**Massively parallel digital transcriptional profiling of single cells**

Grace X.Y. Zheng[1], Jessica M. Terry[1], Phillip Belgrader[1], Paul Ryvkin[1], Zachary W. Bent[1], Ryan Wilson[1], Solongo B. Ziraldo[1], Tobias D. Wheeler[1], Geoff P. McDermott[1], Junjie Zhu[1], Mark T. Gregory[2], Joe Shuga[1], Luz Montesclaros[1], Donald A. Masquelier[1], Stefanie Y. Nishimura[1], Michael Schnall-Levin[1], Paul W Wyatt[1], Christopher M. Hindson[1], Rajiv Bharadwaj[1], Alexander Wong[1], Kevin D. Ness[1], Lan W. Beppu[7], H. Joachim Deeg[7], Christopher McFarland[8], Keith R. Loeb[5,7], William J. Valente[2,3,4], Nolan G. Ericson[2], Emily A. Stevens[7], Jerald P. Radich[7], Tarjei S. Mikkelsen[1], Benjamin J. Hindson[1*], Jason H. Bielas[2,4,5,6,*]

**Supplementary Figure and Table Legends**

**Supplementary Figure 1. Multiplet rate and sensitivity of the GemCode single cell platform from scRNA-seq of 50:50 mixing of 293Ts and 3T3s. (a)** Inferred multiplet rate as a function of recovered cell number. **(b)** Expected (Poisson sampling) and observed (manual counting) number of cells per GEM. Ncell, number of cells in each GEM. **(c)** UMI count distribution of 293T cells (left), and 3T3 cells (right) in the 293T and 3T3 cell mixing sample. **(d)** CV and $CV^2$ of UMIs from 293Ts and 3T3s of 4 independent experiments. Distribution of normalized UMI counts vs. GC content **(e)** and gene length **(f)** in 293T cells. UMI counts were normalized by RNA content (Online Methods). Distribution of normalized UMI counts vs. GC content **(g)** and gene length **(h)** in 3T3 cells. Only genes with at least 1 UMI count detected in at least 1 cell are used. UMI normalization was performed by first dividing UMI counts by the total UMI counts in each cell, followed by multiplication with the median of the total UMI counts across cells. If there are multiple transcripts for a gene, the maximum length of the transcripts is used. Mean of GC content is calculated for each gene.

**Supplementary Figure 2. Conversion efficiency of the GemCode single cell platform. (a)** Distribution of Pearson correlation coefficient between expected vs. observed UMI counts for all GEMs, mean=0.94, sd=0.005. **(b)** Expected ERCC molecules per GEM vs. observed UMI counts at ERCC2 dilution of 1:50. **(c)** Conversion efficiency of each ERCC molecule as a function of their transcript GC content. **(d)** Conversion efficiency of each ERCC molecule as a function of their transcript length. **(e)** Conversion efficiency estimated from ddPCR assay of 8 genes. **(f)** $CV^2$ vs. mean UMI counts, where CV is the coefficient of variation, defined as the ratio of the standard deviation to the mean (on a log-log scale). The dashed line represents $CV^2$=1/mean.

**Supplementary Figure 3. Secondary analysis performed by the Cell Ranger pipeline (a), and custom analysis workflow (b).**

**Supplementary Figure 4. Expected proportions of Jurkat and 293T cells can be detected in Jurkat:293T cell mixture. (a)** Expected cell proportion is well correlated with observed cell proportion among 12 independent experiments. **(b)** Principal

component 1 vs. 3 of normalized scRNA-seq data, with each cell colored by normalized expression of XIST. **(c)** Distribution of filtered SNVs/cell detected in 293Ts.

**Supplementary Figure 5. Conversion efficiency and expression of marker genes in fresh PBMCs. (a)** Median number of genes (left) and UMI counts (right) detected per cell as a function of raw reads per cell. **(b)** Total RNA (pg/cell) in PBMCs, 293Ts and 3T3s. (n=7 for PBMC, n=4 for 293T, n=4 for 3T3 cells, mean ± s.e.m.). **(c)** Normalized dispersion vs. mean UMI counts. Black dots represent top most variable genes used for PCA. **(d)** Within groups sum of squares vs. number of clusters for k-means clustering. **(e-h)** tSNE projection of 68k PBMCs, colored by normalized expression of CD79A, CD4, CCR10 and PF4 in each cell, respectively. UMI normalization was performed by first dividing UMI counts by the total UMI counts in each cell, followed by multiplication with the median of the total UMI counts across cells. Then we took the natural log of the UMI counts. Finally, each gene was normalized such that the mean signal for each gene is 0, and standard deviation is 1. **(i)** Seurat's tSNE projection of 68k PBMCs, colored by the inferred cell type assignment from purified PBMCs.

**Supplementary Figure 6. FACS analysis of bead enriched sub-populations of PBMCs.**

**Supplementary Figure 7. tSNE projection of bead enriched sub-populations of PBMCs. (a)** 11 purified sub-populations of PBMCs were used. Correlation was calculated using their average expression profile and grouped by hierarchical clustering. The heatmap displays the correlation coefficient in the pairwise comparison of sub-populations. **(b-k)** tSNE projection of each purified population. In **b, h, j, k**, each cell is colored by normalized expression of marker genes FTL, CLEC9A, CD8A, CD34 and CD27 respectively. UMI normalization was performed by first dividing UMI counts by the total UMI counts in each cell, followed by multiplication with the median of the total UMI counts across cells. Then we took the natural log of the UMI counts. Finally, each gene was normalized such that the mean signal for each gene is 0, and standard deviation is 1. When more than 1 population was detected in a sample (**b** and **j**), only the population showing the correct marker expression was selected (marked by a dotted polygon).

**Supplementary Figure 8. Comparison between fresh vs. frozen PBMCs from Donor A. (a)** Scatterplot of mean UMI counts per gene across all cells between fresh vs. matched frozen PBMCs. Red dots represent genes that show 2-fold upregulation in frozen PBMCs. **(b)** Median genes (left) and UMI counts (right) detected per cell between fresh and frozen PBMCs (n=3). Black points correspond to fresh PBMCs, whereas grey points correspond to frozen PBMCs. Wilcoxon ranksum test was used to test whether the number of genes and UMI counts from fresh and frozen PBMCs were significantly different. **(c)** Proportion of major cell types detected in fresh and frozen PBMCs (n=3).

**Supplementary Figure 9. SNV analysis of scRNA-seq data from Donor B and Donor C PBMCs. (a)** Distribution of filtered SNVs in each PBMC from donor B. **(b)** Distribution of filtered SNVs in each PBMC from donor C. **(c)** % minor populations that can be confidently detected (PPV and sensitivity >0.95) vs. base error rate. **(d)** tSNE projection of PBMCs from Donor B and Donor C in 50:50 PBMC B:C sample, where each cell is colored based on their clustering (k-means) assignment. **(e)** Expression comparison between 5 clusters of PBMCs from donors B and C, with red indicating high similarity and blue indicating lower similarity. 100 cells were sampled from each cluster of PBMCs from donors B and C, and their pairwise gene expression was compared against each other.

**Supplementary Figure 10. Expression and clustering analyses of transplant samples. (a)** Median number of genes (left) and UMIs (right) detected per cell for pre-transplant, post-transplant and BMMCs from 2 healthy donors. **(b)** Distribution of filtered SNV counts per cell in AML027 pre-transplant sample. **(c)** Distribution of filtered SNV counts per cell in AML035 pre-transplant sample. **(d)** tSNE projection of pooled 6 samples (2 healthy donors, 2 AML027 host and 2 AML035), colored by k-means clustering assignment. **(e)** Normalized expression (centered) of the top variable genes (rows) from each of 9 clusters (columns) is shown in a heatmap. Numbers on the right side indicate cluster number in **d**, with connecting lines indicating the hierarchical relationship between clusters. Representative markers from each cluster are shown on the top. **(f)** tSNE projection of all cells, with each cell colored by normalized expression of HBA1, AZU1, IL8, CD34, GATA1 and CD71 respectively. UMI normalization was performed by first dividing UMI counts by the total UMI counts in each cell, followed by multiplication with the median of the total UMI counts across cells. Then we took the natural log of the UMI counts. Finally, each gene was normalized such that the mean signal for each gene is 0, and standard deviation is 1.

**Supplementary Table 1.** Comparison between GemCode single cell technology and representative single cell RNA-seq approaches. Table attached.
**Supplementary Table 2.** Sequencing metrics summary of all the scRNA-seq data. Table attached.
**Supplementary Table 3.** Cell capture rate from 4 cell lines, and 17 independent samples. Table attached.
**Supplementary Table 4.** First 10 PCs of principal component analysis on combined samples of 293T, Jurkat, 50:50 293T:Jurkat and 99:1 293T:Jurkat. Table attached.
**Supplementary Table 5.** Total number of filtered SNVs and median number of filtered SNV/cell. Table attached.
**Supplementary Table 6.** Comparison of barcode assignment between marker-based and SNV-based approaches in 50:50 293T:Jurkat mixture. Table attached.
**Supplementary Table 7.** Cluster-specific genes from all 10 clusters of 68k PBMCs, and 3 clusters identified within myeloid cells (cluster 9). Table attached.
**Supplementary Table 8.** Bead-purification strategy of bead enriched PBMCs from Donor A. Table attached.
**Supplementary Table 9.** List of genes that show 2-fold upregulation in scRNA-seq data of frozen PBMCs from Donor A. Table attached.
**Supplementary Table 10.** Cluster-specific genes from all 10 clusters identified from transplant samples. Table attached.