nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01066-3

Check for updates

OPEN

# Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation

Blagoje Soskic[1,2,5], Kiki Cano-Gamez[1,2,5], Deborah J. Smyth[1], Kirsty Ambridge[1], Ziying Ke[1], Julie C. Matte[1], Lara Bossini-Castillo[1], Joanna Kaplanis[1,2], Lucia Ramirez-Navarro[1], Anna Lorenc[1], Nikolina Nakic[3], Jorge Esparza-Gordillo[3], Wendy Rowan[3], David Wille[3], David F. Tough[3], Paola G. Bronson[4] and Gosia Trynka[1,2] ✉

**During activation, T cells undergo extensive gene expression changes that shape the properties of cells to exert their effector function. Understanding the regulation of this process could help explain how genetic variants predispose to immune diseases. Here, we mapped genetic effects on gene expression (expression quantitative trait loci (eQTLs)) using single-cell transcriptomics. We profiled 655,349 CD4+ T cells, capturing transcriptional states of unstimulated cells and three time points of cell activation in 119 healthy individuals. This identified 38 cell clusters, including transient clusters that were only present at individual time points of activation. We found 6,407 genes whose expression was correlated with genetic variation, of which 2,265 (35%) were dynamically regulated during activation. Furthermore, 127 genes were regulated by variants associated with immune-mediated diseases, with significant enrichment for dynamic effects. Our results emphasize the importance of studying context-specific gene expression regulation and provide insights into the mechanisms underlying genetic susceptibility to immune-mediated diseases.**

Translating variants from genome-wide association studies (GWASs) to function provides insights into disease biology and improves treatment options[1]. Disease-associated variants from GWASs are enriched within active chromatin regions[2,3], implicating regulation of gene expression. These effects can be discovered using expression quantitative trait loci (eQTLs), which link variants to gene expression changes[4]. However, most currently available eQTL maps use bulk tissues and thus fail to capture gene expression dynamics, such as changes associated with a developmental stage[5,6] or external stimulus[7,8] in a given cell type[9,10]. Mapping dynamic gene expression changes at a single-cell level could overcome these limitations and provide insights into the molecular mechanisms underlying disease.

Variants associated with immune-mediated diseases are enriched in enhancers and promoters whose activity is upregulated upon CD4+ T cell activation[11,12]. However, CD4+ T cells comprise naive cells, which have not yet encountered an antigen, and memory cells, which have previously undergone activation, both of which respond differently to activation[13–15]. Furthermore, memory cells consist of several subpopulations such as central memory ($T_{CM}$), effector memory ($T_{EM}$), and effector memory cells re-expressing CD45RA ($T_{EMRA}$), which differ in proliferative capacity and effector potential[16–18]. Additionally, regulatory T cells ($T_{reg}$), a subset of CD4+ T cells, control T cell activation and prevent excessive inflammation. Transcriptionally, these subpopulations form a continuum of phenotypes[19]. This cellular heterogeneity further complicates interpretation of immune disease-associated variants.

Given the dynamic nature of T cell activation and the heterogeneity of CD4+ T cells, we mapped gene expression regulation using single-cell transcriptomes spanning four time points of CD4+ T cell activation. We reconstructed activation trajectories for naive and memory CD4+ T cells and identified eQTL effects manifesting at different time points and across different subpopulations of cells. We identified 127 genes with colocalizing eQTL and GWAS signals for immune-mediated diseases. Colocalizing genes were enriched in time-dependent eQTLs. Our data suggest that dysregulation of gene expression during T cell activation could underlie immune disease and emphasize the importance of context-specific gene expression regulation.

## Results

**Single-cell response of CD4+ T cells to activation.** We isolated and stimulated naive and memory CD4+ T cells from 119 individuals and performed single-cell RNA sequencing (scRNA-seq)[20] (Fig. 1a, Supplementary Tables 1 and 2 and Supplementary Fig. 1a,b). We profiled cells in resting state, before dividing (16 h), after the first cell division (40 h) and after acquiring effector functions (5 d)[19]. This process resulted in high-quality data for 655,349 cells (Methods and Supplementary Fig. 1c–g).

We performed dimensionality reduction and embedding using uniform manifold approximation (UMAP) (ref. [21]) (Methods) and observed that cells separated by time point of stimulation, forming a gradual progression from resting to the most activated cell state (cells collected at 5 d) (Fig. 1b). This progression was accompanied by changes in activation markers. For example, an early activation marker, *CD69*, was upregulated at 16 h but downregulated at later time points, whereas expression of *IL2RA*, a marker of late activation, peaked at 40 h, remaining present at 5 d (Fig. 1c). A population of cells localized at the intermediate point between resting and 16 h-stimulated cells (Fig. 1b), and was composed of cells from the 16 h (74%) and 40 h (26%) time points. We hypothesized that this intermediate group represented an early activation state. By analyzing cells from these two time points independently, we observed that

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [2]Open Targets, Wellcome Genome Campus, Cambridge, UK. [3]GSK, R&D, Stevenage, UK. [4]R&D Translational Biology, Biogen, Cambridge, MA, USA. [5]These authors contributed equally: Blagoje Soskic, Kiki Cano-Gamez. ✉e-mail: gosia@sanger.ac.uk

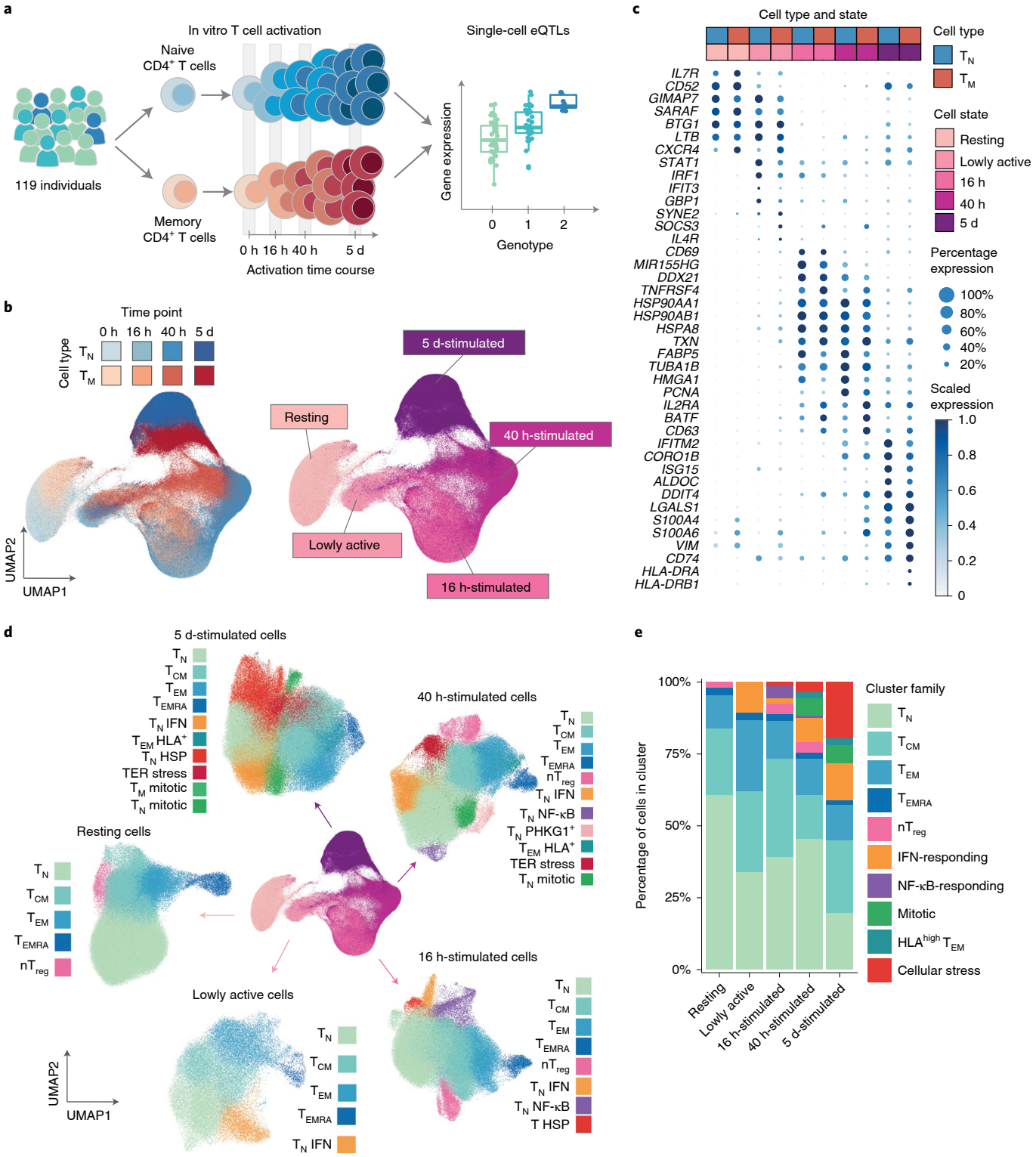



**Fig. 1 | A single-cell transcriptional map of CD4⁺ T cell activation. a**, Schematic of the study design. **b**, UMAP embedding of scRNA-seq data for unstimulated CD4⁺ T cells and at three time points after activation. Colors represent cell types (blue, naive T cell ($T_N$); red, memory T cell ($T_M$)), and shades of colors indicate time points (lighter shades for early time points and darker shades for late time points). Right panel represents the five broad cell states. **c**, Dot plot of highly variable gene expression throughout T cell activation. Shades of blue represent average expression in each cell population, and dot sizes represent the proportion of cells expressing the gene. **d**, Separate UMAP embeddings for the five broad cell states. Colors represent cell populations derived from unsupervised clustering. **e**, Proportion of different cluster groups present at each time point. Cell populations defined from clustering were classified into one of ten families, represented in different colors. ER, endoplasmic reticulum; HLA, human leukocyte antigen; HSP, heat shock protein; NF-κB, nuclear factor κB; nTreg, natural (i.e. thymus-derived) regulatory T cells.

at each of these time points cells separated into two clear groups, one corresponding to the early activation state (Supplementary Fig. 3). Cells in the early activation group expressed fourfold fewer genes compared to other cells at their respective activation time points and showed lower expression of T cell activation markers[19] (Supplementary Fig. 3). Furthermore, they showed a unique profile characterized by high expression of *STAT1*, *IFIT3* and *GBP1* (Fig. 1c). Therefore, these cells represent a distinct, early activation state, and we refer to them as lowly active.

Next, we performed unsupervised clustering of cells throughout the activation time course. This revealed a total of 51 cell clusters, which were merged into 38 cell populations based on their correlated patterns of gene expression (Supplementary Fig. 4 and Methods). This included 25 stable subpopulations consistently detected at multiple time points and 13 transient cell states only detected at specific time points (Fig. 1d and Supplementary Table 3). Stable subpopulations belonged to one of five phenotypes: naive ($T_N$), $T_{CM}$, $T_{EM}$, $T_{EMRA}$ and $T_{reg}$ CD4$^+$ T cells (Fig. 1d,e). The memory pool consisted on average of 60% $T_{CM}$, 30% $T_{EM}$, 5% $T_{reg}$ and 5% $T_{EMRA}$ (Supplementary Fig. 6a). The percentage of $T_{EM}$ cells decreased, whereas $T_{CM}$ and $T_{EMRA}$ increased with age (Supplementary Fig. 6b). We observed no significant differences in subpopulations between sexes (Supplementary Fig. 6c).

Additionally, we observed transient cell states that were only detected at specific activation time points (Fig. 1d,e), such as a population of cells expressing high levels of interferon (IFN)-induced genes (e.g., *IFI6*, *IFIT3*, *ISG15* and *MX1*) during early activation (Supplementary Fig. 5). Another subpopulation expressed high levels of nuclear factor κB response genes (e.g., *NFKBID*, *REL* and *BCL2A1*) (Supplementary Fig. 5) and was dominant at midstages of activation. Additionally, during late activation, we observed a population of mitotic cells and a group of cells expressing high levels of heat shock protein family members (for example *HSPA1A*, *HSPA1B* and *DNAJB1*; Supplementary Fig. 5). Notably, heat shock proteins have been implicated in controlling T cell responses to fever[22]. We also observed a subset of $T_{EM}$ cells that upregulated HLA molecules (e.g., *HLA-DRA*, *HLA-DPA1* and *HLA-DRB1*) during late activation (Supplementary Fig. 5). Importantly, all individuals contributed uniformly to each cluster, with more variability observed in $T_{EMRA}$, as previously described[17,23] (Supplementary Fig. 5f).

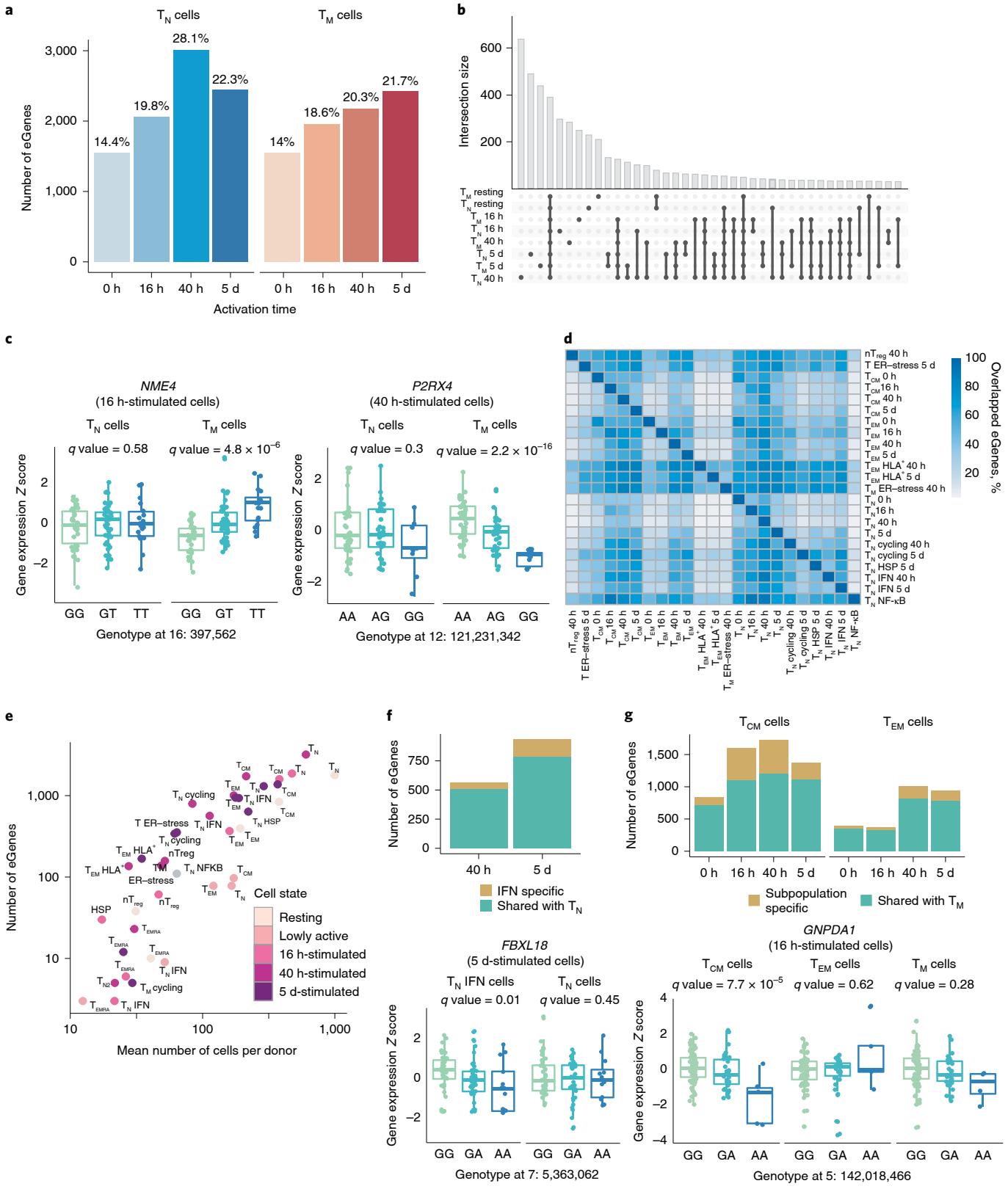
**A temporal eQTL map of CD4$^+$ T cell activation.** To study the genetic regulation of gene expression throughout T cell activation, we performed *cis*-eQTL mapping. For each time point, we reconstructed average transcriptional profiles per cell type and individual (i.e., pseudobulk transcriptomes) corresponding to $T_N$ and $T_M$ CD4$^+$ T cells (Methods). We detected 1,545–3,006 genes with significant *cis*-eQTL effects (eGenes) at different activation time points (Fig. 2a), of which 210–640 eGenes were only detected in individual cell states (Fig. 2b). For example, the kinase gene *NME4* and the purinoceptor gene *P2RX4* only showed effects in $T_M$ at 16 h and 40 h of activation, respectively (Fig. 2c). The multivariate adaptive shrinkage (mashR) method[24] revealed a higher level of eQTL sharing across cell types within the same time point (Supplementary Fig. 7a) than across different time points, suggesting that eQTL effect sizes change throughout activation. We also observed a high replicability (0.67–0.75) of our results with publicly available CD4$^+$ T cell eQTLs from bulk RNA sequencing[25,26] (Supplementary Fig. 7c). However, eQTL sharing was reduced when taking into account both the direction and the magnitude of eQTL effects (0.28–0.34) (Supplementary Fig. 7c), suggesting that effect sizes might differ between different transcriptomic profiling strategies, naive and memory cells and across T cell activation time points.

To gain a more granular view of gene expression regulation throughout T cell activation, we mapped eQTLs in the 38 cell populations (Fig. 1). As expected, we observed a high overlap between eGenes detected in different subpopulations (Fig. 2d). Nevertheless, $T_{EM}$ and cells expressing HLA genes ($T_{EM}$ HLA$^+$) had a higher number of specific eGenes (62%–97%) compared to other populations, suggesting that they are more transcriptionally different than other subsets. Small subpopulations, such as $T_{EMRA}$, yielded a low number of eGenes (3–23), which suggested that the statistical power to detect eGenes correlates with the number of cells profiled ($R^2 = 0.82$, $P = 4.8 \times 10^{-10}$) (Fig. 2e). Indeed, when we subsampled different numbers of $T_{CM}$ cells and repeatedly performed eQTL mapping, we observed that eGene discovery increased proportionally to the number of cells analyzed (Supplementary Fig. 7b). Despite this, we identified subpopulation-specific eGenes absent from the whole $T_N$ or $T_M$ populations. For example, 56–153 eGenes (10%–16%) were found in the subpopulation of naive cells characterized by expression of high levels of IFN-induced genes, but these effects were absent from the whole activated $T_N$ cells (Fig. 2f). Similarly, 47–528 (13%–31%) eGenes were detected in either of the two largest $T_M$ subpopulations ($T_{CM}$ and $T_{EM}$), but not in the whole $T_M$ population (Fig. 2g). For example, *GNPDA1* was an eGene in $T_{CM}$, but not in $T_{EM}$ or in whole memory cells (Fig. 2g). These genes were only detected as eQTLs in specific cell clusters, and we observed that many were not detected in the Database of Immune Cell eQTLs (DICE) data set, which includes different subsets of CD4$^+$ T cells in resting and stimulated states[26]. For example, *VAMP8 and AIMP1* eQTLs ($T_{CM}$-specific $P_{adj} = 5 \times 10^{-4}$ and $P_{adj} = 1.04 \times 10^{-4}$, respectively) and *RNF168* (specific to IFN-expressing cell cluster

**Fig. 2 | eQTL mapping in resting and activated CD4$^+$ T cells. a**, Number of significant eGenes detected at each activation time point. Colors represent cell types (blue, $T_N$; red, $T_M$). **b**, Number of significant eGenes shared between cells sampled at each time point. **c**, Example of T memory cell-specific eQTLs detected at 16 h and 40 h. Box plots show mean expression value of the gene in each sample (*Z*-scored), stratified by genotype at the genomic position of the lead eQTL variant (X axis). Each dot represents a measurement from a separate individual. Central lines indicate the median, with boxes extending from the 25th to the 75th percentiles. Whiskers further extend by ±1.5 times the interquartile range from the limits of each box. *N* of biologically independent samples: $T_N$ *NME4*: 99, $T_M$ *NME4*: 96, $T_N$ *P2RX4*: 89, $T_M$ *P2RX4*: 89. *P* values were derived using tensorQTL and corrected as described in Methods. **d**, Pairwise comparison of eGenes shared between cell subpopulation. Only subpopulations with >100 eGenes were analyzed. **e**, Scatter plot showing the correlation between number of cells per donor and number of detected eGenes in each cluster. **f**, Subpopulation-specific eQTLs detected in IFN-responsive clusters. Bar plot (top) indicates the number of eGenes detected in the IFN-responsive subpopulation that are shared with naive T cells as a whole. Boxplots (bottom) show an example eQTL specific to this subpopulation. Each dot represents a measurement from a separate individual. Central lines indicate the median, with boxes extending from the 25th to the 75th percentiles. Whiskers further extend by ±1.5 times the interquartile range from the limits of each box. *N* of biologically independent samples: $T_N$ IFN *FBXL18*: 96, $T_N$ *FBXL18*: 87, *P* values were derived using tensorQTL and corrected as described in Methods. **g**, Number of subpopulation-specific eQTLs detected in $T_{CM}$ and $T_{EM}$ cells. Bar plots (top) indicate numbers of eGenes detected in $T_{CM}$ and $T_{EM}$ subpopulations that are shared with memory T cells as a whole. Boxplots (bottom) show an example eQTL specific to the $T_{CM}$ subpopulation. Each dot represents a measurement obtained from a separate individual. Central lines indicate the median, with boxes extending from the 25th to the 75th percentiles. Whiskers further extend by ±1.5 times the interquartile range from the limits of each box. *N* of biologically independent samples: $T_{CM}$ *GNPDA1*: 100, $T_{EM}$ *GNPDA1*: 103, $T_M$ *GNPDA1*: 97. *P* values were derived using tensorQTL and corrected as described in Methods. ER, endoplasmic reticulum; nTreg, natural regulatory T cells.

$P_{adj} = 6.2 \times 10^{-3}$) were not detected in any of the T cell populations in DICE. Therefore, as more studies emerge, the power to detect cluster-specific eQTLs will increase, uncovering eQTLs that were previously undetected in bulk tissues.

**Cell-type-specific coexpression gene modules.** We next sought to understand which transcriptional programs shape the T cell response to activation and whether eGenes regulate T cell functions. We computed pairwise gene expression correlations[27] of 11,130 highly expressed and variable genes across 106 individuals and the 38 identified cell populations (Fig. 3a, Supplementary Fig. 8 and Methods). We identified 12 gene modules that represent key cellular functions involved in T cell activation (Fig. 3b and Supplementary Table 4 and 5). For example, module 4 contained genes involved in



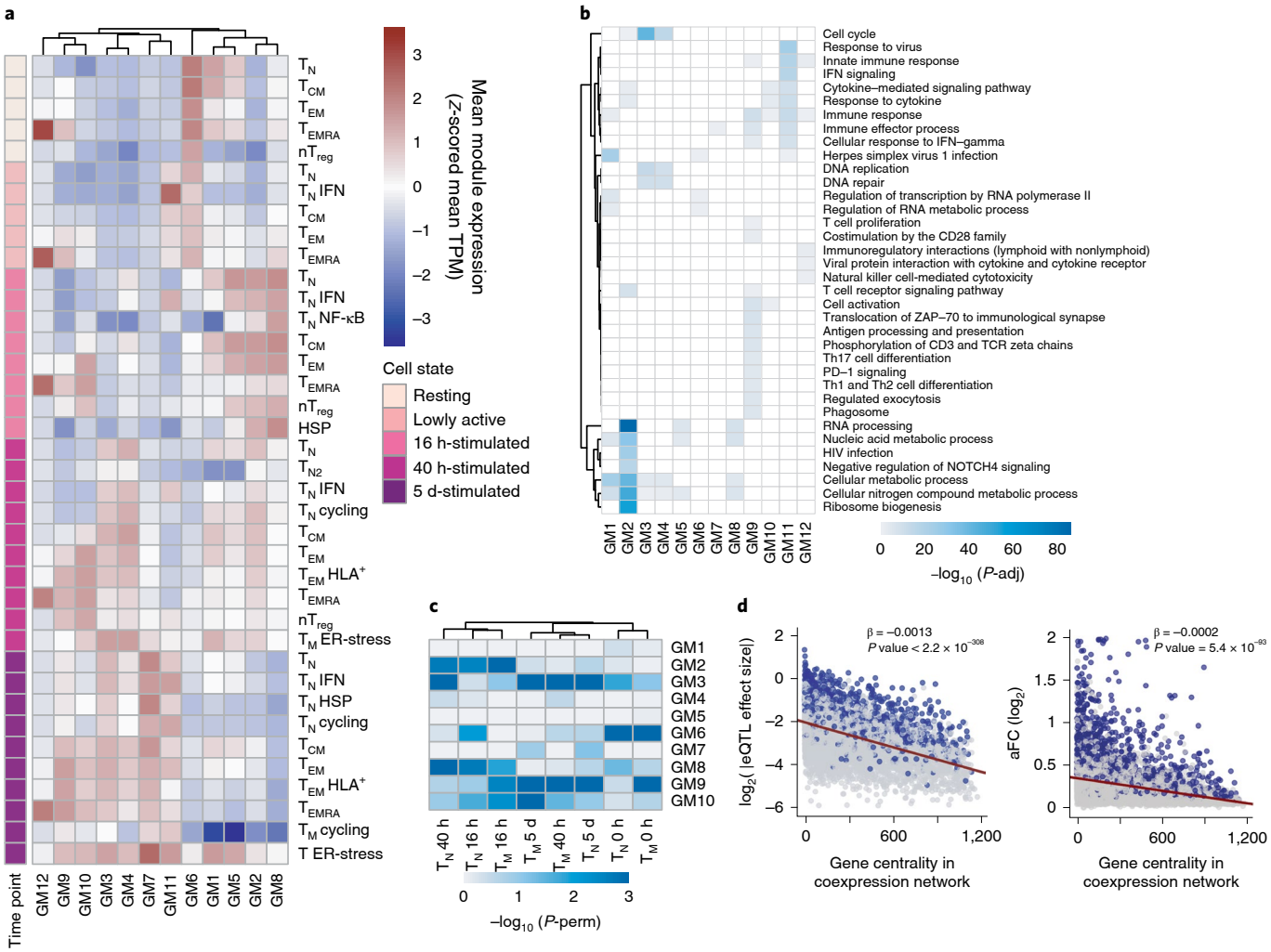



**Fig. 3 | eQTLs are enriched in proliferation and immune response gene modules. a**, Heatmap showing the expression pattern of the 12 identified gene modules. Rows correspond to cell subpopulations. Colors represent the scaled (*Z*-scored) average expression of all genes belonging to a module in a given subpopulation. Gene coexpression network was built using weighted gene coexpression network analysis to identify gene modules. TPM, transcripts per million. **b**, Pathways enriched in each gene module (GM). Shades of blue represent the $\log_{10}$-transformed enrichment *P* values. **c**, Enrichment of eGenes in gene modules. Shades of blue represent $\log_{10}$-transformed *P* values. *P* values were estimated by repeatedly permuting group labels and quantifying the proportion of times an enrichment equal to or larger than the observed one was obtained. *P*-perm, permutation *P* value. **d**, Relationship between a gene's connectivity and the effect size of its lead eQTL variant (left) or allelic fold change (right). All eQTL effect sizes were $\log_2$ transformed. Blue dots represent significant eGenes, whereas gray dots represent genes that do not pass the multiple testing correction. Lines represent the best linear fits obtained from linear regression. *P* values were estimated by testing the null hypothesis of zero intercepts using an *F* test. *P*-adj, adjusted *P* value.

the regulation of cell cycle checkpoints and DNA repair and was highly expressed at 40 h and 5 d after activation, consistent with the timing of the first cell division[28]. Furthermore, module 11 included genes whose expression peaked in lowly active and 16 h-stimulated cells and remained high at later time points. These genes were involved in IFN-induced antiviral mechanisms such as OAS and ISG15-signaling, which are induced rapidly upon viral infection.

In addition to separating genes by temporal dynamics, the coexpression networks also highlighted subpopulation-specific gene expression modules, corresponding to effector T cell functions. For example, genes involved in cytokine secretion and interleukin signaling were highly expressed in $T_{EM}$ and $T_{EMRA}$, but not $T_{CM}$ or $T_N$ cells (Fig. 3a,b), reflecting the potential of $T_{EM}$ and $T_{EMRA}$ cells to respond quickly[18,19]. Consistent with this observation, $T_{EM}$ and $T_{EMRA}$ showed high expression of T cell receptor (TCR)-induced genes (i.e., targets of ZAP-70 and downstream of CD3 zeta chain phosphorylation) at an earlier stage of activation, whereas other subpopulations did not express these genes until 40 h after stimulation (Fig. 3a,b). Furthermore, we observed that module 12, which included genes important for cytotoxic function and chemokine signaling, was most highly expressed in $T_{EMRA}$ (Fig. 3a). This cytotoxic capacity distinguishes $T_{EMRA}$ from other T cell subpopulations.

Next, using a permutation strategy (Methods), we showed that eGenes detected in activated T cells were particularly enriched in modules 2 (metabolism), 3 (cell division) and 9 (immune processes) (Fig. 3c). In contrast, eGenes detected in resting cells showed strongest enrichment in module 6 (RNA metabolism and herpes infection) (Fig. 3c). Finally, we observed that eQTL effect sizes, as well as log-transformed allelic fold changes[29], negatively correlated with the centrality values of their corresponding eGenes in the coexpression network; that is, eGenes with larger eQTL effects were less connected in the network (Fig. 3d and Supplementary Fig. 8b,c). This suggests that genes at the edges of the coexpression network are more tolerant to variation in gene expression.

**Modeling of time-dependent eQTL effects.** Previous studies showed that eQTLs can be context specific[7,30]. Therefore, we assessed the role of genetic variation on the regulation of gene expression dynamics throughout T cell activation (dynamic eQTLs). We used trajectory inference[31] (Methods) to model activation time as a continuous variable (Fig. 4a). The inferred trajectory agreed with the time points profiled experimentally, and T cell activation markers such as *IL7R* (reduced expression upon activation), *CD69* (early activation) and *IL2RA* (early and late activation) (Fig. 4b) also followed their expected expression patterns. In total, we identified 5,090 genes for which expression changed as a function of pseudotime (Supplementary Table 6). For example, *IRF1* and *TOP2A* were respectively downregulated and upregulated at late stages of activation (Fig. 4b). Dynamically regulated genes were enriched in pathways related to T cell activation, such as DNA replication and regulation of cell cycle, mRNA transcription and processing, protein translation, signaling downstream of the TCR and signaling by interleukins (Supplementary Table 7). Finally, we observed that $T_M$ cells were characterized by lower pseudotime values than $T_N$ cells sampled at the same time points. This is a consequence of $T_M$ cells showing a shorter activation path, likely reflecting a faster response.

To model dynamic eQTLs, we divided the pseudotime trajectory into ten bins and averaged the expression of genes per individual in each bin (Methods). Splitting the trajectory enabled us to control for the numbers of cells and therefore to reliably estimate mean gene expression values. We then used mixed models to identify eQTLs for which the effect size changed as a function of activation time (Fig. 4c and Methods). We identified 2,265 genes with dynamic eQTL effects, which comprised 34% of eGenes in our data set (Supplementary Table 8 and Supplementary Fig. 9a). We used a permutation-based strategy to validate that this method was well calibrated (Methods and Supplementary Fig. 9b). We applied both linear and quadratic models and observed that most eQTLs followed linear dynamics (74% and 76% in $T_N$ and $T_M$ cells, respectively; Fig. 4e). However, for 502 and 495 genes in naive and memory cells, respectively, we detected a nonlinear interaction with activation pseudotime. For example, *GBP7* and *CFLAR* demonstrated eQTL effects only upon activation, and their magnitude peaked at midstages of the pseudotime trajectory (Fig. 4d). In contrast, the magnitude of an eQTL for *SERINC5* peaked at early stages of the trajectory and diminished throughout activation (Fig. 4d), whereas an eQTL for the INF-α-inducible gene *IFI27L1* showed an effect size that linearly increased along the activation trajectory.

Finally, linear eQTLs were enriched in metabolic pathways, whereas nonlinear eQTLs were enriched in both metabolic and immune processes (e.g., T cell proliferation and leukocyte degranulation) (Fig. 4f). This suggests that for many immune genes, genetic regulation is only evident during specific stages of T cell activation.

**Colocalization at GWAS loci identifies immune disease genes.** We obtained summary statistics for 13 immune-mediated diseases available in the GWAS catalog[32] (Methods) and tested for colocalization[33,34] (Methods) with the eQTLs mapped to $T_N$, $T_M$ and the subpopulations. We identified 471 unique colocalizations (PP4 > 0.8), corresponding to 247 GWAS loci for 11 diseases and 314 SNP–gene pairs (Supplementary Tables 9 and 10). This enabled us to prioritize 127 candidate disease-causal genes (Fig. 5a). Importantly, 77 (60%) colocalizing genes were detected upon activation and would have been missed by profiling only steady state ex vivo cells. Out of those, 47 (37%) were captured specifically in later time points of activation (40 h + 5 d) (Fig. 5b). This finding is important, as previous eQTL studies have relied on either resting cells or a single, usually early activation time point[2,26].

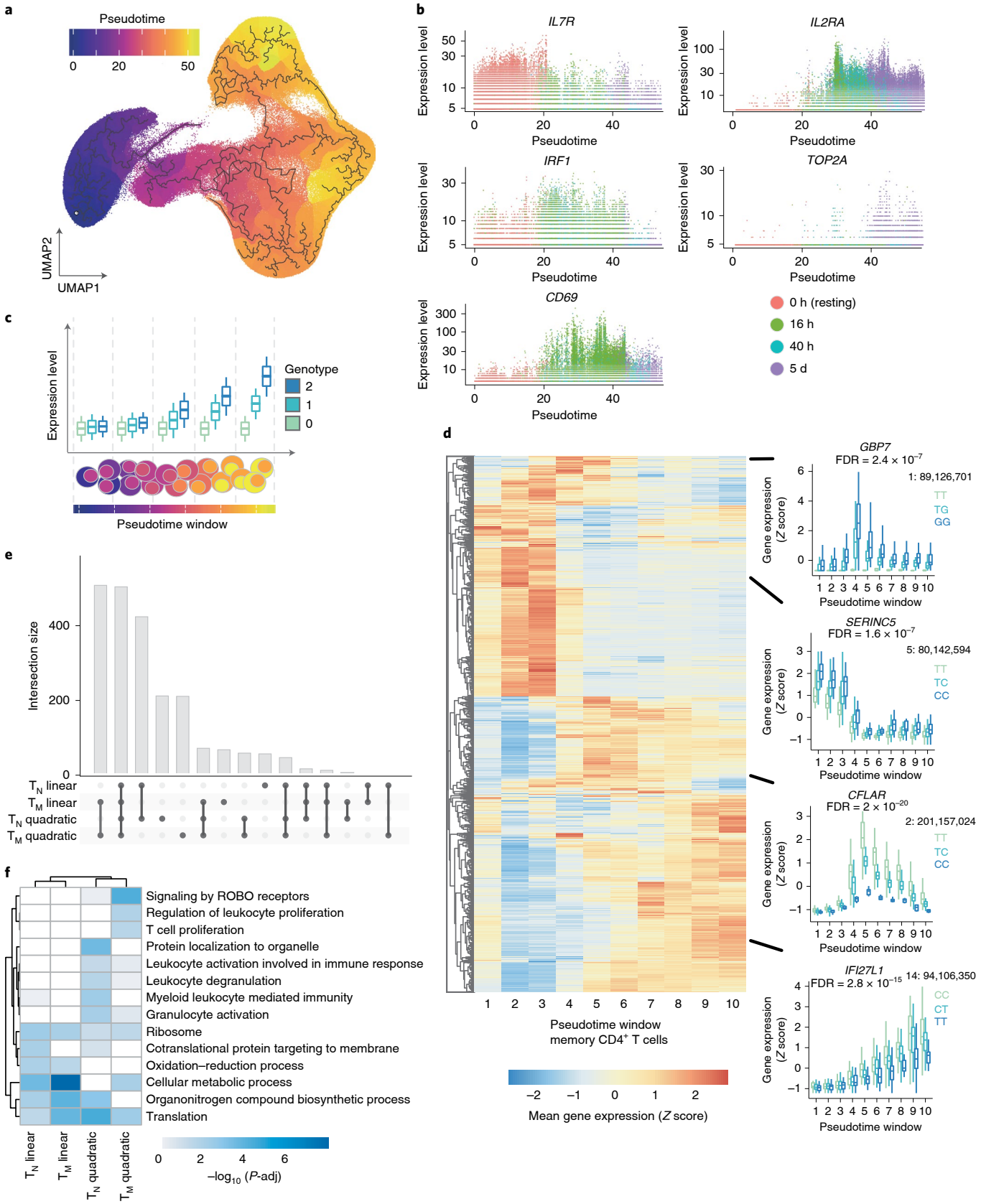
Generally, we observed more colocalizations in larger cell populations (for which we were more powered to detect eQTLs) and in traits with larger numbers of reported GWAS signals (Fig. 5a). The traits with the highest number of colocalizations were Crohn's disease and ulcerative colitis, followed by allergic diseases, in agreement with their proposed T cell-driven biology[11,12,35]. Nevertheless, higher number of colocalizations was not only a consequence of more powered GWAS. Systemic lupus erythematosus, although characterized by a higher number of loci compared to type 1 diabetes, had a smaller proportion of colocalizing variants, in line with studies pointing towards B cells as drivers of systemic lupus erythematosus[11,36]. We found that 72% of genes colocalized only with one trait, 14% with two traits and 14% with three or more diseases (Supplementary Fig. 10a). Overall, 220 disease loci (89%) regulated a single gene, whereas 22 (9%) and 5 (2%) loci regulated two and three genes in the associated regions, respectively.

Although most colocalizing genes were detectable in broad cell types (i.e., total $T_N$ or $T_M$ cells per time point; median per trait = 66%), we observed between 2 and 15 genes per disease (median per trait = 25%) that were only detected in individual subpopulations (Fig. 5c). For example, an eQTL for *TYK2* specifically detected in 16 h-stimulated $T_{EM}$ cells colocalized with a Crohn's disease association (Supplementary Fig. 10b). Similarly, we identified a colocalization between a Crohn's disease locus and a *ZMIZ1* eQTL specific to 16 h-stimulated $T_{CM}$ cells (Supplementary Fig. 10c). This eQTL is absent in other memory T cell populations such as $T_{EM}$, which leads to the eQTL being masked in bulk memory cells, where it is no longer detectable (Supplementary Fig. 10c). Both of these colocalizations are subpopulation and time-point specific, which highlights the importance of measuring gene expression regulation with cell type and state resolution. We observed no differences in the network connectivity of colocalizing genes compared to other eGenes (Supplementary Fig. 10d).

Given that the majority of colocalizations were detected in activated T cells, we asked if these genes showed dynamic genetic regulation. Dynamic eQTLs were enriched in colocalizing eGenes in both naive and memory T cells (36/73 and 44/72 colocalizing genes in naive and memory cells, Fisher's test $P$ values $7.9 \times 10^{-5}$ and $2.6 \times 10^{-7}$, respectively). The expression patterns of most colocalizing eGenes were similar between naive and memory cells (Fig. 5d). An example of a gene whose genetic regulation differs between naive

**Fig. 4 | eQTLs with dynamic effects during CD4⁺ T cell activation. a**, Cells were ordered into a branched pseudotime trajectory using *monocle3*. The UMAP embedding shows all cells, colored by their estimated pseudotime values. Black lines indicate the inferred branched trajectory. **b**, Example genes that significantly change as a function of activation pseudotime. Each dot corresponds to a cell, and colors represent experimental time points. **c**, Schematic of the analysis approach. Cells were split into ten windows of equal cell numbers according to their estimated pseudotime values. Linear and quadratic mixed models were applied to each previously identified eGene to test for an interaction between genotypes and T cell activation pseudotime. **d**, Heatmap showing the expression pattern of each dynamic eGene in memory T cells. Boxplots show examples of nonlinear and linear dynamic eQTLs. The average expression of the gene within each pseudotime window was stratified by genotype. Central lines indicate the median, with boxes extending from the 25th to the 75th percentiles. Whiskers further extend by ±1.5 times the interquartile range from the limits of each box. *N* of biologically independent samples: 106. *P* values were derived and corrected as described in Methods. **e,** Number of eGenes with evidence of a significant genotype–pseudotime interaction (i.e., dynamic eQTLs) in a linear or quadratic mixed model. **f**, Pathways enriched in linear and quadratic eGenes. Shades of blue represent $\log_{10}$-transformed enrichment *P* values. Enrichment *P* values were estimated using a hypergeometric test, and multiple testing correction was performed using the set counts and sizes (SCS) method, as implemented in gprofiler2 version 0.2.0. FDR, false discovery rate. ROBO, roundabout receptors.

and memory cells is the gene encoding the interleukin-18 receptor (IL18R1), a dynamic eQTL in memory T cells. *IL18R1* is highly expressed during early activation of memory cells and, conversely, during late activation of naive cells (Fig. 5e). Another example is *CTLA4*, a dynamic eQTL in both memory and naive T cells but with different regulation in the two cell types (Fig. 5f); naive cells upregulated and maintained high expression of *CTLA4* upon activation, whereas memory cells highly expressed *CTLA4* only during early



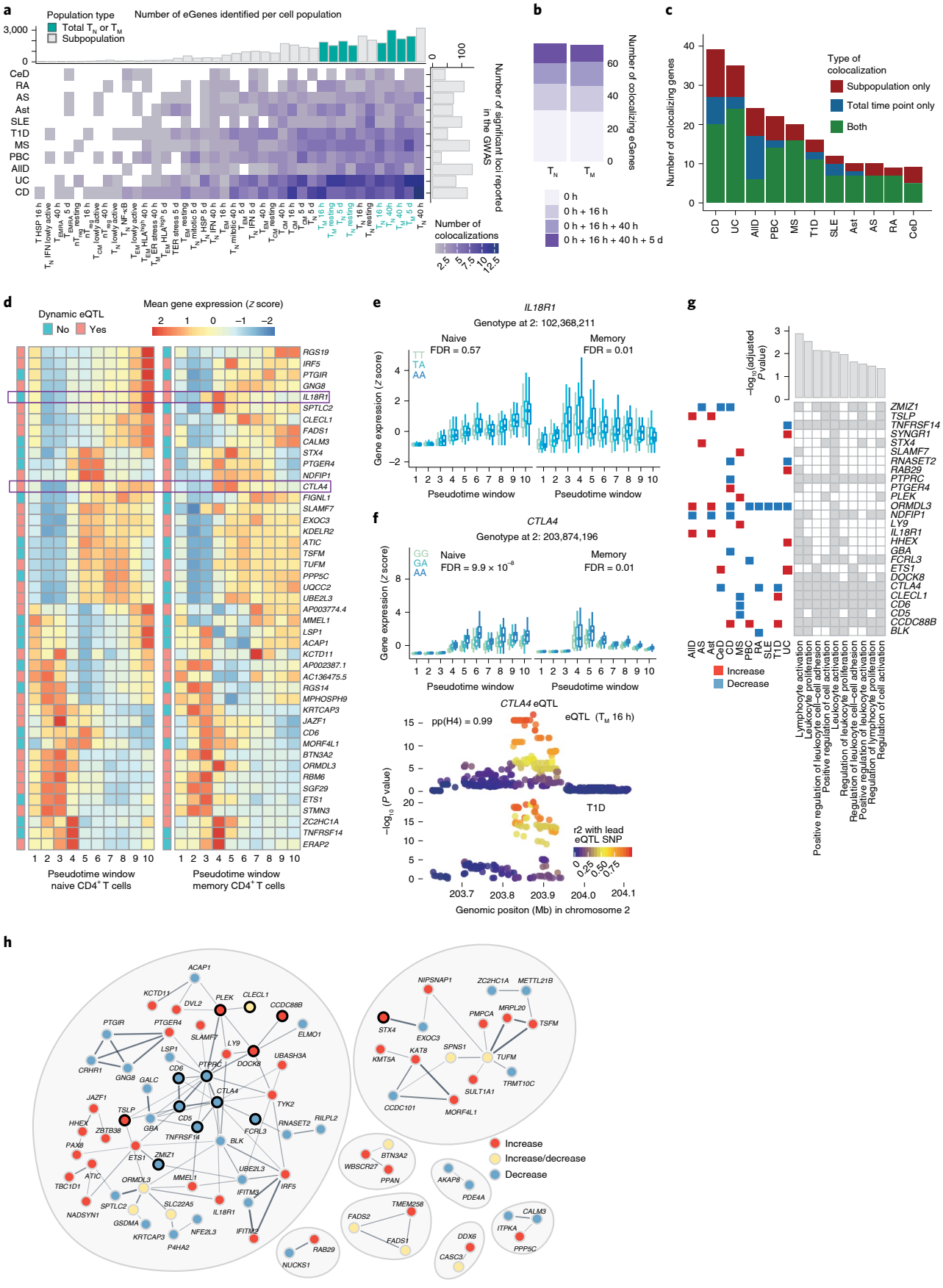

a
Population type
Total $T_N$ or $T_M$
Subpopulation
Number of eGenes identified per cell population
3,000
0
CeD
RA
AS
Ast
SLE
T1D
MS
PBC
AIID
UC
CD
Number of significant loci reported in the GWAS
Number of colocalizations
2.5
5
7.5
10
12.5
b
Number of colocalizing eGenes
60
40
20
0
$T_N$
$T_M$
0 h
0 h + 16 h
0 h + 16 h + 40 h
0 h + 16 h + 40 h + 5 d
c
Number of colocalizing genes
Type of colocalization
Subpopulation only
Total time point only
Both
CD
UC
AIID
PBC
MS
T1D
SLE
Ast
AS
RA
CeD
d
Dynamic eQTL
No
Yes
Mean gene expression (z score)
Pseudotime window naive CD4+ T cells
Pseudotime window memory CD4+ T cells
e
IL18R1
Genotype at 2: 102,368,211
Naive
FDR = 0.57
Memory
FDR = 0.01
Gene expression (z score)
Pseudotime window
f
CTLA4
Genotype at 2: 203,874,196
Naive
FDR = 9.9 × 10−8
Memory
FDR = 0.01
CTLA4 eQTL
pp(H4) = 0.99
eQTL ($T_M$ 16 h)
T1D
r2 with lead eQTL SNP
Genomic positon (Mb) in chromosome 2
g
−log10(adjusted P value)
Increase
Decrease
Lymphocyte activation
Leukocyte activation
Leukocyte cell–cell adhesion
Cell activation
Positive regulation of leukocyte activation
Regulation of leukocyte activation
Regulation of leukocyte proliferation
Positive regulation of leukocyte proliferation
Regulation of lymphocyte proliferation
Regulation of cell activation
h
Increase
Increase/decrease
Decrease

**Fig. 5 | Colocalization of CD4⁺ T cell eQTLs with GWAS associations for immune diseases. a**, Number of significant colocalizations between an eQTL and a GWAS signal identified for each cell type–trait combination. Marginal bar plots represent the number of independent associations reported in the GWAS (*x* axis) and the number of eGenes detected per subpopulation (*y* axis). Light and dark bars indicate whole-cell populations ($T_N$ or $T_M$ cells at a specific time point) and subpopulations, respectively. **b**, Number of additional colocalizing genes detected in stimulated cells. **c,** Number of colocalizing genes observed in whole-cell populations, subpopulations or both. **d**, Heatmap showing the expression pattern of colocalizing eGenes in naive and memory T cells. The color of annotation boxes shows genes that are dynamic and static eQTLs. **e**, Boxplot shows *IL18R1* dynamic eQTLs. The average expression of the gene within each pseudotime window was stratified by genotype. Central lines indicate the median, with boxes extending from the 25th to the 75th percentiles. Whiskers further extend by ±1.5 times the interquartile range from the limits of each box. *N* of biologically independent samples: 106. *P* values were derived and corrected as described in Methods. **f**, Boxplot shows *CTLA4* dynamic eQTLs. The average expression of the gene within each pseudotime window was stratified by genotype. Locus plot for a colocalization between a *CTLA4* dynamic eQTL and a GWAS association for type 1 diabetes. Each dot represents a variant, with colors indicating their linkage disequilibrium with the lead eQTL variant. Central lines indicate the median, with boxes extending from the 25th to the 75th percentiles. Whiskers further extend by ±1.5 times the interquartile range from the limits of each box. *N* of biologically independent samples: 106. *P* values were derived and corrected as described in Methods. **g**, Tile plot shows enriched pathways within colocalizing genes as well as genes driving the enrichment. Bar plots show adjusted *P* values from the enrichment test. Squares on left show the colocalizing disease. Red, disease variant increases gene expression; blue, variant decreases gene expression. **h**, STRING network of colocalizing genes. Red, disease variant increases gene expression; blue, decreases; yellow, effect on gene expression is disease dependent. Black outline highlights genes belonging to the top enriched pathway (GO.0050867: positive regulation of cell activation). GWAS abbreviations: AllD, allergic disease; AS, ankylosing spondylitis; Ast, asthma; CeD, celiac disease; CD, Crohn's disease; MS, multiple sclerosis; PBC, primary biliary cirrhosis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; T1D, type 1 diabetes; UC, ulcerative colitis.

activation. This eQTL colocalized with a type 1 diabetes-associated locus, and individuals carrying the disease risk allele showed lower expression of *CTLA4*. Reduced expression of *CTLA4* at early stages of activation could result in impaired ability to suppress T cells, thus contributing to excessive activation in disease. Additionally, the same eQTL variant colocalized with association signals for rheumatoid arthritis and celiac disease, in agreement with the CTLA4-based therapies used in rheumatoid arthritis[37] (Supplementary Table 9).

Finally, we asked whether immune disease loci affected specific cellular functions. Colocalizing genes were enriched in pathways involved in the regulation of T cell activation and proliferation (Fig. 5g). There were 26 genes driving this enrichment, including genes with association signals shared across two or more diseases. For 24 out of 26 genes, the direction of effect of the risk allele on gene expression was consistent between traits. Colocalizing genes also clustered into connected modules based on the information in STRING[38]; that is, the genes were coexpressed across tissues or the proteins they coded for were physically interacting (Fig. 5h). Furthermore, neighboring genes within these modules tended to be perturbed in the same direction by immune disease variants. For example, we observed a module of interconnected genes, 12 of which were involved in the regulation of T cell activation and proliferation. Among these, *PTPRC* was directly connected to *CD6*, *CD5*, *CTLA4* and *TNFRSF14*. Notably, all of these genes were downregulated by risk alleles, suggesting that their reduced expression may increase disease risk. Our results demonstrate that immune disease loci colocalize with genes involved in the regulation of T cell activation and that genes with similar functions tend to be perturbed in the same direction by disease risk alleles.

## Discussion

Dysregulation of T cell activation can result in poor response to infections, development of inflammatory diseases or primary immunodeficiencies. By using single-cell profiling across 655,349 CD4⁺ T cells, our study provides an unbiased view of the T cell response to activation, revealing 38 distinct subpopulations. This single-cell resolution provides an explanation of previous results from bulk gene expression. For example, we recapitulated the up-regulation of IFN-related genes early upon CD4⁺ TCR engagement[39] and further resolved it to a specific subpopulation of naive cells. We also demonstrated that the previously described modulation of HLA molecules upon T cell activation[40] is driven by $T_{EM}$ cells. Therefore, our data provide a resource for the interpretation of studies of T cell function.

Often, eQTLs obtained from bulk RNA-seq mask cell-type specific effects[41], which can be mapped with single-cell transcriptomics[42]. Many immune cell eQTL resources[25,26], including those capturing T cell activation[39], rely on sorting cells based on surface markers. However, these approaches cannot capture the full cellular heterogeneity. Here, scRNA-seq allowed us to map eQTLs within clusters unbiasedly, providing insights into genetic regulation in different cell subsets. Our study will help infer the effects of genetic regulation on the development of effector T cell functions and could inform cell engineering approaches.

eQTLs can be context specific, including those resulting from responses to stimuli[7,30]. However, current eQTL resources mostly include cells in steady state. Although these resources are instrumental in interpreting GWAS signals, the proportion of GWAS-eQTL colocalizations remains low[43]. In contrast, our study captured context-specific gene expression regulation. In particular, had we only focused on the resting state, we would have missed most disease-relevant eQTLs, as only 40% of colocalizations are detectable in resting cells. Furthermore, colocalizing eQTLs were enriched for eGenes with dynamic regulation, which could explain why at present eQTLs have only explained a small proportion of GWAS associations.

Finally, our results could inform drug target discovery. For example, a *CTLA4* eQTL colocalizes with GWAS associations for three immune diseases, where the disease risk alleles decrease gene expression. CTLA4 removes costimulatory molecules from the surface of antigen-presenting cells, downregulating T cell activation[44]. Thus, a partial reduction in CTLA4 function could impair immune regulation and increase the risk of autoimmunity[45]. This is supported by existing therapies in which a CTLA4 fusion protein is administered to patients with rheumatoid arthritis to help reduce inflammation[46]. Importantly, we show that the expression of *CTLA4* is dynamically regulated, peaking during early activation. Similarly, a *TYK2* eQTL detected in $T_{EM}$ cells colocalizes with a Crohn's disease GWAS association. The *TYK2* locus is associated with ten different immune disorders, with three independent signals reported[1,47,48]. One of these signals is explained by a missense variant, which reduces signaling downstream of several cytokine receptors, resulting in protection from disease[1]. Here, we show a similar effect, where individuals carrying a protective allele for Crohn's disease have lower expression of *TYK2* in $T_{EM}$ cells at 16 h of activation. Inhibition of TYK2 as a treatment for inflammatory diseases is in clinical trials[49,50]. These examples illustrate how colocalizing genes could have therapeutic value.

We note that a limitation of our study is that we profiled healthy individuals. Although this enabled us to identify eQTLs involved in disease susceptibility, we are likely missing eQTL colocalizations relevant for disease progression. Future studies in disease cohorts will be required to understand genetic regulation after disease onset.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01066-3.

Received: 4 August 2021; Accepted: 30 March 2022;
Published online: 26 May 2022

## References

1. Dendrou, C. A. et al. Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* **8**, 363ra149 (2016).
2. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
3. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
4. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
5. 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).
6. Strober, B. J. et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (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. Jerber, J. et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021).
9. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 955 (2020).
10. Regev, A. et al. The human cell atlas.*Elife* **6**, e27041 (2017).
11. Soskic, B. et al. Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases.*Nat. Genet.* **51**, 1486–1493 (2019).
12. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells.*Nat. Genet.* **51**, 1494–1505 (2019).
13. Glinos, D. A. et al. Genomic profiling of T-cell activation suggests increased sensitivity of memory T cells to CD28 costimulation. *Genes Immun.* **21**, 390–408 (2020).
14. Borowski, A. B. et al. Memory CD8$^+$ T cells require CD28 costimulation. *J. Immunol.* **179**, 6494–6503 (2007).
15. Fröhlich, M., Gogishvili, T., Langenhorst, D., Lühder, F. & Hünig, T. Interrupting CD28 costimulation before antigen rechallenge affects CD8$^+$ T-cell expansion and effector functions during secondary response in mice. *Eur. J. Immunol.* **46**, 1644–1655 (2016).
16. Sallusto, F., Geginat, J. & Lanzavecchia, A. Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annu. Rev. Immunol.* **22**, 745–763 (2004).
17. Tian, Y. et al. Unique phenotypes and clonal expansions of human CD4 effector memory T cells re-expressing CD45RA. *Nat. Commun.* **8**, 1473 (2017).
18. Sallusto, F., Lenig, D., Förster, R., Lipp, M. & Lanzavecchia, A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* **401**, 708–712 (1999).
19. Cano-Gamez, K. et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4$^+$ T cells to cytokines. *Nat. Commun.* **11**, 1801 (2020).
20. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
21. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1802.03426 (2018).
22. Wang, X. et al. Febrile temperature critically controls the differentiation and pathogenicity of T helper 17 cells. *Immunity* **52**, 328–341 (2020).
23. Weiskopf, D. et al. Dengue virus infection elicits highly polarized CX3CR1$^+$ cytotoxic CD4$^+$ T cells associated with protective immunity. *Proc. Natl Acad. Sci. U S A* **112**, E4256–E4263 (2015).
24. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2018).
25. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
26. Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune. *Cell Gene Expr. Cell* **175**, 1701–1715 (2018).
27. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* **9**, 559 (2008).
28. Hawkins, E. D. et al. Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. *Nat. Protoc.* **2**, 2057–2067 (2007).
29. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).
30. 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).
31. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
32. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
33. Giambartolomei, C. et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).
34. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, e1008720 (2020).
35. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
36. Scharer, C. D. et al. Epigenetic programming underpins B cell dysfunction in human SLE. *Nat. Immunol.* **20**, 1071–1082 (2019).
37. Gardner, D., Jeffery, L. E. & Sansom, D. M. Understanding the CD28/CTLA-4 (CD152) pathway and its implications for costimulatory blockade. *Am. J. Transpl.* **14**, 1985–1991 (2014).
38. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2018).
39. Ye, C. J. et al. Intersection of population variation and autoimmunity genetics in human T cell activation.*Science* **345**, 1254665 (2014).
40. Gutierrez-Arcelus, M. et al. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat. Genet.* **52**, 247–253 (2020).
41. Kim-Hellmuth, S. et al. Cell type–specific genetic regulation of gene expression across human tissues.*Science* **369**, eaaz8528 (2020).
42. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
43. Chun, S. et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
44. Qureshi, O. S. et al. Trans-endocytosis of CD80 and CD86: a molecular basis for the cell-extrinsic function of CTLA-4. *Science* **332**, 600–603 (2011).
45. Schubert, D. et al. Autosomal dominant immune dysregulation syndrome in humans with CTLA4 mutations. *Nat. Med.* **20**, 1410–1416 (2014).
46. Buch, M. H., Vital, E. M. & Emery, P. Abatacept in the treatment of rheumatoid arthritis. *Arthritis Res. Ther.* **10**, S5 (2008).
47. Diogo, D. et al. TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PLoS One* **10**, e0122271 (2015).
48. López-Isac, E. et al. Influence of TYK2 in systemic sclerosis susceptibility: a new locus in the IL-12 pathway. *Ann. Rheum. Dis.* **75**, 1521–1526 (2016).
49. Papp, K. et al. Phase 2 trial of selective tyrosine kinase 2 inhibition in psoriasis. *N. Engl. J. Med.* **379**, 1313–1321 (2018).
50. Lonial, S. et al. Elotuzumab therapy for relapsed or refractory multiple myeloma. *N. Engl. J. Med.* **373**, 621–631 (2015).

## Methods

**Cell isolation and stimulation.** Blood samples were obtained from 119 healthy individuals of British ancestry. Of these, 67 were male (53.7%) and 52 female (56.3%), and the mean age of the cohort was 47 years (standard deviation = 15.61 years) (Supplementary Fig. 1a). Human biological samples were sourced ethically, and their research use was in accord with the terms of informed consent under an institutional review board/ethics committee-approved protocol (15/NW/0282).

Peripheral blood mononuclear cells (PBMCs) were isolated using Ficoll-Paque PLUS (GE Healthcare) density gradient centrifugation. Naive ($CD25^-$ $CD45RA^+$ $CD45RO^-$) and memory ($CD25^-$ $CD45RA^-$ $CD45RO^+$) $CD4^+$ T cells were isolated from the PBMC fraction using EasySep naive $CD4^+$ T cell isolation kits and memory $CD4^+$ T cell enrichment kits (StemCell Technologies) according to the manufacturer's instructions. Naive and memory T cells were then stimulated with anti-CD3/anti-CD28 human T-Activator Dynabeads (Invitrogen) at a 1:2 beads-to-cells ratio. Cells were harvested after 16 h, 40 h and 5 d of stimulation. In addition, unstimulated cells kept in culture without any beads for 16 h were used as a negative control (i.e., 0 h of activation).

**scRNA-seq.** Upon harvesting, cells were resuspended in RPMI media to obtain a single-cell suspension. Next, cells were stained with the live/dead dye 4,6-diamidino-2-phenylindole, and dead cells were removed from the suspension using fluorescence-activated cell sorting. Live cells were resuspended in phosphate-buffered saline, at which point cells obtained from different individuals but belonging to the same experimental condition were mixed together at equal ratios to form a single-cell suspension (i.e., pool). Each pool corresponded to a mix of cells from four to six different individuals (median = 6), and we processed a total of 172 pools.

Cells were next processed for scRNA-seq using the 10x Genomics 3' v2 kit[20], as specified by the manufacturer's instructions. Namely, $1 \times 10^4$ cells were loaded into each inlet of a 10x Genomics Chromium controller to create Gel Bead-in-emulsions (GEMs). Each experimental condition was loaded in a separate inlet. The targeted recovery was 6,000 cells per pool. Reverse transcription was performed on the emulsion, after which cDNA was purified, amplified and used to construct RNA-sequencing libraries. These libraries were sequenced using the Illumina HiSeq 4000 platform, with 75-bp paired-end reads and one cell pool per sequencing lane.

**Genotyping.** Genomic DNA was isolated from a suspension of $1 \times 10^6$ PBMCs from each individual in the study using a DNA isolation kit (Qiagen). Genotyping was then performed using the Infinium CoreExome-24 (v1.3) chip (Illumina). Genotype data were analyzed as detailed in Supplementary Notes.

**scRNA-seq data analysis.** *Data processing and quality controls.* Raw scRNA-seq data were processed using the Cell Ranger Single-Cell Software Suite[20] (v3.0.0, 10x Genomics). In brief, reads were first assigned to cells and then aligned to the human genome using STAR[51], with the hg38 build of the human genome (GRCh38) as a reference for alignment. Ensembl (v93) was used as a reference for gene annotation, and gene expression was quantified using reads assigned to cells and confidently mapped to the genome.

Results from RNA quantification in Cell Ranger were imported into Python (v3.8.1) and analyzed using scanpy (v1.4.4) (ref. [52]). Samples with less than 70% of reads mapping to cells were discarded. This resulted in 142 (82%) cell pools and 106 (89%) individuals being kept after quality filters. In addition, any cells with fewer than 200 detected genes, an unusually high number of genes (defined as over four standard deviations above the mean number of detected genes), or more than 10% of reads mapping to mitochondrial genes, were removed from the data set. Finally, any genes detected in fewer than ten cells were discarded. This resulted in 713,403 cells (96.77% of total) and 23,360 genes passing quality filters.

*Deconvolution of single cells by genotype.* Each scRNA-seq sample comprised a mix of cells from unrelated individuals. Thus, natural genetic variation was used to assign cells to their respective individuals. First, a list of common exonic variants was compiled from the 1000 Genomes Project phase 3 exome-sequencing data[53]. This list included any variants with a minor allele frequency of at least 5% in the European population. Next, cellSNP (v0.99) (ref. [54]) was used to generate pileups at the genomic location of these variants. These pileups, in combination with the variants called from genotyping in each individual, were used as an input for Vireo (v1) (ref. [54]). Vireo uses a Bayesian approach to infer which cells belong to the same individual based on the genetic variants detected within scRNA-seq reads. Any cells labelled as 'unassigned' (less than 0.9 posterior probability of belonging to any individual) or 'doublets' (containing mixed genotypes) by Vireo were discarded. On average, 92% of the cells in each pool were unambiguously assigned to a single individual in the cohort (Supplementary Fig. 2).

*Cell cycle scoring.* After quality control, the number of unique molecular identifiers (UMIs) mapping to each gene in each single cell were normalized for library size and log-transformed using scanpy's default normalization parameters[52]. Next, a publicly available list of cell cycle genes[55] was used in combination with scanpy to perform cell cycle scoring and assign cells to their respective stage of the cell cycle.

*Exploratory data analysis and removal of cellular contaminations.* We performed exploratory analysis at each experimental time point independently. Cells collected at the same time point were first loaded into scanpy, where normalized log-transformed UMI counts were used to identify highly variable genes. Between 701 and 1,668 highly variable genes were detected at each time point (mean = 1,301). Only highly variable genes were used as a basis for the remaining analyses in this section.

Technical covariates (cell culture batch) and unwanted sources of biological variation (i.e., number of UMIs per cell, proportion of reads mapping to mitochondrial genes, cell cycle scores and reported sex) were regressed out using scanpy's regress_out() function. Next, log-UMI counts were scaled (setting 10 as the maximum value) and used as an input for principal-component analysis (PCA). The first 40 principal components were used to build a $k$-nearest neighbors (kNN) graph (with $k = 15$), which was used as an input for embedding and visualization with the UMAP algorithm[21]. This kNN graph was further used for unsupervised clustering using the Leiden algorithm[56].

At this stage, cell clustering revealed a low proportion of three contaminating cell types that were consistently detected at each time point: B cells, $CD8^+$ T cells and antigen-presenting cells. Furthermore, two additional sources of contamination ($SOX4^+$ precursor cells and cells expressing hallmarks of cell culture stress) were detected at 0 h of activation (Supplementary Fig. 3). Cell contaminations were removed from the data set, resulting in 655,349 (91.86% of total) high-quality cells kept and successfully annotated as $CD4^+$ T cells.

*Identification of a lowly active T cell subpopulation.* Having removed cellular contaminations, highly variable genes were recalculated and the analysis described in the previous section (i.e., batch regression, scaling, PCA, graph construction, embedding and clustering) was repeated using $CD4^+$ T cells only. Cells sampled at 16 h and 40 h showed a clear separation into two groups, one of which expressed a significantly lower number of genes and showed comparatively lower levels of previously described T cell activation markers[19] (Supplementary Fig. 4a). This population of lowly active cells was separated from its original time point and treated as an independent group for clustering.

*Clustering and cluster annotation.* Unsupervised clustering was applied independently to the five cell groups of cells identified in the study (resting, lowly active, 16 h, 40 h and 5 d) based on their respective kNN graphs and using the Leiden algorithm[56]. This method resulted in 51 cell clusters. The similarity of these clusters to each other was assessed by performing PCA on the full data set (i.e., all cells) and estimating the Euclidean distance between pairs of clusters (from cluster center to cluster center) based on the first 100 principal components. Clusters with high levels of similarity or overlapping biological characteristics were merged together (Supplementary Fig. 5b). This method resulted in 38 distinct groups of cells. Gene markers for each of these groups were identified using scanpy's built-in function for gene ranking, which uses a $t$-test to compare the average expression of a gene in a cluster versus its expression outside the cluster. Each cell group was annotated by comparing its inferred marker genes with known cell-type markers reported in the literature.

*Ordering of cells in a pseudotime trajectory.* To perform trajectory inference, raw gene expression measurements for all $CD4^+$ T cells in the study (i.e., 655,349 cells spanning all time points) were imported into R (v3.6.1) and analyzed using monocle3 (v0.2.0) (ref. [31]). As opposed to other analyses, where cells from each time point were treated independently, here, some unwanted sources of variation such as cell cycle scores correlated with the biological process of interest (i.e., T cell activation). Thus, we implemented a hierarchical batch regression approach, where cell cycle scores were first regressed within each time point, followed by batch regression in the full data set. In brief, PCA was performed based on all cells using monocle3's PCA implementation. Next, a matrix containing the first 100 principal component coordinates for each cell was split by time point. Cell cycle effects were then regressed from each submatrix independently using limma's lmFit function[57]. Finally, these cell cycle-corrected matrices were merged back into a full PCA matrix, and cell culture batch effects were regressed based on the full data set using the mutual nearest neighbors algorithm[58].

After batch correction, the first 100 principal components were used to build a kNN graph, and this graph was embedded into a two-dimensional space using UMAP. Finally, UMAP coordinates were used to infer a branched pseudotime trajectory using monocle3's learn_graph function. To identify genes that changed as a function of pseudotime, monocle3's graph test was applied to all genes. This test assesses whether cells adjacent in the trajectory show more correlated expression of a gene than cells which are far apart (i.e., autocorrelation). Correction for multiple testing was performed using the $q$ value procedure[59]. A gene was considered as significantly associated with pseudotime if it had a $q$ value $\leq 0.05$ and a Moran's $I$ (a measurement of the magnitude of autocorrelation) larger than 0.05 (ref. [60]).

*Coexpression network analysis.* Coexpression networks were created using the weighted gene coexpression network analysis package (v1.69). For more details, please see the Supplementary Notes.

*Mapping of eQTLs.* For each gene, we calculated mean expression per cluster per donor. To ensure the high-quality eQTL mapping, we only kept genes with non-zero expression in at least 10% of donors and mean count per million higher than one. We retained between 8,940 and 11,516 genes. To identify *cis*-eQTLs, we used tensorQTL (v1.0.3) (ref. [61]) to run a linear regression for each SNP–gene pair, using a 500-kb window within the transcription start site of each gene (i.e., *cis*_nominal mode). We regressed the first 15 gene expression principal components from this analysis so as to capture the confounders within our data set. To correct for the number of association tests performed per gene, we used a *cis* permutation pass per gene with 1,000 permutations. Finally, to correct for the number of genes tested and identify significant eGenes, we performed a *q*-value correction[62] for the top associated SNP–gene pair, setting a *q*-value threshold of 0.1.

*Analysis of eQTL sharing across cell types.* To assess the sharing between eQTLs, we performed a meta-analysis across cell types and cell states using the multivariate adaptive shrinkage (mashR) method[24]. Please see the Supplementary Notes for details.

*Modeling eQTL effect sizes as a function of network centrality.* The effect size of each gene's lead eQTL variant was modeled as a function of the gene's centrality value in the coexpression network described above. This was first done assuming a linear relationship. However, substantial heteroskedasticity was observed, which suggested a nonlinear relationship, as confirmed using a Breusch–Pagan heteroskedasticity test[63]. Thus, we log-transformed the eQTL effect sizes, which resulted in homoskedastic data and a strong linear relationship between the variables. All linear models were built and tested using base R's lm() function.

*Allelic fold-change computation.* To further verify the relationship between a gene's genetic regulation and network centrality, we calculated the allelic fold change according to Mohammadi et al.[29] using publicly available software (https://github.com/secastel/aFC).

*Modeling of dynamic pseudotime-dependent eQTL effects.* To identify pseudotime-dependent eQTL effects, we divided the activation trajectory into ten windows containing roughly equal numbers of cells (i.e., pseudotime deciles) and averaged the expression of each gene per individual within each window. To facilitate the interpretation of coefficients, pseudotime windows were scaled from 0 to 1 before this analysis. To account for the higher correlation in expression values derived from the same individual at multiple pseudotime windows, we applied linear (1) and quadratic (2) mixed models, with individuals modeled as random intercepts. We used these models to test for a significant interaction between genotypes (i.e., the genetic dosage carried by each individual at the lead eQTL variant for that gene) and pseudotime as follows:

$$\begin{aligned} &Z\_\text{score} \sim \text{genotype} + \text{pseudotime} + \text{cell\_culture\_batch} \\ &+\text{sex} + \text{age} + \text{genotype} * \text{pseudotime} + (1|\text{donor}) \end{aligned} \quad (1)$$

$$\begin{aligned} &Z\_\text{score} \sim \text{genotype} + \text{pseudotime} + \text{pseudotime}^2 \\ &+\text{cell\_culture\_batch} + \text{sex} + \text{age} \\ &+\text{genotype} * \text{pseudotime} + \text{genotype} * \text{pseudotime}^2 + (1|\text{donor}) \end{aligned} \quad (2)$$

In both cases, the null model was computed using the same parameters while excluding the genotype*pseudotime and genotype*pseudotime$^2$ terms. *P* values were calculated by comparing each model to its respective null model using analysis of variance. All models were implemented in R using the lmer() function. To reduce the burden imposed by multiple testing, we only applied this approach to variants previously identified as significant lead eQTL variants for a gene by tensorQTL in at least one time point. This was done separately for naive and memory T cells.

To ensure that the method is robust, we permuted the pseudotime windows per donor and tested for an interaction between genotype and pseudotime. A similar permutation has previously been used to test for an interaction effect between a drug and an eQTL[64]. Briefly, as the genotypes remain fixed, this strategy maintains eQTL effects while disrupting the interaction between genotype and pseudotime. By permuting the pseudotime windows 100 times (this generates a random distribution of pseudotime windows), we tested how often a dynamic eQTL would be detected in each permutation. If a test was well calibrated, then one would not expect to observe a large proportion of significant effects in the permuted data. Of the 7,105 and 6,304 significant static gene–SNP pairs from naive and memory T cells, respectively, we observed on average 92 and 90 significant dynamic eQTLs per each permutation round. In contrast, the number of detected dynamic eGenes in our analysis was 1,475 in naive and 1,551 in memory T cells.

**Estimation of pairwise linkage disequilibrium (LD).** We performed LD calculations based on the individual-level genotype information for the individuals in this study obtained from genotyping. Please see the Supplementary Notes for details.

**Integration of eQTLs with GWAS signals.** *Preprocessing of GWAS summary statistics.* Full summary statistics files from previous GWAS studies were downloaded from the GWAS catalogue[65–77]. The GWAS were processed as described in Supplementary Notes.

*Colocalization analysis.* Genomic loci of interest were identified by intersecting eQTL signals in each cell type with GWAS loci for 13 immune-mediated diseases. For each trait–cell type pair, we applied colocalization to any locus where a lead variant for a significant eQTL (*q* value < 0.1) was located within 100 kb and in high LD ($r^2 > 0.5$) with a significant GWAS variant (i.e., any GWAS variant with nominal *P* value < $1 \times 10^{-5}$, which enabled us to capture suggestive association signals). In addition, we required at least 50 variants to be available for testing at each candidate locus. At each of these loci, coloc (v4.0.4) was used to test for colocalization between the eQTL and the GWAS signals. Importantly, these analyses were based on the recently developed masking approach, which relaxes coloc's previous assumption of a single causal variant per locus[34]. This process is similar to performing conditional analyses at each locus. In brief, we defined a 500-kb window centered on the lead eQTL variant and tested for colocalization using all common variants located in the window and present in both the eQTL and the GWAS summary statistics. We used the pairwise LD calculations from our cohort as a basis for masking, setting an $r^2$ threshold of 0.01 to separate independent signals. coloc's prior parameters were set to their recommended values in the most recent publication[34] ($p1 = 1 \times 10^{-4}$, $p2 = 1 \times 10^{-4}$ and $p12 = 5 \times 10^{-6}$). Significant colocalizations were defined as any instances where the estimated posterior probability of a shared causal variant (PP4) was ≥0.8. To discard potential false positives due to noisy association signals, we only kept for further analysis traits with more than one significant colocalization (11 out of 13 traits).

To infer the relationship between gene expression and disease risk at each locus, we estimated the GWAS and eQTL effect sizes (i.e., $\log_e$ of odds ratio and gene expression *Z* score) for the GWAS variant in highest LD with the lead eQTL variant at the locus. We concluded that a variant increased disease risk via an increase in gene expression if the variant had the same direction of effects in both studies. In the opposite case, we concluded that the variant increased disease risk via a decrease in gene expression. If the same variant had different estimates of eQTL effect size in different T cell populations, then we required that all effect sizes had the same direction.

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

## Data availability

The raw scRNA-seq data study have been deposited in the European Genome-phenome Archive with accession number EGAD00001008197. Genotypes have been deposited in the European Genome-phenome Archive with accession number EGAD00010002291. Processed single-cell data and summary statistics are available at https://trynkalab.sanger.ac.uk.

## References


51. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
52. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
53. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
54. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).
55. Kowalczyk, M. S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).
56. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
57. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
58. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
59. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).
60. Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
61. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
62. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci.* **100**, 9440–9445 (2003).
63. Breusch, T. S. & Pagan, A. R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**, 1287 (1979).

64. Davenport, E. E. et al. Discovering in vivo cytokine-eQTL interactions from a lupus clinical trial.*Genome Biol.* **19**, 168 (2018).
65. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
66. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
67. Eyre, S. et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–1340 (2012).
68. International Genetics of Ankylosing Spondylitis Consortium (IGAS). et al. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat. Genet.* **45**, 730–738 (2013).
69. Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50**, 42–53 (2018).
70. Bentham, J. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
71. Onengut-Gumuscu, S. et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
72. International Multiple Sclerosis Genetics Consortium (IMSGC). et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
73. Cordell, H. J. et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019 (2015).
74. Ferreira, M. A. et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).
75. Hinks, A. et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet.* **45**, 664–669 (2013).
76. Tsoi, L. C. et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.* **44**, 1341–1348 (2012).
77. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

## Acknowledgements

This work was funded by Open Targets grant OTAR040 awarded to G.T. This research was funded in whole or in part by the Wellcome Trust (grant WT206194). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. K.C.-G. is supported by a Gates Cambridge Scholarship (OPP1144). We also thank all the donors who participated in this study. We thank the Wellcome Sanger Institute Flow Cytometry facility for their assistance in cell sorting and the Sequencing facility and Cellular Genetics Informatics team for their contribution to data generation and processing. We thank I. Dunham and L. Taylor for critical feedback on the manuscript.

## Author contributions

G.T., B.S. and K.C.-G. conceived and designed the project. B.S., K.C.-G., D.J.S. and K.A. carried out the experimental work. B.S., K.C.-G., Z.K., J.C.M. and A.L. performed the data analysis. G.T., B.S., K.C.-G., Z.K., J.C.M., L.B.-C., J.K., L.R.-N., N.N., J.E.-G., W.R., D.W., D.F.T. and P.G.B. interpreted the results. G.T. supervised the analysis. G.T., B.S., K.C.-G., N.N., J.E.-G., W.R., D.W., D.F.T. and P.G.B. wrote the manuscript.

## Competing interests

All authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01066-3.

**Correspondence and requests for materials** should be addressed to Gosia Trynka.

**Peer review information** *Nature Genetics* thanks Keishi Fujio and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

nature portfolio

Corresponding author(s): Gosia Trynka

Last updated by author(s): Mar 1, 2022

# 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<br>*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<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

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

## Software and code

Policy information about availability of computer code

Data collection: No software was used for data collecti

Data analysis

Imputation was performed using BEAGLE 4.1 with a reference panel consisting of the 1000 Genomes Phase 3 and the UK10K samples. LD calculations were performed on the individual-level genotype using PLINK (v1.90b4).
Raw scRNA-seq data were processed using the Cell Ranger Single-Cell Software Suite (v3.0.0, 10X-Genomics). Reads were first assigned to cells and then aligned to the human genome using STAR v2.5.1b with the hg38 build. Ensembl (v93) was used as a reference for gene annotation. Results from RNA quantification were imported into Python (v3.8.1) and analysed using scanpy (v1.4.4).
To assign cells to their respective individuals, a list of common exonic variants was compiled from the 1000 genomes project phase 3 exome-sequencing data. CellSNP (v0.99) was used to generate pileups at the genomic location of these variants. These pileups, in combination with the variants called from genotyping in each individual, was used as an input for Vireo (v1).
Unsupervised clustering was applied based on kNN graphs using the Leiden algorithm. The similarity of clusters to each other was assessed by performing PCA and estimating the Euclidean distance between pairs of clusters based on the first 100 principal components. Clusters with high levels of similarity or overlapping biological characteristics were merged together.
To perform trajectory inference, raw gene expression measurements were imported into R (v3.6.1) and analysed using monocle3 (v0.2.0).
For co-expression analysis, we used WGCNA package (v1.69). Gene modules were inferred from this dendrogram using R's dynamicTreeCut package (v1.63.1).
All pathway enrichment analyses were performed using gprofiler2 (v 0.1.9), setting the gene list of interest as an unordered query and using all genes detected in the study as the background. Enriched pathways were visualized in R using the pheatmap package (v1.0.12).
To identify cis-eQTLs we used tensorQTL (v1.0.3). To assess the sharing between eQTLs, we performed a meta-analysis across cell types and cell states using the multivariate adaptive shrinkage (mashR) method. For dynamic eQTL mapping, models were implemented in R using the lmer() function.
Coloc (v4.0.4) was used to test for colocalization between the eQTL and the GWAS signals.

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

## Data

Policy information about availability of data

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

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

The raw single-cell RNA-sequencing data study is deposited in the European Genome-Phenome Archive (EGA), accession number EGAD00001008197. Genotypes are deposited in the EGA, accession number EGAD00010002291. Processed single cell data and summary statistics are available at https://trynkalab.sanger.ac.uk. Summary statistics from previously performed GWAS studies for 13 immune-mediated diseases were downloaded from the GWAS catalogue.
To compare the findings from our study with publicly available CD4+ T cell eQTLs we used eQTLs from Chen L. et al. 2016 (PMID: 27863251 ftp://ftp.ebi.ac.uk/pub/databases/blueprint/blueprint_Epivar//qtl_as/QTL_RESULTS/) and Schmiedel, B. J. et al. 2018 (PMID: 30449622 https://dice-database.org/downloads#eqtl_download).

# 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

Sample size was based on results from previous human eQTL studies (e.g. PMID: 29022597, 33664506), where 100-200 samples provide sufficient power to map eQTL.

Data exclusions

Variants derived from imputation were quality filtered using the following parameters: allelic R-squared (AR2) >= 0.8, HWE p-value < 0.001, and MAF > 10%.
Samples with less than 70% of reads mapping to cells were discarded. In addition, any cells with less than 200 detected genes, an unusually high number of genes (defined as over four standard deviations above the mean number of detected genes), or more than 10% of reads mapping to mitochondrial genes were removed from the data set. Any genes detected in less than 10 cells were discarded. Any cells labelled as "unassigned" (less than 0.9 posterior probability of belonging to any individual) or "doublets" (containing mixed genotypes) by Vireo were discarded. Finally, cell contaminations were removed from the data set, resulting in 655,349 (91.86% of total) high quality cells kept and successfully annotated as CD4+ T cells.
In co-expression analysis, only genes with ≥ 1 TPM in at least 30 samples were used. In addition, genes were filtered by their level of variability, with only genes showing a standard deviation ≥ 0.1 across samples being kept. This resulted in 11,130 genes taken forward for network construction.
To ensure the high quality eQTL mapping, we kept genes with non-zero expression in at least 10% of donors and mean count per million (cmp) higher than one. We retained between 8,940 and 11,516 genes. We also removed related individuals.
In GWAS studies, any signals coming from the X or Y chromosomes, as well as from the MHC region (ch6:28,510,120 – chr6:33,480,577) were discarded.
For each trait-cell type pair, we applied colocalization to any locus where a lead variant for a significant eQTL (q value < 0.1) was located

within 100 kb and in high LD (r2 > 0.5) with a significant GWAS variant. In addition, we required at least 50 variants to be available for testing at each candidate locus.

Replication

We used previously reported results from bulk RNAseq data and replicated markers of clusters. We also replicated that T cells form a continuum (Cano-Gamez et al. 2020; Kiner et al. 2021) and demonstrated that this continuum can be taken into account when mapping response eQTLs (as previously shown in different cell types by Strober et al. 2019; Cuomo et al. 2020). There is no single cell RNA-seq data available that would be suitable for replication of activation eQTLs.
No other experiments were performed in this study.

Randomization

Donors were allocated to experimental batches at random.

Blinding

No blinding was applied, as all samples were processed by the same operators.

# 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 |

## Human research participants

Policy information about studies involving human research participants

Population characteristics

Blood samples were obtained from 119 healthy individuals of British ancestry. Of these, 67 were male (53.7%) and 52 female (56.3%), and the mean age of the cohort was 47 years (sd = 15.61 years).

Recruitment

Participants were recruited from GSK blood resource at Addenbrooke's Hospital. Blood was processed within 3h of collection. To our knowledge there was no selection bias.

Ethics oversight

Human biological samples were sourced ethically and their research use was in accord with the terms of informed consent under an IRB/EC approved protocol (15/NW/0282). Ethics was approved by Wellcome Sanger Institute ethicks committee.

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