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.