ARTICLES
https://doi.org/10.1038/s41587-022-01452-6

nature biotechnology

OPEN

Mostly natural sequencing-by-synthesis for scRNA-seq using Ultima sequencing

Sean K. Simmons[1,2,7], Gila Lithwick-Yanai[3,7], Xian Adiconis[1,2,7], Florian Oberstrass[3], Nika Iremadze[3], Kathryn Geiger-Schuller[1,6], Pratiksha I. Thakore[1,6], Chris J. Frangieh[1,4], Omer Barad[3], Gilad Almogy[3], Orit Rozenblatt-Rosen[1,6], Aviv Regev[1,5,6], Doron Lipson[3] and Joshua Z. Levin[1,2] ✉

Here we introduce a mostly natural sequencing-by-synthesis (mnSBS) method for single-cell RNA sequencing (scRNA-seq), adapted to the Ultima genomics platform, and systematically benchmark it against current scRNA-seq technology. mnSBS uses mostly natural, unmodified nucleotides and only a low fraction of fluorescently labeled nucleotides, which allows for high polymerase processivity and lower costs. We demonstrate successful application in four scRNA-seq case studies of different technical and biological types, including 5′ and 3′ scRNA-seq, human peripheral blood mononuclear cells from a single individual and in multiplex, as well as Perturb-Seq. Benchmarking shows that results from mnSBS-based scRNA-seq are very similar to those using Illumina sequencing, with minor differences in results related to the position of reads relative to annotated gene boundaries, owing to single-end reads of Ultima being closer to gene ends than reads from Illumina. The method is thus compatible with state-of-the-art scRNA-seq libraries independent of the sequencing technology. We expect mnSBS to be of particular utility for cost-effective large-scale scRNA-seq projects.

Single-cell RNA sequencing (scRNA-seq) enables the study and characterization of cellular states and pathways at ever-growing experimental scales, including the Human Cell Atlas[1], cell atlases for tumors[2] and other diseases[3,4], and large-scale Perturb-Seq screens of millions of cells under genetic[5,6] or drug[7] perturbations. Methods for capturing and processing single-cell libraries have been radically scaled in the past few years[8–11], but sequencing itself has largely relied on Illumina technology. Here we describe the development of a sequencing technology intended to facilitate large-scale studies. Mostly natural sequencing-by-synthesis (mnSBS) is a new sequencing chemistry that relies on a low fraction of labeled nucleotides, combining the efficiency of non-terminating chemistry with the throughput and scalability of optical endpoint scanning within an open fluidics system to enable high-throughput sequencing, and has been demonstrated on Genome-in-a-Bottle reference samples and samples from the 1000 Genomes project[12]. To benchmark mnSBS with scRNA-seq, we performed experiments with four library types, sequenced in parallel on an Illumina sequencer and on an Ultima Genomics (Ultima) prototype sequencer implementing mnSBS (Fig. 1a).

To implement mnSBS for massively parallel, droplet-based scRNA-seq, we converted a typical scRNA-seq work flow to be compatible with Ultima sequencing (Fig. 1b–d; Methods). Focusing on 10x Chromium scRNA-seq (Methods), a popular method, we first added adapters to cDNA libraries specific for Ultima sequencing (Fig. 1b). Next, we address the fact that droplet-based scRNA-seq relies on pairing each cDNA read with a cell barcode (CBC) and a unique molecular identifier (UMI) (Methods). With Illumina sequencing, the two ends of the library are sequenced separately by paired-end sequencing, but for single-end Ultima sequencing, we capture all the information in a single read of 200–250 bases (Fig. 1d and Extended Data Fig. 1), such that the CBC and UMI are read first and followed by the cDNA. For those reads derived from the 3′ end of the transcript, we sequence through poly(T) bases, which are the result of the mRNA poly(A) tail, adjacent to the cDNA sequence.

To evaluate mnSBS with scRNA-seq, we carried out experiments with four libraries, spanning different technical and biological use cases, and sequenced each in parallel on both Ultima and Illumina sequencers (Methods). Three libraries were from peripheral blood mononuclear cells (PBMCs) of healthy human donors, spanning 3′ scRNA-seq (~7,000 cells, 1 individual), 5′ scRNA-Seq (~7,000 cells, 1 individual) and a library generated in multiplex by pooling cells from eight donors (~24,000 cells, 8 individuals, 5′ scRNA-seq). We chose PBMCs because they are primary human cells, include diverse cell types of various sizes and frequencies and have been used for previous benchmarking[13,14]. The fourth library was from a Perturb-Seq[5,6] experiment, where ~20,000 cells were profiled after clustered regularly interspaced short palindromic repeats (CRISPR)–Cas9 pooled genetic perturbation, followed by scRNA-seq to detect both the profile of the cell and the associated guide RNA. Together, the four libraries span three major use cases—individual patient atlas, multiplex patient profiling, and large-scale screens, and the two most commonly used library types for scRNA-seq.

We first tested the feasibility of mnSBS for scRNA-seq, with matched Ultima and 5′ and 3′ droplet-based scRNA-seq of PBMCs. Initial analysis (Methods) showed that the number of UMIs generated at a given sequencing depth was comparable between Ultima and Illumina in the 5′ libraries, while for the 3′ libraries we obtained more UMIs with Illumina than Ultima (Fig. 2a), owing to differences in sequence quality. While Ultima and Illumina data for 5′ libraries were similar, for the 3′ data there was lower quality for Ultima in the bases flanking the poly(T) region—the 3′ end of the

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Ultima Genomics, Newark, CA, USA. [4]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA. [5]Department of Biology, MIT, Cambridge, MA, USA. [6]Present address: Genentech, South San Francisco, CA, USA. [7]These authors contributed equally: Sean K. Simmons, Gila Lithwick-Yanai, Xian Adiconis. ✉e-mail: jlevin@broadinstitute.org

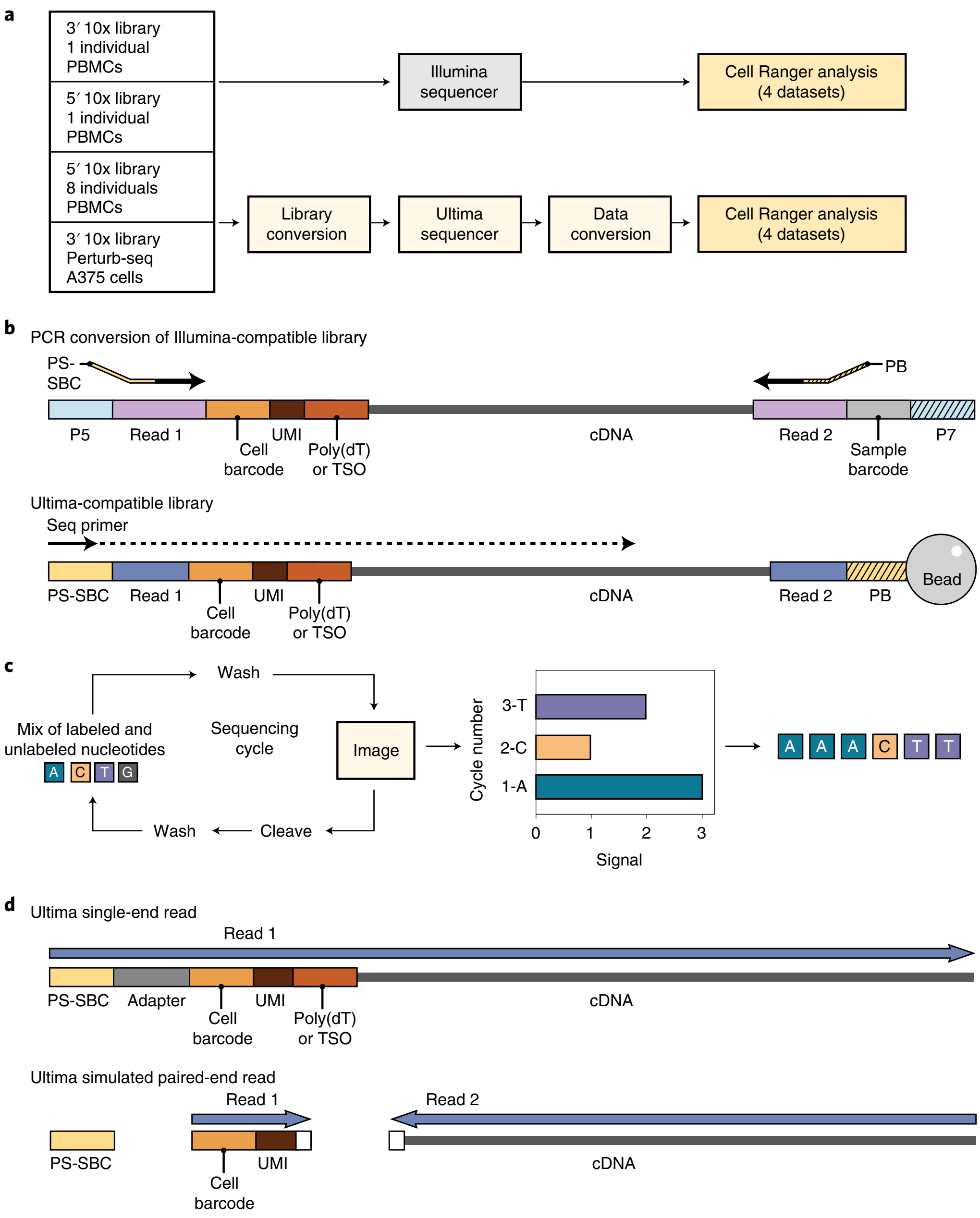


Fig. 1 | Experimental design. **a**, Work flow showing four samples used and adjustments made for Ultima sequencing. **b**, Library conversion showing PCR process to change adapters from Illumina (P5 and P7, parts of Read 1 and 2) to Ultima (Primer for Sequencing + Sample Barcode (PS-SBC) and Primer for Bead (PB), parts of Read 1 and 2). The 5′ libraries have TSO and 3′ libraries have poly(dT). Our Ultima libraries did not require index sequences for combining libraries together, though this feature can be added in the future. **c**, mnSBS schematic. **d**, Data conversion of single-end reads to simulated paired-end reads needed for Cell Ranger analysis. White box shows five bases trimmed from cDNA and three bases trimmed from UMI adjacent to the poly(dT) sequence in 3′ libraries. In 5′ libraries, only three bases were trimmed from the cDNA next to the TSO. PS-SBC read is used to deconvolute multiplexed libraries.

UMI and the 5′ end of the cDNA (Extended Data Fig. 2a). Indeed, filtering out reads that have bases with quality <10 in their UMI (the filter applied by the pre-processing pipeline we used, Cell Ranger[15]) yields similar rarefaction curves for Illumina and Ultima (Extended Data Fig. 2b). Thus, much of the difference in the observed number of UMIs per sequenced read for 3′ libraries is explained by the lower sequence quality UMIs in the Ultima data caused by the need to sequence through the poly(T) bases. To overcome this, for 3′ libraries we trimmed five bases from the cDNA adjacent to the poly(T) bases and then explored how best to trim the UMI. As we shortened the UMIs, UMIs that differed only in the trimmed bases ‘collapsed’ into a single UMI leading to decreases in the fraction of UMI–CBC pairs that occur in only one gene at roughly the same rate in Illumina and Ultima data (Extended Data Fig. 2c). Shortening the UMIs for Illumina had a minimal effect at nine bases or more (Extended Data Fig. 2d), suggesting that the challenges with Ultima reads were caused by lower base quality and that trimming to nine bases was reasonable.

We further investigated whether trimming the last three bases of the UMI impacted the results owing to increased ‘collisions’ when different full UMIs collapsed together to the same trimmed UMI. This could be the case when high UMI complexity is required, for

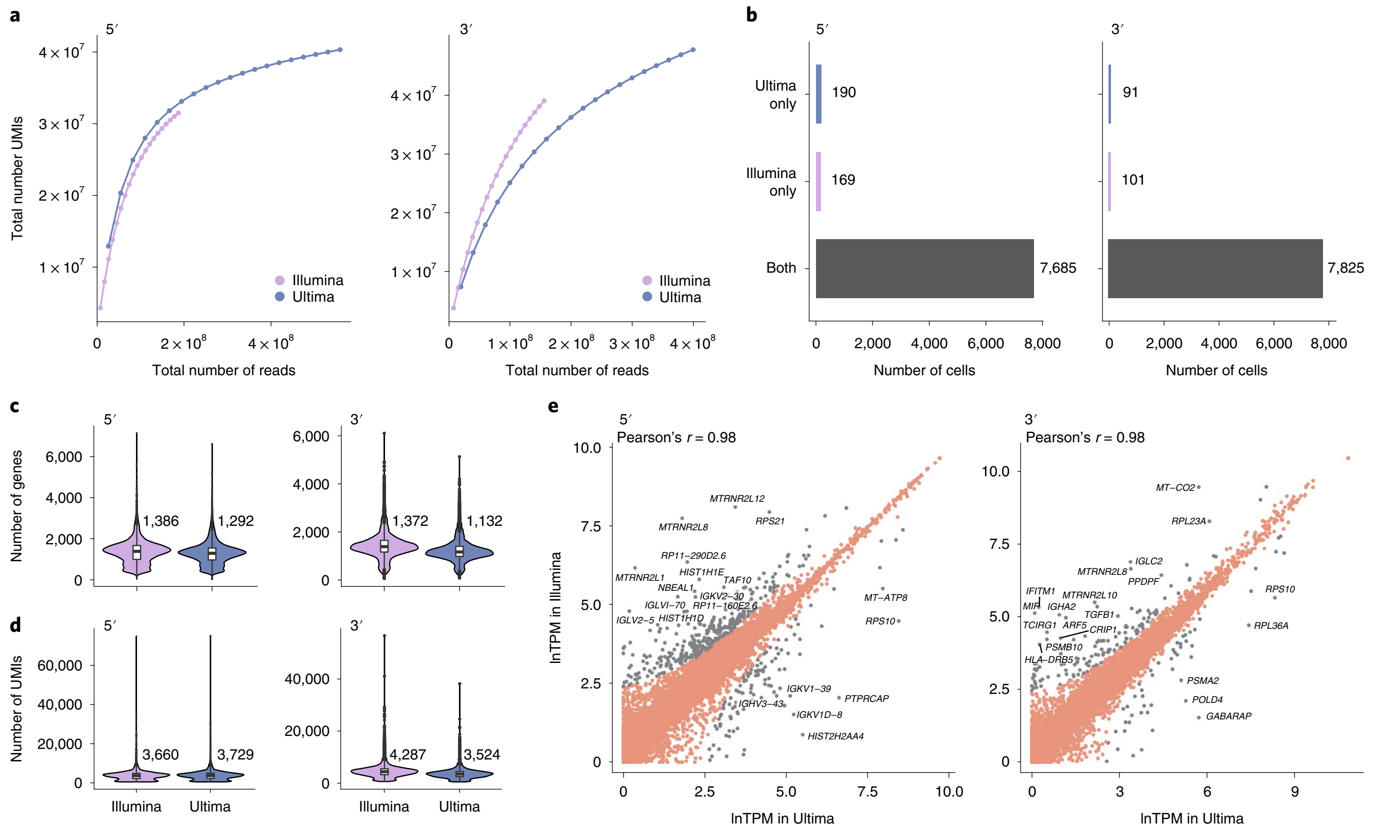


Fig. 2 | Quality metrics for matched 5′ and 3′ libraries. a, Total number of UMIs detected per cell at different sequencing depths. For **b**–**e**, reads were sampled so that Illumina and Ultima have the same number of reads. **b**, Number of cells identified by Cell Ranger only in Ultima, only in Illumina or both. **c**,**d**, Distribution of the number of genes (**c**) or UMIs (**d**) per cell. Box plots show the 25% and 75% quantiles with the median marked in between. **e**, Scatter plots with one point for each gene. Labeled genes (gray) have a high FC (FC > 2 using a pseudocount of 10 TPM). The 20 genes with the highest FC are labeled in each plot. For all 3′ libraries, the last three UMI bases were trimmed for quality reasons.

a cell with many UMIs detected or for a highly expressed gene. To explore this, we examined the ratio of the number of trimmed to untrimmed UMIs for cells and for genes with different numbers of UMIs, in the Illumina 3′ PBMC dataset (with the higher quality UMIs). At the cell level, the ratio reduces as the coverage increases, but only modestly, by less than 10% for all but a few dozen cells with very high number of UMIs (Supplementary Fig. 1a). At the gene level, very few of the highly expressed genes (8 of 3,908 genes with >1,000 UMIs) show a reduction of >10% (Supplementary Fig. 1b). Conversely, some very lowly expressed genes have lower ratios, likely because, for these genes, losing even one UMI will lead to a smaller ratio. Taken together, our analyses show that shortening UMIs has only a modest effect on highly expressed genes and high-complexity cell profiles. This led us to exclude the last three bases of each UMI in Ultima 3′ data in subsequent downstream analysis (Methods).

Next, comparing the performance of these PBMC 3′ and 5′ matched libraries, we obtained similar overall performance for both sequencing technologies. First, to correct for differences in sequencing depths, which were higher in Ultima than Illumina, we randomly sampled Ultima reads, so that we used the same number of reads for each sequencing platform (Methods). Both technologies identified nearly all the same CBCs (Fig. 2b; 7,916 cells (Ultima) versus 7,926 cells (Illumina) in the 3′ data, and 7,875 cells (Ultima) versus 7,854 cells (Illumina) in the 5′ data), with the same number of UMIs and genes per cell for 5′ libraries and slightly lower numbers for 3′ libraries with Ultima (as expected) (Fig. 2c,d). When we sampled reads to have the same number of UMIs (Methods), we obtained a similar number of genes per cell in Illumina and Ultima also for 3′ libraries (Extended Data Fig. 3). Other metrics (Supplementary Table 1) also showed similar overall performance, with slightly higher genome mapping rates in Ultima but comparable transcriptome mapping rates.

The two sequencing technologies yielded highly correlated expression levels for the matched 5′ and 3′ PBMC libraries, albeit with some outlier genes and minor differences (Pearson's $r = 0.98$ in all cases; Fig. 2e and Extended Data Fig. 3c). As expected, when a single sequencing run was randomly split into two datasets, we see even higher correlation of expression levels (Extended Data Fig. 3d). Specifically, there was a modest bias, particularly in the 3′ libraries, towards genes with higher GC content having higher expression in Illumina and the longest genes having higher expression in Ultima 3′ libraries (Extended Data Fig. 4a,b). Of the 166 genes with differences in expression for 3′ PBMC between the two sequencing platforms, most (130 genes, 78.3%) differed in the fraction of reads that were assigned by Cell Ranger to the gene out of all the reads mapped to that gene region (Extended Data Fig. 4c). This is likely related to how Ultima and Illumina reads map to different locations relative to the transcript, as expected from the difference in single-end versus paired-end reads (Fig. 1d). In 5′ data, Ultima reads map closer to the 5′ end than Illumina reads, while in 3′ data, Ultima reads map closer to the 3′ end than Illumina reads (Extended Data Fig. 4d,e). Because Cell Ranger excludes reads that do not fully map within annotated gene boundaries, more Ultima reads are excluded from analysis as they are closer to gene ends (Extended Data Fig. 4d,e), as shown, for example, for *LILRA5* and *HIST1H1D* (Extended Data Fig. 4f,g). This difference in location can also lead to more multimapping or ambiguous reads (Extended Data Fig. 4h and

Supplementary Table 2). For example, four (*ARF5*, *MIF*, *IFITM1* and *TCIRG1*) of the 20 genes with the largest log fold change (FC) (all logs are the natural logarithm (base *e*) in this study, unless otherwise noted) between Ultima and Illumina in the 3′ data (labeled in Fig. 2e) have higher expression in Illumina and a much higher rate of mapped ambiguous reads in the Ultima than the Illumina data (>50 versus <10 ambiguous reads per non-ambiguous read for each gene, respectively) (Supplementary Table 2), possibly explaining the difference in their expression levels. Shortening Ultima Read 2 to the same length as Illumina Read 2 had a small effect on the fraction of assigned reads (Extended Data Fig. 4h) and other metrics (Supplementary Table 1)—suggesting read length is not a major factor in the differences we observed.

To further explore the effects of gene annotation on Ultima and Illumina-based scRNA-seq, we extended the standard reference using RNA-seq data, as we have previously shown this can recover the expression of a gene with an alternative 3′ end compared to the annotation[16]. We created a pipeline that extends the annotated gene boundaries based on reads that overlap a gene but are not completely contained in any of its annotated exons (Methods). We generated three such references, extended with either (1) published bulk PBMC data[13], (2) the Ultima 3′ scRNA-seq data, or (3) the Ultima 5′ scRNA-seq data (with Ultima and Illumina data sampled to the same number of reads). We compared the expression of genes in Ultima data processed with the extended references to those in Illumina data either with or without the extended reference (Extended Data Fig. 5 and Supplementary Tables 1 and 3).

Analyzing the 5′ PBMC data with the extended reference decreased the number of differentially expressed (DE) genes between Ultima and Illumina by 22 to 23% (absolute logFC > ln2) compared with the standard reference, while other overall metrics were largely unchanged. In the 3′ data, there were a similar number of DE genes in analyses with the extended and standard references, although the expression of some genes, for example, *LILRA5* and *MT-CO2*, agreed much more closely using the extended reference. Comparing gene expression levels for the same sequencing dataset processed with the standard or an extended reference shows that most levels are very similar, though a sizeable number (23 to 83) are higher and a few (1 to 3) are lower (Extended Data Fig. 5c). Also, some of the top genes that differ between the extended and standard references are genes that differ between Ultima and Illumina with the standard reference, for example, *MT-CO2* and *LILRA5* in the 3′ data and *HIST1H1D* and *HIST1H1E* in the 5′ data (Fig. 2e and Extended Data Fig. 4f,g). This suggests that a data-driven extended reference might help recover expression in Ultima scRNA-seq data, particularly when using 5′ data. Alternatively, one could consider modifying the way Cell Ranger counts UMIs to better take advantage of reads that overlap genes but are not completely contained within them.

We examined the impact of the single-end Ultima versus paired-end Illumina data by sequencing the 5′ PBMC library with single-ended Illumina sequencing. Applying a similar pipeline to the one we used for Ultima with minor required modifications (Methods), we sampled the Ultima data to have the same number of reads as the single-end Illumina data and compared them (Supplementary Fig. 2). The two methods showed very high agreement in terms of the number of UMIs per cell and genes per cell, with far fewer outlier genes between Ultima and the single-ended Illumina data than observed when comparing to the paired-end data (Supplementary Table 3). Overall, the quality control metrics of single-end Illumina and Ultima sequencing are much more similar, particularly the mapping metrics (Supplementary Table 1).

To compare the biological insights derived from scRNA-seq using the two technologies, we turned to analyze 5′ scRNA-seq of PBMCs from eight individuals processed together and sequenced with both Ultima and Illumina (Methods). Both methods have roughly the same number of UMIs in this dataset (<1% difference) and performed similarly (using all reads; Supplementary Table 1 and Extended Data Fig. 6). We also generated matched T cell Receptor (TCR) and B cell Receptor (BCR) Illumina sequencing data (Methods). Ultima sequencing was not used for this, because the 10x Chromium constructs specifically require paired-ends or much longer single-end reads to cover the entirety of these genes.

In the eight individuals PBMC dataset, the two sequencing platforms produced very similar results for the common tasks of genotype-based assignment, cell-type labeling and identification of DE genes and were well embedded together. First, we used Vireo[17], which finds genotype clusters in the data without prior knowledge of the genotypes of individuals in the experiment, to assign reads to each individual in the mixture (Methods). Both Ultima and Illumina data returned highly concordant labels (Fig. 3a), with 92% agreement in label if we include those cells declared doublets or unassigned (χ^2 test for independence gives a $P < 2.2 \times 10^{-16}$ and $\chi^2 = 199{,}127$ with degrees of freedom = 81) and >99.9% agreement if the cell is assigned singlet by both technologies (only five cells differ; χ^2 test for independence gives a $P < 2.2 \times 10^{-16}$ and $\chi^2 = 146{,}879$ with degrees of freedom = 49). Next, we clustered the cells for each of the two datasets separately (Methods) and used Azimuth[18] to automatically label cell types in each (Methods). In both sequencing datasets, we identify the major cell types expected for PBMCs, with the expected cell-type markers (Extended Data Fig. 7a,b), and cells are comparably well-mixed among individuals (Fig. 3b), with low adjusted mutual information (AMI) between cell type and individual in both Ultima (0.026) and Illumina (0.025) (AMI = 0 corresponds to no relation between individual and cell type; AMI = 1 corresponds to the case of perfect agreement between the two labelings). The two sequencing datasets also had high agreement on proportions of each cell type from each individual, both for the main cell-type categories (Fig. 3c; 95% agreement in cell-type labels between Ultima and Illumina; χ^2 test for independence gives a $P < 2.2 \times 10^{-16}$ and $\chi^2 = 123{,}891$ with degrees of freedom = 49, AMI = 0.88) and for finer cell subsets, such as subclusters of T cells (Fig. 3d). They further agreed on differential expression between cell types (Fig. 3e; $r = 0.93$–0.95), such that 67.9% of genes that are significantly DE (FDR < 0.05 with Presto[19]; Methods) in one cell type in one of the two datasets are significant in both for that cell type. We found similar results with 5′ and 3′ PBMC datasets from a single individual (Extended Data Fig. 8). Moreover, the two PBMC mixture datasets were co-embedded well together into a joint two-dimensional space using uniform manifold approximation and projection (UMAP) after regressing out dataset of origin (Methods), with good mixing between datasets (AMI = 0.00068 between the joint clustering and dataset of origin), and good separation of cell types (Fig. 3f). Thus, data generated by the two sequencing technologies are compatible and can be combined easily in a single analysis.

For B and T cells, where we had clonotype assignment only by Illumina sequencing of TCRs and BCRs (Methods), we found good concordance between Ultima and Illumina cell-type assignments. Most T cells called by either method had TCR sequences (76% in both Ultima and Illumina) with only a very small percent of cells of other cell types having a TCR sequence (3.7% in Ultima and 3.5% Illumina), with similar results for B cells and BCR sequences (Extended Data Fig. 7c; 93% of B cells in both Ultima and Illumina were assigned a BCR clonotype while only 0.72% of non-B cells in Ultima and 0.73% of non-B cells in Illumina were assigned a BCR clonotype). The distribution of T cell subsets to top TCR clonotypes for each individual was also largely concordant between Ultima and Illumina sequencing (Extended Data Fig. 7d), with small differences in cell-type labeling. CD8 T effector memory (TEM) cells were by far the most likely to be expanded, as expected[20]. Thus, Ultima sequencing for scRNA-seq can be combined with Illumina

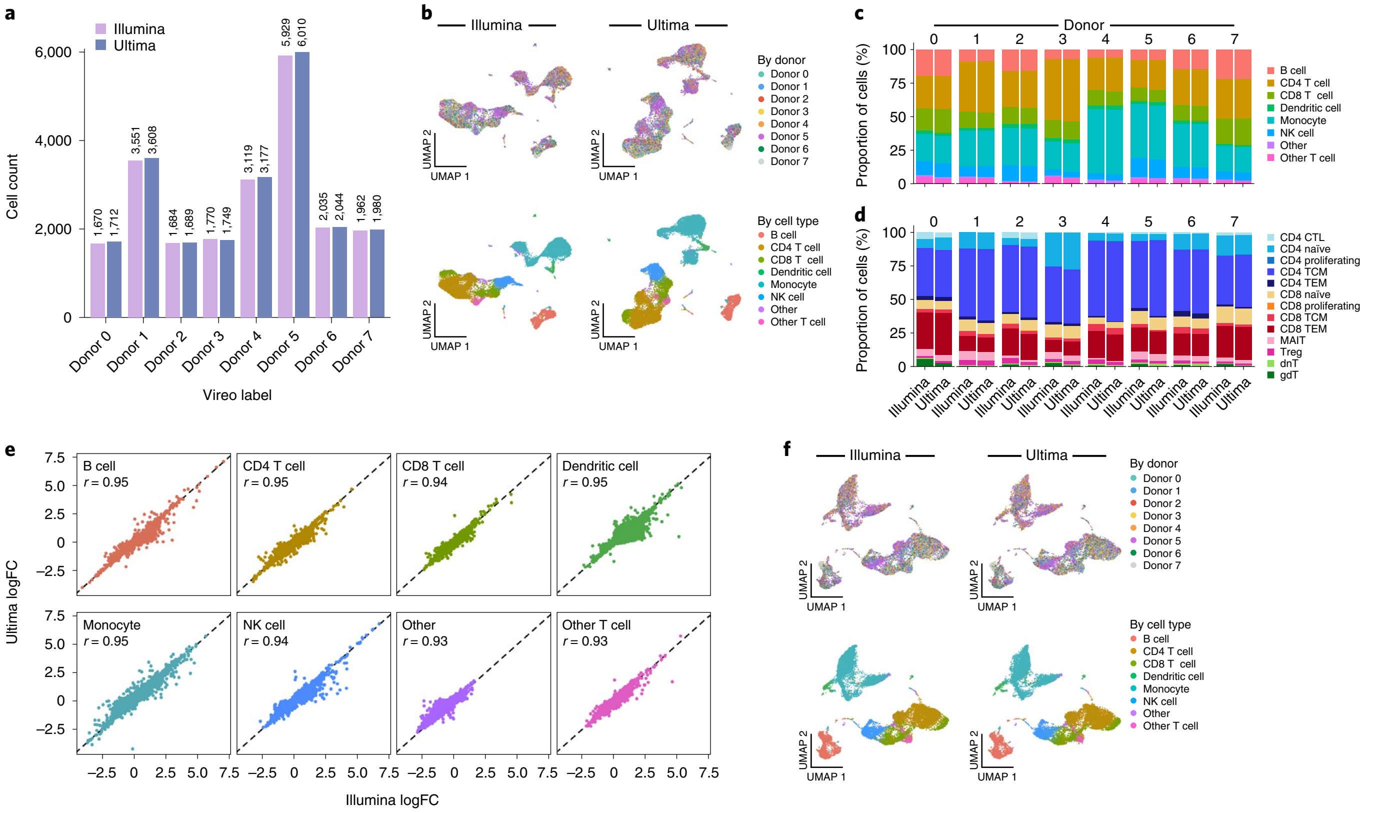


Fig. 3 | Cell-type identification and characterization of a mixture of PBMCs. a, Number of cells assigned to each donor by Vireo. Donors were renamed to match between Ultima and Illumina. **b**, UMAP plots for Ultima (right) and Illumina (left) colored by donor (top) and cell type (bottom). **c**, Bar plots of the proportion of each Azimuth-defined cell type in each donor for Ultima and Illumina. **d**, Bar plots of the proportion of each Azimuth-defined T cell subtype in each donor for Ultima and Illumina. Strong agreement can be seen. **e**, Scatter plots of logFC from performing DE between cell-type clusters with Presto. **f**, Joint UMAP of Ultima and Illumina data colored as in **b**. We did not sample the exact same number of reads from Illumina or Ultima data since they have approximately the same number of total UMIs. NK, natural killer cells; CTL, cytotoxic T cells; TCM, central memory T cells; MAIT, mucosal-associated invariant T cells; Treg, regulatory T cells; dnT, double negative T cells; gdT, gamma delta T cells.

sequencing of TCR and BCR genes to generate comparable results to those found with only Illumina sequencing.

To explore finer signals, we compared the two datasets for continuous cell states—such as activation status or the cell cycle—recovered by unsupervised non-negative matrix factorization (NMF). Each NMF factor can reflect a gene program, defined by a non-negative score for each gene (referred to as gene loadings) and a non-negative score for each cell (referred to as cell loadings). Because NMF runs are not identical even when re-run on the same data, to compare NMF models from Ultima and Illumina data, we fit NMF on Ultima data, Illumina data and a null of randomly permuted Illumina expression values (Methods) and then measured how well cell or gene loadings fit each dataset. Cell loadings from the model learned on Ultima data fit the Illumina data almost as well as cell loadings from the Illumina-learned model and vice versa, while loadings from the permuted (null) dataset led to a much poorer fit (Extended Data Fig. 9a). For gene loadings, there was lower performance when fitting data from one sequencing technologies with loadings from a model learned on the data from the other technologies, each to a comparable extent, and both far better than random permutations (Extended Data Fig. 9b). Consensus NMF (cNMF)[21] (Methods), which reduces variability owing to random sampling between NMF runs, showed high correlations of cell (Extended Data Fig. 9c) or gene (Extended Data Fig. 9d) loadings between models learned on different runs. The correspondence was comparable to that observed between two independent cNMF runs on the same dataset (Extended Data Fig. 9e,f), and lower than when comparing a single run to itself (Extended Data Fig. 9g,h), as expected. It was also much stronger than comparing cNMF models of two different biological systems (5′ PBMC mixture data and Perturb-Seq; see below for details of this experiment) (Extended Data Fig. 9i,j). Notably, the same cell subsets score highly for Ultima (Extended Data Fig. 9k) and Illumina (Extended Data Fig. 9l) data-derived programs on a joint UMAP embedding. For example, factor 13 in Illumina and factor 1 in Ultima scored in the same cells (Extended Data Fig. 9k,l) and were correspondingly highly correlated on both cell (Extended Data Fig. 9c) and gene (Extended Data Fig. 9d) loadings, indicating that they correspond to the same program. Moreover, other factors that differed between Ultima and Illumina were highly related—for example, factor 5 in the Ultima dataset was roughly decomposed into factors 5 and 11 in the Illumina dataset. Overall, we conclude that there is a high correspondence between cell states in Ultima and Illumina data.

As a final test, we evaluated performance with a Perturb-Seq screen, where heavy sequencing requirements are particularly limiting for scale[5,6], and used a design that also tested for CITE-Seq[22] and Cell Hashing[23] performance. Specifically, we used a library from a pilot screen of an ongoing genome-wide Perturb-Seq study (PIT, KGS, CJF and AR, unpublished results) to identify regulators of MHC Class I in melanoma A375 cells (Fig. 4a). In this pilot, we introduced 6,127 guides targeting 1,902 transcription factors and chromatin modifiers (Supplementary Table 4) along with both intergenic and non-targeting control guides, enriched for cells with low human leukocyte antigen (HLA) levels, and followed by scRNA-seq

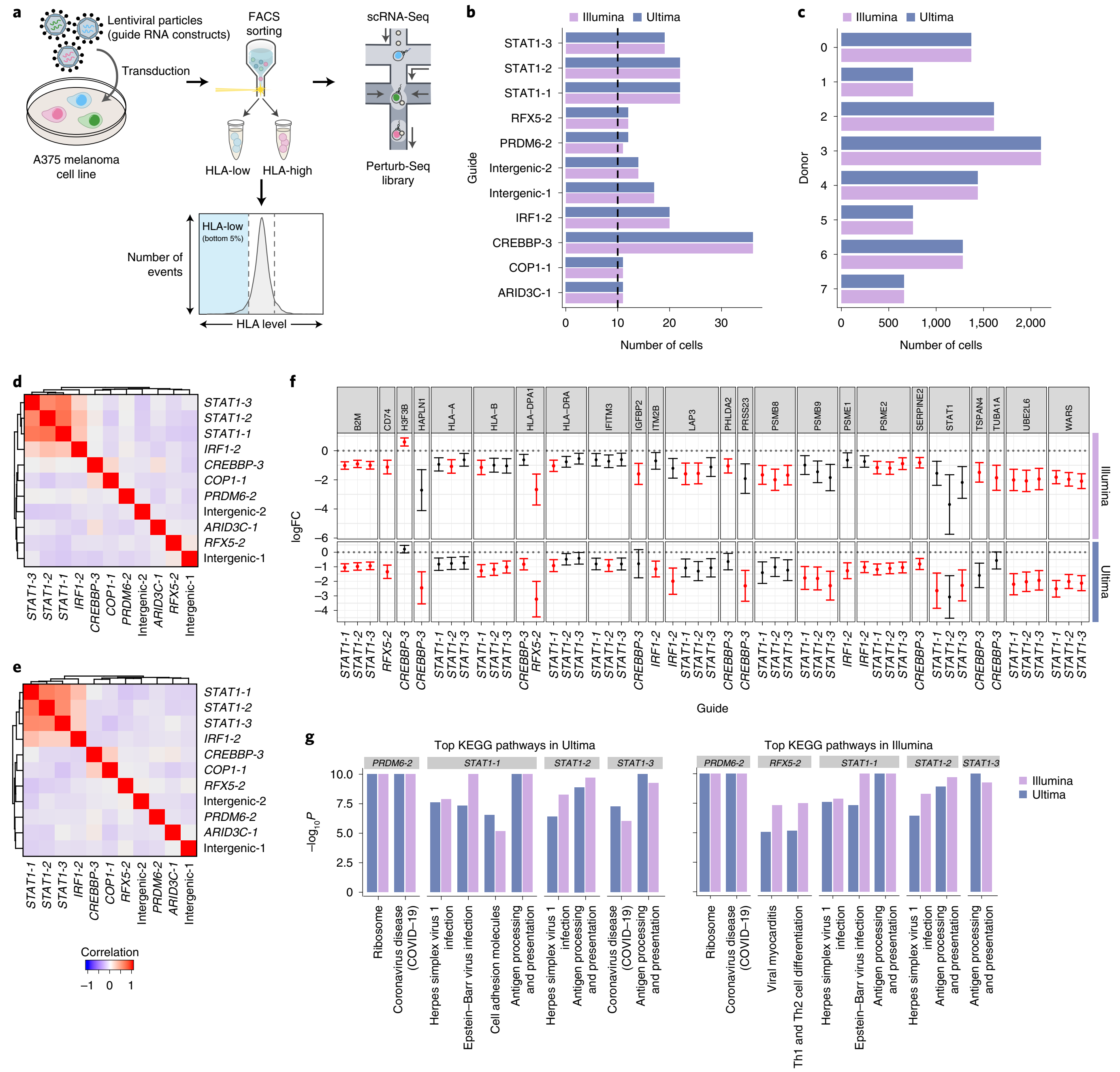


Fig. 4 | Perturb-Seq. a, Perturb-Seq to find regulators of MHC Class I in melanoma. We transduced A375 melanoma cells with a genome-wide library, and cells with low HLA expression were enriched by flow cytometry before scRNA-seq. **b**, Number of cells with each perturbation, only plotting those with >10 cells, excluding non-targeting and background guides. **c**, Number of cells with each Cell Hashing label. **d**,**e**, Guide similarity heat maps in Ultima (**d**) and Illumina (**e**). Effects of each guide on each gene in the Illumina data were calculated with an elastic-net-based approach as in MIMOSCA. The matrix of guides by genes with (uncorrected) $P < 0.05$ was extracted; the correlation between guides was calculated and plotted as a heat map. **f**, We extracted all gene–guide pairs from our DE analysis with FDR < 0.05 in either Illumina or Ultima. For each of these guide–gene pairs, we plotted the logFC on the *y*-axis (with 95% confidence intervals) and guide on the *x*-axis, with each box being a different gene, for both Illumina and Ultima. We included all guides that targeted a gene with any significantly different guides. The dots are colored red if significant (FDR < 0.05) and black if not. **g**, KEGG enrichment analysis for each guide. The ten pathways with smallest *P* values in both Illumina (left) and Ultima (right) are plotted as their $-\log_{10}P$ in both Illumina and Ultima. All pathways were significant in both Illumina and Ultima at FDR of 0.05.

of 20,000 cells that included CITE-seq[22] and Cell Hashing[23,24] (Methods). We sequenced the resulting scRNA-seq libraries with Illumina and Ultima, but the targeted PCR amplification ('dial-out') libraries used for guide detection, CITE-seq and Cell Hashing were only sequenced with Illumina (Extended Data Fig. 10a–c) because the read length was not sufficient for guide detection and the others were not attempted. Initial pre-processing of the Perturb-Seq scRNA-seq data showed similar performance for Ultima and Illumina, after sampling reads to have the same number of UMIs in each dataset (Extended Data Fig. 6a,b and Supplementary Table 1), as before, as well as in terms of cell assignment to guides (Fig. 4b and Extended Data Fig. 10c), Cell Hashing barcodes (Fig. 4c) and cell clustering and marker gene expression (Extended Data Fig. 10d–i).

Importantly, the Ultima and Illumina datasets identified similar relationships between perturbations and similar regulatory effects. For this analysis, we included the 335 cells in Illumina and 336 cells in Ultima, coming from 11 perturbations and 10 control guides in this pilot screen that were assigned to a single perturbation that had more than 10 assigned cells (the same perturbations were found by Illumina and Ultima). We then fit a regularized linear model (with elastic net, similar to previous studies[6,24]; Methods) of the mean impact of each perturbation on each gene, selected genes with nominal $P < 0.05$ using a permutation-based approach (Methods) and clustered the guides by these regulatory profiles. Analyses that were based on Ultima (Fig. 4d) and Illumina (Fig. 4e) yielded very similar guide relationships, both between multiple guides to the same gene (for example, *STAT1* guides) and between guides to different, functionally related genes (for example, *STAT1* and *IRF1* or *COP1_1* and *CREBBP_3*). Moreover, there was very high agreement in the effects on individual genes in both datasets, when comparing DE genes between each guide and an intergenic control (intergenic_1) in each dataset, in both significance and effect size (Fig. 4f and Extended Data Fig. 10j,k). Many such gene–guide pairs were significant in both the Ultima and Illumina datasets (Fig. 4f), and those significant only in one had highly similar effect sizes, showing consistent signal. Moreover, the KEGG pathways enriched with DE genes between each guide and an intergenic control were highly similar between the datasets (Fig. 4g).

Finally, we leveraged the Perturb-Seq data to assess any impact that our use of shorter Illumina Read 2 in the PBMC data may have had on our results. In the Perturb-Seq data, the length of Illumina Read 2 (96 bases) is longer than in the Illumina PBMC data (~55 bases). Reanalyzing the Illumina Perturb-Seq data after trimming Read 2 to 55 bases showed only slight reductions in the number of genes and UMIs per cell, and overall very similar results to the full-length Perturb-Seq data (Supplementary Fig. 3).

In conclusion, the two sequencing platforms generally perform similarly for scRNA-seq, across two main protocols for droplet-based scRNA-seq (3′ and 5′), two different sample types (primary cells and a cell line) and multiple experimental designs (simplex and multiplex, Perturb-Seq, CITE-Seq and Cell Hashing). One key explanation for the minor differences we observed is the position of reads relative to annotated gene boundaries (Extended Data Fig. 4d–g), as a consequence of Ultima single-end reads being closer to gene ends. Additionally, we currently recommend 5′ over 3′ libraries, given the small penalty in lost reads in 3′ libraries (Fig. 2a) owing to lower sequencing quality adjacent to the poly(T) sequence (Extended Data Fig. 2). A similar comparison of BGI MGISEQ-2000 and Illumina sequencing of 10x Chromium scRNA-seq libraries also found highly comparable results[25].

Lower-cost Ultima sequencing should make it possible to sequence more reads, cells and/or samples in the context of large-scale tissue atlasing projects, such as the Human Cell Atlas[1], the BRAIN Initiative[26], the Cancer Moonshot Human Tumor Atlas Network[2] as well as perturbation screens[5,6]. It should also be possible to design droplet-based scRNA-seq reagents, and methods for other large-scale single-cell and spatial genomics[27–31] customized to Ultima sequencing to directly generate libraries and eliminate the need for library conversion (Fig. 1b). With appropriate adaptations for Ultima sequencing, single-cell ATAC-seq should be possible even with the current sequencing lengths. Additionally, with longer Ultima reads or different construct designs, sequencing of other library types including TCR/BCR and targeted PCR amplification can be enabled. Finally, such reduced sequencing costs could open the way to use scRNA-seq in clinical applications, including diagnostics (as in next-generation blood tests, 'CBC2.0'[1]) or for therapeutics screens of small molecules, antibodies or cell therapies, impacting both basic biological discoveries and their clinical translation.

Online content

Any methods, additional references, Nature Research 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/s41587-022-01452-6.

Received: 7 February 2022; Accepted: 1 August 2022;
Published online: 15 September 2022

References

1. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
2. Rozenblatt-Rosen, O. et al. The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181**, 236–249 (2020).
3. Smillie, C. S. et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714–730 (2019).
4. Delorey, T. M. et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595**, 107–113 (2021).
5. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).
6. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
7. Srivatsan, S. R. et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45–51 (2020).
8. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
9. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
10. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
11. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
12. Almogy, G. et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.29.493900 (2022).
13. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
14. Mereu, E. et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).
15. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
16. Wallrapp, A. et al. The neuropeptide NMU amplifies ILC2-driven allergic lung inflammation. *Nature* **549**, 351–356 (2017).
17. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol* **20**, 273 (2019).
18. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
19. Korsunsky, I., Nathan, A., Millard, N. & Raychaudhuri, S. Presto scales Wilcoxon and auROC analyses to millions of observations. Preprint at *bioRxiv* https://doi.org/10.1101/653253 (2019).
20. Wang, P. et al. Single-cell transcriptome and TCR profiling reveal activated and expanded T cell populations in Parkinson's disease. *Cell Discov.* **7**, 52 (2021).
21. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, e43803 (2019).
22. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
23. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
24. Frangieh, C. J. et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat. Genet.* **53**, 332–341 (2021).
25. Senabouth, A. et al. Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. *NAR Genom. Bioinform.* **2**, lqaa034 (2020).

26. Network, B. I. C. C. A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).
27. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
28. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
29. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
30. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
31. Stahl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

PBMC library preparation. All biospecimens were collected with informed consent by a commercial vendor. Use of all de-identified biospecimens for sequencing at the Broad Institute was further approved by the Broad's Office of Research Subject Protection (ORSP), which determined that the research did not involve human subjects according to U.S. federal regulations (45CFR46.102f)—determination ORSP-3635. This study complied with all relevant ethical regulations.

We purchased nine cryopreserved human PMBC samples (AllCells). We thawed PBMC vials in a 37 °C water bath for ~2 min. A quick counting revealed high viability (>90%) in all samples. We added 1 ml RPMI1640 (Thermo Fisher Scientific, 11875093) with 10% fetal bovine serum (Thermo Fisher Scientific, 16140-071), transferred the cells to 15-ml conical tubes and then added another 9 ml of this medium slowly dropwise. We spun down the samples for 10 min at 300 *g* at room temperature. After supernatant removal, we flicked each tube to dislodge the pellet and carefully added 10 ml medium dropwise, followed by another spin under the same conditions. After supernatant removal, we flicked the tubes to dislodge pellets, re-suspended the cells in 500 μl phosphate-buffered saline 0.4% bovine serum albumin (Sigma, B8667) and transferred to 1.5-ml tubes. We then spun down the samples for 5 min at 300 *g* at room temperature. We washed the cells with 500 μl phosphate-buffered saline 0.4% bovine serum albumin an additional two times and filtered through 40-μm cell strainers (Falcon, 352340). We counted cells with a TC-20 cell counter (Bio-Rad) and observed high viability (>90%) for all samples.

For one sample with matched 5′ and 3′ libraries, we loaded one channel of 10x 3′ V3.1 (10x Genomics, 1000128) onto a G chip (10x Genomics, 1000127) and one channel of 5′ V2 (10x Genomics, 1000265) onto a K chip (10x Genomics, 1000286), respectively, aiming to recover 7,000 cells from each. With the other eight samples, we pooled them equally and loaded onto one channel of 10x 5′ V2 assay aiming to recover a total of 24,000 cells. We generated the 10x 3′ and 5′ scRNA-seq libraries following the manufacturer's protocols, as well as the TCR and BCR libraries from the 5′ assay with the 10x Chromium Single Cell Human TCR (10x Genomics, 1000252) and BCR Amplification kits (10x Genomics, 1000253), respectively. We performed each experiment once ($n = 1$ biological replicate). The two 5′ experiments are biological replicates for each other in some, though not all, ways.

Perturb-Seq screen. To generate a large Perturb-Seq library targeting all transcription factors and chromatin regulators, we designed a 5,706-guide library targeting 1,902 genes identified as either transcription factors or chromatin regulators with three guide RNAs per gene taking sequences from the Broad Institute Genetic Perturbation Platform Web short guide RNA Designer (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design)[32]. We included two different types of control guide RNAs either guides that cut in a non-gene region (intergenic control) or guides that do not bind any genomic region (non-targeting control) each at 5% of the total guide count. The pooled CRISPR library was cloned as previously described in the CROPseq mKate2 vector backbone[24]. We transduced Cas9-expressing A375 cells (ATCC CRL-1619) with a transcription factor and chromatin regulator guide RNA library (Supplementary Table 4). We selected perturbed cells for 3 days using 2 μg ml^{-1} puromycin. After selection, we treated cells with 2 ng ml^{-1} recombinant interferon-γ for 16 hours. Following interferon-γ treatment, we stained cells with CITE-seq and hashing antibodies as previously described[24] (Supplementary Table 4) along with a 1:500 dilution of fluorescent HLA antibody (BioLegend 311415). The 5% lowest expressing HLA cells were selected via fluorescence activated cell sorting (FACS) (with FlowJo v.10.5.3 software), and 40,000 cells were loaded onto one 10x 3′ V3 Chromium channel. We performed this experiment once ($n = 1$ biological replicate).

10x Chromium Illumina sequencing. We sequenced the PBMC libraries on Illumina NextSeq 500 flowcells with at least 20,000 reads per cell for scRNA-seq libraries and 5,000 reads per cell for TCR and BCR libraries. For 3′ libraries, we sequenced 28 bases for Read 1, 55 bases for Read 2 and 8 bases for Index 1. For 5′ libraries, we sequenced 26 bases for Read 1, 45 bases for Read 2 and 10 bases each for Index 1 and Index 2. For TCR and BCR libraries, we sequenced 26 bases for Read 1, 90 bases for Read 2 and 10 bases each for Index 1 and Index 2. We sequenced the Perturb-Seq library on Illumina HiSeq X flowcells with 14,000 reads per cell for scRNA-seq libraries, 5,000 reads per cell for CITE-seq libraries and 1,000 reads per cell for Hashing libraries. For Perturb-Seq libraries, we sequenced 28 bases for Read 1, 96 bases for Read 2 and 8 bases for Index 1.

10x Chromium library conversion and Ultima sequencing. Our 10x Chromium libraries were converted using a library conversion PCR work flow (Fig. 1b) to enable sequencing on the Ultima platform. In brief, library concentration was verified using Qubit (Thermo Fisher Scientific), with conversion PCR library input being 7 ng. Conversion was facilitated through two overhang primers. Primer 1 anneals in the Read 1 region of the 10x library and contains a Ultima-specific overhang (index adapter sequence, IA). It contains primer binding sites for clonal amplification and sequencing. The index adapter sequence also includes an in-line Ultima-specific sample barcode (PS-SBC). Primer 2 anneals in the Read 2 region and contains a Ultima-specific overhang (unique bead adapter sequence, UBA) necessary for clonal amplification (Supplementary Table 5). We used the Q5 Hot Start High-Fidelity kit (New England Biolabs) with ten PCR cycles for amplification, followed by DNA Clean Concentrator tubes (Zymo Research) as per the manufacturer's instructions for PCR product purification, and quantification of the purified library.

After pooling, we seeded libraries, clonally amplified them on sequencing beads using a high-scale emulsion amplification tool and sequenced them on a prototype Ultima Sequencer[12].

For sequencing of 10x Chromium 3′ libraries, we used a modified sequencing protocol that accommodates the high consumption of dT nucleotides in the poly(dT) stretch of the cDNA. Specifically, we included additional T injections when sequencing cycles 28 to 32, which were predicted to include the poly(dT) stretch: $(TGCA)_{27}$ $(T_{10}GCA)_5$ $(TGCA)_{60}$.

Initial Ultima read processing. To enable standard scRNA-seq analysis of single-end reads, we first converted Ultima data to create paired-end data (Fig. 1d). To this end, we removed conversion adapters, and quality trimmed reads using Cutadapt v.2.10 (ref. [33]), using a threshold of 30. We discarded reads not containing at least eight Ts for the expected poly(T) for 3′ libraries or the template switch oligo (TSO) for 5′ libraries. We split reads into two sequences: one containing the CBC and UMI and the other containing the reverse complement of the cDNA using Cutadapt and SeqKit v.0.15.0 (ref. [34]) (see Supplementary Table 6 for commands).

We removed reads with a cDNA sequence <50 bases. For 3′ libraries, we clipped the first five bases after the poly(T) and masked the last three bases of the UMI. For 5′ libraries, we clipped the first three bases after the TSO. We trimmed cDNA sequences to 90 bases.

Initial single-end Illumina read processing. For the single-end Illumina reads, we used a similar pipeline to that used for Ultima data. After de-multiplexing the FASTQ file, we quality trimmed reads using Cutadapt v.2.10, using a threshold of 30, and discarded reads not containing the TSO. We then split reads into two sequences: one containing the CBC and UMI and the other containing the reverse complement of the cDNA using Cutadapt and SeqKit v.0.15.0 (see Supplementary Table 6 for commands). We removed reads with a cDNA sequence <50 bases. We clipped the first three bases after the TSO. We trimmed cDNA sequences to 90 bases.

Extracting expression information from FASTQ files. We used Cell Ranger v.5 (ref. [15]) to pre-process data for both Illumina and Ultima (for Ultima using simulated Read 1 and Read 2 as extracted above). For all datasets, we used the GRCh38 human reference from 10x Genomics unless otherwise stated and set '--expect-cells' to the expected number of cells (7,000 cells for the single sample 3′ and 5′ PBMC data, 24,000 cells for the 5′ mixture PBMC sample and 20,000 cells for the Perturb-Seq sample). To process 3′ Ultima data, unless otherwise stated, we modified the last three bases of the UMI using awk by replacing them with A's and setting the last three quality values to be equal to I.

For the 5′ mixture data, we de-multiplexed it by first calculating single-nucleotide polymorphism (SNP) coverage data for SNPs in the 1000 Genomes Project[35] with cellsnp-lite v1.2.0 (ref. [36]) (using '--minMAF 0.1 --minCOUNT 20') followed by Vireo v.0.5.5 (ref. [17]) to get labels for the sample of origin.

For Perturb-Seq data, we also passed Cell Ranger FASTQ files for targeted PCR amplification data (using the CRISPR Guide Capture keyword), hash tag oligo (HTO) data (using the Custom keyword) and antibody-derived tag (ADT) reads for CITE-seq (using the Antibody Capture keyword), as well as feature barcode information for each. We further processed HTO data with DemuxEM v0.1.7 (ref. [37]) to obtain sample labels.

Sampling reads. For each Ultima dataset, unless otherwise stated, we sampled it to have both the same number of reads and the same number of total UMI as the corresponding Illumina dataset. This was performed by sampling the FASTQ files with seqtk v.1.0 sample (https://github.com/lh3/seqtk) passing it the argument '-s 100', the FASTQ file to sample and the proportion to sample by. To sample the same number of UMIs for Ultima and Illumina, we first estimated the proportion of Ultima reads that needs to be sampled in order to match the number of UMIs in the Illumina data. We performed this as a two-step procedure to reduce the number of sampling steps required. Specifically, first we used DropletUtils v1.10.3 (ref. [38]) to sample Ultima data to different levels (in 5% increments) and calculated the total number of UMIs for each. We chose an initial estimate of the sampling proportion as the largest proportion that gave fewer UMIs in Ultima than were present in Illumina. To improve this estimate, we performed a refinement step where we follow the same procedure, except with 1% steps starting from the initial estimate and up to the initial estimate plus 5%. We chose the final proportion used for sampling from this range to be the largest proportion that gave fewer UMIs in Ultima than were present in Illumina. We did not sample the 5′ mixture Ultima data because it had roughly the same number of UMIs as the 5′ mixture Illumina data. After sampling, we processed data in a similar fashion to non-downsampled data (see above).

Extracting FASTQ quality control metrics. To extract base quality information from each FASTQ file, we randomly selected 1,000,000 reads with seqtk sample using the parameter '-s 100' to set a random seed. We then used the SeqIO.parse function from Biopython v.1.79 (ref. [39]) to read the FASTQ into Python v.3.7.7. We then extracted the quality information with the letter_annotations function and recorded the resulting information to a file with one line per read and one column for each base in that read. This was used for downstream visualizations.

To explore the effects of shortening UMIs on number of reads, we loaded the molecular information h5 file generated by Cell Ranger into R with DropletUtils and saved the resulting data frame. We then loaded this into Python, resulting in a table with one entry for each UMI counted by Cell Ranger, which included CBC, UMI and Gene. We shortened the UMI to different lengths and used the pandas[40] groupby function to count the number of UMIs that collapsed together after this shortening and the number of UMIs from different genes that collapsed together after this shortening.

To explore how the number of UMIs per cell or per gene affects the performance of UMI trimming, we calculated the number of UMIs per cell and per gene returned by Cell Ranger with both full-length (12 bases) and trimmed (9 bases) UMIs. We then looked at the ratio between the two (the number of UMIs returned with 9 base UMIs divided by the number of UMIs returned with 12 base UMIs) for each cell and each gene. A value of 1.0 corresponds to the trimming having no effect, and a value of 0.8 corresponding to an 80% reduction, etc.

Analysis of PBMC data. We loaded filtered PBMC count data from Cell Ranger (located in the outs/filtered_feature_bc_matrix subdirectory output) into R v.4.0.3 (https://www.r-project.org/) using Seurat v.4.0.0 (ref. [18]). To avoid biases introduced by using slightly different sets of cells, we used only the intersection of the sets of cells found in Ultima and Illumina for each analysis. For the 5′ mixture data, we also uploaded the labels from Vireo and removed doublets and unassigned cells. We processed data through the standard Seurat pipeline, as follows. We normalized data to log transcripts per million (TPM) with NormalizeData, found variable genes with FindVariableFeatures (using 2,000 variable genes), scaled data and regressed out the number of genes per cell with ScaleData and performed principal component analysis with RunPCA. We performed UMAP embedding and clustering using Seurat FindNeighbors followed by FindClusters with 20 principal components and otherwise default parameters (including FindClusters using Louvain clustering with a resolution of 0.8). Cell types were assigned using Azimuth v.0.3.2 (ref. [18]) with the built-in PBMC dataset. In particular, we assigned cell types at two different levels of granularity, denoted in the Azimuth labeling by l1 (general cell types) and l2 (refined cell types). AMI was calculated with the AMI function in the aricode package v.1.0.0 (https://github.com/jchiquet/aricode). For calculating joint embeddings, we combined and processed the two datasets through the same Seurat pipeline, except the sequencing technology used (Ultima or Illumina) was regressed out before principal component analysis with ScaleData. Presto v.1.0.0 (ref. [19]) was used to calculate DE genes between cell types with default parameters using a false discovery rate (FDR) cutoff of 0.05.

Analysis of method-specific biases. We calculated logFCs comparing expression levels between Ultima and Illumina by creating a pseudobulk profile for an entire dataset using count data, which was then normalized to TPM. We calculated logFC values by taking the difference between the log of the corresponding TPM values in Ultima and Illumina with a pseudocount of 10. All plots of logTPM have a pseudocount of 1 added.

For the sampling analysis to compare reads with each other from the same sequencing run, we used a modified version of the downsampleReads code from DropletUtils applied to the molecule_info.h5 file from Cell Ranger to split reads into two disjoint sets of equal size (up to rounding) and calculated the associated gene by cell UMI count matrix for each set of reads.

We performed DE analysis between Ultima and Illumina with Presto[19]. To explore GC and length biases in DE results, we used a Python script to process the GTF file used in Cell Ranger and collapse overlapping exons from the same gene into one genomic interval. This information was written to a BED file and used to calculate the total length of sequence in each gene covered by at least one exon (the gene length used in our analysis). We used the 'bedtools getfasta' command in BEDTools version 2.26.0 (ref. [41]) with arguments '-s -name -tab' to extract the associated sequences of each of the regions in the above BED file. We processed the resulting FASTA file with a Python script that extracted gene-level GC information.

Classification of gene expression and read assignment bias was performed as follows. For each protein-coding gene with a total count >100 in at least one of the platforms, we calculated read assignment rate (see below) differences for unfiltered (total reads, including reads from cell barcodes not in the filtered list) Illumina and downsampled Ultima reads and normalized the FCs of counts by the median difference. We partitioned the genes into three categories: (1) more than fivefold higher in Illumina ('Higher count in Illumina'), (2) more than fivefold higher in Ultima ('Higher count in Ultima') and (3) all remaining genes ('Similar counts') (Extended Data Fig. 4c). A read assignment rate higher in Illumina ('Higher read assignment in Illumina') was defined if the ratio of Cell Ranger gene-assigned reads out of the total reads mapped to the annotated gene body plus a flanking 100 bp was higher in Illumina (more than threefold higher ratio with $P < 0.01$, binomial test). The read assignment rate higher in Ultima ('Higher read assignment in Ultima') was defined analogously.

To explore 3′ and 5′ bias, the GTF file used by Cell Ranger was processed as described above to collapse overlapping exons together. We then used pysam v.0.15.3 (https://github.com/pysam-developers/pysam) to load each alignment (selecting 1% of alignments at random), excluding those without an assigned gene, CBC or UMI, as well as excluding multimappers. We then calculated the distance along the exonic regions of the gene (normalized by gene length) from the 5′ end of the gene to the 3′ and 5′ ends of the read using the overlapping exon representation we generated earlier. We then recorded this information for each alignment. For plotting, this was divided into bins of length 1% labeled from 0 to 100 (including the bottom of each bin but not the top—meaning that bin 100 was empty for 5′ reads and bin 0 was empty for 3′ reads) and normalized by the number of reads mapping to that gene to avoid highly expressed genes biasing the results.

To extract information about the number of reads falling into different categories (unmapped, ambiguous, etc.), we took the BAM file from Cell Ranger and applied FeatureCount v1.6.2 (ref. [42]) with settings '-t exon -g gene_name –fracOverlap 0'. For 3′ data, we set '-s 1' to denote sense reads, while for 5′ data, we used '-s 2' to denote antisense reads. In addition, for gene-level information about the number of reads in different categories, we reran FeatureCount once with the flag '-M' (for multimappers) and once with the flag '-O' (for ambiguous reads).

For IGV v.2.3.80 (ref. [43]) plots, we used SAMtools v.1.8 (ref. [44]) to extract a region around each gene of interest from the associated 10x BAM file. We then used grep to extract reads that were assigned to a given gene and those that were not.

Extension of the standard reference. We built a Nextflow-based[45] pipeline that takes in a pre-existing reference GTF file and RNA-seq BAM file (from paired-end or single-end RNA-seq) and outputs a new GTF file that extends the old one using RNA-seq data. The first step in the pipeline annotates reads in the BAM file overlapping genes in the standard reference using FeatureCount with the parameters '-t exon', '-R BAM' and '-g gene_id', as well as using the '-s 1' or '-s 2' flag depending on the strandedness of the RNA-seq data. For paired-end data, we also used the flag '-p'. We then used pysam-based start and end coordinates as the start and end coordinates of that read, with an extra field recording the assigned gene for the read. We excluded reads with large gaps (>10 bases labeled as N in the CIGAR string) and, for paired-end reads, only include properly paired reads. We then sorted the BED file with bedops[46], clustered the entries in this BED file using bedtools cluster with the '-s' flag and used bedtools groupby to merge BED entries from the same cluster and gene. We then sorted the resulting BED file with bedops again, use a Python script to turn the BED file into a GTF with one exon per entry in the BED file. We combined this GTF file with the GTF for the standard reference and sorted the results with BEDTools, yielding the extended references. These new GTFs were then used to generate references for Cell Ranger with cellranger mkfastq. We used this approach to create three references, one using published bulk data[13] (using the BAM file generated in that publication) and two using scRNA-seq Ultima PBMC data—one generated with 3′ data and the other with 5′ data. We then processed PBMC data with each of these references using cellranger count as described above and performed downstream analysis. We did not process the 3′ data with the 5′ reference or vice versa.

Processing TCR and BCR data. We processed FASTQ files for 10x Chromium TCR and BCR data with Cell Ranger v5 using the vdj command and the prebuilt 10x reference (refdata-cellranger-vdj-GRCh38-alts-ensembl-5.0.0). We then loaded data into the associated Seurat object with the djvdj package v.0.0.0.9000 (https://rdrr.io/github/rnabioco/djvdj/man/djvdj-package.html) where they were used for downstream visualizations.

Gene program analysis by NMF. We calculated all NMF models with RcppML v.0.3.7 (ref. [47]) using 15 factors and with the logTPM matrix as input including genes expressed in more than 1% of cells. NMF returns a cell-loading matrix, with one row per cell and one column per factor, and a gene-loading matrix, with one row per gene and one column per factor. To test how well NMF factors from one data type fit another, we split our data into a training set (with 5,000 cells) and a test set (all other cells) to avoid data overfitting when testing the gene-loading matrix. We then fit NMF models separately on the Ultima data, Illumina data and a permuted version of the Illumina data (where the values of each gene were randomly scrambled between cells) using the 5,000 cell training set. To test the accuracy of gene loadings of each NMF model for each data type, we used the project function from RcppML to generate a cell-loading matrix on the training data, and the mean squared error was measured. For cell loadings, we used a similar approach, except that testing was performed on the test dataset. In all cases, we repeated the analysis ten times with different random seeds to account for variability in NMF solutions.

For consensus NMF (cNMF)[21], a modification of NMF meant to improve robustness and reproducibility, we implemented an R version with two minor changes: (1) instead of performing NMF on the count data with each gene normalized by its standard deviation, we performed it on logTPM data, and (2) we performed each NMF round used by cNMF with the variable genes from our

Seurat analysis instead of the variable gene selection procedure in the cNMF package. We ran cNMF on all cells in each dataset, using 100 iterations of NMF and 15 factors. For Perturb-Seq-based cNMF, we used the project function in RcppML to project the gene loadings onto the PBMC mixture dataset. Similarity matrices were calculated with Pearson correlation.

Analysis of Perturb-Seq data. We processed Perturb-Seq data through a similar pipeline to the PBMC data (see above), except the HTO, ADT, and guide-count matrices were also uploaded as additional assays, while the DemuxEM labels for Hash ID and the Cell Ranger labels for guide assignment were added to the metadata. After initial processing with Seurat, we removed cells assigned to multiple hash tags or multiple guides, as were cells assigned to guides with 10 or fewer cells assigned to them.

We generated a guide similarity matrix with a slightly modified version of the MIMOSCA package[6]. We extracted a cell-by-gene logTPM expression matrix from the Seurat object, selecting the cells filtered as described above (one hash tag and guide assignment, with guides with more than 10 assigned cells) and genes that were expressed in >5% of cells. We also extracted a covariate matrix consisting of the scaled number of UMIs per cell, as well as one-hot encoded versions of the perturbation assignments, Hash ID assignments and cluster assignments, which are based on clustering all cells with Seurat's FindClusters function at a resolution of 0.2 with 20 principal components and otherwise default settings. We loaded data into Python using pandas, and an elastic-net model was fit modeling expression as a linear model of the covariates using sklearn.linear_model.ElasticNet[48] with parameters 'l1_ratio=0.5', 'alpha=0.0005' and 'max_iter=10000'. The coefficient matrix from this model was saved. We randomly permuted guide labels 100 times (while preserving the number of guides assigned to each hash barcode and vice versa) followed by the same elastic-net-based analysis. We loaded the resulting gene by covariate coefficient matrices into R and discarded columns that did not correspond to guide labels, along with columns corresponding to non-targeting control guides (those labeled as NO_SITE in our feature data) and the Background control guide. For each gene, we calculated a P value on the basis of the resulting matrices by scoring each gene by the maximum absolute value for that gene across all guides. We partitioned genes into 20 bins of equal size on the basis of average expression. We calculated a P value for each gene by comparing the score of that gene in the non-permuted data to the score of all genes in the same bin as it in all 100 permuted datasets. We retained all genes in the coefficient matrix with uncorrected $P<0.05$ and calculated the Pearson correlation between guides on the basis of this matrix.

Perturb-Seq differential expression analysis of genes regulated by each guide. We performed DE between the profiles of each guide and the profiles of the intergenic guide with Nebula v1.1.7 (ref. [49]), with the assigned Hash ID as the sample of origin, and the Intergenic_1 guide as reference. We calculated Benjamini–Hochberg FDR[50] on the resulting P values. We performed KEGG enrichment analysis with the KEGG ontology[51] using ClusterProfiler v.3.18.1 (ref. [52]) and performing GSEA with the fgsea package v.1.16.0 (ref. [53]). KEGG terms with fewer than 20 genes were filtered before visualization.

Visualization. Most of the visualization was performed using ggplot2 v.3.3.3 (ref. [54]) and cowplot v.1.1.1 (https://github.com/wilkelab/cowplot) packages in R. The major exception to this was the heat maps, which were produced with the ComplexHeatmap package v.2.6.2 (ref. [55]) and NMF v.0.30.1 package (ref. [56]), and histograms that were produced with the base R hist function.

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

Data availability

RNA-seq data generated in this project are available from the Gene Expression Omnibus under accession GSE197452 and the Single Cell Portal under accession SCP1759. Figures 2–4 have underlying data available in source files. For the reference human genome, we used: https://cf.10xgenomics.com/supp/cell-exp/refdata-cellranger-GRCh38-1.2.0.tar.gz. For Azimuth[18], we used the built-in PBMC reference: https://azimuth.hubmapconsortium.org/references/#Human%20-%20PBMC. Source data are provided with this paper.

Code availability

The code used for processing and analyzing our sequence data is freely available on Github (https://github.com/seanken/CompareSequence).

References

32. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
33. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
34. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
35. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
36. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* **37**, 4569–4571 (2021).
37. Gaublomme, J. T. et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat. Commun.* **10**, 2907 (2019).
38. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
39. Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
40. McKinney, W. Data structures for statistical computing in Python. In *Proc. 9th Python in Science Conference*. (eds van der Walt, S. & Millman, J.) 56–61 (SciPy, 2010).
41. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
42. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
43. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
45. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
46. Neph, S. et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
47. DeBruine, Z.J., Melcher, K. & Triche, T.J. Fast and robust non-negative matrix factorization for single-cell experiments. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.01.458620 (2021).
48. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
49. He, L. et al. NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun. Biol.* **4**, 629 (2021).
50. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
51. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
52. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
53. Korotkevich, G. et al. Fast gene set enrichment analysis. Preprint at *bioRxiv* https://doi.org/10.1101/060012 (2021).
54. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis (Use R!)* 2nd Edn (Springer, 2009).
55. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
56. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinf.* **11**, 367 (2010).
57. Tukey, J.W. *Exploratory Data Analysis* (Addison-Wesley, 1977).

Acknowledgements

We thank E. Torlai Triglia and O. Ashenberg for advice on TCR/BCR analysis, D. Phillips for BL2+ room training, the Broad Genomics Platform for Illumina sequencing and A. Hupalowska and L. Gaffney for help with illustrations and figures. This work was supported by the Klarman Cell Observatory, Broad Institute of MIT and Harvard (A.R., J.Z.L.), Howard Hughes Medical Institute (A.R.), Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard (J.Z.L.), National Human Genome Research Institute (NHGRI) grant (5RM1HG006193-09) (K.G.S.), NIH F32 Ruth L. Kirschstein Postdoctoral Fellowship (5F32AI138458) (P.I.T.) and NHGRI grants R44HG010558 and R44HG011060 (Ultima Genomics).

Author contributions

J.Z.L., A.R., G.A. and O.R.R. conceptualized the approaches used. X.A. performed PBMC experimentation and Illumina sequencing. P.I.T., K.G.S., C.J.F. and A.R. contributed the Perturb-Seq design, experimentation and Illumina sequencing. F.O., N.I. and O.B. performed the Ultima sequencing. S.K.S. and G.L.Y., with guidance from A.R., performed the computational analysis. J.Z.L., D.L. and A.R. supervised the study. S.K.S., J.Z.L., D.L., G.L.Y., P.I.T., K.G.S. and A.R. wrote the original draft of the paper, and all authors reviewed and edited it.

Competing interests

A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas and until July 31, 2020 was a member of the scientific advisory boards of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov.

From August 1, 2020, A.R. is an employee of Genentech, a member of the Roche Group and has equity in Roche. From November 16, 2020, K.G.S. is an employee of Genentech. From March 22, 2021, P.I.T. is an employee of Genentech. From October 19, 2020, O.R.R. is an employee of Genentech and has equity in Roche. O.R.R. is a co-inventor on patent applications filed by the Broad Institute for inventions related to single-cell genomics. She has given numerous lectures on the subject of single-cell genomics to a wide variety of audiences and, in some cases, has received remuneration to cover time and costs. G.L.Y., F.O, N.I., O.B., G.A. and D.L. are employees and shareholders of Ultima Genomics. The other authors declare that they have no competing interests.

Additional information

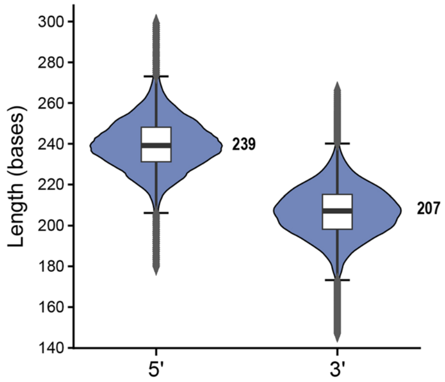
Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01452-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01452-6.

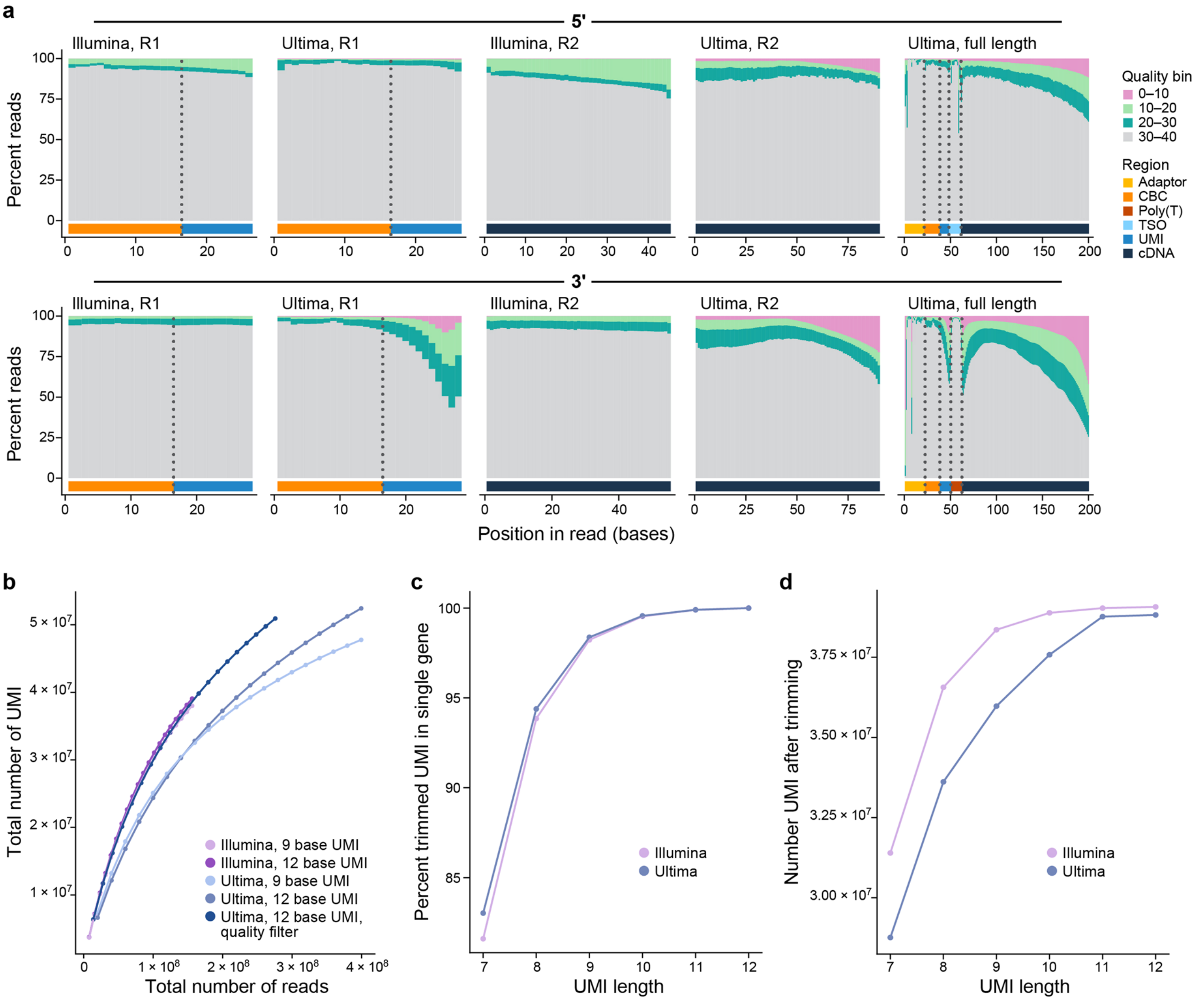
Correspondence and requests for materials should be addressed to Joshua Z. Levin.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

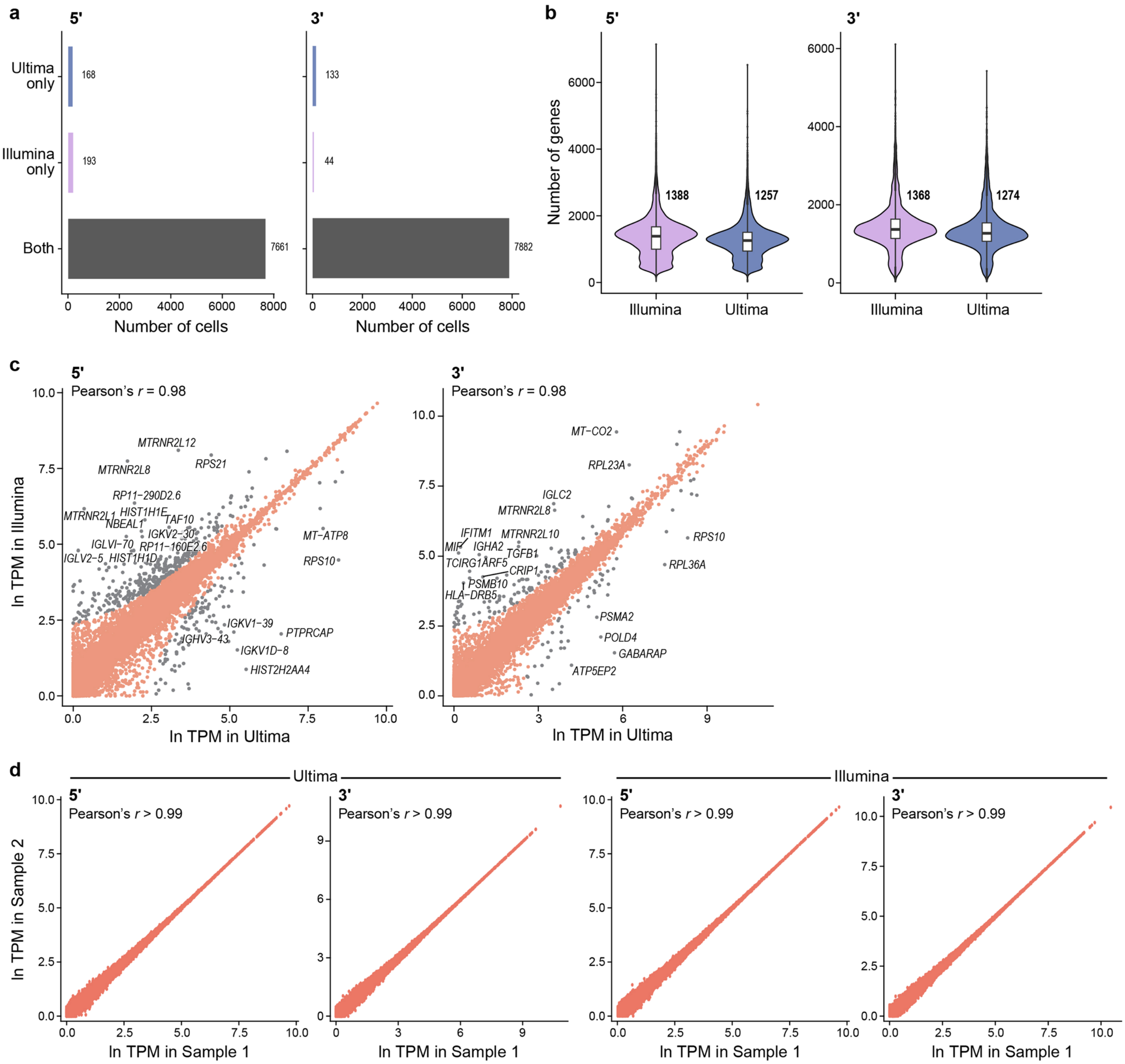
Reprints and permissions information is available at www.nature.com/reprints.



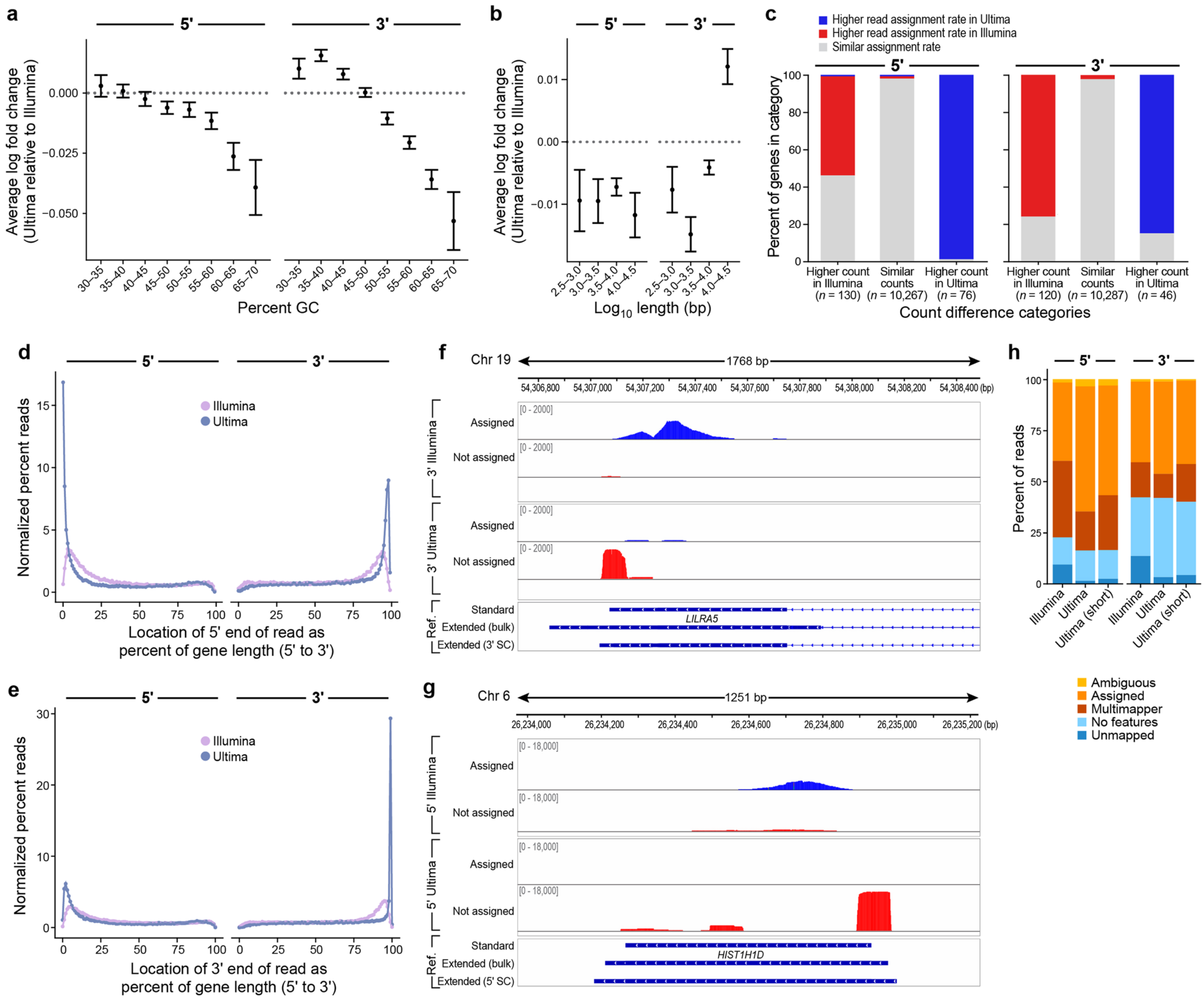
Extended Data Fig. 1 | Ultima read length. Violin plot of read length distribution of unfiltered reads from the matched 5′ and 3′ PBMC libraries. Boxplots are overlaid on the violin plot, displaying the 25% quantile, the 75% quantile, and with the median (also written as a number) marked in between. Whiskers extend from the edge of the box by 1.5x Inter Quartile Range (IQR). For visualization purposes, points beyond 3x IQR ('far out' outliers[57]) were excluded from the plot (1.6% removed for 5′; 1.9% removed for 3′), but not from the read processing pipeline.



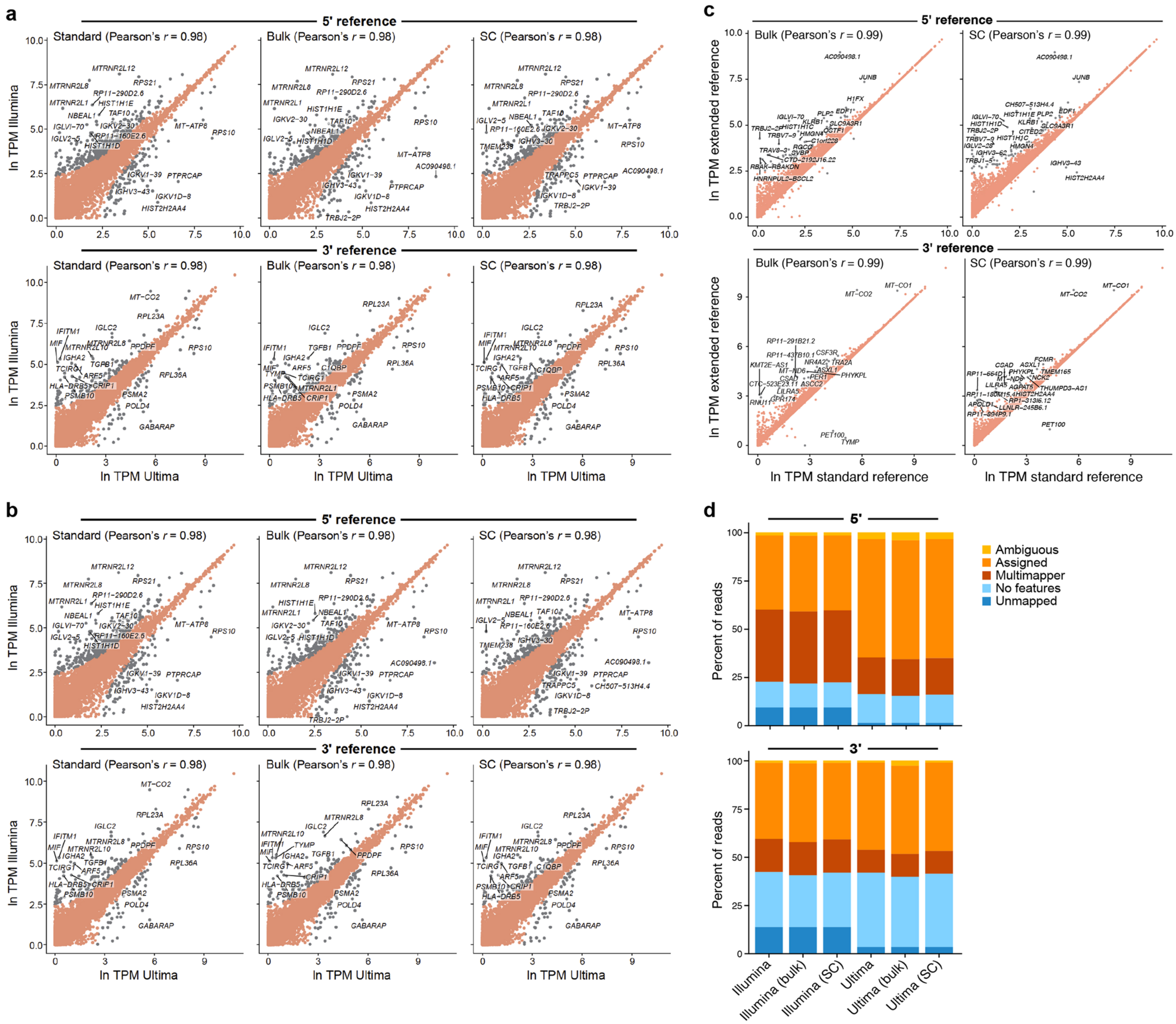
Extended Data Fig. 2 | Sequence read quality for matched 5′ and 3′ PBMC libraries. (**a**) Bar plot of the percent of reads at different quality levels for each sequencing platform and library type along the length of the read, for either Read 1, or Read 2 (or, for Ultima, the subsection of the read corresponding to Read 2), or the full read (up to 200 bases, Ultima only). Annotation of the reads is shown at the bottom. Information to guide UMI trimming from the UMI 3′ end for 3′ libraries (**b-d**). (**b**) Rarefaction curves at different UMI lengths and with and without filtering low quality reads. (**c**) At different UMI lengths, UMI/CBC pairs that have all Read 2s in the same gene. Lower levels suggest UMIs are too short. (**d**) Number of UMIs at different UMI lengths – a measure of similar UMIs collapsing as length decreases.



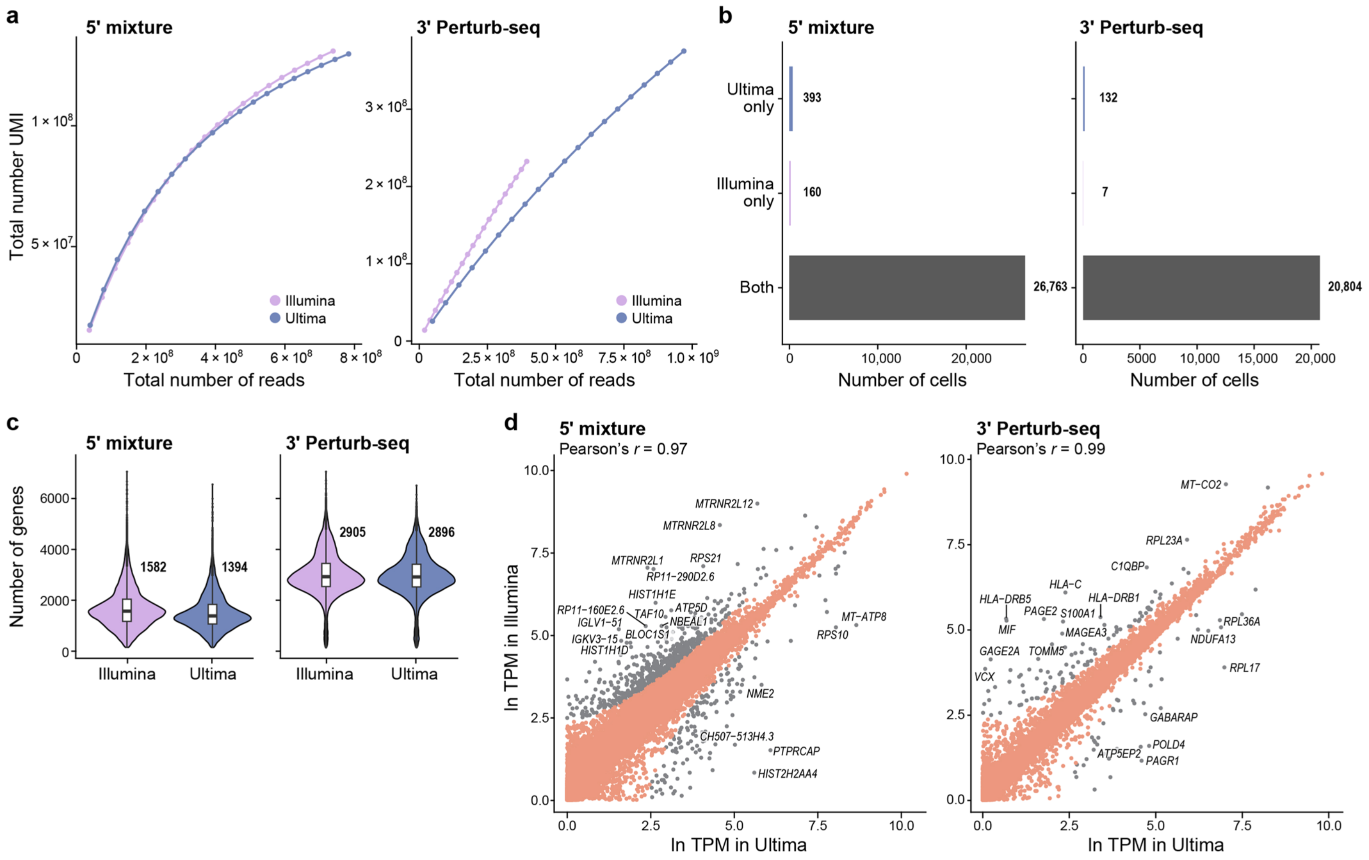
Extended Data Fig. 3 | Quality metrics for matched 5′ and 3′ PBMC libraries sampling reads to the same number of UMIs. (**a**) Number of cells identified by Cell Ranger only in Ultima, only in Illumina, or both. (**b**) Distribution of the number of genes per cell. (**c**) Scatter plots comparing gene expression in Ultima and Illumina sequencing. (**d**) Scatterplots comparing reads sampled from the same sequencing run. For (**c**) and (**d**), one point for each gene as in Fig. 2e. For all 3′ libraries, the last three UMI bases were trimmed for quality reasons.



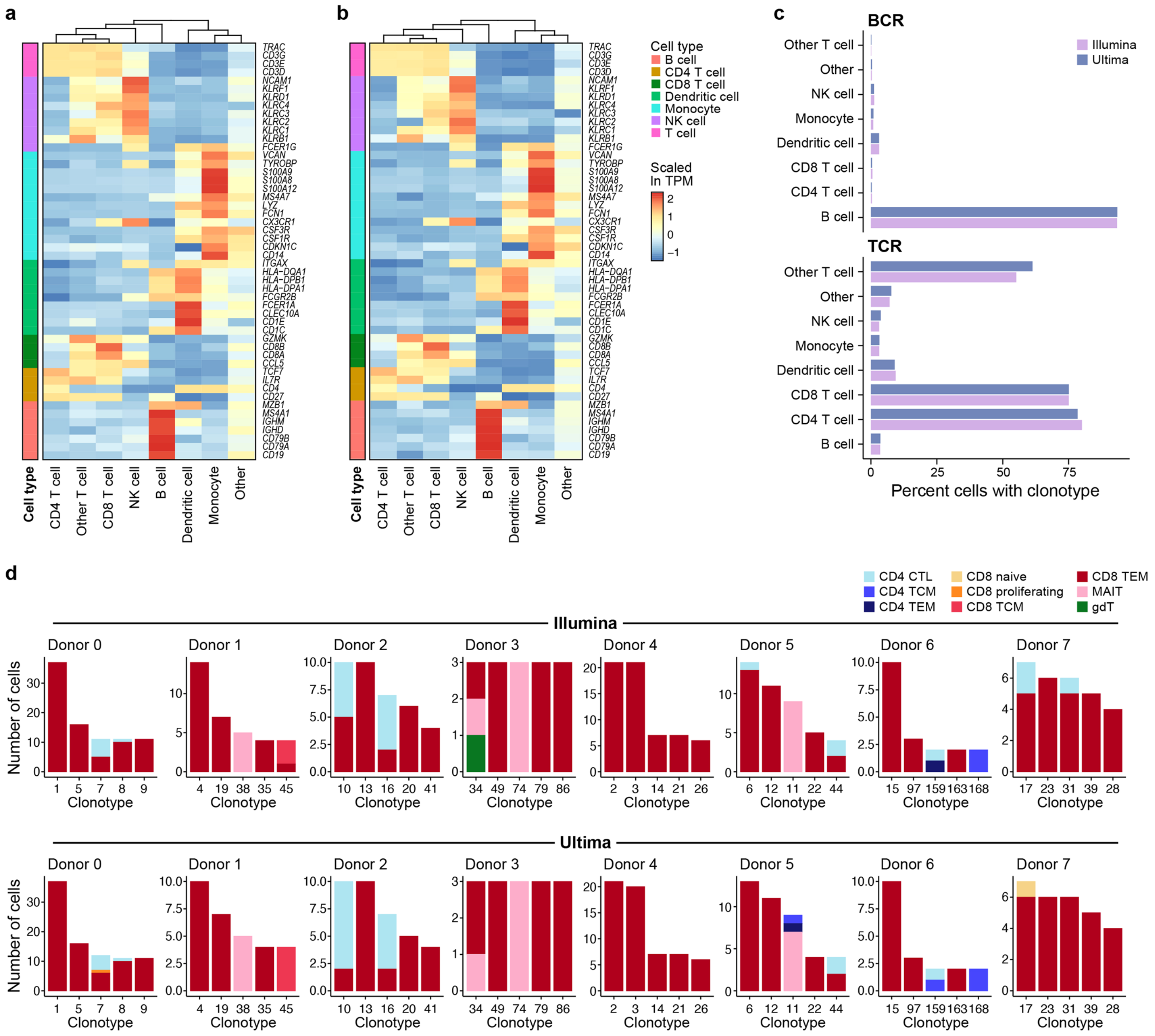
Extended Data Fig. 4 | Sequencing biases in matched 5′ and 3′ PBMC libraries. (**a**) Genes were assigned to bins by GC content (exonic sequences only) and the average logFC for each bin between Ultima and Illumina is shown with 95% CI. Positive logFC indicates higher expression in Ultima. (**b**) Genes were assigned to bins by $\log_{10}$ length (total number of bp in at least one exon for a given gene) and the average logFC for each bin between Ultima and Illumina is shown with 95% CI. Positive logFC indicates higher expression in Ultima. (**c**) Differences in read assignment ratio for genes categorized based on their expression being similar or higher in one platform than the other. Number of genes in each category shown on top of each bar. See **Methods** for additional details. (**d**) For each library (5′ PBMC, left, and 3′PBMC, right), the 5′ end of each read is shown at relative positions along the length of each gene. (**e**) Relative position of the 3′ end of reads as in (**d**). (**f**) *LILRA5* sequence coverage of 3′ libraries with Illumina (top two tracks) and Ultima (next two tracks). Standard gene annotation with introns as lines, exons as boxes, and arrows for direction of transcription shown in bottom track. Extended reference annotation using bulk or single cell (SC) Ultima data shown at bottom. Reads assigned (blue) or not (red) to *LILRA5* by Cell Ranger. (**g**) *HIST1H1D* sequence coverage of 5′ libraries shown as in (**f**). (**h**) Bar plots showing classification of reads by FeatureCount. Ultima (short) reads had Read 2 trimmed to 45 bases for 5′ and 55 bases for 3′ to match Illumina read lengths. Assigned: assigned to genes (uniquely mapped reads that overlap a gene; not all of these are assigned to a gene by Cell Ranger, which only assigns reads that completely overlap a gene); ambiguous: uniquely mapped reads that overlap the exons of multiple genes; no feature: uniquely mapped reads that do not overlap any exons; multimappers: map to multiple genomic locations (note Cell Ranger is able to recover multimappers if they only overlap one gene); unmapped: map to no genomic locations. Each read can only fit in one category. Ultima data sampled to have the same number of UMIs ((**a**) and (**b**)) or reads (other panels) as the Illumina data.



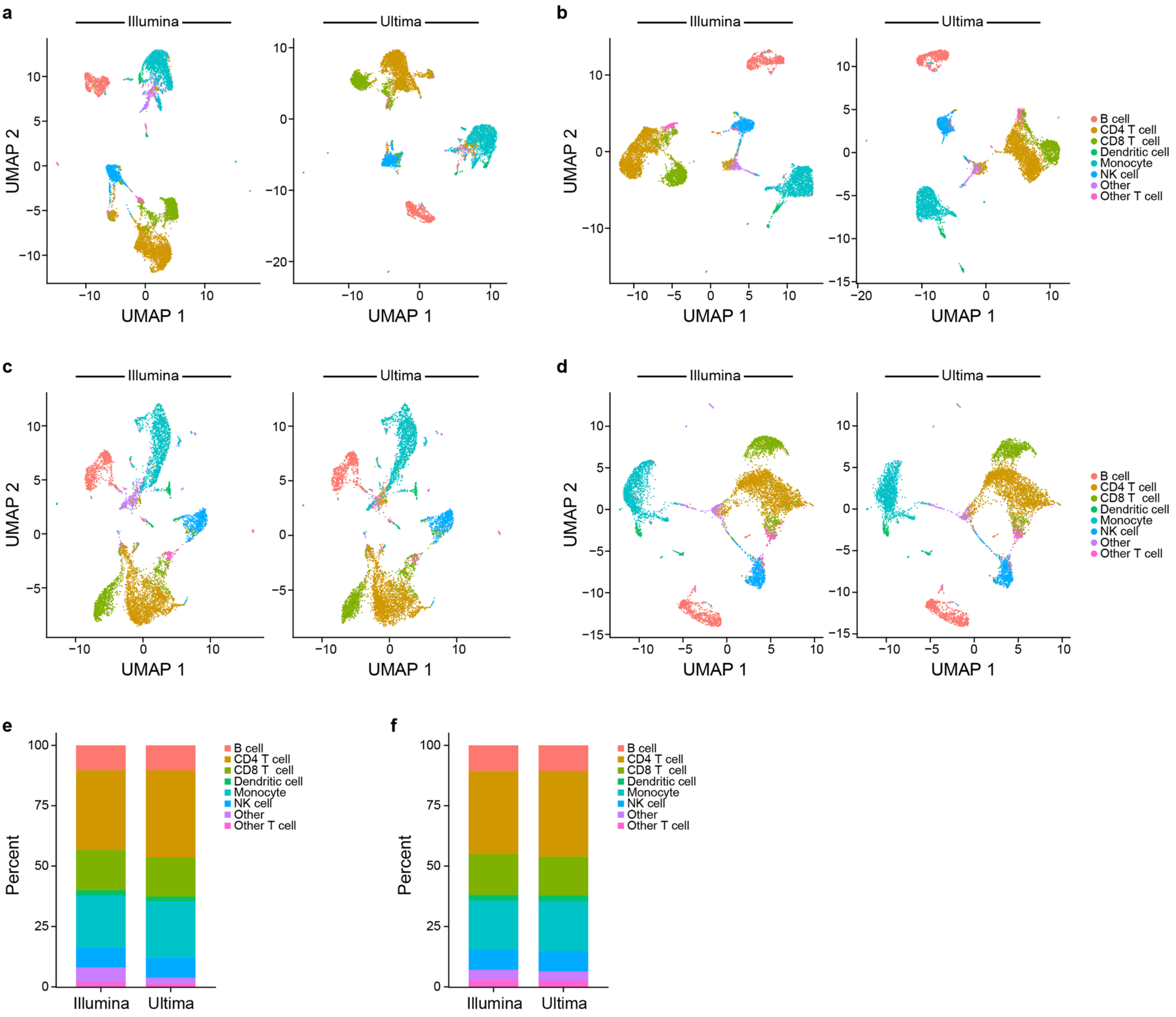
Extended Data Fig. 5 | Gene expression and read assignment with extended references. (**a**) Scatter plots comparing gene expression of Illumina and Ultima datasets using extended references for both. Extended references from bulk RNA-seq (bulk) or Ultima scRNA-seq (SC). Labeling as in Fig. 2e. For 5′ data we used the reference extended by 5′ Ultima data, while for 3′ data we used the reference extended by 3′ Ultima data. (**b**) Same as (**a**), except using the standard reference for Illumina in all cases. (**c**) Same as (**a**), except comparing standard with extended references for the same dataset. (**d**) Bar plots showing classification of reads by FeatureCount, both with the standard reference and with the extended references. Categories as in Extended Data Fig. 4. Reads were sampled so that Illumina and Ultima have the same number of reads.



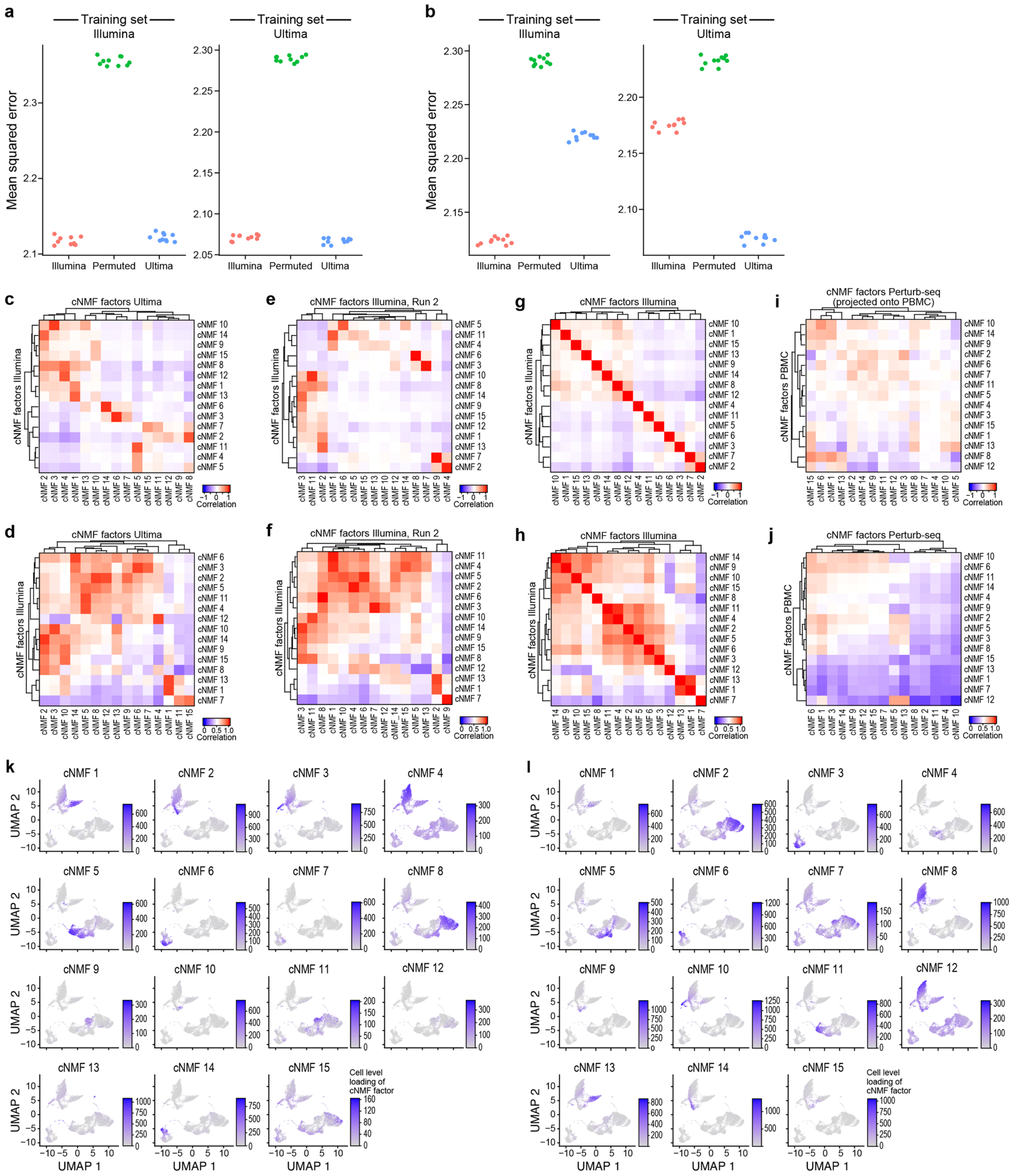
Extended Data Fig. 6 | Quality metrics for libraries from a mixture of PBMCs (5′) and from Perturb-Seq (3′). (**a**) The total number of UMIs detected per cell at different sequencing depths. (**b**) The number of cells found by Cell Ranger only in Ultima, only in Illumina, or in both. (**c**) Distribution of the number of genes per cell. (d) Scatter plots with one point for each gene as in Fig. 2e. For the 3′ library, the last three UMI bases were trimmed for quality reasons. For panels (**b**) to (**d**), reads sampled to have the same number of UMIs for Illumina and Ultima for 3′ libraries, but sampling not needed as already nearly the same for 5′ libraries.



Extended Data Fig. 7 | PBMC marker gene and TCR/BCR expression. Heatmaps showing expression of known cell type markers[13] in Illumina (**a**) and Ultima (**b**) libraries in Azimuth-assigned cell types. We did not use markers for plasma cells or markers selected because they are downregulated in a given cell type. (**c**) Percent of cells in each Azimuth-assigned cell type with associated TCR or BCR clonotype information. (**d**) Top TCR clonotypes among T cells, stratified by individual, colored by more refined cell type in Illumina (top row) and Ultima (bottom row).

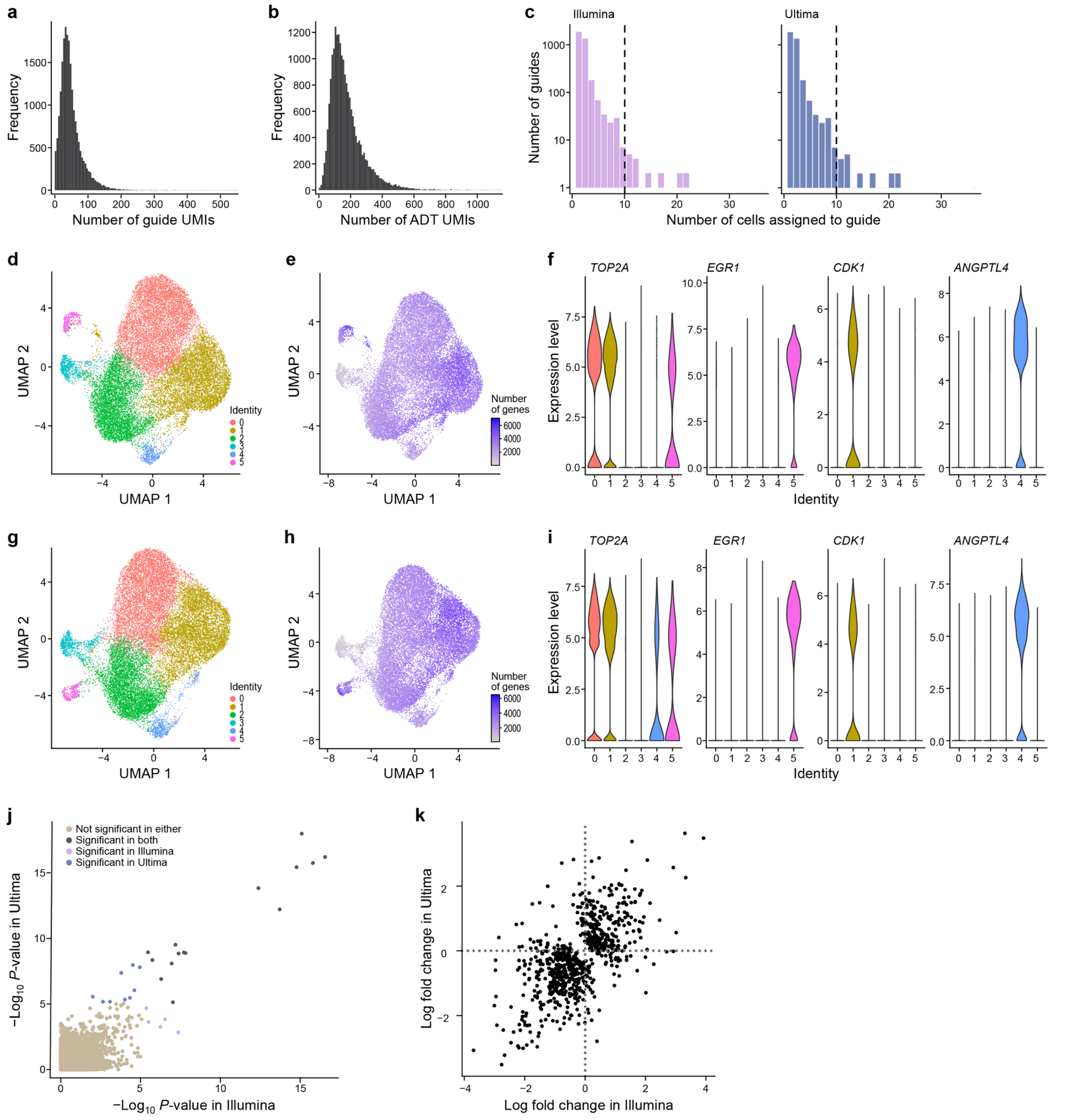


Extended Data Fig. 8 | Cell type characterization of matched 5′ and 3′ PBMC libraries. UMAP plots of Azimuth-assigned cell types in 5′ (**a**) or 3′ (**b**) libraries. UMAP plots of Azimuth-assigned cell types on a joint embedding in 5′ (**c**) or 3′ (**d**) libraries. (**e**) Bar plots of cell type proportions for the 5′ libraries. There is 92% agreement in the cell type labels between the two sequencing platforms. (**f**) Bar plots of cell type proportions for the 3′ libraries. There is 96% agreement in the cell type labels between the two sequencing platforms.



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

Extended Data Fig. 9 | Comparisons of cell state for the PBMC mixture between sequencing platforms. (**a**) NMF applied to either Illumina, Ultima, or Illumina data with permuted genes, to extract the cell loadings, and test how well they fit either the Illumina (left) or Ultima (right) data, measured by Mean Squared Error (MSE; lower MSE is better). Dots show analysis repeated ten times with different seeds. (**b**) NMF applied to either Illumina, Ultima, or Illumina data with permuted genes, to extract the gene loadings, and test how well those loadings fit either the Illumina (left) or Ultima (right) data, measured by MSE. Dots as in (**a**). (**c**) Correlation between cell level loadings shown after applying cNMF on Illumina and Ultima data with 15 factors. (**d**) Correlation between gene level loadings shown after applying cNMF on Illumina and Ultima data with 15 factors. (**e**) Correlation between cell level loadings of both runs after applying cNMF on the same Illumina data twice with 15 factors. (**f**) Correlation between gene level loadings of both runs after applying cNMF on the same Illumina data twice with 15 factors. (**g**) Correlation between cell level loadings in the Illumina data after applying cNMF. (**h**) Correlation between gene level loadings in the Illumina data after applying cNMF. (**i**) Correlation between cell level loadings after performing cNMF on the PBMC Illumina and Perturb-Seq Illumina data with 15 factors and projecting the Perturb-Seq gene loadings onto the PBMC data to get cell loadings. (**j**) Correlation of gene level loadings after performing cNMF on PBMC Illumina and Perturb-Seq Illumina data with 15 factors. Feature plots of the cell level loading cNMF factors for Ultima (**k**) and Illumina (**l**) in the joint UMAP space. All correlations here are Pearson correlations.



Extended Data Fig. 10 | Additional Perturb-Seq analysis. (**a**) Histogram of the number of guide UMIs per cell from targeted PCR amplification PCR. (**b**) Histogram of the number of CITE-seq ADT UMIs per cell. (**c**) Histogram showing the number of guides with a given number of cells assigned to them. Note the y axis is $\log_{10}$ scaled. Dotted line: 10 cells per guide. (**d**) UMAP of gene expression for Illumina. (**e**) Feature plot of the number of genes for Illumina. (**f**) Violin plots of DE genes among clusters in the Illumina data. Shown are clusters showing active cell cycling (cluster 0, 1, and 5) and a cluster with high immediate early gene levels (cluster 5). (**g**) UMAP of gene expression for Ultima. (**h**) Feature plot of the number of genes for Ultima. (**i**) Violin plots of DE genes among clusters in the Ultima data. (**j**) Scatterplot of -$\log_{10}$ (p-value) for each guide/gene pair in Illumina vs. Ultima, where the p-values are the output of DE analysis comparing each guide to the control Intergenic_1 guide. (**k**) Scatterplot of logFC for each guide/gene pair as in (**j**). Only gene/guide pairs with uncorrected p-value < 0.01 in either Ultima or Illumina are shown.

nature portfolio

Corresponding author(s): Joshua Z. Levin

Last updated by author(s): Jul 18, 2022

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	No software was used for data collection.
Data analysis	The custom code used for processing and analyzing our sequence data is freely available from https://github.com/seanken/CompareSequence. FlowJo v10.5.3 software was used to process the FACS data. Cell Ranger v5 was used to process the data. For the mixture experiment, cellsnp-lite v1.2.0 and Vireo v0.5.5 was used to demultiplex, while DemuxEM v0.1.7 was used on the Perturb-seq data. The tool seqtk v1.0 was used for downsampling. Cutadapt v2.10 and SeqKit v0.15.0 were used for processing raw Ultima reads and single-end Illumina reads. DropletUtils v1.10.3 was used for some downsampling analysis. Biopython v1.79 and Python v3.7.7 were used for looking at FASTQ quality scores. R v4.0.3 and Seurat v4.0.0 were used for downstream analysis of the data while Azimuth v0.3.2 was used for annotation. AMI was calculated with aricode v1.0.0, DE genes between clusters with Presto v1.0.0, and DE between perturbations with Nebula v1.1.7. BEDTools version 2.26.0, SAMtools v1.8, and pysam v0.15.3 were used to access BAM files, and FeatureCount v1.6.2 to annotate reads. IGV v2.3.80 was used to view mapped reads. Nextflow was used to make a pipeline for reference extension. The djvdj was used to analyze TCR/BCR data. RcppML v0.3.7 was used to perform NMF. sklearn was used to fit an elastic net model. ClusterProfiler v3.18.1 and fgsea package v1.16.0 were used for enrichment analysis. For visualization we used ggplot2 v3.3.3, cowplot v1.1.1, ComplexHeatmap package v2.6.2, and the NMF v0.30.1 package.

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

RNA-seq data generated in this project are available from the Gene Expression Omnibus with accession number GSE197452 and the Single Cell Portal with accession numbers SCP1759.
For the reference human genome, we used: https://cf.10xgenomics.com/supp/cell-exp/refdata-cellranger-GRCh38-1.2.0.tar.gz.
For Azimuth, we used the built in PBMC reference: https://azimuth.hubmapconsortium.org/references/#Human%20-%20PBMC.

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 size calculations were performed. The number of cells per experiment was chosen to be in line with the number of cells in standard practice for 10x scRNA-seq experiments. In particular, we considered our previous work such as Ding et al. Nat Biotechnol 38, 737-746 (2020).

Data exclusions

We excluded cells filtered out by the Cell Ranger pipeline, a pre-established exclusion criteria built into the software.
For downstream analysis we also filtered out doublets identified by nuclei hashing or vireo (detail in Online Methods). These exclusions are relatively standard practice in the field and were applied in the course of data processing. It is not possible to pre-determine exactly which data will need to be removed without an initial analysis.

Replication

We used different sample types and methods to verify that Ultima and Illumina gave similar results by numerous metrics. Generally, our results were comparable among these experiments. Specific experiments are not replicates and further details are provided in Online Methods.

Randomization

This was not relevant because we did not do anything like select research participants or assign samples to one particular treatment. Each experiment was performed with both Illumina and Ultima sequencing. Randomization was used in sampling equal numbers of reads, UMIs, or cells as described in the Online Methods.

Blinding

This is not relevant because there is no way to blind for sequencing approach either during the experiment or analysis.

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
☒	☐	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

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

Antibodies

Antibodies used	We stained cells with CITE-seq and hashing antibodies as previously described 22 (table S4) along with a fluorescent HLA antibody (BioLegend 311415).
Validation	Hashing antibodies were used in reference 22. The HLA antibody validation is shown in vendor website. https://www.biolegend.com/en-us/search-results/alexa-fluor-488-anti-human-hla-a-b-c-antibody-2899.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	A375 cells (ATCC CRL-1619)
Authentication	Vendor website indicates STR authentication (https://www.atcc.org/products/crl-1619#detailed-product-information). No additional authentication was done.
Mycoplasma contamination	A375 cells used in this study tested negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	Not used.

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	Cells were harvested, blocked in 2% FBS in PBS and stained with an anti-HLA-A,B,C FITC-conjugated antibody prior to flow cytometry.
Instrument	Sony MA900
Software	Flow cytometry data was analyzed using FlowJo v10.5.3 software.
Cell population abundance	The sorted cell population was the lowest 5% of cells based on HLA-FITC antibody binding signal.
Gating strategy	Live cells were sorted based on FSC and SSC. From this live population, mKate2+ cells were selected. Lastly, the lowest 5% of FITC+ cells was selected based on signal from an HLA-FITC antibody.

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