Open access • Posted Content • DOI:10.1101/239749

Equivalent high-resolution identification of neuronal cell types with single-nucleus and single-cell RNA-sequencing — Source link

Trygve E. Bakken, Rebecca D. Hodge, Jeremy A. Miller, Zizhen Yao ...+22 more authors

Institutions: Allen Institute for Brain Science, J. Craig Venter Institute

Published on: 25 Dec 2017 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Nucleus, RNA and Cell type

PLOS ONE

RESEARCH ARTICLE

Single-nucleus and single-cell transcriptomes compared in matched cortical cell types

Trygve E. Bakken[1], Rebecca D. Hodge[1], Jeremy A. Miller[1], Zizhen Yao[1], Thuc Nghi Nguyen[1], Brian Aevermann[2], Eliza Barkan[1], Darren Bertagnolli[1], Tamara Casper[1], Nick Dee[1], Emma Garren[1], Jeff Goldy[1], Lucas T. Graybuck[1], Matthew Kroll[1], Roger S. Lasken[2], Kanan Lathia[1], Sheana Parry[1], Christine Rimorin[1], Richard H. Scheuermann[2], Nicholas J. Schork[2], Soraya I. Shehata[1], Michael Tieu[1], John W. Phillips[1], Amy Bernard[1], Kimberly A. Smith[1], Hongkui Zeng[1], Ed S. Lein[1], Bosiljka Tasic[1]*

1 Allen Institute for Brain Science, Seattle, WA, United States of America, **2** J. Craig Venter Institute, La Jolla, CA, United States of America

* bosiljkat@alleninstitute.org

OPEN ACCESS

Citation: Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, et al. (2018) Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS ONE 13(12): e0209648. https://doi.org/10.1371/journal.pone.0209648

Editor: Eduardo Soriano, University of Barcelona, Parc Cientific de Barcelona and CIBERNED (ISCIII)., SPAIN

Received: February 23, 2018

Accepted: December 10, 2018

Published: December 26, 2018

Data Availability Statement: Data and code to reproduce figures are publicly available from GitHub at https://github.com/AllenInstitute/NucCellTypes. Single-cell and single-nucleus transcriptomic data are available at the NCBI Gene Expression Omnibus (GEO) under accession number GSE123454.

Funding: The authors received no specific funding for this work.

Abstract

Transcriptomic profiling of complex tissues by single-nucleus RNA-sequencing (snRNA-seq) affords some advantages over single-cell RNA-sequencing (scRNA-seq). snRNA-seq provides less biased cellular coverage, does not appear to suffer cell isolation-based transcriptional artifacts, and can be applied to archived frozen specimens. We used well-matched snRNA-seq and scRNA-seq datasets from mouse visual cortex to compare cell type detection. Although more transcripts are detected in individual whole cells (~11,000 genes) than nuclei (~7,000 genes), we demonstrate that closely related neuronal cell types can be similarly discriminated with both methods if intronic sequences are included in snRNA-seq analysis. We estimate that the nuclear proportion of total cellular mRNA varies from 20% to over 50% for large and small pyramidal neurons, respectively. Together, these results illustrate the high information content of nuclear RNA for characterization of cellular diversity in brain tissues.

Introduction

Cell types in mammalian brain have been defined based on various properties including their morphology, electrophysiology, and gene expression [1–3]. scRNA-seq has emerged as a high-throughput method for quantification of the majority of transcripts in thousands of cells [4]. Similarities and differences in gene expression at the single cell level characterized by scRNA-seq have revealed diverse cell types in many mouse brain regions, including neocortex [5–7], hypothalamus [8], and retina [9,10].

However, scRNA-seq profiling does not provide an unbiased survey of neural cell types. Some cell types are more vulnerable to the tissue dissociation process and are underrepresented in the final data set. For example, in mouse neocortex, fast-spiking parvalbumin-positive interneurons and subcortically projecting glutamatergic neurons in layer 5 are observed in

Competing interests: The authors have declared that no competing interests exist.

lower proportions than expected and need to be selectively enriched using Cre-driver lines to achieve sufficient sampling [6]. In adult human neocortex, non-neuronal cells survive dissociation better than neurons and are over-represented in single cell suspensions [11]. In contrast to whole cells, nuclei are more resistant to mechanical assaults and can be isolated from frozen tissue [12,13]. Individual nuclei have been shown to provide sufficient gene expression information to define relatively broad cell classes in adult human brain [14,15] and mouse hippocampus [16].

However, previous studies have not investigated if individual nuclei contain sufficient diversity and number of transcripts to enable discrimination of closely related cell types at a resolution comparable to whole cells. A recent study compared clustering results for single nuclei and whole cells isolated from the mouse somatosensory cortex [17], but it only showed similar ability to distinguish two highly distinct cell classes: superficial- and deep-layer excitatory neurons. Additionally, a recent study using droplet-based high throughput sequencing of human nuclei showed that broad cell classes in human brain could be successfully mapped to corresponding broad classes in mouse [18].

In this study, we investigated differences in mRNA composition and information content between nuclei and whole cells, and the ability to detect cell types by high-depth RNA-sequencing of single cells and single nuclei. For this purpose, we focused on well-matched sets of cells and nuclei: 463 nuclei and 463 whole cells from layer 5 of adult mouse primary visual cortex (VISp). We selected VISp because it contains a known variety of distinguishable yet highly similar cell types [5] that would reveal the cell type detection limit of RNA-seq data obtained from single cells or nuclei. The cells and nuclei were processed by the same experimental and computational methods. We find that although the nuclear content and proportion of mRNA vary among cell types, nuclei contain enough informative transcripts to identify highly related neuronal cell types with resolution similar to whole cells.

Results

RNA-seq profiling of single nuclei and single cells

We isolated 487 NeuN-positive single nuclei from layer 5 of mouse VISp using fluorescence activated cell sorting (FACS). Anti-NeuN staining was performed to enrich for neurons. In parallel, we isolated 12,866 tdT-positive single cells by FACS from all layers of mouse VISp and a variety of Cre-driver lines. Whole cells were collected as part of a larger study on cortical cell type diversity, which contains a complete description of all Cre-driver lines used for cell collection [6]. For both single nuclei and cells, poly(A)-transcripts were reverse transcribed and amplified with SMART-Seq v4, cDNA was tagmented by Nextera XT, and resulting libraries were sequenced to an average depth of 2.5 million reads (Fig 1A). RNA-seq reads were mapped to the mouse genome using the STAR aligner [19]. Gene expression was quantified as the sum of intronic and exonic reads per gene and was normalized as counts per million (CPM) and log2-transformed. For each nucleus and cell, the probabilities of gene detection dropouts were estimated as a function of average expression level based on empirical noise models [20].

463 out of 487 single nuclei (95%) passed quality control metrics. Each nucleus was matched to the most similar nucleus and cell based on the maximum correlated expression of all genes, weighted for gene dropouts based on noise models estimated for each nucleus and cell. Nuclei had similarly high pairwise correlations to cells as to other nuclei, suggesting that cells and nuclei were well matched (Fig 1B). As expected, matched cells were derived almost exclusively from layer 5 and adjacent layers 4 and 6, and from Cre-driver lines that labeled

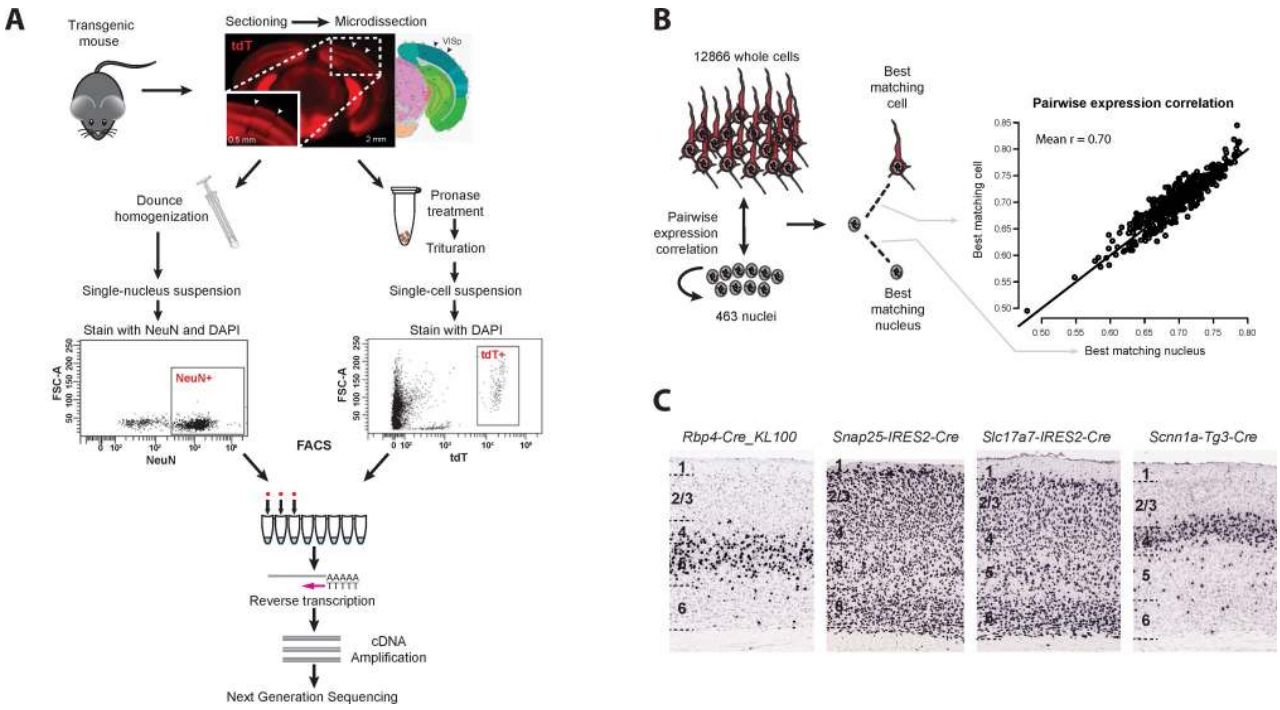

Fig 1. Identification of an expression-matched set of single nuclei and whole cells from mouse primary visual cortex (VISp). (A) Whole brains were dissected from transgenic mice, sectioned into coronal slices, and individual layers of VISp were microdissected. Nuclei were dissociated from layer 5, stained 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-positive cells were isolated from multiple different Cre-driver lines. Single nucleus and cell mRNA were reverse-transcribed, amplified, and sequenced to measure genome-wide gene expression levels. **(B)** Left: 463 nuclei from layer 5 and 12,866 whole cells from all layers, which passed quality control metrics were used to determine expression correlation between each nucleus and every other nucleus and cell. 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 average correlation equal to 0.70 and 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) for tdTomato mRNA in VISp of transgenic mice (Cre-lines crossed to Ai14 Cre reporter [21]). Shown are the tissue sections from 4 Cre-driver lines from which the majority of the best-matching cells to L5 nuclei were derived. As expected, all Cre-lines label cells in layer 5 and adjacent layers.

https://doi.org/10.1371/journal.pone.0209648.g001

cells in layer 5 (Fig 1C and S1 Fig). The small minority of matched cells isolated from superficial layers were GABAergic interneurons that have been detected in many layers [6].

Comparison of nuclear and whole cell transcriptomes

scRNA-seq profiles nuclear and cytoplasmic transcripts, whereas snRNA-seq profiles mostly nuclear transcripts (although some transcripts may be attached to the rough endoplasmic reticulum and partially retained in nuclear preps). Therefore, we expect that RNA-seq reads will differ between nuclei and cells. In nuclei, more than 50% of reads that aligned to the mouse genome did not map to known spliced transcripts but mapped within gene boundaries but outside exons. They were therefore annotated as intronic reads (Fig 2A). In contrast, more than half of cells had less than 30% intronic reads. A minority of cells had close to 50% intronic reads, similar to nuclei. Median gene detection based on exonic reads was lower for nuclei (~5,000 genes) than for cells (~9,500). Including both intronic and exonic reads increased gene detection for nuclei (~7,000) and cells (~11,000), demonstrating that intronic reads provided additional information not captured by exons. Gene detection was largely saturated using 2.5 million reads per sample and was consistently higher for cells than nuclei at lower read depths (S2A Fig). Whole-brain control RNA displayed higher read mapping to exons and

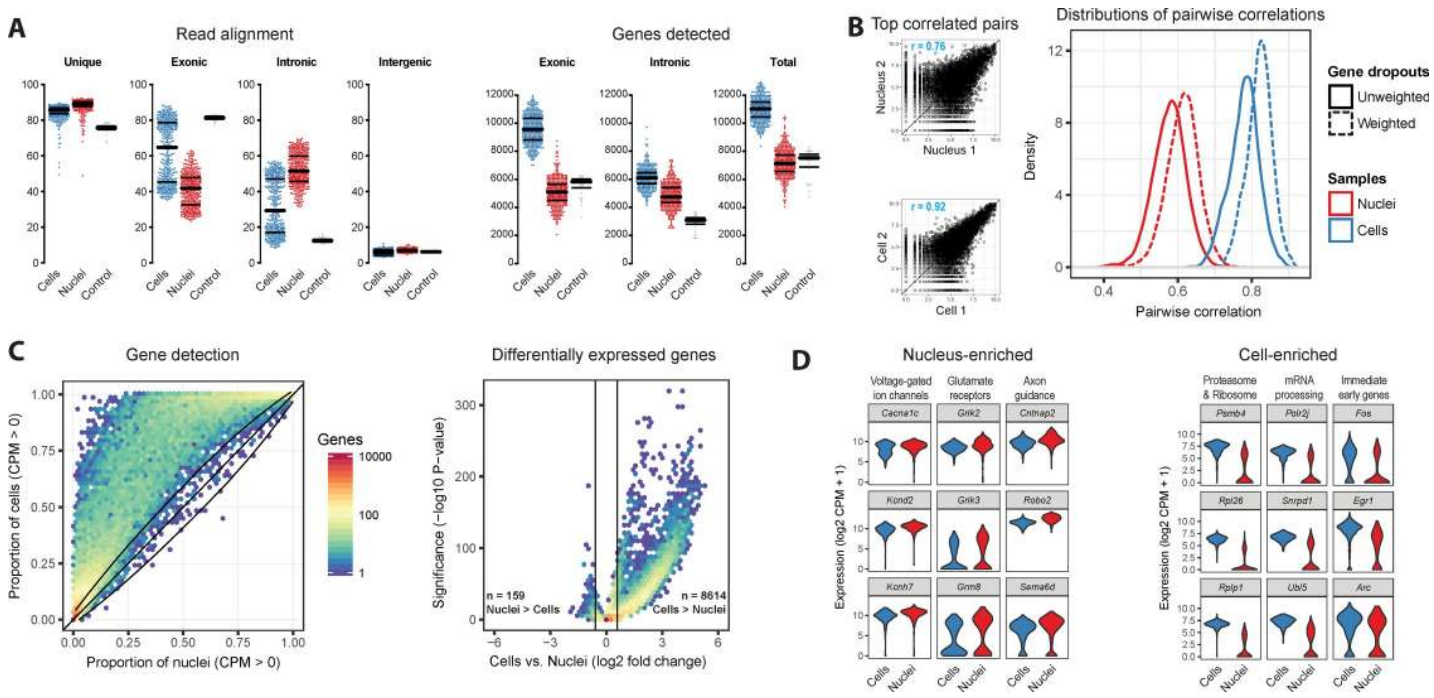

Fig 2. Comparison of nuclear and whole cell transcriptomes. (A) Left: Percentage of RNA-seq reads mapping to genomic regions for cells, nuclei, and whole-brain control RNA. Bars indicate median and 25th and 75th quantiles. Among cells, exonic and intronic read percentages display bimodal distributions. Right: Gene detection (counts per million, CPM > 0) based on read 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 in cells. Right: Cells have consistently more similar expression to each other than nuclei, even after correcting for gene dropouts based on expression noise models. **(C)** Left: Binned scatter plot showing all genes are detected (CPM > 0) with equal or greater reliability in cells than in nuclei. Black 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 in nuclei, and 20.5% of genes are more highly expressed in cells (for both comparisons, fold change > 1.5, adjusted P-value < 0.05). The log-transformed color scale indicates the number of genes in each bin. **(D)** Examples of nucleus-enriched transcripts involved in neuronal connectivity, synaptic transmission, and intrinsic firing properties and cell-enriched transcripts related to mRNA processing and protein translation and degradation. In addition, expression of immediate early genes is up to 10-fold higher in cells.

https://doi.org/10.1371/journal.pone.0209648.g002

lower mapping to introns as compared to cells, showing that non-nuclear transcripts composed a larger fraction of control RNA than cells (Fig 2A). Control RNA may include more non-nuclear transcripts because this RNA is isolated from bulk tissue and may be dominated by shared highly expressed cytoplasmic transcripts. Moreover, bulk RNA likely captures more transcripts in distal processes that are lost in dissociated single cells.

Gene expression levels measured in cells of the same type can vary due to biological factors, for example differences in cell state and stochastic transcription [22–24], and due to missed detection also known as "gene dropouts". Gene dropouts were higher in nuclei than in cells (S2B Fig), and expression correlations were higher and variability lower between pairs of cells than between pairs of nuclei (Fig 2B). To assess the contribution of gene dropouts to expression variability, empirical noise models were fit to nuclei and cells, and correlations were adjusted to account for dropouts. Correlations similarly increased for both cells and nuclei suggesting that biological effects were a major contributor to higher expression variability among nuclei, and this is consistent with nuclei acting as a transcriptional buffer to dampen gene expression noise in the cell [25].

A majority of expressed genes (21,279; 63%) showed similar detection (<10% difference) in nuclei and cells, whereas 7,217 genes (21%) were detected in at least 25% more cells than nuclei (Fig 2C and S1 Table). For nuclei, including intronic reads increased detection of 1334 genes by more than 25% compared to using exonic reads alone. Surprisingly, 83% of these genes

were protein-coding genes that were significantly enriched for neuronal functions, including synaptic transmission, axon projection, and cell adhesion (data not shown). These genes had low to medium expression in cells and were more likely to be markers of cell types than expressed genes overall (S2C Fig).

8,614 genes have significantly higher expression in cells than nuclei (>1.5 fold expression; FDR < 0.05), and many are involved in house-keeping functions such as mRNA processing and translation (S2D Fig). Genetic markers of neuronal activity, such as immediate early genes *Fos*, *Egr1*, and *Arc* also displayed up to 10-fold increased expression in cells, potentially a byproduct of tissue dissociation [13]. 159 genes displayed significantly higher expression in nuclei (Fig 2D and S2 Table), and they appear relevant to neuronal identity as they include connectivity and signaling genes (S2D Fig and S3 Table). Based on the sum of intronic and exonic reads, these 159 nucleus-enriched genes are on average more than 10-fold longer than cell-enriched genes (S2E Fig). A similar observation was recently reported for mRNAs enriched in single nuclei in mouse somatosensory cortex [17]. We did not normalize expression levels for gene length because effective lengths can be highly variable due to differential transcript processing among cells. Therefore, these 159 genes may appear nucleus-enriched because of many reads mapping to long intronic reads rather than an increase in the absolute number of transcripts in the nucleus. Indeed, when only exonic reads were used to quantify expression in nuclei and cells, a different set of 146 genes was significantly enriched in nuclei (S4 Table). These genes were only slightly longer than cell-enriched genes, so likely reflected a true nuclear enrichment of transcripts. They were not associated with neuron-specific functions, and were significantly enriched for genes that participate in pre-mRNA splicing. These differences in gene expression enrichment based on the inclusion of intronic reads highlighted the need to more directly estimate the nuclear fraction of transcripts, which we address in a later section.

Intronic reads are required for high-resolution cell type identification from snRNA-seq

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

Clustering includes two steps–selection of differentially expressed (DE) genes and distance measurement–that are particularly sensitive to expression quantification. We repeated clustering using intronic and exonic reads or only exonic reads for these steps, and ordered co-clustering matrices to match the results using all reads for both steps. When using introns and exons, we found 11 distinct clusters of nuclei and cells, and clusters had similar cohesion (average within cluster co-clustering) and separation (average co-clustering difference with the closest cluster) (Fig 3C). Including intronic reads for either clustering step increased the number of clusters detected for nuclei but not cells and was likely due to improved detection of cell type informative genes (S2C Fig). Therefore, accounting for intronic reads in snRNA-seq was critical to enable high-resolution cluster detection comparable to that observed with scRNA-seq.

Comparable cell types identified with nuclei and cells

We used hierarchical clustering of median gene expression values for each cluster to determine the relationships between clusters. This analysis revealed that cluster relationships represented

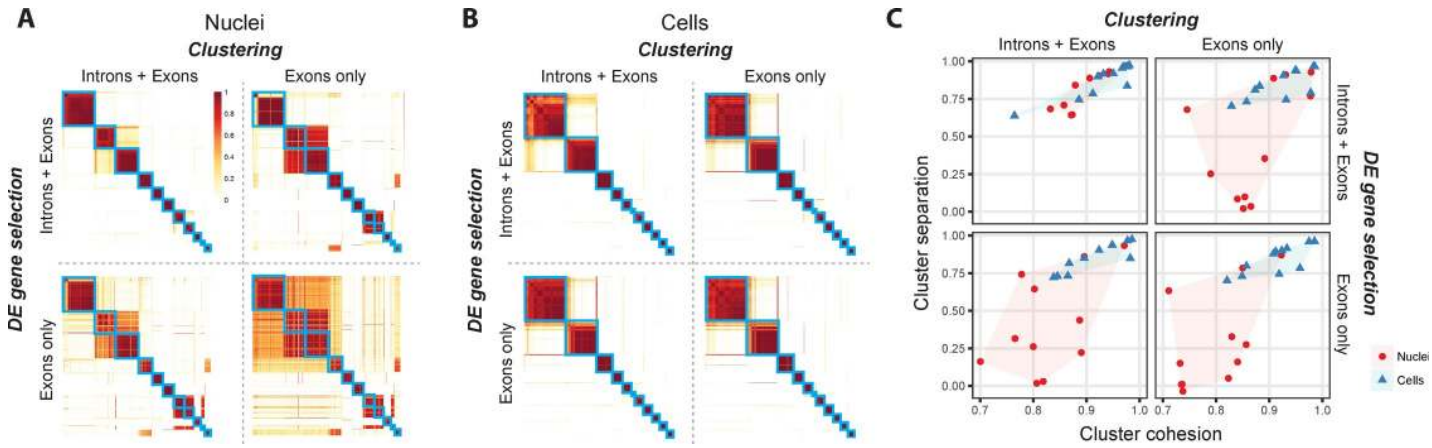

Fig 3. Single nuclei provide comparable clustering resolution to cells when intronic reads are included. (A) Co-clustering heatmaps show the proportion of 100 clustering iterations that each pair of nuclei were assigned to the same cluster. Clustering was performed using gene expression quantified with 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.

https://doi.org/10.1371/journal.pone.0209648.g003

as dendrograms are similar for nuclei and cells (Fig 4A). We compared the 11 clusters identified from both the single nucleus and single cell datasets to previously reported cell types in mouse VISp [5]. Based on highly correlated ($r > 0.85$) expression of hundreds of marker genes, each cluster corresponds to a reported cell type (S4A Fig). Conserved marker gene expression (Fig 4B and S4B Fig) confirmed that the same 11 cell types were identified with nuclei and cells (Fig 4C). These cell types included nine excitatory neuron types from layers 4–6 and two inhibitory interneuron types. Matched cluster proportions were mostly consistent, except that two closely related layer 5a subtypes were under- (L5a Batf3) or over-represented (L5a Hsd11b1) among cells (S4C Fig). This result demonstrated that the initial matching of cells to nuclei was relatively unbiased.

Since most cytoplasmic transcripts are spliced, intronic reads should be derived from nuclear transcripts. We hypothesized that we could estimate transcript abundance in nuclei based on intronic reads from whole-cell RNA-seq. Indeed, average expression levels of genes were highly correlated in cells and nuclei when using intronic but not exonic reads (S4D Fig). Furthermore, matching pairs of nucleus and cell clusters were nearest neighbors in a dendrogram based on the median expression (quantified using only intronic reads), except for two closely related layer 5b subtypes (Fig 4D). Therefore, intronic reads facilitate comparisons between data sets derived from snRNA-seq and scRNA-seq, although some expression differences remain. A dendrogram based on exonic reads grouped clusters first by sample type (nuclei and cells) and then by broad cell class (inhibitory and excitatory neurons). Grouping of samples by type was likely due to differences in cytoplasmic transcripts that were profiled in cells but not in nuclei.

While we detected comparable cell types using nuclei and cells, we expected that gene expression captured with cells likely included additional information from cytoplasmic transcripts. We compared the separation of matched pairs of clusters based on co-clustering and found that most nuclei and cell clusters were similarly distinct, but using single cell data significantly increased the separation of two pairs of similar types: 1. L4 Arf5 and L5a Hsd11b1, and

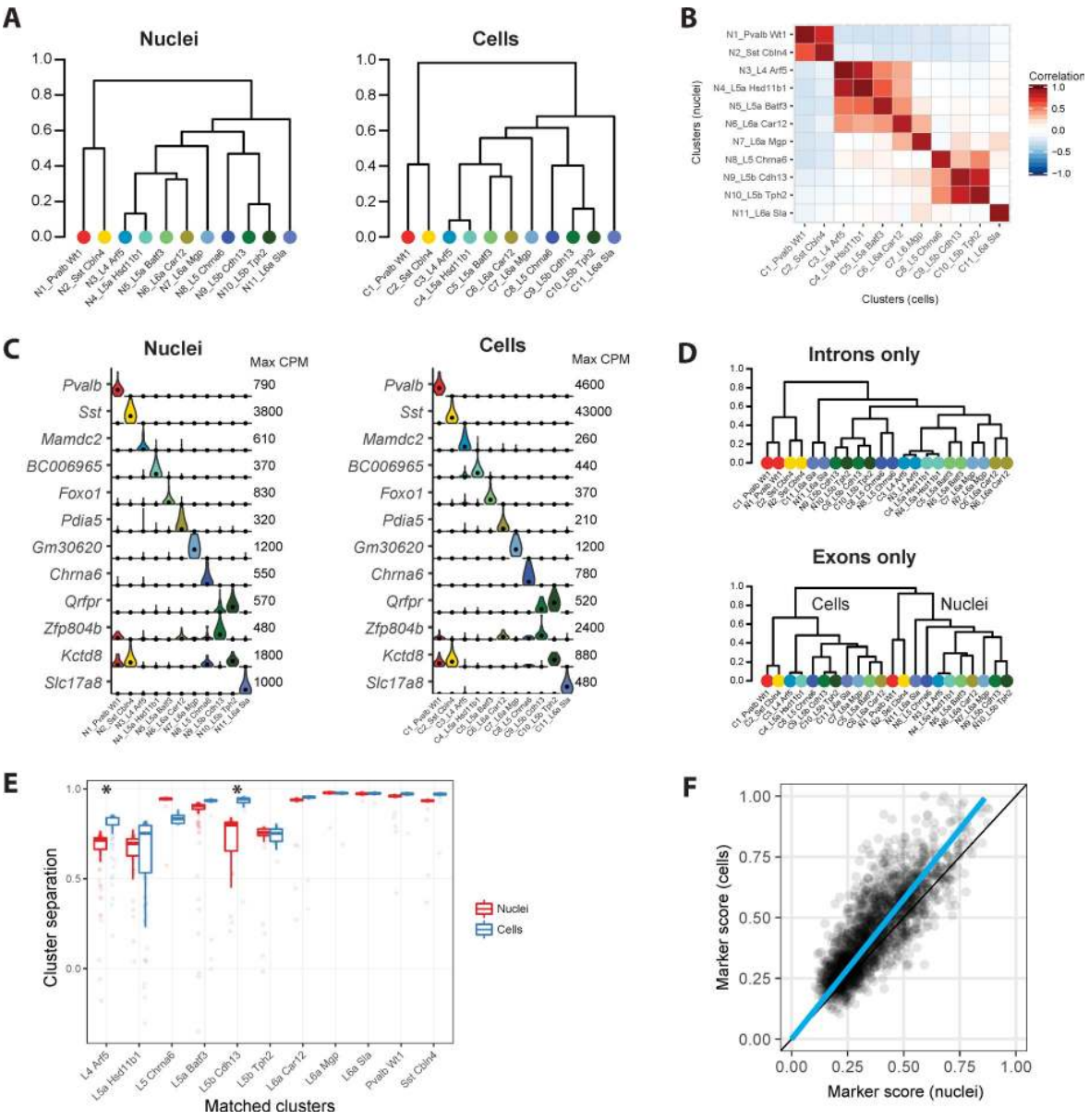

Fig 4. Similar neuronal cell types identified with nuclei and cells. (A) Cluster dendrograms for nuclei and cells based on hierarchical clustering of average expression of the top 1200 cluster marker genes. 11 clusters are labeled based on dendrogram leaf order and the closest matching mouse VISp cell type described in based on correlated marker gene expression (see S4 Fig). **(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 both nuclei and cells, although marker scores (see Methods) are on average 15% higher for cells.

https://doi.org/10.1371/journal.pone.0209648.g004

2. L5b Cdh13 and L5b Tph2 (Fig 4E). Next, we scored each gene for its ability to differentiate cell types by defining a marker score. Briefly, for each gene, the proportion of nuclei or cells with expression above background noise (CPM > 1) was calculated for each cluster, and the

marker score was calculated as the sum of the squared differences in proportions divided by the sum of the differences in proportions. Scores range from zero (ubiquitous or no expression) to one (perfectly binary expression in a subset of clusters). On average, marker scores were 15% higher in cells than nuclei due to reduced expression dropouts in cells (Fig 4F). This better detection of marker genes contributed to the mildly improved cluster separation when single cell data were used.

Nuclear content varies among cell types and for different transcripts

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

Based on the comparison of scRNA-seq and snRNA-seq data, we estimate that L4 types have high nuclear to cell volume ratio (~50%), whereas L5 types have lower nuclear to cell volume (~20%). To evaluate this finding, we measured nucleus and soma sizes of different cell types *in situ*. These types were labeled by fluorescent proteins in transgenic mice containing different Cre-transgenes and a Cre-reporter. *Nr5a1*-Cre and *Scnn1a-Tg3*-Cre mice almost exclusively label two cell types (L4 Arf5 and L5a Hsd11b1), whereas *Rbp4*-Cre mice label all layer 5 cell types including L5a Hsd11b1 (S5B Fig and S5 Table) [5]. We measured the nucleus and cell body sizes *in situ*, calculated their volumes, and derived the nuclear volume proportion (S5C Fig). Nuclear proportions of layer 5 neurons were systematically higher than those predicted based on RNA-seq data (Fig 5D, *Rbp4*), likely due to under-estimation of cell body volume based on cross-sectional area measurements of these large non-spherical (pyramidal) neurons. Despite this, we found that layer 5 neurons had lower average nuclear volume proportions than layer 4 neurons (Fig 5D), consistent with a lower nuclear transcript proportions in layer 5 neurons (Fig 5C).

To test whether layer 5 neurons were exceptional compared to neurons in other layers, we performed an unbiased survey of nuclear volume proportions across the full depth of the cortex. We found that for most cortical cells the nucleus fills more than half (average 0.63; standard deviation 0.12) of the cell body (S5D Fig). A minority of neurons in deep layer 3 and layer 5 have large cell bodies and proportionally smaller nuclei (proportion <0.4) than other cortical neurons. Interestingly, layer 5 neurons with similar morphology were recently described in rat primary visual cortex and found to be polyploid [26].

Next, we investigated transcript localization for individual genes independent of cell type. The nuclear proportion of 11,932 transcripts was estimated by the ratio of nuclear to whole cell expression multiplied by the overall nuclear fraction of each cell type and averaged across cell types (S6 Table). Different functional classes of genes had strikingly different nuclear proportions (Fig 5E). Many non-coding transcripts were localized in the nucleus, but some were abundantly expressed in the cytoplasm, such as the long non-coding RNA (lncRNA) *Tunar* that is highly enriched in the brain, is conserved across vertebrates, and has been associated with striatal pathology in Huntington's disease [27]. Most protein-coding transcripts were expressed in both the nucleus and cytoplasm with a small number restricted to the nucleus, including the Parkinson's risk gene *Park2*. We found that many genes with transcripts

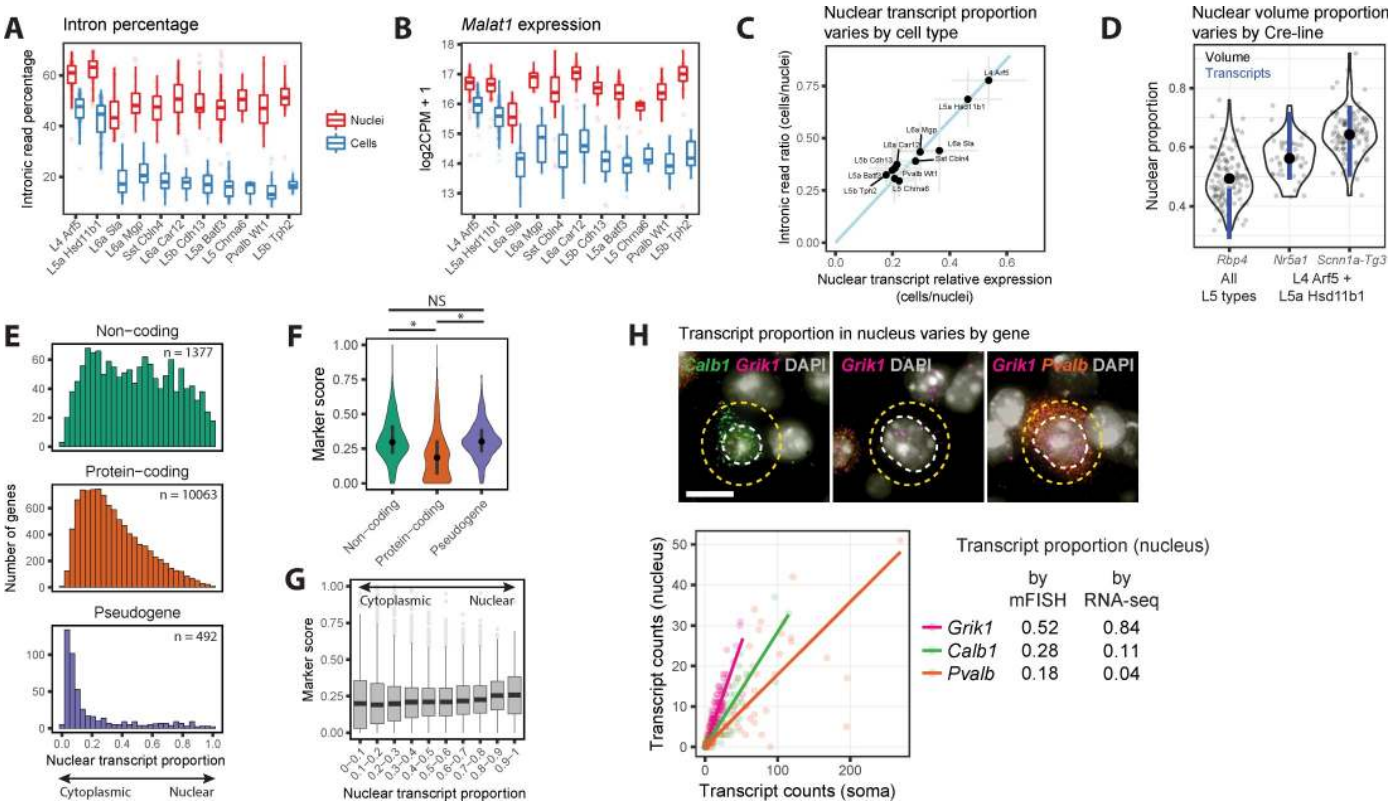

Fig 5. Nuclear transcript content varies among cell types and genes. (A) Box plots showing median (bars), 25th and 75th quantiles (boxes), and range (whiskers) of percentages of reads mapping to introns for matched nuclei and cell clusters. **(B)** Box plots of $\log_2$-transformed expression of the 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 S5D Fig). 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: $^{*}P < 1 \times 10^{-50}$ (Bonferroni-corrected), NS, not significant. **(G)** Box plots of cell type marker scores for genes grouped by estimated nuclear transcript proportion. **(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.

https://doi.org/10.1371/journal.pone.0209648.g005

restricted to the cytoplasm were involved in house-keeping functions including mitochondrial genes and, surprisingly, pseudogenes. The functional paralogs of pseudogenes are often house-keeping genes because many pseudogenes were inserted into the genome during mammalian evolution by retrotransposition of house-keeping transcripts that are highly expressed during embryonic development [28]. Finally, we found that 2550 gene transcripts shown to be localized to remote neuronal dendrites [29] were not enriched in the nucleus. These transcripts were similarly distributed between the nucleus and cytoplasm as transcripts overall and had a median nuclear fraction of 27%.

We compared our estimates of nuclear transcript proportions in cortex to mouse liver and pancreas [25]. We found moderately high correlation (r = 0.61) between 4,373 mostly house-keeping genes that were expressed in all three tissues (S5E Fig). Moreover, the shapes of the distributions of nuclear transcript proportions were highly similar between tissues with slightly

higher proportions estimated in this study (S5F Fig). These results suggest that the mechanisms regulating the intracellular localization of these transcripts–for example, rates of nuclear export and cytoplasmic degradation [25]–are generally conserved across disparate cell types.

We investigated if marker score varied based on gene function or subcellular transcript localization. Surprisingly, non-coding genes and pseudogenes are better markers of cell types, on average, than protein-coding genes (Fig 5F). lncRNAs are known to have specific expression among diverse human cell lines [30], and we show that this is also true for neuronal types in the mouse cortex. Many pseudogene transcripts, most of which are enriched in the cytoplasm (Fig 5E), were selectively depleted in the two cell types, L4 Arf5 and L5a Hsd11b1. This is consistent with our previous analysis that showed that neurons of these types have relatively less cytoplasm. Transcript localization does not appear to be correlated with marker score (Fig 5G).

Finally, we compared nuclear transcript proportions determined by RNA-seq for three genes–*Calb1*, *Grik1*, and *Pvalb*–to proportions estimated by RNA fluorescence *in situ* hybridization (FISH). Despite differences in the absolute values for nuclear transcript proportions between the two methods, we found the same trend for the three genes (Fig 5H). Both methods confirmed that *Pvalb* transcripts were mostly excluded from the nucleus, and this explained why 2 out of 35 nuclei in the Pvalb-positive interneuron type (Pvalb Wt1) had no detectable *Pvalb* mRNA expression, whereas all cells of this cell type had robust *Pvalb* mRNA expression.

Discussion

As large scale initiatives begin to characterize transcriptomic cell types in the whole brain [33] and whole organism [34], it is important to understand the strengths and limitations of different mRNA profiling techniques. Unlike scRNA-seq, snRNA-seq enables transcriptomic profiling of tissues that are refractory to whole-cell dissociation and of archived frozen specimens (e.g., banked human tissue). snRNA-seq is also less susceptible to perturbations of gene expression that occur during cell isolation, such as increased expression of immediate early genes that can obscure transcriptional signatures of neuronal activity [13]. Nuclear profiling is likely to be less cell type biased than scRNA-seq. For example, single cell profiling of adult human cortex isolated 75% interneurons and 25% excitatory neurons [11], whereas single nucleus profiling of the same tissue type isolated 30% interneurons and 70% excitatory neurons [14], close to the proportions found in situ. However, these advantages come at the cost of profiling less mRNA, and it was unclear if the nucleus contained sufficient number and diversity of transcripts to distinguish highly related cell types.

To directly address this question, we profiled a well-matched set of 463 nuclei and 463 cells from layer 5 of mouse primary visual cortex and identified 11 matching neuronal types: 2 interneuron types and 9 excitatory neuron types. Including intronic reads in gene expression quantification was necessary to achieve high-resolution cell type identification from single nuclei. Intronic reads substantially increased gene detection to 7000 genes per nucleus, including cell type-informative genes with robust expression in whole cells. In addition, intronic reads were more frequently derived from long genes that are known to have brain-specific expression [31] and that help define neuronal connectivity and signaling. Intronic reads may also reflect other cell type specific features, such as retained introns or alternative isoforms. For example, intron retention provides a mechanism for the nuclear storage and rapid translation of long transcripts in response to neuronal activity [32].

We found that nuclei contain at least 20% of all cellular transcripts, and this percentage varies among cell types. These findings are consistent with our *in situ* measurements of nuclear volumes and proportions relative to cell volumes. For example, two small pyramidal neuron

types have large nuclei relative to cell size, and these nuclei contain more than half of the cellular transcripts captured by scRNA-seq.

On average, we detect 4000 more genes in a single cell than a single nucleus, yet over 20,000 genes are detected equally well in matched cells and nuclei. Cell type marker scores were 15% higher in cells than nuclei due to reduced expression dropouts in cells (Fig 4F). This better detection of marker genes in cells contributed to mildly improved cluster separation between two pairs of highly similar cell types when using scRNA-seq data. However, sampling more nuclei may potentially compensate for decreased gene detection and result in comparable separation of cell types. In summary, we show that deep snRNA-seq is well suited for large-scale surveys of cellular diversity in various tissues as it provides similar resolution for cell type detection to scRNA-seq.

Materials and methods

Animals and tissue preparation

All procedures were approved by the Institutional Animal Care and Use Committee at the Allen Institute for Brain Science (Protocol no. 1511). Animals were provided food and water *ad libitum* and were maintained on a regular 12-h day/night cycle. Mice were housed at no more than 5 adults per cage, with various enrichment materials added, including nesting materials, gnawing materials, and plastic shelters. Nutritional and foraging enrichment was provided in the form of foods such as sunflower seeds and sucrose pellets. Mice were maintained on the C57BL/6J background.

Tissue samples were obtained from adult (postnatal day (P) 53–59)) male and female transgenic mice carrying a Cre transgene and a Cre-reporter transgene. Prior to euthanasia, to avoid pain and distress, mice were anesthetized with 5% isoflurane. While still under anesthesia, they were intracardially perfused with either 25 or 50 ml of ice cold, oxygenated artificial cerebral spinal fluid (ACSF) at a flow rate of 9 ml per minute until the liver appeared clear, or the full volume of perfusate had been flushed through the vasculature. The ACSF solution consisted of 0.5mM $CaCl_2$, 25mM D-Glucose, 98mM HCl, 20mM HEPES, 10mM $MgSO_4$, 1.25mM NaH_2PO_4, 3mM Myo-inositol, 12mM N-acetylcysteine, 96mM N-methyl-D-glucamine, 2.5mM KCl, 25mM $NaHCO_3$, 5mM sodium L-Ascorbate, 3mM sodium pyruvate, 0.01mM Taurine, and 2mM Thiourea. The brain was then rapidly dissected and mounted for coronal slice preparation on the chuck of a Compresstome VF-300 vibrating microtome (Precisionary Instruments). Using a custom designed photodocumentation configuration (Mako G125B PoE camera with custom integrated software), a blockface image was acquired before each section was sliced at 250 μm intervals. The slice was then hemisected along the midline, and both hemispheres were then transferred to chilled, oxygenated ACSF.

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

For whole cell dissociation, after all regions of interest were dissected, the ACSF was removed and 1 ml of a 2 mg/ml pronase in ACSF solution was added. Tissue was digested at room temperature (approximately 22˚C) for a duration that consisted of adding 15 minutes to the age of the mouse (in days; *i.e.*, P53 specimen had a digestion time of 68 minutes). After digestion, the pronase solution was removed and replaced by 1 ml of ACSF supplemented with 1% Fetal Bovine Serum (FBS). The tissue was washed two more times with the same solution and the sample was then triturated using fire-polished glass pipettes of decreasing bore sizes (600, 300, and 150 μm). The cell suspension was incubated on ice in preparation for fluorescence-activated cell sorting (FACS). FACS preparation involved adding 4'-6-diamidino-2-phenylindole (DAPI) at a final concentration of 4 μg/ml to label dead (DAPI+) versus live (DAPI-) cells. The suspension was then filtered through a fine-mesh cell strainer to remove cell aggregates. Cells were sorted by excluding DAPI positive events and debris, and gating to include red fluorescent events (tdTomato-positive cells). Single cells were collected into strip tubes containing 11.5μl of collection buffer (SMART-Seq v4 lysis buffer 0.83x, Clontech #634894), RNase Inhibitor (0.17U/μl), and ERCCs (External RNA Controls Consortium, MIX1 at a final dilution of 1x10-8) [35,36]. After sorting, strip tubes containing single cells were centrifuged briefly and then stored at -80˚C.

For nuclei isolation, dissected regions of interest were transferred to microcentrifuge tubes, snap frozen in a slurry of dry ice and ethanol, and stored at -80˚C until the time of use. To isolate nuclei, frozen tissues were placed into a homogenization buffer that consisted of 10mM Tris pH 8.0, 250mM sucrose, 25mM KCl, 5mM MgCl2, 0.1% Triton-X 100, 0.5% RNasin Plus RNase inhibitor (Promega), 1X protease inhibitor (Promega), and 0.1mM DTT. Tissues were placed into a 1ml dounce homogenizer (Wheaton) and homogenized using 10 strokes of the loose dounce pestle followed by 10 strokes of the tight pestle to liberate nuclei. Homogenate was strained through a 30μm cell strainer (Miltenyi Biotech) and centrifuged at 900xg for 10 minutes to pellet nuclei. Nuclei were then resuspended in staining buffer containing 1X PBS supplemented with 0.8% nuclease-free BSA and 0.5% RNasin Plus RNase inhibitor. Mouse anti-NeuN antibody (EMD Millipore, MAB377, Clone A60) was added to the nuclei at a final dilution of 1:1000 and nuclei suspensions were incubated at 4˚C for 30 minutes. Nuclei suspensions were then centrifuged at 400xg for 5 minutes and resuspended in clean staining buffer (1X PBS, 0.8% BSA, 0.5% RNasin Plus). Secondary antibody (goat anti-mouse IgG (H+L), Alexa Fluor 594 conjugated, ThermoFisher Scientific) was applied to nuclei suspensions at a dilution of 1:5000 for 30 minutes at 4˚C. After incubation in secondary antibody, nuclei suspensions were centrifuged at 400xg for 5 minutes and resuspended in clean staining buffer. Prior to FACS, DAPI was applied to nuclei suspensions at a final concentration of 0.1μg/ml and nuclei suspensions were filtered through a 35μm nylon mesh to remove aggregates. Single nuclei were captured by gating on DAPI-positive events, excluding debris and doublets, and then gating on Alexa Fluor 594 (NeuN) signal. Strip tubes containing FACS isolated single nuclei were then briefly centrifuged and frozen at -80˚C.

RNA amplification and library preparation for RNA-seq

The SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech #634894) was used per the manufacturer's instructions for reverse transcription of single cell RNA and subsequent cDNA synthesis. Single cells were stored in 8-strips at -80˚C in 11.5 μl of collection buffer (SMART-Seq v4 lysis buffer at 0.83x, RNase Inhibitor at 0.17 U/μl, and ERCC MIX1 at a final dilution of 1x10-8 dilution). Twelve to 24 8-well strips were processed at a time (the equivalent of 1–2 96-well plates). At least 1 control strip was used per amplification set, containing 2 wells without cells but including ERCCs, 2 wells without cells or ERCCs, and either 4 wells of 10 pg

of Mouse Whole Brain Total RNA (Zyagen, MR-201) or 2 wells of 10 pg of Mouse Whole Brain Total RNA (Zyagen, MR-201) and 2 wells of 10 pg Control RNA provided in the Clontech kit. Mouse whole cells were subjected to 18 PCR cycles after the reverse transcription step, whereas mouse nuclei were subjected to 21 PCR cycles. AMPure XP Bead (Agencourt AMPure beads XP PCR, Beckman Coulter A63881) purification was done using the Agilent Bravo NGS Option A instrument. A bead ratio of 1x was used (50 μl of AMPure XP beads to 50 μl cDNA PCR product with 1 μl of 10x lysis buffer added, as per Clontech instructions), and purified cDNA was eluted in 17 μl elution buffer provided by Clontech. All samples were quantitated using PicoGreen on a Molecular Dynamics M2 SpectraMax instrument. A portion of the samples, and all controls, were either run on the Agilent Bioanalyzer 2100 using High Sensitivity DNA chips or the Advanced Analytics Fragment Analyzer (96) using the High Sensitivity NGS Fragment Analysis Kit (1bp-6000bp) to qualify cDNA size distribution. An average of 7.3 ng of cDNA was synthesized across all non-control samples. Purified cDNA was stored in 96-well plates at -20˚C until library preparation.

Sequencing libraries were prepared using NexteraXT (Illumina, FC-131-1096) with NexteraXT Index Kit V2 Set A (FC-131-2001). NexteraXT libraries were prepared at 0.5x volume, but otherwise followed the manufacturer's instructions. An aliquot of each amplified cDNA sample was first normalized to 30 pg/μl with Nuclease-Free Water (Ambion), then this normalized sample aliquot was used as input material into the NexteraXT DNA Library Prep (for a total of 75pg input). AMPure XP bead purification was done using the Agilent Bravo NGS Option A instrument. A bead ratio of 0.9x was used (22.5 ul of AMPure XP beads to 25 ul library product, as per Illumina protocol), and all samples were eluted in 22 μl of Resuspension Buffer (Illumina). All samples were run on either the Agilent Bioanalyzer 2100 using High Sensitivity DNA chips or the Advanced Analytics Fragment Analyzer (96) using the High Sensitivity NGS Fragment Analysis Kit (1bp-6000bp) to for sizing. All samples were quantitated using PicoGreen using a Molecular Dynamics M2 SpectraMax instrument. Molarity was calculated for each sample using average size as reported by Bioanalyzer or Fragment Analyzer and pg/μl concentration as determined by PicoGreen. Samples (5 μl aliquot) were normalized to 2–10 nM with Nuclease-free Water (Ambion), then 2 μl from each sample within one 96-index set was pooled to a total of 192 μl at 2–10 nM concentration. A portion of this library pool was sent to an outside vendor for sequencing on an Illumina HS2500. All of the library pools were run using Illumina High Output V4 chemistry. Covance Genomics Laboratory, a Seattle-based subsidiary of LabCorp Group of Holdings, performed the RNA-Sequencing services. An average of 229 M reads were obtained per pool, with an average of 2.0–3.1 M reads/cell across the entire data set.

RNA-Seq data processing

Raw read (fastq) files were aligned to the GRCm38 mouse genome sequence (Genome Reference Consortium, 2011) with the RefSeq transcriptome version GRCm38.p3 (current as of 1/15/2016) and updated by removing duplicate Entrez gene entries from the gtf reference file for STAR processing. For alignment, Illumina sequencing adapters were clipped from the reads using the fastqMCF program [37]. After clipping, the paired-end reads were mapped using Spliced Transcripts Alignment to a Reference (STAR) [19] using default settings. STAR uses and builds its own suffix array index which considerably accelerates the alignment step while improving sensitivity and specificity, due to its identification of alternative splice junctions. Reads that did not map to the genome were then aligned to synthetic constructs (i.e. ERCC) sequences and the *E.coli* genome (version ASM584v2). Quantification was performed using summerizeOverlaps from the R package GenomicAlignments [38]. Read alignments to the genome (exonic, intronic, and intergenic counts) were visualized as beeswarm plots using the R package *beeswarm*.

Expression levels were calculated as counts per million (CPM) of exonic plus intronic reads, and $\log_2(\text{CPM} + 1)$ transformed values were used for a subset of analyses as described below. Gene detection was calculated as the number of genes expressed in each sample with CPM > 0. CPM values reflected absolute transcript number and gene length, i.e. short and abundant transcripts may have the same apparent expression level as long but rarer transcripts. Intron retention varied across genes so no reliable estimates of effective gene lengths were available for expression normalization. Instead, absolute expression levels were estimated as fragments per kilobase per million (FPKM) using only exonic reads so that annotated transcript lengths could be used. CPM expression values were used for all analyses, figures, and tables except for calculating the average expression (FPKM) of clusters with maximum expression for each gene listed in S6 Table.

Selection of single nuclei and matched cells

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

Gene expression was more likely to drop out in samples with lower quality cDNA libraries and for low expressing genes. To estimate gene dropouts due to stochastic transcription or technical artifacts [20], expression noise models were fit separately to single nuclei and cells using the "knn.error.models" function of the R package *scde* (version 2.2.0) with default settings and eight nearest neighbors. Noise models were used to calculate a dropout weight matrix that represented the likelihood of expression dropouts based on average gene expression levels of similar nuclei or cells using mode-relative weighting [39]. The probability of dropout for each sample (s) and gene (g) was estimated based on two expression measurements: average expected expression level of similar samples, $p(x_{\bar{g}})$, and observed expression levels, $p(x_{sg})$, using the "scde.failure.probability" and "scde.posteriors" functions. The dropout weighting was calculated as a combination of these probabilities: $W_{sg} = 1 - \sqrt{p(x_{sg}) \cdot \sqrt{p(x_{sg}) \cdot p(x_{\bar{g}})}}$.

Dropout weighted Pearson correlations were calculated between all pairs of nuclei and cells using 42,003 genes expressed in at least one nucleus and one cell. The cell with the highest correlation to any nucleus was selected as the best match, and this cell and nucleus were removed from further analysis. This process was repeated until 463 best matching cells were selected, and the expression correlations were compared to correlations of the best matching pairs of nuclei (Fig 1B). The Cre-lines and dissected cortical layers of origin of the best matching cells were summarized as bar plots (S1 Fig). Unweighted Pearson correlations were also calculated between all pairs of nuclei and cells to test the effect of accounting for dropouts on sample similarities (Fig 2B).

Differential expression analysis

Gene detection was estimated as the proportion of cells and nuclei expressing each gene (CPM > 0). In order to estimate the expected variability of gene detection as a result of population sampling, cells were randomly split into two sets of 231 and 232 cells and genes were grouped into 50 bins based on detection in the first set of cells. For each bin of genes, the 97.5 percentile of detection was calculated for the second set of cells. A 95% confidence interval of gene detection was constructed by reflecting this these binned quantiles across the line of unity. Data were summarized with a hexagonal binned scatter plot and a log-transformed color scale using the R package *ggplot2* [40].

Differential expression between nuclei and cells was calculated with the R package *limma* [41] using default settings and $\log_2$(CPM + 1) expression defined based on two sets of reads: introns plus exons and only exons. Significantly differentially expressed were defined as having >1.5-fold change and a Benjamini-Hochberg corrected P-value < 0.05. Gene expression distributions of nuclei or cells within a cluster were visualized using violin plots, density plots rotated 90 degrees and reflected on the Y-axis.

Differences in alignment statistics and gene counts were calculated between cells, nuclei, and total RNA controls (or just cells and nuclei) with analysis of variance using the "aov" function in R [42]. P-values for all comparisons were $P<10^{-13}$.

Two sets of nucleus- and cell-enriched genes (introns plus exons and exons only) were tested for gene ontology (GO) enrichment using the ToppGene Suite [43]. Significantly enriched (Benjamini-Hochberg false discovery rate < 0.05) GO terms were summarized as tree maps with box sizes proportional to -log10(P-values) using REVIGO [44] (S2 Fig).

Clustering

Nuclei and cells were grouped into transcriptomic cell types using an iterative clustering procedure based on community detection in a nearest neighbor graph as described in Levine et al. [45]. Clustering was performed using gene expression quantified with exonic reads only or intronic plus exonic reads for two key clustering steps: selecting significantly variable genes and calculating pairwise similarities between nuclei. Four combinations of expression quantification for nuclei and cells resulted in eight independent clustering runs.

For each gene, $\log_2$(CPM + 1) expression was centered and scaled across samples. Noise models were used to select significantly variable genes (adjusted variance > 1.25). Dimensionality reduction was performed with principal components analysis (PCA) on variable genes, and the covariance matrix was adjusted to account for gene dropouts using the product of dropout weights across genes for each pair of samples. A maximum of 20 principal components (PCs) were retained for which more variance was explained than the broken stick null distribution, a conservative method of PC retention [46].

Nearest-neighbor distances between all samples were calculated using the "nn2" function of the R package *RANN*, and Jaccard similarity coefficients between nearest-neighbor sets were computed. Jaccard coefficients measured the proportion of nearest neighbors shared by each sample and were used as edge weights in constructing an undirected graph of samples. Louvain community detection was used to cluster this graph with 15 nearest neighbors. Considering more than 15 neighbors reduced the power to detect small clusters due to the resolution limit of community detection [47]. Considering fewer than 15 neighbors increased over-splitting, as expected based on simulations by [48]. Fewer nearest neighbors were used only when there were 15 or fewer samples total.

Clustering significance was tested by comparing the observed modularity to the expected modularity of an Erdös-Rényi random graph with a matching number of nodes and average connection probability. Expected modularity was calculated as the maximum estimated by two reported equations [48,49]. Samples were split into clusters only if the observed modularity was greater than the expected modularity, and only clusters with distinct marker genes were retained. Marker genes were defined for all cluster pairs using two criteria: 1) significant differential expression (Benjamini-Hochberg false discovery rate < 0.05) using the R package *limma* and 2) either binary expression (CPM > 1 in >50% samples in one cluster and <10% in the second cluster) or >100-fold difference in expression. Pairs of clusters were merged if either cluster lacked at least one marker gene.

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

To assess the robustness of clusters, the iterative clustering procedure described above was repeated 100 times for random sets of 80% of samples. A co-clustering matrix was generated that represented the proportion of clustering iterations that each pair of samples were assigned to the same cluster. Average-linkage hierarchical clustering was applied to this matrix followed by dynamic branch cutting using "cutreeHybrid" in the R package *WGCNA* [50] with cut height ranging from 0.01 to 0.99 in steps of 0.01. A cut height was selected that resulted in the median number of clusters detected across all 100 iterations. Cluster cohesion (average within cluster co-clustering) and separation (difference between within cluster co-clustering and maximum between cluster co-clustering) was calculated for all clusters. Marker genes were defined for all cluster pairs as described above, and clusters were merged if they had a co-clustering separation <0.25 or either cluster lacked at least one marker gene.

Scoring marker genes based on cluster specificity

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

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

Cluster dendrograms

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

Matching clusters based on marker gene expression

Nuclei and cell clusters were independently compared to published mouse VISp cell types [5]. The proportion of nuclei or cells expressing each gene with CPM > 1 was calculated for all clusters. Approximately 400 genes were markers in both data sets (beta score > 0.3) and were expressed in the majority of samples of between one and five clusters. Markers expressed in more than five clusters were excluded to increase the specificity of cluster matching. Weighted correlations were calculated between all pairs of clusters across these genes and weighted by beta scores to increase the influence of more informative genes. Heatmaps were generated to visualize all cluster correlations. All nuclei and cell clusters had reciprocal best matching clusters from Tasic et al. and were labeled based on these reported cluster names.

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

Nuclei and cell clusters were also compared by calculating average cluster expression based only on intronic or exonic reads and calculating a correlation-based distance using the top 1200 marker genes as described above. Hierarchical clustering was applied to all clusters quantified using the two sets of reads. In addition, the average log2(CPM + 1) expression across all nuclei and cells was calculated using intronic or exonic reads.

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

Estimating proportions of nuclear transcripts

The nuclear proportion of transcripts was estimated in two ways. First, all intronic reads were assumed to be from transcripts localized to the nucleus so that the proportion of intronic reads measured in cells should decrease linearly with the nuclear proportion of the cell as nuclear reads are diluted with cytoplasmic reads. For each cell type, the nuclear proportion was estimated as the proportion of intronic reads in cells divided by the proportion of intronic reads in matched nuclei. Second, the nuclear proportion was estimated as the average ratio of cell to nuclear expression (CPM) using only exonic reads of three highly expressed nuclear genes (*Snhg11*, *Malat1*, and *Meg3*). The standard deviation of nuclear proportion estimates were calculated based on standard error propagation of variation in intronic read proportions and expression levels. Nuclear proportion estimates were compared with linear regression, and the estimate based on relative expression levels was used for further analysis.

The nuclear proportion of transcripts for all genes was estimated for each cell type as the ratio of average expression (CPM) using only exonic reads in nuclei versus matched cells multiplied by the nuclear proportion of all transcripts. Estimated proportions greater than 1 were set equal to 1 for each cell type, and a weighted average proportion was calculated for each gene with weights equal to the average $\log_2$(CPM + 1) expression in each cell type. 11,932 genes were expressed in at least one nuclear or cell cluster (>50% samples expressed with CPM > 1) and were annotated as one of three gene types–protein-coding, protein non-coding, or pseudogene–using gene metadata from NCBI (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 nuclear proportions were generated. Next, beta marker score distributions were visualized as violin plots, and differences across gene types were compared with a Kruskal-Wallis rank sum test followed by Wilcoxon signed rank unpaired tests. Finally, genes were grouped into 10 bins of estimated nuclear proportions, from high cytoplasmic enrichment to high nuclear enrichment, and beta marker score distributions were visualized as box plots. A linear regression was fit to marker scores versus nuclear proportion.

Nuclear transcript proportions were compared to nuclear proportions estimated for mouse liver and pancreatic beta cells based on data from Halpern et al. [25]. Ratios of normalized

nuclear and cytoplasmic transcript counts were calculated in four tissue replicates. Average ratios were calculated for genes with at least one count in either fraction in at least one tissue. Nuclear proportion estimates for all genes with data from both data sets (n = 4373) were compared with Pearson correlation, a linear model with intercept set equal to zero, and histograms with a bin width of 0.02.

Colorimetric in situ hybridization

In situ hybridization data for mouse cortex was from the Allen Mouse Brain Atlas [51]. All data is publicly accessible through www.brain-map.org. Data was generated using a semi-automated technology platform as described in Lein et al. [51]. Mouse ISH data shown is from primary visual cortex (VISp) in the Paxinos Atlas [52].

Multiplex fluorescence RNA in situ hybridization and quantification of nuclear versus cytoplasmic transcripts

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

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

In situ quantification of nucleus and soma size

Coronal brain slices from *Nr5a1-Cre;Ai14*, *Scnn1a-Tg3-Cre;Ai14*, and *Rbp4-Cre_KL100;Ai14* mice were stained with anti-dsRed (Clontech #632496) to enhance tdTomato signal in red channel and DAPI to label nuclei. Maximum intensity projections from six confocal stacks of 1-μm intervals were processed for analysis. Initial segmentation was performed by CellProfiler [53] to identify nuclei from the DAPI signal and soma from the tdTomato signal. Segmentation results were manually verified and any mis-segmented nuclei or somata were removed or re-segmented if appropriate. Area measurement of segmented nuclei and somata was performed in CellProfiler in Layer 4 from *Nr5a1-Cre;Ai14* and *Scnn1a-Tg3-Cre;Ai14* mice, and in Layer 5 from *Rbp4-Cre_KL100;Ai14* mice. A linear regression was fit to nuclear versus soma area to highlight the differences between Cre-lines.

For measurements of nucleus and soma size agnostic to Cre driver, we used 16 μm-tissue sections from P56 mouse brain. To label nuclei, DAPI was applied to the tissue sections at a

final concentration of 1mg/ml. To label cell somata, tissue sections were stained with Neurotrace 500/525 fluorescent Nissl stain (ThermoFisher Scientific) at a dilution of 1:100 in 1X PBS for 5 minutes, followed by brief washing in 1X PBS. Sections were coverslipped with Fluoromount-G (Southern Biotech) and visualized on a Nikon TiE epifluorescence microscope using a 40x oil objective. Soma and nuclei area measurements were taken by tracing the boundaries of the Nissl-stained soma or DAPI-stained nucleus, respectively, using cell measurement tools available in the Nikon TiE microscope software. All cells with a complete nucleus clearly present within the section were measured, except that we excluded glial cells which had very small nuclei and scant cytoplasm. Measurements were taken within a 40x field of view across an entire cortical column encompassing layers 1–6, and the laminar position of each cell (measured as depth from the pial surface) was tracked along with the nucleus and soma area measurements for each cell.

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

Code availability

Data and code to reproduce figures are publicly available from GitHub at https://github.com/AllenInstitute/NucCellTypes. Single-cell and single-nucleus transcriptomic data are available at the NCBI Gene Expression Omnibus (GEO) under accession number GSE123454.

Supporting information

S1 Fig. 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. All recombinase lines were crossed to either *Ai14* or *Ai110* [54], except *Chrna2-Cre_OE25;Pvalb-T2A-Dre;Ai66D* [55], and *Trib2--F2A-CreERT2;Snap25-LSL-F2A-GFP* [55], for which the reporters are indicated.
(TIFF)

S2 Fig. Nuclear versus whole cell transcript dropouts and intron retention. (A) Gene detection violin plots for nuclei and cells at different sub-sampled read depths. Note that while gene detection does not fully saturate, 90% as many genes are detected with 1 million versus approximately 2.5 million ("All") reads. **(B)** Rate of gene dropouts in nuclei versus cells (i.e. proportion of nuclei/cells with zero expression) as compared to the average gene expression level across all nuclei and cells. Loess fits to dropout rates of genome-wide genes. **(C)** Density plots showing the properties of all expressed genes (black lines) and 1334 genes (red lines) that have >25% detection in nuclei using intronic plus exonic reads versus only exonic reads. Mean expression was calculated using only exonic reads in cells, and beta marker scores were calculated for cell clusters as described in the **Methods**. **(D)** REVIGO 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. **(E)** 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).
(TIFF)

S3 Fig. Overview of single nucleus RNA-seq clustering pipeline. See methods for a detailed description of clustering steps.
(TIFF)

S4 Fig. Nuclear and cell clusters are well matched based on marker gene expression. (A) Pairwise correlations between previously reported mouse VISp cell type clusters 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 Fig 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 L5 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.
(TIFF)

S5 Fig. 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 μm. **(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 μm. Right: Nuclear volume proportion was quantified based on nucleus and soma area measurements and plotted as a function of cortical depth. Size and color of points are proportional to soma volume. **(E)** Average nuclear proportions of 4,373 genes (mostly house-keeping) also expressed in mouse pancreatic beta-cells and liver cells 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 Fig 5D.
(TIFF)

S1 Table. Average gene expression and detection in matched nuclei and cells.
(XLSX)

S2 Table. Differentially expressed genes in cells versus nuclei using intronic plus exonic reads.
(XLSX)

S3 Table. Gene ontology (GO) enrichment of differentially expressed genes in cells and nuclei based on intronic and exonic reads or only exonic reads.
(XLSX)

S4 Table. Differentially expressed genes in cells versus nuclei using only exonic reads.
(XLSX)

S5 Table. Cre-driver line composition of cell clusters.
(XLSX)

S6 Table. Gene properties including the number of clusters with any expression, maximum cluster expression (FPKM of exonic reads only), cell type marker score (beta), and estimated nuclear proportion of transcripts.
(XLSX)

Acknowledgments

The authors thank the Allen Institute founder, P. G. Allen, for his vision, encouragement and support.

Author Contributions

Conceptualization: Trygve E. Bakken, Amy Bernard, Hongkui Zeng, Ed S. Lein, Bosiljka Tasic.

Data curation: Trygve E. Bakken, Jeremy A. Miller.

Formal analysis: Trygve E. Bakken, Jeremy A. Miller, Zizhen Yao, Jeff Goldy.

Investigation: Rebecca D. Hodge, Thuc Nghi Nguyen, Eliza Barkan, Darren Bertagnolli, Tamara Casper, Nick Dee, Emma Garren, Lucas T. Graybuck, Matthew Kroll, Kanan Lathia, Sheana Parry, Christine Rimorin, Soraya I. Shehata, Michael Tieu, Kimberly A. Smith.

Methodology: Rebecca D. Hodge, Brian Aevermann, Roger S. Lasken, Richard H. Scheuermann, Nicholas J. Schork.

Project administration: John W. Phillips, Amy Bernard, Kimberly A. Smith, Hongkui Zeng, Bosiljka Tasic.

Supervision: John W. Phillips, Amy Bernard, Hongkui Zeng, Ed S. Lein, Bosiljka Tasic.

Validation: Rebecca D. Hodge, Thuc Nghi Nguyen, Eliza Barkan, Emma Garren.

Visualization: Trygve E. Bakken, Rebecca D. Hodge, Jeremy A. Miller.

Writing – original draft: Trygve E. Bakken, Rebecca D. Hodge, Jeremy A. Miller, Ed S. Lein, Bosiljka Tasic.

Writing – review & editing: Trygve E. Bakken, Rebecca D. Hodge, Jeremy A. Miller, Ed S. Lein, Bosiljka Tasic.

References

1. Poulin J, Tasic B, Hjerling-Leffler J, Trimarchi JM, Awatramani R. Disentangling neural cell diversity using single-cell transcriptomics. Nat Neurosci. 2016; 19: 1131–41. https://doi.org/10.1038/nn.4366 PMID: 27571192
2. Zeng H, Sanes JR. Neuronal cell-type classification: challenges, opportunities and the path forward. Nat Rev Neurosci. Nature Publishing Group; 2017; https://doi.org/10.1038/nrn.2017.85 PMID: 28775344
3. Bernard A, Sorensen SA, Lein ES. Shifting the paradigm: new approaches for characterizing and classifying neurons. Current Opinion in Neurobiology. 2009. https://doi.org/10.1016/j.conb.2009.09.010 PMID: 19896835
4. Tasic B. Single cell transcriptomics in neuroscience: cell classification and beyond. Current Opinion in Neurobiology. 2018. https://doi.org/10.1016/j.conb.2018.04.021 PMID: 29738987
5. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016; 19: 335–346. https://doi.org/10.1038/nn.4216 PMID: 26727548
6. Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature. Springer US; 2018; 563: 72–78. https://doi.org/10.1038/s41586-018-0654-5 PMID: 30382198
7. Zeisel A, Manchado ABM, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-). 2015; science.aaa1934-. https://doi.org/10.1126/science.aaa1934 PMID: 25700174
8. Campbell JN, Macosko EZ, Fenselau H, Pers TH, Lyubetskaya A, Tenen D, et al. A molecular census of arcuate hypothalamus and median eminence cell types. Nat Neurosci. 2017; 20: 484–496. https://doi.org/10.1038/nn.4495 PMID: 28166221
9. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell. Elsevier Inc.; 2016; 166: 1308–1323.e30. https://doi.org/10.1016/j.cell.2016.07.054 PMID: 27565351
10. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. Elsevier Inc.; 2015; 161: 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002 PMID: 26000488
11. Darmanis S, Sloan S a., Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci. 2015; 201507125. https://doi.org/10.1073/pnas.1507125112 PMID: 26060301
12. Krishnaswami SR, Grindberg R V, Novotny M, Venepally P, Lacar B, Bhutani K, et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. Nat Protoc. 2016; 11: 499–524. https://doi.org/10.1038/nprot.2016.015 PMID: 26890679
13. Lacar B, Linker SB, Jaeger BN, Krishnaswami S, Barron J, Kelder M, et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. Nat Commun. 2016; 7: 11022. https://doi.org/10.1038/ncomms11022 PMID: 27090946
14. Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Intergovernmental Panel on Climate Change, editor. Science (80-). Cambridge: American Association for the Advancement of Science; 2016; 352: 1586–1590. https://doi.org/10.1126/science.aaf1204 PMID: 27339989
15. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2017; 36: 70–80. https://doi.org/10.1038/nbt.4038 PMID: 29227469
16. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta JJ, et al. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science (80-). 2016; 353: 925–928. https://doi.org/10.1126/science.aad7038 PMID: 27471252
17. Lake BB, Codeluppi S, Yung YC, Gao D, Chun J, Kharchenko P V., et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. Sci Rep. Springer US; 2017; 7: 6031. https://doi.org/10.1038/s41598-017-04426-w PMID: 28729663
18. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017; 14: 955–958. https://doi.org/10.1038/nmeth.4407 PMID: 28846088
19. Dobin A, Davis C a., Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

20. Kharchenko P V, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014; 11: 740–2. https://doi.org/10.1038/nmeth.2967 PMID: 24836921

21. Madisen L, Zwingman TA, Sunkin SM, Oh SW, Zariwala HA, Gu H, et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. Nat Neurosci. 2010; https://doi.org/10.1038/nn.2467 PMID: 20023653

22. Li G-W, Xie XS. Central dogma at the single-molecule level in living cells. Nature. 2011; 475: 308–315. https://doi.org/10.1038/nature10315 PMID: 21776076

23. Little SC, Tikhonov M, Gregor T. Precise Developmental Gene Expression Arises from Globally Stochastic Transcriptional Activity. Cell. Elsevier Inc.; 2013; 154: 789–800. https://doi.org/10.1016/j.cell.2013.07.025 PMID: 23953111

24. Munsky B, Neuert G, van Oudenaarden A. Using Gene Expression Noise to Understand Gene Regulation. Science (80-). 2012; 336: 183–7. https://doi.org/10.1126/science.1216379 PMID: 22499939

25. Bahar Halpern K, Caspi I, Lemze D, Levy M, Landen S, Elinav E, et al. Nuclear Retention of mRNA in Mammalian Tissues. Cell Rep. The Authors; 2015; 13: 2653–2662. https://doi.org/10.1016/j.celrep.2015.11.036 PMID: 26711333

26. Sigl-Glöckner J, Brecht M. Polyploidy and the Cellular and Areal Diversity of Rat Cortical Layer 5 Pyramidal Neurons. Cell Rep. 2017; 20: 2575–2583. https://doi.org/10.1016/j.celrep.2017.08.069 PMID: 28903039

27. Lin N, Chang KY, Li Z, Gates K, Rana ZA, Dang J, et al. An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Mol Cell. Elsevier Inc.; 2014; 53: 1005–1019. https://doi.org/10.1016/j.molcel.2014.01.021 PMID: 24530304

28. Zhang Z, Carriero N, Gerstein M. Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 2004; 20: 62–67. https://doi.org/10.1016/j.tig.2003.12.005 PMID: 14746985

29. Cajigas IJ, Tushev G, Will TJ, tom Dieck S, Fuerst N, Schuman EM. The Local Transcriptome in the Synaptic Neuropil Revealed by Deep Sequencing and High-Resolution Imaging. Neuron. 2012; 74: 453–466. https://doi.org/10.1016/j.neuron.2012.02.036 PMID: 22578497

30. Djebali S, Davis C a, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012; 489: 101–8. https://doi.org/10.1038/nature11233 PMID: 22955620

31. Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, et al. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature. 2015; 522: 89–93. https://doi.org/10.1038/nature14319 PMID: 25762136

32. Mauger O, Lemoine F, Scheiffele P. Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity. Neuron. 2016; 92: 1266–1278. https://doi.org/10.1016/j.neuron.2016.11.032 PMID: 28009274

33. Ecker JR, Geschwind DH, Kriegstein AR, Ngai J, Osten P, Polioudakis D, et al. The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. Neuron. 2017. https://doi.org/10.1016/j.neuron.2017.10.007 PMID: 29096072

34. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. Elife. 2017; https://doi.org/10.7554/eLife.27041 PMID: 29206104

35. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, et al. The external RNA controls consortium: A progress report. Nat Methods. 2005; https://doi.org/10.1038/nmeth1005-731 PMID: 16179916

36. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014; https://doi.org/10.1038/nbt.2931 PMID: 25150836

37. Aronesty E. "ea-utils: Command-line tools for processing biological sequencing data." Expr Anal. 2011; http://code.google.com/p/ea-utils

38. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol. 2013; https://doi.org/10.1371/journal.pcbi.1003118 PMID: 23950696

39. SCDE by Kharchenko Lab at Harvard DBMI [Internet]. [cited 6 Jan 2018]. Available: http://hms-dbmi.github.io/scde/diffexp.html

40. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available: http://ggplot2.org

41. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; https://doi.org/10.1093/nar/gkv007 PMID: 25605792

42. Heiberger RM, Freeny AE, Chambers J. M. Analysis of Variance; Designed Experiments. Statistical Model in S. 2017. https://doi.org/10.2143/EP.6.2.505355

43. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009; https://doi.org/10.1093/nar/gkp427 PMID: 19465376

44. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011; https://doi.org/10.1371/journal.pone.0021800 PMID: 21789182

45. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. Elsevier Inc.; 2015; 162: 184–197. https://doi.org/10.1016/j.cell.2015.05.047 PMID: 26095251

46. Jackson DA. Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. Ecology. 1993; 74: 2204–2214. https://doi.org/10.2307/1939574

47. Fortunato S, Barthelemy M. Resolution limit in community detection. Proc Natl Acad Sci. 2007; 104: 36–41. https://doi.org/10.1073/pnas.0605965104 PMID: 17190818

48. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E—Stat Nonlinear, Soft Matter Phys. 2006; https://doi.org/10.1103/PhysRevE.74.016110 PMID: 16907154

49. Guimerà R, Sales-Pardo M, Amaral LAN. Modularity from fluctuations in random graphs and complex networks. Phys Rev E. 2004; 70: 025101. https://doi.org/10.1103/PhysRevE.70.025101 PMID: 15447530

50. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. Bioinformatics. 2008; https://doi.org/10.1093/bioinformatics/btm563 PMID: 18024473

51. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2006/12/08. 2007; 445: 168–176. nature05453 [pii] https://doi.org/10.1038/nature05453 PMID: 17151600

52. Paxinos G, Keith B. J. Franklin MA. Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates [Internet]. Elsevier Science; 2012. Available: https://books.google.com/books?id=8RJZLwEACAAJ

53. Lamprecht MR, Sabatini DM, Carpenter AE. CellProfiler: free, versatile software for automated biological image analysis. Biotechniques. 2007; 42: 71–75. https://doi.org/10.2144/000112257 PMID: 17269487

54. Daigle TL, Madisen L, Hage TA, Valley MT, Knoblich U, Larsen RS, et al. A Suite of Transgenic Driver and Reporter Mouse Lines with Enhanced Brain-Cell-Type Targeting and Functionality. Cell. 2018; 174: 465–480.e22. https://doi.org/10.1016/j.cell.2018.06.035 PMID: 30007418

55. Madisen L, Garner AR, Shimaoka D, Chuong AS, Klapoetke NC, Li L, et al. Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance. Neuron. 2015; 85: 942–958. https://doi.org/10.1016/j.neuron.2015.02.022 PMID: 25741722