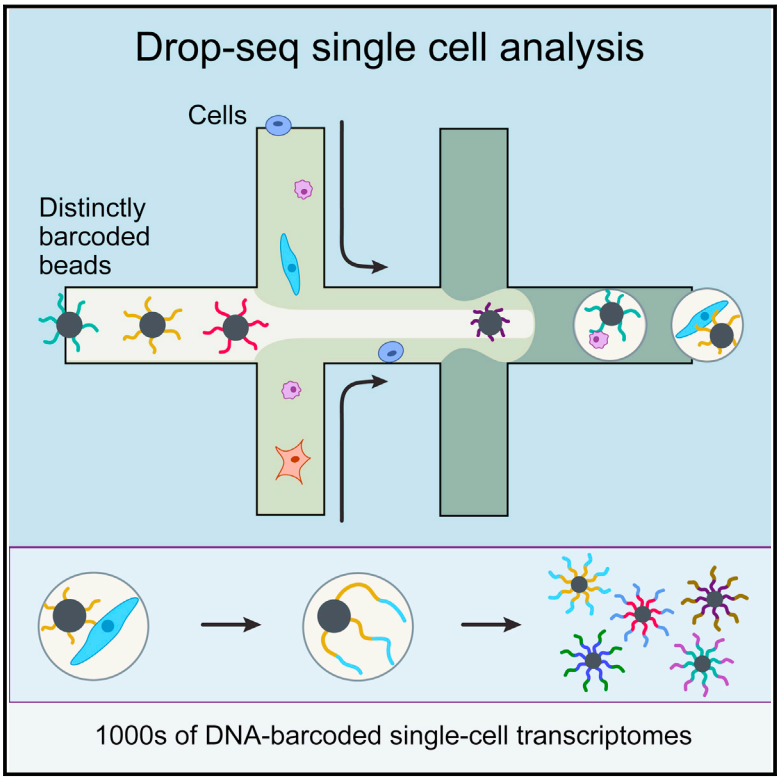
Resource

Cell

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Graphical Abstract



Authors

Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

Correspondence

emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

In Brief

Capturing single cells along with sets of uniquely barcoded primer beads together in tiny droplets enables large-scale, highly parallel single-cell transcriptomics. Applying this analysis to cells in mouse retinal tissue revealed transcriptionally distinct cell populations along with molecular markers of each type.

Highlights

- Drop-seq enables highly parallel analysis of individual cells by RNA-seq
- Drop-seq encapsulates cells in nanoliter droplets together with DNA-barcoded beads
- Systematic evaluation of Drop-seq library quality using species mixing experiments
- Drop-seq analysis of 44,808 cells identifies 39 cell populations in the retina

Accession Numbers

GSE63473

Macosko et al., 2015, Cell *161*, 1202–1214
May 21, 2015
http://dx.doi.org/10.1016/j.cell.2015.05.002

CellPress

Cell

Resource

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan Z. Macosko,[1,2,3,*] Anindita Basu,[4,5] Rahul Satija,[4,6,7] James Nemesh,[1,2,3] Karthik Shekhar,[4] Melissa Goldman,[1,2] Itay Tirosh,[4] Allison R. Bialas,[8] Nolan Kamitaki,[1,2,3] Emily M. Martersteck,[9] John J. Trombetta,[4] David A. Weitz,[5,10] Joshua R. Sanes,[9] Alex K. Shalek,[4,11,12] Aviv Regev,[4,13,14] and Steven A. McCarroll[1,2,3,*]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[2]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[4]Klarman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[5]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[6]New York Genome Center, New York, NY 10013, USA
[7]Department of Biology, New York University, New York, NY 10003, USA
[8]The Program in Cellular and Molecular Medicine, Children's Hospital Boston, Boston, MA 02115, USA
[9]Department of Molecular and Cellular Biology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
[10]Department of Physics, Harvard University, Cambridge, MA 02138, USA
[11]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA
[12]Institute for Medical Engineering and Science and Department of Chemistry, MIT, Cambridge, MA 02139, USA
[13]Department of Biology, MIT, Cambridge, MA 02139, USA
[14]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
*Correspondence: emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)
http://dx.doi.org/10.1016/j.cell.2015.05.002

SUMMARY

Cells, the basic units of biological structure and function, vary broadly in type and state. Single-cell genomics can characterize cell identity and function, but limitations of ease and scale have prevented its broad application. Here we describe Drop-seq, a strategy for quickly profiling thousands of individual cells by separating them into nanoliter-sized aqueous droplets, associating a different barcode with each cell's RNAs, and sequencing them all together. Drop-seq analyzes mRNA transcripts from thousands of individual cells simultaneously while remembering transcripts' cell of origin. We analyzed transcriptomes from 44,808 mouse retinal cells and identified 39 transcriptionally distinct cell populations, creating a molecular atlas of gene expression for known retinal cell classes and novel candidate cell subtypes. Drop-seq will accelerate biological discovery by enabling routine transcriptional profiling at single-cell resolution.

INTRODUCTION

Individual cells are the building blocks of tissues, organs, and organisms. Each tissue contains cells of many types, and cells of each type can switch among biological states. In most biological systems, our knowledge of cellular diversity is incomplete; for example, the cell-type complexity of the brain is unknown and widely debated (Luo et al., 2008; Petilla Interneuron Nomenclature Group, et al., 2008). To understand how complex tissues work, it will be important to learn the functional capacities and responses of each cell type.

A major determinant of each cell's function is its transcriptional program. Recent advances now enable mRNA-seq analysis of individual cells (Tang et al., 2009). However, methods of preparing cells for profiling have been applicable in practice to just hundreds (Hashimshony et al., 2012; Picelli et al., 2013) or (with automation) a few thousand cells (Jaitin et al., 2014), typically after first separating the cells by flow sorting (Shalek et al., 2013) or microfluidics (Shalek et al., 2014) and then amplifying each cell's transcriptome separately. Fast, scalable approaches are needed to characterize complex tissues with many cell types and states, under diverse conditions and perturbations.

Here, we describe Drop-seq, a method to analyze mRNA expression in thousands of individual cells by encapsulating cells in tiny droplets for parallel analysis. Droplets—nanoliter-scale aqueous compartments formed by precisely combining aqueous and oil flows in a microfluidic device (Thorsen et al., 2001; Umbanhowar et al., 2000)—have been used as tiny reaction chambers for PCR (Hindson et al., 2011; Vogelstein and Kinzler, 1999) and reverse transcription (Beer et al., 2008). We sought here to use droplets to compartmentalize cells into nanoliter-sized reaction chambers for analysis of all of their RNAs. A basic challenge of using droplets for transcriptomics is to retain a molecular memory of the identity of the cell from which each mRNA transcript was isolated. To accomplish this, we developed a molecular barcoding strategy to remember the cell-of-origin of each mRNA. We critically evaluated Drop-seq, then used it to profile cell states along the cell cycle. We then applied it to a complex neural tissue, mouse retina, and from 44,808 cell profiles identified 39 distinct populations, each corresponding to one or a group of closely related cell types. Our results demonstrate how large-scale single-cell analysis can help deepen our understanding of the biology of complex tissues and cell populations.

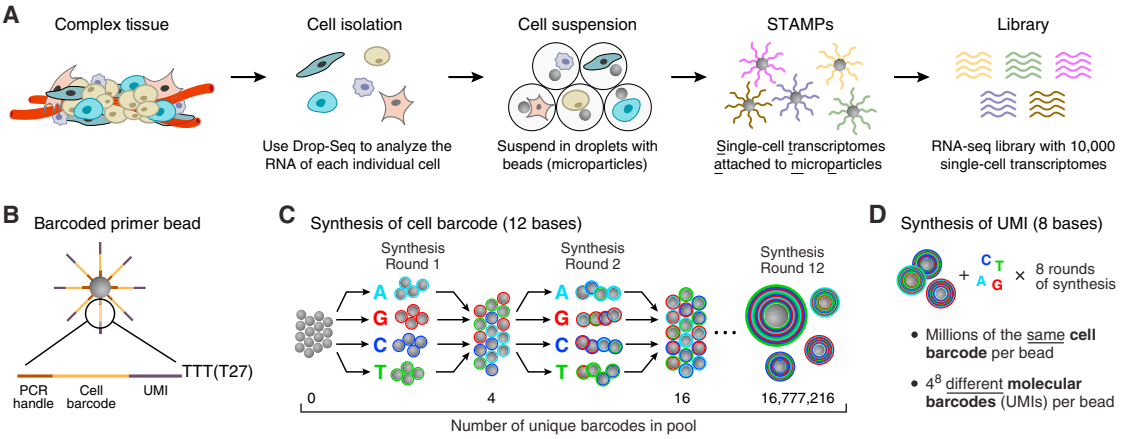


Figure 1. Molecular Barcoding of Cellular Transcriptomes in Droplets

(A) Drop-Seq barcoding schematic. A complex tissue is dissociated into individual cells, which are then encapsulated in droplets together with microparticles (gray circles) that deliver barcoded primers. Each cell is lysed within a droplet; its mRNAs bind to the primers on its companion microparticle. The mRNAs are reverse-transcribed into cDNAs, generating a set of beads called "single-cell transcriptomes attached to microparticles" (STAMPs). The barcoded STAMPs can then be amplified in pools for high-throughput mRNA-seq to analyze any desired number of individual cells.

(B) Sequence of primers on the microparticle. The primers on all beads contain a common sequence ("PCR handle") to enable PCR amplification after STAMP formation. Each microparticle contains more than 10^8 individual primers that share the same "cell barcode" (C) but have different unique molecular identifiers (UMIs), enabling mRNA transcripts to be digitally counted (D). A 30-bp oligo dT sequence is present at the end of all primer sequences for capture of mRNAs.

(C) Split-and-pool synthesis of the cell barcode. To generate the cell barcode, the pool of microparticles is repeatedly split into four equally sized oligonucleotide synthesis reactions, to which one of the four DNA bases is added, and then pooled together after each cycle, in a total of 12 split-pool cycles. The barcode synthesized on any individual bead reflects that bead's unique path through the series of synthesis reactions. The result is a pool of microparticles, each possessing one of 4^{12} (16,777,216) possible sequences on its entire complement of primers (see also Figure S1).

(D) Synthesis of a unique molecular identifier (UMI). Following the completion of the "split-and-pool" synthesis cycles, all microparticles are together subjected to eight rounds of degenerate synthesis with all four DNA bases available during each cycle, such that each individual primer receives one of 4^8 (65,536) possible sequences (UMIs).

RESULTS

Drop-seq consists of the following steps (Figure 1A): (1) prepare a single-cell suspension from a tissue; (2) co-encapsulate each cell with a distinctly barcoded microparticle (bead) in a nanoliter-scale droplet; (3) lyse cells after they have been isolated in droplets; (4) capture a cell's mRNAs on its companion microparticle, forming STAMPs (single-cell transcriptomes attached to microparticles); (5) reverse-transcribe, amplify, and sequence thousands of STAMPs in one reaction; and (6) use the STAMP barcodes to infer each transcript's cell of origin.

A Split-Pool Synthesis Approach to Generate Large Numbers of Distinctly Barcoded Beads

To deliver large numbers of distinctly barcoded primer molecules into individual droplets, we use microparticles (beads). We synthesized oligonucleotide primers directly on beads (from 5′ to 3′, yielding free 3′ ends available for enzymatic priming). Each oligonucleotide is composed of four parts (Figure 1B): (1) a constant sequence (identical on all primers and beads) for use as a priming site for downstream PCR and sequencing; (2) a "cell barcode" (identical across all the primers on the surface of any one bead, but different from the cell barcodes on other beads); (3) a Unique Molecular Identifier (UMI) (different on each primer, to identify PCR duplicates) (Kivioja et al., 2012); and (4) an oligo-dT sequence for capturing polyadenylated mRNAs and priming reverse transcription.

To efficiently generate massive numbers of beads, each with a distinct barcode, we developed a "split-and-pool" DNA synthesis strategy (Figure 1C). A pool of millions of microparticles is divided into four equally sized groups; a different DNA base (A, G, C, or T) is then added to each. All microparticles are then re-pooled, mixed, and re-split at random into another four groups, and then a different DNA base (A, G, C, or T) is added to each of the four new groups. After 12 cycles of split-and-pool DNA synthesis, the primers on any given microparticle possess the same one of 4^{12} = 16,777,216 possible 12-bp barcodes, but different microparticles have different sequences (Figure 1C). The entire microparticle pool then undergoes eight rounds of degenerate oligonucleotide synthesis to generate the UMI on each oligo (Figure 1D); finally, an oligo-dT sequence (T30) is synthesized on the 3′ end of all oligos on all beads.

To confirm that we could distinguish RNAs based on attached barcodes, we reverse-transcribed a pool of synthetic RNAs onto 11 microparticles and sequenced the resulting cDNAs (Figure S1A and Supplemental Experimental Procedures); 11 microparticle barcodes each constituted 3.5%–14% of the resulting sequencing reads, whereas the next-most-abundant 12-mer constituted only 0.06% (Figure S1A). These results suggested that the microparticle-of-origin for most cDNAs can be recognized by sequencing. We also found that each bead contained more than 10^8 barcoded primer sites and that the sequence complexity of the barcodes approached theoretical limits (Figures S1B and S1C, Supplemental Experimental Procedures).

Microfluidics Device for Co-encapsulating Cells with Beads

We designed a microfluidic "co-flow" device (Utada et al., 2007) to co-encapsulate cells with barcoded microparticles (Figures 2A and S2 and Data S1). This device quickly co-flows two aqueous solutions across an oil channel to form more than 100,000 nanoliter-sized droplets per minute. One flow contains the barcoded microparticles suspended in a lysis buffer; the other flow contains a cell suspension (Figure 2A, left, and 2B). The number of droplets created greatly exceeds the number of beads or cells injected, so that a droplet will generally contain zero or one cells, and zero or one beads. Millions of nanoliter-sized droplets are generated per hour, of which thousands contain both a bead and a cell (Movie S1). STAMPs are produced in the subset of droplets that contain both a bead and a cell.

Sequencing and Analysis of Many STAMPs in a Single Reaction

To efficiently process thousands of STAMPs at once, we break droplets, collect the mRNA-bound microparticles, and reverse-transcribe the mRNAs (from the microparticle-attached primers) together in one reaction, forming covalent, stable STAMPs (Figure 2A, step 7, and Experimental Procedures). A scientist can then select any desired number of STAMPs for the preparation of 3′-end digital expression libraries (Figure 2C, Experimental Procedures). We sequence the resulting molecules from each end (Figure 2C) using high-capacity parallel sequencing. We digitally count the number of mRNA transcripts of each gene ascertained in each cell, using the UMIs to avoid double-counting sequence reads that arose from the same mRNA transcript. We thereby create a matrix of digital gene-expression measurements (one measurement per gene per cell) for further analysis (Figure 2D, Experimental Procedures).

The Single-Cell Accuracy and Sensitivity of Drop-Seq Libraries

To measure the accuracy with which Drop-seq remembers the cell-of-origin of each mRNA, we analyzed mixtures of cultured human (HEK) and mouse (3T3) cells, scoring the numbers of human and mouse transcripts that associated with each cell barcode (Figures 3A, 3B, and S3A). We found that the individual STAMPs created by Drop-seq were highly organism-specific (Figures 3A and 3B), indicating high single-cell integrity of the libraries. At saturating levels of sequence coverage, we detected an average of 44,295 mRNA transcripts from 6,722 genes in HEK cells and 26,044 transcripts from 5,663 genes in 3T3 cells (Figures 3C and 3D).

To understand how Drop-seq libraries compare to other single-cell methods, we used three quality metrics: (1) the frequency of cell-cell doublets; (2) single-cell purity; and (3) transcript capture rates.

Cell Doublets

One potential mode of failure in any single-cell method involves cells that stick together or happen to otherwise be co-isolated for library preparation. In Drop-seq, across four conditions spanning 12.5 cells/μl to 100 cells/μl, the fraction of species-mixed STAMPs correlated with cell concentration (Figures 3A, 3B, and S3B; Experimental Procedures), with cell doublet estimates ranging from 0.36% to 11.3% for the various cell concentrations tested (under the assumption that human-mouse doublets account for half of all doublets). This reflects the greater chance at higher cell concentrations that a droplet could encapsulate multiple cells. By comparison, previous studies that used FACS (Jaitin et al., 2014) or a commercial microfluidics platform (Shalek et al., 2014) to isolate single cells reported doublet rates of 2.3% and 11% respectively, based upon examining microscopy images of captured cells. In analyzing the above mouse-human cell suspension mixture in a commercial microfluidics system (Fluidigm C1), we found that 30% of the resulting libraries in that experiment were species-mixed (Figure S3C); about one-third of these doublets were visible in the microscopy images.

Single-Cell Impurity

Species-mixing experiments enabled us to measure single-cell purity across thousands of libraries prepared at different cell concentrations. We found that purity was strongly related to cell concentration, ranging from 98.8% at 12.5 cells / μl to 90.4% at 100 cells / μl (Figure S3B). The largest source of single-cell impurity appeared to be ambient RNA that is present in the cell suspension (a first step of almost all single-cell methods) and presumably results from cells that are damaged during preparation (Figure S3D). We measured a mean single-cell purity of 95.8% for the same cell mixtures in the Fluidigm C1 system (Figure S3C), similar to Drop-seq at 50 cells /μl.

Conversion Efficiency

The use of synthetic RNA "spike-in" controls at known concentrations, together with UMIs to avoid double-counting, allows estimation of capture rates for digital single-cell expression technologies (Brennecke et al., 2013; Islam et al., 2014). We identified evidence that PCR and sequencing errors inflate the numbers of apparently unique UMIs (Table S1 and Supplemental Experimental Procedures), so we developed a more conservative estimation method than has been used in earlier studies (Islam et al., 2014); in our approach, we collapse similar UMI sequences into a single count. Using this approach we calculated a capture rate of 12.8% for Drop-seq (Figure 3G). We corroborated this estimate by making independent digital expression measurements (on bulk RNA from 50,000 HEK cells) on ten genes using droplet digital PCR (ddPCR) (Hindson et al., 2011), calculating an average conversion efficiency of 10.7% (Figures S4A, S4B, and S4C).

To further evaluate how the digital transcriptomes ascertained by Drop-seq related to the underlying mRNA content of cells, we compared Drop-seq log-expression measurements to those made by a commonly used in-solution amplification process, finding strong correlation (r = 0.94, Figure 3E), though Drop-seq ascertained GC-rich transcripts at a lower rate (Figure S4D). We also compared Drop-seq single-cell log-expression measurements with measurements from bulk mRNA-seq, observing a correlation of r = 0.90 (Figures 3F, S4E, and S4F).

Cell States: Drop-Seq Analysis of the Cell Cycle

To evaluate the visibility of cell states in Drop-seq, we first examined cell-to-cell variation among the 589 HEK and 412 3T3 STAMPs shown in Figure 3B. Both cultures consisted of asynchronously dividing cells; principal components analysis (PCA) of the single-cell expression profiles showed the top principal components to be dominated by genes with roles in protein

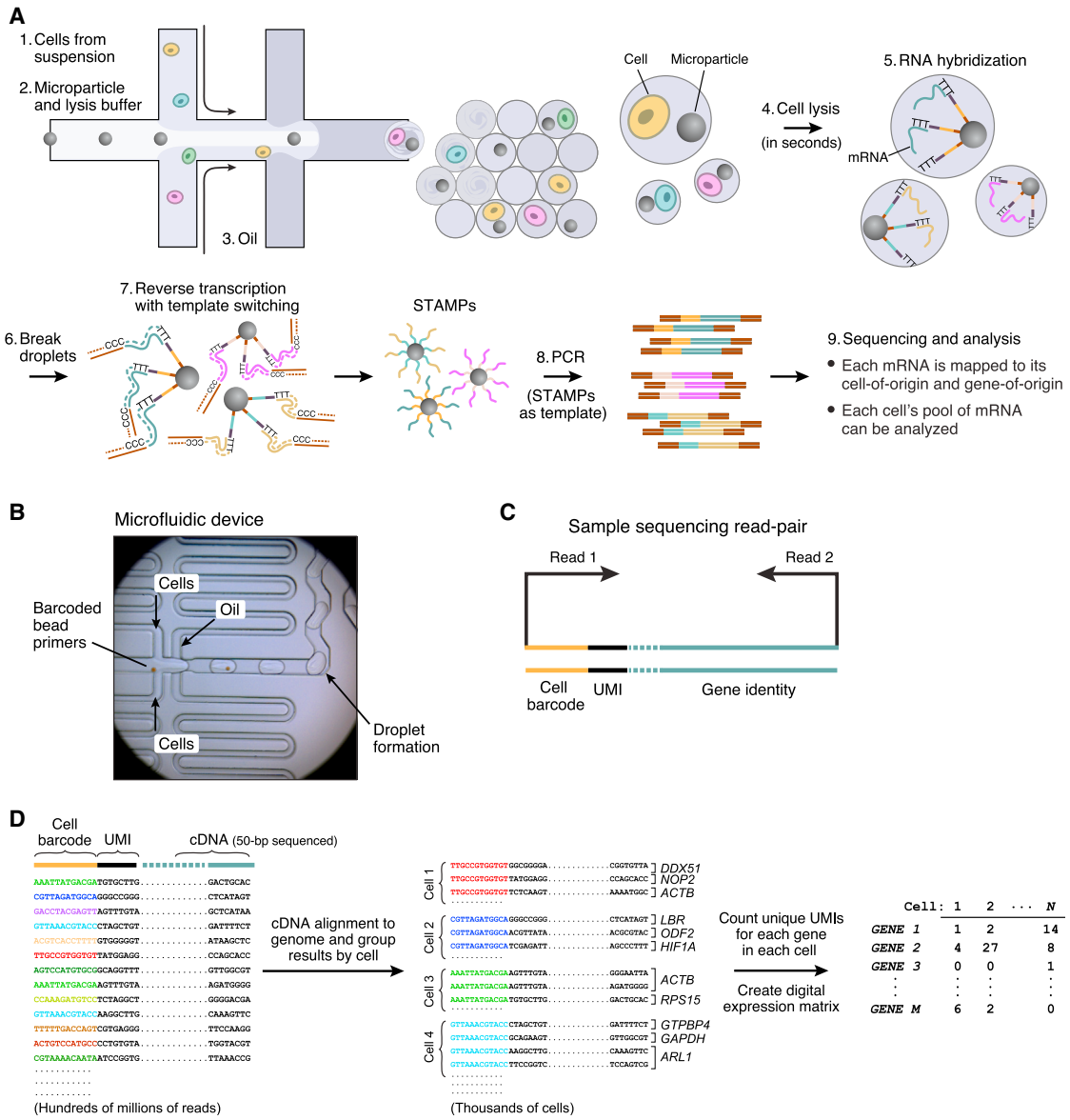


Figure 2. Extraction and Processing of Single-Cell Transcriptomes by Drop-Seq

(A) Schematic of single-cell mRNA-seq library preparation with Drop-seq. A custom-designed microfluidic device joins two aqueous flows before their compartmentalization into discrete droplets. One flow contains cells, and the other flow contains barcoded primer beads suspended in a lysis buffer. Immediately following droplet formation, the cell is lysed and releases its mRNAs, which then hybridize to the primers on the microparticle surface. The droplets are broken by adding a reagent to destabilize the oil-water interface (Experimental Procedures), and the microparticles collected and washed. The mRNAs are then reverse-transcribed in bulk, forming STAMPs, and template switching is used to introduce a PCR handle downstream of the synthesized cDNA (Zhu et al., 2001).

(B) Microfluidic device used in Drop-seq. Beads (brown in image), suspended in a lysis agent, enter the device from the central channel; cells enter from the top and bottom. Laminar flow prevents mixing of the two aqueous inputs prior to droplet formation (see also Movie S1). Schematics of the device design and how it is operated can be found in Figure S2.

(C) Molecular elements of a Drop-seq sequencing library. The first read yields the cell barcode and UMI. The second, paired read interrogates sequence from the cDNA (50 bp is typically sequenced); this sequence is then aligned to the genome to determine a transcript's gene of origin.

(D) In silico reconstruction of thousands of single-cell transcriptomes. Millions of paired-end reads are generated from a Drop-seq library on a high-throughput sequencer. The reads are first aligned to a reference genome to identify the gene-of-origin of the cDNA. Next, reads are organized by their cell barcodes, and individual UMIs are counted for each gene in each cell (Supplemental Experimental Procedures). The result, shown at far right, is a "digital expression matrix" in which each column corresponds to a cell, each row corresponds to a gene, and each entry is the integer number of transcripts detected from that gene, in that cell.

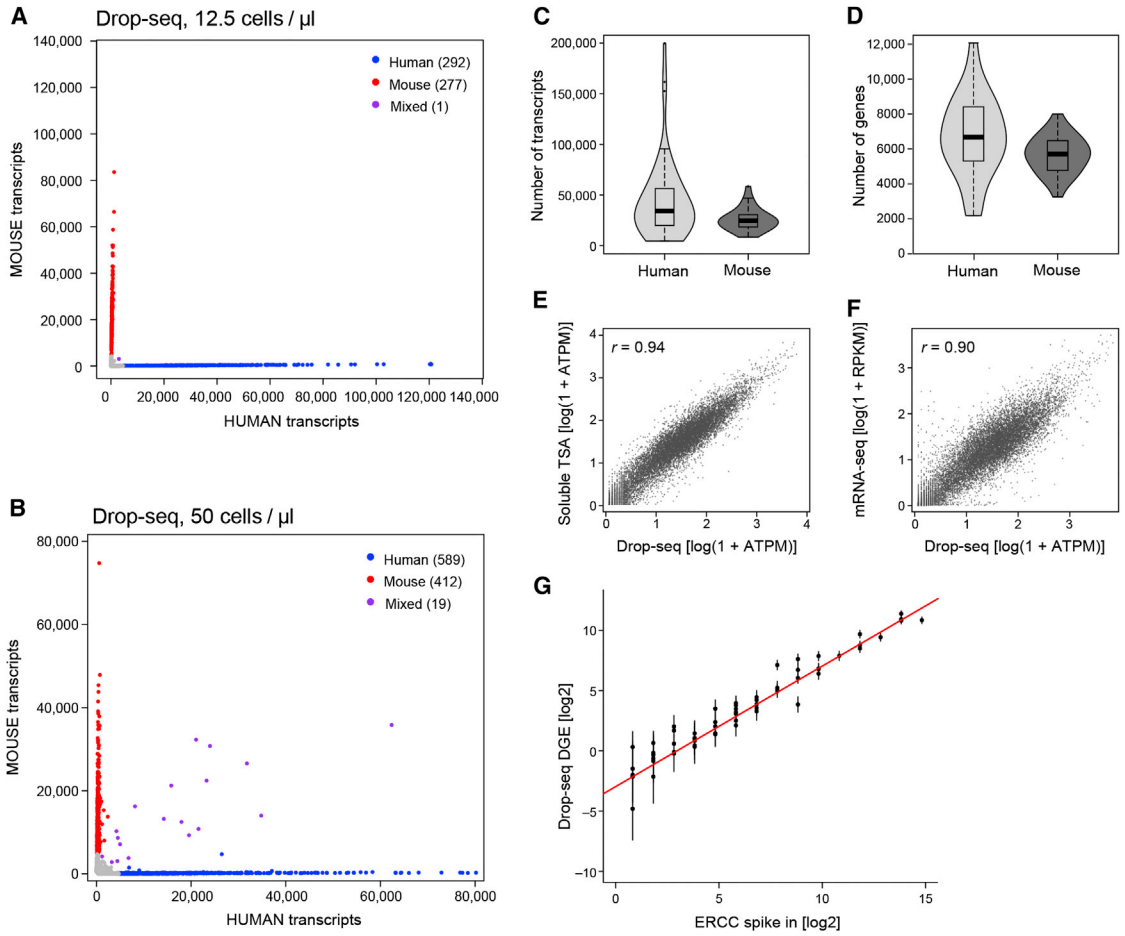


Figure 3. Critical Evaluation of Drop-Seq Using Species-Mixing Experiments

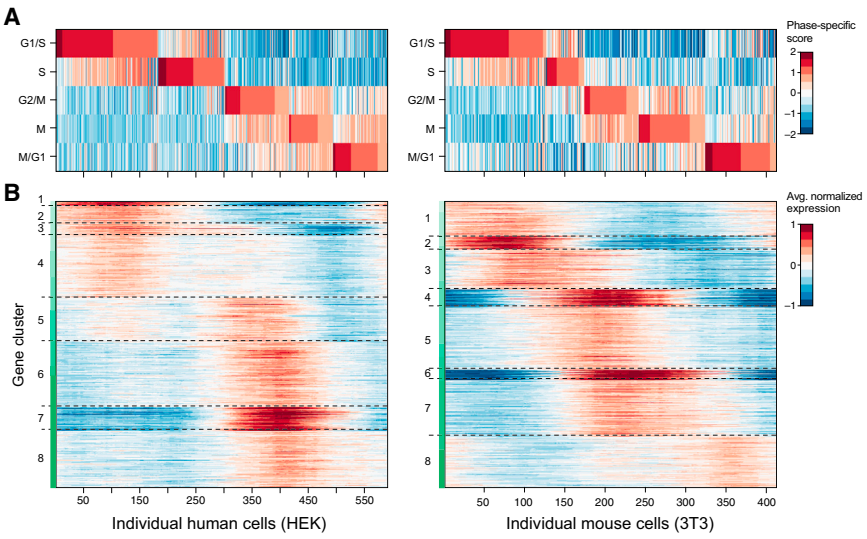
(A and B) Drop-seq analysis of mixutres of mouse and human cells. Mixtures of human (HEK) and mouse (3T3) cells were analyzed by Drop-seq at the concentrations shown. The scatter plot shows the number of human and mouse transcripts associating to each STAMP. Blue dots indicate STAMPs that were designated from these data as human-specifiic (average of 99% human transcripts); red dots indicate STAMPs that were mouse-specific (average 99%). At the lower cell concentration, one STAMP barcode (of 570) associated with a mixture of human and mouse transcripts (A, purple). At the higher cell concentration, about 1.9% of STAMP barcodes associated with mouse-human mixtures (B). Data for other cell concentrations and a different single-cell analysis platform are in Figures S3B and S3C.

(C and D) Sensitivity analysis of Drop-seq at high read-depth. Violin plots show the distribution of the number of transcripts (C, scored by UMIs) and genes (D) detected per cell for 54 HEK (human) STAMPs (blue) and 28 3T3 (mouse) STAMPs (green) that were sequenced to a mean read depth of 737,240 high-quality aligned reads per cell.

(E and F) Correlation between gene expression measurements in Drop-seq and non-single-cell RNA-seq methods. Comparison of Drop-seq gene expression measurements (averaged across 550 STAMPs) to measurements from bulk RNA analyzed by: (E) an in-solution template switch amplification (TSA) procedure similar to Smart-seq2 (Picelli et al., 2013) (Supplemental Experimental Procedures); and (F) Illumina TruSeq mRNA-seq. All comparisons involve RNA derived from the same cell culture flask (3T3 cells). All expression counts were converted to average transcripts per million (ATPM) and plotted as log (1+ATPM).

(G) Quantitation of Drop-seq capture efficiency by ERCC spike-ins. Drop-seq was performed with ERCC control synthetic RNA at an estimated concentration of 100,000 ERCC RNA molecules per droplet. 84 beads were sequenced at a mean depth of 2.4 million reads, aligned to the ERCC reference sequences, and UMIs counted for each ERCC species, after applying a stringent down-correction for potential sequencing errors (Table S1 and Supplemental Experimental Procedures). For each ERCC RNA species above an average concentration of one molecule per droplet, the predicted number of molecules per droplet was plotted in log space (x-axis), versus the actual number of molecules detected per droplet by Drop-seq, also in log space (y-axis). Error bars indicate SD. The intercept of a regression line, constrained to have a slope of 1 and fitted to the seven highest points, was used to estimate a conversion factor (0.128). A second estimation, using the average number of detected transcripts divided by the number of ERCC molecules used (100,000), yielded a conversion factor of 0.125.

synthesis, growth, DNA replication, and other aspects of the cell cycle. We inferred the cell-cycle phase of each of the 1,001 cells by scoring for gene sets (signatures) reflecting five phases of the cell cycle previously characterized in chemically synchronized cells (G1/S, S, G2/M, M, and M/G1) (Figure 4A, Table S2) (Whitfield et al., 2002). We identified 544 human and 668 mouse genes with expression patterns that varied along the cell cycle (at a false discovery rate of 5%; Experimental Procedures) (Figure 4B), including 200 orthologous gene pairs ($p < 10^{-65}$ by hypergeometric test). Of these orthologous gene pairs, most (82.5%)



C

Classic cell cycle genes		Novel, conserved cell cycle genes	
CCNB1	*MCM6*	*ATF4*	*OTUB1*
CCNB2	*MCM7*	*ARHGAP11A*	*PARPBP*
MCM2	*MCM10*	*ARPC2*	*RPL26*
MCM3	*AURKA*	*CDCA4*	*SNHG3*
MCM4	*AURKB*	*E2F7*	*SRP9*
MCM5		*HISTH1E*	*TCF19*
		MCMBP	*WDHD1*
		NCAPG	*ZFHX4*
		NXT1	

Figure 4. Cell-Cycle Analysis of HEK and 3T3 Cells Analyzed by Drop-Seq

(A) Cell-cycle state of 589 HEK cells (left) and 412 3T3 cells (right) measured by Drop-seq. Cells were assessed for their progression through the cell cycle by comparison of each cell's global pattern of gene expression with gene sets known to be enriched in one of five phases of the cycle (horizontal rows). A phase-specific score was calculated for each cell across these five phases (Supplemental Experimental Procedures), and the cells ordered by their phase scores.

(B) Discovery of cell-cycle regulated genes. Heat map showing the average normalized expression of 544 human and 668 mouse genes found to be regulated by the cell cycle. Maximal and minimal expression was calculated for each gene across a sliding window of the ordered cells, and compared with shuffled cells to obtain a false discovery rate (FDR) (Experimental Procedures). The plotted genes (FDR threshold of 5%) were then clustered by k-means analysis to identify sets of genes with similar expression patterns. Cluster boundaries are represented by dashed gray lines.

(C) Representative cell-cycle regulated genes discovered by Drop-seq. Selected genes that were found to be cell-cycle regulated in both the HEK and 3T3 cell sets. Left: genes that are well-known to be cell-cycle regulated. Right: some genes identified in this analysis that were not previously known to be associated with the cell cycle (Experimental Procedures). A complete list of cell-cycle regulated genes can be found in Table S2.

have been previously annotated as related to the cell cycle in at least one species; among the other 17.5%, we found some that would be expected to show cell-cycle variation (e.g., *E2F7* and *PARPBP*) and many that to our knowledge were not previously connected to the cell cycle (Figure 4C and Table S2). Single-cell analysis at this scale enabled characterization of cell-cycle gene expression without chemical synchronization and at high temporal resolution.

Cell Types: Drop-Seq Analysis of the Retina

We selected the retina as the first tissue to study with Drop-seq because decades of work has generated molecular information about many retinal cell types (Masland, 2012; Sanes and Zipursky, 2010), allowing us to relate our RNA-seq data to prior classification. The retina contains five neuronal classes—retinal ganglion, bipolar, horizontal, photoreceptor, and amacrine—each defined by morphological, physiological, and molecular criteria (Figure 5A). Most of the classes are divisible into discrete types—a total currently estimated at about 100—but well under half of these types possess known, distinguishing molecular markers.

We sequenced 49,300 STAMPs prepared from the retinas of 14-day-old mice (STAMPs were collected in seven batches over 4 days). We performed principal components analysis on the 13,155 largest libraries (Figure S5, Table S3), then reduced the 32 statistically significant PCs (Experimental Procedures) to two dimensions using t-Distributed Stochastic Neighbor Embedding (tSNE) (Amir et al., 2013; van der Maaten and Hinton, 2008). We projected the remaining 36,145 cells in the data into the tSNE analysis. We then combined a density clustering approach with post hoc differential expression analysis to divide 44,808 cells among 39 transcriptionally distinct clusters (Supplemental Experimental Procedures) ranging from 50 to 29,400 cells in size (Figures 5B and 5C). Finally, we organized the 39 cell populations into larger categories (classes) by building a dendrogram of similarity relationships among the 39 cell populations (Figure 5D, left).

The cell populations inferred from this analysis were readily matched to the known retinal cell types, including all five neuronal cell classes, based on the specific expression of known markers for these cell types (Figure 5D, right, and Figure S6A). Additional clusters corresponded to astrocytes (associated with retinal ganglion cell axons exiting the retina), resident microglia, endothelial cells (from intra-retinal vasculature), pericytes, and fibroblasts (Figure 5D). The relative abundances of the major cell classes in our data agreed with earlier estimates from microscopy (Jeon et al., 1998) (Table 1).

Replication and Cumulative Power of Drop-Seq Data

Replication across experimental sessions enables the construction of cumulatively powerful datasets—but only if data are replicable and comparable. The retinal STAMPs were generated on 4 different days (weeks apart), utilizing different litters and multiple runs in several sessions, for a total of seven replicates. One of the

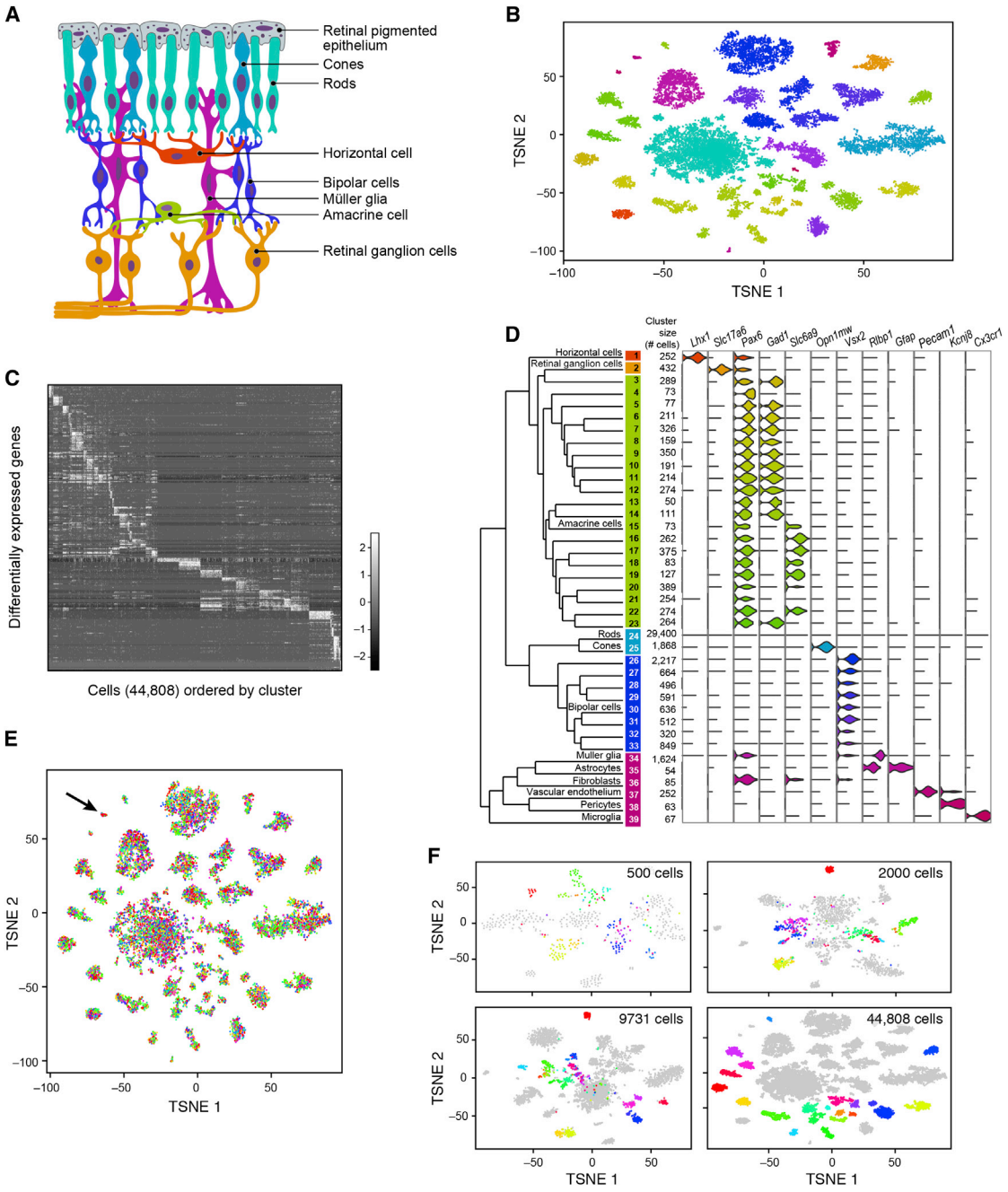


Figure 5. *Ab Initio* Reconstruction of Retinal Cell Types from 44,808 Single-Cell Transcription Profiles Prepared by Drop-Seq

(A) Schematic representation of major cell classes in the retina. Photoreceptors (rods or cones) detect light and pass information to bipolar cells, which in turn contact retinal ganglion cells that extend axons into other CNS tissues. Amacrine, bipolar and horizontal cells are retinal interneurons; Müller glia act as support cells for surrounding neurons.

(B) Clustering of 44,808 Drop-seq single-cell expression profiles into 39 retinal cell populations. The plot shows a two-dimensional representation (tSNE) of global gene expression relationships among 44,808 cells; clusters are colored by cell class, according to Figure 5A.

(C) Differentially expressed genes across 39 retinal cell populations. In this heat map, rows correspond to individual genes found to be selectively upregulated in individual clusters ($p < 0.01$, Bonferroni corrected); columns are individual cells, ordered by cluster (1–39). Clusters with > 1,000 cells were downsampled to 1,000 cells to prevent them from dominating the plot.

(D) Gene expression similarity relationships among 39 inferred cell populations. Average expression across all detected genes was calculated for each of 39 cell clusters, and the relative (Euclidean) distances between gene-expression patterns for the 39 clusters are represented by a dendrogram. The branches of the dendrogram were annotated by examining the differential expression of known markers for retina cell classes and types. Twelve examples are shown at right, using violin plots to represent the distribution of expression within the clusters. Violin plots for additional genes are in Figure S6A.

(legend continued on next page)

Table 1. Ascertainment of Cell Types and Frequencies in the Mouse Retina by Drop-Seq

Cell Class	Percentage of Retina (Jeon et al., 1998) (%)	Percentage of Cell Population in Drop-Seq (%)
Rod photoreceptors	79.9	65.6
Cone photoreceptors	2.1	4.2
Muller glia	2.8	3.6
Retinal ganglion cells	0.5	1.0
Horizontal cells	0.5	0.6
Amacrine cells	7.0	9.9
Bipolar cells	7.3	14.0
Microglia	—	0.2
Retinal endothelial cells	—	0.6
Astrocytes		0.1

The sizes of the 39 annotated cell clusters produced from Drop-seq were used to estimate their fractions of the total cell population. These data were compared with those obtained by microscopy techniques (Jeon et al., 1998).

runs was performed at a particularly low cell concentration (15 cells/μl) and thus high purity, to evaluate whether results were artifacts of cell-cell doublets or single-cell impurity. We found that all 39 clusters contained cells from every experiment. One cluster (arrow in Figure 5E; star in Figure S6B), which drew disproportionately from two replicates, expressed markers of fibroblasts, a non-retinal cell type that is present in tissue surrounding the retina, and hence likely represents imprecise dissection.

We examined how the classification of cells (based on their patterns of gene expression) evolved as a function of the numbers of cells in analysis. We used 500, 2,000, or 9,731 cells from our dataset, and asked how (for example) cells identified as amacrines in the full dataset clustered in analyses of smaller numbers of cells (Figure 5F). As the number of cells in the data increased, distinctions between related clusters become clearer, stronger, and finer in resolution, with the result that a greater number of rare amacrine cell sub-populations (each representing 0.1%–0.9% of the cells in the experiment) could ultimately be distinguished from one another (Figure 5F).

Profiles of Amacrine Cell Types

To characterize distinctions among closely related cell populations, we focused on the 21 clusters of amacrines. Amacrines are the most morphologically diverse neuronal class (Masland, 2012), but the majority of types lack defining molecular markers. Most amacrine cells are inhibitory, utilizing either GABA or glycine as a neurotransmitter. Excitatory amacrine cells that release glutamate have also been identified (Haverkamp and Wässle, 2004). Another amacrine cell population expresses no GABAergic, glycinergic or glutamatergic markers; its neurotransmitter is unidentified (nGnG amacrines) (Kay et al., 2011).

We first identified markers that were most universally expressed by amacrines relative to other cell classes (Figure 6A). We then assessed the expression of known glycinergic and GABAergic markers; their mutually exclusive expression is a fundamental distinction among amacrines. Of the 21 amacrine clusters, 12 were identifiable as GABAergic (*Gad1* and/or *Gad2*-positive) and 5 others were glycinergic (glycine transporter *Slc6a9*-positive) (Figure 6B). An additional cell population was identified as excitatory by its expression of a glutamate transporter, *Slc17a8* (Figure 6B). The remaining three clusters (clusters 4, 20, and 21) had low levels of GABAergic, glycinergic, and glutamatergic markers; these likely include nGnG amacrines.

Among the glycinergic and GABAergic clusters, we found many amacrine types with known markers. The most divergent glycinergic cluster appeared to correspond to the A-II amacrine neurons (Figure 6B, cluster 16), as this was the only cluster to strongly express the *Gjd2* gene encoding the gap junction protein connexin 36 (Feigenspan et al., 2001). *Ebf3*, a transcription factor found in SEG glycinergic as well as nGnG amacrines, was specific to clusters 17 and 20. Starburst amacrine neurons (SACs), the only retinal cells that use acetylcholine as a co-transmitter, were identifiable as cluster 3 by their expression of the cholinergic marker *Chat* (Figure 6B). Unlike other GABAergic cells, SACs expressed *Gad1* but not *Gad2*, as previously observed in rabbit (Famiglietti and Sundquist, 2010).

We then identified selectively expressed markers for each of the 21 amacrine cell populations (Figure 6C and Table S4). We validated two of the markers immunohistochemically. First, we co-stained retinal sections with antibodies to the transcription factor MAF, the top marker of cluster 7, plus antibodies to either GAD1 or SLC6A9, markers of GABAergic and glycinergic transmission, respectively. As predicted by the Drop-seq analysis, MAF was found in a small subset of amacrine cells that were GABAergic and not glycinergic (Figure 6D). Cluster 7 had numerous genes that were enriched relative to its nearest neighbor, cluster 6 (Figure 6E, 16 genes > 2.8-fold enrichment, $p < 10^{-9}$), including *Crybb3*, which belongs to the crystallin family of proteins that are known to be directly upregulated by *Maf* (Yang and Cvekl, 2005), and another, the protease *Mmp9*, which accepts crystallins as substrates (Descamps et al., 2005). Second, we stained sections with antibodies to PPP1R17 (Figure 6F), a nominated marker of cluster 20. Cluster 20 shows weak, infrequent glycine transporter expression and is one of only two clusters (with cluster 21) that express *Neurod6*, a marker of nGnG neurons (Kay et al., 2011). We used a transgenic strain (MitoP) that has been shown to express CFP specifically in nGnG amacrines (Kay et al., 2011). PPP1R17 stained 85% of all CFP-positive amacrines in

(E) Representation of experimental replicates in each cell population. tSNE plot from Figure 2B, with each cell now colored by experimental replicate (for visual clarity, the central rod cluster was downsampled to 10,000 cells). Each of the seven replicates contributes to all 39 cell populations. Cluster 36 (arrow), in which these replicates are unevenly represented, expressed markers of fibroblasts, which are not native to the retina and are presumably a dissection artifact (see also Figure S6B).

(F) Trajectory of amacrine clustering as a function of number of cells analyzed. Three different downsampled datasets were generated: (1) 500, (2) 2,000, or (3) 9,731 cells (Supplemental Experimental Procedures). Cells identified as amacrines (clusters 3–23) in the full analysis are here colored by their cluster identities in that analysis. Analyses of smaller numbers of cells incompletely distinguished these subpopulations from one another.

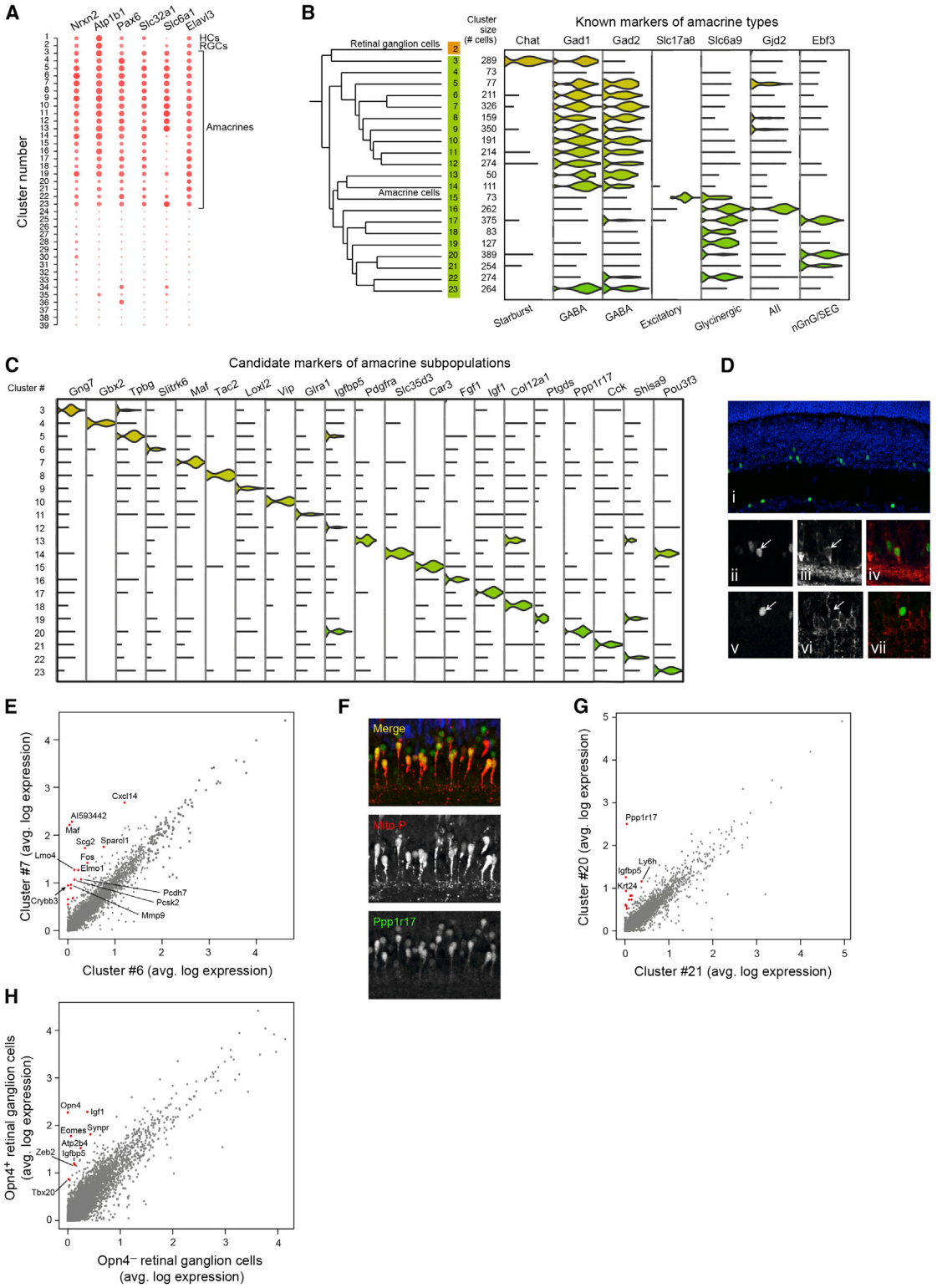


Figure 6. Finer-Scale Expression Distinctions among Amacrine Cells, Cones, and Retinal Ganglion Cells

(A) Pan-amacrine markers. The expression levels of the six genes identified (*Nrxn2*, *Atp1b1*, *Pax6*, *Slc32a1*, *Slc6a1*, *Elavl3*) are represented as dot plots across all 39 clusters; larger dots indicate broader expression within the cluster; deeper red denotes a higher expression level.

(legend continued on next page)

the MitoP line, validating this as a marker of nGnG cells (Figure 6F). PPP1R17 was one of several markers that distinguished Cluster 20 from its closest neighbor, Cluster 21 (Figure 6G; 12 genes > 2.8-fold enrichment, $p < 10^{-9}$). The differences between Clusters 20 and 21 suggest a hitherto unsuspected level of heterogeneity among nGnG amacrines.

Supervised Analysis Reveals Additional Diversity

Our unsupervised analysis grouped cells into 39 transcriptionally distinct populations, but morphological and functional criteria suggest that there are ~100 retinal cell types. We asked whether supervised analysis could reveal multiple types within individual clusters. For example, retinal ganglion cells (RGCs), which consist of about 30 types (Sanes and Masland, 2015), formed a single cluster in our analysis, perhaps because it is a rare cell population (1%, Table 1). Five RGC types, called intrinsically photosensitive RGCs (ipRGCs), express *Opn4*, the gene encoding the photopigment melanopsin. *Opn4*+ RGCs (26/432) expressed nine genes at levels 2-fold higher than *Opn4*- RGCs ($p < 10^{9}$, Figure 6H), including *Tbr2*/*Eomes*, known to be a selective marker for this population (Sweeney et al., 2014). This result reveals additional heterogeneity that may also emerge *ab initio* as analyses expand to include more cells.

DISCUSSION

Ascertaining transcriptional variation across individual cells is a valuable way of learning about complex tissues and functional responses, but single-cell analysis has been limited by the time and cost of preparing libraries from many individual cells. A scientist employing Drop-seq can prepare 10,000 single-cell libraries for sequencing in 12 hr, for about 6.5 cents per cell (Table S5), representing a >100-fold improvement in both time and cost relative to existing methods. A Drop-seq setup can be constructed quickly and inexpensively in a standard biology lab using readily available equipment (Figure S2B and Supplemental Experimental Procedures). We hope that ease, speed, and low cost facilitate exuberant experimentation, careful replication, and many cycles of experiments, analyses, ideas, and more experiments.

In validating Drop-seq, we developed stringent species-mixing experiments to measure single-cell purity and cell doublet rates in our libraries. In another article in this issue, Klein et al. (Klein et al., 2015) describe a droplet-based approach to single-cell RNA-seq and also use species-mixing experiments to evaluate it. Our results indicate that all methods of isolating single cells from a cell suspension, including Drop-seq, fluorescence activated cell sorting (FACS) and microfluidics, are vulnerable to impurities, and highlight the value of performing species mixing experiments to assess single-cell approaches. In our retina analysis, even relatively impure libraries generated in "ultra-high-throughput" modes (100 cells per μl, allowing the processing of 10,000 cells per hour at ~10% doublet and impurity rates) appeared to yield a robust and biologically validated cell classification, but other tissues or applications may require using Drop-seq in purer modes.

Unsupervised computational analysis of Drop-seq data identified 39 transcriptionally distinct retinal cell populations, many representing specific subtypes of the major retinal cell classes (Figures 5 and 6). It is a particular strength of the retina that establishing correspondence between cluster and type was in many cases straightforward; an important direction will be to identify cell types and states in other parts of the brain—as well as in other tissues—about which less is currently known.

We see many applications of Drop-seq, beyond the identification of cell types and cell states. Genome-scale genetic studies are identifying many genes whose variation contributes to disease risk, but biology has lacked similarly high-throughput ways of connecting these genes to specific cell populations and unique functional responses. Drop-seq could be used to provide initial insights into how these genes function in the diverse cell types composing each tissue. In addition, coupling Drop-seq to perturbations—such as small molecules, mutations, pathogens, or other stimuli—could generate an information-rich, multi-dimensional readout of the influence of perturbations on many kinds of cells.

The functional implications of a gene's expression are a product not just of that gene's intrinsic properties, but also of the entire cell-level context in which the gene is expressed. We hope Drop-seq enables the abundant and routine discovery of such relationships in many areas of biology.

(B) Identification of known amacrine types among clusters. The 21 amacrine clusters consisted of 12 GABAergic, five glycinergic, one glutamatergic, and three non-GABAergic non-glycinergic clusters. Starburst amacrines were identified in cluster 3 by their expression of *Chat*; excitatory amacrines by expression of *Slc17a8*; A-II amacrines by their expression of *Gjd2*; and SEG amacrine neurons by their expression of *Ebf3*.

(C) Nomination of novel candidate markers of amacrine subpopulations. Each cluster was screened for genes differentially expressed in that cluster relative to all other amacrine clusters ($p < 0.01$, Bonferroni corrected) (McDavid et al., 2013), and filtered for those with highest relative enrichment. Expression of a single candidate marker for each cluster is shown across all amacrines.

(D) Validation of MAF as a marker for a GABAergic amacrine population. Staining of a fixed adult retina from wild-type mice for MAF (i, ii, v, and green staining in iv and vii), GAD1 (iii and iv, red staining), and SLC6A9 (vi and vii, red staining), demonstrating co-localization of MAF with GAD1, but not SLC6A9.

(E) Differential expression of cluster 7 (*Maf*+) with nearest neighboring amacrine cluster (#6). Average gene expression was compared between cells in clusters 6 and 7; 16 genes (red dots) were identified with >2.8-fold enrichment in cluster 7 ($p < 10^{-9}$).

(F) Validation of PPP1R17 as a marker for an amacrine subpopulation. Staining of a fixed adult retina from Mito-P mice, which express CFP in both nGnG amacrines and type 1 bipolars (Kay et al., 2011). Overlapping labeling by PPP1R17 antibody (green) and Mito-P CFP (red) supports Drop-seq identification of *Ppp1r17* expression in the nGnG amacrine neurons. 85% of CFP+ cells were PPP1R17+ and 50% of the PPP1R17+ cells were CFP−, suggesting a second amacrine type expressing this marker. Blue staining is for VSX2, a marker of bipolar neurons.

(G) Differential expression of cluster 20 (*Ppp1r17*+) with nearest neighboring amacrine cluster (#21). Average gene expression was compared between cells in clusters 20 and 21; 12 genes (red dots) were identified with >2.8-fold enrichment in cluster 20 ($p < 10^{-9}$).

(H) Differential expression of melanopsin-positive and negative RGCs. Average expression was compared between *Opn4*-positive and -negative RGCs in cluster 2. Seven genes were identified as enriched in *Opn4*-positive cells (red dots, > 2-fold, $p < 10^{-9}$).

EXPERIMENTAL PROCEDURES

Device Design and Fabrication

Microfluidic devices were designed using AutoCAD software (Autodesk), and the components tested using COMSOL Multiphysics (COMSOL). Full details are described in Supplemental Experimental Procedures.

Barcoded Microparticle Synthesis

Bead functionalization and reverse-direction phosphoramidite synthesis were performed by Chemgenes Corp (Wilmington, MA). "Split-and-pool" cycles were accomplished by removing the dry resin from each column, hand mixing, and weighing out four equal portions before returning the resin for an additional cycle of synthesis. Full details are described in Supplemental Experimental Procedures.

Drop-Seq Procedure

Monodisperse droplets ~1 nl in size were generated using the microfluidic device described in Supplemental Experimental Procedures, in which barcoded microparticles, suspended in lysis buffer, were flowed at a rate equal to that of a single-cell suspension, so that resulting droplets were composed of an equal amount of each component. As soon as droplet generation was complete, droplets were broken with perfluorooctanol in 30 ml of 6× SSC. The addition of a large aqueous volume to the droplets reduces hybridization events after droplet breakage, because DNA base pairing follows second-order kinetics (Britten and Kohne, 1968; Wetmur and Davidson, 1968). The beads were then washed and resuspended in a reverse transcriptase mix, followed by a treatment with exonuclease I to remove unextended primers. The beads were then washed, counted, aliquoted into PCR tubes, and PCR amplified. The PCR reactions were purified and pooled, and the amplified cDNA quantified on a BioAnalyzer High Sensitivity Chip (Agilent). The cDNA was fragmented and amplified for sequencing with the Nextera XT DNA sample prep kit (Illumina) using custom primers that enabled the specific amplification of only the 3′ ends (Table S6). The libraries were purified, quantified, and then sequenced on the Illumina NextSeq 500. All details regarding reaction conditions, primers used, and sequencing specifications can be found in the Supplemental Experimental Procedures.

Cell-Cycle Analysis of HEK and 3T3 Cells

Gene sets reflecting five phases of the HeLa cell cycle (G1/S, S, G2/M, M and M/G1) were taken from Whitfield et al. (Whitfield et al., 2002) with some modification (Supplemental Experimental Procedures and Table S2). A phase-specific score was generated for each cell, across all five phases, using averaged normalized expression levels ($\log_2$(TPM+1)) of the genes in each set. Cells were then ordered along the cell cycle by comparing the patterns of these five phase scores per cell. To identify cell-cycle-regulated genes, we used a sliding window approach, and identified windows of maximal and minimal average expression, both for ordered cells, and for shuffled cells, to evaluate the false-discovery rate. Full details may be found in Supplemental Experimental Procedures.

Principal Components and Clustering Analysis of Retina Data

The clustering algorithm for the retinal cell data was implemented and performed using Seurat, a recently developed R package for single-cell analysis (Satija et al., 2015). PCA was first performed on a 13,155-cell "training set" of the 49,300-cell dataset, using single-cell libraries in which transcripts from >900 genes were detected. We found this approach was more effective in discovering structures corresponding to rare cell types than performing PCA on the full dataset, which was dominated by numerous, tiny rod photoreceptors (Supplemental Experimental Procedures). Thirty-two statistically significant PCs were identified using a permutation test and independently confirmed using a modified resampling procedure (Chung and Storey, 2015). We projected individual cells within the training set based on their PC scores onto a single two-dimensional map using t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008). The remaining 36,145 single-cell libraries (<900 genes detected) were next projected on this t-SNE map, based on their representation within the PC-subspace of the training set (Berman et al., 2014; Shekhar et al., 2014). This approach mitigates the impact of noisy variation in the lower complexity libraries due to gene dropouts. It was also reliable in the sense that when we withheld from the t-SNE all cells from a given cluster and then tried to project them, these withheld cells were not spuriously assigned to another cluster by the projection (Table S7). Point clouds on the t-SNE map represent candidate cell types; density clustering (Ester et al., 1996) identified these regions. Differential expression testing (McDavid et al., 2013) was then used to confirm that clusters were distinct from each other. Hierarchical clustering based on Euclidean distance and complete linkage was used to build a tree relating the clusters. We noted expression of several rod-specific genes, such as *Rho* and *Nrl*, in every cell cluster, an observation that has been made in another retinal cell gene expression study (Siegert et al., 2012) and likely arises from solubilization of these high-abundance transcripts during cell suspension preparation. Additional information regarding retinal cell data analysis can be found in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for the raw and analyzed data reported in this paper is GEO: GSE63473.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, seven tables, one movie, and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.05.002.

AUTHOR CONTRIBUTIONS

E.Z.M. developed the barcoding and molecular biology analysis, advised by S.A.M. A.B. designed and fabricated the microfluidic devices, advised by D.A.W. and A.R. E.Z.M. and M.G. developed Drop-seq experimental protocols and performed the Drop-seq experiments in S.A.M.'s lab. J.N. developed the methods and software for obtaining digital gene expression measurements for each cell, advised by E.Z.M. and S.A.M. J.N., E.Z.M. and S.A.M. performed the analyses of species-mixing experiments. I.T. performed the cell-cycle analysis. A.R.B. prepared the retinal cell suspensions. R.S., K.S., and A.R. developed and performed the retinal cell type clustering analyses with contribution from N.K. E.Z.M., R.S., K.S., and J.R.S. interpreted the retina expression data. E.M.M. and J.R.S. performed the immunohistochemistry experiments. J.J.T. and A.K.S. performed the Fluidigm C1 experiments. E.Z.M., S.A.M., A.R., A.B., and A.K.S. conceived the study and key ways that Drop-seq works together as an integrated system. E.Z.M. and S.A.M. wrote the manuscript with contributions from all authors.

ACKNOWLEDGMENTS

This work was supported by the Stanley Center for Psychiatric Research (to S.M.), the MGH Psychiatry Residency Research Program and Stanley-MGH Fellowship in Psychiatric Neuroscience (to E.Z.M.), a Stewart Trust Fellows Award (to S.M.), a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT (to A.R., S.M., and D.W.), an NHGRI CEGS P50 HG006193 (to A.R.), the Klarman Cell Observatory (to A.R. and A.B.), NIMH grant U01MH105960 (to S.M., A.R. and J.R.S.), NIMH grant R25MH094612 (to E.M.), NIH F32 HD075541 (to R.S.). AR is an investigator of the Howard Hughes Medical Institute. Microfluidic device fabrication was performed at the Harvard Center for Nanoscale Systems (CNS), a member of the National Nanotechnology Infrastructure Network (National Science Foundation award no. ECS-0335765), with support from the National Science Foundation (DMR-1310266) and the Harvard Materials Research Science and Engineering Center (DMR-1420570). We thank Christina Usher and Leslie Gaffney for contributions to the manuscript figures and Chris Patil for helpful comments on the manuscript. We thank Connie Cepko for helpful conversations about the retina data, Beth Stevens for advice on retinal dissociations, and Assaf Rotem and Huidan Zhang for advice on microfluidics design and fabrication. A.R. is a

member of the Scientific Advisory Board for Thermo Fisher Scientific and Syros Pharmaceuticals and a consultant for Driver Genomics.

Received: November 9, 2014
Revised: March 4, 2015
Accepted: April 30, 2015
Published: May 21, 2015

REFERENCES

Amir, A.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. *31*, 545–552.

Beer, N.R., Wheeler, E.K., Lee-Houghton, L., Watkins, N., Nasarabadi, S., Hebert, N., Leung, P., Arnold, D.W., Bailey, C.G., and Colston, B.W. (2008). On-chip single-copy real-time reverse-transcription PCR in isolated picoliter droplets. Anal. Chem. *80*, 1854–1858.

Berman, G.J., Choi, D.M., Bialek, W., and Shaevitz, J.W. (2014). Mapping the stereotyped behaviour of freely moving fruit flies. J. R. Soc. Interface *11*, 20140672.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

Britten, R.J., and Kohne, D.E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science *161*, 529–540.

Chung, N.C., and Storey, J.D. (2015). Statistical Significance of Variables Driving Systematic Variation in High-Dimensional Data. Bioinformatics *31*, 545–554.

Descamps, F.J., Martens, E., Proost, P., Starckx, S., Van den Steen, P.E., Van Damme, J., and Opdenakker, G. (2005). Gelatinase B/matrix metalloproteinase-9 provokes cataract by cleaving lens betaB1 crystallin. FASEB J. *19*, 29–35.

Ester, M., Kriegel, H.P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise (Menlo Park, Calif: AAAI Press).

Famiglietti, E.V., and Sundquist, S.J. (2010). Development of excitatory and inhibitory neurotransmitters in transitory cholinergic neurons, starburst amacrine cells, and GABAergic amacrine cells of rabbit retina, with implications for previsual and visual development of retinal ganglion cells. Vis. Neurosci. *27*, 19–42.

Feigenspan, A., Teubner, B., Willecke, K., and Weiler, R. (2001). Expression of neuronal connexin36 in AII amacrine cells of the mammalian retina. J. Neurosci. *21*, 230–239.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. *2*, 666–673.

Haverkamp, S., and Wässle, H. (2004). Characterization of an amacrine cell type of the mammalian retina immunoreactive for vesicular glutamate transporter 3. J. Comp. Neurol. *468*, 251–263.

Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., Makarewicz, A.J., Bright, I.J., Lucero, M.Y., Hiddessen, A.L., Legler, T.C., et al. (2011). High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal. Chem. *83*, 8604–8610.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

Jeon, C.J., Strettoi, E., and Masland, R.H. (1998). The major cell populations of the mouse retina. J. Neurosci. *18*, 8936–8946.

Kay, J.N., Voinescu, P.E., Chu, M.W., and Sanes, J.R. (2011). Neurod6 expression defines new retinal amacrine cell subtypes and regulates their fate. Nat. Neurosci. *14*, 965–972.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods *9*, 72–74.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single cell transcriptomics and its application to embryonic stem cells. Cell *161*, this issue, 1187–1201.

Luo, L., Callaway, E.M., and Svoboda, K. (2008). Genetic dissection of neural circuits. Neuron *57*, 634–660.

Masland, R.H. (2012). The neuronal organization of the retina. Neuron *76*, 266–280.

McDavid, A., Finak, G., Chattopadyay, P.K., Dominguez, M., Lamoreaux, L., Ma, S.S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics *29*, 461–467.

Petilla Interneuron Nomenclature Group, Ascoli, G.A., Alonso-Nanclares, L., Anderson, S.A., Barrionuevo, G., Benavides-Piccione, R., Burkhalter, A., Buzsáki, G., Cauli, B., Defelipe, J., Fairén, A., et al. (2008). Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. Nat. Rev. Neurosci. *9*, 557–568.

Picelli, S., Björklund, A.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods *10*, 1096–1098.

Sanes, J.R., and Masland, R.H. (2015). The Types of Retinal Ganglion Cells: Current Status and Implications for Neuronal Classification. Annu. Rev. Neurosci. Published online April 9, 2015.

Sanes, J.R., and Zipursky, S.L. (2010). Design principles of insect and vertebrate visual systems. Neuron *66*, 15–36.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. Published online 13 April, 2015. http://dx.doi.org/10.1038/nbt.3192.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363–369.

Shekhar, K., Brodin, P., Davis, M.M., and Chakraborty, A.K. (2014). Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). Proc. Natl. Acad. Sci. USA *111*, 202–207.

Siegert, S., Cabuy, E., Scherf, B.G., Kohler, H., Panda, S., Le, Y.Z., Fehling, H.J., Gaidatzis, D., Stadler, M.B., and Roska, B. (2012). Transcriptional code and disease map for adult retinal cell types. Nat. Neurosci. *15*, 487–495, S1–S2.

Sweeney, N.T., Tierney, H., and Feldheim, D.A. (2014). Tbr2 is required to generate a neural circuit mediating the pupillary light reflex. J. Neurosci. *34*, 5447–5453.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Thorsen, T., Roberts, R.W., Arnold, F.H., and Quake, S.R. (2001). Dynamic pattern formation in a vesicle-generating microfluidic device. Phys. Rev. Lett. *86*, 4163–4166.

Umbanhowar, P.B., Prasad, V., and Weitz, D.A. (2000). Monodisperse Emulsion Generation via Drop Break Off in a Coflowing Stream. Langmuir *16*, 347–351.

Utada, A.S., Fernandez-Nieves, A., Stone, H.A., and Weitz, D.A. (2007). Dripping to jetting transitions in coflowing liquid streams. Phys. Rev. Lett. *99*, 094502.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Vogelstein, B., and Kinzler, K.W. (1999). Digital PCR. Proc. Natl. Acad. Sci. USA *96*, 9236–9241.

Wetmur, J.G., and Davidson, N. (1968). Kinetics of renaturation of DNA. J. Mol. Biol. *31*, 349–370.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., and Botstein, D. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell *13*, 1977–2000.

Yang, Y., and Cvekl, A. (2005). Tissue-specific regulation of the mouse alphaA-crystallin gene in lens via recruitment of Pax6 and c-Maf to its promoter. J. Mol. Biol. *351*, 453–469.

Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R., and Siebert, P.D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques *30*, 892–897.

Cell

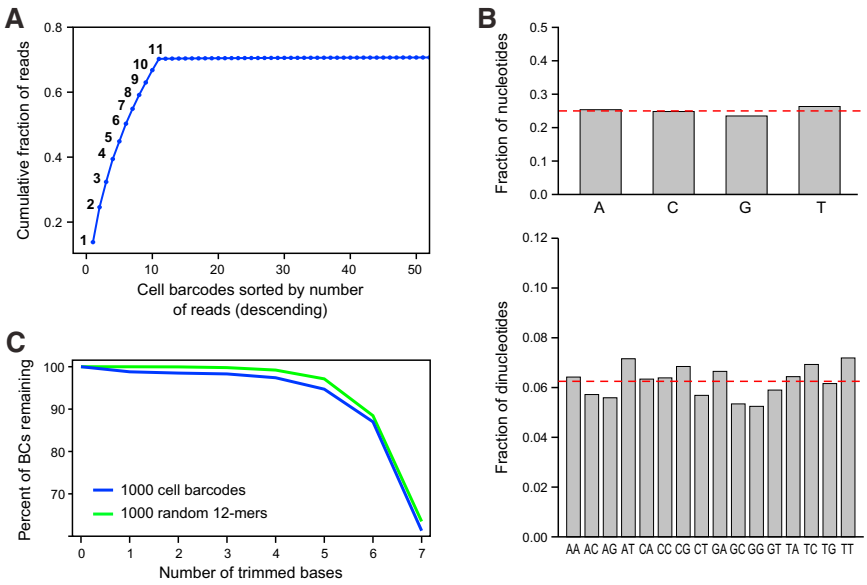


Figure S1. Assessment of the Properties of Barcoded Primers on the Surface of Microparticles, Related to Figure 1

(A) Identification of individual bead barcodes in a multiplexed experiment. A synthetic polyadenylated RNA was reverse transcribed onto the surface of barcoded primer beads. Eleven of these beads were then manually selected and used as a template for construction of a sequencing library (Supplemental Experimental Procedures). The library was sequenced on a MiSeq, and the cell barcode sequences gathered and counted. A sharp distinction was observed between the numbers of reads carrying the eleventh and twelfth most abundant 12mers at the barcode position in the sequencing read, demonstrating that cell barcodes from each bead can be recognized from their high representation in the results of a sequencing experiment.

(B) Base composition analysis of 12-bp cell barcodes. The sequences of 1,000 cell barcodes, ascertained in another sequencing experiment, were assessed for overall nucleotide and dinucleotide composition. Red dotted lines represent the values for completely random barcode sets that would lack any sequence bias.

(C) Computational truncation of 12-bp cell barcodes. The 1,000 cell barcode sequences in (B) were trimmed from the 3′ end, and the number of unique barcodes remaining was calculated at each number of trimmed bases (blue line). The number of unique barcodes at each number of trimmings was compared to a randomly generated set of 1,000 12-mers (green line).

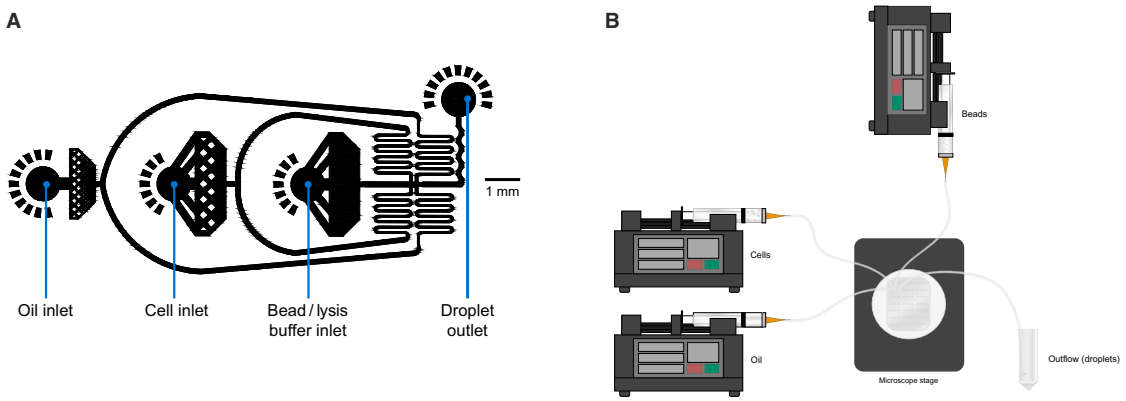


Figure S2. Schematics of Microfluidic Device Design and Operation, Related to Figure 2

(A) Microfluidic co-flow device design. Three inlets—for oil, cell suspension, and microparticles—converge and generate aqueous droplets composed of equal volume contributions from the cell suspension and microparticle channels. A bumpy outlet improves mixing of the droplets to promote hybridization of released RNAs onto the beads. A CAD file of the device can be found in Data S1.

(B) Schematic representation of Drop-seq setup. Three syringe pumps, loaded with oil, cells, and beads, respectively, are connected to the PDMS device in Figure S2A via flexible tubing. The device rests on the stage of an inverted microscope so that droplet generation can be monitored in real-time. Tubing connects the outlet channel to a 50 ml conical tube for collection of droplets.

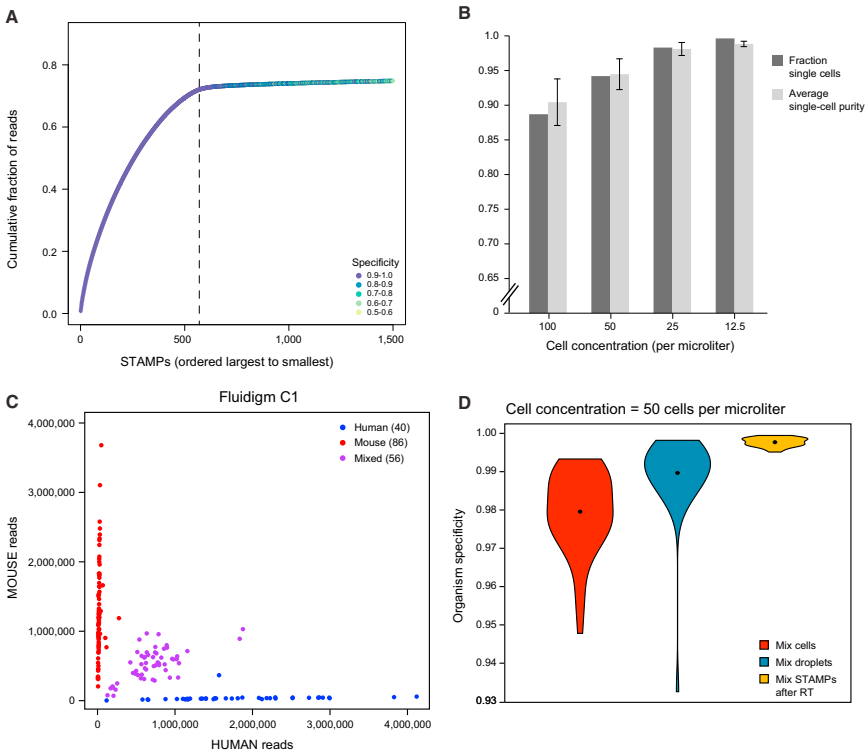


Figure S3. Dissection of Technical Contributions to Single-Cell Impurities in Drop-Seq Library Preparations, Related to Figure 3

(A) Identification of STAMPs in a pool of amplified beads. Drop-seq involves generation of single-cell profiles by diluting cells to poisson-limiting concentrations in droplets; therefore, the great majority of amplified beads (90%–99%) are not exposed to a cell's RNA, only ambient RNA. To identify the cell barcodes corresponding to STAMPs, cell barcodes from the experiment shown in Figure 3A are arranged in decreasing order of size (number of reads), and the cumulative fraction of reads is plotted. An inflection point (vertical dotted line at 570) is observed very close to the number of cells predicted by Poisson statistics for the counted and aliquoted number of beads (~500). We confirmed the significance of this inflection point by plotting the species specificity of individual STAMPs, and observing a dramatic drop in specificity near the inflection point, indicating the transition from beads that sampled cellular RNA, to the beads that sampled ambient RNA.

(B) Concentration dependence of Drop-Seq library purity. STAMPs were prepared using a mixture of human (HEK) and mouse (3T3) cells at four different concentrations (n = 1,150, 690, 595, and 570 STAMPs for 100 cells/μl, 50 cells/μl, 25 cells/μl, and 12.5 cells/μl respectively). The rate of cell doublets was calculated by multiplying by two the number of mixed species STAMPs; single-cell impurity was calculated by summing the mean human-cell and mean mouse-cell impurities. Error bars indicate SD.

(C) Human-mouse experiments on Fluidigm C1. Human (HEK) and mouse (3T3) cells were mixed at equal concentrations and run on two Fluidigm C1 chips according to the manufacturer's instructions. Reads were aligned to a joint human-mouse reference in exactly the same analysis pipeline as Drop-seq. The smallest 10 cells (with less than 100,000 reads each) were removed from analysis. Fifty-six mixed-organism libraries were identified out of 182, placing a lower bound of 30.7% on cell-cell doublets. Twelve C1 ports were identified as possessing >1 cell by microscopy, of which five were mixed species by sequencing.

(D) Single-cell impurity analysis. Drop-seq libraries were prepared from combinations of human and mouse cells pooled at three different stages of Drop-seq library preparation. In the first condition, human and mouse cells were mixed together prior to droplet formation (red violin plot, "Cell Mix"). In the second condition, human and mouse cells were separately encapsulated in droplets, which were then mixed before breaking them and performing subsequent analyses on the mixture (blue, "Droplet Mix"). In the third condition, human and mouse cells were separately encapsulated in droplets, which were broken in separate reactions and then reverse-transcribed to form separate pools of covalent STAMPs, which were mixed prior to PCR amplification (green, "PCR Mix"). The twenty largest STAMPs from each organism were selected for each of the three conditions, downsampled to the same read depth, and the organism purity represented as violin plots. The black dot is the average organism purity of the forty STAMPs in each distribution. The cell mixes used were diluted to a final concentration of 50 cells/μl in droplets. From these data we estimate that (at this cell concentration) cell suspension contributes 48% of impurities, RNA transfer after droplet breakage contributes 40%, and PCR artifacts contribute 12%.

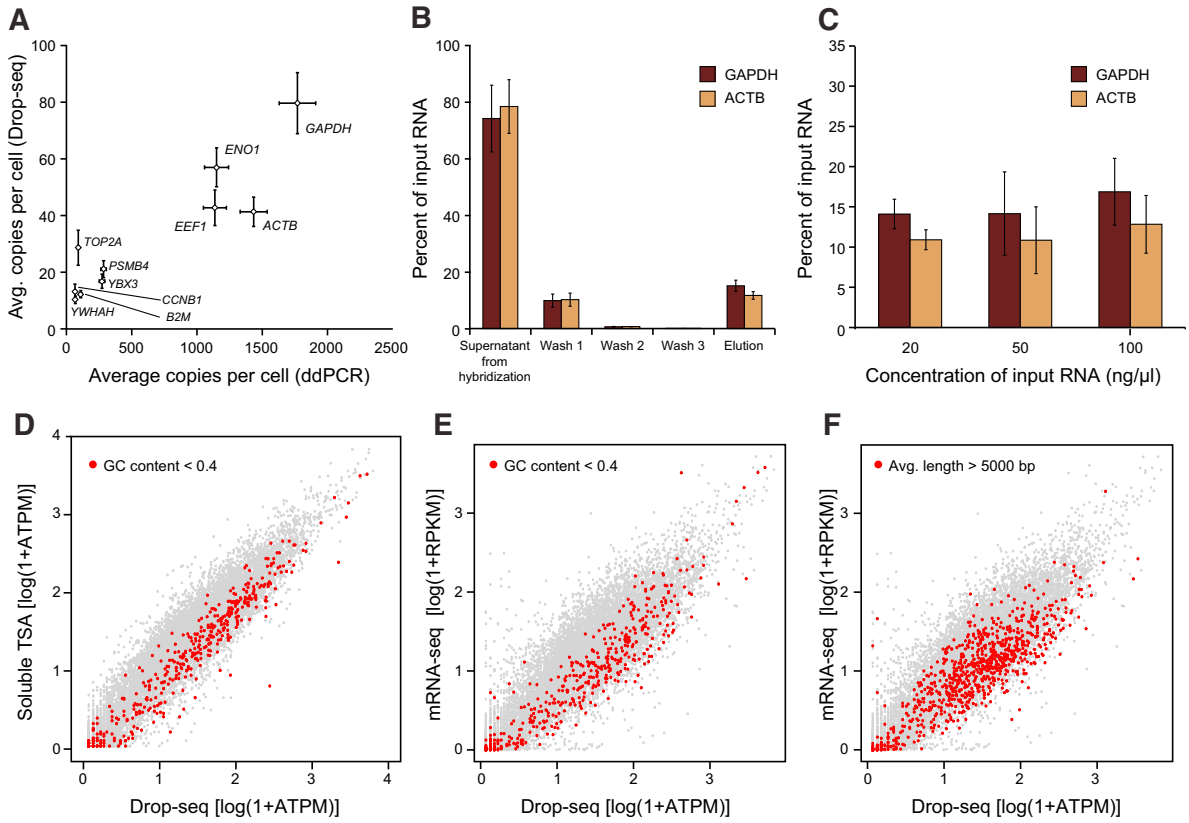


Figure S4. Estimation of Drop-Seq Expression Bias and Capture Efficiency, Related to Figure 3

(A) Sensitivity estimation by ddPCR. RNA was isolated from a culture of 50,000 HEK cells, and levels of ten genes (*ACTB, B2M, CCNB1, GAPDH, EEF2, ENO1, PSMB4, TOP2A, YBX3,* and *YWHAH*) were digitally quantitated in this bulk solution using RT-ddPCR. These transcript counts were then compared to the average number of unique transcripts counted per cell by Drop-seq. Error bars show the SE for individual ddPCR measurements (horizontal bars, n = 3 replicates) or across STAMPs (vertical bars, n = 54). Based upon the mean of these ten gene expression measurements, we estimate that Drop-seq captures approximately 10.7% of cellular mRNAs.

(B) Capture efficiency of barcoded primer beads. The same barcoded primer beads used in Drop-seq were hybridized in solution to purified human brain RNA at a concentration of 20 ng/µl (Supplemental Experimental Procedures). The beads were then spun down and washed three times, and the bound RNA eluted by heating the beads in the presence of water. The concentrations of two mRNA transcripts, GAPDH and ACTB, were measured in each of the five steps by ddPCR. Error bars, SD for three replicate experiments.

(C) Assessment of barcoded bead primer binding saturation. The same procedure described in (B) was performed using three different input RNA concentrations: 20 ng/µl, 50 ng/µl and 100 ng/µl. The fraction of input RNA that was eluted off the beads scaled linearly with input RNA concentration, indicating that hybridization to the beads was not limited by a saturation of mRNA binding sites.

(D) GC content bias between average gene expression in Drop-seq and in-solution template-switch amplification (TSA). Comparison of average gene expression in low GC content genes (< 0.4 average content, red dots) from a library of 550 3T3 STAMPs, and an mRNA-seq library prepared by an in-solution template switch amplification (TSA) procedure similar to Smart-seq2 (Picelli et al., 2013) (Supplemental Experimental Procedures), using RNA derived from an extract of the same cell culture used to provide intact cells for Drop-seq.

(E) GC content bias between average gene expression in Drop-seq and standard mRNA-seq. Comparison of average gene expression in low GC content genes (< 0.4 average content, red dots) from a library of 550 3T3 STAMPs, and an mRNA-seq library prepared by standard methods (Supplemental Experimental Procedures), using RNA derived from the same cell culture flask that was used in Drop-seq.

(F) Length bias between average gene expression in Drop-seq and standard mRNA-seq. Comparison of average gene expression in long transcripts (>5,000 average transcript length, red dots) from a library of 550 3T3 STAMPs, and an mRNA-seq library prepared by standard methods (Supplemental Experimental Procedures), using RNA derived from the same cell culture flask that was used in Drop-seq. The bias observed here was not found in a comparison of Drop-seq and in-solution TSA (data not shown), and could result from the use of template suppression PCR, which preferentially amplifies longer fragments (Matz et al., 2003).

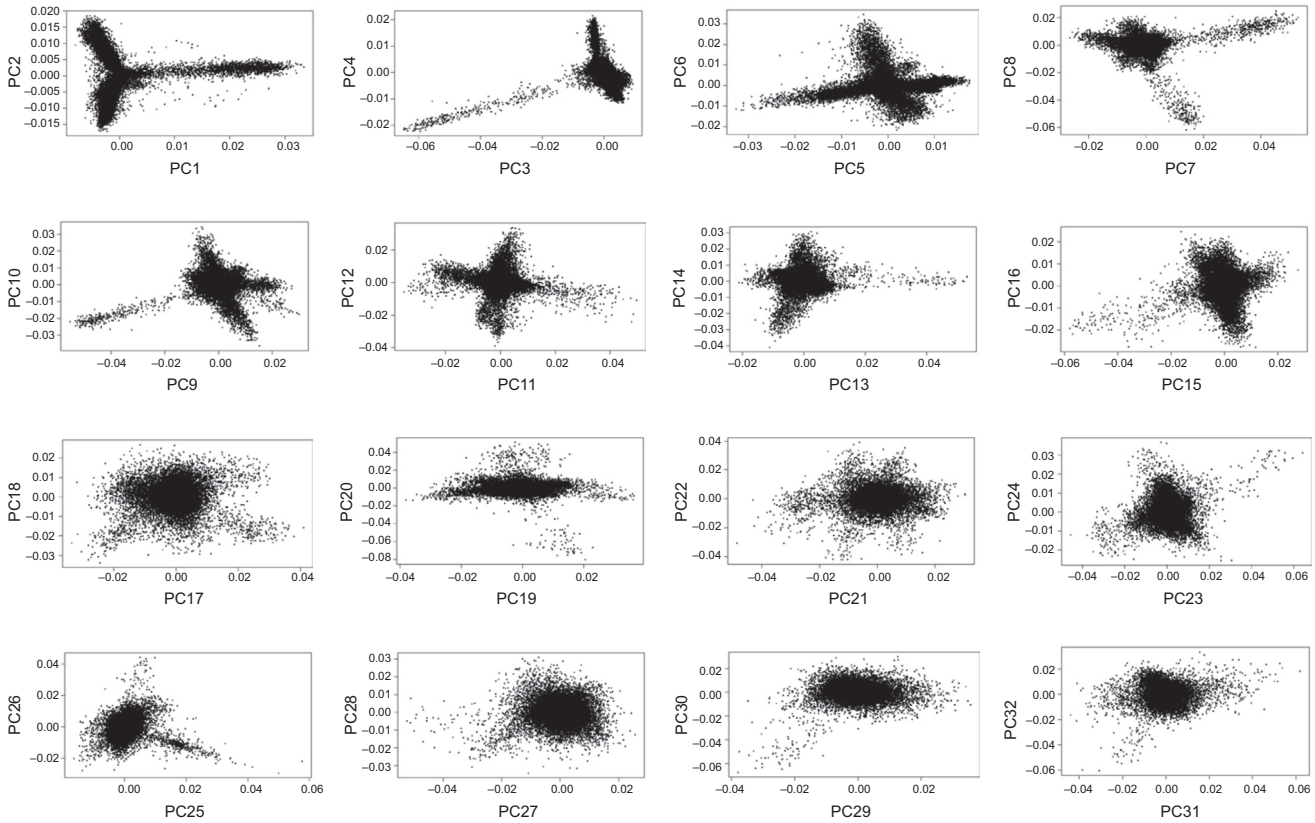


Figure S5. Plots of Principal Components 1-32 of the 44,808 Retinal Cell STAMPs Used in Analysis, Related to Figure 5

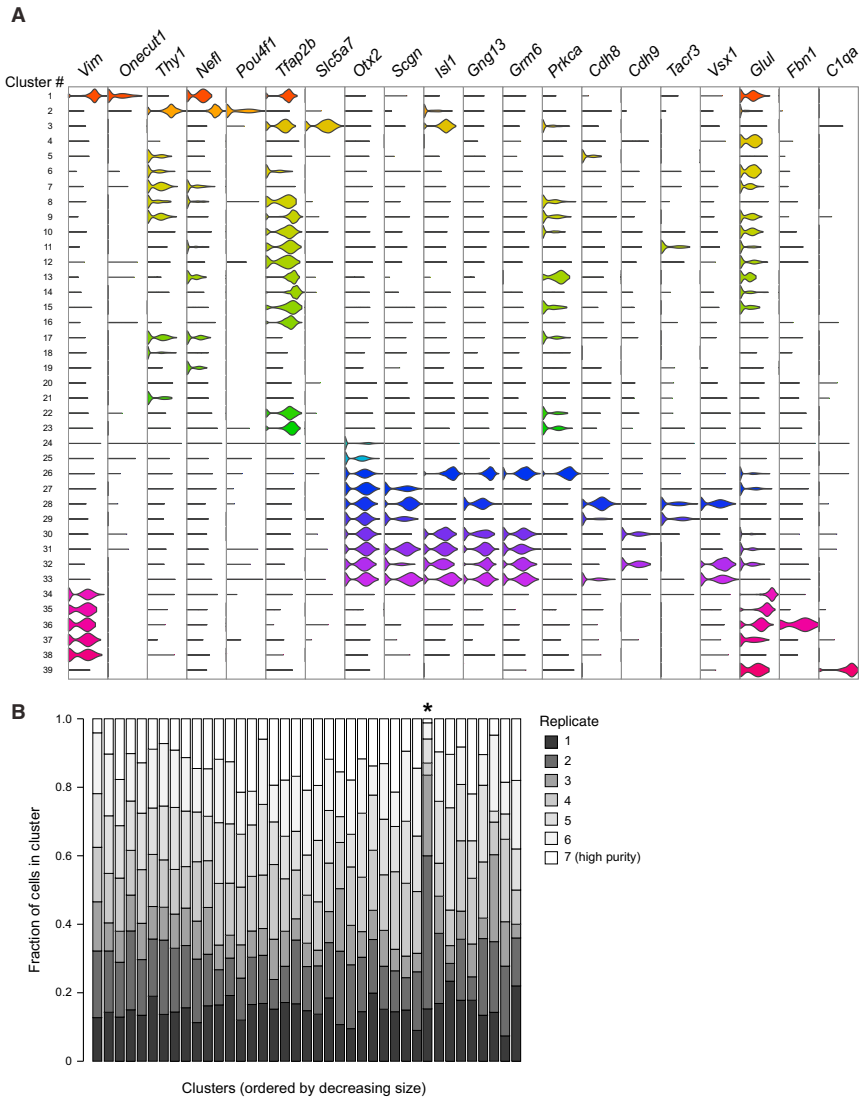


Figure S6. Expression of Additional Genes across Retinal Cell Clusters and Replicate Representation in Each Cluster, Related to Figure 5

(A) Violin plots showing expression of selected marker genes in the 39 retinal cell clusters generated by unsupervised analysis of single-cell gene expression. (B) The fraction of each cluster composed of cells deriving from one of the seven replicates (prepared over four different days, see Supplemental Experimental Procedures), that composed the full 44,808-cell dataset. The fractions of each replicate are represented as a stacked barplot. Replicates 1-6 were prepared in an "aggressive mode" of Drop-seq (~90% single-cell, ~90% purity); replicate 7 was prepared in a "pure mode" (>99% single-cell, 98.6% purity). The star designates an imbalanced cluster, #36, corresponding to contaminating fibroblasts that result from imperfect retinal dissection.

Cell

Supplemental Information

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, Steven A. McCarroll

Supplemental Experimental Procedures

Device Fabrication

Microfluidic devices were designed using AutoCAD software (Autodesk, Inc.), and the components tested using COMSOL Multiphysics (COMSOL Inc.). A CAD file is also available in (**Data S1**).

Devices were fabricated using a bio-compatible, silicon-based polymer, polydimethylsiloxane (PDMS) via replica molding using the epoxy-based photo resist SU8 as the master, as previously described (Mazutis et al., 2013; McDonald et al., 2000). The PDMS devices were then rendered hydrophobic by flowing in Aquapel (Rider, MA, USA) through the channels, drying out the excess fluid by flowing in pressurized air, and baking the device at 65ºC for 10 minutes.

Bead Synthesis

Bead functionalization and reverse direction phosphoramidite synthesis (5' to 3') were performed by Chemgenes Corp. Toyopearl HW-65S resin (~30 micron mean particle diameter) was purchased from Tosoh Biosciences (catalog #19815, Tosoh Bioscience), and surface hydroxyls were reacted with a PEG derivative to generate an 18-carbon long, flexible-chain linker. The functionalized bead was then used as a solid support for reverse-direction phosphoramidite synthesis (5'→3') on an Expedite 8909 DNA/RNA synthesizer using DNA Synthesis at 10 micromole scale and a coupling time of 3 minutes. Amidites used were: N^6-Benzoyl-*3'*-*O*-DMT-2'- deoxyadenosine-5'-cyanoethyl-*N*,*N*-diisopropyl-phosphoramidite (dA-N^6-Bz-CEP); N^4-Acetyl-3'-*O*-DMT-2'-deoxycytidine-5'-cyanoethyl-*N*,*N*-diisopropyl-phosphoramidite (dC-N^4-Ac-CEP); N^2-DMF-*3'*-*O*-DMT-2'- deoxyguanosine-5'-

cyanoethyl-*N*,*N*-diisopropyl-phosphoramidite (dG-N^2-DMF-CEP); and *3'-O*-DMT-2'- deoxythymidine-5'-cyanoethyl-*N*,*N*-diisopropyl-phosphoramidite (T-CEP). Acetic anhydride and N-methylimidazole were used in the capping step; ethylthio-tetrazole was used in the activation step; iodine was used in the oxidation step, and dichloroacetic acid was used in the deblocking step. After each of the twelve split-and-pool phosphoramidite synthesis cycles, beads were removed from the synthesis column, pooled, hand-mixed, and apportioned into four equal portions by mass; these bead aliquots were then placed in a separate synthesis column and reacted with either dG, dC, dT, or dA phosphoramidite. This process was repeated 12 times for a total of 4^12 = 16,777,216 unique barcode sequences. For complete details regarding the barcoded bead sequences used, see **Table S6.**

Cell Culture

Human 293 T cells were purchased from ATCC (cat # CRL-11268); murine NIH/3T3 cells were purchased from ATCC (cat # CRL-1658).

293T and 3T3 cells were grown in DMEM purchased from Invitrogen (cat # 11965092) supplemented with 10% FBS (Life Technologies, cat # 10437-028) and 1% penicillin-streptomycin (cat # 15070-063).

Cells were grown to a confluence of 30-60% and treated with TrypLE (Invitrogen, cat #12604013) for five min, quenched with equal volume of growth medium, and spun down at 300 x g for 5 min. The supernatant was removed, and cells were resuspended in 1 mL of 1x PBS + 0.2% BSA (Sigma cat #A8806) and re-spun at 300 x g for 3 min. The supernatant was again removed, and the cells re-suspended in 1 mL of 1x PBS, passed through a 40-micron cell strainer (Falcon, VWR cat #21008-949), and counted. For Drop-Seq, cells were diluted to the final concentration in 1x PBS + 200 μg/mL BSA (NEB, cat # B9000S).

Generation of Whole Retina Suspensions

Single-cell suspensions were prepared from P14 mouse retinas by adapting previously described methods for purifying retinal ganglion cells from rat retina (Barres et al., 1988). Briefly, mouse retinas were digested in a papain solution (40U papain / 10mL DPBS) for 45 minutes. Papain was then neutralized in a trypsin inhibitor solution (0.15% ovomucoid in DPBS) and the tissue was triturated to generate a single-cell suspension. Following trituration, the cells were pelleted, resuspended, and filtered through a 20μm Nitex mesh filter to eliminate any clumped cells. The cells were then diluted in DPBS + 0.2% BSA (Sigma #A8806) to either 200 cells / μL (replicates 1-6) or 30 cells / μL (replicate 7).

Retina suspensions were processed through Drop-Seq on four separate days. One library was prepared on day 1 (replicate 1); two libraries on day 2 (replicates 2 and 3); three libraries on day 3 (replicates 4-6); and one library on day 4 (replicate 7, high purity). To replicates 4-6, human HEK cells were spiked in at a concentration of 1 cell / μL (0.5%) but the wide range of cell sizes in the retina data made it impossible to calibrate single-cell purity or doublets by cross-species comparison. Each of the seven replicates was sequenced separately.

Experiments were approved by the institutional animal use and care committee at Harvard Medical School in accordance with NIH guidelines for the humane treatment of animals.

Drop-Seq

Preparation of beads

Beads (either Barcoded Bead SeqA or Barcoded Bead SeqB; **Table S6** and see note at end of **Supplemental Experimental Procedures)** were washed twice with 30 mL of 100% EtOH and twice with 30 mL of TE/TW (10 mM Tris pH 8.0, 1 mM EDTA, 0.01% Tween). The bead pellet was resuspended in 10 mL TE/TW and passed through a 100 μm filter (BD Falcon, cat # 352360) into a 50 mL Falcon tube for long-term storage at 4 ^{o}C. The stock concentration of beads (in beads/μL) was assessed using a Fuchs-Rosenthal cell counter purchased from INCYTO (cat # DHC-F01). For Drop-Seq, an aliquot of beads was removed from the stock tube, washed in 500 μL of Drop-Seq Lysis Buffer (DLB, 200 mM Tris pH 7.5, 6% Ficoll PM-400, 0.2% Sarkosyl, 20 mM EDTA), then resuspended in the appropriate volume of DLB + 50 mM DTT for a bead concentration of ~120 beads/μL.

Droplet Generation

The two aqueous suspensions—the single-cell suspension and the bead suspension—were loaded into 3 mL plastic syringes (BD cat #309657). To the bead syringe, we added a 6.4 mm magnetic stir disc (V&P Scientific, VP cat # 782N-6-150). Droplet generation oil (Biorad, cat # 186-4006) was loaded into a 10 mL plastic syringe (BD #309604). The three syringes were connected to a 125 μm co-flow device (**Figure S2A)** by 0.38 mm inner-diameter polyethylene tubing (Scientific Commodities, inc cat # BB31695-PE/2), and injected using syringe pumps (KD Scientific, Legato 100) at flow rates of 4.1 mL/hr for each aqueous suspension, and 14 mL/hr for the oil, resulting in ~125 μm emulsion drops with a volume of ~1 nanoliter each. For movie generation, the flow was visualized under an optical microscope (Olympus IX83) at 10x magnification and imaged at ~1000-2000 frames per second using a FASTCAM SA5 color camera (Photron, Japan). Droplets were collected in 50 mL falcon tubes; the collection tube was changed out after every 1 mL of combined aqueous flow volume.

During droplet generation, the beads were kept in suspension by continuous, gentle magnetic stirring (V&P Scientific, cat # VP710D2). The uniformity in droplet size and the occupancy of beads were evaluated by observing aliquots of droplets under an optical microscope with bright-field illumination; in each experiment, greater than 95% of the bead-occupied droplets contained a single bead.

Droplet Breakage

The oil from the bottom of each aliquot of droplets was removed with a P1000 pipette, after which 30 mL 6X SSC (Life Technologies, cat # 15557-036) at room temperature was added.

To break droplets, we added 600 μL of Perfluoro-1-octanol (Sigma-Aldrich, cat # 370533-25G), and shook the tube vigorously by hand for about 20 seconds. The tube was then centrifuged for 1 minute at 1000 x g. To reduce the likelihood of annealed mRNAs dissociating from the beads, samples were kept on ice for the remainder of the breakage protocol. The supernatant was removed to roughly 5 mL above the oil-aqueous interface, and the beads washed with an additional 30 mL of room temperature 6X SSC, the aqueous layer transferred to a new tube, and centrifuged again. The supernatant was removed, and the bead pellet transferred to non-stick 1.5 mL microcentrifuge tubes (VWR, cat # 20170-650). The pellet was then washed twice with 1 mL 6X SSC, and once with 300 μL of 5x Maxima H-RT buffer (EP0751).

Reverse Transcription and Exonuclease I Treatment

To a pellet of up to 90,000 beads, 200 μL of RT mix was added, where the RT mix contained 1x Maxima RT buffer, 4% Ficoll PM-400 (GE Healthcare, cat # 17-0300-05), 1 mM dNTPs (Clontech, cat # 639125), 1 U/μL Rnase Inhibitor (Lucigen, cat # 30281-2), 2.5 μM Template_Switch_Oligo (**Table**

S6), and 10 U/μL Maxima H- RT (ThermoScientific cat #EP0751). The beads were incubated at room temperature for 30 minutes, followed by 42 ^{o}C for 90 minutes. The beads were then washed once with 1 mL 1x TE + 0.5% Sodium Dodecyl Sulfate (TE/SDS, Sigma cat# L4522), twice with 1 mL TE/TW, and once with 10 mM Tris pH 7.5. The bead pellet was then resuspended in 200 μL of exonuclease I mix containing 1x Exonuclease I Buffer and 1 U/μL Exonuclease I (NEB cat # B0293S), and incubated at 37 ^{o}C for 45 minutes.

The beads were then washed once with 1 mL TE/SDS, twice with 1 mL TE/TW, once with 1 mL ddH_2O, and resuspended in ddH_2O. Bead concentration was determined using a Fuchs-Rosenthal cell counter. Aliquots of 1000 beads were amplified by PCR in a volume of 50 μL using 1x Hifi HotStart Readymix (Kapa Biosystems, cat #KK2602) and 0.8 μM Template_Switch_PCR primer (**Table S6)**.

The aliquots were thermocycled as follows: 95 ^{o}C 3 min; then four cycles of: 98 ^{o}C for 20 sec, 65 ^{o}C for 45 sec, 72 ^{o}C for 3 min; then X cycles of: 98 ^{o}C for 20 sec, 67 ^{o}C for 20 sec, 72 ^{o}C for 3 min; then a final extension step of 5 min. For the human-mouse experiment using cultured cells, X was 8 cycles; for the dissociated retina experiment, X was 9 cycles. Pairs of aliquots were pooled together after PCR and purified with 0.6x Agencourt AMPure XP beads (Beckman Coulter, cat # A63881) according to the manufacturer's instructions, and eluted in 10 μL of H_2O. Aliquots were pooled according to the number of STAMPs to be sequenced, and the concentration of the pool quantified on a BioAnalyzer High Sensitivity Chip (Agilent Technologies, cat # 5067-4626).

Preparation of Drop-Seq cDNA Library for Sequencing

To prepare 3'-end cDNA fragments for sequencing, four aliquots of 600 pg of cDNA were used as input in four standard Nextera XT tagmentation reactions (Illumina, cat #FC-131-1096), performed

according to the manufacturer's instructions except that 200 nM of the custom primers P5_TSO_Hybrid and Nextera_N701 (**Table S6**) were used in place of the kit's provided oligonucleotides. The samples were then amplified as follows: 95 ^{o}C for 30 sec; 11 cycles of 95 ^{o}C for 10 sec, 55 ^{o}C for 30 sec, 72 ^{o}C for 30 sec; then a final extension step of 72 ^{o}C for 5 min.

Pairs of the 4 aliquots were pooled together, and then purified using 0.6x Agencourt AMPure XP Beads according to the manufacturer's instructions, and eluted in 10 μL of water. The two 10 μL aliquots were combined together and the concentration determined using a BioAnayzer High Sensitivity Chip. The average size of sequenced libraries was between 450 and 650 bp.

The libraries were sequenced on the Illumina NextSeq 500 using 4.67 pM in a volume of 3 mL HT1, and 3 mL of 0.3 μM Read1CustSeqA or Read1CustSeqB (**Table S6** and see note at the end of **Supplemental Experimental Procedures**) for priming of read 1. Read 1 was 20 bp (bases 1-12 cell barcode, bases 13-20 UMI); read 2 (paired end) was 50 bp for the human-mouse experiment, and 60 bp for the retina experiment.

Species Contamination Experiment

To determine the origin of off-species contamination of STAMP libraries (**Figure S3D**), we: (1) performed Drop-Seq exactly as above (control experiment) with a HEK/3T3 cell suspension mixture of 100 cells / μL in concentration; (2) performed the microfluidic co-flow step with HEK and 3T3 cells separately, each at a concentration of 100 cells / μL, and then mixed droplets prior to breakage; and (3) performed STAMP generation through exonuclease digestion, with the HEK and 3T3 cells separately, then mixed equal numbers of STAMPs prior to PCR amplification. A single 1000 microparticle aliquot was amplified for each of the three conditions, then purified and quantified on a BioAnalyzer High

Sensitivity DNA chip. 600 pg of each library was used in a single Nextera Tagmentation reaction as described above, except that each of the three libraries was individually barcoded with the primers Nextera_N701 (condition 1), Nextera_N702 (condition 2), or Nextera_N703 (condition 3), and a total of 12 PCR cycles were used in the Nextera PCR instead of 11. The resulting library was quantified on a High Sensitivity DNA chip, and each was loaded at a concentration of 8 pM on a single, multiplexed MiSeq run using 0.5 μM Read1CustSeqA as a custom primer for read 1 (see note at end of this section).

Soluble RNA Experiments

To quantify the number of primer annealing sites, 20,000 beads were incubated with 10 μM of polyadenylated synthetic RNA (synRNA, **Table S6**) in 2x SSC for 5 min at room temperature, and washed three times with 200 μL of TE-TW, then resuspended in 10 μL of TE-TW. The beads were then incubated at 65 ^{o}C for 5 minutes, and 1 μL of supernatant was removed for spectrophotometric analysis on the Nanodrop 2000. The concentration was compared with beads that had been treated the same way, except no synRNA was added.

To determine whether the bead-bound primers were capable of reverse transcription, and to measure the homogeneity of the cell barcode sequence on the bead surface, beads were washed with TE-TW, and added at a concentration of 100 / μL to the reverse transcriptase mix described above. This mix was then co-flowed into the standard Drop-Seq 125 μm co-flow device with 200 nM SynRNA in 1x PBS + 0.02% BSA. Droplets were collected and incubated at 42 ^{o}C for 30 minutes. 150 μL of 50 mM EDTA was added to the emulsion, followed by 12 μL of perfluorooctanoic acid to break the emulsion. The beads were washed twice in 1 mL TE-TW, followed by one wash in H_2O, then resuspended in TE. Eleven beads were handpicked under a microscope into a 50 μL PCR mix containing 1x Kapa HiFi Hotstart PCR mastermix, 400 nM P7-TSO_Hybrid, and 400 nM TruSeq_F (**Table S6**). The PCR

reaction was cycled as follows: 98 ^{o}C for 3 min; 12 cycles of: 98 ^{o}C for 20 s, 70 ^{o}C for 15 s, 72 ^{o}C for 1 min; then a final 72 ^{o}C incubation for 5 min. The resulting amplicon was purified on a Zymo DNA Clean and Concentrator 5 column, and run on a BioAnalyzer High Sensitivity Chip to estimate concentration. The amplicon was then sequenced on an Illumina MiSeq at a final concentration of 6 pM. Read 1, primed using the standard Illumina TruSeq primer, was a 20 bp molecular barcode on the SynRNA, while Read 2, primed with CustSynRNASeq, contained the 12 bp cell barcode and 8 bp UMI.

To estimate the efficiency of Drop-Seq, we used a set of external RNAs (ERCC Spike-ins, Life Technologies #4456740). We diluted the ERCC spike-ins to 0.32% of the stock in 1x PBS + 1 U/μL RNase Inhibitor (Lucigen) + 200 μg/ mL BSA (NEB), and used this in place of the cell flow in the Drop-Seq protocol, so that each bead was incubated with ~100,000 ERCC mRNA molecules per nanoliter droplet. Sequence reads were aligned to a dual ERCC-human (hg19) reference, using the human sequence as "bait," which dramatically reduced the number of low-quality alignments to ERCC transcripts reported by STAR compared with alignment to an ERCC-only reference.

Standard mRNA-Seq and In-Solution Template Switch Amplification

To compare Drop-Seq average expression data to standard mRNAseq data, we used 1.815 ug of purified RNA from 3T3 cells, from which we also prepared and sequenced 550 STAMPs. The RNA was used in the TruSeq Stranded mRNA Sample Preparation kit (Illumina, # RS-122-2101) according to the manufacturer's instructions. For NextSeq 500 sequencing, 0.72 pM of Drop-Seq library was combined with 0.48 pM of the mRNAseq library in a final volume of 3 mL Buffer HT1.

To compare Drop-Seq average expression data to mRNAseq libraries prepared by a standard, in-solution template switch amplification approach, 5 ng of the same purified 3T3 RNA used above was

diluted in 2.75 μL of H2O. To the RNA, 1 μL of 10 μM UMI_SMARTdT primer was added (**Table S6**) and heated to 72 C, followed by incubation at 4 C for 1 min, after which we added 2 μL 20% Ficoll PM-400, 2 μL 5x RT Buffer (Maxima H- kit), 1 μL 10 mM dNTPs (Clontech), 0.5 μL 50 μM Template_Switch_Oligo (**Table S6**), and 0.5 μL Maxima H- RT. The RT was incubated at 42 C for 90 minutes, followed by heat inactivation for 5 min at 85 C. An RNase cocktail (0.5 μL RNase I, Epicentre N6901K, and 0.5 μL RNase H, Life Tech 18021071) was added to remove the terminal riboGs from the template switch oligo, and the sample incubated for 30 min at 37 C. Then, 0.4 μL of 100 μM Template_Switch_PCR primer was added, along with 25 μL 2x Kapa Hifi supermix, and 13.6 μL H2O. The sample was cycled as follows: 95 C 3 min; 14 cycles of: 98 C 20 s, 67 C 20 s, and 72 C 3 min; then 72 C 5 min. The samples were purified with 0.6x AMPure XP beads according to the manufacturer's instructions, and eluted in 10 μL H2O. 600 pg of amplified cDNA was used as input into a Nextera XT reaction. 0.6 pM of library was sequenced on a NextSeq 500, multiplexed with three other samples; Read1CustSeqB was used to prime read 1.

Droplet Digital PCR (ddPCR) Experiments

To quantify the efficiency of Drop-Seq (**Figure S4A**), 50,000 HEK cells, prepared in an identical fashion as in Drop-Seq, were pelleted and RNA purified using the Qiagen RNeasy Plus Kit according to the manufacturer's protocol. The eluted RNA was diluted to a final concentration of 1 cell-equivalent per microliter in an RT-ddPCR reaction containing RT-ddPCR supermix (BioRad, # 186-3021), and a gene primer-probe set. Droplets were produced using BioRad ddPCR droplet generation system, and thermocycled with the manufacturer's recommended protocol, and droplet fluorescence analyzed on the BioRad QX100 droplet reader. Concentrations of RNA and confidence intervals were computed by BioRad QuantaSoft software. Three replicates of 50,000 HEK cells were purified in parallel, and the

concentration of each gene in each replicate was measured two independent times. The probes (Life Technologies #4331182) used were: ACTB (hs01060665_g1), B2M (hs00984230_m1), CCNB1 (mm03053893), EEF2 (hs00157330_m1), ENO1 (hs00361415_m1), GAPDH (hs02758991_g1), PSMB4 (hs01123843_g1), TOP2A (hs01032137_m1), YBX3 (hs01124964_m1), and YWHAH (hs00607046_m1).

To estimate the RNA hybridization efficiency of Drop-Seq (**Figures S4B** and **S4C**), human brain total RNA (Life Technologies #AM7962) was diluted to 40 ng / μL in a volume of 20 μL and combined with 20 μL of barcoded primer beads resuspended in Drop-Seq lysis buffer (DLB, composition shown above) at a concentration of 2,000 beads / μL. The solution was incubated at 15 minutes with rotation, then spun down and the supernatant transferred to a fresh tube. The beads were washed 3 times with 100 μL of 6x SSC, resuspended in 50 μL H2O, and heated to 72 C for 5 min to elute RNA off the beads. The elution step was repeated once and the elutions pooled. All steps of the hybridization (RNA input, hybridization supernatant, three washes, and combined elution) were separately purified using the Qiagen RNeasy Plus Mini Kit (cat #74134) according to the manufacturers' instructions. Various dilutions of the elutions were used in RT-ddPCR reactions with primers and probes for either ACTB or GAPDH.

Fluidigm C1 Experiments

C1 experiments were performed as previously described (Shalek et al., 2014). Briefly, suspensions of 3T3 and HEK cells were stained with calcein violet and calcein orange (Life Technologies) according to the manufacturer's recommendations, diluted down to a concentration of 250,000 cells per mL, and mixed 1:1. This cell mixture was then loaded into two medium C1 cell capture chips from Fluidigm and, after loading, caught cells were visualized and identified using DAPI and TRITC fluorescence. Bright

field images were used to identify ports with > 1 cell (a total of 14 were identified from the two C1 chips used, out of 192 total). After C1-mediated whole transcriptome amplification, libraries were made using Nextera XT (Illumina), and loaded on a NextSeq 500 at 2.2 pM. Single-read sequencing (60 bp) was performed to mimic the read structure in DropSeq, and the reads aligned as per below. Ten of the 192 cells, containing fewer than 100,000 reads per cell, were excluded from analysis.

Read Alignment and Generation of Digital Expression Data

Raw sequence data was first filtered to remove all read pairs with a barcode base quality of less than 10. The second read (50 or 60 bp) was then trimmed at the 5' end to remove any TSO adapter sequence, and at the 3' end to remove polyA tails of length 6 or greater, then aligned to either the mouse (mm10) genome (retina experiments) or a combined mouse (mm10) –human (hg19) mega-reference (species mixing experiments), using STAR v2.4.0a with the default settings.

Uniquely mapped reads were grouped by cell barcode. To digitally count gene transcripts, a list of UMIs in each gene, within each cell, was assembled, and UMIs within ED = 1 were merged together. The total number of distinct UMI sequences was counted, and this number was reported as the number of transcripts of that gene for a given cell.

To generate the digital expression matrices in this paper, we performed UMI merging at ED=1, including insertions and deletions. However, a subsequent comparison of UMI edit distance relationships within and across genes showed that inclusion of indels resulted in excessive merging (**Table S1**). For our ERCC sensitivity analysis, we therefore used substitution-only UMI merging, and plan to also use this approach in future experiments. Without any edit distance correction (or using the corrective approach described in Islam et al., 2014), we obtained an efficiency estimate of 47% for the

ERCC dataset shown in **Figure 3G**, though we believe (from the analysis in **Table S1**) that for our data, our own correction approach, and the lower capture-rate estimate derived from it, are more accurate.

To distinguish cell barcodes arising from STAMPs, rather than those that corresponded to beads never exposed to cell lysate, we ordered our digital expression matrix by the total number of transcripts per cell barcode, and plotted the cumulative fraction of all transcripts in the matrix for each successively smaller cell barcode. Empirically, our data always displays a "knee" at a cell barcode number close to the estimated number of STAMPs amplified (**Figure S3A**). All cell barcodes larger than this cutoff were used in downstream analysis, while the remaining cell barcodes were discarded.

Cell Cycle Analysis of HEK and 3T3 Cells

Gene sets reflecting five phases of the HeLa cell cycle (G1/S, S, G2/M, M and M/G1) were taken from Whitfield et al. (Whitfield et al., 2002) (**Table S2**), and refined by examining the correlation between the expression pattern of each gene and the average expression pattern of all genes in the respective gene-set, and excluding genes with a low correlation ($R<0.3$). This step removed genes that were identified as phase-specific in HeLa cells but did not correlate with that phase in our single-cell data. The remaining genes in each refined gene-set were highly correlated (not shown). We then averaged the normalized expression levels ($\log_2(TPM+1)$) of the genes in each gene-set to define the phase-specific scores of each cell. These scores were then subjected to two normalization steps. First, for each phase, the scores were centered and divided by their standard deviation. Second, the normalized scores of each cell were centered and normalized.

To order cells according to their progression along the cell cycle, we first compared the pattern of phase-specific scores of each cell to eight potential patterns along the cell cycle: only G1/S is on, both G1/S and S, only S, only G2/M, G2/M and M, only M, only M/G1, M/G1 and G1. We also added a ninth pattern for equal scores of all phases (either all active or all inactive). Each pattern was defined simply as a vector of ones for active programs and zeros for inactive programs. We then classified the cells by the defined patterns based on the maximal correlation of the phase-specific scores with these potential patterns. Importantly, none of the cells were classified to the ninth pattern of equal activity, while multiple cells were assigned to each of the other patterns. To further order the cells within each class, we sorted the cells based on their relative correlation with the preceding and succeeding patterns, thereby smoothing the transitions between classes (**Figure 4A**).

To identify cell cycle-regulated genes we used the cell cycle ordering defined above and a sliding window approach with a window size of 100 cells. We identified the windows with maximal average expression and minimal average expression for each gene and used a two-sample t-test to assign an initial p-value for the difference between maximal and minimal windows. A similar analysis was performed after shuffling the order of cells to generate control p-values that can be used to evaluate false-discovery rate (FDR). Specifically, we examined for each potential p-value threshold, how many genes pass that threshold in the cell cycle ordered and in the randomly ordered analyses to assign FDR. Genes were defined as being previously known to be cell-cycle regulated if they were included in a cell cycle GO/KEGG/REACTOME gene set, or reported in a recent genome-wide study of gene expression in synchronized replicating cells (Bar-Joseph et al., 2008).

Unsupervised Dimensionality Reduction and Clustering Analysis of Retina Data

P14 mouse retina suspensions were processed through Drop-Seq in seven different replicates on four separate days, and each sequenced separately. Raw digital expression matrices were generated for the seven sequencing runs. The inflection points in the cumulative distribution plot, corresponding to the number of cells in each sample replicate, were: 6,600, 9,000, 6,120, 7,650, 7,650, 8280, and 4000. The full 49,300 cells were merged together in a single matrix, and normalized by dividing by the total number of UMIs per cell, then multiplying by 10,000. All calculations and data were then performed in log space (i.e. ln(transcripts-per-10,000 +1)).

Initial Downsampling and Identification of Highly Variable Genes

Rod photoreceptors constitute 60-70% of the retinal cell population. Furthermore, they are significantly smaller than other retinal cell types (Carter-Dawson and LaVail, 1979), and as a result yielded significantly fewer genes (and higher levels of noise) in our single cell data. In our preliminary computational experiments, performing unsupervised dimensionality reduction on the full dataset resulted in representations that were dominated by noisy variation within the numerous rod subset; this compromised our ability to resolve the heterogeneity within other cell-types that were comparatively much rarer (e.g. amacrines, microglia). Thus, to increase the power of unsupervised dimensionality reduction techniques for discovering these types we first downsampled the 49,300-cell dataset to extract single-cell libraries where 900 or more genes were detected, resulting in a 13,155-cell "training set". We reasoned that this "training set" would be enriched for rare cell types that are larger in size at the expense of "noisy" rod cells. The remaining 36,145 cells (henceforth "projection set") were then directly embedded onto the two-dimensional representation learned from the training set (see below). This enabled us to leverage the full statistical power of our data to define and annotate cell types.

We first identified the set of genes that was most variable across our training set, after controlling for the relationship between mean expression and variability. We calculated the mean and a dispersion

measure (variance/mean) for each gene across all 13,155 single cells, and placed genes into 20 bins based on their average expression. Within each bin, we then z-normalized the dispersion measure of all genes within the bin, in order to identify outlier genes whose expression values were highly variable even when compared to genes with similar average expression. We used a z-score cutoff of 1.7 to identify 384 highly variable genes.

Principal Components Analysis

We ran Principal Components Analysis (PCA) on our training set as previously described (Shalek et al., 2013), using the prcomp function in R, after scaling and centering the data along each gene. We used only the previously identified "highly variable" genes as input to the PCA in order to ensure robust identification of the primary structures in the data.

While the number of principal components returned is equal to the number of profiled cells, only a small fraction of these components explain a statistically significant proportion of the variance, as compared to a null model. We used two approaches to identify statistically significant PCs for further analysis: (1) we performed 10000 independent randomizations of the data such that within each realization, the values along every row (gene) of the scaled expression matrix are randomly permuted. This operation randomizes the pairwise correlations between genes while leaving the expression distribution of every gene unchanged. PCA was performed on each of these 10000 "randomized" datasets. Significant PCs in the un-permuted data were identified as those with larger eigenvalues compared to the highest eigenvalues across the 10000 randomized datasets ($p < 0.01$, Bonferroni corrected). (2) We modified a randomization approach ('jack straw') proposed by Chung and Storey (Chung and Storey, 2014) and which we have previously applied to single-cell RNA-seq data (Shalek et al., 2014). Briefly, we performed 1,000 PCAs on the input data, but in each analysis, we randomly 'scrambled' 1% of the genes to empirically estimate a null distribution of scores for every gene. We

used the joint-null criterion (Leek and Storey, 2011) to identify PCs that had gene scores significantly different from the respective null distributions ($p<0.01$, Bonferroni corrected). Both (1) and (2) yielded 32 'significant' PCs. Visual inspection confirmed that none of these PCs was primarily driven by mitochondrial, housekeeping, or hemoglobin genes. As expected, markers for distinct retinal cell types were highly represented among the genes with the largest scores (+ve and –ve) along these PCs (**Table S3**).

t-SNE Representation and Post-Hoc Projection of Remaining Cells

Because canonical markers for different retinal cell types were strongly represented along the significant PCs (**Figure S5),** we reasoned that the loadings for individual cells in our training set along the principal eigenvectors (also "PC subspace representation") could be used to separate out distinct cell types in our data. We note that these loadings leverage information from the 384 genes in the PCA, and therefore are more robust to technical noise than single-cell measurements of individual genes. We used these PC loadings as input for t-Distributed Stochastic Neighbor Embedding (tSNE) (van der Maaten and Hinton, 2008), as implemented in the tsne package in R with the "perplexity" parameter set to 30. The t-SNE procedure returns a two-dimensional embedding of single cells. Cells with similar expression signatures of genes within our variable set, and therefore similar PC loadings, will likely localize near each other in the embedding, and hence distinct cell types should form two-dimensional point clouds across the tSNE map.

Prior to identifying and annotating the clusters, we projected the remaining 36,145 cells (the projection set) onto the tSNE map of the training set by the following procedure:

(1) We projected these cells onto the subspace defined by the significant PCs identified from the training set. Briefly, we centered and scaled the 384 x 36,145 expression matrix corresponding

to the projection set, considering only the highly variable genes; the scaling parameters of the training set were used to center and scale each row. We then multiplied the transpose of this scaled expression matrix with the 384 x 32 gene scores matrix learned from the training set PCA. This yields a PC "loading" for the cells in the projection set along the 32 significant PCs learned on the training set.

(2) Based on its PC loadings, each cell in the projection set was independently embedded on to the tSNE map of the training set introduced earlier using a mathematical framework consistent with the original tSNE algorithm (Shekhar et al., 2014). We note that while this approach does not discover novel clusters outside of the ones identified from the training set, it sharpens the distinctions between different clusters by leveraging the statistical power of the full dataset. Moreover, the cells are projected based on their PC signatures, not the raw gene expression values, which makes our approach more robust against technical noise in individual gene measurements.

See section *"Embedding the projection set onto the tSNE map"* below for full details.

One potential concern with this "post-hoc projection approach" was the possibility that a cell type that is completely absent from the training set might be spuriously projected into one of the defined clusters. We tested our projection algorithm on a control dataset to explore this possibility, and placed stringent conditions to ensure that only cell types adequately represented within the training set are projected to avoid spurious assignments (see *'"Out of sample" projection test'*). Using this approach, 97% of the cells in the projection set were successfully embedded, resulting in a tSNE map consisting of 48296 out of 49300 sequenced cells (**Table S7**).

As an additional validation of our approach, we note that the relative frequencies of different cell types

identified after clustering the full data (see below) closely matches estimates in the literature (**Table 1**). With the exception of the rods, all the other cell types were enriched at a median value of 2.3X in the training set compared to their frequency of the full data. This strongly suggests that our downsampling approach indeed increases the representation of other cell types at the expense of the rod cells, enabling us to discover PCs that define these cells.

Density Clustering to Identify Cell-Types

To identify putative cell types on the tSNE map, we used a density clustering approach implemented in the DBSCAN R package (Ester et al., 1996), initially setting the reachability distance parameter (eps) to 1.0, and removing clusters less than 20 cells, then setting eps to 1.9, and removing clusters less than 50 cells. The first step (eps=1) resulted in an over-partitioning of the data, but enabled us to easily identify and remove singleton cells that were located along the interfaces of bigger clusters. Following this "pruning" step, we re-clustered the data with a larger eps value (1.9) to identify a smaller set of 49 clusters involving 44808 cells (91% of our data) with each cluster containing at least 50 cells. This two-step pruning strategy enabled us to avoid over-partitioning of the data, while at the same time suppress the co-option of outlier cells into a neighboring cluster. The 49 clusters were further interrogated through stringent differential expression tests (see below).

We next examined the 49 total clusters to ensure that our identified clusters truly represented distinct cellular classifications, as opposed to over-partitioning. We performed a *post-hoc* test where we searched for differentially expressed genes (McDavid et al., 2013) between every pair of clusters (requiring at least 10 genes, each with an average expression difference greater than 1 natural log value between clusters with a Bonferroni corrected $p<0.01$). We iteratively merged cluster pairs that did not satisfy this criterion, starting with the two most related pairs (lowest number of differentially expressed genes). This process resulted in 10 merged clusters, leaving 39 remaining.

We then computed average gene expression for each of the 39 remaining clusters, and calculated Euclidean distances between all pairs, using this data as input for complete-linkage hierarchical clustering and dendrogram assembly. We then compared each of the 39 clusters to the remaining cells using a likelihood-ratio test (McDavid et al., 2013) to identify marker genes that were differentially expressed in the cluster.

Embedding the Projection Set onto the tSNE Map

We used the computational approach in Shekhar et al. (Shekhar et al., 2014) and Berman et al. (Berman et al., 2014) to project new cells onto an existing tSNE map. First, the expression vector of the cell is reduced to include only the set of highly variable genes, and subsequently centered and scaled along each gene using the mean and standard deviation of the gene expression in the training set. This scaled expression vector *z* (dimensions 1 x 384) is multiplied with the scores matrix of the genes *S* (dimensions 384 x 32), to obtain its "loadings" along the significant PCs *u* (dimensions 1 x 32). Thus,

$$u' = z'.S$$

u (dimensions 1 x 32) denotes the representation of the new cell in the PC subspace identified from the training set. We note a point of consistency here in that performing the above dot product on a scaled expression vector of a cell *z* taken from the training set recovers its correct subspace representation *u*, as it ought to be the case.

Given the PC loadings of the cells in the training set $\{u^i\}$ $(i=1,2,\ldots N_{train})$ and their tSNE coordinates $\{y^i\}$ $(i=1,2,\ldots N_{train})$, the task now is to find the tSNE coordinates *y'* of the new cell based on its loadings vector *u'*. As in the original tSNE framework (van der Maaten and Hinton, 2008), we "locate" the new cell in the subspace relative to the cells in the training set by computing a set of transition probabilities,

$$p(u'|u^i) = \frac{\exp\left(-d(u',u^i)^2/2\sigma_{u'}^2\right)}{\sum_{\{u^i\}}\exp(-d(u',u^i)^2/2\sigma_{u'}^2)}$$

Here, *d(. , .)* represents Euclidean distances, and the the bandwidth $\sigma_{u'}$ is chosen by a simple binary search in order to constrain the Shannon entropy associated with $p(u'|u^i)$ to $\log_2(30)$, where 30 corresponds to the value of the perplexity parameter used in the tSNE embedding of the training set. Note that $\sigma_{u'}$ is chosen independently for each cell.

A corresponding set of transition probabilities in the low dimensional embedding are defined based on the Student's t-distribution as,

$$q(y'|y^i) = \frac{\left(1 + d(y', y^i)^2\right)^{-1}}{\sum_{\{y^i\}}(1 + d(y', y^i)^2)^{-1}}$$

where *y'* are the coordinates of the new cell that are unknown. We calculate these by minimizing the Kullback-Leibler divergence between $p(u'|u^i)$ and $q(y'|y^i)$,

$$y' = \ argmin \sum_i p(u'|u^i) \log \frac{p(u'|u^i)}{q(y'|y^i)}$$

This is a non-convex objective function with respect to its arguments, and is minimized using the Nelder-Mead simplex algorithm, as implemented in the Matlab function fminsearch. This procedure can be parallelized across all cells in the projection set.

A few notes on the implementation,

1. Since this is a post-hoc projection, and $p(u'|u^i)$ is only a relative measure of pairwise similarity in that it is always constrained to sum to 1, we wanted to avoid the possibility of new cells being embedded on the tSNE map by virtue of their high relative similarity to one or two training cells ("short circuiting"). In other words, we chose to project only those cells that were drawn from regions of the PC subspace that were well represented in the training set by at least a few cells.

   Thus, we retained a cell *u'* for projection only if $p(u'|u^i) > p_{thres}$ was true for at least N_{min} cells in the training set ($p_{thres} = 5 \times 10^{-3}$, $N_{min} = 10$). We calibrated the values for p_{thres} and

N_{min} by testing our projection algorithm on cases where the projection set was known to be completely different from the training set to ensure that such cells were largely rejected by this constraint. (see Section '"*Out of sample" projection test'*)

2. For cells that pass the constraint in pt. 1., the initial value of the tSNE coordinate y'_0 is set to,

$$y'_0 = \sum_i p(u'|u^i)y^i$$

i.e. a weighted average of the tSNE coordinates of the training set with the weights set to the pairwise similarity in the PC subspace representation.

3. A cell satisfying the condition in 1. is said to be "successfully projected" to a location y'* when a minimum of the KL divergence could be found within the maximum number of iterations. However since the program is non-convex and is guaranteed to only find local minima, we wanted to explore if a better minima could be found. Briefly, we uniformly sampled points from a 25 x 25 grid centered on y'* to check for points where the value of the KL-divergence was within 5% of its value at y'* or lower. Whenever this condition was satisfied (< 2%) of the time, we re-ran the optimization by setting the new point as the initial value.

"Out of Sample" Projection Test

In order to test our post-hoc projection method, we conducted the following computational experiment wherein each of the 39 distinct clusters on the tSNE map was synthetically "removed" from the tSNE map, and then reprojected cell-by-cell on the tSNE map of the remaining clusters using the procedure outlined above. Only cells from the training set were used in these calculations.

Assuming our cluster distinctions are correct, in each of these 39 experiments, the cluster that is being reprojected represents an "out of sample" cell type. Thus successful assignments of these cells into one of the remaining 38 clusters would be spurious. For each of the 39 clusters that was removed and reprojected, we classified the cells into three groups based on the result of the projection method:

(1) Cells that did not satisfy the condition 1. in the previous section (i.e. did not have a high relative similarity to at least N_{min} training cells), and therefore "failed" to project.

(2) Cells that were successfully assigned a tSNE coordinate y', but that could not be assigned into any of the existing clusters according to the condition below.

(3) Cells that were successfully assigned a tSNE coordinate y', and which were "wrongly assigned" to one of the existing clusters. A cell was assigned to a cluster whose centroid was closest to y' if and only if the distance between y' and the centroid was smaller than the cluster radius (the distance of the farthest point from the centroid).

Encouragingly for all of the 39 "out of sample" projection experiments, only a small fraction of cells were spuriously assigned to one of the clusters, i.e. satisfied (3) above with the parameters p_{thres} = 5×10^{-3} and N_{min} = 10 (**Table S7**). This gave us confidence that our post-hoc embedding of the projection set would not spuriously assign distinct cell types into one of the existing clusters.

Downsampling Analyses of Retina Data

To generate the 500-cell and 2000-cell downsampled tSNE plots shown in **Figure 5F**, the largest 500 or 2000 cells were sampled from the high-purity replicate (replicate 7), and used as input for PCA and tSNE. Two extreme outlier points were removed from the 500-cell tSNE prior to plotting. To generate the 9,731-cell downsampled tSNE plot, 10,000 cells were randomly sampled from the full dataset, and the cells expressing transcripts from more than 900 genes were used in principal components analysis and tSNE; the remaining (smaller) cells were projected onto the tSNE embedding.

Immunohistochemistry

Wild-type C57 mice or Mito-P mice, which express CFP in nGnG amacrine and Type 1 bipolar cells (Kay et al., 2011), were euthanized by intraperitoneal injection of pentobarbital. Eyes were fixed in 4% PFA in PBS on ice for one hour, followed by dissection and post-fixation of retinas for an additional 30 minutes, then rinsed with PBS. Retinas were frozen and sectioned at 20 μm in a cryostat. Sections were incubated with primary antibodies (chick anti-GFP [Abcam], rabbit anti-PPP1R17 [Atlas], or goat anti-VSX2 [Santa Cruz]) overnight at 4°C, and with secondary antibodies (Invitrogen and Jackson ImmunoResearch) for 2 hours at room temperature. Sections were then mounted using Fluoromount G (Southern Biotech) and viewed with an Olympus FVB confocal microscope.

Note on Bead Surface Primers and Custom Sequencing Primers

During the course of experiments for this paper, we used two batches of beads that had two slightly different primer sequences (Barcoded Bead SeqA and Barcoded Bead SeqB, **Table S6)**. Barcoded Bead SeqA was used in the human-mouse experiments, and in replicates 1-3 of the retina experiment. Replicates 4-7 were performed with Barcoded Bead SeqB. To prime read 1 for Drop-Seq libraries produced using Barcoded Bead SeqA beads, Read1CustSeqA was used; to prime read 2 for Drop-Seq libraries produced using Barcoded Bead SeqB beads, Read1CustSeqB was used. ChemGenes plans to manufacture beads harboring the Barcoded Bead SeqB sequence. These beads should be used with Read1CustSeqB.

Additional Notes Regarding Drop-Seq Implementation

Cell and Bead Concentrations

Our experiments have shown that the cell concentration used in Drop-Seq has a strong, linear relationship to the purity and doublet rates of the resulting libraries (**Figures 3A, 3B, and S3B**). Cell concentration also linearly affects throughput: ~10,000 single-cell libraries can be processed per hour when cells are used at a final concentration of 100 cells / μL, and ~1,200 can be processed when cells are used at a final concentration of 12.5 cells / μL. The trade-off between throughput and purity is likely to affect users differently, depending on the specific scientific questions being asked. Currently, for our standard experiments, we use a final concentration of 50 cells / μL, tolerating a small percentage of doubles and cell contaminants, to be able to easily and reliably process 10,000 cells over the course of a couple of hours. As recommended above, we currently favor loading beads at a concentration of 120 / μL (final concentration in droplets = 60 / μL), which empirically yields a < 5% bead doublet rate.

Drop-Seq Start-Up Costs

The main pieces of equipment required to implement Drop-Seq are three syringe pumps (KD Legato 100 pumps, list price ~$2,000 each) a standard inverted microscope (Motic AE31, list price ~$1,900), and a magnetic stirrer (V&P scientific, #710D2, list price ~$1,200). A fast camera (used to monitor droplet generation in real time) is not necessary for the great majority of users (droplet quality can be monitored by simply placing 3 μL of droplets in a Fuchs-Rosenthal hemocytometer with 17 μL of droplet generation oil to dilute the droplets into a single plane of focus).

Table S1. Analysis of edit distance relationships among UMIs, Related to Figure 3

UMI Sampling	% Reduction in UMI counts	
	Substitution-only collapse	Indel and substitution collapse
Within a gene	68.2%	76.1%
Across genes	19.1%	45.7%

Edit distance relationships among UMIs. For the data in **Figure 3G**, the sequences of the UMIs for each ERCC gene detected in each cell barcode were collapsed at an edit distance of 1, including only substitutions (left column) or with both substitutions and insertions/deletions (right column). A control UMI set was prepared for each gene, using an equal number of UMIs sampled randomly across all genes/cells. The table shows the percent of the original UMIs that were collapsed for each condition.

Table S5. Cost Analysis of Drop-Seq, Related to Figure 5

Reagents	Supplier	Catalog #	Cost for 10,000 cells ($)
Microfluidics costs (tubing, syringes, droplet generation oil, device fabrication)	N/A	N/A	35.00
DropSeq lysis buffer (Ficoll, Tris, Sarkosyl, EDTA, DTT)	N/A	N/A	9.35
Barcoded microparticles	Chemgenes	N/A	137.20
Maxima H– Reverse Transcriptase	Thermo	EP0753	59.15
dNTP mix	Clontech	639125	7.78
RNase inhibitor	Lucigen	30281-2	3.80
Template switch oligo	IDT	N/A	7.60
Perfluorooctanol	Sigma	370533	11.90
Exonuclease I	NEB	M0293L	3.84
KAPA Hifi HotStart ReadyMix	KAPA BioSystems	KK2602	210.00
Nextera XT DNA sample preparation kit	Illumina	FC-131-1096	120.80
Ampure XP beads	Beckman Coulter	A63882	37.35
BioAnalyzer High Sensitivity Chips	Agilent	5067-4626	9.64
Total cost:			**$653.41**
Cost per cell:			**$0.065**

Table S6. Oligonucleotide Sequences Used in This Study

synRNA	rCrCrUrArCrArCrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrNrNrNrNrNrNrNrNrNrNrNrNrNrNrBrArArArArArArArArArArArArArArArArArArArArA
Barcoded Bead SeqA	5′ –Bead–Linker-TTTTTTTAAGCAGTGGTATCAACGCAGAGTACGTJJJJJJJJJJJNNNNNNNN TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT-3′
Barcoded Bead SeqB	5′ –Bead–Linker-TTTTTTTAAGCAGTGGTATCAACGCAGAGTACJJJJJJJJJJJNNNNNNNN TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT-3′
Template_Switch_Oligo	AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG
TSO_PCR	AAGCAGTGGTATCAACGCAGAGT
P5-TSO_Hybrid	AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*A*C
Nextera_N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG
Nextera_N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGG
Nextera_N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGG
Read1CustomSeqA	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTACGT
Read1CustomSeqB	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC
P7-TSO_Hybrid	CAAGCAGAAGACGGCATACGAGATCGTGATCGGTCTCGGCGGAAGCAGTGGTATCAACGCAGAGT*A*C
TruSeq_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
CustSynRNASeq	CGGTCTCGGCGGAAGCAGTGGTATCAACGCAGAGTAC
UMI_SMARTdT	AAGCAGTGGTATCAACGCAGAGTACNNNNNNNNNTTTTTTTTTTTTTTTTTTTTTTT

Table S7. "Out-of-Sample" Projection Test

Cluster #	# Cells in Cluster	# failed to project	# Projected	# Wrongly Assigned	% Wrongly Assigned
1	153	153	0	0	0.00
2	271	271	0	0	0.00
3	201	201	0	0	0.00
4	46	46	0	0	0.00
5	63	62	1	0	0.00
6	173	156	17	9	5.20
7	277	272	5	5	1.81
8	115	115	0	0	0.00
9	275	275	0	0	0.00
10	155	153	2	2	1.29
11	165	162	3	3	1.82
12	175	175	0	0	0.00
13	46	40	6	5	10.87
14	89	89	0	0	0.00
15	52	44	8	6	11.54
16	179	179	0	0	0.00
17	284	284	0	0	0.00
18	64	63	1	1	1.56
19	108	107	1	0	0.00
20	206	206	0	0	0.00
21	154	154	0	0	0.00
22	180	180	0	0	0.00
23	183	182	1	1	0.55
24	3712	3417	295	180	4.85
25	1095	1071	24	18	1.64
26	1213	1212	1	0	0.00
27	323	318	5	4	1.24
28	339	330	9	7	2.06
29	332	324	8	6	1.81
30	447	426	21	18	4.03
31	346	340	6	3	0.87
32	235	233	2	2	0.85
33	453	450	3	3	0.66
34	784	784	0	0	0.00
35	27	27	0	0	0.00
36	43	43	0	0	0.00
37	145	139	6	5	3.45
38	30	30	0	0	0.00
39	17	17	0	0	0.00

For each cluster, the "training" cells were removed from the tSNE plot, and then projected onto the tSNE. The number of cells that successfully projected into the embedding, and the number of cells that were inappropriately incorporated into a different cluster were tabulated.

References

Bar-Joseph, Z., Siegfried, Z., Brandeis, M., Brors, B., Lu, Y., Eils, R., Dynlacht, B.D., and Simon, I. (2008). Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. Proceedings of the National Academy of Sciences of the United States of America *105*, 955-960.

Barres, B.A., Silverstein, B.E., Corey, D.P., and Chun, L.L. (1988). Immunological, morphological, and electrophysiological variation among retinal ganglion cells purified by panning. Neuron *1*, 791-803.

Berman, G.J., Choi, D.M., Bialek, W., and Shaevitz, J.W. (2014). Mapping the stereotyped behaviour of freely moving fruit flies. Journal of the Royal Society, Interface / the Royal Society *11*.

Carter-Dawson, L.D., and LaVail, M.M. (1979). Rods and cones in the mouse retina. I. Structural analysis using light and electron microscopy. The Journal of comparative neurology *188*, 245-262.

Chung, N.C., and Storey, J.D. (2014). Statistical Significance of Variables Driving Systematic Variation in High-Dimensional Data. Bioinformatics.

Ester, M., Kriegel, H.P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. (Menlo Park, Calif.: AAAI Press).

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lonnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nature methods *11*, 163-166.

Kay, J.N., Voinescu, P.E., Chu, M.W., and Sanes, J.R. (2011). Neurod6 expression defines new retinal amacrine cell subtypes and regulates their fate. Nature neuroscience *14*, 965-972.

Leek, J.T., and Storey, J.D. (2011). The joint null criterion for multiple hypothesis tests. Applications in Genetics and Molecular Biology *10*, 1-22.

Matz, M.V., Alieva, N.O., Chenchik, A., and Lukyanov, S. (2003). Amplification of cDNA ends using PCR suppression effect and step-out PCR. Methods in molecular biology *221*, 41-49.

Mazutis, L., Gilbert, J., Ung, W.L., Weitz, D.A., Griffiths, A.D., and Heyman, J.A. (2013). Single-cell analysis and sorting using droplet-based microfluidics. Nature protocols *8*, 870-891.

McDavid, A., Finak, G., Chattopadyay, P.K., Dominguez, M., Lamoreaux, L., Ma, S.S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics *29*, 461-467.

McDonald, J.C., Duffy, D.C., Anderson, J.R., Chiu, D.T., Wu, H., Schueller, O.J., and Whitesides, G.M. (2000). Fabrication of microfluidic systems in poly(dimethylsiloxane). Electrophoresis *21*, 27-40.

Picelli, S., Bjorklund, A.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature methods *10*, 1096-1098.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., *et al.* (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236-240.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., *et al.* (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363-369.

Shekhar, K., Brodin, P., Davis, M.M., and Chakraborty, A.K. (2014). Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). Proceedings of the National Academy of Sciences of the United States of America *111*, 202-207.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research *9*, 2579-2605.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., *et al.* (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Molecular biology of the cell *13*, 1977-2000.