nature genetics

Article

https://doi.org/10.1038/s41588-024-02019-8

Single-cell RNA sequencing of peripheral blood links cell-type-specific regulation of splicing to autoimmune and inflammatory diseases

Received: 7 December 2023

Accepted: 30 October 2024

A list of authors and their affiliations appears at the end of the paper

Alternative splicing contributes to complex traits, but whether this differs in trait-relevant cell types across diverse genetic ancestries is unclear. Here we describe cell-type-specific, sex-biased and ancestry-biased alternative splicing in ~1 M peripheral blood mononuclear cells from 474 healthy donors from the Asian Immune Diversity Atlas. We identify widespread sex-biased and ancestry-biased differential splicing, most of which is cell-type-specific. We identify 11,577 independent *cis*-splicing quantitative trait loci (sQTLs), 607 *trans*-sGenes and 107 dynamic sQTLs. Colocalization between *cis*-eQTLs and *trans*-sQTLs revealed a cell-type-specific regulatory relationship between *HNRNPLL* and *PTPRC*. We observed an enrichment of *cis*-sQTL effects in autoimmune and inflammatory disease heritability. Specifically, we functionally validated an Asian-specific sQTL disrupting the 5′ splice site of *TCHP* exon 4 that putatively modulates the risk of Graves' disease in East Asian populations. Our work highlights the impact of ancestral diversity on splicing and provides a roadmap to dissect its role in complex diseases at single-cell resolution.

Genome-wide association studies (GWAS) have identified tens of thousands of loci associated with complex traits[1]. Most GWAS loci lie in noncoding genomic regions and their functional mechanisms are predominantly elusive. Genetic studies on molecular phenotypes such as Genotype-Tissue Expression (GTEx) project[2] and the eQTLGen consortium[3] have identified expression quantitative trait loci (eQTLs) for nearly every gene, yet the proportion of GWAS signals attributable to eQTLs has been modest[4,5]. To identify the genetic effects underlying complex diseases, it is important to investigate disease-relevant cell types during relevant differentiation stages under pertinent perturbations[6]. Single-cell RNA sequencing (scRNA-seq) enables unbiased examination of cell types[7,8] and massively parallel perturbation measurements of transcriptional regulation[9].

Splicing QTLs (sQTLs) are a mediator of genetic effects on complex traits. Growing evidence shows that sQTLs and eQTLs exert largely orthogonal effects on the genetic risk of complex diseases[10,11]. Cell-type-specific sQTLs are critical for dissecting complex diseases, yet they are heavily underexplored compared to eQTLs. FACS has enabled cell-type-specific dissection of sQTLs[12–14], but they are biased toward known cell types defined by established surface markers. Genetics-coupled single-cell RNA sequencing (scRNA-seq) studies[7,8] enabled unbiased detection of cell types and cellularly resolved eQTLs but did not investigate alternative splicing (AS) due to 3′ bias of their scRNA-seq libraries[15,16]. Full-length single-cell technologies, on the other hand, currently suffer from insufficient throughput and elevated costs for population-scale profiling[17,18].

In this study, we leveraged 5′ library preparation (10x Genomics) and stochastic mRNA cleavage and recapping—an endogenous cellular phenomenon that creates multiple 5′ ends per isoform[19,20]—to increase the exon coverage of scRNA-seq. Using ~1 million single-cell transcriptomic profiles from 474 donors of Eastern, Southeastern and South Asian ancestries, we identified 48 sex-biased differential splicing

e-mail: boxiangliu@nus.edu.sg

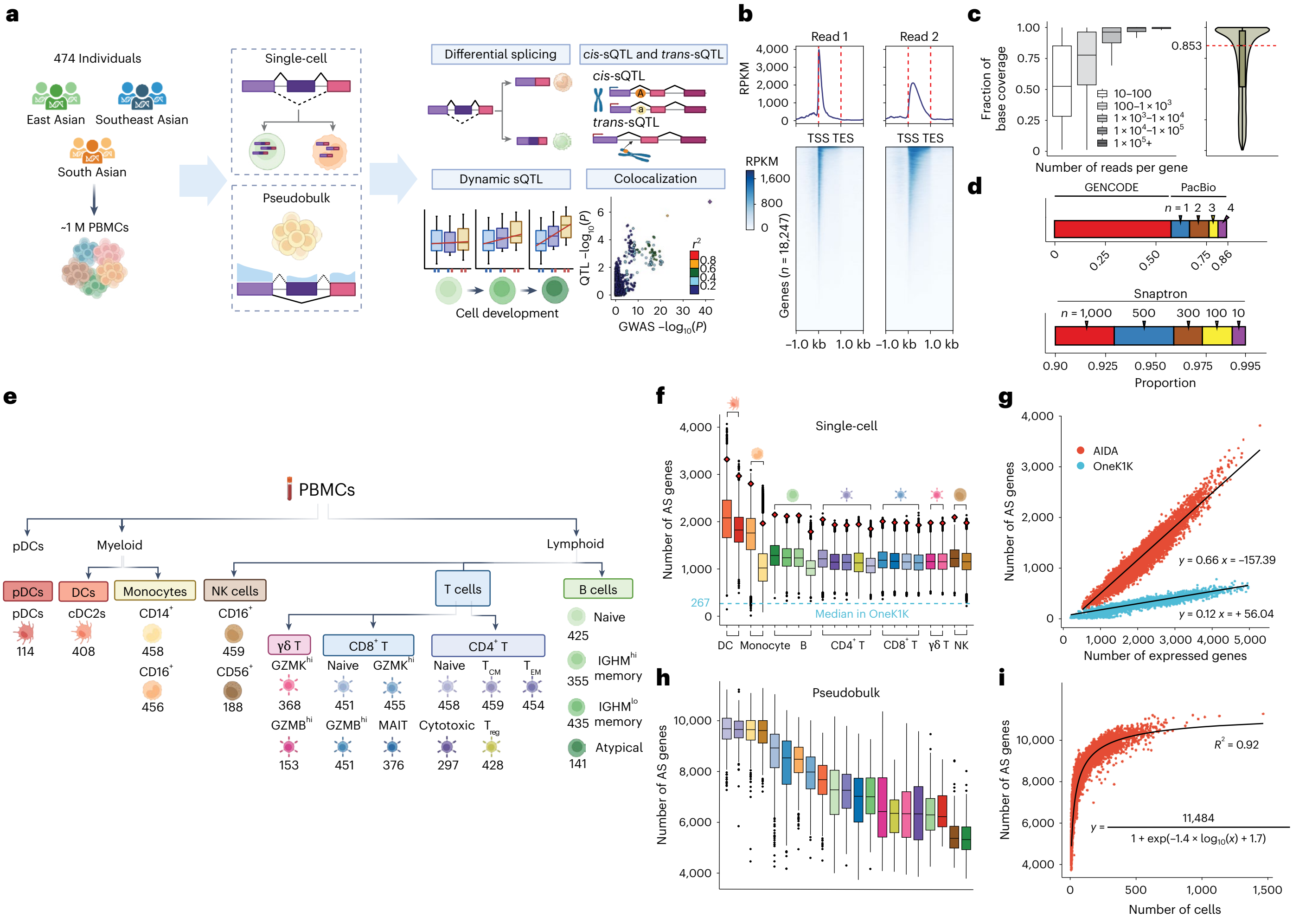


Fig. 1 | Population-scale 5′ scRNA-seq identified 21 cell types and thousands of alternatively spliced genes per cell type. a, The AIDA cohort and study design. **b**, Profile plot and heatmap showing that read 1 of 5′ scRNA-seq was biased toward the transcription start site and read 2 was spread more evenly across the gene body. **c**, The base coverage rate per gene increased with the read count (fraction of base coverage = covered bases/all bases). Left, box plot showing the fraction of base coverage across different read count bins (n = 4,034, 4,114, 4,803, 4,883 and 491, from left to right). Outliers are not shown. Right, box plot showing that a median of 85.3% of exonic bases (red line) are covered across all expressed genes. **d**, Replication of LeafCutter intron discoveries in GENCODE, PacBio MAS-seq and Snaptron. Top, 59.3% of LeafCutter discoveries were annotated in GENCODE and 85.9% replicated in PacBio long-read sequencing from four individuals. Bottom, close to 93% of detected splice junctions appeared in more than 1,000 samples, 98.8% in more than 100 and 99.5% in more than ten. **e**, We examined 21 distinct PBMC subtypes with sufficient cell counts. Cell types are colored according to their hematopoietic lineage. The numbers below the cell type labels indicate the sample size for differential splicing analysis and sQTL calling. **f**, Number of alternatively spliced genes detected per cell across 21 cell types at the single-cell level (see Supplementary Table 1 for the number of cells used (n)). The red diamonds indicate the average number of detected genes (NODGs) per cell. The dashed blue line indicates the number of AS genes detected using the OneK1K dataset. **g**, NODGs positively correlated with the number of AS genes. Linear regression lines (black) are shown for AIDA and OneK1K, respectively. **h**, Number of detected AS genes per pseudobulk cell type (see Supplementary Table 1 for the number of cells used (n)). **i**, Number of detected AS genes scaled with the number of cells in a pseudobulk, plateauing at ~11,500 genes. A sigmoid curve was fitted to the data and plotted. cDC, conventional dendritic cell; GZMB, granzyme B; GZMK, granzyme K; IGHM, immunoglobulin heavy constant Mu; pDC, plasmacytoid dendritic cell; RPKM, reads per kilobase of transcript per million mapped reads; TES, transcription end site; TSS, transcription start site.

events (DSEs) across 32 genes. Specifically, sex-biased splicing of *FLNA* was putatively driven by female-biased expression of the isoform ENST00000498491. We also identified 1,031 ancestry-biased DSEs affecting a total of 509 genes. In particular, ancestry-biased *SPSB2* splicing was probably driven by cross-population allele frequency differences in rs11064437 that disrupted the 3′ splice site of *SPSB2* exon 2. We identified 10,874 and 703 *cis*-sQTLs for protein-coding genes and long noncoding RNAs, respectively, many of which were cell-type-specific or sex-biased. We also identified 865 dynamic intron use events and 107 dynamic sQTLs along B cell differentiation. Our analysis revealed 607 *trans*-sQTL genes (sGenes) and cell-type-specific genetic coregulation between a *trans*-sQTL for *PTPRC*, a protein tyrosine kinase critical for T cell development, and a *cis*-eQTL for *HNRNPLL*, a master regulator of T cell activation-induced AS. Finally, we observed strong enrichment of *cis*-sQTL effects in autoimmune and inflammatory disease heritability. Colocalization analysis identified 563 putative effector genes. In particular, we functionally validated an Asian-specific sQTL that putatively modulated Graves' disease (GD) risk by disrupting the 5′ splice site of *TCHP* exon 4 in East Asian populations.

Results

Quantification of AS across cell types

The Asian Immune Diversity Atlas (AIDA) Data Freeze v.1 contained 474 Eastern, Southeastern and South Asian donors from Japan, Korea and Singapore (Fig. 1a, Extended Data Fig. 1a and Supplementary Note). Each donor was sampled to an average of 1,959 single cells (Extended Data Fig. 1b). Each cell was sampled to an average of 53,846 reads, of which 25,228 (46.9%) overlapped at least one splice junction

(Extended Data Fig. 1d,e). Although read 1 was biased toward the 5′ end, read 2 was spread more evenly because of normally distributed insert lengths[21] (Fig. 1b and Supplementary Fig. 1a). Furthermore, incomplete reverse transcription, along with stochastic mRNA cleavage and recapping—or 'exon painting'—created multiple 5′ ends per isoform[19,20] (Supplementary Fig. 1b) to extend coverage throughout the gene body. The covered bases for each gene increased with its read count, with a median of 85.3% (interquartile range (IQR) = 51.8–97.3%; Fig. 1c and Extended Data Fig. 1f).

We classified scRNA-seq data into 34 cell types based on their gene expression profiles[22] and examined AS in 21 peripheral blood mononuclear cell (PBMC) subtypes at two levels (Methods, Fig. 1e and Extended Data Fig. 1c). At the single-cell level, SpliZ[23] detected a median of 1,146 AS genes per cell (range of medians = 1,013–2,081; Fig. 1f and Supplementary Table 1). Our estimate was 4.3-fold of OneK1K[7] using identical pipelines (1,146 versus 267). The number of AS and expressed genes per cell were highly correlated (Pearson's $r = 0.95$). An average of 66% (ordinary least squares (OLS) 95% confidence interval (CI) = 65.8–66.3%) of expressed genes had detectable AS events in AIDA, compared to 12.1% in OneK1K (OLS 95% CI = 12.0–12.2%; Fig. 1g), suggesting that the difference was not due to sequencing depth but rather to exon painting and other factors. At the pseudobulk level, LeafCutter[24] detected a median of 7,721 AS genes per pseudobulk (range of medians = 5,341–9,683; Fig. 1h and Supplementary Table 1). AS gene counts scaled sigmoidally with cell counts per pseudobulk, saturating at ~11,500 genes (coefficient of determination $R^2 = 0.92$; Fig. 1i).

As a quality control (QC), we compared pseudobulk introns to GENCODE (v.32): 59.3% were canonical introns, while 38.9% and 1.8% contained one and two new junctions, respectively, which is consistent with findings from the Transcriptomic Resource of Immune Cells using Long-read Sequencing[25]. New introns had a higher Gini index, indicating cell-type-specific expression, than canonical introns ($P < 2.2 \times 10^{-308}$, $t = -55.874$, d.f. = 40,909; Extended Data Fig. 2a). Furthermore, we sequenced complementary DNA (cDNA) from four donors with PacBio multiplexed arrays sequencing (MAS-seq); 85.9% of short-read junctions were replicated in long-read data, while 13.9% and 0.17% contained one and two new junctions, respectively (Fig. 1d and Extended Data Fig. 2b). Lastly, compared with Snaptron[26], ~93%, 98.8% and 99.5% of pseudobulk junctions appeared in more than 1,000, more than 100 and more than ten Snaptron samples, respectively (Fig. 1d and Extended Data Fig. 2c). For both canonical and new junctions defined by GENCODE, 5′ splice sites were highly enriched for the canonical motif 5′-AG|GU; 3′ splice sites were highly enriched for AG|G-3′ (Extended Data Fig. 2d).

Cell-type-dependent and context-dependent AS

AS is ubiquitous across immune cells[27], so we asked whether AS-based clustering recapitulated our current understanding of the hematopoietic lineage. Hierarchical clustering using pseudobulk quantifications separated myeloid from lymphoid cells, which further clustered into B, cytotoxic T and natural killer (NK), and noncytotoxic T cells (Fig. 2a and Extended Data Fig. 3a). To verify the robustness, we repeated hierarchical clustering using single-cell quantifications and observed congruent clustering patterns (Fig. 2a).

We asked whether well-known isoforms were captured using 10x 5′ scRNA-seq assays. *PTPRC* encodes the CD45 transmembrane protein tyrosine phosphatase, which is critical in T cell differentiation[28]. The mRNA encoding the CD45RA$^+$ isoform, which included one or more of exons 4, 5 and 6, was preferentially expressed in naive T cells to facilitate T cell activation. In contrast, the mRNA encoding the CD45RO isoform, which skipped exons 4–6, was preferentially expressed in activated and memory T cells[29]. We observed that use of junctions corresponding to the CD45RA$^+$ isoform was highest in naive T cells and decreased in activated and memory T cells, whereas the CD45RO isoform exhibited an opposite trend (Fig. 2b). Another example is *CD44*, which encodes a key surface glycoprotein during T cell activation and homing. The mRNA encoding the standard CD44 isoform (CD44s) was most abundant in naive T cells, whereas isoforms containing exons 2–10 were more likely to be expressed after activation[30]. We observed elevated use of junctions corresponding to the CD44s isoform in naive T cells compared to other T cell subtypes (Fig. 2c).

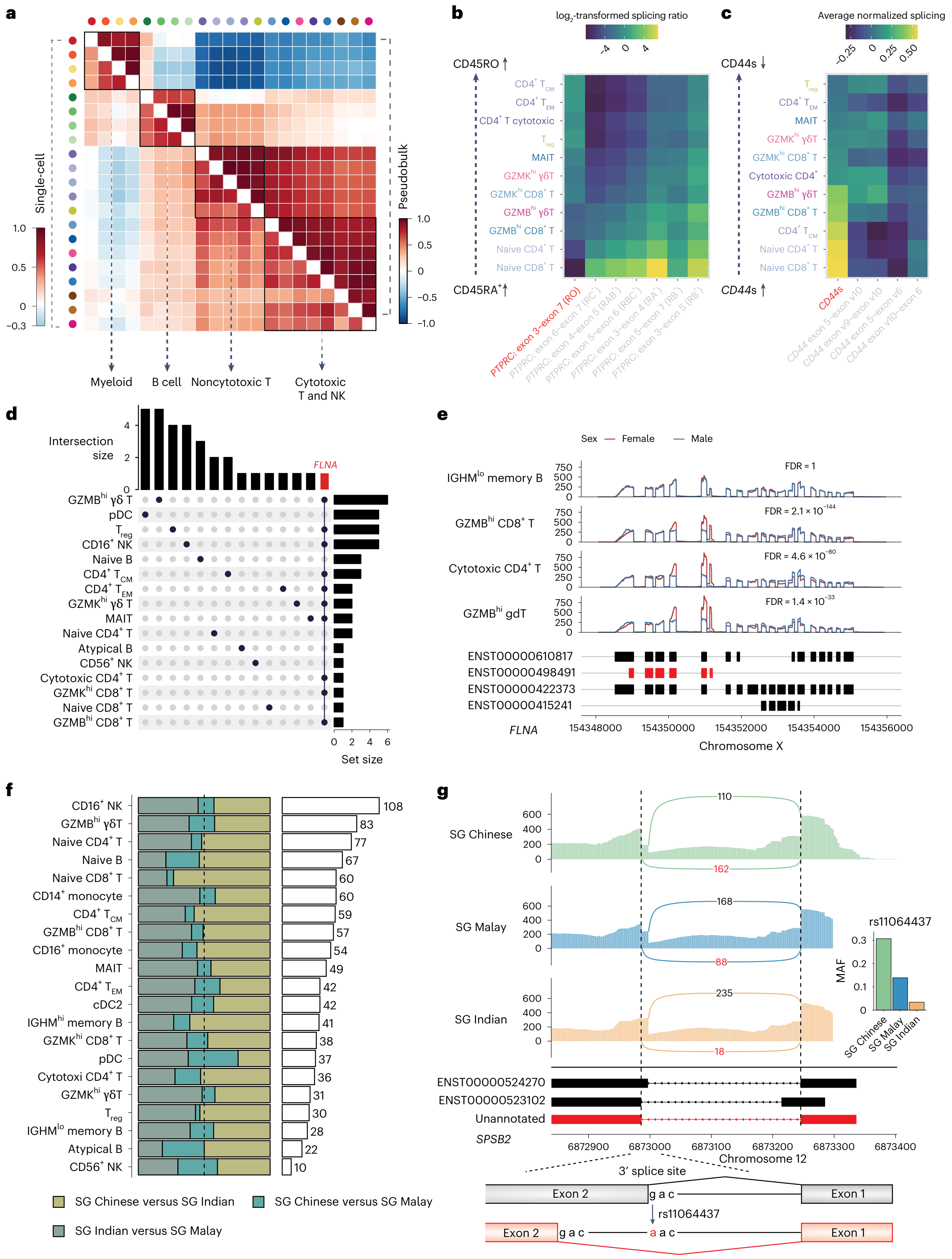
Sexual dimorphism influences a broad range of autoimmune and inflammatory diseases. Using LeafCutter, we identified 48 sex-biased DSEs in 32 genes across all cell types (false discovery rate (FDR) < 0.05; Fig. 2d and Supplementary Table 2). An example of shared sex-biased splicing involved filamin A (*FLNA*), an auto-antigen targeted by T and B cells in more than 50% of patients with rheumatoid arthritis (RA)[31]. Female donors had higher expression of a short isoform, ENST00000498491, consisting of exons 44–49 in mature T and NK cells (Fig. 2e and Extended Data Fig. 3b).

We next determined the influence of ancestry on AS and identified 1,031 DSEs affecting 509 genes (Fig. 2f and Supplementary Table 3). We observed the smallest difference between Singaporean Chinese and Singaporean Malay individuals, potentially because they are more genetically similar to each other than they are to Singaporean Indian individuals[32]. We identified an SNP, rs11064437, whose minor allele frequency (MAF) decreased from Eastern to Southeastern to South Asian individuals in AIDA (EAS = 30.7%; SEA = 13.9%; SAS = 3.4%; Fig. 2g) and in the 1000 Genomes Project (Extended Data Fig. 3c). Functionally, rs11064437 disrupted the 3′ splice site of *SPSB2* exon 2 (5′-AG|C-3′ to 5′-AA|C-3′), prompting the use of an alternative 3′ splice site 12-bp downstream. Notably, the EAS-biased alternative intron was missing from GENCODE, highlighting the lack of ancestral diversity in a widely used gene annotation database.

cis-sQTLs are cell-type-specific and context-biased

We followed an established pipeline[33,34] to identify *cis*-sQTL within a ± 1-Mb window across 19 cell types (Supplementary Table 1). We detected 10,874 and 703 sQTLs for protein-coding and long noncoding RNA genes, respectively (FDR < 0.05), representing 39.9% of all

Fig. 2 | Cell-type-dependent and context-dependent AS. a, Hierarchical clustering of single-cell and pseudobulk quantification of AS recapitulated well-known hematopoietic lineages. The heatmap shows the Spearman's rank correlation coefficient. Within the T and NK cluster, two subclusters demarcated cytotoxic and noncytotoxic cell types. The cytotoxic cellular cluster contained CD4$^+$ T cytotoxic, mucosal-associated invariant (MAIT), γδ T, NK and CD8$^+$ T (GZMK^{hi} and GZMB^{hi}) cells, whereas CD4$^+$ T cell (naive, T$_{CM}$ and T$_{EM}$), regulatory T (T$_{reg}$) cells and CD8$^+$ T naive cells fell within the noncytotoxic cluster. **b**,**c**, Alternative intron use of *PTPRC* and *CD44* reflected isoform-specific roles in T cell development. In **b**, the mRNA encoding the CD45RO isoform (red) was the lowest in naive T cells and was more abundant in activated and memory T cells. This trend was reversed for the mRNA encoding the CD45RA$^+$ isoforms. log-transformed splicing ratio = $\log_2$(CD45RX/CD45RO), where RX indicates any isoforms other than RO. For CD45RO, log-transformed splicing ratio = $\log_2$(CD45RO/ΣCD45RX). In **c**, the standard *CD44* (CD44s) isoform (red) was highest in naive T cells and was less abundant in activated and memory T cells. **d**, Discovery and sharing of sex-biased differentially spliced genes (DSGs) (FDR < 0.05). **e**, The sex-biased isoform expression of *FLNA* was cell-type-specific. The ENST00000498491 isoform (red boxes) exhibited strong female bias in T cells but not in B cells. **f**, Ancestry-biased DSGs discovered through pairwise comparisons across Eastern, Southeastern and South Asian individuals. Left, relative contributions of the three pairwise comparisons to the total number of DSGs in each cell type. Right, total number of DSGs across all cell types. **g**, Allele frequency difference in rs11064437 led to ancestry-biased isoform use of *SPSB2* in CD8$^+$ T GZMB^{hi}. rs11064437 disrupted the canonical splice site, thereby promoting use of the new splice site. Black, annotated canonical intron; red, new intron missing from GENCODE. Inset, MAF of rs11064437 decreased from Eastern to Southeastern to South Asian individuals.


a
Single-cell
Pseudobulk
1.0
0.5
0
−0.3
1.0
0.5
0
−0.5
−1.0
Myeloid
B cell
Noncytotoxic T
Cytotoxic T and NK
b
log2-transformed splicing ratio
−4 0 4
CD45RO ↑
CD45RA+↑
CD4+ TCM
CD4+ TEM
CD4+ T cytotoxic
Treg
MAIT
GZMKhi γδT
GZMKhi CD8+ T
GZMBhi γδT
GZMBhi CD8+ T
Naive CD4+ T
Naive CD8+ T
PTPRC: exon 3–exon 7 (RO)
PTPRC: exon 6–exon 7 (RC+)
PTPRC: exon 4–exon 5 (RAB+)
PTPRC: exon 5–exon 6 (RBC+)
PTPRC: exon 3–exon 4 (RA+)
PTPRC: exon 5–exon 7 (RB+)
PTPRC: exon 3–exon 5 (RB+)
c
Average normalized splicing
−0.25 0 0.25 0.50
CD44s ↓
CD44s ↑
Treg
CD4+ TEM
MAIT
GZMKhi γδT
GZMKhi CD8+ T
Cytotoxic CD4+
GZMBhi γδT
GZMBhi CD8+ T
CD4+ TCM
Naive CD4+ T
Naive CD8+ T
CD44s
CD44 exon 5–exon v10
CD44 exon v9–exon v10
CD44 exon 5–exon v6
CD44 exon v10–exon 6
d
Intersection size
4
2
0
FLNA
GZMBhi γδ T
pDC
Treg
CD16+ NK
Naive B
CD4+ TCM
CD4+ TEM
GZMKhi γδ T
MAIT
Naive CD4+ T
Atypical B
CD56+ NK
Cytotoxic CD4+ T
GZMKhi CD8+ T
Naive CD8+ T
GZMBhi CD8+ T
0 2 4 6
Set size
e
Sex — Female — Male
IGHMlo memory B
FDR = 1
GZMBhi CD8+ T
FDR = 2.1 × 10−144
Cytotoxic CD4+ T
FDR = 4.6 × 10−80
GZMBhi gdT
FDR = 1.4 × 10−33
750
500
250
0
ENST00000610817
ENST00000498491
ENST00000422373
ENST00000415241
FLNA
154348000
154350000
154352000
154354000
154356000
Chromosome X
f
CD16+ NK 108
GZMBhi γδT 83
Naive CD4+ T 77
Naive B 67
Naive CD8+ T 60
CD14+ monocyte 60
CD4+ TCM 59
GZMBhi CD8+ T 57
CD16+ monocyte 54
MAIT 49
CD4+ TEM 42
cDC2 42
IGHMhi memory B 41
GZMKhi CD8+ T 38
pDC 37
Cytotoxi CD4+ T 36
GZMKhi γδT 31
Treg 30
IGHMlo memory B 28
Atypical B 22
CD56+ NK 10
SG Chinese versus SG Indian
SG Chinese versus SG Malay
SG Indian versus SG Malay
g
SG Chinese
110
162
SG Malay
168
88
SG Indian
235
18
600
400
200
0
rs11064437
MAF
0.3
0.2
0.1
0
SG Chinese
SG Malay
SG Indian
ENST00000524270
ENST00000523102
Unannotated
SPSB2
6872900
6873000
6873100
6873200
6873300
6873400
Chromosome 12
3′ splice site
Exon 2
g a c
Exon 1
rs11064437
Exon 2
g a c
a a c
Exon 1

tested genes (Fig. 3a and Supplementary Table 4). Lead sQTL variants (sVariants) were enriched near splice sites, introns and splice regions (Fig. 3b,c). Sample size was positively correlated with sQTL discovery (Pearson's $r = 0.95$; Fig. 3d), but were inversely correlated with sQTL effect size (Pearson's $r = -0.95$; Extended Data Fig. 4a), indicating that large sample sizes were better-powered to detect small-effect variants[33]. sQTLs discovery was also positively correlated with pseudobulk library size (Pearson's $r = 0.96$; Extended Data Fig. 4b). Sample size and library size independently contributed to sQTL discovery power (likelihood ratio test; sample size: $P = 3.68 \times 10^{-10}$; library size: $P = 1.14 \times 10^{-10}$). Across all cell types, we observed a maximum of 16.1% of sQTL genes (sGenes) with two or more regulatory variants (Fig. 3a). We observed that allelic heterogeneity scaled with the number of significant sGenes to detect sQTLs (Fig. 3e).

To replicate *cis*-sQTLs with external datasets, we used four publicly available PBMC-related bulk RNA-seq datasets: (1) the BLUEPRINT dataset[13] ($n = 197$); (2) the DICE dataset[12] ($n = 91$); (3) the GTEx whole-blood ($n = 670$) and lymphoblastoid cell lines (LCLs) ($n = 147$); and (4) ImmuNexUT[35] ($n = 416$) (Extended Data Fig. 5). We used the $\pi 1$ statistics to assess replication between AIDA and matching cell types from the aforementioned datasets. $\pi 1$ estimates ranged from 0.91 to 0.93 for BLUEPRINT, 0.83 to 0.91 for DICE, 0.70 to 1 for GTEx whole-blood (except for cDC2; $\pi 1 = 0.48$), 0.71 to 0.86 for GTEx LCL and 0.70 to 1 for ImmuNexUT (except for CD56$^+$ NK; $\pi 1 = 0.42$) (Fig. 3f and Supplementary Fig. 2), suggesting that AIDA *cis*-sQTLs replicated well. Despite high replication rates, we identified an average of 14.31% (range = 4.76–19.32%) new sQTLs compared with ImmuNexUT, the largest existing PBMC dataset (Extended Data Fig. 5).

To identify cell-type-specific sQTLs, we quantified sQTL sharing in terms of sign and magnitude using mash[36]. sQTLs tended to have the same direction of effects but different effect size magnitudes (Fig. 3g). Neighboring cell-type pairs on the hematopoietic lineage shared sQTLs more than other pairs (Fig. 3h). For example, myeloid and B cells formed two distinct clusters. We identified many cell-type-specific sQTLs in known autoimmune risk genes[37]. One example was *CCL4*, which encodes a cytokine upregulated in systemic lupus erythematosus (SLE)-associated autoimmune hemolytic anemia[38]. Its sQTL effects preferentially appeared in T cell subtypes (Extended Data Fig. 6).

Sex disparities in disease risk might also be attributed to sex-biased sQTLs (sQTL)[39]. For each independent lead *cis*-sQTL, we performed sex-biased sQTL analysis with a linear model to test for genotype-by-sex (G × S) interaction. We identified a total of 27 sex-biased sQTLs across 20 genes (FDR < 0.05; Fig. 3i and Supplementary Table 5). Among these, two sex-biased sQTLs had significant effects in only one sex. For example, the lead variant rs930090 modulated *TECR* intron use (chromosome 19: 14529711–14562525) in CD16$^+$ NK cells for females ($P = 6.42 \times 10^{-19}$; $\hat{\beta} = 0.359$) but not males ($P = 0.043$; $\hat{\beta} = 0.125$; Extended Data Fig. 4d). Another 25 sex-biased sQTLs appeared in the two sexes with different effect sizes. The lead sVariant rs17713729 had a significant sQTL effect on the *SH3YL1* intron (chromosome 2: 253115–264782) in CD4$^+$ central memory T (T_{CM}) cells for both males ($P = 6.41 \times 10^{-30}$) and females ($P = 2.49 \times 10^{-14}$), but its effect size was larger in males ($\hat{\beta} = 0.523$) than in females ($\hat{\beta} = 0.319$; Extended Data Fig. 4e). Using 20 GWAS traits conducted in East Asian populations[40–44] (Supplementary Table 6), we identified a sex-biased colocalization between lymphocyte count and *CLEC2D* sQTL in CD4$^+$ effector memory T (T_{EM}) cells. The sQTL had a stronger effect on *CLEC2D* splicing in females ($\hat{\beta} = 0.817$, $P = 5.5 \times 10^{-21}$) than males ($\hat{\beta} = 0.499$, $P = 3.3 \times 10^{-6}$) and stronger colocalization with lymphocyte count in females (H4 = 0.935) than males (H4 = 0.676; Fig. 3j). *CLEC2D* encodes a transmembrane C-type lectin receptor that causes inflammation and tissue injury[45]. Further experimental evidence is needed to understand its function affecting lymphocyte count.

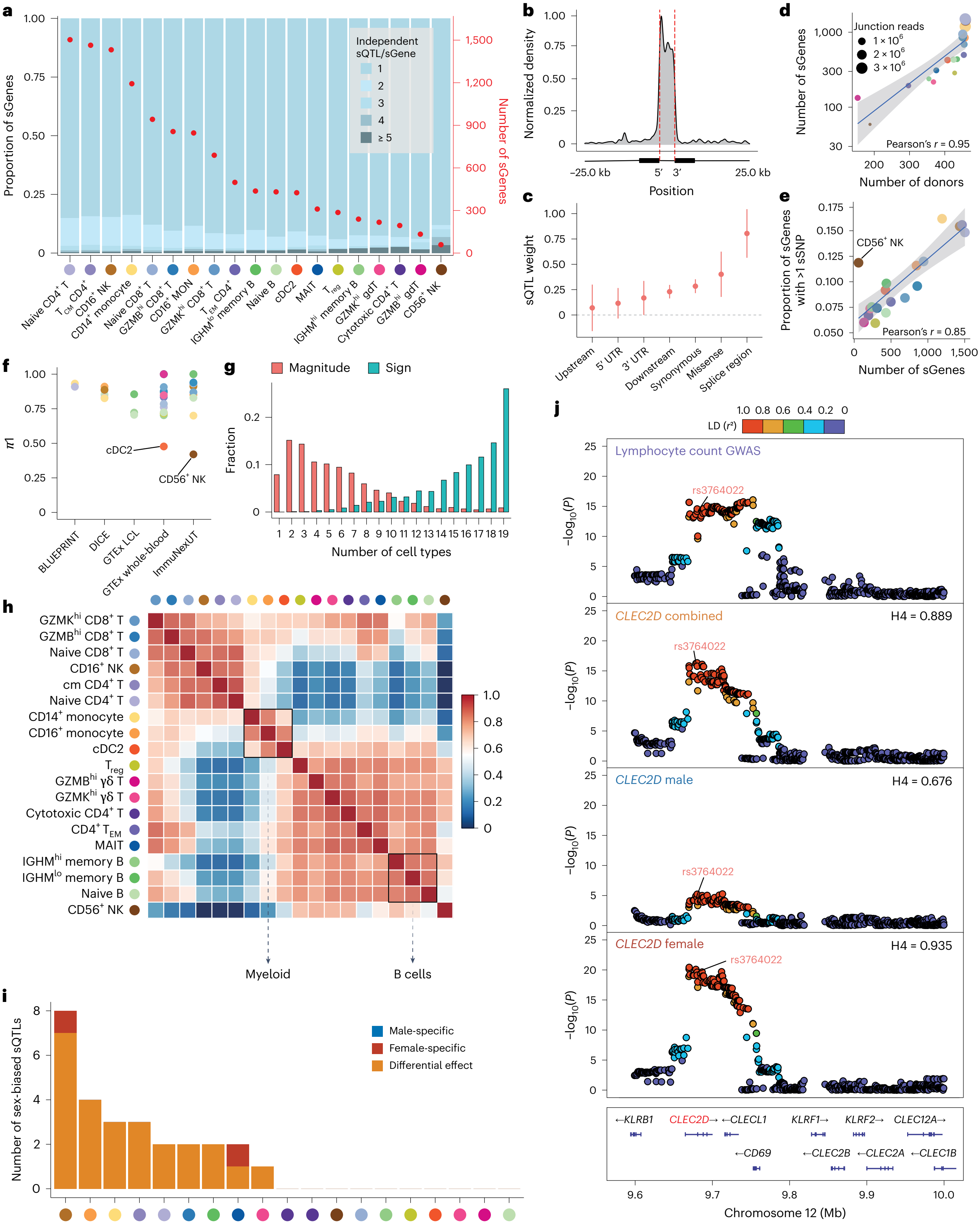
We performed ancestry-biased sQTL detection by stratifying the AIDA cohort into East Asian, Southeast Asian and South Asian. We observed eight Malay-biased *cis*-sQTLs and 19 Indian-biased *cis*-sQTLs (Supplementary Table 7). One example of Malay-biased sQTL is the lead variant rs492083, which modulates *ATP5MPL* intron use (chromosome 14: 103914633–103915066) in CD16$^+$ monocytes ($P = 2.10 \times 10^{-8}$; $\hat{\beta} = -0.995$; Extended Data Fig. 4f). One example of Indian-biased sQTL is the lead variant rs6576010, which modulates *POLB* intron use (chromosome 8: 42338685–42344953) in naive CD4$^+$ T cells ($P = 1.31 \times 10^{-7}$; $\hat{\beta} = -0.946$; Extended Data Fig. 4g).

Dynamic intron use and sQTL across B cell development

B cell differentiation is crucial for the immune system[46]. We focused on naive ($n = 26{,}617$), IGHM$^{\mathrm{lo}}$ memory ($n = 16{,}280$) and IGHM$^{\mathrm{hi}}$ memory B cells ($n = 10{,}067$) to explore how splicing regulation correlates with the different stages of B cell development (Fig. 4a). We assigned pseudotime to each cell along a differentiation trajectory (Fig. 4b) and split the population into six quantiles based on pseudotime values (Fig. 4c). In this trajectory, memory B cells underwent class switch recombination from IgM (IGHM$^{\mathrm{hi}}$) to other isotypes (IGHM$^{\mathrm{lo}}$), in agreement with previous studies (Fig. 4d)[47]. We identified 865 introns whose use correlated with pseudotime (FDR < 0.05; Supplementary Table 8). Dynamic introns showed three distinct modes of pseudotime-dependent use (Fig. 4e). The first mode was a stepwise change between naive and memory B cells. *PAX5* is a B cell-specific transcription factor that has a critical role in B cell development (Fig. 4f and Extended Data Fig. 7)[48,49]. Its two isoforms were expressed throughout B cell development, but there was no consensus regarding their isoform-specific roles in the current literature[49–51]. We observed *PAX5A* downregulation and *PAX5B* upregulation as B cells matured. The second mode showed a steady linear change in intron junction use, like the decrease in exon 4 use for *PTPRC* during naive to memory B cell differentiation. The final quadratic mode involved transient intron use, such as transient upregulation of exon 1 in *DOCK8* during B cell development (Fig. 4f and Extended Data Fig. 7).

Next, we sought to identify dynamic genetic effects on splicing across the B cell differentiation trajectory. We found 107 lead sQTLs with significant pseudotime interactions (FDR < 0.05, Fig. 4g and

Fig. 3 | Single-cell sQTL analysis revealed cell-type-specific and sex-biased regulation of splicing. a, Numbers of sGenes (red dots) and proportions of sGenes (stacked bars) with various numbers of independent sQTLs across 19 cell types (adjusted beta-approximated $P < 0.05$). **b**, *cis*-sVariants preferentially located near splice junctions and in the affected introns. **c**, A Bayesian hierarchical model revealed that sVariants were enriched in the splice region and as missense and synonymous variants. The dot plot shows the mean ± s.e.m. of functional annotations (n of sVariants = 11,577). **d**, Number of sGenes scaled with the number of donors and junction read count across 19 cell types. The shaded area on either side of the linear regression line represents the 95% CI. **e**, The proportion of sGenes with more than one independent sVariant increased with the power of sGene discovery across 19 cell types. The shaded area on either side of the linear regression line represents the 95% CI. **f**, AIDA *cis*-sQTLs were well replicated in BLUEPRINT, DICE, GTEx LCL, GTEx whole-blood and ImmuNexUT. Each dot represents one cell type from AIDA, colored as in **a**. **g**, Fractions of lead *cis*-sQTLs shared according to sign and magnitude in one or more cell types. Sharing according to sign was defined as a *cis*-sQTL sharing the same sign with the top *cis*-sQTL across 19 cell types. Sharing according to magnitude was defined as the effect size of a *cis*-sQTL being within a factor of two of the top *cis*-sQTLs across 19 cell types. **h**, Pairwise sQTL sharing according to magnitude across 19 cell types. A total of 2,488 sQTLs that were significant (linear feedback shift register (LFSR) < 0.05) in at least one cell type were considered to avoid random noise in association testing. **i**, Number of sex-biased sQTLs discovered in 19 cell types (FDR < 0.05). Cell type coloring as in **a**. **j**, *CLEC2D* sQTLs in CD4$^+$ T_{EM} cells colocalized with the GWAS of lymphocyte count. This colocalization was primarily driven by a female-biased sQTL. The sQTL lead variant rs3764022 was an exonic variant located in the splice region of *CLEC2D* exon 2. The unadjusted two-sided P value was calculated using QTLtools.


a
Proportion of sGenes
1.00
0.75
0.50
0.25
0
Independent sQTL/sGene
1
2
3
4
≥ 5
Number of sGenes
1,500
1,200
900
600
300
0
Naive CD4+ T
TCM CD4+
CD16+ NK
CD14+ monocyte
Naive CD8+ T
GZMBhi CD8+ T
CD16+ MON
GZMKhi CD8+ T
TEM CD4+
IGHMlo memory B
Naive B
cDC2
MAIT
Treg
IGHMhi memory B
GZMKhi gdT
Cytotoxic CD4+ T
GZMBhi gdT
CD56+ NK
b
Normalized density
1.00
0.75
0.50
0.25
0
−25.0 kb
5′
3′
25.0 kb
Position
c
sQTL weight
1.00
0.75
0.50
0.25
0
Upstream
5′ UTR
3′ UTR
Downstream
Synonymous
Missense
Splice region
d
Number of sGenes
1,000
300
100
30
Junction reads
1 × 10^6
2 × 10^6
3 × 10^6
Pearson's r = 0.95
200
300
400
Number of donors
e
Proportion of sGenes with >1 sSNP
0.175
0.150
0.125
0.100
0.075
0.050
CD56+ NK
Pearson's r = 0.85
0
500
1,000
1,500
Number of sGenes
f
π1
1.00
0.75
0.50
0.25
0
cDC2
CD56+ NK
BLUEPRINT
DICE
GTEx LCL
GTEx whole-blood
ImmuNexUT
g
Magnitude
Sign
Fraction
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Number of cell types
h
GZMKhi CD8+ T
GZMBhi CD8+ T
Naive CD8+ T
CD16+ NK
cm CD4+ T
Naive CD4+ T
CD14+ monocyte
CD16+ monocyte
cDC2
Treg
GZMBhi γδ T
GZMKhi γδ T
Cytotoxic CD4+ T
CD4+ TEM
MAIT
IGHMhi memory B
IGHMlo memory B
Naive B
CD56+ NK
1.0
0.8
0.6
0.4
0.2
0
Myeloid
B cells
i
Number of sex-biased sQTLs
8
6
4
2
0
Male-specific
Female-specific
Differential effect
j
LD (r²)
1.0 0.8 0.6 0.4 0.2 0
Lymphocyte count GWAS
rs3764022
−log10(P)
25
20
15
10
5
0
CLEC2D combined
H4 = 0.889
CLEC2D male
H4 = 0.676
CLEC2D female
H4 = 0.935
←KLRB1
CLEC2D→
←CLECL1
KLRF1→
KLRF2→
CLEC12A→
←CD69
←CLEC2B
←CLEC2A
←CLEC1B
9.6
9.7
9.8
9.9
10.0
Chromosome 12 (Mb)

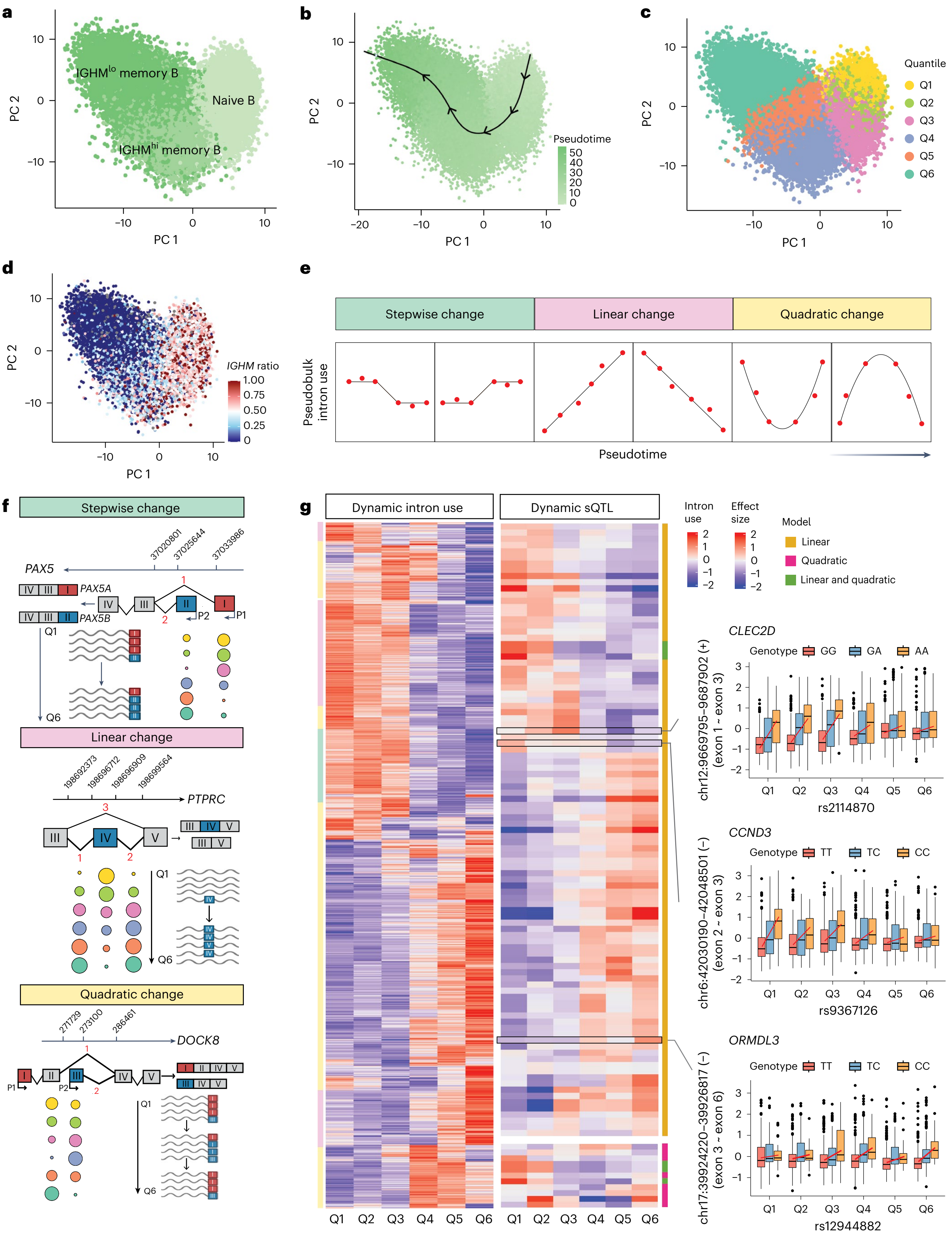

a
IGHM^{lo} memory B
Naive B
IGHM^{hi} memory B
PC 1
PC 2
b
Pseudotime
50
40
30
20
10
0
c
Quantile
Q1
Q2
Q3
Q4
Q5
Q6
d
IGHM ratio
1.00
0.75
0.50
0.25
0
e
Stepwise change
Linear change
Quadratic change
Pseudobulk intron use
Pseudotime
f
Stepwise change
PAX5
37020801
37025644
37033986
PAX5A
PAX5B
P1
P2
Q1
Q6
Linear change
PTPRC
198692373
198696712
198696909
198699564
Quadratic change
DOCK8
271729
273100
286461
g
Dynamic intron use
Dynamic sQTL
Intron use
Effect size
Model
Linear
Quadratic
Linear and quadratic
CLEC2D
Genotype GG GA AA
chr12:9669795–9687902 (+) (exon 1 ~ exon 3)
rs2114870
CCND3
Genotype TT TC CC
chr6:42030190–42048501 (−) (exon 2 ~ exon 3)
rs9367126
ORMDL3
Genotype TT TC CC
chr17:39924220–39926817 (−) (exon 3 ~ exon 6)
rs12944882
Q1 Q2 Q3 Q4 Q5 Q6

Fig. 4 | Dynamic intron use and sQTLs identified through B cell development. **a**, Principal component (PC) projections of single-cell gene expression for naive, IGHM^{hi} memory and IGHM^{lo} memory B cells. **b**, Pseudotime projection of 52,964 B cells. The direction of the curve and the intensity of the green color indicate the dynamic process of B cell maturation from naive to IGHM^{hi} memory and to IGHM^{lo} memory B cells. **c**, B cells were partitioned into six quantiles according to pseudotime values. **d**, Dynamic expression of *IGHM* during cellular development agreed with B cell class switch recombination from producing IgM to other isotypes. *IGHM* ratio: *IGHM* expression level/(*IGHM* + *IGHG1* + *IGHG2* + *IGHG3* + *IGHG4* + *IGHA1* + *IGHA2* + *IGHD* + *IGHE*) expression level. **e**, Three distinct patterns were identified for pseudotime-dependent intron use: stepwise, linear and quadratic. **f**, Dynamic intron use across six quantiles of B cell development. Three example genes with different dynamic intron use patterns (top, stepwise change in *PAX5*; middle, linear change in *PTPRC*; bottom, quadratic change in *DOCK8*). The dot color corresponds to the six quantiles in **c** and the dot size reflects the mean intron usage in that quantile. **g**, Left, heatmap of scaled mean intron use across pseudotime, with the color bar corresponding to the three dynamic intron use patterns in **e**. sVariant–intron pairs with significant interaction effects with B cell pseudotime are shown. Both linear (genotype × time) and quadratic (genotype × time^2) models were used to assess the interaction between genetic and pseudotime quantiles. Middle, scaled effect size estimates of sVariant–intron pairs. Right, three example genes (*CLEC2D*, *CCND3*, *ORMDL3*) with dynamic effect sizes across pseudotime. The samples sizes for each quantile are: Q1 ($n = 419$), Q2 ($n = 425$), Q3 ($n = 427$), Q4 ($n = 450$), Q5 ($n = 448$) and Q6 ($n = 449$).

Supplementary Table 9). Of these dynamic sQTLs, the effect sizes of 62 (57.94%) grew stronger, 37 (35.58%) grew weaker and 11 (10.28%) showed transient effects along the differentiation trajectory. Several dynamic sGenes were known to be immune-related, including *CLEC2D*[45], *CCND3* (ref. 52) and *ORMDL3* (ref. 53). Notably, we found several dynamic sQTLs that colocalized with autoimmune disease GWAS loci. For example, the dynamic sQTL for *CD83*, whose effect of lead variant rs6936285 weakened during B cell maturation, strongly colocalized with RA (COLOC H4 = 0.980) (Extended Data Fig. 8a). Additionally, the dynamic sQTL for *BCL2A1*, whose effect of lead variant rs16971619 increased during B cell maturation, colocalized with lymphocyte count (COLOC H4 = 0.848). These findings suggest a functional link between dynamic sQTLs and autoimmune diseases (Extended Data Fig. 8b).

trans-sQTLs are highly cell-type-specific

We performed *trans*-sQTL mapping to identify distal genetic effects on splicing (Methods) and identified 607 *trans*-sGenes (FDR < 0.01; Fig. 5a). The number of detected *trans*-sGenes was positively correlated with sample size (Spearman's $\rho = 0.74$) and with the number of *cis*-sGenes (Spearman's $\rho = 0.62$; Fig. 5b).

Most detected *trans*-sGenes were cell-type-specific. Out of all *trans*-sGenes, 393 (64.7%) appeared in only one cell type, which was similar to the OneK1K estimate (63.6%). *Trans*-sQTLs $\pi 1$ replication between cell types was significantly lower than *cis*-sQTLs (two-sided *t*-test $P = 2.91 \times 10^{-79}$, $t = -28.1$, d.f. = 250; Fig. 5c). Additionally, we noticed that the effect size differences between discovery and replication cell types were larger for *trans*-sQTLs than *cis*-sQTLs (two-sided *t*-test $P = 2.81 \times 10^{-91}$, $t = -31.4$, d.f. = 250; Supplementary Fig. 3). This suggests that *trans*-sQTLs are more cell-type-specific than *cis*-sQTLs.

We asked whether the *cis*-effects on RNA-binding proteins could underlie *trans*-sQTLs. We identified 17 colocalization events between AIDA cell-type-specific *cis*-eQTLs and *trans*-sQTLs (COLOC H4 > 0.75; Fig. 5d and Supplementary Table 10). Notably, *PTPRC trans*-sQTLs exhibited T cell-biased colocalization with *HNRNPLL cis*-eQTLs (Fig. 5e,f and Extended Data Fig. 9a,b). Furthermore, single-cell quantifications also suggested that higher *HNRNPLL* expression led to shorter *PTPRC* isoforms (two-sided *t*-test $P = 3.07 \times 10^{-197}$, $t = -30.0$, d.f. = 81578.31; Fig. 5g).

Notably, we observed a lead switch across CD4^+ T cell subtypes. In naive CD4^+ T cells, rs6751481 was the lead SNP for *HNRNPLL cis*-eQTLs and *PTPRC trans*-sQTLs (H4 = 1.00). However, the lead SNP rs74258942 in CD4^+ T_{EM} cells (H4 = 0.998) was 36,158 bp away and in modest linkage disequilibrium (LD) with rs6751481 ($r^2 = 0.28$). Fine-mapping with SuSiE[54] confirmed that both lead SNPs were the most likely causal SNPs for their respective cell types (Fig. 5j). We observed that both lead SNPs overlapped functional elements on HaploReg: rs6751481 overlapped a blood-specific promoter region, while rs74258942 overlapped a blood-specific enhancer region. In addition, we also replicated rs6751481 in OneK1K. It was the lead SNP for *HNRNPLL* in CD4^+ naive T cells (FDR < 0.15) but not in CD4^+ T_{EM} cells (FDR > 0.9). The variant rs74258942 was Asian-biased (OneK1K MAF = 0.07; AIDA MAF = 0.32; 1KG_EUR = 0.07; 1KG_EAS = 0.29) and was not significant in OneK1K (FDR > 0.99).

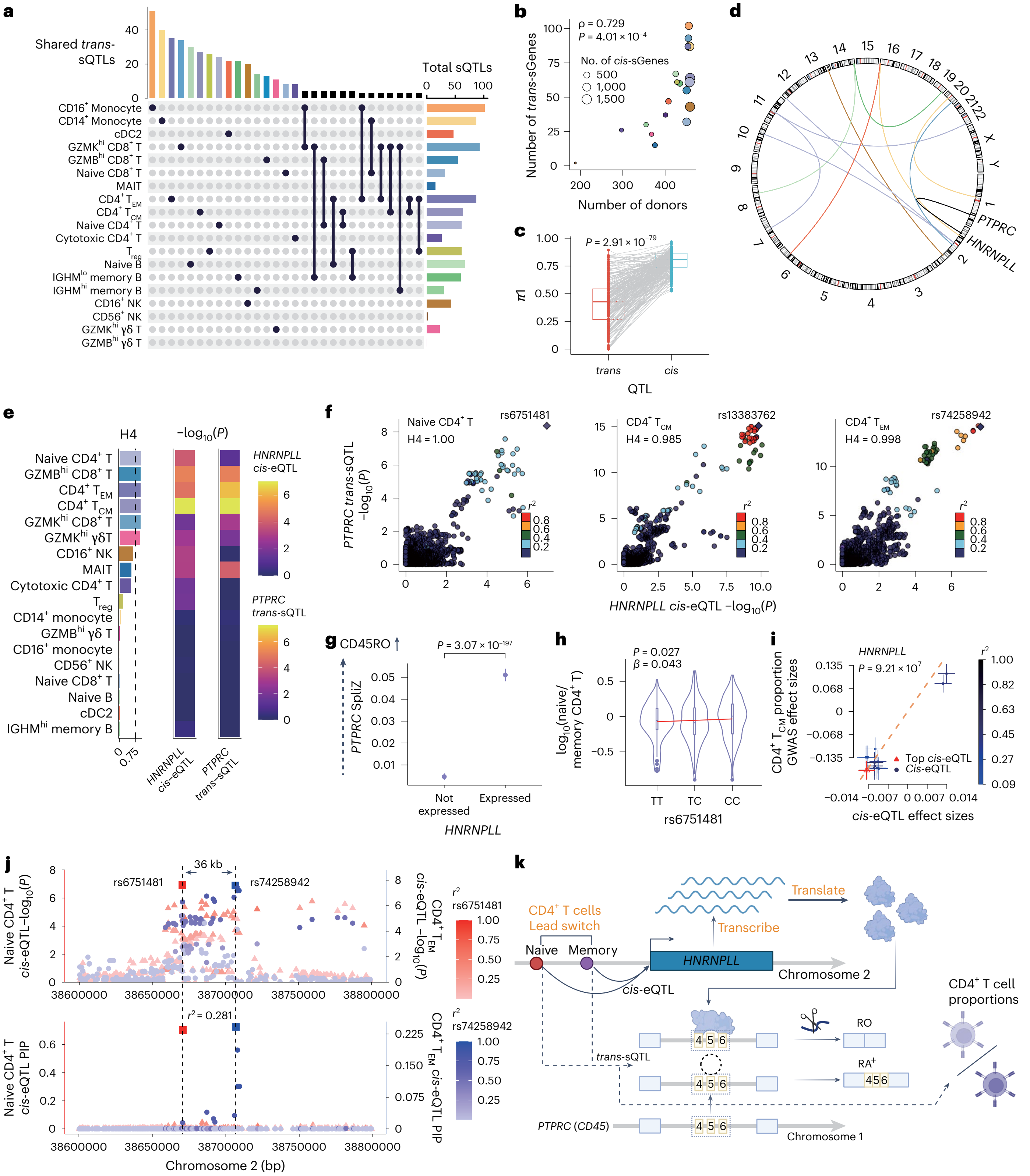
Isoform-specific expression of *PTPRC* (*CD45*) is associated with naive CD4^+ T cell differentiation. Because rs6751481 was associated with the CD45RA^+ and RO isoform ratios in CD4^+ T naive cells, we tested whether it would be associated with CD4^+ T cell proportions and observed a modest but significant association ($\hat{\beta} = 0.043$; $P = 0.027$; Fig. 5h). As an orthogonal validation, we leveraged a previous GWAS study on T cell proportions[55] and conducted summary-based Mendelian randomization (SMR)[56]. We observed strong pleiotropy between *HNRNPLL* expression and the proportion of memory CD4^+ T cells ($P = 9.21 \times 10^{-7}$; Fig. 5i) without significant linkage effect (HEIDI $P = 0.103$). These results showed that germline variations influencing *HNRNPLL* expression and *PTPRC* splicing also influenced T cell proportions (Fig. 5k).

Fig. 5 | *trans*-sQTL analysis revealed a regulatory relationship between *HNRNPLL* and *PTPRC*. **a**, Upset plot showing discovery and sharing of *trans*-sQTLs across cell types. Right, the bar plot shows the number of *trans*-sQTLs per cell type. Top, the bar plot shows the number of *trans*-sQTLs in each category. The *x* axis is truncated at a minimum of five sQTLs. **b**, The number of *trans*-sGenes scaled with the number of donors. The two-sided *P* value was calculated using Spearman's rank correlation. **c**, Box plot of the $\pi 1$ statistics for *cis*-sQTLs and *trans*-sQTLs. The *P* value was calculated using a two-sided paired *t*-test ($n = 251$ for *trans*-sQTLs; $n = 251$ for *cis*-sQTLs). **d**, Circos plot revealing the *cis*-regulatory effects (*cis*-eQTLs) underlying *trans*-sQTLs (links colored according to cell type as in **a**). A link is black if a colocalization event occurred in multiple cell types. **e**, Bar plot and heatmaps showing the colocalization probability (COLOC PP: H4) between *HNRNPLL cis*-eQTL and *PTPRC trans*-sQTL and QTL *P* values. In **e**,**f**,**j**, Unadjusted *P* values were obtained using Matrix eQTL (*cis*-eQTL) and QTLtools (*trans*-sQTL). **f**, LocusCompare plots showing the colocalization between *HNRNPLL cis*-eQTL and *PTPRC trans*-sQTL in CD4^+ T (naive, T_{CM} and T_{EM}) cells. **g**, Higher SpliZ scores (representing more isoforms with longer intron length) were observed in single cells with greater *HNRNPLL* expression. The dot plot shows the mean and 95% CI. The *P* value was calculated using a two-sided *t*-test ($n = 214{,}504$ for 'not expressed'; $n = 53{,}064$ for 'expressed'). **h**, Violin and box plots showing that rs6751481 was associated with the ratio between naive and memory CD4^+ T cells across AIDA donors. The *P* value and β were determined using linear regression (red line; $n = 96$ for TT; $n = 217$ for TC; $n = 114$ for CC). **i**, SMR revealed strong pleiotropy between *HNRNPLL cis*-eQTLs and GWAS on activated T cell proportion. The *P* value was obtained using SMR ($n = 3579$ for all the input variants). The SMR effect plot shows the mean ± s.e.m. of the variant effects. **j**, LocusZoom plot showing that naive and CD4^+ T_{EM} cells harbored two independent lead SNPs for *HNRNPLL cis*-eQTLs (square: lead SNPs for naive and CD4^+ T_{EM} cells; triangle: remaining SNPs for naive CD4^+ T cells; circle: remaining SNPs for CD4^+ T_{EM} cells). Bottom, SuSiE posterior inclusion probability (PIP). The LD between rs6751481 and rs74258942 was modest ($r^2 = 0.28$). **k**, Schematic showing the proposed regulatory relationship between *HNRNPLL cis*-eQTLs and *PTPRC trans*-sQTLs.

https://doi.org/10.1038/s41588-024-02019-8

Aberrant splicing mediates complex diseases

To identify the AS mechanisms that mediated polygenic disease risk, we compiled GWAS summary statistics for 20 traits focused on Asian populations (Supplementary Table 6). We conducted colocalization between *cis*-sQTLs from 19 cell types and GWAS for 20 complex traits with COLOC[57] (Fig. 6a and Extended Data Fig. 10a). The proportion of colocalized GWAS loci varied across different GWAS traits, with immune-related diseases having the highest proportion, while nonimmunological traits had the lowest proportion of colocalization. We applied stratified LD score regression (S-LDSC) to estimate heritability enrichment from GWAS summary statistics[58]. The heritability of SLE, atopic dermatitis (AD), GD and RA showed higher enrichment in PBMC sQTLs than other traits (Fig. 6b and Extended Data Fig. 10e,f).



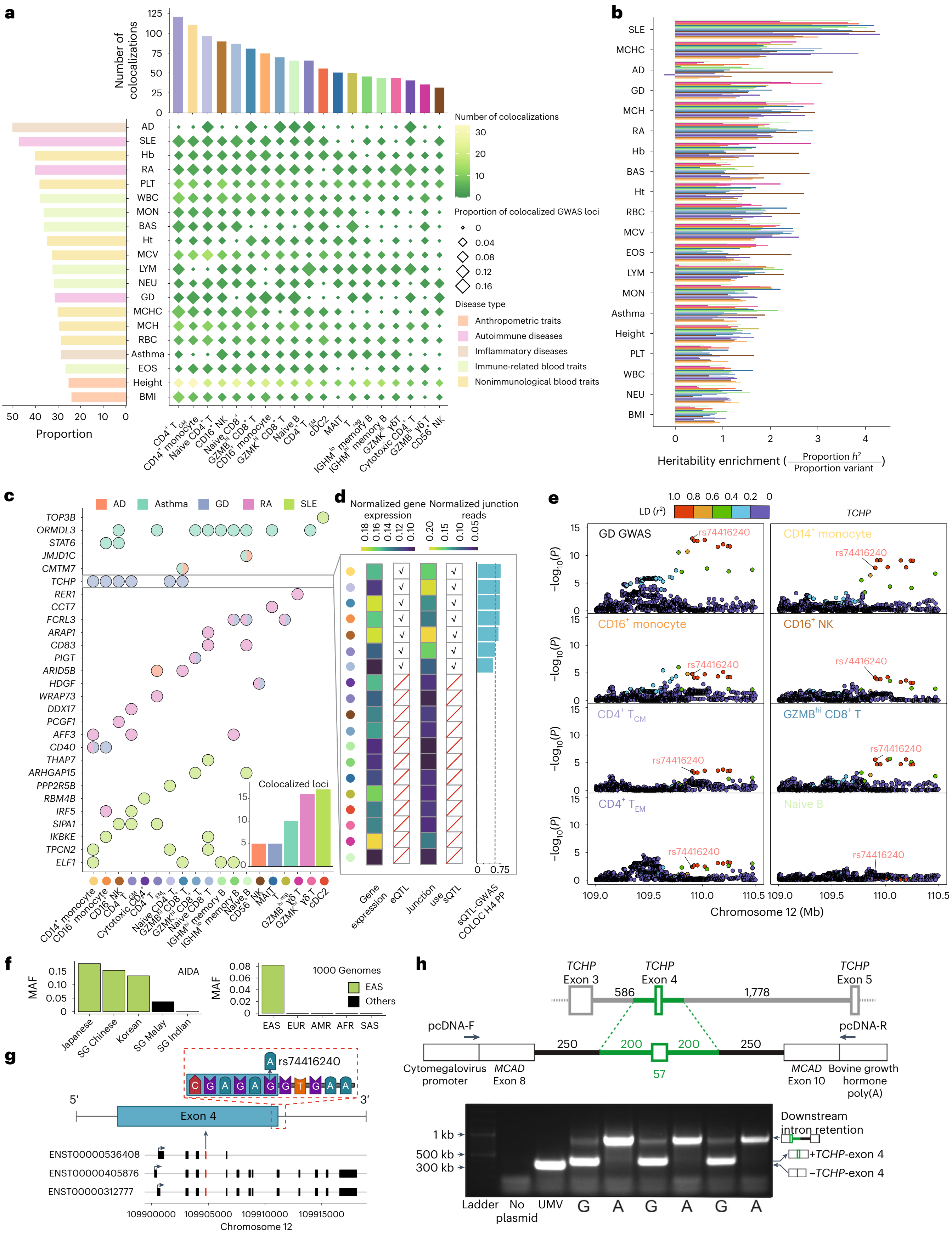

a
Number of colocalizations
125
100
75
50
25
0
AD
SLE
Hb
RA
PLT
WBC
MON
BAS
Ht
MCV
LYM
NEU
GD
MCHC
MCH
RBC
Asthma
EOS
Height
BMI
50 40 30 20 10 0
Proportion
CD4+ TCM
CD14+ monocyte
Naive CD4+ T
CD16+ NK
Naive CD8+ T
GZMBhi CD8+ T
CD16+ monocyte
GZMKhi CD8+ T
Naive B
CD4+ TEM
cDC2
MAIT
Treg
IGHMlo memory B
IGHMhi memory B
GZMKhi γδT
Cytotoxic CD4+ T
GZMBhi γδT
CD56+ NK
Number of colocalizations
30
20
10
0
Proportion of colocalized GWAS loci
0
0.04
0.08
0.12
0.16
Disease type
Anthropometric traits
Autoimmune diseases
Imflammatory diseases
Immune-related blood traits
Nonimmunological blood traits
b
SLE
MCHC
AD
GD
MCH
RA
Hb
BAS
Ht
RBC
MCV
EOS
LYM
MON
Asthma
Height
PLT
WBC
NEU
BMI
0 1 2 3 4
Heritability enrichment (Proportion h2 / Proportion variant)
c
AD
Asthma
GD
RA
SLE
TOP3B
ORMDL3
STAT6
JMJD1C
CMTM7
TCHP
RER1
CCT7
FCRL3
ARAP1
CD83
PIGT
ARID5B
HDGF
WRAP73
DDX17
PCGF1
AFF3
CD40
THAP7
ARHGAP15
PPP2R5B
RBM4B
IRF5
SIPA1
IKBKE
TPCN2
ELF1
Colocalized loci
15
10
5
0
CD14+ monocyte
CD16 monocyte
CD16 NK
CD4+ TCM
Cytotoxic CD4+ T
CD4+ TEM
Naive CD4+ T
GZMBhi CD8+ T
GZMKhi CD8+ T
Naive CD8 T
IGHMhi memory B
IGHMlo memory B
Naive B
CD56 NK
MAIT
Treg
GZMBhi γδT
GZMKhi γδT
cDC2
d
Normalized gene expression
0.18 0.16 0.14 0.12 0.10
Normalized junction reads
0.20 0.15 0.10 0.05
√
0 0.75
Gene expression
eQTL
Junction use
sQTL
sQTL-GWAS COLOC H4 PP
e
LD (r2)
1.0 0.8 0.6 0.4 0.2 0
TCHP
GD GWAS
rs74416240
CD14+ monocyte
CD16+ monocyte
CD16+ NK
CD4+ TCM
GZMBhi CD8+ T
CD4+ TEM
Naive B
−log10(P)
15
10
5
0
109.0 109.5 110.0 110.5
Chromosome 12 (Mb)
f
MAF
0.15
0.10
0.05
0
AIDA
Japanese
SG Chinese
Korean
SG Malay
SG Indian
0.08
0.06
0.04
0.02
0
1000 Genomes
EAS
Others
EAS EUR AMR AFR SAS
g
A rs74416240
C G A G A G G T G A A
5′
3′
Exon 4
ENST00000536408
ENST00000405876
ENST00000312777
109900000 109905000 109910000 109915000
Chromosome 12
h
TCHP Exon 3
TCHP Exon 4
TCHP Exon 5
586
1,778
pcDNA-F
pcDNA-R
250
200
200
250
57
Cytomegalovirus promoter
MCAD Exon 8
MCAD Exon 10
Bovine growth hormone poly(A)
1 kb
500 kb
300 kb
Ladder
No plasmid
UMV
G A G A G A
Downstream intron retention
+TCHP-exon 4
−TCHP-exon 4

Fig. 6 | Aberrant splicing mediates complex diseases. a, Cell-type-specific colocalization between *cis*-sQTLs from 19 cell types and 20 complex traits. **b**, Heritability enrichment (proportion h^2/proportion variant) for 20 traits mediated by *cis*-sQTLs from 19 cell types. Autoimmune and inflammatory diseases are highlighted in bold. **c**, Colocalization for 28 example sGenes across 19 cell types in the five disease traits. The color of each circle indicates the associated diseases. The inset shows the total number of colocalized loci across the five diseases. **d**, Gene expression, eQTLs, junction reads, sQTLs and H4 posterior probability (sQTL-GWAS colocalization) for *TCHP* across 19 cell types. High junction use between exons 4 and 5 led to sQTL and sQTL-GWAS colocalization. **e**, Cell-type-specific colocalization of GD GWAS and *TCHP* sQTLs in seven cell types. rs74416240 was the lead GWAS risk variant. The unadjusted, two-sided *P* value was calculated using QTLtools. **f**, MAF of rs74416240 in five AIDA populations and five major populations in the 1000 Genomes Project showed an East Asian bias of the rs74416240 minor allele. **g**, Gene model of *TCHP* with three isoforms. rs74416240 was located in the 5′ splice site of the intron junction between exons 4 and 5. **h**, Minigene experiment to validate the effect of rs74416240 on *TCHP* exon 4 splicing in K562 cells. The universal minigene vector (UMV) backbone alone corresponded to the band with the smallest molecular weight on the gel image. The test region, containing the 57-nt long exon 4 plus the 200-bp flanking sequences, was cloned into the UMV. Two identical minigene constructs with one nucleotide difference at rs74416240 (reference = G; alternative = A) were transfected into K562 cells. The reference allele (G) predominantly led to the normal isoform; the alternative allele (A) led to intron retention. BAS, basophil count; BMI, body mass index; EOS, eosinophil; Hb, hemoglobin; Ht, hematocrit; MCH, mean corpuscular Hb; LYM, lymphocyte; MCHC, MCH concentration; MCV, mean corpuscular volume; MON, monocyte count; NEU, neutrophil; PLT, platelet count; RBC, red blood cell; WBC, white blood cell count.

We identified 53 colocalized loci among the five autoimmune and inflammatory diseases (COLOC H4 > 0.75; Fig. 6c and Supplementary Tables 11 and 12). Our results identified cell-type-specific colocalization between the SLE GWAS and an sQTL that regulates the splicing of *IRF5* exon 1 (Extended Data Fig. 10b,c). *IRF5* is a well-known risk gene for SLE[59]. The putative causal SNP rs2004640 disrupted the 5′ splice site of exon 1B, leading to nonsense-mediated decay and downregulation of *IRF5* expression (Extended Data Fig. 10d).

We turned our attention to GD, an autoimmune hyperthyroidism with a significantly higher incidence rate in Asian than European populations[60]. Colocalization analysis captured *TCHP* as a cell-type-specific risk gene for GD. Cell types with high use of the intron junction between *TCHP* exons 4 and 5 harbored *cis*-sQTLs, which led to cell-type-specific sQTL-GWAS colocalization (Fig. 6d). Furthermore, *cis*-sQTL effects led to intron retention between exons 4 and 5 and nonsense-mediated decay, which manifested as *cis*-eQTLs (Fig. 6d). A recent study in the Japanese population identified a GWAS locus associated with GD ($P = 8.6 \times 10^{-14}$)[42]. The GWAS lead variant rs74416240 had a significant genetic effect on *TCHP* exon 4 use in monocytes, NK, CD4$^+$ and CD8$^+$ T cells (FDR < 0.05) and strongly colocalized with *TCHP cis*-sQTLs (Fig. 6d,e). Notably, the MAF of rs74416240 was high in the East Asian population (Japanese = 0.18; Chinese = 0.15, Korean = 0.13), modest in the Southeast Asian (Malay = 0.04) population and absent in the South Asian population in AIDA. We observed the same pattern of allele frequencies in 1000 Genomes Project populations (Fig. 6f).

The lead variant rs74416240 resided in the last nucleotide of *TCHP* exon 4 and was predicted to disrupt the 5′ splice junction (Fig. 6g). To validate this effect, we conducted a minigene experiment to test the effect of rs74416240 on *TCHP* exon 4 splicing in K562 cells. The test region contained a 57-nt-long exon 4 plus the 200-bp flanking sequence (Fig. 6h). We transfected two identical minigene constructs with one nucleotide difference at rs74416240. A low level of intron retention was observed for the reference allele (G) construct. The alternative allele (A) construct, where the 5′ splice site was disrupted, revealed a nearly complete intron retention isoform (Fig. 6h). These results suggest that rs74416240 is a causal variant for the *TCHP* sQTL and possibly contributes to GD risk via this differential splicing.

Discussion

AS is a key mediator of genetic effects on complex diseases. Our study leveraged droplet-based 5′ scRNA-seq to investigate cell-type-specific AS and sQTLs. Unlike FACS-based bulk RNA-seq studies[12,13,35], our study conducted unbiased sampling to capture low-abundance PBMC cell types. Compared to previous population-scale scRNA-seq PBMC studies[7,8], our study demonstrated a 4.3-fold increase in exon coverage by leveraging 5′ scRNA-seq, enabling better power for sQTL detection. In addition, our data filled a critical gap in expanding the catalog of under-represented Asian populations, especially Southeast and South Asians.

The ancestral diversity of our cohort enabled the identification of ancestry-biased splicing and sQTLs. Like previous studies[61], we found that ancestry had a stronger influence than sex on AS. Such influence can be partially attributed to differences in allele frequency: ancestry-biased *SPSB2* splice site use was caused by an MAF gradient from Eastern to Southeastern to South Asian populations. In addition, most Malay-biased and Indian-biased sQTLs had higher MAFs in the respective discovery populations, while the rest were explained by differences in effect sizes. Notably, we captured an Asian-specific causal variant that disrupted the 5′ splice site of *TCHP* exon 4 and led to intron retention. This variant colocalized with an Asian-specific GD risk locus identified in the Japanese population[42].

We identified 607 *trans*-sGenes across 19 cell types, which was significantly higher than previously identified by tissue-level bulk analysis[33]. Most *trans*-sGenes are specific to one cell type, which probably explains the scarcity of *trans*-sGenes when studying homogenized tissue samples. Notably, *trans*-sQTLs for *PTPRC* colocalized with *cis*-eQTLs for *HNRNPLL* in a cell-type-specific manner. Lead SNPs in naive and CD4$^+$ T_{EM} cells were more than 36-kbp apart and in modest LD; fine-mapping analysis suggested distinct causal variants across the T cell subtypes.

Our study highlights mechanistic findings and disease implications from single-cell sQTLs and underscores the importance of ancestral diversity. However, our findings must be interpreted within the limitations of our study. One limitation is that dynamic B cell analysis missed transitional states such as activated B cells and plasmablasts, which are typically observed during the adaptive immune response and not in healthy donors. Individuals undergoing active immune challenges are better suited to study B cell development. A second limitation of this study is that rare cell types are underpowered for sQTL discovery. Thus, we probably missed true positive sQTLs in these rare cell types. A third limitation was the uneven sequencing depth across the length of the gene. This problem could be addressed using full-length scRNA-seq library preparation methods, such as SmartSeq, or using long-read sequencing, such as Oxford Nanopore and PacBio. Because of the higher cost per sample and lower throughput, these methods have not been widely applied to large-scale sQTL studies. With the rapid advancement in genomic technologies, high-throughput and low-cost full-length scRNA-seq will enable unbiased and comprehensive coverage of the transcriptomic landscape.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-02019-8.

References

1. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
2. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
3. 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).
4. Yao, D. W., O'connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
5. Chun, S. et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
6. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
7. Yazar, S. et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
8. Perez, R. K. et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
9. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
10. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
11. Qi, T. et al. Genetic control of RNA splicing and its distinct role in complex trait variation.*Nat. Genet.* **54**, 1355–1363 (2022).
12. Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715 (2018).
13. Martens, J. H. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98**, 1487–1489 (2013).
14. Mu, Z. et al. The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol.* **22**, 122 (2021).
15. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
16. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
17. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
18. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
19. Fejes-Toth, K. et al. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
20. Kanamori-Katayama, M. et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**, 1150–1159 (2011).
21. Meyer, E., Chaung, K., Dehghannasiri, R. & Salzman, J. ReadZS detects cell type-specific and developmentally regulated RNA processing programs in single-cell RNA-seq. *Genome Biol.* **23**, 226 (2022).
22. Kock, K.H. et al. Single-cell analysis of human diversity in circulating immune cells. Preprint at *bioRxiv* https://doi.org/10.1101/2024.06.30.601119 (2024).
23. Olivieri, J. E., Dehghannasiri, R. & Salzman, J. The SpliZ generalizes 'percent spliced in' to reveal regulated splicing at single-cell resolution. *Nat. Methods* **19**, 307–310 (2022).
24. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
25. Inamo, J. et al. Long-read sequencing for 29 immune cell subsets reveals disease-linked isoforms. *Nat. Commun.* **15**, 4285 (2024).
26. Wilks, C., Gaddipati, P., Nellore, A. & Langmead, B. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics* **34**, 114–116 (2018).
27. Ergun, A. et al. Differential splicing across immune system lineages. *Proc. Natl Acad. Sci. USA* **110**, 14324–14329 (2013).
28. Janeway, C. A. Jr The T cell receptor as a multicomponent signalling machine: CD4/CD8 coreceptors and CD45 in T cell activation. *Annu. Rev. Immunol.* **10**, 645–674 (1992).
29. Oberdoerffer, S. et al. Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science* **321**, 686–691 (2008).
30. Lynch, K. W. Consequences of regulated pre-mRNA splicing in the immune system. *Nat. Rev. Immunol.* **4**, 931–940 (2004).
31. Pianta, A. et al. Two rheumatoid arthritis-specific autoantigens correlate microbial immunity with autoimmune responses in joints. *J. Clin. Investig.* **127**, 2946–2956 (2017).
32. Wu, D. et al. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* **179**, 736–749 (2019).
33. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
34. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
35. Ota, M. et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006–3021 (2021).
36. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
37. Stelzer, G. et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics* **54**, 1.30.1–1.30.33 (2016).
38. Ghafouri-Fard, S., Shahir, M., Taheri, M. & Salimi, A. A review on the role of chemokines in the pathogenesis of systemic lupus erythematosus. *Cytokine* **146**, 155640 (2021).
39. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).
40. Shirai, Y. et al. Multi-trait and cross-population genome-wide association studies across autoimmune and allergic diseases identify shared and distinct genetic component. *Ann. Rheum. Dis.* **81**, 1301–1312 (2022).
41. Tsuo, K. et al. Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. *Cell Genom.* **2**, 100212 (2022).
42. Sakaue, S. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
43. Ishigaki, K. et al. Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).
44. Wang, Y.-F. et al. Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. *Nat. Commun.* **12**, 772 (2021).
45. Lai, J.-J., Cruz, F. M. & Rock, K. L. Immune sensing of cell death through recognition of histone sequences by C-type lectin-receptor-2d causes inflammation and tissue injury. *Immunity* **52**, 123–135 (2020).
46. Inoue, T. & Kurosaki, T. Memory B cells.*Nat. Rev. Immunol.* **24**, 5–17 (2024).
47. Xu, Z., Zan, H., Pone, E. J., Mai, T. & Casali, P. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nat. Rev. Immunol.* **12**, 517–531 (2012).

48. O'Brien, P., Morin, P. Jr, Ouellette, R. J. & Robichaud, G. A. The *Pax-5* gene: a pluripotent regulator of B-cell differentiation and cancer disease. *Cancer Res.* **71**, 7345–7350 (2011).
49. Robichaud, G. A., Perreault, J.-P. & Ouellette, R. J. Development of an isoform-specific gene suppression system: the study of the human *Pax-5B* transcriptional element. *Nucleic Acids Res.* **36**, 4609–4620 (2008).
50. Kikuchi, H. et al. Paired box gene 5 isoforms A and B have different functions in transcriptional regulation of B cell development-related genes in immature B cells. *Microbiol. Immunol.* **59**, 426–431 (2015).
51. Cresson, C. et al. *PAX5A* and *PAX5B* isoforms are both efficient to drive B cell differentiation. *Oncotarget* **9**, 32841–32854 (2018).
52. Cooper, A. B. et al. A unique function for cyclin D3 in early B cell development. *Nat. Immunol.* **7**, 489–497 (2006).
53. Ma, X. et al. ORMDL 3 and its implication in inflammatory disorders. *Int. J. Rheum. Dis.* **21**, 1154–1162 (2018).
54. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
55. Orrù, V. et al. Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. *Nat. Genet.* **52**, 1036–1045 (2020).
56. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
57. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
58. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
59. Graham, R. R. et al. A common haplotype of interferon regulatory factor 5 (*IRF5*) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat. Genet.* **38**, 550–555 (2006).
60. McLeod, D. S. A., Caturegli, P., Cooper, D. S., Matos, P. G. & Hutfless, S. Variation in rates of autoimmune thyroid disease by race/ethnicity in US military personnel. *JAMA* **311**, 1563–1565 (2014).
61. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

Chi Tian[1,42], Yuntian Zhang[2,42], Yihan Tong[1,42], Kian Hong Kock[3], Donald Yuhui Sim[4], Fei Liu[1], Jiaqi Dong[2], Zhixuan Jing[2], Wenjing Wang[1,2], Junbin Gao[1], Le Min Tan[3], Kyung Yeon Han[5], Yoshihiko Tomofuji[6,7,8], Masahiro Nakano[9,10], Eliora Violain Buyamin[3], Radhika Sonthalia[3], Yoshinari Ando[11,12], Hiroaki Hatano[10], Kyuto Sonehara[6,7,8], Asian Immune Diversity Atlas Network*, Xin Jin[13,14,15,16], Marie Loh[3,17,18], John Chambers[17], Chung-Chau Hon[19], Murim Choi[20], Jong-Eun Park[21], Kazuyoshi Ishigaki[10], Tomohisa Okamura[22], Keishi Fujio[22], Yukinori Okada[6,7,8,23,24], Woong-Yang Park[5], Jay W. Shin[3,11,25], Xavier Roca[4], Shyam Prabhakar[3,17] & Boxiang Liu[1,2,3,26,27,28] ✉

[1]Department of Pharmacy and Pharmaceutical Sciences, Faculty of Science, National University of Singapore, Singapore, Singapore. [2]Department of Biomedical Informatics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [3]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [4]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. [5]Samsung Genome Institute, Samsung Medical Center, Seoul, South Korea. [6]Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama City, Japan. [7]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. [8]Department of Genome Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [9]Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama City, Japan. [10]Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, Yokohama City, Japan. [11]Laboratory for Advanced Genomics Circuit, RIKEN Center for Integrative Medical Sciences, Yokohama City, Japan. [12]Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences, Yokohama City, Japan. [13]BGI Research, Shenzhen, China. [14]The Innovation Centre of Ministry of Education for Development and Diseases, School of Medicine, South China University of Technology, Guangzhou, China. [15]Shanxi Medical University-BGI Collaborative Center for Future Medicine, Shanxi Medical University, Taiyuan, China. [16]Shenzhen Key Laboratory of Transomics Biotechnologies, BGI Research, Shenzhen, China. [17]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [18]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [19]Laboratory for Genome Information Analysis, RIKEN Center for Integrative Medical Sciences, Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima, Japan. [20]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, South Korea. [21]Graduate School of Medical Science and Engineering, KAIST, Daejeon, South Korea. [22]Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [23]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan. [24]Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, Suita, Japan. [25]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [26]Precision Medicine Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [27]NUS Centre for Cancer Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [28]Cardiovascular-Metabolic Disease Translational Research Programme, Yong Loo Lin School of

Medicine, National University of Singapore, Singapore, Singapore. [42]These authors contributed equally: Chi Tian, Yuntian Zhang, Yihan Tong. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: boxiangliu@nus.edu.sg

Asian Immune Diversity Atlas Network

Varodom Charoensawan[29,30,31,32,33,34,35], Chung-Chau Hon[19], Partha P. Majumder[36,37], Ponpan Matangkasombut[31,38], Woong-Yang Park[5], Shyam Prabhakar[3,17], Jay W. Shin[3,11,25], Piero Carninci[12,39], John C. Chambers[17], Marie Loh[3,17,18], Manop Pithukpakorn[33,40], Bhoom Suktitipat[30,33], Kazuhiko Yamamoto[9], Deepa Rajagopalan[3], Nirmala Arul Rayan[3], Shvetha Sankaran[3], Juthamard Chantaraamporn[29,30,31], Ankita Chatterjee[36], Supratim Ghosh[41], Kyung Yeon Han[5], Damita Jevapatarakul[31,38], Sarintip Nguantad[29,30,31], Sumanta Sarkar[41], Narita Thungsatianpun[31,38], Mai Abe[9], Seiko Furukawa[9], Gyo Inoue[9], Keiko Myouzen[9], Jin-Mi Oh[5], Akari Suzuki[9], Yoshinari Ando[11,12], Miki Kojima[12], Tsukasa Kouno[11], Jinyeong Lim[5], Arindam Maitra[41], Le Min Tan[3], Prasanna Nori Venkatesh[3], Murim Choi[20], Jong-Eun Park[21], Eliora Violain Buyamin[3], Kian Hong Kock[3], Quy Xiao Xuan Lin[3], Jonathan Moody[19], Radhika Sonthalia[3], Kazuyoshi Ishigaki[10], Masahiro Nakano[9,10], Yukinori Okada[6,7,8,23,24] & Yoshihiko Tomofuji[6,7,8]

[29]Department of Biochemistry, Faculty of Science, Mahidol University, Bangkok, Thailand. [30]Integrative Computational BioScience (ICBS) Center, Mahidol University, Nakhon Pathom, Thailand. [31]Systems Biology of Diseases (SyBiD) Research Unit, Faculty of Science Mahidol University, Bangkok, Thailand. [32]Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. [33]Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. [34]Siriraj Genomics, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. [35]School of Chemistry, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima, Thailand. [36]John C. Martin Centre for Liver Research and Innovations, Kolkata, India. [37]Indian Statistical Institute, Kolkata, India. [38]Department of Microbiology, Faculty of Science, Mahidol University, Bangkok, Thailand. [39]Genomics Research Center, Fondazione Human Technopole, Milano, Italy. [40]Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. [41]Biotechnology Research and Innovation Council-National Institute of Biomedical Genomics, Kalyani, India.

Methods

Inclusion and ethics statement

Local researchers from the AIDA consortium member countries were involved in study design, study implementation, data ownership, intellectual property and authorship of publications throughout the research process. All participants were approved by local ethics review committees before study enrollment.

All study protocols were approved by the institutional review boards (IRBs) of the institutions that the laboratories are affiliated with (Genome Institute of Singapore: IRB-2020-012 and IRB-2022-051; Nanyang Technological University: IRB-2016-11-030-01, IRB-2016-11-030 and IRB-18IC4698; RIKEN: IRB-H30-9; Samsung Genome Institute: IRB-2019-09-121; Faculty of Medicine Siriraj Hospital, Mahidol University: IRB-725/2563(IRB3); National Institute of Biomedical Genomics: IRB-NIBMG/2022/1/0022) before dataset generation. All donors provided written informed consent for sample and metadata collection and subsequent analyses.

The AIDA cohort

The AIDA Data Freeze v.1 performed 10x 5′ scRNA-seq on 503 donors of Eastern (Singaporean Chinese, Japanese, Korean), Southeastern (Singaporean Malay) and South Asian (Singaporean Indian) descent. After removing donors whose samples failed genotyping or scRNA-seq QC, and related donors (up to third-degree cousins), 484 nonrelated individuals remained. Taking the intersection between genotype and scRNA-seq resulted in 474 individuals for sQTL analysis. The donor characteristics are provided in Supplementary Table 1.

Genotype data processing and QC

Genotyping was performed using the Illumina Infinium Global Screening Array v.3 according to the manufacturer's protocol. We exported genotype data into PLINK data format using the GenomeStudio PLINK Input Report plug-in v.2.1.4 to perform QC and imputation. A total of 477,889 post-QC variants were used for genome-wide imputation with prephasing using the Michigan imputation server[62]. All populations in 1000 Genomes Project high-coverage (hg38) were selected as reference panel. After imputation, only imputed variants with an imputation quality of $R^2 > 0.8$ were retained. Variants with missingness greater than 0.05, MAF < 0.05 or Hardy–Weinberg equation $P < 1 \times 10^{-6}$ were excluded from further analysis. Monoallelic and multiallelic variants in each population were removed, leaving 5,065,361 genetic variants for downstream analysis. We extracted autosomal nonpalindromic variants and merged the AIDA genotypes with the 1000 Genomes ($n = 2,504$) for PC analysis (PCA). We merged the AIDA samples with the 1000 Genomes samples using bcftools[63] v.1.9 and performed PCA using PLINK[64] v.1.9.

scRNA-seq data processing and QC

The single-cell experiments used the 10x Genomics 5′ v.2 RNA-seq method. Each batch included a pooling of 15 Asian donors and one European control sample. The initial data preprocessing and QC steps, such as doublet identification and demultiplexing, were conducted using DRAGEN v.3.8.4, Cell Ranger v.7.0.1 and Freemuxlet (https://github.com/statgen/popscle;v0.1-beta). We used the GENCODE release 32 (GRCh38, Ensembl 98, 5 September 2019) as our gene annotation reference. Cells with fewer than 300 GENCODE release 32 genes detected (NODG < 300) were filtered out. Subsequent analyses involved Seurat v.4.1.1R[65] and RCA v.2.0 (ref. 65) for the initial cell type annotation for the doublet identification workflow. Further QC, both at the library-specific and cell-type-specific levels, was performed to remove low-quality cells. AIDA Data Freeze v.1 included 1,058,909 PBMCs from 503 Asian donors and five European controls profiled in Japan, Singapore and South Korea. Cells expressing heightened platelet marker genes in B cells, plasmacytoid dendritic cells, myeloid cells, innate lymphoid cells, NK cells and T cells were removed during cell-population-specific QC. Data integration was achieved using the Seurat anchor integration reciprocal PCA algorithm. Subsequent subclustering annotation was based on marker genes curated from the literature and an examination of gene expression across clusters in our dataset. In the final annotation, 34 cell types were identified. In these 34 subtypes, we focused on 21 PBMC subtypes, each including more than 3,500 cells. Red blood cells, platelets and clusters with ambiguous identities (for example, expression of marker genes of other cell types) were excluded from this analysis (the details are described in ref. 22).

Quantification of mRNA AS

RNA-seq data were aligned to the human reference genome GRCh38 primary assembly and GENCODE v.32 using STARsolo[66] v.2.7.10a with the options --soloCBmatchWLtype 1MM --soloUMIdedup 1MM_Directional_UMItools. The cell barcode whitelist is available inside the Cell Ranger installation (cellranger-x.y.z/lib/python/cellranger/barcodes/737K-august-2016.txt). A two-pass mode was used to enable new splice junction discovery. The --waspOutputMode option was used to reduce allelic mapping bias. For each sample, the corresponding post-QC VCF file was used for WASP filtering. Deduplication of reads was performed based on cell barcodes and unique molecular identifier tags from the BAM files using the MarkDuplicates function from Picard. Only uniquely mapped reads in proper pairs that passed the WASP filter were retained for downstream analysis. To calculate the percentage of exonic bases covered by sequencing reads, BEDTools[67] v.2.27.1 merge was used to merge overlapping exon regions. To vizualize read distribution, deepTools[68] v.3.5.1 was used.

Reads from the same cell type of each donor were extracted using custom scripts to make pseudobulk BAM files; reads from the same individuals were pooled using SAMtools[69] v.1.16.1 merge. We used RegTools[70] v.0.0.1 to extract intron junctions and LeafCutter[24] v.0.2.9 to quantify intron use levels. The prepare_phenotype_table.py script from LeafCutter was used to generate phenotype files in sQTL mapping. Introns with zero read counts in more than 40% of the samples or with insufficient variation (s.d. < 0.005) were removed. Replication of intron junctions obtained using RegTools is detailed in the Supplementary Methods. SpliZ was used to quantify the gene-level splicing of each cell. This pipeline generated a scalar score for each gene–cell pair. A larger negative score indicates that the introns for the gene in a given cell are shorter than average, while a larger positive score indicates the opposite. The GENCODE v.32 annotation file was used as the reference annotation file. The --lower_bound parameter was set to 1 to include all junction reads in the calculation of SpliZ.

Differential splicing analysis across sex and ancestry

Sex-biased differential splicing was performed with LeafCutter v.0.2.7 using default parameters, with age, sequencing center and five genotype PCs as covariates. We calculated FDRs with the Benjamin–Hochberg method and used an FDR < 0.05 as a significance cutoff. The R package gggenes v.0.5.1 was used to plot the gene model and visualize the read coverage on *FLNA*.

Ancestry-biased differential splicing was performed with LeafCutter using default parameters, with sex and age as covariates. To minimize the batch effects caused by different sequencing centers, we focused on the Singapore cohort ($n = 75$ for Singaporean Chinese; $n = 54$ for Singaporean Malay; $n = 60$ for Singaporean Indian) for this analysis. We determined the influence of ancestry on AS by performing one-versus-one differential splicing analysis between the Singaporean Chinese, Malay and Indian populations. We calculated FDRs with the Benjamin–Hochberg method and used an FDR < 0.05 as a significance cutoff. Sashimi plots used to show the junction use difference between ancestries were generated using ggsashimi[34] v.1.1.5.

cis-sQTL mapping

cis-sQTL mapping was performed using QTLtools[71] v.1.2, using intron excision ratios and a *cis*-window of 1 Mb on both sides of the junction. Each cell type needed a minimum of ten cells per individual. Eight PCs derived from splicing ratios, five genotype PCs, sex and age information were used as covariates in the linear model. The number of phenotypes and genotype PCs were chosen to maximize sQTL discovery. Grouped permutations (--grp option) were used to jointly compute an empirical *P* value over all intron clusters of a gene. QTLtools was run using the permutation mode (1,000 permutations); beta-approximated permutation *P* values were adjusted for multiple testing correlation using the qvalue package[72] v.2.30.0. The significance threshold was set at FDR < 0.05. Conditional sQTL analysis was done by forward stepwise regression followed by a backward selection step. The gene-level significance threshold was set to the maximum beta-adjusted *P* value over all sGenes in a given cell type. Scanning for *cis*-sQTLs using QTLtools was performed to correct for all previously discovered variants and all covariates. If the beta-adjusted *P* value for the lead variant was insignificant at the gene-level threshold, the forward stage was complete and the procedure moved on to the backward stage. If this *P* value was significant, the lead variant was added to the list of discovered *cis*-QTLs as an independent signal and the forward step moved on to the next iteration. The backward stage consisted of testing each variant separately, controlling for all other discovered variants. Enrichment of lead sQTL variants in the annotations was calculated using qtlBHM v.1.0 (https://github.com/rajanil/qtlBHM).

To evaluate the replication of AIDA *cis*-sQTLs in the external datasets, we used five data sources: (1) BLUEPRINT[13], which had CD14$^+$ monocytes and CD4$^+$ T cells; (2) DICE[12], which had CD14$^+$ monocytes, CD16$^+$ monocytes, CD16$^+$ NK cells, CD4$^+$ naive T cells, CD8$^+$ naive T cells and naive B cells; (3) GTEx whole-blood; (4) GTEx LCL; and (5) ImmuNexUT[35]. During replication, we selected *cis*-sQTLs from the corresponding cell types and queried their *P* value in the aforementioned datasets. The replication was quantified using the $\pi 1$ statistic as implemented in the qvalue package.

We used mashr[36] v.0.2.79 to estimate sQTL sharing across cell types. We used as input the nominal *P* values from sQTLs (sIntron-sVariant) for each cell type. We defined the *cis*-sQTL with the most significant *P* value in each gene as the top *cis*-sQTL for this gene and combined all the top *cis*-sQTLs as the strong tests. Random tests were chosen by randomly selecting 20,000 *cis*-sQTLs from sQTL nominal *P* values. Strong tests were used to learn data-driven covariance matrices; random tests were used to learn correlation structures. Finally, we used the aforementioned fitted model to compute the posterior means and LFSR for the strong tests.

Sex-biased *cis*-sQTL

Sex-biased sQTLs were defined as those *cis*-sQTLs with a significant genotype-by-sex (G × S) interaction effect. For each significant intron–variant pair identified in the independent sQTL analysis, a linear regression model was fitted for each cell type to test for genotype-by-sex (G × S) interaction while adjusting for known and unknown confounders. Male-specific or female-specific means that the allelic effect of the sQTL was significant in the discovery sex ($P < 5 \times 10^{-4}$) and nonsignificant in the other sex ($P > 0.05$). Differential effect means that the allelic effects of the sQTL were significant in both sexes, but their effect sizes differed.

Dynamic intron use and dynamic sQTL mapping in B cell development

We inferred the pseudotime trajectory from naive to memory B cells using slingshot[73] v.2.10.0 and divided the trajectory into six discrete quantiles. To identify dynamic splicing junctions during B cell development, we used analysis of variance to identify intron junctions with a significant change (FDR < 0.05) through six quantiles. To identify dynamic sQTLs, we used linear mixed models with the R package lme4 (ref. 74) v.1.1-35.3. The model assessed the interaction between genotype and pseudotime quantiles, incorporating random effects to account for individual identity and fixed effects for genotype, pseudotime, sex, age and ancestry PCs, and phenotype PCs from each quantile (Supplementary Methods). We tested both linear (genotype × pseudotime) and quadratic interactions (genotype × pseudotime^2). An FDR of 0.05 was used as a threshold to call dynamic sQTLs.

trans-sQTL mapping

We followed a previously established permutation-based pipeline[33] to map regulatory variants at least 5 Mb away from their *trans*-sGenes and applied stringent mappability filters to minimize false positive findings[74]. Only variants with a 50-mer mappability greater than 0.9 were retained as test variants. To exclude genes susceptible to mapping artifacts, we excluded genes with a mappability of less than 0.8 and any variant–gene pair where the *trans*-sGene cross-mapped with any gene within the *cis*-window of the variant. Mappability was calculated with a *k*-mer length of 75 for exons and 36 for untranslated regions[75]. We performed *trans*-sQTL mapping using QTLtools[34] to test for associations between the test genes and all variants beyond 5 Mb of the same chromosome and interchromosomal associations. For gene-level FDR control, we obtained beta-approximated empirical *P* values based on 50,000 permutations. To correct for multiple intron phenotypes per gene, we used the most significant empirical *P* values of variant–phenotype pairs to compute the distribution for the *P* value across *k* phenotypes using $1 - (1 - F(x))^k$, where F(x) is the empirical cumulative distribution function. Benjamin–Hochberg correction was used across all genes and an FDR of 0.01 was used as a threshold to define *trans*-sQTLs. The circos plot was generated using circlize[76] v.0.4.15.

Complex trait associations

We obtained publicly available GWAS summary statistics of 20 phenotypes covering a broad range of categories including blood traits (hemoglobin, platelet count, white blood cell count, monocyte count, basophil count, hematocrit, mean corpuscular volume, lymphocyte, neutrophil, mean corpuscular hemoglobin concentration, mean corpuscular hemoglobin, red blood cell, eosinophil), anthropometric traits (height, BMI) and immune-related traits (AD, asthma, GD, RA, SLE). We estimated sQTL enrichment in the 20 traits using S-LDSC v.1.0.1 with default parameters from the partitioned heritability pipeline (https://github.com/bulik/ldsc/wiki/Partitioned-Heritability) in a window of ±100 kb around each sGene with at least one independent *cis*-sQTL. To assess colocalization between GWAS loci and *cis*-sQTLs, we tested colocalization between GWAS and sQTL signals for genes with at least ten variants using the coloc.abf function of the COLOC R package[57] v.5.2.3 (with default prior, using beta coefficients from the GWAS and sQTL analysis). We used the MAFs of our sQTL study to estimate the standard deviation of the quantitative trait. H4 > 0.75 was set as the threshold for colocalization. The colocalization results were visualized using LocusCompare[77] (http://locuscompare.com/) v.1.0.0.

Experimental validation

To validate the effect of rs74416240 on *TCHP* exon 4 splicing, we built two minigene constructs that were identical except at rs74416240, with a G allele in the reference construct and an A allele in the alternative construct (the complete sequences can be found in the Supplementary Information). The minigenes consisted of three exon and two intron fragments cloned into an expression plasmid, whereby the middle exon may be subject to splicing changes by the variant. We inserted the sequences into the *MCAD* minigene we built previously[78], which had two outer constitutive exons with their intronic portions and an intronic multiple cloning site to clone the fragments. We cloned both reference and alternative alleles using PCR amplification of genomic DNA fragments. We transfected minigene constructs into the K562 cell line. Forty-eight hours after transfection, we examined the splicing patterns using RNA extraction, PCR with reverse transcription and agarose gels. All oligonucleotides were synthesized by Integrated DNA Technologies (Supplementary Data 1).

Statistics and reproducibility

We included all donors with both quality-controlled genotype and scRNA-seq data. For each cell type, donors with fewer than ten cells were excluded from the analysis. This filtering process resulted in the final sample sizes for each cell type, as detailed in Supplementary Table 1. Sample sizes ranged from 114 to 459 (>400 for most cell types). These numbers are comparable to those reported in previous studies. Donors whose samples failed genotyping or scRNA-seq QC, as well as related donors (up to third-degree cousins), were removed. Genetic variants with a missingness >0.05, MAF < 0.05 or Hardy–Weinberg equation $P < 0.000001$ were excluded from the analysis. The experiments were not randomized. The investigators were not blinded to allocation during the experiments and outcome assessment. All statistical tests described in the article were two-sided. Box plots generated using the R ggplot2 function show the median and IQR; the whiskers are 1.5 times the IQR. Any data points outside the whiskers were considered outliers and were plotted as individual points.

Reporting summary

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

Data availability

The AIDA Data Freeze v.1 gene-cell matrix (1,058,909 cells from 503 Japanese, Singaporean Chinese, Singaporean Malay, Singaporean Indian and South Korea Asian donors and five distinct Lonza commercial controls), and donor age, sex and self-reported ethnicity metadata are available via the Chan Zuckerberg CELLxGENE data portal at https://cellxgene.cziscience.com/collections/ced320a1-29f3-47c1-a735-513c7084d508. The Japanese genotypes and scRNA-seq FASTQ files are available via Gene Expression Omnibus (accession no. GSE280974). The Korean genotypes and scRNA-seq FASTQ files are available via Gene Expression Omnibus (accession no. GSE281106). The Singaporean genotypes and scRNA-seq FASTQ files require a data access application to the HELIOS Data Access Committee (helios_science@ntu.edu.sg). All sQTL summary statistics are available via Zenodo at https://doi.org/10.5281/zenodo.8343364 (ref. 79). The nucleotide sequences of the GRCh38 primary genome assembly and gene annotation are available at https://www.gencodegenes.org/human/release_32.html. The OneK1K scRNA-seq data are available via the Gene Expression Omnibus (accession no. GSE196830). The junctions assembled using Snaptron can be downloaded from the Snaptron website (https://snaptron.cs.jhu.edu/data/srav2/junctions.bgz). The GTEx sQTL summary statistics can be downloaded from the GTEx portal (https://gtexportal.org/home/downloads/adult-gtex/qtl). Full summary statistics of the BLUEPRINT and DICE sQTL data[14] were acquired from Y. I. Li (yangili1@uchicago.edu). The T cell proportion GWAS data[55] are available at http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90001001-GCST90002000/GCST90001538/. The sources of the GWAS summary statistics[40–44] used for the colocalization analysis are outlined in Supplementary Table 6. Source data are provided with this paper.

Code availability

The source code used in this study is available via GitHub at https://github.com/boxiangliulab/AIDA_phase1_sQTL (ref. 78).

References

62. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
63. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
64. Schmidt, F. et al. RCA2: a scalable supervised clustering algorithm that reduces batch effects in scRNA-seq data. *Nucleic Acids Res.* **49**, 8505–8519 (2021).
65. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.05.442755 (2021).
66. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
67. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
68. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
69. Cotto, K. C. et al. Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat. Commun.* **14**, 1589 (2023).
70. Garrido-Martín, D., Palumbo, E., Guigó, R. & Breschi, A. ggsashimi: sashimi plot revised for browser- and annotation-independent splicing visualization. *PLoS Comput. Biol.* **14**, e1006360 (2018).
71. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
72. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
73. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
74. Saha, A. & Battle, A. False positives in *trans*-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res.* **7**, 1860 (2019).
75. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
76. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).
77. Tan, J. et al. Noncanonical registers and base pairs in human 5′ splice-site selection. *Nucleic Acids Res.* **44**, 3908–3921 (2016).
78. Zhang, Y., Tian, C., Tong, Y., Liu, B. & Liu, F. boxiangliulab/AIDA_phase1_sQTL: AIDA Phase1 sQTL FreezeV1. *Zenodo* https://zenodo.org/records/13170522 (2024).
79. Zhang, Y., Tian, C., Tong, Y. & Liu, B. AIDA PBMC sQTL. *Zenodo* https://doi.org/10.5281/zenodo.8343364 (2023).

Acknowledgements

We thank J. Salzman and J. E. Olivieri for technical assistance on SpliZ, and Z. Mu and Y. I. Li for technical assistance with sQTL replication. B.L. is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (FY2023; 23-0434-A0001; 22-5800-A0001) and Tier 2 (MOE-T2EP30123-0015), the Precision Medicine Translational Research Programme Core Funding (NUHSRO/2020/080/MSC/04/PM), NUS ODPRT Seed Funding and NUS YLLSoM Seed Funding. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg) and was partially supported by the NUS IT Cloud Credits for Research Programme. X.R. acknowledges support from Singapore's National Research Foundation (grant no. NRF2019-NRF-ISF003-3104). Y. Tong and W.W. are supported by an MGI IRP scholarship. Y.O. was supported by JSPS KAKENHI (22H00476); AMED (JP22ek0410075, JP23km0405211, JP23km0405217, JP23ek0109594, JP23ek0410113, JP223fa627002, JP223fa627010, JP233fa627011, JP23zf0127008); JST Moonshot R&D (JPMJMS2021, JPMJMS2024); the Takeda Science

Foundation; the Bioinformatics Initiative of Osaka University Graduate School of Medicine; the Institute for Open and Transdisciplinary Research Initiatives; the Center for Infectious Disease Education and Research; and the Center for Advanced Modality and DDS, Osaka University. We thank all donors and participants in the studies constituting AIDA. The Singapore donor samples were obtained through the Health for Life in Singapore (HELIOS) study (Lee Kong Chian School of Medicine, Nanyang Technological University; National Healthcare Group, Singapore; Imperial College London). We thank participants of the HELIOS study and the HELIOS operation team for recruitment, organization and data and sample collection. This study (NTU IRB: 2016-11-030) is supported by the National Medical Research Council of the Singapore Ministry of Health (MOH) under its OF-LCG funding scheme (MOH-000271-00) and intramural funding from Nanyang Technological University, the Lee Kong Chian School of Medicine and the National Healthcare Group. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This project was made possible in part by grant no. CZF2019-002446 to S.P., W.-Y.P., J.W.S. and J.C. from the Chan Zuckerberg Foundation, and grant nos. 2020-224570 (to S.P., V.C., P.M. and P.P.M.) and 2021-240178 (to S.P., W.-Y.P., J.W.S., J.C., V.C., P.M. and P.P.M.) from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. This project was also supported by the Thailand Program Management Unit for National Competitiveness Enhancement (PMU-C; C10F650132) to V.C., P.M., M.P. and B.S. and Mahidol University's Basic Research Fund (fiscal year 2021; BRF1-017/2564) to V.C. and B.S. We thank J. Zamanian, J. Chien and J. Hilton from the Human Cell Atlas Lattice team (Stanford University) for their help with and work on data deposits and coordination for community access. This publication is part of the Human Cell Atlas (www.humancellatlas.org/publications/).

Author contributions

C.T. performed the scRNA-seq data processing and QC, quantified the splicing, and carried out the differential splicing analysis and *trans*-sQTL mapping. Y.Z. quantified the splicing, carried out the *cis*-sQTL calling, identified cell-type-specific and sex-biased and ancestry-biased sQTLs, and analyzed the colocalization. Y.Tong performed QC, processed the genotype data and analyzed dynamic intron use and dynamic sQTL. K.H.K., E.V.B. and R.S. performed the scRNA-seq QC, and clustering and cell type annotation under the supervision of S.P. D.Y.S. performed the minigene experiments and the in silico analysis under the supervision of X.R. F.L. performed the sQTL fine-mapping. J.D. conducted the splice junction replication using the PacBio MAS-seq data. Z.J. performed the *cis*-eQTL and *trans*-sQTL colocalization. W.W. processed the genotyping data. J.G. assisted with pseudobulk data processing and the SpliZ pipeline. L.M.T., K.Y.H. and Y.A. performed the scRNA-seq experiments and extracted the genomic DNA for genotyping. Y.Tomofuji, K.S. and Y.O. analyzed the genotype data, including whole-genome genotype imputation and the collection and curation of the GWAS summary statistics. M.N. and K.I. conducted the collection and curation of the GWAS summary statistics. H.H., T.O. and K.F. conducted the sQTL analysis in ImmuNexUT. X.J. cosupervised Y.Tong, M.L. and J.C. (cohort leads) and sample collection. C.-C.H., W.-Y.P., J.W.S. and S.P. led the generation of the single-cell experimental dataset. M.C. and J.-E.P. were responsible for the generation of the single-cell experimental data and data analysis. The PacBio MAS-seq data were provided by J.W.S. B.L. designed the study, performed the analyses, interpreted the results, wrote the paper and obtained the funding. All authors critically revised the paper.

Competing interests

X.J. is an employee of BGI Research. Y. Tong is undertaking a PhD scholarship partially supported by BGI Research. The other authors declare no competing interests.

Additional information

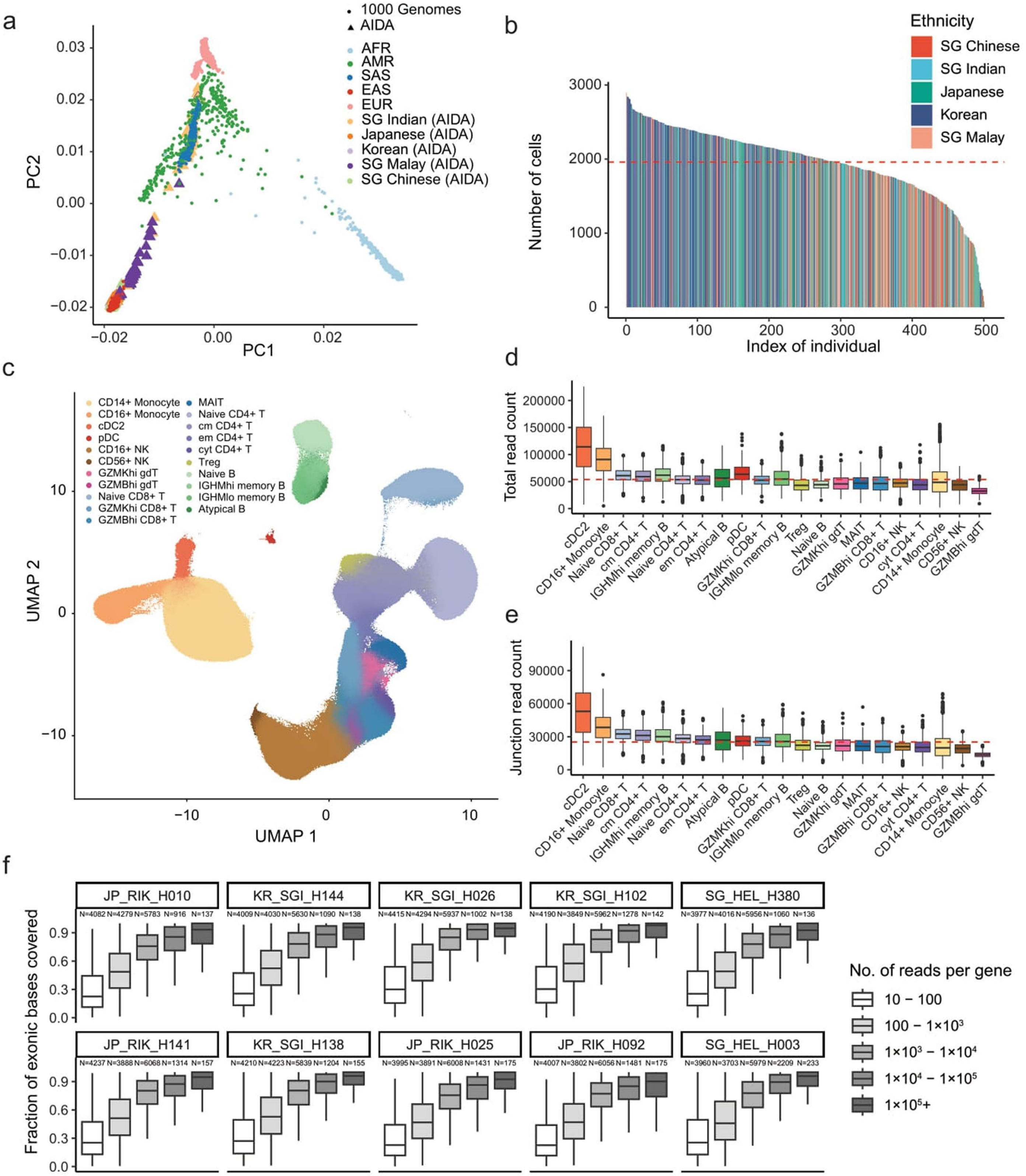
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-02019-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-02019-8.

Correspondence and requests for materials should be addressed to Boxiang Liu.

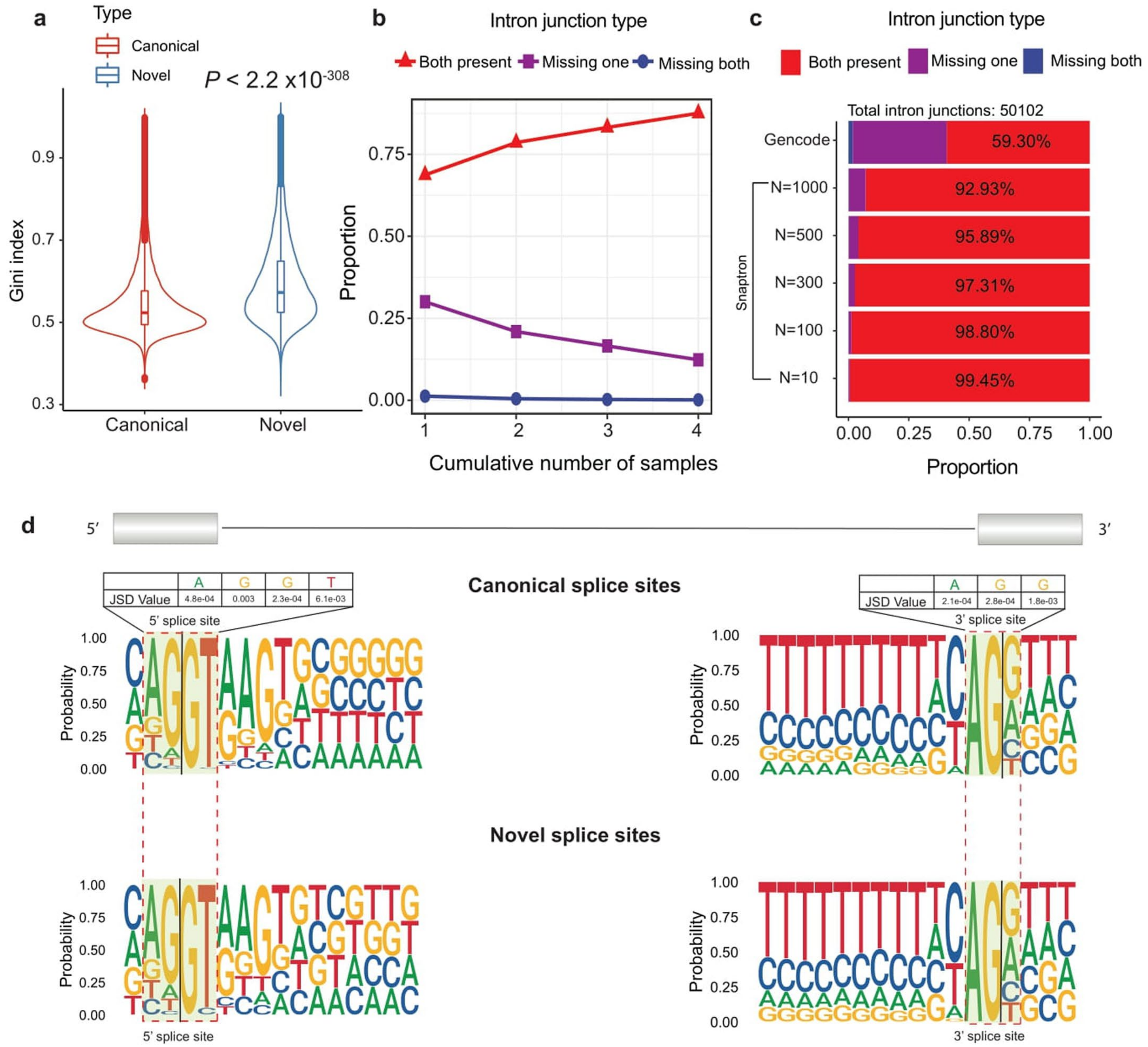
Peer review information *Nature Genetics* thanks Wei Li, Pandurangan Vijayanand and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

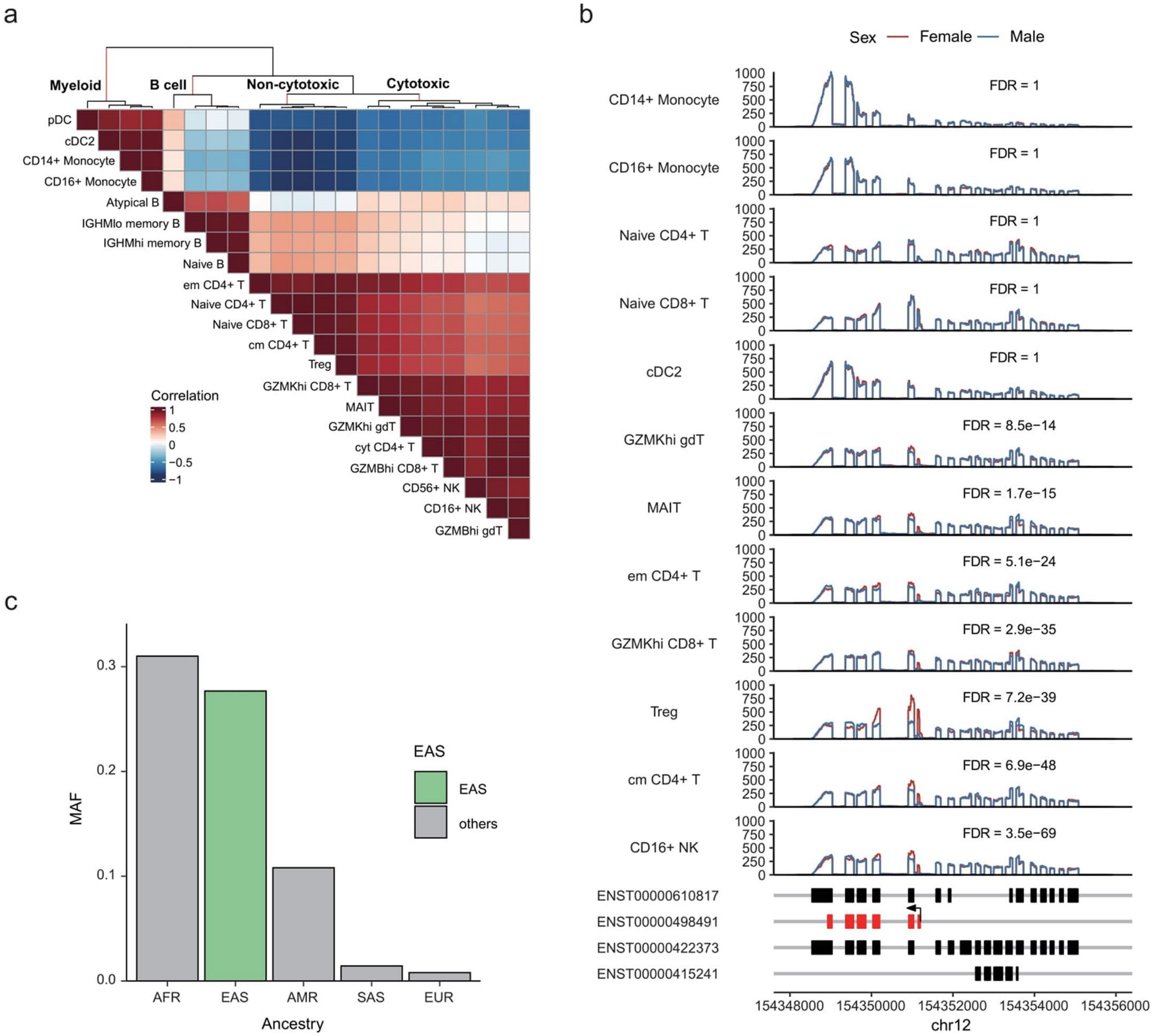


Extended Data Fig. 1 | See next page for caption.

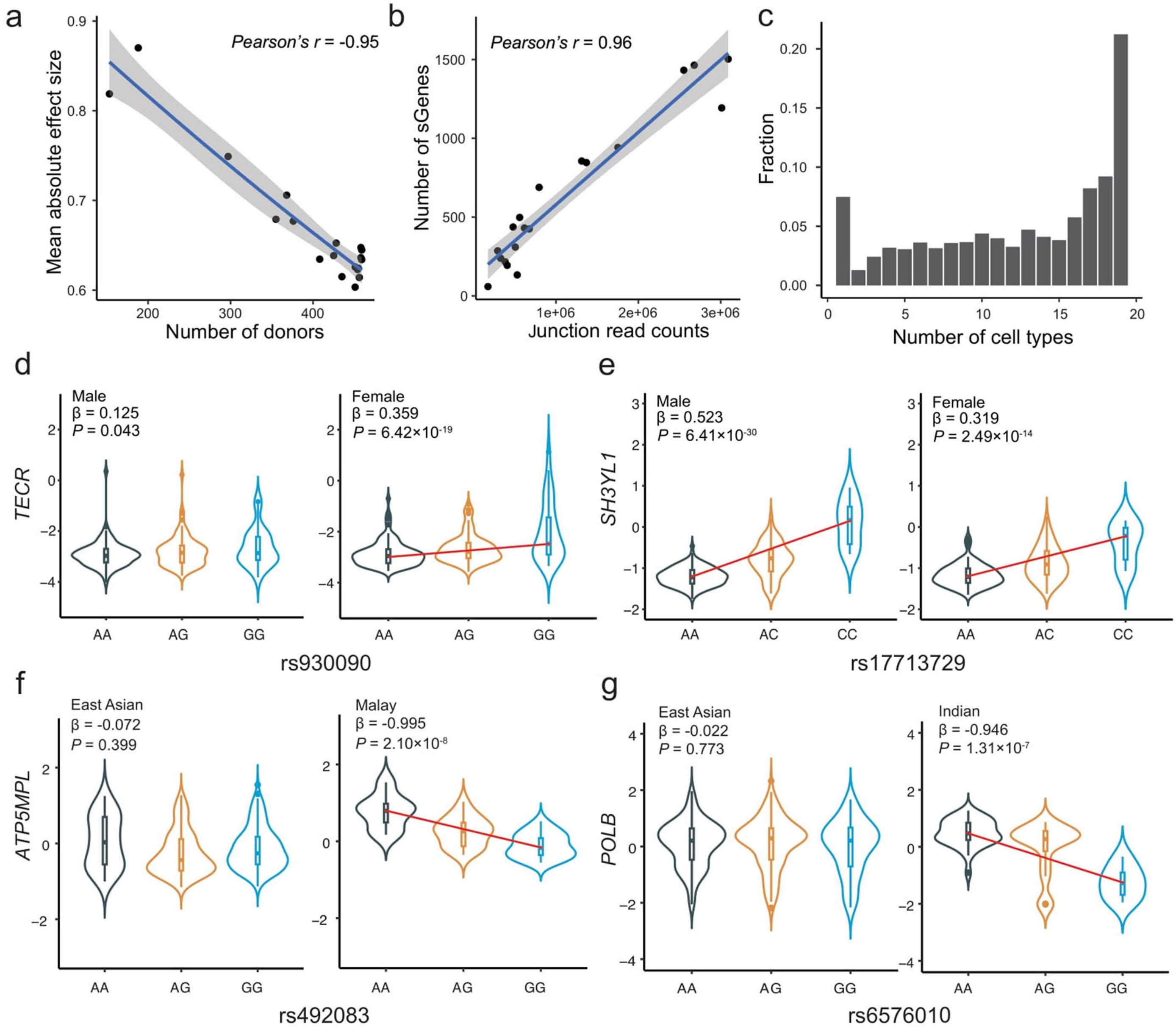
Extended Data Fig. 1 | Overview of the AIDA dataset. (a) PC1 and PC2 of AIDA and 1000 Genomes individuals. East Asian individuals from AIDA (Singaporean Chinese, Japanese, Korean) overlapped with the 1000 Genomes EAS individuals. South Asian individuals from AIDA (Singaporean Indian) overlapped with the 1000 Genomes SAS individuals. Southeast (Singaporean Malay) individuals form a continuum between EAS and SAS individuals from 1000 Genomes. **(b)** The number of single cells across ancestry groups averaged 1,959 cells per donor. The red line shows the mean across all individuals. **(c)** UMAP of 21 PBMC subtypes in AIDA Data Freeze v1, colored by cell types. **(d)** The total number of reads per cell, grouped by cell types. The cell number (N) in (d) and (e): cDC2 (N = 197), CD16+ Monocyte (N = 508), Naive CD8 + T (N = 699), cm CD4 + T (N = 1026), IGHMhi memory B (N = 263), Naive CD4 + T (N = 1976), em CD4 + T (N = 333), atypical B (N = 143), pDC (N = 210), GZMKhi CD8 + T (N = 343), IGHMlo memory B (N = 423), Treg (N = 314), Naive B (N = 513), GZMKhi gdT (N = 199), MAIT (N = 426), GZMBhi CD8 + T (N = 809), CD16 + NK (N = 1244), cyt CD4 + T (N = 638), CD14+ Monocyte (N = 3145), CD56 + NK (N = 157), GZMBhi gdT (N = 437). The red line shows the mean across all cell types. The box plots show median and IQR, and whiskers are 1.5-fold IQR. **(e)** The total number of splice junction reads per cell, grouped by cell types. The red line shows the mean across all cell types. The box plots show median and IQR, and whiskers are 1.5-fold IQR. **(f)** We ranked and divided all donor libraries into ten quantiles according to library size and randomly selected one individual from each quantile. These donors are labeled as Q1-Q10, and the number of genes (N) for each bin and each donor is shown above each box plot. The box plots show median and IQR, and whiskers are 1.5-fold IQR. We observed base coverage across genes increased with read count for all ten quantiles. Fraction of base coverage = covered bases / all bases.



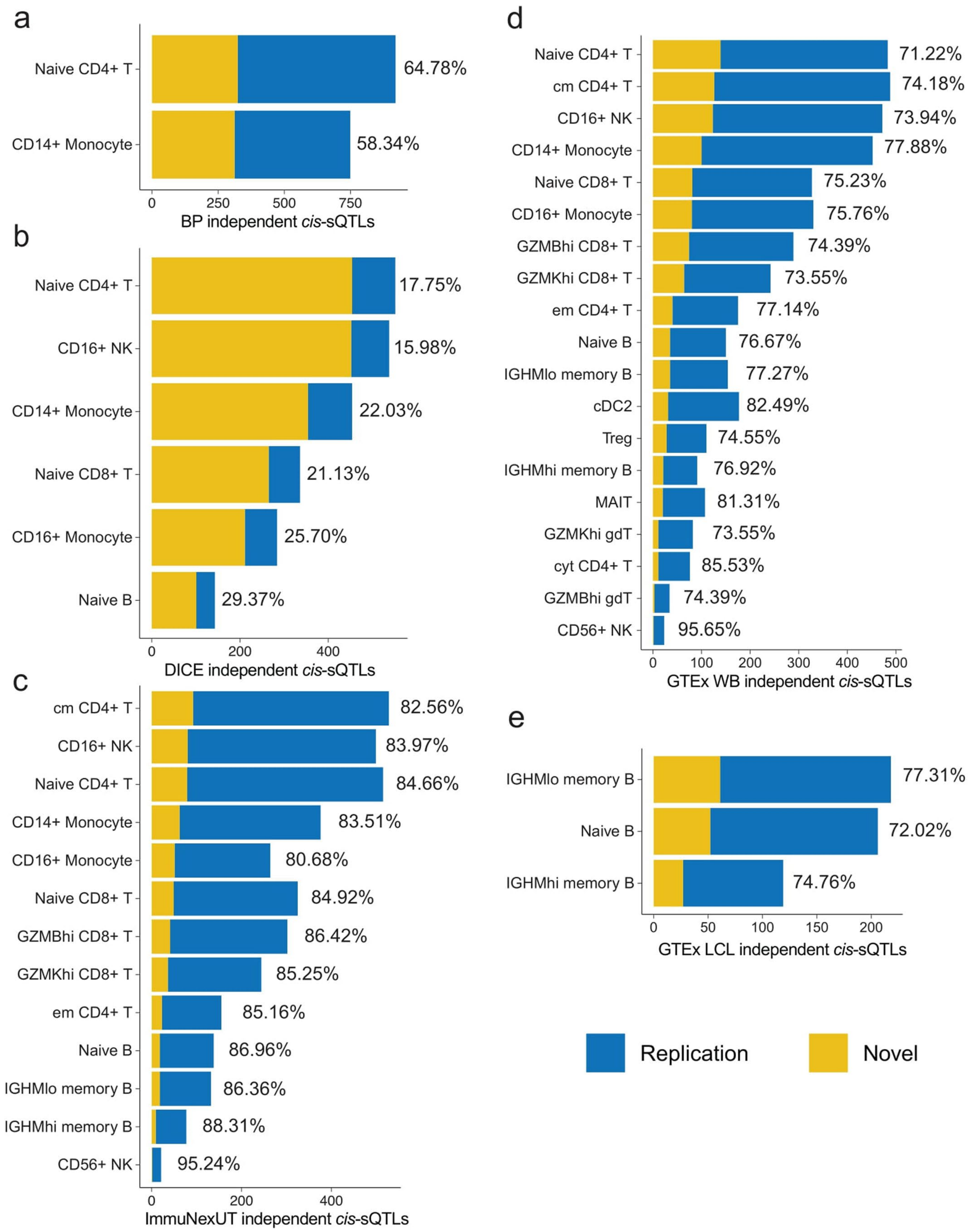
Extended Data Fig. 2 | Quality control of splice junctions. (a) Canonical introns had a significantly lower Gini index than novel introns, indicating that the expression levels of canonical introns were more homogeneous across cell types. *P* value was calculated using t-test (two-sided, $N_{novel} = 53{,}653$, $N_{canonical} = 59{,}400$). The boxes show median and IQR, and whiskers are 1.5-fold IQR. **(b)** Replication of LeafCutter junction discoveries in PacBio MAS-seq long-read dataset. The proportion of replicated junctions increased with the number of PacBio MAS-seq libraries. **(c)** Replication of LeafCutter junction discoveries in GENCODE and Snaptron. The number of replicated introns increased as we relaxed the threshold for Snaptron. **(d)** Position-weight matrices for canonical splice sites and novel splice sites. Both canonical and novel splice sites were highly enriched for canonical splice site motifs. JSD value refers to the Jensen-Shannon divergence value: positive JSD values imply that the given base is more prevalent in canonical splice sites' Position Probability Matrix (PPM) compared to novel splice sites' PPM. Canonical and novel splice sites were assigned based on whether they appeared in GENCODE.



Extended Data Fig. 3 | Context-dependent differentially spliced genes. **(a)** Hierarchical clustering of pseudobulk quantification of alternative splicing. Hierarchical clustering revealed four distinct clusters: myeloid cells, B cells, non-cytotoxic T cells, cytotoxic T / NK cells. The heatmap shows Spearman's rank correlation coefficient. **(b)** Cell-type-specific differential splicing analysis identified female-biased expression of the isoform ENST00000498491 (highlighted in red) in GZMKhi γδ T, MAIT, GZMKhi CD8$^+$ T, Treg, CD4$^+$ (em and cm), and CD16$^+$ NK cells. **(c)** Minor allele frequency (MAF) of rs11064437 in 1000 Genome populations. MAF of rs11064437 is higher in African and East Asians than in other populations.

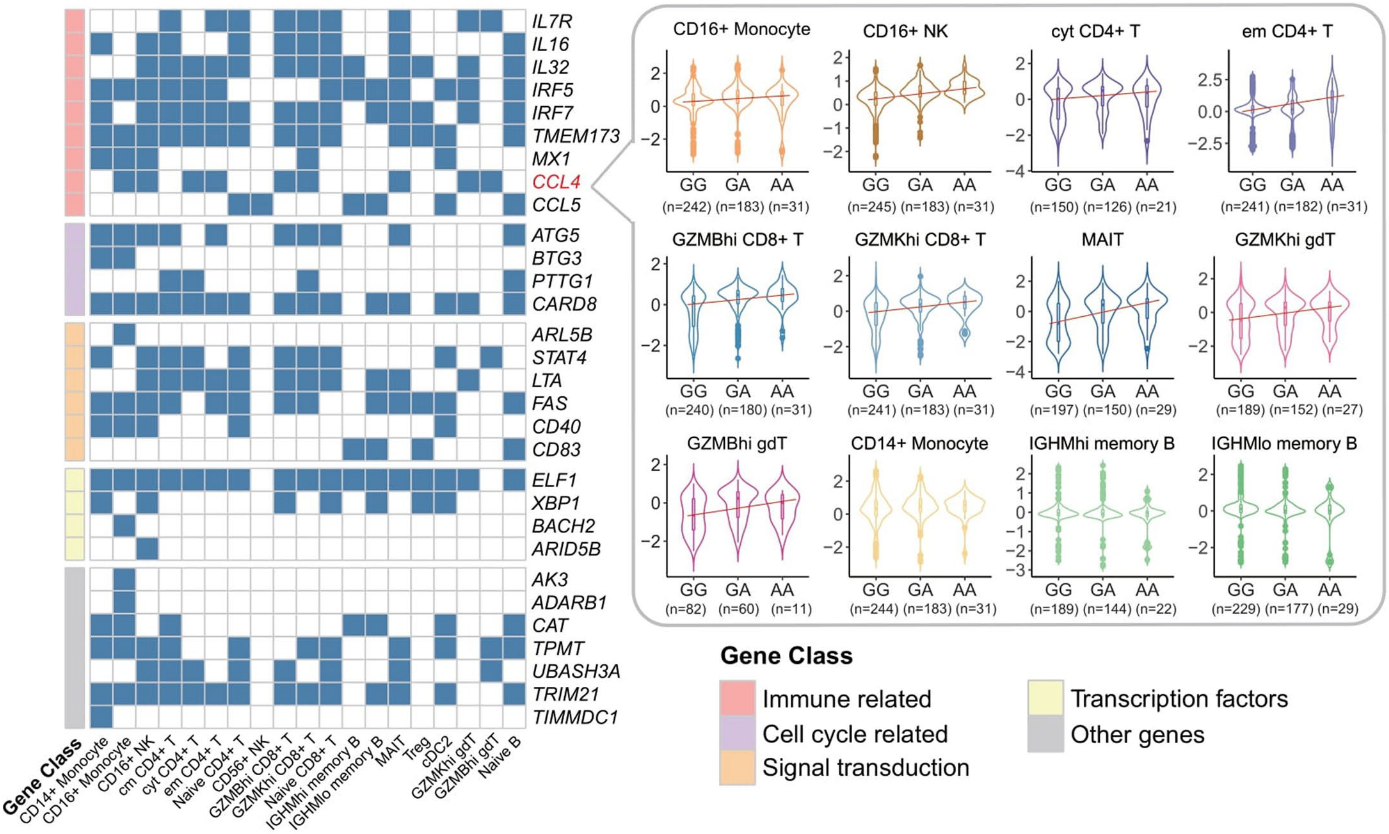


Extended Data Fig. 4 | sQTL power, sharing, and sex-biases. (a) The inverse relationship between the mean absolute effect size of *cis*-sQTLs (y-axis) and the number of donors (x-axis) across 19 cell types (*Pearson's r* = -0.95). Each black dot represents one cell type. The dark blue line represents the fitted linear regression model, and the grey shadow represents the 95% confidence interval in the linear regression. **(b)** The positive relationship between the number of sGenes and the total junction read counts across 19 cell types (*Pearson's r* = 0.96). Each black point represents one cell type. The shaded area represents 95% confidence interval. **(c)** Fractions of cell-type-specific sQTLs detected by mashr using a threshold of LFSR < 0.05 shared by various numbers of cell types. LFSR = local false sign rate. **(d)** An example of single-sex sQTLs (rs930090 modulated *TECR* intron chr19:14529711-14562525; N = 459). The allelic effect in CD16$^+$ NK was only significant in females but not males. **(e)** An example of sex-differential sQTLs (rs17713729 modulated *SH3YL1* intron chr2: 253115-264782; N = 459). The allelic effect in cm CD4$^+$ T was significant in both males and females but larger in males than in females. **(f)** An example of Malay-specific sQTLs (rs492083 modulated *ATP5MPL* intron chr14: 103914633-103915066; N = 456). The allelic effect in CD16+ Monocyte was significant in Malay but not significant in East Asian. **(g)** An example of Indian-specific sQTLs (rs6576010 modulated *POLB* intron chr8: 42338685-42344953; N = 458). The allelic effect in Naive CD4$^+$ T was significant in Indian but not significant in East Asian. Note: The box plots show median and interquartile range (IQR), and whiskers are 1.5-fold IQR in (d), (e), (f) and (g). Unadjusted two-sided *P* value was calculated by QTLtools in (d), (e), (f) and (g). Red lines in (d), (e), (f) and (g) indicate significant linear relationship between intron usage and genotype.

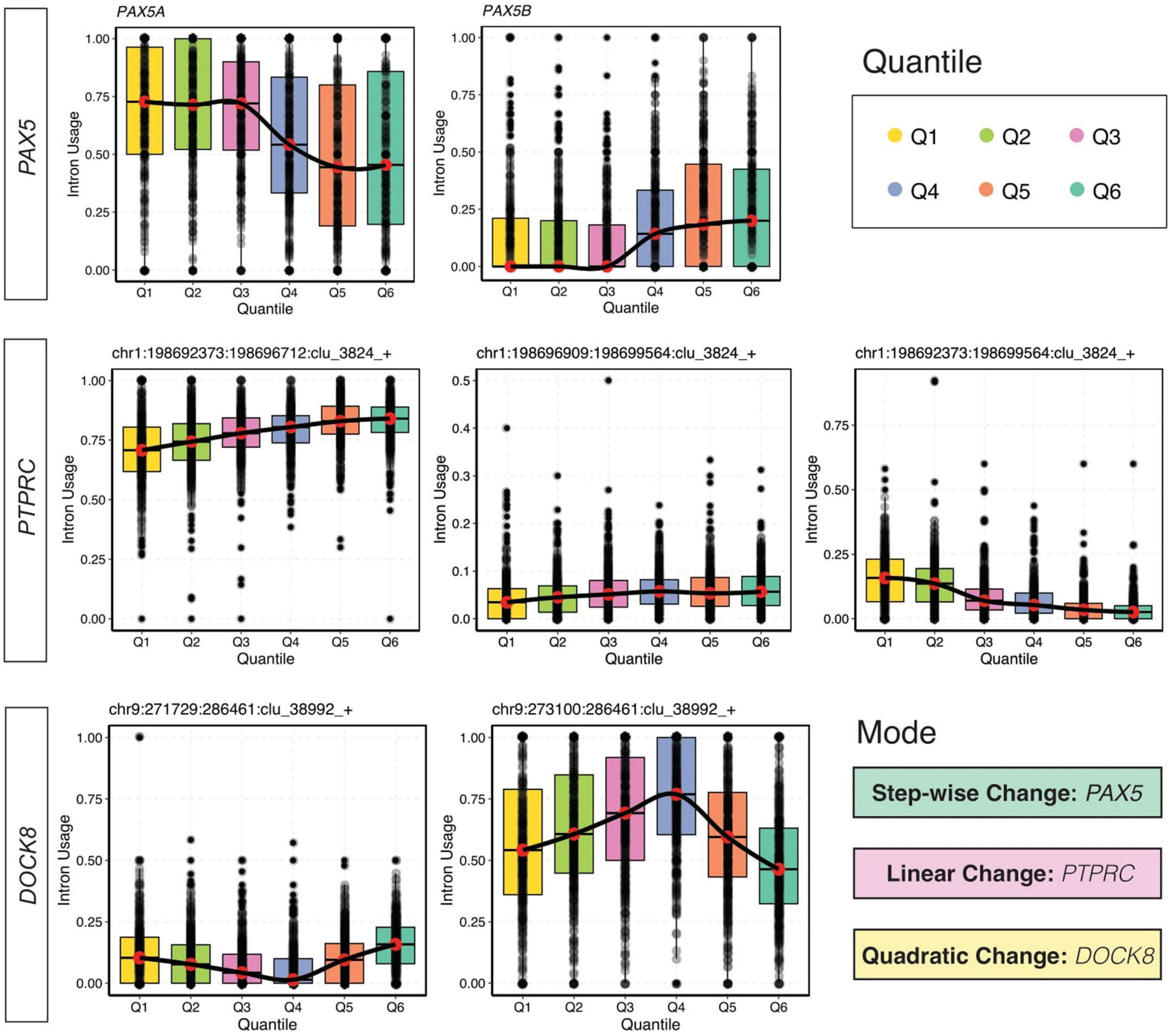


Extended Data Fig. 5 | See next page for caption.

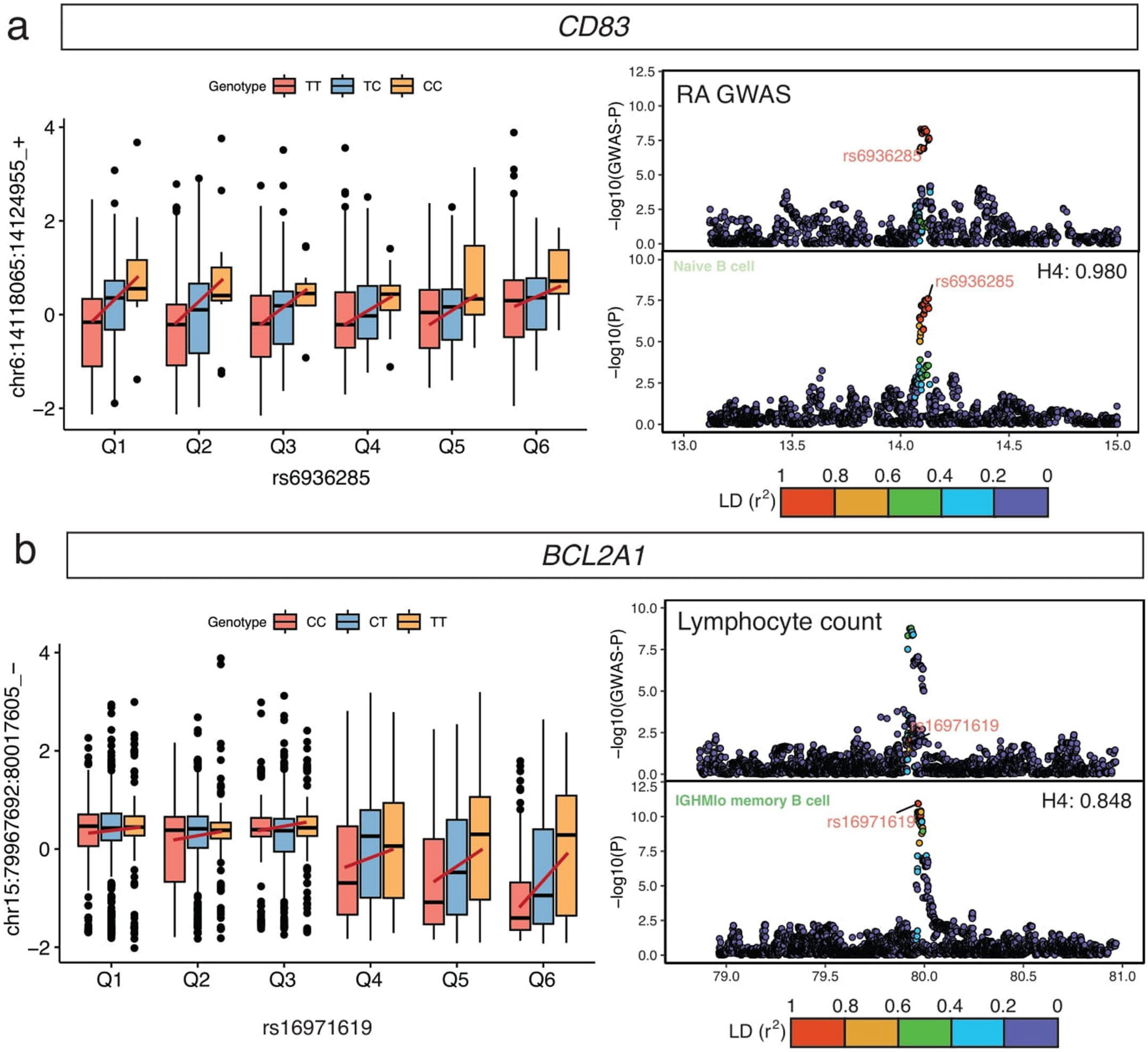
Extended Data Fig. 5 | sQTL replication. The results of AIDA *cis*-sQTLs were replicated in BLUEPRINT (**a**), DICE (**b**), ImmuNexUT (**c**), GTEx whole blood (**d**), and GTEx lymphoblastoid cell lines (**e**). The proportions of replicated sQTLs were used to quantify the replication of independent *cis*-sQTLs in BLUEPRINT (BP), DICE, GTEx and ImmuNexUT for all matched cell types. Replicated sQTLs mean the AIDA independent *cis*-sQTLs with summary statistics available in BP, DICE, and GTEx and are significant with FDR < 0.05. Each bar plot represents the replicated sQTLs' proportions in all the *cis*-sQTLs which have summary statistics in corresponding databases.



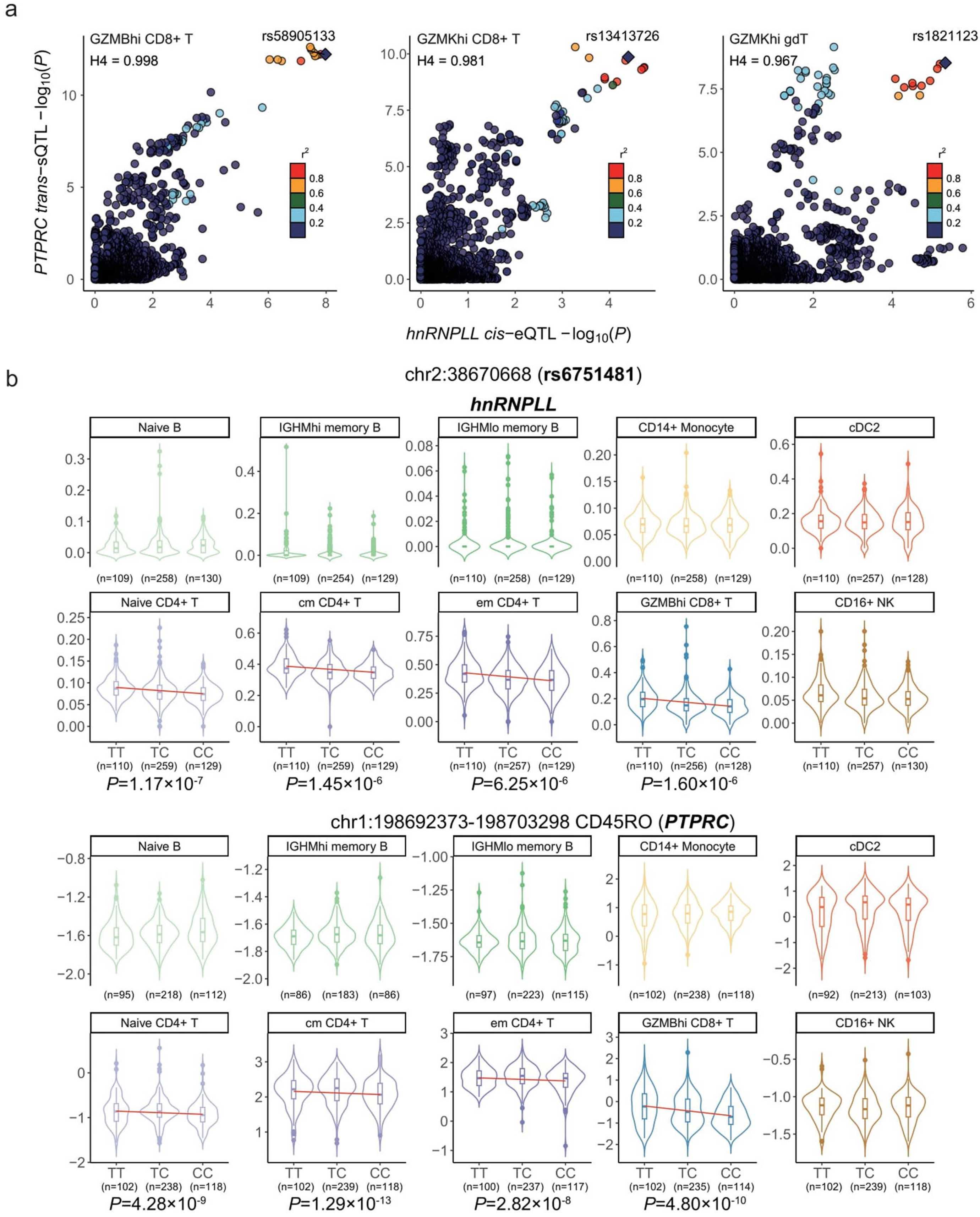
Extended Data Fig. 6 | Examples of cell-type-specific sQTLs in known SLE risk genes. A total of 30 cell-type-specific *cis*-sQTLs affecting known risk genes in Systemic Lupus Erythematosus. The alternate allele of the lead SNP rs147291617 upregulated an intron junction (chr17:36103981-36104528) of *CCL4* in a cell-type-specific fashion. Dark blue blocks in the left panel indicates the existence of *cis*-sQTL. Red lines in violin plots in the right panel indicate the significant linear relationships between the junction ratios of chr17:36103981-36104528 and the genotype of rs147291617 in CD16+ Monocyte, CD16 + NK, cyt CD4 + T, em CD4 + T, GZMBhi CD8 + T, GZMKhi CD8 + T, MAIT, GZMKhi gdT and GZMBhi gdT. The lack of red lines in the violin plot of CD14+ Monocyte, IGHMhi memory B, and IGHMlo memory B indicates no significant relationship between the junction ratios of the intron and the genotype of rs147291617. The box plots show median and interquartile range (IQR), and whiskers and 1.5-fold IQR.



Extended Data Fig. 7 | Examples of dynamic intron usage. Boxplot of dynamic intron usage change of *PAX5*, *PTPRC*, and *DOCK8*. Each data point within the boxplot corresponds to the intron usage measurement of an individual, and these points are organized into six different quantiles. The box plots show median and interquartile range (IQR), and whiskers are 1.5-fold IQR. The samples sizes N for each quantile are: Q1(N = 4190), Q2(N = 4250), Q3(N = 427), Q4(N = 450), Q5(N = 448), Q6(N = 449). To enhance clarity, the bars in the boxplot are color-coded to represent various quantiles. The curve displayed within each bar plot provides insight into the three patterns (step-wise change, linear change, and quadratic change) of intron usage changes from the first quantile (Q1) to the sixth quantile (Q6), offering a visual representation of how intron usage varies across these quantiles. Red dot shows the median intron usage of each quantile.

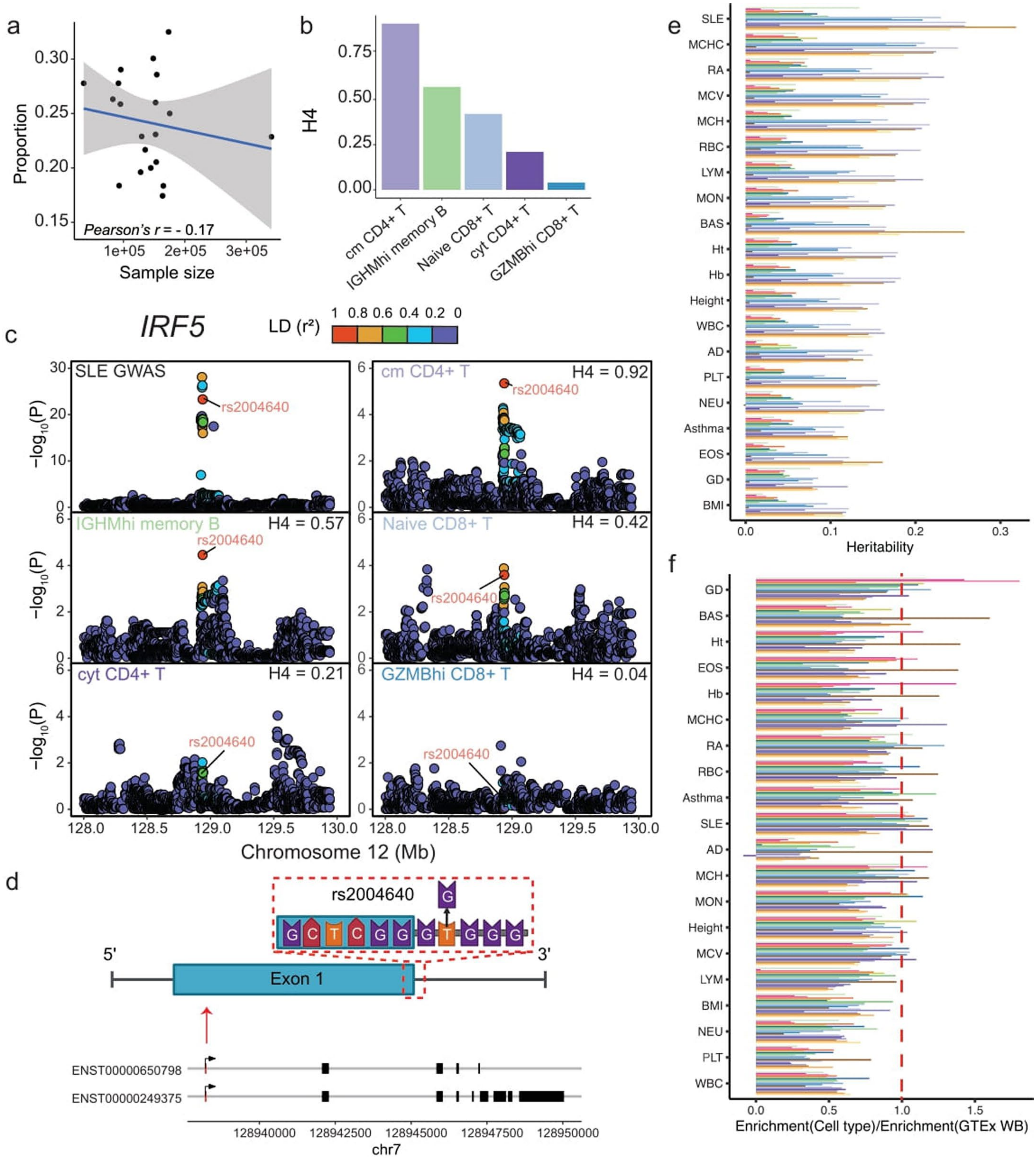


Extended Data Fig. 8 | Examples of dynamic sQTLs colocalization results. (**a**) The first dynamic sQTL example involves rs6936285. rs6936285 shows a decreased effect on CD83 splicing during the B cell maturation and is highly colocalized with RA in naïve B cells. Unadjusted two-sided *P* value was calculated by QTLtools (right panel). Red lines in box plots indicate the effect trend of genotype on intron usage. (**b**) The second dynamic sQTL example of rs16971619, which inserts increased effect on BCL2A1 splicing. It is found to be colocalized with lymphocyte count. The box plots show median and interquartile range (IQR), and whiskers are 1.5-fold IQR. The samples sizes N for each quantile are: Q1(N = 419), Q2(N = 425), Q3(N = 427), Q4(N = 450), Q5(N = 448), Q6(N = 449). Unadjusted two-sided *P* value was calculated by QTLtools (right panel). Red lines in box plots indicate the effect trend of genotype on intron usage.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | *Trans*-sQTL analysis revealed a regulatory relationship between *hnRNPLL* and *PTPRC*. **(a)** Colocalization between *hnRNPLL cis*-eQTL and *PTPRC trans*-sQTL. We identified colocalization (H4 > 0.75) in GZMBhi CD8$^+$ T, GZMKhi CD8$^+$ T, and GZMKhi γδ T cells. Unadjusted two-sided *P* values were obtained using Matrix eQTL (eQTL) and QTLtools (sQTL). **(b)** Violin plot showing the cell-type-specific effect of *hnRNPLL cis*-eQTL and *PTPRC trans*-sQTL. The minor allele of rs6751481 leads to a lower expression of *hnRNPLL* (upper panel) and a lower expression of *CD45RO* isoform (lower panel). Unadjusted two-sided *P* values were obtained using Matrix eQTL (upper) and QTLtools (lower). The number of donors for each genotype is shown under each violin plot. The box plots show median and IQR, and whiskers are 1.5-fold IQR. Red lines indicate significant linear relationship between intron usage and genotype.



Extended Data Fig. 10 | Aberrant splicing mediated complex autoimmune. **(a)** Correlation between GWAS sample size (x-axis) and proportion of colocalized loci (y-axis). A low correlation (*Pearson's r* = -0.17) was observed between the proportion of colocalization events and GWAS sample size across 20 traits. Each black dot in the panel represents a trait. The dark blue line indicates the linear relationship between the proportion of colocalized loci and GWAS sample size. The shaded area on either side of regression line represents 95% confidence interval. **(b)** H4 posterior probability of *IRF5* in five cell types. H4 posterior probability measures the association level between *cis*-sQTLs and SLE GWAS. H4 > 0.75 was used as the threshold for the colocalization. **(c)** Cell-type-specific colocalization results of *IRF5* in SLE GWAS. *IRF5* sQTL colocalized with SLE GWAS in cm CD4$^+$ T but not in IGHMhi memory B, Naïve CD8 + T, cyt CD4 + T and GZMBhi CD8$^+$ T. Unadjusted two-sided *P* value was calculated by QTLtools. **(d)** Schematic to show how causal SNP rs2004640 disrupts the 5′ splice site of exon 1B, leading to nonsense-mediated decay (NMD) and downregulation of *IRF5* expression. **(e)** Absolute heritability for 20 traits mediated by *cis*-sQTLs from 19 cell types. **(f)** The ratio between Heritability enrichment for 20 traits mediated by *cis*-sQTLs from 19 cell types and Heritability enrichment for 20 traits mediated by *cis*-sQTLs in GTEx whole blood. Red dash line represents the ratio equals to 1.

nature portfolio

Corresponding author(s): Boxiang Liu

Last updated by author(s): Aug 14, 2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Sequel IIe (ICS version 11.0.1.162970), skera (v0.1.0-8), SMRT Link (version 11.1.0.166339), Lima (v2.6.0-6), Isoseq3 (v3.8.2), pbmm2 (v1.10.0)
Data analysis	bcftools (v1.9), PLINK (v1.9), STARsolo (v2.7.10a), bedtools (v2.27.1), samtools (v1.16.1), RegTools (0.0.1), LeafCutter (0.2.9), QTLtools (v1.2), mashr (0.2.79), slingshot (v2.10.0), lme4 (1.1-35.3), COLOC (v5.2.3), SMR (v1.3.1), SpliZ (v1.0), deepTools (v3.5.1), qtlBHM (v1.0), susieR (0.12.10), LocusCompare (v1.0.0), LDSC (v1.0.1), ggplot2 (v3.4.2), qvalue (2.30.0), ggsashimi (v1.1.5), circlize (0.4.15). Custom code can be found at Github (https://github.com/boxiangliulab/AIDA_phase1_sQTL).

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

Data

Policy information about availability of data

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

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

The AIDA Data Freeze v1 gene-cell matrix (1,058,909 cells from 503 Japan, Singaporean Chinese, Singaporean Malay, Singaporean Indian, and South Korea Asian donors and 5 distinct Lonza commercial controls), with BCR-seq and TCR-seq metadata, and donor age, sex, and self-reported ethnicity metadata, is available via the Chan Zuckerberg CELLxGENE data portal at https://cellxgene.cziscience.com/collections/ced320a1-29f3-47c1-a735-513c7084d508. The open-access AIDA datasets are available via the Human Cell Atlas Data Coordination Platform at https://data.humancellatlas.org/explore/projects/f0f89c14-7460-4bab-9d42-22228a91f185. The managed-access AIDA datasets are available via data access applications to Shyam Prabhakar (prabhakars@gis.a-star.edu.sg). The PBMC sQTL data is deposited in Zenodo (doi: https://zenodo.org/record/8343365).

Research involving human participants, their data, or biological material

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

Reporting on sex and gender	Most findings of this study will apply to both male and female because sQTL mapping was performed using partipants of both sex while adjusting sex as a covariate. Sex balance was considered in the design of sequencing batches. Sex is self-reported, but participants with discordant data-driven sex were excluded in the analysis. The AIDA cohort incorporates a balance of female and male sex. There are 284 females (56.5%) and 219 males (43.5%). Disaggregated sex of AIDA participants can be found at https://explore.data.humancellatlas.org/projects/f0f89c14-7460-4bab-9d42-22228a91f185/project-metadata. Sex-biased differential splicing and sQTL analysis were performed in this study to assess the impact of sex on splicing and genetic regulation of splicing.
Reporting on race, ethnicity, or other socially relevant groupings	Donor self-reported ethnicity was collected from donors via questionnaires. In the AIDA Data Freeze v1 dataset, 75 Singaporean Chinese, 60 Singaporean Indian, 54 Singaporean Malay, 149 Japan Japanese, and 165 South Korea Korean donors were profiled for a total of 503 Asian donors. For ancestry-related analysis, Singaporean Chinese, Japan Japanese and South Korea Korean were grouped as East Asian, Singaporean Indian was categorized as South Asian, and Singaporean Malay was categorized as Southeast Asian.
Population characteristics	Population characteristics are described in Supplementary Note Table 1.
Recruitment	Healthy donors were included in the current version of atlas through applying the following exclusion criteria for donor datasets: 1) A person unable to provide informed consent. 2) A person with active infection or fever. 3) A person on regular medication (consumption of dietary supplements and / or herbal remedies was not considered in the exclusion of participants from our study). 4) A person with autoimmune disease. 5) A person with haemoglobin A1c (HbA1c) ≥ 6%. In addition, persons who had received any vaccines in the 8 weeks prior to the date of blood draw were excluded.
Ethics oversight	Local researchers from AIDA consortium member countries were involved in study design, study implementation, data ownership, intellectual property, and authorship of publications throughout the research process. All participants have been approved by local ethics review committees prior to study enrollment. All study protocols were approved by the Institutional Review Boards (IRBs) of the institutions that the laboratories are affiliated with (Genome Institute of Singapore: IRB 2020-012 and 2022-051; Nanyang Technological University: IRB-2016-11-030-01, IRB-2016-11-030, and 18IC4698; RIKEN: IRB H30-9; Samsung Genome Institute: IRB 2019-09-121; Faculty of Medicine Siriraj Hospital, Mahidol University: IRB 725/2563(IRB3); National Institute of Biomedical Genomics: IRB NIBMG/2022/1/0022) prior to dataset generation. All donors provided written informed consent for sample and metadata collection and subsequent analyses.

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We included all donors with both quality-controlled genotype and single-cell RNA sequencing (scRNA-seq) data. For each cell type, donors with fewer than 10 cells were excluded from the analysis. This filtering process resulted in the final sample sizes for each cell type, as detailed in Supplementary Table 1. The sample sizes range from 114 to 459 (> 400 for most cell types). These numbers are comparable to those reported in previous studies.
Data exclusions	Donors whose samples failed genotyping or scRNA-seq quality control, as well as related donors (up to third-degree cousins) were removed. Genetic variants with missingness > 0.05, MAF < 0.05, or HWE-P < 0.000001 were excluded from analysis. The exclusion criteria were not pre-established.
Replication	To replicate cis-sQTLs with external datasets, we used four publicly available PBMC-related RNA-seq datasets: (1) The BLUEPRINT dataset

Replication	(N=197); (2) The DICE dataset (N=91); (3) GTEx whole blood (N = 670) and lymphoblastoid cell lines (LCLs; N=147); (4) ImmuNexUT33 (N = 416). We used π1 statistics to assess replication between AIDA and matching cell types from above datasets. π1 estimates ranged from 0.91 – 0.93 for BLUEPRINT, 0.83 – 0.91 for DICE, 0.70 – 1 for GTEx whole blood (except for cDC2; π1 = 0.48), 0.71 – 0.86 for GTEx LCL and 0.70 – 1 for ImmuNexUT (except for CD56+ NK; π1 = 0.42), suggesting that AIDA cis-sQTLs replicated well.
Randomization	Randomization is not relevant because this is an observational study rather than an experimental study. There is no experimental group in the study design.
Blinding	Blinding is not relevant because this study does not involve group allocation.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	K562 Cell Line is from ATCC (CCL-243™).
Authentication	The cell line was not re-authenticated after purchasing from ATCC.
Mycoplasma contamination	The cell line was not re-tested for Mycoplasma after purchasing from ATCC.
Commonly misidentified lines (See ICLAC register)	N/A

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*