

# Equivalent high-resolution identification of neuronal cell types with single-nucleus and single-cell RNA-sequencing

Trygve E. Bakken<sup>1</sup>, Rebecca D. Hodge<sup>1</sup>, Jeremy M. Miller<sup>1</sup>, Zizhen Yao<sup>1</sup>, Thuc N. Nguyen<sup>1</sup>, Brian Aevermann<sup>2</sup>, Eliza Barkan<sup>1</sup>, Darren Bertagnolli<sup>1</sup>, Tamara Casper<sup>1</sup>, Nick Dee<sup>1</sup>, Emma Garren<sup>1</sup>, Jeff Goldy<sup>1</sup>, Lucas T. Gray<sup>1</sup>, Matthew Kroll<sup>1</sup>, Roger S. Lasken<sup>2</sup>, Kanan Lathia<sup>1</sup>, Sheana Parry<sup>1</sup>, Christine Rimorin<sup>1</sup>, Richard H. Scheuermann<sup>2</sup>, Nicholas J. Schork<sup>2</sup>, Soraya I. Shehata<sup>1</sup>, Michael Tieu<sup>1</sup>, Kimberly A. Smith<sup>1</sup>, Hongkui Zeng<sup>1</sup>, Ed S. Lein<sup>1</sup>, and Bosiljka Tasic<sup>1</sup>

<sup>1</sup>Allen Institute for Brain Science, Seattle, WA, USA

<sup>2</sup>J. Craig Venter Institute, La Jolla, CA, USA

December 24, 2017

## 1 Abstract

2 Transcriptional profiling of complex tissues by RNA-sequencing of single nuclei presents some advantages over whole cell  
3 analysis. It enables unbiased cellular coverage, lack of cell isolation-based transcriptional effects, and application to archived  
4 frozen specimens. Using a well-matched pair of single-nucleus RNA-seq (snRNA-seq) and single-cell RNA-seq (scRNA-seq)  
5 SMART-Seq v4 datasets from mouse visual cortex, we demonstrate that similarly high-resolution clustering of closely related  
6 neuronal types can be achieved with both methods if intronic sequences are included in nuclear RNA-seq analysis. More  
7 transcripts are detected in individual whole cells (~11,000 genes) than nuclei (~7,000 genes), but the majority of genes have  
8 similar detection across cells and nuclei. We estimate that the nuclear proportion of total cellular mRNA varies from 20% to  
9 over 50% for large and small pyramidal neurons, respectively. Together, these results illustrate the high information content of  
10 nuclear RNA for characterization of cellular diversity in brain tissues.

## 11 Introduction

12 Understanding neural circuits requires characterization of their cellular components. Cell types in mam-  
13 malian brain have been defined based on shared morphological, electrophysiological and, more recently,  
14 molecular properties (Poulin et al., 2016; Zeng and Sanes, 2017; Bernard et al., 2009). scRNA-seq has  
15 emerged as a high-throughput method for quantification of the majority of transcripts in thousands of cells.  
16 scRNA-seq data have revealed diverse cell types in many mouse brain regions, including neocortex (Tasic  
17 et al., 2016, 2017; Zeisel et al., 2015), hypothalamus (Campbell et al., 2017), and retina (Shekhar et al.,  
18 2016; Macosko et al., 2015).

19 However, scRNA-seq profiling does not provide an unbiased survey of neural cell types. Some cell types are  
20 more vulnerable to the tissue dissociation process and are underrepresented in the final data set. For exam-  
21 ple, in mouse neocortex, fast-spiking parvalbumin-positive interneurons and deep-projecting glutamatergic  
22 neurons in layer 5b are observed in lower proportions than expected and need to be selectively enriched  
23 using Cre-driver lines (Tasic et al., 2017) for sufficient sampling. In adult human neocortex, neurons largely

24 do not survive dissociation thereby causing over-representation of non-neuronal cells in single cell suspen-  
25 sions (Darmanis et al., 2015). In contrast to whole cells, nuclei are more resistant to mechanical assaults  
26 and can be isolated from frozen tissue (Krishnaswami et al., 2016; Lacar et al., 2016). Single nuclei have  
27 been shown to provide sufficient gene expression information to define relatively broad cell classes in adult  
28 human brain (Lake et al., 2016, 2017a) and mouse hippocampus (Habib et al., 2016).

29 Previous studies have not addressed if the nucleus contains sufficient diversity and number of transcripts to  
30 enable discrimination of closely related cell types at a resolution comparable to whole cells. A recent study  
31 compared clustering results for single nuclei and whole cells isolated from mouse somatosensory cortex (Lake  
32 et al., 2017b), but it only showed similar ability to distinguish two very different cell classes: superficial- and  
33 deep-layer excitatory neurons.

34 In this study, we compared 463 matched nuclei and whole cells from layer 5 of mouse primary visual cortex  
35 (VISp) to investigate differences in single nucleus and single cell transcriptomes. We selected this brain  
36 region because it contains a known variety of distinguishable yet highly similar cell types that would reveal  
37 the cell-type detection limit of RNA-seq data obtained from single cells or nuclei (Tasic et al., 2016). We  
38 used the same primary cell source and processed cells and nuclei with the same transcriptomic profiling  
39 method to directly compare the resolution limit of cell type detection from well-matched sets of single cells  
40 and nuclei.

## 41 Results

### 42 RNA-seq profiling of single nuclei and single cells

43 We isolated 487 NeuN-positive single nuclei from layer 5 of mouse VISp using fluorescence activated cell  
44 sorting (FACS). Anti-NeuN staining was performed to enrich for neurons. In parallel, we isolated 12,866  
45 tdT-positive single cells by FACS from all layers of mouse VISp and a variety of Cre-driver lines, as part of  
46 a larger study on cortical cell type diversity (Tasic et al., 2017). For both single nuclei and cells, poly(A)-  
47 transcripts were reverse transcribed and amplified with SMART-Seq v4, cDNA was tagged by Nextera  
48 XT, and resulting libraries were sequenced to an average depth of 2.5 million reads (Figure 1A). RNA-seq  
49 reads were mapped to the mouse genome using the STAR aligner (Dobin et al., 2013). Gene expression was  
50 quantified as the sum of intronic and exonic reads per gene and was normalized as counts per million (CPM)  
51 and log<sub>2</sub>-transformed. For each nucleus and cell, the probabilities of gene detection dropouts were estimated  
52 as a function of average expression level based on empirical noise models (Kharchenko et al., 2014).

53 463 out of 487 single nuclei (95%) passed quality control metrics, and each nucleus was matched to the  
54 most similar nucleus and cell based on the maximum correlated expression of all genes, weighted for gene  
55 dropouts. Nuclei had similarly high pairwise correlations to cells as to other nuclei suggesting that cells and  
56 nuclei were well matched (Figure 1B). As expected, matched cells were derived almost exclusively from layer  
57 5 and adjacent layers 4 and 6 (Figure S1B), and from Cre-driver lines that labeled cells in layer 5 (Figure 1C  
58 and Figure S1A,C). The small minority of matched cells isolated from superficial layers were GABAergic  
59 interneurons that have been detected in many layers (Tasic et al., 2017).

### 60 Comparison of nuclear and whole cell transcriptomes

61 scRNA-seq profiles nuclear and cytoplasmic transcripts, whereas snRNA-seq profiles nuclear transcripts.  
62 Therefore, we expect that RNA-seq reads will differ between nuclei and cells. In nuclei, more than 50%

63 of reads that aligned to the mouse genome did not map to known spliced transcripts but to non-exonic  
64 regions within gene boundaries. They were therefore annotated as intronic reads (Figure 2A). In contrast,  
65 the majority of cells had less than 30% intronic reads with a minority of cells having closer to 50% intronic  
66 reads, similar to nuclei. Median gene detection based on exonic reads was lower for nuclei (~5,000 genes) than  
67 for cells (~9,500). Including both intronic and exonic reads increased gene detection for nuclei (~7,000) and  
68 cells (~11,000), demonstrating that intronic reads provided additional information not captured by exons.  
69 Whole brain control RNA displayed a read mapping distribution similar to cells, which is consistent with  
70 dissociated single cells capturing the majority of transcripts in the whole cell.

71 Transcript dropouts likely result from both technical and biological variability, and both effects are more  
72 pronounced in nuclei than in cells. When transcript dropouts were adjusted based on empirical noise models,  
73 correlations between pairs of nuclei and pairs of cells increased, although cell-cell similarities remained sig-  
74 nificantly higher (Figure 2B). A majority of expressed genes (21,279; 63%) showed similar detection (<10%  
75 difference) in nuclei and cells, whereas 7,217 genes (21%) were detected in at least 25% more cells than  
76 nuclei (Figure 2C and Table S1). 8,614 genes have significantly higher expression in cells than nuclei (>1.5  
77 fold expression; FDR < 0.05) and many are involved in house-keeping functions such as mRNA processing  
78 and translation (Figure 2D). Genetic markers of neuronal activity, such as immediate early genes *Fos*, *Egr1*,  
79 and *Arc* also displayed up to 10-fold increased expression in cells, potentially a byproduct of tissue dissocia-  
80 tion (Lacar et al., 2016). 159 genes have significantly higher expression in nuclei (Figure 2D and Table S2),  
81 and they appear relevant to neuronal identity as they include connectivity and signaling genes (Figure S2A  
82 and Table S4). Based on the sum of intronic and exonic reads, these 159 nucleus-enriched genes are on aver-  
83 age more than 10-fold longer than cell-enriched genes (Figure S2B), as recently reported for single nuclei in  
84 mouse somatosensory cortex (Lake et al., 2017b). When only exonic reads were used to quantify expression  
85 in nuclei and cells, a different set of 146 genes were significantly enriched in nuclei (Table S3) and were only  
86 slightly longer than cell-enriched genes. These genes were not associated with neuron-specific functions and  
87 were significantly enriched for genes that participate in pre-mRNA splicing.

## 88 **Intronic reads are required for high-resolution cell type identification from snRNA-** 89 **seq**

90 Next, we applied an iterative clustering procedure (see Methods and Figure S3) to identify clusters of single  
91 nuclei and cells that share gene expression profiles. To assess cluster robustness, we repeated clustering  
92 100 times using 80% random subsets of nuclei and cells and calculated the proportion of clustering runs in  
93 which each pair of samples clustered together. Co-clustering matrices were reordered using Ward's hierar-  
94 chical clustering and represented as heatmaps with coherent clusters ordered as squares along the diagonal  
95 (Figure 3A,B).

96 Clustering includes two steps – differentially expressed (DE) gene selection and distance measurement – that  
97 are particularly sensitive to expression quantification. We repeated clustering using intronic and exonic reads  
98 or only exonic reads for these steps, and ordered co-clustering matrices to match the results using all reads  
99 for both steps. When using introns and exons, we found 11 distinct clusters of nuclei and cells, and clusters  
100 had similar cohesion (average within cluster co-clustering) and separation (average co-clustering difference  
101 with the closest cluster) (Figure 3C). Including intronic reads for either clustering step increased the number  
102 of clusters detected for nuclei but not cells. Therefore, accounting for intronic reads in snRNA-seq was  
103 critical to enable high-resolution cluster detection equivalent to that observed with scRNA-seq.

## 104 Equivalent cell types identified with nuclei and cells

105 We used hierarchical clustering of median gene expression values in each cluster to determine the relationships  
106 between clusters. We find that cluster relationships represented as dendrograms are remarkably similar for  
107 nuclei and cells (Figure 4A). We compared the 11 clusters identified with single nuclei and cells to reported  
108 cell types in mouse VISp (Tasic et al., 2016). Each nucleus and cell cluster could be linked to a reported cell  
109 type (Figure S4A) and to each other (Figure 4B) based on correlated expression of marker genes. Many genes  
110 contributed to high expression correlations ( $r > 0.85$ ) for all cluster pairs (Figure S4B). Conserved marker  
111 gene expression confirmed that the same 11 cell types were identified with nuclei and cells (Figure 4C).  
112 These cell types included nine excitatory neuron types from layers 4-6 and two inhibitory interneuron types.  
113 Matched cluster proportions were mostly consistent, except two closely related layer 5a subtypes were under-  
114 (L5a Batf3) or over-represented (L5a Hsd11b1) among cells (Figure S4C). This demonstrated that the initial  
115 matching of cells to nuclei was relatively unbiased.

116 We hypothesized that most intronic reads were mapped to nuclear transcripts, so quantifying gene expression  
117 in cells using only introns would approximate nuclear expression. This was supported by higher correlations  
118 of average expression across all nuclei and cells using only intronic reads as compared to only exonic reads  
119 (Figure S4D). Thus, a dendrogram based on the median expression (quantified using only intronic reads) of  
120 nuclei and cell clusters paired all matching cell types, except for two closely related layer 5b subtypes (Fig-  
121 ure 4D). Therefore, intronic reads can help facilitate comparisons between data sets derived from snRNA-seq  
122 and scRNA-seq although small expression differences remain. A dendrogram based on exonic reads grouped  
123 clusters first by sample type (nuclei and cells) and then by broad cell class (inhibitory and excitatory neu-  
124 rons). Samples grouped by sample type likely due to differences in cytoplasmic transcripts that were profiled  
125 in cells but not nuclei. A dendrogram based on intronic reads did not show this grouping because most  
126 cytoplasmic transcripts are spliced so were quantified by exonic but not intronic reads.

127 While we detected the same cell types using nuclei and cells, we expected that gene expression captured with  
128 cells included additional information from cytoplasmic transcripts. We compared the separation of matched  
129 pairs of clusters based on co-clustering and found that all nuclei and cell clusters were similarly distinct,  
130 except using single cells significantly increased the separation of two pairs of similar types: L4 Arf5 from L5a  
131 Hsd11b1 and L5b Cdh13 from L5b Tph2 (Figure 4E). Next, we compared how well genes marked cell types  
132 by calculating the degree of binary expression. Cell marker scores were, on average, 15% higher than nuclei  
133 scores due to fewer expression dropouts in cells (Figure 4F), and this was consistent with mildly improved  
134 cluster separation.

## 135 Nuclear content varies among cell types and for different transcripts

136 We estimated the nuclear proportion of mRNA for each cell type in two ways. Transcripts in the cytoplasm  
137 are spliced so intronic reads should be restricted to the nucleus. First, we estimated the nuclear proportion  
138 by calculating the ratio of the percentage of intronic reads in cells to the percentage of intronic reads in nuclei  
139 (Figure 5A). Second, we estimated nuclear proportions by selecting three genes (*Malat1*, *Meg3*, and *Snhg11*)  
140 with the highest expression in nuclei (Figure S4D) and calculating the ratio of the average expression in cells  
141 versus nuclei (Figure 5B and Figure S5A). Both methods predicted that L4 Arf5 and L5a Hsd11b1 had a  
142 significantly larger proportion of transcripts located in the nucleus compared to other cell types (Figure 5C).

143 Based on the comparison of scRNA-seq and snRNA-seq data, we estimate that L4 types have high nuclear  
144 to cell volume (~50%), whereas L5 types have lower nuclear to cell volume. To evaluate this finding, we  
145 measured nucleus and soma sizes of different cell types *in situ*. These types were labeled by different Cre-  
146 transgenes and a Cre-reporter. *Nr5a1*-Cre and *Scnn1a-Tg3*-Cre mice almost exclusively label two cell types

147 (L4 *Arf5* and L5a *Hsd11b1*), whereas *Rbp4*-Cre mice label all layer 5 cell types including L5a *Hsd11b1*  
148 (Figure S5B and Table S5). We measured the nuclear and cell sizes *in situ*, and calculated the nuclear  
149 proportion of each cell as the ratio of nuclear to soma volume (Figure S5C). We found that the average  
150 nuclear proportion was significantly lower for layer 5 cells compared to layer 4 cells, as predicted based on  
151 RNA-seq data (Figure 5D).

152 In addition, nuclear proportion estimates based on *in situ* size measurements were systematically higher than  
153 predicted for layer 5 but not layer 4 neurons. This could be the result of under-estimating the soma volume  
154 based on cross-sectional area measurements of these large non-spherical (pyramidal) neurons. Alternatively,  
155 layer 5 neuronal nuclei may have a lower density of nuclear transcripts or there may be cell type-specific biases  
156 in our RNA-seq based estimates. We then performed an unbiased survey of nuclear proportions across the  
157 full depth of cortex to test whether layer 4 or layer 5 neurons were exceptional compared to neurons in other  
158 layers. We found that layer 5 neurons tend to be larger and have proportionally smaller nuclei (Figure S5D)  
159 than other cortical neurons, as was recently reported in rat primary visual cortex (öckner2017?).

160 Next, we determined the nuclear versus cytoplasmic distribution of transcripts for individual genes. The  
161 nuclear proportion of 11,932 transcripts was estimated by the ratio of nuclear to whole cell expression mul-  
162 tiplied by the overall nuclear fraction of each cell type and averaged across cell types (Table S6). Different  
163 functional classes of genes had strikingly different nuclear proportions (Figure 5E). Many non-coding trans-  
164 cripts were localized in the nucleus, but some were abundantly expressed in the cytoplasm, such as the long  
165 non-coding RNA (lncRNA) *Tunax* that is highly enriched in the brain, is conserved across vertebrates, and  
166 has been associated with striatal pathology in Huntington’s disease (Lin et al., 2014). Most protein-coding  
167 transcripts were expressed in both the nucleus and cytoplasm with a small number restricted to the nucleus,  
168 including the Parkinson’s risk gene *Park2*. We found that pseudogenes were almost exclusively cytoplasmic  
169 and were highly enriched for house-keeping functions.

170 We compared our estimates of nuclear enrichment in cortex to mouse liver and pancreas based on data  
171 from Halpern et al. (2015) and found moderately high correlation ( $r = 0.61$ ) between 4,373 mostly house-  
172 keeping genes that were expressed in all three tissues. Moreover, the shape of the distributions of nuclear  
173 transcript proportions was highly similar between tissues with slightly higher proportions estimated in this  
174 study. These results suggest that the mechanisms regulating the spatial localization of these transcripts – for  
175 example, rates of nuclear export and cytoplasmic degradation (Halpern et al., 2015) – are conserved across  
176 cell types.

177 Surprisingly, non-coding genes and pseudogenes are better markers of cell types, on average, than protein-  
178 coding genes (Figure 5F). lncRNAs are known to have more specific expression among diverse human cell  
179 lines (Djebali et al., 2012), and we show that this is also true for neuronal types in the mouse cortex. Many  
180 pseudogene transcripts, most of which are enriched in the cytoplasm, were selectively depleted in the two  
181 cell types, L4 *Arf5* and L5a *Hsd11b1*. This is consistent with our previous analysis that showed that neurons  
182 of these types have relatively less cytoplasm. We also find that nucleus-enriched transcripts are slightly  
183 better cell-type markers than cytoplasm-enriched transcripts, although this is highly variable across genes  
184 (Figure 5G).

185 Finally, we compared our estimates of nuclear localization of transcripts for three genes – *Calb1*, *Grik1*,  
186 and *Pvalb* – to relative counts of transcripts in nuclei and cytoplasm using multiplex RNA fluorescence *in*  
187 *situ* hybridization (mFISH). We found that the relative nuclear proportions estimated by scRNA-seq and  
188 mFISH were consistent although the absolute levels were quite variable (Figure 5H). Both methods confirmed  
189 that *Pvalb* transcripts were mostly excluded from the nucleus, and this explained why 2 out of 35 nuclei  
190 in the *Pvalb*-positive interneuron type (*Pvalb* Wt1) had no detectable *Pvalb* expression, whereas all cells of

191 this cell type had robust *Pvalb* expression.

## 192 Discussion

193 Unlike scRNA-seq, snRNA-seq enables transcriptomic profiling of tissues that are refractory to whole cell  
194 dissociation and archived frozen specimens. snRNA-seq is also less susceptible to perturbations in gene  
195 expression that occur during cell isolation, such as increased expression of immediate early genes that can  
196 obscure transcriptional signatures of neuronal activity (Lacar et al., 2016). However, these advantages come  
197 at the cost of profiling less mRNA, and until this study, it was unclear if the nucleus contained sufficient  
198 number and diversity of transcripts to distinguish highly related cell types.

199 To directly address this question, we profiled a well-matched set of 463 nuclei and 463 cells from layer 5 of  
200 mouse primary visual cortex and identified 11 matching neuronal types: 2 interneuron types and 9 similar  
201 excitatory neuron types. Including intronic reads in gene expression quantification was necessary to achieve  
202 high-resolution cell type identification from single nuclei. Intronic reads substantially increased gene detection  
203 to 7000 genes per nucleus. In addition, intronic reads were more frequently derived from long genes that  
204 are known to have brain-specific expression (Gabel et al., 2015) and that help define neuronal connectivity  
205 and signaling. Intronic reads may also reflect other cell-type specific features, such as retained introns or  
206 alternative isoforms. For example, intron retention provides a mechanism for the nuclear storage and rapid  
207 translation of long transcripts in response to neuronal activity (Mauger et al., 2016).

208 We found that nuclei contain at least 20% of all cellular transcripts, and this percentage varies among cell  
209 types. Two small pyramidal neuron types have large nuclei relative to cell size that contain more than half  
210 of all transcripts. We detect 4000 more genes in single cells than single nuclei, but the majority of genes are  
211 detected equally well in both. Cytoplasm-enriched transcripts are missed by profiling single nuclei but include  
212 mostly house-keeping genes and pseudogenes, which are not related to neuronal identity. Nucleus-enriched  
213 transcripts include protein-coding and non-coding genes that are more likely to be cell-type markers than  
214 cytoplasmic transcripts. Overall, single cells do provide somewhat better detection of cell-type marker genes,  
215 thereby resulting in slightly better cluster separation for two pairs of highly similar cell types. Therefore, as  
216 more nuclei and cells are profiled, it is possible that finer discrimination of cell types may require single cell  
217 profiling. However, the benefits of profiling single nuclei may outweigh potential loss in the finest cell type  
218 resolution.

219 snRNA-seq is well suited for large-scale surveys of cellular diversity in various tissues and has the potential to  
220 be less cell-type biased. For example, single cell profiling of adult human cortex isolated more interneurons  
221 than excitatory neurons (Darmanis et al., 2015), whereas single nucleus profiling of the same tissue type  
222 isolated 30% interneurons and 70% excitatory neurons (Lake et al., 2016), close to the proportions found *in*  
223 *situ*. snRNA-seq also enables the use of stored frozen specimens to study cell types that will inform our  
224 understanding of human diversity and disease. As large scale initiatives begin to characterize transcriptomic  
225 cell types in the whole brain (Ecker et al., 2017) and whole organism (Regev et al., 2017), it is important to  
226 understand the strengths and limitations of each mRNA profiling technique.



## 227 Materials and Methods

### 228 Tissue preparation

229 Tissue samples were obtained from adult (postnatal day (P) 53-59)) male and female transgenic mice carrying  
230 a Cre transgene and a Cre-reporter transgene. Mice were anesthetized with 5% isoflurane and intracardially  
231 perfused with either 25 or 50 ml of ice cold, oxygenated artificial cerebral spinal fluid (ACSF) at a flow  
232 rate of 9 ml per minute until the liver appeared clear, or the full volume of perfusate had been flushed  
233 through the vasculature. The ACSF solution consisted of 0.5mM CaCl<sub>2</sub>, 25mM D-Glucose, 98mM HCl, 20mM  
234 HEPES, 10mM MgSO<sub>4</sub>, 1.25mM NaH<sub>2</sub>PO<sub>4</sub>, 3mM Myo-inositol, 12mM N-acetylcysteine, 96mM N-methyl-  
235 D-glucamine, 2.5mM KCl, 25mM NaHCO<sub>3</sub>, 5mM sodium L-Ascorbate, 3mM sodium pyruvate, 0.01mM  
236 Taurine, and 2mM Thiourea. The brain was then rapidly dissected and mounted for coronal slice preparation  
237 on the chuck of a Compresstome VF-300 vibrating microtome (Precisionary Instruments). Using a custom  
238 designed photodocumentation configuration (Mako G125B PoE camera with custom integrated software),  
239 a blockface image was acquired before each section was sliced at 250 μm intervals. The slice was then  
240 hemisected along the midline, and both hemispheres were then transferred to chilled, oxygenated ACSF.

241 Each slice-hemisphere was transferred into a Sylgard-coated dissection dish containing 3 ml of chilled, oxy-  
242 genated ACSF. Brightfield and fluorescent images between 4X and 20X were obtained of the intact tissue with  
243 a Nikon Digital Sight DS-Fi1 or a Sentech STC-SC500POE camera mounted to a Nikon SMZ1500 dissecting  
244 microscope. To guide anatomical targeting for dissection, boundaries were identified by trained anatomists,  
245 comparing the blockface image and the slice image to a matched plane of the Allen Reference Atlas. In  
246 general, three to five slices were sufficient to capture the targeted region of interest, allowing for expression  
247 analysis along the anterior/posterior axis. The region of interest was then dissected and both brightfield and  
248 fluorescent images of the dissections were acquired for secondary verification. The dissected regions were  
249 transferred in ACSF to a microcentrifuge tube, and stored on ice. This process was repeated for all slices  
250 containing the target region of interest, with each region of interest deposited into a new microcentrifuge  
251 tube.

252 For whole cell dissociation, after all regions of interest were dissected, the ACSF was removed and 1 ml of  
253 a 2 mg/ml pronase in ACSF solution was added. Tissue was digested at room temperature (approximately  
254 22°C) for a duration that consisted of adding 15 minutes to the age of the mouse (in days; *i.e.*, P53 specimen  
255 had a digestion time of 68 minutes). After digestion, the pronase solution was removed and replaced by  
256 1 ml of ACSF supplemented with 1% Fetal Bovine Serum (FBS). The tissue was washed two more times  
257 with the same solution and the sample was then triturated using fire-polished glass pipettes of decreasing  
258 bore sizes (600, 300, and 150 μm). The cell suspension was incubated on ice in preparation for fluorescence-  
259 activated cell sorting (FACS). FACS preparation involved adding 4'-6-diamidino-2-phenylindole (DAPI) at  
260 a final concentration of 4 μg/ml to label dead (DAPI+) versus live (DAPI-) cells. The suspension was then  
261 filtered through a fine-mesh cell strainer to remove cell aggregates. Cells were sorted by excluding DAPI  
262 positive events and debris, and gating to include red fluorescent events (tdTomato-positive cells). Single  
263 cells were collected into strip tubes containing 11.5μl of collection buffer (SMART-Seq v4 lysis buffer 0.83x,  
264 Clontech #634894), RNase Inhibitor (0.17U/μl), and ERCCs (External RNA Controls Consortium, MIX1  
265 at a final dilution of 1x10<sup>-8</sup>) (Baker et al., 2005; Risso et al., 2014). After sorting, strip tubes containing  
266 single cells were centrifuged briefly and then stored at -80°C.

267 For nuclei isolation, dissected regions of interest were transferred to microcentrifuge tubes, snap frozen in a  
268 slurry of dry ice and ethanol, and stored at -80°C until the time of use. To isolate nuclei, frozen tissues were  
269 placed into a homogenization buffer that consisted of 10mM Tris pH 8.0, 250mM sucrose, 25mM KCl, 5mM  
270 MgCl<sub>2</sub>, 0.1% Triton-X 100, 0.5% RNasin Plus RNase inhibitor (Promega), 1X protease inhibitor (Promega),  
271 and 0.1mM DTT. Tissues were placed into a 1ml dounce homogenizer (Wheaton) and homogenized using 10

272 strokes of the loose dounce pestle followed by 10 strokes of the tight pestle to liberate nuclei . Homogenate  
273 was strained through a 30 $\mu$ m cell strainer (Miltenyi Biotech) and centrifuged at 900xg for 10 minutes to pellet  
274 nuclei. Nuclei were then resuspended in staining buffer containing 1X PBS supplemented with 0.8% nuclease-  
275 free BSA and 0.5% RNasin Plus RNase inhibitor. Mouse anti-NeuN antibody (EMD Millipore, MAB377,  
276 Clone A60) was added to the nuclei at a final dilution of 1:1000 and nuclei suspensions were incubated at  
277 4°C for 30 minutes. Nuclei suspensions were then centrifuged at 400xg for 5 minutes and resuspended in  
278 clean staining buffer (1X PBS, 0.8% BSA, 0.5% RNasin Plus). Secondary antibody (goat anti-mouse IgG  
279 (H+L), Alexa Fluor 594 conjugated, ThermoFisher Scientific) was applied to nuclei suspensions at a dilution  
280 of 1:5000 for 30 minutes at 4°C. After incubation in secondary antibody, nuclei suspensions were centrifuged  
281 at 400xg for 5 minutes and resuspended in clean staining buffer. Prior to FACS, DAPI was applied to nuclei  
282 suspensions at a final concentration of 0.1 $\mu$ g/ml and nuclei suspensions were filtered through a 35 $\mu$ m nylon  
283 mesh to remove aggregates. Single nuclei were captured by gating on DAPI-positive events, excluding debris  
284 and doublets, and then gating on Alexa Fluor 594 (NeuN) signal. Strip tubes containing FACS isolated  
285 single nuclei were then briefly centrifuged and frozen at -80°C.

## 286 RNA amplification and library preparation for RNA-seq

287 The SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech #634894) was used per the ma-  
288 nufacturer's instructions for reverse transcription of single cell RNA and subsequent cDNA synthesis. Single  
289 cells were stored in 8-strips at -80°C in 11.5  $\mu$ l of collection buffer (SMART-Seq v4 lysis buffer at 0.83x,  
290 RNase Inhibitor at 0.17 U/ $\mu$ l, and ERCC MIX1 at a final dilution of 1x10<sup>-8</sup> dilution). Twelve to 24 8-well  
291 strips were processed at a time (the equivalent of 1-2 96-well plates). At least 1 control strip was used per  
292 amplification set, containing 2 wells without cells but including ERCCs, 2 wells without cells or ERCCs, and  
293 either 4 wells of 10 pg of Mouse Whole Brain Total RNA (Zyagen, MR-201) or 2 wells of 10 pg of Mouse  
294 Whole Brain Total RNA (Zyagen, MR-201) and 2 wells of 10 pg Control RNA provided in the Clontech  
295 kit. Mouse whole cells were subjected to 18 PCR cycles after the reverse transcription step, whereas mouse  
296 nuclei were subjected to 21 PCR cycles. AMPure XP Bead (Agencourt AMPure beads XP PCR, Beckman  
297 Coulter A63881) purification was done using the Agilent Bravo NGS Option A instrument. A bead ratio of  
298 1x was used (50  $\mu$ l of AMPure XP beads to 50  $\mu$ l cDNA PCR product with 1  $\mu$ l of 10x lysis buffer added, as  
299 per Clontech instructions), and purified cDNA was eluted in 17  $\mu$ l elution buffer provided by Clontech. All  
300 samples were quantitated using PicoGreen® on a Molecular Dynamics M2 SpectraMax instrument. A por-  
301 tion of the samples, and all controls, were either run on the Agilent Bioanalyzer 2100 using High Sensitivity  
302 DNA chips or the Advanced Analytics Fragment Analyzer (96) using the High Sensitivity NGS Fragment  
303 Analysis Kit (1bp-6000bp) to qualify cDNA size distribution. An average of 7.3 ng of cDNA was synthesized  
304 across all non-control samples. Purified cDNA was stored in 96-well plates at -20°C until library preparation.

305 Sequencing libraries were prepared using NexteraXT (Illumina, FC-131-1096) with NexteraXT Index Kit  
306 V2 Set A (FC-131-2001). NexteraXT libraries were prepared at 0.5x volume, but otherwise followed the  
307 manufacturer's instructions. An aliquot of each amplified cDNA sample was first normalized to 30 pg/ $\mu$ l  
308 with Nuclease-Free Water (Ambion), then this normalized sample aliquot was used as input material into  
309 the NexteraXT DNA Library Prep (for a total of 75pg input). AMPure XP bead purification was done using  
310 the Agilent Bravo NGS Option A instrument. A bead ratio of 0.9x was used (22.5  $\mu$ l of AMPure XP beads  
311 to 25  $\mu$ l library product, as per Illumina protocol), and all samples were eluted in 22  $\mu$ l of Resuspension  
312 Buffer (Illumina). All samples were run on either the Agilent Bioanalyzer 2100 using High Sensitivity DNA  
313 chips or the Advanced Analytics Fragment Analyzer (96) using the High Sensitivity NGS Fragment Analysis  
314 Kit (1bp-6000bp) to for sizing. All samples were quantitated using PicoGreen using a Molecular Dynamics  
315 M2 SpectraMax instrument. Molarity was calculated for each sample using average size as reported by  
316 Bioanalyzer or Fragment Analyzer and pg/ $\mu$ l concentration as determined by PicoGreen. Samples (5  $\mu$ l  
317 aliquot) were normalized to 2-10 nM with Nuclease-free Water (Ambion), then 2  $\mu$ l from each sample within  
318 one 96-index set was pooled to a total of 192  $\mu$ l at 2-10 nM concentration. A portion of this library pool  
319 was sent to an outside vendor for sequencing on an Illumina HS2500. All of the library pools were run using



320 Illumina High Output V4 chemistry. Covance Genomics Laboratory, a Seattle-based subsidiary of LabCorp  
321 Group of Holdings, performed the RNA-Sequencing services. An average of 229 M reads were obtained per  
322 pool, with an average of 2.0-3.1 M reads/cell across the entire data set.

### 323 RNA-Seq data processing

324 Raw read (fastq) files were aligned to the GRCm38 mouse genome sequence (Genome Reference Consortium,  
325 2011) with the RefSeq transcriptome version GRCm38.p3 (current as of 1/15/2016) and updated by remov-  
326 ing duplicate Entrez gene entries from the gtf reference file for STAR processing. For alignment, Illumina  
327 sequencing adapters were clipped from the reads using the fastqMCF program (Aronesty, 2011). After clip-  
328 ping, the paired-end reads were mapped using Spliced Transcripts Alignment to a Reference (STAR) (Dobin  
329 et al., 2013) using default settings. STAR uses and builds its own suffix array index which considerably  
330 accelerates the alignment step while improving sensitivity and specificity, due to its identification of alterna-  
331 tive splice junctions. Reads that did not map to the genome were then aligned to synthetic constructs (i.e.  
332 ERCC) sequences and the *E.coli* genome (version ASM584v2). Quantification was performed using summer-  
333 izeOverlaps from the R package GenomicAlignments (Lawrence et al., 2013). Read alignments to the genome  
334 (exonic, intronic, and intergenic counts) were visualized as beeswarm plots using the R package *beeswarm*.

335 Expression levels were calculated as counts per million (CPM) of exonic plus intronic reads, and  $\log_2(\text{CPM}$   
336  $+ 1)$  transformed values were used for a subset of analyses as described below. Gene detection was calculated  
337 as the number of genes expressed in each sample with  $\text{CPM} > 0$ . CPM values reflected absolute transcript  
338 number and gene length, i.e. short and abundant transcripts may have the same apparent expression level  
339 as long but rarer transcripts. Intron retention varied across genes so no reliable estimates of effective gene  
340 lengths were available for expression normalization. Instead, absolute expression levels were estimated as  
341 fragments per kilobase per million (FPKM) using only exonic reads so that annotated transcript lengths  
342 could be used.

### 343 Selection of single nuclei and matched cells

344 463 of 487 (95%) of single nuclei isolated from layer 5 of mouse VISp passed quality control criteria:  $>500,000$   
345 genome-mapped reads,  $>75\%$  reads aligned, and  $>50\%$  unique reads. 12,866 single cells isolated from layers  
346 1-6 of mouse VISp passed quality control criteria:  $>200,000$  transcriptome mapped reads and  $>1000$  genes  
347 detected ( $\text{CPM} > 0$ ).

348 Gene expression was more likely to drop out in samples with lower quality cDNA libraries and for low ex-  
349 pressing genes. To estimate gene dropouts due to stochastic transcription or technical artifacts (Kharchenko  
350 et al., 2014), expression noise models were fit separately to single nuclei and cells using the “knn.error.models”  
351 function of the R package *scde* (version 2.2.0) with default settings and eight nearest neighbors. Noise models  
352 were used to calculate a dropout weight matrix that represented the likelihood of expression dropouts based  
353 on average gene expression levels of similar nuclei or cells using mode-relative weighting (*dbm*). The prob-  
354 ability of dropout for each sample (*s*) and gene (*g*) was estimated based on two expression measurements:  
355 average expected expression level of similar samples,  $p(x_{\bar{g}})$ , and observed expression levels,  $p(x_{sg})$ , using  
356 the “scde.failure.probability” and “scde.posterior” functions. The dropout weighting was calculated as a  
357 combination of these probabilities:  $W_{sg} = 1 - \sqrt{p(x_{sg}) \cdot \sqrt{p(x_{sg}) \cdot p(x_{\bar{g}})}}$ .

358 Dropout weighted Pearson correlations were calculated between all pairs of nuclei and cells using 42,003  
359 genes expressed in at least one nucleus and one cell. The cell with the highest correlation to any nucleus

360 was selected as the best match, and this cell and nucleus were removed from further analysis. This process  
361 was repeated until 463 best matching cells were selected, and the expression correlations were compared to  
362 correlations of the best matching pairs of nuclei (Figure 1B). The Cre-lines and dissected cortical layers of  
363 origin of the best matching cells were summarized as bar plots (Figure S1). Unweighted Pearson correlations  
364 were also calculated between all pairs of nuclei and cells to test the effect of accounting for dropouts on  
365 sample similarities (Figure 2B).

## 366 Differential expression analysis

367 Gene detection was estimated as the proportion of cells and nuclei expressing each gene ( $\text{CPM} > 0$ ). In order  
368 to estimate the expected variability of gene detection as a result of population sampling, cells were randomly  
369 split into two sets of 231 and 232 cells and genes were grouped into 50 bins based on detection in the first  
370 set of cells. For each bin of genes, the 97.5 percentile of detection was calculated for the second set of cells.  
371 A 95% confidence interval of gene detection was constructed by reflecting these binned quantiles across  
372 the line of unity. Data were summarized with a hexagonal binned scatter plot and a log-transformed color  
373 scale using the R package *ggplot2* (Wickham, 2009).

374 Differential expression between nuclei and cells was calculated with the R package *limma* (Ritchie et al.,  
375 2015) using default settings and  $\log_2(\text{CPM} + 1)$  expression defined based on two sets of reads: introns plus  
376 exons and only exons. Significantly differentially expressed were defined as having  $>1.5$ -fold change and  
377 a Benjamini-Hochberg corrected P-value  $< 0.05$ . Gene expression distributions of nuclei or cells within a  
378 cluster were visualized using violin plots, density plots rotated 90 degrees and reflected on the Y-axis.

379 Differences in alignment statistics and gene counts were calculated between cells, nuclei, and total RNA  
380 controls (or just cells and nuclei) with analysis of variance using the “aov” function in R (Chambers et al.,  
381 1992). P-values for all comparisons were  $P < 10^{-13}$ .

382 Two sets of nucleus- and cell-enriched genes (introns plus exons and exons only) were tested for gene ontology  
383 (GO) enrichment using the ToppGene Suite (Chen et al., 2009). Significantly enriched (Benjamini-Hochberg  
384 false discovery rate  $< 0.05$ ) GO terms were summarized as tree maps with box sizes proportional to  $-\log_{10}(\text{P-}$   
385 values) using REVIGO (Supek et al., 2011b) (Figure S2).

## 386 Clustering

387 Nuclei and cells were grouped into transcriptomic cell types using an iterative clustering procedure based  
388 on community detection in a nearest neighbor graph as described in Levine et al. (2015). Clustering was  
389 performed using gene expression quantified with exonic reads only or intronic plus exonic reads for two key  
390 clustering steps: selecting significantly variable genes and calculating pairwise similarities between nuclei.  
391 Four combinations of expression quantification for nuclei and cells resulted in eight independent clustering  
392 runs.

393 For each gene,  $\log_2(\text{CPM} + 1)$  expression was centered and scaled across samples. Noise models were used to  
394 select significantly variable genes (adjusted variance  $> 1.25$ ). Dimensionality reduction was performed with  
395 principal components analysis (PCA) on variable genes, and the covariance matrix was adjusted to account  
396 for gene dropouts using the product of dropout weights across genes for each pair of samples. A maximum  
397 of 20 principal components (PCs) were retained for which more variance was explained than the broken stick  
398 null distribution, a conservative method of PC retention (Jackson, 1993).

399 Nearest-neighbor distances between all samples were calculated using the “nn2” function of the R pack-  
400 age *RANN*, and Jaccard similarity coefficients between nearest-neighbor sets were computed. Jaccard coeffi-  
401 cients measured the proportion of nearest neighbors shared by each sample and were used as edge weights in  
402 constructing an undirected graph of samples. Louvain community detection was used to cluster this graph  
403 with 15 nearest neighbors. Considering more than 15 neighbors reduced the power to detect small clusters  
404 due to the resolution limit of community detection (Fortunato and Barthelemy, 2007). Considering fewer  
405 than 15 neighbors increased over-splitting, as expected based on simulations by Reichardt and Bornholdt  
406 (2006). Fewer nearest neighbors were used only when there were 15 or fewer samples total.

407 Clustering significance was tested by comparing the observed modularity to the expected modularity of an  
408 Erdős-Rényi random graph with a matching number of nodes and average connection probability. Expected  
409 modularity was calculated as the maximum estimated by two reported equations ( $\Omega$ ; Reichardt and Born-  
410 holdt, 2006). Samples were split into clusters only if the observed modularity was greater than the expected  
411 modularity, and only clusters with distinct marker genes were retained. Marker genes were defined for all  
412 cluster pairs using two criteria: 1) significant differential expression (Benjamini-Hochberg false discovery  
413 rate  $< 0.05$ ) using the R package *limma* and 2) either binary expression (CPM  $> 1$  in  $>50\%$  samples in one  
414 cluster and  $<10\%$  in the second cluster) or  $>100$ -fold difference in expression. Pairs of clusters were merged  
415 if either cluster lacked at least one marker gene.

416 Clustering was applied iteratively to each sub-cluster until the occurrence of one of four stop criteria: 1)  
417 fewer than six samples (due to a minimum cluster size of three); 2) no significantly variable genes; 3) no  
418 significantly variable PCs; 4) no significant clusters.

419 To assess the robustness of clusters, the iterative clustering procedure described above was repeated 100 times  
420 for random sets of 80% of samples. A co-clustering matrix was generated that represented the proportion of  
421 clustering iterations that each pair of samples were assigned to the same cluster. Average-linkage hierarchical  
422 clustering was applied to this matrix followed by dynamic branch cutting using “cutreeHybrid” in the R  
423 package *WGCNA* (Langfelder et al., 2007) with cut height ranging from 0.01 to 0.99 in steps of 0.01. A cut  
424 height was selected that resulted in the median number of clusters detected across all 100 iterations. Cluster  
425 cohesion (average within cluster co-clustering) and separation (difference between within cluster co-clustering  
426 and maximum between cluster co-clustering) was calculated for all clusters. Marker genes were defined for  
427 all cluster pairs as described above, and clusters were merged if they had a co-clustering separation  $<0.25$   
428 or either cluster lacked at least one marker gene.

## 429 Scoring marker genes based on cluster specificity

430 Many genes were expressed in the majority of nuclei or cells in a subset of clusters. A marker score (beta)  
431 was defined for all genes to measure how binary expression was among clusters, independent of the number  
432 of clusters labeled. First, the proportion ( $x_i$ ) of samples in each cluster that expressed a gene above back-  
433 ground level (CPM  $> 1$ ) was calculated. Then, scores were defined as the squared differences in proportions  
434 normalized by the sum of absolute differences plus a small constant ( $\epsilon$ ) to avoid division by zero. Scores  
435 ranged from 0 to 1, and a perfectly binary marker had a score equal to 1.

$$\beta = \frac{\sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2}{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j| + \epsilon}.$$

## 436 Cluster dendrograms

437 Clusters were arranged by transcriptomic similarity based on hierarchical clustering. First, the average  
438 expression level of the top 1200 marker genes (i.e. highest beta scores) was calculated for each cluster.  
439 A correlation-based distance matrix ( $D_{xy} = \frac{1-\rho(x,y)}{2}$ ) was calculated, and complete-linkage hierarchical  
440 clustering was performed using the “hclust” R function with default parameters. The resulting dendrogram  
441 branches were reordered to show inhibitory clusters followed by excitatory clusters, with larger clusters first,  
442 while retaining the tree structure. Note that this measure of cluster similarity is complementary to the  
443 co-clustering separation described above. For example, two clusters with similar gene expression patterns  
444 but a few binary marker genes may be close on the tree but highly distinct based on co-clustering.

## 445 Matching clusters based on marker gene expression

446 Nuclei and cell clusters were independently compared to published mouse VISp cell types (Tasic et al.,  
447 2016). The proportion of nuclei or cells expressing each gene with CPM > 1 was calculated for all clusters.  
448 Approximately 400 genes were markers in both data sets (beta score > 0.3) and were expressed in the  
449 majority of samples of between one and five clusters. Markers expressed in more than five clusters were  
450 excluded to increase the specificity of cluster matching. Weighted correlations were calculated between all  
451 pairs of clusters across these genes and weighted by beta scores to increase the influence of more informative  
452 genes. Heatmaps were generated to visualize all cluster correlations. All nuclei and cell clusters had reciprocal  
453 best matching clusters from Tasic et al. and were labeled based on these reported cluster names.

454 Next, nuclei and cell clusters were directly compared using the above analysis. All 11 clusters had reciprocal  
455 best matches that were consistent with cluster labels assigned based on similarity to published types. The  
456 most highly conserved marker genes of matching clusters were identified by selecting genes expressed in a  
457 single cluster (>50% of samples with CPM > 1) and with the highest minimum beta score between nuclei  
458 and cell clusters. Two additional marker genes were identified that discriminated two closely related clusters.  
459 Violin plots of marker gene expression were constructed with each gene on an independent, linear scale.

460 Nuclei and cell clusters were also compared by calculating average cluster expression based only on intronic  
461 or exonic reads and calculating a correlation-based distance using the top 1200 marker genes as described  
462 above. Hierarchical clustering was applied to all clusters quantified using the two sets of reads. In addition,  
463 the average  $\log_2(\text{CPM} + 1)$  expression across all nuclei and cells was calculated using intronic or exonic  
464 reads.

465 Cluster separation was calculated for individual nuclei and cells as the average within cluster co-clustering  
466 of each sample minus the maximum average between cluster co-clustering. Separations for matched pairs of  
467 clusters were visualized with box plots and compared using a Student’s *t*-test, and significance was tested  
468 after Bonferroni correction for multiple testing. Finally, a linear model was fit to beta marker scores for  
469 genes that were expressed in at least one but not all cell and nuclear clusters, and the intercept was set to  
470 zero.

## 471 Estimating proportions of nuclear transcripts

472 The nuclear proportion of transcripts was estimated in two ways. First, all intronic reads were assumed to  
473 be from transcripts localized to the nucleus so that the proportion of intronic reads measured in cells should  
474 decrease linearly with the nuclear proportion of the cell as nuclear reads are diluted with cytoplasmic reads.  
475 For each cell type, the nuclear proportion was estimated as the proportion of intronic reads in cells divided

476 by the proportion of intronic reads in matched nuclei. Second, the nuclear proportion was estimated as the  
477 average ratio of cell to nuclear expression (CPM) using only exonic reads of three highly expressed nuclear  
478 genes (*Snhg11*, *Malat1*, and *Meg3*). The standard deviation of nuclear proportion estimates were calculated  
479 based on standard error propagation of variation in intronic read proportions and expression levels. Nuclear  
480 proportion estimates were compared with linear regression, and the estimate based on relative expression  
481 levels was used for further analysis.

482 The nuclear proportion of transcripts for all genes was estimated for each cell type as the ratio of average  
483 expression (CPM) in nuclei versus matched cells multiplied by the nuclear proportion of all transcripts.  
484 Estimated proportions greater than 1 were set equal to 1 for each cell type, and a weighted average proportion  
485 was calculated for each gene with weights equal to the average  $\log_2(\text{CPM} + 1)$  expression in each cell type.  
486 11,932 genes were expressed in at least one nuclear or cell cluster (>50% samples expressed with CPM >  
487 1) and were annotated as one of three gene types – protein-coding, protein non-coding, or pseudogene –  
488 using gene metadata from NCBI ([ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE\\_INFO/Mammalia/Mus\\_](ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Mus_musculus.gene_info.gz)  
489 [musculus.gene\\_info.gz](ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Mus_musculus.gene_info.gz); downloaded 10/12/2017). For each type, histograms of gene counts with different  
490 nuclear proportions were generated. Next, beta marker score distributions were visualized as violin plots,  
491 and differences across gene types were compared with a Kruskal-Wallis rank sum test followed by Wilcoxon  
492 signed rank unpaired tests. Finally, genes were grouped into 10 bins of estimated nuclear proportions, from  
493 high cytoplasmic enrichment to high nuclear enrichment, and beta marker score distributions were visualized  
494 as box plots. A linear regression was fit to marker scores versus nuclear proportion.

495 Nuclear transcript proportions were compared to nuclear proportions estimated for mouse liver and pan-  
496 creatic beta cells based on data from [Halpern et al. \(2015\)](#). Ratios of normalized nuclear and cytoplasmic  
497 transcript counts were calculated in four tissue replicates. Average ratios were calculated for genes with at  
498 least one count in either fraction in at least one tissue. Nuclear proportion estimates for all genes with data  
499 from both data sets ( $n = 4373$ ) were compared with Pearson correlation, a linear model with intercept set  
500 equal to zero, and histograms with a bin width of 0.02.

501

## 502 **Colorimetric *in situ* hybridization**

503 *In situ* hybridization data for mouse cortex was from the Allen Mouse Brain Atlas ([Lein et al., 2007](#)). All data  
504 is publicly accessible through [www.brain-map.org](http://www.brain-map.org). Data was generated using a semiautomated technology  
505 platform as described in [Lein et al. \(2007\)](#). Mouse ISH data shown is from primary visual cortex (VISp) in  
506 the Paxinos Atlas ([Paxinos et al., 2013](#)).

## 507 **Multiplex fluorescence RNA *in situ* hybridization and quantification of nuclear** 508 **versus cytoplasmic transcripts**

509 The RNAscope multiplex fluorescent kit was used according to the manufacturer's instructions for fresh  
510 frozen tissue sections (Advanced Cell Diagnostics), with the exception that 16 $\mu\text{m}$  tissue sections were fixed  
511 with 4% PFA at 4°C for 60 minutes and the protease treatment step was shortened to 15 minutes at room  
512 temperature. Probes used to identify nuclear and cytoplasmic enriched transcripts were designed antisense  
513 to the following mouse genes: *Calb1*, *Grik1*, and *Pvalb*. Following hybridization and amplification, stained  
514 sections were imaged using a 60X oil immersion lens on a Nikon TiE epifluorescence microscope.

515 To determine if spots fell within the nucleus or cytoplasm, a boundary was drawn around the nucleus to  
516 delineate its border using measurement tools within Nikon Elements software. To delineate the cytoplasmic



517 boundary of each cell, a circle with a diameter of 15 $\mu$ m was drawn and centered over the cell (Fig. 5). RNA  
518 spots in each channel were quantified manually using counting tools available in the Nikon Elements software.  
519 Spots that fell fully within the interior boundary of the nucleus were classified as nuclear transcripts. Spots  
520 that fell outside of the nucleus but within the circle that defined the cytoplasmic boundary were classified  
521 as cytoplasmic transcripts. Additionally, if spots intersected the exterior boundary of the nucleus they were  
522 classified as cytoplasmic transcripts. To prevent double counting of spots and ambiguities in assigning spots  
523 to particular cells, labeled cells whose boundaries intersected at any point along the circumference of the  
524 circle delineating their cytoplasmic boundary were excluded from the analysis. A linear regression was fit to  
525 nuclear versus soma probe counts, and the slope was used to estimate the nuclear proportion.

## 526 *In situ* quantification of nucleus and soma size

527 Coronal brain slices from *Nr5a1-Cre;Ai14*, *Scnn1a-Tg3-Cre;Ai14*, and *Rbp4-Cre-KL100;Ai14* mice were stain-  
528 ed with anti-dsRed (Clontech #632496) to enhance tdTomato signal in red channel and DAPI to label nuclei.  
529 Maximum intensity projections from six confocal stacks of 1- $\mu$ m intervals were processed for analysis. Initial  
530 segmentation was performed by CellProfiler (Lamprecht et al., 2007) to identify nuclei from the DAPI signal  
531 and soma from the tdTomato signal. Segmentation results were manually verified and any mis-segmented  
532 nuclei or somata were removed or re-segmented if appropriate. Area measurement of segmented nuclei and  
533 somata was performed in CellProfiler in Layer 4 from *Nr5a1-Cre;Ai14* and *Scnn1a-Tg3-Cre;Ai14* mice, and  
534 in Layer 5 from *Rbp4-Cre-KL100;Ai14* mice. A linear regression was fit to nuclear versus soma area to  
535 highlight the differences between Cre-lines.

536 For measurements of nucleus and soma size agnostic to Cre driver, we used 16  $\mu$ m-tissue sections from P56  
537 mouse brain. To label nuclei, DAPI was applied to the tissue sections at a final concentration of 1mg/ml.  
538 To label cell somata, tissue sections were stained with Neurotrace 500/525 fluorescent Nissl stain (Ther-  
539 moFisher Scientific) at a dilution of 1:100 in 1X PBS for 5 minutes, followed by brief washing in 1X PBS.  
540 Sections were coverslipped with Fluoromount-G (Southern Biotech) and visualized on a Nikon TiE epiflu-  
541 orescence microscope using a 40x oil objective. Soma and nuclei area measurements were taken by tracing  
542 the boundaries of the Nissl-stained soma or DAPI-stained nucleus, respectively, using cell measurement tools  
543 available in the Nikon TiE microscope software. All cells with a complete nucleus clearly present within the  
544 section were measured, except that we excluded glial cells which had very small nuclei and scant cytoplasm.  
545 Measurements were taken within a 40x field of view across an entire cortical column encompassing layers  
546 1-6, and the laminar position of each cell (measured as depth from the pial surface) was tracked along with  
547 the nucleus and soma area measurements for each cell.

548 For each cell in the experiments above, the nuclear proportion was estimated as the ratio of nucleus and soma  
549 area raised to the  $3/2$  power. This transformation was required to convert area to volume measurements  
550 and assumed that the 3-dimensional geometries of soma and nuclei were reflected by their cross-sectional  
551 profiles. This is true for approximately symmetrical shapes such as most nuclei and some somata, but will  
552 lead to under- or over-estimates of nuclear proportions for asymmetrical cells. Therefore, the estimated  
553 nuclear proportion of any individual cell may be inaccurate, but the average nuclear proportion for many  
554 cells should be relatively unbiased.

## 555 Code availability

556 Data and code to reproduce all figures are publicly available from GitHub at [https://github.com/AllenInstitute/  
557 NucCellTypes](https://github.com/AllenInstitute/NucCellTypes).

558 **Competing interests**

559 The authors declare no competing interests.

560 **Acknowledgements**

561 The authors thank the Allen Institute for Brain Science founders, P. G. Allen and J. Allen, for their vision,  
562 encouragement, and support.

563 **Figures**

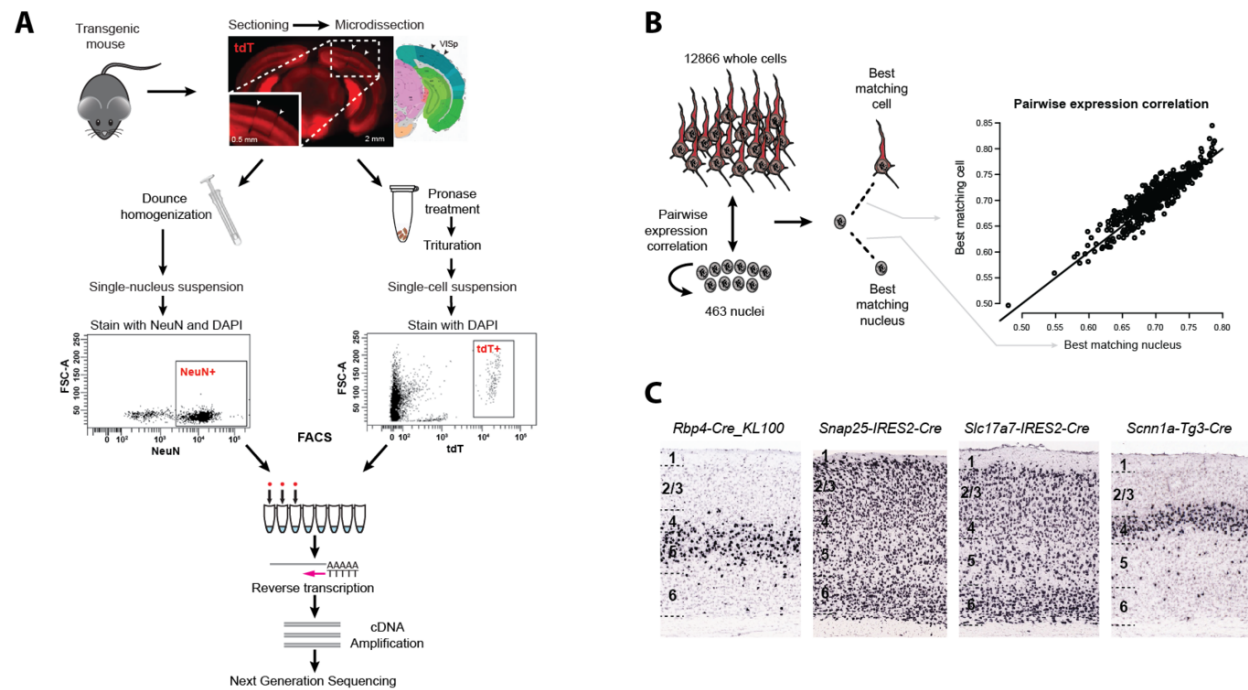


Figure 1: Identification of an expression-matched set of single nuclei and whole cells from mouse primary visual cortex (VISp). **(A)** Whole brains were dissected from transgenic mice, coronal slices were sectioned, and individual layers of VISp were microdissected. Nuclei were dissociated from layer 5, stained with DAPI and against the neuronal marker NeuN. Single NeuN-positive nuclei were isolated by fluorescence-activated cell sorting (FACS). In parallel, whole cells were dissociated from all layers, and single td-Tomato reporter-positive cells were isolated. Single nucleus and cell mRNA were reverse transcribed, amplified, and sequenced to measure transcriptome-wide expression levels. **(B)** Left: 463 nuclei from layer 5 and 12,866 whole cells from all layers passed quality control metrics, and the expression correlation was calculated between each nucleus and all other nuclei and cells. Expression similarity can vary based on sample quality, so nuclei were compared to each other to provide a baseline expected similarity. For each nucleus, the best matching nucleus and cell were selected based on maximal correlation. Right: Cells and nuclei displayed comparable expression similarities to all nuclei, with 95% of correlations between 0.63 and 0.78. This suggested that nuclei and cells were well matched. **(C)** Chromogenic RNA *In situ* hybridization (ISH) images of all VISp layers from four mouse Cre-lines from which the best matching cells were most commonly derived. As expected, all Cre-lines label cells in layer 5 and adjacent layers.

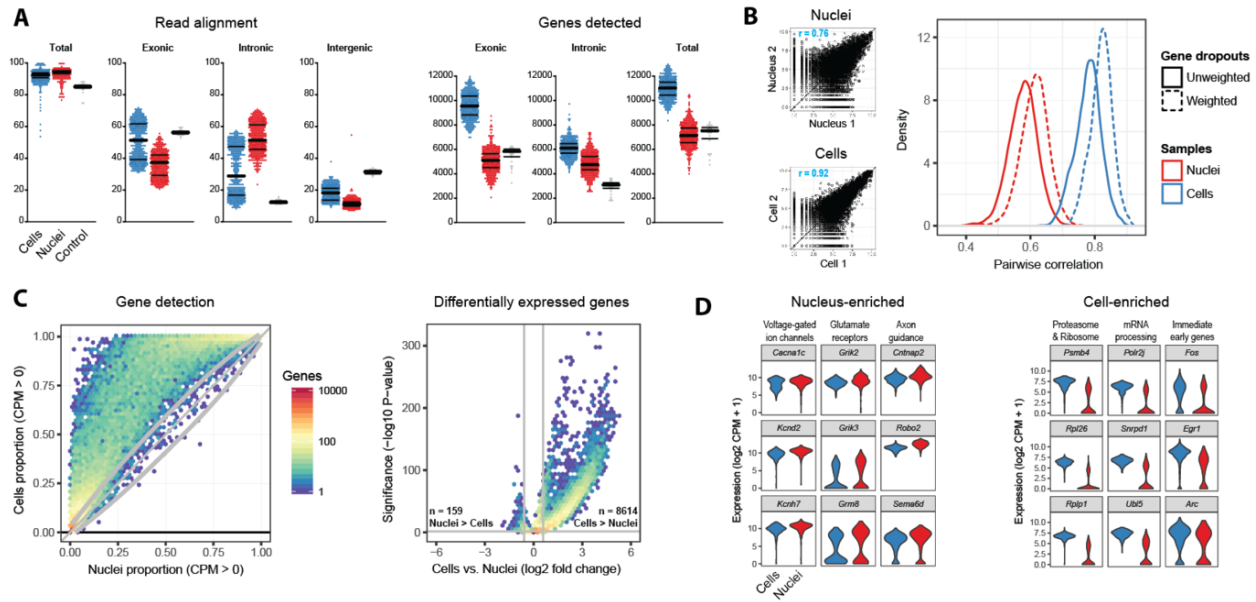


Figure 2: Comparison of nuclear and whole cell transcriptomes. **(A)** Left: Percentage of RNA-seq reads mapping to genomic regions for cells, nuclei, and whole brain control RNA. Bars indicate median and 25<sup>th</sup> and 75<sup>th</sup> quantiles. Note that among cells exonic and intronic read alignment is bimodal. Right: Gene detection (counts per million, CPM > 0) based on reads mapping to exons, introns, or both introns and exons. **(B)** Left: The most similar pair of cells have more highly correlated gene expression ( $r = 0.92$ ) than the most similar pair of nuclei ( $r = 0.76$ ), due to fewer gene dropouts. Right: Cells have consistently more similar expression to each other than nuclei, even after correcting for gene dropouts based on an expression noise model. **(C)** Left: Binned scatter plot showing all genes are detected (CPM > 0) with equal or greater reliability in cells than nuclei. Grey lines show the variation in detection that is expected by chance (95% confidence interval). Right: Binned scatter plot showing 0.4% of genes are significantly more highly expressed (fold change > 1.5, adjusted P-value < 0.05) in nuclei, and 20.5% of genes are more highly expressed in cells. The log-transformed color scale indicates the number of genes in each bin. **(D)** Nuclear enriched genes are highly enriched for genes involved in neuronal connectivity, synaptic transmission, and intrinsic firing properties. Cell enriched genes are predominantly related to mRNA processing and protein translation and degradation. In addition, immediate early gene expression is increased up to 10-fold in cells, despite comparable isolation protocols for cells and nuclei.

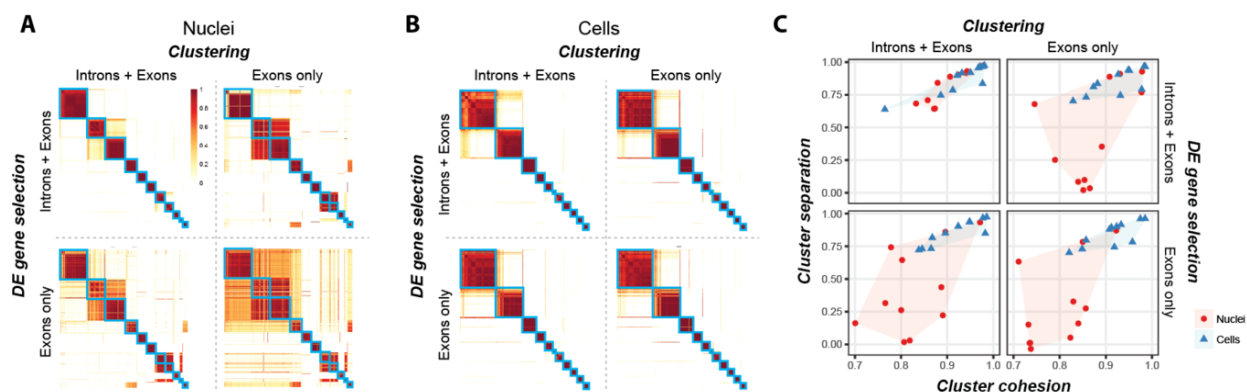


Figure 3: Single nuclei provide comparable clustering resolution to cells with inclusion of intronic reads. **(A)** Co-clustering heatmaps show the proportion of 100 clustering iterations that each pair of nuclei were assigned to the same cluster. Clustering was performed using gene expression quantified with exonic reads or intronic plus exonic reads for two key clustering steps: selecting significantly differentially expressed (DE) genes and calculating pairwise similarities between nuclei. Co-clustering heatmaps were generated for each combination of gene expression values, and blue boxes highlight 11 clusters of nuclei that consistently co-clustered using introns and exons (upper left heatmap) and were overlaid on the remaining heatmaps. The row and column order of nuclei is the same for all heatmaps. **(B)** Co-clustering heatmaps were generated for cells as described for nuclei in **(A)**, and blue boxes highlight 11 clusters of cells. **(C)** Cluster cohesion (average within cluster co-clustering) and separation (difference between within cluster co-clustering and maximum between cluster co-clustering) are plotted for nuclei and cells and all combinations of reads. Including introns in gene expression quantification dramatically increases cohesion and separation of nuclei but not cell clusters.



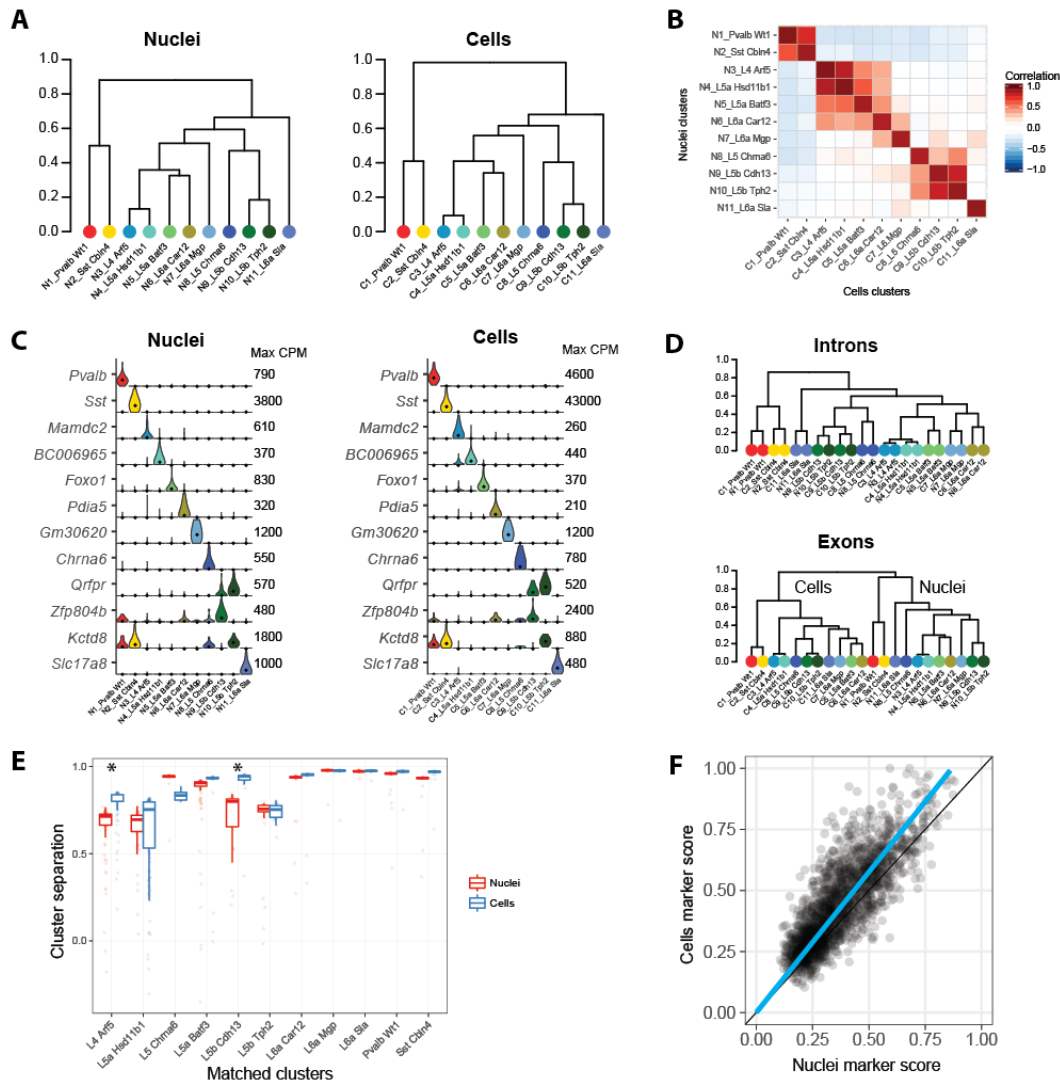


Figure 4: Equivalent neuronal cell types identified with nuclei and cells. **(A)** Cluster dendrograms for nuclei and cells based on hierarchical clustering of average expression of the top 1200 cluster marker genes. 11 clusters are labeled based on dendrogram leaf order and the closest matching mouse VISp cell type described in Tasic et al. (2016) based on correlated marker gene expression (see Figure S4). **(B)** Pairwise correlations between nuclear and cell clusters using average cluster expression of the top 490 shared marker genes. **(C)** Violin plots of cell type specific marker genes expressed in matching nuclear and cell clusters. Plots are on a linear scale, max CPM indicates the maximum expression of each gene, and black dots indicate median expression. **(D)** Hierarchical clustering of nuclear and cell clusters using the top 1200 marker genes with expression quantified by intronic or exonic reads. Intronic reads group nine matching nuclear and cell clusters together at the leaves, while two closely related deep layer 5 excitatory neuron types group by sample type. In contrast, exonic reads completely segregate clusters by sample type. **(E)** Box plots of cluster separations for all samples in matched nuclear and cell clusters. Clusters are equally well separated for all but two cell types, L4 Arf5 and L5b Cdh13, that are moderately but significantly (Wilcoxon signed rank unpaired tests; Bonferroni corrected P-value < 0.05) more distinct with cells than nuclei. **(F)** Cell type marker genes are consistently detected in nuclei and cells, although marker scores (see Methods) were on average 15% higher for cells.

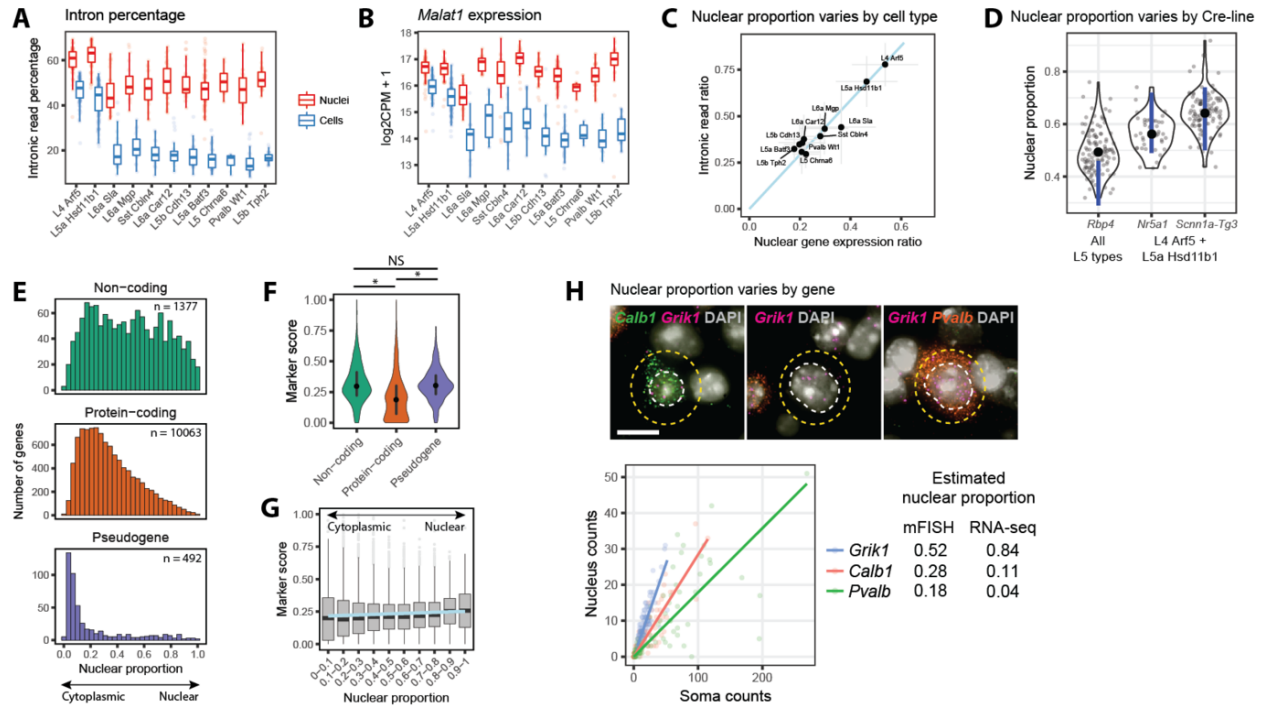


Figure 5: Nuclear transcript content varies among cell types and genes. **(A)** Box plots showing median (bars), 25<sup>th</sup> and 75<sup>th</sup> quantiles (boxes), and range (whiskers) of percentages of reads mapping to introns for matched nuclei and cell clusters. **(B)** Box plots of log<sub>2</sub>-transformed expression of the nuclear non-coding RNA, *Malat1*, in matched nuclei and cell clusters. **(C)** The nuclear fraction of transcripts in cell types was estimated with two methods: the ratio of intronic read percentages in cells compared to nuclei; and the average ratio of expression in cells compared to nuclei of three highly expressed genes (*Snhg11*, *Meg3*, and *Malat1*) that are localized to the nucleus. The relative ranking of nuclear fractions was consistent (Spearman rank correlation = 0.84), although estimates based on the intronic read ratio were consistently 50% higher. **(D)** Estimated nuclear proportion (ratio of nucleus and soma volume) of neurons labeled by three mouse Cre-lines in Layers 4 and 5 (see Supplementary Figure S5D). Single neuron measurements (grey points) were summarized as violin plots, and average nuclear proportions (black points) were compared to the range of estimated proportions (blue lines) based on intronic read ratios and nuclear gene expression. **(E)** Histograms of nuclear fraction estimates for 11,932 genes expressed (CPM > 1) in at least one nuclear or cell cluster and grouped by type of gene. **(F)** Violin plots of marker score distributions with median and inter-quartile intervals. Non-coding genes and pseudogenes are on average better markers of cell types than protein-coding genes. Kruskal–Wallis rank sum test, post hoc Wilcoxon signed rank unpaired tests: \*P < 1 × 10<sup>-50</sup> (Bonferroni-corrected), NS, not significant. **(G)** Box plots of cell type marker scores for genes grouped by estimated nuclear enrichment. Nucleus-enriched genes have significantly higher marker scores (linear regression; P = 2.3 × 10<sup>-8</sup>). **(H)** Validation of the estimated nuclear proportion of transcripts for *Calb1*, *Grik1*, and *Pvalb* using multiplex fluorescent *in situ* hybridization (mFISH). Top: For each gene, transcripts were labeled with fluorescent probes and counted in the nucleus (white) and soma (yellow). Bottom: Probe counts in the nucleus and soma across all cells with linear regression fits to estimate nuclear transcript proportions for each gene. Estimated proportions based on mFISH and RNA-seq data are summarized on the right.

564 Supplemental Figures

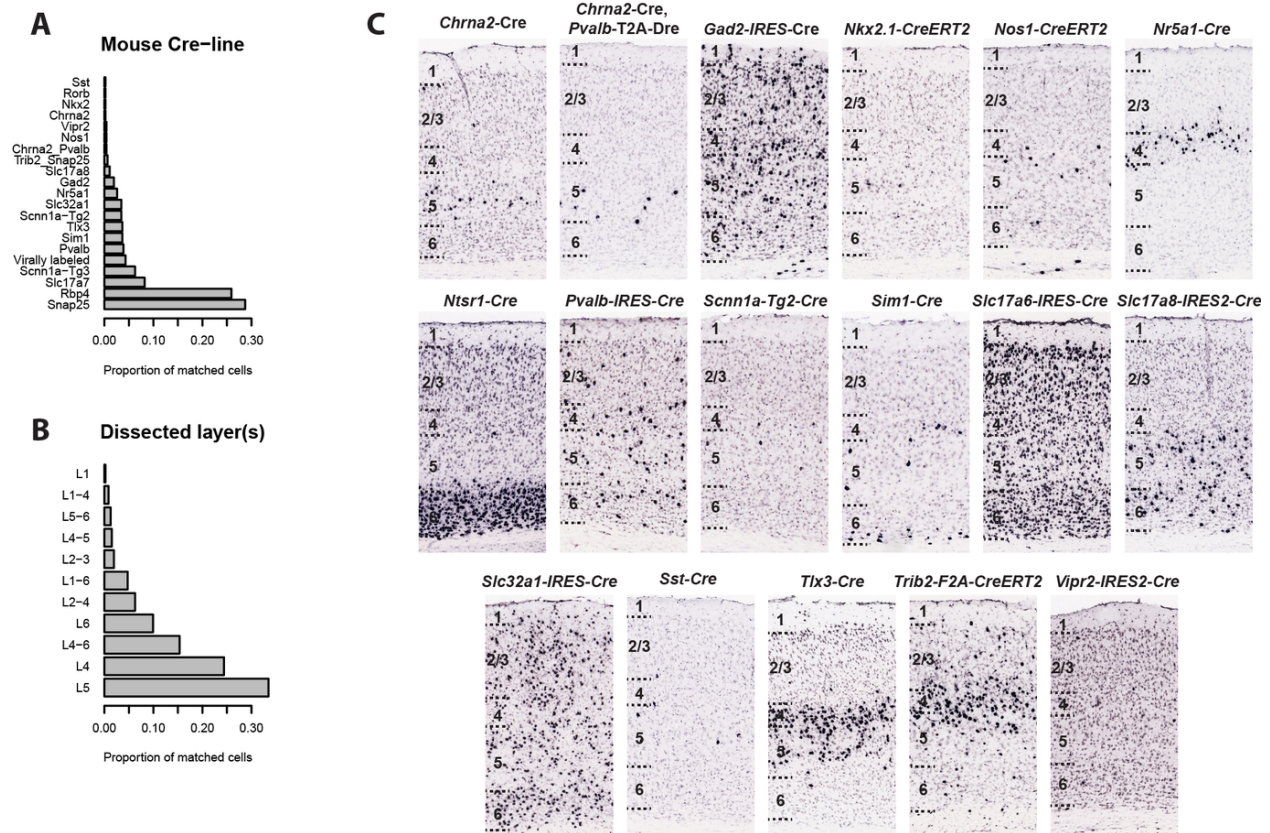


Figure S1: [Figure 1 - supplemental] Properties of 463 cells matched to nuclei. **(A)** Proportion of matched cells isolated from transgenic mouse lines that label different subsets of cortical neurons. Note that a small number of “virally labeled” cells (<5%) were FAC sorted from wild-type mice based on retrograde labeling by viral injections into various cortical and subcortical structures. **(B)** Proportion of matched cells dissected from one or more adjacent layers of cortex. **(C)** ISH images from additional mouse Cre-lines from which the best matching cells were most commonly derived. ISH images show all cortical layers within VISp.

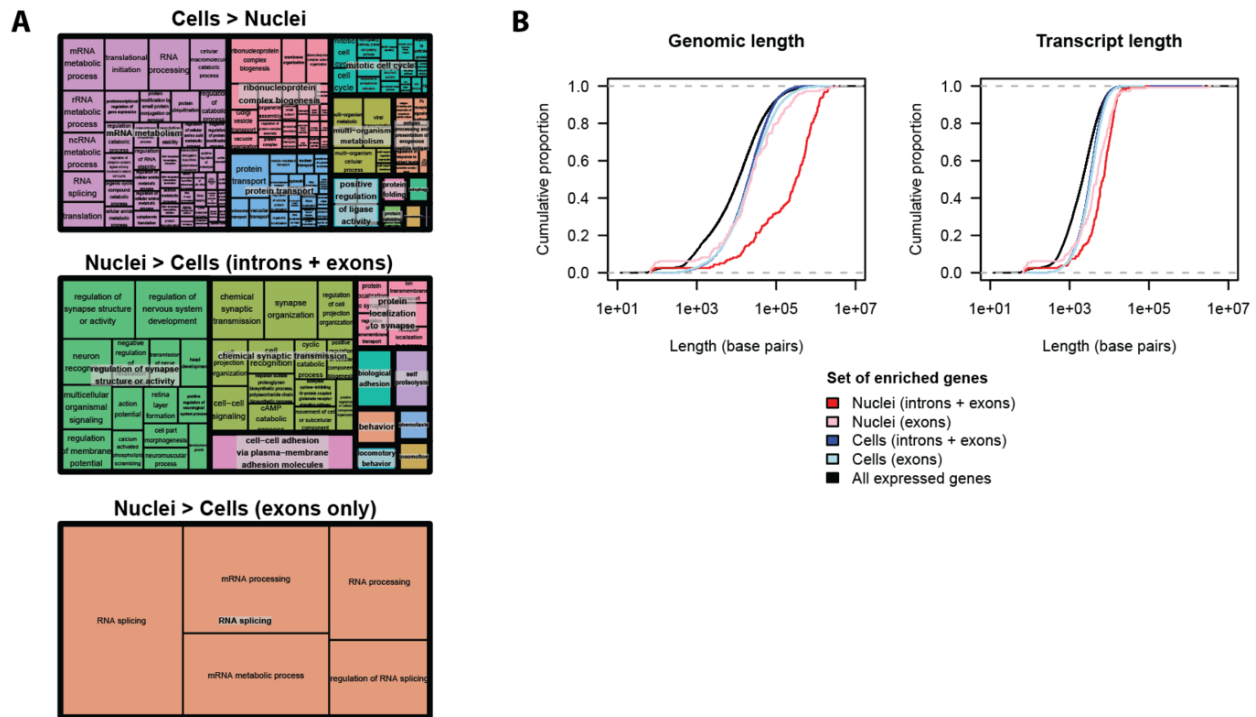


Figure S2: [Figure 2 - supplemental] Nuclear enrichment of transcripts related to neuron function can be explained by nuclear intron retention of long genes. **(A)** REVIGO (Supek et al., 2011a) summaries of gene ontology (GO) enrichment of genes enriched in cells or nuclei. Including introns dramatically changes the functional categories of nuclear but not cell enriched genes. **(B)** Cumulative distribution of genomic and transcript lengths for genes enriched in nuclei and cells (fold change > 1.5) based on expression of exons or introns plus exons. Using introns plus exons, the median genomic length of nuclear enriched genes is 16-fold longer than cell enriched genes. Using exons only, there is no significant difference in genomic lengths (Kolmogorov-Smirnov test P-value = 0.27).

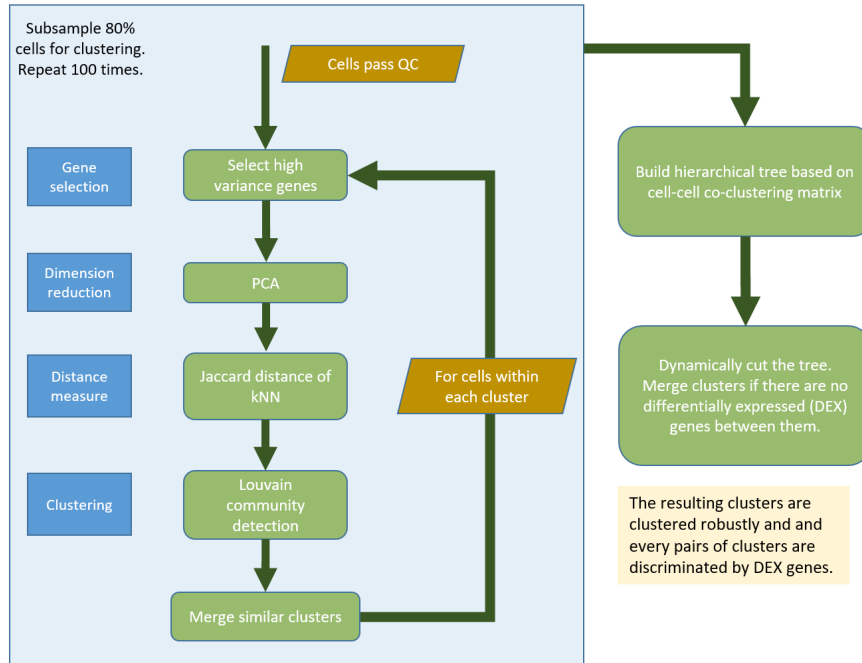


Figure S3: [Figure 3 - supplemental] Overview of single nucleus RNA-seq clustering pipeline. See methods for a detailed description of clustering steps.



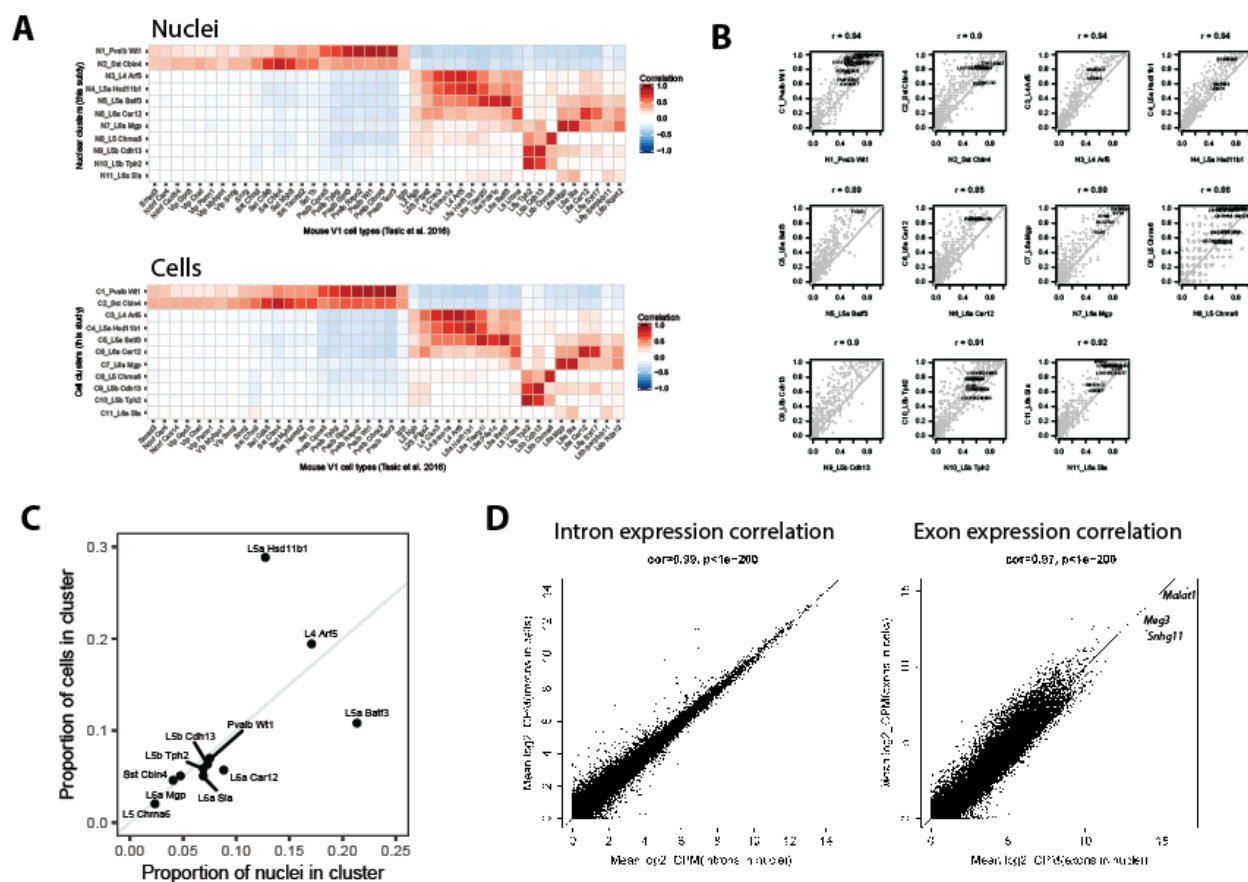


Figure S4: [Figure 4 - supplemental] Nuclear and cell clusters are well matched based on marker gene expression. **(A)** Pairwise correlations between previously reported mouse VISp cell type clusters (Tasic et al., 2016) and nuclear and cell clusters using average cluster expression of the top shared marker genes. Heatmaps show remarkably similar correlation patterns, supporting the existence of a well matched set of nuclear and cell clusters. Nuclear and cell clusters were annotated based on the reciprocal best matching published cluster name and mapped to two interneuron types and five of eight layer 5 excitatory neuron types. **(B)** Comparisons of the proportion of nuclei or cells expressing marker genes (CPM > 1) for matched pairs of clusters. Correlations are reported at the top of each scatter plot, and cell type specific markers are labeled. As expected based on Figure 2C, gene detection is consistently higher in cells than nuclei. **(C)** Matched clusters have similar proportions of nuclei and cells (except for two closely related cell types, L5a Hsd11b1 and L5a Batf3), which supports the accuracy of the initial correlation based mapping of single nuclei to cells. **(D)** Average gene expression quantified based on intronic reads is more highly correlated between cells and nuclei than expression quantified based on exonic reads, particularly for highly expressed genes. *Malat1*, *Meg3*, and *Snhg11* are the three highest expressing genes in nuclei and have consistently lower expression in cells, as expected based on their reported nuclear localization.

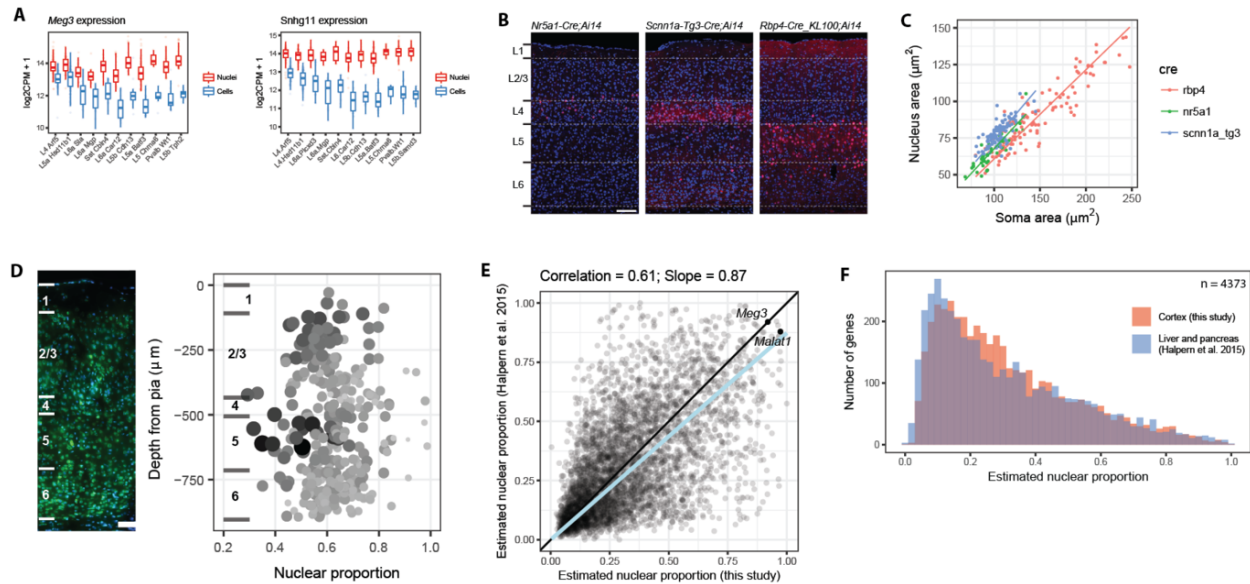


Figure S5: [Figure 5 - supplemental] Nuclear proportion estimates are supported by multiple genes and consistent with previously reported values. **(A)** Box plots of  $\log_2$ -transformed expression of two nuclear transcripts, *Meg3* and the small nucleolar RNA *Snhg11*, in matched nuclear and cell clusters. **(B)** Representative sections of VISp from three Cre-driver mouse lines with layer boundaries, nuclei labeled with DAPI (blue), and subsets of neurons labeled with tdTomato (red). Scale bar is  $100 \mu\text{m}$ . **(C)** Nucleus and soma area measurements from three Cre-lines, and linear regressions to estimate nuclear proportions. **(D)** Left: Section of VISp from wild type mouse labeled with DAPI and Neurotrace 500 fluorescent Nissl stain with layer boundaries indicated by white lines. Scale bar is  $100 \mu\text{m}$ . Right: Nuclear proportion was quantified based on nucleus and soma area measurements and plotted as a function of cortical depth. Size and darkness of points are proportional to soma area. **(E)** Average nuclear proportions of 4,373 genes (mostly house-keeping) also expressed in mouse pancreatic beta-cells and liver cells (Halpern et al., 2015) are moderately correlated with and approximately 13% less than estimated proportions in this study. **(F)** The distributions of nuclear proportions are highly similar with slightly higher reported cytoplasmic enrichment for reported genes. Note that the matched set of genes includes 99% protein-coding genes so the distributions more closely resemble those genes in Figure 5D.

## 565 **Supplemental Tables**

566 Table S1 [Figure 2 - supplemental]. Average gene expression and detection in matched nuclei and cells.

567 Table S2 [Figure 2 - supplemental]. Differentially expressed genes in cells versus nuclei using intronic plus  
568 exonic reads.

569 Table S3 [Figure 2 - supplemental]. Differentially expressed genes in cells versus nuclei using only exonic  
570 reads.

571 Table S4 [Figure 2 - supplemental]. Gene ontology (GO) enrichment of differentially expressed genes in cells  
572 and nuclei.

573 Table S5 [Figure 4 - supplemental]. Cre-driver line composition of cell clusters.

574 Table S6 [Figure 5 - supplemental]. Gene properties including the number of clusters with any expression,  
575 maximum cluster expression, cell type marker score, and estimated nuclear proportion of transcripts.

## 576 References

- 577 SCDE by Kharchenko Lab at Harvard DBMI. <http://hms-dbmi.github.io/scde/diffexp.html>. URL <http://hms-dbmi.github.io/scde/diffexp.html>. Accessed on Tue, October 24, 2017.
- 578
- 579 Erik Aronesty. ea-utils : Command-line tools for processing biological sequencing data; <https://github.com/ExpressionAnalysis/ea-utils>. 2011.
- 580
- 581 SC Baker, SR Bauer, RP Beyer, JD Brenton, B Bromley, J Burrill, H Causton, MP Conley, R Elespuru, M Fero, C Foy, J Fuscoe, X Gao, DL Gerhold, P Gilles, F Goodsaid, X Guo, J Hackett, RD Hockett, P Ikonomi, RA Irizarry, ES Kawasaki, T Kaysser-Kranich, K Kerr, G Kiser, WH Koch, KY Lee, C Liu, ZL Liu, A Lucas, CF Manohar, G Miyada, Z Modrusan, H Parkes, RK Puri, L Reid, TB Ryder, M Salit, RR Samaha, U Scherf, TJ Sendera, RA Setterquist, L Shi, R Shippy, JV Soriano, EA Wagar, JA Warrington, M Williams, F Wilmer, M Wilson, PK Wolber, X Wu, and R Zadro. The External RNA Controls Consortium: a progress report. *Nat Methods*, 2:731–4, Oct 2005.
- 582
- 583
- 584
- 585
- 586
- 587
- 588 Amy Bernard, Staci A Sorensen, and Ed S Lein. Shifting the paradigm: new approaches for characterizing and classifying neurons. *Current Opinion in Neurobiology*, 19(5):530–536, oct 2009. doi: 10.1016/j.conb.2009.09.010. URL <https://doi.org/10.1016%2Fj.conb.2009.09.010>.
- 589
- 590
- 591 JN Campbell, EZ Macosko, H Fenselau, TH Pers, A Lyubetskaya, D Tenen, M Goldman, AM Verstegen, JM Resch, SA McCarroll, ED Rosen, BB Lowell, and LT Tsai. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat Neurosci*, 20:484–496, Mar 2017.
- 592
- 593
- 594 J. M. Chambers, A. Freeny, and R. M. Heiberger. *Analysis of variance; designed experiments*. Wadsworth & Brooks/Cole, 1992.
- 595
- 596 J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(Web Server):W305–W311, may 2009. doi: 10.1093/nar/gkp427. URL <https://doi.org/10.1093%2Fnar%2Fgkp427>.
- 597
- 598
- 599 S Darmanis, SA Sloan, Y Zhang, M Enge, C Caneda, LM Shuer, Gephart MG Hayden, BA Barres, and SR Quake. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A*, 112:7285–90, Jun 2015.
- 600
- 601
- 602 S Djebali, CA Davis, A Merkel, A Dobin, T Lassmann, A Mortazavi, A Tanzer, J Lagarde, W Lin, F Schlesinger, C Xue, GK Marinov, J Khatun, BA Williams, C Zaleski, J Rozowsky, M Röder, F Kokocinski, RF Abdelhamid, T Alioto, I Antoshechkin, MT Baer, NS Bar, P Batut, K Bell, I Bell, S Chakraborty, X Chen, J Chrast, J Curado, T Derrien, J Drenkow, E Dumais, J Dumais, R Duttgupta, E Falconnet, M Fastuca, K Fejes-Toth, P Ferreira, S Foissac, MJ Fullwood, H Gao, D Gonzalez, A Gordon, H Gunawardena, C Howald, S Jha, R Johnson, P Kapranov, B King, C Kingswood, OJ Luo, E Park, K Persaud, JB Preall, P Ribeca, B Risk, D Robyr, M Sammeth, L Schaffer, LH See, A Shahab, J Skancke, AM Suzuki, H Takahashi, H Tilgner, D Trout, N Walters, H Wang, J Wrobel, Y Yu, X Ruan, Y Hayashizaki, J Harrow, M Gerstein, T Hubbard, A Reymond, SE Antonarakis, G Hannon, MC Giddings, Y Ruan, B Wold, P Carninci, R Guigó, and TR Gingeras. Landscape of transcription in human cells. *Nature*, 489:101–8, Sep 2012.
- 603
- 604
- 605
- 606
- 607
- 608
- 609
- 610
- 611
- 612
- 613 A Dobin, CA Davis, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and TR Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29:15–21, Jan 2013.
- 614
- 615 JR Ecker, DH Geschwind, AR Kriegstein, J Ngai, P Osten, D Polioudakis, A Regev, N Sestan, IR Wickersham, and H Zeng. The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. *Neuron*, 96:542–557, Nov 2017.
- 616
- 617
- 618 Santo Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, Jan 2007. ISSN 0027-8424. doi: 10.1073/pnas.0605965104. URL <http://www.pnas.org/cgi/doi/10.1073/pnas.0605965104>.
- 619
- 620

- 621 HW Gabel, B Kinde, H Stroud, CS Gilbert, DA Harmin, NR Kastan, M Hemberg, DH Ebert, and ME Green-  
622 berg. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, 522:  
623 89–93, Jun 2015.
- 624 N Habib, Y Li, M Heidenreich, L Swiech, I Avraham-Davidi, JJ Trombetta, C Hession, F Zhang, and  
625 A Regev. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*,  
626 353:925–8, Aug 2016.
- 627 Keren B Halpern, I Caspi, D Lemze, M Levy, S Landen, E Elinav, I Ulitsky, and S Itzkovitz. Nuclear  
628 Retention of mRNA in Mammalian Tissues. *Cell Rep*, 13:2653–62, Dec 2015.
- 629 Donald A. Jackson. Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and  
630 Statistical Approaches. *Ecology*, 74(8):2204–2214, Dec 1993. ISSN 00129658. doi: 10.2307/1939574. URL  
631 <http://doi.wiley.com/10.2307/1939574>.
- 632 PV Kharchenko, L Silberstein, and DT Scadden. Bayesian approach to single-cell differential expression  
633 analysis. *Nat Methods*, 11:740–2, Jul 2014.
- 634 Suguna Rani Krishnaswami, Rashel V Grindberg, Mark Novotny, Pratap Venepally, Benjamin Lacar, Ku-  
635 nal Bhutani, Sara B Linker, Son Pham, Jennifer A Erwin, Jeremy A Miller, Rebecca Hodge, James K  
636 McCarthy, Martin Kelder, Jamison McCarrison, Brian D Aevermann, Francisco Diez Fuertes, Richard H  
637 Scheuermann, Jun Lee, Ed S Lein, Nicholas Schork, Michael J McConnell, Fred H Gage, and Roger S  
638 Lasken. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nature Pro-*  
639 *tocols*, 11(3):499–524, feb 2016. doi: 10.1038/nprot.2016.015. URL [https://doi.org/10.1038%2Fnpot.](https://doi.org/10.1038%2Fnpot.2016.015)  
640 [2016.015](https://doi.org/10.1038%2Fnpot.2016.015).
- 641 Benjamin Lacar, Sara B. Linker, Baptiste N. Jaeger, Suguna Krishnaswami, Jerika Barron, Martijn Kelder,  
642 Sarah Parylak, Apuã Paquola, Pratap Venepally, Mark Novotny, Carolyn O'Connor, Conor Fitzpatrick,  
643 Jennifer Erwin, Jonathan Y. Hsu, David Husband, Michael J. McConnell, Roger Lasken, and Fred H. Gage.  
644 Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nature Communications*, 7:  
645 11022, apr 2016. doi: 10.1038/ncomms11022. URL <https://doi.org/10.1038%2Fncoms11022>.
- 646 BB Lake, R Ai, GE Kaeser, NS Salathia, YC Yung, R Liu, A Wildberg, D Gao, HL Fung, S Chen, R Vi-  
647 jayaraghavan, J Wong, A Chen, X Sheng, F Kaper, R Shen, M Ronaghi, JB Fan, W Wang, J Chun, and  
648 K Zhang. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain.  
649 *Science*, 352:1586–90, Jun 2016.
- 650 BB Lake, S Chen, BC Sos, J Fan, GE Kaeser, YC Yung, TE Duong, D Gao, J Chun, PV Kharchenko, and  
651 K Zhang. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain.  
652 *Nat Biotechnol*, Dec 2017a.
- 653 BB Lake, S Codeluppi, YC Yung, D Gao, J Chun, PV Kharchenko, S Linnarsson, and K Zhang. A com-  
654 parative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type  
655 expression from nuclear RNA. *Sci Rep*, 7:6031, Jul 2017b.
- 656 MR Lamprecht, DM Sabatini, and AE Carpenter. CellProfiler: free, versatile software for automated bio-  
657 logical image analysis. *Biotechniques*, 42:71–5, Jan 2007.
- 658 Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the  
659 Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720, 2007.
- 660 M Lawrence, W Huber, H Pagès, P Aboyoun, M Carlson, R Gentleman, MT Morgan, and VJ Carey. Software  
661 for computing and annotating genomic ranges. *PLoS Comput Biol*, 9:e1003118, 2013.
- 662 ES Lein, MJ Hawrylycz, N Ao, M Ayres, A Bensinger, A Bernard, AF Boe, MS Boguski, KS Brockway,  
663 EJ Byrnes, L Chen, L Chen, TM Chen, MC Chin, J Chong, BE Crook, A Czaplinska, CN Dang, S Datta,  
664 NR Dee, AL Desaki, T Desta, E Diep, TA Dolbeare, MJ Donelan, HW Dong, JG Dougherty, BJ Duncan,  
665 AJ Ebbert, G Eichele, LK Estin, C Faber, BA Facer, R Fields, SR Fischer, TP Fliiss, C Frensley, SN Gates,



- 666 KJ Glattfelder, KR Halverson, MR Hart, JG Hohmann, MP Howell, DP Jeung, RA Johnson, PT Karr,  
667 R Kawal, JM Kidney, RH Knapik, CL Kuan, JH Lake, AR Laramée, KD Larsen, C Lau, TA Lemon,  
668 AJ Liang, Y Liu, LT Luong, J Michaels, JJ Morgan, RJ Morgan, MT Mortrud, NF Mosqueda, LL Ng,  
669 R Ng, GJ Orta, CC Overly, TH Pak, SE Parry, SD Pathak, OC Pearson, RB Puchalski, ZL Riley, HR Rock-  
670 ett, SA Rowland, JJ Royall, MJ Ruiz, NR Sarno, K Schaffnit, NV Shapovalova, T Sivisay, CR Slaughter-  
671 beck, SC Smith, KA Smith, BI Smith, AJ Sodt, NN Stewart, KR Stumpf, SM Sunkin, M Sutram, A Tam,  
672 CD Teemer, C Thaller, CL Thompson, LR Varnam, A Visel, RM Whitlock, PE Wohnoutka, CK Wolkey,  
673 VY Wong, M Wood, MB Yaylaoglu, RC Young, BL Youngstrom, XF Yuan, B Zhang, TA Zwingman, and  
674 AR Jones. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445:168–76, Jan 2007.
- 675 JH Levine, EF Simonds, SC Bendall, KL Davis, el-AD Amir, MD Tadmor, O Litvin, HG Fienberg, A Jager,  
676 ER Zunder, R Finck, AL Gedman, I Radtke, JR Downing, D Pe'er, and GP Nolan. Data-Driven Phenotypic  
677 Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162:184–97, Jul 2015.
- 678 N Lin, KY Chang, Z Li, K Gates, ZA Rana, J Dang, D Zhang, T Han, CS Yang, TJ Cunningham, SR Head,  
679 G Duyster, PD Dong, and TM Rana. An evolutionarily conserved long noncoding RNA TUNA controls  
680 pluripotency and neural lineage commitment. *Mol Cell*, 53:1005–19, Mar 2014.
- 681 EZ Macosko, A Basu, R Satija, J Nemes, K Shekhar, M Goldman, I Tirosh, AR Bialas, N Kamitaki,  
682 EM Martersteck, JJ Trombetta, DA Weitz, JR Sanes, AK Shalek, A Regev, and SA McCarroll. Highly  
683 Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161:1202–  
684 1214, May 2015.
- 685 O Mauger, F Lemoine, and P Scheiffele. Targeted Intron Retention and Excision for Rapid Gene Regulation  
686 in Response to Neuronal Activity. *Neuron*, 92:1266–1278, Dec 2016.
- 687 George Paxinos et al. *Paxinos and Franklin's the mouse brain in stereotaxic coordinates*. Academic Press,  
688 2013.
- 689 Jean-Francois Poulin, Bosiljka Tasic, Jens Hjerling-Leffler, Jeffrey M Trimarchi, and Rajeshwar Awatramani.  
690 Disentangling neural cell diversity using single-cell transcriptomics. *Nature Neuroscience*, 19(9):1131–1141,  
691 aug 2016. doi: 10.1038/nn.4366. URL <https://doi.org/10.1038%2Fnn.4366>.
- 692 Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bod-  
693 denmiller, Peter J Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dun-  
694 ham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens,  
695 Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung K Kim, Paul Klenerman, Arnold Kriegstein,  
696 Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundberg, Partha Majumder, John C Marioni,  
697 Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Philli-  
698 pakis, Chris P Ponting, Stephen R Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua R Sanes, Rahul  
699 Satija, Ton N Schumacher, Alex K Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle,  
700 Michael R Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oude-  
701 naarden, Allon Wagner, Fiona M Watt, Jonathan S Weissman, Barbara J Wold, Ramnik J Xavier, and  
702 Nir Yosef and. Science Forum: The Human Cell Atlas. *eLife*, 6, dec 2017. doi: 10.7554/elife.27041. URL  
703 <https://doi.org/10.7554%2Felifelife.27041>.
- 704 Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74  
705 (1):016110, 2006.
- 706 D Risso, J Ngai, TP Speed, and S Dudoit. Normalization of RNA-seq data using factor analysis of control  
707 genes or samples. *Nat Biotechnol*, 32:896–902, Sep 2014.
- 708 Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth.  
709 limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids  
710 research*, 43(7):e47–e47, 2015.

- 711 K Shekhar, SW Lapan, IE Whitney, NM Tran, EZ Macosko, M Kowalczyk, X Adiconis, JZ Levin, J Nemesl,   
712 M Goldman, SA McCarroll, CL Cepko, A Regev, and JR Sanes. Comprehensive Classification of Retinal   
713 Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166:1308–1323.e30, Aug 2016.
- 714 F Supek, M Bošnjak, N Škunca, and T Šmuc. REVIGO summarizes and visualizes long lists of gene ontology   
715 terms. *PLoS One*, 6:e21800, 2011a.
- 716 Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. REVIGO Summarizes and Visualizes Long   
717 Lists of Gene Ontology Terms. *PLoS ONE*, 6(7):e21800, jul 2011b. doi: 10.1371/journal.pone.0021800.   
718 URL <https://doi.org/10.1371%2Fjournal.pone.0021800>.
- 719 B Tasic, V Menon, TN Nguyen, TK Kim, T Jarsky, Z Yao, B Levi, LT Gray, SA Sorensen, T Dolbeare,   
720 D Bertagnolli, J Goldy, N Shapovalova, S Parry, C Lee, K Smith, A Bernard, L Madisen, SM Sunkin,   
721 M Hawrylycz, C Koch, and H Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcrip-   
722 tomics. *Nat Neurosci*, 19:335–46, Feb 2016.
- 723 Bosiljka Tasic, Zizhen Yao, Kimberly A Smith, Lucas T Graybuck, and Thuc Nghi Nguyen. Shared and   
724 distinct transcriptomic cell types across neocortical areas. *bioRxiv*, 2017. doi: dx.doi.org/10.1101/229542.
- 725 Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing Company, Incorporated,   
726 2nd edition, 2009. ISBN 0387981403, 9780387981406.
- 727 A Zeisel, AB Muñoz-Manchado, S Codeluppi, P Lönnerberg, Manno G La, A Juréus, S Marques, H Munguba,   
728 L He, C Betsholtz, C Rolny, G Castelo-Branco, J Hjerling-Leffler, and S Linnarsson. Cell types in the   
729 mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347:1138–42, Mar 2015.
- 730 Hongkui Zeng and Joshua R. Sanes. Neuronal cell-type classification: challenges opportunities and the   
731 path forward. *Nature Reviews Neuroscience*, 18(9):530–546, aug 2017. doi: 10.1038/nrn.2017.85. URL   
732 <https://doi.org/10.1038%2Fnrn.2017.85>.