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

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

[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

[†] Contributed equally

* Corresponding authors:

Eliver E.B. Ghosn | e-mail: eliver.ghosn@emory.edu

Peng Qiu | e-mail: peng.qiu@bme.gatech.edu

Abstract

Single-cell transcriptomics enables the definition of diverse human immune cell types across multiple tissues and disease contexts. Still, 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-seq, a novel analysis workflow to increase the resolution and accuracy of clustering and allow for the discovery of previously hidden cell subsets. SuPERR-seq 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 in the human immune system, including a novel subset 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[1-3], library preparation[4,5], and sequencing methods[1,4,6,7]. These increasingly more widely adopted technologies have significantly improved the understanding of cell heterogeneity in health and disease[4-8]. However, reliance on cellular transcriptomics alone limits the comprehensive identification of heterogenous cell populations[9]. 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[7], REAP-seq[10], and others[11], simultaneously measure gene expression (mRNA) and cell surface proteins. However, additional heterogeneity of immune cell subsets can be revealed by combining single-cell gene expression with simultaneous TCR/BCR repertoire sequencing using techniques such as RAGE-seq and DART-seq[12-16]. Moreover, we show that simultaneous measurement and comprehensive integration of transcriptomics, cell-surface protein, and cell-receptor repertoire will provide improved resolution to reveal heterogeneous cell types relevant to disease mechanisms and healthy homeostasis.

However, multi-omics technologies also present computational challenges for data integration and analysis[17-19]. Challenges include high dimensionality of the data[20], sparsity of the data[21], diversity across various omics data types[22], and technical effects between different sample batches[23]. Several algorithms have been developed to integrate and analyze multi-omics measurements, including linear regression models[24] and multi-omics factor analysis (MOFA)[25]. Correlated features across different -omics have been exploited by both LIGER[26] and canonical correlation analysis (CCA), as employed in the Seurat pipeline[22]. 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. While

data-driven methods have been successful for clustering and identifying cell types, significant improvements can be made by incorporating robust prior knowledge, such as well-established marker genes and protein surface markers that define cell types.

This study combined our extensive expertise on high-dimensional flow cytometry data analysis[27] with our multi-omics single-cell data sets to develop the SuPERR-seq (**Su**rface **P**rotein **E**xpression, m**R**NA- and **R**epertoire-seq) workflow. SuPERR-seq is a novel, biologically motivated approach towards the integration and analysis of multi-omics single-cell data matrices. By combining a flow cytometry-derived “gating strategy” with the high-dimensional analysis of scRNA-seq, SuPERR-seq increases the resolution and accuracy in clustering algorithms and allows the discovery of new biologically relevant cell subsets. We first applied the “gating strategy” based on a combination of cell-surface markers and immunoglobulin-specific transcript counts to identify major immune cell lineages. Next, we explore 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-seq workflow to multi-omics data sets (cell-surface protein, gene expression, VDJ repertoire) generated in-house from human blood and bone marrow cells and directly compare its performance to existing methods. We show that SuPERR-seq can leverage the power of each “omics” to identify major immune lineages more accurately and a greater number of cell subsets 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 individual healthy adult donors. As shown in Fig. 1A, we simultaneously captured the following three omics: total gene expression (GEX), 31 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-seq analysis pipeline can be described in two major steps shown in Fig. 1B: 1) manual biaxial gating based on the expression levels of cell-surface proteins (ADT) and total immunoglobulin (Ig) transcript counts, and 2) sub-clustering using the GEX matrix for each manually-gated population.

For the first step of the SuPERR-seq pipeline, we normalized the cell-surface (ADT) data using the DSB normalization method[28], followed by concatenating the normalized ADT matrix with the total Ig-specific UMI counts from the VDJ matrix. 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 sequential gating strategy on biaxial plots similar to conventional flow cytometry data analysis (Figs. 2A and 3A). As antibody-secreting cells (ASC), also known as plasma cells, produce and secrete large quantities of immunoglobulin, they could be accurately identified based on their Ig-specific transcript counts (Figs. 2A and 3A). Of note, the SuPERR-seq pipeline was able to readily identify a rare cell cluster containing as little 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 workflows.

In the analysis of PBMC samples, we defined gates for six major immune cell lineages:

plasma cells, B cells, NK/NKT cells, Monocytes, CD4+ T cells, and CD8+ T cells (Fig. 2A, black borders). In the analysis of the BM samples, we defined gates for five major lineages: CD138- Plasma cells, CD138+ Plasma cells, B cells, Myeloid cells, and Hematopoietic Stem and Progenitor Cells (HSPCs) (Fig. 3A, black borders). In addition to these main cell lineages, our gating strategy revealed additional cell sub-clusters (Figs. 2A and 3A, gray borders), which we gated to validate downstream sub-clustering based on the GEX data. These results were confirmed by well-established high-dimensional flow cytometry analysis on a separate aliquot of the same samples taken before single-cell encapsulation (Fig. S4).

Identifying the *major cell lineages* as the first step of the SuPERR-seq pipeline also allows for accurate cell-doublet discrimination. We applied flow cytometry-style biaxial feature plots on the ADT data to identify and remove cell doublets (Figs. S1 and S2). Cell barcodes containing surface marker signals known to belong to more than one immune lineage (e.g., co-expression of B-cell/CD19 and T-cell/CD3 markers) were considered doublets and removed from the downstream analysis. Figs. S1 and S2 illustrate the gating strategy used for removing doublets from the PBMC and BM data sets, respectively.

Notably, by assigning major cell lineages via a manual-gating approach before performing the principal component analysis (PCA) of the GEX matrix, we determined that much of the data variation observed in scRNA-seq is cell-lineage specific (Fig. 4). Key variables, including ribosomal transcript content, gene dropout rates, mRNA abundance, and total unique gene counts, vary significantly among the cell lineages (Fig. 4, ANOVA $p<2.2e-16$). We show that we can capture biologically meaningful gene variation that occurs only *within* each pre-defined major lineage by gating major cell lineages before assessing variation in the GEX matrix. For example, the top 30 highly-variable genes (HVGs) of total B cells are not present in the top 300 HVGs of the ungated total PBMC sample. (Fig. 4D). In contrast, conventional workflows capture a range of genes dominated by variations occurring mostly *between* cell

lineages (Fig. 4D, right panel), giving less weight to genes that would otherwise reveal further heterogeneity *within* each major lineage. Hence, pre-selection of major cell lineages before GEX clustering enables targeted selection of HVGs that are unique to the biology of each major cell lineage, resulting in the identification of more and novel cell subsets.

SuPERR-seq reveals greater heterogeneity within major immune lineages

***Human Peripheral Blood*.** Gating for major cell lineages revealed six distinct populations within the PBMC sample (Fig. 2A, black outlines). We then further explored each major lineage by generating 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 neighbors' graph (KNN, K = 30), followed by Louvain community detection (see Methods section for detailed description). Following these steps, we obtained the subclusters for each major lineage, herein called SuPERR-seq clusters. At a Louvain resolution of 0.8, SuPERR-seq identified 38 clusters in human PBMC, with each major lineage broken out into two or more subclusters, representing sub-lineage heterogeneity (Fig. 2B).

We further investigated the 38 SuPERR-seq clusters by exploring the expression levels of cell-surface proteins (ADT matrix) (Figs. 2C and 2D). Some ADT markers were lineage-specific (e.g., CD19 used to classify B cells and Ig-specific transcript counts to classify plasma cells) and were previously used to gate major cell lineages and confirm SuPERR-seq 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-seq cluster identification (Figs. 2D and S4-S9). We confirmed the heterogenous clusters as biologically meaningful using established cell-surface lineage markers. For example, four of the nine NK subclusters were identified as NKT cells based on the co-expression of NK markers and CD3. Similarly, the monocytes were divided into three classical and two non-classical subtypes (Fig. S4). SuPERR-seq was also able to accurately identify T-regulatory cells (Treg) as a subtype of CD4+ T cells based on their high expression of CD25 (ADT matrix) and FOXP3 (GEX matrix), and low expression of CD127 (ADT matrix) (Figs. 2A and S5).

Remarkably, SuPERR-seq readily identified a cluster of rare plasma cells (antibody-secreting cells) in the blood even though we captured only eight plasma cells within this cluster (Fig. 2). This level of resolution and accuracy in identifying rare plasma cells was possible by analyzing the total Ig-specific transcript counts from the VDJ matrix. As plasma cells produce large quantities of immunoglobulin transcripts, they were readily identified based on a ~2.5 $\log_{10}$-fold increase in total Ig-specific transcripts than B cells (Fig. 2A).

Five sub-clusters were identified within the B-cell major lineage, all of which could be mapped back to previously described B-cell subsets[29,30] (Fig. 2D). Specifically, clusters 1 and 2 lack surface CD27, a canonical B-cell memory marker[31], and the presence of IGHD and IGHM (IgD and IgM transcripts, respectively) identify these clusters as naïve B cells[32] (Fig. 2D). Clusters 3 and 4 (CD27+/TACI+) are memory B cells. CD95 expression and the absence of IGHM further characterize cluster 3 as class-switched memory B cells[32]. Finally, cluster 5 is unswitched (IgM+) memory B cells (Fig. 2D). All other major lineages were classified using the same approach, and the results are described in Table 1.

Human Bone Marrow. SuPERR-seq identified 34 unique clusters in human BM cells (Figs. 3A and 3B). Of note, SuPERR-seq matched the B-cell subclusters to the different stages of their developmental pathway[33] (Fig. 5). The cell-surface expression of CD10, along 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 were defined by their surface expression of CD21[34]. Finally, the mature class-switched memory B cells were identified by the expression of IGHA1/IGHG1 immunoglobulin and surface CD27[35] (Fig. 5). The SuPERR-seq workflow also was able to identify the developmental pathway of neutrophils, monocytes, and erythrocytes starting from the hematopoietic stem cells (HSC) (Figs. S8 and S9). The classification results for the other subclusters are summarized in Table 2.

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

The SuPERR-seq pipeline readily and unambiguously identified four biologically distinct subsets of human antibody-secreting cells (a.k.a., plasma cells) in the adult bone marrow (BM) of healthy donors (Figs. 3 and 6).

The determinant feature of a plasma cell is its ability to produce and secrete large quantities of immunoglobulins (Ig) (i.e., antibodies). The SuPERR-seq 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 (Fig. 3A). Unlike the GEX matrix, the V(D)J matrix provides a more accurate count of total Ig-specific transcripts produced by plasma cells because the V(D)J data matrix is generated from a separate library using Ig-specific primers[8].

Remarkably, the number of total Ig-specific transcripts (Ig-specific UMI counts) detected in plasma cells is ~2.5 $\log_{10}$-fold higher than in B cells (Figs. 3A and 3C), allowing for unambiguous and straightforward identification of total plasma cells. Next, we separated the

total plasma cells into two subsets, based on the cell-surface protein expression (ADT matrix) of CD138, I canonical plasma-cell marker expressed by some, but not all, BM plasma cells[29,36]. Finally, we used the GEX data matrix to further subcluster the CD138+ and CD138- plasma cell subsets based on their transcriptomic profile, revealing a total of 4 distinct subsets of human plasma cells (Figs. 3 and 6).

To facilitate the identification of the differential gene expression (DGE) that distinguish each of the four plasma cell subsets, we first removed from the GEX matrix all the mRNA transcripts (UMI) derived from immunoglobulin genes (immunoglobulin genes represent more than 50% of the total transcripts/UMI recovered from I plasma cell). We then log-normalized the immunoglobulin-depleted GEX matrix and performed DGE analysis using the Wilcoxon Rank-Sum Test. The resulting differentially-expressed genes readily defined unique biological processes for each plasma cell cluster (Fig. 6). Cluster 1: CD138-/Ig^{hi}/PRDM1-; cluster 2: CD138-/HLA-DR^{hi}/SDC1^{lo}/G2-M phase; cluster 3: CD138+/IRF4^{hi}; cluster 4: CD138+/XBP1^{hi}/SDC1^{hi} (Figs. 6A and 6C). Clusters 2 and 4 show characteristics similar to previously identified human BM plasma cell subsets, in which they were described as Fraction I and Fraction B, respectively[36]. Notably, cluster 1 represents a unique plasma cell subset, in that >75% of its total mRNA transcripts represent immunoglobulin genes (Fig. 6B). This large proportion of immunoglobulin gene transcripts suggests these cells are allocating most of their metabolic activity to produce and secrete antibodies. Indeed, pathway analysis (Reactome Pathway Database)[37] using the DGE list for cluster 1 revealed signals mainly for the metabolism of proteins (Fig. 6F).

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[36]. 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). Surprisingly, we found that not all plasma cell clusters express the canonical plasma cell genes. Of note, cluster 1 does not express IRF4 and PRDM1. The absence of these canonical genes in cluster 1 was not due to the overrepresentation of immunoglobulin (Ig) genes because IRF4 and PRDM1 still were not detected after normalizing the GEX matrix without the immunoglobulin genes (Fig. 6C).

By comparing the immunoglobulin isotypes and subclasses (IgM, IgD, IgG1-4, IgA1-2, IgE) expressed by each plasma cell subset, we found that cluster 2 contains plasma cells of multiple isotypes/subclasses. In contrast, cluster 1 is more homogeneous, containing mainly IgG1 and IgA1 (Fig. 6D). Furthermore, we observed that cluster 1 is composed of plasma cell repertoire clones (defined by their BCR CDR3 amino acid sequences) shared among cluster 3 and cluster 4 (Fig. 6E). Finally, pathway analysis (Reactome Pathway Database) revealed unique biological processes and genetic programs that define each plasma cell subset (Fig. 6F). Notably, cluster 4 expresses genes involved in the cell cycle and programmed cell death, while cluster 3 appears to be actively responding to environmental stimulation.

In sum, the SuPERR-seq pipeline 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 types as done by existing approaches in the literature. In the following section, we provide specific examples of how the SuPERR-seq workflow outperforms existing approaches.

Benchmarking of SuPERR-seq against conventional methods reveals reduced cell-type misclassification, and superior ability to resolve cell doublets and classify new cell types

As we described above**,** the first step of the SuPERR-seq workflow, in which we use cell-surface proteins in sequential biaxial plots to identify major cell lineages, also allows for accurate cell-doublet discrimination (Figs. S1, S2, 7A, and 7B). Our SuPERR-seq approach successfully identified and removed 370 cell doublets. In contrast, the conventional approach[38] of trimming cells with very high mRNA transcripts (i.e., cell barcodes with greater than the mean UMI value + 4 standard deviation) as potential cell doublets identified only 42 doublets in PBMC. For BM, the SuPERR-seq approach identified 108 cell doublets, and the conventional approach identified only one cell doublet. (Fig. 7A). Strikingly, every plasma cell we identified in PBMCs would have been trimmed/removed by the conventional approach of removing cells with mean + 4 SD UMI counts (Fig. 7A). However, unlike the standard/conventional approach, the SuPERR-seq workflow readily recognized the PBMC plasma cells as single cells (confirmed by their unique VDJ repertoire) containing high UMI counts (Fig. 7A). Thus, the true cell doublets removed by SuPERR-seq would otherwise be missed by conventional approaches and erroneously included as single cells in downstream GEX analysis, as visualized on the UMAP plot (Fig. 7C).

We further compared our approach with current computational methods explicitly designed to remove cell doublets in scRNA-seq datasets. First, we tested the standard workflow of DoubletFinder[39] and compared it with the cell doublets defined by the SuPERR-seq method (Fig. 7D). Although DoubletFinder outputs a similar number of cells compared to the SuPERR-seq method, the results of the two methods did not overlap, mainly due to false positives and false negatives from DoubletFinder. Notably, SuPERR-seq identified hundreds of true cell doublets (i.e., cell barcodes that co-express markers of more than one cell lineage) missed by the DoubletFinder workflow (Fig. 7D). Interestingly, DoubletFinder also classified PBMC plasma cells as cell doublets due to their high UMI counts. However, we confirmed that PBMC plasma cells represent singlets by both their surface markers (ADT matrix) and their unique V(D)J

repertoire (VDJ matrix).

Next, we compared our SuPERR-seq pipeline with the commonly used Seurat package V3[40] using default parameters. For the PBMC sample, Seurat identified 24 clusters compared to 38 clusters identified by the SuPERR-seq pipeline (Fig. S10). Seurat workflow identified 22 clusters for the BM sample compared to 34 clusters identified by the SuPERR-seq pipeline (Fig. S11). When we directly compared the cluster assignment of the same cell barcodes from each pipeline (Figs. S10 and S11), multiple Seurat clusters were further divided by the SuPERR-seq workflow into several biologically meaningful clusters (Figs. 5, 6, S4-S9). The ability of the SuPERR-seq pipeline 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 Fig. 4, pre-gating major lineages reveal the gene variation that occurs only *within* the pre-defined lineage instead of gene variation observed *across all* major lineages. The select list of highly variable genes (HVGs) that occur only within the pre-defined lineages generates clusters of biologically meaningful cell types dominated by lineage-specific genes (Fig. 4D).

Importantly, our benchmarking analysis reveals a significant number of cell misclassifications (i.e., a cluster containing cells from different major lineages) generated by the standard Seurat workflow. For example, the Seurat workflow generated three clusters (5, 9, and 15) (Fig. S10) containing a mixture of cell lineages, including CD8+ T cells and NK cells. In contrast, SuPERR-seq correctly clustered and classified these cells separately. Similarly, Seurat cluster 3 mixed CD4+ and CD8+ T cells, while SuPERR-seq correctly identified and separated these different cell types (Fig. S10). Notably, such cell misclassification artifacts are not rare and commonly occur when clustering cells in high-dimensional space using the GEX data matrix as done by most, if not all, scRNA-seq analysis workflow[41,42].

To quantify the extent of cell misclassification in conventional data analysis approaches,

we developed a new scoring system named Cell Misclassification Statistics (CMS) score[43]. Briefly, we consider the biaxial feature plots on the cell surface (ADT) protein data (Figs. S1 and S2) to represent a “gold standard,” as this iterative nature of SuPERR-seq clustering and cell-type assignment 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. We then compare these cell-type identities to those from a standard Seurat workflow and consider the proportion of cells that change classification. This proportion is reported as a “CMS score,” which provides a statistical measure of uncertainty in the cell-type assignment steps of the Seurat workflow. For example, the PBMC and BM clusters generated by Seurat showed a CMS score of 0.0984 and 0.0594, respectively, indicating that 9.84% of PBMCs and 5.94% of BM cells were misclassified by the conventional approach (Fig. 8A).

To determine whether cell misclassification in Seurat clusters could confound interpretation of downstream differential gene expression (DGE) analysis, we explored the PBMC cluster 5 generated by Seurat (Fig. 8C). DGE analysis is a commonly used method to interpret biological differences between clusters or cell types[44]. We hypothesized that running DGE analysis on a cluster with contaminating cell types could generate misleading results even at low contamination numbers. For example, the Seurat PBMC cluster 5 contains mainly NK cells, but it is also contaminated/mixed with NKT and T cells. Indeed, by analyzing the DGE list of cluster 5 before removing the contaminating cells (NKT and T), the *TRDV1* gene appeared as highly expressed for this cell population (Fig. 8C). However, *TRDV1* is expressed on gamma-delta T cells[45], not on NK cells. These results suggest that widespread cell misclassification often observed in conventional data analysis workflow can confound data interpretation.

Finally, when comparing the new SuPERR-seq-identified plasma cells with conventional Seurat workflow, we observed major differences. While SuPERR-seq readily and accurately identified a cluster of eight plasma cells in PBMCs (Fig. 2A), the conventional workflow fails to

identify these rare cells. The inability of conventional approaches to identifying these rare plasma cells can be explained by: a) the conventional approach of removing cell doublets based on total UMI counts incorrectly removes plasma cells due to their naturally high mRNA content (Fig. 7A), and b) even when plasma cells are not removed as cell doublets and instead included in downstream clustering analysis, the plasma cells are assigned to a cluster of mixed cell types (Fig. S10B).

Unlike the PBMC, BM cells contain significantly more plasma cells. In this case, SuPERR-seq identified four biologically distinct plasma cell subsets, while Seurat workflow identified one cluster (Figs. 6 and S11). Of note, the plasma cell cluster generated by the Seurat workflow contained a mixture of cell types, including B cells and HSPCs (Fig. S11B). In addition, the Seurat workflow misclassified a fraction of the plasma cells identified by the SuPERR-seq based on Ig-specific transcripts (VDJ matrix). As a result, they spread across multiple clusters on the UMAP (Fig. S11B). To determine whether the conventional workflow could still identify the four SuPERR-seq plasma cell subsets, we performed a second iteration of clustering (re-cluster) on the original Seurat plasma cell cluster. Surprisingly, even after the second iteration of clustering, the conventional workflow was able to identify only three BM plasma cell subsets (Fig. S11C). More importantly, the three plasma cells clusters identified by the conventional approach remained a mixture of B cells and HSPCs (Fig. S11C). For the fairness of comparison, the parameters used for running the Seurat and SuPERR-seq workflows were set as the default parameters described in Methods. And the re-clustering approach on the original Seurat clusters was limited to two iterations.

Taken together, the results from the benchmarking analysis show that the SuPERR-seq 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 any single-cell RNA-seq analysis is the identification and classification of cell types. The most popular and widely-used approach relies on principal component analysis (PCA) derived from a set of highly variable genes (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.) (Fig. 4). Thus, an improved method to select HVGs *within* each cell lineage is needed to discover new biologically meaningful cell types.

Moreover, selecting genes 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 misclassification (Fig. 8). 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[41,42], inadvertently misguiding biological interpretations (Fig. 8, see CMS scores). Another contributing factor to cell 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[2,71]. Current approaches to remove cell doublets bioinformatically simply remove droplets that contain very high mRNA transcript counts (i.e., outliers)[22]. However, as we show here, many cell doublets contain average mRNA counts and can mistakenly be carried

over to downstream analysis (Fig. 7).

To address these pitfalls of existing methods, we presented here the SuPERR-seq pipeline. In SuPERR-seq, we simultaneously apply information gained from triple-omics sequencing, namely gene expression (GEX), surface-protein markers (ADT), and antibody production/repertoire (VDJ). As a quality control measure and to prevent downstream misclassification, the first step of SuPERR-seq is to perform a “manual gating” similar to standard flow cytometry analysis. ADT and V(D)J matrices represent cell-surface protein and antibody production measurements, respectively, that can reliably classify cells into clearly distinct major lineages by canonical markers. This gating step ensures that any subsequent cell clusters generated are composed of a single lineage, allowing for a more accurate investigation of clusters representing cell types. The second step of the SuPERR-seq 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 shows that the SuPERR-seq approach reveals additional cell types (and likely cell states) compared to a conventional approach. We reason that this is mainly due to the ability of SuPERR-seq 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.

We recognize the difficulty and expense of generating triple-omics data for every sample, as well as the limitations of applying the SuPERR-seq approach to older data sets 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 matrix. Thus, to broaden the application of SuPERR-seq principles to datasets without triple-omics available, we propose a gene-based (GEX) recursive analysis approach, which carries some of the same benefits of SuPERR-seq.

We applied this approach to the same BM dataset explored by SuPERR-seq and identified improvements to exisiting approaches and limitations. Focusing on the plasma cells, the recursive application of Seurat, which only identified one contaminated (misclassified) plasma cell cluster in the first iteration, now generated three sub-clusters of plasma cells. However, this recursive application of Seurat was not able to accurately isolate the plasma cells from the cell contaminants. The three resulting sub-clusters of plasma cells represented a mixture of plasma cells, B cells, and HSPCs (Fig. S11C).

In contrast, SuPERR-seq readily and accurately identified four subsets of plasma cells that were confirmed based on the three omics (GEX, ADT, VDJ), including the high antibody transcript counts (~2.5 $\log_{10}$-fold higher antibody transcripts than B cells). The additional plasma cell subset identified by SuPERR-seq is of great biological importance, as revealed by the Reactome pathway analysis generated on the set of DGE that distinguishes each plasma cell subset (Fig. 6). Therefore, integrating three omics in the SuPERR-seq provides unique information that recursive strategies for defining HVGs cannot achieve. This conclusion agrees with the latest implementation of Seurat Version 4, which also integrates ADT measurements in the analysis pipeline[22], further supporting the SuPERR-seq approach of multi-omics integration for defining cell clusters.

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

Methods

Subjects and specimen collection

Peripheral blood was collected from healthy adult donors through Emory University's Children's Clinical and Translational Discovery Core (CCTDC). Bone marrow samples were collected from healthy adult donors through Emory University under IRB00066294 or obtained from AllCells (Alameda, CA).

Cell preparation

Peripheral blood mononuclear cells (PBMCs) were isolated from peripheral venous blood using EasySep™ Direct Human PBMC Isolation Kit (# 19654). B cells were enriched from hip bone marrow (BM) using EasySep™ Direct Human PBMC Isolation Kit (# 19654) followed by EasySep Human B-Cell Enrichment Kit II without CD43 Depletion (# 17963) following manufacture's protocol. Up to $1X10^6$ cells per donor were incubated with Fc block (Myltenyi) 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 deficient in biotin, L-glutamine, phenol red, riboflavin, and sodium bicarbonate, and containing 3% newborn calf serum (def-RPMI-1640), 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 encapsulation, library generation, and sequencing

Cells were loaded to target encapsulation of 6000 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 CellRanger 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," while 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 highly variable genes for each sample using regularized negative binomial regression[72], 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 XXX.

Antibody-derived tags (ADT)

The surface protein expression (ADT) data were normalized using the DSB normalization[28]. Cell

barcodes in the CellRanger unfiltered matrix containing less than 40 genes and less than 100 total UMIs were considered background. The barcodes from CellRanger "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[28], the ADT values representing background expression levels were centered around zero. ADT raw count matrices are available through NCBI GEO, accession number XXX.

Antibody repertoire (VDJ)

We used the CellRanger-generated "all_contig_annotations.csv" (unfiltered) and the "filtered_contig_annotations.csv" files to generate new features merged with the corresponding ADT matrix. Briefly, we summed the total UMIs of each cell barcode in the "all_contig_annotations.csv" 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 V(D)J information is contained in the CellRanger "filtered" annotation file and hence considered true B cell/plasma cell. B-cell receptor repertoire V(D)J raw count matrices are available through NCBI GEO, accession number XXX.

Manual gating

Manual gating was performed based on the two VDJ features (i.e., total Ig-specific transcripts and productive VDJ) and the 31 surface protein features in the normalized ADT data. The major cell lineages (six for PBMCs and five for BM cells) were identified and manually gated using a customized strategy of biaxial plots implemented by MATLAB script (https://github.com/LindaXucongmin/SuPERR-seq.git). The markers and the gating hierarchy were determined by prior knowledge of variations of cell-surface protein expression across various major cell lineages (Figs. 2A and 3A). Cell doublets were manually "gated out" based on the expression of two or more major cell-lineage markers (Figs. S1 and S2).

Clustering on gene expression matrix

Each major cell lineage identified by the manual gating was next clustered based on the integrated gene expression matrix[72]. First, highly variable genes were computed for cells in each major cell lineage using the variance-stabilizing transformation (VST) method[72], and expression data of each gene was transformed to zero mean and unit variance. We then performed PCA on the highly variable genes 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 derive a refined version of the cell-cell similarity matrix, followed by Louvain community detection algorithm[73] to define cell clusters in each major cell lineage.

Differential gene expression analysis

We used the “FindAllMarkers” function of Seurat V3 to perform the differential gene expression (DGE) 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.

Pathway analysis

The DGE list for each BM plasma cell cluster was used as input data to the Reactome pathway database[37] (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.

Comparison of cell doublets identified by DoubletFinder and SuPERR-seq

The DoubletFinder workflow was run independently on the gene expression matrices from each PBMC and BM sample, following the standard workflow described in the original publication[39]. Briefly, we used the bimodality coefficient (BC_{MVN}) maximization to select the optimal proportion of K-nearest neighbors (pK) for each sample. We then set the proportion of artificial doublets to pN=25% and selected the top 10 PCs to identify cell doublets. The estimated doublet rate from the Chromium microfluidics was determined for each sample based on the 10X Genomics user guide[74]. Next, the clustering results generated by the Seurat integration workflow were used to reduce the false-positive rate. Finally, since we did not explicitly classify every cell barcode during manual gating, we removed the cell barcodes not present in the final singlet and doublet population defined by the SuPERR-seq workflow.

Cell Misclassification Statistic

To generate the Cell Misclassification Statistic (CMS) score[43], we cross-check the cell-type identities generated by the SuPERR-seq (reference) versus conventional workflow (test). We apply the CMS to measure the loss in cell-type fidelity caused by relying on clustering algorithms based solely on the gene expression matrix to classify cell types, as performed by conventional workflows. CMS relies on the tenet that if different workflows identify a single cell as different biological types, both cannot be correct, and therefore some information has been lost. To generate a CMS score, we first produce clusters following a conventional workflow (Seurat V3, default settings). To assign cell-type labels to each cluster, we explore cell-type-specific genes as well as follow a strategy of biaxial plots for surface proteins identical to the SuPERR-seq gating strategy; only instead of applying cutoff thresholds to assign individual cells to each lineage, we assign entire clusters to a cell type based on the expression of the cell-specific markers. After cell-type identities are recorded for each cluster generated by the conventional workflow (test sample), we match a vector of cell identities from the same sample but processed by the SuPERR-seq workflow (reference sample). The CMS score is directly the percent of cells

that change classification between the two workflows. As both classifications cannot be correct, we consider this percentage to be cells that are misclassified by one, or both, systems. CMS is reported as a fraction from 0-1, where 0 means that no cells were misclassified, and a score of 0.2 means that 20% of cells were misclassified.

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 NIH UL1TR002378 (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.); 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 CMS scores. All authors contributed to the writing of the manuscript and approved the final version.

Competing Interests

The authors declare no competing interests.

References

1 Evan *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

2 Allon *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).

3 Utada, A. S., Fernandez-Nieves, A., Stone, H. A. & Weitz, D. A. Dripping to Jetting Transitions in Coflowing Liquid Streams. *Physical Review Letters* **99**, doi:10.1103/physrevlett.99.094502 (2007).

4 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, 1096-1098, doi:10.1038/nmeth.2639 (2013).

5 Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports* **2**, 666-673, doi:10.1016/j.celrep.2012.08.003 (2012).

6 Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods* **14**, 955-958, doi:10.1038/nmeth.4407 (2017).

7 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**, 865-868, doi:10.1038/nmeth.4380 (2017).

8 Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049, doi:10.1038/ncomms14049 (2017).

9 Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5**, doi:10.12688/f1000research.7223.1 (2016).

10 Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology* **35**, 936-939, doi:10.1038/nbt.3973 (2017).

11 Lee, J., Hyeon, D. Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine* **52**, 1428-1442, doi:10.1038/s12276-020-0420-2 (2020).

12 Singh, M. *et al.* High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nature Communications* **10**, doi:10.1038/s41467-019-11049-4 (2019).

13 Horns, F., Dekker, C. L. & Quake, S. R. Memory B Cell Activation, Broad Anti-influenza Antibodies, and Bystander Activation Revealed by Single-Cell Transcriptomics. *Cell Reports* **30**, 905-913.e906, doi:10.1016/j.celrep.2019.12.063 (2020).

14 Zemmour, D. *et al.* Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nature Immunology* **19**, 291-301, doi:10.1038/s41590-018-0051-0 (2018).

15 Yermanos, A. *et al.* 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, doi:10.1098/rspb.2020.2793 (2021).

16 Meyer, K. D. DART-seq: an antibody-free method for global m6A detection. *Nature Methods* **16**, 1275-1280, doi:10.1038/s41592-019-0570-0 (2019).

17 Luecken, M. D. & Theis, F. J. Current best practices in single‐cell RNA‐seq

analysis: a tutorial. *Molecular Systems Biology* **15**, e8746, doi:10.15252/msb.20188746 (2019).

18 Stuart, T. & Satija, R. Integrative single-cell analysis. *Nature Reviews Genetics* **20**, 257-272, doi:10.1038/s41576-019-0093-7 (2019).

19 Colomé-Tatché, M. & Theis, F. J. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology* **7**, 54-59, doi:10.1016/j.coisb.2018.01.003 (2018).

20 Yu, P. & Lin, W. Single-cell Transcriptome Study as Big Data. *Genomics, Proteomics & Bioinformatics* **14**, 21-30, doi:10.1016/j.gpb.2016.01.005 (2016).

21 Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications* **11**, doi:10.1038/s41467-020-14976-9 (2020).

22 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529, doi:10.1016/j.cell.2021.04.048 (2021).

23 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821, doi:10.1016/j.cell.2019.05.031 (2019).

24 Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (New York, N.Y.)* **361**, 1380-1385, doi:10.1126/science.aau0730 (2018).

25 Argelaguet, R. *et al.* Multi‐Omics Factor Analysis—a framework for unsupervised integration of multi‐omics data sets. *Molecular Systems Biology* **14**, e8124, doi:10.15252/msb.20178124 (2018).

26 Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e1817, doi:10.1016/j.cell.2019.05.006 (2019).

27 Meehan, S. *et al.* Automated subset identification and characterization pipeline for multidimensional flow and mass cytometry data clustering and visualization. *Communications Biology* **2**, 229, doi:10.1038/s42003-019-0467-6 (2019).

28 Mulè, M. P., Martins, A. J. & Tsang, J. S. Normalizing and denoising protein expression data from droplet-based single cell profiling. *bioRxiv*, 2020.2002.2024.963603, doi:10.1101/2020.02.24.963603 (2021).

29 Garimalla, S. *et al.* Differential transcriptome and development of human peripheral plasma cell subsets. *JCI Insight* **4**, doi:10.1172/jci.insight.126732 (2019).

30 Glass, D. R. *et al.* An Integrated Multi-omic Single-Cell Atlas of Human B Cell Identity. *Immunity* **53**, 217-232 e215, doi:10.1016/j.immuni.2020.06.013 (2020).

31 Kaminski, D. A., Wei, C., Qian, Y., Rosenberg, A. F. & Sanz, I. Advances in human B cell phenotypic profiling. *Front Immunol* **3**, 302, doi:10.3389/fimmu.2012.00302 (2012).

32 Tiller, T. *et al.* Autoreactivity in Human IgG+ Memory B Cells. *Immunity* **26**, 205-213, doi:10.1016/j.immuni.2007.01.009 (2007).

33 Mehtonen, J. *et al.* 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 Medicine* **12**, doi:10.1186/s13073-020-00799-2 (2020).

34 Zhou, Y. *et al.* Transitional B cells involved in autoimmunity and their impact on neuroimmunological diseases. *Journal of translational medicine* **18**,

doi:10.1186/s12967-020-02289-w (2020).

35 Becker, S. C. *et al.* A comparative analysis of human bone marrow–resident and peripheral memory B cells. *Journal of Allergy and Clinical Immunology* **141**, 1911-1913.e1917, doi:10.1016/j.jaci.2017.12.983 (2018).

36 Halliley, J. L. *et al.* Long-Lived Plasma Cells Are Contained within the CD19(-)CD38(hi)CD138(+) Subset in Human Bone Marrow. *Immunity* **43**, 132-145, doi:10.1016/j.immuni.2015.06.016 (2015).

37 Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498-d503, doi:10.1093/nar/gkz1031 (2020).

38 Ocasio, J. *et al.* scRNA-seq in medulloblastoma shows cellular heterogeneity and lineage expansion support resistance to SHH inhibitor therapy. *Nature Communications* **10**, doi:10.1038/s41467-019-13657-6 (2019).

39 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324, doi:10.1016/j.cels.2019.03.003 (2019).

40 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

41 Orlova, D. Y., Herzenberg, L. A. & Walther, G. Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets. *Nature Reviews Immunology* **18**, 77-77, doi:10.1038/nri.2017.150 (2018).

42 Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nature Methods* **15**, 399-400, doi:10.1038/s41592-018-0019-x (2018).

43 Babcock, B. R., Kosters, A., Yang, J., White, M. L. & Ghosn, E. E. B. Data Matrix Normalization and Merging Strategies Minimize Batch-specific Systemic Variation in scRNA-Seq Data. *bioRxiv* (2021).

44 Anjum, A. *et al.* Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. *Journal of Computational Biology* **23**, 239-247, doi:10.1089/cmb.2015.0205 (2016).

45 Uldrich, A. P. *et al.* CD1d-lipid antigen recognition by the gammadelta TCR. *Nat Immunol* **14**, 1137-1145, doi:10.1038/ni.2713 (2013).

46 Shlomchik, M. J. & Weisel, F. Germinal center selection and the development of memory B and plasma cells. *Immunological Reviews* **247**, 52-63, doi:10.1111/j.1600-065x.2012.01124.x (2012).

47 Poli, A. *et al.* CD56bright natural killer (NK) cells: an important NK cell subset. *Immunology* **126**, 458-465, doi:10.1111/j.1365-2567.2008.03027.x (2009).

48 Evans, J. H. *et al.* 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, doi:10.1002/eji.201041180 (2011).

49 He, Y. *et al.* EBV promotes human CD8 NKT cell development. *PLoS pathogens* **6**, e1000915, doi:10.1371/journal.ppat.1000915 (2010).

50 Lawand, M., Déchanet-Merville, J. & Dieu-Nosjean, M. C. Key Features of Gamma-Delta T-Cell Subsets in Human Diseases and Their Immunotherapeutic Implications. *Front Immunol* **8**, 761, doi:10.3389/fimmu.2017.00761 (2017).

51 Cai, Y. *et al.* Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis. *EBioMedicine* **53**, 102686,

doi:10.1016/j.ebiom.2020.102686 (2020).

52 Wong, E. B. *et al.* TRAV1-2(+) CD8(+) T-cells including oligoconal expansions of MAIT cells are enriched in the airways in human tuberculosis. *Commun Biol* **2**, 203, doi:10.1038/s42003-019-0442-2 (2019).

53 Villani, A. C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science (New York, N.Y.)* **356**, doi:10.1126/science.aah4573 (2017).

54 Blaser, C. *et al.* Beta-galactoside-binding protein secreted by activated T cells inhibits antigen-induced proliferation of T cells. *Eur J Immunol* **28**, 2311-2319, doi:10.1002/(sici)1521-4141(199808)28:08<2311::Aid-immu2311>3.0.Co;2-g (1998).

55 Hashimoto, K. *et al.* Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 24242-24251, doi:10.1073/pnas.1907883116 (2019).

56 Juno, J. A. *et al.* Cytotoxic CD4 T Cells-Friend or Foe during Viral Infection? *Front Immunol* **8**, 19, doi:10.3389/fimmu.2017.00019 (2017).

57 Kumar, B. V., Connors, T. J. & Farber, D. L. Human T Cell Development, Localization, and Function throughout Life. *Immunity* **48**, 202-213, doi:10.1016/j.immuni.2018.01.007 (2018).

58 Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu Rev Immunol* **28**, 445-489, doi:10.1146/annurev-immunol-030409-101212 (2010).

59 Braun, J., Frentsch, M. & Thiel, A. Hobit and human effector T-cell differentiation: The beginning of a long journey. *Eur J Immunol* **45**, 2762-2765, doi:10.1002/eji.201545959 (2015).

60 Martin, M. D. & Badovinac, V. P. Defining Memory CD8 T Cell. *Front Immunol* **9**, 2692, doi:10.3389/fimmu.2018.02692 (2018).

61 Jin, X. *et al.* Characterization of dendritic cell subtypes in human cord blood by single-cell sequencing. *Biophysics Reports* **5**, 199-208, doi:10.1007/s41048-019-00096-5 (2019).

62 Kuramasu, A., Saito, H., Suzuki, S., Watanabe, T. & Ohtsu, H. 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, doi:10.1074/jbc.273.47.31607 (1998).

63 Källberg, E. & Leanderson, T. A subset of dendritic cells express joining chain (J-chain) protein. *Immunology* **123**, 590-599, doi:10.1111/j.1365-2567.2007.02733.x (2008).

64 Dzierzak, E. & Philipsen, S. Erythropoiesis: development and differentiation. *Cold Spring Harb Perspect Med* **3**, a011601, doi:10.1101/cshperspect.a011601 (2013).

65 Pellin, D. *et al.* A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat Commun* **10**, 2395, doi:10.1038/s41467-019-10291-0 (2019).

66 Xie, X. *et al.* Single-cell transcriptomic landscape of human blood cells. *National Science Review* **8**, doi:10.1093/nsr/nwaa180 (2020).

67 Lai, S. *et al.* Mapping Human Hematopoietic Hierarchy at Single Cell Resolution by Microwell-seq. *bioRxiv*, 127217, doi:10.1101/127217 (2017).
68 Yang, J., Zhang, L., Yu, C., Yang, X. F. & Wang, H. Monocyte and macrophage differentiation: circulation inflammatory monocyte as biomarker for inflammatory diseases. *Biomark Res* **2**, 1, doi:10.1186/2050-7771-2-1 (2014).
69 Kawamura, S. *et al.* Identification of a Human Clonogenic Progenitor with Strict Monocyte Differentiation Potential: A Counterpart of Mouse cMoPs. *Immunity* **46**, 835-848.e834, doi:10.1016/j.immuni.2017.04.019 (2017).
70 Evrard, M. *et al.* Developmental Analysis of Bone Marrow Neutrophils Reveals Populations Specialized in Expansion, Trafficking, and Effector Functions. *Immunity* **48**, 364-379.e368, doi:10.1016/j.immuni.2018.02.002 (2018).
71 Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (New York, N.Y.)* **357**, 661-667, doi:10.1126/science.aam8940 (2017).
72 Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296, doi:10.1186/s13059-019-1874-1 (2019).
73 Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, doi:10.1038/s41598-019-41695-z (2019).
74 Chromium Single Cell V(D)J Reagent Kits with Feature Barcoding technology for Cell Surface Protein, Document Number CG000186 Rev A, 10x Genomics, (2019, July 25).

Main lineage	SuPERR-seq cluster	Cell classification	Phenotype	Refs
Plasma cells	1	Plasma cells	CD19+/CD20-/highest levels of Ig-specific transcripts	46
B cells (CD19+, CD20+)	2	Naïve-I	IgD+/IgM+/CD27-/CD95-	30-33
	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/gdT (CD56+)	7	NK-I CD16+	KLRC3+/$CD11b^{hi}$	47-52
	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 ($CD14^{lo/+}$)	17	Classical-I	CD16-/CD68+/HLA-Drhi	53
	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 (CD3+, CD4+)	22	Naïve-I	CCR7+/SELL+/CD27+/CD95-/$CD3^{hi}$	54-58
	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 Naïve	FOXP3+/CD45RA+/CD95-	
	30	Temra	CD45RA+/NKG7+/GNLY+/GZMB+	
CD8 T cells (CD3+, CD8a+)	31	Naïve-I	CCR7+/CD45RA+/CD27+/CD127+	59,60
	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+	

Table 1. Cell classification and phenotype of PBMC clusters generated by the SuPERR-seq pipeline. Gene expression (GEX) and cell-surface (ADT) markers used to classify the lymphoid and myeloid cell populations in the human peripheral blood. NK: Nature Killer cells; NKT: NK T cells; iNKT: invariant NKT; γδ T: gamma-delta T cells. Tcm: central memory; Tem: effector memory; CTL: cytotoxic T lymphocytes; Treg: regulatory T cells; Temra: effector memory expressing CD45RA; TLE: long-lived effector memory; MAIT: mucosal-associated invariant T cells.

Main lineage	SuPERR-seq cluster	Cell classification	Phenotype	Refs
Plasma cells (CD138-)	1	Ig^{hi}/PRDM1-	Ig^{hi}/PRDM1-	30, 37
	2	HLA-DR^{hi}/$SDC1^{lo}$/G2-M phase	HLA-DR^{hi}/$SDC1^{lo}$/G2-M phase	
Plasma cells (CD138+)	3	$IRF4^{hi}$	$IRF4^{hi}$	
	4	$XBP1^{hi}$/$SDC1^{hi}$	$XBP1^{hi}$/$SDC1^{hi}$	
B cells (CD19+, CD20+)	5	T3/Naïve	CD21+/IL4R+	34-36
	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 (Lineage-, CD34+/-)	14	Pre-reticulocytes	$GYPA^{int}$/HBB+	61-67
	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 (CD11c+)	26	Neutropoiesis	MPO^{hi}/$ELANE^{hi}$	68-70
	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+	

Table 2. Cell classification and phenotype of BM clusters generated by the SuPERR-seq pipeline. 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. Ig: immunoglobulin production; 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.

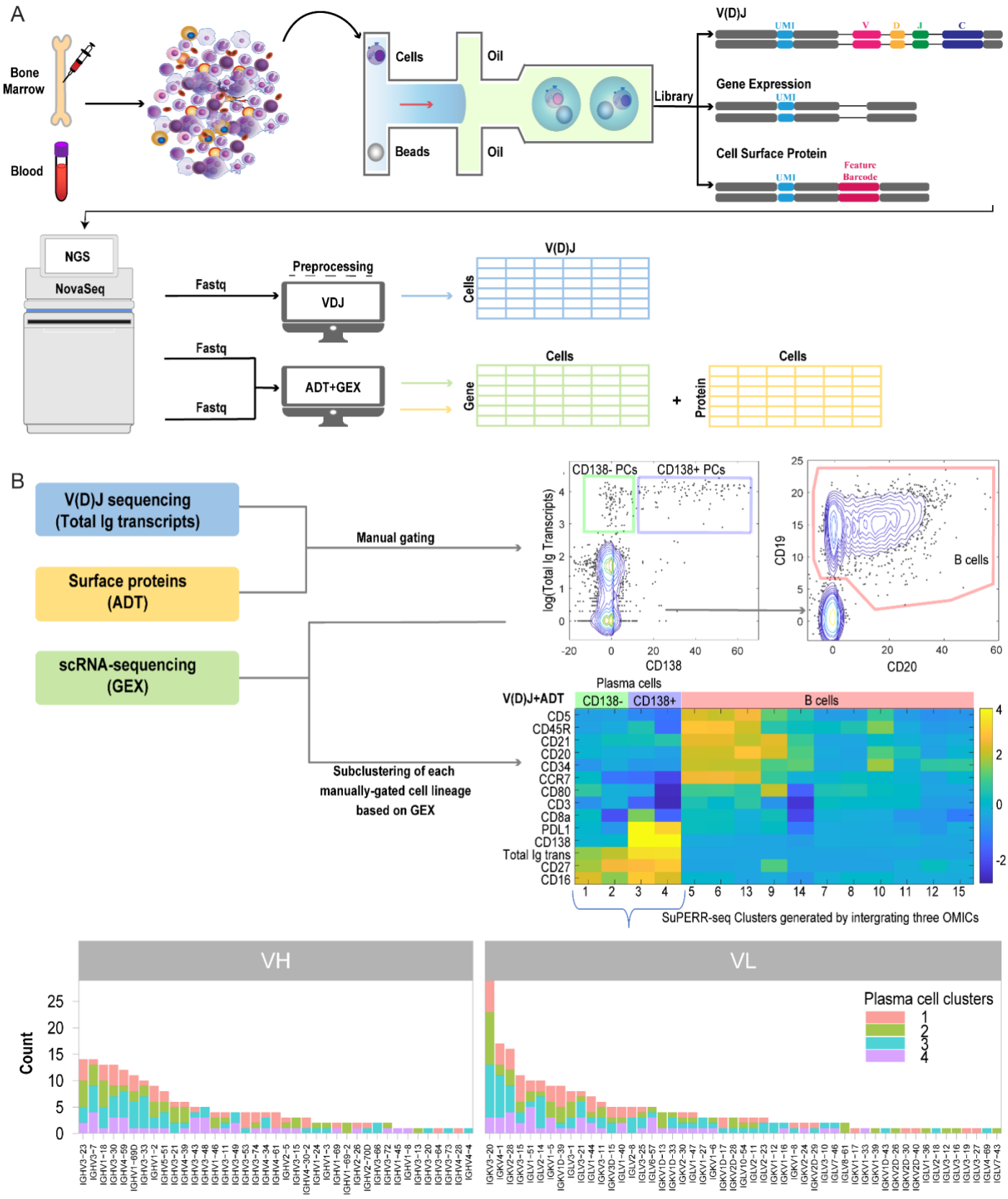

Figure 1. SuPERR-seq workflow. (A) Schematic overview of the experimental design. Blood and bone marrow aspirates were encapsulated with barcoded microspheres. We generated three libraries for each sample corresponding to gene expression (GEX), cell-surface protein/antibody-derived tags (ADT), and cell-receptor repertoire (VDJ). Libraries were sequenced to a target depth, and count matrices were assembled for each -omic data separately. (B) SuPERR-seq 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.

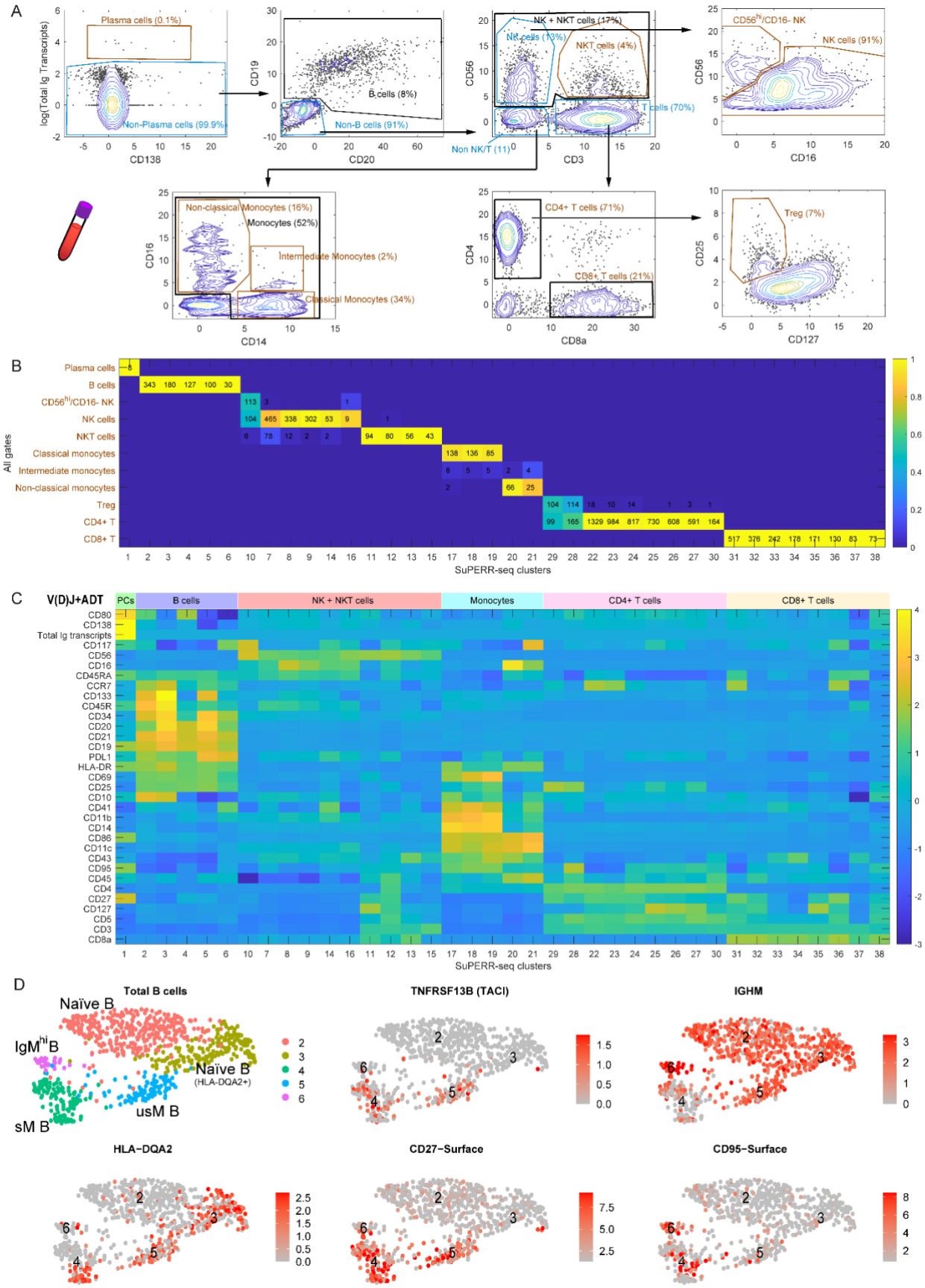

Figure 2. SuPERR-seq workflow applied to PBMCs. (A) "Gating strategy" approach to identify major cell lineages on biaxial plots based on surface markers (ADT) and V(D)J data. Gates for major lineages are indicated as black outlines and black text. Gates for downstream cell identity validation are indicated as golden outlines and golden text. (B) Cross comparison between the manually-gated major lineages and the final SuPERR-seq clusters. (C) The average expression levels of surface markers (ADT) and Ig-specific transcripts (V(D)J) for the final SuPERR-seq clusters. (D) UMAP representation of the B-cell clusters identified by the SuPERR-seq workflow. PCs: plasma cells. usM B: unswitched memory B cells. sM B: switched memory B cells.

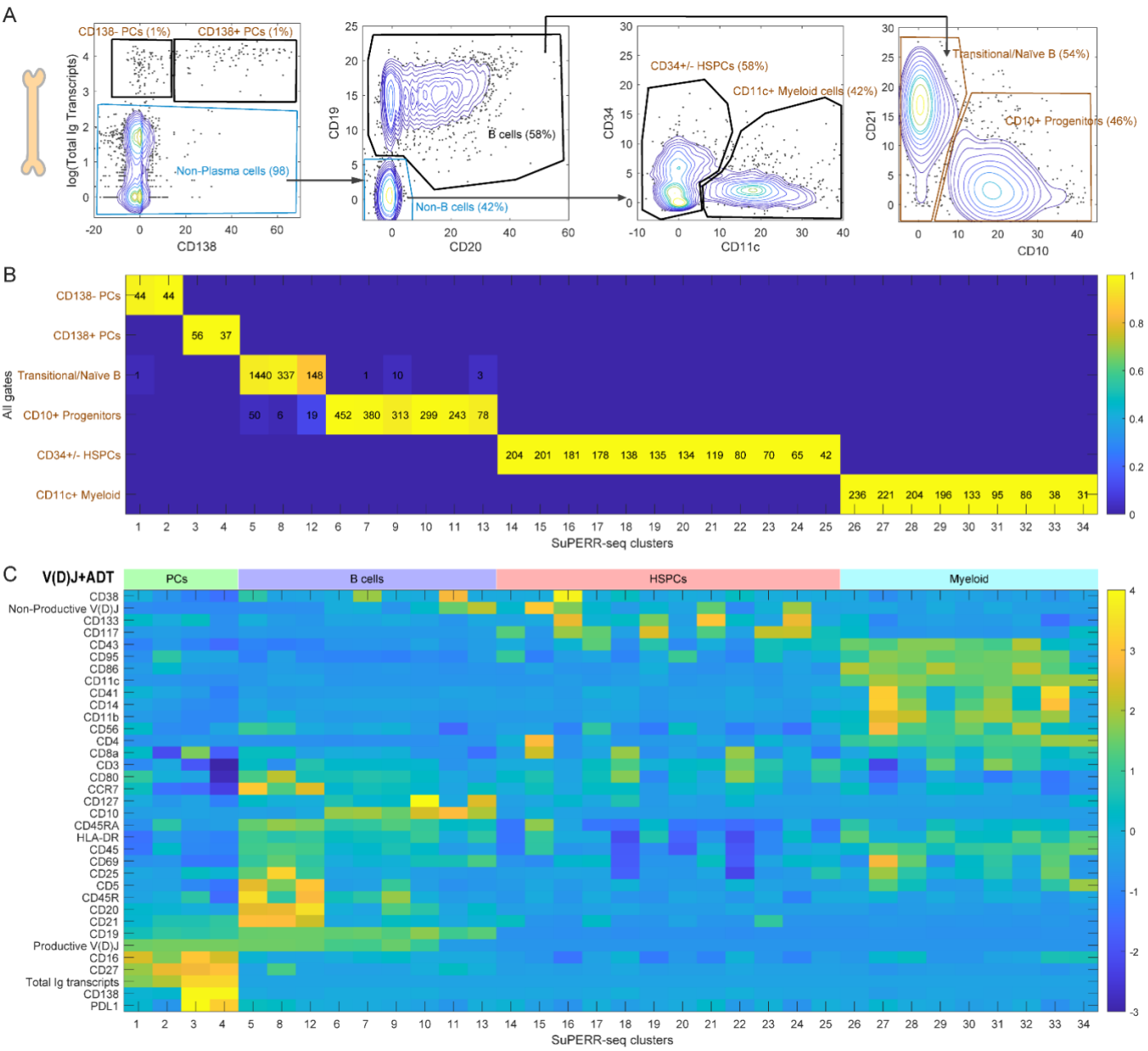

Figure 3. SuPERR-seq workflow applied to BM cells. (A) "Gating strategy" approach to identify major cell lineages on biaxial plots based on surface markers (ADT) and V(D)J data. Gates for major lineages are indicated as black outlines and black text. Gates for downstream cell identity validation are indicated as golden outlines and golden text. (B) Cross comparison between the manually-gated major lineages and the final SuPERR-seq clusters. (C) The average expression levels of surface markers (ADT) and Ig-specific transcripts (V(D)J) for the final SuPERR-seq clusters. PCs: plasma cells.

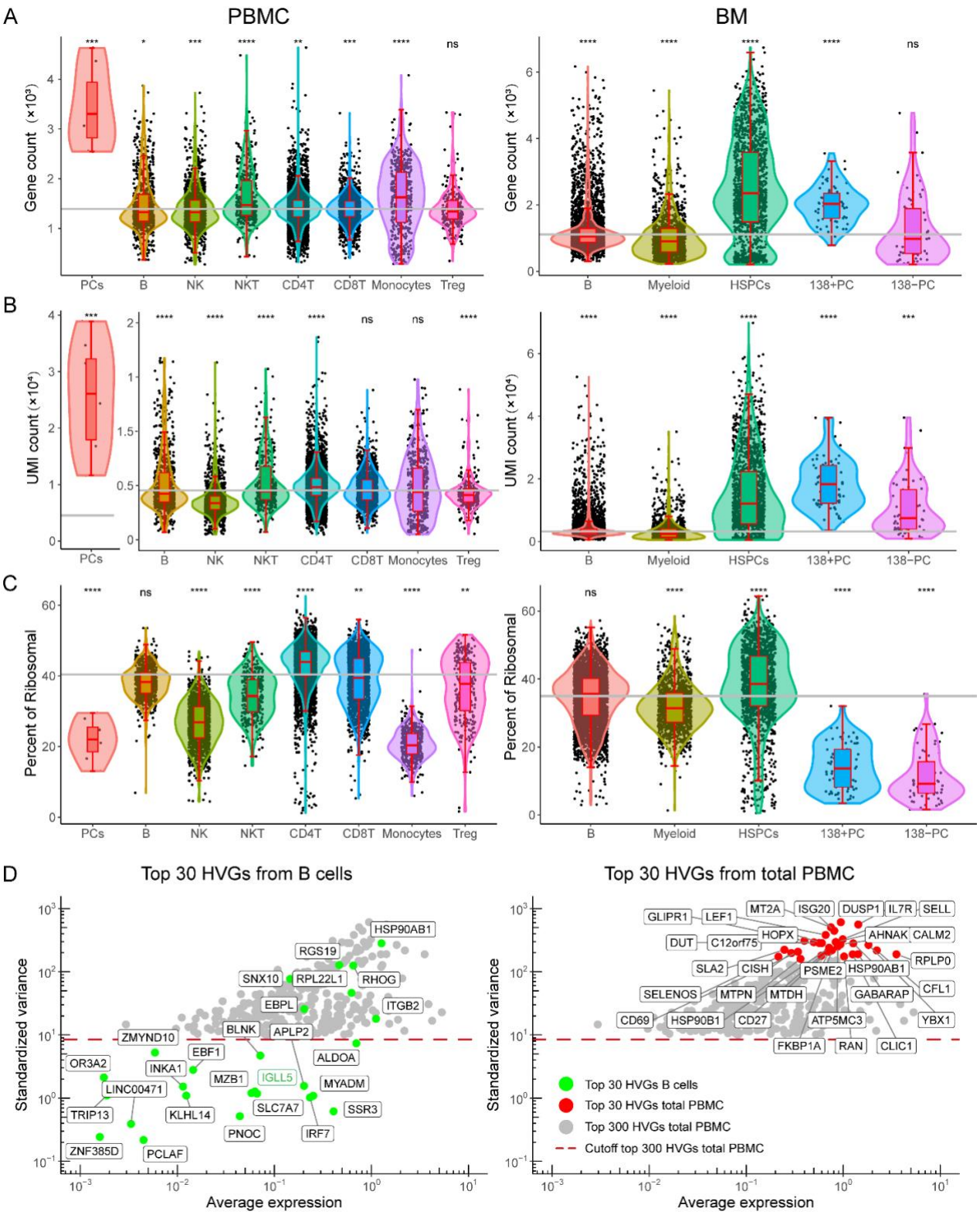

Figure 4. Cell-type-specific variations in gene expression count matrices. (A) "Gene count" represents the number of unique genes expressed by each cell type. (B) "UMI count" represents the total mRNA abundance expressed by each cell type. (C) "Percent of Ribosomal" represents the percentage of ribosomal gene-derived 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: top 30 highly-variable genes from PBMC-derived B cells (green dots) displayed with the top 300 highly-variable genes from total PBMC (grey dots). Right panel: top 30 highly-variable genes from total PBMC (red dots) displayed and the top 300 highly-variable genes from total PBMC (grey dots). The dots under the red dashed line fall below the top 300 HVGs of the total PBMC. 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. HVGs: highly-variable genes.

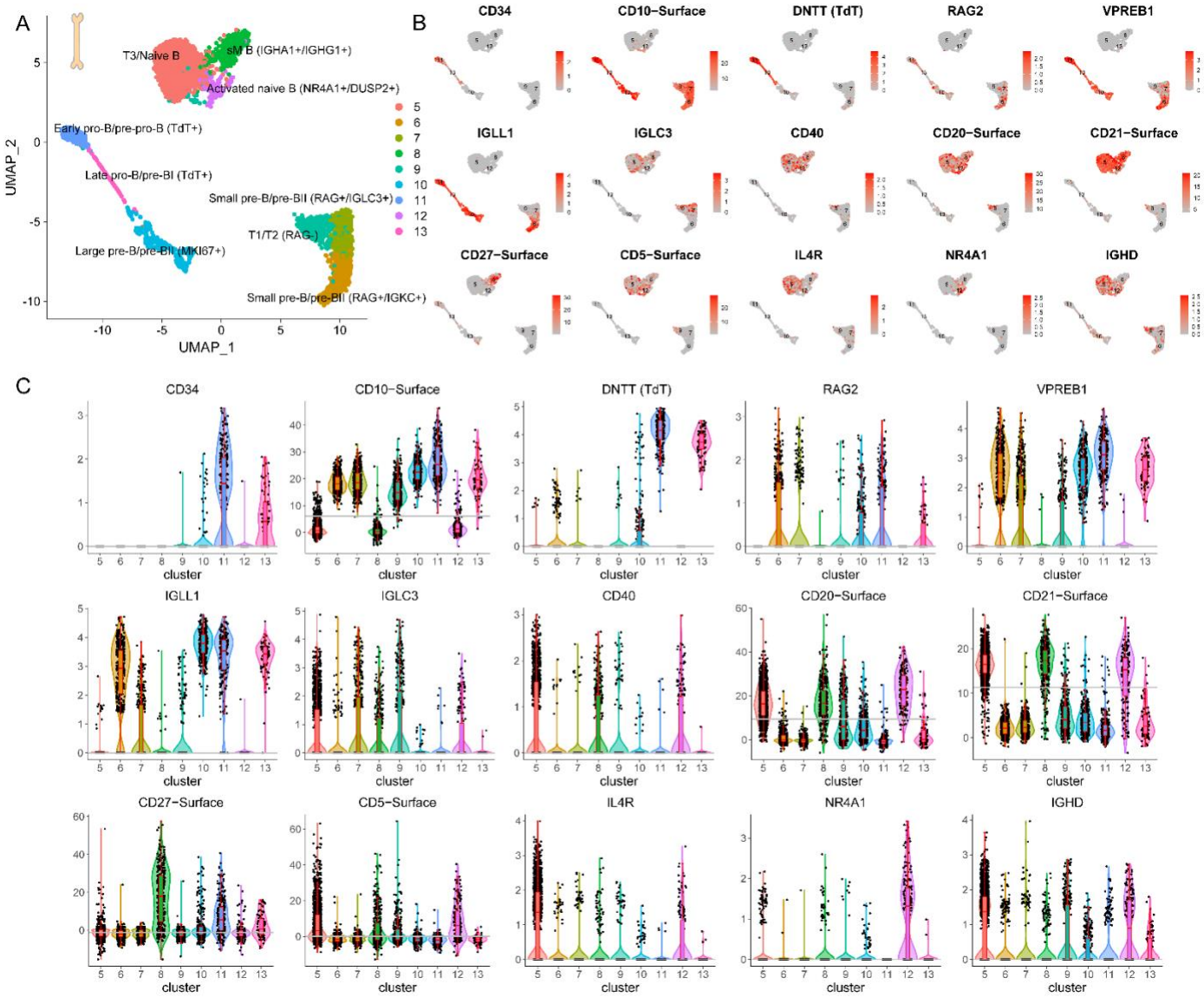

Figure 5. B-cell developmental pathway in human adult BM identified by the SuPERR-seq workflow. (A) Cell-type classification of BM B cells based on three -omics, surface protein (ADT), V(D)J matrix, and gene expression profile (GEX). (B) The expression level of surface proteins and genes used to define the B-cell developmental stages. (C) Violin plots illustrate the differences in expression levels (ADT and GEX) between clusters. The grey line shows the mean expression level across all clusters.

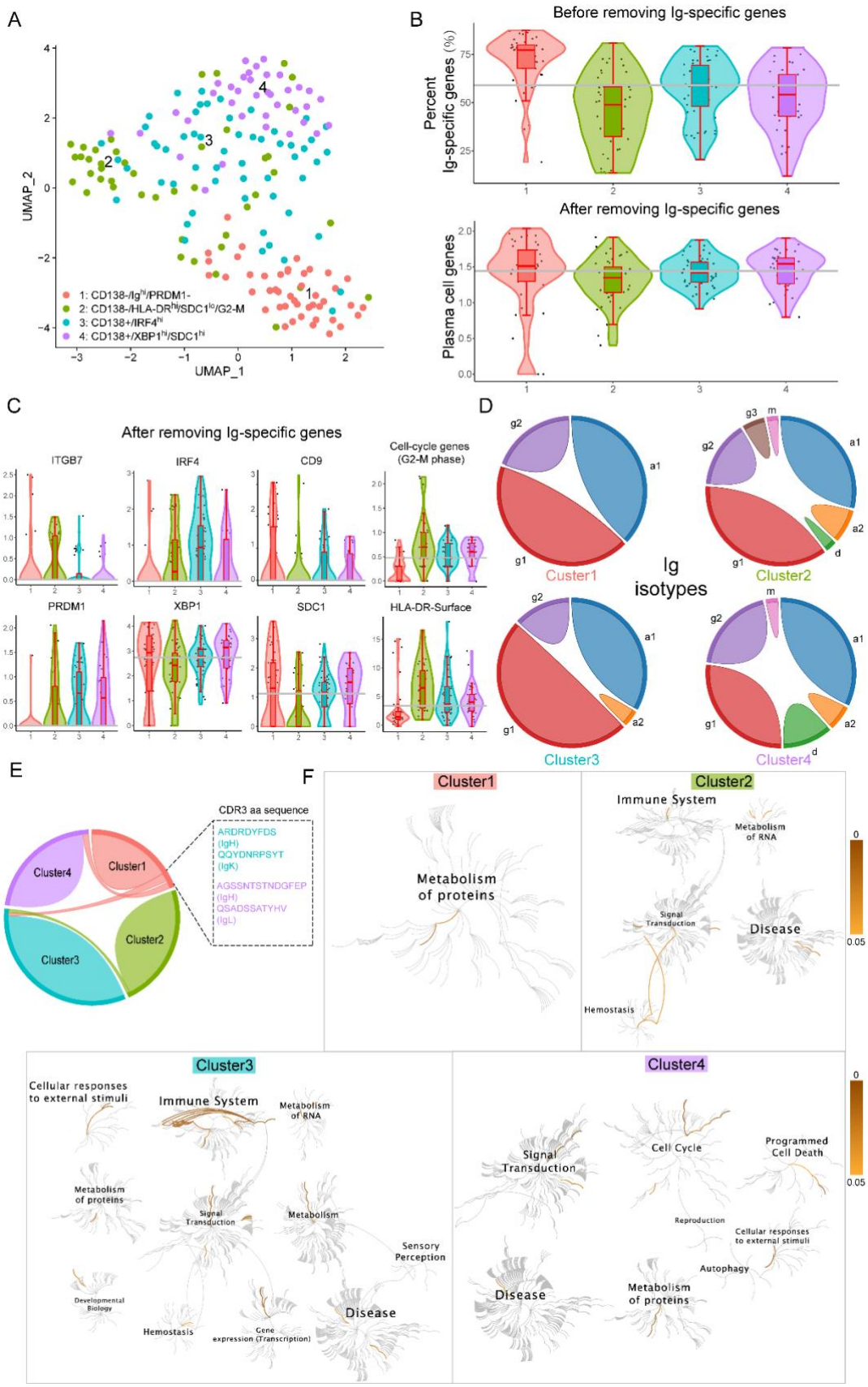

Figure 6. SuPERR-seq workflow identifies four subsets of human plasma cells in the BM. (A) UMAP representation of the four BM plasma cells. (B) Top panel: percentage of Ig-specific transcripts (UMI) expressed in each plasma cell subset. Bottom panel: expression levels (sum) of plasma cell genes (IRF4, PRDM1, XBP1, SDC1) after removing Ig-specific UMIs and re-normalizing the data matrix. (C) Expression levels of individual plasma cell genes after removing Ig-specific UMIs and re-normalizing the data matrix. The grey line shows the mean expression level across all clusters. (D) The antibody isotypes and subclasses expressed by each plasma cell subset. (E) Circus plot showing 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) The Reactome Pathway Database analysis shows the unique biological processes that define each plasma cell subset.

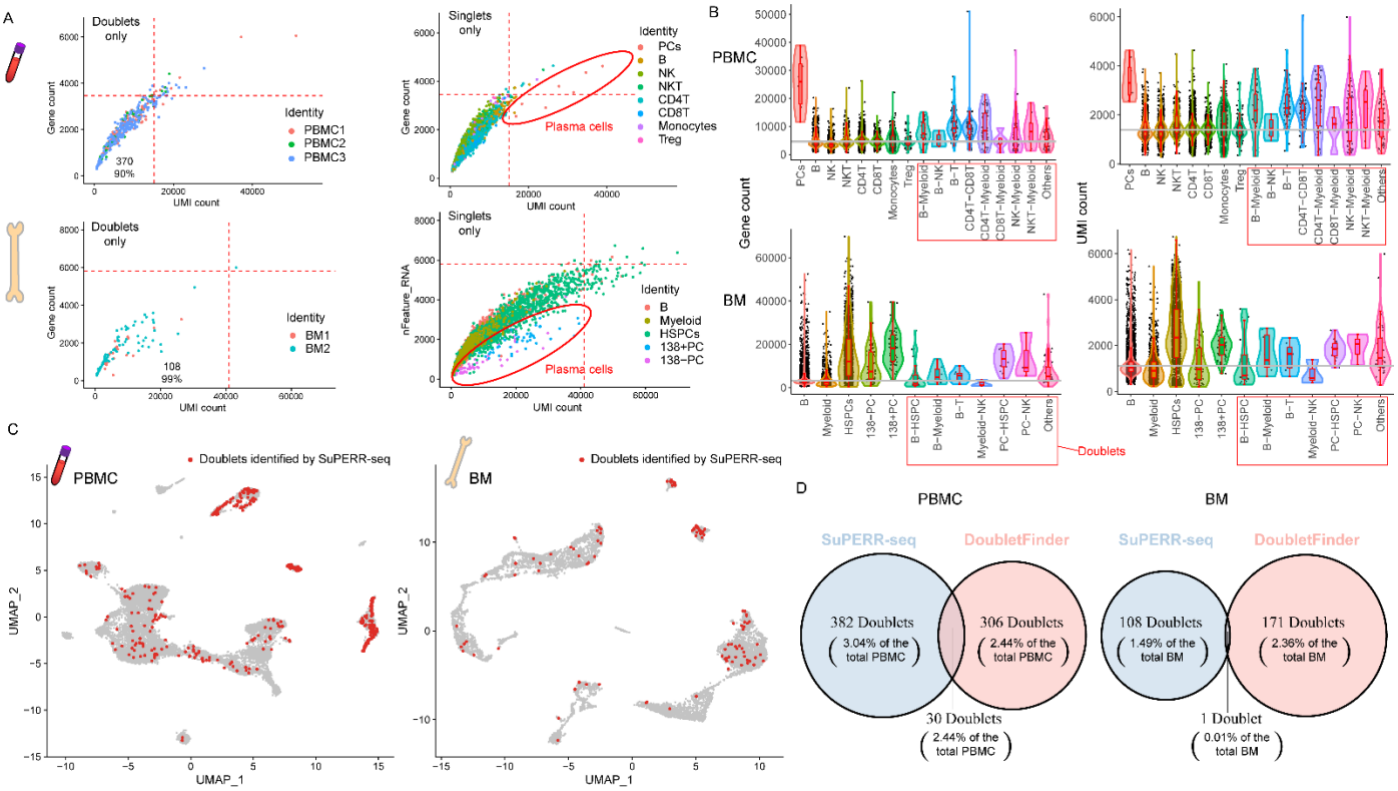

Figure 7. Cell-doublet identification by SuPERR-seq and other conventional approaches. (A) Distribution (unique gene count x total UMI) of cell doublets (left) and singlets (right) detected by the SuPERR-seq approach. The red dashed lines show the threshold used by the conventional approaches to exclude 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 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. (C) Cell doublets identified by SuPERR-seq and projected on a UMAP show these cells are spread across multiple clusters. (D) Venn diagram comparing the cell doublets identified by the SuPERR-seq and DoubletFinder. DoubletFinder missed many cell-doublets identified by SuPERR-seq (blue).

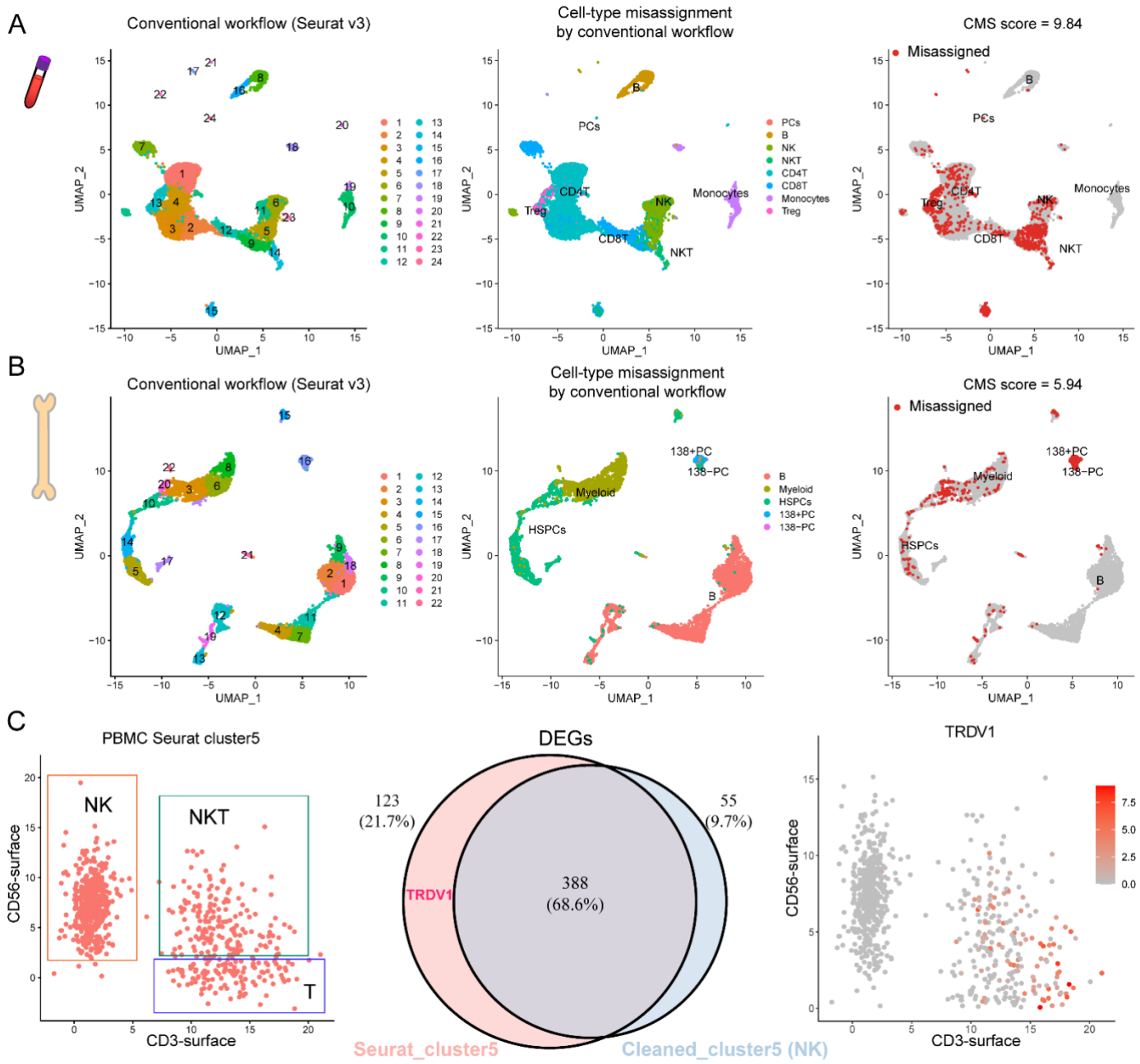

Figure 8. Cell-type misclassification by conventional workflow. (A) and (B) *Left panels*: PBMC (A) and BM (B) clustering result by conventional workflow (i.e., based solely on gene expression matrix). *Middle panels*: PBMC (A) and BM (B) major cell lineages identified by SuPERR-seq and projected onto the UMAP plot of the conventional workflow. *Right panels*: red dots show the PBMC (A) and BM (B) cells misassigned/misclassified by the conventional Seurat workflow. The Cell Misclassification Statistic (CMS) score calculates the percentage of misclassified cells (9.84% in PBMC and 5.94% in BM) by the conventional workflow. (C) The Seurat cluster 5 contains misassigned/misclassified cells (i.e., a mixture of NK, NKT, and T cells) and was further explored using the cell-surface markers CD56 and CD3 (left panel). The differential gene expression (DEG) analysis for cluster 5 (pink circle) shows *TRDV1* as a top hit (Venn diagram). TRDV1 is expressed on CD3+ gamma-delta T cells only (right panel) and absent in NK cells, showing that the contaminated cells in cluster 5 have contributed to the top DEGs.

Antigen	Clone	Vendor	Cat. no.
CCR7 (CD197)	G043H7	Biolegend	353251
CD11b	ICRF44	Biolegend	301359
CD127 (IL-7Ra)	A019D5	Biolegend	351356
CD14 (LPS receptor)	M5E2	Biolegend	301859
CD16 (FcγRIII)	3G8	Biolegend	302065
CD19 (B4)	HIB19	Biolegend	302265
CD20 (B1)	2H7	Biolegend	302363
CD21 (C3dR)	Bu32	Biolegend	354923
CD25	BC96	Biolegend	302649
CD27 (S152)	O323	Biolegend	302853
CD3 (T3, CD3ε)	UCHT1	Biolegend	300479
CD38 (T10)	HIT2	Biolegend	303543
CD4	RPA-T4	Biolegend	300567
CD45	2D1	Biolegend	368545
CD45RA	HI100	Biolegend	304163
CD56	QA17A16	Biolegend	392425
CD69	FN50	Biolegend	310951
CD80	2D10	Biolegend	305243
CD86	IT2.2	Biolegend	305447
CD8a	RPA-T8	Biolegend	301071
CD95 (APO-1)	DX2	Biolegend	305651
HLA-DR	L243	Biolegend	307663
PDL1 (CD274)	29E.2A3	Biolegend	329751
Custom oligo-conjgated in house			
Antigen	**Clone**	**Vendor**	**Cat. no.**
CD11c (αX integrin)	S-HCL-3	Biolegend	371502
CD133 (Prominin-1)	clone 7	Biolegend	372802
CD138 (Syndecan-1)	MI15	Biolegend	356502
CD34 (gp105-120)	581	Biolegend	343602
CD41 (gpIIb)	HIP8	Biolegend	303702
CD43 (Ly-48)	CD43-10G7	Biolegend	343202
CD45R (B220)	RA3-6B2	Biolegend	103202
CD5 (Ly-1)	UCHT2	Biolegend	300602
CD20 (B1)	2H7	Biolegend	302302

Table S1. List of antibody-derived tags (ADT) used in the multi-omics single-cell assays. Cell-surface protein antigens measured in both human peripheral blood and bone marrow samples.

	Sample ID	Mean Reads per Cell	Cell count
Gene expression	PBMC1	138,274	4,034
	PBMC2	150,868	2,302
	PBMC3	45,805	6,423
	BM1	149,030	3,480
	BM2	149,622	3,946
Surface protein expression	PBMC1	1,399	4,034
	PBMC2	2,273	2,302
	PBMC3	928	6,423
	BM1	2,762	3,480
	BM2	1,703	3,946
V(D)J expression	PBMC1	126,518	273
	PBMC2	601,049	64
	PBMC3	56,584	582
	BM1	22,991	1,669
	BM2	17,418	2,553

Table S2. Sequencing depth. Mean reads per cell and the total number of cells analyzed in each sample. PBMC: peripheral blood mononuclear cells; BM: bone marrow cells.

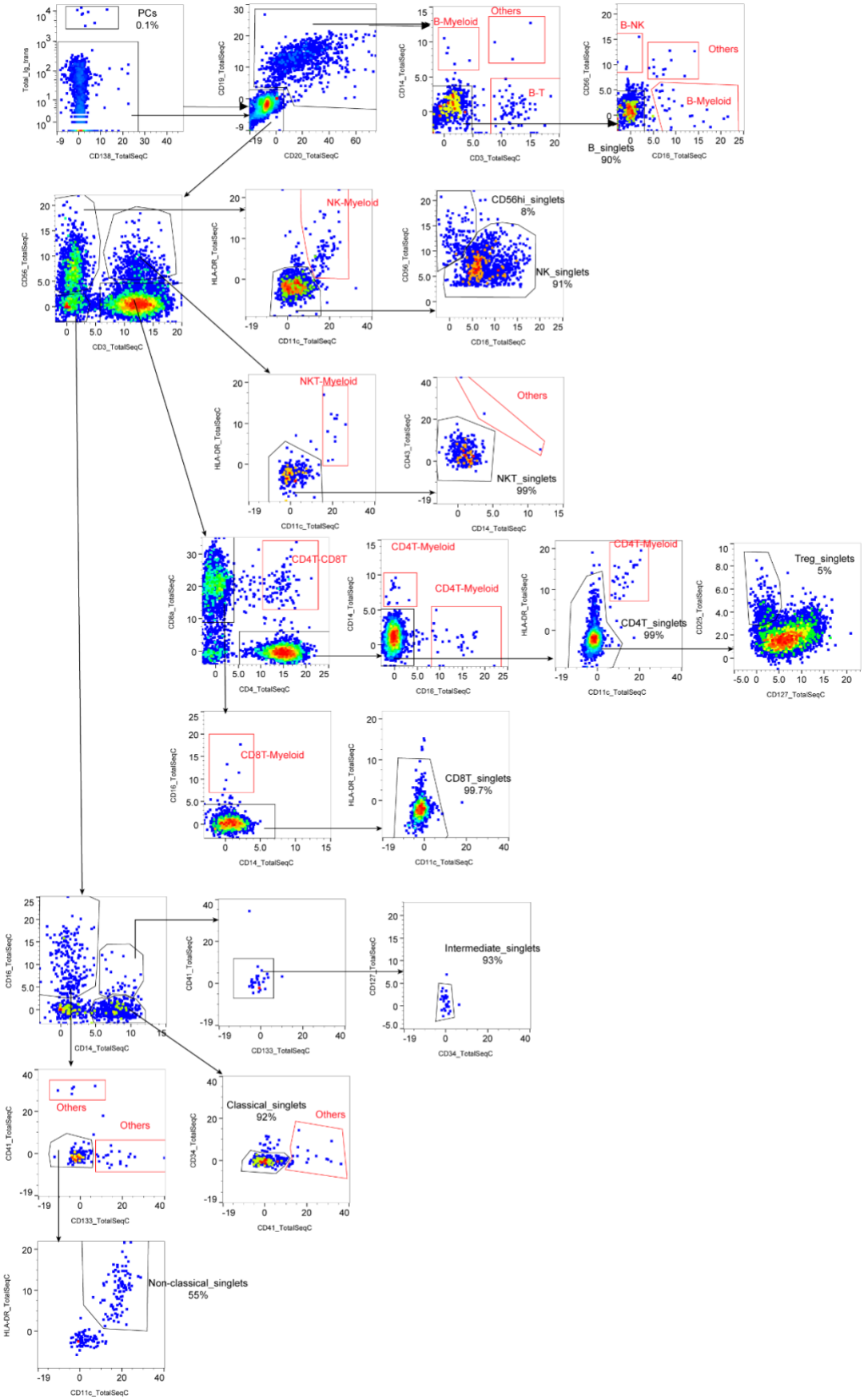

Figure S1. Biaxial feature plots to identify and remove cell doublets from PBMCs. Supplement for Figure 2. Cell doublets were gated using red boxes. The visualization was generated by SeqGeq™ (FlowJo, LLC).

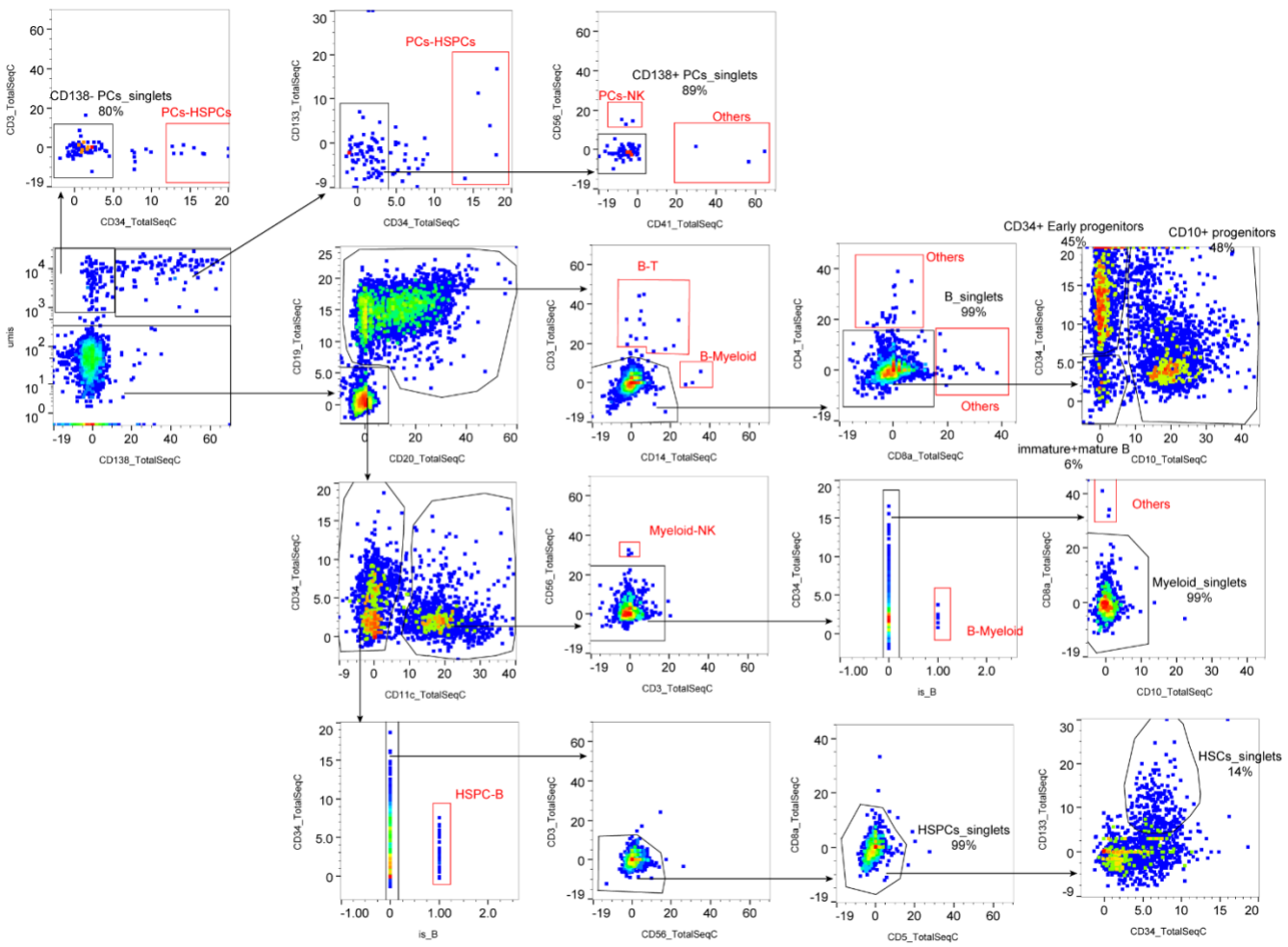

Figure S2. Biaxial feature plots to identify and remove cell doublets from BM cells. Supplement for Figure 3. Cell doublets were gated using red boxes. The visualization was generated by SeqGeq™ (FlowJo, LLC).

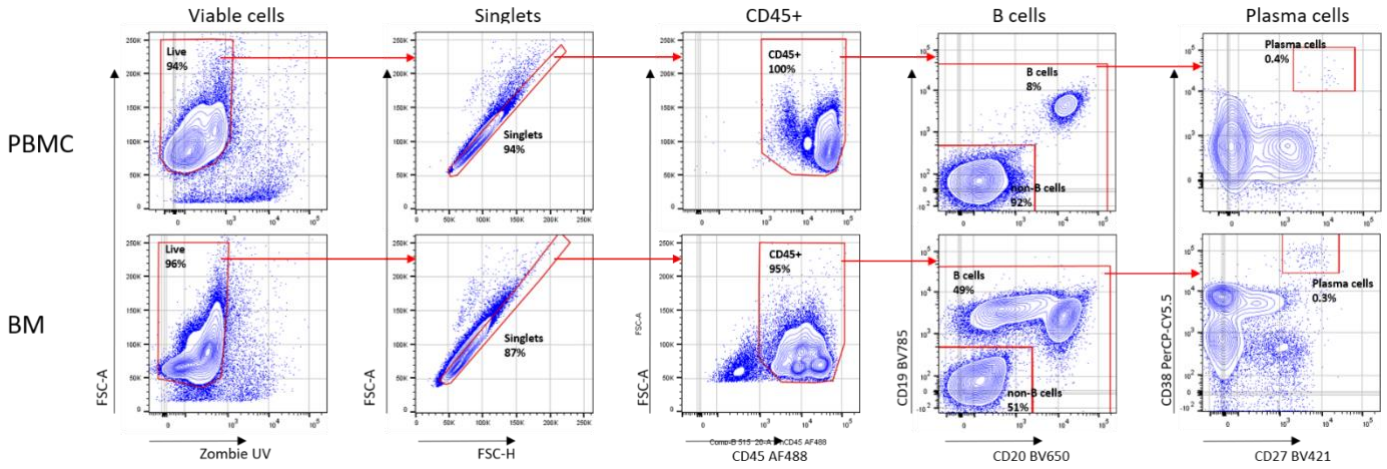

Figure S3. Gating strategy to identify major lineages by high-dimensional flow cytometry. An aliquot of each PBMC (top row) and BM (bottom row) sample was stained with a panel of fluorescently-labeled antibodies in parallel to the SuPERR-seq workflow. The gating strategy (red gates) shows the percentage of live, single B cells, and plasma cells.

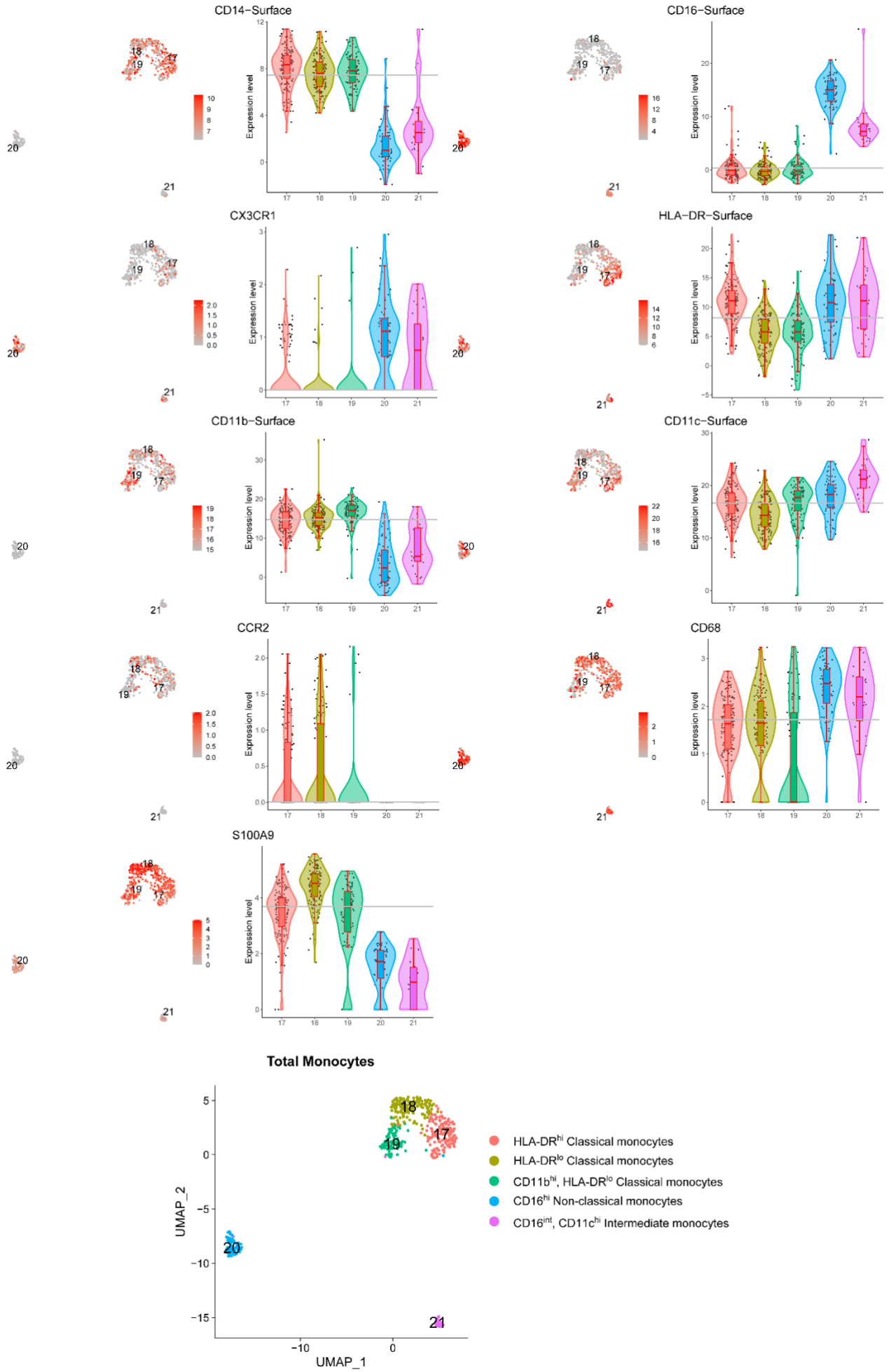

Figure S4. Blood monocytes identified by the SuPERR-seq workflow. UMAP and violin plots show surface markers and gene expression levels used to identify and classify five subsets of blood (PBMC) monocytes. The grey line shows the mean expression level across all clusters.

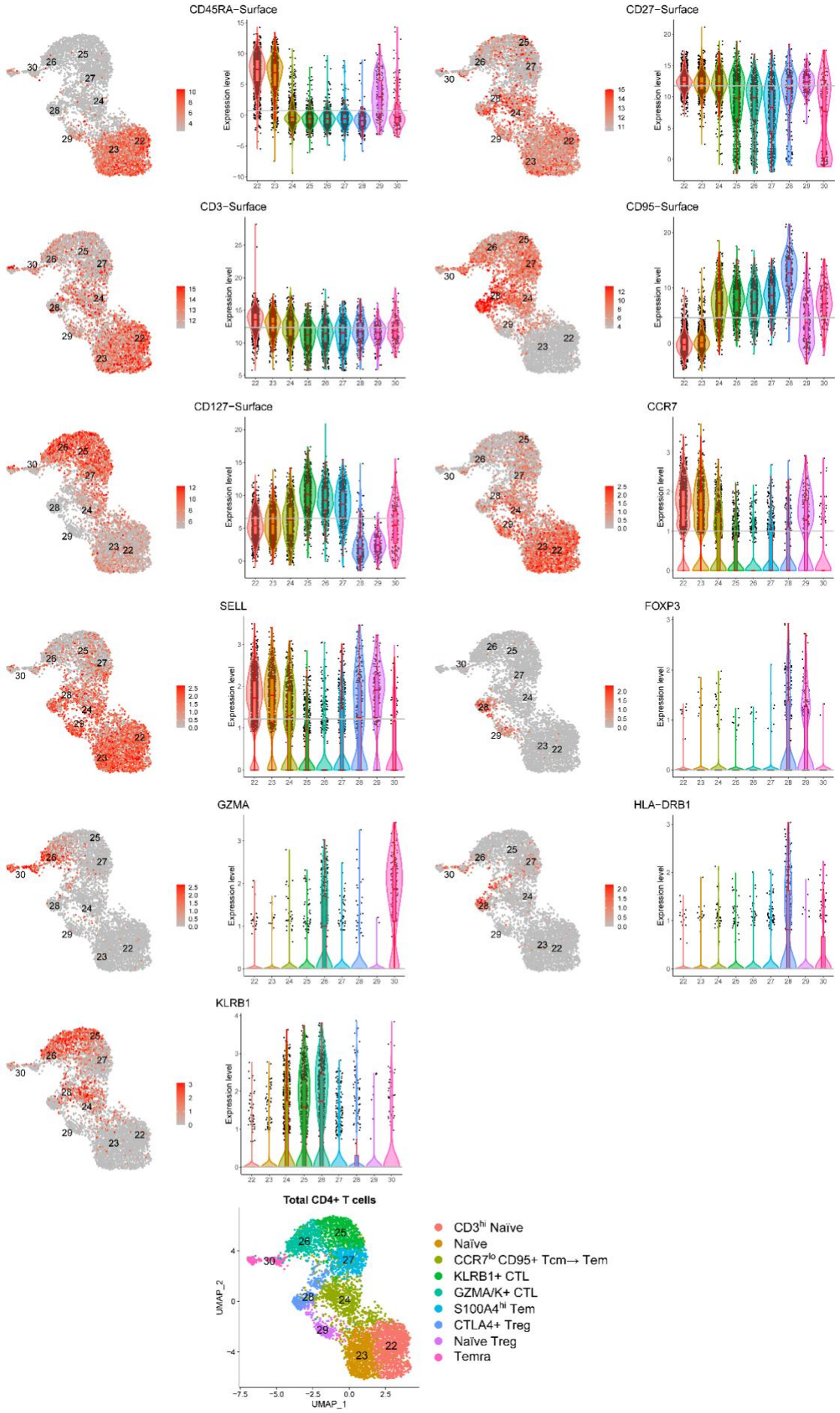

Figure S5. Blood CD4 T-cell subsets identified by the SuPERR-seq workflow. UMAP and violin plots show surface markers and gene expression levels used to identify and classify nine subsets of blood (PBMC) CD4 T cells. The grey line shows the mean expression level across all clusters. Tcm: central memory; Tem: effector memory; CTL: cytotoxic T lymphocyte; Treg: regulatory T cells; Temra: effector memory expressing CD45RA.

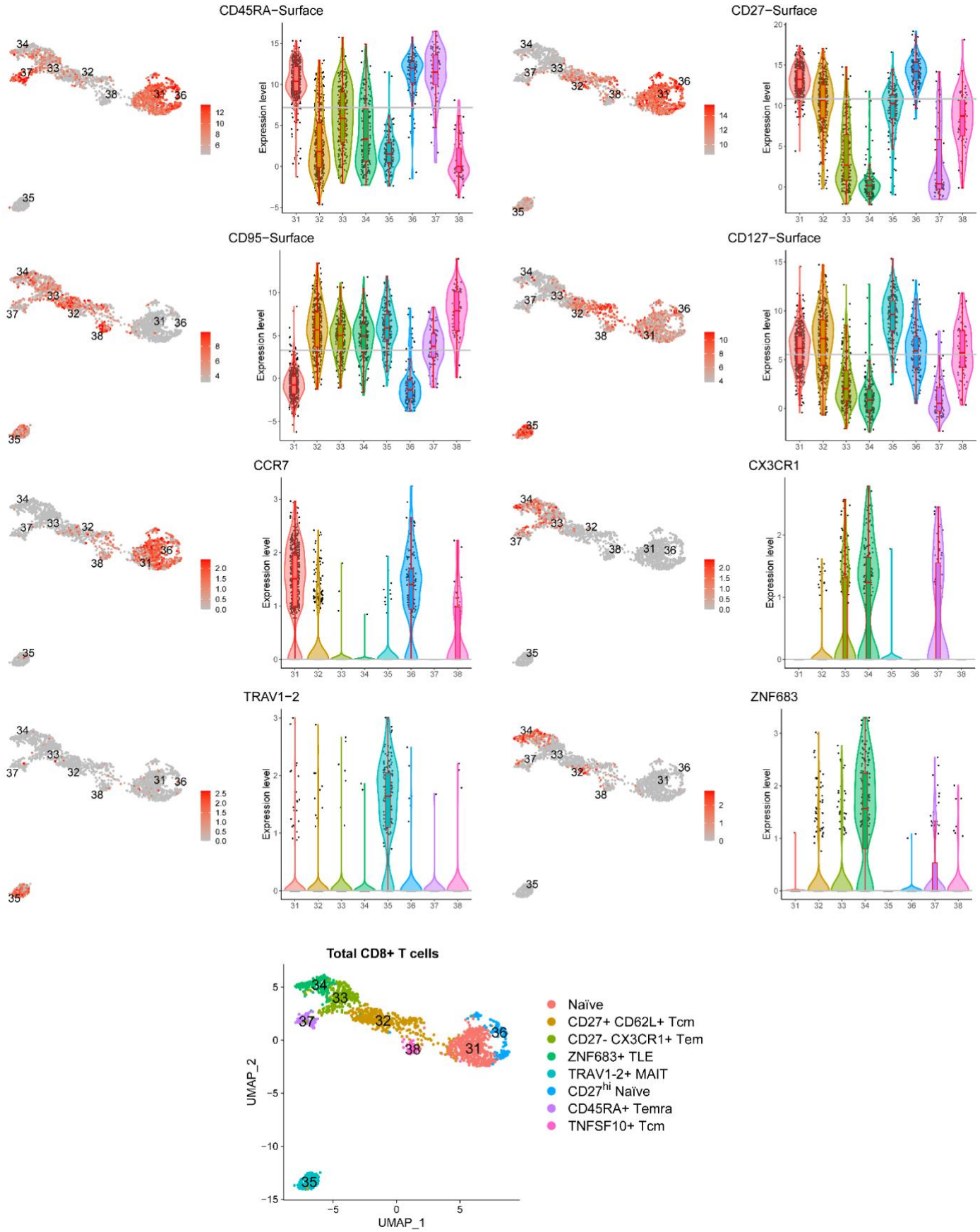

Figure S6. Blood CD8 T-cell subsets identified by the SuPERR-seq workflow. UMAP and violin plots show surface markers and gene expression levels used to identify and classify eight subsets of blood (PBMC) CD8 T cells. The grey line shows the mean expression level across all clusters. Tcm: central memory; Tem: effector memory; TLE: long-lived effector memory; MAIT: mucosal-associated invariant T cells; Temra: effector memory expressing CD45RA.

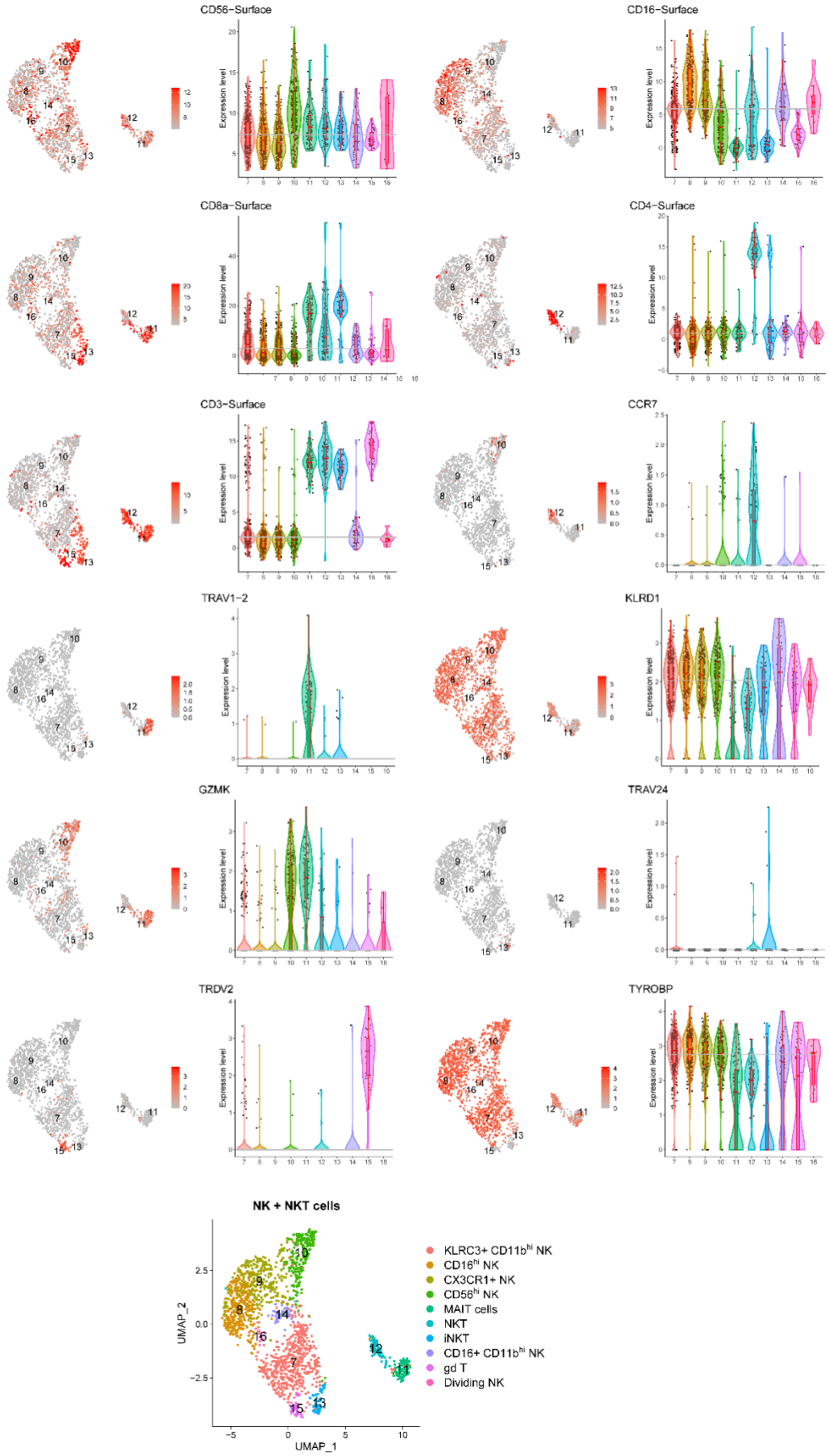

Figure S7. Blood NK, NKT, and gd T cell subsets identified by the SuPERR-seq workflow. UMAP and violin plots show surface markers and gene expression levels used to identify and classify ten subsets of blood (PBMC) NK, NKT, and gd T cells. The grey line shows the mean expression level across all clusters. MAIT: mucosal-associated invariant T cells; NKT: natural killer T cells; iNKT: invariant NKT; gd T: gamma-delta T cells.

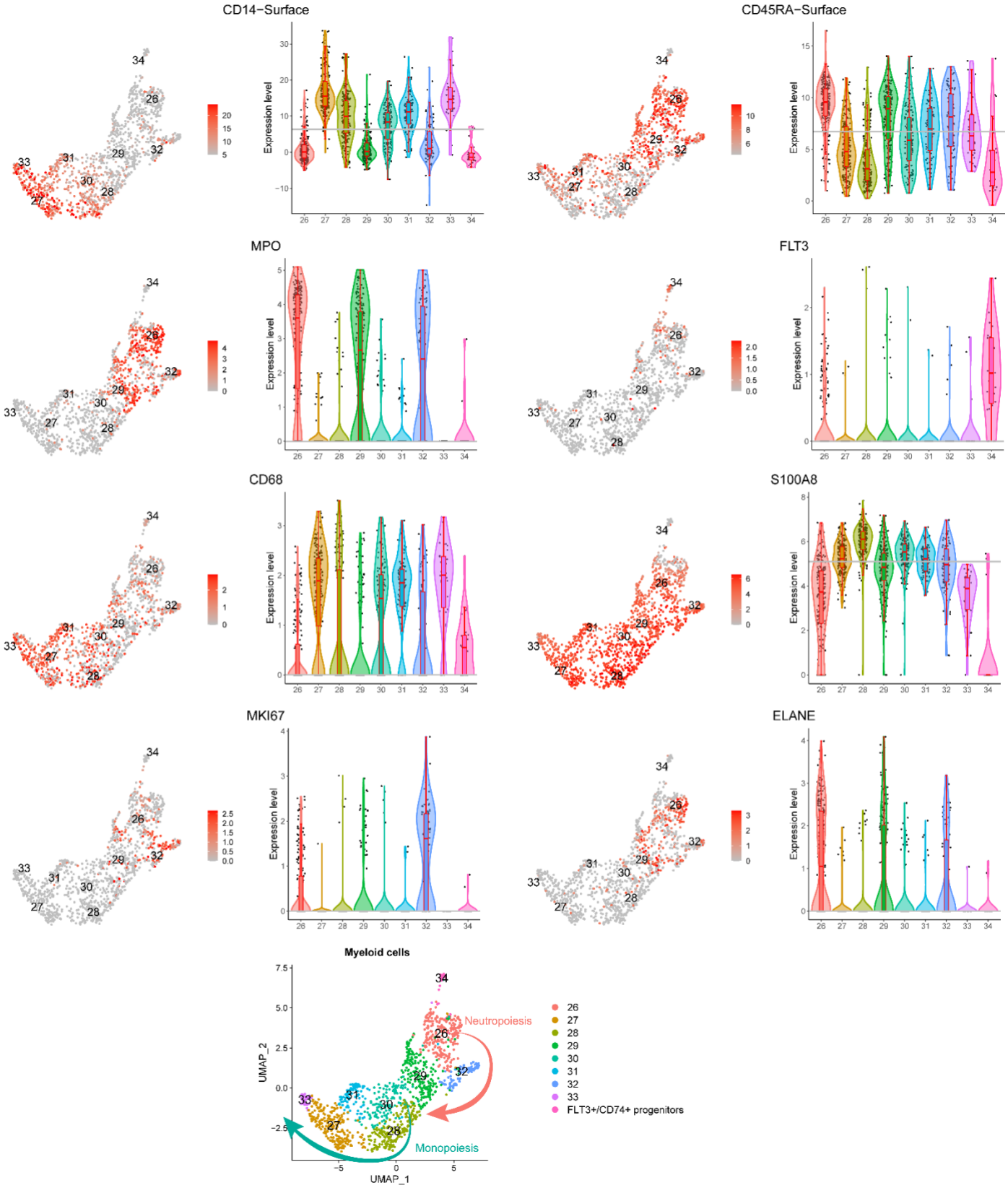

Figure S8. Monocyte and neutrophil development in the BM identified by the SuPERR-seq workflow. UMAP and violin plots show surface markers and gene expression levels used to identify and classify nine subsets of BM monopoiesis and granulopoiesis. The grey line shows the mean expression level across all clusters. The red arrow shows the potential direction of the developmental pathway for neutrophils. The cyan arrow shows the potential direction of the developmental pathway for monocytes.

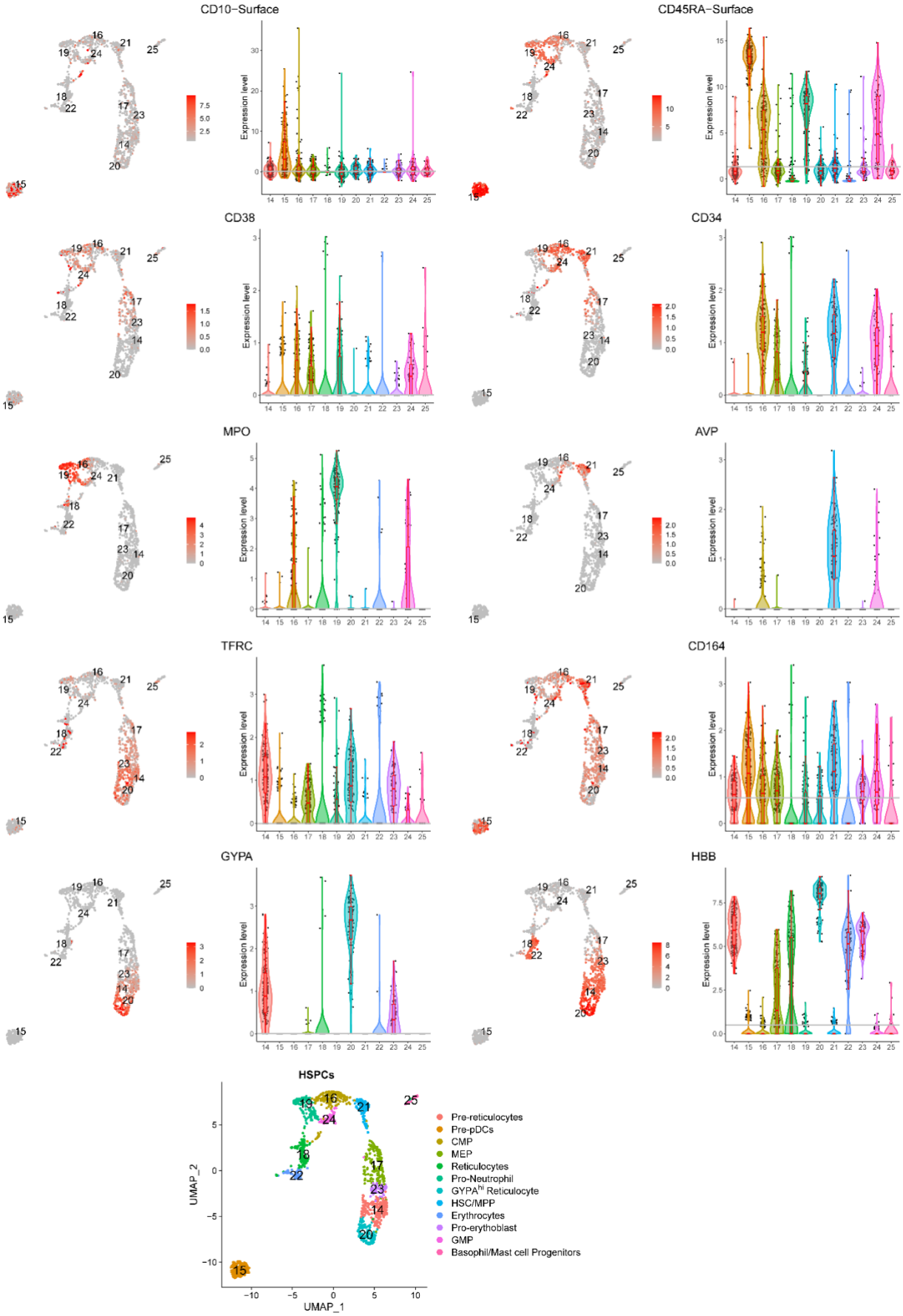

Figure S9. BM Hematopoietic Stem and Progenitor Cells (HSPCs) identified by the SuPERR-seq workflow. UMAP and violin plots show surface markers and gene expression levels used to identify and classify twelve subsets of BM HSPCs. The grey line shows the mean expression level across all clusters. pDCs: plasmacytoid dendritic cells; GMP: granulocyte-monocyte progenitor; MEP: megakaryocyte-erythroid progenitor; HSC: hematopoietic stem cell; MPP: multipotent progenitor; CMP: common myeloid progenitor.

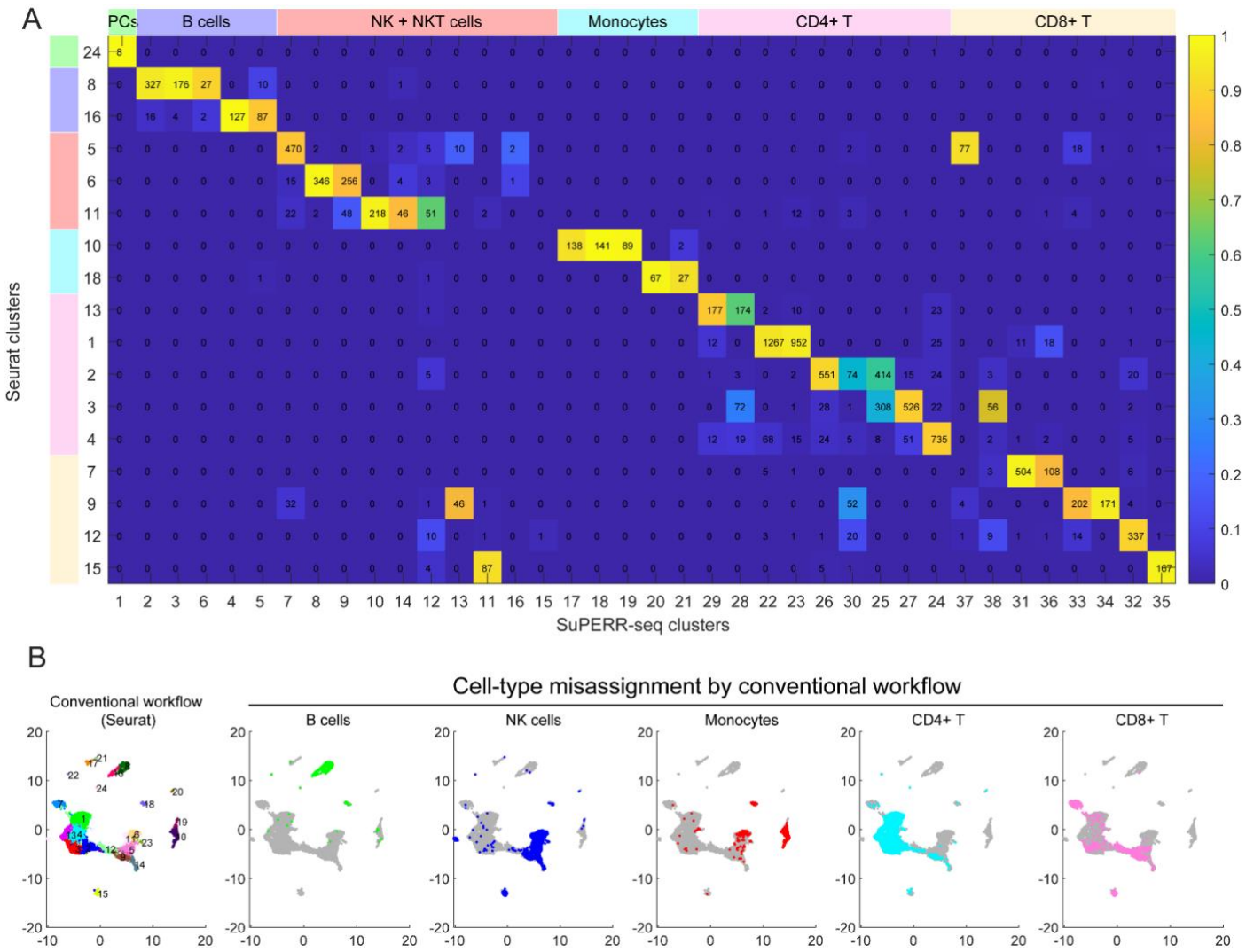

Figure S10. PBMC cell-type misassignment by conventional workflow. (A) Cross comparison of PBMC clustering results between conventional (Seurat) and SuPERR-seq workflows. For the sake of comparison, each workflow analyzed the same set of cell barcodes. The plasma cells identified by the SuPERR-seq "manual gating" using the Ig-specific transcripts were mostly removed from the conventional workflow (see Fig. 7A). Still, they were included here as SuPERR-seq cluster 1. (B) UMAP, and clustering results from the conventional workflow (Seurat). SuPERR-seq-defined major lineages were projected onto UMAP to reveal the misassignment/misclassification of cells by the conventional workflow. For example, cluster 9 from the conventional workflow (Seurat) represents a mixture of NK cells (blue), monocytes (red), CD4+ T cells (aqua), and CD8+ T cells (pink). Similarly, the conventional workflow misclassified NK cells, illustrated here by the spreading of the well-defined SuPERR-seq NK cells (blue) across multiple clusters in the conventional workflow.

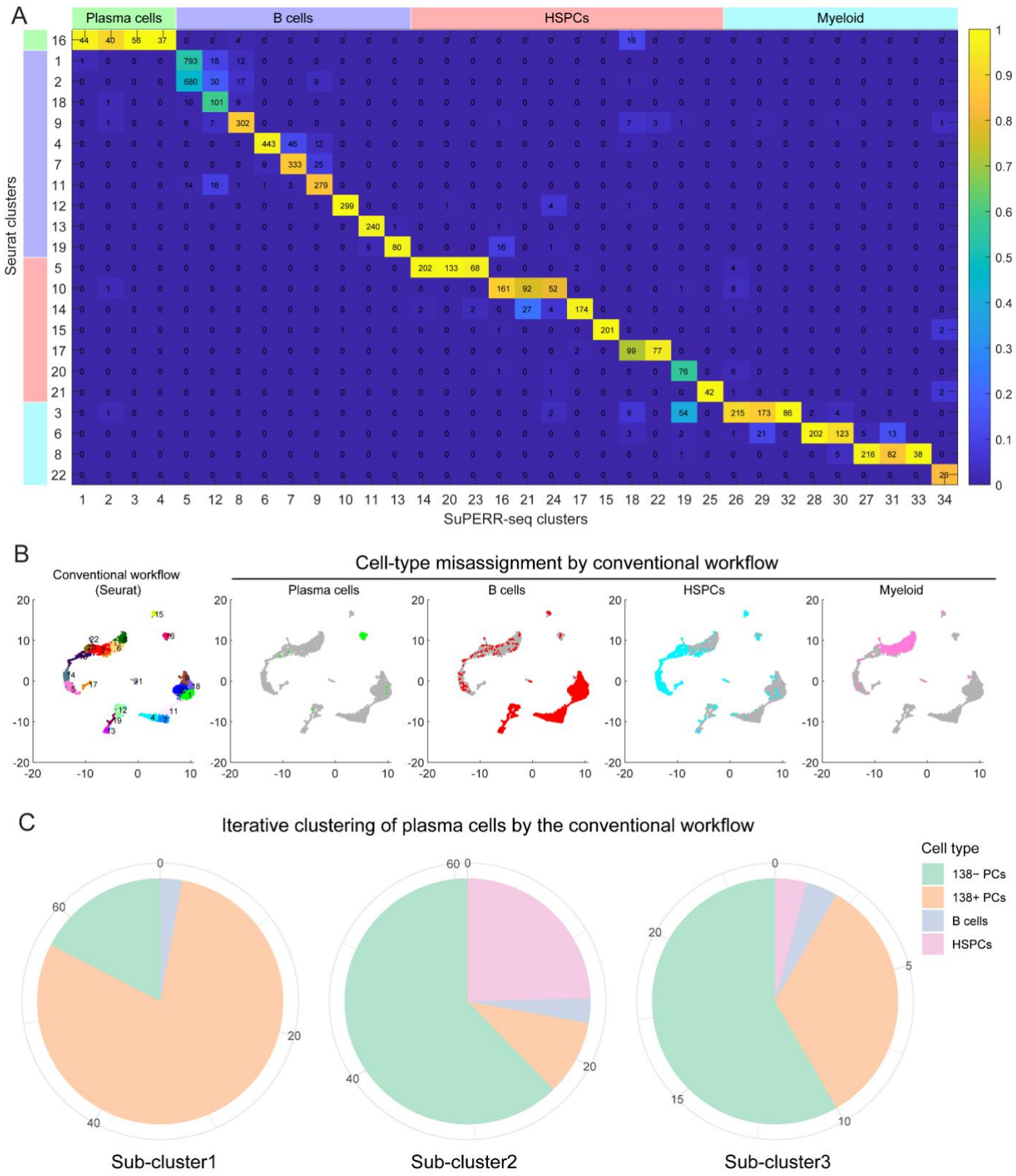

Figure S11. BM cell-type misassignment by conventional workflow. (A) Cross comparison of BM clustering results between conventional (Seurat) and SuPERR-seq workflows. For the sake of comparison, each workflow analyzed the same set of cell barcodes. (B) UMAP and clustering results from the conventional workflow (Seurat). SuPERR-seq-defined major lineages were projected onto UMAP to reveal the misassignment/misclassification of cells by the conventional workflow. For example, cluster 3 from the conventional workflow (Seurat) represents a mixture of plasma cells (green), B cells (red), and various HSPCs (blue), and myeloid cells (pink). Similarly, the conventional workflow misclassified B cells, illustrated here by the spreading of the well-defined SuPERR-seq B cells (red dots) across multiple clusters in the conventional workflow. (C) Cell-type composition in each “plasma cell” sub-cluster generated by iterative/recursive Seurat clustering.