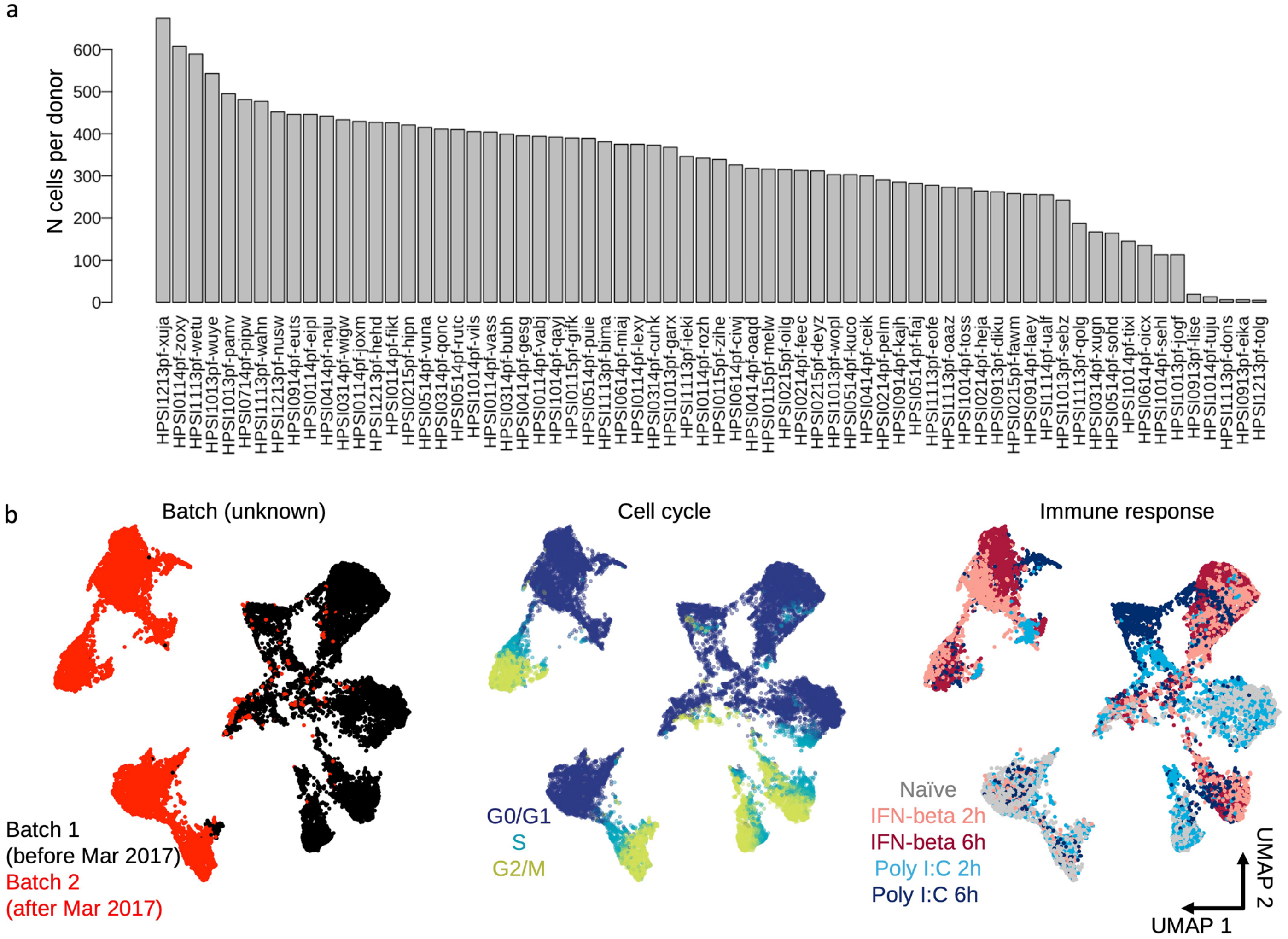
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01421-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01421-y.

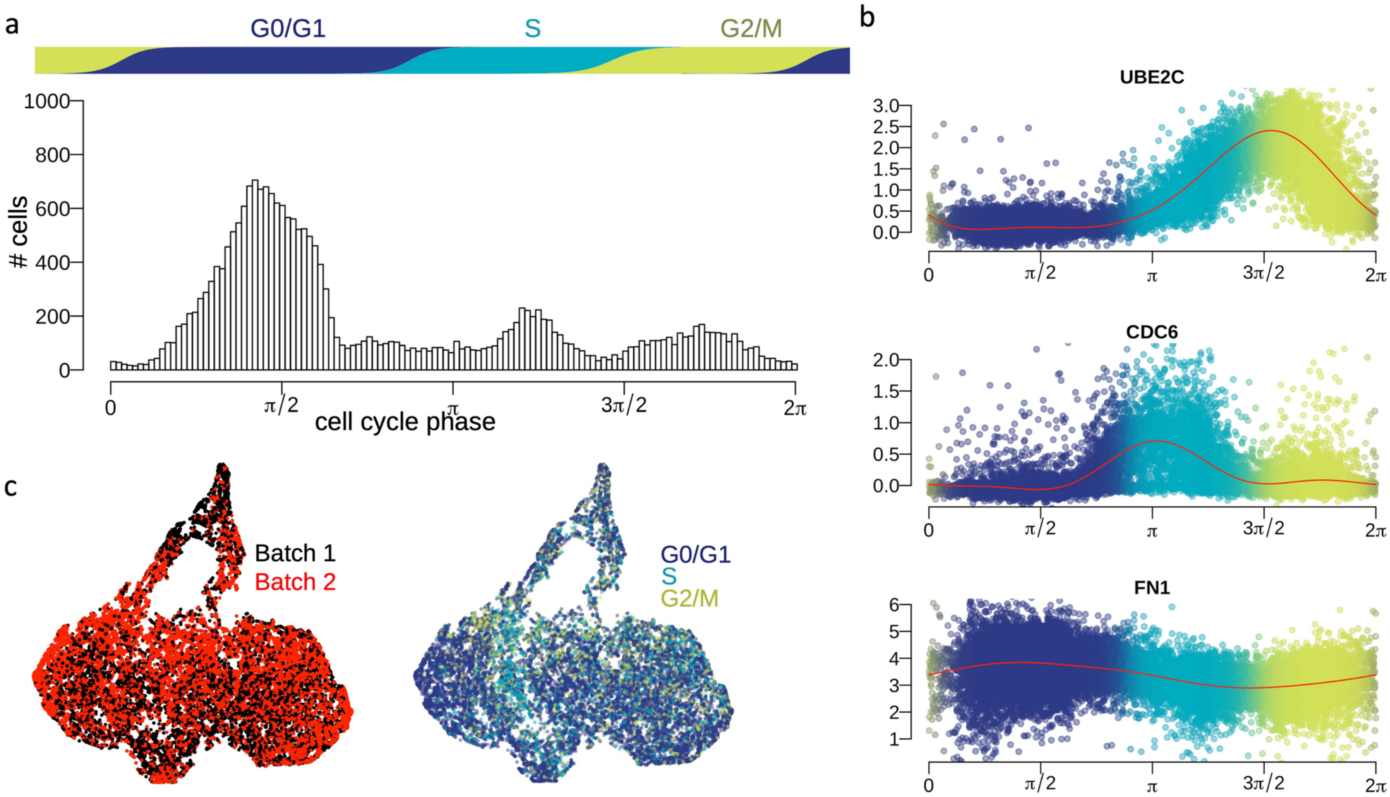
Correspondence and requests for materials should be addressed to Tzachi Hagai or Sarah A. Teichmann.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

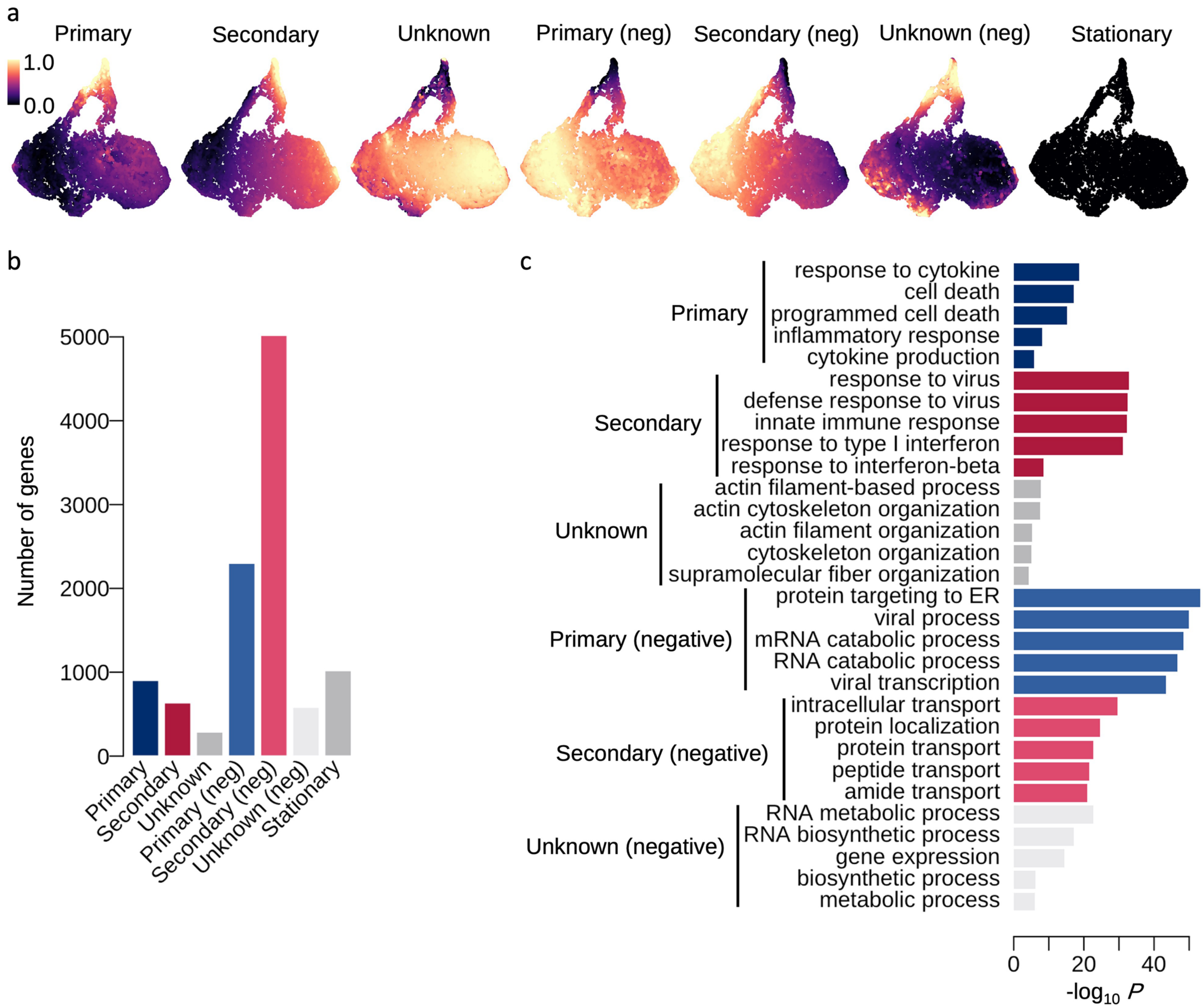
Reprints and permissions information is available at www.nature.com/reprints.



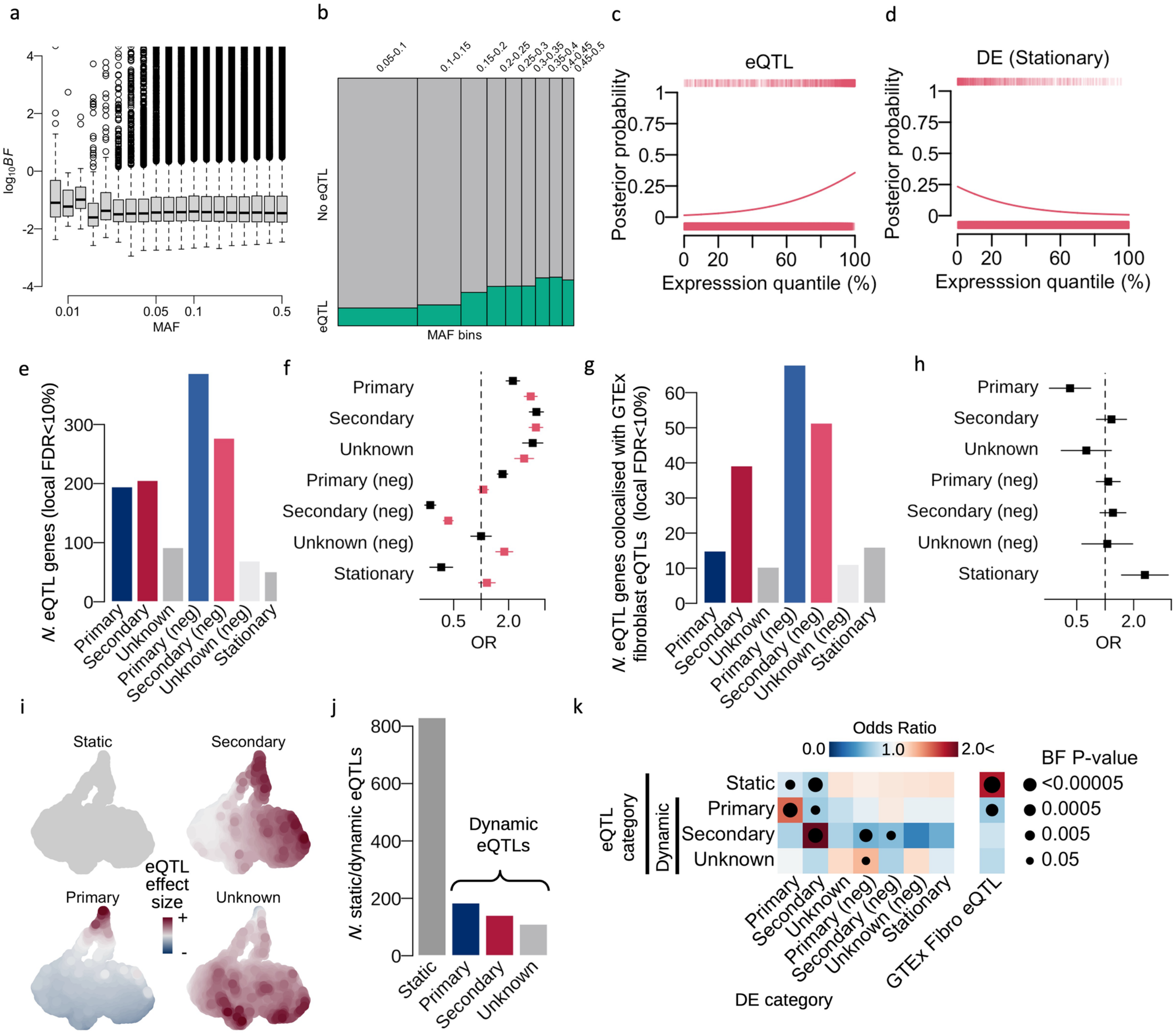
Extended Data Fig. 1 | Data and quality controls. a. Barplot shows the number of cells for each donor (cell line). **b**. UMAPs calculated from the first 10 principal components from the data. Points are coloured by unknown batch effect (well correlated with experimental date), cell cycle phase estimated from known marker genes (Online Methods) and experimental conditions.



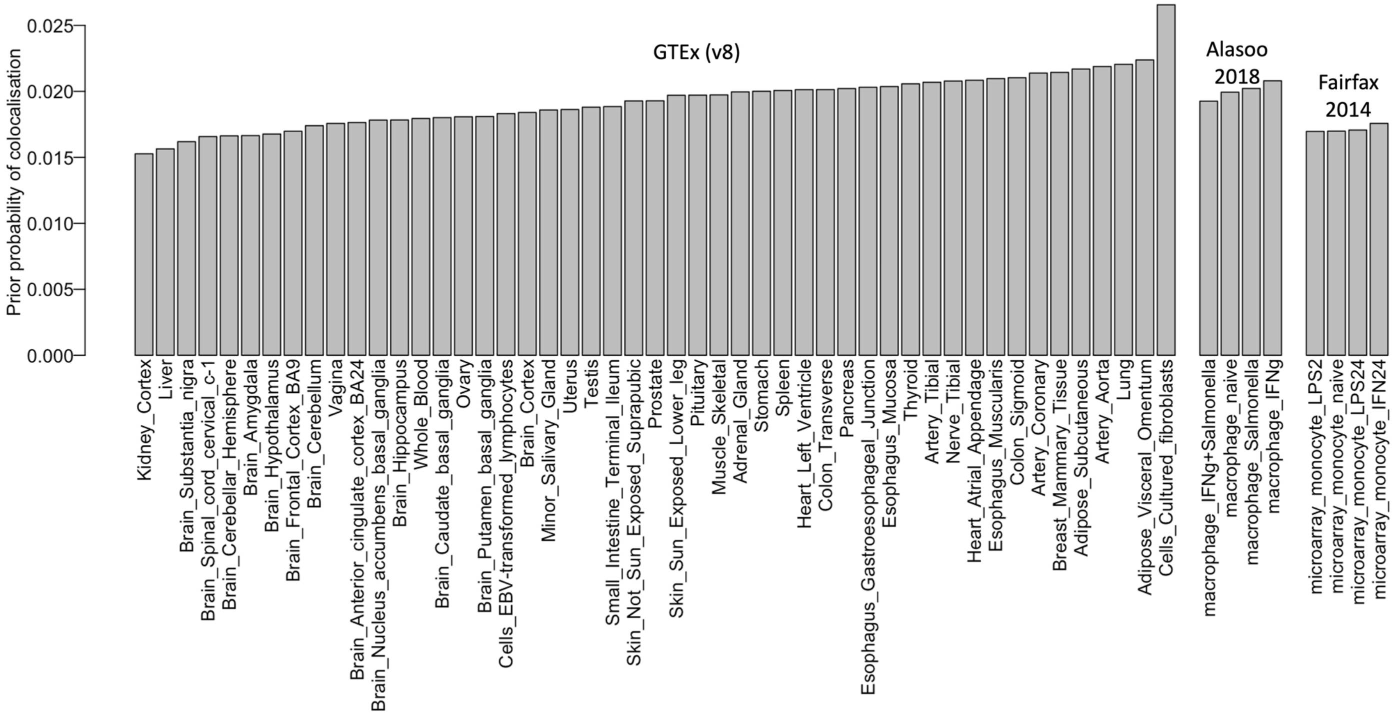
Extended Data Fig. 2 | Observed cellular states in the fibroblast single cell data. a. Histogram shows the distribution of estimated cell cycle phase by GPLVM (Online Methods). **b**. Scatterplots show scaled expression of known cell cycle genes (UBE2C and CDC6) and a gene highly expressed in G0/G1 phase (FN1). The red curves show the posterior mean estimates of expression levels by GPLVM (Online Methods). **c**. UMAPs of the target cell states coloured by unknown batch or cell cycle phase. UMAP coordinates are identical to Fig. 2a.



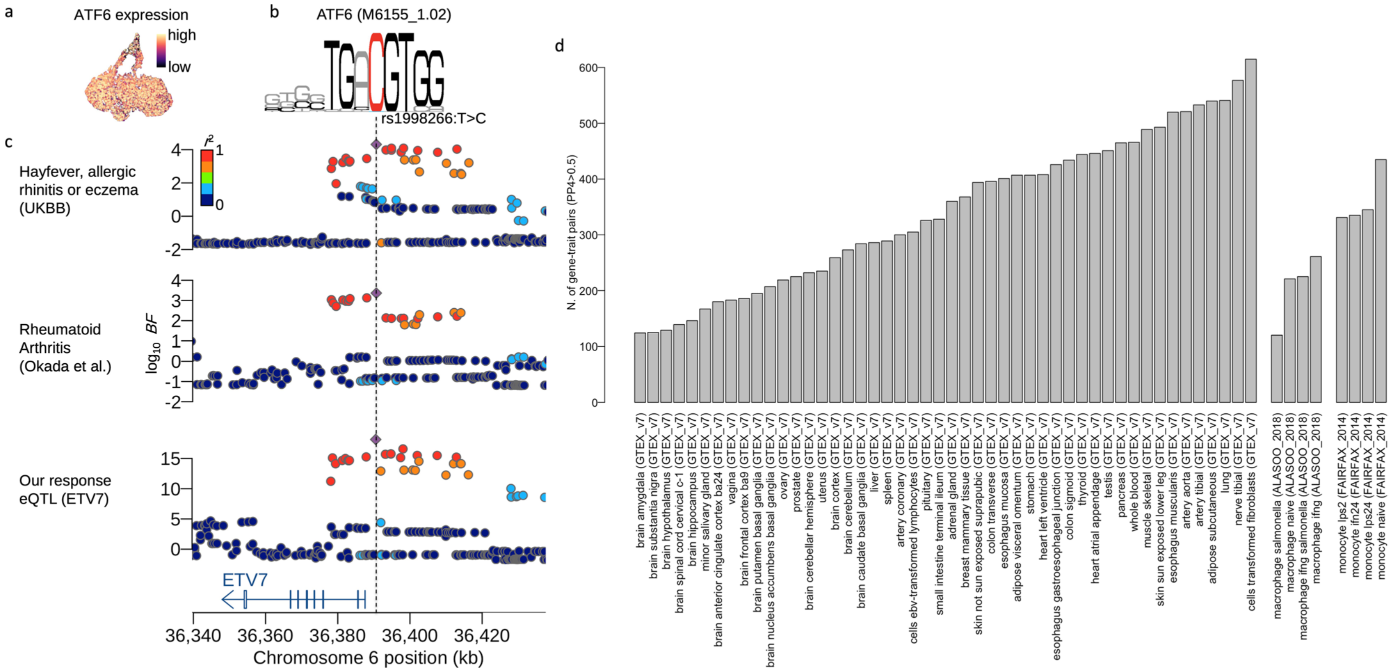
Extended Data Fig. 3 | Spatial differential expression analysis to classify genes into different innate immune responses. d. UMAPs of the seven different mixture components estimated from the GP mixture model (Online Methods). The values are quantile-normalised as in [0,1], and also treated as pseudotime for the first two components (primary and secondary). UMAP coordinates are identical to Fig. 2a. **e.** The number of genes categorised in each of the seven components. **f.** Top GO terms enriched with the genes detected in each of the seven DE categories.



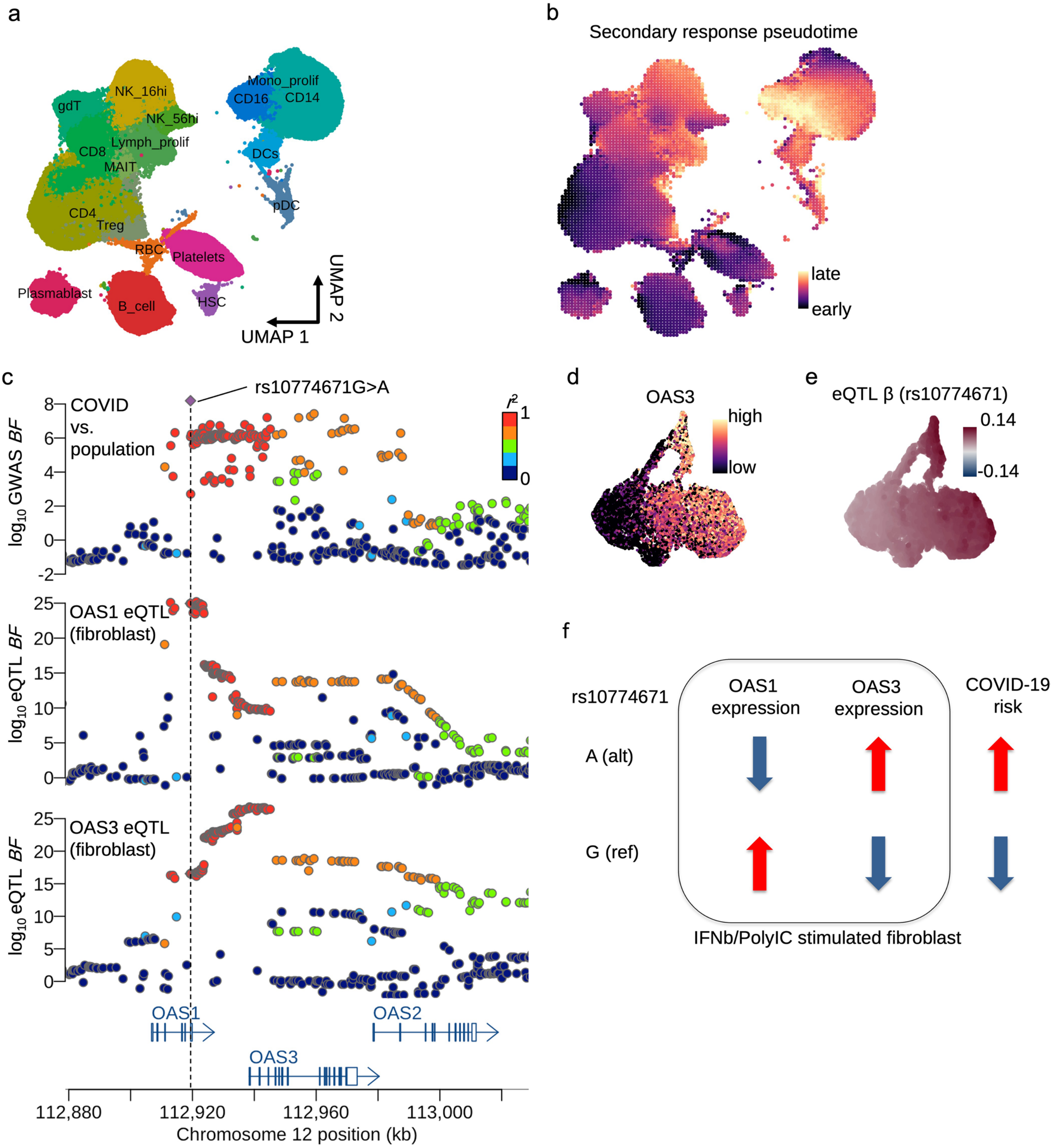
Extended Data Fig. 4 | Characteristics of mapped response eQTLs. a. Distributions of Bayes factors from our fibroblast data for $N = 1$ million random gene-variant pairs. Here we used all variants with MAF > 0% for the Bayes factor calculation (not only variants with MAF > 5%). Here The bottom and top of each box (Q1 and Q3) are the 25th and 75th percentile (the lower and upper quartiles, respectively), and the band near the middle of the box is the 50th percentile (the median). The ends of the whiskers are defined as follows: upper whisker = min(max value, Q3 + 1.5*IQR) and lower whisker = max(min value, Q1−1.5*IQR), where IQR = Q3-Q1 is the box length. **b.** Mosaic plot shows the number of eQTLs (local FDR = 10%) in different MAF bins (column). We included the number of non-eQTL (the top row of the plot coloured by gray) in comparison to the number of eQTLs (bottom row coloured by green). **c-d.** The trend of posterior probabilities of eQTLs or stationary genes (one of the 7 differential expression categories) against gene expression levels. The line was estimated using logistic regression where the response variable is the posterior probability greater than 0.9 against the expression quantile based on the average expression across all cells for each gene. **e.** The number of eQTL genes stratified by the spatial DE genes demonstrated in Extended Data Fig. 2d. **f.** Forest plot showing the enrichment of the 1,275 eQTL genes in each of the 7 DE categories. The black dots show non-adjusted eQTL enrichment, and the red dots show the enrichment following adjustment for gene expression levels (Online Methods). The error bars in the forest plot show 95% confidence intervals (standard errors) of odds ratios using $N = 10,748$ genes as independent samples (see Online Methods for details). **g.** The number of eQTL genes colocalised with GTEx fibroblast eQTLs in the 7 different DE gene categories. **h.** Forest plot shows the enrichment of the eQTL colocalised with GTEx fibroblast eQTLs in the 7 different DE gene categories. The error bars in the forest plot show 95% confidence intervals (standard errors) of odds ratios using $N = 10,748$ genes as independent samples (see Online Methods for details). **i.** UMAPs showing spatial distributions of eQTL effect sizes estimated from the GP mixture model with 4 different spatial eQTL categories (static, primary, secondary and unknown; Online Methods). UMAP coordinates are identical to Fig. 2a. **j.** The numbers of static and dynamic eQTLs estimated from the GP mixture model. **k.** Heatmap showing the enrichment of spatial eQTL categories for the 7 DE categories and the GTEx fibroblast eQTLs. Colour scale shows the odds ratio of enrichment and the size of dots denotes the magnitude of Bonferroni-corrected *P*-values of the enrichment. The odds ratios and corresponding *P*-values (from a one-sided Chi-square test) were computed using $N = 10,748$ genes as independent samples (see Online Methods for details). Dots with $P > 0.05$ were omitted.



Extended Data Fig. 5 | eQTL sharing with GTEx and other immune cell eQTLs. Prior probability inferred by the pairwise hierarchical model (Online Methods) fitted between our eQTLs and those for the 48 GTEx tissues as well as immune cells (monocytes and iPSC-derived macrophages) under different stimulation conditions.



Extended Data Fig. 6 | eQTL and disease GWAS colocalisation. a. ATF6 expression on the UMAP of fibroblast data. UMAP coordinates are identical to Fig. 2a. **b**. ATF motif (M6155_1.02; CIS-BP version 1.02). The nucleotide C coloured by red indicates the location of the eQTL variant rs1998266T>C. **c**. Locus zoom plots of hayfever, allergic rhinitis or eczema, rheumatoid arthritis and the ETV7 eQTL. **d**. The numbers of colocalised loci between our eQTLs and OpenTarges GWAS traits that were overlapping with the OpenTarges colocalisation resource for GTEx eQTLs (48 tissues) as well as monocytes and iPSC derived macrophages under various stimulation conditions.



Extended Data Fig. 7 | Fine-mapping OAS locus. a. UMAP shows the 18 different cell types annotated previously in the PBMC data[22]. **b**. UMAP shows the secondary response pseudotime calculated from the secondary response genes discovered in the fibroblast data. **c**. Locus zoom plot shows the COVID-19 GWAS, OAS1 and OAS3 eQTL (both in fibroblasts) associations. UMAP coordinates are identical to Extended Data Fig. 7a. **d**. UMAP shows the OAS3 gene expression. UMAP coordinates are identical to Fig. 2a. **e**. UMAP shows the eQTL effect size at rs10774671. UMAP coordinates are identical to Fig. 2a. **f**. Effect directions of OAS1/3 eQTLs and the risk allele of COVID-19 GWAS at the lead variant rs10774671.

nature portfolio

Corresponding author(s): Sarah Teichmann
Tzachi Hagai

Last updated by author(s): 19/4/22

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	CEVIChE (App 2019): https://saezlab.shinyapps.io/ceviche/
Data analysis	R 3.6.1 (https://www.r-project.org/) Beagle 4.0 (https://faculty.washington.edu/browning/beagle/b4_0.html) skewer 0.1.127 (https://github.com/relipmoc/skewer) STAR 2.5.3a (https://github.com/alexdobin/STAR/releases) featureCounts 2.0.0 (http://subread.sourceforge.net/) demuxlet 0.1.0 (https://github.com/statgen/demuxlet) CrossMap 0.5.2 (https://crossmap.readthedocs.io/en/latest/) cellRegMap 0.0.3 (https://limix.github.io/CellRegMap/) GASPACHO 0.0.1 (https://github.com/natsuhiko/GASPACHO)

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

All Smart-seq2 cram files of our fibroblast data are available from the European Nucleotide Archive (Accession ID: PRJEB20147). The genotype data of fibroblast samples is under managed access and available through the HIPSCI portal (https://www.hipsci.org/data). The lines used in this study have the identifiers: HPSI0114pf-eipl, HPSI0114pf-fikt, HPSI0114pf-joxm, HPSI0114pf-lexy, HPSI0114pf-rozh, HPSI0114pf-vabj, HPSI0114pf-vass, HPSI0114pf-zoxy, HPSI0115pf-gifk, HPSI0115pf-melw, HPSI0115pf-zihe, HPSI0214pf-feec, HPSI0214pf-heja, HPSI0214pf-pelm, HPSI0215pf-deyz, HPSI0215pf-fawm, HPSI0215pf-hipn, HPSI0215pf-oilg, HPSI0314pf-bubh, HPSI0314pf-cuhk, HPSI0314pf-qonc, HPSI0314pf-wigw, HPSI0314pf-xugn, HPSI0414pf-ceik, HPSI0414pf-gesg, HPSI0414pf-naju, HPSI0414pf-oaqd, HPSI0514pf-fiaj, HPSI0514pf-kuco, HPSI0514pf-puie, HPSI0514pf-rutc, HPSI0514pf-sohd, HPSI0514pf-vuna, HPSI0614pf-ciwj, HPSI0614pf-miaj, HPSI0614pf-oicx, HPSI0714pf-pipw, HPSI0913pf-diku, HPSI0913pf-eika, HPSI0913pf-lise, HPSI0914pf-euts, HPSI0914pf-kajh, HPSI0914pf-laey, HPSI1013pf-garx, HPSI1013pf-jogf, HPSI1013pf-pamv, HPSI1013pf-sebz, HPSI1013pf-wopl, HPSI1013pf-wuye, HPSI1014pf-qayj, HPSI1014pf-sehl, HPSI1014pf-tixi, HPSI1014pf-toss, HPSI1014pf-tuju, HPSI1014pf-vils, HPSI1113pf-bima, HPSI1113pf-dons, HPSI1113pf-eofe, HPSI1113pf-ieki, HPSI1113pf-oaaz, HPSI1113pf-qolg, HPSI1113pf-wahn, HPSI1113pf-wetu, HPSI1114pf-ualf, HPSI1213pf-hehd, HPSI1213pf-nusw, HPSI1213pf-tolg and HPSI1213pf-xuja. The genotype data of COVID-19 PBMC samples is under managed access and available from European Genome-Phenome Archive (Accession ID: XXXXXXX). The annotation of the CpG site was downloaded from the UCSC website (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/cpgIslandExt.txt.gz). The position weight matrices (PWM) of transcription factor motifs were obtained from CIS-BP (http://cisbp.ccbr.utoronto.ca/bulk.php). The PWM used to find TATA-box in the gene promotor have the following identifiers: M1641_1.02, M2191_1.02, M4011_1.02, M4266_1.02, M6502_1.02, M1642_1.02, M4010_1.02, M4014_1.02 and M4708_1.02. OpenTargets GWAS summary statistics are available from GWAS Catalog (https://www.ebi.ac.uk/gwas/), FINNGEN (https://www.finngen.fi/en/access_results) and UK Biobank (https://www.nealelab.is/uk-biobank). COVID-19 GWAS summary statistics (release 5) are available from the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/results/r5/). The Open Targets colocalisation data is obtained from the website (https://ftp.ebi.ac.uk/pub/databases/opentargets/genetics/210608/). The eQTL summary statistics of GTEx 48 tissues as well as immune cells (iPSC derived macrophages and monocytes) under different stimulation conditions were obtained from the eQTL catalogue (http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/). The 1000 Genomes Project VCF data (version: shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased) is obtained from (http://hgdownload.soe.ucsc.edu/gbdb/hg38/1000Genomes/). The summary statistics of fibroblast eQTLs are available from ZENODO (https://doi.org/10.5281/zenodo.7680146).

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 total sample size is 68 fibroblast lines. The number was determined so that the probability to observe at least one heterozygous/minor homozygous patient at a genetic variant with minor allele frequency 0.05 is greater than 0.9. With N=68, the probability was 0.9990659.
Data exclusions	We haven't excluded any sample.
Replication	We discovered 1,275 expression QTLs in the fibroblast scRNA-seq data obtained from 68 unrelated donors of HIPSCI. Of which, we replicated the OAS1 eQTL using two independent model systems: (1) the PBMC scRNA-seq data from 112 donors including 84 COVID-19 positive individuals; and, (2) the scRNA-seq data of nasal brushing samples obtained from 33 adult COVID-19 positive patients.
Randomization	Randomization is not applicable, because this is a population based eQTL mapping study. Based on the principles of mendelian inheritance, it is hypothesized that our study design is protected against typical forms of confounding factors on gene expression, because any confounding factor is unable to alter genotypes of common germline genetic variants.
Blinding	Blinding is not applicable, because this is a population based eQTL mapping study. In principle, samples were recruited blind with respect to their genotypes and any bias in the association between a genetic variant and gene expression is unexpected.

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
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

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