nature genetics

Analysis

https://doi.org/10.1038/s41588-023-01586-6

Mapping the dynamic genetic regulatory architecture of *HLA* genes at single-cell resolution

In the format provided by the authors and unedited

Table of Contents

Supplementary Note 1

This section contains additional results from assessing the performance of the scHLApers pipeline.

Examining reads switching alignments from *HLA-B* to *HLA-C*

Given the shared evolutionary history of class I genes[1], we hypothesized that the observed decrease in *HLA-B* expression after personalization was due to reads aligned to *HLA-B* in the standard pipeline aligning to a different gene in scHLApers. By tracking where individual reads aligned before and after personalization for Synovium and PBMC-cultured using the BAM files generated by STARsolo, we found that in both datasets, 99% of reads that previously aligned to *HLA-B* (but aligned to a different location after personalization) aligned instead to *HLA-C*. We then tracked where the read alignments to *HLA-C* in scHLApers "came from" in the standard pipeline (**Extended Data Fig. 1d**). For Synovium, 14.8% came from *HLA-B* in the standard pipeline, 75.1% were originally also aligned to *HLA-C* in the standard pipeline, 8.3% came from unmapped reads, and the remaining 1.8% came from other genomic regions. For PBMC-cultured, the breakdown was 2.5% *HLA-B*, 51.7% *HLA-C*, 44.4% unmapped, and 1.4% other.

Interestingly, an individual's change in *HLA-B* counts depended on their *HLA-C* genotype, supporting the observed decrease in *HLA-B* after personalization. Performing a multiple sequence alignment on the *HLA-C* alleles present in our cohorts showed that the reference allele (*HLA-C*07:02*) grouped with a set of similar "reference-like" alleles (**Extended Data Fig. 1e**): *HLA-C*04:04*, *C*04:01*, *C*18:01*, *C*14:02*, *C*14:03*, *C*01:02*, *C*03:04*, *C*03:03*, *C*03:05*, *C*03:02*, *C*17:01*, *C*07:01*, and *C*07:04*. For individuals with both *HLA-C* alleles similar to the reference allele (*HLA-C*07:02*), *HLA-B* was less affected by personalization (**Extended Data Fig. 1e**). However, for individuals with at least one "non-reference-like" *HLA-C* allele (i.e., other than *HLA-C*07:02*), some reads aligned to *HLA-B* before personalization aligned better to *HLA-C* after personalization, leading to decreased *HLA-B* counts.

Application of scHLApers to 10x 5' data shows different trends for *HLA-A* and *HLA-B* compared to 3' data

For the main analyses in the study, we used datasets sequenced with 10x 3'-based libraries. As a proof-of-concept analysis, we also demonstrated the feasibility of scHLApers on 5'-based data. We applied scHLApers to a separate dataset from a subset of Synovium individuals with matching 10x 5' data (n=9 individuals, 26,638 cells). We found that in 5' data, estimates for all eight classical *HLA* genes increased after personalization (**Supplementary Fig. 3b**). We sought to understand why the estimated expression increased for *HLA-A* and *HLA-B* in 5' data but not 3' data. We determined that the difference in trends are a result of *HLA-A* and *HLA-B* alleles having increased dissimilarity from the reference allele on the 5' end of the genes compared to the 3' end (*HLA-A* mean Levenshtein distance = 4.1 at the 3' end vs. 16.3 at the 5' end; *HLA-B* = 4.7 at the 3' end vs. 19.0 at the 5' end, **Supplementary Table 5, Supplementary Fig. 3c,d**). In 3' data, where the dissimilarity to the reference allele is low, the estimated counts remained unchanged for *HLA-A*. For *HLA-B*, changes in expression were driven by realignment from *HLA-C*. However, in 5' data, the greater dissimilarity to the reference allele for *HLA-A* and *HLA-B* resulted in greater increases in their estimated expression levels after personalization, due to rescuing of unmapped reads and improved read assignment. In contrast, *HLA-C* expression estimates increased after personalization for both 3' and 5' data because *HLA-C* alleles had comparable dissimilarity from the reference on both ends of the gene (12.1 at 3' end vs. 11.8 at 5' end).

Technical note on *HLA* allele calling and allele-specific expression

The scHLApers pipeline requires *HLA* allele calls per individual, which can be obtained directly by sequence-based typing or by HLA imputation using genotyped variants. There have been efforts to use bulk RNA-seq to infer *HLA* alleles without orthogonal genotype data[2,3]; however, inferring alleles from single-cell reads with high accuracy may prove challenging beyond one-field resolution[4]. Allele-specific expression (ASE) analysis is an alternative way to detect regulatory effects. However, it is challenging to map reads unambiguously between alleles using short-read 3'-based sequencing data because it largely excludes the highly variable 5' region of the gene (**Supplementary Fig. 3**). In contrast, 5'-based data may be more effective for ASE.

Supplementary Note 2

See the separate document (appended to the end of this PDF), which contains additional methods regarding the removal of suspected doublets for the PBMC-blood dataset (OneK1K cohort). We present these methods as a separate supplementary note to allow it to serve as a standalone entity and be referenced by subsequent work.

Supplementary Note 3

This section discusses how an eQTL interacting with cell states can manifest in pseudobulk analysis when testing for interactions with contextual variables that themselves influence cell state abundances.

As a motivating example, consider the proportion of cell states as a sample-level variable. Our single-cell analysis showed that eQTLs for *HLA-DQA1* (rs3104371) and *HLA-DQB1* (rs9272271) interact with T cell states. Consequently, one would expect to identify a similar signal when conducting bulk-level analysis when testing for interactions with the proportion of pertinent cell states. To test this, we used PBMC-blood T cells and calculated each sample's proportion of cytotoxic T cells (prop_cyto) and naïve T cells (prop_naive). We then tested these variables for interaction with T cell *HLA* eQTLs using a pseudobulk model (**Methods, Supplementary Table 14**). As expected, we found that *HLA-DQA1* eQTL significantly interacted with prop_cyto (LRT P=9.62x10^{-4}) and prop_naive (LRT P=2.24x10^{-3}), and the *HLA-DQB1* eQTL interacted with prop_naive (LRT P=8.47x10^{-3}), which is consistent with the cell states implicated in single-cell eQTL analysis. This highlights that pseudobulk analysis shows evidence of eQTL interaction with a factor—in this case, cell state proportions—which mediates its effect through a cell-state-dependent eQTL. However, the significance of these interactions is diminished compared to the dramatically significant dynamic eQTL effects detected using the single-cell NBME model, where interaction significance reached values of $P < 1x10^{-28}$.

We next performed a pseudobulk analysis to explicitly test for the interaction between eQTLs and age, sex, and IFN response using the PBMC-blood dataset (n=909 samples) in each cell type (**Methods**). We found that out of the 24 lead eQTLs tested, 2 had a nominally significant interaction with age (LRT P<0.05), none interacted with sex, and 4 had nominally significant interactions with IFN response (**Supplementary Table 14**). The strongest signals were observed for T cell *HLA-DQB1* eQTL (rs9272271, age LRT P = 0.00297, **Supplementary Fig. 15a**), which showed stronger eQTL effect at older ages, and B cell *HLA-DPA1* eQTL

(rs2163472, IFN response LRT P = 0.00661, **Supplementary Fig. 15e**), which showed weaker eQTL effect with higher IFN response.

We sought to link these interactions to underlying shifts in cell state abundance, analogous to the motivating example of cell state proportion above. We determined that the *HLA-DQB1* age-interacting eQTL can be explained by a shift from naïve to cytotoxic T cells in the peripheral blood with increasing age. Naïve cells decrease in abundance with age (Pearson r = -0.53, P<2.2x10^{-16}; **Supplementary Fig. 15b**), and cytotoxic cells increase (r = 0.28, P<2.2x10^{-16}). Our single-cell eQTL analysis testing for cell state interaction (NBME model) showed that the eQTL effect (β_{total}) is weaker in naïve cells compared to cytotoxic cells (**Supplementary Fig. 15c,d**), confirming the observed interaction effect with age can be explained by the underlying age-associated T cell states. Similarly, we were able to explain the interaction between IFN-response and the *HLA-DPA1* eQTL at least partially in terms of cell state shifts. Among B cells, plasmablasts have both the highest IFN scores and weakest eQTL effects as estimated in our NBME model (**Supplementary Fig. 15f,g)**. At the sample level, higher IFN response tracks with greater abundance of plasmablasts (r = 0.16, P=9.32x10^{-7}), consistent with the weaker eQTL effect observed with higher IFN response. These examples show how dynamic eQTLs, which vary in strength across cell states, can underlie eQTL interactions with sample-level factors that are themselves associated with shifts in cell states.

Supplementary Note 4

This section contains supplementary methods.

Details of four cohorts

Synovium

The original study[5] collected synovial biopsies from patients with RA (n=70) and control patients with osteoarthritis (n=9) as part of the Accelerating Medicines Partnership (AMP) RA/SLE Phase 2 Consortium. In this study, one RA sample was excluded due to lack of genotyping, and nine more individuals were excluded because we could not impute phased alleles for all classical *HLA* genes (see “HLA imputation”), resulting in a final cohort of 69 individuals. For three individuals with repeat biopsies, we included only initial biopsies. The original study used CITE-seq to simultaneously measure single-cell RNA and 58 protein markers, but this study uses only the RNA data. RNA libraries were generated using the 10x Genomics 3’ v3 protocol, and sample-level FASTQ files were used for input to scHLApers.

Intestine

The original study included intestinal biopsies from 30 individuals[6]. We accessed the read-level scRNA-seq data via the Broad Data Use Oversight System (https://duos.broadinstitute.org; dataset: Ulcerative_Colitis_in_Colon_Regev_Xavier). After removing five individuals without genotyping data and three individuals for whom we could not impute phased alleles for all classical *HLA* genes, the final cohort consisted of 22 individuals for this study. RNA libraries were generated using the 10x 3’ v1 protocol (n=12) or 3’ v2 protocol (n=10). Sample BAM files were used for input to scHLApers.

PBMC-cultured

The original study collected whole blood from healthy male donors (n=90) of African and European ancestry[7]. Samples from each individual were treated with influenza A virus or mock conditions (90*2 = 180 samples). Libraries were prepared using the 10x 3' v2 protocol and multiplexed in 30 experimental batches for sequencing. We downloaded the batch-level scRNA-seq FASTQ files from GEO (GSE162632). We removed one individual with missing genotype data, one based on our IBD threshold (IBD PI_HAT > 0.9), and 15 individuals for whom we could not impute phased alleles for all classical *HLA* genes, leading to a final cohort of 73 individuals. To demultiplex the batches, we obtained a barcode-to-sample mapping file from the original authors. We used the *filterbarcodes* function from sinto (v0.8.4) and the mapping file to demultiplex the batch-level BAM file into sample-level BAM files for input to scHLApers.

PBMC-blood

The OneK1K cohort[8] was recruited from patients and relatives at clinical sites and retirement villages in Australia. PBMCs were collected and sequenced using the 10x 3' v2 protocol. We obtained multiplexed read-level data in BAM file format across 77 batches from the original authors. We used the *filterbarcodes* function from sinto (v0.8.4) to demultiplex each batch-level BAM file into constituent sample-level BAM files using a batch-to-sample cell barcode mapping provided by the original authors. Starting with an initial cohort of 973 individuals with paired genotyping and scRNA-seq data, we excluded three samples due to elevated missing data rates during genotype quality control (QC) and one sample missing in the barcode mapping file. We also removed 60 samples where we could not impute phased alleles for all classical *HLA* genes, leading to a final cohort of 909 samples for this study.

Details of quality control of genotyping data

Synovium

We genotyped donors in the AMP RA/SLE Network (including the 69 donors in this study) using the Illumina Multi-Ethnic Genotyping Array split across three batches. We lifted over the data from hg38 to hg19 coordinates, converted reverse strand variants to forward alleles with snpflip (v.0.0.6), and removed duplicated variants with a custom script. We performed initial QC to remove variants with high missing call rates or violating Hardy-Weinberg equilibrium (--geno 0.1 --hwe 1e-8) and removed 1 sample with high missingness (--mind 0.01). We then applied variant filters (--hwe 1e-6, --geno 0.01) and removed indels. We merged all three batches, removed multi-allelic variants, and removed variants with MAF <1% (--maf 0.01) and >1% missingness (--geno 0.01) across all samples. A total of 820,019 genome-wide variants (10,159 in MHC) and 788 individuals passed genotype QC.

Intestine

All study subjects were recruited under IRB protocols from local institutions[6]. The genotype data (including the 22 donors in this study) comprised three batches, denoted batch1 (n=774 individuals), batch2 (n=874), and batch3 (n=1,335). The batch1 data was generated using an Illumina Infinium Global Screening Array, and the others used a custom GWAS SNP array. Each batch was quality controlled separately. We first removed SNP duplicates, performed an initial variant filtering (--geno 0.1 --hwe 1e-10), then removed samples with high missingness (--mind 0.01) or high sample-relatedness (IBD PI_HAT > 0.9). We removed variants with >2% missingness (--geno 0.02) or >0.35 allele frequency difference from 1000 Genomes project phase 3 v.5 data. We then applied final filters (--maf 0.01 --hwe 1e-10 --mind 0.01 --geno 0.02). For batch1, 488,343 genome-wide variants (7,544 in MHC) and 765

individuals passed QC. We merged the batch2 and batch3 datasets by their 222,030 shared SNPs genome-wide (772 in MHC), and 2,121 individuals passing QC.

PBMC-cultured

The imputed WGS data (VCF, n=91) was obtained directly from the original authors (data available at SRA accession PRJNA736483). To match the HLA imputation reference panel, we lifted over the variants from GRCh38 to hg19 positions using CrossMap (v0.6.1) with chain file http://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz. We marked genotypes with missing calls using bcftools (v1.9) if posterior genotype probabilities were <95%. We removed duplicated and multiallelic variants using PLINK v1.90 and a custom script. After initial variant QC (--geno 0.1 and –hwe 1e-10), we performed sample-level QC based on genotype missingness (--mind 0.01) and genetic relatedness using IBD comparisons on variants pruned for high LD. One sample (HMN52561) was excluded based on high genetic relatedness to another sample (HMN17122; IBD PI_HAT > 0.9). We removed variants with >2% missingness or >0.25 allele frequency difference from 1000 Genomes phase 3 v.5, resulting in 5,133,858 variants passing QC. As a final step, we subset variants to those in our HLA imputation reference panel and removed variants with MAF<5% (10,814 MHC variants and 90 individuals passing QC).

PBMC-blood

The raw genotyping data in hg19 coordinates was obtained from the original authors. We used snpflip (v0.0.6) to flip 247,054 variants to the forward strand and removed 682 duplicates. We performed an initial SNP QC (--geno 0.1, --hwe 1e-10) and removed three individuals with elevated missingness rates (--mind 0.01). We removed variants with high missingness (--geno 0.02), 2,508 ambiguous (A/T or C/G) SNPs, and variants with >0.3 allele frequency difference from 1000 Genomes phase 3 v.5. After applying final filters (--maf 0.01 --hwe 1e-10 --mind 0.01 --geno 0.02), 972 individuals and 487,471 genome-wide variants (7,046 in MHC) passed QC.

Details of power analysis for NBME model

For these analyses we used the PBMC-blood myeloid data and *HLA-DQA1*. To summarize our approach, we took the following steps:

1. Fit the NBME model to the actual data without the effect of genotype.

2. Using the fitted model, simulate *d*=1000 new sets of single-cell expression values for *HLA-DQA1*. Use the same number of cells and individuals to mimic the original dataset as closely as possible.

3. Given an allele frequency (*AF*), simulate genotype dosages (0, 1, or 2) for all individuals, assuming Hardy-Weinberg equilibrium.

4. “Spike in” the effect of genotype to the simulated expression values, assuming a fixed eQTL effect size (β_{NBME}). For every 1-unit increase in allele dosage, the mean log(counts) is increased by a factor of β_{NBME}. Hence, $E_{sim\_new} = \exp(\log(E_{sim\_orig}) + \beta_{NBME} * \text{dosage})$.

5. Determine whether the genotype effect is detectable by fitting a full NBME model (including genotype). Obtain significance value by likelihood ratio test comparing to a reduced model missing genotype.

6. Determine power across 1000 trials (proportion of times null rejected at $\alpha = 5 \times 10^{-8}$).

7. Repeat this process across different values for *AF* (0.01, 0.05, 0.1, 0.2, 0.35, 0.5) and β_{NBME} (0.1, 0.25, 0.5, 0.75, 1).

Testing for eQTL interactions with contextual factors

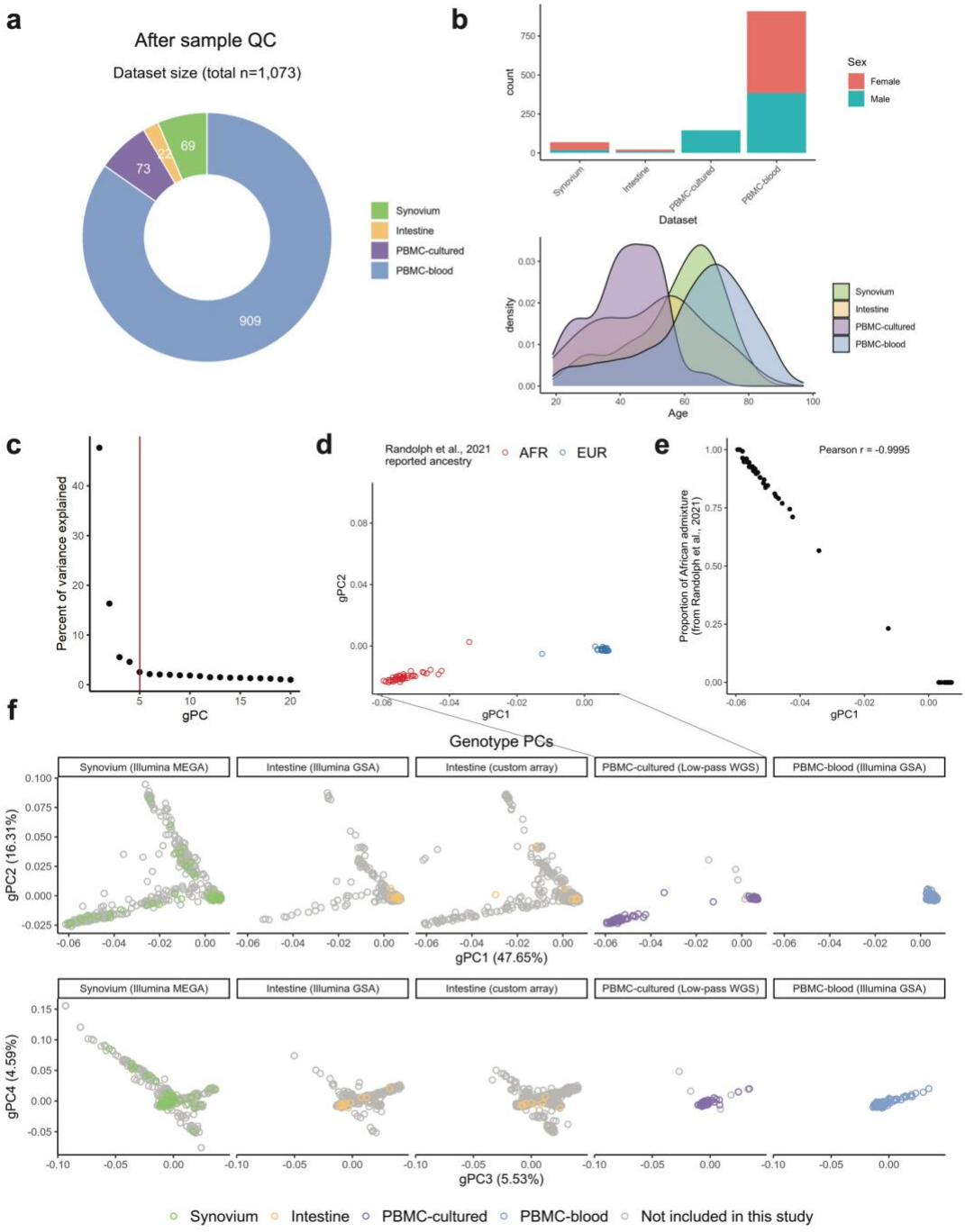
To explore whether cell-state-dependent effects identified with the NBME model are captured in sample-level contextual factors, we used the pseudobulk eQTL linear modeling framework to test for eQTL interactions with age, sex, and IFN response on the PBMC-blood dataset (n=909). IFN response was defined as an 11-gene signature of IFN response from Davenport et al. (*Genome Bio*, 2018): *HERC5, IFI27, IRF7, ISG15, LY6E, MX1, OAS2, OAS3, RSAD2, USP18, GBP5*. For each sample, we scaled each gene's normalized pseudobulk expression across samples, took the sum of the 11 scaled genes per sample, then scaled the resulting score. We performed the interaction analysis in each cell type (myeloid, B, and T), determining significance by LRT compared to a null model without the interaction. We tested age and sex interactions separately. For example, for testing the age interaction, we fit the model in **Eq. 9**. Note that the E_{resid_PEER} term denotes expression after residualizing out PEER factors, but not age, sex, or ancestry, as these were included as covariates. We tested the IFN response interaction using **Eq. 10**, which is the same except that the dependent variable E_{resid} is the expression residual after regressing out PEER factors, age, sex, and ancestry.

$$E_{resid\_PEER} = \beta_G X_G + \beta_{sex} X_{sex} + \sum_{k=1}^{5} \beta_{gPC_k} X_{gPC_k} + \beta_{age} X_{age} + \beta_{G\times age} X_{G\times age} + \varepsilon \quad (9)$$

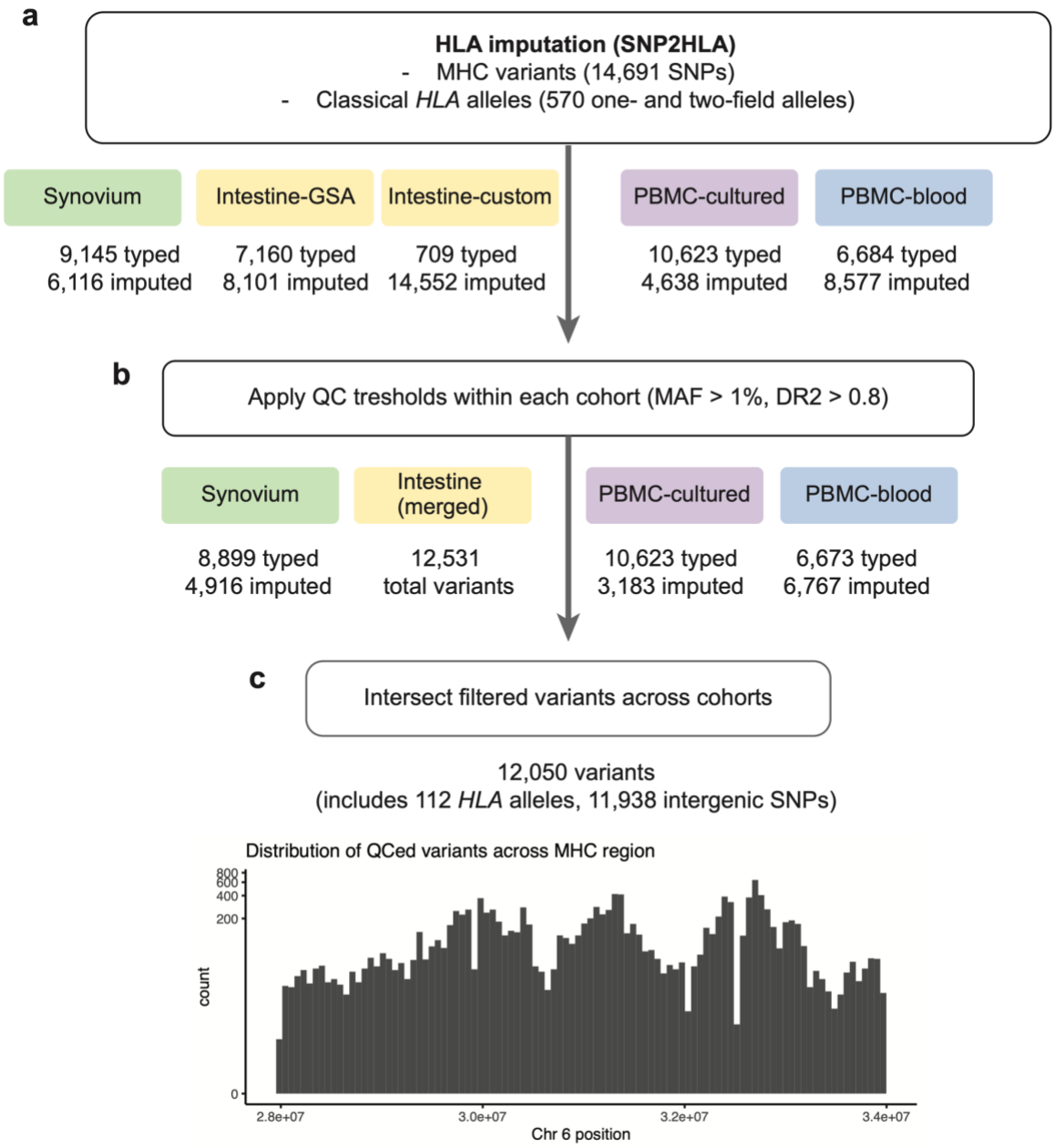
$$E_{resid} = \beta_G X_G + \beta_{IFN\_score} X_{IFN\_score} + \beta_{G\times IFN\_score} X_{G\times IFN\_score} + \varepsilon \quad (10)$$

Using the pseudobulk framework, we tested for eQTL interactions with proportion of cytotoxic T cells (prop_cyto) and proportion of naïve (prop_naive) T cells in PBMC-blood. Cytotoxic states were defined as T cells annotated "CD4+ Cytotoxic", "CD8+ Cytotoxic", or "gdT". Naïve were those labeled "CD4+ Naïve" and "CD8+ Naïve". We determined significance by LRT compared to a null model without the interaction. We tested prop_cyto and prop_naive interactions separately. For example, for testing prop_cyto interaction, we fit the model in **Eq. 11**.

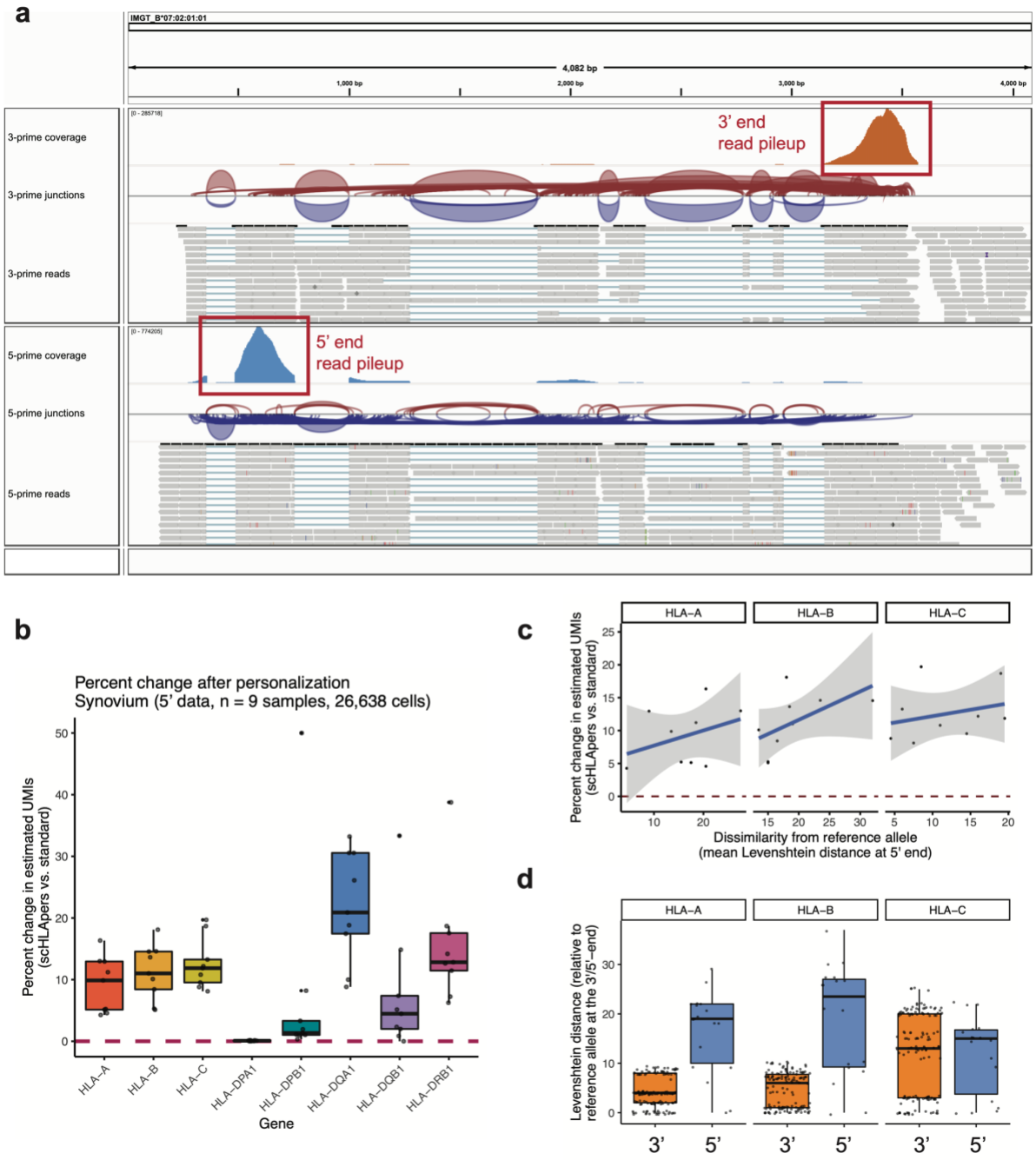
$$E_{resid} = \beta_G X_G + \beta_{prop\_cyto} X_{prop\_cyto} + \beta_{G\times prop\_cyto} X_{G\times prop\_cyto} + \varepsilon \quad (11)$$



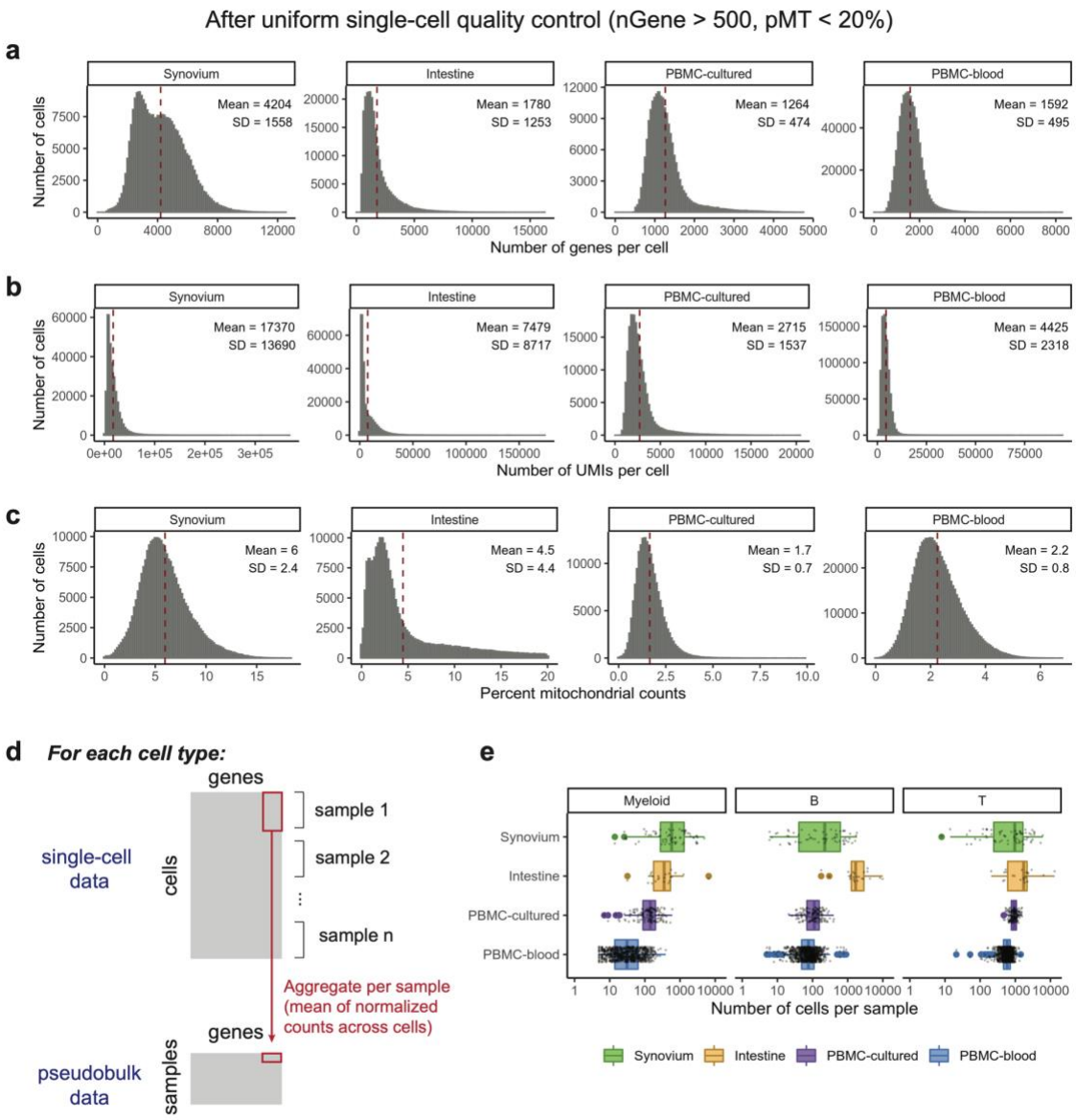
Supplementary Fig. 1. Cohort demographics. (a) Cohort sizes after QC. **(b)** Sex and age distributions. Note: PBMC-cultured is male-only. **(c-f)** Genotype PCs (gPC) capturing genetic ancestry, calculated on the intersecting genome-wide variants across all cohorts. **(c)** Percentage of variance explained by each gPC. Red line denotes five gPCs used in the eQTL analysis. **(d)** Reported ancestry (African = red, European = blue) of PBMC-cultured individuals. **(e)** PBMC-cultured individuals along gPC1 (*x*-axis) versus estimated proportion of African admixture (as reported by original study, *y*-axis). **(f)** Top four gPCs across individuals. Colors denote individuals included in the eQTL analysis, whereas gray individuals were genotyped on the same array (used in PCA) but did not have available scRNA-seq data. Note: all PBMC-blood individuals are European ancestry; Intestine cohort was genotyped across two arrays.



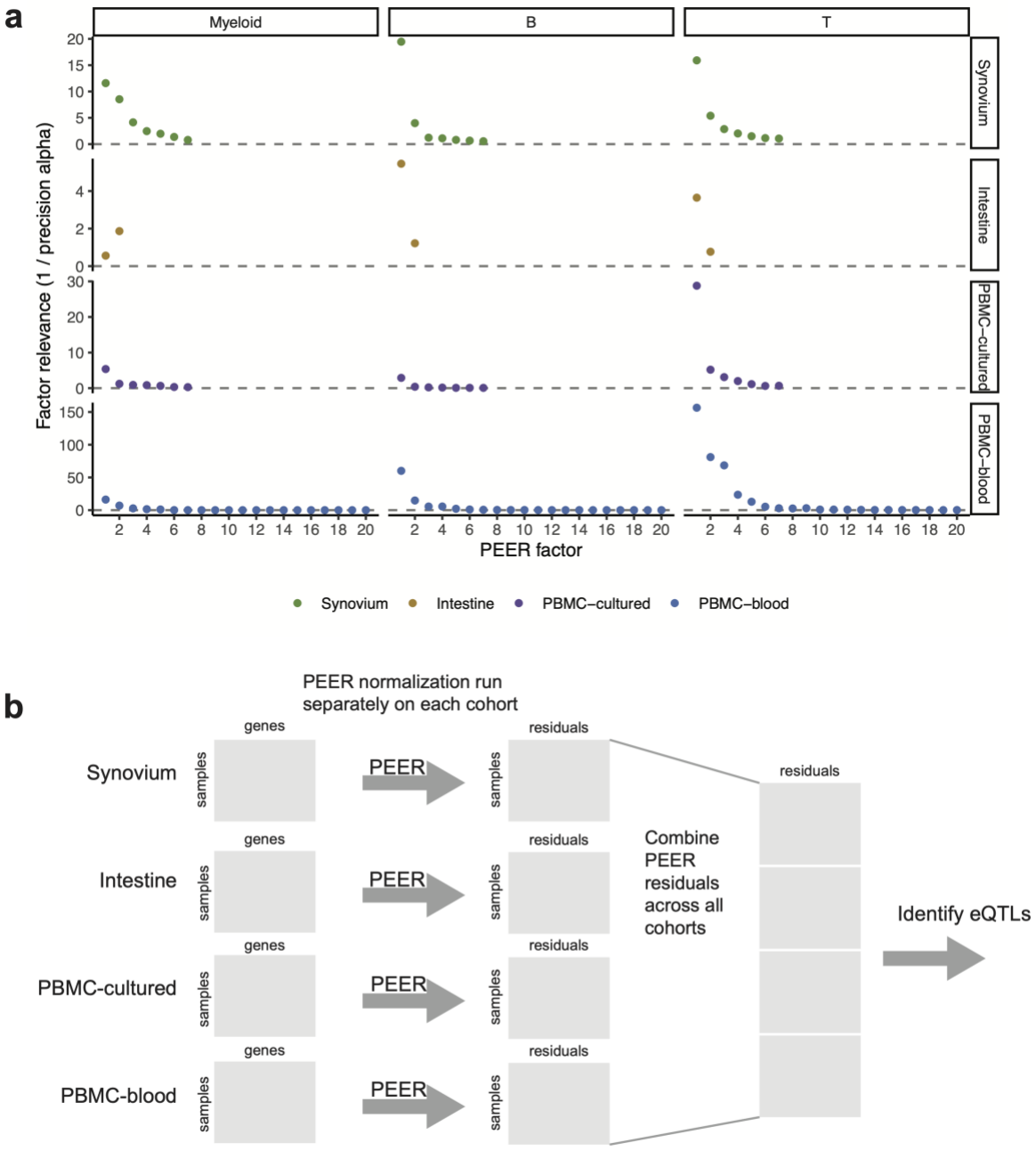
Supplementary Fig. 2. Imputation and quality control of MHC variants for eQTL analysis. (a) The number of starting typed and imputed MHC variants in each cohort; the Intestine dataset was genotyped on two arrays. **(b)** The number of variants remaining after filtering for MAF > 1% and DR2 > 0.8 within each cohort separately. **(c)** The final number of variants used in eQTL analysis after taking the intersection of variants passing QC across cohorts. The histogram shows the distribution of variants across the MHC region (*x*-axis).



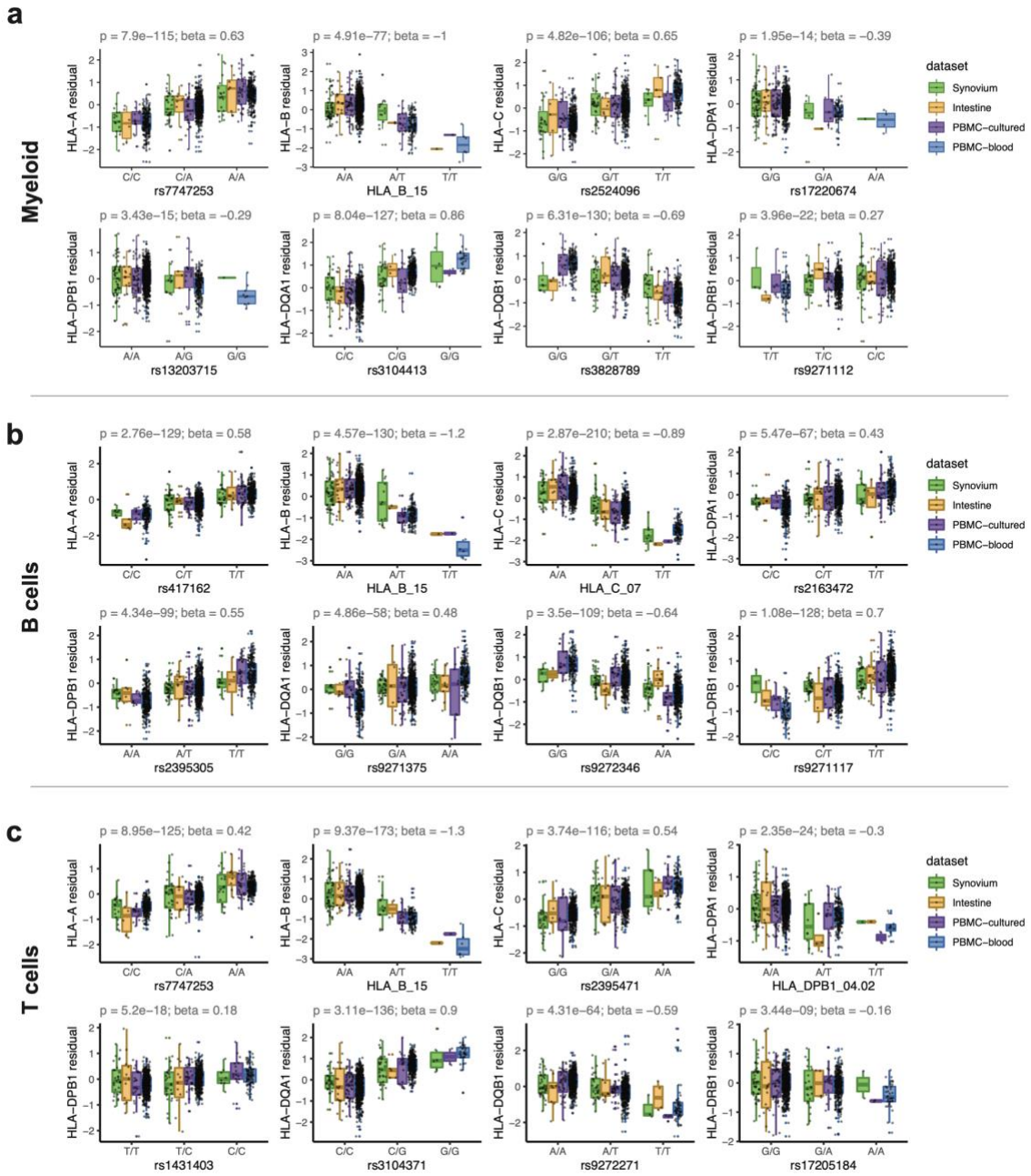
Supplementary Fig. 3. Effect of personalization on 10x 5' scRNA-seq data compared to 3' data. (a) Integrative Genomics Viewer (IGV) screenshot showing read alignments from scHLApers pipeline to a representative *HLA-B* allele for two samples (from the same Synovium individual), sequenced with 10x 3' assay (top, orange) and 5' assay (bottom, blue). Additional tracks show inferred splice junctions and example individual read alignments. **(b)** Boxplot showing percent change in estimated UMI counts (y-axis) in scHLApers vs. standard pipeline in 5'-based data (n=9 samples, 26,638 cells, subset of Synovium cohort), across each gene (x-axis). **(c)** Percentage change in estimated expression (total UMIs for *HLA* gene per individual, *y*-axis) in Synovium 5' data (n=9) as a function of the mean (between the individual's two alleles) Levenshtein distance relative to the GRCh38 reference allele at the 5' end of each gene (*x*-axis). Fitted linear regression line (blue) shown with 95% confidence region. **(d)** Boxplot of allele dissimilarities (Levenshtein distance relative to the reference allele) for class I genes in Synovium 3' (n=69 individuals, 138 alleles) vs. 5' (n=9 individuals, 18 alleles) data, where each dot is 1 allele. In panels **(b)** and **(d)**, Boxplot center line represents median, lower/upper box limits represent 25/75% quantiles, whiskers extend to box limit ±1.5 × IQR, and outlying points are plotted individually.



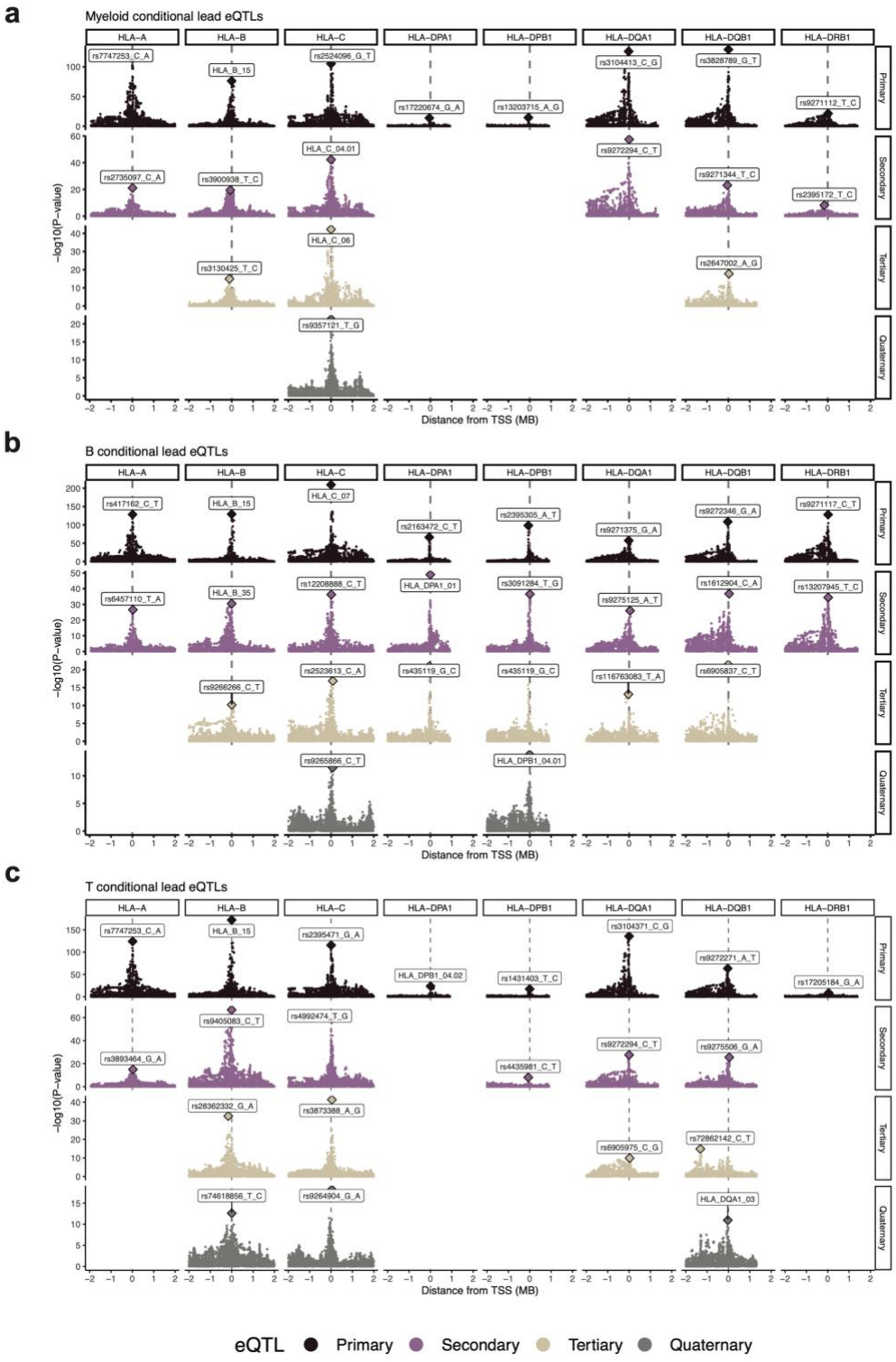
Supplementary Fig. 4. Single-cell dataset metrics after QC. Metrics for scRNA-seq data for each cohort after uniform QC (removing cells with fewer than 500 genes or greater than 20% mitochondrial UMIs). **(a)** The number of genes per cell. **(b)** The number of UMIs per cell. **(c)** The percentage of mitochondrial UMIs per cell. The red dotted line indicates the mean value across cells; mean and standard deviation (SD) are listed. **(d)** Schematic showing the aggregation process to generate pseudobulk profiles. For each sample and cell type, we take the mean of the log(CP10k+1)-normalized expression across cells for the sample. **(e)** Number of cells per sample in eQTL analysis by cell type and cohort (colors), after removing individuals with fewer than 5 cells of the cell type: myeloid n=1,025, B n=1,069, and T n=1,072 individuals total across all four datasets, see **Supplementary Table 2** for dataset breakdown. Boxplot center line represents median, lower/upper box limits represent 25/75% quantiles, whiskers extend to box limit ±1.5 × IQR, and outlying points are plotted individually.



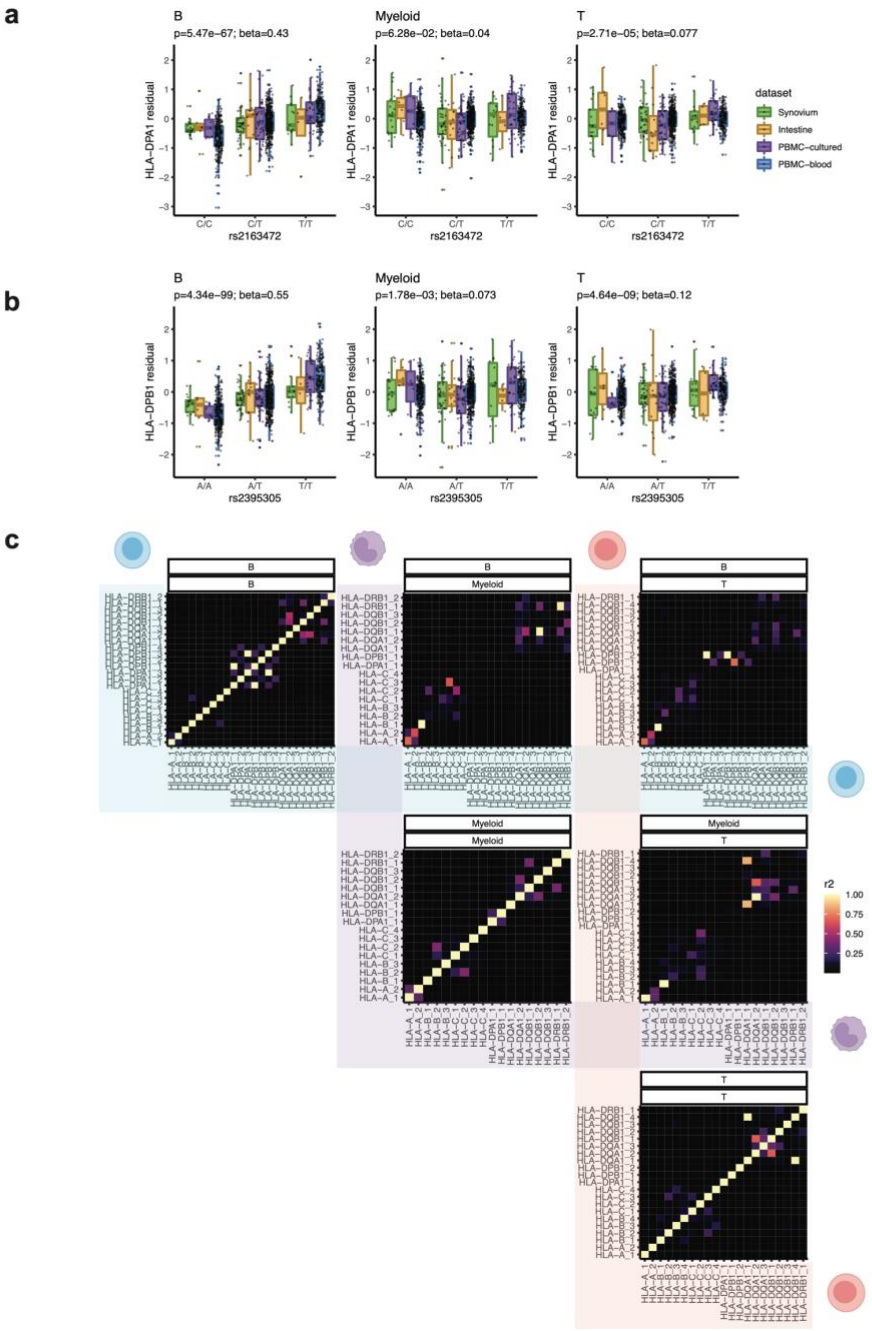
Supplementary Fig. 5. PEER factor relevance and schematic of the multi-cohort model. (a) Relevance of each PEER factor (*y*-axis) for each dataset and cell type. Different numbers of PEER factors were used for each cohort (K=7 for Synovium, 2 for Intestine, 7 for PBMC-cultured, and 20 for PBMC-blood). **(b)** Schematic of pseudobulk eQTL multi-cohort analysis strategy. Cells were mean-aggregated within each sample to obtain a samples-by-genes matrix for each cohort and cell type. Then, we ran inverse normal transformation and PEER factor normalization separately within each cohort to obtain a samples-by-residuals matrix for each cohort. We concatenated these matrices into a single matrix across all cohorts. We identified eQTLs for *HLA* genes using a single linear model, modeling the residual as a function of genotype and cohort across all individuals.



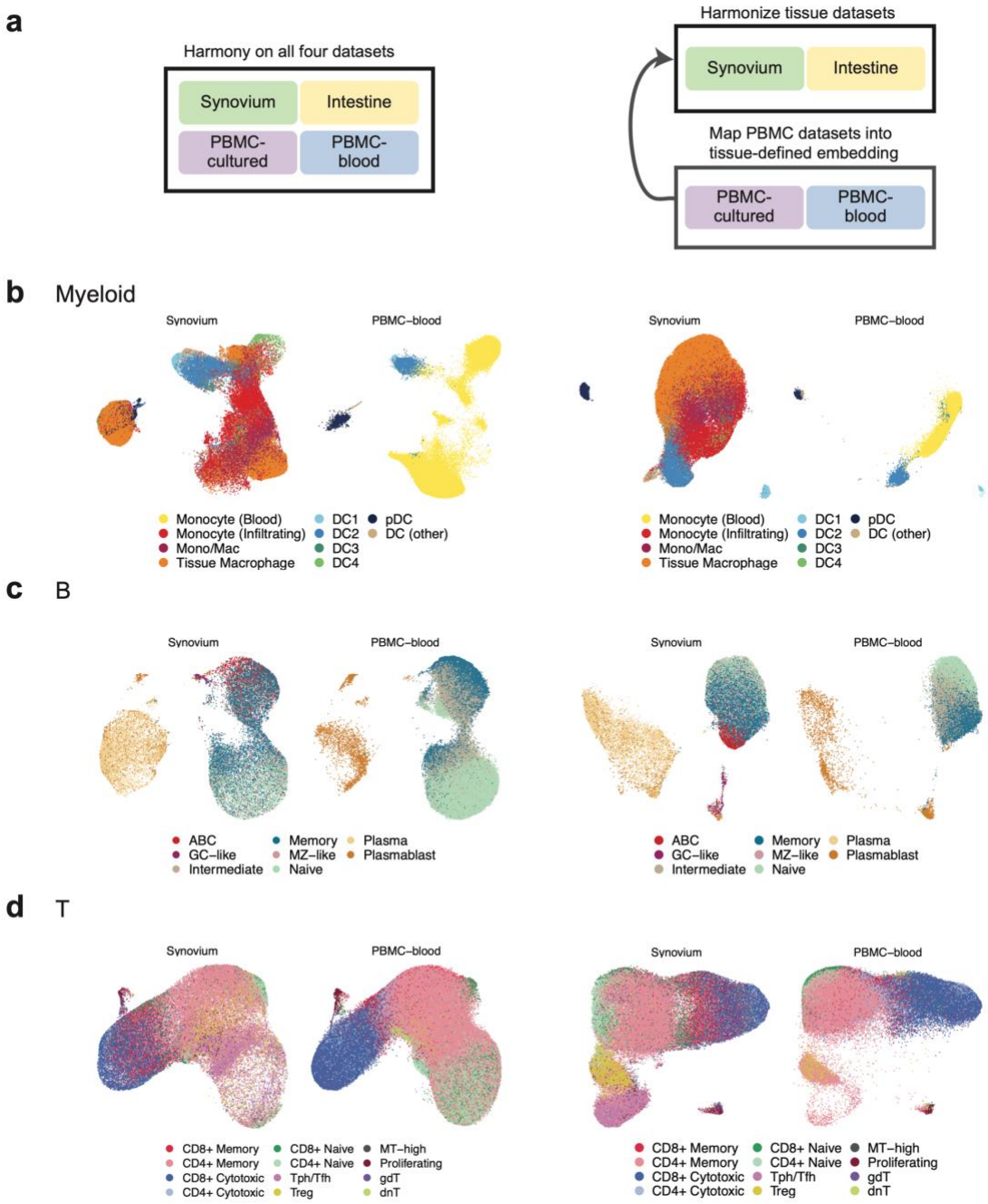
Supplementary Fig. 6. Pseudobulk model lead eQTLs in each cell type. Boxplots showing the effect of the lead eQTL for each gene from the multi-cohort model for **(a)** myeloid (n=1,025 individuals across all datasets), **(b)** B (n=1,069), and **(c)** T (n=1,072) cells. See **Supplementary Table 2** for number of samples per dataset for each cell type. The genotype of each individual (*x*-axis) is plotted against the inverse-normal transformed residual of the gene's expression (after adjusting for covariates, *y*-axis), plotted by dataset (color). Variants starting with "HLA" denote *HLA* alleles. Nominal Wald *P*-values are derived from linear regression (two-sided test). Boxplot center line represents the median; lower and upper box limits represent the 25% and 75% quantiles, respectively; whiskers extend to box limit ±1.5 × IQR; outlying points are plotted individually.



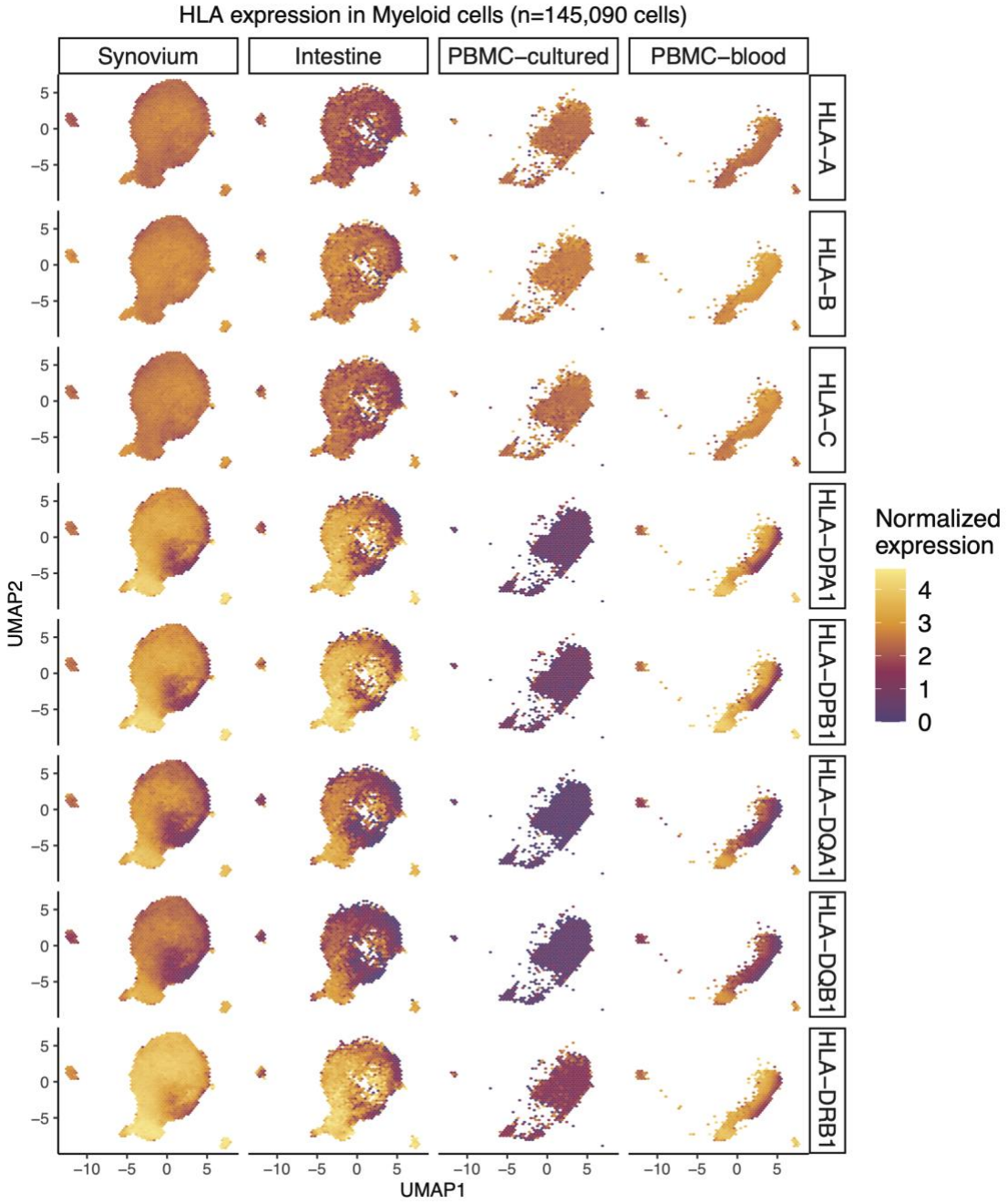
Supplementary Fig. 7. Conditional analysis using pseudobulk model in each cell type. We performed up to three additional rounds of conditional analysis to identify independent eQTLs for **(a)** myeloid, **(b)** B, and **(c)** T cells. Manhattan plots showing the distance from TSS (*x*-axis, TSS ± 2MB of each gene) versus the significance of association with gene expression (*y*-axis). Nominal Wald P-values are derived from linear regression (two-sided test). Each row represents one round of conditional analysis, and each subsequent round controls for the lead effects from the previous rounds. Blank elements in the grid indicate that no variants reach *P*-value < 5e-8.



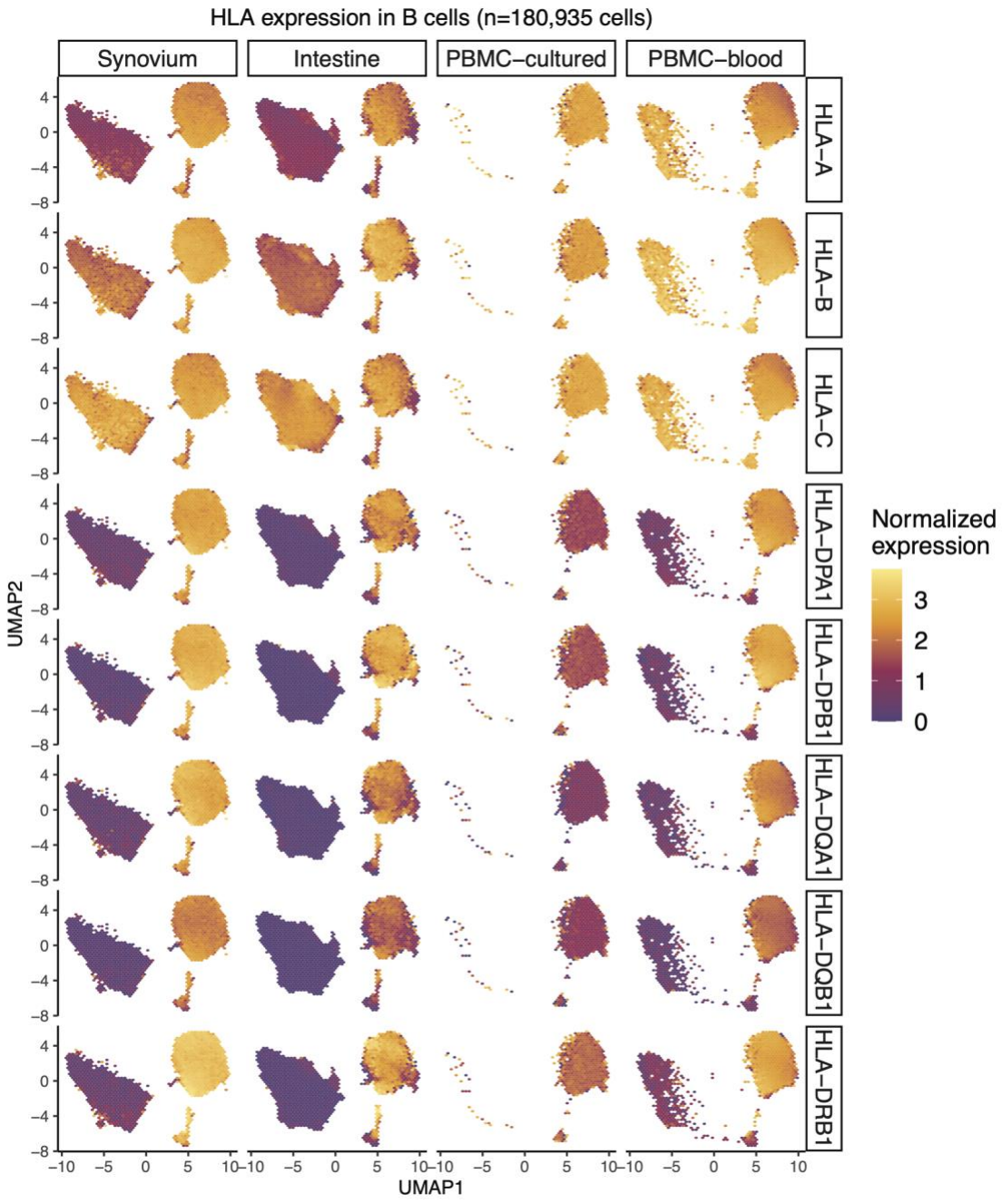
Supplementary Fig. 8. Cell-type-dependent eQTLs and LD among independent eQTLs. (a-b) Boxplots across cell types (columns) comparing the effects of B cell lead eQTLs (rows) for **(a)** *HLA-DPA1* and **(b)** *HLA-DPB1*, with myeloid n=1,025, B cells n=1,069, and T cells n=1,072 individuals across all datasets. For both examples, eQTL was weaker in myeloid and T cells. The genotype of each individual (*x*-axis) is plotted against the inverse-normal transformed residual of the gene's expression (after adjusting for covariates, *y*-axis). Nominal Wald *P*-values are derived from linear regression (two-sided test). Boxplots colored by dataset, with individual points overlaid. Boxplot center line represents median, lower/upper box limits represent 25/75% quantiles, whiskers extend to box limit ±1.5 × IQR, and outlying points are plotted individually. **(c)** Heatmaps showing LD among lead eQTLs identified in multiple rounds of conditional analysis. For each pair of cell types among B (blue), myeloid (purple), and T cells (red) (including self-pairs), plot shows the LD (r^2, color) between each pair of eQTLs. Each eQTL is labeled as HLA-X_Y, where X is the gene and Y is the round of conditional analysis (e.g., HLA-B_3 represents the tertiary eQTL for *HLA-B*). LD calculated using multi-ancestry MHC reference used for HLA imputation.



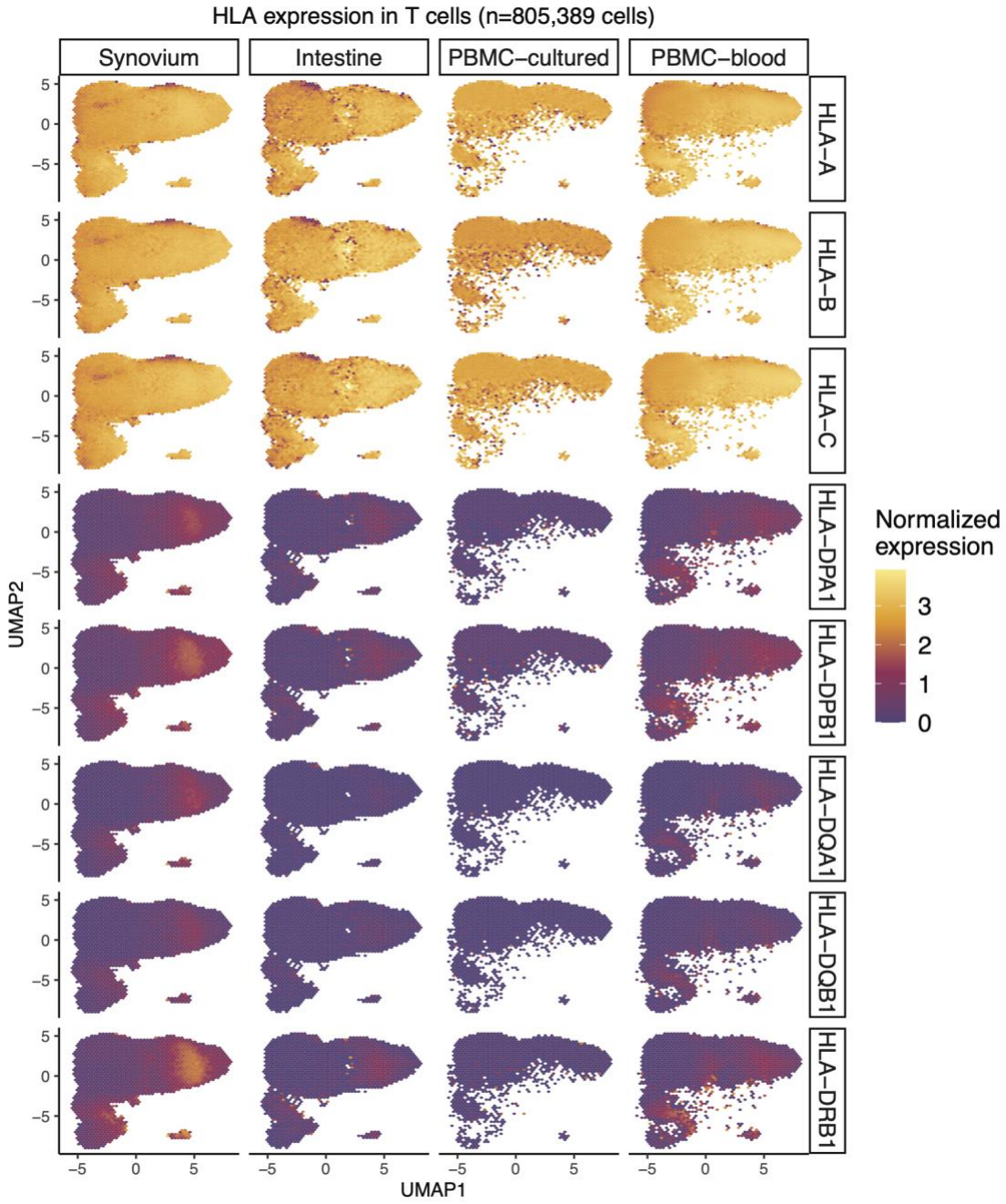
Supplementary Fig. 9. Two strategies for embedding cells from multiple datasets. (a) Schematic of *de novo* integration of all datasets using Harmony (left) versus a reference-mapping-based approach where the two solid tissue datasets were used to construct the embedding, and PBMC datasets were mapped into the same coordinate space using Symphony (right). **(b-d)** The resulting UMAP embeddings for Synovium and PBMC-blood datasets using each approach for **(b)** myeloid, **(c)** B, and **(d)** T cells, colored by merged cell state annotations.



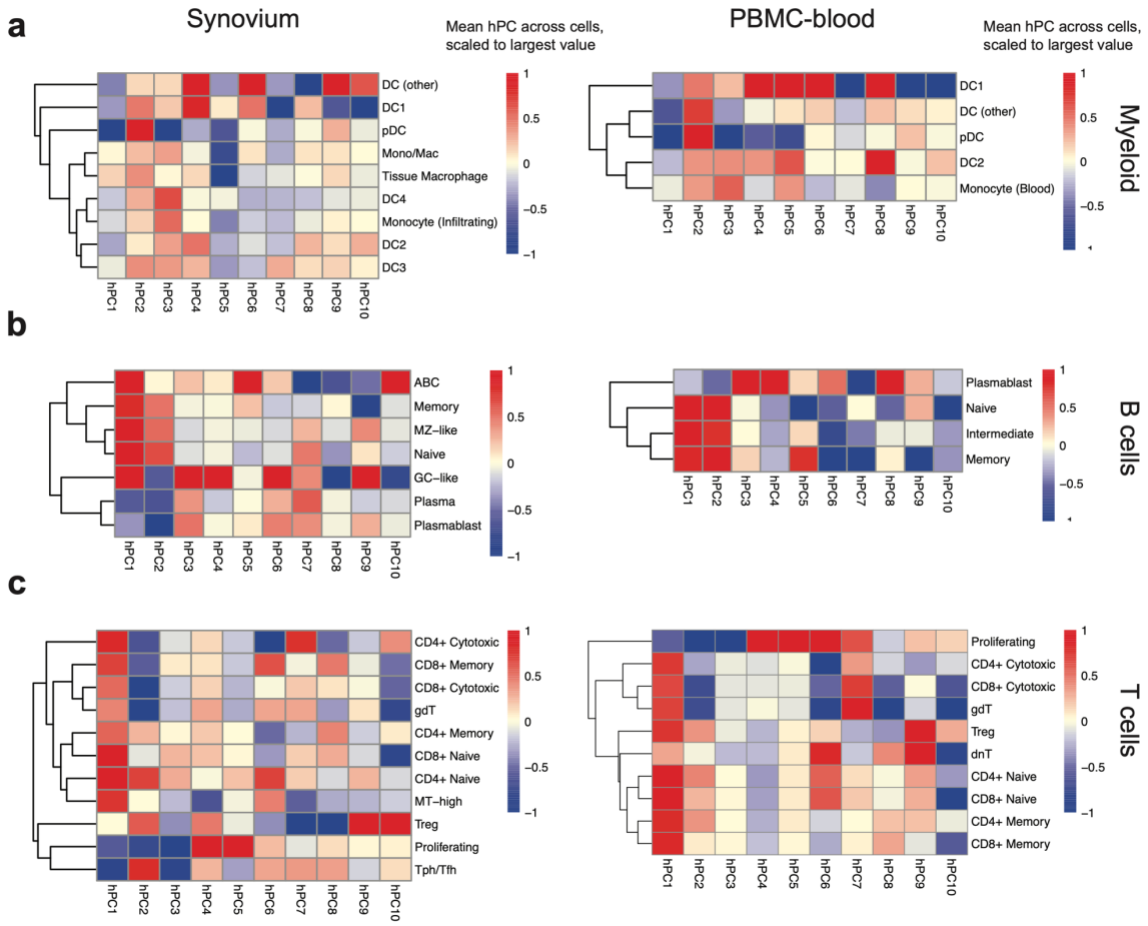
Supplementary Fig. 10. Atlas of *HLA* gene expression in myeloid cells across four datasets. Expression of eight classical *HLA* genes (rows) in myeloid cells in Synovium (n=66,789 cells), Intestine (n=14,492 cells), PBMC-cultured (n=23,241 cells), and PBMC-blood (n=40,568 cells) plotted on a hexagon-binned UMAP to address overplotting (50 bins per both horizontal and vertical directions), with each bin colored by mean log(CP10k+1)-normalized expression of the gene.



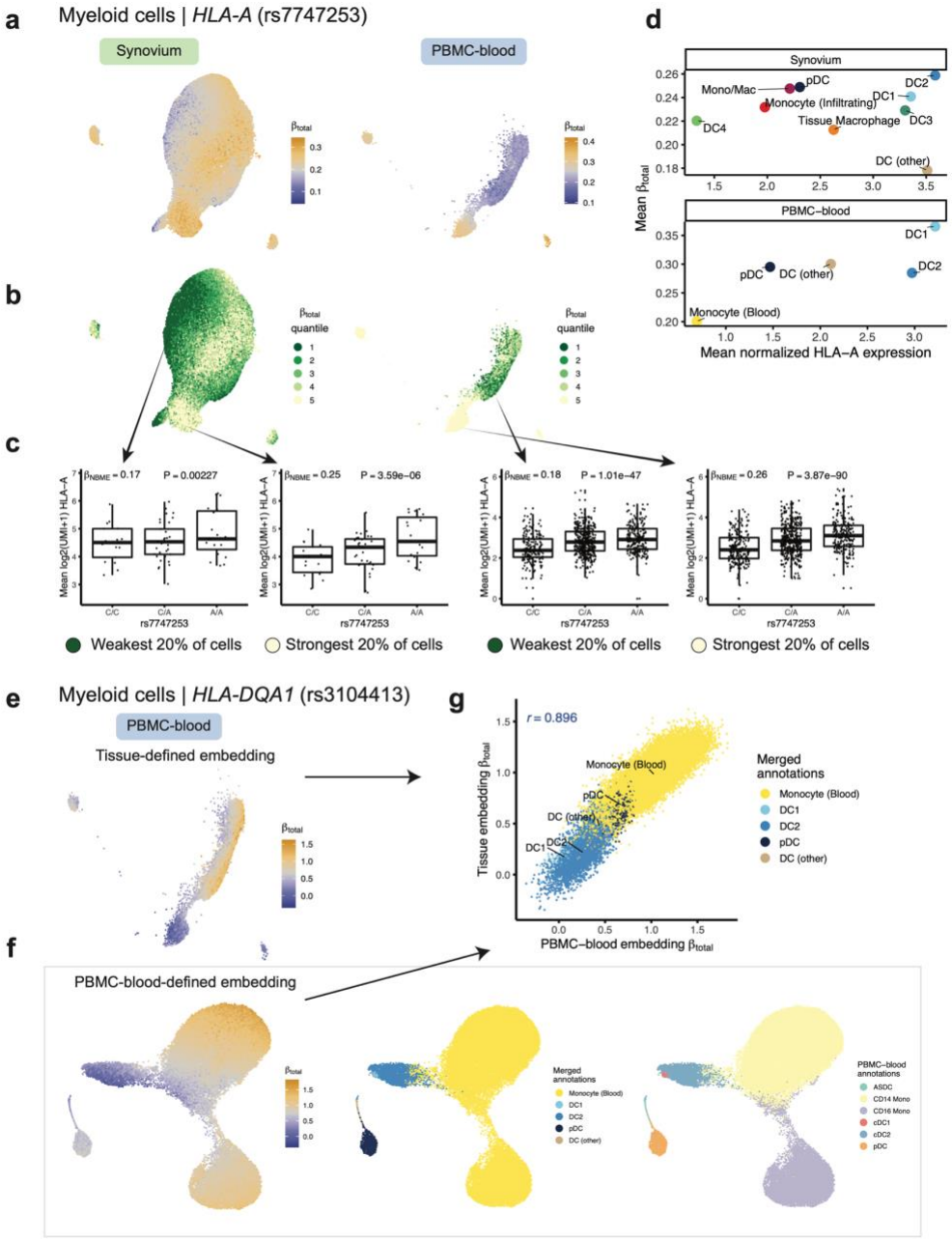
Supplementary Fig. 11. Atlas of *HLA* gene expression in B cells across four datasets. Same as **Supplementary Fig. 10** but for B cells in Synovium (n=25,917 cells), Intestine (n=56,572 cells), PBMC-cultured (n=17,662 cells), and PBMC-blood (n=80,784 cells).



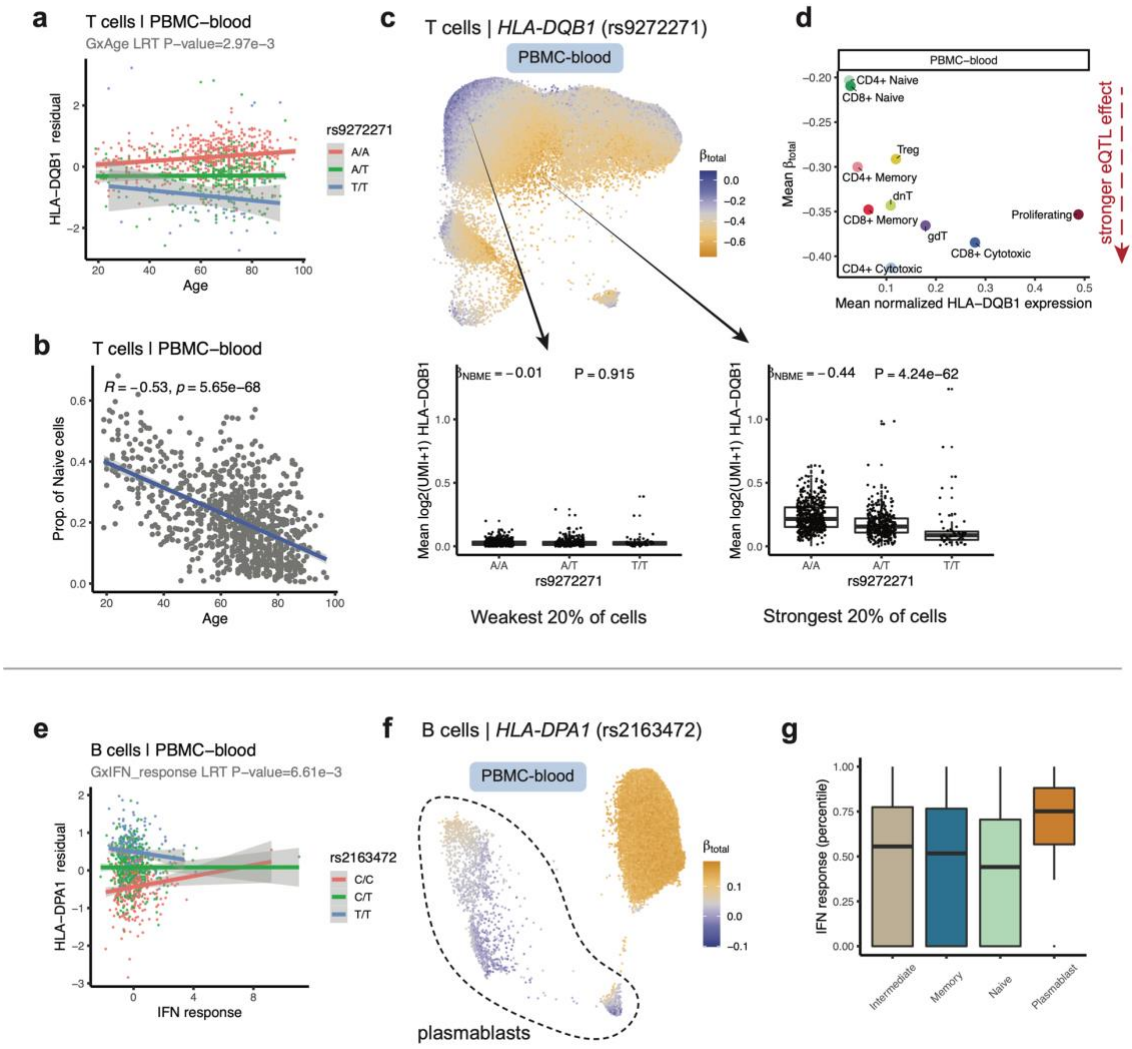
Supplementary Fig. 12. Atlas of *HLA* gene expression in T cells across four datasets. Same as **Supplementary Fig. 10** but for T cells in Synovium (n=82,423 cells), Intestine (n=47,868 cells), PBMC-cultured (n=136,519 cells), and PBMC-blood (n=538,579 cells).



Supplementary Fig. 13. Linking hPCs to annotated cell states. Heatmaps showing the mean value for each hPC across cells for each discrete cell state within each major cell type: **(a)** myeloid, **(b)** B, and **(c)** T cells. The discrete cell states were defined by standardizing the labels provided by the PBMC-blood and Synovium studies to the set of merged annotations. Mean hPC values (color) are scaled for each cell state relative to the most extreme value across cell states.



Supplementary Fig. 14. Dynamic eQTLs in myeloid cells. Lead *HLA-A* eQTL (rs7747253) in myeloid cells (n=66,789 cells in Synovium, 40,568 in PBMC-blood). **(a)** UMAP colored by per-cell estimated eQTL effect (β_{total}), from blue (weakest) to orange (strongest). **(b)** Cells colored by quintiles of β_{total}. **(c)** Boxplot showing the eQTL effect across individuals in the top and bottom quintiles of cells. Labeled β_{NBME} and *P*-value are derived from fitting the NBME model without cell state interaction terms on cells from the discrete quintile and comparing to a null model without genotype using an LRT. Mean log2(UMI+1) across cells per individual (*y*-axis) by each genotype. Boxplot center line represents median, lower/upper box limits represent 25/75% quantiles, whiskers extend to box limit ±1.5 × IQR, and outlying points are plotted individually. **(d)** Scatter plot showing the mean estimated β_{total} (*y*-axis) compared to the mean log(CP10k+1)-normalized expression of *HLA-A* (*x*-axis) across annotated cell states (color). **(e-g)** Comparing myeloid *HLA-DQA1* eQTL (rs3104413) effects in two different cell embeddings. UMAP of PBMC-blood myeloid cells (n=40,568 cells) in **(e)** tissue-defined hPCs versus **(f)** hPCs defined using PBMC-blood alone, colored by β_{total} (left), merged cell annotations (middle), and dataset annotations (right). **(g)** Concordance between per-cell β_{total} values in tissue-defined (*y*-axis) versus PBMC-blood embedding (*x*-axis); Pearson *r* is labeled. Abbreviations: LRT, likelihood ratio test (one-sided).



Supplementary Fig. 15. Age and IFN-associated eQTLs. (a-d) Age-associated eQTL. **(a)** Age-genotype interaction for *HLA-DQB1* eQTL (rs9272271) in T cells using pseudobulk model in PBMC-blood. Each dot is one individual, colored by eQTL genotype. Higher dosage of T allele is associated with lower *HLA-DQB1* expression (*y*-axis). The eQTL becomes stronger with increasing age (*x*-axis). Fitted linear regression lines by genotype shown with 95% confidence regions. **(b)** Negative relationship between age and proportion of naïve T cells per sample, with Pearson correlation (and two-sided *P*-value) labeled. **(c)** UMAP colored by per-cell estimated eQTL effect (β_{total}) from the NBME model, from blue (weakest) to orange (strongest), with boxplots showing eQTL effect across individuals in the top and bottom quintiles of cells, analogous to **Fig. 5d,e**. Note: in this example, the eQTL effect is negative as defined by the ALT allele, so more negative β_{NBME} corresponds to stronger effect. **(d)** Scatterplot showing mean β_{total} (*y*-axis) compared to the mean log(CP10k+1)-normalized expression of *HLA-DQB1* (*x*-axis) across annotated cell states (color). **(e-g)** IFN-associated eQTL. **(e)** Interaction between sample-level IFN-response (*x*-axis) and *HLA-DPA1* eQTL (rs2163472) in PBMC-blood B cells identified using pseudobulk eQTL model. Each dot is an individual (colored by eQTL genotype); eQTL becomes weaker with higher IFN response. Fitted linear regression lines by genotype shown with 95% confidence regions. **(f)** UMAP colored by β_{total} from NBME model, from weakest (blue) to strongest (orange). **(g)** Boxplots showing the distribution of IFN-response percentile across B cell states (m=20,894 intermediate, 19,224 memory, 38,414 naïve, 2,252 plasmablast cells), where IFN-response is the per-cell sum of scaled normalized expression of 11-gene IFN signature. Center line is median, lower/upper box limits are 25/75% quantiles, whiskers extend to box limit ±1.5 × IQR, and outlying points are plotted individually. For **(a)** and **(e)**, *P*-values are from LRT comparing linear regression models with and without genotype interaction.

Supplementary Table legends

See separate file, Kang_etal_SuppTables.xlsx. Legends below:

Supplementary Table 1. Datasets included in the study. Cohort characteristics include reference publication, sampled tissue, biological conditions (if any), number of individuals, number of single cells, and genetic ancestry. The numbers of individuals and cells are shown after removing individuals with uncertain HLA allele calls and low-quality cells. All PBMC-cultured individuals had samples from both conditions (treated with influenza A virus and mock conditions). Dataset technical details include type of genotype data (and # of variants in the MHC as input to HLA imputation), single-cell assay, read length(s), input format, barcode and UMI length, and whitelist used for STARsolo. Abbreviations: PBMCs, peripheral blood mononuclear cells; RA, rheumatoid arthritis; OA, osteoarthritis; UC, ulcerative colitis; HC, healthy control; IAV, influenza A virus; MHC, major histocompatibility complex; WGS, whole-genome sequencing; GSA, Global Screening Array; MEGA, Multi-ethnic Genotyping Array.

Supplementary Table 2. Sample and cell numbers before and after QC. Top section includes number of individuals per dataset before QC and during each step. Middle section includes number of cells before QC per dataset, cell counts for each major cell type after QC, sum of cells in all cell types, and sum of cells used in eQTL analysis (myeloid, B, and T cells). Bottom section shows the number of individuals used in eQTL analysis per cell type.

Supplementary Table 3. SNP2HLA imputation quality for *HLA* alleles. Mean imputation dosage R^2 (DR2) for two-field *HLA* alleles with AF > 5% (top) and >1% (bottom) for each *HLA* gene (row) across each array dataset (columns), as well as mean across datasets (rightmost column).

Supplementary Table 4. Percent change in estimated *HLA* expression after scHLApers. For each classical *HLA* gene in each dataset (rows), the mean, median, 25^{th} and 75^{th} quantile of percent change in total UMI counts (sum across all cells per individual) using scHLApers relative to a standard pipeline without personalization.

Supplementary Table 5. Dissimilarity from the reference at 3' vs. 5' end of class I genes. Comparisons were made using the 500-bp region at the 3'/5' end of the multiple sequence alignment of all alleles and the reference allele.

Supplementary Table 6. Merging cell annotations across datasets to shared labels. Mapping between the cell annotations provided by the original dataset, major cell types in this study (B, myeloid, T, NK, fibroblast, or endothelial), and merged finer-grained annotations (for B, myeloid, and T cells in PBMC-blood and Synovium datasets).

Supplementary Table 7. Characteristics of MHC variants used for eQTL testing. Information regarding the 12,050 variants across the MHC used for eQTL testing, including chromosome 6 genomic position in GRCh38 (POS), REF and ALT alleles (for one- and two-field *HLA* alleles, A denotes absent, and T denotes present), imputation quality in Synovium (DR2), MAF in each cohort, and hg19 position (hg19_POS; as output by SNP2HLA based on MHC reference). For variant names, "rs" prefix indicates variant in the MHC region (dbSNP name), and "HLA" prefix indicates classical *HLA* allele.

Supplementary Table 8. Multi-cohort pseudobulk lead eQTL results for myeloid, B, and T cells. The lead eQTLs for each *HLA* gene and cell type in the multi-cohort pseudobulk linear

model. Columns list the effect size (beta), standard error of beta estimate, nominal Wald *P*-value from linear regression (two-sided test), and REF and ALT alleles.

Supplementary Table 9. Comparison of effect sizes from multi-cohort vs. single-cohort pseudobulk eQTL models. Data for **Fig. 3d**. Columns list the lead variants from multi-cohort analysis, cell type, gene, dataset (either one of four single-dataset cohorts or the combined multi-dataset cohort), effect size of variant on covariate-corrected standardized gene expression (beta), standard error of beta estimate, nominal Wald *P*-value from linear regression (two-sided test), and REF and ALT alleles. See **Supplementary Table 7** for metadata about each tested variant.

Supplementary Table 10. Grouping of classical *HLA* alleles by lead eQTLs. For each lead eQTL variant that was not itself an *HLA* allele, we determined the co-occurrence pattern between the eQTL variant REF and ALT alleles versus two-field classical *HLA* alleles for the eQTL-associated gene. Each row corresponds to one [eQTL]-[*HLA*-allele] pair, listing the number of haplotypes in the multi-ethnic HLA reference panel with the two-field *HLA* allele and the REF eQTL allele (nHaplos_wREF), number of haplotypes with the two-field allele and ALT eQTL allele (nHaplos_wALT), and proportion of total reference haplotypes with the two-field allele (nHaplos_withAllele) with the ALT version (prop_ALT).

Supplementary Table 11. Multi-cohort cell-type-interaction analysis results. Results from testing lead *HLA* eQTLs from multi-cohort pseudobulk analysis for cell-type interaction. For the lead eQTL in each gene/cell type pair, table lists the effect size (beta), standard error, and Wald *P*-value for the cell type it was the lead eQTL for, the LRT (one-sided) *P*-value from the mixed-effects model testing for cell type interaction, and the betas and nominal Wald *P*-values (two-sided test) in each cell type (myeloid, B, and T, from the original multi-cohort pseudobulk linear regression model without cell type interaction) for comparison.

Supplementary Table 12. Proportion of gene expression variance explained by cell state. The estimated proportion of variance in each classical *HLA* gene explained by cell state (first 10 tissue-defined hPCs) in Synovium and PBMC-blood in myeloid, B, and T cells. Columns indicate estimated R^2 for the full NBME model (full_rsq), R^2 for a model without cell state (nostate_rsq), and the difference (full_rsq - nostate_rsq) representing variance explained by cell state.

Supplementary Table 13. Testing eQTLs for cell-state interaction with single-cell NBME model. Results from testing lead *HLA* eQTLs for cell-state dependence using the single-cell NBME model with cell state defined using the top 10 tissue-defined hPCs per cell type. For each gene, lead eQTL variant, dataset, and cell type (row), column E lists the significance (LRT *P*-value) of the genotype main effect as determined using a NBME model with genotype but without cell state terms (used to define 58 variant-gene pairs with robust main effects). Columns F-M show the results from the NBME model testing for cell-state interactions: the hPC with the most significant interaction with genotype ($\beta_{G\times hPC}$, max_int_term), the interaction effect size and nominal Wald *P*-value, the genotype main effect (β_G, G_main_Estimate) and its nominal Wald *P*-value, the size of the maximum interaction effect size in proportion to the genotype main effect (int_prop_main), and the significance of cell-state-dependency (LRT *P*-value and Chi-Square statistic comparing the full model for all hPCs to a null model without cell state interaction terms). Abbreviations: LRT, likelihood ratio test (one-sided).

Supplementary Table 14. Degree of cell-state-dependency by gene. The mean LRT χ^2 statistic value from testing for cell-state-dependence across all variant-dataset-cell-type tests (and number of tests performed) for each gene.

Supplementary Table 15. Testing for eQTL interaction with sample-level factors in PBMC-blood. (cols A-L) Results from testing lead HLA eQTLs for interaction with age (in years), sex (effect defined as female relative to male), and IFN response (defined using 11 gene signature from Davenport et al.) using pseudobulk model. For the lead eQTL in each gene/cell type pair, shows the age main effect (age_beta), age interaction with genotype (agexG_beta), and significance of interaction (agexG_LRT_pval), and analogous columns for sex and IFN score. Each interaction model was run separately. **(cols M-R)** Testing for eQTL interaction with cell state proportion in PBMC-blood T cells. Testing lead HLA eQTLs for interaction with proportion of cytotoxic T cells (prop_cyto) and proportion of naïve T cells (prop_naive) using pseudobulk model. Shows the cell state proportion main effect (prop_cyto_beta), interaction with genotype (prop_cytoxG_beta), and significance of interaction (prop_cytoxG_LRT_pval), and analogous columns for prop_naive. Prop_cyto and prop_naive interaction models were run separately. LRT *P*-values are obtained by comparing model with interaction term of interest to model without interaction (one-sided). Nominally significant P-values are bolded.

Supplementary Data legends

Supplementary Data 1. Multi-cohort pseudobulk eQTL full results for myeloid, B, and T cells. (See separate file Kang_etal_Data1.csv)
Results from testing each of 12,050 for association with classical *HLA* gene expression in each cell type (total 8 genes x 3 cell types x 12,050 variants = 289,200 tests) in the multi-cohort pseudobulk linear model. Columns list the variants in multi-cohort analysis, cell type, gene, effect size of variant on covariate-corrected standardized gene expression (beta), standard error of beta estimate, nominal Wald *P*-value from linear regression (two-sided test), and REF and ALT alleles. See **Supplementary Table 7** for metadata about each tested variant.

Supplementary Data 2. Multi-cohort pseudobulk conditional analysis results. (See separate file Kang_etal_Data2.csv)
Results from conditional analysis identifying eQTLs, conditioning on the lead variant(s) from previous round(s). Columns list the variant, cell type, gene, round of conditional analysis (conditional_iter, ranging from 1 to 4 for primary to quaternary effects), effect size of eQTL (beta), standard error of beta estimate, and nominal Wald *P*-value from linear regression (two-sided test). Includes only variants with nominal $P < 0.05$ to reduce file size. See **Supplementary Table 7** for metadata about each tested variant.

Supplementary References

1. Hughes, A. L. Origin and evolution of HLA class I pseudogenes. *Mol. Biol. Evol.* **12**, 247–258 (1995).
2. Aguiar, V. R. C., César, J., Delaneau, O., Dermitzakis, E. T. & Meyer, D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLoS Genet.* **15**, e1008091 (2019).
3. Orenbuch, R. *et al.* arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* **36**, 33–40 (2020).
4. Solomon, B. D. *et al.* Prediction of HLA genotypes from single-cell transcriptome data. *bioRxiv* (2022) doi:10.1101/2022.06.09.495569.
5. Zhang, F. *et al.* Cellular deconstruction of inflamed synovium defines diverse inflammatory phenotypes in rheumatoid arthritis. *bioRxiv* (2022) doi:10.1101/2022.02.25.481990.
6. Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-730.e22 (2019).

7. Randolph, H. E. *et al.* Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science* **374**, 1127–1133 (2021).
8. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).

Supplementary Note 2: More Stringent Quality Control of OneK1K Cohort Single-cell RNA-seq Dataset

Laurie Rumker (Laurie_Rumker@hms.harvard.edu)
Joyce B. Kang (Joyce_Kang@hms.harvard.edu)
Soumya Raychaudhuri (soumya@broadinstitute.org)

August 25, 2023

1 Motivation

In collaboration with the authors of the OneK1K dataset index publication [1], we applied more stringent quality control to these PBMC scRNA-seq profiles before employing the dataset in our own analyses. Our additional dataset processing, summarized in this document, was prompted by our observations of isolated populations with mixed type assignments that expressed unexpected marker genes. We initially observed these putative doublet populations when performing a standard PCA-based analysis on each major cell type separately (e.g. PCA on cells labeled B cells). We observed fragmented cell populations with mixed type assignments (e.g. mixed B naive and B memory labels in a population separate from the major populations for these B cell subtypes) that also contained expression of unexpected marker genes that did not match the assigned labels (e.g. CD3 among these "B cells"). We we found that these populations corresponded to droplets identified as doublets by Demuxlet [2] or Scrublet [3] but not previously removed from the dataset.

2 Approach

We received scRNA-seq profiling (cells-by-counts matrix), as well as Demuxlet and Scrublet method output, directly from the study authors. Cell type assignments provided by the study authors were based on Azimuth mapping to a PBMC reference dataset [4]. After affirming basic per-profile QC thresholds were met (>200 genes, <8% mitochondrial gene reads), and removing 7680 genes that appeared in fewer than three profiles, we subdivided the profiles by major cell type using the following mapping from the 31 available labels to 7 major types:

- CD4+ T = [CD4 TCM, CD4 Naive, CD4 TEM, Treg, CD4 CTL, CD4 Proliferating]
- Other T = [CD8 TEM, CD8 Naive, CD8 TCM, MAIT, CD8 Proliferating, gdT, dnT]
- NK = [NK, NK_CD56bright, NK Proliferating, ILC]
- Monocyte = [CD14 Mono, CD16 Mono]
- DC = [cDC1, cDC2, pDC, ASDC]
- B = [B naive, B memory, B intermediate, Plasmablast]
- Other = [HSPC, Platelet, Eryth]

For each major cell type, we followed standard processing using scanpy (with parameters as described in the "Preprocessing and clustering 3k PBMCs" tutorial unless otherwise specified [5]) to total-count normalize to 10,000 reads per profile, logarithmize the data, retain only highly-variable genes and compute principal components (PCs). For each major cell type, we corrected these PCs for batch with harmony (batch = "pool", nclust = 50, sigma = 0.2, max_iter_harmony = 50) to generate hPCs. Resuming the scanpy pipeline, we used these hPCs to construct a nearest-neighbor graph and UMAP embedding per major cell type.

The index publication authors had previously removed any droplet identified as a doublet by both Scrublet and Demuxlet, but retained all droplets identified as doublets by only one of these two

methods. Of the 1,249,037 profiles provided by the OneK1K dataset authors, 22,662 were identified as doublets by Scrublet (predicted_doublet_mask==True) and 382,464 were called as doublets by Demuxlet ('BEST' assignment to 'DBL-'). We chose to remove these profiles. None of the cells included in the published dataset provided by the original authors had been classified by Demuxlet as ambiguous. Given that many profiles identified as doublets by Scrublet or Demuxlet were observed to cluster together transcriptionally in the dataset (Figures 1, 2, 3, 4, 5, 6, 7), we performed fine-grained clustering within each major cell type and removed any clusters for which >2/3 profiles were identified as doublets (by either Demuxlet or Scrublet).

We used Wilcoxon rank-sum tests to identify differentially-expressed genes per fine-grained cluster (scanpy's rank_gene_groups function with method = 'wilcoxon'). For major cell type groups besides "Other"—which contains the profiles assigned by Azimuth to the Platelet type—we also removed fine-grained clusters for which differential expression analysis identified PPBP, PF4, GP1BB and NRGN among the top 6 cluster-characteristic markers, suggestive of platelet doublets.

Finally, 1803 profiles lacked results from Demuxlet and Scrublet. Of these profiles, 131 were labeled "Doublet" by the publication authors and the remainder corresponded to an individual who also failed genotype data quality control in our analyses (not described here). We removed these 1083 profiles.

In summary, we removed each profile if:

- The profile was identified as a doublet by either Demuxlet or Scrublet OR
- The profile was assigned to a doublet-dominated fine-grained cluster OR
- The profile was labeled as a non-platelet type but assigned to a fine-grained cluster characterized by platelet-related genes OR
- The profile lacked doublet-calling results

Finally, we reassigned cell type labels to our retained cells, applying the same approach used by the publication authors for the initially-provided cell type labels: Azimuth reference mapping to the Azimuth PBMC reference. To accommodate Azimuth data volume limitations, we split the total dataset into 15 subsets by batch pool group and applied Azimuth separately to each subset. The major cell type classifications for the retained cells (i.e. among T, B, NK, and Myeloid groups) were unchanged for the vast majority of cells when compared to each cell's original major type assignment.

3 Results

Of the 1,249,037 profiles provided by the study authors from the published dataset, we chose to remove 416,556 (33%), the vast majority of which (405,126 profiles, 97%) were identified as doublets by either Scrublet or Demuxlet, and the remainder selected using our other two criteria (Tables 1, 2).

We found that the droplets identified as doublets by Scrublet or Demuxlet largely explained the isolated cell populations with mixed assigned types that we had observed, and platelet-contaminated populations explained some remaining fragmented populations (Figures 1, 2, 3, 4, 5, 6, 7, 8).

Major Type	Profiles	Resolution	Demuxlet	Scrublet	Fraction Removed
DC	6648	1.0	0.2	0.04	0.28
Mono	51876	2.0	0.22	0.02	0.25
B	129588	3.0	0.29	0.02	0.32
NK	172397	4.0	0.29	0.02	0.33
CD4+ T	624592	6.0	0.32	0.01	0.34
Other T	259893	6.0	0.31	0.03	0.34
Other	3912	0.2	0.44	0.04	0.61

Table 1: Profiles selected for removal, by major type. For each major type group, the total number of profiles assigned to that group ("Profiles") is shown, along with the resolution used for fine-grained clustering ("Resolution"), the fraction of all profiles identified as doublets by Demuxlet or Scrublet ("Demuxlet" and "Scrublet", respectively), and the fraction selected for removal based on all criteria ("Fraction Removed")

Removed	Scrublet Dblt.	Demuxlet Dblt.	Platelet Clust.	Dblt. Clust.	Count
F	F	F	F	F	832481
T	F	F	F	T	7734
T	F	F	T	F	1893
T	F	T	F	F	358161
T	F	T	F	T	23198
T	F	T	T	F	1105
T	T	F	F	F	18602
T	T	F	F	T	4033
T	T	F	T	F	27
T	NA	NA	NA	NA	1803

Table 2: Profiles selected for removal, by criterion. "Removed": T if the profile was selected for removal. "Scrublet Dblt.": T if Scrublet identified the profile as a doublet. "Demuxlet Dblt.": T if Demuxlet identified the profile as a doublet. "Platelet Clust.": T if the profile was assigned to a fine-grained cluster characterized by platelet-related genes. "Dblt. Clust.": T if the profile was assigned to a fine-grained cluster with <2/3 doublets. "Count": The number of profiles matching the combination of features captured by the corresponding row. The authors had previously removed profiles called as doublets by both Scrublet and Demuxlet. The final row captures profiles for which doublet-calling results were not available.

We make available a table containing the results of this data processing. This table contains one row per cell, indexed by barcode. In addition to cell-level metadata provided in the published dataset, we have added the following columns:

- demuxlet_DBL: True iff the cell was assigned as a doublet by Demuxlet
- demuxlet_AMB: True iff the cell was assigned as ambiguous by Demuxlet
- scrublet_DBL: True iff the cell was assigned as a doublet by Scrublet
- scrublet_score: Score assigned by Scrublet
- preQC_Azimuth_type: Azimuth-based cell types shared by the publication authors
- DBL_cluster: True iff the cell belonged to a cluster with >2/3 cells assigned as doublets by Scrublet or Demuxlet
- Platelet_cluster: True iff the cell belonged to a cluster characterized by platelet-associated marker genes
- remove_cellQC: True iff the cell met one of the four critera for removal described here
- remove_sampleQC: True iff the cell was associated with a sample we removed for our analyses (samples with low-quality or missing genotyping data, or labeled as ethnic outliers)
- fail_QC: True iff remove_cellQC or remove_sampleQC is True
- celltype: Azimuth cell type assignments for retained cells
- majortype: Major cell type assignments, aggregated from celltype

4 Discussion

Identification and removal of doublet droplets is a crucial quality control step in single-cell data analysis. Scrublet and Demuxlet are two of many available methods to accomplish this task. Scrublet simulates doublet transcriptional profiles as combinations of observed profiles and compares the observed profiles to these simulates. Demuxlet identifies droplets whose transcripts reflect a combination of genetic variants unlikely to arise from a single individual in the dataset. Because these methods have contrasting failure modes, applying both to the same dataset can enable the detection of droplets

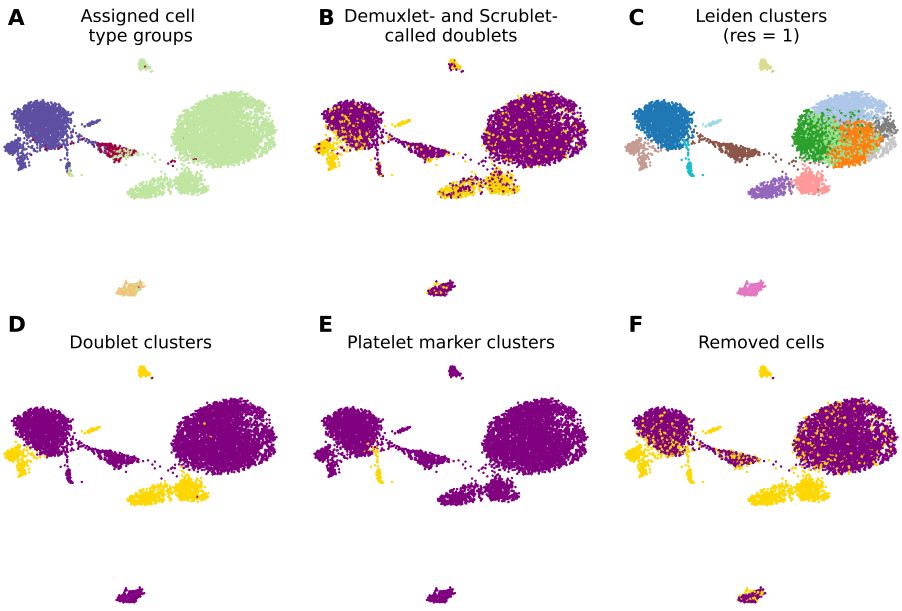


Figure 1: **Dendritic cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

by one method that were missed by the other. In the original publication of the OneK1K dataset, only cells called as doublets on the basis of both Scrublet and Demuxlet were removed, a small fraction of all identified doublets. Of the retained profiles identified as doublets, the vast majority were flagged by Demuxlet (i.e. on the basis of contrasting genotypes detected in the same droplet). We found that the retained doublets were transcriptionally perturbed in the dataset relative to cells identified as singlets and have chosen a more stringent quality control approach to remove these cells. In collaboration with the OneK1K dataset authors, we make available a table indicating which cells were selected for removal in our more stringent quality control.

References

[1] Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R. P. Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L. Cook, Chun Jimmie Ye, Kirsten A. Fairfax, Alex W. Hewitt, and Joseph E. Powell. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, April 2022.

[2] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M. Lanata, Rachel E. Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A. Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, January 2018. Number: 1 Publisher: Nature Publishing Group.

[3] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, 8(4):281–291.e9, April 2019.

[4] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael

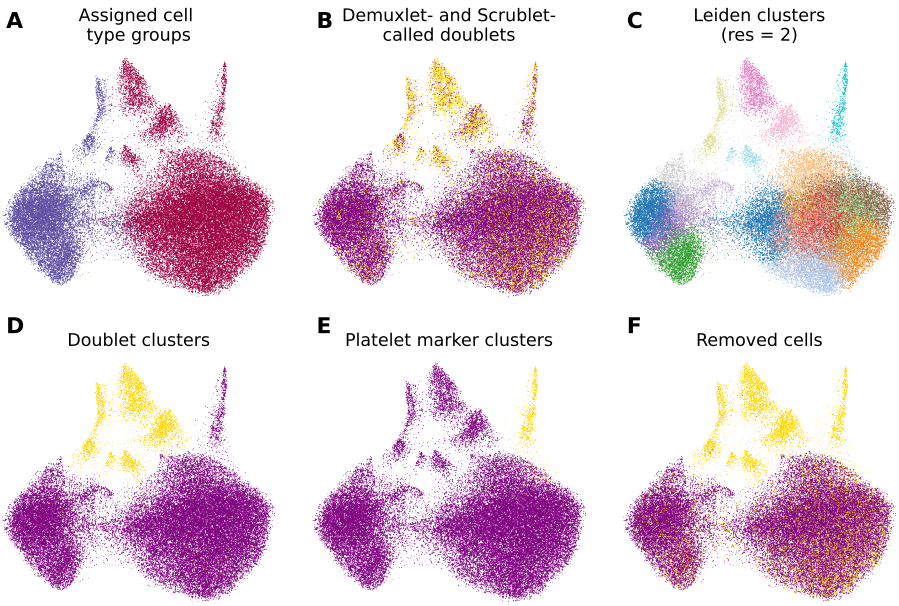


Figure 2: **Monocytes.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

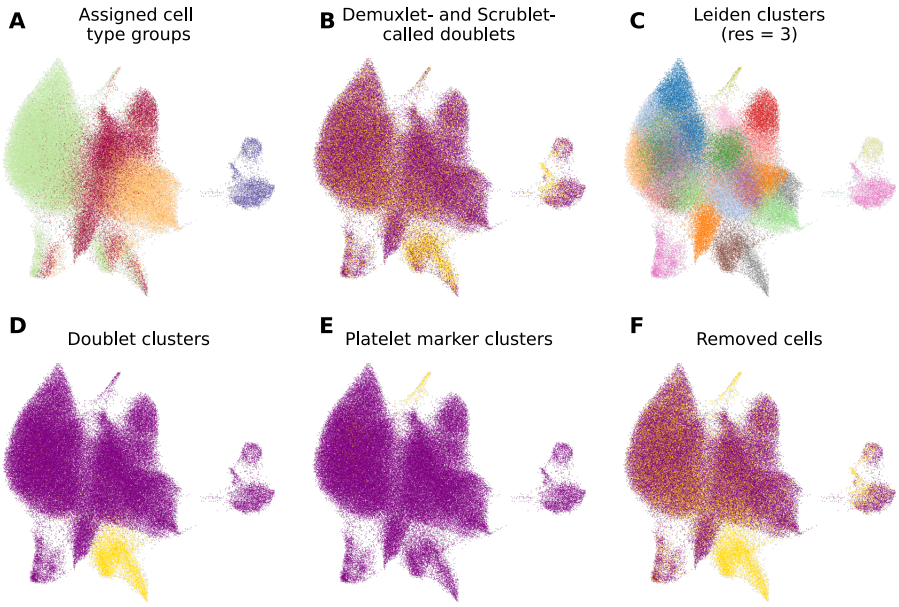


Figure 3: **B cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

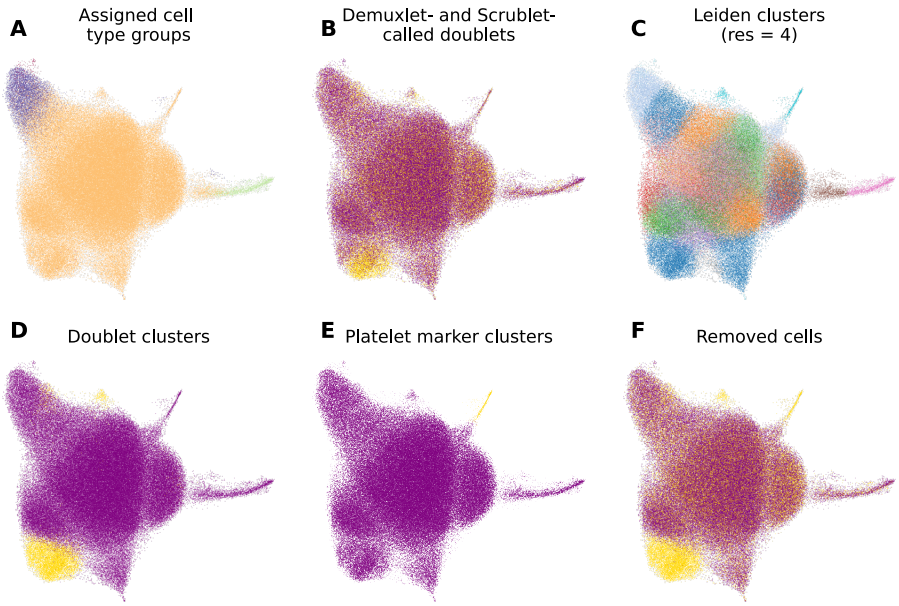


Figure 4: **NK cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

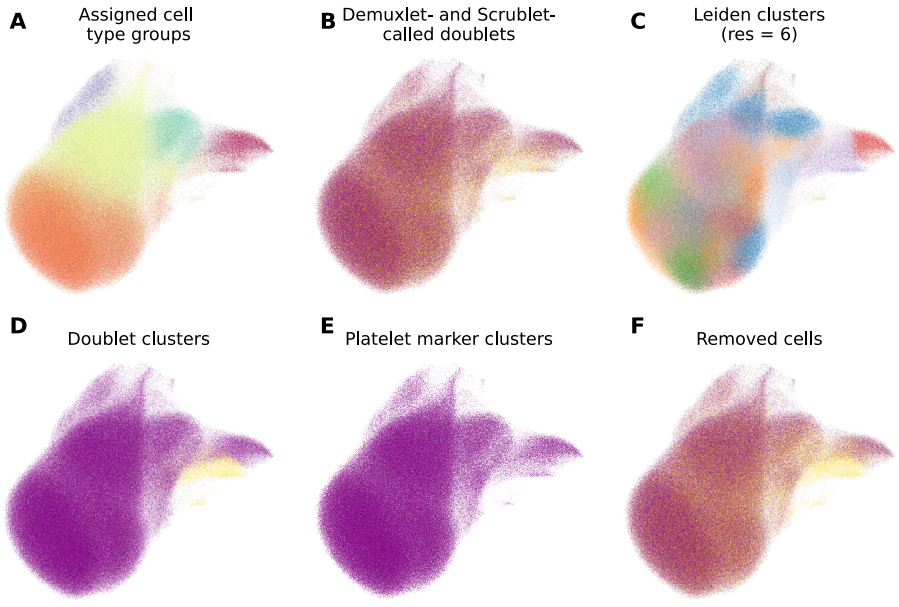


Figure 5: **CD4+ T cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

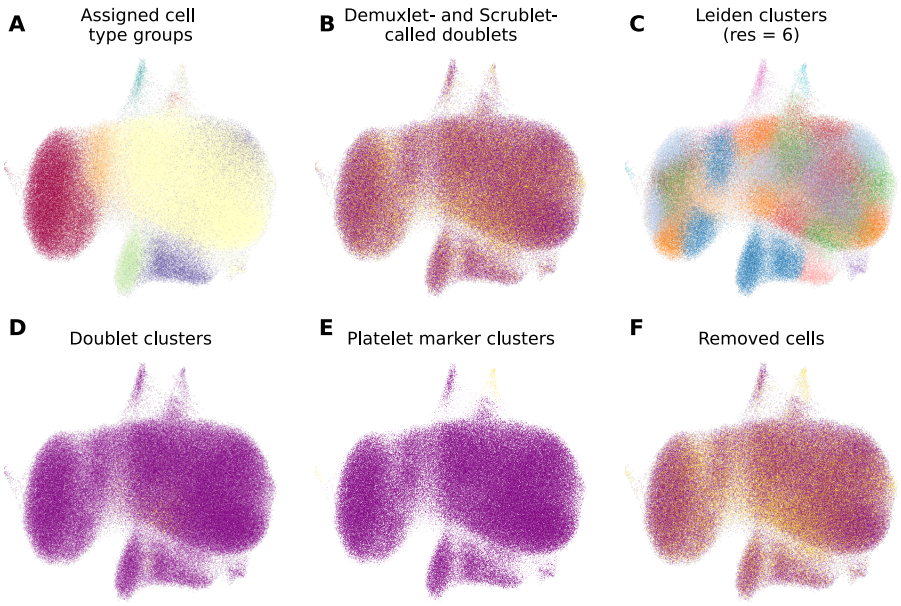


Figure 6: **Other T cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

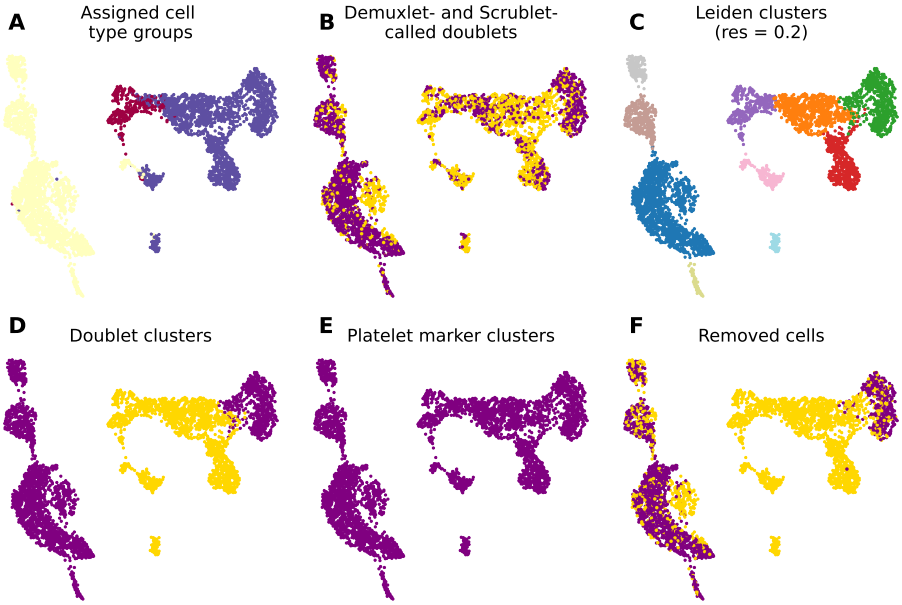


Figure 7: **All other cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

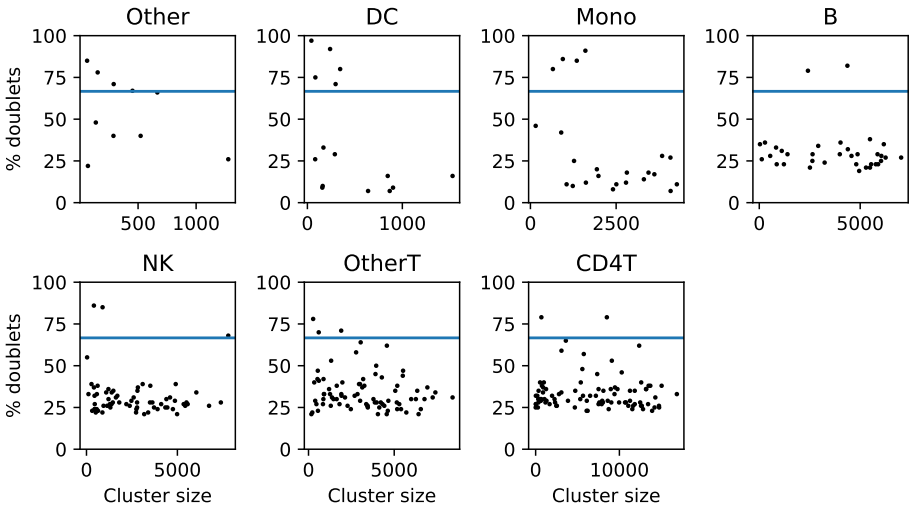


Figure 8: **Doublet cluster identification.** The profiles within each major type were clustered at a fine-grain resolution to identify and remove doublet-predominant clusters, in addition to isolated doublet profiles. The fraction of profiles called as doublets (by Scrublet or Demuxlet) for each fine-grained cluster is shown along the y axis, while the size of each cluster (number of profiles) is shown along the x axis, with plots separated by major type. Clusters with >2/3 doublets, above the blue line, were selected for removal.

Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021. Publisher: Elsevier.

[5] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, May 2015.