Impact of Rare and Common Genetic Variation on Cell Type-Specific Gene Expression

Anna S.E. Cuomo[1,2,3,4,5], Eleanor Spenceley[1,4,5], Hope A. Tanudisastro[2,3,4,6], Blake Bowen[1,4,5], Albert Henry[1,4,5,7], Hao Lawrence Huang[1,4,5], Angli Xue[1,4,5], Wei Zhou[8,9], Matthew J. Welland[2,3,4], Arthur S. Lee[8,9], Kristof Wing[10,11,12], Owen Tang[6,13,14], Michael P. Gray[6,13,14], Michael Franklin[2,3,4], Michael Harper[2,3,4], Michael Silk[2,3,4], Katalina Bobowik[2,3,4], Alexander Stuckey[2,3,4], John Marshall[2,3,4], Vivian Bakiris[2,3,4], Caitlin Uren[2,3,4], Bindu Swapna Madala[2,3,4], Amy Miniter[2,3,4], Caitlin Bartie[1,4,5], Drew R. Neavin[1], Zhen Qiao[4,15], Eyal Ben-David[16], Ling Chen[16], Kyle Kai-How Farh[16], Stuart M. Grieve[13,14], Tung Nguyen[6,13,14], Jennifer Piscionere[2,3,4], Owen M. Siggs[4,15], Hannah Nicholas[2,3,4], Katrina M. de Lange[2,3,4], Alex H. Hewitt[10,11,12], Gemma A. Figtree[6,13,14], Daniel G. MacArthur[2,3,4,*], Joseph E. Powell[1,4,5,*]

[1]Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia
[2]Centre for Population Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia
[3]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, VIC, Australia
[4]Faculty of Medicine and Health, University of New South Wales, Sydney, NSW, Australia
[5]UNSW Cellular Genomics Futures Institute, University of New South Wales, Sydney, NSW, Australia
[6]Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia
[7]Institute of Cardiovascular Science, University College London, London, UK
[8]Massachusetts General Hospital, Boston, MA, USA
[9]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[10]Menzies Institute for Medical Research, University of Tasmania, Hobart, TAS, Australia
[11]Department of Ophthalmology, Royal Hobart Hospital, Hobart, TAS, Australia
[12]Centre for Eye Research Australia, University of Melbourne, East Melbourne, VIC, Australia
[13]Charles Perkins Centre, The University of Sydney, Sydney, NSW, Australia
[14]Kolling Institute of Medical Research, Royal North Shore Hospital, Sydney, NSW, Australia
[15]Genomics and Inherited Diseases Program, Garvan Institute of Medical Research, Sydney, NSW, Australia
[16]Illumina Artificial Intelligence Laboratory, Illumina, San Diego, CA, USA

[*]co-senior authors

Abstract

Understanding the genetic basis of gene expression can shed light on the regulatory mechanisms underlying complex traits and diseases. Single-cell resolved measures of RNA levels and single-cell expression quantitative trait loci (sc-eQTLs) have revealed genetic regulation that drives sub-tissue cell states and types across diverse human tissues. Here, we describe the first phase of TenK10K, the largest-to-date dataset of matched whole-genome sequencing (WGS) and single-cell RNA-sequencing (scRNA-seq). We leverage scRNA-seq data from over 5 million cells across 28 immune cell types and matched WGS from 1,925 individuals. This provides power to detect associations between rare and low-frequency genetic variants that have largely been uncharacterised in their impact on cell-specific gene expression. We map the effects of both common and rare variants in a cell type specific manner using SAIGE-QTL. This newly developed method increases power by modelling single cells directly using a Poisson model rather than relying on aggregated 'pseudobulk' counts. We identify putative common regulatory variants for 83% of all 21,404 genes tested and cumulative rare variant signals for 47% of genes. We explore how genetic effects vary across cell type and state spectra, develop a framework to determine the degree to which sc-eQTLs are cell type specific, and show that about half of the effects are observed only in one or a few cell types. By integrating our results with functional annotations and disease information, we further characterise the likely molecular modes of action for many disease-associated variants. Finally, we explore the effects of genetic variants on gene expression across different cell states and functions, as well as effects that directly vary cell state abundance.

Introduction

Genome-wide association studies (GWASs) have revealed hundreds of thousands of association signals between genetic variants, human diseases, and phenotypic traits[1]. Yet, the modes of action by which most of these variants affect human disease and phenotypes remain to be characterized, especially for variants outside protein-coding regions. The characterisation of the impact of genetic variants on gene expression - known as expression quantitative trait locus (eQTL) mapping - has already uncovered molecular mechanisms for many of these variants by linking them to the putative genes that they regulate[2–4].

eQTL mapping can help identify the genes and pathways involved in disease pathogenesis, which is a critical step in identifying opportunities for therapeutic intervention. Yet, important limitations remain. First, eQTL mapping has traditionally been limited to common genetic variation (population minor allele frequency [MAF] > 5%) due to relatively small sample sizes and the widespread use of genotyping arrays, which capture variants in linkage disequilibrium (LD) with known common genetic variants, but do not accurately genotype most rare genetic variants and a substantial fraction of complex structural variants[5]. Generating a more complete picture of the impact of genetic variation on human molecular phenotypes requires using DNA sequencing approaches.

Second, eQTLs are extremely context-specific, with multiple studies demonstrating that genetic variants frequently impact gene expression levels in ways that can only be detected in specific tissues[3], sorted cell types[6], or upon stimulation[7]. The recent introduction of single cell-resolved measures of RNA levels allows expression phenotypes to be measured in these more granular cell types and states, and single-cell eQTL mapping has demonstrated that the effects of many genetic variants – including disease-associated ones – on gene expression can only be detected when mapping eQTLs within specific cell types and even transient cellular states[8–12].

To overcome these challenges, we introduce phase 1 of the TenK10K (up to 10,000 cells for 10,000 individuals) study. TenK10K phase 1 encompasses 5,920,025 cells from 2,103 individuals, from which we selected 5.4 million cells from 1,925 individuals of European ancestry to deeply explore the impact of rare and common genetic variation on gene expression phenotypes. In accompanying papers, we leverage the same dataset to investigate the impact of tandem repeat variation on gene expression [Tanudisastro *et al.* 2025] and the causal links between gene expression measurements and human diseases and traits [Senabouth *et al.* 2025], and an overlapping set of samples to explore the impact of genetic variation on chromatin accessibility [Xue *et al.* 2025].

Here we use a recently developed association method, SAIGE-QTL[13], to map single-cell expression quantitative trait loci across cell types for both common and rare variants called from whole genome sequencing (WGS) data and identify over 150,000 common variants and over 30,000 rare variant effects on gene expression, including common or rare variant signals for nearly 5,000 genes (~25%) with no previously reported eQTLs in the largest blood eQTL study to date[2]. We propose a framework for quantifying the extent of cell type specificity vs sharedness of reported eQTLs. We show that about half of the effects are found in only one or a few cell types, and almost none are ubiquitously found across all. We perform statistical colocalization between our eQTLs and 58 disease phenotypes and blood traits and identify over 30,000 colocalization events for over 2,000 genes, providing novel potential molecular mechanisms in almost 400 genes associated with human phenotypic variation. We use a novel biologically-driven deep learning method, scDeepID, to identify continuous cell states related to cell

function and map over 6,000 eQTLs that vary dynamically along those states. Finally, we identify 165 independent genetic variants that regulate the abundance of cell states.

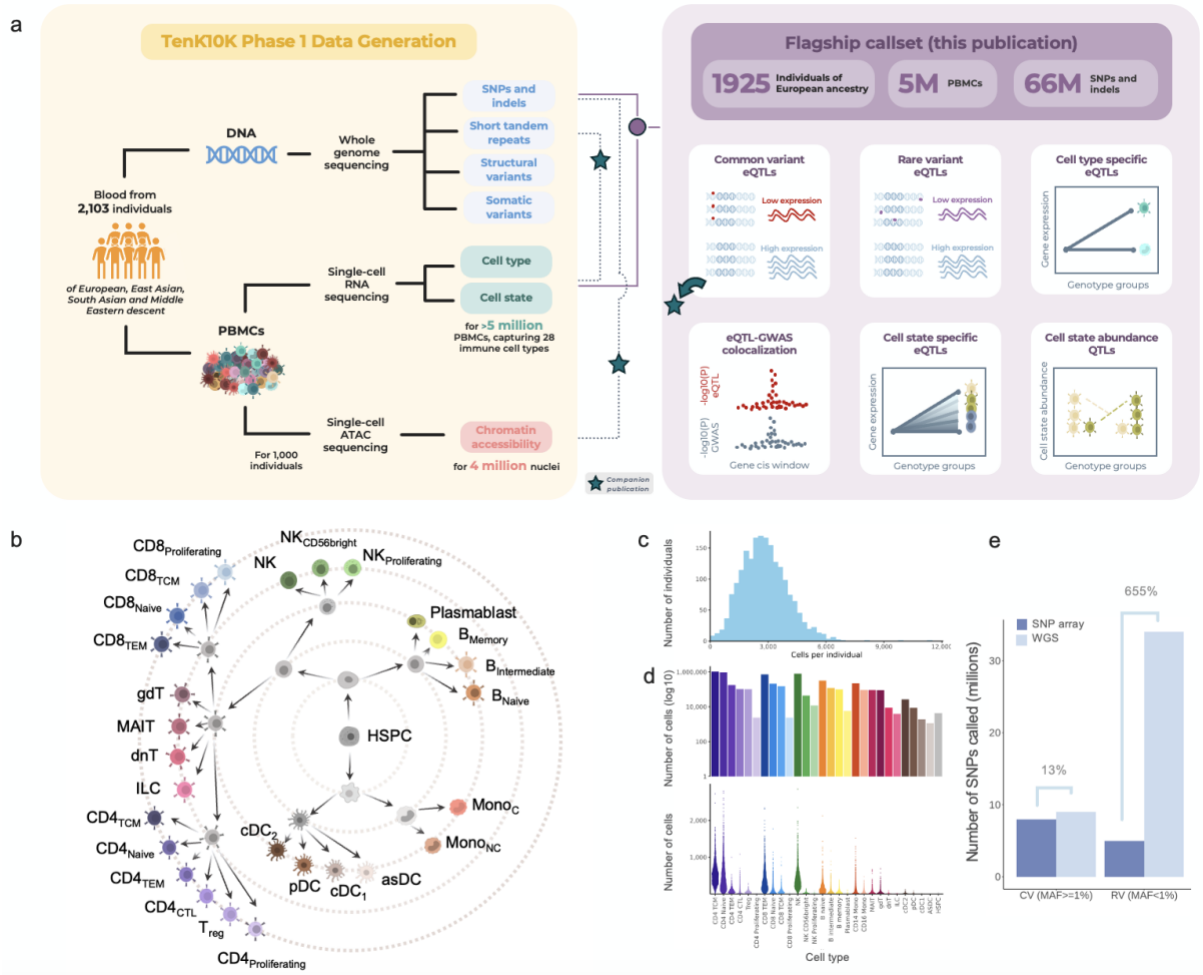


Figure 1. Population-scale single-cell RNA sequencing and matched whole-genome sequencing from 1,925 individuals identified 28 immune cell types and 66 million variants. a) An overview of the overall cohort and the specific data used in this paper. High-quality whole-genome sequencing (WGS) and single-cell RNA sequencing (scRNA-seq) data were generated for 2,103 individuals (after QC), and single-cell ATAC-seq data was additionally generated for a 1,000 individual subset. In this paper, we consider common and rare SNPs and indels called from WGS and matched blood scRNA-seq from 1,925 individuals of European ancestry. After computational analysis and processing, we conducted both common and rare variant eQTL analysis and identified cell type and state specific effects of genetic regulation in immune cells. Accompanying analyses of these data described in companion manuscripts are highlighted with a green star. **b)** Schematic of the 28 immune cell types identified, organised by cell lineage. The grey cells represent important steps in the lineage but are not cell types in the study (except HSPC). **c)** Distribution of the number of cells per donor after scRNA-seq, demultiplexing, and quality-control filtering (mean=2,825, range=124-11,348). **d)** The number of cells available for analysis by cell type for the 28 immune cell types is summarised in d. Top: bar plot displaying total cell numbers per cell type on a log 10 scale, bottom: violin plots displaying the distribution across individuals. **e)** For a subset of 960 individuals, SNP array data and WGS were available. Bar plots representing the number of biallelic SNPs called using the two protocols, split by frequency groups. CV: common variant (minor allele frequency; MAF≥1%), RV: rare variant (MAF<1%).

Results

High throughput blood scRNA-seq and matched WGS data of 1,925 individuals

After quality control, a total of 5,438,679 peripheral mononuclear blood cells (PBMCs) from 1,925 donors were considered for analysis (**Fig. 1, Supplementary Table 1; Methods**). We collected and sequenced PBMCs from all individuals using the high-throughput 3' sequencing kit from 10x Genomics, achieving 2,825 cells on average per individual after quality control (**Fig. 1d**, **Supplementary Fig. 1**; **Methods**).

Each cell was assigned to their donor of origin via demultiplexing and to an immune cell type using a combined approach of reference-based classification and hierarchical clustering as recommended by the single-cell eQTLGen Consortium[14] (**Methods**). Briefly, we applied a supervised classification using a reference panel of protein markers from Hao *et al*[15] to assign cells to distinct lineages, including myeloid and lymphoid, with the latter further classified into B, natural killer (NK), and CD4+ and CD8+ T cells, and the former into monocytes and dendritic cells. Next, we used unsupervised clustering to further split cells into sub-cell types, as depicted in **Fig. 1c**. Each cell was assigned to one of 28 different cell types, spanning the key expected immune lineages (**Fig. 1, Supplementary Table 2**).

Additionally, we generated WGS data for all 1,925 individuals, which was processed and quality-controlled using established pipelines (**Methods**). We called a total of 66,658,242 sites between single nucleotide polymorphisms (SNPs) and indels, of which 53,784,018 (80.6%) were rare (MAF<1%), approximately 4.5 times as many as called from SNP array data (34.4m vs 4.6m when considering SNPs only for 960 individuals with matched data; **Fig. 1B; Methods**); and 4,317,810 (6.5%) were low-frequency (1%≤MAF<5%).

Single-cell eQTL mapping reveals common variant impact on gene expression

To assess the effects of common genetic variation on gene expression changes, we mapped *cis* eQTLs, testing for the effects of genotypes at common (MAF>1%) variant sites within a window of +/-100kb around the gene body on the single-cell expression profile of the target gene (**Methods**). We used SAIGE-QTL[13] to map eQTL effects using single-cell profiles, modelled as Poisson counts, rather than relying on 'pseudobulk' aggregation. We mapped eQTLs separately in each of the 28 cell types and confirmed our Type I error was controlled using permutations (**Fig. 2a**). At FDR<5% after multiple testing correction, we identified a total of 154,932 eQTLs in at least one of the 28 cell types from 17,674 unique eGenes (83% of all genes tested; **Fig. 2b, f** and **Supplementary Fig. 2**). The number of eQTLs identified varied substantially across cell types, largely due to the number of cells available for each cell type affecting statistical power, as previously shown[13,16] (**Supplementary Fig. 3**).

Similarly to what has been shown previously[2], eQTL variants were concentrated around the transcription start and end sites (TSS and TES, respectively) and declined in number proportionally to the distance from the gene body, likely indicating their role in genetic regulation (**Supplementary Fig. 4a**). Indeed, we found that 83% of all 154,932 eQTL variants were located within 50kb of either side of the gene, and 94% within +/- 80kb. Furthermore, they were enriched in promoter and enhancer-like genomic regions and regions estimated to be open chromatin (**Supplementary Fig. 4b**).

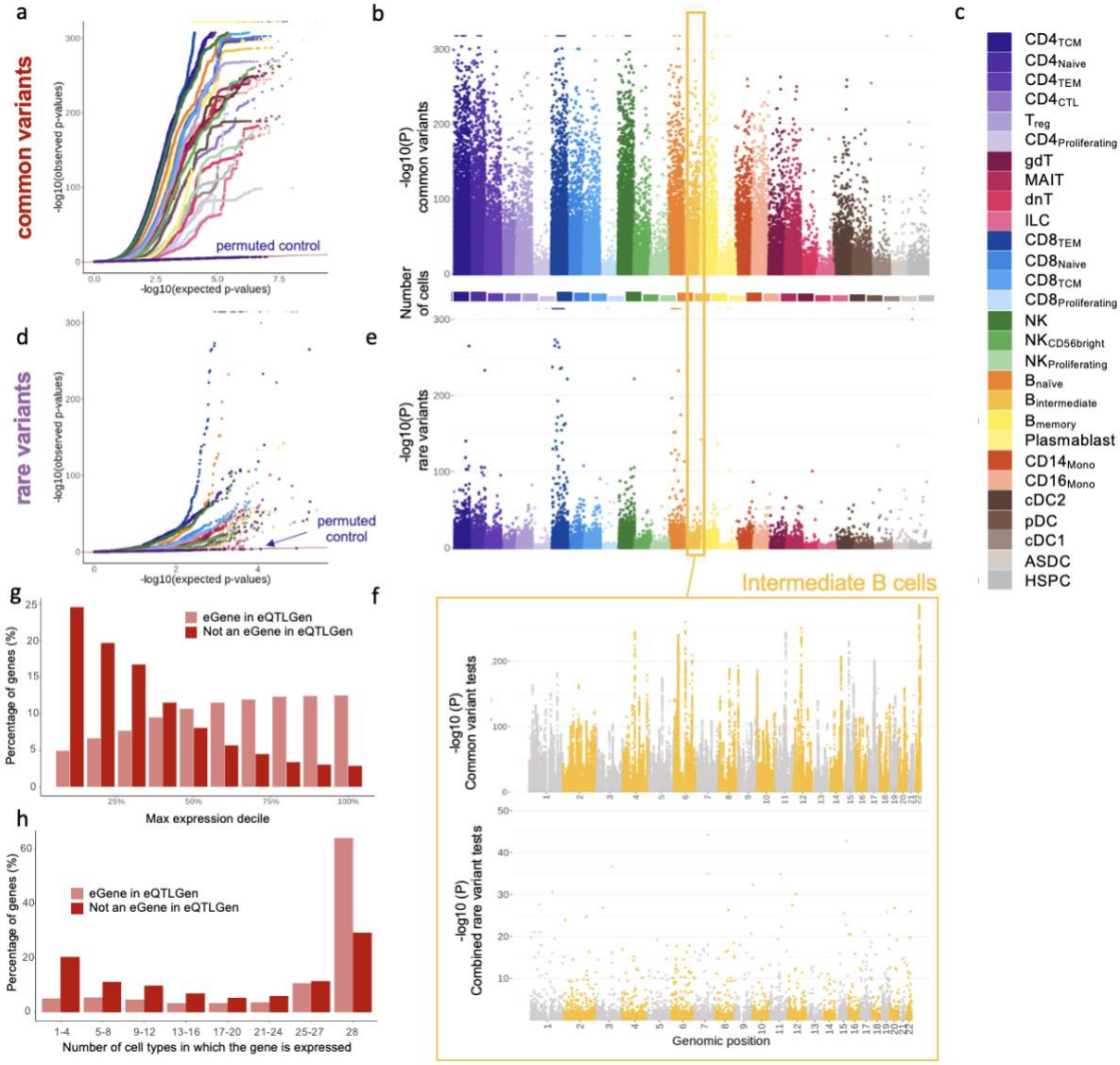


Figure 2. Common and rare eQTL mapping in 28 immune cell types. a) Quantile-quantile (QQ) plot for all common variants (MAF>1%) eQTL results across all 28 immune cell types. For the most abundant cell type, $CD4_{TCM}$, we shuffled individual IDs and re-performed the test to assess calibration (labelled 'permuted control'). Color represents cell type, as described in panel c. **b)** Cell type-specific Manhattan plot. We display the negative log10 gene-level Cauchy corrected p-values for each cell type for each gene tested. Below, the number of cells available for each cell type are highlighted. **c)** Legend, mapping colors to the 28 cell types. **d)** and **e)** are equivalent to a) and b) but considering rare variants (MAF≤1%). The p-values in c and d are gene-level combined p-values of the burden and SKAT tests. **f)** As an example, zoom in from the plots in b) and d) to display the full Manhattan plot split by chromosomes for one representative cell type, $B_{Intermediate}$ (Manhattan plots for all other cell types are included in Supplementary Figures 2 and 7 for common and rare variants, respectively). **g)** Common variant eGenes from this study that are found or not found to be eGenes in the eQTLGen study[2], stratified by max expression across cell types. eGenes novel to this study are more lowly expressed. **h)** Common variant eGenes from this study that are found or not found to be eGenes in the eQTLGen study, stratified by the number of cell types in which they are sufficiently expressed (>1% of all cells). eGenes novel to this study are expressed in a more cell type specific manner.

We compared the eGenes we identified with those identified by the whole-blood eQTLGen consortium study[2] and the OneK1K study[16], as the currently largest eQTL studies in blood using bulk and single-cell RNA-sequencing data, respectively. We identified a comparable number of eGenes compared to eQTLGen (17,674 vs 16,987) despite the 15-fold lower sample size (1,925 vs over 30,000 total

individuals). Over 1 in 4 of our significant eGenes were not identified in eQTLGen (n=4,572, 25.9%); those were genes more lowly expressed and with more cell type-specific expression, demonstrating the importance of mapping eQTLs at cellular resolution (**Fig. 2g, 2h**). We identified nearly three times more eGenes compared to OneK1K (17,674 vs 6,469), demonstrating the increased power from both larger sample size (about twice the number of individuals: 1,925 vs 982) and number of cells per individual (2,825 vs 1,291) and the consequent more granular cell type resolution (28 vs 14 cell types). The eGenes found in our study and not OneK1K were more lowly expressed and more likely to be expressed in fewer cell type populations (**Supplementary Fig. 5**).

We correlated our identified effect sizes and scores from PromoterAI (a convolutional neural network that predicts promoter activity of variants, their target gene and their effect on gene expression; **Methods**) for the top variant for each gene across the different cell types, finding good agreement (median correlation=0.66; **Supplementary Fig. 6**), demonstrating concordance of both the magnitude and direction of effects.

Mapping the effects of rare variation on single-cell gene expression

Most bulk and single-cell eQTL studies to date[3,10,16–18] have focused on common genetic variants (population minor allele frequency; MAF≥ 5%) due to their relatively small sample size and the fact that they rely on genotype arrays and imputation. Studies investigating the effects of rare variants (MAF < 1%) on tissue-level gene expression[4,19] have conducted "outlier analyses", where rare variants are explicitly sought in individuals with extreme expression levels for a given gene rather than performing systematic assessments of the effects of *cis* rare variants on expression levels genome-wide across individuals.

By combining whole-genome sequencing and single-cell data at population scale, TenK10K phase 1 is uniquely positioned to assess the effects of rare genetic variation on cell type-specific expression genome-wide. Using SAIGE-QTL, we mapped the effects of rare variation (MAF<1%) within a *cis* window (+/-100kb) around each gene's body, using gene-set based tests for increased power (**Methods**). For each cell type, we tested for the effects of local rare variants on all sufficiently expressed genes (genes expressed in at least 1% of all cells from the cell type of interest) using two different types of tests designed to capture different regulatory mechanisms: the burden test, and the sequence kernel association test (SKAT)[20].

Overall, we observe 34,376 rare variant signals across all cell types for 10,005 unique "rare eGenes" (47% of all genes tested) after FDR correction on the combined p-values from the two tests (burden test and SKAT; FDR<5%, **Fig. 2d, f, Supplementary Fig. 7; Methods**). Type I error was well controlled for the rare variant tests (**Fig. 2c),** and the number of cells for each cell type was highly correlated with the number of eGenes identified (**Supplementary Fig. 3**).

When comparing eGenes with a rare variant signal to those with at least one common variant eQTL, we found that "rare eGenes" were more ubiquitously expressed across all cell types, whereas "common eGenes" were expressed in a more cell type specific manner (**Supplementary Fig. 8a**). Single-variant test results show the expected relationship between allele frequency and effect sizes with rarer cell types displaying larger effects on gene expression (**Supplementary Fig. 8b**). Both common and rare eGenes were enriched for putative promoter activity as estimated by PromoterAI (**Methods**), with on average a larger number of variants with promoter-like activity (on the specific gene) compared to non-eGenes (23.14 vs 19.12; 23.45 vs 20.83 for common and rare eGenes respectively; **Supplementary Fig. 8c**).

Rare variant effects on cell type specific gene expression reveal novel regulation

While common variant effects on gene expression have been well characterised, rare variants remain less studied, especially in the non-coding genome. To further characterise our rare variant signals, we overlapped our signals with a combined set of functional annotations from publicly available to characterise genomic regions as putative enhancers, promoters, TAD boundaries, as well open chromatin regions using scATAC-seq for a subset of our samples (**Methods**).

As an example, we identified the significant downregulation of *SH2D2A* in $CD4_{TCM}$ (Cauchy combined p-value=1.48×10^{-2}) with the effects driven by several rare variants located in functional regions, including three 5' UTR variants, one 3' UTR variant, three putative enhancer variants (almost certainly on a single haplotype, given all are found in the same individuals) and one intron variant (**Fig. 3a**). None of these variants has been previously reported to affect the expression of *SH2D2A*, likely due to their low frequency (MAF< 0.005), highlighting the value of whole genome sequencing. The top 5' UTR variant (1:156816831:A:AC) is an indel which additionally overlapped with an open chromatin peak in the 5' UTR region of *SH2D2A* using scATAC-seq data generated for a subset of the individuals from our cohort (**Fig. 3c, Supplementary Fig. 9a; Methods**). The signal became more significant upon conditioning on a common variant signal for the same gene in the opposite direction of effect (p-value=2.6×10^{-4} after conditioning on the top common variant). *SH2D2A* encodes the T cell specific adaptor protein (TSAd) that induces cytokine production in CD4+ T cells specifically and has been linked to autoimmune conditions, including hypothyroidism[21–23].

We also found that the expression of *PRDM4* was downregulated by the combined effects of multiple rare variants in most T cell subpopulations, most significantly in CD4+ naive T cells ($CD4_{Naïve}$; p-value=1.2×10^{-17}). While one of these variants (12:107695720:A:G) was previously reported (in whole blood, with a consistent direction of effect[2]), we report two novel rare variants here for the first time. We hypothesise that one variant (12:107761146:T:G; MAF=0.0099) in high LD with 12:107695720:A:G is likely causal, as it overlaps a CTCF binding site and likely promoter region according to multiple sources (Encode[24–26], chromHMM[27], UCSF TF-ChIP data[28]), had high activity-by-contact (ABC[29,30]) score and was predicted to be in a promoter region and to result in expression downregulation (PromoterAI score=-0.77; **Fig. 3b,d, Supplementary Fig. 9b**). *PRDM4* encodes a transcription factor that mediates cell proliferation and has been linked to several cancers[31].

Additionally, we found that rare genetic variants regulate the expression levels of *SLC26A3* in $CD4_{Naïve}$ cells. *SLC26A3* encodes an ion transporter, which is downregulated in adenoma (DRA) and is involved in mucus formation. Variants regulating *SLC26A3* have been associated with ulcerative colitis (UC) and inflammatory bowel disease (IBD)[32], and rare variants in the *SLC26A3* gene body have been associated with a relevant rare disease (congenital chloride diarrhea)[33]. The most significant variant is a missense variant within *SLC26A3*, while several other variants are located in open chromatin regions (**Supplementary Fig. 10**). *SLC26A3* was one of only 88 genes with a significant effect driven by rare variation and no common variant signal.

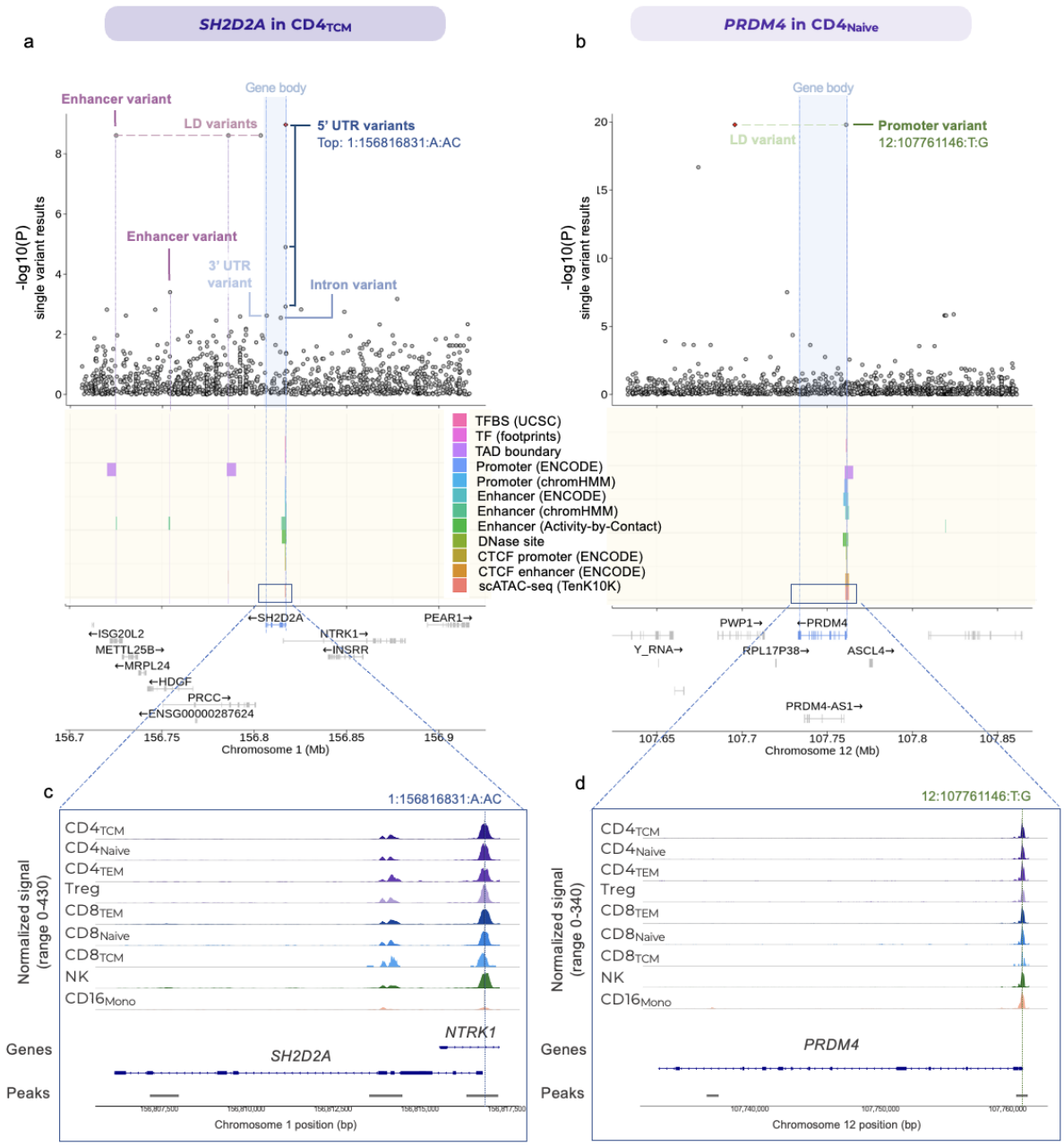
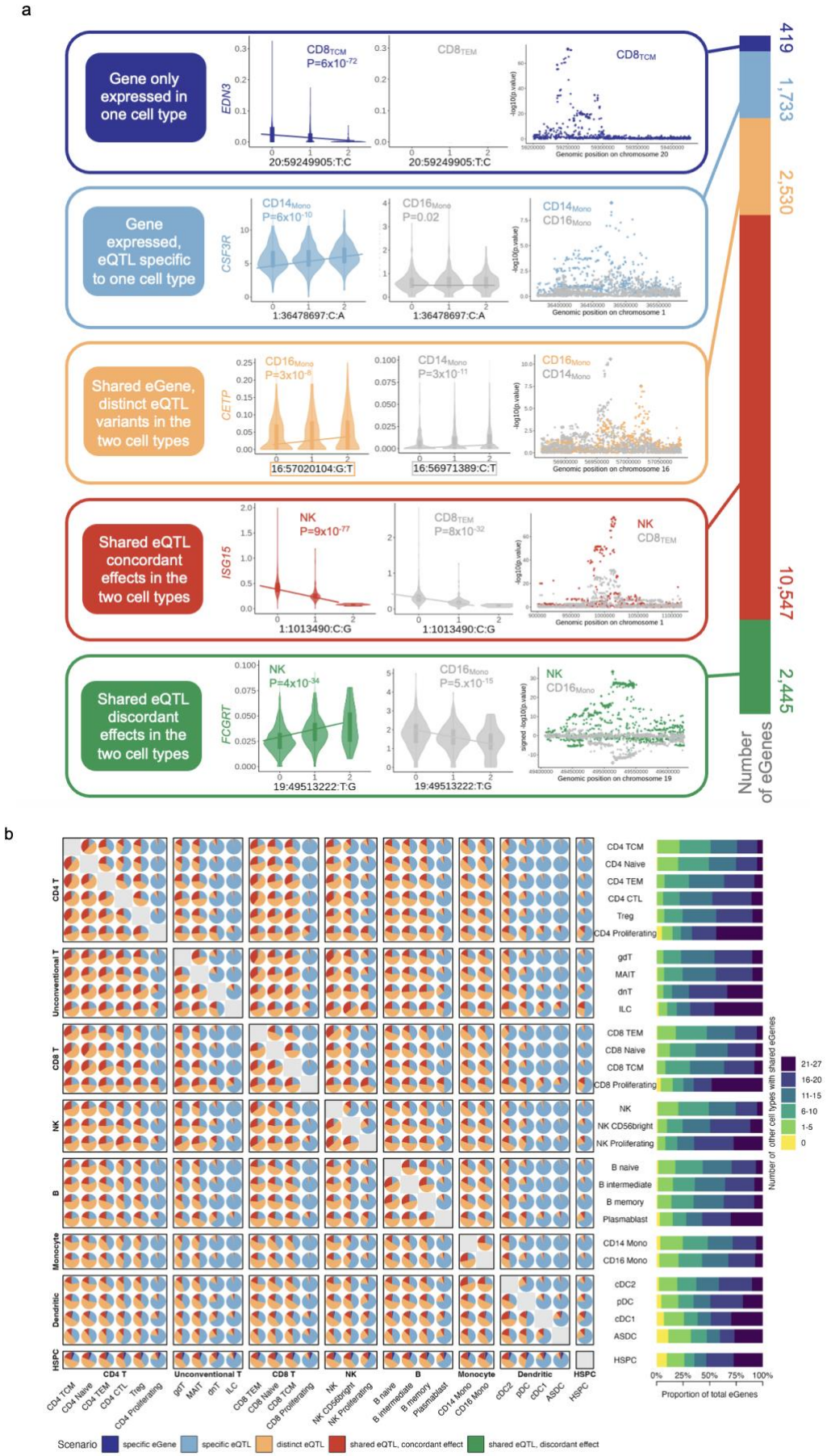


Figure 3. Rare variation shapes cell type specific gene expression in blood. a) Locus zoom plot of rare genetic variants associated with the *SH2D2A* gene on chromosome 1 in $CD4_{TCM}$, with the leading genetic variants and overlapping genomic regions annotated. "LD variants" are variants in perfect LD with the annotated putatively causal ones. **b)** Locus zoom plot of rare genetic variants associated with the *PRDM4* gene on chromosome 12 in $CD4_{Naive}$. **c)** Zoom in on the chromatin accessibility profile across selected cell types around the *SH2D2A* gene and with the top variant highlighted in **a** (1:156816831:A:AC) shown as a vertical line. **d)** Similar to **c** but for *PRDM4* and 12:107761146:T:G. Peaks for all 28 cell types are included for both genes in Supplementary Fig. 9. LD: linkage disequilibrium; TF: transcription factor, TFBS: transcription factor binding site; UTR: untranslated region.


a
Gene only expressed in one cell type
EDN3
CD8TCM P=6x10-72
CD8TEM
CD8TCM
20:59249905:T:C
20:59249905:T:C
Genomic position on chromosome 20
Gene expressed, eQTL specific to one cell type
CSF3R
CD14Mono P=6x10-10
CD16Mono P=0.02
CD14Mono CD16Mono
1:36478697:C:A
1:36478697:C:A
Genomic position on chromosome 1
Shared eGene, distinct eQTL variants in the two cell types
GETP
CD16Mono P=3x10-6
CD14Mono P=3x10-11
CD16Mono CD14Mono
16:57020104:G:T
16:56971389:C:T
Genomic position on chromosome 16
Shared eQTL concordant effects in the two cell types
ISG15
NK P=9x10-77
CD8TEM P=8x10-32
NK CD8TEM
1:1013490:C:G
1:1013490:C:G
Genomic position on chromosome 1
Shared eQTL discordant effects in the two cell types
FCGRT
NK P=4x10-34
CD16Mono P=5.x10-15
NK CD16Mono
19:49513222:T:G
19:49513222:T:G
Genomic position on chromosome 19
419
1,733
2,530
10,547
2,445
Number of eGenes
b
CD4 T
Unconventional T
CD8 T
NK
B
Monocyte
Dendritic
HSPC
CD4 TCM
CD4 Naive
CD4 TEM
CD4 CTL
Treg
CD4 Proliferating
gdT
MAIT
dnT
ILC
CD8 TEM
CD8 Naive
CD8 TCM
CD8 Proliferating
NK
NK CD56bright
NK Proliferating
B naive
B intermediate
B memory
Plasmablast
CD14 Mono
CD16 Mono
cDC2
pDC
cDC1
ASDC
HSPC
Number of other cell types with shared eGenes
21-27
16-20
11-15
6-10
1-5
0
0%
25%
50%
75%
100%
Proportion of total eGenes
Scenario
specific eGene
specific eQTL
distinct eQTL
shared eQTL, concordant effect
shared eQTL, discordant effect

280

Figure 4. Cell type specificity analysis. a) Examples of common variant eQTLs illustrating the five scenarios of cell type specificity between two cell types (cell type A in each scenario's color, cell type B in grey). For each scenario, from left to right: a brief description of the scenario, box plots for the lead SNPs in each of the two cell types, locus zoom plots of all tested variants around the gene, with signal for both cell types overlaid together. For scenario 5, the locus zoom plot displays the -log10(P) as in the others, but this time weighted by the sign of the effect beta to emphasise the opposite direction of the effect. The bar chart at the end summarises each scenario at the level of eGenes, using the maximum from all pairwise cell type comparisons. **b)** For each pair of cell types, pie charts counting the proportion of eQTLs in each scenario represented by color as in a. Bar plots to the right contain the proportion of eGenes binned by the number of cell types the effect is shared (scenarios 3 to 5) with. Cell types are grouped by major cell type categories.

Quantifying the cell type specificity of immune cell eQTLs

An important question with single-cell eQTL analysis is the degree to which the loci exhibit genetic effects in specific cell types or subsets; in other words, how 'cell type-specific' are they? For several reasons, this is a complex question to address directly: genes are variably expressed across cell types, which can lead to true eQTLs being missed for lowly expressed genes; differences in cell numbers lead to variable power to detect eQTLs in a given cell type, making like for like comparison between cell types complex; and linkage disequilibrium (LD) between loci means we cannot always directly infer if alternative top eQTL variants for the same gene in different cells are tagging the same or different causal loci.

We defined five possible scenarios that can occur for a given eQTL identified at FDR<5% in one cell type (cell type A) for any of the pairwise comparisons with another cell type (cell type B; **Figure 4a**). In *scenario 1 (specific eGene)*, the eGene corresponding to the eQTL is not sufficiently expressed (<1% cells) in cell type B. In *scenario 2 (specific eQTL)*, the eGene is sufficiently expressed in cell type B, but no eQTL is identified. *Scenario 3 (distinct eQTL)* represents a case where two eQTL variants in low LD (r^2 <0.5) for the same eGene are identified in cell types A and B. Conversely, we define a case where the same eQTL or eQTL in LD ($r^2 \geq 0.5$) for the same eGene are found in cell types A and B as *scenario 4 (shared eQTL, concordant effect)* when the direction of allelic effect per additional minor allele on gene expression is concordant, or *scenario 5 (shared eQTL, discordant effect)* when the direction is discordant. Noting that an eQTL in *scenario 2* may have an underpowered association signal in another cell type that is rarer than the primary cell type, we performed a pairwise random-effect meta-analysis and reclassified eQTL with a meta-analysis p-value 10^{-5} lower than the p-value in the primary cell type as scenario 4 or 5 depending on the concordance of effect direction (**Methods**).

We assigned each scenario to each pair of the primary cell type – 27 other cell types for each of the identified 154,932 common variant eQTLs (**Supplementary Table 3**). Next, we aggregated the eQTL scenarios for each pairwise comparison into a single score for 17,674 common variant eGenes by taking the *maximum* scenario value across 27 other eQTLs under contrast (**Figure 4a**, bar plot; **Methods**).

In the overall eGene set aggregation, we found 419 of 17,674 (2.4%) eGenes with a maximum score of 1, representing specific expression in only one cell type. For example, *EDN3* is only expressed in CD8 positive central memory T cells ($CD8_{TCM}$) and has an eQTL at 20:59249905:T:C (rs259962), such that more copies of the C allele results in higher expression in $CD8_{TCM}$ (**Fig. 4a**). A higher percentage of eGenes (805 out of 17,674, 4.6%) were expressed in at most two cell types, 7% in at most three cell types, and 15% in at most 7 (1/4 of all cell types). Just over half (9,723 out of 17,674, 55%) of all eGenes were sufficiently expressed in all 28 cell types (**Supplementary Fig. 11**).

Amongst eGenes that are sufficiently expressed in two or more cell types, 1,733 (9.8% of all eGenes) had a maximum score of 2, indicating the presence of an eQTL in only one of the 28 cell types. For example, rs3917929 (1:36478697) is an eQTL for *CSF3R* only in classical monocytes ($CD14_{Mono;}$ **Fig. 4a**). About half (8,506) of eGenes were significant in at most a quarter of all cell types (7/28), with only 35 eGenes (0.2%) found to have a significant eQTL in all 28 cell types (**Supplementary Fig. 12**).

A maximum overall score of 3, representing the presence of at least two distinct eQTLs (LD $r^2 < 0.5$) for the same eGene in multiple cell types, was observed for 2,530 (~14%) eGenes. For example, *CETP* is upregulated by different variants in classical (16:57020104:G:T, rs2052880) and non-classical (16:56971389:C:T, rs1532625) monocytes (**Fig. 4a**). The remaining eGenes showed evidence of a shared eQTL (same variant or two different variants with LD $r^2 \geq 0.5$) in two or more cell types, where 10,547 (~60%) eGenes showed concordant direction of allelic effect on gene expression and 2,445 (~14%) showed discordance. For example, the *FCGRT* gene was upregulated by 19:49513222:T:G (rs111981233) in NK cells but downregulated in non-classical monocytes ($CD16_{Mono;}$ **Fig. 4a**).

As expected, eQTL sharedness between cell types was grouped by cell lineage and broad cell type categories (**Supplementary Fig. 13**). Sharedness within both lymphoid and myeloid lineages and major cell types (T, NK, B, Monocyte, Dendritic) was higher than outside of the lineage and major cell type, with the highest sharing in Monocytes (average correlation 0.72), and the lowest in dendritic cells (average correlation 0.49; **Supplementary Fig. 14a**). Excluding rare cell types (number of cells < 6,000; **Methods**), which are underpowered and skew the distributions, we found higher correlations for B and T cells (average correlations 0.82 and 0.75 respectively, **Supplementary Fig. 14b**).

An equivalent analysis on rare variants identified 2,560 cell type-specific effects (26%), either cell type-specific expression (n=75, ~1%) or true cell type-specific regulation (n=2,485, 25%) and 7,445 shared (or likely shared; **Methods**) effects across cell types (74%; **Fig. 4, Supplementary Fig. 15; Methods; Supplementary Table 4**). While the cell type specificity percentage is higher than the common variant results (~12%), we note that scenario 3 (distinct eQTLs for the same gene) cannot be assessed because these results are at the gene-level. The percentages become similar when the scenario is included (both ~26%).

Colocalization analysis with disease phenotypes and blood traits

We used colocalization analysis to identify the relationship between the cell type specific eQTLs identified and genetic risk loci for several diseases and blood traits. We performed colocalization with 14 diseases, including autoimmune conditions (inflammatory bowel disease (IBD)[34], rheumatoid arthritis (RA)[35], systemic lupus erythematosus (SLE)[36], IgA nephropathy[37], steroid sensitive nephrotic syndrome[38]), cancer (breast[39], colorectal[40], lung[41], lymphoma[21], prostate[42], Non-Hodgkin's lymphoma (NHL)[43], and lymphocytic leukemia[43]), COVID-19[44], and neuroinflammatory conditions (Alzheimer's[45] and Parkinson's disease[46]) as well as 44 blood counts and traits from the UK BioBank[47] (**Methods**). We restricted the analysis to common eGenes only (FDR<5%) and to genomic regions where at least one variant was genome-wide significant in the respective GWAS (p-value$<5x10^{-8}$).

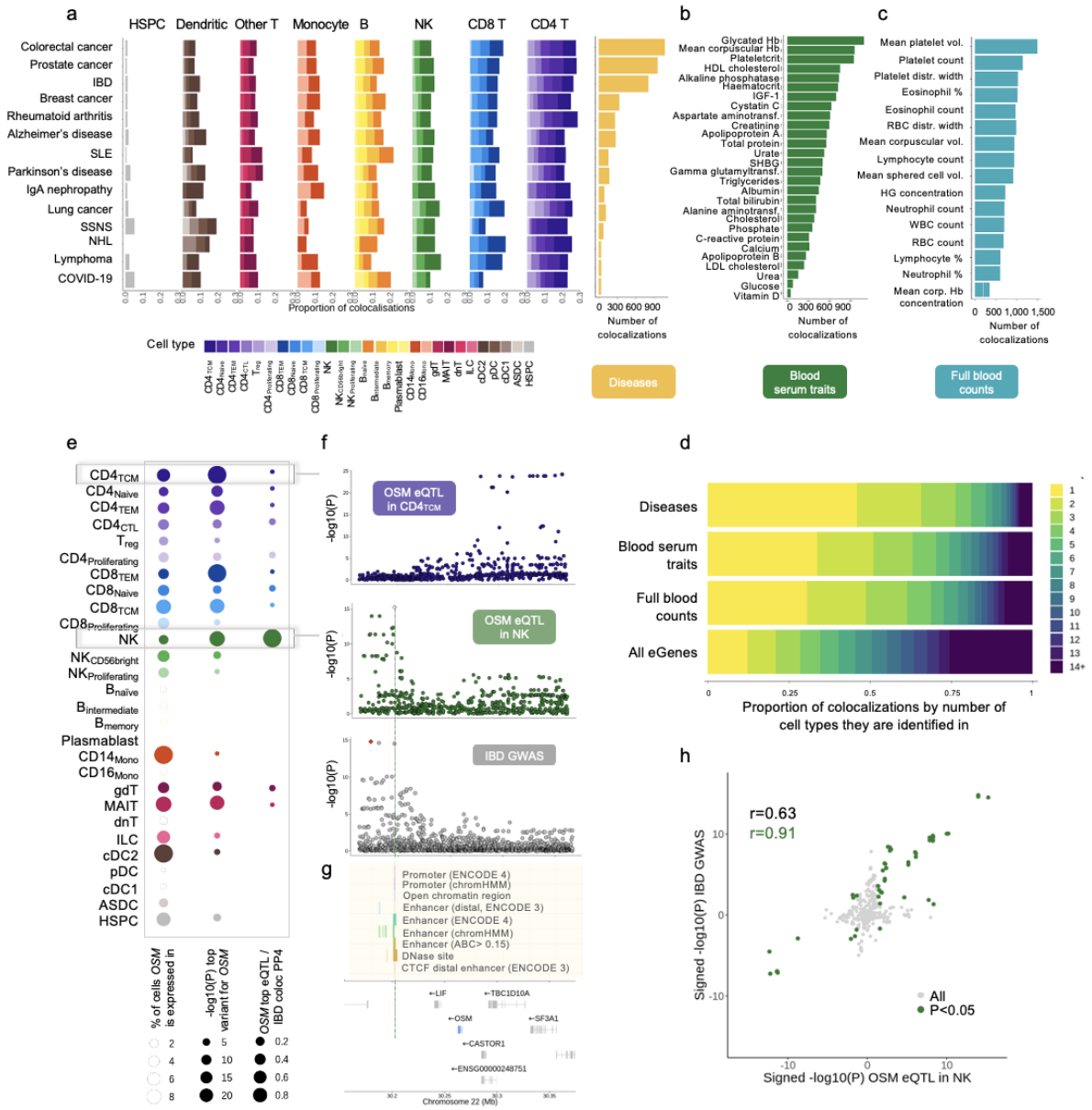


Figure 5. Colocalization analysis with disease phenotypes and blood traits. a) Different cell types contribute to different diseases. Left: bar plots showing the proportion of colocalization events for each trait, across the 28 cell types, colored by cell type. Right: bar plots showing the number of colocalization events (PP4>0.8) identified for each disease trait. **b,c)** Equivalent to the right most plot in a but for blood serum traits, and full blood counts respectively. **d)** Number of cell types each colocalization event was found in, stratified by trait category. All eGenes are included for reference: the colocalization events identified are very cell type specific, especially the disease ones, much more than eQTLs without shown disease relevance. **e-h)** Cell type specific colocalization between an eQTL signal for *OSM* in NK cells and inflammatory bowel disease (IBD). **e)** Patterns of expression, eQTL signal, and colocalization with IBD across cell types. *OSM* is broadly expressed and has eQTLs for many cell types, but the colocalization with IBD is specific to NK cells. **f)** Locus zoom plots for *OSM* eQTL signal for $CD4_{TCM}$ (top), *OSM* eQTL signal for NK (middle), IBD GWAS (bottom). **g)** Annotations for the likely causal *OSM* eQTL in NK are highlighted, and the location of *OSM* within the region is shown at the bottom. **h)** Scatter plot of negative log10 p-values multiplied by the direction of effect for the *OSM* eQTL in NK and the IBD GWAS signals, confirming the relationship between increased *OSM* expression in NK cells and increased IBD risk. IBD: inflammatory bowel disease, SLE: systemic erythematosus lupus, SSNS: steroid sensitive nephrotic syndrome, NHL: non-Hodgkin's lymphoma, Hb: haemoglobin, IGF-1: Insulin-like Growth Factor 1, SHBG: Sex Hormone-Binding Globulin, aminotransf.: aminotransferase, RBC: red blood cell, WBC: white blood cell, vol.: volume, distr.: distribution, corp.: corpuscular.

We found strong colocalization, indicating that the trait and cell type specific gene expression are driven by the same variant (posterior probability of hypothesis 4, that the same variant drives both traits: PP H4 ≥ 0.8), for 2,299 genes in 32,584 gene-cell type-trait combinations (**Supplementary Table 5**). Of these, 4,468 colocalization events (1,103 unique genes) were found for disease traits, 14,222 (1,029 genes) for blood serum traits, and 13,899 (982 genes) for full blood counts. Different cell types contributed to different traits and diseases; for example, B cells most contributed to SLE, CD4+ T cells to rheumatoid arthritis, and dendritic cells to Alzheimer's disease (**Fig. 5a; Supplementary Fig. 16**).

These colocalization events were remarkably cell type specific: out of 6,802 unique gene-trait colocalization pairs, 2,292 (34%) were uniquely found in one of the 28 cell types, and 4,828 (71%) identified in at most four cell types. This was especially pronounced for disease traits (43%), followed by blood serum traits (33%) and full blood counts (30%); when considering all eGenes the percentage is only 12% (**Fig. 5d; Supplementary Fig. 17**). Out of 2,299 genes with a significant colocalization event, 381 (17%) were not previously reported as eGenes by the eQTLGen consortium, and of those, 42% were specific to a single cell type (and 61% to at most 2/28), highlighting the importance of cell type-specific maps of genetic regulation to identify novel putative molecular and cellular mechanisms of disease variants.

For example, we identified a novel colocalization event between expression levels of *OSM* and both inflammatory bowel disease (IBD; PP4=0.99) specific to natural killer (NK) cells. The *OSM* gene encodes the Oncostatin M, a cytokine that contributes to many inflammatory diseases[48] and has recently been shown to play a role in modifying the risk of IBD, specifically in NK cells[49]. The top eQTL identified leads to overexpression of *OSM* in NK cells in individuals with the IBD risk allele, consistent with previous observations that *OSM* levels are higher in IBD patients[50,51]. We show that *OSM* is expressed broadly (20/28 cell types) and has evidence of genetic regulation of expression in over a third of all cell types (11/28), yet only the regulatory variants active in NK (downstream of the gene) colocalize with the disease-associated variant (PP4 for all other cell types < 0.06), with the other signal affecting *OSM* expression in other cell types is located several base pairs away, upstream of the gene and not in LD (r^2=0.0005 between top NK and top $CD4_{TCM}$ variant, an example of a *scenario 3* eGene in our cell type specificity framework). Effect sizes were broadly concordant between eQTL and GWAS signals, confirming that increased expression of *OSM* in NK cells was associated with increased IBD risk (**Fig. 5**).

We also identified a colocalization event between Alzheimer's disease (AD) and changes in expression levels of *IGHG4* specific to memory B cells (PP4=0.99). The *IGHG4* gene encodes the Immunoglobulin heavy constant gamma 4 (IGHG4); while the involvement of the immune system in AD is relatively recent, it has been shown that significantly higher levels of immunoglobulin occur in late-stage AD patients[52]. *IGHG4* is expressed in a subset of memory B cells, which have been linked to AD: a possible mechanism is that depleted memory B cells may initiate Aβ cascade-related neuropathological events leading to AD[53]. In our data, *IGHG4* is expressed in $B_{intermediate}$ and B_{memory} cells as well as in plasmablasts - consistent with the observation that class-switch recombination from IgM/IgD to IgG/A/E only happens after B cell activation[54] - and the colocalization with AD specific to B_{memory} alone. The direction of effect is consistent, indicating that higher *IGHG4* expression corresponds to higher AD risk (**Supplementary Fig. 18**).

Additionally, we identified a colocalization event between IBD and changes in expression levels of *TNFSF15* in dendritic cells, specifically Plasmacytoid dendritic cells (pDC, PP4=0.99), type 2 conventional dendritic cells (cDC2, PP4=0.98) and AXL+SIGLEC6+ dendritic cells (ASDC,

PP4=0.95). *TNFSF15* encodes the TNF-like cytokine 1A (TL1A)[55], a key regulator of inflammation and the only known ligand of the death-domain receptor 3 (DR3, encoded by *TNFRSF25*). *TNFSF15* was one of the first risk loci defined by the GWAS approach for IBD[56], and has the largest effect size in populations of East Asian ancestry[57]. In addition, a trans-ancestry meta-analysis has implicated a GTEx cis-eQTL for *TNFRSF25* (rs2986751), in IBD risk[58]. TL1A inhibition is an active area of drug development[59], with the anti-TL1A monoclonal antibody *tulisokibart* having recently shown efficacy in phase II trials for moderate to severe ulcerative colitis[60]. In our data, *TNFSF15* is uniquely expressed (>1% of cells) in the four dendritic cell types, and has an eQTL in three out of four. The effect size of the top variants is negative, indicating that lower expression of *TNFSF15* corresponds to higher IBD risk, consistent with similar results previously reported in monocytes[61] **Supplementary Fig. 19**).

Other examples of novel colocalization events identified were between the expression of the *PYCARD* gene in $CD8_{TEM}$ and both SLE (PP4=0.91) and IgAN (PP4=0.9) and a colocalization between a non-classical monocyte ($CD16_{Mono}$)-specific eQTL for *PPARGC1B* and SLE (PP4=0.94; **Supplementary Fig. 20**).

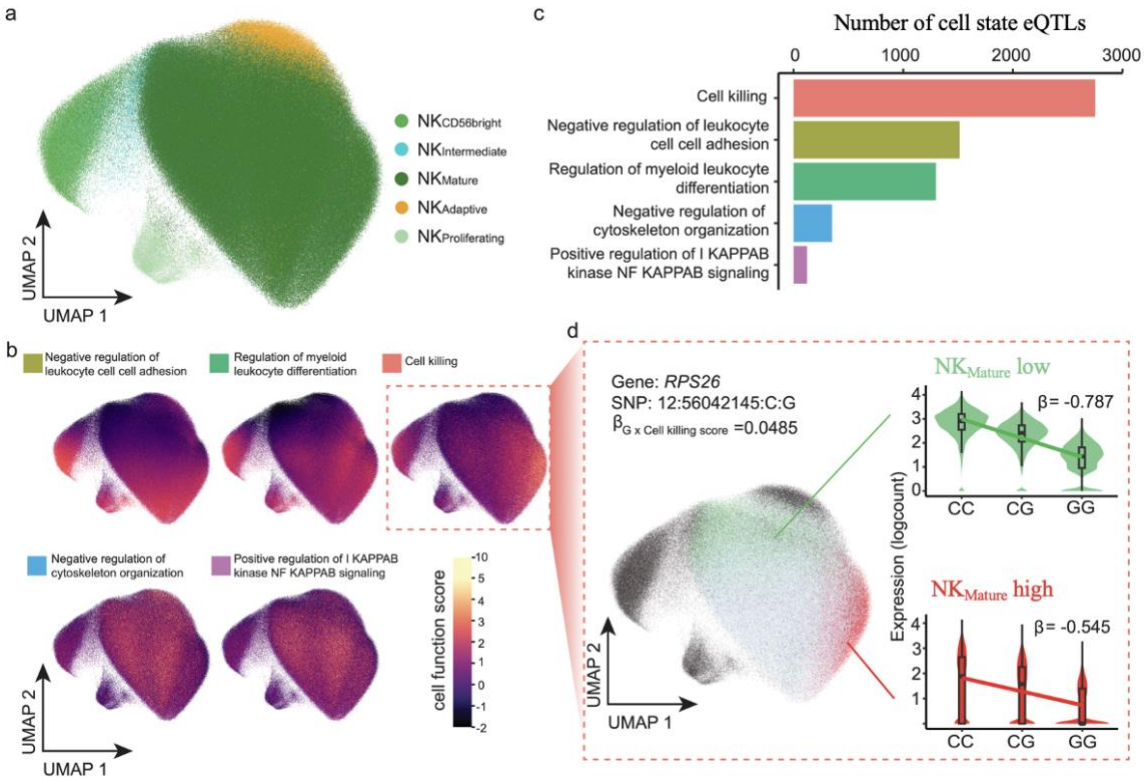


Figure 6. Identifying cell state eQTLs in NK cells. a) Cell function score-based UMAP, colored by different subtypes within NK cells, broadly matching those described in Rebuffet *et al*[62]. **b)** Visualisation of five representative cell functions as continuous cell states. **c)** Number of cell state eQTLs identified with significant interaction with the five cell states from **b**. **d)** An example of a cell state-specific sc-eQTL. Fine-grouping of NK_{mature} cells based on the cell killing scores multiplied by the effect size of the interaction term between genotype and cell killing scores (5% and 95% quantile); Violin plots show the expression of *RPS26* within each of the high and low cell state cell populations and grouped by different genotypes, with fitted trend line across different genotype groups.

Identification of cell-state eQTLs

We aimed to identify eQTL effects modulated by continuous cellular states beyond discrete cell types. A challenge with these analyses is the definition of cell states, with existing studies typically going one

of two ways: an unsupervised approach to determine axes of variability (often using principal components or related methods[9,10]) or using one or few very well defined variables such as disease state[63,64] or sex and age[63]. In the first case, these factors tend to be difficult to interpret; the latter approach misses out on novel states and contexts yet to be studied. Here, we used an in-house deep learning method - scDeepID - to identify biologically informed sub-cell types, states, and phenotypes. Collectively, scDeepID is a multi-task transformer capable of cell type annotation, cell function identification and cell simulation (additional detail in **Methods**).

Taking NK cells as an example, we performed label transfer with scDeepID from an integrated scRNA-seq data of NK cells[62] and further identify three subtypes within the macro NK cells defined previously: $NK_{adaptive}$ (most involved in cell activation), NK_{mature} (most cytotoxic) and $NK_{intermediate}$ (showing signatures of both NK_{mature} and $NK_{CD56bright}$ subtypes), with a good match to the original publication based on the expression of key markers (**Supplementary Fig. 21**). Together, we have five different subtypes for NK cells, including the $NK_{CD56bright}$ and $NK_{proliferating}$ cells. We trained a scDeepID model with these NK subtypes and used five GO biological process (GOBP) terms previously identified as key functions in NKs. They are: cell killing, positive regulation of I-kappaB kinase/NF-kappaB signaling, negative regulation of cytoskeleton organisation, negative regulation of leukocyte cell-cell adhesion and regulation of myeloid leukocyte differentiation to identify the cell state variation in NK cells (**Fig. 6a,b; Methods**).

To identify genetic regulation of gene expression mediated by these cell states, we mapped dynamic eQTLs using a linear mixed model with an interaction term and the cell function scores described above representing continuous cell states (**Methods**). We confirmed controlled type 1 error using permutations (**Supplementary Fig. 22**). We consider all NK cells (NK, now further split into three subtypes, plus $NK_{Proliferating}$, $NK_{CD56bright}$, n=850,833); all gene-SNP pairs that displayed evidence of eQTL effects (FDR<5%) in any of the three NK subtypes (16,759 unique gene-SNP pairs; **Methods**) and the five cell states described above to identify gene-by-context (GxC) effects. We identify 6,198 cell-state eQTLs in NK cells (for 3,723 unique gene-SNP pairs, 22% **Supplementary Fig. 23**), across the five selected cell functions, with cell killing, negative regulation of leukocyte cell-cell adhesion and regulation of myeloid leukocyte differentiation having the most dynamic eQTLs (**Fig. 6c**). For example an eQTL for *RPS26* (12:56042145:C:G) is modulated in strength by the cell killing function, with the gene regulatory effect of allele G decreasing *RPS26* expression levels being strongest in less mature or less cytotoxic cells (**Fig. 6d**).

Cell state abundance QTLs

We sought to identify genetic variants that directly modulate the composition of states, *i.e.*, cell state abundance. We applied Genotype–Neighborhood Associations (GeNA[65]) to test for cell state abundance QTLs (csaQTLs). We mapped csaQTLs in 8 major cell type categories (B cells, CD4+ T cells, CD8+ T cells, unconventional T cells, NK cells, monocytes, dendritic cells, and HSPCs; **Fig. 7a, Supplementary Fig. 24**), filtering to common genetic variants with MAF>1% and AAF>5% (**Methods**).

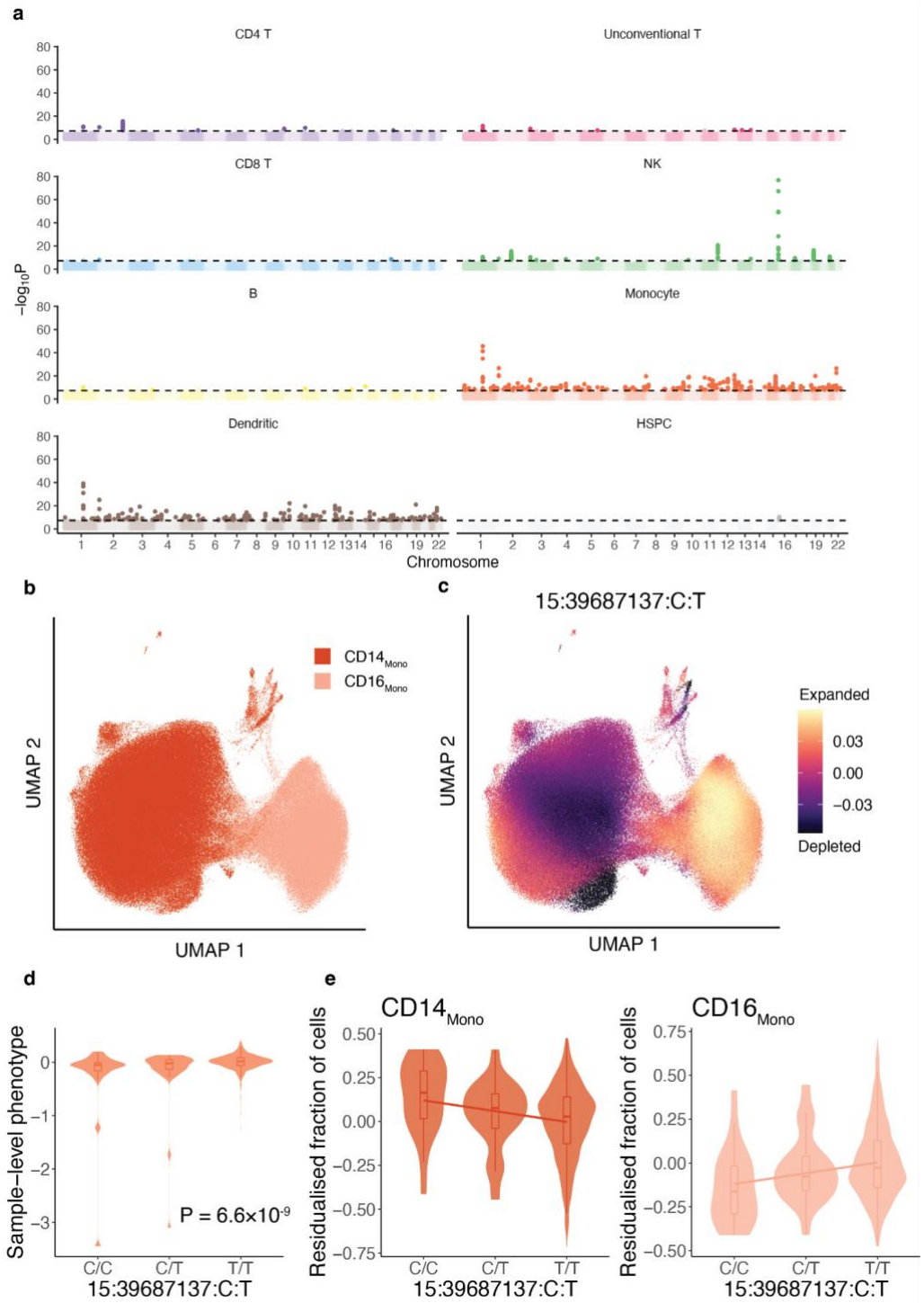


Figure 7. Identifying cell state abundance QTLs. a) Manhattan plot illustrating csaQTLs identified with GeNA for all major cell types. Dashed lines indicate the genome-wide significance threshold (p-value < 5×10^{-8}), and loci passing this threshold are colored according to the cell type they were significant in. **b)** UMAP projection of monocytes, colored by cell type. **c)** UMAP projection of monocytes, colored by GeNA neighbourhood-level phenotype values (correlation between cell abundance and alternative allele dose per neighbourhood) for 15:39687137:C:T. These values describe which cell populations are expanded or depleted. Colors are scaled between the 5th and 95th percentile. **d)** Sample-level phenotype values (linear combination of sample loadings on NAM-PCs used in the fitted model for the lead SNP) for 15:39687137:C:T. **e)** Fraction of total monocytes belonging to each monocyte subtype, corrected for covariates (age, sex, cohort and genotype PCs 1-7), for each genotype at 15:39687137:C:T**.**

In total, we identified 165 independent csaQTLs passing genome-wide significance (p-value<$5x10^{-8}$) and confirmed our Type I error was controlled using permutations (**Supplementary Fig. 24a; Methods**). Monocytes, dendritic cells and NK cells had the highest number of csaQTLs identified (**Supplementary Fig. 24b**). The discordance between csaQTL discovery power and cell type abundance may reflect the differential contribution of environmental and genetic factors to cell state abundance traits within immune subpopulations. NK, dendritic, and monocyte populations have been previously described as displaying high heritability for cell state abundance traits compared to other PBMC subsets[66–70]**.** Our analysis replicated four out of five total csaQTLs previously identified in the OneK1K dataset[65](**Supplementary Fig. 24c**). We found three csaQTL variants significant in both analyses and one locus significant in the OneK1K analysis that was suggestive in our analysis (p-value = 7.1×10^{-8}), in moderate LD with a lead variant (r^2=0.46) displaying a matching association with depletion of $NK_{CD56bright}$ cells (**Supplementary Table 6**).

We discovered novel csaQTLs associated with abundance shifts for discrete cell types and continuous cell states. These included 15:39687137:C:T for which alternative allele dose corresponded to depletion of CD14 monocytes and expansion of CD16 monocytes (p-value = 6.6×10^{-9}) (**Fig 7b-e**). While this locus was not an eQTL, it is located within an *FSIP1* intron and has been linked to *ANKRD3* and *PAK6* by promoter capture Hi-C, though only in T and B cells[71]. Many of the csaQTLs were associated with more complex alterations in the abundance of continuous cell states and shared gene expression programs within one or more cell types. These included monocyte csaQTL 12:69350234:C:A (p-value = 1.2×10^{-17}, **Supplementary Fig. 25a**), also an eQTL for *LYZ* in CD14 and CD16 Monocytes. *LYZ* encodes the antimicrobial enzyme lysozyme, which is secreted by phagocytes and catalyses peptidoglycan hydrolysis in bacterial cell walls, forming a critical component of innate immunity. We also identified a monocyte csaQTL 12:9953308:T:TG (p-value = 7.0×10^{-14}, **Supplementary Fig. 25b**) associated with *CLEC12A* expression in CD16 monocytes. *CLEC12A* encodes a C-type lectin, which inhibits monocyte and granulocyte activation[72] and may regulate inflammation in response to cell death[73]. *CLEC12A* has also been described as a marker for leukemic stem cells and is a possible drug target in acute myeloid leukaemia[74].

Discussion

We present a comprehensive cell type–specific map of the effects of common and rare genetic variation on gene expression in circulating immune cells assayed by scRNA-seq from 1,925 individuals. By mapping the effects of common and rare SNPs and indels called from whole-genome sequencing in 28 immune cell types, we identified more than 150,000 common variant eQTLs for 83% of the genes tested and over 30,000 gene-level rare variant signals for nearly 50% of all genes.

For common variants, especially for the most abundant cell types, we find an eQTL for the majority of genes tested, with results comparable to the largest eQTL map using whole blood 'bulk' transcriptomics from >30,000 people, the eQTLGen consortium (eQTLGen[2]). Testing similar numbers of genes, the eQTLGen authors found eQTLs for 88% of genes tested (n=16,987) compared to our 83% overall (17,674) and even slightly higher for our most abundant cell type ($CD4_{TCM}$; 85%). The cell type resolution is critical to identify effects that are masked in bulk genomic data: over 25% of the eGenes (4,572/17,674) identified in this study were not reported in eQTLGen, and these novel genes were generally expressed at a lower level and in a more cell type-specific manner. We note that the expected relationship between the number of cells and the number of eGenes begins to plateau, *i.e.*, increasing less steeply for the four most abundant cell types (number of cells >500,000), suggesting that this may be the necessary number to identify common variant effects within a cell type. For example, when

comparing to OneK1K[16], the currently largest single-cell eQTL study, we identify nearly three times as many eGenes (17,674 vs 6,469) due to a combination of factors: approximately twice the number of individuals, almost three times the number of cells per individual, twice as many cell types, and new methods for eQTL mapping that, rather than relying on pseudobulk, model single-cell profiles directly, increasing power and allowing us to test more genes in total.

When mapping eQTLs across increasing numbers of cell types, an important and often non-trivial question is the extent to which these eQTLs are specific or shared. We propose a framework to distinguish between five possible scenarios, which start from the pairwise comparison between two cell types: an eQTL may be specific to a cell type because the corresponding target gene is only expressed in that cell type or because the regulatory effect is specific despite significant expression in both cell types; eQTLs for the same gene may be found in both cell types, but with distinct top variants or the same; and in some cases the same variant may affect gene expression in opposite directions between the two cell types. Aggregation of all pairwise combinations can be used to build a specificity score for each identified eQTL, which will be a useful framework to quantify cell type specificity in future studies. In this study, about 50% of eQTLs were specific to one or a subset of cell types, and very few were ubiquitous to all, even when accounting for differences in power, demonstrating the importance of eQTL maps across different cell types, especially rare ones, that would be missed in tissue-level maps.

Colocalization analysis with 14 diseases and 44 blood traits identified over 30,000 colocalization events across all cell types and traits for over 2,000 unique genes. Of these, 381 (17%) had not been described as having any eQTLs by the largest blood eQTL resource. A third of all colocalizations were specific to a single cell type, and over 70% were found in at most four cell types. This is significantly higher (~3X) than the equivalent number for all eGenes (with or without colocalization), suggesting that variants associated with diseases or traits are especially likely to operate in a cell type specific manner. For the 381 novel colocalization events, the cell type specificity is even higher: 41% were specific to a single cell type (and 60% to at most two (out of 28)), suggesting that eQTL maps in increasingly granular and rarer cell types (and states) may be key to bridging the gap between eQTL and GWAS variants[75,76].

Rare and low-frequency genetic variants are expected to contribute to disease risk, but their molecular modes of action are under-explored. Using a recently developed tool, we can test for the combined effects of cis-regulatory rare variants using set-based tests, and a combined approach of using both a burden test and the sequence-kernel association test (SKAT) provides us with the power to identify effects driven by rare variation for over 10,000 genes: nearly half of the genes tested. 20% of these genes (n=1,982) had not previously been reported to have eQTL effects – common or rare – by eQTLGen. These genes were often expressed in fewer cell types on average, likely explaining why they were missed by whole blood bulk transcriptomics. Rare variants tend to have larger effect sizes and be more directly interpretable mechanistically, so their discovery is important, but they will require increasingly larger cohorts with whole genome sequencing data. Such data also allows the characterisation of other variant types not visible from genotyping array data, such as the tandem repeat variation explored in the accompanying manuscript by Tanudisastro *et al*.

This dataset also allows us to explore genetic effects modulated not just by cell type but by cell states or functions. Using natural killer cells and a biologically-driven deep learning method, we first define cellular states corresponding to specific functions (for example, cell killing) and then test for the effects

that these have on eQTLs, identifying over 6,000 cell state eQTLs (for 22% of all NK eQTLs) across five cell phenotypes in NK cells.

Finally, in addition to exploring the impact of genetic variation on the expression of genes *within* defined cell states, we map the effects of genetic variants on the composition of cell states within each major cell type. Previous GWAS studies incorporating flow cytometry data have found that variants associated with cell type abundance traits often colocalize with disease risk[67,69,70,77–79]. However, to date, few studies have applied statistical genetic analysis of cell type abundance in single cell data. Using a recently proposed method[65], we identify 165 cell state abundance QTLs (csaQTLs) across 8 major cell types. We detected most of our csaQTLs for monocytes, dendritic cells, and NK cells, consistent with previous reports[65]. This could be explained by the higher estimated heritability of NK and myeloid cell types in abundance compared to other blood cell populations[67–70]. Alternatively, cell surface markers may better represent T and B cell states, which scRNA-seq does not capture. These results suggest that monocyte, dendritic, and NK cell state composition is more influenced by genetic variation. In contrast, environmental and higher-order physiological factors, such as past exposure to infection more substantially drive cell state proportions within the other major cell types.

In three accompanying TenK10K manuscripts, we leverage extracts of the same dataset to explore additional aspects of human cell type-specific molecular variation. In one[80], we investigate the impact of an under-studied class of genetic variation, tandem repeat length polymorphism, on cell type-specific gene expression, identifying over 60,000 eQTLs, many of them independent of the eQTLs described in this manuscript, emphasising the substantial molecular impact, high mutation rates, and weak LD structure of this important class of variation. In other manuscripts (in preparation), we further characterise the causal relationship between molecular QTLs and human phenotypic associations and the impact of genetic variation on measures of cell type-specific chromatin accessibility, a critical upstream marker of regulatory impacts on gene expression.

In sum, these publications demonstrate both the power of population-scale cellular genomics[8], and the importance of multiple aspects of study design for future resources. We show that sequencing-based approaches to characterising genetic variation, ideally whole-genome sequencing, are critical for accessing both low-frequency variants and complex classes of structural variation and that single-cell methods for studying both gene expression and chromatin accessibility phenotypes provide both a massive increase in discovery power and the opportunity to access cell type-specific effects, which constitute the majority of genetic impacts on molecular phenotypes. Population cellular genomic datasets provide an important lens on the impact of gene perturbation on whole-organism molecular physiology, complementary to high-throughput experimental perturbation approaches. They will be critical to building increasingly accurate models of human biology.

Further expansions in the sample size of population cellular genomic studies and increasing the diversity of participants and tissues - including individuals of diverse ancestry and from a wide variety of active disease states and a much wider variety of tissue and cell types - will be important for building a comprehensive understanding of the molecular origins of human disease.

Methods

Cohorts

Blood samples were collected from 2,339 consenting individuals from two Australian biomedical research cohorts: the Tasmanian Ophthalmic Biobank (TOB) [ethics protocol 2020/ETH02479] and the BioHEART study [ethics protocol 2019/ETH08376] cohort.

The TOB cohort consists of 983 individuals (562 females, 421 males) recruited between January and February 2018 by the Royal Hobart Hospital and Menzies Institute of Medical Research in Tasmania, Australia. The selection criteria for inclusion in the cohort were as follows: (1) age 18 years and over, (2) no signs of ocular disease, and (3) self-reported British, Scottish, or Irish ancestry.

The BioHEART-CT cohort included 1,356 individuals (561 females and 795 males) recruited between March 29th, 2016 and May 2nd, 2023, presenting to cohort clinical sites for clinically indicated cardiovascular imaging. The selection criteria for inclusion in the BioHEART-CT cohort were as follows: (1) age 18 years and over, and (2) clinical indication for coronary CT angiography for the assessment of suspected or previously diagnosed CAD.

After donor, variant, and cell QC, including selection to individuals of European ancestry only, 1,925 individuals remained: 950 from TOB and 975 from BioHEART-CT (**Supplementary Fig. 1**).

Blood collection and DNA extraction

For the TOB cohort, blood samples were collected from each participant and stored in 1 x 8mL heparin cell preparation tubes (CPT) and 1x 9mL EDTA tubes. For the BioHEART-CT cohort, aliquots were stored at -80C.

DNA extraction from BioHEART-CT buffy coat samples was performed by the Australian Genome Research Facility (AGRF) using QIAamp96 DNA QIAcube HT kit. DNA extraction from TOB samples was performed at the Menzies Institute of Medical Research (University of Tasmania) using the QIAmp Blood Mini Kit. For a subset of TOB samples that failed initial quality control checks during sequencing, DNA was extracted from PBMCs a second time using the Qiagen DNeasy Blood and Tissue Kit and eluted in an AE buffer by the Garvan Molecular Genetics lab service. DNA concentration was measured for all samples using fluorometric quantitation with the Quantifluor from Promega prior to sequencing.

Whole-genome sequencing data

Whole-genome sequencing (WGS) data was generated for all individuals from the TOB cohort (n=983) and BioHEART-CT (n=1,356).

Sequencing library preparation and whole-genome sequencing

Sequencing library preparation was performed using the KAPA PCR-free v2.1 kit (Roche) for the TOB samples, and the DNA PCR-free kit (Illumina) for the BioHEART-CT samples. Short-read whole genome sequencing was performed using Illumina 2 x 150 bp chemistry on the NovaSeq 6000 instrument to achieve mean 30x coverage. Library preparation and sequencing were performed by the Garvan Sequencing Platform.

Alignment and variant calling

Alignment and variant calling were performed in line with GATK-DRAGEN Best Practices. Reads were aligned from FASTQ (BioHEART) or CRAMs (TOB) against the GRCh38 reference genome (masked version, as available at gs://gcp-public-data--broad-references/hg38/v0/dragen_reference/) using a fork of DRAGMAP v1.3.1 (https://github.com/populationgenomics/DRAGMAP). When aligning from CRAMs, bazam (https://github.com/ssadedin/bazam) was used to first convert the reads back to FASTQ format. Calling of SNPs and indels was then performed using GATK4 HaplotypeCaller in DRAGEN mode. The resulting gVCF files were combined using hail's VDS Combiner and stored in the variant dataset (VDS) format.

Sample QC

Hard sample filtering was applied to remove individuals with: high rates of contamination (FREEMIX ≥ 5%); an elevated percentage of chimeras (PCT_CHIMERAS ≥ 6.3%); median insert size < 250bp; and sex aneuploidy or discordance between reported and observed sex. The dataset was also inspected for outlier samples based on the number of singletons, het/hom ratio, Ts/Tv ratio, and ins/del ratio, taking care not to exclude ancestry outliers based on these metrics.

A set of high-quality sites, defined as an LD-pruned subset of autosomal, biallelic, allele frequency > 0.01, call rate > 99% and InbreedingCoeff > -0.25, were extracted from 1000 Genomes and HGDP. Samples with call rate < 0.85 across these sites were also excluded.

Following hard sample filtering, related individuals were identified using hail's implementation of PC-Relate and confirmed using somalier relate v0.2.15 (https://github.com/brentp/somalier). For family sets consisting of second-degree or closer relatives, as defined by a kinship coefficient > 0.088, the sample with the highest quality single cell RNA-seq data was retained, as determined by the number of cells and cell type composition. When equal quality single cell data was available, the sample with the highest mean genome sequencing depth on chromosome 20 was retained.

Ancestry inference was then performed on the high-quality, unrelated set of samples, by combining this set with background samples from 1000 Genomes and HGDP, then computing principal components using gnomAD's run_pca_with_relateds method and the same set of high quality sites defined for call rate estimation above. Ancestry labels were assigned based on genetic similarity, as identified through principal component analysis, to those of known ancestry in the background datasets. For association testing, only samples assigned to the European cluster were retained.

Variant QC

Allele-specific Variant Quality Score Recalibration or AS-VQSR (GATK v4.2.6.1) was run using all QC-pass, unrelated samples (regardless of ancestry), using the default training resources and priors. For SNPs this is hapmap (prior=15), omni (prior=12), 1000G (prior=10) and dbsnp (prior=7). For indels this is mills (prior=12), axiomPoly (prior=10), and dbsnp (prior=2). The following features were included for all SNPs and indels: AS_ReadPosRankSum, AS_MQRankSum, AS_QD, AS_FS, AS_SOR. AS_MQ was used for SNPs only. All variants filtered by AS-VQSR were removed.

Within the European-only association callset, we then removed any variants that had no high-quality genotype calls (defined as GQ>=20, DP>=10 and allele balance > 0.2 for heterozygotes) or with an excess of heterozygotes at that site compared to Hardy-Weinberg expectations (InbreedingCoeff < -0.3). A final filter to remove sites with missingness > 15% was then applied prior to association testing.

Single-cell RNA sequencing data

Single-cell RNA-seq data was collected from peripheral blood mononuclear cells (PBMCs) from samples from individuals from the Tasmanian Ophthalmic Biobank (TOB) (OneK1K[16]; n=1,036) and BioHEART[81] (n=1,012) cohorts. **Supplementary Fig. 26** illustrates the experimental workflow and key steps.

Sample pooling and cell count estimation

TOB samples were previously pooled during preparation of the OneK1K project, with pools made up of 12-14 individuals[16]. In order to capture more cells per individual (aiming for up to 5,000 cells per individual), two pools (n=24-28) at a time were further pooled (weighted by cell viability) and loaded into 3 sequencing lanes (aiming for 40,000 cells each, so 120,000 in total). Additionally, "maxi pools" of 4 combined pools (96-112 individuals) were sequenced in two separate lanes to compensate for possible low cell numbers.

In contrast, BioHEART-CT samples were pooled (weighted by cell viability) for the first time for this project, each pool comprises 8 individuals and is sequenced in a separate sequencing lane, aiming for 40,000 cells (~5,000 cells per individual). In some cases, pools with low cell viability were sequenced again to rescue cell numbers and get closer to the target.

Capture and sequencing

Cells were captured using the single-cell 3' gene expression high throughput (3' GEX HT) kit v3.1 from 10X on a Chromium X machine. Library preparation and short-read sequencing were performed using illumina's NovaSeq 6,000 machine.

Alignment and ambient RNA detection

scRNA-seq reads were aligned to a genome reference (GRCh38, Gencode release 44, Ensembl 110) using STAR[82], as implemented within 10X's CellRanger software (v7.2.0) which also estimates the number of cells captured (**Supplementary Fig. 27**). Next, CellBender[83] v0.3.0 was used to identify residual ambient RNA in each sequencing library.

Demultiplexing and doublet detection

To assign cells to their donor of origin, as well as to remove doublets (droplets containing more than one cell), we used Demuxafy[84] (version: 2.1.0), a tool to runs a number of demultiplexing and double detection software and outputs a consensus call for each individual cell. Using *Demuxafy*, we ran *vireo*[85] to demultiplex cells to their donor of origin within each sequencing library, using genotypes extracted from either SNP array data (available as part of the OneK1K data; for TOB) or using common (MAF>5%) biallelic SNPs from WGS (for BioHEART). A handful of individuals did not have available genotype data, but *vireo* allows us to include the expected number of individuals in each pool with or without genotypes, which makes the donor assignment and doublet detection more accurate. For doublet detection, we used *Demuxafy* to run scds[86] and *ScDblFinder*[87], and used majority voting between all three tools (*vireo*, *scds* and *scDblFinder*) to determine the final set of doublets for each sequencing library (**Supplementary Fig. 28**).

Cell typing

To assign individual cells to specific immune cell types we follow the recommendations of the single-cell eQTLGen consortium[14]. We use a combined approach that uses *scPred*[88], hierarchical progressive learning, scHPL[89], and Azimuth as the reference dataset[15] as detailed in https://powellgenomicslab.github.io/WG2-pipeline-classification-docs/index.html. To run this step, we converted our single-cell Scanpy object into Seurat[90] objects. We also used CellTypist[91,92] to confirm cell typing of individual cells, with good concordance (**Supplementary Fig. 29**). Since the former approach provided more granularity (28 cell types vs 18 identified by CellTypist), we proceeded with the 28 cell types for the remainder of the manuscript.

Quality Control

Next, we performed quality control (QC) with Scanpy (v1.8.2)[93]. For each library, we removed cells with >20% mitochondrial reads, <800 reads detected, and <1000 or >10,000 genes detected (**Supplementary Fig. 30**). We also removed individuals with abnormal cell type composition and individuals with fewer than 100 cells.

Overall, after cell and sample QC (based on scRNA-seq and WGS), a total of 5,438,679 cells from 1,925 individuals were left for consequent analyses (**Supplementary Fig. 1**).

Association testing

SAIGE-QTL

To test for the effects of sets of both common and rare variant associations with single-cell gene expression, we used the recently proposed SAIGE-QTL method[13]. Briefly, SAIGE-QTL implements a Poisson mixed model, thus effectively modelling single-cell counts and repeat cells for the same individual. Additionally, several computational and statistical speed-ups are implemented to improve speed, scalability, and storage usage. Finally, implementation of both single-variant tests (where effects of the genotype at a single locus are assessed), as well as set-based tests (where signals from multiple variants within a region are assessed together in their combined effect), is available, as briefly described below.

We mapped eQTLs separately in each of the 28 cell types. We confirmed that our tests were calibrated by shuffling the individual IDs and re-performing the test for our most abundant cell type, CD4 positive central memory T cells ($CD4_{TCM}$), the results of which show no evidence of false positives for both common and rare variant tests (**Fig. 2a, c**).

Fitting the null

For every gene and cell type, we fit a single null model using SAIGE-QTL, the same for both subsequent tests (common and rare variant tests alike). Briefly, we fit a Poisson model using the raw counts from the single-cell expression data as the outcome variable (phenotype), and the following as covariates: sex and age of the individuals, 7 genotype PCs, 5 cell type-specific expression PCs, cohort status (TOB or BioHEART), percentage of counts from mitochondrial genes and total counts. We provide all additional parameters and flags as part of the available code in GitHub. We only test genes that are expressed in at least 1% of cells in each cell type. For a subset of gene-cell type combinations the model failed to converge (0.27%) and these were excluded from the analysis.

Common variant eQTLs

We test for the effects of common genetic variants on single-cell expression using the single-variant test as implemented in SAIGE-QTL. We consider all SNPs and indels with MAF>=1% and missingness < 15% in our sample. We test for *cis* (variants within a +/-100kb window around the gene body) effects only. We apply a two-step multiple testing correction as recommended in the SAIGE-QTL manuscript[13], using the Aggregated Cauchy Association Test (ACAT[94]) at gene-level, and Storey's q-value method (*qvalue* package in R) across genes. At FDR<5%, we identify a total of 154,932 eQTLs in at least one of the 28 cell types, from 17,674 unique genes (83% of all genes tested).

The number of eQTLs identified varied substantially across cell types, ranging from 11,491 eQTLs in $CD4_{TCM}$ (85% of all genes tested in this cell type) to 449 eQTLs in innate lymphoid cells (ILC; 3%). This was likely explained by differences in the statistical power across cell types, largely due to the number of cells available for each cell type (correlation between log10 number of cells and number of eGenes detected = 0.98).

Rare variant eQTLs

We test for the effects of rare variants on single-cell expression using the set-based test as implemented in SAIGE-QTL. We consider all SNPs and indels with MAF<1% and missingness < 15%. We test for the combined effects of all variants within a *cis* region around the gene (+/-100kb around the gene body). As described in the SAIGE-QTL publication, we consider a burden test and the SKAT test, and use the Cauchy method to combine p-values at the gene-level. Similar to the approach used for the common variant results above, we use the q-value method and report significant results at FDR<5%.

Overall, we observe 34,376 rare variant signals across all cell types, for 10,005 unique genes after FDR correction using Storey's qvalue on the Cauchy combined p-values from the two tests (burden test and SKAT). The percentage of genes with a rare-variant combined signal out of all genes tested ranged from 0.09% (ILC) to 39% ($CD4_{TCM}$). Once again, the tests were well calibrated when shuffling individual IDs and the number of cells for each cell type is highly correlated with the number of eGenes identified.

PromoterAI to predict promoter activity of variants

PromoterAI is a convolutional neural network that uses ~20 kb of sequence context around a promoter variant to predict its effect on gene expression. It is pretrained on histone modifications, DNA accessibility, TF occupancy, and strand-specific CAGE value around TSS at base-pair resolution and fine-tuned with a curated set of rare non-coding promoter variants associated with unusually high or low expression in GTEx v8. PromoterAI scores range from -1 to 1, with negative values indicating a predicted under-expression effect and positive values indicating a predicted over-expression effect.

Single-cell ATAC-sequencing processing

Single-cell ATAC data was obtained from a subset of the TOB cohort which includes 30,829 nuclei from 8 individuals (Xue *et al.* 2025). Sequencing reads were aligned to the Human Genome Reference GRCh38-1.2.0 (gencode v44) using Cell Ranger ATAC (v2.1.0) and the chromatin peaks were called using the built-in MACS2[95]. Cell type labels were annotated based on a CITE-seq reference[15] with a combined Multiome bridging dataset (Xue *et al.* 2025), consistent with the 28 PBMC cell types

annotated for scRNA-seq data. After quality control, 17,099 nuclei were retained to annotate the chromatin peaks in each cell type for the eQTL signals.

Cell type specificity analysis

To quantify cell-type specificity, for each of the 154,932 identified eQTLs, we first performed pairwise comparisons of association estimates in the primary cell type and in each of the other 27 cell types. We classified these pairwise comparisons into five scenarios based on eQTL association, eGene expression, and direction of allelic effect on gene expression in the two cell types: 1) eGene is expressed only in the primary cell type, 2) eGene is expressed in both cell types, but eQTL identified only in the primary cell type, 3) independent eQTL variants (LD $r^2 < 0.5$) for the same eGene are identified in each cell type, 4) same eQTL variants or different variants in LD $r^2 \geq 0.5$) for the same eGene are identified in each cell type with concordant direction of effect, and 5) same eQTL variants or different variants in LD $r^2 \geq 0.5$) for the same eGene are identified in each cell type with discordant direction of effect. These scenarios are summarised in the table below

Scenario	**eQTL**	**eGene**	**Direction of effect on gene expression**
1	Found in cell type A only	Expressed in cell type A only	–
2	Found in cell type A only	Expressed in cell type A & B	–
3	Independent eQTL variants (LD r^2 <0.2) found in cell type A and B	Expressed in cell type A & B	–
4	Shared eQTL (same variant or different variants with LD $r^2 \geq 0.2$) found in cell type A and B	Expressed in cell type A & B	Concordant
5	Shared eQTL found in cell type A and B	Expressed in cell type A & B	Discordant

Further, we used a pairwise meta-analysis approach for all eQTLs that are initially classified as in *scenario 2* to take into account potential type 2 error due to insufficient statistical power for association analysis in rarer cell types. This was done for each significant eQTL (FDR<5%), by comparing its effect size in the primary cell type to its effect sizes in each of the other cell types. Meta-analysis was performed using a random-effects model with the *metafor* package in R[96] (v4.0.0). We compared the resulting meta-analysis p-value to the primary p-value, denoting an empirically chosen threshold of 10^5-fold reduction (**Supplementary Fig. 31**) as evidence of sub-discovery eQTL signal in cell type 2, in which case the eQTL in question would be reclassified as scenario 4 / 5 depending on concordance of allelic effect direction.

Next, we calculated an aggregate score for each of the 17,674 common variant eGenes by taking the maximum value of the scenarios above across all pairs of cell types for each eGene (results presented in **Figure 4a, bar plot on the right**). The following table described interpretation of the scenario / aggregated values in each of the comparison:

Scenario / Score	Pairwise comparison (pie charts inside cell type by cell type matrix in Figure 4b)	Aggregated eGene score (bar plots on Figure 4a)
1	eQTL found in the cell type A (rows / y-axis in Figure 4b) and eGene not expressed in cell type B (columns / x-axis in Figure 4b)	eQTL found and eGene expressed only in 1 cell type
2	eGene expressed in cell type A and B, but eQTL identified in cell type A only	eGene expressed in ≥2 cell types, but eQTL found only in 1 cell type
3	Independent eQTL variants (LD r^2 <0.5) for the same eGene found in cell type A and B	≥2 independent eQTL variants (LD r^2 < 0.5) for the same eGene found in ≥2 cell types, without evidence of sharedness across cell type pairs
4	Shared eQTL (same variant or different variants with LD $r^2 \geq 0.5$) found in cell type A and B with concordant allelic effect direction on gene expression	eQTL is shared between ≥2 cell types (LD $r^2 \geq 0.5$), all of which show concordant allelic effect direction
5	Shared eQTL found in cell type A and B with discordant allelic effect direction on gene expression	eQTL is shared between ≥2 cell types (LD $r^2 \geq 0.5$), and ≥2 cell types show discordant allelic effect direction

Given that correlation of effect estimates are proportional to LD r^2 between variants, we note that variants in lower LD are more likely to exhibit a discordant allelic effect direction and therefore may result in overestimation of aggregated count for scenario 5. We repeated this approach using the gene-level rare variant association results. For each gene in a given cell type identified at FDR <5% based on the Cauchy-aggregation p-value (from burden test and the SKAT test), we compared the estimates for the same gene against those derived from each of the other 27 cell types. Similar to the common variant results above, each pairwise comparison is initially classified into one of the following scenarios: 1) gene is expressed in the primary cell type only, 2) gene is expressed in both cell types, but rare variant association signal is identified only in cell type 1, and 3) rare variant association signal is identified in both cell types. To account for potential misclassification of scenario 2 genes where rare variant association signal is absent due to lack of statistical power in the association test in the secondary cell type, we meta-analysed effect estimates from Burden test results of both cell cell types using a random-effects model. Genes with meta-analysis p-value 10^5-fold smaller than the original meta-analysis were reclassified as *likely shared* eGenes. Unlike the common variant results, we did not perform comparison of direction of effect sizes given that the discovery was done using unsigned p-values from Cauchy aggregation of Burden and SKAT tests.

Statistical colocalization

We use the COLOC package[97] to perform statistical colocalization between our eQTLs and a number of traits and diseases using GWAS summary stats from publicly available sources.

Specifically, we downloaded GWAS summary statistics for

- inflammatory bowel disease[34] (from https://www.ibdgenetics.org/),
- systemic lupus erythematosus[36] (https://www.ebi.ac.uk/gwas/studies/GCST003156),

- rheumatoid arthritis[35] (https://www.ebi.ac.uk/gwas/studies/GCST90132223),
- COVID-19[44] (https://www.ebi.ac.uk/gwas/studies/GCST011071),
- lung cancer[41] (https://www.ebi.ac.uk/gwas/studies/GCST004748),
- breast cancer[39] (https://www.ebi.ac.uk/gwas/studies/GCST004988),
- Parkinson's disease[46] (https://www.ebi.ac.uk/gwas/studies/GCST009325),
- malignant lymphoma[21] (https://www.ebi.ac.uk/gwas/studies/GCST90018878),
- prostate cancer[42] (https://www.ebi.ac.uk/gwas/studies/GCST90274713),
- Alzheimer's disease[45] (https://www.ebi.ac.uk/gwas/studies/GCST90027158),
- colorectal cancer[40] (https://www.ebi.ac.uk/gwas/studies/GCST90129505),
- IgA nephropathy[37] (https://www.columbiamedicine.org/divisions/kiryluk/study_gwas_stat_IgA_summary.php),
- lymphocytic leukemia[43] (https://www.ebi.ac.uk/gwas/studies/GCST90011814),
- steroid sensitive nephrotic syndrome[38] (https://www.ebi.ac.uk/gwas/studies/GCST90258619).

We downloaded UKBB GWAS summary statistics for 44 blood cell count and serum traits[47] (SNPs and indels only) from (https://gymreklab.com/science/2023/09/08/Margoliash-et-al-paper.html).

Colocalization was performed using coloc.abf function from coloc (v5.2.3)[97]. Colocalization was performed using SNP GWAS summary statistics and common variant eQTLs in our dataset with eQTL type set to 'quant' and providing N and MAF rather than sdY as single-cell expression was not normalized and modelled as a Poisson distribution by SAIGE-QTL.

Cell state definition

scDeepID is a multi-task transformer, where a biologically-guided masked-encoder is leveraged to generate cell function embedding along the supplied database, followed by combining a classification token [CLS] for cell type identification and reconstruction [REC] token for cell simulation. The reparameterization trick from variational autoencoders[98] was applied to the REC token before the cell reconstruction decoder. The attention scores from each cell function to CLS token form an attention-based latent space after training. scDeepID can be flexibly used with different supplied databases such as Reactome, gene ontology (GO), and more.

To define the cell states in the major NK cells of the TenK10K cohort, we first train our scDeepID model on a publicly available dataset that contains different NK subtypes[62], then perform label transfer by the trained model on our data. Specifically, we use Scanpy to automatically select ~600 highly variable genes (min_mean=0.0125, max_mean=3, min_disp=0.5) in the reference dataset, then only keep the shared genes between the reference and TenK10K cohort to train the model (*i.e.*, 554 genes). In particular, we further define $NK_{intermediate}$, $NK_{adaptive}$ and NK_{mature} cells in the NK cell type. Then we train a separate scDeepID model with the three fine-grained cell types as well as $NK_{CD56bright}$ and $NK_{proliferating}$ cells (with 2,292 highly variable genes, selected by Scanpy with min_mean=0.0125, max_mean=3, min_disp=0.5). We use the human GOBP database as a guide to train the scDeepID models above. We then extract the attention-based latent space and normalise each cell function score with a zero mean and unit variance. Following the cell state definition, we use lme4[99] to test the significance of the interaction term *cell state* * *genotype* while correcting for age, sex, cohort (bioHEART or TOB), and total counts.

csaQTL GWAS analysis

We performed csaQTL associations using GeNA[65] v1.0.0. We generated single cell objects for each major cell type group and applied the following processing.to each object. Following a standard Scanpy pipeline, we normalised read counts per 10,000, log-transformed normalised counts, and selected approximately 3000 highly variable genes per major cell type (min_mean=0.0125, max_mean=3, min_disp=0.5). The scanpy.pp.regress_out function was used to obtain residualised counts, removing technical variation associated with total counts (library size) and percentage of mitochondrial reads. Expression for each gene was scaled to unit variance across cells. We then created MultiAnnData objects containing the processed counts and donor covariate information. We applied principal component analysis with scanpy using default parameters. Harmony[100] (harmonypy v0.0.10) was run with default parameters, using sequencing library pools as batch assignment to correct principal components for technical batch effects between sequencing libraries. The harmonized principal components were used to construct a nearest neighbor graph and UMAP using scanpy default parameters.

Next, we generated a neighbourhood abundance matrix (NAM) and its principal components (NAM-PCs) with CNA[101] (v0.1.6) using default parameters. We then applied GeNA, controlling for age, sex, cohort and, the top seven genotype principal components. We tested associations for all common variants (MAF >= 0.01 and percent missing genotypes < %15). To calibrate the model, we also ran GeNA on permuted genotypes generated by random label swapping (one permutation performed per cell type). Based on this permutation testing, we observed poor calibration for variants below 0.05 AAF, and removed them from the analysis. We considered variants passing the genome-wide significance threshold (p-value<$5x10^{-8}$) as statistically significant. Lead SNPs were defined using the clumping method as described in the GeNA publication[65]. Briefly, at each locus the most significant SNP was considered the lead SNP, and all SNPs with LD > 0.8 to the lead SNP or within a 1 MB window were considered LD-SNPs.We considered the number of unique lead-SNPs for csaQTL loci as the total independent csaQTLs discovered. To visualise cell counts adjusted for covariates, we calculated the fraction of cells within each major subtype belonging to each minor subtype. Using these cell fractions, we fit a linear model with the lm function in R, including age, sex, cohort, and the first 7 genotype principal components as covariates, and visualised the residuals from the fitted model.

Acknowledgements

We acknowledge Vladislav Savelyev and Leonhard Gruenschloss for their contributions in prototyping the whole-genome sequencing production pipeline, and Garvan Genomics Platform for providing sequencing services and technical support. We acknowledge Samantha Bryen's contribution to the variant curation of selected examples. We acknowledge Michael Geaghan from the Garvan Data Science Platform for support. Data generation and analysis were partly supported by funding from National Health and Medical Research Council (NHMRC) investigator grants (2009982 and 2034556). The contents of this manuscript are solely the responsibility of the authors and do not reflect the views of the NHMRC.

Author contributions

A.S.E.C., J.E.P., and D.G.M. designed the study. A.S.E.C. designed and performed the single-cell processing and eQTL mapping pipelines and led the writing of the manuscript. E.S. and C.B. designed and performed the single-cell RNA sequencing experiments. H.A.T. first established the colocalization

pipeline. A.S.E.C and B.B. processed the single-cell RNA sequencing data. B.B. designed and performed the cell state abundance QTL analysis. A.S.E.C, A.H. and J.E.P. designed and performed the cell type specificity analysis. H.L.H. designed and applied the scDeepID tool for cell state identification. H.L.H., A.S.E.C and J.E.P. designed and performed the cell state specific eQTL mapping. A.X. contributed scATAC-seq to support annotations of selected variants. A.X., A.H., H.A.T., K.M.dL. contributed to downstream interpretation. W.Z. provided support for running SAIGE-QTL. M.J.W, H.A.T., M.F., M.H., M.S., A.S., J.M. and V.B provided support with the eQTL mapping pipeline. A.S.L. contributed a curated list of regulatory annotations used to annotate the results. K.W., O.T., M.P.G., S.M.G., and T.N. assisted with participant recruitment and sample collection. M.H., K.B., M.S. and K.M.dL. processed the whole genome sequencing data. B.S.M. and C.U. developed the ethics protocol for the study and coordinated the generation of WGS data. A.M. and M.S. processed the metadata needed to process the WGS data. D.R.N. assisted with the scRNA-seq sample demultiplexing. Z.Q., O.M.S, and H.R.N. contributed to the study design. K.M.dL. and J.P. led the processing of whole genome sequencing data. E.B.D., L.C., and K.K.H.F. designed and performed the annotation of variants using PromoterAI and SpliceAI. A.W.H. designed the Tasmanian Ophthalmic Biobank study. G.A.F. designed the BioHEART study.

All authors contributed to editing the manuscript. J.E.P and D.G.M conceived the project and supervised all aspects of the work.

Competing interests

E.B.D., L.C., and K.K.H.F. are employed at Illumina Inc. D.G.M. is a paid advisor to Insitro and GSK, and receives research funding from Google and Microsoft, unrelated to the work described in this manuscript. G.A.F reports grants from National Health and Medical Research Council (Australia), grants from Abbott Diagnostic, Sanofi, Janssen Pharmaceuticals, and NSW Health. G.A.F reports honorarium from CSL, CPC Clinical Research, Sanofi, Boehringer-Ingelheim, Heart Foundation, and Abbott. G.A.F serves as Board Director for the Australian Cardiovascular Alliance (past President), Executive Committee Member for CPC Clinical Research, Founding Director and CMO for Prokardia and Kardiomics, and Executive Committee member for the CAD Frontiers A2D2 Consortium. In addition, G.A.F serves as CMO for the non-profit, CAD Frontiers, with industry partners including, Novartis, Amgen, Siemens Healthineers, ELUCID, Foresite Labs LLC, HeartFlow, Canon, Cleerly, Caristo, Genentech, Artyra, and Bitterroot Bio, Novo Nordisk and Allelica. In addition, G.A.F has the following patents: “Patent Biomarkers and Oxidative Stress” awarded USA May 2017 (US9638699B2) issued to Northern Sydney Local Health District, “Use of P2X7R antagonists in cardiovascular disease” PCT/AU2018/050905 licensed to Prokardia, “Methods for treatment and prevention of vascular disease” PCT/AU2015/000548 issued to The University of Sydney/Northern Sydney Local Health District, “Methods for predicting coronary artery disease” AU202290266 issued to The University of Sydney, and the patent “Novel P2X7 Receptor Antagonists” PCT/AU2022/051400 (23.11.2022), International App No: WO/2023/092175 (01.06.2023), issued to The University of Sydney.

The other authors declare no competing interests.

Code availability

All scripts to reproduce the analyses presented here are available in the following GitHub repositories:

- Single-cell expression processing and downstream analyses: https://github.com/powellgenomicslab/tenk10k_phase1/

- WGS data processing: https://github.com/populationgenomics/production-pipelines/tree/main/cpg_workflows/large_cohort
- Processing scripts and association testing pipeline using SAIGE-QTL: https://github.com/populationgenomics/saige-tenk10k

References

1. Cerezo, M. *et al.* The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Res* **53**, D998–D1005 (2025).
2. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).
3. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
4. Bonder, M. J. *et al.* Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. *Nat. Genet.* **53**, 313–321 (2021).
5. Mn, W. *et al.* Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation. *BMJ* **372**, n214 (2021).
6. Chen, L. *et al.* Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
7. Kim-Hellmuth, S. *et al.* Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat Commun* **8**, 266 (2017).
8. Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nat. Rev. Genet.* **24**, 535–549 (2023).
9. 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).
10. Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* **606**, 120–128 (2022).
11. Schmiedel, B. J. *et al.* Single-cell eQTL analysis of activated T cell subsets reveals activation and cell type-dependent effects of disease-risk variants. *Sci Immunol* **7**, eabm2508 (2022).
12. Soskic, B. *et al.* Immune disease risk variants regulate gene expression dynamics during CD4 T cell activation. *Nat. Genet.* **54**, 817–826 (2022).
13. Zhou, W. *et al.* Efficient and accurate mixed model association tool for single-cell eQTL analysis. *medRxiv* (2024) doi:10.1101/2024.05.15.24307317.
14. van der Wijst, M. *et al.* The single-cell eQTLGen consortium. *Elife* **9**, (2020).
15. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
16. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of

autoimmune disease. *Science* **376**, eabf3041 (2022).

17. Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
18. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
19. Ferraro, N. M. *et al.* Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, (2020).
20. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82–93 (2011).
21. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
22. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).
23. Dönertaş, H. M., Fabian, D. K., Valenzuela, M. F., Partridge, L. & Thornton, J. M. Common genetic associations between age-related diseases. *Nat Aging* **1**, 400–412 (2021).
24. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
25. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882–D889 (2020).
26. Hitz, B. C. *et al.* The ENCODE Uniform Analysis Pipelines. *bioRxiv* 2023.04.04.535623 (2023) doi:10.1101/2023.04.04.535623.
27. Vu, H. & Ernst, J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biol* **23**, 9 (2022).
28. Euskirchen, G. M. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res* **17**, 898–909 (2007).
29. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664–1669 (2019).
30. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
31. Yang, W.-T., Chen, M., Xu, R. & Zheng, P.-S. PRDM4 inhibits cell proliferation and tumorigenesis by inactivating the PI3K/AKT signaling pathway through targeting of PTEN in cervical carcinoma. *Oncogene* **40**, 3318–3330 (2021).
32. Asano, K. *et al.* A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat Genet* **41**, 1325–1329 (2009).
33. Höglund, P. SLC26A3 and congenital chloride diarrhoea. *Novartis Found Symp* **273**, 74–86; discussion 86–90, 261–4 (2006).

34. Liu, Z. *et al.* Genetic architecture of the inflammatory bowel diseases across East Asian and European ancestries. *Nat. Genet.* **55**, 796–806 (2023).
35. Ishigaki, K. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).
36. 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).
37. Kiryluk, K. *et al.* Genome-wide association analyses define pathogenic signaling pathways and prioritize drug targets for IgA nephropathy. *Nat Genet* **55**, 1091–1105 (2023).
38. Barry, A. *et al.* Multi-population genome-wide association study implicates immune and non-immune factors in pediatric steroid-sensitive nephrotic syndrome. *Nat Commun* **14**, 2481 (2023).
39. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
40. Fernandez-Rozadilla, C. *et al.* Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat. Genet.* **55**, 89–99 (2023).
41. McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017).
42. Wang, A. *et al.* Characterizing prostate cancer risk through multi-ancestry genome-wide discovery of 187 novel risk variants. *Nat. Genet.* **55**, 2065–2074 (2023).
43. Rashkin, S. R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat Commun* **11**, 4423 (2020).
44. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet.* **28**, 715–718 (2020).
45. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
46. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
47. Margoliash, J. *et al.* Polymorphic short tandem repeats make widespread contributions to blood and serum traits. *Cell Genom* **3**, 100458 (2023).
48. Wolf, C. L., Pruett, C., Lighter, D. & Jorcyk, C. L. The clinical relevance of OSM in inflammatory diseases: a comprehensive review. *Front Immunol* **14**, 1239732 (2023).
49. Gilchrist, J. J. *et al.* Natural Killer cells demonstrate distinct eQTL and transcriptome-wide disease associations, highlighting their role in autoimmunity. *Nat Commun* **13**, 4073 (2022).

50. Yang, Y., Fu, K.-Z. & Pan, G. Role of Oncostatin M in the prognosis of inflammatory bowel disease: A meta-analysis. *World J Gastrointest Surg* **16**, 228–238 (2024).
51. Zhang, Y.-M., Zhou, X.-J. & Zhang, H. What Genetics Tells Us About the Pathogenesis of IgA Nephropathy: The Role of Immune Factors and Infection. *Kidney Int Rep* **2**, 318–331 (2017).
52. Lekhraj, R. *et al.* Altered abundances of human immunoglobulin M and immunoglobulin G subclasses in Alzheimer's disease frontal cortex. *Sci Rep* **12**, 6934 (2022).
53. Kim, K. *et al.* Therapeutic B-cell depletion reverses progression of Alzheimer's disease. *Nat Commun* **12**, 2185 (2021).
54. Liu, J.-C. *et al.* Immunoglobulin class-switch recombination: Mechanism, regulation, and related diseases. *MedComm (2020)* **5**, e662 (2024).
55. Migone, T. S. *et al.* TL1A is a TNF-like ligand for DR3 and TR6/DcR3 and functions as a T cell costimulator. *Immunity* **16**, 479–492 (2002).
56. Yamazaki, K. *et al.* Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum Mol Genet* **14**, 3499–3506 (2005).
57. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979–986 (2015).
58. Cordero, R. Y. *et al.* Trans-ancestry, Bayesian meta-analysis discovers 20 novel risk loci for inflammatory bowel disease in an African American, East Asian and European cohort. *Hum Mol Genet* **32**, 873–882 (2023).
59. Bamias, G., Menghini, P., Pizarro, T. T. & Cominelli, F. Targeting TL1A and DR3: the new frontier of anti-cytokine therapy in IBD. *Gut* (2024) doi:10.1136/gutjnl-2024-332504.
60. Sands, B. E. *et al.* Phase 2 Trial of Anti-TL1A Monoclonal Antibody Tulisokibart for Ulcerative Colitis. *N Engl J Med* **391**, 1119–1129 (2024).
61. Richard, A. C. *et al.* Reduced monocyte and macrophage TNFSF15/TL1A expression is associated with susceptibility to inflammatory bowel disease. *PLoS Genet* **14**, e1007458 (2018).
62. Rebuffet, L. *et al.* High-dimensional single-cell analysis of human natural killer cell heterogeneity. *Nat Immunol* **25**, 1474–1488 (2024).
63. Yao, C. *et al.* Sex- and age-interacting eQTLs in human complex diseases. *Hum Mol Genet* **23**, 1947–1956 (2014).
64. Haglund, A. *et al.* Cell state-dependent allelic effects and contextual Mendelian randomization analysis for human brain phenotypes. *Nat Genet* **57**, 358–368 (2025).
65. Rumker, L. *et al.* Identifying genetic variants that influence the abundance of cell states in single-cell data. *Nat Genet* **56**, 2068–2077 (2024).
66. Mangino, M., Roederer, M., Beddall, M. H., Nestle, F. O. & Spector, T. D. Innate and adaptive immune traits are differentially affected by genetic and environmental factors. *Nat Commun* **8**, 13850 (2017).
67. Aguirre-Gamboa, R. *et al.* Differential effects of environmental and genetic factors on T and B

cell immune traits. *Cell Rep* **17**, 2474–2487 (2016).

68. Patin, E. *et al.* Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. *Nat Immunol* **19**, 302–314 (2018).
69. Chen, M.-H. *et al.* Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198–1213.e14 (2020).
70. Astle, W. J. *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.e19 (2016).
71. Javierre, B. M. *et al.* Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384.e19 (2016).
72. Marshall, A. S. J. *et al.* Identification and characterization of a novel human myeloid inhibitory C-type lectin-like receptor (MICL) that is predominantly expressed on granulocytes and monocytes. *J Biol Chem* **279**, 14792–14802 (2004).
73. Neumann, K. *et al.* Clec12a is an inhibitory receptor for uric acid crystals that regulates inflammation in response to cell death. *Immunity* **40**, 389–399 (2014).
74. van Rhenen, A. *et al.* The novel AML stem cell associated antigen CLL-1 aids in discrimination between normal and leukemic stem cells. *Blood* **110**, 2659–2666 (2007).
75. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet* **37**, 109–124 (2021).
76. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. *bioRxiv* 2022.05.07.491045 (2022) doi:10.1101/2022.05.07.491045.
77. Orrù, V. *et al.* Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. *Nat Genet* **52**, 1036–1045 (2020).
78. Vuckovic, D. *et al.* The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231.e11 (2020).
79. Kachuri, L. *et al.* Genetic determinants of blood-cell traits influence susceptibility to childhood acute lymphoblastic leukemia. *Am J Hum Genet* **108**, 1823–1835 (2021).
80. Tanudisastro, H. A. *et al.* Polymorphic tandem repeats shape single-cell gene expression across the immune landscape. *bioRxiv* 2024.11.02.621562 (2024) doi:10.1101/2024.11.02.621562.
81. Kott, K. A. *et al.* Biobanking for discovery of novel cardiovascular biomarkers using imaging-quantified disease burden: protocol for the longitudinal, prospective, BioHEART-CT cohort study. *BMJ Open* **9**, e028649 (2019).
82. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
83. Fleming, S. J. *et al.* Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* **20**, 1323–1335 (2023).
84. Neavin, D. *et al.* Demuxafy: improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. *Genome Biol.* **25**, 94 (2024).

85. 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).
86. Bais, A. S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150–1158 (2020).
87. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using. *F1000Res.* **10**, 979 (2021).
88. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* **20**, 264 (2019).
89. Michielsen, L., Reinders, M. J. T. & Mahfouz, A. Hierarchical progressive learning of cell identities in single-cell data. *Nat. Commun.* **12**, 2799 (2021).
90. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
91. Xu, C. *et al.* Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell* **186**, 5876–5891.e20 (2023).
92. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
93. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
94. Liu, Y. *et al.* ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am J Hum Genet* **104**, 410–421 (2019).
95. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
96. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Soft.* **36**, 1–48 (2010).
97. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
98. Kingma, D. P. & Welling, M. *An Introduction to Variational Autoencoders*. (2019).
99. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models using lme4. (2014).
100. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
101. Reshef, Y. A. *et al.* Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nat Biotechnol* **40**, 355–363 (2022).