nature portfolio

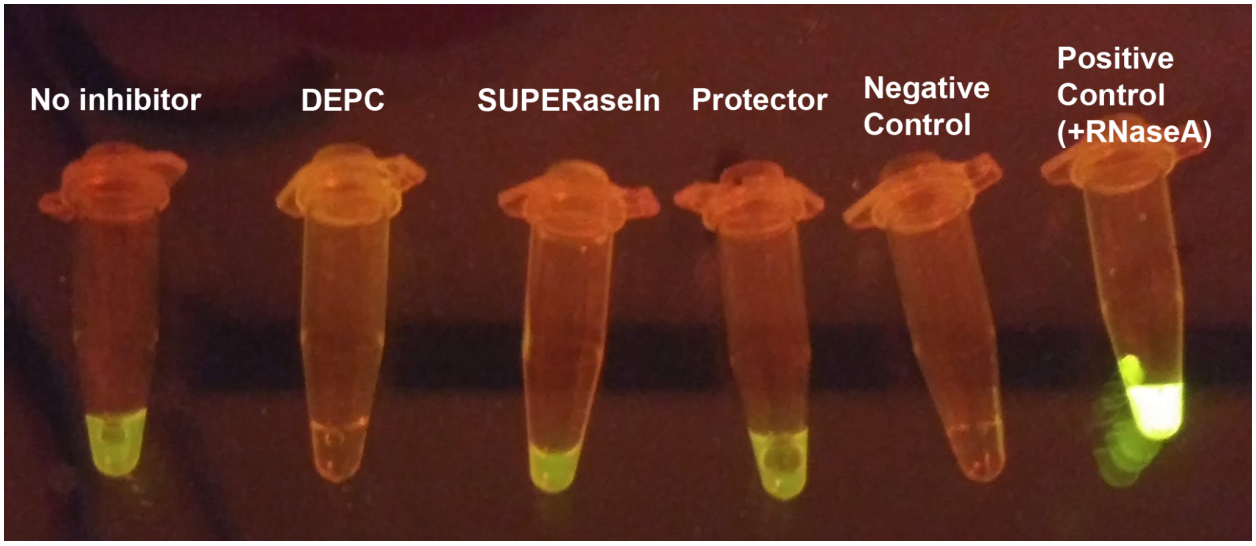
https://doi.org/10.1038/s41596-022-00752-0

Supplementary information

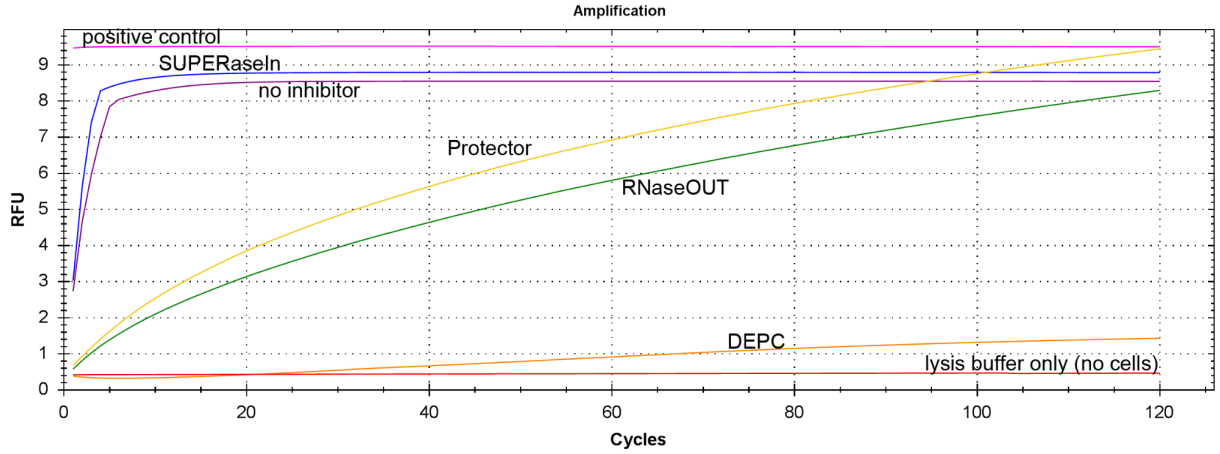
Optimized single-nucleus transcriptional profiling by combinatorial indexing

In the format provided by the authors and unedited

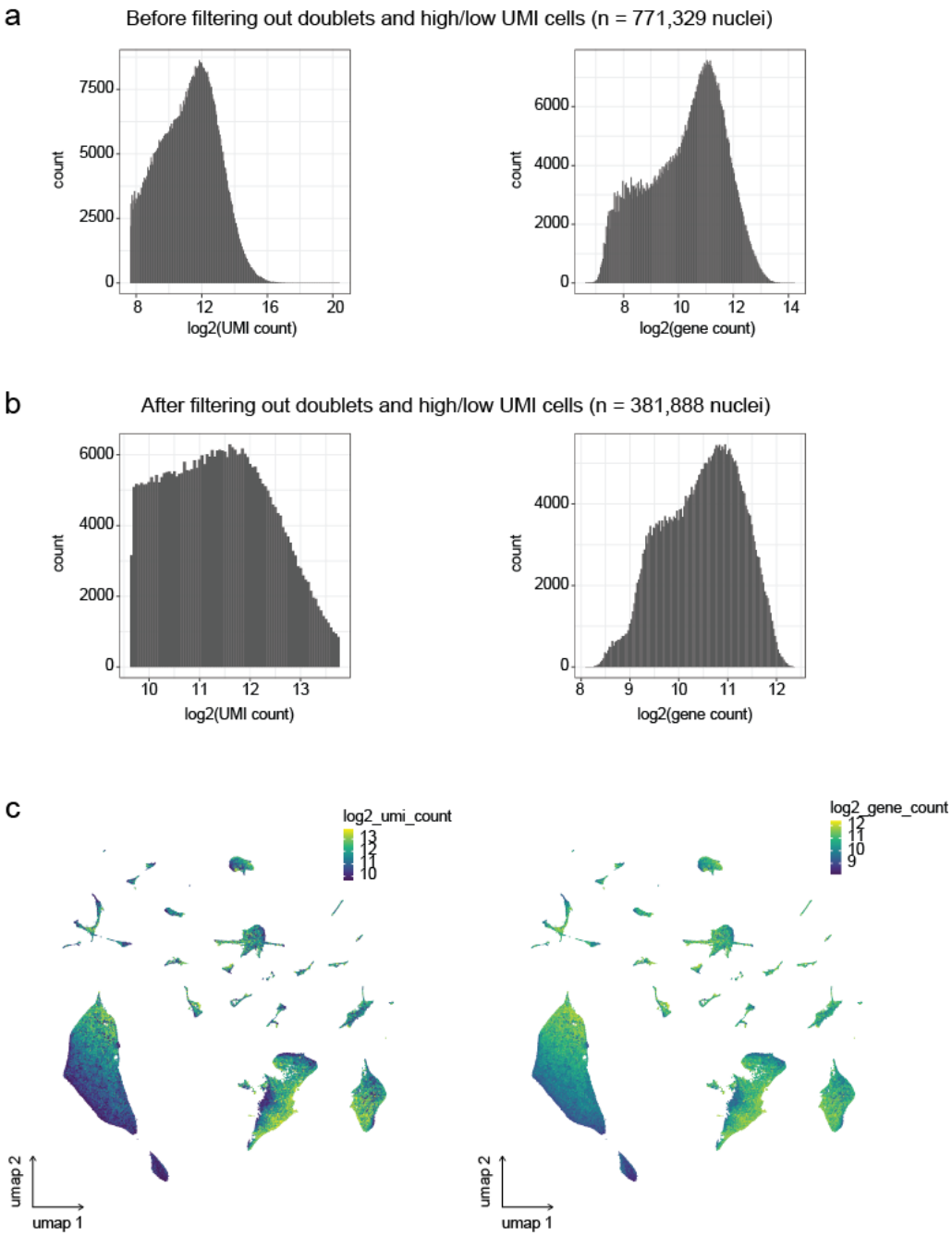
Supplementary Information



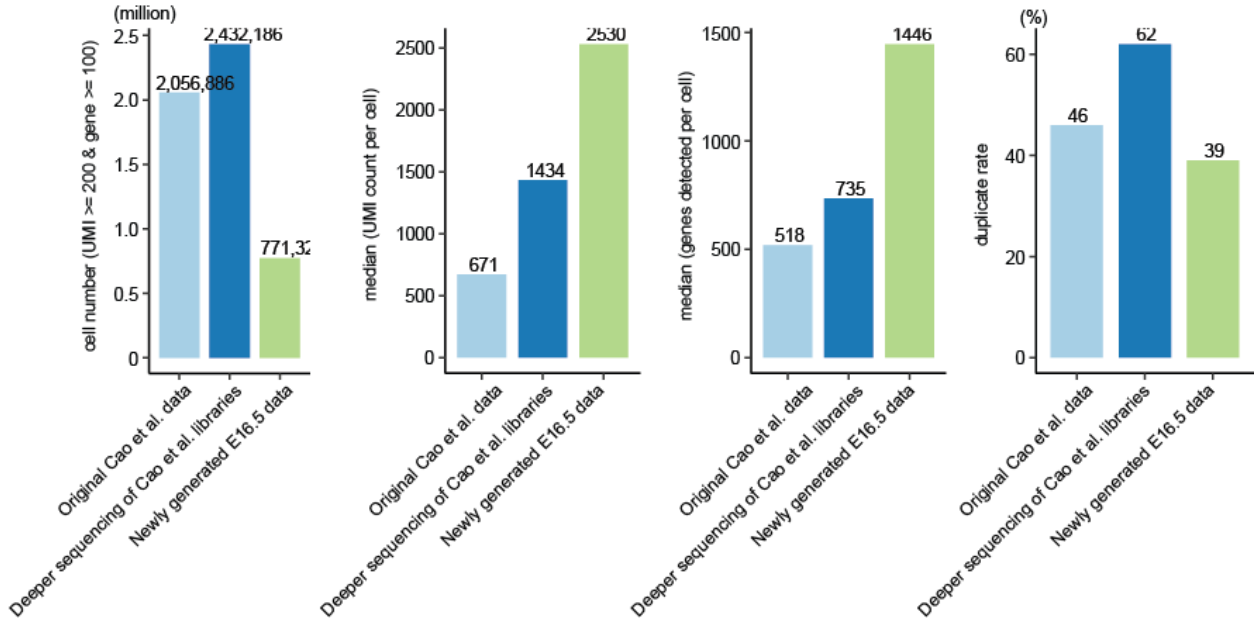
Supplementary Figure 1: IDT RNaseAlert test of lysed E13.5 mouse nuclei with various RNase inhibitors: DEPC, Superasin (ThermoFisher), Protector (Roche) (10 µl inhibitor to 1ml of lysis buffer). An increase in fluorescence indicates cleavage of the test's RNA oligo. DEPC added to lysis buffer is the only one that was able to inactivate the RNases.



Supplementary Figure 2: Real-time tracking of IDT RNaseAlert test of lysed E18.5 mouse nuclei with various RNase inhibitors. An E18.5 mouse was lysed in 20ml lysis buffer A without inhibitors. The lysate was split into 1ml aliquots and 10ul of RNase inhibitors were added. An increase in fluorescence indicate cleavage of the test's RNA oligo. Readings were taken every 30 seconds for a total of one hour. Pink: positive control - lysate with RNaseA added. Purple: cell lysate with no inhibitor added. Blue: SUPERaseIn (Thermo). Green: RNaseOUT (Thermo). Yellow: Protector (Roche). Orange: DEPC. Red: Lysis buffer only without cells added.



Supplementary Figure 3 Summary of quality of the sci-RNA-seq3 data (E16.5 mouse embryo) generated by application of the optimized protocol. a, Histograms of log2(UMI count) per cell (left) and log2(gene count detected) per cell (right) for the initial dataset after basically filtering out low quality cells (UMI < 200 or detected gene > 100 or unmatched_rate ≥ 0.4). **b**, Histograms of log2(UMI count) per cell (left) and log2(gene count detected) per cell (right) for the final dataset after further filtering out doublet cells and extremely low/high cells **c**, 2D UMAP of cells, colored by log2(UMI count) per cell (left) and log2(gene count detected) per cell (right).



Supplementary Figure 4. Comparison with previous method. The cell number, median UMI count per cell, median genes detected per cell, and duplicate rate, are shown for a previously published dataset on E9.5 - E13.5 embryos (light blue bars)[1], deeper sequencing and reanalysis of those same sequencing libraries (dark blue bars) or data newly generated on E16.5 embryo using the optimized sci-RNA-seq3 protocol (green bars).

Application of Optimized sci-RNA-seq3 Protocol to E16.5 Mouse Embryo

To showcase the optimized sci-RNA-seq3 protocol, we describe here its application to a whole mouse embryo from embryonic day 16.5 (E16.5) of development. As described in Biological Materials section, on E16.5, dams were euthanized and embryos were collected. Embryos were rinsed in 1xPBS buffer, dabbed dry, and immediately flash frozen in a foil pouch in liquid nitrogen until ready for dissociation.

Nuclei were isolated with a hypotonic, phosphate-based, lysis buffer containing sucrose and DEPC, and fixed with a combination of DSP and methanol. Nuclei were washed and resuspended in SPBSTM buffer. ~4 million of these nuclei were processed with 2 plates of RT indexes, 2 plates of ligation indexes, and 3.5 plates of PCR indexes, as per the protocol described above. About 2000 nuclei per well were distributed into the final PCR plates. All 3.5 plates of PCR reactions were pooled and sequenced on an Illumina NovaSeq 6000 with dual index reads using the S4-200 kit and standard primers (Read1 34 cycles, Index1 10 cycles, Index2 10 cycles, Read2 100 cycles), resulting in about 7.15 billion reads in total, or 4.34 billion after removal of PCR duplicates.

After processing the sequencing data using the original sci-RNA-seq3 pipeline[1], we obtained profiles for 771,329 nuclei with UMI count per cell > 200. Even though these new data are still sequenced to a lower duplication rate relative to Cao *et al* (2019)[1] (39% and 46%, respectively), the optimized sci-RNA-seq3 method has markedly improved data quality, with ~4-fold higher

UMIs and ~3-fold higher gene detection per nucleus (median UMI count 2,530; median genes detected 1,446; **Supplementary Figure 4**). We further filtered out cells which were detected as potential doublets, and then we set upper and lower thresholds of UMI counts used for quality filtering, which correspond to the mean +/- 2 standard deviations of log2-scaled values (except for the lower bound of 800, which was manually assigned). The resulting high-quality dataset included 381,888 cells that were further analyzed.

Read alignment and gene count matrix generation was performed using the pipeline that we previously developed for sci-RNA-seq3[1] (posted in https://github.com/JunyueC/sci-RNA-seq3_pipeline). Of note, the reads of mouse embryo nuclei were mapped to the mouse reference genome (mm10), and GENCODE VM12 was used for gene annotations.

The single cell gene count matrix was first filtered to remove cells with low quality (UMI < 200 or detected gene < 100 or unmatched_rate ≥ 0.4), resulting in 771,329 cells retained. To detect the potential doublet cells, we applied the scrublet/v0.1 pipeline[2] and annotated cells with doublet score > 0.2 as detected doublets (2% in the whole data set). To further detect the doublet-derived subclusters, we applied an iterative clustering strategy based on Scanpy/v.1.6.0[3]. Briefly, we selected the top 1,000 genes with the highest variance to perform the dimensionality reduction by PCA on the normalized gene expression matrix (normalized by the total UMI count per cell followed by log transformation), and then applied the Louvain clustering on the top 30 principal components. We repeated the similar strategy for each major cluster to detect the subclusters, and subclusters with a detected doublet ratio (by Scrublet) over 15% were annotated as doublet-derived subclusters.

We noticed that the above two steps were limited in marking cell doublets between abundant cell clusters and rare cell clusters (e.g. less than 1% of total cell population). To address this, we applied an additional step to remove such doublet cells using Monocle/3-alpha. First, we excluded cells identified as doublets or doublet-derived subclusters in the above two steps, and only retained protein-coding genes, lincRNA genes and pseudogenes. Then, we performed dimension reduction by PCA (50 components) on the top 5,000 most highly dispersed genes followed by UMAP, and identified the cell clusters using Louvain clustering. Next, we downsampled each cell cluster to 2,500 cells, and computed differentially expressed genes across cell clusters using the top_markers function implemented in Monocle/v3. We selected a gene set combining the top ten gene markers for each cell cluster (filtering out genes with fraction_expressing < 0.1 and then ordering by pseudo_R2), which represented the most variable gene features across cell clusters, and then we leveraged such a gene set to perform dimension reduction and Louvain clustering for each cell cluster. Finally, subclusters showing low expression of target cell cluster-specific markers and enriched expression of non-target cell cluster-specific markers were annotated as doublet-derived subclusters. All the doublet cells detected by these three steps were excluded in the downstream analysis.

We further filtered out the potential low-quality cells by investigating the numbers of UMIs and the proportion of reads mapping to the exonic regions per cell, resulting in a set of 381,888 cells that were used for performing dimension reduction.

We took the unique molecular identifiers (UMI) count matrix (feature × nuclei) to perform conventional single-cell RNA-seq data processing using *Monocle*/v3: 1) normalizing the UMI counts by the estimated size factor per cell followed by log-transformation; 2) applying PCA and then using the top 50 PCs to perform UMAP dimension reduction (umap.n_neighbors = 50, umap.min_dist = 0.01, max_components = 2); 3) performing louvain clustering using cluster_cells function in *Monocle*/v3.

We next performed manual annotation of individual clusters based on marker gene expression. As shown in a 2D UMAP, cells were generally assigned with one of 20 major developmental trajectories (**Figure 8**). As compared to trajectories described in Cao *et al.* (2019)[1], which were generated on mice ranging from E9.5 to E13.5, we identified several new cell types (*e.g.* adipocytes) and relatively dense substructures for some major trajectories. Focussing in further on white blood cells, these can be further separated into multiple sub-trajectories, including T cells, B cells, different types of macrophages, *etc.* (**Figure 8 box**). Of note, we observe closely related but distinct populations of cells corresponding to border-associated macrophages (*Lyve1*+, *F13a1*+) and microglia (*Sall1*+, *Sall3*+), consistent with a previous study of the same stage of mouse development[4]. We generally find that the above Scrublet and iterative clustering based approach is limited in marking cell doublets between abundant cell clusters and rare cell clusters (e.g. less than 1% of total cell population). To further remove such doublet cells, we took the cell clusters identified by Monocle/3, downsampled each cell cluster to 2,500 cells, and computed differentially expressed genes across cell clusters with the top_markers function of Monocle/3 (reference_cells=1000). We then selected a gene set combining the top ten gene markers for each cell cluster (filtering out genes with fraction_expressing < 0.1 and then ordering by pseudo_R2). Cells from each main cell cluster were selected for dimension reduction by PCA (10 components) first on the selected gene set of top cluster specific gene markers, and then by UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'), followed by clustering identification using the Louvain algorithm implemented in Monocle/3 (res = 1e-04 for most clustering analysis). Subclusters showing low expression of target cell cluster-specific markers and enriched expression of non-target cell cluster-specific markers were annotated as doublets derived subclusters and filtered out in visualization and downstream analysis. We further filtered out the potential low-quality cells by investigating the numbers of UMIs and the proportion of reads mapping to the exonic regions per cell, resulting in a set of 381,888 cells that were used for performing dimension reduction.

We took the unique molecular identifiers (UMI) count matrix (feature × nuclei) to perform conventional single-cell RNA-seq data processing using *Monocle*/v3: 1) normalizing the UMI counts by the estimated size factor per cell followed by log-transformation; 2) applying PCA and then using the top 50 PCs to perform UMAP dimension reduction (umap.n_neighbors = 50, umap.min_dist = 0.01, max_components = 2); 3) performing louvain clustering using cluster_cells function in *Monocle*/v3.

We next performed manual annotation of individual clusters based on marker gene expression. As shown in a 2D UMAP, cells were generally assigned with one of 20 major developmental trajectories (Figure 8). As compared to trajectories described in Cao *et al.* (2019)[1], which were generated on mice ranging from E9.5 to E13.5, we identified several new

cell types (*e.g.* adipocytes) and relatively dense substructures for some major trajectories. Focussing in further on white blood cells, these can be further separated into multiple sub-trajectories, including T cells, B cells, different types of macrophages, *etc.* (Figure 8). Of note, we observe closely related but distinct populations of cells corresponding to border-associated macrophages (*Lyve1*+, *F13a1*+) and microglia (*Sall1*+, *Sall3*+), consistent with a previous study of the same stage of mouse development[4].

1. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
2. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281–291.e9 (2019).
3. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
4. Utz, S. G. *et al.* Early Fate Defines Microglia and Non-parenchymal Brain Macrophage Development. *Cell* **181**, 557–573.e18 (2020).