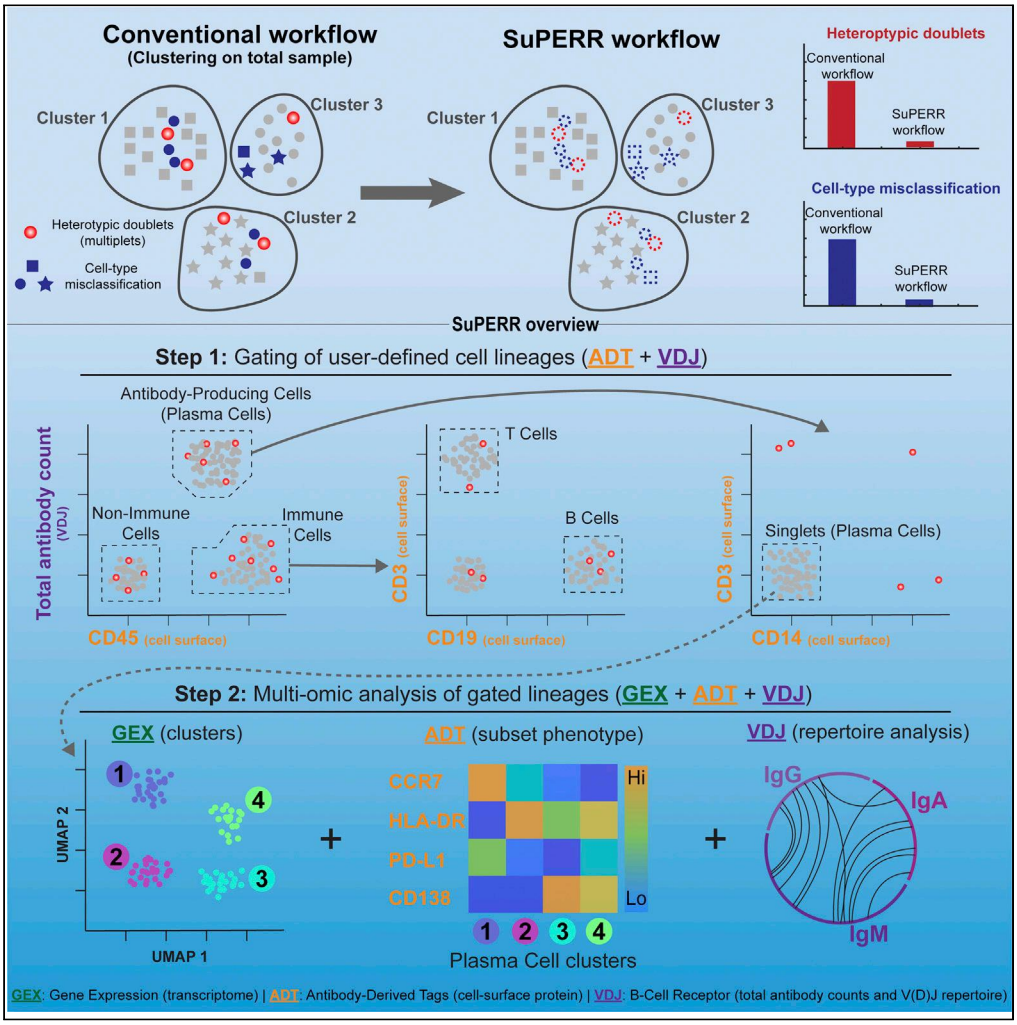
iScience

CellPress
OPEN ACCESS

Article

Comprehensive multi-omics single-cell data integration reveals greater heterogeneity in the human immune system



Congmin Xu, Junkai Yang, Astrid Kosters, Benjamin R. Babcock, Peng Qiu, Eliver E.B. Ghosn

peng.qiu@bme.gatech.edu (P.Q.)
eliver.ghosn@emory.edu (E.E.B.G.)

Highlights

SuPERR removes heterotypic doublets and cell-type misclassifications in scRNA-seq

Sequential gating on cell-surface proteins resolves major cell lineages in scRNA-seq

Defining major cell lineages before clustering reduces cell-type misclassifications

Antibody counts from single-cell V(D)J matrix accurately identify plasma cells

Xu et al., iScience 25, 105123
October 21, 2022
https://doi.org/10.1016/j.isci.2022.105123

iScience

CellPress
OPEN ACCESS

Article

Comprehensive multi-omics single-cell data integration reveals greater heterogeneity in the human immune system

Congmin Xu,[1,4] Junkai Yang,[2,4] Astrid Kosters,[2] Benjamin R. Babcock,[2] Peng Qiu,[1,*] and Eliver E.B. Ghosn[2,3,5,*]

SUMMARY

Single-cell transcriptomics enables the definition of diverse human immune cell types across multiple tissues and disease contexts. Further deeper biological understanding requires comprehensive integration of multiple single-cell omics (transcriptomic, proteomic, and cell-receptor repertoire). To improve the identification of diverse cell types and the accuracy of cell-type classification in multi-omics single-cell datasets, we developed SuPERR, a novel analysis workflow to increase the resolution and accuracy of clustering and allow for the discovery of previously hidden cell subsets. In addition, SuPERR accurately removes cell doublets and prevents widespread cell-type misclassification by incorporating information from cell-surface proteins and immunoglobulin transcript counts. This approach uniquely improves the identification of heterogeneous cell types and states in the human immune system, including rare subsets of antibody-secreting cells in the bone marrow.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) technologies have rapidly advanced in the last decade, including advances to cell-capture approaches (Evan et al., 2015; Klein et al., 2015; Utada et al., 2007), library preparation (Picelli et al., 2013; Hashimshony et al., 2012), and sequencing methods (Evan et al., 2015; Picelli et al., 2013; Habib et al., 2017; Stoeckius et al., 2017). These increasingly more widely adopted technologies have significantly improved the understanding of cell heterogeneity in health and disease (Hashimshony et al., 2012; Zheng et al., 2017; Habib et al., 2017; Stoeckius et al., 2017; Picelli et al., 2013). However, reliance on cellular transcriptomics alone limits the comprehensive identification of heterogenous cell populations (Liu and Trapnell 2016). This limitation has propelled the development of multi-omics single-cell sequencing technologies to increase the resolution and accuracy for cell subset classification.

Multi-omics single-cell sequencing technologies, such as CITE-seq (Stoeckius et al., 2017), REAP-seq (Peterson et al., 2017), and others (Lee et al. 2020), simultaneously measure gene expression (mRNA) and cell-surface proteins. Additional heterogeneity of immune cell subsets can be revealed by combining single-cell gene expression with simultaneous T- and B-cell receptor (TCR and BCR) repertoire sequencing using techniques such as RAGE-seq and DART-seq (Meyer 2019; Singh et al., 2019; Horns et al. 2020; Zemmour et al., 2018; Yermanos et al., 2021). Thus, simultaneous measurement and comprehensive integration of transcriptomics, cell-surface protein, and cell-receptor repertoire can reveal heterogeneous cell types relevant to disease mechanisms and homeostasis.

However, multi-omics technologies also present computational challenges for data integration and analysis (Colomé-Tatché and Theis 2018; Luecken and Theis 2019; Stuart and Satija 2019). Challenges include high dimensionality of the data (Yu and Lin 2016), sparsity of the data (Qiu 2020), diversity across various omics data types (Hao et al., 2021), and technical effects between different sample batches (Stuart et al., 2019). Several algorithms have been developed to integrate and analyze multi-omics measurements, including weighted nearest neighbor (WNN) implemented in Seurat v4 (Hao et al., 2021), similarity network fusion (SNF) in CiteFuse (Kim et al., 2020), among others (Wang et al., 2020; Gayoso et al., 2021; Jin et al. 2020; Argelaguet et al., 2018). The commonality of these methods is to utilize the shared signals among different omics data types to align their distributions and achieve integration, which is an unsupervised data-driven approach. Although unsupervised data-driven methods have been successful for clustering and identifying cell types, significant improvements can be made by incorporating robust prior knowledge

[1]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA

[2]Department of Medicine, Division of Immunology, Lowance Center for Human Immunology, Emory University School of Medicine, Atlanta, GA 30322, USA

[3]Emory Vaccine Center, Yerkes National Primate Research Center, Emory University School of Medicine, Atlanta, GA 30322, USA

[4]These authors contributed equally

[5]Lead contact

*Correspondence: peng.qiu@bme.gatech.edu (P.Q.), eliver.ghosn@emory.edu (E.E.B.G.)

https://doi.org/10.1016/j.isci.2022.105123

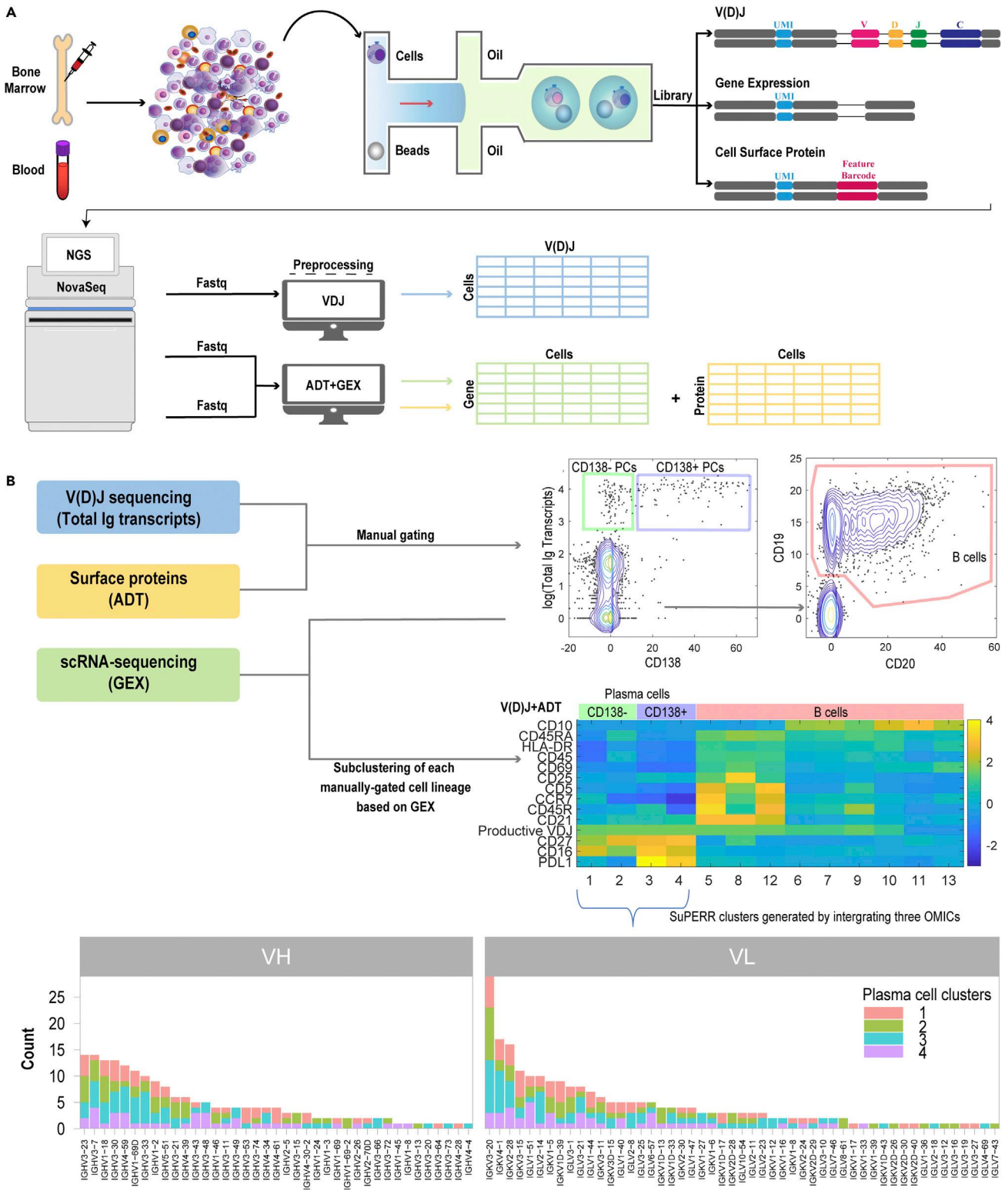


Figure 1. SuPERR workflow

(A) Schematic overview of the experimental design. Peripheral blood and bone marrow aspirates were processed, surface-stained with barcoded antibodies, and then encapsulated with barcoded microspheres. We generated three libraries for each sample corresponding to gene expression (GEX), cell-surface

Figure 1. *Continued*
protein/antibody-derived tags (ADT), and cell-receptor repertoire (VDJ). Libraries were sequenced to a target depth, and count matrices were assembled for each-omic data separately.
(B) SuPERR workflow is composed of two main steps. Major cell lineages are manually gated at the first step by integrating information from both the ADT and V(D)J data matrices. Then, the manually-gated cell lineages are further sub-clustered based on information from the GEX data. The V(D)J matrix can be used to further identify the diversity of heavy (VH) and light (VL) variable genes among the plasma cell clusters. PCs: plasma cells. See also Tables S1 and S2.

such as well-established marker genes and cell-surface protein markers that can accurately define cell types (Aran et al., 2019; Mahnke et al. 2010).

Here, to address the challenges of multi-omics analysis, we combined our extensive expertise on high-dimensional flow cytometry data analysis (Meehan et al., 2019) with our multi-omics single-cell data sets to develop the SuPERR (Surface Protein Expression, mRNA and Repertoire) workflow. SuPERR is a novel, semi-supervised, biologically-motivated approach towards the integration and analysis of multi-omics single-cell data matrices. By combining a robust prior knowledge of flow cytometry-based cell-surface markers (gating strategy) (Mahnke et al. 2010) with the high-dimensional analysis of scRNA-seq, SuPERR increases the resolution and accuracy in clustering algorithms and allows the discovery of new biologically relevant cell subsets. We first applied the flow cytometry-based "gating strategy" on a combination of cell-surface markers and immunoglobulin-specific transcript counts to identify major immune cell lineages. Next, we explored the gene expression matrix following this gating strategy to resolve cell subsets within each major immune lineage. The inclusion of this atypical "gating strategy" step also allows for cell-doublet discrimination and dramatically enhances lineage-specific variation, which helps better capture biological signals *within* each cell lineage. Finally, we apply the SuPERR workflow to human blood and bone marrow cells and directly compare its performance to existing methods. We demonstrate that SuPERR can leverage the power of each "omics" to identify major immune lineages more accurately and reveal biologically-meaningful heterogeneity within each lineage that can be confirmed by flow cytometry, facilitating the discovery of novel immune cell types.

RESULTS

Cell-surface proteins and immunoglobulin transcript counts identify major immune lineages in human blood and bone marrow

We generated multi-omics single-cell datasets from 12,759 human peripheral blood mononuclear cells (PBMC) and 7,426 human bone marrow (BM) cells from five healthy adult donors. As shown in Figure 1A, we simultaneously captured the following three omics: Total gene expression (GEX), 32 cell-surface proteins/antibody-derived tags (ADT) (Table S1), and B-cell receptor (BCR) heavy and light chain V(D)J repertoire (VDJ). The SuPERR analysis workflow can be described in two major steps shown in Figure 1B: (1) a manual biaxial gating based on the expression levels of well-stablished (Mahnke et al. 2010) cell-surface proteins (ADT) and total immunoglobulin (Ig) transcript counts to accurately identify antibody-secreting cells (ASCs), and (2) a subsequent sub-clustering of each manually-gated lineage/population identified in step 1, using the GEX matrix.

For the first step of the SuPERR workflow, we normalized the cell-surface protein (ADT) data using the DSB normalization method (Mulè et al. 2021). Next, we concatenated the normalized ADT matrix with the total Ig-specific unique molecular identifier (UMI) counts from the V(D)J matrix, which describes the total number of immunoglobulin-derived transcripts per cell. The integrated ADT/Ig matrix was used to identify major immune cell lineages before assessing their gene expression profile. Major immune cell lineages were identified and classified using a well-established sequential gating strategy on biaxial plots (Figures 2 and 3A) widely used in conventional flow cytometry data analysis and readily available through the Optimized Multicolor Immunofluorescence Panel (OMIP) publications (Mahnke et al. 2010). Because ASCs, also known as plasma cells, produce and secrete large quantities of immunoglobulin, they could be accurately identified based on their Ig-specific transcript counts (Figures 2 and 3A). Of note, the semi-supervised SuPERR workflow was able to readily identify a rare cell cluster containing as few as eight plasma cells in the human PBMC sample. As we show below, such a rare population of plasma cells cannot be identified using current conventional unsupervised workflows.

In the analysis of PBMC samples, we defined gates for six major immune cell lineages based on well-established markers (see Table 1): plasma cells, B cells, NK/NKT/MAIT/γδT cells, Monocytes, $CD4^+$T cells, and $CD8^+$T cells (Figure 2A, black borders). In the analysis of the BM samples, we defined gates for five major lineages (see Table 2): CD138− plasma cells, CD138+ plasma cells, B cells, myeloid cells, and

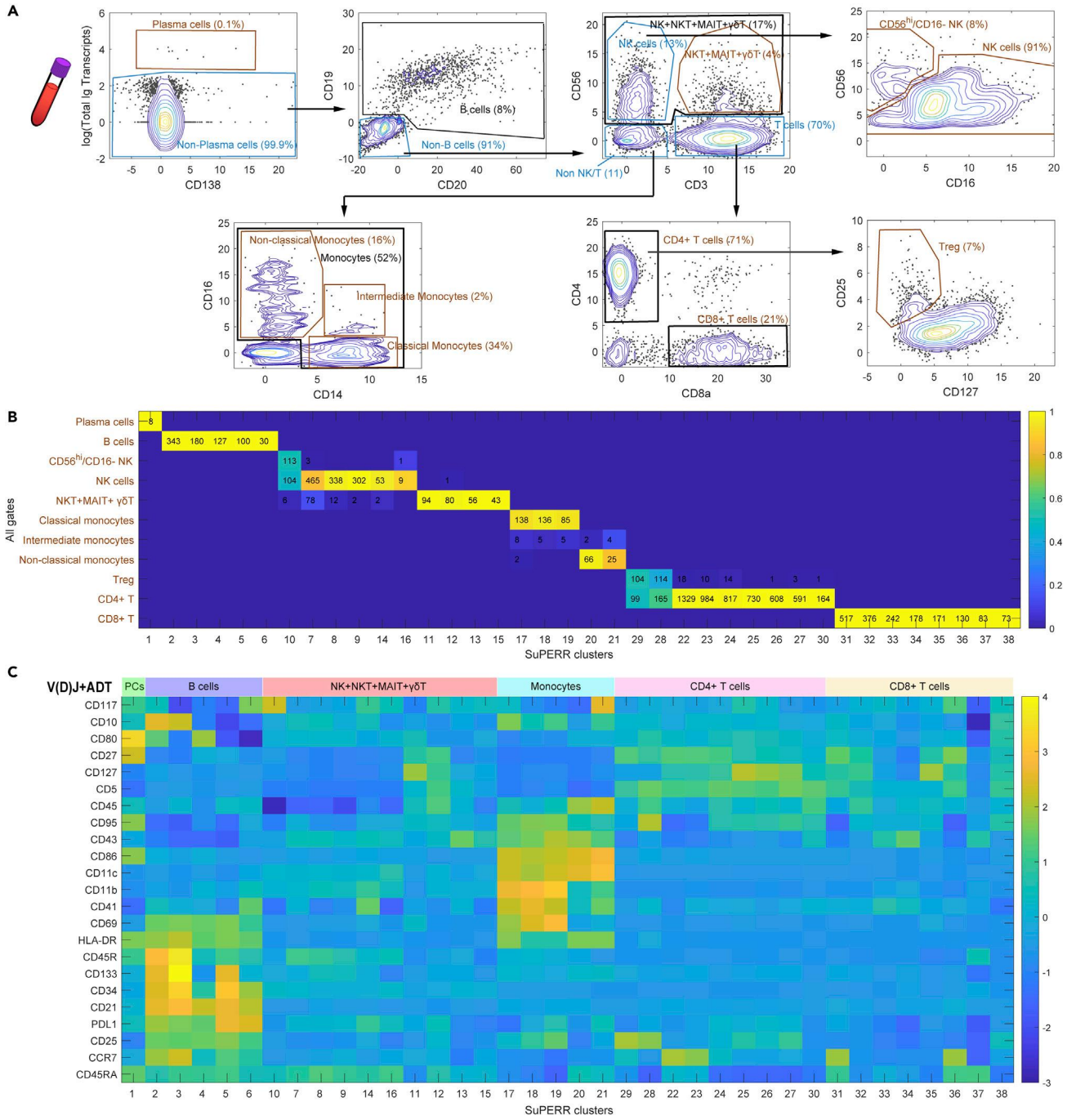


Figure 2. SuPERR workflow applied to peripheral blood mononuclear cells (PBMCs)

(A) "Gating strategy" approach to identify major cell lineages on biaxial plots based on surface markers (ADT) and V(D)J data. Total Ig transcript: sum of Ig UMIs in the VDJ matrix. Gates for major lineages are indicated as black outlines and black text. Gates for downstream cell-identity validation are indicated as golden outlines and golden text.

(B) Cross comparison between the manually-gated major lineages and the final SuPERR clusters.

(C) The average expression levels of surface markers (ADT) and VDJ features for the final SuPERR clusters. Only the ADTs/VDJ features that were not used for sequential gating are included. All gates: all cell types defined by sequential gating. SUPERR clusters: clusters generated by clustering on each major cell types. PCs: plasma cells. See also Figures S1 and S3.

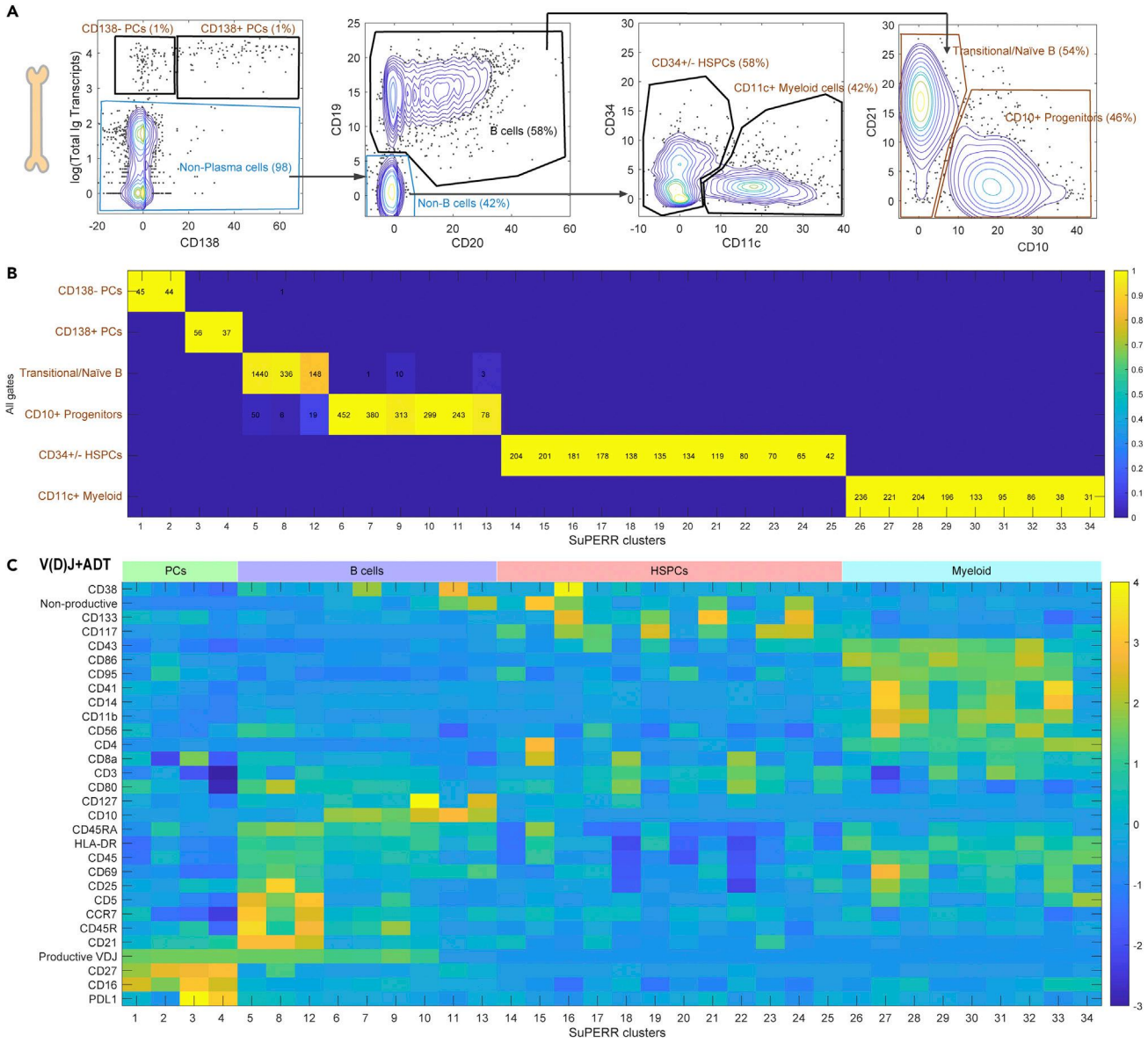


Figure 3. SuPERR workflow applied to bone marrow (BM) cells

(A) "Gating strategy" approach to identify major cell lineages on biaxial plots based on surface markers (ADT) and V(D)J data. Total Ig transcript: sum of Ig UMIs in the VDJ matrix. Gates for major lineages are indicated as black outlines and black text. Gates for downstream cell-identity validation are indicated as golden outlines and golden text.

(B) Cross comparison between the manually-gated major lineages and the final SuPERR clusters.

(C) The average expression levels of surface markers (ADT) and VDJ features for the final SuPERR clusters. Only the ADTs/VDJ features that were not used for sequential gating are included. Non-productive: 1, if a cell were labeled as non-productive in the VDJ matrix and 0 if not. Productive VDJ: 1, if a cell was labeled as productive in the VDJ matrix and 0 if not. All gates: all cell types defined by sequential gating. SUPERR clusters: clusters generated by clustering on each major cell types. PCs: plasma cells. See also Figures S2 and S3.

hematopoietic stem and progenitor cells (HSPCs) (Figure 3A, black borders). Notably, the B cells identified by the manual gating strategy using the cell-surface markers (ADT) were also present within the V(D)J matrix, which validated our strategy of using ADT for cell-lineage classification (Figures S1, S2 and S3A). In addition to the main cell lineages, our manual-gating strategy revealed other sub-clusters (Figures 2 and 3A, gray borders), which we used to validate the results from the downstream GEX-based clustering analysis. Our manual-gating strategy was further validated by high-dimensional flow cytometry analysis using an aliquot of the same samples taken before the single-cell encapsulation (Figure S3B).

Table 1. Cell-type classification and phenotype of PBMC clusters generated by the SuPERR workflow

Main lineage	SuPERR clusters	Cell classification	Phenotype	Refs.
Plasma cells	1	Plasma cells	CD19+/CD20-/highest levels of Ig-specific transcripts	(Shlomchik and Weisel 2012)
B cells	2	Naïve-I	IgD+/IgM+/CD27-/CD95-	(Glass et al., 2020; Kaminski et al., 2012; Tiller et al., 2007; Mehtonen et al., 2020)
($CD19^+$, $CD20^+$)	3	Naïve-II	IgD+/IgM+/CD27-/CD95-/HLA-DQA2+	
	4	Switched Memory	IgD-/IgM-/CD27+/CD95+	
	5	Unswitched Memory	IgD+/IgM+/CD27+	
	6	IgM^{hi}	IgM^{hi}/CD27-/CD95+/CD24-	
NK/NKT/MAIT/γδT	7	NK-I $CD16^+$	KLRC3+/$CD11b^{hi}$	(Poli et al., 2009; Evans et al., 2011; He et al., 2010; Lawand et al. 2017; Cai et al., 2020; Wong et al., 2019)
($CD56^+$)	8	NK-II $CD16^{hi}$	CX3CR1+	
	9	NK-III $CD16^+$	CX3CR1+	
	10	NK $CD56^{hi}$	$CD16^-$	
	11	MAIT cells	TRAV1-2+	
	12	NKT	CD4+/CCR7+	
	13	iNKT	TRAV24+/CD8+	
	14	NK-IV $CD16^+$	$CD11b^{hi}$	
	15	γδ T cells	TRGV9+/TRDV2+	
	16	NK-V $CD16^+$ (dividing)	MKI67+	
Monocytes	17	Classical-I	CD16-/CD68+/HLA-Drhi	(Villani et al., 2017)
($CD14^{lo/+}$)	18	Classical-II	CD16-/CD68+/HLA-DR^{lo}	
	19	Classical-III	CD16-/CD68+/HLA-DR^{lo}/$CD11b^{hi}$	
	20	Non-classical	CD16+/$CD14^{-/lo}$/CD68+/HLA-DR^{hi}	
	21	Intermediate/DCs	$CD16^{int}$/$CD14^{int}$/CD68+/HLA-DR^{hi}/CD11c+	
CD4 T cells	22	Naïve-I	CCR7+/SELL+/CD27+/CD95-/$CD3^{hi}$	(Blaser et al., 1998; Hashimoto et al., 2019; Juno et al., 2017; Kumar et al. 2018; Zhu et al. 2010)
($CD3^+$, $CD4^+$)	23	Naïve-II	CCR7+/SELL+/CD27+/CD95-	
	24	Tcm → Tem	$CCR7^{lo}$/SELL+/CD27+/CD95+	
	25	CTL-I	KLRB1+	
	26	CTL-II	KLRB1+/GZMA+/GZMK+	
	27	Tem	$CCR7^{lo}$/SELL-/CD27-/CD95+/LGALS1+/$S100A4^{hi}$	
	28	Treg	FOXP3+/CTLA4+/CD95+/HLA-DRB1+	
	29	Treg Naive	FOXP3+/CD45RA+/CD95-	
	30	Temra	CD45RA+/NKG7+/GNLY+/GZMB+	

(*Continued on next page*)

Table 1. *Continued*

Main lineage	SuPERR clusters	Cell classification	Phenotype	Refs.
CD8 T cells	31	Naïve-I	CCR7+/CD45RA+/CD27+/CD127+	(Braun et al. 2015; Martin and Badovinac 2018)
($CD3^+$, CD8a+)	32	Tcm-I	CCR7+/SELL+/CD27+/CX3CR1-	
	33	Tem	CCR7-/SELL+/CX3CR1+	
	34	TLE	CCR7-/SELL+/CX3CR1+/ZNF683+	
	35	MAIT cells	TRAV1-2+	
	36	Naïve-II	CCR7+/CD45RA+/$CD27^{hi}$/CD127+	
	37	Temra	CD45RA+/CCR7-/CD27-/CD127-	
	38	Tcm-II	CCR7+/SELL+/CD27+/CX3CR1-/TNFSF10+	

Gene expression (GEX) and cell-surface (ADT) markers used to classify the lymphoid and myeloid cell populations in the human peripheral blood mononuclear cells (PBMC). NK: Nature Killer cells; NKT: NK T cells; iNKT: invariant NKT; γδ T: gamma-delta T cells; Tcm: central memory T cells; Tem: effector memory T cells; CTL: cytotoxic T lymphocytes; Treg: regulatory T cells; Temra: effector memory T cells expressing CD45RA; TLE: long-lived effector memory T cells; MAIT: mucosal-associated invariant T cells. See also Figures S5–S9.

Cell-doublet discrimination

By identifying the *major cell lineages* as the first step of the SuPERR workflow we were able to accurately identify and remove cell doublets. For example, we applied flow cytometry-style biaxial feature plots on the ADT data to identify cell barcodes that co-expressed features of two or more major lineages (Figures S1 and S2). Cell barcodes containing ADT signals (cell-surface markers) known to belong to more than one well-defined immune lineage (e.g., co-expression of B-cell/CD19 and T-cell/CD3 markers) were considered doublets and removed from the downstream analysis. Figures S1 and S2 illustrate the gating strategy used for removing doublets from the PBMC and BM data sets, respectively. In our benchmarking analysis below, we further validate and compare our cell-doublet discrimination method here to other recently-developed algorithms.

Selection of major cell lineages prior to GEX analysis enhances true biological signals

Notably, the assignment of major cell lineages via a supervised manual-gating approach (ADT + V(D)J matrices) before performing the principal component analysis (PCA) on the GEX matrix, revealed that much of the data variation captured by the PCA is cell-lineage specific (Figure 4). Key variables, including ribosomal transcript content, mRNA abundance, and total unique gene counts, vary significantly among the cell lineages (Figure 4, ANOVA $p < 2.2e\text{-}16$). Even though our initial step of sample integration only outputted data for 2000 highly variable genes (HVGs), meaning that the subsequent PCA for the various lineages were performed on the same set of HVGs, the PCA can be interpreted as a feature selection step. For example, the PCA analysis focusing only on a pre-defined subset of cells (i.e., pre-gated major lineages) is able to produce principal components driven by biologically-meaningful variations that occur only *within* that lineage. In contrast, the principal components computed based on all cells are primarily driven by variations among different lineages. As such, enhancing the lineage-specific biological signals should also capture variations originated from differences in cellular states within a particular major lineage. Thus, lineage-specific variations, including variations because of differences in cellular state, will be reflected in the corresponding PCA, which in turn will be reflected in the downstream clustering and visualization (UMAP/t-SNE) results for the corresponding lineage. In sum, our pre-selection of major cell lineages before GEX clustering generates lineage-specific principal components that are biologically-meaningful, improving the resolution of the downstream GEX clustering analysis.

SuPERR reveals greater heterogeneity within major immune lineages

Human peripheral blood

Gating for major cell lineages using the ADT and V(D)J matrices revealed six distinct populations within the PBMC sample (Figure 2A, black outlines). We then further explored each major lineage by generating

Table 2. Cell-type classification and phenotype of BM clusters generated by the SuPERR workflow

Main lineage	SuPERR clusters	Cell classification	Phenotype	Refs.
Plasma cells	1	Ig^{hi}/PRDM1-	Ig^{hi}/PRDM1-	(Glass et al., 2020; Halliley et al., 2015)
(CD138-)	2	HLA-DR^{hi}/$SDC1^{lo}$/G2-M phase	HLA-DR^{hi}/$SDC1^{lo}$/G2-M phase	
Plasma cells	3	$IRF4^{hi}$/$PDL1^{+}$	$IRF4^{hi}$/$PDL1^{+}$	
(CD138+)	4	$XBP1^{hi}$/$SDC1^{hi}$/$PDL1^{hi}$	$XBP1^{hi}$/$SDC1^{hi}$/$PDL1^{hi}$	
B cells	5	T3/Naive	CD21+/IL4R+	(Zhou et al., 2020; Becker et al., 2018)
($CD19^{+}$, $CD20^{+}$)	6	Small pre-B/pre-BII-I	RAG+/IGKC+	
	7	Small pre-B/pre-BII-II	RAG+/IGLC3+	
	8	Switched Memory	CD21+/CD27+/IGHA1+/IGHG1+	
	9	T1/T2	RAG-/CD10+/$CD20^{-/lo}$	
	10	Large pre-B/pre-BII	CD34-/DNTT-/MKI67+	
	11	Early pro-B/pre-pro-B	CD34+/$DNTT^{hi}$	
	12	Activated naïve	CD21+/NR4A1+/DUSP2+	
	13	Late pro-B/pre-BI	CD34+/$DNTT^{int}$	
HSPCs	14	Pre-reticulocytes	$GYPA^{int}$/HBB+	(Jin et al., 2019; Kuramasu et al., 1998; Källberg and Leanderson 2008; Dzierzak and Philipsen 2013; Xie et al., 2020; Lai et al., 2017)
(Lineage-, $CD34^{+/-}$)	15	Pre-pDCs	CD123+/CD304+/CD303+/CSF2RA+	
	16	GMP	CD34+/FLT3+/CD164+/$CD45RA^{hi}$/Cell cycle	
	17	MEP	CD34+/GYPA-/ITGA2B+	
	18	Reticulocytes	GYPA-/HBB+/HBM-	
	19	Pro-Neutrophil	CD34+/MPO+/ELANE+	
	20	Reticulocytes $GYPA^{hi}$	$GYPA^{hi}$/HBB+	
	21	HSC/MPP	CD34+/AVP+/CD38-	
	22	Erythrocytes	GYPA-/HBB+/HBM-	
	23	Pro-erythroblast	$GYPA^{lo}$/HBB+/HBM+	
	24	CMP	CD34+/FLT3+/CD164+	
	25	Basophil/Mast cell progenitors	CD34+/CLC+/HDC+	
Myeloid/Granuloid	26	Neutropoiesis	MPO^{hi}/$ELANE^{hi}$	(Yang et al., 2014; Kawamura et al., 2017; Evrard et al., 2018)
(CD11c+)	27	Monopoiesis	$CD68^{hi}$/$CD14^{hi}$	
	28	Monopoiesis	$CD68^{lo}$/$CD14^{int}$	
	29	Neutropoiesis	MPO^{int}/$ELANE^{int}$	
	30	Monopoiesis	$CD68^{lo}$/$CD14^{lo}$	
	31	Monopoiesis	$CD68^{int}$/$CD14^{int}$	
	32	Neutropoiesis	MPO^{lo}/$ELANE^{lo}$/$MKI67^{hi}$	
	33	Monopoiesis	$CD68^{hi}$/$CD14^{hi}$	
	34	Progenitors	FLT3+/CD74+	

Gene expression (GEX) and cell-surface (ADT) markers used to classify the plasma cells, B cells, myeloid and granuloid cells, and the hematopoietic stem and progenitor cells (HSPCs) in the human bone marrow (BM). Ig: immunoglobulin transcripts; G2-M phase: genes involved in the cell cycle; T1/2/3: Transitional B cells; pDCs: plasmacytoid dendritic cells; GMP: granulocyte-monocyte progenitor; MEP: megakaryocyte-erythroid progenitor; HSC: hematopoietic stem cell; MPP: multipotent progenitor; CMP: common myeloid progenitor. See also Figures S4, S10, and S11.

subclusters using information from the GEX matrix. Briefly, we selected a set of HVGs from within the pre-gated population. The counts of selected genes for each cell were normalized by library size and then natural-log transformed, followed by per-gene Z-score scaling. We then applied a singular value decomposition (SVD) implementation of principal component analysis (PCA) to reduce the dimensionality. Left singular values were taken as gene scores and right singular values as cell scores. Next, we generated a K-nearest

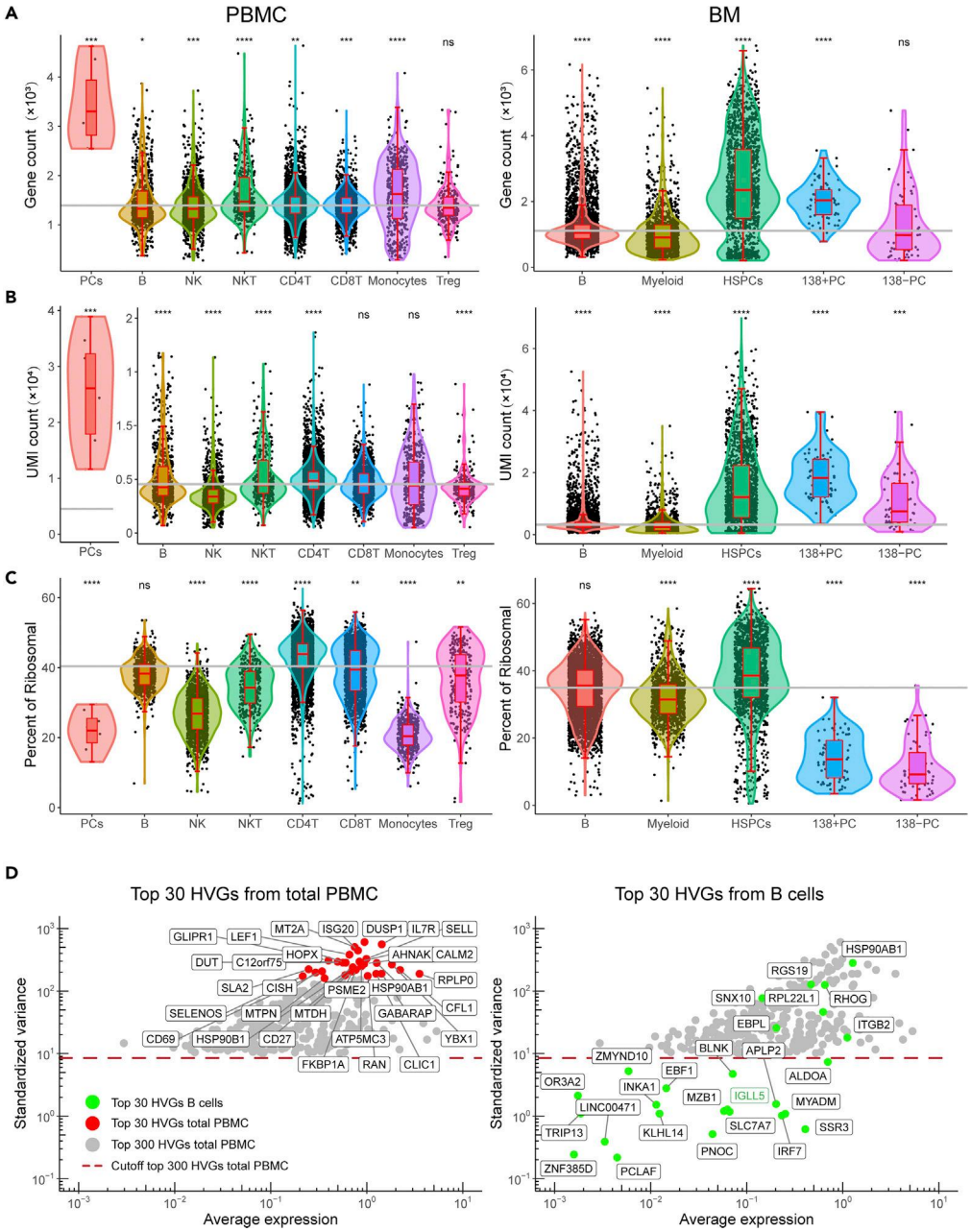


Figure 4. Cell-type-specific variations in gene expression

(A) "Gene count" represents the number of unique genes expressed by each cell type. Error bars in boxplots are the 95% confidence interval.

(B) "UMI count" represents the total mRNA abundance expressed by each cell type.

(C) "Percent of Ribosomal" represents the percentage of ribosomal gene UMI counts expressed by each cell type. The grey line shows the mean expression level for each feature in the total PBMC and BM samples.

(D) Left panel: the top 30 (red points) and the top 300 (grey points) highly variable genes (HVGs) from total PBMC. The points under the red dashed line fall below the top 300 HVGs of total PBMC. Right panel: the top 30 HVGs from PBMC-derived B cells (green points) displayed with the top 300 HVGs from total PBMC (grey points). Student's t-test was used to compare the mean of each cell type with the mean of the total PBMC/BM. *p<0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, unpaired, two-tailed. Multiple-group ANOVA test for (A), (B), and (C): p < 2.2e-16. PCs: plasma cells.

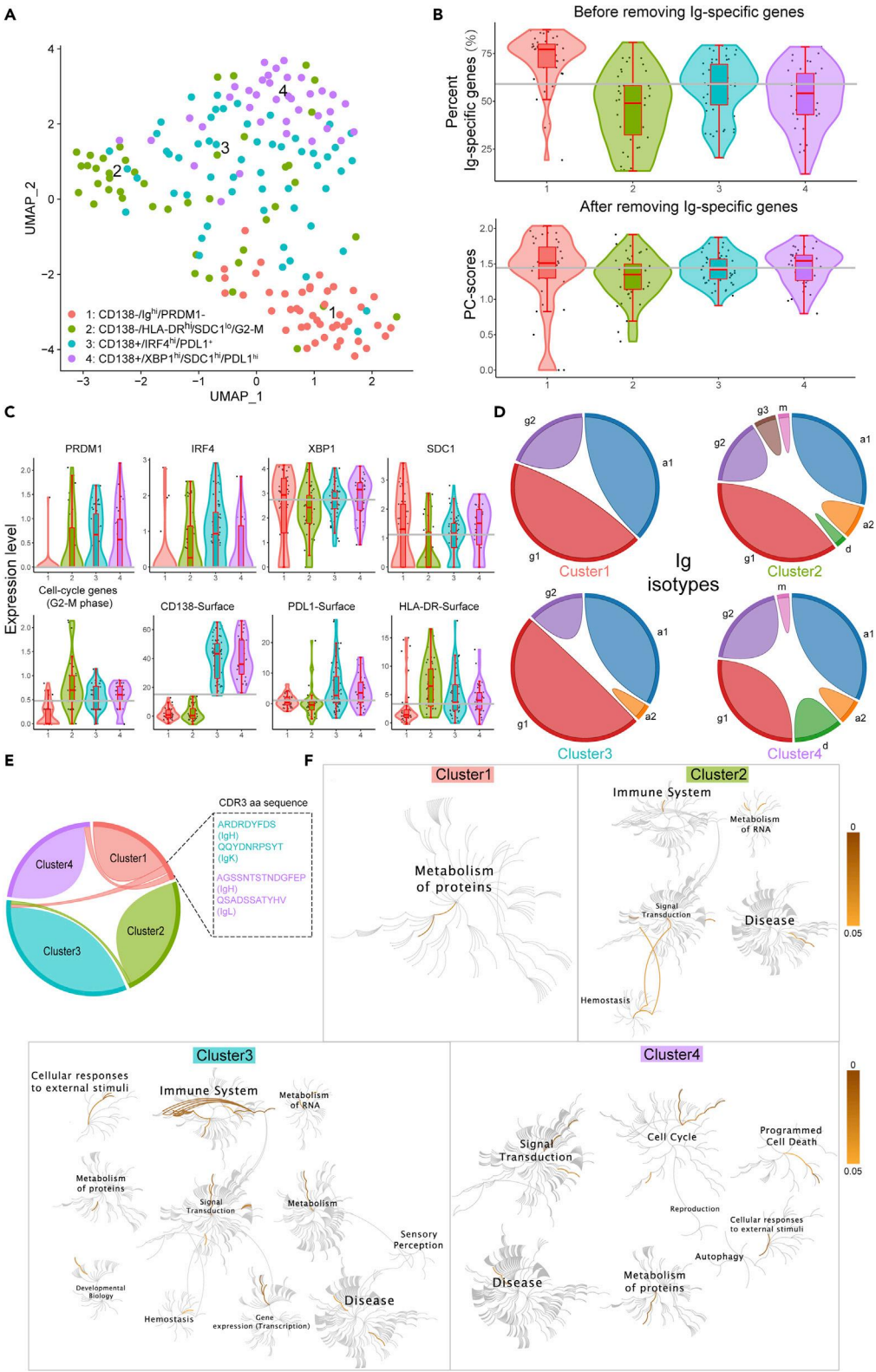


Figure 5. SuPERR workflow identifies four subsets of human plasma cells in the BM

(A) UMAP representation of the four bone marrow (BM) plasma cell clusters.

Figure 5. *Continued*

(B) Top panel: percentage of Ig-specific transcripts (UMI) expressed in each plasma cell subset. Bottom panel: expression levels (sum) of plasma cell genes (see STAR Methods) after removing Ig-specific UMIs and re-normalizing the data matrix. Error bars in boxplots are the 95% confidence interval.

(C) Expression levels of individual plasma cell genes, cell-cycle score after removing Ig-specific UMIs (See STAR Methods) and ADT. The grey line shows the mean expression level across all clusters.

(D) The antibody isotypes and subclasses expressed by each plasma cell subset.

(E) The connected lines on the Circus plot describe shared clones between clusters (clonal lineage was identified by the identical V and J gene usage, identical CDR3 nucleotide length, and ≥85% homology within the CDR3 nucleotide sequence).

(F) Reactome Pathway Database analysis (see STAR Methods) shows unique biological processes that define each plasma cell subset.

neighbors' graph, followed by Louvain community detection (see STAR Methods section for detailed description). Following these steps, we obtained the subclusters for each major lineage, herein called SuPERR clusters. At a Louvain resolution of 0.8, SuPERR identified 38 clusters in human PBMC, with each major lineage broken out into two or more subclusters, representing sub-lineage heterogeneity (Figure 2B).

We further investigated the 38 SuPERR clusters by exploring the expression levels of cell-surface proteins (ADT matrix) (Figures 2C and S5). Some ADT markers were lineage-specific (e.g., CD19 was used to classify B cells and Ig-specific transcript counts were used to classify plasma cells) and used to confirm SuPERR cell classification accuracy. Other markers displayed heterogeneous expression within lineages and were primarily used for defining and confirming subcluster identity. We integrated the ADT and GEX data matrices by simple concatenation to generate a joint matrix for SuPERR cluster identification (Figures 5 and S4–S11). We confirmed the heterogenous clusters as biologically meaningful using well-established cell-surface lineage markers (Table 1). For example, five subclusters were identified within the B-cell major lineage, all of which could be mapped back to previously described B-cell subsets (Garimalla et al., 2019; Glass et al., 2020; Kaminski et al., 2012; Tiller et al., 2007) (Figure S5). Similarly, several subclusters were identified within the T-cell and NK-cell major lineages (Figures S7–S9), including T-regulatory cells (Treg) defined as a subtype of $CD4^{+}$T cells with high expression of surface CD25 (ADT matrix) and FOXP3 transcripts (GEX matrix), and low expression of surface CD127 (ADT matrix) (Figures 2A and S7). Finally, five subclusters of monocytes were identified, including the previously described classical, non-classical, and intermediate monocyte subsets (Figure S6).

Remarkably, SuPERR readily identified a cluster of rare plasma cells (ASCs) in the blood even though we captured only eight plasma cells within this cluster (Figure 2). This level of resolution and accuracy in identifying rare plasma cells was possible by analyzing the total Ig-specific transcript counts from the V(D)J matrix, which unlike the analogous Ig transcripts from the GEX matrix, provide a more accurate count of the total productive Ig expression per cell. As plasma cells are defined by their unique ability to produce large quantities of Ig transcripts, they were readily identified based on a ~2.5 $\log_{10}$-fold increase in total Ig-specific UMI counts compared to B cells (Figure 2A).

Human bone marrow

SuPERR identified 34 unique clusters in human BM cells (Figures 3A and 3B). Of note, SuPERR matched the various B-cell subclusters to the different stages of B-cell development known to occur in the human BM (Mehtonen et al., 2020) (Figure S4). For example, the cell-surface expression of CD10, with DNTT and CD34 gene expression (GEX) transcripts, classified cluster 11 and cluster 13 as Early Pro-B (a.k.a., pre-pro-B) and Late Pro-B (a.k.a., pre-BI), respectively. The lack of DNTT mRNA transcripts in cluster 10 indicates a Large Pre-B stage (a.k.a., pre-BII), which is followed by the Small Pre-B stage (a.k.a., pre-BII) represented by clusters 6 ($VPREB1^{hi}$) and 7 ($VPREB1^{lo}$). The Transitional (T)1/T2/T3 and Naive B cells could be identified by their surface expression of CD21 (Zhou et al., 2020). Finally, the mature class-switched memory B cells were identified based on their expression of IGHA1/IGHG1 immunoglobulin transcripts and cell-surface CD27 (Becker et al., 2018) (Figure S4). Similarly, the SuPERR workflow also identified the developmental pathway of neutrophils, monocytes, and erythrocytes starting from the most undifferentiated population of hematopoietic stem cells (HSC) and multipotent progenitors (MPP) expressing CD34 and AVP transcripts (but lacking CD38) (Figures S10 and S11). The classification results for the BM subclusters are summarized in Table 2.

SuPERR workflow reveals new subsets of antibody-secreting cells in the human bone marrow

The SuPERR workflow readily and unambiguously identified four biologically distinct subsets of human ASCs (a.k.a., plasma cells) in the adult bone marrow (BM) of healthy donors (Figures 3 and 5). The determinant feature of a plasma cell is its ability to produce and secrete large quantities of immunoglobulins (Ig) (i.e., antibodies). The SuPERR workflow leverages this unique plasma cell feature by quantifying the absolute UMI counts of Ig-specific genes (IGH + IGL) from the V(D)J repertoire matrix (Figure 3A). Next, we integrate the Ig UMI count matrix to the ADT matrix we used to pre-gate major lineages (see above) and apply the same semi-supervised gating strategy on biaxial plots to identify cells with high Ig UMI counts. Unlike the GEX matrix, the Ig-specific UMI counts from V(D)J matrix provides a more accurate count of the productive Ig transcripts produced by plasma cells as the V(D)J data matrix is generated from a separate library using only Ig-specific primers (Zheng et al., 2017).

Remarkably, the number of total Ig-specific transcripts (UMI counts) detected in plasma cells is ~2.5 $\log_{10}$-fold higher than in B cells (Figures 3A and 3C), allowing for an unambiguous identification of total plasma cells. Next, we separated the total plasma cells into two subsets, based on the cell-surface expression of CD138 (ADT matrix), a canonical plasma-cell marker expressed by some, but not all, BM plasma cells (Halliley et al., 2015). Finally, we used the GEX matrix to further subdivide the CD138+ and CD138- plasma cell subsets based on their transcriptomic profile, revealing a total of 4 distinct subsets of human plasma cells (Figures 3 and 5).

To facilitate the identification of the differential gene expression that distinguish each of the four plasma cell subsets, we first removed from the GEX matrix all the mRNA transcripts derived from immunoglobulin genes (i.e., we removed IGHV, IGKV and IGLV genes). The rationale is that immunoglobulin genes represent more than 50% of the total mRNA transcripts (UMI counts) recovered from plasma cells. We then log-normalized the immunoglobulin-depleted GEX matrix and performed differential gene expression (DGE) analysis using the Wilcoxon Rank-Sum Test followed by Bonferroni correction. The resulting differentially-expressed genes readily defined unique biological processes for each plasma cell cluster (Figure 5). Cluster 1: CD138-/Ig^{hi}/PRDM1-; cluster 2: CD138-/HLA-DR^{hi}/$SDC1^{lo}$/G2-M phase; cluster 3: CD138+/IRF4-/PDL1+, cluster 4: CD138+/$XBP1^{hi}$/$SDC1^{hi}$/$PDL1^{hi}$ (Figures 5A and 5C). Clusters 2 and 4 show characteristics similar to previously identified human BM plasma cell subsets, described as Fraction A and Fraction B, respectively (Halliley et al., 2015). Notably, cluster 1 represents a unique plasma cell subset, in that >75% of its total mRNA transcripts represent immunoglobulin genes (Figure 5B). This large proportion of immunoglobulin gene transcripts indicates a high metabolic activity that is geared towards producing and secreting antibodies. Indeed, pathway analysis (Reactome Pathway Database) (Jassal et al., 2020) using the DGE list for cluster 1 revealed signals mainly for the metabolism of proteins (Figure 5F).

Surprisingly, we found that not all plasma cell clusters express the canonical plasma cell genes. Plasma cells develop from activated B cells through a dynamic cell-differentiation process, leading to the down regulation of B-cell identity genes, such as PAX5, and up-regulation of well-described plasma-cell genes, including PRDM1, SDC1, XBP1, and IRF4 (Halliley et al., 2015). These genes are considered canonical plasma-cell genes and, hence, they are used in scRNA-seq experiments to identify and classify plasma cells based on their transcriptomics (GEX matrix). However, the plasma cell cluster 1 does not express detectable levels of IRF4 and PRDM1. The absence of these canonical genes in cluster 1 was not because of the overrepresentation of immunoglobulin (Ig) genes because IRF4 and PRDM1 were not detected even after normalizing the GEX matrix without the immunoglobulin genes (Figure 5C).

By comparing the immunoglobulin isotypes and subclasses (IgM, IgD, IgG1-4, IgA1-2, IgE) expressed by each plasma-cell subset identified by SuPERR, we found that cluster 2 contains plasma cells of multiple isotypes/subclasses. In contrast, cluster 1 is more homogeneous, containing mainly IgG1 and IgA1 (Figure 5D). Furthermore, we observed that cluster 1 is composed of clonal plasma cells (defined by their IGH CDR3 amino acid sequences) that are shared among clusters 3 and 4 (Figure 5E). Finally, pathway analysis (Reactome Pathway Database) revealed unique biological processes and genetic programs that define each plasma cell subset (Figure 5F). Notably, cluster 4 expresses genes involved in cell cycle and programmed cell death, whereas cluster 3 appears to be actively responding to environmental stimulation.

In sum, the SuPERR workflow readily and unambiguously identified four biologically-distinct human plasma cell subsets in the adult BM. These findings further support the need for comprehensive

multi-omics single-cell data integration and reveal the potential shortcomings of relying solely on one omics data type (i.e., transcriptomics) to identify and classify cell (sub)types. In the following sections, we provide specific examples in which SuPERR workflow can outperform existing approaches.

Benchmarking of SuPERR against current methods developed to remove cell doublets

As we described above, the first step of the SuPERR workflow, in which we use cell-surface markers (ADT) and V(D)J gene counts in sequential biaxial plots to identify major cell lineages, also allows for accurate cell-doublet discrimination (Figures S1 and S2 and 6A and 6B). Our SuPERR approach successfully identified and removed 370 cell doublets in the PBMC sample. In contrast, standard/conventional approaches of trimming out cell barcodes with very high mRNA transcripts (e.g., removing cell barcodes with greater than the mean UMI value +4 standard deviation) (Ocasio et al., 2019) identified only 42 doublets. In the BM, SuPERR identified 108 cell doublets, and the conventional approach identified only one cell doublet (Figure 6A). Strikingly, every plasma cell we identified in PBMCs would have been trimmed/removed by the conventional approach of removing cells with very high mRNA transcripts (i.e., mean +4 SD UMI counts) (Figure 6A). In contrast, the SuPERR workflow readily recognized the PBMC plasma cells as single cells containing high UMI counts (Figure 6A). The plasma cell identity was further confirmed by the presence of a single productive V(D)J repertoire usage and the lack of other major lineage markers. Thus, the true cell doublets removed by SuPERR would otherwise be missed by conventional approaches and erroneously included as single cells in downstream GEX analysis, as visualized on the UMAP plot (Figure 6C).

We further compared our approach with current computational methods specifically designed to remove cell doublets in scRNA-seq datasets. First, we tested the standard workflow of scDblFinder (Germain et al., 2021) and compared it with the cell doublets defined by the SuPERR method (Figure 6D). To validate the cell doublets identified by both SuPERR and scDblFinder doublets, we used a set of well-defined gene markers to calculate a cell-type score (see STAR Methods) and then projected these scores for each cell barcode identified as doublets (Figures 6E and 6F). We showed that SuPERR and scDblFinder had good agreement on doublets that expressed high scores for multiple major lineages. Although scDblFinder outputted a larger number of cell doublets compared to the SuPERR workflow, the scDblFinder-specific doublets did not show heterotypic patterns when compared against the singlets defined by both methods (Figure S12), indicating potential false positives in the scDblFinder workflow. Most importantly, SuPERR identified hundreds of validated cell doublets (i.e., cell barcodes that co-express markers, both mRNA and cell-surface protein, of more than one cell lineage) that were missed by the scDblFinder workflow (Figure 6D).

Benchmarking of SuPERR against commonly-used methods reveal reduced cell-type misclassification, and superior ability to resolve and classify new cell types

We first compared our SuPERR workflow with the commonly-used Seurat package v3 (Stuart et al., 2019) using default parameters. In the PBMC sample, Seurat v3 identified 24 clusters compared to 38 clusters identified by the SuPERR workflow (Figure S14A). In the BM sample, Seurat v3 identified 22 clusters compared to 34 clusters identified by SuPERR (Figure S15A). When we directly compared the cluster assignment of each cell barcode (Figures S14A and S15A), multiple Seurat v3 clusters were further subdivided by the SuPERR workflow into several biologically-meaningful clusters (Figure 5 and S4–S11). The ability of the SuPERR workflow to identify greater and novel subsets of immune cells can be explained by the advantages of pre-gating major immune cell lineages based on cell-surface markers and Ig-specific transcripts before exploring the GEX matrix. As shown in Figure 4, pre-gating major lineages reveals the gene variation that occurs only *within* the pre-defined lineage instead of gene variation observed *across all* major lineages (Figure 4D).

Importantly, our benchmarking analysis reveals a significant number of cell-type misclassifications (i.e., a single cluster containing cells from different major lineages) generated by the default Seurat v3 workflow. For example, Seurat v3 generated three clusters (5, 9, and 15) containing a mixture of cell lineages, including $CD8^+$T cells and NK cells (Figure S14). In contrast, SuPERR correctly clustered and classified these cells separately. Similarly, Seurat v3 cluster 3 mixed $CD4^+$ and $CD8^+$T cells, whereas SuPERR correctly identified and separated these different cell types (Figure S14). Notably, such cell-type misclassification artifacts are not rare and commonly occur when simultaneously clustering all cells in high-dimensional space (Orlova et al. 2018; Altman and Krzywinski 2018) using the GEX data matrix as performed by most, if not all, scRNA-seq analysis workflow.

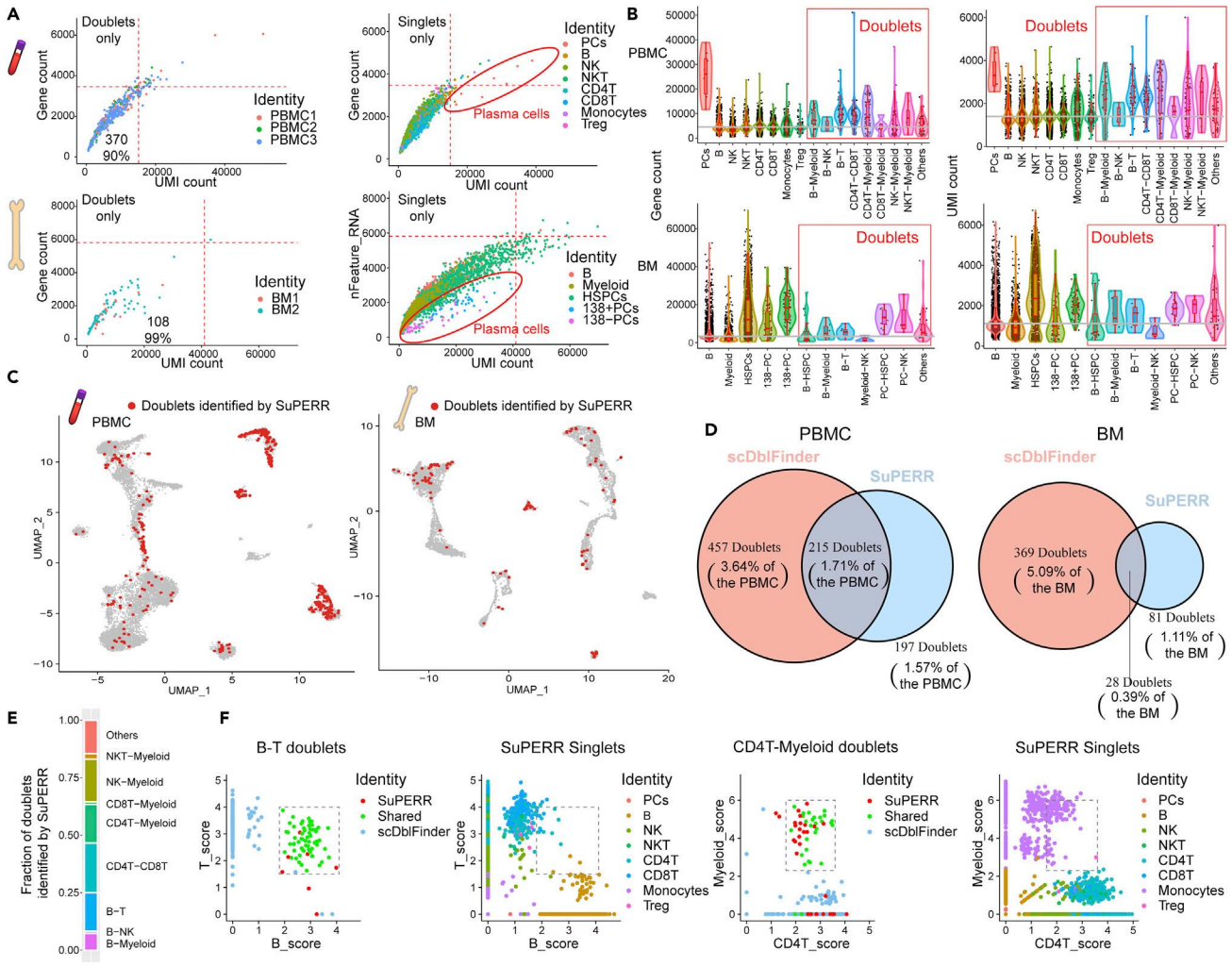


Figure 6. Cell-doublet identification by SuPERR using both surface markers and gene expression data matrices

(A) Distribution (gene count x total UMI) of cell doublets (left) and singlets (right) detected by the SuPERR approach. The red dashed lines show the threshold used by some conventional approaches to exclude cells that express higher than mean+4SD of gene count and total UMI. Only the cells above the dashed lines would have been excluded from the downstream analysis in conventional approaches (i.e., plasma cells in PBMCs, highlighted in the red circle, would have been incorrectly excluded from downstream analysis).

(B) The number of unique genes (left panels) and the number of total UMIs (right panels) expressed by singlets and doublets in PBMC (top panels) and BM (bottom panels). The grey line shows the mean expression level across all clusters. Error bars in boxplots are the 95% confidence interval.

(C) Cell doublets identified by the SuPERR workflow and projected on a UMAP, showing the cell doublets are spread across multiple clusters.

(D) Venn diagram comparing the cell doublets identified by the SuPERR workflow and the ScDblFinder pipelines.

(E) Proportion of heterotypic doublets identified and classified by SuPERR in PBMC.

(F) Expression level of gene signatures (see STAR Methods) of heterotypic doublets defined by SuPERR and scDblFinder to confirm their cell identities. Red points represent SuPERR-defined doublets. Green points are the cell doublets identified by both SuPERR and scDblFinder. Blue points represent scDblFinder-defined doublets, which were identified as singlets by SuPERR. The immune cell types were annotated by the SuPERR workflow. See also Figures S1, S2, and S12.

One might attribute the improved performance of the SuPERR workflow to the fact that it utilizes additional input data from other omics (i.e., ADT and V(D)J). Therefore, we further compared the SuPERR workflow with two recently-developed pipelines that also integrate the information from both GEX and ADT data for clustering cells. We compared the SuPERR workflow to the weighted nearest neighbor (WNN) approach as implemented in Seurat v4 (Hao et al., 2021) and the similarity network fusion (SNF) as implemented in CiteFuse (Kim et al., 2020) (Figures S14B and S14C and S15B and S15C). To better quantify the performance of each data analysis workflow and to reveal the extent of unwanted cell-type misclassifications from each

approach, we developed a new scoring system named Cell Fidelity Statistics (CFS) score (Babcock et al., 2021). Briefly, we consider the biaxial feature plots of cell-surface markers (ADT) and V(D)J-derived Ig-transcript counts (Figures S1 and S2) to represent a "gold standard" as this iterative nature of SuPERR prohibits cells from inappropriately co-clustering with a separate lineage. This approach is borrowed from the well-established biaxial gating strategy of flow cytometry analyses (Mahnke et al. 2010). We then compare the cell-type identity assigned to each cell barcode to those from a different workflow and consider the proportion of cells that change identity, generating a cell fidelity metric. The proportion of misclassified cells is reported as a "1-CFS score," which provides a statistical measure of uncertainty in the cell-type assignment steps of the compared workflow. For example, the PBMC and BM clusters generated by Seurat v3 showed a 1 - CFS score of 0.0694 and 0.0531, respectively, indicating that 6.94% of PBMCs and 5.31% of BM cells were misclassified by the Seurat v3 (Figures 7A and 7B). Notably, the"1-CFS scores"for the WNN (Seurat v4) and SNF (CiteFuse) approaches were lower than Seurat v3, indicating better agreement between Seurat v4, CiteFuse and SuPERR and further highlighting the benefits of integrating additional omics for single-cell analysis (Figures 7A and 7B). However, it is important to note that, similar to Seurat v3, WNN and SNF approaches still generated cell-type misclassifications (Figures S14–S17).

To compare the performance of SuPERR against WNN (Seurat v4) and SNF (CiteFuse), we calculated the Average Silhouette Width (ASW) (Rousseeuw 1987) score to determine the optimal cluster resolution for WNN, and k for SNF. However, the optimal ASW scores resulted in low resolution/k with exceedingly few clusters, revealing insufficient heterogeneity and preventing direct comparisons with SuPERR (Figure S13). Therefore, we instead intentionally increased the cluster resolution to generate a total number of clusters that is comparable across all methods. For example, the default Seurat v4 pipeline at a cluster resolution 0.5 generated 34 and 29 clusters versus 38 and 34 clusters generated by SuPERR in the PBMC and BM, respectively. In contrast, at a higher cluster resolution of 3, WNN generated 40 and 36 clusters in the PBMC and BM, respectively. Our rationale was that by generating additional clusters from the WNN approach we could observe improved cluster agreement with SuPERR. However, even at higher cluster resolution and larger number of clusters, WNN was not able to identify a plasma cell cluster in the PMBC, and could not distinguish plasma cell subclusters in the BM (Figures 7A and 7B and S14B and S15B and S17). Similarly, the CiteFuse workflow was not able to identify plasma cells in the PBMC or plasma-cell heterogeneity in the BM even after manually increasing the number of k (which is similar to increasing cluster resolution) for their spectral clustering approach (Figures 7A and 7B and S14C and S15C and S17). These results further support the SuPERR workflow and its ability to generate biologically-meaningful subclusters of cells while preventing cell-type misclassifications.

To further validate the ability of SuPERR to prevent cell-type misclassifications, we explored the purity of the cell clusters generated by each approach independently. We explored the gene expression profile (GEX) and cell-surface protein (ADT) expression for cells included in the most abundant PBMC cluster (CD4 T cells) generated by SuPERR, Seurat v3, Seurat v4, and CiteFuse (Figure S16A). As we show in Figure S16B, the CD4 T cell clusters generated by Seurat v3, Seurat v4, and CiteFuse contained substantial cell-type misclassifications dominated by CD8 T cells and NK/NKT cells. In contrast, SuPERR-defined CD4 T cell clusters showed higher overall purity (Figure S16B).

To determine whether cell-type misclassifications generated by the other approaches could confound interpretation of downstream differential gene expression (DGE) analysis, we explored the PBMC clusters generated by the WNN approach (Seurat v4) (Figure 7C). Because DGE analysis is a commonly-used method to interpret biological differences between clusters or cell types, we hypothesized that running DGE analysis on a WNN (Seurat v4) cluster containing cell-type misclassification could generate misleading results even at low contamination (i.e., cell-type mixing) numbers. For example, the WNN-derived PBMC cluster 4 contains mainly NK cells, but it is also contaminated/mixed with NKT and T cells. Indeed, by analyzing the DGE list of the cluster 4 before removing the contaminating (NKT and T) cells, the *TRGV10* gene appeared as highly expressed for this cell population (Figure 7C). However, *TRGV10* is expressed on gamma-delta T cells (Aliseychik et al., 2020), not on NK cells. These results indicate that the widespread cell-type misclassification that is often observed in conventional data-analysis workflows can confound data interpretation.

Finally, when comparing the new SuPERR-identified plasma cells with the other workflows, we observed major differences. Although SuPERR readily and accurately identified a cluster of eight plasma cells in

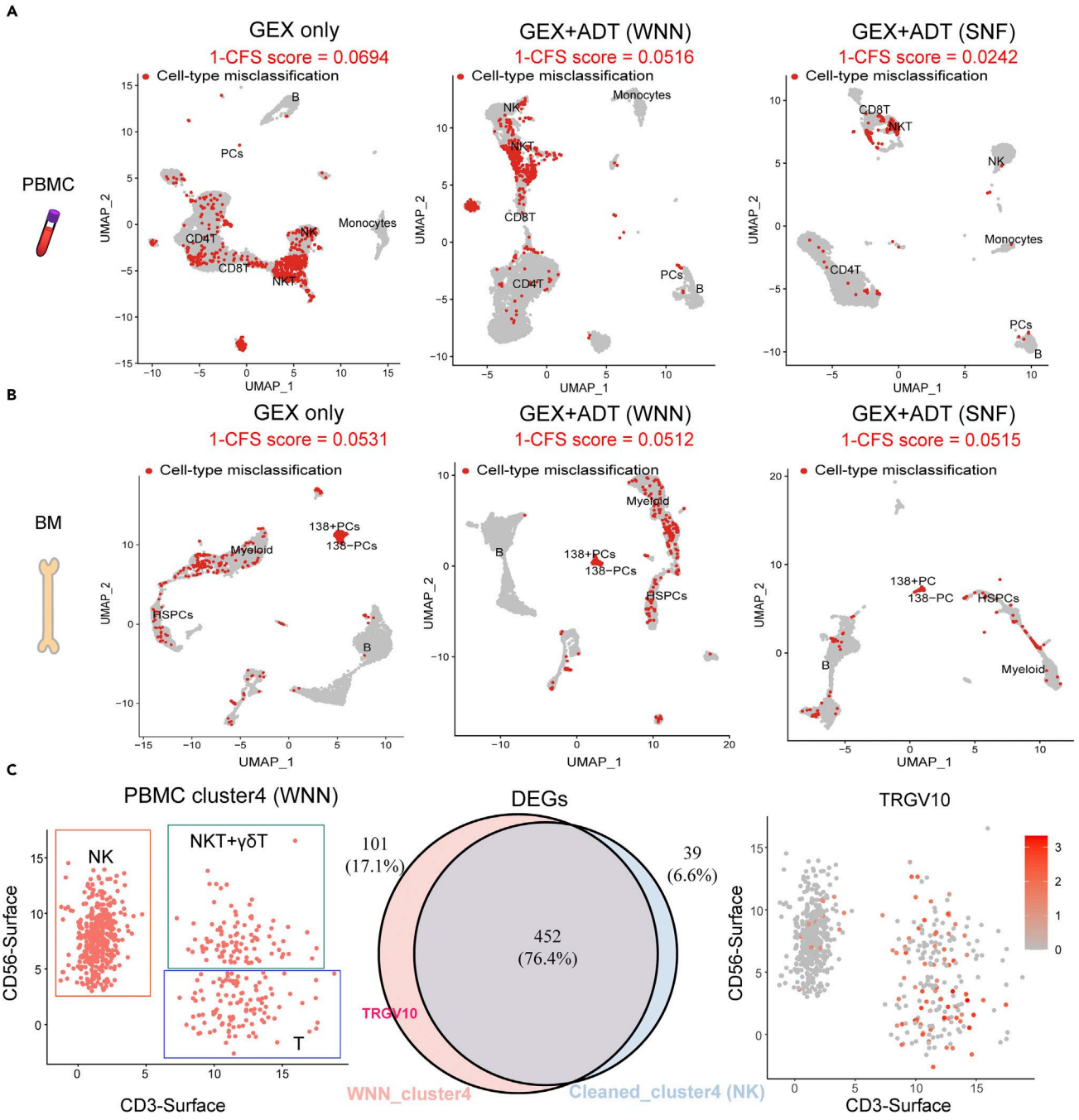


Figure 7. SuPERR identifies significant cell-type misclassifications in other commonly-used approaches

(A) Red points represent the peripheral blood mononuclear cells (PBMC) that were misclassified by either the conventional approach using GEX data only (i.e., Seurat v3), or by more recent approaches using both GEX and ADT data, such as the WNN in Seurat v4, and the SNF in CiteFuse. The Cell Fidelity Statistic (CFS, see STAR Methods) reports the fraction of correctly classified cells, the inverse of which is the fraction of misclassified cells (6.94% by Seurat v3, 5.16% by Seurat v4, 2.42% by CiteFuse).

(B) Red points represent the bone marrow (BM) cells that were misclassified by Seurat v3 (5.31%), WNN/Seurat v4 (5.12%), and SNF/CiteFuse (5.15%) as determined by CFS. CFS scores show a progressive improvement in cell-type classification from Seurat v3 (GEX only) to Seurat v4 and CiteFuse, revealing higher agreement between CiteFuse and gold-standard biaxial gating of cell lineages.

(C) The PBMC cluster 4 generated by the WNN method (Seurat v4) contains misclassified cells (i.e., a mixture of NK, NKT, and T cells) and was further explored using the cell-surface (ADT) markers CD56 and CD3 (left panel). The Differential Gene Expression (DGE) analysis for cluster 4 (pink circle) compared to "cleaned" NK cells (Venn diagram) shows the *TRGV10* gene as a top hit. However, the *TRGV10* gene is mostly expressed in $CD3^{+}$ gamma-delta T cells and absent in NK cells (right panel). See also Figures S14–S17.

PBMCs (Figure 2A), the other pipelines (Seurat v3, WNN/Seurat v4, and SNF/CiteFuse) failed to identify these rare cells (Figure S14). Unlike PBMC, BM cells contain significantly more plasma cells. In this case, SuPERR identified four biologically distinct plasma cell clusters, whereas the other workflows identified only one cluster (Figures 5 and S15). Of note, the plasma cell cluster generated by the Seurat v3 workflow contained a mixture of cell types (i.e., cell-type misclassification), including B cells and HSPCs (Figure S15A). In addition, all the workflows we tested have misclassified a fraction of the SuPERR-identified plasma cells, as they appeared spread across multiple clusters (Figure S15).

To determine whether further optimizations in the analysis workflow would allow the other approaches (Seurat v3, WNN and SNF) to identify the four SuPERR plasma-cell clusters in the BM, we performed a second iteration of clustering (i.e., recursive approach/re-clustering) on the original BM plasma cell cluster identified by both Seurat (v3 and v4) and CiteFuse. After the second iteration of clustering, Seurat (v3 and v4) and CiteFuse were able to identify more BM plasma cell clusters (Figure S17). However, the plasma cells clusters remained as a mixture of contaminating B cells and HSPCs (Figure S17). For the fairness of comparison, the parameters used for running the Seurat v3, Seurat v4, CiteFuse, and SuPERR workflows were set as the default parameters described in STAR Methods (unless noted otherwise). And the re-clustering approach on the original Seurat v3, Seurat v4 and CiteFuse clusters was limited to two iterations.

Taken together, the results from the benchmarking analysis show that the SuPERR workflow not only identifies more and novel clusters but also improves the clustering purity and the accuracy of biological interpretation.

DISCUSSION

A core component of a single-cell RNA-seq analysis is the identification and classification of cell types. The most popular and widely-used approaches rely on principal component analysis (PCA) derived from a set of HVGs, followed by nearest neighbors graph-based clustering. The resulting clusters are then assigned to a cell type and biological function via manual annotation, using the mRNA transcript counts as a guide. In existing approaches, the selection of HVGs for downstream clustering analysis is often dominated by lineage-specific markers. These markers often capture variance that exists *between* cell lineages (e.g., the variance that distinguishes T cells from B cells) but often misses variance that exists *within* each cell lineage (e.g., the variance that distinguishes the various T-cell subsets, including effector memory from central memory, stem memory, naïve, etc.) (Figure 4). Thus, an improved method to select HVGs *within* each cell lineage is needed to discover new biologically-meaningful cell subtypes with high accuracy.

Moreover, selecting HVGs with high variance across all lineages (instead of *within* lineages) as the very first step in data analysis (as in many existing approaches) carries the risk of inappropriately grouping cells with similar gene expression but arising from separate cell lineages, resulting in cell-type misclassification (Figure 7). Cell-type misclassification often occurs in conventional scRNA-seq data analyses. It can be caused by a mathematical phenomenon known as the curse of dimensionality (Altman and Krzywinski 2018; Trevor Hastie et al., 2009; Orlova et al. 2018), inadvertently misguiding biological interpretations (Figure 7, 1-CFS scores). The integration of additional omics, including cell-surface markers used in WNN (Hao et al., 2021) and SNF (Kim et al., 2020) approaches, can improve cell-type classification, however, it doesn't eliminate cell-type misclassification. Another contributing factor to cell-type misclassification is the presence of cell doublets in the final dataset. Despite recent advances in microfluidics that precisely generate droplets containing single cells, a significant fraction of the droplets can still contain more than one cell (Klein et al., 2015; Cao et al., 2017). Current approaches to remove cell doublets rely on overly high mRNA transcript counts (i.e., outliers) (Ocasio et al., 2019) and gene marker co-expression. However, as we show here, many cell doublets contain average mRNA counts and can mistakenly be carried over to downstream analysis (Figure 6). More recent and sophisticated computational algorithms to identify cell doublets such as scDblFinder (Germain et al., 2021) can still ignore true doublets and captures false positives (Figures 6 and S12).

To address these pitfalls of existing methods, we presented here the semi-supervised SuPERR workflow. In SuPERR, we simultaneously apply information gained from triple-omics sequencing, namely gene expression (GEX), cell-surface proteins measured by antibody-derived tags (ADT), and immunoglobulin transcript counts from the V(D)J repertoire matrix. As a quality control measure and to prevent downstream cell-type misclassification, the first step of SuPERR is to perform a "manual gating" similar to the standard flow

cytometry analysis. ADT and V(D)J matrices can reliably classify cells into clearly distinct major lineages using well-established canonical markers. This manual gating step ensures that any downstream subclusters are composed of a single lineage, allowing for a more accurate cell-type classification and data interpretation. The second step of the SuPERR is to examine each lineage independently, both from one another and from the first gating steps. Rather than selecting genes with high variance across all cell lineages, we select HVGs only from within each manually-gated lineage. These HVGs inform a PCA and subsequent nearest-neighbors graph-based clustering. Our PBMC and BM samples analysis show that the SuPERR approach can reveal additional cell types (and likely cell states) compared to other existing approaches. We reason that this is mainly because of the ability of SuPERR to select only the relevant HVGs for each well-defined lineage and ignore extra sources of variance that may not be informative within each cell lineage.

To ensure the generalizability of our biaxial gating approach so that SuPERR can be broadly applicable regardless of the type of sample/tissue, we included in our ADT panel well-established cell-surface markers that have been previously validated using the gold-standard flow cytometry approach (Figure S3). These cell-surface markers can readily (and accurately) identify major cell lineages, including immune lineages (i.e., $CD45^{+}$), non-immune lineages (i.e., $CD45^{-}$), epithelial lineages (i.e., EPCAM+), and other lineages regardless of the origin of the sample. Of note, many lineage-specific cell-surface markers have been validated and published in open-source journals, including the collection of cell-type classifications available as optimized multicolor immunophenotyping panels (OMIPs) (Mahnke et al. 2010). These well-characterized cell-surface markers can be included as part of the ADT panels used in any experimental design to identify major immune and non-immune lineages. Thus, new datasets with different combinations of antibody (ADT) panels can still follow a similar SuPERR gating strategy as long as the panel includes known lineage markers. Furthermore, we expect the SuPERR approach to improve the ability to identify novel cell subsets and cellular states (within each major lineages) for which cell-surface markers have not yet been characterized. For example, a novel subset (cluster) of B cells can be discovered within the major B-cell lineage (i.e., within cells expressing surface CD19). In the (unlikely) event that the novel B-cell subset doesn't express CD19, the novel subset should still form a separate cluster within another pre-defined gate based on its cell-surface marker expression. In this scenario, the additional omics (GEX and VDJ) in our workflow would reveal the B-cell identity of the novel B-cell subset.

Future studies to optimize the initial biaxial manual gating step might include developing an automated semi-supervised clustering algorithm to readily identify major cell lineages using the ADT and V(D)J data matrices. Automated algorithms that still rely on biaxial projections, such as the exhaustive projection pursuit (EPP) approaches (Friedman and Tukey 1974) might provide more accurate results compared to fully unsupervised methods in high-dimensional space, which might suffer from the curse of dimensionality (Altman and Krzywinski 2018; Trevor Hastie et al., 2009; Orlova et al. 2018). In fact, we recently implemented an automated EPP approach to identify cell subsets using cell-surface markers in high-dimensional flow and mass cytometry datasets (Meehan et al., 2019). Although this new subset identification and characterization (SIC) pipeline was specifically developed for flow and mass cytometry datasets (Meehan et al., 2019), future studies should aim to optimize such pipelines to process multi-omics single-cell ADT and V(D)J matrices as an alternative automated step in the SuPERR workflow.

We recognize the difficulty and expense of generating triple-omics data for every sample, as well as the limitations of applying the SuPERR approach to older datasets for which ADT and V(D)J matrices are unavailable. Also, some samples might not include B and/or T cells, which means there will not be V(D)J matrices. Thus, to broaden the application of the SuPERR principles to datasets without triple-omics available, we propose a gene-based (GEX) recursive analysis approach (i.e., re-clustering), which carries some of the same benefits of SuPERR. When we applied this recursive approach to the same BM dataset explored by the SuPERR we identified improvements, but also limitations (Figure S17). Focusing on the plasma cells, the recursive approach applied to other existing workflows, which only identified one cluster of plasma cells in the first clustering iteration, now generated more subclusters of plasma cells. However, the re-clustering approach was not able to accurately identify biologically-meaningful plasma cell clusters and, most importantly, it was not able to isolate the plasma cells from the other contaminating cells. Thus, the resulting subclusters of plasma cells represented a mixture of plasma cells, B cells, and HSPCs (Figure S17). In contrast, SuPERR readily and accurately identified four subsets of plasma cells that were confirmed based on the three omics (GEX, ADT, and VDJ), including the high antibody transcript counts (~2.5 $\log_{10}$-fold higher

antibody transcripts than B cells). Moreover, the plasma cell subsets identified by SuPERR are of great biological importance as revealed by the Reactome pathway analysis (Figure 5). Therefore, integrating three omics in the SuPERR workflow provides unique information that recursive clustering strategies for defining HVGs cannot achieve. Our conclusions here agree with the latest implementation of Seurat v4 (Hao et al., 2021) and CiteFuse (Kim et al., 2020), which also integrate ADT measurements in their analysis pipeline, further supporting the SuPERR approach of multi-omics integration for defining cell clusters.

Taken together, we developed a comprehensive multi-omics single-cell data integration and analysis workflow that mitigates or resolves pitfalls in existing approaches and allows for the discovery of novel and biologically-meaningful cell subsets in the human immune system.

Limitations of the study

The SuPERR approach was designed to minimize cell-type misclassifications and reveal additional heterogeneity in multi-omics single-cell assays that include at least GEX and ADT data. Hence, the SuPERR workflow is limited to datasets in which cell-surface protein (ADT) data in available. One of the steps in the SuPERR workflow is the user-defined sequential gating performed on select cell-surface markers (ADT) that are highly expressed and lineage-specific, similar to Hi-D flow cytometry. However, this sequential manual gating could introduce user bias. Automating the sequential gating step could overcome this limitation. In fact, we have developed an automated sequential gating approach for Hi-D flow and CyTOF data (Meehan et al., 2019), which could be implemented for ADT data in future versions of SuPERR. Finally, the ADT datasets from multiple experiments may contain batch effects, preventing data concatenation before manual gating. In this scenario, we recommend performing the sequential manual gating in each sample individually before data integration, even though it could be time consuming when there are multiple samples. In our studies here, the ADT datasets did not show any significant batch effects. Hence, we were able to concatenate all tissue-specific samples before manual gating.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Subjects and specimen collection
- METHOD DETAILS
  - Cell preparation
  - Single-cell RNA-sequencing
  - Multi-omics single-cell data preprocessing
  - Gene expression (GEX)
  - Antibody-derived tags (ADT)
  - Antibody repertoire (VDJ)
  - Manual gating
  - Clustering on gene expression matrix
  - Pathway analysis
  - Cell-doublet identification
  - Calculation of cell-type scores
  - Cell fidelity statistic
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Differential gene expression analysis
  - Student's t-test

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105123.

ACKNOWLEDGMENTS

This study was supported in part by Georgia Clinical and Translational Science Alliance (CTSA) through the National Center for Advancing Translational Sciences of the National Institutes of Health under Award number NIHUL1TR002378 (E.E.B.G.); Pediatric Research Alliance, Center for Transplantation, and Immune-Mediated Disorders (E.E.B.G.); Lowance Center for Human Immunology (E.E.B.G.); NIH-NIAIDR01-AI123126-05S1 (E.E.B.G.), and the National Science Foundation awards CCF1552784 and CCF2007029 (P.Q.). P.Q. is an ISAC Marylou Ingram Scholar and a Carol Ann and David D. Flanagan Faculty Fellow. We thank Sachin Kumar (Emory University) for providing the Python script used to generate the Circos plots and Matthew C. Woodruff (Emory University) for the R script used to generate the V(D)J input files for the Circos plots. Flow cytometry data were collected at the Emory's Pediatrics/Winship Flow Cytometry Core (access supported in part by Children's Healthcare of Atlanta). Single-cell libraries were sequenced at the Emory Integrated Genomics Core (EIGC), which is subsidized by the Emory University School of Medicine and is one of the Emory Integrated Core Facilities; the Parker H. Petit Institute for Bioengineering and Bioscience at the Georgia Institute of Technology; and the PerkinElmer Genomics Inc. We thank Dr. F. Eun-Hyung Lee (Emory University), Sang N. Le, and Mindy R. Hernández for kindly providing the bone marrow aspirates from healthy adult donors through Dr. Lee's Emory IRB protocol, and Emory University's Children's Clinical and Translational Discovery Core (CCTDC) for providing peripheral blood samples from healthy donors.

AUTHOR CONTRIBUTIONS

E.E.B.G. and P.Q. conceived the idea and experimental design, interpreted the data, and wrote the final draft. C.M. and J.Y. generated, analyzed, and interpreted the multi-omics single-cell datasets and generated the figures. J.Y. performed the preprocessing workflows to obtain final count matrices for each single-cell dataset. A.K. received and processed the human samples, generated the single-cell libraries and flow cytometry datasets, and analyzed the data. B.R.B. provided input into data analysis and generated the CFS scores. All authors contributed to the writing of the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research. We worked to ensure gender balance in the recruitment of human subjects. We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. One or more of the authors of this paper self-identifies as a gender minority in their field of research. One or more of the authors of this paper self-identifies as a member of the LGBTQIA + community.

Received: April 21, 2022
Revised: July 12, 2022
Accepted: September 9, 2022
Published: October 21, 2022

REFERENCES

Aliseychik, M., Patrikeev, A., Gusev, F., Grigorenko, A., Andreeva, T., Biragyn, A., and Rogaev, E. (2020). Dissection of the human T-cell receptor γ gene repertoire in the brain and peripheral blood identifies age- and alzheimer's disease-associated clonotype profiles. Front. Immunol. *11*, 12.

Altman, N., and Krzywinski, M. (2018). The curse(s) of dimensionality. Nat. Methods *15*, 399–400.

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). 'Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, e8124.

Babcock, B.R., Kosters, A., Yang, J., White, L., and Ghosn, E.B. (2021). Data matrix normalization and merging strategies minimize batch-specific systemic variation in scRNA-seq data. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.18.456898.

Becker, S.C., Szyska, M., Mensen, A., Hellwig, K., Otto, R., Olfe, L., Volk, H.-D., Dörner, T., Dörken, B., Scheibenbogen, C., et al. (2018). A comparative analysis of human bone marrow–resident and peripheral memory B cells. J. Allergy Clin. Immunol. *141*, 1911–1913.e7.

Blaser, C., Kaufmann, M., Müller, C., Zimmermann, C., Wells, V., Mallucci, L., and Pircher, H. (1998). Beta-galactoside-binding protein secreted by activated T cells inhibits antigen-induced proliferation of T cells. Eur. J. Immunol. *28*, 2311–2319.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). 'Fast unfolding of

communities in large networks. J. Stat. Mech. *2008*, P10008.

Braun, J., Frentsch, M., and Thiel, A. (2015). Hobit and human effector T-cell differentiation: the beginning of a long journey. Eur. J. Immunol. *45*, 2762–2765.

Cai, Y., Dai, Y., Wang, Y., Yang, Q., Guo, J., Wei, C., Chen, W., Huang, H., Zhu, J., Zhang, C., et al. (2020). Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis. EBioMedicine *53*, 102686.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661–667.

Colomé-Tatché, M., and Theis, F.J. (2018). Statistical single cell multi-omics integration. Curr. Opin. Struct. Biol. *7*, 54–59.

Dzierzak, E., and Philipsen, S. (2013). Erythropoiesis: development and differentiation. Cold Spring Harb. Perspect. Med. *3*, a011601.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Evans, J.H., Horowitz, A., Mehrabi, M., Wise, E.L., Pease, J.E., Riley, E.M., and Davis, D.M. (2011). A distinct subset of human NK cells expressing HLA-DR expand in response to IL-2 and can aid immune responses to BCG. Eur. J. Immunol. *41*, 1924–1933.

Evrard, M., Kwok, I.W.H., Chong, S.Z., Teng, K.W.W., Becht, E., Chen, J., Sieow, J.L., Penny, H.L., Ching, G.C., Devi, S., et al. (2018). Developmental analysis of bone marrow neutrophils reveals populations specialized in expansion, trafficking, and effector functions. Immunity *48*, 364–379.e8.

Friedman, J.H., and Tukey, J.W. (1974). A projection Pursuit algorithm for exploratory data analysis. IEEE Trans. Comput. *C-23*, 881–890.

Garimalla, S., Nguyen, D.C., Halliley, J.L., Tipton, C., Rosenberg, A.F., Fucile, C.F., Saney, C.L., Kyu, S., Kaminski, D., Qian, Y., et al. (2019). Differential transcriptome and development of human peripheral plasma cell subsets. JCI Insight, 126732.

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A., and Yosef, N. (2021). 'Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat. Methods *18*, 272–282.

Germain, P.L., Lun, A., Macnair, W., and Robinson, M.D. (2021). Doublet identification in single-cell sequencing data using scDblFinder [version 1; peer review: 1 approved. F1000Res. *10*, 10.

Glass, D.R., Tsai, A.G., Oliveria, J.P., Hartmann, F.J., Kimmey, S.C., Calderon, A.A., Borges, L., Glass, M.C., Wagar, L.E., Davis, M.M., and Bendall, S.C. (2020). An integrated multi-omic single-cell atlas of human B cell identity. Immunity *53*, 217–232.e5.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat. Methods *14*, 955–958.

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

Halliley, J.L., Tipton, C.M., Liesveld, J., Rosenberg, A.F., Darce, J., Gregoretti, I.V., Popova, L., Kaminiski, D., Fucile, C.F., Albizua, I., et al. (2015). 'Long-Lived plasma cells are contained within the CD19(-)CD38(hi)CD138(+) subset in human bone marrow. Immunity *43*, 132–145.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.

Hashimoto, K., Kouno, T., Ikawa, T., Hayatsu, N., Miyajima, Y., Yabukami, H., Terooatea, T., Sasaki, T., Suzuki, T., Valentine, M., et al. (2019). Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. Proc. Natl. Acad. Sci. USA *116*, 24242–24251.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). 'CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. Cell Rep. *2*, 666–673.

He, Y., Xiao, R., Ji, X., Li, L., Chen, L., Xiong, J., Xiao, W., Wang, Y., Zhang, L., Zhou, R., et al. (2010). EBV promotes human CD8 NKT cell development. PLoS Pathog. *6*, e1000915.

Horns, F., Dekker, C.L., and Quake, S.R. (2020). Memory B cell activation, broad anti-influenza antibodies, and bystander activation revealed by single-cell transcriptomics. Cell Rep. *30*, 905–913.e6.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res. *48*. D498–d503.

Jin, S., Zhang, L., and Nie, Q. (2020). scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biol. *21*, 25.

Jin, X., Meng, L., Yin, Z., Yu, H., Zhang, L., Liang, W., Wang, S., Liu, G., and Zhang, L. (2019). Characterization of dendritic cell subtypes in human cord blood by single-cell sequencing. Biophys. Rep. *5*, 199–208.

Juno, J.A., van Bockel, D., Kent, S.J., Kelleher, A.D., Zaunders, J.J., and Munier, C.M.L. (2017). Cytotoxic CD4 T cells-friend or foe during viral infection? Front. Immunol. *8*, 19.

Källberg, E., and Leanderson, T. (2008). A subset of dendritic cells express joining chain (J-chain) protein. Immunology *123*, 590–599.

Kaminski, D.A., Wei, C., Qian, Y., Rosenberg, A.F., and Sanz, I. (2012). Advances in human B cell phenotypic profiling. Front. Immunol. *3*, 302.

Kawamura, S., Onai, N., Miya, F., Sato, T., Tsunoda, T., Kurabayashi, K., Yotsumoto, S., Kuroda, S., Takenaka, K., Akashi, K., and Ohteki, T. (2017). 'Identification of a human clonogenic progenitor with strict monocyte differentiation potential: a counterpart of mouse cMoPs. Immunity *46*, 835–848.e4.

Kim, H.J., Lin, Y., Geddes, T.A., Yang, J.Y.H., and Yang, P. (2020). CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics *36*, 4137–4143.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Leonid Peshkin, Da., and Marc. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Kumar, B.V., Connors, T.J., and Farber, D.L. (2018). Human T cell development, localization, and function throughout life. Immunity *48*, 202–213.

Kuramasu, A., Saito, H., Suzuki, S., Watanabe, T., and Ohtsu, H. (1998). Mast cell-/basophil-specific transcriptional regulation of human L-histidine decarboxylase gene by CpG methylation in the promoter region. J. Biol. Chem. *273*, 31607–31614.

Lai, S., Xu, Y., Huang, W., Jiang, M., Chen, H., Ye, F., Wang, R., Qiu, Y., Jiang, X., Huang, D., et al. (2017). Mapping human hematopoietic hierarchy at single cell resolution by microwell-seq. Preprint at bioRxiv, 127217. https://doi.org/10.1101/127217.

Lawand, M., Déchanet-Merville, J., and Dieu-Nosjean, M.C. (2017). Key features of gamma-delta T-cell subsets in human diseases and their immunotherapeutic implications. Front. Immunol. *8*, 761.

Lee, J., Hyeon, D.Y., and Hwang, D. (2020). Single-cell multiomics: technologies and data analysis methods. Exp. Mol. Med. *52*, 1428–1442.

Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Res *5*.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746.

Mahnke, Y., Chattopadhyay, P., and Roederer, M. (2010). Publication of optimized multicolor immunofluorescence panels. Cytometry A. *77*, 814–818.

Martin, M.D., and Badovinac, V.P. (2018). Defining memory CD8 T cell. Front. Immunol. *9*, 2692.

Meehan, S., Kolyagin, G.A., Parks, D., Youngyunpipatkul, J., Herzenberg, L.A., Walther, G., Ghosn, E.E.B., and Orlova, D.Y. (2019). Automated subset identification and characterization pipeline for multidimensional flow and mass cytometry data clustering and visualization. Commun. Biol. *2*, 229.

Mehtonen, J., Teppo, S., Lahnalampi, M., Kokko, A., Kaukonen, R., Oksa, L., Bouvy-Liivrand, M., Malyukova, A., Mäkinen, A., Laukkanen, S., et al. (2020). Single cell characterization of B-lymphoid differentiation and leukemic cell states during chemotherapy in ETV6-RUNX1-positive pediatric

leukemia identifies drug-targetable transcription factor activities. Genome Med. *12*, 99.

Meyer, K.D. (2019). DART-seq: an antibody-free method for global m6A detection. Nat. Methods *16*, 1275–1280.

Mulè, M.P., Martins, A.J., and Tsang, J.S. (2021). Normalizing and denoising protein expression data from droplet-based single cell profiling. Preprint at bioRxiv. https://doi.org/10.1101/2020.02.24.963603.

Ocasio, J.K., Babcock, B., Malawsky, D., Weir, S.J., Loo, L., Simon, J.M., Zylka, M.J., Hwang, D., Dismuke, T., Sokolsky, M., et al. (2019). scRNA-seq in medulloblastoma shows cellular heterogeneity and lineage expansion support resistance to SHH inhibitor therapy. Nat. Commun. *10*, 5829.

Orlova, D.Y., Herzenberg, L.A., and Walther, G. (2018). Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets. Nat. Rev. Immunol. *18*, 77.

Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. *35*, 936–939.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods *10*, 1096–1098.

Poli, A., Michel, T., Thérésine, M., Andrès, E., Hentges, F., and Zimmer, J. (2009). 'CD56bright natural killer (NK) cells: an important NK cell subset. Immunology *126*, 458–465.

Qiu, P. (2020). 'Embracing the dropouts in single-cell RNA-seq analysis. Nat. Commun. *11*, 1169.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

Shlomchik, M.J., and Weisel, F. (2012). Germinal center selection and the development of memory B and plasma cells. Immunol. Rev. *247*, 52–63.

Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J.M., Blackburn, J., Barton, K., Roden, D., Luciani, F., Giang Phan, T., Junankar, S., et al. (2019). High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat. Commun. *10*, 3120.

Stoeckius, M., Christoph Hafemeister, W.S., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). 'Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21.

Stuart, T., and Satija, R. (2019). 'Integrative single-cell analysis. Nat. Rev. Genet. *20*, 257–272.

Tiller, T., Tsuiji, M., Yurasov, S., Velinzon, K., Nussenzweig, M.C., and Wardemann, H. (2007). Autoreactivity in human IgG+ memory B cells. Immunity *26*, 205–213.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning (Springer).

Utada, A.S., Fernandez-Nieves, A., Stone, H.A., and Weitz, D.A. (2007). Dripping to Jetting transitions in coflowing liquid streams. Phys. Rev. Lett. *99*, 094502.

Villani, A.C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science *356*.

Waltman, L., and Jan van Eck, N. (2013). A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B *86*.

Wang, X., Sun, Z., Zhang, Y., Xu, Z., Xin, H., Huang, H., Duerr, R.H., Chen, K., Ding, Y., and Chen, W. (2020). 'BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic Acids Res. *48*, 5814–5824.

Wong, E.B., Gold, M.C., Meermeier, E.W., Xulu, B.Z., Khuzwayo, S., Sullivan, Z.A., Mahyari, E., Rogers, Z., Kløverpris, H., Sharma, P.K., et al. (2019). 'TRAV1-2(+) CD8(+) T-cells including oligoconal expansions of MAIT cells are enriched in the airways in human tuberculosis. Commun. Biol. *2*, 203.

Xie, X., Liu, M., Zhang, Y., Wang, B., Zhu, C., Wang, C., Li, Q., Huo, Y., Guo, J., Xu, C., et al. (2020). 'Single-cell transcriptomic landscape of human blood cells. Natl. Sci. Rev. *8*, nwaa180.

Yang, J., Zhang, L., Yu, C., Yang, X.F., and Wang, H. (2014). 'Monocyte and macrophage differentiation: circulation inflammatory monocyte as biomarker for inflammatory diseases. Biomark. Res. *2*, 1.

Yermanos, A., Neumeier, D., Sandu, I., Borsa, M., Waindok, A.C., Merkler, D., Oxenius, A., and Reddy, S.T. (2021). 'Single-cell immune repertoire and transcriptome sequencing reveals that clonally expanded and transcriptionally distinct lymphocytes populate the aged central nervous system in mice. Proc. Biol. Sci. *288*, 20202793.

Yu, P., and Lin, W. (2016). Single-cell transcriptome study as big data. Dev. Reprod. Biol. *14*, 21–30.

Zemmour, D., Zilionis, R., Kiner, E., Klein, A.M., Mathis, D., and Benoist, C. (2018). 'Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. Nat. Immunol. *19*, 291–301.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). 'Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Zhou, Y., Zhang, Y., Han, J., Yang, M., Zhu, J., and Jin, T. (2020). 'Transitional B cells involved in autoimmunity and their impact on neuroimmunological diseases. J. Transl. Med. *18*, 131.

Zhu, J., Yamane, H., and Paul, W.E. (2010). 'Differentiation of effector CD4 T cell populations (*). Annu. Rev. Immunol. *28*, 445–489.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
TotalSeq-C0148 anti-human CD197 (CCR7) (clone G043H7)	BioLegend	Cat# 353251; RRID: AB_2800943
TotalSeq-C0390 anti-human CD127 (IL-7Ralpha) (clone A019D5)	BioLegend	Cat# 351356; RRID: AB_2800937
TotalSeq-C0161 anti-human CD11b (clone ICRF44)	BioLegend	Cat# 301359; RRID: AB_2800732
TotalSeq-C0081 anti-human CD14 (clone M5E2)	BioLegend	Cat# 301859; RRID: AB_2800736
TotalSeq-C0083 anti-human CD16 (clone 3G8)	BioLegend	Cat# 302065; RRID: AB_2800738
TotalSeq-C0050 anti-human CD19 (clone HIB19)	BioLegend	Cat# 302265; RRID: AB_2800741
TotalSeq-C0100 anti-human CD20 (clone 2H7)	BioLegend	Cat# 302363; RRID: AB_2800743
TotalSeq-C0181 anti-human CD21 (clone Bu32)	BioLegend	Cat# 354923; RRID: AB_2800953
TotalSeq-C0085 anti-human CD25 (clone BC96)	BioLegend	Cat# 302649; RRID: AB_2800745
TotalSeq-C0154 anti-human CD27 (clone O323)	BioLegend	Cat# 302853; RRID: AB_2800747
TotalSeq-C0034 anti-human CD3 (clone UCHT1)	BioLegend	Cat# 300479; RRID: AB_2800723
TotalSeq-C0389 anti-human CD38 (clone HIT2)	BioLegend	Cat# 303543; RRID: AB_2800758
TotalSeq-C0072 anti-human CD4 (clone RPA-4)	BioLegend	Cat# 300567; RRID: AB_2800725
TotalSeq-C0048 anti-human CD45 (clone 2D1)	BioLegend	Cat# 368545; RRID: AB_2801014
TotalSeq-C0063 anti-human CD45RA (clone HI100)	BioLegend	Cat# 304163; RRID: AB_2800764
TotalSeq-C0084 anti-human CD56 (NCAM) (clone QA17A16)	BioLegend	Cat# 392425; RRID: AB_2801024
TotalSeq-C0146 anti-human CD69 (clone FN50)	BioLegend	Cat# 310951; RRID: AB_2800810
TotalSeq-C0005 anti-human CD80 (clone 2D10)	BioLegend	Cat# 305243; RRID: AB_2800783
TotalSeq-C0006 anti-human CD86 (clone IT2.2)	BioLegend	Cat# 305447; RRID: AB_2800786
TotalSeq-C0080 anti-human CD8a (clone RPA-T8)	BioLegend	Cat# 301071; RRID: AB_2800730
TotalSeq-C0156 anti-human CD95 (Fas) (clone DX2)	BioLegend	Cat# 305651; RRID: AB_2800787
TotalSeq-C0159 anti-human HLA-DR (clone L243)	BioLegend	Cat# 307663; RRID: AB_2800795
TotalSeq-C0007 anti-human CD274 (B7-H1, PD-L1) (clone 29E.2A3)	BioLegend	Cat# 329751; RRID: AB_2800860

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Purified anti-human CD11c (clone S-HCL-3)	BioLegend	Cat# 371502; RRID: AB_2616755
Purified anti-human CD133 (Prominin-1) (clone 7)	BioLegend	Cat# 372802; RRID: AB_2728390
Purified anti-human CD138 (Syndecan-1) (clone MI15)	BioLegend	Cat# 356502; RRID: AB_2561790
Purified anti-human CD34 (clone 581)	BioLegend	Cat# 343602; RRID: AB_1732014)
Purified anti-human CD41 (clone HIP8)	BioLegend	Cat# 303702; RRID: AB_314372)
Purified anti-human CD43 (clone CD43-10G7)	BioLegend	Cat# 343202; RRID: AB_1659198
Purified anti-mouse/human CD45R/B220 (clone RA3-6B2)	BioLegend	Cat# 103202; RRID: AB_312987
Purified anti-human CD5 (clone UCHT2)	BioLegend	Cat# 300602; RRID: AB_314088)
Alexa Fluor 488 anti-human CD45 (clone HI30)	BioLegend	Cat# 368536; RRID: AB_2721364
PerCP/Cyanine5.5 anti-human CD38 (clone HIT2)	BioLegend	Cat# 303522; RRID: AB_893314
Alexa Fluor 647 anti-human CD14 (clone HCD14)	BioLegend	Cat# 325612; RRID: AB_830685
Alexa Fluor 647 anti-human CD16 (clone 3G8)	BioLegend	Cat# 302020; RRID: AB_492976
Alexa Fluor 700 anti-human IgM (clone MHM-88)	BioLegend	Cat# 314538; RRID: AB_2566615
Brilliant Violet 421 anti-human CD27 (clone M-T271)	BioLegend	Cat# 356418; RRID: AB_2562599
Brilliant Violet 570 anti-human CD16 (clone 3G8)	BioLegend	Cat# 302036; RRID: AB_2632790
Brilliant Violet 570 anti-human CD3 (clone UCHT1)	BioLegend	Cat# 300436; RRID: AB_2562124
Brilliant Violet 650 anti-human CD20 (clone 2H7)	BioLegend	Cat# 302336; RRID: AB_2563806
Brilliant Violet 785 anti-human CD19 (clone HIB19)	BioLegend	Cat# 302240; RRID: AB_2563442
ZombieUV Viability Dye	BioLegend	423107
PE Mouse Anti-Human CD34 (clone 8G12)	BD Biosciences	Cat# 348057; RRID: AB_400371
PE-Cy7 Mouse Anti-Human IgD (clone IA6-2)	BD Biosciences	Cat# 561314; RRID: AB_10642457
Biological samples		
Human Adult Peripheral Blood	Emory University's Children's Clinical and Translational Discovery Core	N/A
Human Adult Bone Marrow	Emory University; AllCells	N/A
Critical commercial assays		
Chromium Single Cell 5′ Library & Gel Bead Kit	10X Genomics	Cat# 1000006
Chromium Single Cell A Chip Kit	10X Genomics	Cat# 120236
Chromium Single Cell V(D)J Enrichment Kit, Human B Cell	10X Genomics	Cat# 1000016
Chromium Single Cell 5′ Library Construction Kit	10X Genomics	Cat# 1000020
Chromium Single Cell 5′ Feature Barcode Library Kit	10X Genomics	Cat# 1000080

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
EasySep Direct Human PBMC Isolation Kit	Stem Cell Technologies	Cat# 19654
EasySep Human B-Cell Enrichment Kit II without CD43 Depletion	Stem Cell Technologies	Cat# 17963
Custom oligonucleotide antibody conjugation	Expedeon	Custom
Deposited data		
Raw and analyzed data	This paper	GEO: GSE181543
Oligonucleotides		
/5AmMC12/CGGAGATGTGTATAAGA GACAGNNNNNNNNN-GGACGCAAC TTAAGA -NNNNNNNNNCCCATATAAGA*A*A	IDT	Custom
/5AmMC12/ CGGAGATGTGTATAAGAGACAGNN NNNNNNNN-GTGCAAGAGTTGGCG -NNNNNNNNNCCCATATAAGA*A*A	IDT	Custom
/5AmMC12/CGGAGATGTGTATAA GAGACAGNNNNNNNNNN-GAA GAAGCGTTATTC -NNNNNNNNNCCCATATAAGA*A*A	IDT	Custom
/5AmMC12/CGGAGATGTGTATA AGAGACAGNNNNNNNNNN-TAG TTGACATGCCAT -NNNNNNNNNCCCATATAAGA*A*A	IDT	Custom
/5AmMC12/CGGAGATGTGTATAAG AGACAGNNNNNNNNNN-CGAGGT ACATCTTGT -NNNNNNNNNCCCATATAAGA*A*A	IDT	Custom
/5AmMC12/CGGAGATGTGTATA AGAGACAGNNNNNNNNNN-CAC TCCTTGACAGGT -NNNNNNNNNCCCATATAAGA*A*A	IDT	Custom
/5AmMC12/CGGAGATGTGTATA AGAGACAGNNNNNNNNNN-GCC AAGATCAGGTCC -NNNNNNNNNCCCATATAAGA*A*A	IDT	Custom
3/5AmMC12/CGGAGATGTGT ATAAGAGACAGNNNNNNNNNN-GGACGCAACTTAAGA -NNNNNNNNNCCCATATAAGA*A*A	IDT	Custom
Software and algorithms		
FlowJo v10.6	BD Biosciences	https://www.flowjo.com/solutions/flowjo/downloads/
Cell Ranger v3.1.0	10X Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/3.1/what-is-cell-ranger
Seurat v4.0	Hao et al. 2021	https://github.com/satijalab/seurat/
DSB	Mulè et al. (2021)	https://github.com/niaid/dsb
Reactome pathway database	Jassal et al., (2020)	https://reactome.org/PathwayBrowser/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
scDblFinder	Germain et al., (2021)	https://github.com/plger/scDblFinder
Cell Fidelity Statistic (CFS) scores	Babcock et al., (2021)	https://github.com/Ghosn-Lab/BatchNorm
Custom scripts	This paper	Zenodo: https://doi.org/10.5281/zenodo.7045077

RESOURCE AVAILABILITY

Lead contact

Further information and request for resources should be directed to and will be fulfilled by the lead contact, Eliver E.B. Ghosn (eliver.ghosn@emory.edu)

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Single-cell RNA-seq data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact on request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects and specimen collection

Peripheral blood samples (n = 3; 2F/1M) were collected from healthy adult donors through Emory University's Children's Clinical and Translational Discovery Core (CCTDC). Bone marrow samples (n = 2; 1M/1F) were collected from healthy adult donors through Emory University Hospitals under IRB00066294 or obtained from AllCells (Alameda, CA). Subjects were between 20 and 50 years of age. All subjects provided written informed consent before sample collection.

METHOD DETAILS

Cell preparation

Peripheral blood mononuclear cells (PBMCs) were isolated from peripheral venous blood using EasySep™ Direct Human PBMC Isolation Kit (Cat # 19654). B cells were enriched from hip bone marrow (BM) using EasySep™ Direct Human PBMC Isolation Kit (Cat # 19654) followed by EasySep Human B-Cell Enrichment Kit II without CD43 Depletion (Cat # 17963) following manufacture's protocols. Up to 1×10^6 cells per donor were incubated with Fc block (Miltenyi) on ice for 10 min, followed by staining with a mix of 31 oligo-conjugated (barcoded) antibodies (see Table S1 for full list) for 30 min on ice in custom-made RPMI-1640 media (def-RPMI-1640; deficient in biotin, L-glutamine, phenol red, riboflavin, and sodium bicarbonate), and containing 3% newborn calf serum , followed by two washes in def-RPMI-1640/0.04% BSA. Cells were resuspended at a concentration of 1200–1500 cells/uL in custom RPMI-1640/0.04% BSA and passed through a 20-um nylon filter before loading onto a Chromium Controller (10X Genomics, Pleasanton, CA).

Single-cell RNA-sequencing

Cells were loaded to target encapsulation of 6,000 cells. Gene expression (GEX), Antibody-derived tag (ADT), and V(D)J libraries were generated using the Chromium Single Cell 5′ Library & Gel Bead Kit v1 with feature barcoding (10X Genomics, Pleasanton, CA) following the manufacturer's instructions. Gene expression libraries were pooled and sequenced on a NovaSeq 6000 platform (Illumina, San Diego, CA). ADT and V(D)J libraries were sequenced separately on a Next-seq platform (Illumina, San Diego, CA). The sequencing depths are shown in Table S2.

Multi-omics single-cell data preprocessing

10X Genomics Cell Ranger v3.1.0 was used to perform barcode processing and single-cell 5′ unique molecular identifier (UMI) counting. Reads from GEX and ADT libraries were processed simultaneously using "cellranger count," whereas reads from V(D)J libraries were aligned by running "cellranger vdj."

Gene expression (GEX)

The scRNA-seq expression datasets were integrated into two final matrices, one for PBMC and another for BM. To remove potential batch effects, we used the canonical correlation analysis (CCA) to integrate the three individual PBMC samples into one final PBMC dataset and combine the two individual BM samples into one final BM dataset. Before integration, we performed library-size scaling and log-transformation for each sample individually. Next, we identified the top 2,000 HVGs for each sample using regularized negative binomial regression (Hafemeister and Satija 2019), followed by CCA to find anchors and integrate cells from individual samples into one integrated dataset. Finally, we performed per-gene z-score normalization for each integrated dataset. We removed from the downstream analysis the genes expressed in less than three cells, the cells with less than 200 genes, and the cells with more than 20% of mitochondrial gene counts. GEX raw count matrices are available through NCBI GEO, accession number GSE181543.

Antibody-derived tags (ADT)

The cell-surface protein expression (ADT) matrices were normalized separately for each individual sample using the DSB normalization (Mulè et al. 2021). Cell barcodes in the Cell Ranger unfiltered matrix containing less than 40 genes and less than 100 total UMIs were considered background. The barcodes from Cell Ranger "filtered_feature_bc_matrix" were considered true cells. By subtracting the mean value of the background population and then regressing out the cell-cell technical variation (Mulè et al. 2021), the ADT values representing background expression levels were centered around zero. ADT raw count matrices are available through NCBI GEO, accession number GSE181543.

Antibody repertoire (VDJ)

We used the Cell Ranger-generated "all_contig_annotations.csv" (unfiltered) and the "filtered_contig_annotations.csv" files to generate new V(D)J features merged with the corresponding ADT matrix. Briefly, we summed the total UMIs of each cell barcode in the "all_contig_annotations.csv" V(D)J file to generate a feature called "Total Ig Transcripts" that represent the total UMIs of immunoglobulin heavy and light chains. In addition, we generated a binary feature called "Productive VDJ" that identifies whether a cell barcode with a productive V(D)J information is contained in the CellRanger "filtered" annotation file and hence considered a true B cell or plasma cell. B-cell receptor repertoire V(D)J raw count matrices are available through NCBI GEO, accession number GSE181543.

Manual gating

Manual gating was performed based on the two V(D)J features (i.e., Ig-specific transcript counts and productive VDJ sequences) and the 31 cell-surface protein features in the DSB normalized ADT data. Because the DSB normalized data showed similar distribution within the same tissues, the 3 PBMC samples and 2 BM samples were concatenated, respectively. The major cell lineages (six for PBMCs and five for BM cells) were then identified and manually gated using a customized strategy of biaxial plots implemented in a MATLAB script (https://github.com/Ghosn-Lab/SuPERR.git). The markers and the gating hierarchy were determined by prior knowledge of variations of cell-surface protein expression across various major cell lineages (Figures 2A and 3A). Cell doublets were manually "gated out" based on the co-expression of two or more major lineage markers (Figures S1 and S2).

Clustering on gene expression matrix

Each major cell lineage identified by the manual gating was next clustered based on the integrated GEX matrix (Stuart et al., 2019). First, HVGs were computed for cells in each major cell lineage using the variance-stabilizing transformation (VST) method (Hafemeister and Satija 2019), and expression data of each gene was transformed to zero mean and unit variance. We then performed PCA on the HVGs of each main lineage and selected the top 30 PCs for downstream analysis. Next, a K-nearest neighbor (KNN) graph was constructed in the low-dimensional PCA space based on the Euclidean distance between cells, with K = 30. Jaccard further converted the KNN graph to a cell-cell similarity matrix, followed by Louvain community detection algorithm (Blondel et al., 2008) to define cell clusters in each major cell lineage.

The weighted nearest neighbor (WNN) approach as implemented in Seurat v4 and the similarity network fusion (SNF) as implemented in CiteFuse were separately applied to generate cell clusters cells based on an intergrated matrix containing both GEX and ADT data. For Seurat v4, we used the graph-based smart local moving (SLM) algorithm (Waltman and Jan van Eck 2013), and set the clustering resolution to 3 to generate as many (or more) clusters as SuPERR. For CiteFuse, cells from different samples were first integrated using CCA as described above and then the integrated dataset was used as input. We applied the spectral clustering algorithm as suggested by the authors and the K was set to 39 for PBMC and 35 for BM to align the number of cell clusters generated by SuPERR.

Pathway analysis

The DGE list for each BM plasma cell cluster was used as input data to the Reactome pathway database (Jassal et al., 2020) (https://reactome.org/PathwayBrowser/) to visualize active biological pathways and regulatory processes in each plasma cell cluster. The "firework" diagrams were cropped so that only the hit pathways were preserved.

Cell-doublet identification

The scDblFinder (Germain et al., 2021) workflow was run independently on the GEX matrix from the PBMC and BM samples, following the workflow described on github (https://github.com/plger/scDblFinder). First, we performed log-normalization, selection of HVGs and PCA analysis on the GEX matrix for each sample. Next, we ran the scDblFinder pipeline using the default parameters.

Calculation of cell-type scores

Each cell-type specific score was calculated by summing the raw UMI counts from the gene list we generated based on prior knowledge and log-normalized the data. The following gene lists were used for each cell-type score. *B cells*: CD79A, MS4A1, CD19, VPREB3; *CD4T*: CD4, CD3D, CD3E, CD5, IL7R; *CD8T*: CD8A, CD8B, GZMK, CD3D, CD3E; *Myeloid*: LYZ, S100A8,S100A9, S100A12, CD68, CD14, CYBB; *NK*: GNLY, NKG7, GZMB, KLRD1, GZMA; *NKT*: CCL5, GNLY, NKG7, GZMH, KLRB1, CST7; *Treg*: FOXP3, CTLA4, IL2RA, IL32; *T cells*: CD3D, CD3E, CD4, CD8A, CD8B, CD5, IL7R, GZMK, GZMH; and *PC* (Plasma cell score): ITGB7, IRF4, CD9, PRDM1, XBP1, SDC1, VCAM1, CD38.

Cell fidelity statistic

To generate the cell fidelity statistic (CFS) score (Babcock et al., 2021), we cross-check the cell lineages identities generated by a biaxial gating approach (reference) versus cell cluster assignment (test). CFS relies on the tenet that if different workflows identify a single cell as belonging to multiple lineages, both cannot be correct and therefore some information has been lost. We apply CFS to measure the loss in cell classification fidelity caused by relying on clustering algorithms to discriminate cell lineages, as is the norm in conventional workflows. To generate a CFS score, we first produce clusters (Seurat v3, Seurat v4, CiteFuse). We then count the number of cells which are grouped into majority out-of-lineage clusters. The CFS score is expressed as a fraction of total cells which were grouped into the inappropriate lineage, or the cells which have a disagreement between cluster assignment and biaxial gate. CFS is reported as a fraction from 0-1, where 1 means that no cells were misclassified, and a score of 0.8 means that 20% of cells were misclassified.

QUANTIFICATION AND STATISTICAL ANALYSIS

Differential gene expression analysis

We used the "FindAllMarkers" function of Seurat v3 to perform the differential gene expression analysis. Statistical significance was tested using the Wilcoxon Rank-Sum test, with p_val_adj<0.05 (p-value after Bonferroni correction). For the DGE analysis of the BM plasma cell clusters, we first removed the immunoglobulin-specific UMIs from the GEX matrix. Then, we log-normalized the resulting matrix to discover and explore non-immunoglobulin plasma cell genes.

Student's t-test

To quantify the difference of biological signals (total gene count, total UMI count and percentage of ribosomal UMI) among major cell types, student's t-test was performed that compared the mean of each cell type with the mean of the total PBMC/BM. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, unpaired, two-tailed.