Supplementary Materials for

Atlas of cell type-specific post-transcriptional genetic impacts in immune-related diseases

Ting Zhang[1]†, Hui Chen[1,2]†, Shuxin Chen[1]†, Ruofan Ding[1], Dinglin Luo[1], Xudong Zou[1], Kewei Xiong[1], Yinuo Wang[1], Shibin Hu[3], Jian Yang[4], Gao Wang[5], Stephen Montgomery[6], Qin Li[7]*, Lei Li[1]*

Corresponding authors: Lei Li, lei.li@szbl.ac.cn; Qin Li, qinli@pennmedicine.upenn.edu

The PDF file includes:

Materials and Methods
Figs. S1 to S22
References

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S8

Materials and Methods

Data collection and processing

We analyzed single-cell transcriptomic and genotype data from seven cohorts (*1*-*7*): (1) **Popcell**: 222 healthy donors of Central African (AFR), European (EUR), and East Asian (EAS) descent. Samples were either resting (non-stimulated condition; NS) or stimulated with SARS-CoV-2 (COV) or influenza A virus (IAV) for six hours; (2) **OneK1K** with 982 healthy donors; (3) **scSLE** with 162 systemic lupus erythematosus cases and 99 healthy controls of Asian or European ancestry; (4) **scIAV** with 90 healthy donors of European or African ancestry, including mock-exposed or IAV-infected cells; (5) **scTB,** T-cell scRNA-seq data from 259 healthy Peruvian individuals with a previously resolved *Mycobacterium tuberculosis* infection; (6) **DICE** with 89 healthy donors of diverse ancestries; (7) **scTcell**, scRNA-seq data of $CD4^+$ T cells from 119 healthy individuals of British ancestry under unstimulated treatment and 16h, 40h, and 5d stimulation with anti-CD3/anti-CD28 human T-Activator Dynabeads (Invitrogen).

Single-cell RNA sequencing analysis

Raw sequencing data were processed using CellRanger (v6.1.1, 10x Genomics) (*8*) with the GRCh38 reference human genome. Briefly, raw reads were aligned to the Ensembl v104 reference genome using STAR (v2.7.10b) (*9*), and filtered feature-barcode matrices were generated from confidently assigned cells. For the scRNA-seq data with virus stimulation, we implemented a specialized mapping strategy where reads were mapped to a concatenated reference genome combining human GRCh38, SARS-COV-2 (hCoV-19/France/GE1973/2020) and IAV (A/Puerto Rico/8/1934(H1N1)) genome, while scIAV cohort data were aligned to merged human GRCh38 and the IAV (Cal/04/09) reference genome. These concatenated reference genomes were generated using the *mkref* function in CellRanger.

Quality control of mapped data followed a systematic two-step process. First, potential doublets were identified and removed using Scrublet (v0.2.3) (*10*) with optimized parameters (expected_doublet_rate = 0.06; min_counts = 2; min_cells = 3; min_gene_variability_pctl = 85, n_prin_comps = 30). Second, a Seurat (v5.0.1) (*11*) object was created containing the expression matrix, cell barcode metadata, and Scrublet-inferred doublet information. This object was used to calculate per-cell mitochondrial RNA content. Cells were filtered out if they met any of the following criteria (*1*): fewer than 1,000 UMI counts, fewer than 500 detected genes, greater than 5,000 detected genes, greater than 20% mitochondrial content, or Scrublet score exceeding 0.3.

Cell type annotation and integration

We developed a computational pipeline for cell type annotation and data integration. Initial data processing was performed using the Scanpy pipeline (v1.9.6) (*12*), RNA quantification from CellRanger and curated annotations from original publications were merged. Raw count data were log-normalized, followed by the identification of 3,000 highly variable genes. These selected genes were scaled and used for PCA generation. Cell type annotations were harmonized across the seven scRNA-seq datasets using CellHint (v0.1.1) (*13*). The hint.harmonize function was applied with specific parameters ('dataset = source', 'cell_type = cell_type', 'use_rep = X_pca', and 'use_pct = True') to achieve unified annotations. Integration was performed using the cellhint.integrate function, which constructs an annotation-aware cell neighborhood graph. This graph facilitated dimensionality reduction while preserving cell type relationships and was subsequently embedded into two-dimensional space using UMAP.

To validate the annotation accuracy, we examined the expression of canonical marker genes for major immune cell types (*14*), including *CD3D* for T cells, *CD4* for $CD4^+$ T cells, *CD8A* and *GNLY* for $CD8^+$ T cells, *NCAM1* and *GNLY* for NK cells, *MS4A1* and *HLA-DRA* for B cells, *MZB1* for Plasma B cells, and *CD14* and *HLA-DRA* for monocytes.

Implementation of SERA for molecular phenotype definition

Single-cell gene expression quantification. Gene expression was quantified using a pseudo-bulk approach based on optimized eQTL mapping workflows (*15*). Normalization was performed using the NormalizeData function in Seurat (v5.0.1) (*11*) through a multi-step process. First, we adjusted for sequencing depth variation by dividing the UMI counts of each gene by the total UMI counts per cell. These values were then scaled by a factor of 10,000 and log-transformed (log1p). Pseudo-bulk samples were generated using the AverageExpression function. We excluded genes detected in fewer than 1% of cells for each cell type and removed donors containing fewer than 10 cells to ensure robust quantification.

Identifying alternative poly(A) sites (PASs) from scRNA-seq data. To detect PASs from immune scRNA-seq datasets, we first extracted and collapsed overlapping last exons from the RefSeq annotation (GCF_000001405.40-RS_2016_10), resulting in 68,542 distinct last exons representing 54,132 genes. We then compiled PAS annotations from multiple sources: RefSeq gene annotations (174,310 PASs), 35,779 PolyA_DB (*16*) annotations from UCSC (https://genome.ucsc.edu/cgi-bin/hgTables), and previously published 3′-end sequencing data (*17*) from 102 human primary cell types across 60 tissue types. We also analyzed 4.3 billion reads containing A/T stretches. All PAS candidates within 300 nt were merged, retaining the PAS with the maximum value. The potential false-positive annotations caused by internal priming were filtered out. Finally, we obtained a non-redundant list of PASs. In addition, we classified genes and last exons into three categories according to the PASs distribution (*18*), including the last exon with single PAS (SPAS), last exon with Tandem 3′UTR (TUTR), and gene with alternative last exon (ALE), and only TUTR and ALE classes were used for APA quantification.

APA quantification. We used scUTRquant (*17*) to calculate APA usage with some modifications. Considering the peak size of scRNA-seq data is approximately 400bp (*19*), we constructed our single-cell PAS annotation by extending 400 bp upstream from the merged PASs as the final PAS reference. Read count in each PAS and each cell was quantified using umi_tools (v1.1.4) (*20*). We first aggregated the read counts of each PAS *m* across all cells within cell type *n* for individual *k*, denoted as P_{mnk}. Similarly, the read counts for each last exon (hereafter referred to as isoform) *m* were aggregated across all cells within the same cell type and individual, denoted as I_{mnk}. For simplicity, the subscripts *n* and *k* are omitted in subsequent descriptions. PDUI (percentage of distal PAS usage index) was originally introduced by the DaPars algorithm (*21*) to quantify APA events with a range of 0−1. A higher PDUI value indicates more distal usage of a transcript. In addition to the PDUI, we further developed PPUI (percentage of PAS usage index), PIUI (percentage of isoform usage index), and PDIUI (percentage of distal isoform usage index) to enhance the accuracy of APA usage quantification. Specifically, for a TUTR containing a total of *D* PASs, PPUI and PDUI were calculated as follows:

$$PPUI = \frac{P_i}{\sum_{l=1}^{D} P_l}$$

$$PDUI = \sum_{i=1}^{D} w_i \frac{P_i}{\sum_{l=1}^{D} P_l},\ w_i = \frac{i-1}{D-1}$$

where the index *i* of a PAS is determined by its distance from the transcription start site (TSS), with *i* increasing as the distance from the TSS increases. For PDUI, different weights w_i were assigned to each PAS usage, with larger distances from the TSS corresponding to higher weights.

Similarly, for an ALE-type gene containing a total of *D* isoforms, PIUI and PDIUI were calculated as follows:

$$PIUI = \frac{I_j}{D\sum_{l=1}^{D} I_l}$$

$$PDIUI = \sum_{i=1}^{D} w_j \frac{I_j}{\sum_{l=1}^{D} I_l},\ w_j = \frac{j-1}{D-1}$$

where the index *j* of an isoform (the total count of all included PASs) is determined by its distance from the transcription start site (TSS), with *j* increasing as the distance from the TSS increases. For PDIUI, we also assigned different weights w_j to each isoform usage, where larger distances from the TSS are associated with higher weights. Based on the quantitative results of all cell types and all individuals, we filtered the independent PPUI and PIUI by filtering the correlations between PPUI and PDUI and between PIUI and PDIUI for the same gene, respectively (the Pearson's *r* of both in all samples was less than 0.5). For each cell type, an APA matrix containing the above four types of quantification for each individual was generated. APA events detected in fewer than 10% of individuals in each cell type were considered to be lowly confident and were excluded from further analyses, leading to a final set of 25,851 APA events at the cell type level.

RNA editing site identification. The reference of 15,681,871 RNA editing sites was downloaded from REDIportal (V2, http://srv00.recas.ba.infn.it/atlas/index.html). To minimize false positives, we excluded 2,316,608 editing sites that overlapped with germline SNPs annotated in dbSNP (build 151), resulting in 13,365,263 unique editing sites. RNA editing sites were identified from our collected single-cell RNA-sequencing data using the REDItoolKnown.py function in REDItools (v1.0.0) (*22*) with the following parameters: *-c* 5 (minimal read coverage), *-v 2* (minimum number of reads supporting the variation), *-q 30* (minimum read quality score) *-m 30* (minimum mapping quality score), *-O 5* (homopolymeric-span length)*, -s 1* (strand inference), while all other parameters were set to default. We retained only sites from A to G and removed sites that were multiallelic, overlapped with genomic variants in individual genotype data, or were supported by less than 50% of individuals in each dataset.

RNA editing quantification. To quantify RNA editing levels at the detected sites for each cell type in each individual, we first extracted the read pileup at each editing site using the *mpileup* function in SAMtools (v1.10) (*23*). The RNA editing level at each site was then calculated as the ratio of edited reads to the total number of reads. RNA editing detected in fewer than 10% of individuals in each cell type was excluded from further analyses, resulting in 62,362 RNA editing sites across cell types.

Genotype data processing and quality control

The genotype data were collected and processed through a standardized pipeline across all cohorts. For the popCell cohort, genome-wide genotype data from 473 samples were

provided by investigators and subset to 222 samples with matching scRNA-seq data. For the OneK1K and scTcell cohorts, IDAT format genotype data were obtained from GEO (accession: GSE196829) and EGA (accession: EGAD00010002291), respectively. These IDAT files were processed using Illumina's IAAP Genotyping Command Line Interface (v1.1.0) and BCFtools (v.1.16) (*24*) gtc2cvf plugin for VCF conversion, followed by export to PLINK format using GenomeStudio PLINK Input Report Plug-in (v2.00a3). For the remaining cohorts, curated VCF format genotype data were obtained from dbGaP (scSLE: phs002812.v1.p1; DICE: phs001703.v4.p1; scTB: phs002025.v1.p1) and Zenodo (scIAV: doi:10.5281/zenodo.4273998). To ensure data consistency, the popCell and scIAV VCF files were normalized using bcftools (v1.8) with reference allele checking. For other datasets, coordinates were lifted from hg19 to GRCh38 using Picard-LiftoverVcf with the 'RECOVER_SWAPPED_REF_ALT' option. We then performed chromosome-wise splitting of VCF files and conducted imputation using the Michigan Imputation Server (*25*) with the 1000 Genomes Project (hg38 build) reference panel. Stringent quality control criteria were applied to retain high-confidence variants: MAF > 0.05, imputation accuracy $R^2 > 0.3$, Hardy-Weinberg equilibrium $p > 10^{-6}$, and call rate > 99%. This yielded high-quality variant sets for each cohort: popCell (10,229,106), OneK1K (6,320,570), scSLE (6,827,122), scIAV (7,499,155), DICE (6,177,822), scTB (5,451,861), and scTcell (6,226,949), representing a 2.72-fold increase in variants post-imputation. Cross-cohort analysis revealed 39.3% of variants were shared across all datasets, while 21.42% were unique to individual studies. Population structure was assessed using principal components analysis via QTLtools (v1.2) (*26*) in pca mode, identifying 1,403 samples of European (EUR) ancestry, 255 of American (AMR) ancestry, 189 of East Asian (EAS) ancestry, and 131 of African (AFR) ancestry.

sc-xQTL mapping

For sc-eQTL, sc-aQTL, and sc-edQTL (sc-xQTL) discovery, we grouped samples by cell type and condition. All molecular phenotypes (gene expression values, APA usages, and RNA editing levels) were inverse-normal rank-transformed to a standard normal distribution. We retained only phenotype matrices containing at least 50 individuals for QTL mapping. The top five genotype principal components (PCs) were sufficient to account for the population structure within each cohort. Additional hidden covariates were identified from phenotype data using PCAForQTL (*27*). The final covariate set for sc-xQTL mapping included these top five genotype PCs, phenotype-derived PCs, chromosomal sex, age, and cell composition information. We tested associations for each adjusted phenotype with variants around 1 Mb using a standard linear model via tensorQTL (v1.0.8) (*28*). The most highly associated variant per gene/APA/RNA editing was identified using empirical *P* values based on beta-distribution with a permutation test (nperm = 1,000). These *P*-values were corrected for multiple testing across genes using Storey's q value. Genes with significant cis-xQTL (q value < 0.05) were called xGenes. All possible phenotype-variant pairs in cis were calculated using the *cis.map_nominal* model. These nominal p-values were adjusted using the *qvalue* (v.2.28.0) package.

Identification of cell-type-specific genetic regulation

The number of detected sc-xQTL per cell type highly depended on the number of cells and sample size (*2, 4*). To overcome sc-xQTL mapping limited power due to cell proportion and sample size, we adopted the mash method (*29*) to increase power using sharing information across cell types. For each type of xQTL, the estimated z-score of the lead variant per gene from tensorQTL permutation (the 'strong' set). We set missing z-scores as zero and standard error as 10^6 via the zero_Bhat_Shat_reset = 10e6. We randomly selected 1,000,000 variant-gene pairs to produce a null input (the 'random' set). We estimated the

strong set learned covariance matrices using the top 5 principal components and the extreme deconvolution method in mashr and the random set learned canonical covariance matrix (as recommended in https://stephenslab.github.io/mashr/articles/eQTL_outline.html). We then ran the MASH model to fit a mixture model and compute posterior summaries for the strong set. Finally, we obtained the posterior effect size, posterior standard deviations, and local false sign rate (lfsr) from the MASH model. We used the estimated lfsr to discover xGene as a complement of tensorQTL and considered xQTL was significant if the lfsr was < 0.05, which we defined as an additional xGene (*30*).

QTL enrichment in functional annotation

We generated comprehensive genomic annotations for our sc-xQTL variants using Ensembl's Variant Effect Predictor (VEP). These annotations included genomic locations (3′UTR, 5′UTR, intron, exon, splice site), functional predictions (missense, stop gained, etc.), regulatory features (promoter), transcription factor binding peaks, DNaseI hypersensitivity sites, transcription factor motif overlaps, and ADAR1 binding sites. To test for functional enrichment of sc-xQTLs within these annotated regions, we employed TORUS (*31*) (QTL Discovery utilizing Genomic Annotations), which is freely available at https://github.com/xqwen/torus. The enrichment analysis was performed using the command (*torus -d qtl_statistics -annot annotation_file -est)*.

Comparison between sc-xQTL and bulk xQTL

The bulk xQTL data of whole blood, including cis-eQTL, cis-aQTL, and cis-edQTL, were from the Genotype-Tissue Expression (GTEx) project. The cis-aQTL and cis-edQTL data were obtained from our prior research (*32, 33*). The performance of sc-xQTL mapping is measured by the number of significant sc-xQTLs and their replication with bulk xQTL data.

Fine-mapping of sc-xQTLs

Fine mapping can narrow the genetic association signals to a few potential causal variants. We use the Sum of Single Effects (SuSiE) (*34*) algorithm for fine-mapping analysis. SusieR implements the Susie algorithm and individual-level genotype and phenotype data as input. For each cohort, we utilized the genotype file to estimate LD. We set the number of putative causal signals to 10 and defined credible sets consisting of variants with a cumulative 95% posterior probability.

Bayesian-based mediation analysis

For each X-M-Y triplet, where X is a genetic variant significantly associated with both M and Y molecular phenotypes, M is the APA, and Y is the gene expression or RNA editing, respectively. We applied bmediatR (v.0.1.3) (*35*) for mediation analysis, and the covariates for each QTL data corresponding to population structure and technical factors were used. To test the causal forward model, the ln_prior_c parameter for the model prior was set to "complete" was set. We defined the relationship as causal forward if the sum of the posterior probabilities of complete and partial mediation was higher than 0.5, and as reactive if the sum of the posterior probabilities of complete reactive and partial mediation models was higher than 0.5.

Identification of cell type-specific xQTLs

To assess xQTL sharing across cell types and identify cell type-specific effects, we utilized the lfsr statistic calculated by mashr. We classified eQTL as lineage-specific and cell type-specific according to their significance across cell types. A cell type-specific xQTL was defined by meeting all of the following criteria: (1) significant lfsr in the target cell type (lfsr

< 0.05), (2) *P* value < 10^{-3} in the target cell type, and (3) expressed in other cell types. Similarly, we defined lineage-specific xQTL as those significant only within a specific lineage.

Identification of population-biased and population-specific eQTLs

To identify population-biased and population-specific xQTLs, we applied a two-tiered analytical framework. First, population-biased xQTLs were defined as genetic variants showing significant interaction effects between genotype and population on three transcriptional phenotypes. The tensorQTL interaction model (*28*) was used to test genotype-population interaction terms while adjusting for principal components and covariates. Variants with FDR < 0.05 were retained. To ensure robust allele frequency representation, only variants with MAF > 0.1 across all analyzed populations were included in this analysis. For population-specific eQTLs, we prioritized variants exhibiting allele frequency divergence and evidence of causality from fine mapping. Specifically, variants were required to meet the following criteria: (1) MAF > 0.05 in one population and MAF < 0.01 in all other populations, ensuring substantial frequency differentiation; (2) inclusion in a 95% credible set from fine-mapping results, and (3) > 50% of variants belong to one specific population. These filters collectively ensured that identified variants were statistically robust and biologically plausible as population-specific regulatory drivers.

Sex-biased sc-xQTL

Sex-biased xQTLs with a significant genotype-by-sex (G × S) interaction effect were identified using the tensorQTL interaction model (*28*). For all sc-xQTL mapping inputs, we remove sex as a covariate and use a linear regression model to test for genotype-by-sex (G × S) interaction while adjusting for known and unknown confounders. Variant-gene pairs with adjusted *P*-value < 0.05 were sex-biased sc-xQTLs.

Dynamic sc-xQTL in each cell type

We computed gene expressions to construct a two-dimensional space using PHATE (potential of heat diffusion for affinity-based transition embedding). Based on the principal component of PHATE, we computed the continual progression (pseudo-time or trajectory) to order cells using slingshot. We checked the pseudo-time curve, grouped each cell type into six quantile bins, and quantified gene expression/APA usage/RNA editing levels of each bin with the pseudo-bulk method. We used both linear and nonlinear models to identify dynamic sc-xQTLs as previously described. The linear model used a Gaussian linear mixed model with individual-specific random effects that assessed the interaction between genotype and quantile representing the pseudo-time trajectory, while controlling for the linear effects of both genotype and quantile.

$$Y = \beta_0 + \beta_1 * geno + \beta_2 * sex + \beta_3 * age + \sum_{j=1}^{5} \beta_{3+j} * PC_j + \sum_{k=1}^{5} \beta_{9+j} * PC_j + + \beta_{14} * pseudotime + \beta_{15} * geno * pseudotime + \varepsilon$$

Where Y corresponds to the gene expression/APA usage, *geno* is the genotype information, pseudotime is the number of six quantiles, and genotype * pseudotime is the interaction term. In addition, the nonlinear model adds a second-order polynomial to the linear dynamic model.

$$Y = \beta_0 + \beta_1 * geno + \beta_2 * sex + \beta_3 * age + \sum_{j=1}^{5} \beta_{3+j} * PC_j + \sum_{k=1}^{5} \beta_{9+j} * PC_j ++ \beta_{14} * pseudotime + \beta_{15} * pseudotime^2 + \beta_{16} * geno * pseudotime + \beta_{17} * geno * pseudotime^2 + \varepsilon$$

Where pseudotime^2 is the vector of quadratic values of corresponding quantile, and genotype*pseudotime^2 is the quadratic interaction term between genotype and pseudotime. We fitted a reduced model by dropping the genotype * pseudotime interaction in the linear model and genotype * pseudotime + genotype * pseudotime^2 in the nonlinear model, and computed a likelihood ratio test for the full and reduced models using R anova. Finally, a multiple testing correction for the *P*-values and FDR < 0.05 was used to identify significant dynamic sc-xQTL.

Preprocessing of GWAS summary statistics

We collected GWAS Summary statistics files from previous GWAS studies for 15 autoimmune diseases (allergic disease, AS, asthma, AITD, celiac disease, CD, eczema, IBD, MS, primary biliary cholangitis (PBC), psoriasis, RA, SLE, T1D, and ulcerative colitis (UC)), one infectious disease (COVID-19), two anthropometric measurements (body mass index and height), two brain disorders (Alzheimer's disease and schizophrenia), one cancer (breast cancer), one metabolic disease (Type 2 diabetes), and one heart disease (Coronary artery disease) (table S4). To ensure consistent formatting and compatibility across different cohorts, the raw GWAS summary statistics were harmonized and imputed using combined genotypes, including 1000 Genomes reference genotypes (*36*) and seven cohorts. The code for harmonization and imputation was available from https://github.com/hakyimlab/summary-gwas-imputation/wiki.

Heritability enrichment of sc-xQTLs in complex traits

We curated publicly available GWAS summary statistics of 23 diseases covering various diseases, including anthropometric traits, autoimmune diseases, brain disorders, cancer, and infectious diseases. We utilized the S-LDSC v.1.0.1 (*37*) to estimate the heritability enrichment attributable to sc-xQTLs fitting with 52 other functional categories in the baseline-LD model for the 23 phenotypes mentioned above. Briefly, the most significant sc-xQTL variants were assigned an annotation value of 1, and the remaining variants were assigned a zero value in each cell type. The LD scores of the SNPs were calculated using the European 1000 Genome Project (PHASE 3) with a window size of 1cM. The heritability enrichment of sc-xQTLs was calculated as the ratio of the proportion of heritability explained by the cell type to the proportion of xVariants within the cell type.

Colocalization analysis

Genome-wide significant loci were extracted from GWAS summary statistics using PLINK v1.9 (*38*) with the –clump function (parameters: --clump-r2 0.6 --clump-p1 5e-8 --clump-p2 0.05 --clump-kb 1000) to identify independent variants. Independent loci were identified using PLINK with a linkage disequilibrium threshold of $r^2 > 0.6$. Adjacent loci separated by < 250 kb were merged to account for potential overlap. For each GWAS locus, we applied SuSiE-coloc (*34*) and standard coloc (*39*) to evaluate shared causal variants between GWAS and cis-QTL signals and identify gene/APA/RNA editing sites near risk SNPs within a 1 Mb region. PP4 score, representing the potential same causal variants shared by GWAS variants with sc-xQTLs, was used to define significant colocalization (PP4 > 0.75).

Transcriptome-wide association study

We extended traditional gene expression to the single-cell level by conducting sc-xTWAS across three molecular phenotypes in each cell type using FUSION (*40*). For feature prediction model construction, we employed a mixed-linear model incorporating SNPs within 1 Mb regions and the same covariates used in QTL mapping to estimate cis-SNP heritability. Only models with significant *P*-values ($P < 0.05$) and reliable heritability estimates (cis-h2) were retained for subsequent analyses. Within the FUSION framework, we evaluated four different weight calculation models: best linear unbiased predictor (BLUP), elastic-net regression (Elastic Net), lasso regression (LASSO), and single best eQTL (Top1). Model selection was performed using fivefold cross-validation to identify the best prediction accuracy. Finally, we applied high-quality prediction models (prediction accuracy > 50%) to GWAS summary statistics from 16 immune-related diseases and identified significant TWAS associations (FDR < 0.05).

Association score of gene in 16 immune-related diseases

We used the Open Targets Platform database (https://platform.opentargets.org/) to estimate the gene-disease associations (*41*). We restricted diseases to above 16 immune-related diseases and then obtained gene-disease association scores ranging from 1,145 to 10,638 genes. The association score ranges from 0 to 1 and summarizes all the aggregated evidence using the harmonic sum, including genetics, genomics, transcriptomics, drug information, animal models, and scientific literature evidence.

Identification of shared genetic architecture between protein, transcriptional phenotypes, and immune-related diseases

To assess the downstream effects of transcriptional phenotypes, we analyzed three genome-proteome-wide association studies for genetic variant–protein target associations (*42-44*), which were obtained from Synapse (https://www.synapse.org/#!Synapse:syn51824537 and https://www.synapse.org/#!Synapse:syn51365303) and GWAS Catalog (Accession ID: GCST90274758 to GCST90274848). For each proteo-genomic summary data, we defined cis-pQTLs as variants within a 1Mb window of the gene's transcription start site (TSS) and retained only variants present in our genotype data. After comparing Pearson's *r* of *P*-values for overlapping variant-protein pairs across datasets and removing redundant targets, we identified 2,653 genes with proteo-genomic summary data.

We employed statistical colocalization (*34*) to identify shared genetic signals between circulating protein abundance and expression/APA/RNA editing across all cell types. Significant shared signals were defined with PP4 > 0.5, *P*-value < 5e-8 in pQTLs, and FDR < 0.05 in sc-xQTLs. In addition, we used the same colocalization method to analyze 16 immune-related disease GWAS data and 2,653 proteo-genomic summary data. We conducted multi-trait colocalization analysis to detect shared genetic regulation across disease, protein, and transcriptional phenotypes, with posterior probability > 50% indicating significant colocalization.

Drug target identification

We compiled approved and potential druggable targets from two public resources: the Therapeutic Target Database (https://db.idrblab.net/ttd) and the druggable genome (*45, 46*). The Therapeutic Target Database encompasses 3,395 genes, including 426 successful targets, 1,014 clinical trial targets, 212 preclinical/patented targets, and 1,479 literature-reported targets. The druggable genome comprises 4,479 genes categorized into three tiers: tier 1 (1,427 genes corresponding to approved drugs or clinical-phase candidates), tier 2 (682 genes

encoding targets with drug-like compounds sharing ≥ 50% identity with approved drug targets), and tier 3 (2,370 genes representing druggable targets with lower similarity to approved drug targets).

Dual-luciferase reporter construction

Renilla luciferase in psiCHECK-2 Vector (Promega, cat#: C8021) was used as a primary reporter gene. The firefly reporter cassette served as an intra-plasmid transfection normalization reporter (*47*). To investigate the role of 3′aQTLs in regulating their target genes, the 3′UTR containing reference and alternative variants of *SACM1L*, *STAT6*, and *DLD* was cloned downstream of the Renilla luciferase translational stop codon of the psiCHECK-2 Vector, respectively. To investigate the role of poly(A) site usage in regulating target genes, the short 3′UTR and long 3′UTR of *SACM1L*, *STAT6,* and *DLD* were cloned downstream of the Renilla luciferase translational stop codon of the psiCHECK-2 Vector, respectively.

Rapid amplification of cDNA ends (RACE)

Human embryonic kidney (HEK) 293T cells (Cell Resource Center of Shanghai Institutes for Biological Sciences) were cultured in Dulbecco's Modified Eagle Medium (DMEM; Gibco, cat #: C11995500BT), containing 10% fetal bovine serum (FBS; Gibco, cat #: C10010500BT), 100-IU/mL penicillin, and 100-μg/mL streptomycin (Gibco, cat#:15140-122). Cells were cultured at 37 °C in a 5% CO_2 atmosphere with 100% humidity. HEK 293T cells were seeded 1 day before transfection. The previously described luciferase psiCHECK-2 vector was transfected into cells using Lipofectamine 3000 Transfection Reagent (Invitrogen, cat#: L3000015). After 48h, cells were harvested for RNA extraction. Total RNA was extracted using TRIzol reagent (Sigma, cat#: T9424) according to the manufacturer's instructions. Next, the full length of 3′ UTR was identified and amplified from the total RNA using the GoScript™ Reverse Transcriptase (Promega, cat#: A5001) with oligo dT18-XbaKpnBam primer following the manufacturer's protocol. Finally, 3′ RACE was carried out using the check-luc forward primer and XbaKpnBam reverse primer to distinguish exogenous RNA from endogenous RNA. RACE-PCR products were separated on a 2% agarose gel. Bands were excised and extracted using the TIANgel mini purification kit (Tiangen, cat#: DP209) were cloned into the PCE2 vector by the 5min TOPO-Blunt Cloning Kit (Vazyme, cat#: C602) for Sanger sequencing.

Dual-luciferase reporter assay

HEK 293T cells were seeded 1 day before transfection. The previously described luciferase psiCHECK-2 vector was transfected into cells. Forty-eight hours post-transfection, firefly and renilla luciferase activities were measured by Dual-Luciferase Assay System (Promega, #E1980) on a BioTek Synergy H1 plate reader with full waveband. Each assay was measured in three independent replicates.

The sc-xQTL atlas portal

sc-xQTL atlas (https://bioinfo.szbl.ac.cn/sc-xQTLatlas), a single-cell xQTL knowledge portal, was constructed to present detailed information on genetic regulation of gene expression/APA/RNA editing in various cell types. The interactive web pages were implemented using HTML, CSS, JavaScript, and Python, with several JavaScript libraries (react-ploty.js, DataTable.js, and IGV.js) and the Bootstrap framework (a popular framework for developing interactive websites). The sc-xQTL atlas portal is freely accessible online, and there is no requirement to register or log in.

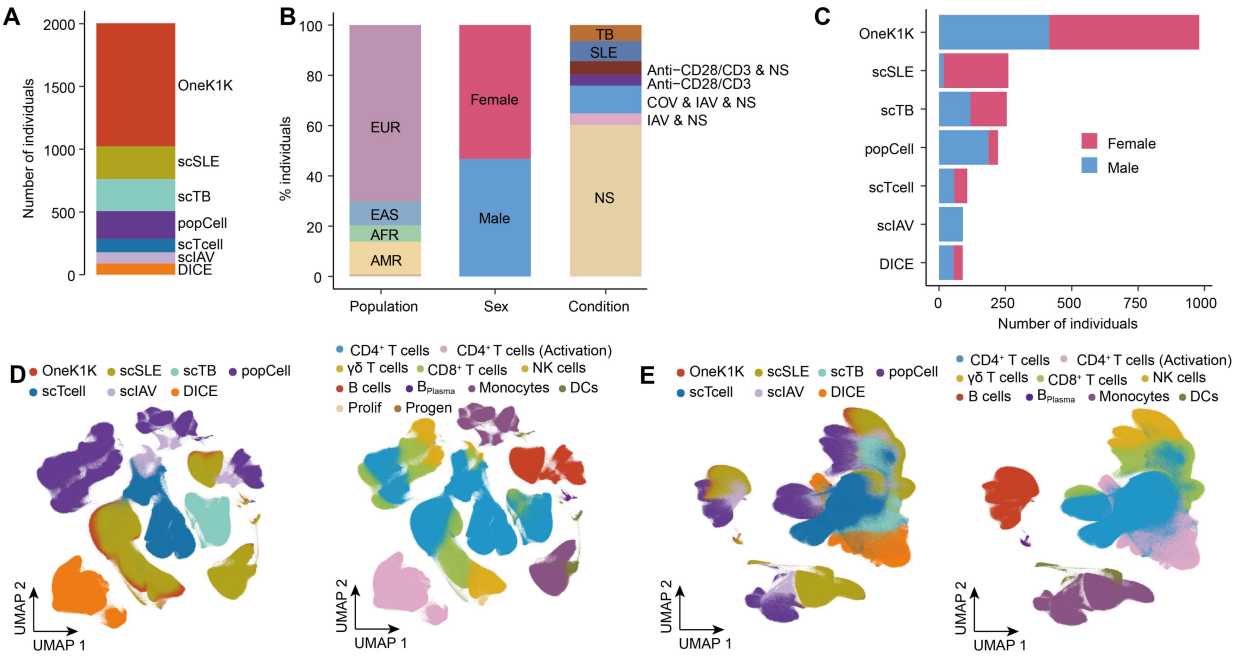

Fig. S1. Multi-cohort harmonization and integration for immune cell profiling. (**A**) Distribution of sample sizes across seven independent immune datasets. (**B**) Demographic composition of collected datasets. Stacked bar plots indicate proportions of ancestral populations (EUR—European, EAS—East Asian, AFR—African, and AMR—Admixed American), biological sex (male/female), and experimental/clinical conditions: NS—Non-stimulated, IAV—Influenza A virus, COV—SARS-CoV-2, anti-CD28/CD3—T cell receptor stimulation, SLE—systemic lupus erythematosus and TB—tuberculosis. (**C**) Histogram of sex for different cohorts. (**D**) Uncorrected single-cell transcriptomic landscape. Uniform Manifold Approximation and Projection (UMAP) visualization of all cells before batch correction, colored by source cohort (left) and ten annotated cell types (right). (**E**) Batch-corrected cellular atlas following CellHint integration. Coordinated UMAP embedding mitigates technical variability (left), with refined clustering identifying nine conserved cell types (right).

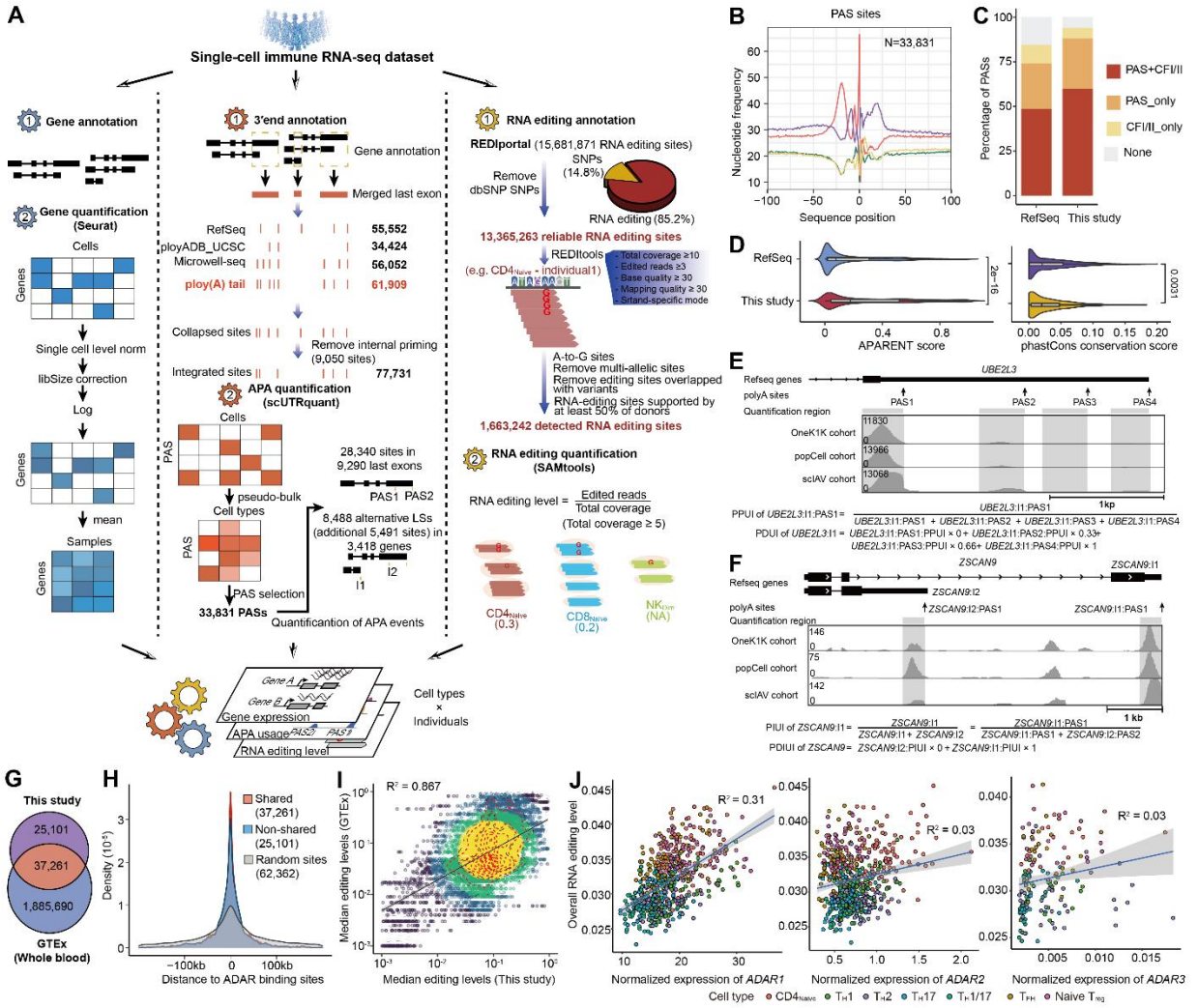

Fig. S2. Atlas of single-cell RNA-based molecular traits. (**A**) Pipeline for scRNA-based molecular traits. Expressed genes were quantified using Seurat (left). APA events were identified via poly(A) site (PAS) annotation and quantification (middle), and RNA editing sites were called using REDItools and SAMtools (right). (**B**) Positional frequency of polyadenylation sites (PASs). Nucleotide composition surrounding cleavage sites (positions -100 to +100 relative to PAS) shows canonical patterns. (**C**) Quality control of PAS identification showing comparison of motifs. (**D**) Comparison of APARENT score and conservation scores between Refseq and processed PASs. (**E**) Tandem APA regulation in *UBE2L3*. scRNA 3′ coverage (top) reveals four PASs (arrows) detected in our dataset. The calculation process of APA quantification is annotated below the figure. (**F**) Intronic APA switching in *ZSCAN9*. *ZSCAN9* has multiple last exons and can be detected in our dataset. The quantitative analysis of two exons is annotated at the bottom of the figure. (**G**) Comparison of RNA editing sites between this study and GTEx whole blood. (**H**) Distribution of RNA editing to ADAR1 binding sites. (**I**) Quality control of RNA editing identification showing conservation scores in different RNA editing sources. (**J**) Correlation between RNA editing levels and three *ADAR* genes.

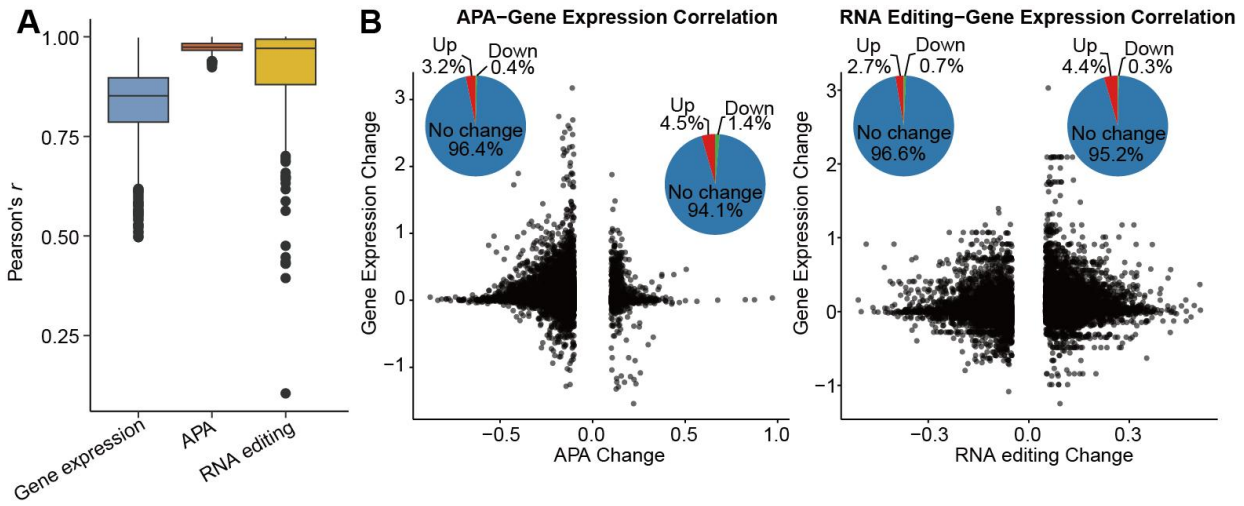

Fig. S3. Cross-cohort consistency and regulatory decoupling of three molecular traits. (**A**) Inter-cohort correlation of single-cell molecular traits. Boxplots depict Pearson's *r* for gene expression (blue), APA usage (red), and RNA editing (orange) across seven cohorts. (**B**) Differential molecular events independent of transcriptional changes. Left: Volcano plot of differential APA events ($|\Delta PDUI| > 0.1$, FDR < 0.05), with no significant correlation to gene expression fold changes. Right: Volcano plot of differential RNA editing ($|\Delta editing\ level| > 0.05$, FDR < 0.05), showing minimal overlap with differentially expressed genes.

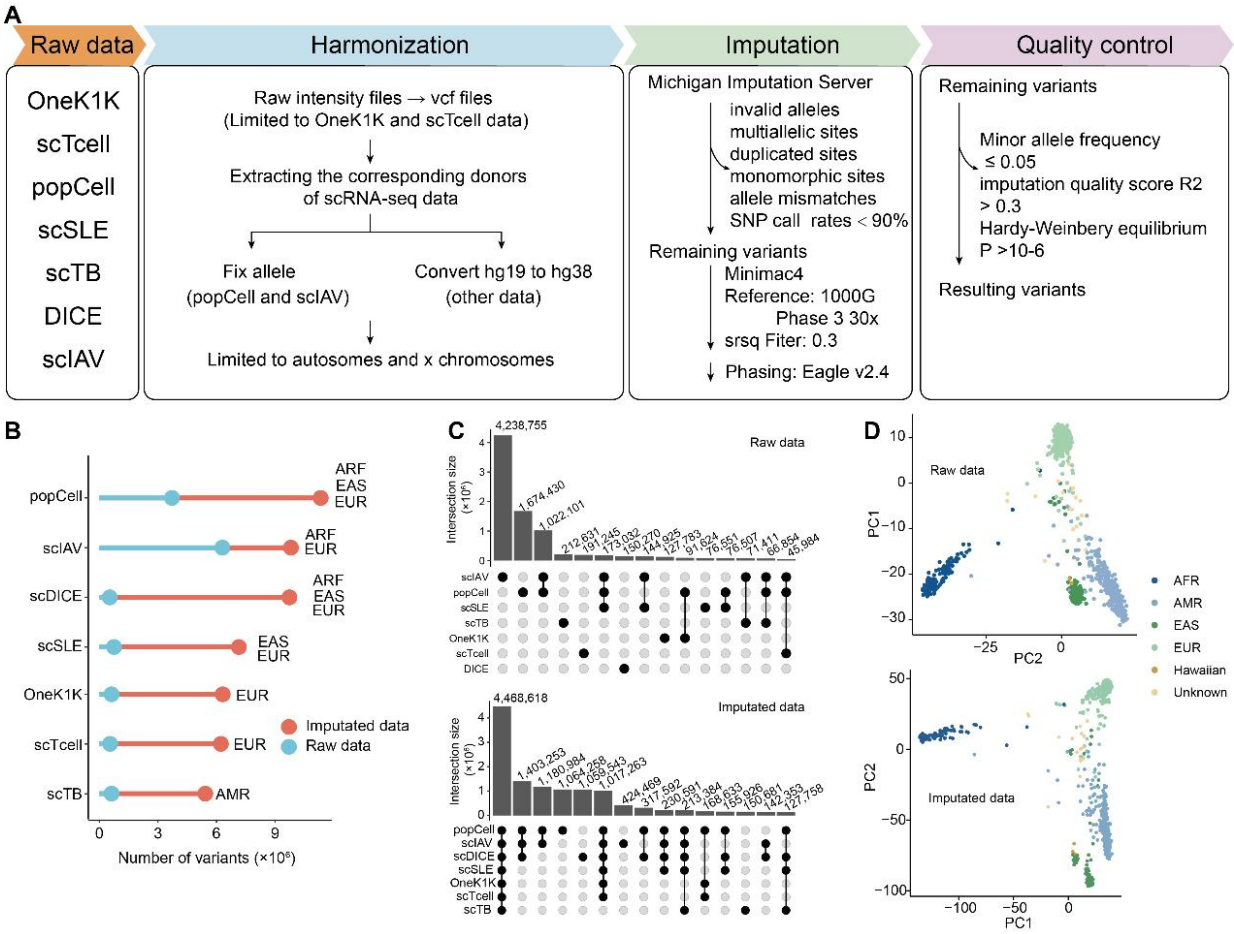

Fig. S4. Genotype imputation and quality control. (**A**) Genotyping data processing and quality control. (**B**) The number of variants with or without imputation based on the 1000 Genomes reference. (**C**) Comparison of the overlapping variants among different cohorts before (top) and after imputation (bottom). The raw data were collected from published datasets. (**D**) Principal component analysis of raw and imputed genotype data. EAS, East Asian. AFR, African. AMR, Ad Mixed American. EUR, European.

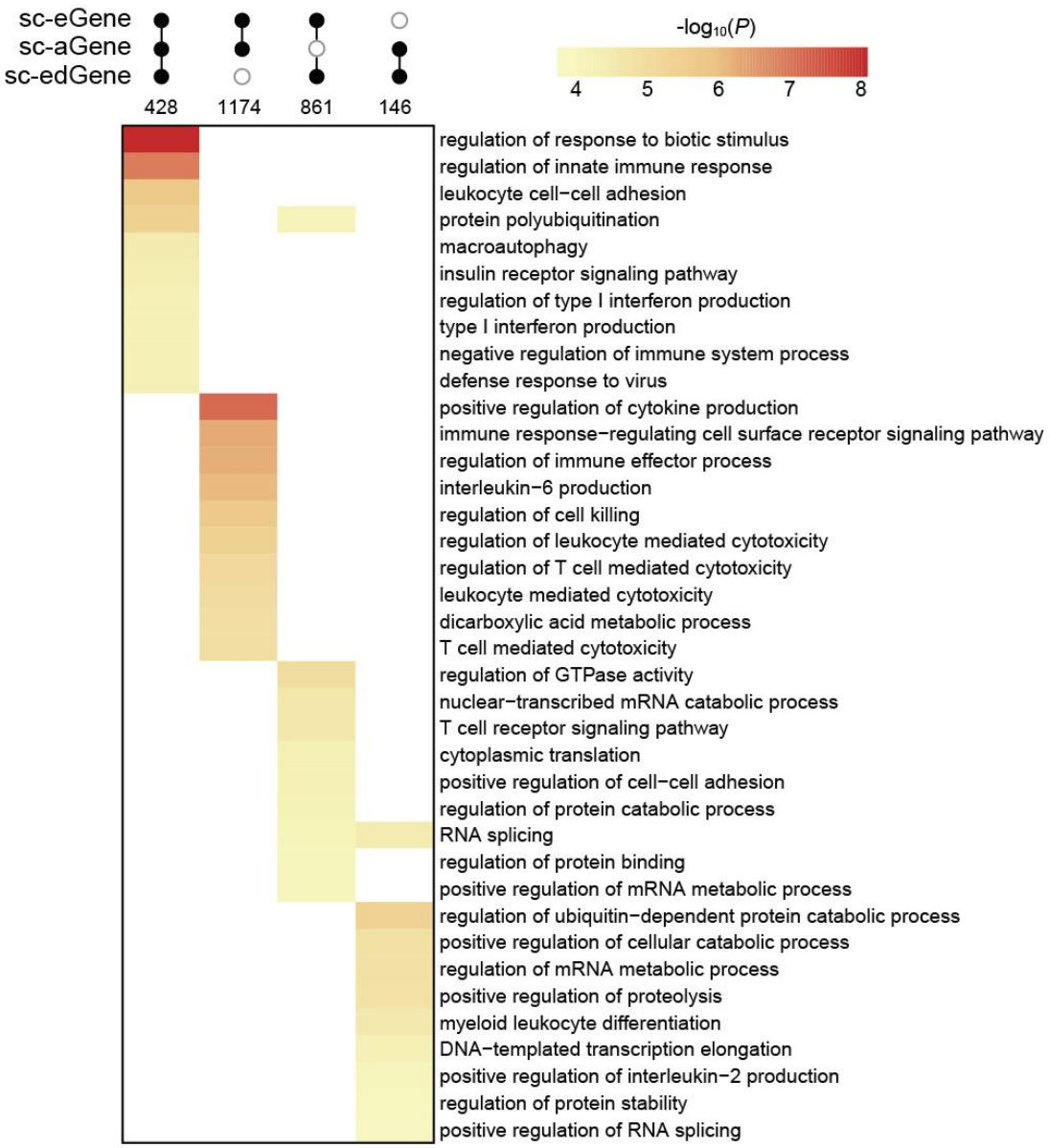

Fig. S5. GO enrichment analysis of overlapped sc-xGenes. The color represents the log-transformed enrichment *P*-value of each GO term among four types of overlapped sc-xGenes.

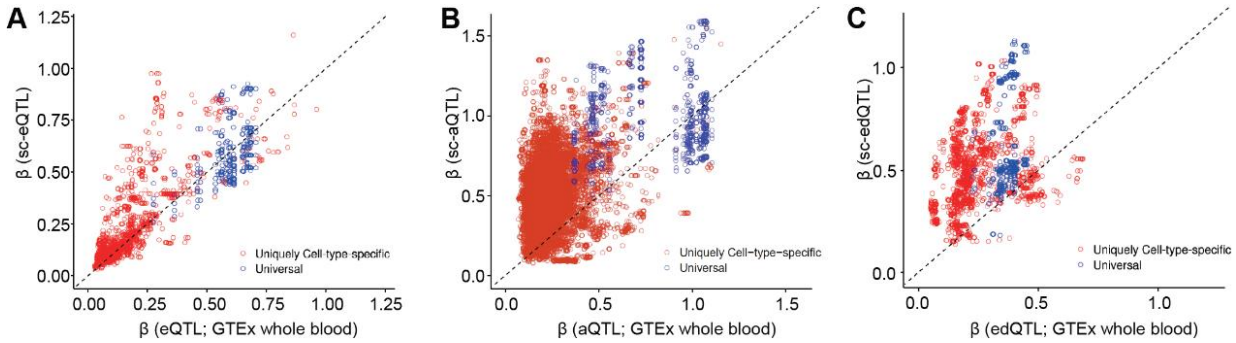

Fig. S6. Replication of sc-xQTL and bulk xQTL. (**A**) Correlation of shared eQTLs between sc-eQTL and GTEx whole blood eQTL. (**B**) Correlation of shared aQTL between sc-aQTL and GTEx whole blood aQTL. (**C**) Correlation of shared edQTL between sc-edQTL and GTEx whole blood edQTL. xQTLs shared across > 5 cell types (blue) and cell type-specific QTLs (red) are highlighted.

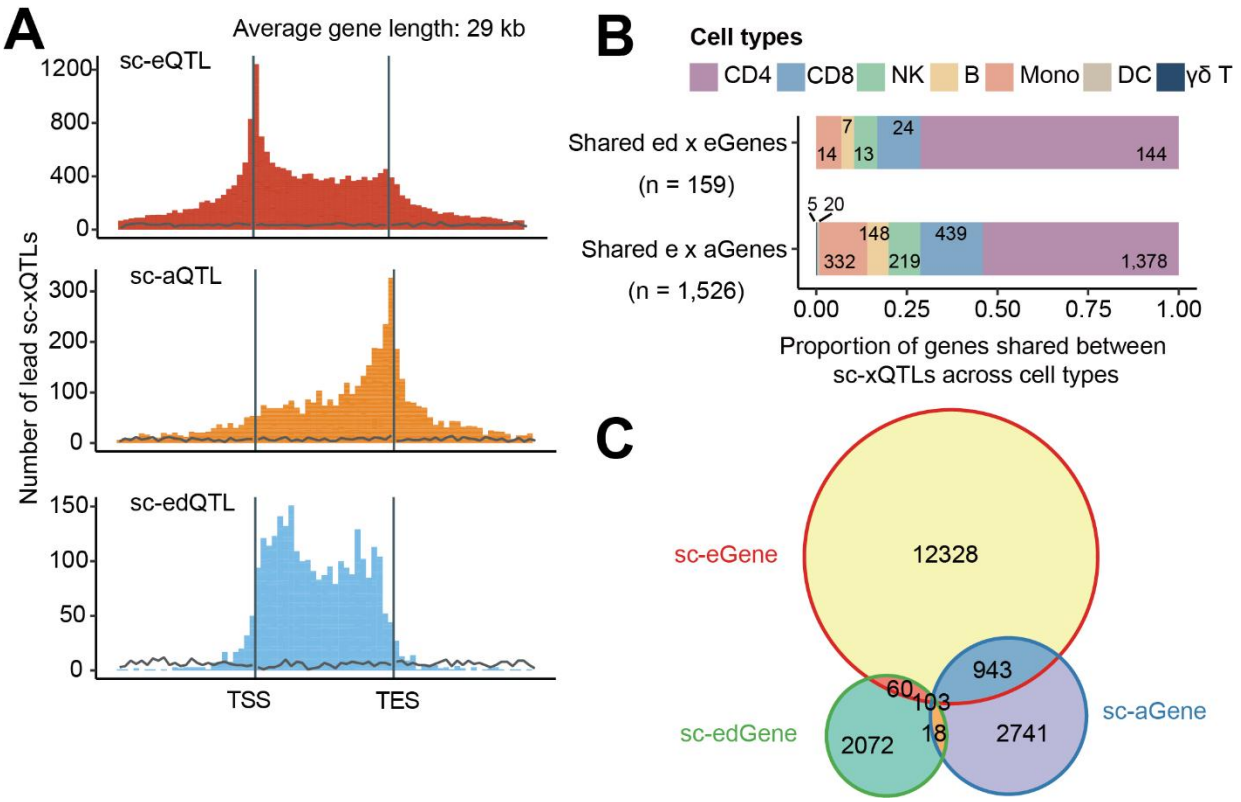

Fig. S7. Distribution and overlapping of different sc-xQTLs. (**A**) Distribution of different sc-xQTLs in gene regions. For each sc-xQTL, the distance between lead SNP and transcription start site (TSS) and transcription end site (TES) was calculated. (**B**) Number of shared molecular phenotypes with overlapped genetic variants. (**C**) Overlap analysis among eGenes, aGenes, and edGenes based on multi-layer sc-xQTL colocalization result.

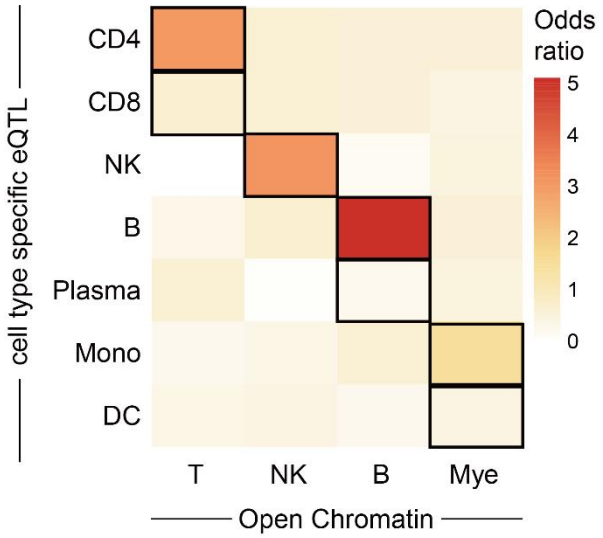

Fig. S8. Enrichment of cell type-specific sc-eQTLs in cell type hyper-accessible peaks. Color scale reflects odds ratio (OR) values from 0 (no enrichment, pale yellow) to 5 (strongest enrichment, dark red), calculated using a 2×2 contingency table including the number of cell type-specific eQTL, number of hyper-accessible peaks (ATAC-seq peaks, FDR < 0.05), and number of overlaps.

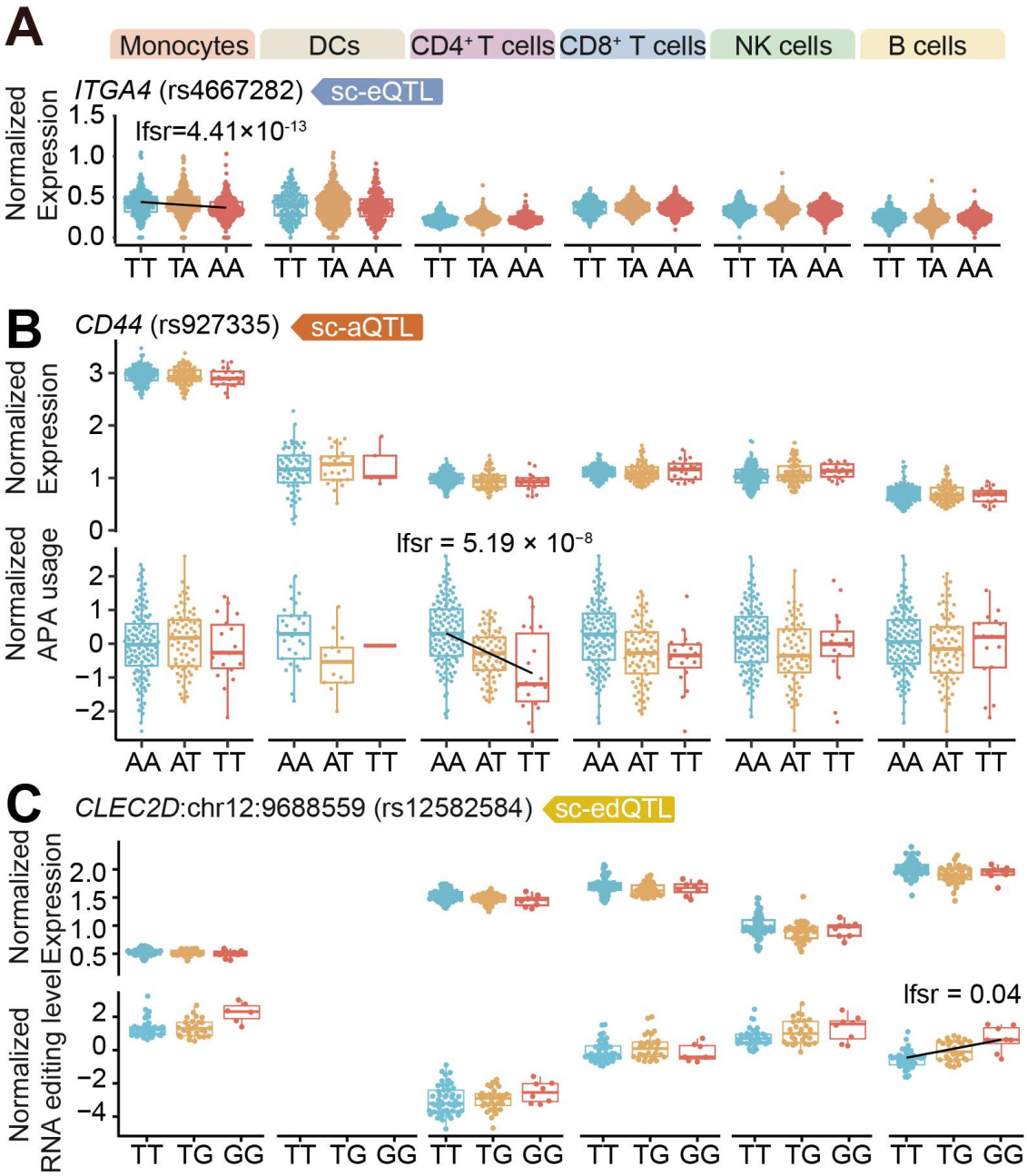

Fig. S9. Allelic effect plots of multiple molecular phenotypic variation. (**A**) Allelic effect plots for *ITGA4* showing cell type-specific expression patterns. The Y-axis shows normalized expression values for each genotype. (**B**) Allelic effect plots comparing expression and APA usage patterns of *CD44* between CD4⁺ and CD8⁺ T cells. (**C**) Allelic effect plots for *CLEC2D* editing site (chr12:9688559) showing expression patterns and editing levels across cell types.

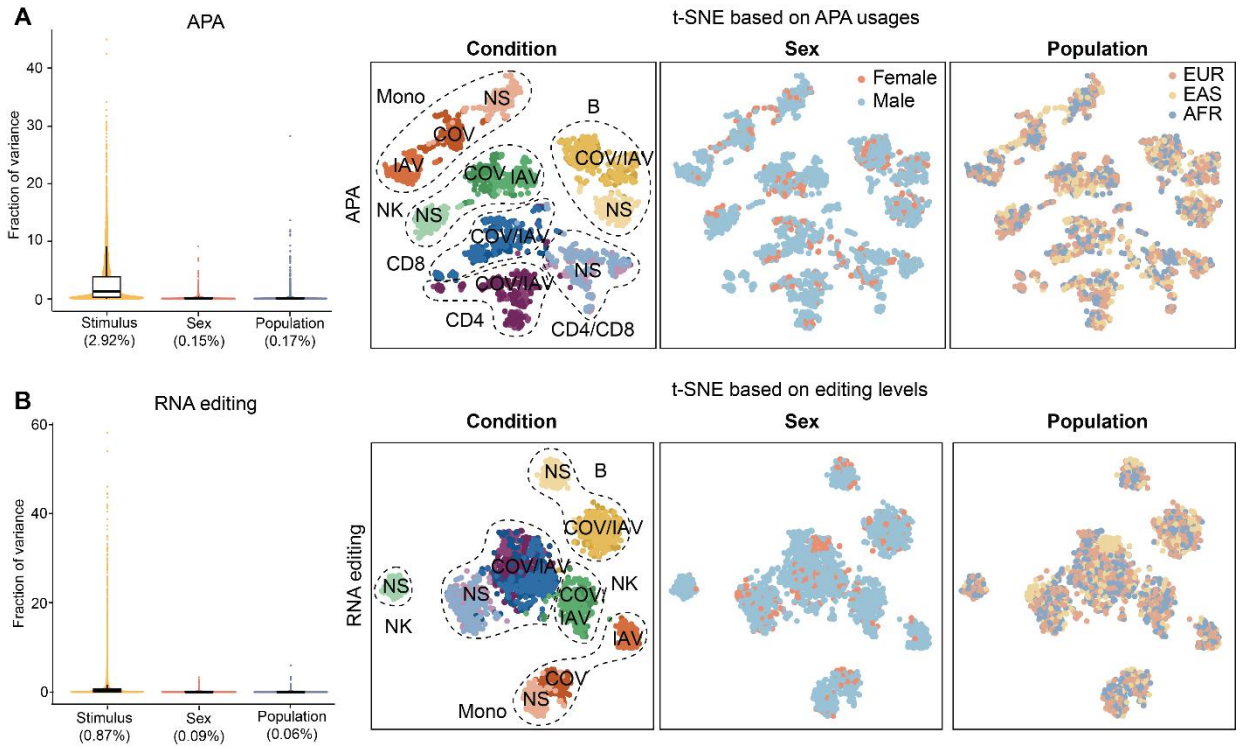

Fig. S10. Molecular phenotypic variation analysis using different contexts. (**A**) Quantification and t-SNE visualization of popCell data based on APA usages. The left panel shows quantification of the drivers of variation, and the right panel shows the distribution of spots by stimulus, sex, and population. (**B**) Quantification and t-SNE visualization of popCell data based on RNA editing levels. The left panel shows quantification of the drivers of variation, and the right panel shows the distribution of spots by stimulus, sex, and population.

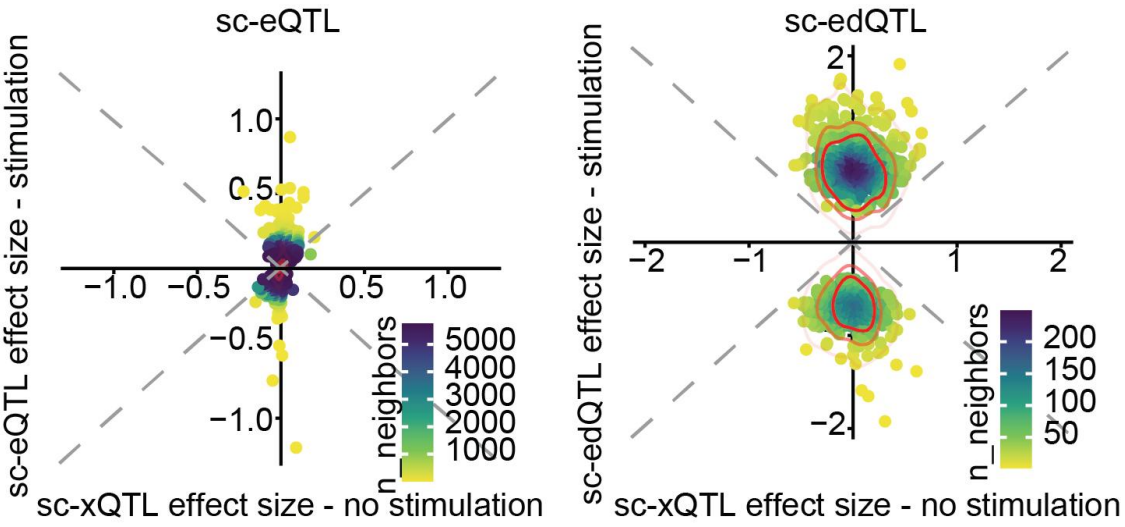

Fig. S11. Effect sizes of stimulus-dependent sc-eQTLs (left) and sc-edQTLs (right) compared to baseline. Each dot represents a stimulus-dependent sc-xQTL with a different effect size between stimulated (*x*-axis) and non-stimulated conditions (*y*-axis).

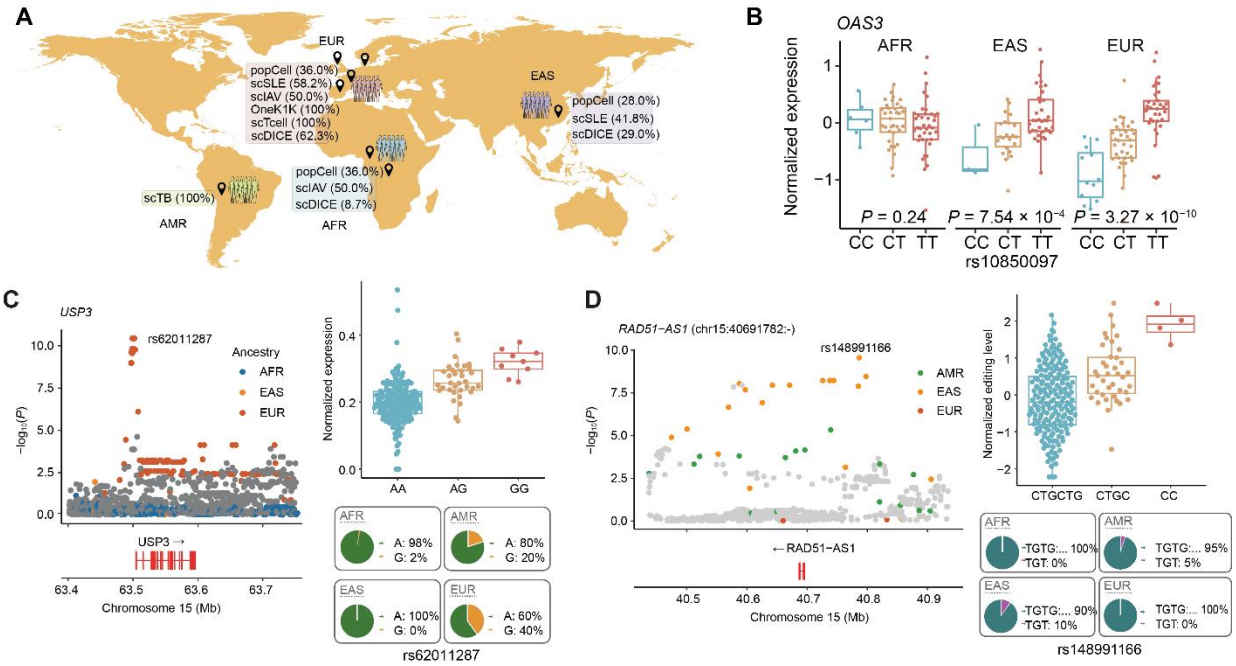

Fig. S12. Population-specific sc-xQTL profiles. (**A**) Geographical distribution of our collected samples. Numbers in parentheses indicate the percentage of that population for that source. (**B**) Population-biased genetic regulation of *OAS3* in AFR, EAS, and EUR populations. (**C**) Manhattan plot and boxplot showing the EUR-specific genetic regulation of *USP3* expression. (**D**) Manhattan plot and boxplot showing the EAS-specific genetic regulation of *RAD51-AS1* (chr15:40691782:-) RNA editing. The allele frequencies of rs62011287 and rs148991166 in four populations were extracted from the 1000 Genomes Project Phase 3.

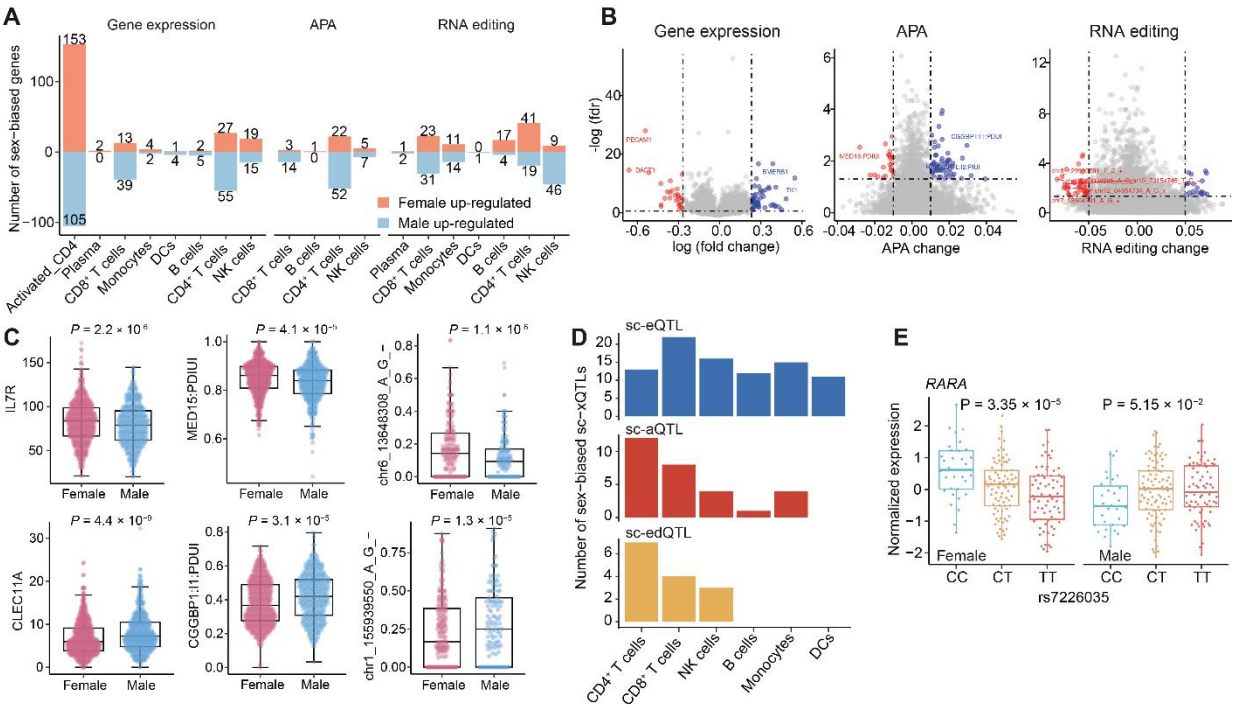

Fig. S13. Sex-biased sc-xQTL profiles. (**A**) Number of sex-biased expression, APA usages, and RNA editing levels in seven major cell types. (**B**) Volcano plots showing differential expression levels, APA usages, and RNA editing levels between different sexes. (**C**) Examples of sex-biased expression, APA usages, and RNA editing levels. The *P*-values were based on empirical Bayes statistics. (**D**) Bar plot showing the number of sex-biased sc-xQTLs identified in major immune cell populations. (**E**) Examples of sex-biased sc-xQTLs. The *P*-values were based on an interaction model using tensorQTL.

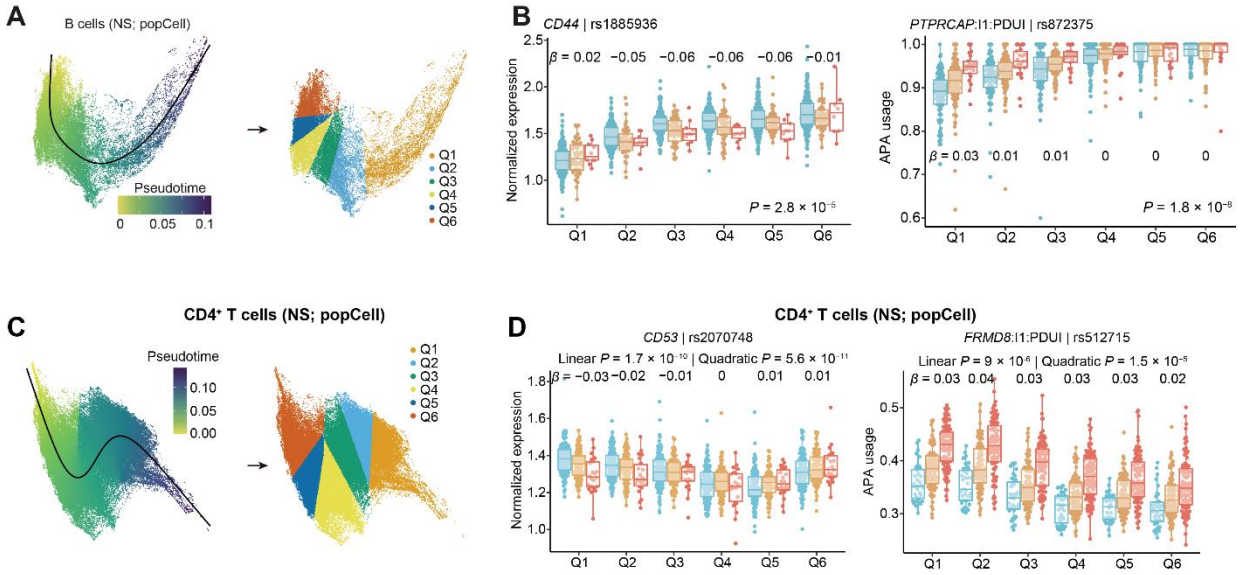

Fig. S14. Dynamic sc-xQTL profiles. (**A**) UMAP plot showing the pseudo-time projection of B cells. Left cells were colored by a pseudo-time curve with a color scale from 0 (the earliest pseudo-time) to 1 (the latest pseudo-time), and right cells were colored by six quantiles across the pseudo-time trajectory. (**B**) Examples of sc-xQTLs showing varying effect sizes (β values) across developmental quantiles. (**C**) UMAP plot showing the pseudotime projection of non-stimulated $CD4^+$ T cells from popCell data. (**D**) Examples of dynamic sc-xQTL across cell quantiles in $CD4^+$ T cells. The β values are shown for each quantile.

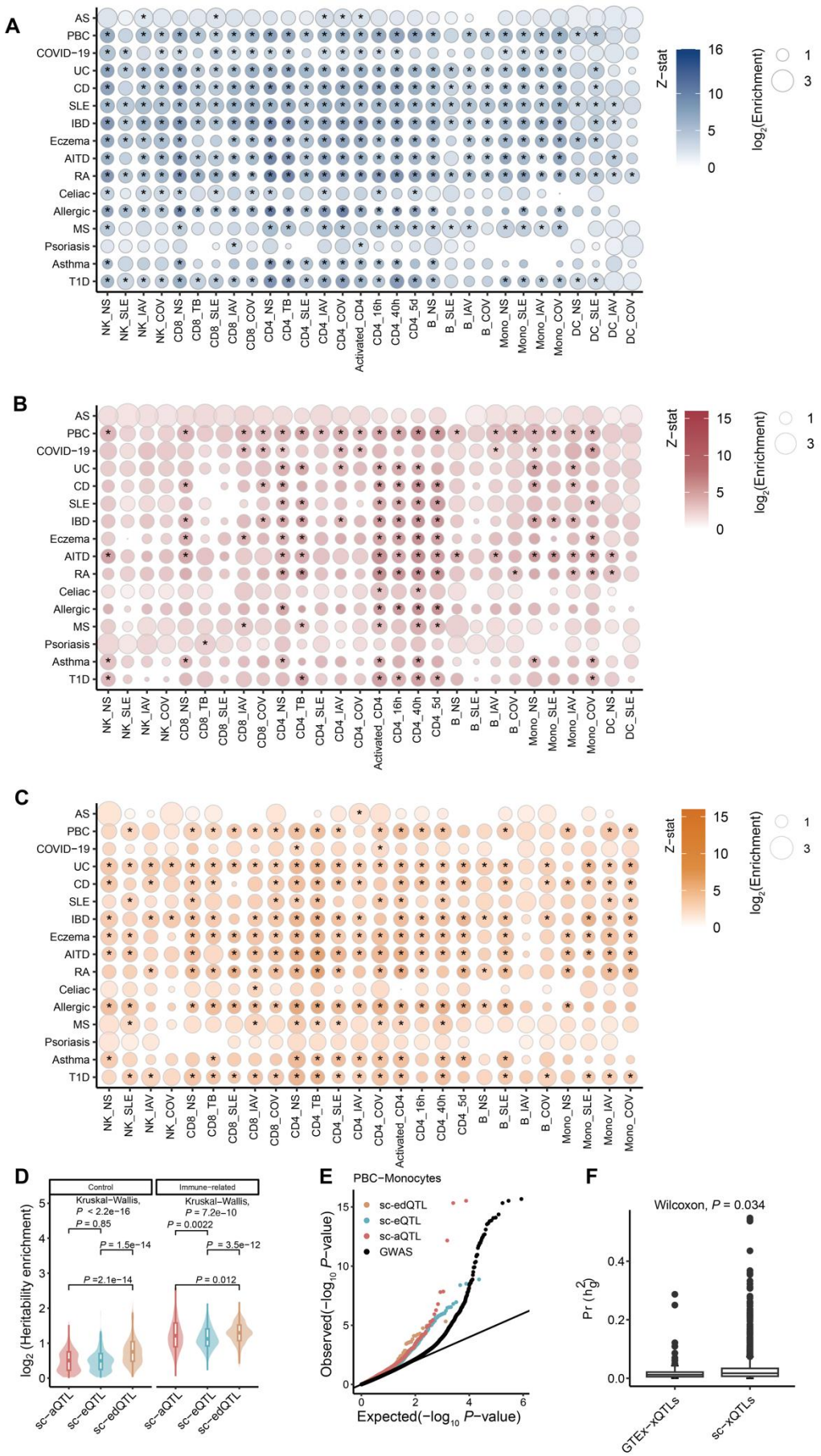
A
B
C
AS
PBC
COVID−19
UC
CD
SLE
IBD
Eczema
AITD
RA
Celiac
Allergic
MS
Psoriasis
Asthma
T1D
NK_NS
NK_SLE
NK_IAV
NK_COV
CD8_NS
CD8_TB
CD8_SLE
CD8_IAV
CD8_COV
CD4_NS
CD4_TB
CD4_SLE
CD4_IAV
CD4_COV
Activated_CD4
CD4_16h
CD4_40h
CD4_5d
B_NS
B_SLE
B_IAV
B_COV
Mono_NS
Mono_SLE
Mono_IAV
Mono_COV
DC_NS
DC_SLE
DC_IAV
DC_COV
Z−stat
log2(Enrichment)
D
Control
Immune−related
Kruskal−Wallis, P < 2.2e−16
Kruskal−Wallis, P = 7.2e−10
P = 0.85
P = 0.0022
P = 1.5e−14
P = 3.5e−12
P =2.1e−14
P = 0.012
log2 (Heritability enrichment)
sc−aQTL
sc−eQTL
sc−edQTL
E
PBC−Monocytes
sc-edQTL
sc-eQTL
sc-aQTL
GWAS
Observed(−log10 P−value)
Expected(−log10 P−value)
F
Wilcoxon, P = 0.034
Pr (h2g)
GTEx−xQTLs
sc−xQTLs

Fig. S15. Heritability enrichment for xQTL SNPs. Heritability enrichment of (**A**) sc-eQTL, (**B**) sc-aQTL, and (**C**) sc-edQTL variants at the cell-type level. Color indicates the z-statistic of the null hypothesis of no enrichment, while point size indicates log2 enrichment. Tests with FDR < 5% are indicated by "*". (**D**) Comparisons of the heritability enrichment between control and immune-related diseases, as well as across the sc-xQTLs categories. (**E**) Quantile-quantile (QQ) plot comparing PBC GWAS *P*-values in monocytes sc-xQTL against genome-wide SNPs. GWAS SNPs are binarized based on annotation to sc-eQTLs, sc-aQTLs, or sc-edQTLs at 5% FDR. (**F**) Pr (h_g^2) is a ratio of heritability attributable to the SNPs in query to the overall SNP-based heritability (h_{xQTL}^2/h_{SNP}^2). Pr (h_g^2) attributed by sc-xQTL and bulk xQTL were compared. Statistical significance between two groups was determined by an unpaired two-tailed Wilcoxon test, while differences among three groups were assessed using the Kruskal-Wallis test.

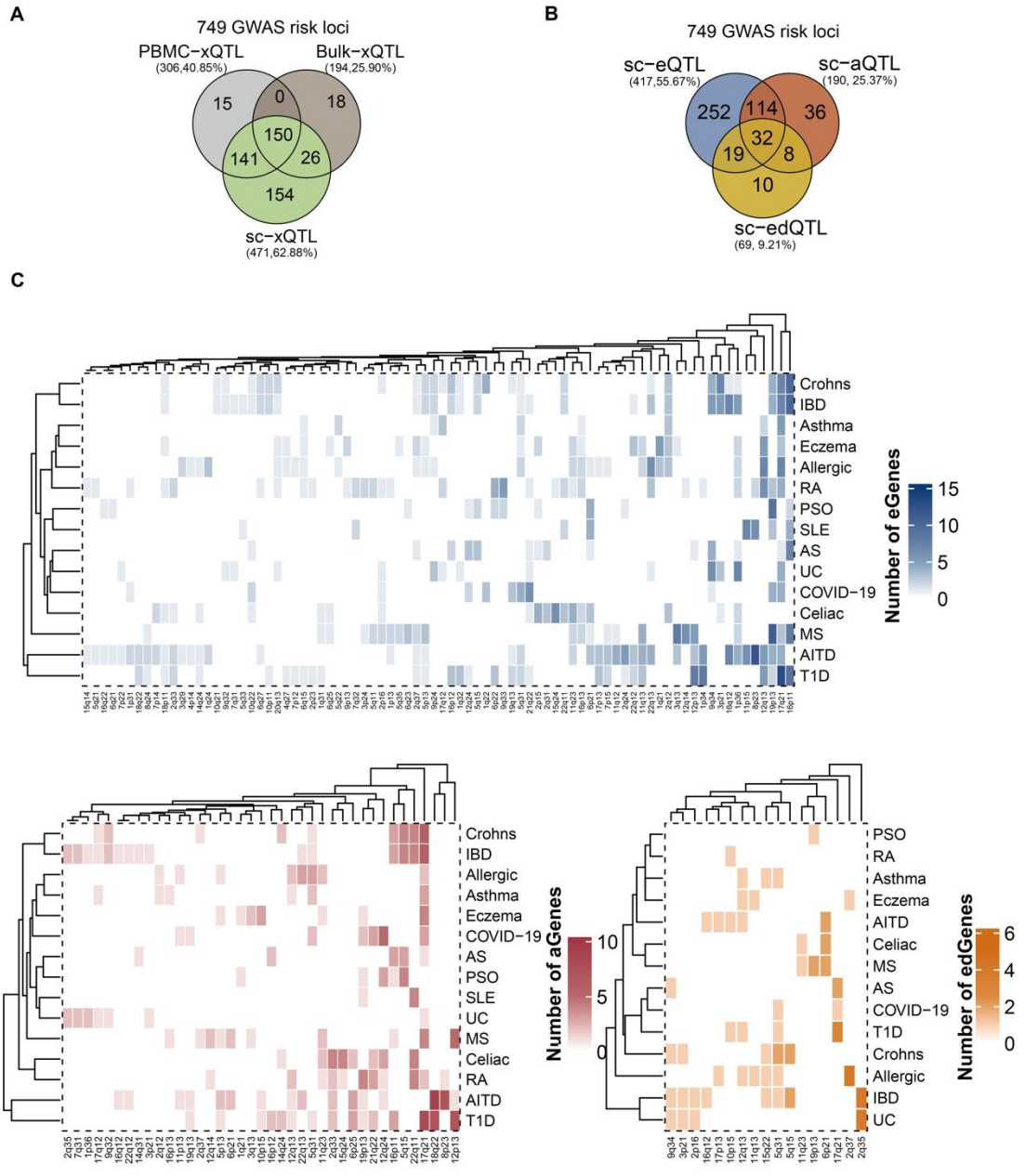

Fig. S16. The colocalization of sc-xQTLs and GWAS highlights distinct genes mediated by specific molecular mechanisms. (**A**) Venn diagrams showing overlap between colocalized loci identified in pseudo-bulk PBMC, bulk tissue (GTEx whole blood), and single-cell analyses. (**B**) Venn diagrams showing overlap between colocalized loci identified among three layers of sc-xQTLs. (**C**) The proportion of colocalized loci across different cell types for 16 immune related diseases. Shared colocalized genes across different immune-related diseases, mediated by molecular effects of gene expression (blue), APA (red) and RNA-editing (orange).

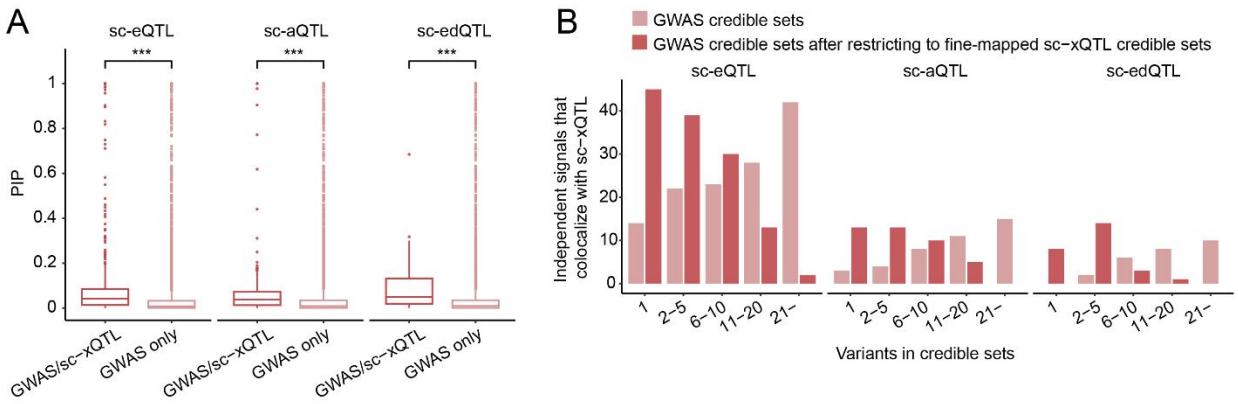

Fig. S17. Comparison of fine-mapping GWAS variants and sc-xQTL variants. (**A**) For all immune-associated loci that colocalize with an sc-xQTL, we examined all fine-mapped variants (99% credible sets in GWAS, GWAScred) and compared the distribution of causal posterior probabilities for variants that are also fine-mapped sc-xQTL variants (QTLcred) vs. those that were not fine-mapped sc-xQTL variants. Mann-Whitney *P*-values are provided. Boxplots show IQR without outliers, although *P*-values were calculated using all data points. *** *P*-value < 0.001. (**B**) Integration of immune-related GWAS credible set variants with credible sets from colocalizing sc-xQTLs increases fine mapping resolution.

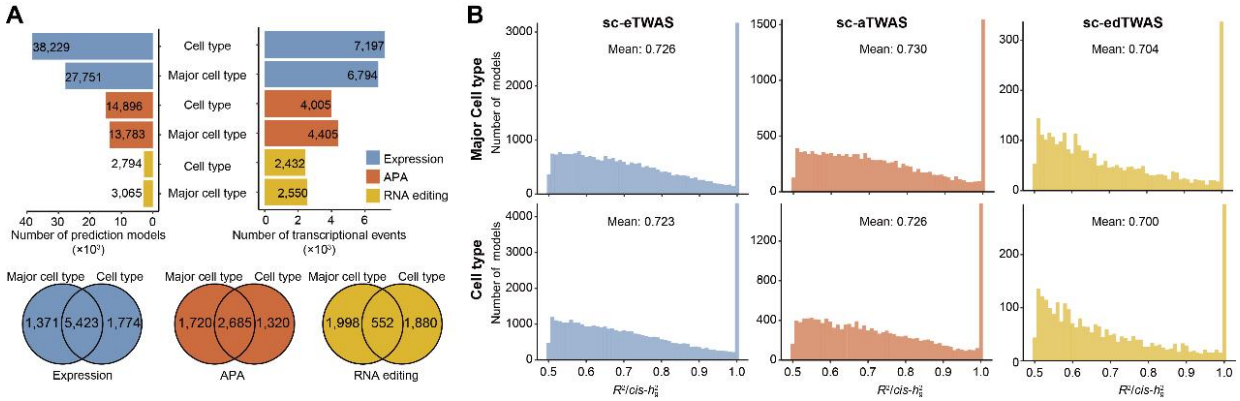

Fig. S18. TWAS model statistics. (**A**) The number of three transcriptional phenotypes in terms of models and events in all cell types. The bottom panel shows the intersection of major cell types and cell types. (**B**) Barplot showing the distribution of model prediction accuracy. We selected the models with a prediction accuracy greater than 0.5 for subsequent analysis.

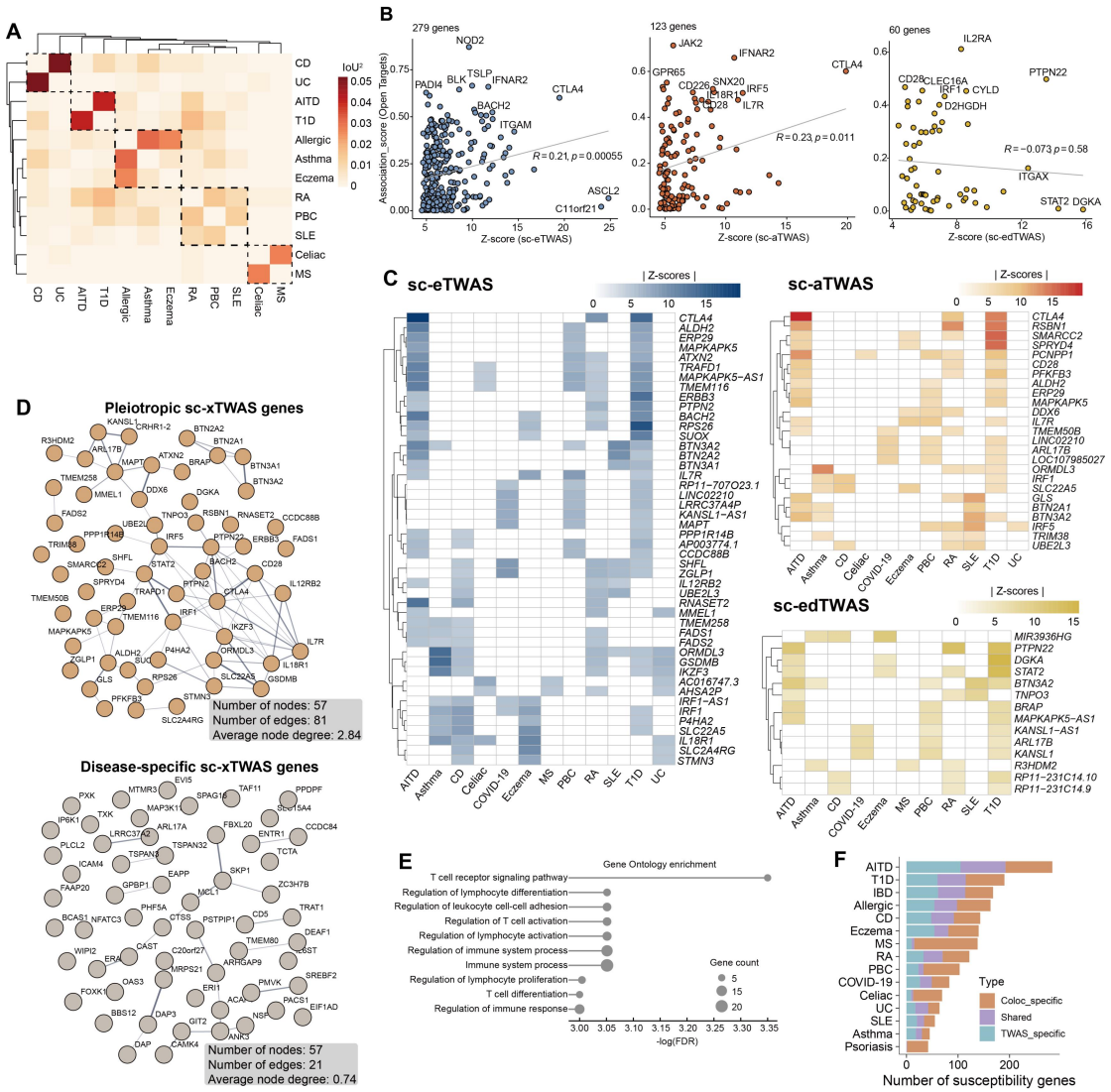

Fig. S19. Characteristics of susceptibility genes in scTWAS. (**A**) Heatmap showing the intersection over union of susceptibility genes among diseases. The high IoU represents a high sharing of susceptibility genes between them. (**B**) Comparison between the association score from the Open Target Platform and the TWAS z-score of reported disease genes. The top ten with higher distances were labeled. R: Pearson's correlation coefficient. *P*-value was based on Pearson's correlation test. Left, eTWAS, middle, aTWAS, and right, edTWAS. (**C**) Susceptibility genes with pleiotropic effects across three transcriptional phenotypes, colored by z-score. (**D**) PPI networks of pleiotropic genes (from panel C) and disease-specific genes without pleiotropy. (**E**) GO enrichment annotation of pleiotropic genes, with dot size indicating gene count. (**F**) The number of susceptibility genes supported by both colocalization and sc-xTWAS.

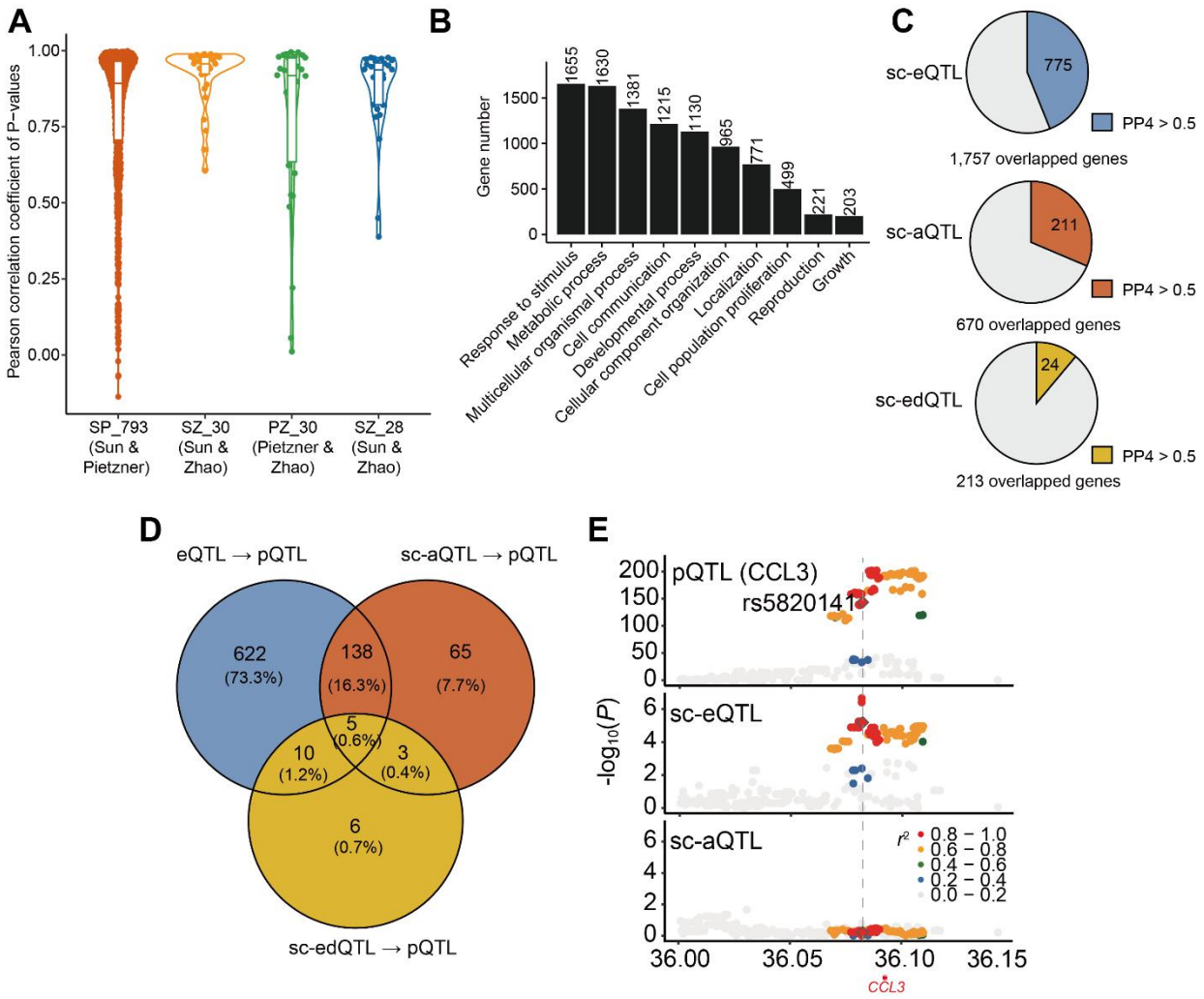

Fig. S20. Characteristics of the pQTL datasets. (**A**) Violin plot showing the Pearson correlation coefficient of -log10 (*P*-value) for shared variants between overlapping datasets. (**B**) GO enrichment of unique 2,653 genes. The color of each node indicates the gene number. (**C**) Pie plot showing the number of genes with colocalized signals between sc-xQTLs and pQTLs. (**D**) Venn diagram showing the number of overlapping genes with significant colocalization of sc-xQTL-pQTL (PP4 > 0.5). (**E**) Regional association plot showing colocalization between sc-eQTL and pQTL for *CCL3*.

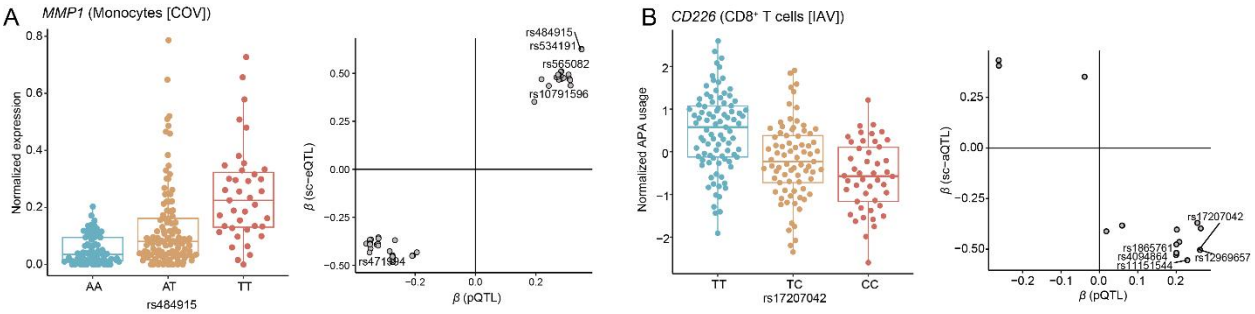

Fig. S21. Directional effects of sc-xQTL on pQTL. (**A**) Boxplot and scatter plot showing the positive correlation between *MMP1* gene expression and protein abundance associated with rs484915. Left, the T-allele of rs484915 is associated with increasing *MMP1* gene expression. Right, expression and protein abundance regulated by significant variants show a positive correlation. The top five variants with large distances are labeled. (**B**) Boxplot and scatter plot showing the negative correlation between APA usages of *CD226* and protein abundance associated with rs17207042. Left, the C-allele of rs17207042 is associated with *CD226* APA shortening. Right, APA shortening and protein abundance regulated by significant variants show a negative correlation. The top five variants with large distances are labeled.

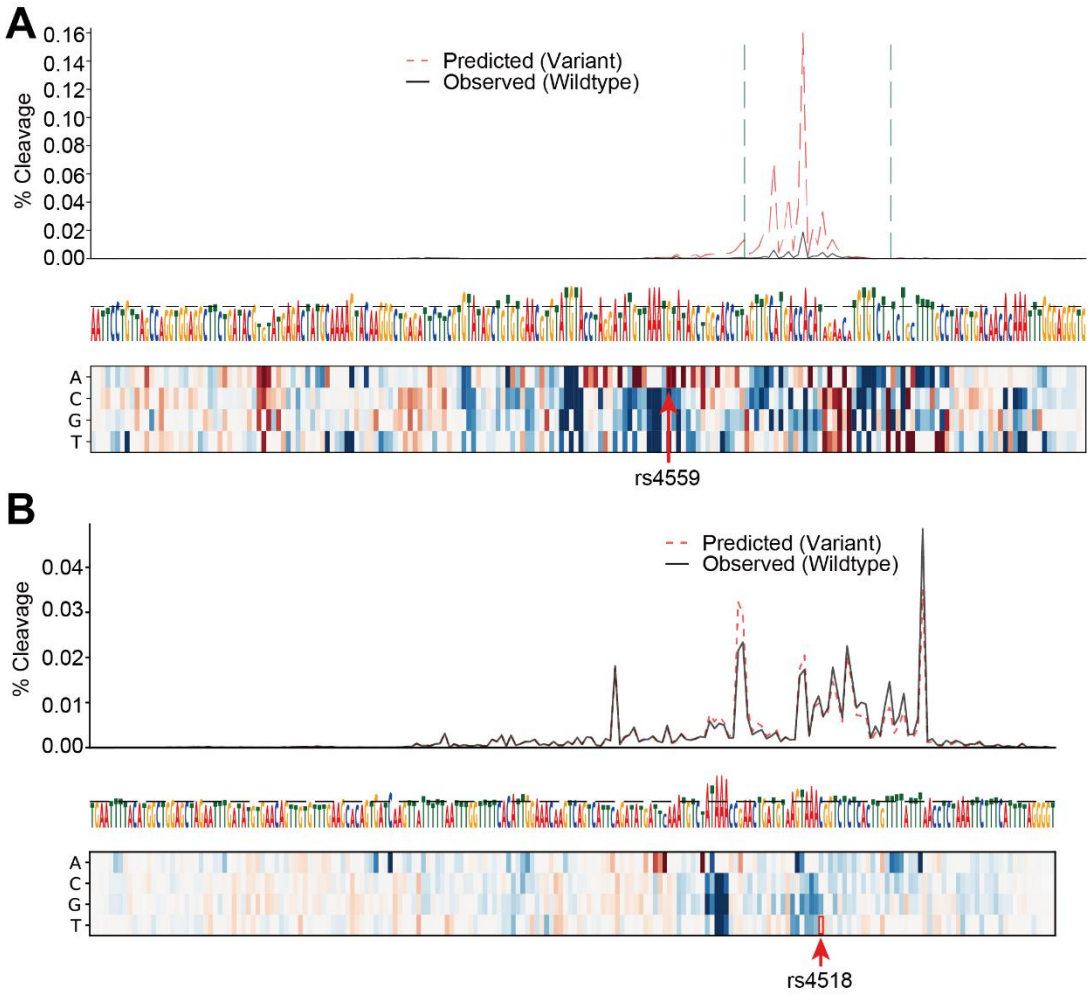

Fig. S22. Predicted allelic effects on poly(A) site selection of *STAT6* and *DLD*. (**A**) Predicted effect of causal variant rs4559 on PAS score based on APARENT. The G to A variant affects AATGTA to AATATA and increases PAS binding. The black line represents the predicted PAS score under the G allele, while the red line represents the PAS score under the A allele. (**B**) Predicted effect of causal variant rs4518 on PAS score based on APARENT. The black line represents the predicted PAS score under the C allele, while the red line represents the PAS score under the T allele.

Table S1. (separate file)
Dynamic changes of APA and RNA editing under different experimental conditions

Table S2. (separate file)
Cell type-specific sc-xQTLs.

Table S3. (separate file)
List of DNA oligos used in this study.

Table S4. (separate file)
List of GWAS summary statistics from literature.

Table S5. (separate file)
Colocalization results (PP4 > 0.75) between 16 immune-related diseases and sc-xQTLs

Table S6. (separate file)
Overlap between GWAS and context-dependent sc-xQTLs

Table S7. (separate file)
sc-xTWAS results.

Table S8. (separate file)
Colocalization results between pQTLs and sc-xQTLs.