# Massively parallel single-nucleus RNA-seq with DroNc-seq

Naomi Habib[1–3,8], Inbal Avraham-Davidi[1,8], Anindita Basu[1,4,7,8], Tyler Burks[1], Karthik Shekhar[1], Matan Hofree[1], Sourav R Choudhury[2,3], François Aguet[2], Ellen Gelfand[2], Kristin Ardlie[2], David A Weitz[4,5], Orit Rozenblatt-Rosen[1], Feng Zhang[2,3] & Aviv Regev[1,6]

**Single-nucleus RNA sequencing (sNuc-seq) profiles RNA from tissues that are preserved or cannot be dissociated, but it does not provide high throughput. Here, we develop DroNc-seq: massively parallel sNuc-seq with droplet technology. We profile 39,111 nuclei from mouse and human archived brain samples to demonstrate sensitive, efficient, and unbiased classification of cell types, paving the way for systematic charting of cell atlases.**

Single cell RNA-seq (scRNA-seq) has become instrumental for interrogating cell types, dynamic states, and functional processes in complex tissues[1,2]. However, the current requirement for single-cell suspensions to be prepared from fresh tissue is a major roadblock to assessing clinical samples, archived materials, and tissues that cannot be readily dissociated. The harsh enzymatic dissociation needed for brain tissue is particularly problematic because it harms the integrity of neuronal RNA, biases proportions of recovered cell types, and only works on samples from younger organisms, which precludes the use of, for example, those from deceased patients with neurodegenerative disorders. To address this challenge, we[3] and others[4–6] developed methods to analyze RNA in single nuclei from fresh, frozen, or lightly fixed tissues. Methods such as sNuc-Seq[3], Div-Seq[3], and others[4,5] can handle minute samples of complex tissues that cannot be dissociated, thereby providing access to archived samples. However, these methods rely on sorting nuclei by FACS into 96- or 384-well plates[3,5] or on C1 microfluidics[4], neither of which scales to tens of thousands of nuclei (needed for human brain tissue) or large numbers of samples (for example, tumor biopsies from patients). Conversely, massively parallel scRNA-seq methods, such as Drop-seq[7] and related methods[8–10], can be readily applied at scale[11] in a cost-effective manner[12] but require intact single-cell suspension as input.

Here, we develop DroNc-seq (**Supplementary Fig. 1a**), a massively parallel single-nucleus RNA-seq method that combines the advantages of sNuc-seq and Drop-seq to profile nuclei at low cost and high throughput. We modified Drop-seq[7] to accommodate the lower amount of RNA in nuclei compared to cells, including a modified microfluidic design and changes in the nuclei isolation protocol (**Supplementary Fig. 1**, **Supplementary Table 1**, **Supplementary Data 1**, and Online Methods).

We used DroNc-seq to robustly generate high-quality expression profiles of nuclei from a mouse cell line (3T3, 5,636 nuclei), adult frozen mouse brain tissue (19,561 nuclei), and archived frozen adult human post-mortem tissue (19,550 nuclei). DroNc-seq (for samples sequenced at 160,000 reads per nucleus, Online Methods) detected on average 3,295 genes (4,643 transcripts) for 3T3 nuclei, 2,731 genes (3,653 transcripts) for mouse brain, and 1,683 genes (2,187 transcripts) for human brain (**Supplementary Fig. 2**). Using down sampling, we estimate that 19,000–26,000 transcriptome-mapped reads per nucleus are required for saturation (**Supplementary Fig. 2f**,**g**).

To assess throughput and sensitivity, we sequenced single 3T3 cells (with Drop-seq) and nuclei (with DroNc-seq) deeply to ~160,000 reads per nucleus or cell. Both methods yielded high-quality libraries, detecting an average of 5,134 and 3,295 genes for cells and nuclei, respectively (**Supplementary Fig. 2b**,**c**). DroNc-seq had similar throughput to that of Drop-seq with efficiencies of 78% for 3T3 nuclei, 89% for mouse brain, and 95% for human brain (1,003, 1,251, and 1,333 high-quality nuclei per library out of 1,400 expected nuclei, given our loading parameters for cell lines, mouse brain, and human brain, respectively), compared to 72% high-quality cells per library (1,444 nuclei out of 2,000 expected) (Online Methods). Notably, libraries were sampled from a pool of 20,000 STAMPs (single transcriptome-associated microparticles[7]), which can be resampled multiple times if a user wishes to sequence additional nuclei from the same input (Online Methods).

The average expression profile of single nuclei correlated well with that of single cells (Pearson $r = 0.87$, **Supplementary Fig. 2d**). Expression profiles for genes with significantly higher expression in nuclei (such as those encoding lncRNAs Malat1 and Meg3) or cells (mitochondrial genes *mt-Nd1*, *mt-Nd2*, and

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]McGovern Institute, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. [5]Department of Physics, Harvard University, Cambridge, Massachusetts, USA. [6]Howard Hughes Medical Institute, Department of Biology, Koch Institute of Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [7]Present address: Department of Medicine, University of Chicago, Chicago, Illinois, USA, and Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to F.Z. (zhang@broadinstitute.org) or A.R. (aregev@broadinstitute.org).

RECEIVED 3 MARCH; ACCEPTED 13 JULY; PUBLISHED ONLINE 28 AUGUST 2017; DOI:10.1038/NMETH.4407

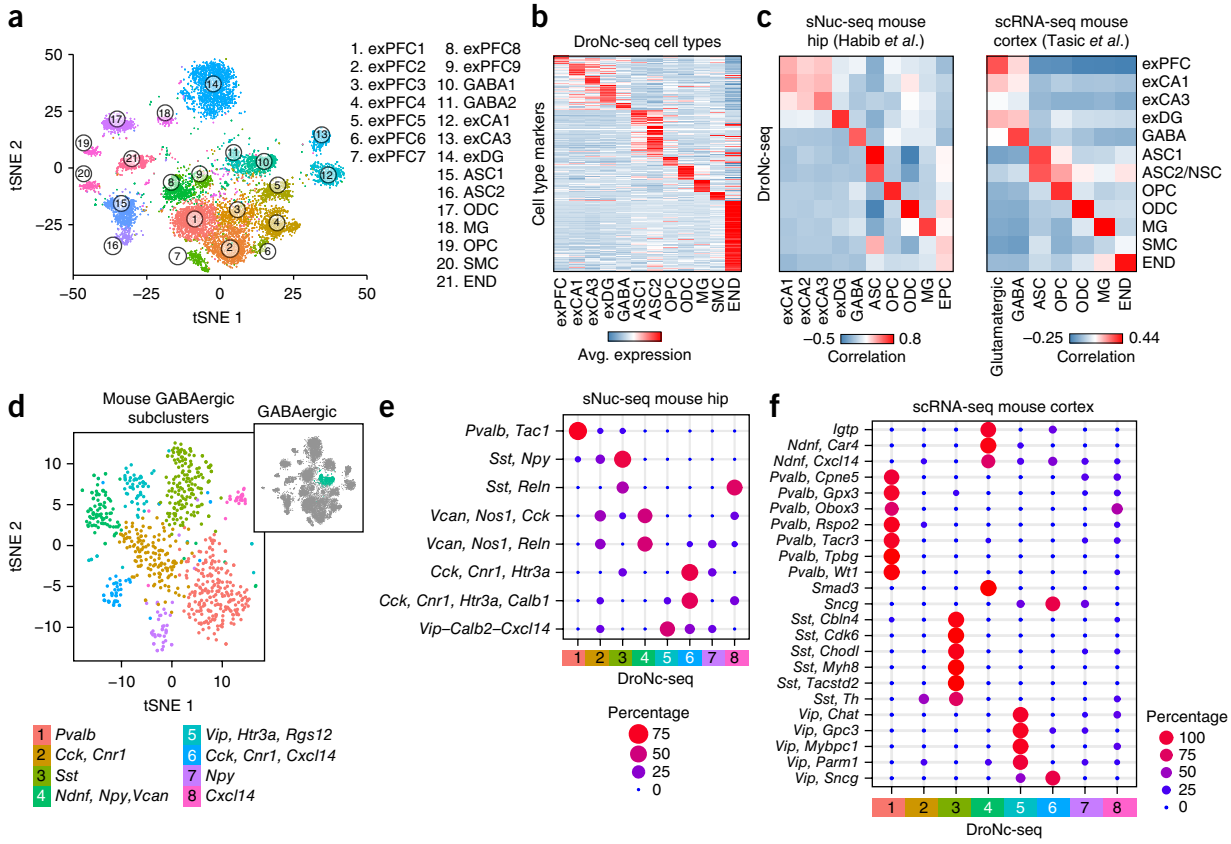


**Figure 1** | DroNc-seq: massively parallel sNuc-seq. (**a**) DroNc-seq of adult frozen mouse hippocampus and prefrontal cortex. A tSNE plot of 13,133 DroNc-seq nuclei profiles (>10,000 reads and >200 genes per nucleus) from hippocampus (hip; 4 samples) and prefrontal cortex (PFC; 4 samples). Nuclei (dots) are colored by cluster membership and labeled *post hoc* according to cell type and anatomical distinctions. exPFC, glutamatergic neurons from the PFC; GABA, GABAergic interneurons; exCA1/3, pyramidal neurons from the hip CA region; exDG, granule neurons from the hip dentate gyrus region; ASC, astrocytes; MG, microglia; ODC, oligodendrocytes; OPC, oligodendrocyte precursor cells; NSC, neuronal stem cells; SMC, smooth muscle cells; END, endothelial cells. Clusters are grouped by cell type as in **Supplementary Figure 3a**. Flagged clusters (**Supplementary Fig. 3b** and **Supplementary Table 3**, Online Methods) were removed. (**b**) Cell-type signatures. The average expression of differentially expressed signature genes (rows, Online Methods) in each DroNc-seq mouse brain cell subset (columns). (**c**) DroNc-seq cell-type expression signatures in the mouse brain agree with previous studies. Pairwise correlations of the average expression (Online Methods) for the genes in each cell-type signature defined by DroNc-seq and cell types defined by sNuc-seq in the hippocampus[3] (left) and scRNA-seq in the visual cortex[17] (right). (**d**) Subsets of mouse GABAergic neurons. tSNE embedding of 816 DroNc-seq nuclei profiles from the GABAergic neurons cluster (clusters 10 and 11 in **a**; inset, blue), color coded by subcluster membership. (**e**,**f**) Congruence of GABAergic neurons subclusters defined here (from **d**) with subsets defined from nuclei profiles in the mouse hippocampus[3] (**e**) and single-cell profiles in the mouse visual cortex[17] (**f**). Dot plot shows the proportion of cells in each cluster defined by the other two data sets that were classified to each DroNc-seq cluster using a multiclass random forest classifier (as in ref. 11, Online Methods) trained on the DroNc-seq subclusters.

*mt-Nd4*) were consistent with known distinct enrichment in these compartments (**Supplementary Table 2**). In both methods, over 84% of reads align to the genome (in a representative example), but in cells, 75.2% of these genomic reads map to exons and 9.1% map to introns, whereas in nuclei, 46.2% of genomic reads map to exons and 41.8% to introns (**Supplementary Fig. 2e**), thus reflecting the enrichment of nascent transcripts in the nucleus[3,13–16]. To allow comparison with previous studies, we used only exonic reads subsequently, although intronic reads can be leveraged in future[13].

Clustering of 13,313 nuclei profiled from frozen adult mouse hippocampus ($n$ = 4 mice) and prefrontal cortex (PFC, $n$ = 4) (sequenced at low depth of >10,000 reads and >200 genes detected per nucleus), with an average of 1,810 genes in neurons and 1,077 in non-neuronal cells (Online Methods), revealed groups of nuclei corresponding to known cell types (for example, GABAergic neurons) and to anatomically distinct brain regions or subregions (for example, CA1 and CA3 within the hippocampus; **Fig. 1a**, **Supplementary Figs. 3** and **4** and **Supplementary Table 3**). Each had a distinct expression signature (**Fig. 1b** and **Supplementary Table 4**) and was supported by nuclei from all mice (**Supplementary Fig. 5a**). GABAergic neurons of the same class but from different brain regions (and different samples) grouped together, as did non-neuronal cells (**Fig. 1b**, **Supplementary Figs. 3e** and **5**). Among non-neuronal cells, different glial cell types, including astrocytes, microglia, oligodendrocytes, and oligodendrocyte precursor cells, readily partitioned into separate clusters (**Fig. 1a**) despite their relatively low RNA levels and correspondingly lower numbers of detected genes (**Supplementary Fig. 5c**,**d**). Finally, despite the lower number of genes detected per nucleus in this setting, the cell types and their signatures from DroNc-seq were comparable to those obtained previously with sNuc-seq of mouse hippocampus[3] and scRNA-seq of the visual cortex[17] (**Fig. 1c** and Online Methods).

We also captured finer distinctions between closely related cells, congruent with results of earlier lower-throughput studies. For example, we distinguished eight subsets of GABAergic neurons (**Fig. 1d** and **Supplementary Fig. 6a**,**b**), each expressing a unique combination of canonical marker genes and signatures (**Supplementary Fig. 6c**,**d** and **Supplementary Table 5**). To determine the congruence between cell subtypes obtained from DroNc-seq and those in previous data sets, we trained a multiclass

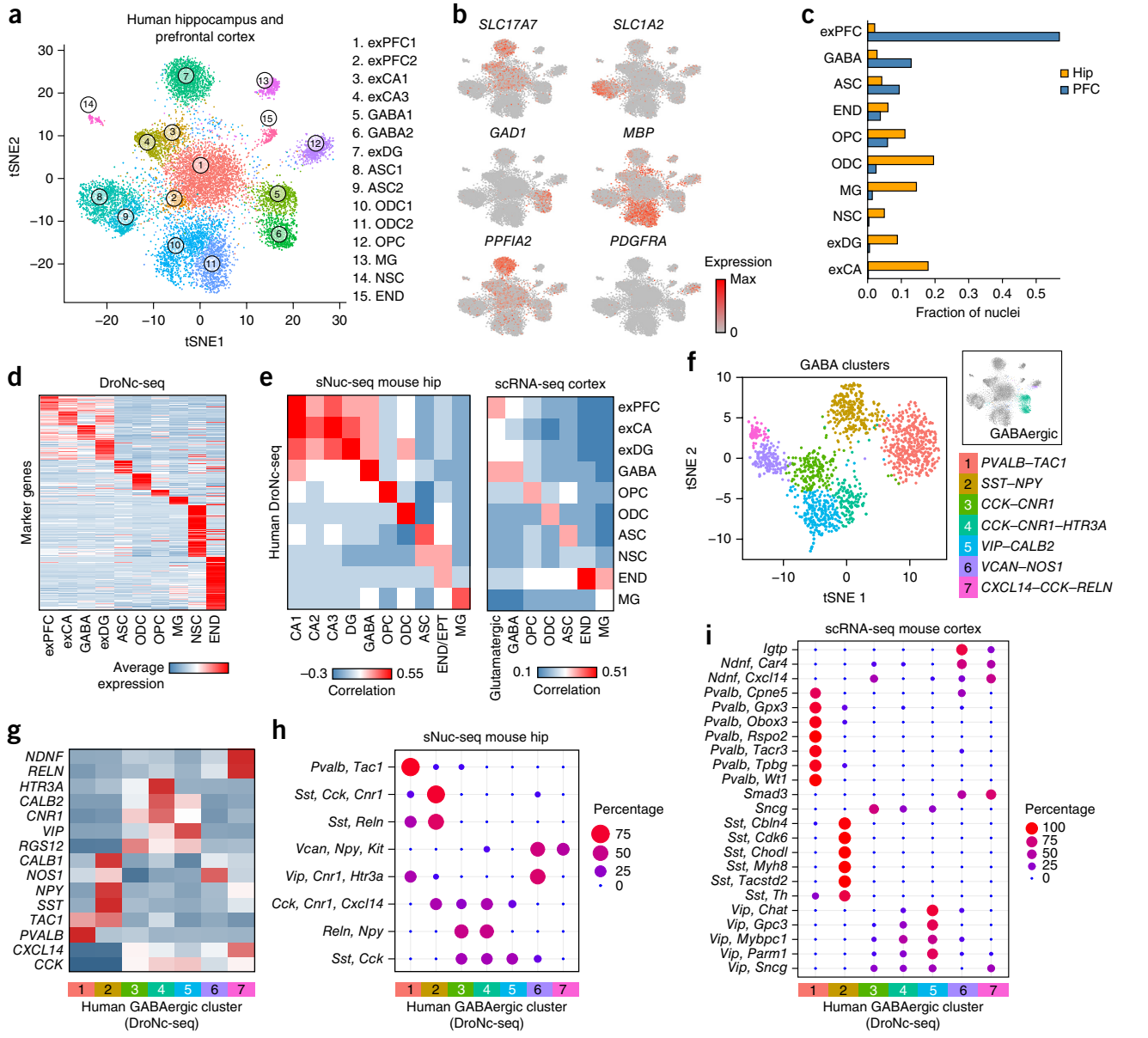


**Figure 2** | DroNc-seq distinguishes cell types and signatures in adult post-mortem human brain tissue. (**a**) Cell-type clusters. tSNE embedding of 14,963 DroNc-seq nuclei profiles (each with >10,000 reads and >200 genes) from adult frozen human hippocampus (Hip, 4 samples) and prefrontal cortex (PFC, 3 samples) from five donors. Nuclei are color-coded by cluster membership and clusters are labeled *post-hoc* (abbreviations as in **Fig. 1a**). (**b**) Marker genes. Plots are as in **a** but with nuclei colored according to the expression level of known cell-type marker genes. (*SLC17A7*, excitatory neurons; *GAD1*, GABAergic neurons; *PPFIA2*, exDG; *SLC1A2*, ASC; *MBP*, ODC; *PDGFRA*, OPC). (**c**) Fraction of nuclei from each brain region associated with each cell type. Cell types are defined as in **Supplementary Figure 7a** and sorted from left by types enriched in PFC versus Hip. (**d**) Cell-type expression signatures. The average expression of differentially expressed signature genes (Online Methods, rows) in each DroNc-seq human brain cell subset (columns; defined as in **Supplementary Fig. 7a**). (**e**) DroNc-seq cell-type expression signatures in the human brain agree with previous mouse data sets. Pairwise correlations of the average expression (Online Methods) for the genes in each cell-type signature defined by DroNc-seq (rows), cell types defined by sNuc-Seq in the mouse hippocampus[3] (left, columns), and scRNA-seq in the visual cortex[17] (right, columns). (**f**–**i**) GABAergic neurons subclusters. (**f**) tSNE embedding of 1,500 DroNc-seq nuclei profiles from the GABAergic neurons cluster (clusters 5 and 6 in **Fig. 2a**; inset), color coded by subcluster membership. (**g**) Average expression of canonical GABAergic marker genes (rows) in each of the nuclei subclusters (columns) defined in **f**. (**h**,**i**) Mapping of human GABAergic neurons subcluster defined here (columns, from **f**) to subsets defined from nuclei profiles in the mouse hippocampus[3] (**h**) and single-cell profiles in the mouse visual cortex[17] (**i**; rows). Dot plot shows the proportion of cells in each cluster defined by the other two data sets that were classified to each DroNc-seq cluster (as in **Fig. 1e**,**f**).

random forest classifier[11] on the DroNc-seq GABAergic subclusters and used it to map GABAergic neuronal cells[17] or nuclei[3] from other data sets (**Fig. 1e**,**f** and Online Methods). Despite the different brain regions and experimental methods and the lower number of genes detected, the DroNc-seq subclusters mapped nearly one-to-one with subclusters defined by sNuc-seq[3] in hippocampus and matched satisfactorily to sets of fine-resolution subclusters defined by scRNA-seq of the visual cortex[17] (**Fig. 1e**,**f** and **Supplementary Fig. 6**).

To demonstrate the utility of DroNc-seq on archived human tissue, we profiled seven frozen post-mortem samples of human hippocampus and PFC from five adults (40–65 years old), archived for 3.5–5.5 years by the GTEx project[18] (**Supplementary Table 6**). Our analysis of 14,963 low-depth sequenced nuclei (>10,000 reads per nucleus, with an average of 1,238 genes in neurons and 607 in non-neuronal cells; **Fig. 2a–d** and **Supplementary Fig. 7**) revealed distinct clusters corresponding to known cell types (**Fig. 2a**, **Supplementary Fig. 7a**, and **Supplementary Table 7**). Although the human archived samples varied in quality, DroNc-seq yielded high-quality libraries of both neurons and glia cells from each sample (**Supplementary Fig. 7c**,**d**). By analyzing a large number of cells, we were able to recover rare cell types, such as that in cluster 14 (**Fig. 2a**), a cluster of hippocampal cells probably comprised of neural stem cells based on marker gene expression (**Supplementary Fig. 7f**).

The cell-type-specific gene signatures we determined for each human cell-type cluster (**Fig. 2d**, **Supplementary Table 8**) agreed well with previously defined signatures in mouse hippocampus[3] and cortex[17] (**Fig. 2e**) and highlighted specific pathways (**Supplementary Fig. 7e**). Moreover, we captured finer distinctions between closely related cells, including subtypes of CA pyramidal neurons, reflecting anatomical distinctions within the hippocampus (**Supplementary Fig. 8**), subtypes of glutamatergic neurons in the PFC expressing unique cortical layer marker genes, such as *RORB* (layer 4–5, refs. 4,17) (**Supplementary Fig. 9**, **Supplementary Table 9**), and subtypes of GABAergic neurons (**Fig. 2f** and **Supplementary Fig. 10a–c**), each associated with a distinct combination of canonical markers and signatures (**Fig. 2g**, **Supplementary Fig. 10d**,**e**, and **Supplementary Table 9**), as previously reported[3,4,17,19]. Notably, we found good congruence between our GABAergic subclusters and those previously defined[3,4,17] in mouse and human using a classifier trained on one data set and tested on the other (Online Methods). Human GABAergic subclusters mapped well to previously defined clusters in the mouse hippocampus[3] (sNuc-seq, **Fig. 2h**), mouse visual cortex[17] (scRNA-seq, **Fig. 2i**), and human cortex[4] (sNuc-seq, **Supplementary Fig. 11**), with the same assignment of canonical marker genes to each cluster (for example, *PVALB*, *SST*, and *VIP*; **Supplementary Table 9**) despite the different species, experimental methods, and brain regions used in each study, as well as the lower number of genes detected in DroNc-seq.

DroNc-seq is a massively parallel sNuc-seq method that is robust, cost effective, and easy to use. Profiling of mouse and human frozen archived brain tissues successfully identified cell types and subtypes, rare cells, expression signatures, and activated pathways. Classifications and signatures derived from DroNc-seq profiles were congruent with those from prior studies in human and mouse (despite the lower number of detected genes per nucleus) but were derived with considerably improved throughput and cost. Moreover, DroNc-seq readily identified rare cell types without the need for enrichment. Nuclei grouped primarily by cell type and not by individual, indicating that cell-type signatures are largely consistent across individuals. Future studies with larger numbers of individuals should assess interindividual variations, which may increase with aging and pathological conditions[20]. DroNc-seq opens the way to systematic single-nucleus analysis of complex tissues that are inherently challenging to dissociate or already archived, thereby helping create vital atlases of human tissues and clinical samples.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

### ACKNOWLEDGMENTS

We thank R. Macare, A. Rotem, C. Muus, and E. Drokhlyansky for helpful discussions, T. Habib for babysitting, T. Tickle and A. Bankapur for technical support, and L. Gaffney and A. Hupalowska for help with graphics. Work was supported by the Klarman Cell Observatory, National Institute of Mental Health (NIMH) grant U01MH105960, National Cancer Institute (NCI) grant 1R33CA202820-1 and NIAID grant U24AI118672-01 (to A.R.), and Koch Institute Support (core) grant P30-CA14051 from the NCI. Microfluidic devices were fabricated at the Center for Nanoscale Systems, Harvard University, supported by National Science Foundation award no. 1541959. N.H. is supported by HHMI through the HHWF, A.R. is supported by HHMI, and F.Z. is supported by the New York Stem Cell Foundation. F.Z. is supported by NIMH (5DP1-MH100706 and 1R01-MH110049), NSF, HHMI, and the New York Stem Cell, Simons, Paul G. Allen Family, and Vallee Foundations, and by J. and P. Poitras, R. Metcalfe, and D. Cheng. D.A.W. thanks NSF DMR-1420570, NSF DMR-1310266, and NIH P01HL120839 grants for their support. GTEx is supported by the NIH Common Fund (Contract HHSN268201000029C to K.A.).

### AUTHOR CONTRIBUTIONS

N.H., I.A.D., A.B., O.R., F.Z., and A.R. conceived the study. A.R. and N.H. devised analyses. N.H., K.S., M.H., and F.A. analyzed the data. A.B. designed and fabricated the microfluidics device. D.A.W. devised the microfluidics design. A.B., I.A.D., N.H., and T.B. designed and conducted the experiments. S.R.C. provided mouse brain tissue. E.G. and K.A. provided human brain tissue. N.H., I.A.D., A.B., and A.R. wrote the paper with input from all of the authors.

### COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.



1. Wagner, A., Regev, A. & Yosef, N. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
2. Tanay, A. & Regev, A. *Nature* **541**, 331–338 (2017).
3. Habib, N. *et al. Science* **353**, 925–928 (2016).
4. Lake, B.B. *et al. Science* **352**, 1586–1590 (2016).
5. Lacar, B. *et al. Nat. Commun.* **7**, 11022 (2016).
6. Grindberg, R.V. *et al. Proc. Natl. Acad. Sci. USA* **110**, 19802–19807 (2013).
7. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
8. Dixit, A. *et al. Cell* **167**, 1853–1866 (2016).
9. Adamson, B. *et al. Cell* **167**, 1867–1882 (2016).
10. Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015).
11. Shekhar, K. *et al. Cell* **166**, 1308–1323 (2016).
12. Ziegenhain, C. *et al. Cell* **65**, 631–643 (2017).
13. Rabani, M. *et al. Nat. Biotechnol.* **29**, 436–442 (2011).
14. Rabani, M. *et al. Cell* **159**, 1698–1710 (2014).
15. Schwanhäusser, B. *et al. Nature* **473**, 337–342 (2011).
16. Cheadle, C. *et al. BMC Genomics* **6**, 75 (2005).
17. Tasic, B. *et al. Nat. Neurosci.* **19**, 335–346 (2016).
18. GTEx Consortium. *Science* **348**, 648–660 (2015).
19. Zeisel, A. *et al. Science* **347**, 1138–1142 (2015).
20. Tirosh, I. *et al. Science* **352**, 189–196 (2016).

## ONLINE METHODS

See Protocol Exchange[21] and **Supplementary Protocol** for a step-by-step protocol for DroNc-seq.

**Microfluidic device design.** Microfluidic devices were designed using AutoCAD (AutoDESK, USA), tested using COMSOL Multiphysics as well as empirically, and fabricated using soft lithographic techniques[22] (**Supplementary Data 1**). The devices were tested on a Drop-seq setup, using bare beads (Tosoh, Japan, Cat # HW-65s) in Drop-Seq Lysis Buffer (DLB[7]; 10 ml stock consists of 4 ml of nuclease-free $H_2O$, 3 ml 20% Ficoll PM-400 (Sigma, Cat # F5415-50ML), 100 µl 20% Sarkosyl (Teknova, Cat # S3377), 400 µl 0.5 M EDTA (Life Technologies), 2 ml 1M Tris pH 7.5 (Sigma), and 500 µl 1M DTT (Teknova, Cat # D9750), where the DTT is added fresh) and 1× PBS, to optimize flow and bead occupancy parameters in drops. Droplet generation was assessed under a microscope in real time using a fast camera (Photron, Model # SA5) and later by sampling the emulsion using a disposable hemocytometer (Life Technologies, Cat # 22-600-100) to check droplet integrity, size, and bead occupancy. The device design is provided in **Supplementary Data 1** and **Supplementary Figure 1b**. The unit in the CAD provided is 1 unit = 1 µm; channel depth on device is 75 µm.

**Cell culture.** 3T3 and HEK293 cells were prepared as described[7]. TF1 cells were cultured according to ATCC's instructions. For DroNc-seq, cells were washed once with PBS, scraped with 2 ml nuclease- and protease-free Nuclei EZ lysis buffer (Sigma, Cat # EZ PREP NUC-101) and processed as tissues, described below.

**Dissection of mouse hippocampus and prefrontal cortex (PFC).** Microdissections of mouse hippocampus and PFC were performed using a stainless steel coronal adult mouse brain matrice and sterile biopsy tissue punch (Braintree Scientific). Dissected subregions were flash frozen on dry ice and stored at −80 °C until processed for nuclei isolation. To validate DroNc-seq for fixed tissue (**Supplementary Fig. 1f**), subregions were placed in ice-cold RNA*later* (ThermoFisher Scientific, Cat # AM7020) and stored at 4 °C overnight, after which RNA*later* was removed and samples were stored at −80 °C until processing.

**Human hippocampus and PFC samples.** Human hippocampus and PFC samples were obtained from the Genotype-Tissue Expression (GTEx) project. Samples were originally collected from recently deceased, non-diseased donors[18,23]. For this study, we selected samples of frozen hippocampus and PFC from five male donors, aged 40–65 (including three samples of PFC and four samples of hippocampus). We used RNA quality from tissues as a proxy for tissue quality and selected tissues with RNA Integrity Number (RIN) values of 6.9 or higher (average RIN was 7.3). The average post-mortem ischemic interval for tissues was 12.4 h (**Supplementary Table 6**).

**Nuclei isolation.** Nuclei were isolated with EZ PREP buffer (Sigma, Cat #NUC-101). Tissue samples were cut into pieces <0.5 cm or cell pellets were homogenized using a glass dounce tissue grinder (Sigma, Cat #D8938) (25 times with pastel A and 25 times with pastel B) in 2 ml of ice-cold EZ PREP and incubated on ice for 5 min, with an additional 2 ml of ice-cold EZ PREP. Nuclei were centrifuged at 500 × *g* for 5 min at 4 °C, washed with 4 ml ice-cold EZ PREP and incubated on ice for 5 min. After centrifugation, the nuclei were washed in 4 ml Nuclei Suspension Buffer (NSB; consisting of 1× PBS, 0.01% BSA and 0.1% RNase inhibitor (Clontech, Cat #2313A)). Isolated nuclei were resuspended in 2 ml NSB, filtered through a 35-µm cell strainer (Corning, Cat # 352235) and counted. A final concentration of 300,000 nuclei per ml was used for DroNc-seq experiments.

For comparison experiments of nuclei isolation protocols (**Supplementary Fig. 1d**,**e**), nuclei were also isolated using the sucrose gradient centrifugation method described for sNuc-Seq[3]. The nuclei isolation protocol used here is more efficient than the gradient-centrifugation-based method and does not require ultracentrifugation. This reduced processing time and minimized RNA degradation, facilitating processing of multiple samples.

**Coencapsulation of nuclei and barcode beads.** 10 µl of the single nuclei suspension in NSB (described above) was stained with DAPI (Fisher, Cat # D1306), loaded on a hemocytometer, and checked under a microscope to ensure that nuclei were adequately isolated into singletons. The nuclei were suspended in NSB at ~300,000 nuclei per ml. Using ~75-µm droplets, a loading concentration of 300,000 nuclei per ml and ~4.5 million drops per ml amounts to a Poisson loading parameter, $\lambda$ ~300,000/4,500,000 = 0.07.

Barcoded beads (Chemgenes, Cat # Macosko-2011-10) were prepared as in ref. 7. Because the channels of the DroNc-seq microfluidic device are narrow (~70 µm), they are more likely to clog from large beads compared to Drop-seq. We therefore size selected beads <40 µm diameter using a strainer (PluriSelect, Cat # 43-50040-03); in our experience, these smaller beads comprise roughly 55% of the purchased bead pool. The barcoded beads were suspended in DLB (described above) and counted at 1:1 dilution in 20% PEG solution using a hemocytometer (VWR, Cat # 22-600-102)[7], at concentrations between 325,000 and 350,000 beads per ml.

The nuclei and barcoded bead suspension were loaded[7] and flown at 1.5 ml/h each, along with carrier oil (BioRad Sciences, Cat # 186-4006) at 16 ml/h, to coencapsulate single nuclei and beads in ~75-µm drops (vol. ~200 pl) at 4,500 drops/s and double Poisson loading concentrations. The smaller droplet volume in DroNc-seq results in higher mRNA concentration in drops (>5×) compared to 125-µm drops in Drop-seq.

The theoretical Poisson loading concentration at 1/10 bead and nuclei occupancy for devices with channels 70 µm wide and 75 µm deep is ~520,000/ml, and 100 µm depth (also tested) is 340,000/ml. We tested bead and cell loading at this and other concentrations using species-mixing experiments[7] (for example, **Supplementary Fig. 1g** and **Supplementary Table 1**) and ease of bead flow as metrics and found that beads at 350,000/ml and nuclei at 300,000/ml concentrations performed best, in terms of low human–mouse doublet rate and fewer clogging events during droplet generation. At the nuclei loading concentrations used, the occurrence of one or more nuclei in a drop follows a Poisson distribution, $P(x) = \lambda^x e^{-\lambda}/x!$, where $\lambda$ = Poisson parameter and $x = 2$ for doublet estimation. As a theoretical lower bound, increasing nuclei concentration will increase doublet rate as $\lambda^2 e^{-\lambda}/2$; for example, if nuclei loading is increased by 10%, the probability of getting two nuclei in a drop will increase from 0.21% to 0.25%.

However, the probability of getting two or more nuclei in a drop, i.e., doublets, triplets, etc., all of which would be indistinguishable in species-mixing experiments, is $P(x \geq 2; \lambda = 0.07) = 0.5\%$. In practice, nuclei that stick together or cellular debris could also contribute to doublets or doublet-like phenomena. Empirical doublet rates in experiments ranged from ~1% (mouse tissue; clustering analysis) to ~5% (species mixing).

For nuclei experiments on human and mouse tissue, 75-µm DroNc-seq devices were used, except for when a 125-µm Drop-seq device was used for comparison (**Supplementary Fig. 1c**). Note that for 3T3 nuclei, both 125-µm Drop-seq and 75-µm DroNc-seq devices yielded similar results, whereas profiling 3T3 cells by Drop-seq had better efficiency and complexity.

**Droplet breaking, washes, and reverse transcription (RT).** Microfluidic emulsion was collected into 50-ml Falcon tubes for ~22 min each and left at room temperature for up to 45 min before breaking drops[7] and performing RT[7].

**Post-RT wash, exonuclease I treatment, PCR, and library preparation.** Post RT, each barcoded bead had cDNA barcoded with the bead's unique barcode bound onto it, also referred to as a STAMP[7]. STAMPs from multiple collections of a given sample were pooled at this point, resuspended in 1 mL $H_2O$, and a 10-µl aliquot of the suspension was mixed with 10 µl of 20% PEG solution and counted. Aliquots of 5,000 beads were amplified[7] using the following PCR steps: 95 °C for 3 min, then four cycles of: 98 °C for 20 s, 65 °C for 45 s, 72 °C for 3 min, then *X* cycles of: 98 °C for 20 s, 67 °C for 20 s, 72 °C for 3 min, and finally, 72 °C for 5 min, in which *X* was adjusted according to sample quality. STAMPs from mouse tissue were amplified for $X = 10$ cycles, and PCR products were pooled in batches of four wells or 16 wells. STAMPS from human tissue were amplified for $X = 10$ or 12 cycles. Human PCR products were pooled in batches of four wells ($X = 12$) or 16 wells ($X = 10$). Supernatants from each well were combined in a 1.5-ml Eppendorf tube and cleaned with 0.6× SPRI beads (Ampure XP, Beckman Coulter, Cat # A63881).

Notably, the number of PCR wells from a DroNc-seq run depends on the number of STAMPs obtained. A user may access the STAMPs in different ways, depending on the number of nuclei they wish to sequence. One would either access the pool one time or more, each time taking only a portion of the STAMPs to generate a library, and repeat the process if more is desired. For mouse and human brain, it was optimal to use 5,000 STAMPs in each PCR reaction and then pool four PCR wells together for library preparation, which is expected to yield 1,400 nuclei profiles based on our loading and flow parameters. Depending on the desired number of reads per nucleus and sequencing yield, one can pool higher numbers of PCR wells in a single Illumina Nextera™ library, as demonstrated here using 16–32 wells for libraries used in the clustering analysis of mouse and human brain tissue.

Purified cDNA was quantified[7] and 550 pg of each sample was fragmented, tagged, and amplified in each Nextera reaction[7].

**Sequencing.** The libraries were sequenced at 2.2 pM (mouse, 16-well pool), 2.7 pM (mouse, 4-well pool), and 2.3 pM (human) on an Illumina NextSeq 500. We used NextSeq 75 cycle v3 kits to sequence 20-bp and 64-bp paired-end reads, with Custom Read1 primer[7]. The sequencing cluster density and percent passing filter number from different experiments varied according to the quality of nuclei samples used but were optimized around cluster density of 220 and 90% passing filter.

**Preprocessing of DroNc-seq data.** *Read filtering and alignment.* Paired-end sequence reads were processed mostly as previously described[7,11]. Briefly, the left read was used to infer both the cell of origin, based on the first 12 bases (the Nucleus Barcode or NB), and the molecule of origin, based on the next eight bases (Unique Molecular Index or UMI). Reads were first filtered by quality score, and the right mate of each read pair was trimmed and aligned to the genome (mouse mm10 UCSC, human hg19 UCSC) using STAR v2.4.0a, ref. 24. Reads mapping to exonic regions of genes as per the mouse UCSC genome (version mm10) or the human UCSC genome (version hg19) were recorded.

*Digital gene expression.* Nucleus (cell) barcodes that represent genuine nuclei RNA libraries rather than technical and sequencing errors were distinguished as previously described[7,11] as true or 'core' nucleus barcodes. Briefly, barcodes were first filtered on the basis of a minimum number of transcripts associated with them and then barcodes were checked for synthesis errors and collapsed to core barcodes if they were within an edit distance of 1. To account for amplification bias, gene counts were collapsed within each sample, using UMI sequences (within an edit distance of 1, substitutions only), as previously described[7,11]. The expression count (or number of transcripts) for a given gene in a given nucleus was determined by counting unique UMIs and compiled into a digital gene expression (DGE) matrix. The DGE matrix was scaled by total UMI counts, multiplied by the mean number of transcripts (calculated for each data set separately), and the values were log transformed. To reduce the effects of library quality and complexity on cluster identity, a linear model was used to regress out effects of the number of transcripts and genes detected per nucleus (using the 'RegressOut' function in the Seurat software package).

**Gene detection and quality controls.** *Additional filtering of the expression matrix.* Nuclei with less than 200 detected genes and less than 10,000 usable reads were filtered out. We note that, as for scRNA-seq, depending on the cell type in question, the cutoff may need to be set on a case-by-case basis, on the basis of the characteristic RNA content of the cell type. A gene is considered detected in a cell if it has at least two unique UMIs (transcripts) associated with it. For each analysis, genes were removed that were detected in less than 10 nuclei. After filtering, the number of cells and nuclei were as follows: (1) 1,710 cells from the 3T3 single cell libraries (collected by Drop-seq) across two replicates, (2) 5,636 3T3 nuclei across six replicates, (3) 19,561 nuclei from the mouse brain (four PFC samples and four hippocampus samples from four mice used for cell-type analysis and an additional eight cortical samples from four mice used for quality-control experiments), and (4) 19,550 nuclei from the human brain (three PFC samples and four hippocampus samples from five donors). Clusters and cell-type classification were robust for different gene-detection thresholds. The above threshold was used in all of the clustering analyses. For the quality-control experiments (specifically, testing the performance with RNA*Later*, different nuclei isolation protocols, and different microfluidic devices; **Supplementary Fig. 1**), at least 20,000 usable reads per nucleus




doi:10.1038/nmeth.4407

were required (the number of reads at which we estimated sample saturation; **Supplementary Fig. 2f**,**g**). For the assessment of the complexity and sensitivity of DroNc-seq, at least 80,000 usable reads per nucleus were required; this analysis was performed with only the samples sequenced deeply to an average of 160,000 reads per nucleus, as required for saturation analysis.

*QC metrics.* A list of quality metrics was obtained for all DroNc-seq data sets using Samtools (http://samtools.sourceforge.net/), Picard Tools (http://broadinstitute.github.io/picard/), and in-house scripts. For each single-nucleus profile, we calculated the total number of reads mapped to coding regions and UTRs, number of genes detected per nucleus, and the percentage of the total number of reads assigned to nucleus barcode that were from: (1) coding regions, (2) UTRs, (3) intronic regions, (4) intergenic regions, (5) ribosomal RNA (rRNA), and (6) transcripts derived from the mitochondrial genome.

**Comparison of Drop-seq (cells) and DroNc-seq (nuclei).** We compared DroNc-seq (nuclei) and Drop-seq (cells) using several measures. (1) We compared the capture-rate efficiency of DroNc-seq and Drop-seq in libraries derived from pooling four PCR wells, followed by sequencing to an average depth of 160,000 usable reads per nucleus or cell. The efficiency is defined as the percent of nuclei actually observed out of the proportion expected per library, given the Poisson loading of 0.07 for DroNc-seq and 0.1 for Drop-seq. For example, at 100% efficiency, a DroNc-seq pool of 20,000 beads is expected to contain 1,400 nuclei (2,000 cells in Drop-seq). On average, we observed 87% efficiency for DroNc-seq (78%, 89%, and 95% efficiency for cell lines, mouse brain, and human brain tissue, respectively) and 72% for Drop-seq on cell lines. (2) We compared the means and the distributions of the number of genes and transcripts detected for all cells and nuclei that pass our quality filter (**Supplementary Fig. 2b**,**c**). (3) We compared the expression profiles of nuclei and cells (3T3 cell line) by computing the average expression for each gene (average log transformed UMI counts) in each replicate and then the Pearson correlation coefficients between technical replicates of cells or nuclei (all have $r = 0.99 \pm$ s.d. $= 0.0023$), then between nuclei and cells ($r = 0.81 \pm$ s.d. $= 0.0024$) (**Supplementary Fig. 2d**). (4) We tested for genes differentially expressed between cells and nuclei (3T3 cell lines) after pooling technical replicates. We defined differentially expressed genes using Student's *t* test, requiring FDR < 0.001, log ratio > 1, and an average expression across all nuclei or cell samples log(UMI count) > 3. We found only two genes upregulated in the nuclei (encoding lncRNAs Malat1 and Meg3) and 57 genes up regulated in cells, including those encoding many mitochondrial RNAs and ribosomal protein RNAs (known to be stable and thus enriched in cells compared to nuclei[13,14]) (**Supplementary Table 2**). (5) We compared the fraction of the total number of reads that were mapped to (i) coding regions, (ii) UTRs, (iii) intronic regions, (iv) intergenic regions, and (v) ribosomal RNA (as described above) (**Supplementary Fig. 2e**).

**Principal components analysis (PCA), clustering, and tSNE visualization.** *Finding variable genes.* To select highly variable genes, we fit a relationship between mean counts and coefficient of variation using a gamma distribution on the data from all of the genes[19,25] and ranked genes by the extent of excess variation as a function of their mean expression (using a threshold of at least 0.2 difference in the coefficient of variation between the empirical and the expected and a minimal mean transcript count of 0.005).

*Dimensionality reduction using PCA.* We used a DGE matrix consisting only of variable genes as defined above, scaled and log transformed, and then reduced its dimensions with PCA. We used the fast 'rpca' function in R (package 'rsvd') and chose the most significant principal components (or PCs) based on the largest eigen value gap[3] (separately for each data set) to use as input in downstream analysis.

*Graph clustering.* We partitioned the profiles into clusters of transcriptionally similar nuclei using the top significant PCs as an input to a graph-based clustering algorithm, as previously described[11]. Briefly, in the first step, we computed a *k*-nearest neighbor (*k*-NN) graph and connected each nucleus to its *k*-nearest neighbors (based on Euclidean distance, using the 'nng' function of the 'igraph' package in R). We next used the *k*-NN graph as an input to the Infomap algorithm[26], which decomposes an input graph into modules using the 'cluster_infomap' function in R). The clustering results were visualized by coloring a tSNE[27] 2D map *post hoc* (described below). We used $k = 100$ for clustering of each full data set and $k = 80$ for the human brain subset clustering (**Fig. 2f**, **Supplementary Figs. 8**,**9**).

*Subclustering.* To identify subtypes of cells, the same analyses were performed as described above but on a specific subset of nuclei (one or few of the major clusters; as described in the main text) to partition it to subclusters.

*tSNE visualization.* We generated a 2D nonlinear embedding of the nuclei profiles using tSNE. The scores along the top significant PCs estimated above were used as input to the algorithm (using the 'Rtsne' package, with a maximum of 2,000 iterations, disabling the initial PCA step and setting the perplexity parameter to 100 for detection of the major clusters and 60 for subclusters). Because tSNE can produce different visualizations in different runs, we used these coordinates only for visualization and not to identify cell clusters. Interestingly, we can associate nuclei with a distinct known cell type, even for those nuclei with as few as 100 genes detected, suggesting that the cell-type identity in the brain can be encoded by a small set of genes, easily detected with shallow sequencing, as previously observed in other systems[11].

To visualize the expression of known marker genes (for example, subtypes of GABAergic neurons in the hippocampus and cortex[3,19]) or genes found to be upregulated, we visualized the average expression of the markers across each cluster or cell type as violin plots and visualized the distribution of the expression across cells in the tSNE space by color coding the dots based on expression levels.

*Testing for batch and technical effects.* To rule out the possibility that the resulting clusters are driven by batch or other technical effects, we examined the distribution of samples within each cluster and the distribution of the number of genes detected across clusters (as a measure of nuclei quality). Overall, the nuclei separated into distinct point clouds in tSNE space that were not driven by batch; each cluster or cloud was an admixture of cells from all technical and biological replicates, with variable numbers of genes. Related to the number of genes, we note that there is a distinct biological difference in cell size (and expected RNA content) between neuronal and glial cells in the brain.

**Transcript and gene saturation analysis.** To assess the extent of saturation and required read depth of the DroNc-seq libraries, we used nuclei libraries from a mouse cell line (3T3), mouse brain tissue, and human brain tissue (cortex), each sequenced to an average read depth of 160,000 reads per nucleus. We removed nuclei with less than either 200 genes detected or 10,000 reads. We performed saturation analyses for transcripts (UMI) and genes for each nucleus separately by subsampling reads with replacement across the range of reads for that nucleus (from 0.02 to 0.98 of the total read counts within a given nucleus or cell, in 0.02 increments). For each subsampling, we calculated the number of reads and transcripts detected. This sampling procedure was repeated ten times, and the mean values were reported. Saturation limits for UMI and genes were estimated by nonlinear fitting of the following saturation function to all points generated by the sampling procedure:

$$y = \frac{ax}{(b + x)} + c$$

**Cluster annotation, filtering, differential expression, and pathway analysis.** Major cell-type clusters were identified by using a set of known cell-type marker genes from the literature, as previously described[3,19]. In addition, we identified signatures of upregulated genes for each cluster (**Supplementary Tables 4**, **5**, **8** and **9**), which we used to further validate the identity of the cluster by matching these signatures with canonical cell-type marker genes and by testing for enriched pathways. Differentially expressed signatures were calculated using a binomial likelihood ratio test[28] to find genes that are upregulated within each cluster compared to the rest of the nuclei in the data set, with a FDR of 0.01 and requiring genes to be expressed in at least 20% of nuclei in the given cluster and have a minimum difference of 20% in the fraction of nuclei in which they are detected. The differential expression signatures were tested for enriched pathways and gene sets using a hypergeometric test (FDR < 0.01). Pathways were taken from the MSigDB/GSEA resource (combining data from Hallmark pathways, REACTOME, KEGG, GO and BIOCARTA)[29].

We flagged problematic clusters to be disregarded in downstream analysis by any one of three criteria: (1) clusters with dubious quality of nuclei, in which the nuclei associated mainly with one sample did not associate with specific cell-type markers, (2) clusters with nuclei expressing both overlapping markers of two different cell types and having a relatively higher number of transcripts, indicating they might be nuclei doublets, or (3) clusters expressing markers of neighboring brain regions that might be a result of nonspecific tissue dissection (such as genes enriched in the choroid plexus, **Supplementary Fig. 3b**). Several small clusters in the human and mouse brain were discarded from downstream analysis (as annotated in **Supplementary Tables 3** and **7** and in **Supplementary Fig. 3b**).

Cell types were defined by combining clusters of all subtypes (for example, the GABAergic subclusters were combined into one group of GABAergic neurons), which were used in the downstream analysis for testing the number of genes and transcripts in each cell type, defining cell-type-specific expression signatures, subclustering, and comparing cell-type signatures to previous data sets.

**Comparison of DroNc-seq data to previous data sets.** *Comparison of cell-type signatures.* Cell-type-specific expression patterns were compared to signatures previously defined in several relevant data sets by calculating the pairwise Pearson correlations coefficients between each pair of cell types in the other data set and DroNc-seq data sets for the same set of genes. First, we compared to average cell-type-specific signatures from sNuc-Seq analysis in the mouse hippocampus[3] (Supplementary Tables in ref. 3). Second, we compared to the single-cell RNA-seq data set of the mouse visual cortex (Tasic *et al.*[17]), using the previously defined cell-type annotations and expression values per cell (from GEO data set GSE71585 and ref. 17.). Average log transformed TPM counts, FPKM counts, or scaled UMI counts were used to generate the mouse hippocampus[3], mouse visual cortex[17], and DroNc-seq signatures, respectively.

*Comparison of mouse and human GABAergic subclusters to previously defined subclusters in mouse brain.* To determine the congruence of cell subtypes between the DroNc-seq analyses to other neural data sets, we adopted an approach that we previously described in an analysis of retinal neurons[11]. Briefly, we trained a multiclass random forest classifier[30] on the clusters defined on the DroNc-seq data separately for human and mouse GABAergic neurons. In each case, we used the most variable genes (approximately 700–2,000 genes across data sets, as described above) to build a classifier on 60% of the data (training set). For each data set, the classifier was tested on the remaining 40% of the data that was not used for training (test set) to obtain an estimate of the classification accuracy. Nuclei in the test set mapped to their correct classes at a rate of 93% for the human GABAergic neurons and 91% for the mouse GABAergic neurons (expected accuracy based on random assignment was 12.5%). These classifiers were then used to map cells or nuclei in other data sets, including single-nucleus RNA-seq in the mouse hippocampus brain region[3] and single-cell RNA-seq in the mouse visual cortex[17].

*Comparison of human GABAergic subclusters to previously defined subclusters in human brain.* To determine the congruence of neuron subtypes between DroNc-seq analysis of hippocampus and PFC and previous analyses of human visual cortex (Lake *et al.*[4]), we used the classifier previously defined in Lake *et al.*[4] that includes a set of signature genes at each point along a decision tree leading to the classification of eight GABAergic subtypes. To classify the DroNc-seq nuclei profiles, at each branch point in the tree, we scored each nucleus profile using the left and right gene signatures, by the average expression level of all signature genes per nucleus (log transformed UMI counts centered around the mean value), and assigned the tested nucleus by the higher score.

*RNA in situ hybridization data.* RNA *in situ* hybridization images for marker genes was taken from the Allen Institute Brain Atlas[31].

**Data availability.** Raw human sequencing data is available at dbGaP under accession code phs000424.v8.p1, and expression tables are available at http://www.gtexportal.org/home/datasets. Raw and processed mouse sequencing data is available at https://portals.broadinstitute.org/single_cell and at the Gene Expression Omnibus (GEO) database.





doi:10.1038/nmeth.4407

A **Life Sciences Reporting Summary** for this publication is available.


21. Anindita, B. *et al. Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.094 (2017).
22. McDonald, J.C. *et al. Electrophoresis* **21**, 27–40 (2000).
23. Carithers, L.J. *et al. Biopreserv. Biobank.* **13**, 311–319 (2015).
24. Dobin, A. *et al. Bioinformatics* **29**, 15–21 (2013).
25. Brennecke, P. *et al. Nat. Methods* **10**, 1093–1095 (2013).
26. Rosvall, M. & Bergstrom, C.T. *Proc. Natl. Acad. Sci. USA* **105**, 1118–1123 (2008).
27. van der Maaten, L. & Hinton, G. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
28. McDavid, A. *et al. Bioinformatics* **29**, 461–467 (2013).
29. Subramanian, A. *et al. Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
30. Breiman, L. *Mach. Learn.* **45**, 5–32 (2001).
31. Lein, E.S. *et al. Nature* **445**, 168–176 (2007).

nature research

Corresponding author(s): Feng Zhang and Aviv Regev

☐ Initial submission ☒ Revised version ☐ Final submission

# Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

## Experimental design

1. Sample size

Describe how sample size was determined.

We used at least 3 biological replicates for each experiment.
we made sure that our clusters are supported by all biological replicates.

2. Data exclusions

Describe any data exclusions.

No animals/human samples were excluded. Low quality nuclei/ cells were filtered out (Methods, p.11). "Problematic" clusters were remove from downstream analysis (Methods p.19-20)

3. Replication

Describe whether the experimental findings were reliably reproduced.

The experimental findings were supported by the following replicates:
(1) 1,710 cells from the 3T3 single cell libraries (collected by Drop-seq) across two replicates; (2) 5,636 3T3 nuclei across 6 replicates; (3) 19,561 nuclei from the mouse brain (4 PFC samples and 4 hippocampus samples from 4 mice used for cell type analysis, and an additional 8 cortical samples from 4 mice used for quality control experiments); and (4) 19,550 nuclei from the human brain (3 PFC samples and 4 hippocampus samples from 5 donors).

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Human hippocampus and pre-frontal cortex samples were collected as part of the Genotype-Tissue Expression (GTEx) project according to the described criteria (Methods, p.2-3). Mouse hippocampus and pre-frontal cortex samples were collected from wild-type mice and were allocated to one experimental group.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No groups were allocated during data collection. Cluster analysis was blinded to the origin of the analyzed sample.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.







Nature Methods: doi:10.1038/nmeth.4407

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| ☐ | ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ A statement indicating how many times each experiment was replicated |
| ☐ | ☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| ☐ | ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| ☐ | ☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted |
| ☐ | ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) |
| ☐ | ☒ Clearly defined error bars |

*See the web collection on statistics for biologists for further resources and guidance.*

## Software

Policy information about availability of computer code

7. Software

| | |
|---|---|
| Describe the software used to analyze the data in this study. | R code was written and specific functions and packages are described in Methods p.11-22 |

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

## Materials and reagents

Policy information about availability of materials

8. Materials availability

| | |
|---|---|
| Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company. | No restrictions. |

9. Antibodies

| | |
|---|---|
| Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species). | N/A |

10. Eukaryotic cell lines

| | |
|---|---|
| a. State the source of each eukaryotic cell line used. | ATCC |
| b. Describe the method of cell line authentication used. | Cells were cultured according to the ATCC's instructions. |
| c. Report whether the cell lines were tested for mycoplasma contamination. | Random cell batches are tested for mycoplasma contamination rustically by PCR. |
| d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. | N/A |

## Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

| | |
|---|---|
| Provide details on animals and/or animal-derived materials used in the study. | C57B/6 male mice, 10-14 weeks old. |







Nature Methods: doi:10.1038/nmeth.4407

Policy information about studies involving human research participants

12. Description of human research participants

| | |
|---|---|
| Describe the covariate-relevant population characteristics of the human research participants. | Described in Methods, p.2-3 and supplementary table 5. |







Nature Methods: doi:10.1038/nmeth.4407