nature genetics

Technical Report

https://doi.org/10.1038/s41588-024-02036-7

Transcript-specific enrichment enables profiling of rare cell states via single-cell RNA sequencing

Received: 6 May 2024

Accepted: 18 November 2024

Published online: 08 January 2025

Check for updates

Tsion Abay[1,2,3,11], Robert R. Stickels[1,2,11] ✉, Meril T. Takizawa[4,11], Benan N. Nalbant[5], Yu-Hsin Hsieh[5,6,7,8], Sidney Hwang[1,2], Catherine Snopkowski[4], Kenny Kwok Hei Yu[9], Zaki Abou-Mrad[9], Viviane Tabar[9], Brooke E. Howitt[1], Leif S. Ludwig[7,8], Ronan Chaligné[4] ✉, Ansuman T. Satpathy[1,2,10] ✉ & Caleb A. Lareau[5] ✉

Single-cell genomics technologies have accelerated our understanding of cell-state heterogeneity in diverse contexts. Although single-cell RNA sequencing identifies rare populations that express specific marker transcript combinations, traditional flow sorting requires cell surface markers with high-fidelity antibodies, limiting our ability to interrogate these populations. In addition, many single-cell studies require the isolation of nuclei from tissue, eliminating the ability to enrich learned rare cell states based on extranuclear protein markers. In the present report, we addressed these limitations by developing Programmable Enrichment via RNA FlowFISH by sequencing (PERFF-seq), a scalable assay that enables scRNA-seq profiling of subpopulations defined by the abundance of specific RNA transcripts. Across immune populations (n = 184,126 cells) and fresh-frozen and formalin-fixed, paraffin-embedded brain tissue (n = 33,145 nuclei), we demonstrated that programmable sorting logic via RNA-based cytometry can isolate rare cell populations and uncover phenotypic heterogeneity via downstream, high-throughput, single-cell genomics analyses.

Rare cell types are found across diverse biological contexts and often participate critically in tissue function, represent key transitional states or define incipient disease. Single-cell RNA sequencing (scRNA-seq) has powered the discovery of many low-frequency populations defined by cell-type-specific transcript combinations. However, their limited sampling makes such populations difficult to study. For example, *CFTR*$^+$ pulmonary ionocytes likely mediate the pathogenesis of cystic fibrosis but make up only 1 in 200 human lung epithelial cells[1]. Similarly, enteric neurons are present in 1 in 300 cells in the colon[2] while *AXL*$^+$*SIGLEC6*$^+$ dendritic cells (AS DCs) occur in only 1 in 5,000 peripheral blood mononuclear cells (PBMCs)[3,4]. More recently, we identified a rare population of chimeric antigen receptor T cells that reactivate human herpesvirus 6 (HHV-6) at a frequency of ~1 in 10,000 cells in infusion products[5], which may contribute to the etiology of HHV-6 encephalitis in patients receiving cell therapies. Although these anecdotes represent diverse populations and tissue types, the conceptual

[1]Department of Pathology, Stanford University, Stanford, CA, USA. [2]Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA. [3]Program in Biological and Biomedical Sciences, Harvard University, Boston, MA, USA. [4]Single-cell Analytics Innovation Lab, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [5]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [6]Charité Universitätsmedizin Berlin, Berlin, Germany. [7]Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Berlin, Germany. [8]Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association, Berlin Institute for Medical Systems Biology, Berlin, Germany. [9]Department of Neurosurgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [10]Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA. [11]These authors contributed equally: Tsion Abay, Robert R. Stickels, Meril T. Takizawa. ✉e-mail: rstickel@stanford.edu; chalignr@mskcc.org; satpathy@stanford.edu; lareauc@mskcc.org

mode of discovery for these rare populations has been consistent: genomic profiles of ~10^5–10^7 cells were generated, yielding ~10^1–10^3 events of interest. The tremendous resources required to define rare but consequential populations limit the power of downstream analyses, including refining transcriptional heterogeneity within these populations and identification of marker genes.

After rare populations have been identified and defined by marker transcripts via scRNA-seq, challenges remain in isolating them for further characterization. A frequently used approach is the enrichment or depletion of cells expressing specific surface proteins via FACS. However, as these rare populations are defined through transcriptomics analyses, analogous surface proteins may not be defined or high-quality antibodies unavailable. Furthermore, many archived sample preparations require nuclei dissociation steps from frozen or formalin-fixed, paraffin-embedded (FFPE) tissue samples, which eliminates the possibility of enrichiment or depletion based on non-nuclear proteins. Although the profiling of intranuclear proteins and scRNA-seq was recently demonstrated[6], high-quality antibodies recognizing transcription factors (TFs) often do not exist, owing to a lack of highly structured antigens available for targeting[7]. These limitations motivate an approach that enriches for either cells or nuclei based on individual or combinations of RNA markers upstream of additional scRNA-seq profiling.

To this end, we report the development of Programmable Enrichment via RNA FlowFISH by sequencing (PERFF-seq). Our assay utilizes FISH to specifically label RNA(s) that enrich(es) populations of interest and is compatible with downstream high-quality transcriptional profiling via high-throughput, droplet-based scRNA-seq. We identified parameters in currently utilized FISH workflows that diminish downstream scRNA-seq performance and determined conditions that yield high-quality data on a par with libraries derived from commercially available protocols. We demonstrate the broad applicability of PERFF-seq to enrich immune cell subsets and nuclei extracted from frozen or FFPE tissue samples using one or more RNA markers.

Results

Assay rationale and overview

Although tissues preserved in formaldehyde, including FFPE tissue blocks, are broadly available, extracting RNA from these sources has remained a major challenge[8]. Notably, formaldehyde-associated RNA degradation is a major limiting factor in generating high-quality libraries from fixed cells, because the degraded RNA molecules cannot be reliably reverse transcribed for conventional downstream analyses. A recently introduced workflow from 10x Genomics, termed Single Cell Gene Expression Flex[9] ('Flex' hereafter), profiles fixed cells using whole transcriptome probe pairs that hybridize and ligate before single-cell barcoding (Extended Data Fig. 1a). In parallel, advances in FISH technologies, including hybridization chain reaction (HCR), generate an amplified fluorophore signal with high signal-to-noise separation[10] upon hybridization to RNA molecules of interest[11,12], enabling flow cytometry-based detection and enrichment of specific populations at high throughput. Notably, nearly all RNA FISH technologies require an irreversible fixation step, hindering conventional reverse transcription-based, scRNA-seq technologies. Taken together, we envisioned a programmable method for enriching cells via RNA FISH and flow cytometry-based sorting, followed by single-cell transcriptome profiling via the Flex scRNA-seq workflow (Fig. 1a). Such an approach would enable the study of heterogeneity underlying cell states based on any marker gene, including rare cell states[1,2,4,5,13].

PERFF-seq development and quality control

We first tested the concatenation of the HCR–FlowFISH workflow[11,14] with Flex library preparation. After sorting PBMCs for the B cell-specific marker gene, *MS4A1* (encoding CD20), we captured cells and compared quality control measures with a standard Flex library of only PBMCs (Supplementary Table 1 and Methods). Despite targeting a similar number of cells (~3,000), the FlowFISH-sorted cells yielded far fewer cells with overall worse per-cell data quality than the standard Flex library (Fig. 1b and Supplementary Table 2). We noticed that this FlowFISH → Flex v.0 library contained many reads half-mapping to the probe set, indicating poor ligation efficiency and resulting in barcoding of the unligated Flex probes (Fig. 1c and Extended Data Fig. 1b,c). These inferences were corroborated by Bioanalyzer traces, which verified that the extension of the truncated probes limited per-cell data quality (Fig. 1d,e and Methods). Thus, to our surprise, the direct integration of FlowFISH into fixed-cell scRNA-seq is incompatible with attaining high-quality genomics data.

We interrogated the effect of individual FISH workflow components, including dextran sulfate, which is used for background suppression but can suppress enzymatic activity[11,15,16], and reagents for generating fluorescent signal. Although dextran sulfate had no impact, the generation of the HCR polymer was detrimental to data quality (1.2% versus 76% full probe reads; Fig. 1f). As the formation of the gene probe–hairpin polymer is essential for FISH signal amplification, we reasoned that removing the tethered polymer after sorting, but before scRNA-seq library preparation, could rescue transcriptional profiling. Thus, we sought to disassemble the HCR polymer via formamide[11,15] and enzymatic[10,17] stripping, both used in imaging and microscopy analyses (Methods). Although formamide resulted in probe–hairpin polymer stripping (Extended Data Fig. 1d), it inhibited the capture of any ligated product, irrespective of the presence of HCR polymer (Fig. 1g). Alternatively, digestion of the HCR polymer via double-stranded (ds)DNase (Thermo Fisher Scientific) reduced the fluorescence signal as expected from prior microscopy work[17], but yielded markedly higher library complexity and ligation efficiency (Fig. 1g and Extended Data Fig. 1e). Sorting using bovine serum albumin (BSA) with an RNase inhibitor (rather than fetal bovine serum (FBS) as used by prior HCR–FlowFISH applications[14]) further improved scRNA-seq library quality (Fig. 1h and Methods). The result was a FlowFISH-sorted Flex library with 90% of molecules mapping to a full probe set and 4.2% mapping to half-ligated probes, which was confirmed via Bioanalyzer traces (Extended Data Fig. 1f). Performing the full workflow with all modifications except the dsDNase step confirmed that polymer stripping is the critical step for avoiding incomplete probe ligation and generating high-quality scRNA-seq libraries (Extended Data Fig. 1g and Methods). These optimization steps for FlowFISH sorting and profiling via scRNA-seq comprise the PERFF-seq assay, resulting in thousands of unique molecular identifiers (UMIs) and unique genes detected per cell (Fig. 1i).

PERFF-seq benchmarking

Next, we assessed HCR–FlowFISH enrichment across well-defined markers and populations. Staining PBMCs for *ACTB* with AF647- and AF488-labeled amplifiers allowed us to assess HCR–FISH sensitivity. With either fluorophore, 97% of cells were positive for *ACTB*, indicating high sensitivity (Fig. 2a). Similarly, HCR–FISH of *EPCAM* across four different cell lines showed an increase in mean fluorescence intensity (MFI) consistent with *EPCAM* abundance from RNA-seq (Extended Data Fig. 2a). Next, we designed probes against *XIST*, a long noncoding (lnc) RNA expressed in cells with XX sex chromosomes (K-562), but absent from XY cells (Raji). We mixed Raji and K-562 human cell lines in varying proportions that spanned two orders of magnitude (Methods). Flow cytometry confirmed the sensitive and specific detection of *XIST*$^+$ populations in as few as 1 in 500 cells (Fig. 2b and Extended Data Fig. 2b), confirming feasibility with noncoding transcripts and suggesting that PERFF-seq could be used to study a broad range of RNA molecules.

To assess the performance of the scRNA-seq profiles, we multiplexed the various experimental conditions using the native multiplexed barcodes in Flex to compare the changes required to implement PERFF-seq (Fig. 2c and Methods). Specifically, in addition to the standard Flex and PERFF-seq protocols, we assessed the impact of

https://doi.org/10.1038/s41588-024-02036-7

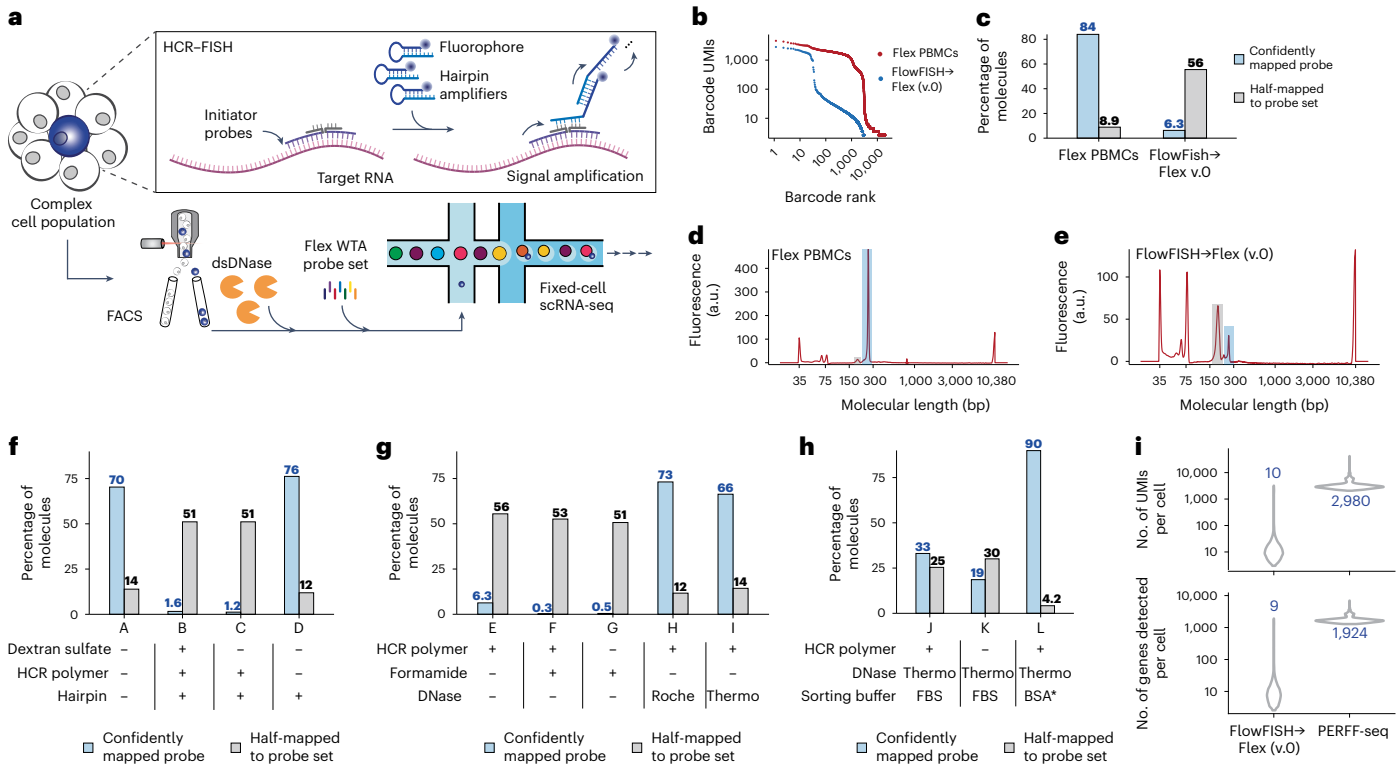


Fig. 1 | Rationale and development of PERFF-seq. a, Schematic of the PERFF-seq assay. Target RNA(s) is(are) bound by pairs of adjacent initiator probes that ensure specificity. Hairpin amplifiers unzip and hybridize iteratively to generate fluorescent signal and enable FACS before single-cell profiling with the droplet-based scRNA-seq Flex kit. **b**, Knee plot of cells profiled with standard Flex versus HCR–FlowFISH-sorted cells. **c**, Fraction of reads fully mapping (blue) or half-mapping (gray) to the reference probe set. **d**, Bioanalyzer traces highlighting the expected product size of the full probe (~260 bp; blue) and half-probe (~190 bp; gray) for a high-quality Flex library. **e**, Same as **d** but for the FlowFISH → Flex v.0 experiment. **f**, Experiments identifying the HCR polymer as the corrupting agent for data quality. **g**, Conditions screened for polymer stripping, including DNase and formamide treatments. **h**, Sorting buffer components analyzed to improve data quality. **i**, UMIs (top) and genes (bottom) detected per cell comparing the initial FlowFISH → Flex v.0 experiment, from **b** to the final PERFF-seq library from **h**. Median values are shown in blue. Values plotted in **c** and **f**–**i** represent overall library values for a single replicate. a.u., arbitrary units; Thermo, Thermo Fisher Scientific.

the changes required for HCR–FlowFISH (for example, buffers and dsDNAse treatment: 'no probe, no sort' (NPNS)) as well as the probe staining steps ('yes probe, no sort' (YPNS)). Across this comparison, we profiled 59,313 cells, recovering the major cell types expected from PBMCs (Extended Data Fig. 2c,d). For the PERFF-seq library, we sorted for *CD3E*$^+$ cells, yielding ~95–97% T cells (Fig. 2d,e, Extended Data Fig. 2e and Methods). We observed that adding *CD3E* probes reduced the transcript expression (~2× reduction of $\log_2$(counts per million)) but with minimal changes across the whole transcriptome (Extended Data Fig. 2f–h and Methods). These results suggest that the transcripts targeted by FlowFISH can still be detected via scRNA-seq but warrants caution for interpreting quantitative gene expression levels.

Next, we integrated PERFF-seq data with a gold-standard public PBMC dataset from 10x Genomics. After downsampling for consistency (to 16,750 reads per cell based on the lowest library), we observed almost no loss in data quality. Key data quality metrics were consistent within ~10% comparing public Flex (median 5,200 UMIs and median 2,857 genes detected per cell) with our Flex (median 4,885 UMIs and median 2,536 genes) and PERFF-seq (median 4,789 UMIs and median 2,535 genes) libraries, together verifying high-quality scRNA-seq data (Fig. 2f, Supplementary Table 3 and Methods).

Multicolor programmable enrichment with PERFF-seq

Many cell states are identifiable by the combinatorial expression of two or more genes. In PBMCs, CD4$^+$ T cells are important coordinators of the immune response that co-express *CD3E* and *CD4* genes. We sought to assess the utility of PERFF-seq in this setting and designed probes against three well-described genes, *CD3E*, *CD4* and *MS4A1* (CD20; B cells), to target cells in the lymphoid lineage (Fig. 3a). We recovered four cell populations using PERFF-seq, including a *CD4*$^+$, *CD3E*$^+$ population that together specifically enrich for CD4$^+$ T cells (Fig. 3b). PERFF-seq profiling of the 4 populations yielded 35,220 cells (Fig. 3c and Methods). We co-embedded these PERFF-seq libraries with Flex PBMCs, confirming the concordance of cell states between methods, and verified the quality of enrichments via three independent methods (Fig. 3c–e, Extended Data Fig. 3a,b and Methods). Our results showed that PERFF-seq has 70–88% accuracy in recovering the desired cell-type labels based on programmable marker-gene analyses.

Next, we performed differential gene expression analyses of *CD3E*$^+$ cells that either co-expressed or were negative for the *CD4* transcript. Reassuringly, the top differential transcripts coincided with key CD4$^+$ and CD8$^+$ T cell markers (Fig. 3f and Methods). Moreover, reclustering of the 9,035 cells from the *CD4*$^+$*CD3E*$^+$ population identified major CD4$^+$ T cell subsets, including regulatory T cells (T_{reg} cells) expressing *FOXP3*, naive T cells marked by *LEF1*, central/effector memory cells marked by *BHLHE40*, interferon-responsive cells expressing *IFIT* genes[18] and cytotoxic CD4$^+$ cells expressing granzyme H (*GZMH*; Extended Data Fig. 3c). Although our analyses confirmed enrichment of CD4$^+$ T cells, we noted that the overall *CD4* HCR–FlowFISH signal was lower in the *CD3E*$^+$ population, reflective of *CD4* messenger RNA expression in myeloid populations but the opposite of surface CD4 protein quantifications (Extended Data Fig. 3d,e). We confirmed this increase of *CD4* mRNA in myeloid cells using a combination of conventional antibody staining and *CD4* HCR–FlowFISH (Extended Data

https://doi.org/10.1038/s41588-024-02036-7

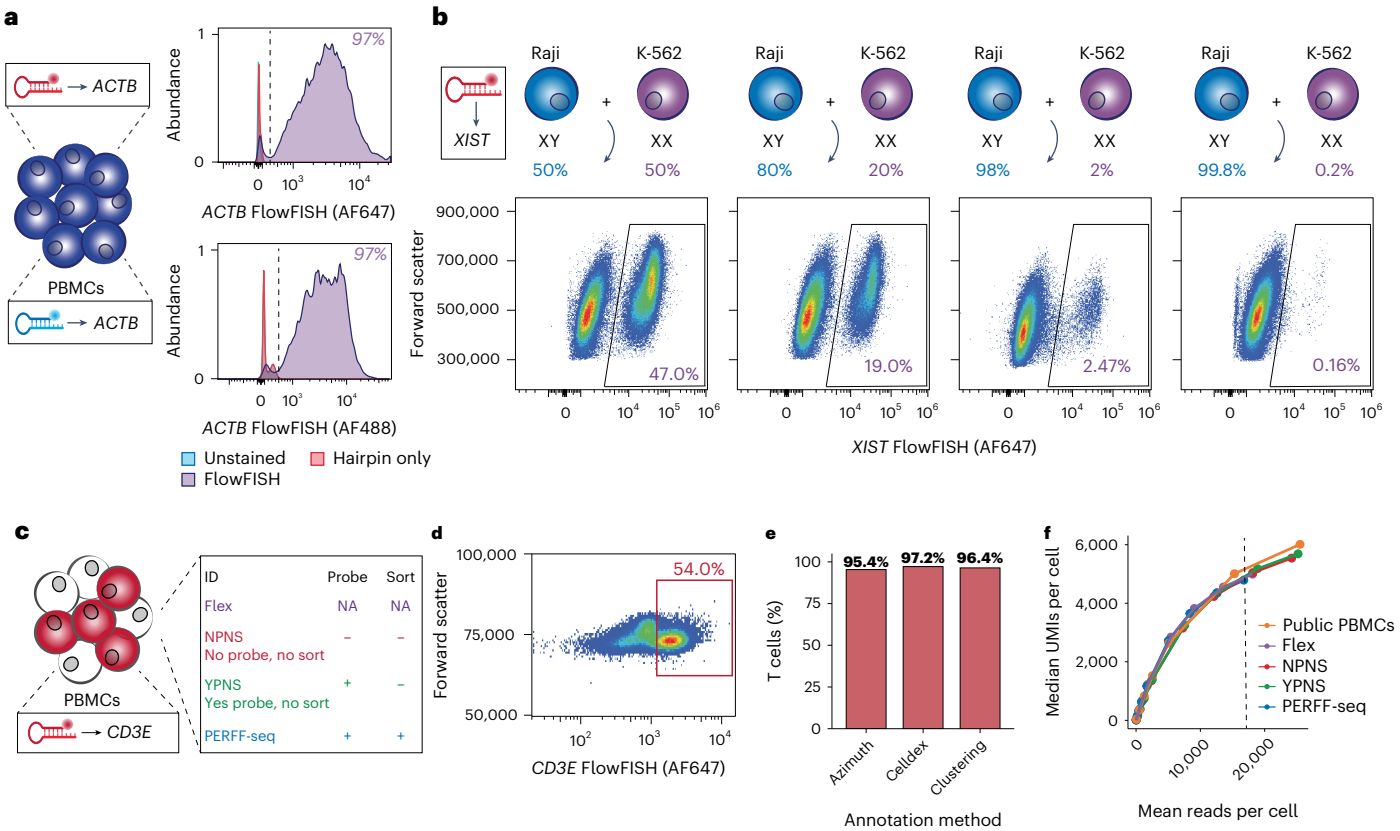


Fig. 2 | Benchmarking of PERFF-seq. a, Sorting of PBMCs stained with AF647- and AF488-labeled *ACTB* probes. The percentage of positive events is shown at the top right. **b**, Benchmarking of lncRNA *XIST* FlowFISH by diluting K-562 cells (XX/*XIST*$^+$) into Raji cells (XY/*XIST*$^-$) at varying ratios. The expected cell ratios per line are noted above the flow plot and the percentage of *XIST*$^+$ events are reported in the gate. **c**, PERFF-seq benchmarking experiment for four libraries, including the standard Flex workflow with or without PERFF-seq probe staining or sorting steps. PERFF-seq was enriched for *CD3E*$^+$ cells. All tested conditions: NPNS and YPNS profiles of all PBMC subpopulations with minor modifications to the reaction. **d**, Sorting strategy for *CD3E*$^+$ cells for the PERFF-seq library. **e**, Proportion of cells from the sorted PERFF-seq library annotated as T cells using three different computational methods for classification. **f**, Downsampling analysis for library saturation and UMI benchmarking. The dashed line represents the mean reads per cell for a final comparison (depth of lowest sample, 16,755 reads per cell). NA, not available.

Fig. 3f), demonstrating that marker expression may not always reflect protein expression.

Rare cell states enriched via transcriptional regulators

Next, we explored using RNA features typically inaccessible to antibody-based flow cytometry. We elected to enrich populations of PBMCs expressing the hematopoietic lineage-defining TFs, *BCL11A* and *SPI1* (encoding PU.1), which regulate lymphoid[19] and myeloid development[20], respectively. These markers exhibit robust expression in public scRNA-seq atlases and we profiled a total of 37,566 cells from *SPI1*$^+$, *BCL11A*$^+$ and double-negative libraries (Fig. 4a,b and Extended Data Fig. 4a,b). Cell states segregated readily based on the sorted TFs, but also as a function of well-established (surface) marker transcripts (Fig. 4c and Extended Data Fig. 4c). Specifically, we observed enrichment as 96.0% and 91.1% of cells sorted for *SPI1* and *BCL11A* had non-zero UMI counts for those genes (compared with 13.3% and 9.5% in the unsorted fraction, respectively), with over half of the positive cells having at least 10 UMIs in sorted libraries (Fig. 4d and Extended Data Fig. 4d). Standard clustering and cell-type annotation showed a clear skewing of cell states in either enriched TF library, including the consistent and desired depletion of T cells that do not express either TF (Fig. 4e,f).

Although *BCL11A* is an essential regulator of B lymphoid development, nearly 75% of the enriched cell population from its PERFF-seq library were plasmacytoid DCs (pDCs). This enrichment was consistent with bulk[21] and single-cell expression data[13] indicating that pDCs express *BCL11A* at one to two orders of magnitude higher levels than B cells (Extended Data Fig. 4a,e and Methods). Co-staining of HCR–FlowFISH and surface antibodies confirmed the marked increase of *BCL11A* in pDCs relative to B cells by nearly 5–6× (Extended Data Fig. 4f and Methods). Although B cells had a modest enrichment in the *BCL11A*$^+$ PERFF-seq library, these B cells had higher *BCL11A* expression and TF target gene module scores than B cells from the other two libraries (Extended Data Fig. 4g and Methods), indicating that PERFF-seq can enrich for cells with higher TF activities within specific cell types. Furthermore, *SPI1* staining primarily enriched monocyte and classic DCs (cDCs), consistent with its expression in PBMCs (Fig. 4f and Extended Data Fig. 4a).

We observed that AS DCs, an ultra-rare cell type present at ~1 in 5,000 PBMCs, had an almost 5-fold enrichment in both *BCL11A*- and *SPI1*-enriched populations (Fig. 4g). Subclustering of the 269 cells from the sorted populations confirmed consistent expression of *AXL* and *SIGLEC6*, verifying their AS DC identity (Fig. 4h). We hypothesized that our FlowFISH-sorted, lineage-defining TFs may be associated with heterogeneity in this cell state. Indeed, differential gene expression analyses between the *SPI1*$^+$- and *BCL11A*$^+$-sorted populations identified key markers of the pDC-like subset (for example, *IL3RA* and *MZB1*), which are distinct from a cDC-like subset (for example, *IFI30* and *ITGAX*; Fig. 4i,j and Supplementary Table 4). PERFF-seq differential analyses further identified genes that may play a functional role in these populations, including surface markers (*CD5* and *TLR9*), granzymes (*GZMB*) and cytokines (*IL1B*; Fig. 4j and Extended Data Fig. 4h). Reanalysis of

https://doi.org/10.1038/s41588-024-02036-7

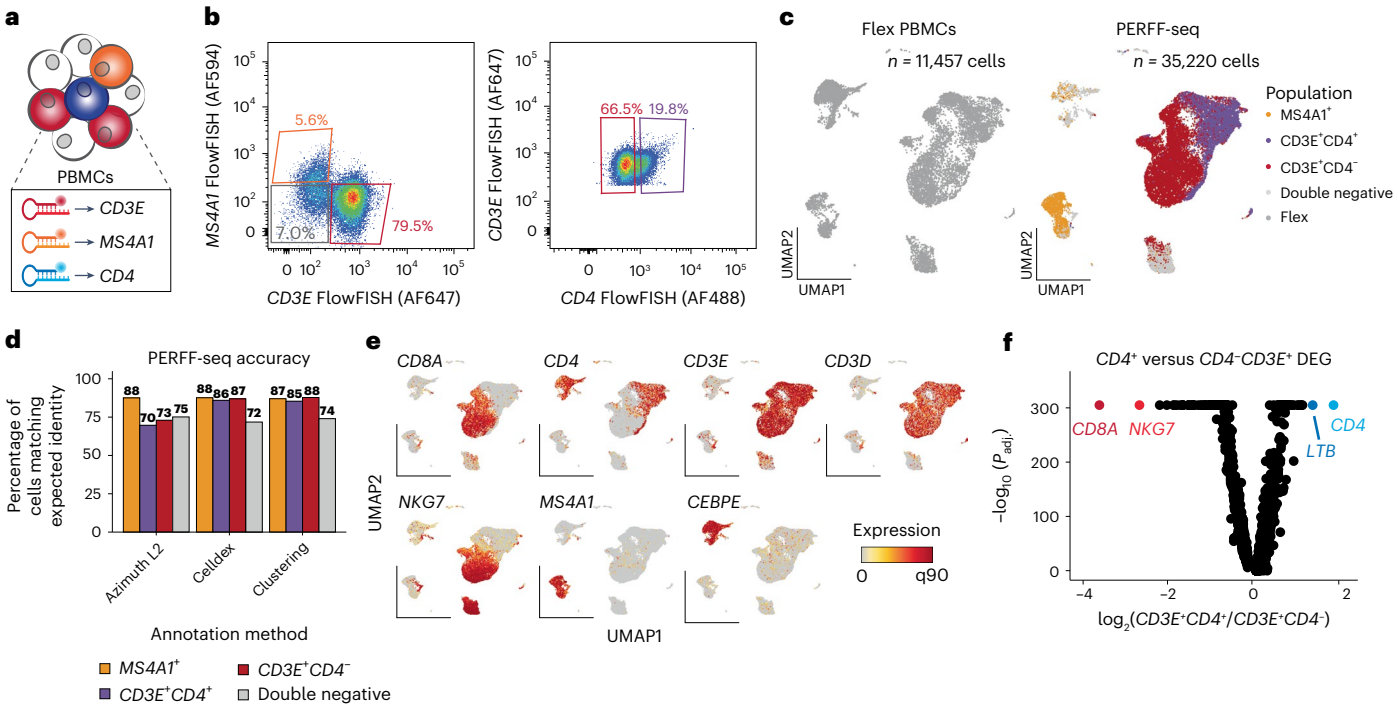


Fig. 3 | Enrichment of cells with multicolor and multigene panels. a, Schematic of experimental design. Probes targeting three indicated genes are each labeled with distinct fluorophore stain-specific populations in PBMCs from a healthy human donor. **b**, FlowFISH signal and sort gates. Percentages represent the overall fraction of events sorted in each gate. **c**, Reduced dimensionality representation of four populations profiled with PERFF-seq (right) co-embedded with a standard Flex library of PBMCs (left). The colors represent the gates drawn from the FlowFISH sort in **b**. **d**, Percentage of high-quality cells from PERFF-seq assigned to expected cell types using three distinct annotation methods. The colors represent the gates drawn from the FlowFISH sort in **b**. **e**, Annotation of relevant marker genes for populations in reduced dimensionality space, including genes used in the FlowFISH panel. **f**, Differentially expressed gene (DEG) analysis comparing *CD4*$^+$ and *CD4*$^-$ populations from the *CD3E*$^+$ sort (**b**, right). Genes corroborating annotation are highlighted. The Bonferroni-adjusted *P* value for the two-sided Wilcoxon's rank-sum test is shown with at a minimum of 1×10^{-314} for machine precision.

published Smart-seq2 data confirmed the co-expression of *BCL11A* and *SPI1* in these AS DC populations with marker genes previously identified for the cDC-like and pDC-like subpopulations[4] (Fig. 4k). PERFF-seq targeting *IL3RA* with an inclusive sort gate yielded 9,178 profiles, including 95 additional AS DCs with similar gene modules to the Smart-seq2 and TF PERFF-seq results, corroborating the lineage-defining *BCL11A* and *SPI1* TFs as candidate regulators of heterogeneity in this rare cell state (Extended Data Fig. 4i–l). Our PERFF-seq profiles of *IL3RA*$^+$ cells included many other T cell, B cell and other myeloid populations, consistent with bulk mRNA profiles in PBMCs[21] (Extended Data Fig. 4m,n). However, *IL3RA* encodes CD123, an AS DC- and pDC-specific surface marker, resulting in different populations sorting on RNA or protein. This result underscores the importance of evaluating transcriptional data for PERFF-seq targets rather than relying on established knowledge or surface protein marker expression alone.

Resolving somatic mosaicism with PERFF-seq

To further evaluate the sensitivity of PERFF-seq to lowly expressed genes, we built an OR-gated panel targeting multiple genes co-expressed in individual cells to boost sensitivity. Conceptually, if any one or combination is expressed at sufficient levels, a fluorescent signal would be generated, allowing for the isolation of the population. We sought to study mosaic loss-of-Y chromosome (LOY), the most common, age-related, somatically acquired mutation in the male genome[22]. When LOY occurs in a hematopoietic stem and progenitor cell, the descendant, terminally differentiated cells present with LOY across multiple lineages with a myeloid skew[23]. To capture LOY events, we designed a ten-gene FlowFISH panel for the male-specific region of the Y chromosome (MSY; Fig. 5a). We elected to put the *MSY* genes together because each is lowly expressed individually (1.5 UMIs per gene per positive cell; sum 5.5 UMIs per positive cell) compared with other targets (13.4 UMIs; Fig. 5b). Profiling four healthy donors aged from 20 years to 51 years via FlowFISH, we observed the expected age-associated accumulation of MSY$^-$ cells in up to 1.9% in our oldest donor (Fig. 5c and Extended Data Fig. 5a). Downstream analyses of PERFF-seq libraries profiling 35,283 cells confirmed a ~2–3× increase in cells with no Y chromosome UMIs and an overall shift in the cumulative distribution across the three aged donors (Extended Data Fig. 5b and Methods). Using Azimuth[13], we recovered the expected major PBMC types, including anticipated skews between libraries from the *MSY* sort (Fig. 5d). Cell-type annotations confirmed anticipated skews[24] between MSY$^-$ and MSY$^+$ PERFF-seq libraries, including MSY$^+$ enrichment of monocytes, CD4$^+$ naive T cells and T_{reg} cells (Fig. 5d,e). Differential gene expression and downstream gene set enrichment analyses for CD14 monocytes revealed enrichment of four hallmark gene sets, including tumor necrosis factor (TNF) signaling by nuclear factor κ-light-chain enhancer of activated B cells (NF-κB), suggesting that these monocytes may be linked to inflammatory dysregulation implicated in aged individuals with LOY[25] (Fig. 5f and Methods). In sum, PERFF-seq can infer somatic changes as long as a sufficient fluorescent signal can be generated via FlowFISH. Although *MSY* genes are lowly expressed, our analyses show PERFF-seq can be applied to a range of gene expression levels with corresponding degrees of enrichment.

Enrichment and heterogeneity of nuclei via PERFF-seq

Given the lack of high-quality antibodies against nuclear protein markers, we asked whether PERFF-seq is compatible with profiling nuclei from archived material. We isolated nuclei from fresh-frozen adult mouse brain cerebellum and FFPE-preserved human glioblastoma multiforme (GBM) tissues (Fig. 6a). Oligodendrocytes in the

https://doi.org/10.1038/s41588-024-02036-7

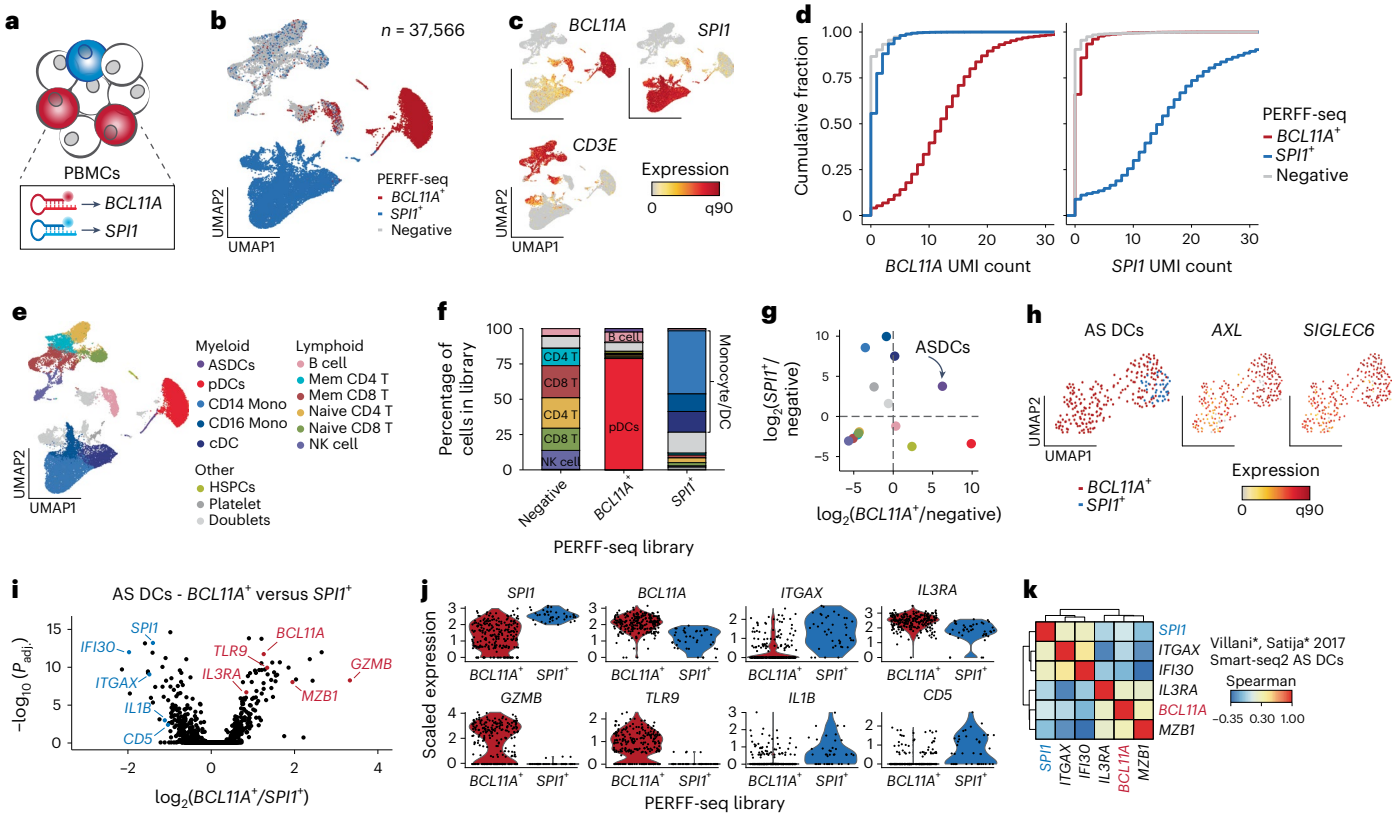


Fig. 4 | Rare cell states enriched via nontraditional cell-type markers. **a**, Schematic of human PBMC staining with probes targeting *BCL11A* and *SPI1*. **b**, Uniform Manifold Approximation and Projection (UMAP) embedding of PERFF-seq profiles from three populations based on TF FlowFISH sorting logic. **c**, Depiction of marker-gene expression across all PERFF-seq profiled cells. **d**, Empirical cumulative distribution plot of raw UMI counts for *BCL11A* (left) and *SPI1* (right) stratified by the captured PERFF-seq library. **e**, Annotated cell states from PERFF-seq profiling. **f**, Proportions of each cell type per library with major cell types labeled. The colors match **e**. **g**, Relative enrichment of each cell type in the *BCL11A*$^+$ sort (*x* axis) or *SPI1*$^+$ sort (*y* axis) relative to the negative population. AS DCs are highlighted as the only enriched population in both sorted populations. The colors match **e**. **h**, UMAP of AS DCs, highlighting the TF FlowFISH library and defining marker-gene expression. **i**, Volcano plot comparing DEGs from the two FlowFISH-sorted populations. Notable marker genes are highlighted, including known and newly identified marker genes for AS DC subsets. The Bonferroni-adjusted *P* value for the two-sided Wilcoxon's rank-sum test is shown. **j**, Violin plots of marker genes, stratified by the FlowFISH library. All genes were significantly differentially expressed at a false discovery rate (FDR) < 0.01. **k**, Gene–gene Spearman's rank correlations of all AS DCs using the original dataset, highlighting the co-occurrence of TFs from our analysis with established marker genes for the AS DC subsets.

cerebellum are sparse (~2% of total nuclei[26]; Extended Data Fig. 6a) and are implicated in the cell-state-specific progression of multiple neurodegenerative disorders[27–29], motivating us to profile them with the population-defining gene, *Mobp*. For the human GBM samples ($n = 2$ donors), we selected three vasculature-associated genes, *DCN*[30], *FN1* (ref. 31) and *VWF*[32], all implicated as biomarkers in glioblastoma pathogenesis or treatment response. Reanalysis of existing GBM FFPE data suggested that these three genes are principally expressed in a rare population (~4.2%) of vascular-derived cells in primary GBM tumors (Extended Data Fig. 6b,c). Similar to our immune cell benchmarking (Fig. 2), downsampling analyses demonstrate that our upstream HCR–FlowFISH workflow had minimal impact on total UMI recovery (Fig. 6b,c, Methods and Supplementary Table 3).

Sorting and profiling the ~4.1% of *Mobp*-expressing cells generated an almost 50-fold enriched population (Fig. 6d–f). Co-embedding of the *Mobp*$^+$ and *Mobp*$^-$ samples and annotation of marker genes confirmed the expected cerebellum cell types, including granule cells (*Gabra6* and *Rbfox3*), interneurons (*Gad1*) and Bergmann glia (*Itih3*; Extended Data Fig. 6d). Subclustering *Mobp*$^+$ oligodendrocytes recovered four major populations with distinct marker profiles (Supplementary Table 5). Cluster 0 expressed markers of terminally differentiated oligodendrocytes (*Il33* and *Ptgds*)[33,34], whereas cluster 1 was characterized by *Klk6* and *S100B*, markers of maturing oligodendrocyte precursors[35–37] (Fig. 6g and Extended Data Fig. 6e). Clusters 2 and 3 were rarer populations ($n = 429$ and $n = 212$) defined by a mix of mature oligodendrocyte, oligodendrocyte precursor and neuronal synapse markers with *Atp1b1*, *Scnb* and *Snap25* marking cluster 2 and *Agt* and *Aqp4* (ref. 26) marking cluster 3, probably a low-frequency *Mobp*$^+$ astrocyte population (Fig. 6g and Extended Data Fig. 6e). Although rare (approximately 0.05–0.10% of all cerebellum nuclei), the two subclusters probably represent distinct cell states—cluster 2 representing differentiating oligodendroglia that prune synapses[38] and cluster 3 representing *Mobp*$^+$ astrocytes in cortical populations[39], but to our knowledge these have not been identified in subcortical structures.

The PERFF-seq profiles from GBM enrichment exhibited a distinct cell cluster enriched primarily for cells from our vasculature OR-gated panel that express relevant marker genes (Fig. 6h,i). Not only did total UMI distributions in the panel clearly separate positive and negative populations (Fig. 6j and Extended Data Fig. 6f), but we also detected an ~23× increase in signal overall (mean panel$^+$: 33.0 UMIs; mean panel$^-$: 1.4 UMIs) where most background expression was driven by promiscuous expression of *FN1*, as expected (Fig. 6i and Extended Data Fig. 6c). Differential expression analyses confirmed that genes enriched in our panel were among the top effects by fold-change with other collagen-associated genes similarly enriching within our vascular-enriched cells (Fig. 6k). Subclustering of the 1,015 vascular

https://doi.org/10.1038/s41588-024-02036-7

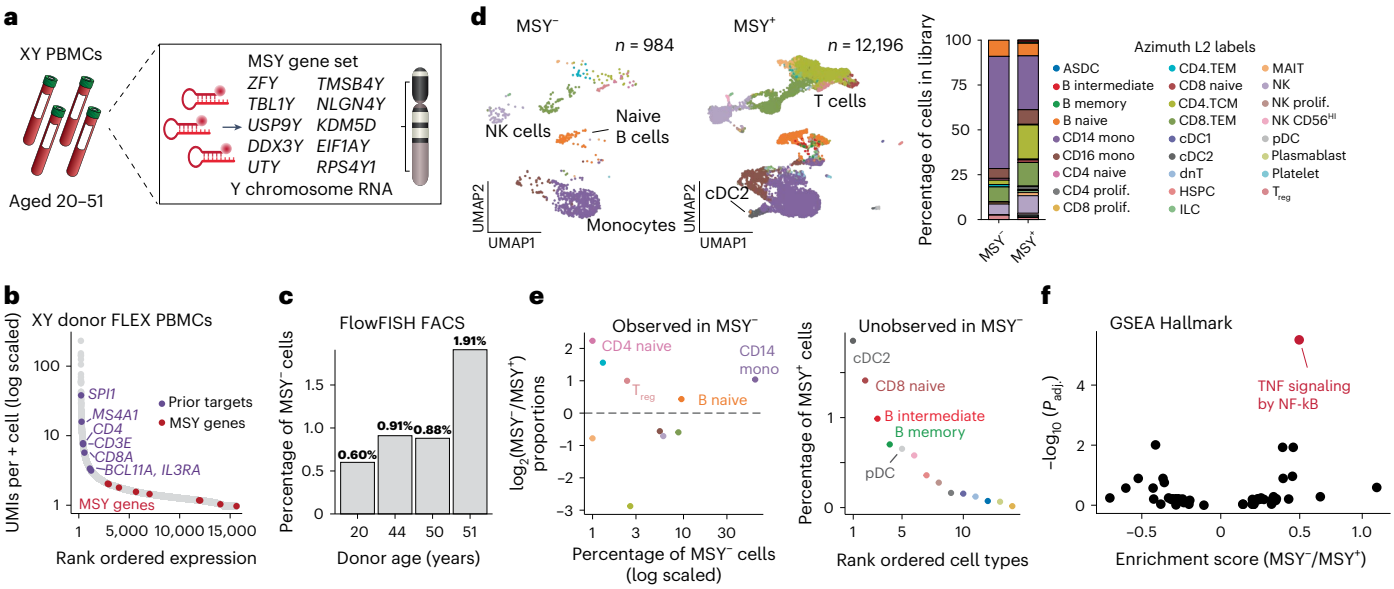


Fig. 5 | Profiling somatic mosaicism with PERFF-seq. a, Schematic of experiment. PBMCs from donors of different ages were sorted for a ten-gene OR-gated panel of MSY. **b**, Mean per-cell expression of all genes detected in Flex with genes analyzed for FlowFISH noted. **c**, Summary of the percentage of MSY⁻ cells, with donor age labels, from the FlowFISH cytometry data. **d**, UMAP embedding of PERFF-seq profiles from the 51-year-old donor based on MSY sorting logic. **e**, Analyses of cell types from scRNA-seq analyses for cell types enriched (left) or depleted (right) in the MSY⁻ library. The colors represent cell types as shown in **c**. **f**, Gene set enrichment analyses of MSY⁻ versus MSY⁺ CD14 monocytes, highlighting TNF signaling by NF-κB. Statistical significance is based on a permutated enrichment score under a two-sided null. prolif., proliferative.

cells from our positive sort confirmed the two major populations, including *VWF*⁺ endothelial cells and *FN1*⁺/*DNC*⁺ mural cells (Fig. 6l,m and Extended Data Fig. 6g). Although our identification of these major populations within the vasculature corroborates prior scRNA-seq analyses[40], our high-quality PERFF-seq profiles allowed for further identification of eight subclusters within the enriched population (Fig. 6m,n, Extended Data Fig. 6g and Supplementary Table 6). These included endothelial cells marked by *PVLAP*⁺, a vascular marker of blood–brain barrier disruption[40], and proliferating endothelial cells marked by *TOP2A*/*MKI67* that have been recently described in fetal and pathological brains[41]. Finally, we identified an ultra-rare population of mural cells expressing *OGN* potentially linked to brain tumorigenesis[42]. In sum, PERFF-seq unlocks enrichment upstream of single-nuclei sequencing, which should expedite and corroborate nuclei profiling from diverse tissue types and heterogeneous input materials.

Discussion

PERFF-seq is a robust approach that can be used to enrich a variety of transcriptomic markers across varying input material, including major immune cell populations (Fig. 3), rare cell types (Figs. 4 and 5) and nuclei derived from fresh-frozen and FFPE material (Fig. 6). As solid tissue often requires or benefits from the isolation of nuclei rather than cells for profiling, the lack of high-quality intranuclear protein targets that meaningfully differentiate cell populations of interest poses a formidable challenge. Furthermore, continued efforts to build cell atlases recover increasingly complex cell types defined by specific marker combinations, including in the mouse cerebellum[26]; thus, we envision a primary use of PERFF-seq to be enriching for populations without the need for laborious genetic engineering, including via Cre-recombinases.

Although prior methods such as Probe-seq[43] have coupled nucleic acid detection and sorting to bulk RNA-seq, PERFF-seq is distinct in its ability to profile single-cell populations at high throughput (~10^4–10^5 cells per capture). Another approach, FIND-seq[44], sorts emulsion droplets to establish nucleic acid cytometry that similarly enriches rare cells based on specific transcripts. However, FIND-seq yields low-throughput (~10^1–10^3 cells per capture) bulk profiles and requires specialized instrumentation. Thus, as a result of its combination of throughput and the commercial availability of reagents, we anticipate straightforward adoption of PERFF-seq by groups proficient in either RNA FISH or scRNA-seq. Although we showed experimental feasibility with up to three fluorophores in the same experiment (Fig. 3), we noted that the HCR–FlowFISH kits contain more colors, allowing for further AND/NOT/OR logic gating of populations before scRNA-seq profiling.

Although our dsDNase stripping effectively removes the HCR polymer, FISH probes probably remain bound to the target gene of interest. In our benchmarking with Fig. 2, we observed a ~24% reduction in cells positive for *CD3E* and a twofold $\log_2$(reduction) in gene expression (Extended Data Fig. 2d,e). Other targets such as *BCL11A*, *SPI1* and *Mobp* showed robust expression because >90% of enriched cells were positive for the gene targeted by FISH. Thus, depending on the target genes of interest, scrutiny is warranted when analyzing the scRNA-seq counts of the targeted genes. In addition, our applications focus on FISH hybridization to RNAs of interest to sort populations based on transcript abundance. In the context of LOY, we successfully showed the potential of PERFF-seq to segregate somatic mosaicism (Fig. 5). Future efforts may also enable the capture of other somatic events, such as single-nucleotide variants or gene fusions, which will require refined FlowFISH strategies.

A primary benefit of coupling HCR–FlowFISH sorting to scRNA-seq is that it enables distinct strategies for isolating populations. Rather than counting events on a cytometer, an inclusive PERFF-seq sorting strategy can be used that allows populations to be subsequently refined using transcriptomic profiles, which may overcome requirements of low cell availability. For example, our application to sorting *BCL11A*⁺ cells allowed for detailed reanalyses of both AS DCs and B cells in the same experiment without pre-specifying the populations with distinct markers during sorting, because these populations could be readily separated by scRNA-seq analysis (Fig. 4). Although many surface markers that we evaluated such as *CD3E* and *MS4A1* (CD20) were

https://doi.org/10.1038/s41588-024-02036-7

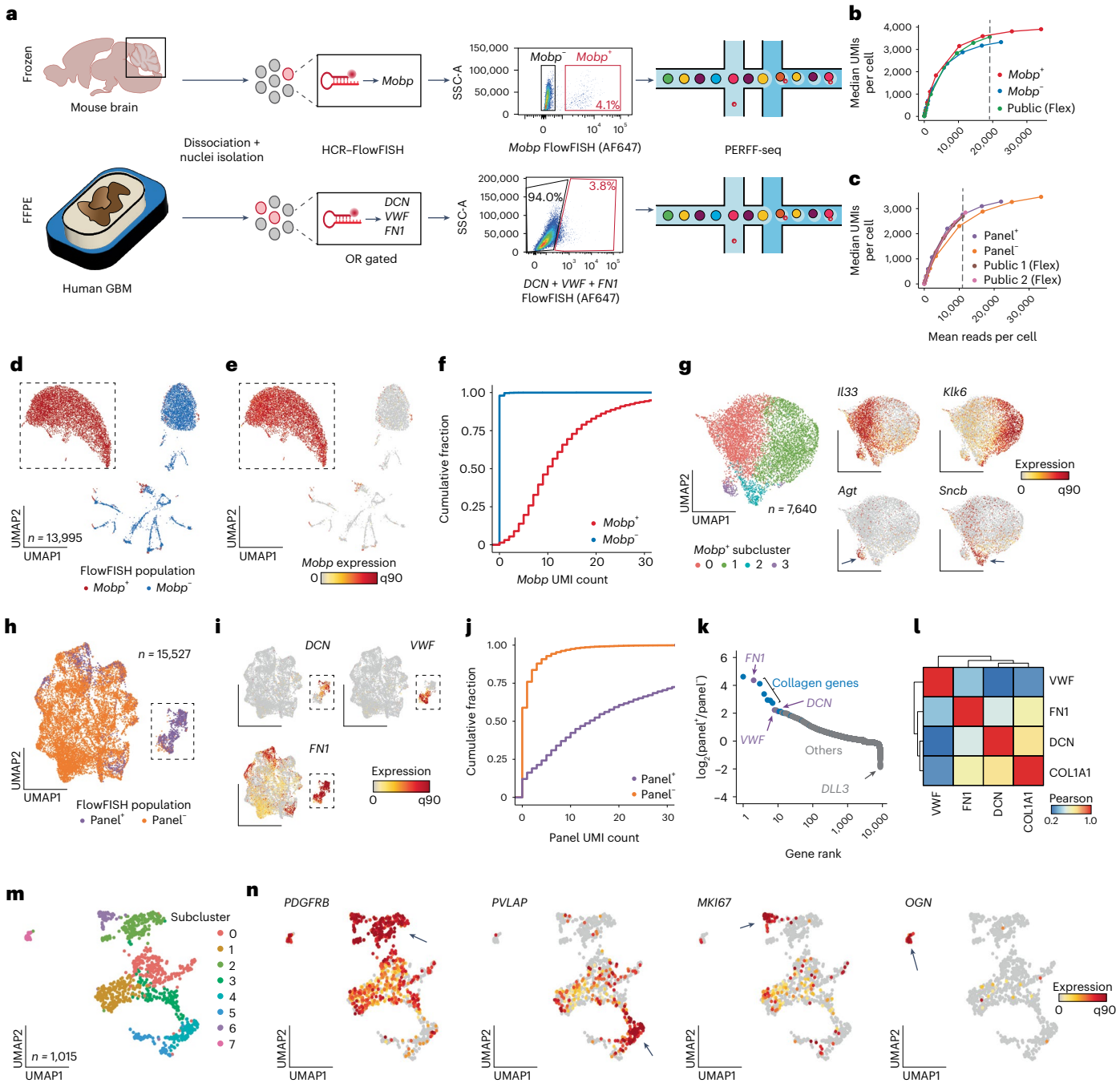


Fig. 6 | Study of rare nuclei from fresh-frozen and FFPE tissue. a, Schematic of PERFF-seq single-nucleus experiments from frozen mouse brain tissue or FFPE human GBM tissue, showing HCR–FlowFISH staining and sorting strategy. Side scatter area (SSC-A) and FISH signal separate populations. **b**, Downsampling analysis for library saturation and UMI benchmarking for the mouse brain nuclei. The dashed line represents the mean reads per cell for a final comparison (depth of lowest sample: top, 19,140 reads per cell; bottom, 11,021 reads per cell). **c**, Same as **b** but for the human FFPE tissue sample. **d**, UMAP embedding of the mouse brain nuclei, FlowFISH-enriched or -depleted populations profiled with PERFF-seq. **e**, Same as **d** but colored by *Mobp* marker-gene expression. The boxed population was further subclustered. **f**, Empirical cumulative distribution plot of raw UMI count for *Mobp* stratified by the captured PERFF-seq library. **g**, Subclustering of the *Mobp*$^+$ population. The arrows highlight top marker genes per cluster. **h**, Reduced dimensionality representation of the human FFPE nuclei FlowFISH-enriched or -depleted populations profiled with PERFF-seq. **i**, Same as **h** but colored by marker genes used in the FlowFISH panel. The boxed population was further subclustered. **j**, Empirical cumulative distribution plot of total UMI count for the sum of the three genes enriched via FlowFISH, stratified by the captured PERFF-seq library. **k**, Top DEGs between the two FFPE populations profiled with PERFF-seq. **l**, Gene–gene Pearson's correlations of relevant marker genes, including those used in the FlowFISH enrichment panel. **m**, Subclustering of the panel$^+$ population with cluster states noted. **n**, Top marker genes enriched in specific subclusters. The arrows indicate critical populations where each gene is highly expressed.

consistent between surface protein and mRNA expression, others such as *CD4* and *IL3RA* (CD123) exhibited gene expression in lineages lacking surface protein expression. Similarly, although *BCL11A* is required for both B cell[19] and pDC[45] development, the higher mRNA expression of this TF in pDCs resulted in a substantial enrichment of this cell type and minimal enrichment of B cells in our PERFF-seq library, which we confirmed with dual staining of RNA FISH and surface-staining reagents. Collectively, these vignettes motivate a careful data-driven

exploration of appropriate marker genes using expression data rather than conventional knowledge derived from established FACS markers. Fortunately, such data-driven explorations are straightforward from high-quality scRNA-seq profiles across many tissues, systems and pathologies.

Ultimately, we anticipate that PERFF-seq will be particularly advantageous in settings where antibodies against marker proteins are unavailable, not applicable or poorly defined for a population of interest. As FlowFISH technologies have been established for viral gene expression[46], ribosomal RNA content[47] or lncRNAs[48], our workflow provides unique access to the understudied populations defined by these markers. We envision an iterative process where populations defined within large-scale, single-cell atlases are rationally enriched to study heterogeneity in rare and newly discovered populations to refine our understanding of cellular diversity with ever greater definition.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-02036-7.

References

1. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
2. Drokhlyansky, E. et al. The human and mouse enteric nervous system at single-cell resolution. *Cell* **182**, 1606–1622.e23 (2020).
3. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
4. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
5. Lareau, C. A. et al. Latent human herpesvirus 6 is reactivated in CAR T cells. *Nature* **623**, 608–615 (2023).
6. Chung, H. et al. Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nat. Methods* **18**, 1204–1212 (2021).
7. Chen, A. & Koehler, A. N. Transcription factor inhibition: lessons learned and emerging targets. *Trends Mol. Med.* **26**, 508–518 (2020).
8. Evers, D. L., Fowler, C. B., Cunningham, B. R., Mason, J. T. & O'Leary, T. J. The effect of formaldehyde fixation on RNA: optimization of formaldehyde adduct removal. *J. Mol. Diagn.* **13**, 282–288 (2011).
9. Janesick, A. et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat. Commun.* **14**, 8353 (2023).
10. Wang, Y. et al. EASI-FISH for thick tissue defines lateral hypothalamus spatio-molecular organization. *Cell* **184**, 6361–6377.e24 (2021).
11. Choi, H. M. T. et al. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, dev165753 (2018).
12. Choi, H. M. T., Beck, V. A. & Pierce, N. A. Next-generation in situ hybridization chain reaction: higher gain, lower cost, greater durability. *ACS Nano.* **8**, 4284–4294 (2014).
13. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
14. Reilly, S. K. et al. Direct characterization of *cis*-regulatory elements and functional dissection of complex genetic associations using HCR-FlowFISH. *Nat. Genet.* **53**, 1166–1176 (2021).
15. Alon, S. et al. Expansion sequencing: spatially precise in situ transcriptomics in intact biological systems. *Science* **371**, eaax2656 (2021).
16. Marshall, J. L. et al. HyPR-seq: single-cell quantification of chosen RNAs via hybridization and sequencing of DNA probes. *Proc. Natl Acad. Sci. USA* **117**, 33404–33413 (2020).
17. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
18. Cano-Gamez, E. et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of $CD4^+$ T cells to cytokines. *Nat. Commun.* **11**, 1801 (2020).
19. Yu, Y. et al. Bcl11a is essential for lymphoid development and negatively regulates p53. *J. Exp. Med.* **209**, 2467–2483 (2012).
20. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–644 (2008).
21. Choi, J. et al. Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans. *Nucleic Acids Res.* **47**, D780–D785 (2018).
22. Guo, X. et al. Mosaic loss of human Y chromosome: what, how and why. *Hum. Genet.* **139**, 421–446 (2020).
23. Bruhn-Olszewska, B. et al. Loss of Y in leukocytes as a risk factor for critical COVID-19 in men. *Genome Med.* **14**, 139 (2022).
24. Mattisson, J. et al. Loss of chromosome Y in regulatory T cells. *BMC Genom.* **25**, 243 (2024).
25. Lleo, A. et al. Y chromosome loss in male patients with primary biliary cirrhosis. *J. Autoimmun.* **41**, 87–91 (2013).
26. Kozareva, V. et al. A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. *Nature* **598**, 214–219 (2021).
27. Siokas, V. et al. Myelin-associated oligodendrocyte basic protein rs616147 polymorphism as a risk factor for Parkinson's disease. *Acta Neurol. Scand.* **145**, 223–228 (2022).
28. Irwin, D. J. et al. Myelin oligodendrocyte basic protein and prognosis in behavioral-variant frontotemporal dementia. *Neurology* **83**, 502–509 (2014).
29. Kon, T. et al. Immunoreactivity of myelin-associated oligodendrocytic basic protein in Lewy bodies. *Neuropathology* **39**, 279–285 (2019).
30. Patel, K. S. et al. Decorin expression is associated with predictive diffusion MR phenotypes of anti-VEGF efficacy in glioblastoma. *Sci. Rep.* **10**, 14819 (2020).
31. Serres, E. et al. Fibronectin expression in glioblastomas promotes cell cohesion, collective invasion of basement membrane in vitro and orthotopic tumor growth in mice. *Oncogene* **33**, 3451–3462 (2014).
32. Mojiri, A. et al. Functional assessment of von Willebrand factor expression by cancer cells of non-endothelial origin. *Oncotarget* **8**, 13015–13029 (2017).
33. Sung, H.-Y. et al. Down-regulation of interleukin-33 expression in oligodendrocyte precursor cells impairs oligodendrocyte lineage progression. *J. Neurochem.* **150**, 691–708 (2019).
34. Huang, H.-T. & Tzeng, S.-F. Interleukin-33 has the protective effect on oligodendrocytes against impairment induced by cuprizone intoxication. *Neurochem. Int.* **172**, 105645 (2024).
35. Floriddia, E. M. et al. Distinct oligodendrocyte populations have spatial preference and different responses to spinal cord injury. *Nat. Commun.* **11**, 5860 (2020).
36. Langlieb, J. et al. The molecular cytoarchitecture of the adult mouse brain. *Nature* **624**, 333–342 (2023).
37. Du, J. et al. S100B is selectively expressed by gray matter protoplasmic astrocytes and myelinating oligodendrocytes in the developing CNS. *Mol. Brain* **14**, 154 (2021).
38. Auguste, Y. S. S. et al. Oligodendrocyte precursor cells engulf synapses during circuit remodeling in mice. *Nat. Neurosci.* **25**, 1273–1278 (2022).

39. Morel, L. et al. Intracortical astrocyte subpopulations defined by astrocyte reporter mice in the adult brain. *Glia* **67**, 171–181 (2019).
40. Xie, Y. et al. Key molecular alterations in endothelial cells in human glioblastoma uncovered through single-cell RNA sequencing. *JCI insight* **6**, e150861 (2021).
41. Wälchli, T. et al. Single-cell atlas of the human brain vasculature across development, adulthood and disease. *Nature* **632**, 603–613 (2024).
42. Mei, Y. et al. Osteoglycin promotes meningioma development through downregulation of NF2 and activation of mTOR signaling. *Cell Commun. Signal.* **15**, 34 (2017).
43. Amamoto, R. et al. Probe-Seq enables transcriptional profiling of specific cell types from heterogeneous tissue by RNA-based isolation. *eLife* **8**, e51452 (2019).
44. Clark, I. C. et al. Identification of astrocyte regulators by nucleic acid cytometry. *Nature* **614**, 326–333 (2023).
45. Ippolito, G. C. et al. Dendritic cell fate is determined by BCL11A. *Proc. Natl Acad. Sci. USA* **111**, E998–E1006 (2014).
46. Warren, C. J. et al. Quantification of virus-infected cells using RNA FISH-Flow. *STAR Protoc.* **4**, 102291 (2023).
47. Antony, C., Somers, P., Gray, E. M., Pimkin, M. & Paralkar, V. R. FISH-Flow to quantify nascent and mature ribosomal RNA in mouse and human cells. *STAR Protoc.* **4**, 102463 (2023).
48. González-Vasconcellos, I., Cobos-Fernández, M. A., Atkinson, M. J., Fernandez-Piqueras, J. & Santos, J. Quantifying telomeric lncRNAs using PNA-labelled RNA-Flow FISH (RNA-Flow). *Commun. Biol.* **5**, 513 (2022).

Methods

Our research complies with all relevant ethical and regulatory guidance, including the institutional review boards at Stanford University and Memorial Sloan Kettering Cancer Center. Our project was exempt from human participant research.

PERFF-seq method and development

A full protocol for executing the combined HCR–FlowFISH and 10x Genomics Flex profiling steps is available on protocols.io (https://doi.org/10.17504/protocols.io.14egn3k6ql5d/v1 and https://doi.org/10.17504/protocols.io.8epv5x8r4g1b/v1) and in the following sections. In brief, based on our optimizations in Fig. 1, we emphasize a few critical aspects of the method development. First, enriching for populations via HCR–FlowFISH proceeded with minimal modifications to the protocol except for RNase inhibitor BSA buffer to preserve RNA quality. After the enriched populations had been isolated, a dsDNase step was used to degrade the HCR polymer, which we found to be critical for 10x Genomics library preparation. After dsDNase digestion, including the 10x Flex, whole transcriptome amplicon (WTA) probes proceeded as is standard before droplet encapsulation on the 10x Genomics Chromium platform. Downstream amplification and sequencing follow the standard Flex guidelines with no modifications. Together, PERFF-seq leveraged the quality-controlled aspects of both workflows with minimal modifications. However, we emphasize that these modifications (dsDNase and RNase-free BSA) can severely limit data quality if not implemented. Complete details for each step are available from the manufacturer and the protocols.io link accompanying this manuscript.

Quality control metrics and optimization

From both Bioanalyzer traces and CellRanger quality control summaries, we observed that the presence of half-ligated probes was a correlate for overall library quality (Fig. 1 and Extended Data Fig. 1). Our interpretation of the half-mapped probes is that conditions for Flex were incompatible with the ligation required for the pairs of gene probes for RNA detection (Extended Data Fig. 1). As a consequence, many left probes became barcoded, resulting in a product that was ~70 bases smaller than the expected Flex barcode product (Extended Data Fig. 1b). To confirm this, we examined one million reads from either of our initial libraries for the presence of the probe capture sequence ('CGGTCCTAGCAA') in the read position where the left probe sequence is contained on R2 (between positions 1 and 25), which resulted in 21.1% of reads containing a perfect match to probe capture sequence (Extended Data Fig. 1c). From our remaining optimization experiments, our interpretation of these data was that the presence of the HCR polymer (or formamide) is sufficient to disrupt the essential probe ligation for high-quality Flex data. In this sense, removing the polymer after sorting via dsDNase removes the polymer that otherwise inhibits ligation.

Immune cell experiments

Protocol development and optimization were performed on cryopreserved human PBMCs sourced from American Type Culture Collection (ATCC) and AllCells. Vials were thawed and viability exceeded 90% for all samples. PBMCs were used as the primary input for developing the assay in Fig. 1 owing to the ease of material availability and well-defined heterogeneity for *MS4A1* and *CD3E*. For applications in Figs. 2–4, the same vials were used but enriched for specific markers as indicated in the experimental overview schematics (Figs. 2c, 3a, 4a and 5a). All experiments started with ~10,000,000 cells, except for the TF sort experiment (Fig. 4), which began with ~25,000,000 cells to yield ample cell numbers for downstream profiling given the rare *BCL11A* population that was sorted.

For the mosaic loss-of-Y experiment, 10 million PBMCs were thawed from donors of varying ages (20, 44, 50 and 51 years), including the youngest donor (20) serving as a control. To account for background fluorescence, we included another negative control comprising cells with mixed populations from the original three donors, labeled with no probe but with the hairpin amplifier. Using this experimental setup, we implemented a gating strategy where the sorted positive population was delineated from the negative population (Extended Data Fig. 5a). The negative populations were then subdivided into two distinct gates: the upper negative gate which exhibited higher levels of autofluorescence and appeared to comprise a separate population (which were excluded), whereas the lower negative gate was characterized by both low fluorescence and autofluorescence levels. Six libraries were profiled across the three donors, both MSY^+ and MSY^-.

Human cell lines

Human cell lines K-562, Raji, MCF-7 and Jurkat lines were procured from ATCC and monitored for *Mycoplasma* contamination. Genotyping at ATCC confirmed the identity of all lines. Analyses of RNA expression in Fig. 2 matched known RNA-seq profiles for these lines.

Cell fixation and permeabilization

Fixation and permeabilization were performed as described in 'HCR RNA flow cytometry protocol for mammalian cells in suspension' provided by Molecular Instruments. Briefly, cells were thawed and fixed in 4% paraformaldehyde (PFA) solution (4% PFA in 1× PBS and 0.1% Tween 20 (PBST)) at room temperature for 1 h at a concentration of 1 million cells per milliliter (1 M ml^{-1}). After fixation, cells were centrifuged at 350*g* for 5 min and resuspended in a PBST solution (1× PBS and 0.1% Tween 20) at 1 M ml^{-1}. This step was repeated once for a total of two washes. After washing, cells were permeabilized with ice-cold 70% EtOH overnight at 1 M ml^{-1}. After permeabilization, cells were centrifuged and resuspended with PBST solution at 1 M ml^{-1} twice.

HCR–FlowFISH

Probes for RNA targets of interest and complementary hairpins were purchased from Molecular Instruments at the highest number of probe pairs available for desired genes.

Most steps were performed as described in the 'HCR RNA flow cytometry protocol for mammalian cells in suspension' provided by Molecular Instruments and Reilly et al.[14] with the following adjustments:

First, all centrifugation was performed at 850*g* for 5 min unless otherwise noted. Second, low-binding plastic-ware tubes and RNase-free, molecular-grade reagents were utilized when possible. Third, during the detection stage, we found that an optimal signal:noise ratio was achieved during fluorescence detection at 16 nM probe concentration per 500,000–1,000,000 cells (8 µl of probe stock per sample). Finally, 37 °C incubations were performed in a heated lid thermomixer with gentle shaking.

HCR–FlowFISH detection stage. Permeabilized cells or nuclei were resuspended in prewarmed 400 µl of hybridization buffer (Molecular Instruments) per 500,000–1,000,000 cells. Cells or nuclei were incubated for 30 min at 37 °C, 300 rpm in a heated lid thermomixer. The probe solution was prepared by mixing 8 µl of 1 µM probe stock and hybridization buffer for a final 100 µl volume per sample. Probe solution was added to each sample for a final probe concentration of 16 nM and cells were incubated at 37 °C for 16–24 h. To pellet cells or nuclei, 500 µl of SSCT solution (5× saline–sodium citrate (SSC) + 0.1% Tween 20) was added to each sample and centrifuged at 850*g* for 15 min. Cells or nuclei were resuspended in 500 µl of prewarmed probe wash buffer (Molecular Instruments), incubated for 10 min at 37 °C and subsequently pelleted at 850*g* for 5 min. This step was repeated three more times for a total of four washes. Then, cells or nuclei were resuspended in 500 µl of SSCT solution and incubated at room temperature for 5 min.

HCR–FlowFISH amplification stage. Cells or nuclei were centrifuged and resuspended in 150 µl of amplification buffer and incubated at room temperature for 30 min. In the meantime, 5 µl of 3 µM h1 and h2

hairpin stock was aliquoted for each probe set and snap cooled by performing a heat shock at 950 °C for 90 s and cooling in the dark at room temperature for 30 min. To prepare the hairpin solution, snap-cooled h1 and h2 hairpins were mixed with an amplification buffer to make a final volume of 100 µl per sample. The hairpin solution was added to appropriate samples for the final hairpin concentration of 60 nM. Cells or nuclei were incubated at room temperature for 16–24 h (this time can be reduced to 4 h). After incubating, samples were washed 6× with 500 µl of SSCT for each sample.

Co-staining with antibodies. PBMCs were thawed and stained with anti-CD123 (BioLegend, cat. no. S18016F), anti-CD19 (BioLegend, HIB19), anti-CD14 (BioLegend. M5E2) and/or anti-CD3 (BioLegend, OKT3). For antibody staining, we followed the manufacturer's protocol and stained with 5 µl of antibody per 1 million cells in 100 µl of staining buffer (1% BSA in 1× PBS). This was immediately followed by fixation or permeabilization and HCR–FlowFISH as described above. We note that surface antibodies conjugated to synthetic dyes result in the most robust signal when used together with the HCR–FlowFISH protocol.

Sample enrichment using FACS

Samples were resuspended in sorting and collection buffer (1× PBS, 5% BSA, Gibco, cat. no. 15260037; 0.13 U µl^{-1} of RNAse inhibitor, Millipore Sigma, cat. no. 3335399001) and filtered through a 35-µm strainer. Cells were kept in the dark and on ice until sorting. Collection tubes were prepared with 300 µl of collection buffer. Cells were sorted using the BD FACSAria III or FACSymphony S6 with a 70-µm nozzle for cells and an 85-µm nozzle for nuclei. Representative gating is shown where appropriate. For multiplexing experiments, compensation was performed with single-color controls.

HCR polymer disassembly

Sorted cells were pelleted and resuspended in 275 µl of 1× dsDNase buffer (Thermo Fisher Scientific, cat. no. EN0771) and incubated for 15 min, after which 25 µl of dsDNase enzyme (Thermo Fisher Scientific, cat. no. EN0771) was added and the sample was incubated at 37 °C for 2 h. After incubation, 3 µl of 1 M dithiothreitol was added to the sample to quench dsDNAse activity and incubated at 55 °C for 5 min for heat inactivation. Samples were pelleted at 850*g* for 5 min, resuspended in 500 µl of prewarmed wash buffer and incubated for 10 min. This step was repeated once for a total of two washes. Samples were then resuspended in 500 µl of SSCT buffer and incubated for 5 min. Samples were centrifuged and resuspended in 1 ml of 0.5× PBS and 0.02% BSA (Thermo Fisher Scientific, cat. no. AM2616) and 0.2 U µl^{-1} of RNase inhibitor. Polymer disassembly was assessed by measurement of fluorescence intensity on the BD FACSAria III (Extended Data Fig. 1d,e). We noted that it is neither typical nor advised to verify fluorescence stripping for PERFF-seq. This was performed only for benchmarking purposes (Extended Data Fig. 1d,e), but was otherwise not examined in other experiments because it is not necessary.

10x Genomics Flex

Preparation of all 10x Genomics Flex libraries were prepared using the manufacturer's instructions because all modifications for PERFF-seq happen upstream of Flex probe hybridization. All libraries were sequenced on an Illumina Nextseq 550, Novaseq 6000, Novaseq X or an Element AVITI with standard dual indexing and demultiplexing. Raw .bcl files were processed using CellRanger v.7.2, and the resulting .fastq files were quantified for the human and mouse probe set to v.1.0.1 using default parameters for the CellRanger pipeline for the hg38 (human) and GRCm39 (mouse).

Cell-line mixing and benchmarking

The XY Burkitt's lymphoma cell line (Raji cells) and the XX chronic myelogenous leukemia cell line K-562 cells were obtained from ATCC. Raji cells were cultured in Roswell Park Memorial Institute (RPMI)-1640 medium, whereas K-562 cells were cultured in Iscove's modified Dulbecco's medium, both supplemented with 10% FBS and 1% penicillin–streptomycin. Both lines were cultured at 18% O_2 and 37 °C. To benchmark the recovery of populations, cells were washed with PBS, centrifuged, and counted. Raji cells were subsequently mixed with decreasing tenfold dilutions of K-562 cells. The mixed cells were then fixed and permeabilized as described in the earlier sections, followed by the HCR–FlowFISH protocol using *XIST* RNA probes for detection and Alexa Fluor-647-conjugated hairpins for amplification. FACS analyses were conducted on a Thermo Fisher Scientific Attune NxT Flow Cytometer.

EpCAM benchmarking

Four cell lines with variable levels of EpCAM (epithelial cell adhesion molecule) were procured from ATCC and *EpCAM* RNA levels were assessed using RNA-seq data from the *Cancer Cell Line Encyclopedia*[49]. The unstained and hairpin-only controls represented an equal mix of all four cell lines to establish a quantification of the background signal. MFI was computed after gating for single cells.

PBMC benchmarking

To assess the impact of the modified features of the PERFF-seq protocol on Flex, we assessed four conditions profiled from the same donor PBMCs (Fig. 2c–f). The Flex condition proceeded on all PBMCs using the manufacturer's recommendations. For the NPNS conditions, three changes are noted: (1) instead of a probe-containing solution used in PERFF-seq, we used PBS, which was followed by multiple washes using a formamide-containing buffer; these changes are consistent with the FlowFISH protocol but absent from Flex; (2) we emulated the HCR–FlowFISH amplification by adding fluorescent hairpins and amplification buffer, but, as there were no gene probes, there were no HCR polymer forms; and (3) we treated with dsDNase as was performed in PERFF-seq to disassemble the HCR–FISH polymer. For the YPNS condition, the HCR–FlowFISH workflow was conducted without sorting for the *CD3E* population. Finally, for the PERFF-seq condition, the full workflow enriching for *CD3E*+ cells was performed.

Mouse tissue sourcing

Male C57BL/6 mouse brain tissue was sourced from Zyagen Inc. as a fresh-frozen whole brain stored in optimal cutting temperature (OCT) compound. On receipt, tissue was stored at −80 °C. Dissection was performed in a cryotome and immediately prepared for nuclei processing.

Mouse brain dissociation and profiling

Mouse brain dissociation was performed as described by 10x Genomics in the tissue fixation and dissociation for chromium fixed RNA profiling. Briefly, fresh-frozen mouse cerebellum was weighed and fixed in 4% PFA solution for 2 h at 2 ml per 25 mg of tissue with periodic agitation. Then, the tissue was centrifuged and resuspended in 1× PBS twice. Washed tissue was resuspended in ice-cold 70% ethanol at 2 ml per 25 mg of tissue and incubated overnight at 4 °C. After incubation, the tissue was centrifuged and resuspended with 1× PBS twice. Tissue was then resuspended in 2 ml of dissociation buffer (160 µl of LiberaseTL enzyme, Millipore Sigma, cat. no. 5401020001 + RPMI) using the gentleMACS OctoDissociator with heaters (Miltenyi Biotec, cat. no. 30-096-427) for 30 min at 50 rpm. Nuclei were washed with 1× PBS + 0.02% BSA (Thermo Fisher Scientific, cat. no. AM2616) + 0.2 U µl of RNase inhibitor and stained with 1 µl ml^{-1} of DAPI (Thermo Fisher Scientific, cat. no. 62248) for 10 min. To remove excess debris, DAPI^{+} nuclei singlets were sorted using FACS before processing by HCR–FlowFISH. Nuclei were either stored for future use according to 10x Genomics recommendation or proceeded directly to HCR–FlowFISH.

GBM FFPE dissociation

FFPE samples were preprocessed on a prototype S2 Singulator system. The sample was automatically processed in a NIC+ cartridge (S2 Genomics, cat. no. 100-215-389) by three 15-min deparaffinization steps (CitriSolv, VWR), rehydrated by successive 1-ml washes of 100%, 100%, 70%, 50% and 30% ethanol, followed by two washes of PBS. The sample was then spun at 1,000*g* for 3 min and resuspended in 0.5 ml of Nuclei Isolation Reagent (NIR, S2 Genomics, cat. no. 100-063-396) with 0.1 U μl^{-1} of RNase inhibitor (Protector, Millipore Sigma, cat. no. 3335399001); all subsequent solutions had RNase inhibitor. The sample was dissociated to single nuclei in a second NIC+ cartridge with 2 ml of NIR for 10 min followed by a 2-ml wash with Nuclei Storage Reagent (NSR, S2 Genomics, cat. no. 100-063-405). The single-nuclei suspension was spun 500*g* for 5 min, resuspended in NSR and counted.

Bioinformatics analyses overview

All bioinformatics analyses were conducted using standard output files from the execution of CellRanger to sequencing data of the Flex libraries. Downstream analyses, including cell filtering, marker-gene analyses and visualization, were performed using Seurat v.4 (ref. 13). In brief, cells were identified via a combination of passing the CellRanger knee plot as well as meeting minimum quality control standards, including at least 1,000 UMIs detected, 500 genes detected and no more than 5% mitochondrial RNA abundance, which are standard thresholds for scRNA-seq analyses. For all subclustering analyses (Figs. 3g, 4h and 6g,m), we required cells to be present in the enriched PERFF-seq library and belonging to the Seurat cluster associated with most of the population. All differential expression and marker-gene analyses were performed using the FindMarkers functionality in Seurat[13]. All customized code to reproduce all customized downstream analyses, including intermediate data files, is available as part of an online repository.

Benchmarking analyses

To examine the loss of data quality in PERFF-seq compared with analogous Flex libraries, all comparisons were made against gold-standard data generated and released by 10x Genomics. Saturation curves were drawn by downsampling the total reads in the library to 0.1%, 1.0%, 2.5%, 5.0%, 10%, 30%, 50%, 75% and 100% of the total sequencing depth. Downsampling proceeded via the sample() function in R on the per-molecule vectors encoded in the _sample_molecule_info.h5 file from the CellRanger processing. To compare median per-cell UMI and gene counts, we selected the read depth of the lowest library in the comparison and downsampled every other library to compare relevant statistics. Thus, these analyses are robust to differences in sequencing depth, UMI collapsing and barcode correction (which occur within the CellRanger processing steps upstream of the h5 output).

To assess the impact of the PERFF-seq workflow compared with Flex, the *CD3E*-targeted transcript was analyzed alongside *CD3D* (Extended Data Fig. 2f–h). For these panels, we subsetted the analyzed cells to the distinct clusters of T cells from the Seurat embedding. Differential gene expression analyses were conducted using the FindMarkers function and requiring a $\log_2$(fold-change) of 0.5 for inclusion in the Venn diagram, resulting in only two genes, *CD3E* and *S100A8*, across the three pairs of comparisons. These results indicate that the overall transcripts of T cells were not changed between the potential conditions, although there is an ~4× reduction in *CD3E* expression where ~70% of T cells had detectable levels of *CD3E* (compared with 95%) from Flex alone. As other genes (for example, *BCL11A*, *SPI1* and *Mobp*) and ~90% of cells sorted via FlowFISH have detectable expression from the Flex library, we suggest that each gene may behave differently, probably as a function of overall gene length and whether the Flex and HCR–FlowFISH probes overlap at the same locus on the mRNA. Thus, special consideration is required for the specific genes used in the FlowFISH panel, but the overall transcripts of the cells should remain stable.

Immune cell analyses

Three methods were used to benchmark the enrichment efficiency (Figs. 2e and 3d). First, reference projections with Azimuth to a gold-standard PBMC atlas were computed, noting the differences in chemistries between the reference and projections (reference: reverse transcription based; here: probe-based Flex chemistry). Qualitatively, the cell-type assignments from Azimuth were sensible but projection on to the two-dimensional space often failed to cover the breadth of the reference, which we attribute to differences in the fundamental sequencing chemistry. As a second reference-based method, we annotated genes using the default celldex[50] workflow for human immune cells. For the logic-gated classification (Fig. 3), we partitioned the output from classification as 'B cells' for *MS4A1*$^+$, 'CD4$^+$ T cells' for *CD4*$^+$ and *CD3E*$^+$ cells, 'CD8$^+$ T cells' and 'T cells' for the *CD4*$^-$*CD3E*$^+$ population, and all other labels as the negative population. Finally, we individually clustered genes and defined cell-type annotations based on standard practice for the presence or absence of individual marker genes. The proportions annotated as accurate classification represent the total number of high-quality cells ($n > 10{,}000$ per comparison) and were consistent between different classification methods, verifying the specificity of our enrichment via sorting strategy and preservation of transcriptomes for downstream analyses. For comparisons with other RNA-seq datasets, normalized data from flow-sorted bulk[21] populations and single-cell annotations[13], a collection of 2,674 target genes of BCL11A was downloaded from Harmonizome 3.0 (ref. 51) using the standard AddModuleScore functionality in Seurat[13]. For the mosaic LOY analyses, all libraries were annotated using the Azimuth reference projection functionality in Seurat[13]. Nine of the ten genes in the MSY region were included in the Flex WTA probe set, which was used for verifying LOY enrichment or depletion (Extended Data Fig. 5b). Gene set enrichment analyses were performed using the standard msigdb gene sets, specifically the Hallmark list.

Nuclei analyses

Mouse brain analyses, including clustering and subclustering analyses, proceeded as described above. The selection of plotted marker genes followed the cerebellum atlas that defined oligodendrocyte subtype markers without any prior enrichment[26]. To compare PERFF-seq performance against frozen nuclei, we downloaded a public eye nuclei Flex library (https://www.10xgenomics.com/datasets/40k-mixture-of-nuclei-isolated-from-4-mouse-tissues-multiplexed-samples-4-probe-barcodes-1-standard) as the closest anatomical tissue to our profiled tissue, an imperfect yet useful comparison (Fig. 6b). Analyses of *Mobp*-expressing cells from the adult mouse cerebellum were performed using the interactive web browser via a recent atlas[26].

For existing GBM FFPE Flex data, the counts matrix was downloaded from the datasets hosted on the 10x Genomics website. The two GBM samples were run as a 4-plex in-line barcode multiplexing with another tumor type (colorectal), which was discarded during pre-processing. These processed data were used both in defining endothelial or mural cell markers (Extended Data Fig. 6b,c) and in the downsampling performance analysis (Fig. 6c). Selection of marker genes was based on prior profiles of endothelial cells from GBM cells[40].

Statistics and reproducibility

Single replicates of the Flex libraries were used to develop the PERFF-seq protocol. No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. A total of 18 total PERFF-seq libraries were completed as part of this work. All customized analyses are available as part of our online resources.[52]

Reporting summary

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

Data availability

Sequencing data associated with this work are available on the Gene Expression Omnibus (GEO) at accession no. GSE262355. A full step-by-step protocol for PERFF-seq is available on protocols.io (https://doi.org/10.17504/protocols.io.14egn3k6ql5d/v1 and https://doi.org/10.17504/protocols.io.8epv5x8r4g1b/v1).

Code availability

Customized code and intermediate code files to reproduce all analyses supporting this manuscript are available at https://github.com/clareaulab/perffseq_reproducibility and via Zenodo at https://doi.org/10.5281/zenodo.14089656 (ref. 52).

References

49. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
50. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
51. Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, baw100 (2016).
52. Lareau, C. A. PERFF-seq reproducibility. *Zenodo* https://doi.org/10.5281/zenodo.14089656 (2024).

Acknowledgements

We thank the Satpathy Lab, Lareau Lab, Gladstone Flow Cytometry Core and Single-cell Analytics Innovation Lab members for helpful discussions. We acknowledge a helpful blog post from 10x Genomics describing the singlet unligated probe set. We thank T. Nawy for helpful feedback on the manuscript, A. Chow for assistance with flow cytometry and N. Pereira with S. Jovanovich of S2 Genomics for support with nuclei isolation from FFPE tissue. This work was supported by National Insitutes of Health grants (nos. P30CA008748, R00HG012579 and U01AT012984 to C.A.L. and UM1HG012076 to A.T.S. and L.S.L.) and a Gates Foundation seed award. A.T.S. is supported by the Burroughs Wellcome Fund Career Award for Medical Scientists, the Parker Institute for Cancer Immunotherapy, a Pew-Stewart Scholars for Cancer Research Award, a Cancer Research Institute Lloyd J. Old STAR Award, a Scholar Award from the American Society of Hematology and a Baxter Foundation Faculty Scholar Award. Y.H.H. is supported by a PhD fellowship from the Hector Fellow Academy. L.S.L. is supported by the Hector Fellow Academy, the Paul Ehrlich Foundation, the European Molecular Biology Organization Young Investigator Programme, an Emmy Noether fellowship (grant no. LU 2336/2-1) and grants by the German Research Foundation (Dnos. LU 2336/3-1, LU 2336/6-1, STA 1586/5-1, TRR241 and SFB1588, and Heinz Maier-Leibnitz Award). The Single-cell Analytics Innovation Lab is supported by the Alan and Sandra Gerry Metastasis and Tumor Ecosystems Center at Memorial Sloan Kettering. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

T.A., R.R.S., M.T.T., R.C., A.T.S. and C.A.L. conceived and designed this work. T.A. and R.R.S. led assay development with input from A.T.S. and C.A.L. T.A., R.R.S., M.T.T., B.N.N., Y.-H.H., S.H. and C.S. performed experiments. K.K.H.Y., Z.A.-M., V.T., B.E.H. and L.S.L. aided in the interpretation of data analyses. R.R.S., R.C., A.T.S. and C.A.L. supervised the work. C.A.L. led bioinformatics analyses with input from T.A. and R.R.S. T.A. led the development of the protocol with input from R.R.S. and M.T.T. T.A., R.R.S., A.T.S. and C.A.L. drafted the manuscript with input from the other authors.

Competing interests

A.T.S. is a founder of Immunai, Cartography Biosciences, Santa Ana Bio and Prox Biosciences, an advisor to Wing Venture Capital and receives research funding from Astellas and Merck Research Laboratories. R.R.S., L.S.L. and C.A.L. are consultants to Cartography Biosciences. R.C. is a consultant for Sanavia Oncology, S2 Genomics and LevitasBio. The other authors declare no competing interests.

Additional information

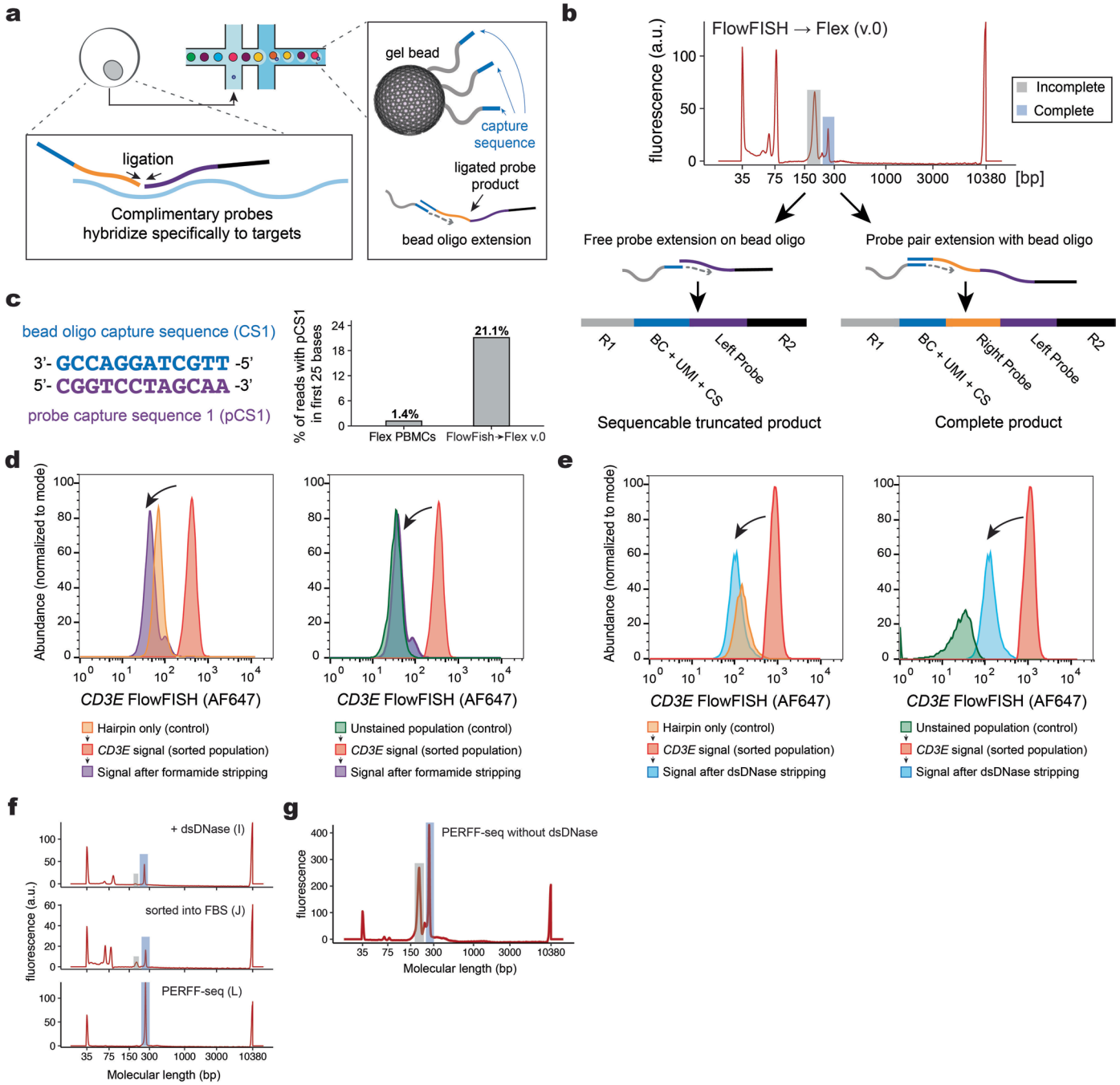
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-02036-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-02036-7.

Correspondence and requests for materials should be addressed to Robert R. Stickels, Ronan Chaligné, Ansuman T. Satpathy or Caleb A. Lareau.

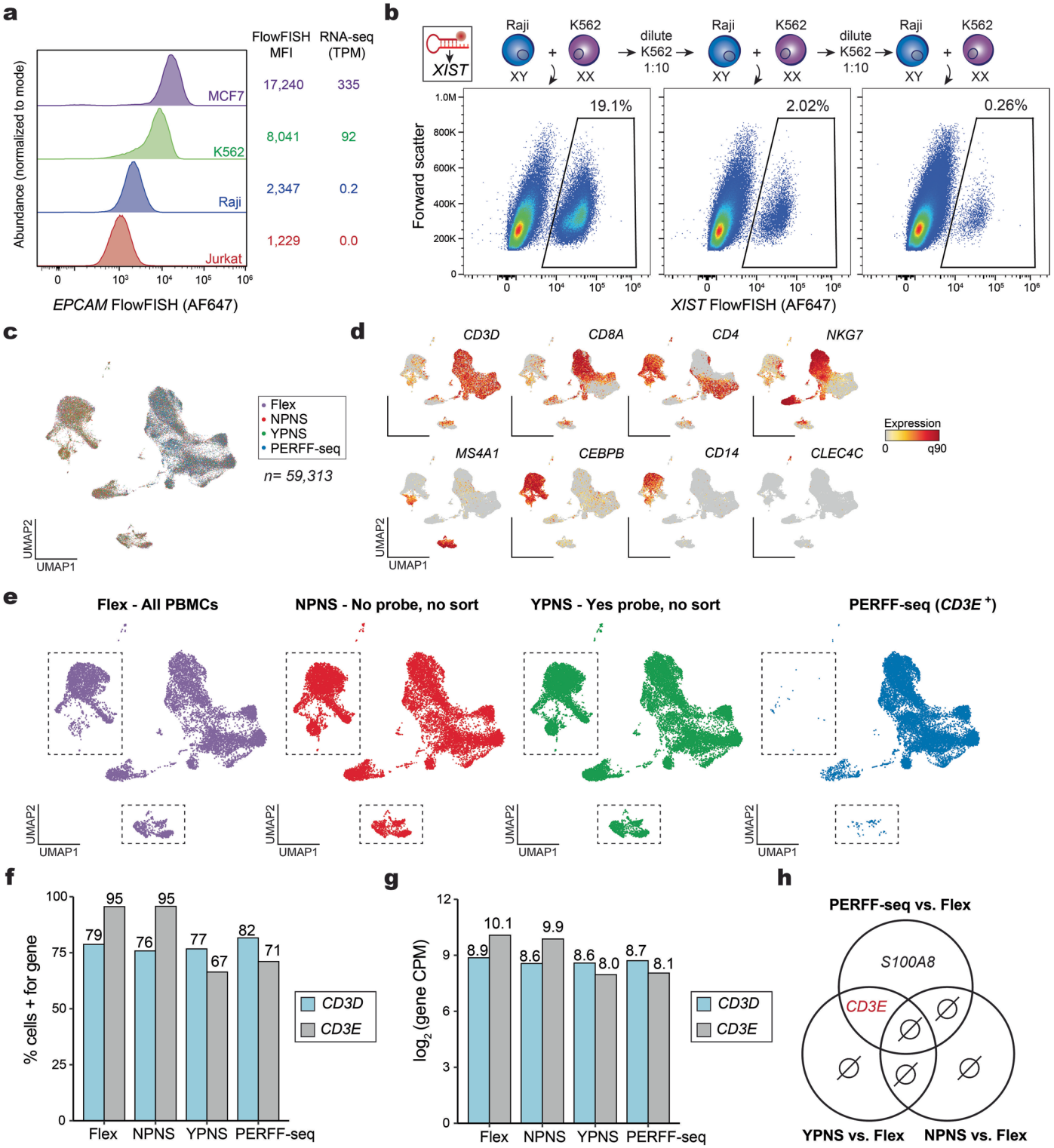
Peer review information *Nature Genetics* thanks Dominic Gruen, Sydney Shaffer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-02036-7



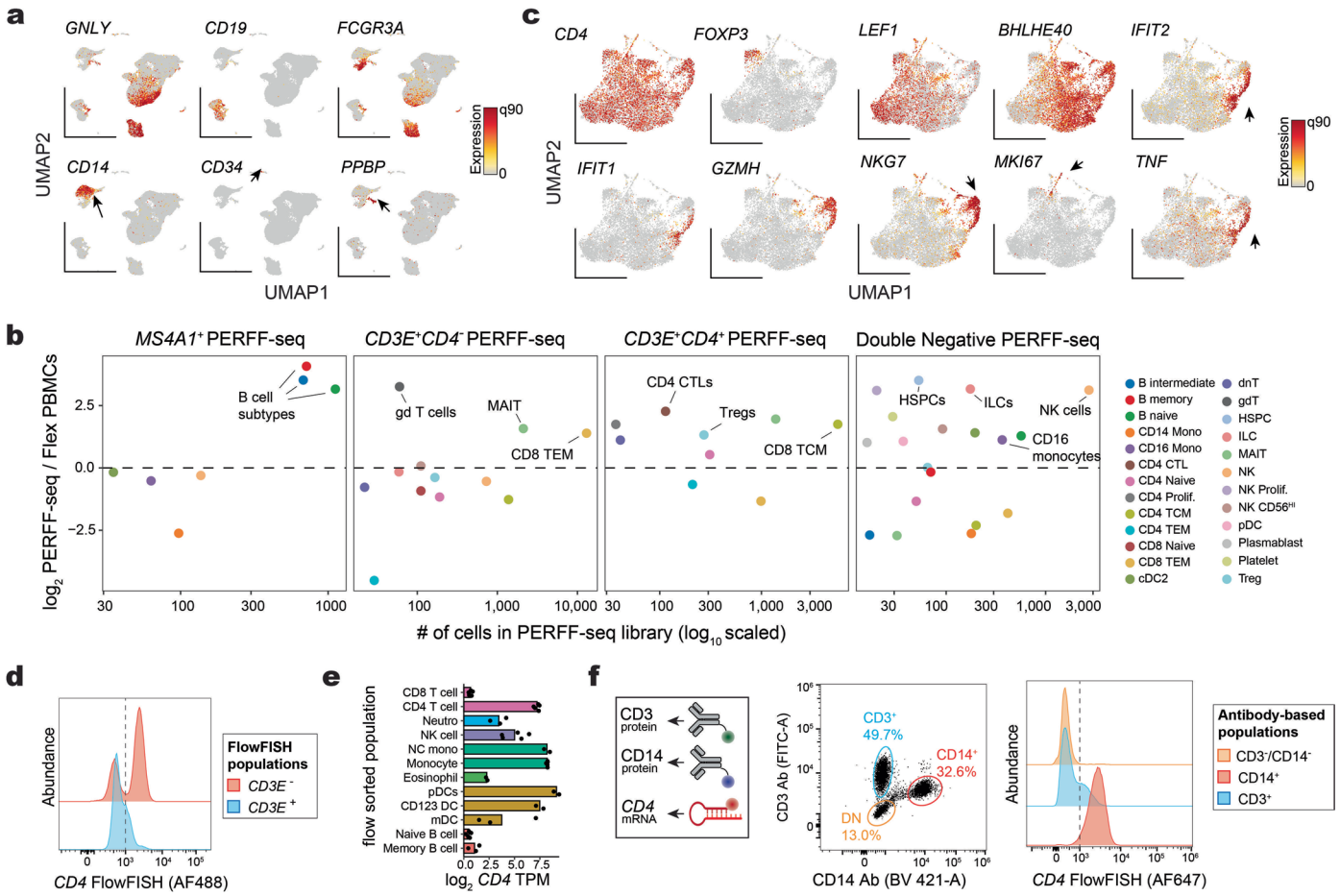
Extended Data Fig. 1 | Analyses supporting PERFF-seq development. (**a**) Schematic overview of Flex workflow, including probe hybridization to transcript fragments in cells upstream of Chromium and bead oligo extension of the ligation product. (**b**) Representative Bioanalyzer (Agilent Technologies) trace outlining complete versus incomplete sequencing molecules. (**c**) Graphical summary of probe capture sequence (pCS1) bead oligo capture sequence (left) and percent of reads with pCS1 detected in first 25 bases. (**d**) Comparison of FlowFISH signal using either unstained cells or the hairpin only comparing the sorted *CD3E*$^{+}$ population and/or stripped via formamide. (**e**) Same as in (**d**) but using dsDNase for stripping. Note: quantifying fluorescence after sorting/stripping (**d**, purple; **e**, blue) is not standard for the PERFF-seq protocol but shown here as part of assay development. (**f**) Bioanalyzer traces for representative libraries from panels in Fig. 1, highlighting half- and fully-mapped probes. (**g**) Bioanalyzer traces of library preparation where the full PERFF-seq workflow was completed except omiting the dsDNAse stripping step.

https://doi.org/10.1038/s41588-024-02036-7



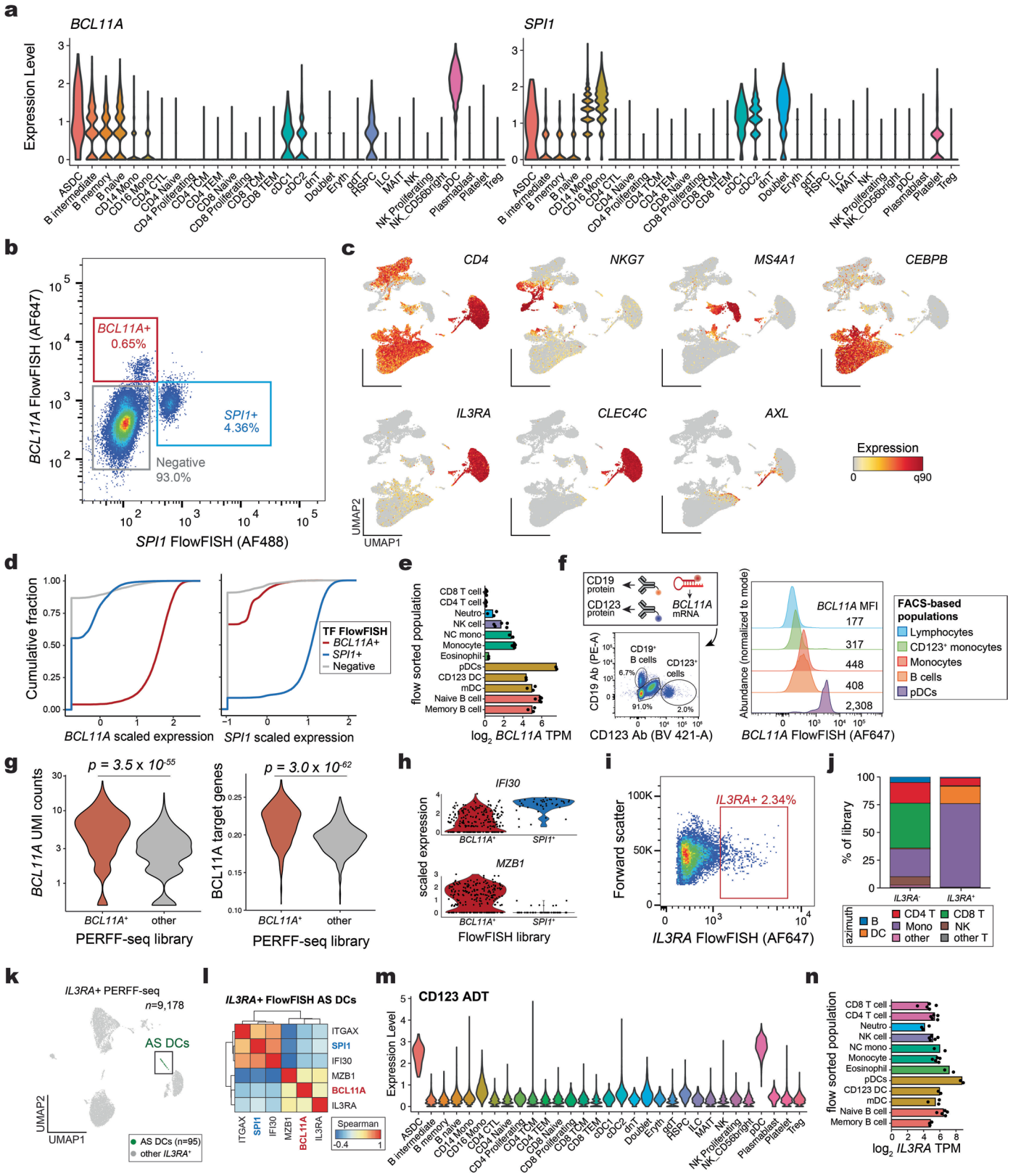
Extended Data Fig. 2 | Supporting analyses of assay benchmarking. (**a**) HCR FISH staining and quantification of *EPCAM* across cell lines. Mean fluorescence intensity (MFI) and bulk RNA-seq transcripts per million (TPM from[49]) are noted for each condition. (**b**) Replication of cell line mixing experiment and staining for *XIST*. (**c**) Reduced dimensionality embedding of all cells in the four-plex benchmarking experiment. PERFF-seq was enriched for *CD3E*+ cells. No probe, no sort (NPNS) and Yes probe, no sort (YPNS) profile all PBMC subpopulations with minor modifications to the reaction. (**d**) Marker genes supporting annotation of key populations. (**e**) Same as (**c**) but stratified by library. Boxes indicate B cell and monocyte populations that are depleted from the PERFF-seq library. (**f**) Percent of T cells from each library with at least 1 UMI for *CD3D* or *CD3E*. (**g**) log2 counts per million (CPM) of *CD3D* and *CD3E* across different library conditions. (**h**) Differentially expressed genes between different facets of PERFF-seq compared to Flex. Two genes were differentially expressed, including *CD3E* in both the YPNS and PERFF-seq conditions. Ø means empty or no genes detected.

https://doi.org/10.1038/s41588-024-02036-7



Extended Data Fig. 3 | Supporting analyses of combinatorial PBMC cell states. (**a**) Additional marker genes for distinct populations from PBMC cell type analyses. Arrows indicate markers for rare populations expected from PBMC profiling. (**b**) Relative enrichment of cell populations (colours) for each PERFF-seq library compared to the Flex PBMC library. (**c**) Subclustering of *CD3E*$^+$/*CD4*$^+$ cells, highlighting rare subclusters marked by relevant genes. (**d**) Summary of *CD4* HCR FISH signal, stratified by *CD3E* populations. (**e**) Bulk RNA-seq expression of *CD4* from FACS-isolated populations[21]. Replicates for **e** are all libraries from Haemopedia[21] with no statistical test. (**f**) Design and results of antibody and HCR FISH co-staining to evaluate *CD4* RNA and protein expression.

https://doi.org/10.1038/s41588-024-02036-7

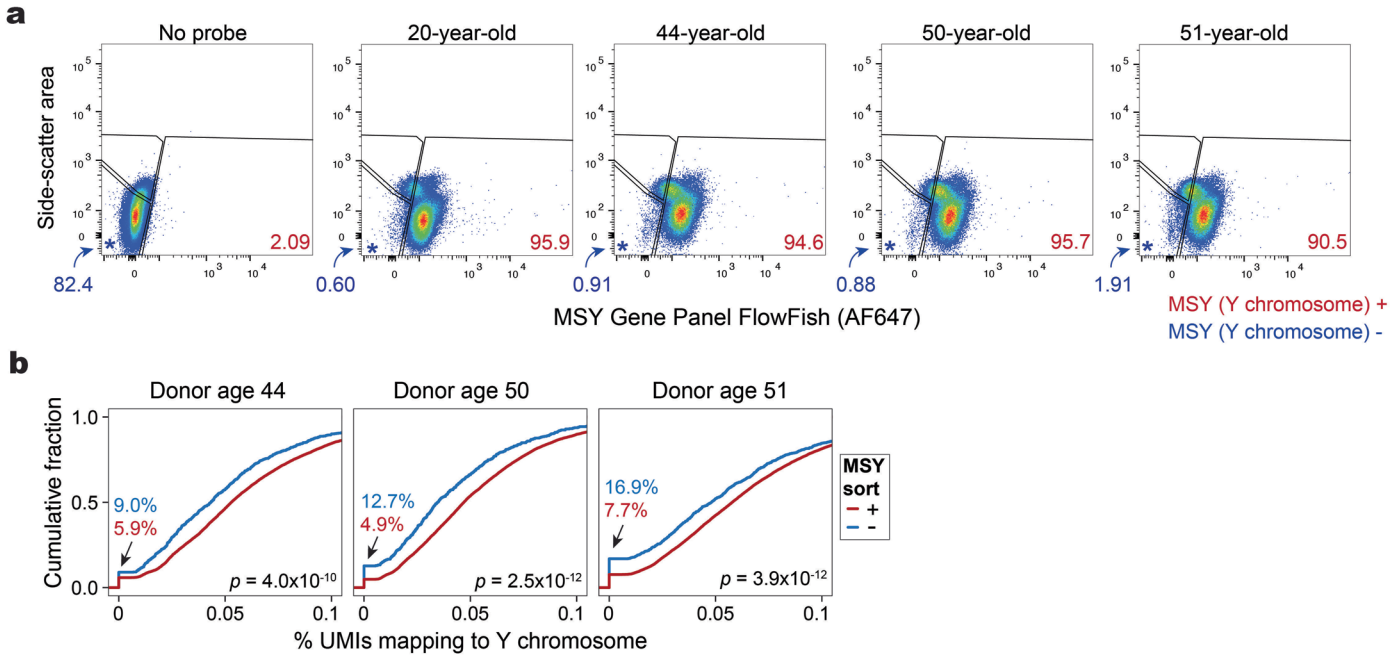


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41588-024-02036-7

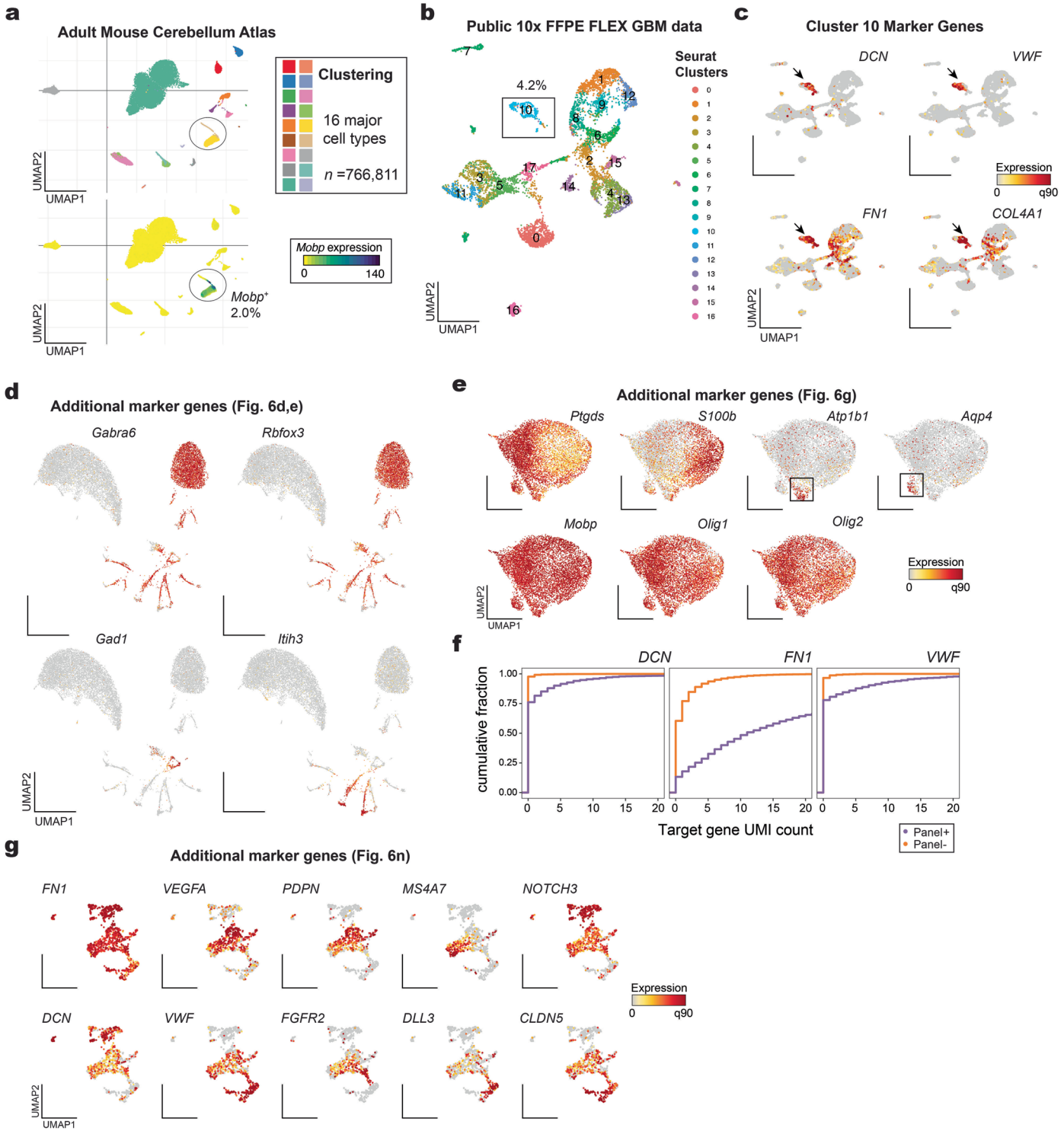
Extended Data Fig. 4 | Supporting analyses of unconventional enrichments of rare cell states. (**a**) Azimuth Violin plots for *BCL11A* and *SPI1* RNA expression across well-annotated populations in peripheral blood mononuclear cell types. (**b**) Summary of FACS populations, including unsorted, *BCL11A*$^+$, and *SPI1*$^+$ populations. (**c**) Additional marker genes supporting cell type annotations. (**d**) Empirical cumulative distribution plot of scaled expression of *BCL11A* (left) and *SPI1* (right) stratified by the captured PERFF-seq library. (**e**) Bulk RNA-seq of sorted populations of *BCL11A*[21]. (**f**) Design (left) and results (right) of cytometry analysis of PBMCs co-stained with *BCL11A* mRNA (via HCR-FISH) and CD19 and CD123 protein (via antibodies). Mean fluorescence intensity (MFI) for *BCL11A* of each population is quantified. (**g**) Comparison of B cells from *BCL11A*+ FlowFISH or negative/*SPI1*$^+$ populations for *BCL11A* expression or BCL11A target gene module scores. Uncorrected p-value for the two-sided Wilcoxon rank-sum test is shown. (**h**) Additional violin plots of marker genes, stratified by the FlowFISH library. All genes were significantly differentially expressed at a false discovery rate (FDR) < 0.01. (**i**) Summary of *IL3RA*$^+$ FACS sort and population characterized with PERFF-seq. (**j**) Proportion of cell types from the Azimuth L1 reference for *IL3RA*+/- PERFF-seq libraries. (**k**) Reduced dimensionality representation of *IL3RA*$^+$ PERFF-seq library, highlighting profiled AS DCs. (**l**) Gene-gene Spearman correlations of all AS DCs from the *IL3RA*$^+$ sort. Genes match those in Fig. 4k. (**m**) Summary of CD123 expression from antibody-derived tags (ADT) of PBMC CITE-seq. (**n**) Bulk RNA-seq of sorted populations of *IL3RA*[21]. Replicates for **e** and **n** are all libraries from Haemopedia[21] with no statistical test.

https://doi.org/10.1038/s41588-024-02036-7



Extended Data Fig. 5 | Supporting analyses for mosaic loss of Y chromosome. (**a**) FlowFISH cytometry gating scheme, including control (no probe, left) and four male (XY) donors of different ages. The percent of cells corresponding to MSY^{+} (red) and MSY^{-} (blue) in each donor gate are shown as numeric values. (**b**) Empirical cumulative distribution plot of the percent of Y chromosome UMIs stratified by the captured PERFF-seq library. Percentages of cells with 0 MSY UMIs from the scRNA-seq library are noted. P-values are from a two-sided Kolmogorov–Smirnov test comparing the distributions of the positive and negative samples.

https://doi.org/10.1038/s41588-024-02036-7



Extended Data Fig. 6 | Supporting analyses of nuclei enrichment from fresh and fixed tissues. (**a**) UMAP of adult mouse cerebellum atlas[26], including cell types (top) and *Mobp* expression (bottom). The *Mobp*$^+$ oligodendrocytes and oligodendrocyte precursors are circled with their frequency noted. (**b**) Reduced dimensionality representation of public GBM FFPE Flex data showing 17 clusters. (**c**) Annotation of marker genes for cluster 10, the population highlighted by the arrow in (**b**). (**d**) Supporting marker genes annotating other subpopulations from the PERFF-seq experiment, including the primary cluster of granule cells. (**e**) Additional marker genes from *Mobp*$^+$ cells were profiled with PERFF-seq. *Atp*-associated genes (*Atp1b1, Aqp4*) supporting rare subclsuters are highlighted in the boxes as well as marker genes highly expressed in all cells. (**f**) Empirical cumulative distribution plot of the raw UMI counts for each of the three genes enriched via FlowFISH, stratified by the captured PERFF-seq library. (**g**) Additional marker genes showing heterogeneity defining subclusters of endothelial cells and pericytes.

nature portfolio

Corresponding author(s): Stickels, Chaligne, Satpathy, Lareau

Last updated by author(s): 10/10/2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Software: CellRanger v7
Data analysis	Custom code and intermediate data files to reproduce all analyses supporting this manuscript are available at https://github.com/clareaulab/perffseq_reproducibility. A full step-by-step protocol for PERFF-seq is available on protocols.io (dx.doi.org/10.17504/protocols.io.14egn3k6ql5d/v1, dx.doi.org/10.17504/protocols.io.8epv5x8r4g1b/v1). Downstream analysis packages include: Seurat v4, Harmonizome 3.0, celldex v1.14, azimuth v1, FlowJo v10.10, Harmonizome 3.0 for TF targets.

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

Data

Policy information about availability of data

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

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

Sequencing data associated with this work is available at GEO accession GSE262355.

Research involving human participants, their data, or biological material

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

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No sample sizes were chosen prospectively. Cell captures were determined based on loading recommendations from 10x Genomics.Single library replicates were chosen for cost efficiency and relying on the single cell data for statistical power. Sample sizes are sufficient as a clear hypothesis is tested via analysis of thousands of cells per experiment. Donor replicates were used for critical experiments, including LOY sort (Fig. 5) and nuclei sort (Fig. 6).
Data exclusions	No data were systematically excluded. Cells were filtered based on well-established quality control methods.
Replication	The PERFF-seq protocol was replicated by four individuals across three different institutions to ensure that the protocol was functioning appropriately.
Randomization	Randomization was not relevant for our study as no experimental groups were determined ahead of time.
Blinding	Blinding was not relevant for our study as most analyses were conducted with a specific hypothesis in mind that was encoded in the experimental condition. The behavior of the biological samples here are not impacted by the unblinded design.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used	For antibody staining, we followed the manufacturer protocol and stained with 5uL of antibody per 1 million cells in 100uL of staining buffer (1% BSA in 1x PBS). Clones used in this work were BioLegend anti-human CD123 (IL3Ra) S18016F, cat #396705; BioLegend anti-human CD19 HIB19, cat #302208; BioLegend anti-human CD3 OKT3, cat # 317444; BioLegend anti-human CD14 M5E2, cat # 301829.
Validation	Specificity of antibodies were verified from manufacturer analyses. All protein + cells were validated by the RNA FISH probe against the same gene within the same cells.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Rajis, Jurkats, MCF7, K562 used in this were purchased from ATCC.
Authentication	Cell line identities were confirmed via STR profiling via ATCC.
Mycoplasma contamination	Cell lines were tested for mycoplasma contamination at MSKCC and were verified as negative.
Commonly misidentified lines (See ICLAC register)	Lines used in this study where not part of the set of misidentified lines.

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Samples were resuspended in sorting and collection buffer (1x PBS, 5% BSA - Gibco #15260037, 0.13U/uL RNAse inhibitor - Millipore Sigma #3335399001) and filtered through a 35um strainer.
Instrument	Cells were sorted using the BD FACSAria III or FACSymphony S6 with 70um nozzle for cells and 85um nozzle for nuclei.
Software	FlowJo v10.10 was used for analysis and making figures
Cell population abundance	Gates were drawn based on live cell staining and FSC and SSC consistent with viable cell inputs. Based on single-cell RNA-seq, the populations were highly enriched, including 95%+ T cells in the relevant figure. We emphasize that our flow application was not counting as is typical for flow analyses but as input into scRNA-seq, which provides an orthogonal validation of the sort logic utilized in this study.
Gating strategy	Gates were drawn using relevant negative controls. For FlowFISH experiments, these are typically the same population with amplifier but no targeted gene probe.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.