SUPPLEMENTARY MATERIALS AND METHODS

Plasmid and virus production for isolation of GABAergic neurons

EGFP-KASH construct was a generous gift of Prof. Worman (Columbia University, NYC) inverted into pAAV-EF1a-DIO-EYFP-WPRE-hGH-polyA (Addgene, MA, #27056) using AscI and NcoI restriction sites, and WPRE was removed using ClaI restricition sites. pAAV-EF1a-Cre-WPRE-hGH-polA was obtained from Addgene (#27056). The pAAV-hSyn-EGFP-KASH-WPRE-hGH-polyA was described [8]. Concentrated adeno-associated virus 1/2 (AAV1/2) and low titer AAV1 particles in DMEM were produced and titered as described previously [8].

Stereotactic injection of AAV1/2 into the mouse brain

Stereotactic injections were approved by the MIT Committee on Animal Care (MIT CAC). 12-16 week old male vGAT-Cre mice (Slc32a1tm2(cre)Lowl, The Jackson Laboratory, ME, #016962) (Rossi J, Cell Metab 13(2):195-204) were anaesthetized by intraperitoneal (i.p.) injection of 100 mg/kg Ketamine and 10 mg/kg Xylazine and pre-emptive analgesia was given (Buprenex, 1 mg/kg, i.p.). 1 ml of high titer AAV1/2 ($\approx 4 \times 10^{12}$ Vg/ml of pAAV-EF1a-DIO-EYFP-WPRE-hGH-polyA) was injected into dorsal and/or ventral hippocampus. The following stereotactic coordinates were used: Dorsal dentate gyrus (anterior/posterior: -1.7; mediolateral: 0.6; dorsal/ventral: -2.15), ventral dentate gyrus (anterior/posterior: -3.52; mediolateral: 2.65; dorsal/ventral: -3), dorsal CA1/2 (anterior/posterior: -1.7; mediolateral: 1.0; dorsal/ventral: -1.35) and ventral CA1/2 (anterior/posterior: -3.52; mediolateral: 3.35; dorsal/ventral: -2.75). After each injection, the pipette was held in place for 5 minutes prior to retraction to prevent leakage. Finally, the incision was sutured and postoperative analgesics (Meloxicam, 1-2 mg/kg) were administered for three days following surgery.

Animal work statement

All animal work was performed under the guidelines of Division of Comparative Medicine (DCM), with protocols (0411-040-14, 0414-024-17 0911-098-11, 0911-098-14 and 0914-091-17) approved by Massachusetts Institute of Technology Committee for Animal Care (CAC), and were consistent with the Guide for Care and Use of Laboratory Animals, National Research Council, 1996 (institutional animal welfare assurance no. A-3125-01).

Immunohistochemistry and Nissl staining

Mice were sacrificed by a lethal dose of Ketamine/Xylazine 3 weeks post viral injection, and transcardially perfused with PBS followed by 4% PFA. Sagittal sections of 30 µm were cut using vibratome (Leica, IL, VT1000S) and sections were boiled for 2 min in sodium citrate buffer (10 mM tri-sodium citrate dehydrate, 0.05% Tween20, pH 6.0) and cooled down to room temperature (RT) for 30 min. Brain sections were blocked in 5% normal goat serum (NGS) (Cell Signaling Technology, MA, #5425) and 5% donkey serum (DS) (Sigma, MO, #D9663) in PBST (PBS, 0.15% Triton-X) for 1 h at RT and stained with chicken anti-GFP (Aves labs, OR, #GFP-1020, 1:400) and mouse anti-parvalbumin (Sigma, #P3088, 1:500) in 2.5% NGS and 2.5% DS in PBST over night at 4°C. Sections were washed 3 times in PBST and stained with secondary antibodies (Alexa Fluor 488 and 568, 1:1000) at RT for 1 h. After washing with PBST 3 times, sections were mounted using VECTASHIELD HardSet Mounting Medium with DAPI (Vector Laboratories, CA, #H-1500) and imaged using confocal microscopy (Zeiss, Jena DE, LSM 710, Ax10 ImagerZ2, Zen 2012 Software). For Nissl staining, mice were perfused with PBS and 4% PFA. Brain samples were dehydrated and paraffin-embedded and 7µm sagittal sections were cut. Nissl staining was performed as described elsewhere [22]. Images were taken with a Zeiss microscope and AxioCam MRm camera.

Nuc-Seq

I. Dissection of mouse hippocampal subregions, nuclei isolation and FACS sorting
Freshly dissected mouse brain samples were placed in ice cold PBS and kept cold during microdissection. Microdissections of dentate gyrus, CA1 and CA2/3 regions were performed under a stereomicroscope as described elsewhere [23]. Dissected subregions were placed into ice-cold RNAlater (Ambion, CA, RNAlater, #7020) and stored at 4°C overnight. Thoracic spinal cord of EdU injected mice were dissected in ice-cold PBS and fixed in RNAlater at 4C over night. Then samples were processed for nuclei isolation immediately or stored in -80°C. Nuclei were isolated by sucrose gradient centrifugation as described [8]

with two modifications: RNAse inhibitor (Clontech, CA, Recombinant Ribonuclease Inhibitor, #2313A, 40 units/µl) was added to the resuspension buffer (final 1U/µl), and nuclei were filtered through a 35µm cell strainer (Corning, NY, Falcon, #352235) before sorting. Nuclei were labeled with ruby dye (Thermo Fisher Scientific, MA, Vybrant DyeCycle Ruby Stain, #V-10309) added to the resuspension buffer at a concentration of 1:800. Nuclei were kept on ice until sorting using Fluorescence Activated Cell Sorting (Harvard University, Bauer Core Facility, Beckman Coulter MoFlo Astrios EQ Cell Sorter) into 96 well plates containing 5µl of TCL lysis buffer (Qiagen, CA, #1031576) added with 10% 2-Mercaptoethanol. FACS gating was set on FSC, SSC, and on fluorescent channels to include only Ruby^+ or $\text{Ruby}^+\text{GFP}^+$ nuclei (for nuclei tagged by GFP-KASH or EdU-GFP). Each 96 well plate included an empty well as a negative control and a population well of 50-100 nuclei as a positive control.

II. Single nucleus RNA library construction and sequencing

Single nucleus RNA was first purified using RNAClean XP beads (Beckman Coulter, IN, Agencourt RNAClean XP, #A63987) at 2.2X beads to sample volume ratio. Single nucleus derived cDNA libraries were generated following a modified Smart-seq2 method [21]. The protocol did not use unique molecular identifiers (UMI). Briefly, beads were eluted into 4µl elution mix made of 1µl RT primer (10µm), 1µl dNTP mix (10 mM each, Thermo Fisher Scientific, #R0191), 1µl RNAse inhibitor diluted at 1:10 in water (final 1U/µl), and 1µl H_2O. Eluted samples were incubated at 72°C for 3 min and immediately placed on ice. Each sample was added with 7 µl reverse transcription (RT) mix made of 0.75µl H_2O, 0.1µl Maxima RNase-minus RT (Thermo Fisher Scientific, Maxima Reverse Transcriptase, #EP0752), 2µl 5x Maxima RT buffer, 2µl Betaine (Sigma Aldrich, 5M, #B0300), 0.9µl $MgCl_2$ (Sigma Aldrich, 100mM, #M1028), 1µl TSO primer (10 µm), 0.25µl RNase inhibitor (40U/µl). The RT reaction was incubated at 42°C for 90 min and followed by 10 cycles of (50°C for 2 min, 42°C for 2 min), then heat inactivated at 70°C for 15 min. Samples were then amplified with an addition of 14µl polymerase chain reaction (PCR) mix made of 1µl H_2O, 0.5µl ISPCR primer (10 µm), 12.5µl KAPA HiFi Hot-Start ReadyMix (KAPA Biosystems, MA, #KK2602). The PCR reaction was performed as follows: 98°C for 3 min, 21 cycles of (98°C for 15 sec, 67°C for 20 sec, 72°C for 6 min), and final extension at 72°C for 5 min. PCR product was purified using AMPure XP (Beckman Coulter, Agencourt AMPure XP, #A63880) twice and eluted in TE buffer (Thermo Fisher Scientific, #AM9849). Purified cDNA libraries were analyzed on Agilent 2100 Bioanalyzer (Agilent, CA, Agilent High Sensitivity DNA Kit,

#5067-4626) and quantified using picogreen (Thermo Fisher Scientific, Quant-iT PicoGreen dsDNA Assay Kit, #P11496) on a plate reader (Biotek, Synergy H4, wavelength at 485nm, 528nm with 20nm bandwidth). Sequencing libraries were prepared using Nextera XT kit (Illumina, CA, #FC-131-1024) as described previously [24]. Single nucleus cDNA libraries were sequenced on an Illumina NextSeq 500 to an average depth of 632,169 reads. Primer sequences: RT primer (Integrated DNA Technologies), /5BiosG/AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN; TSO primer (Exiqon), AAGCAGTGGTATCAACGCAGAGTACrGrG+G; ISPCR primer (Integrated DNA Technologies), /5BiosG/AAGCAGTGGTATCAACGCAG*A*G*T.

Single cell dissociation and cell picking

Cells were dissociated and hand picked as described [25]. Images were taken on dissociated cells.

Sequencing reads initial processing

Tophat [26] was used to align reads to mouse mm10 UCSC genome with default parameters and the mouse gene annotations (RefSeq mm10 and Ensemble GRCm38 merged using Cufflink [27]). The alignment was visualized using integrated genome brower (IGV) [28]. To estimate gene expression, RSEM v1.27 [29] was run with default parameters on alignments created by Bowtie2 [30] (command line options -q –phred33-quals -n 2 -e 99999999 -l 25 -I 1 -X 1000 -a -m 200 -p 4 --chunkmbs 512). Estimated expression levels were multiplied by 10^6 to obtain transcript per million (TPM) estimates for each gene, and TPM estimates were transformed to log-space by taking log(TPM+1). Genes were considered detected if their transformed expression levels are equal to or above 3 (1.1 in log(TPM+1) scale). A library was filtered out if it had less than 2,000 detected genes or more than 8,000 detected genes (threshold set by analysis of 1, 2, 4, and populations of sorted nuclei). 3' and 5' bias was measured using the RNA-SeQC package [31].

Bulk Nuc-Seq and Tissue RNA-Seq

Fresh dorsal and ventral DG tissue was micro-dissected from 4 adult female mice (11 - 13 weeks) and placed in RNA-later for 24 hours. Each sample was cut in half and used as bulk tissue in RNA-Seq or bulk nuclei populations in Nuc-Seq. Nuclei isolation was done as described for Nuc-Seq protocol, except that at the last stage of the isolation, nuclei were transferred to 300µl RLT lysis buffer (QIAGEN) with 10% 2-Mercaptoethanol instead of the resuspension buffer. We proceeded immediately to extract RNA from nuclei using the RNAeasy MinElute kit (QIAGEN, #74204) according to the manufacturer's protocol. For RNA extraction from bulk tissue, the tissue was placed in 300µl RLT lysis buffer (QIAGEN) with 10% 2-Mercaptoethanol, and mechanically dissociated using tissue raptor followed by the RNAeasy MinElute protocol. For each of the 8 nuclei and 8 tissue samples, libraries were made in triplicates, using the SMARTseq2 protocol, as described for the Nuc-Seq protocol with two modifications: (1) the number of PCR cycles in the whole transcriptome amplification stage was reduced to 14 cycles; (2) 1µl of the extracted RNA was used as the initial input to the protocol, replacing 1µl of water in the first RT mix. Libraries were sequenced on the NextSeq 500 to an average read depth of 3 million reads. Correlations were calculated between each pair of samples. Number of genes detected was calculated for each quantile of expression levels by counting the number of genes with expression log(TPM+1) $>$ 1.1. Differential expression was analyzed using student's t-test, with FDR $<$ 0.01, log-ratio $>$ 1, and average expression across all nuclei or tissue samples log(TPM+1) $>$ 3.

Comparison of Nuc-Seq and single neuron RNA-Seq

For comparison of number of genes detected, the expression matrices of published single neuorn RNA-Seq were downloaded from GEO Accessions (GSE60361 on 5/28/2015 [1]; GSE71585 on 6/9/2016 [3]). To count the number of genes detected, TPM estimates were transformed to log-space by taking log(TPM+1). Genes were considered detected if their transformed expression levels are equal to or above 3 (1.1 in log(TPM+1) scale). For data from Zeisel 2015 [1] where UMI was used, genes were considered detected if their UMI counts are equal to or above 1.1.

For comparison of correlation of averaged single neuron/nuclei of CA1 pyramidal neurons, the cells labeled as 'CA1Pyr' from the single neuron RNA-Seq dataset [1] were subsampled to get a dataset (referred to as snRNA-Seq CA1Pyr) that has the same number of cells as the CA1 pyramidal nuclei from Nuc-Seq (referred to as Nuc-Seq CA1). To calculate correlations of averaged 10 single neuron/nuclei, snRNA-Seq CA1Pyr and Nuc-Seq CA1 were separately subsampled 20 times, each time to 10 cells with replacement, and the averaged expressions of these 10 cells were calculated. The Spearman correlation was then calculated on the 20 averaged expressions of subsampled snRNA-Seq CA1Pyr data and those of subsampled Nuc-Seq CA1 data. The same procedure was repeated for averaged 20, 30, 40, 50 single neuron/nuclei.

For comparison of correlation of averaged single neuron/nuclei of CA1 pyramidal neurons and interneurons, the cells labeled as 'CA1Pyr' and 'Int' from the single neuron RNA-Seq dataset [1] were separaty subsampled, each to 100 cells to get two datasets, referred to as snRNA-Seq CA1Pyr and snRNA-Seq Int respectively. The CA1 and GABAergic nuclei from Nuc-Seq were subsampled, each to 100 nuclei to get two datasets, referred to as Nuc-Seq CA1 and Nuc-Seq Int respectively. To calculate correlations of averaged 10 single neuron/nuclei, snRNA-Seq CA1Pyr, snRNA-Seq Int, Nuc-Seq CA1, and Nuc-Seq Int were separately subsampled 20 times, each time to 10 cells with replacement, and the averaged expressions of these 10 cells were calculated. The Spearman correlation was then calculated between the 20 averaged expressions of subsampled snRNA-Seq CA1Pyr and those of snRNA-Seq Int, and between 20 averaged expressions of subsampled Nuc-Seq CA1 and those of Nuc-Seq Int. The same procedure was repeated for averaged 20, 30, 40, 50 single neuron/nuclei.

Analysis of nuclei clusters

Clustering analysis partitions nuclei into groups, such that nuclei from the same group share higher similarity than nuclei from different groups. The quality of the grouping can be measured using the Dunn index [32]

$$DB = \frac{\min_{1\leq i<j\leq n} d(i,j)}{\max_{1\leq k\leq n} d'(k)} \ ,$$

where $d(i,j)$ represents the inter-group distance between group i and j, and $d'(k)$ represents the intra-group distance of group k.

We expect that the coherent structure in transcriptomes of cells of high similarity generates observations that lie on a low-dimentional structure in the high-dimensional measurement space [33]. In this case, data points for cells belonging to the same group would lie on a continuous and smooth low-dimensional structure, and data points for cells from different groups would lie on different structures. We confine distances used in calculating the Dunn index to the low-dimensional structure and define the distance

$d'(k)$ as

$$\hat{\Phi}_{p\,q} = \underset{\Phi_{p\,q}}{\operatorname{argmin}} \max\{d_{mn} \,|\, d_{mn} \in \Phi_{p\,q}\}$$
$$d'(k) = \max\{d_{mn} \,|\, d_{mn} \in \bigcup_{p,\,q} \hat{\Phi}_{p\,q}\} \ ,$$

where p, q, m, and n are data points belonging to the group k, d_{mn} represents the pairwise distance of data points m and n, and $\Phi_{p\,q}$ represents a path connecting p and q through data points belonging to the group k. We define the distance $d(i,j)$ similarly to $d'(k)$ and confine p, q, m, and n to be data points belonging to the union of the groups i and j.

Here, we describe a pipeline of techniques to obtain nuclei clusters. We first normalize data, then we estimate false negatives as to reduce their impact on the calculation of d_{mn}. Next, we perform modified PCA and tSNE [34] to map the low-dimensional structure to a 2-D space, where d_{mn} and $\Phi_{p\,q}$ in the 2-D space represent their high-dimensional counterparts. The mapping transforms each of the low-dimensional structures to dense data clouds in the 2-D space, permitting grouping of cells by a density clustering technique [35]. This non-linear mapping is particularly useful for the data sets, in which the scales of $d'(k)$ for different cell groups are highly different and $d'(k)$ are affected by large noises in the original high dimensional space. Finally, we identify cell sub-clusters within each cell cluster by the biSNE algorithm. The PCA-tSNE, biSNE, and density clustering are applied hierachically to each cell clusters to obtain clusters at finer level. In each iteration, the Dunn index with the defined local distances $d'(k)$ can be used to evaluate the quality of the clustering assignment.

Normalization

Each library of single nuclei was prepared individually. Biases exist among libraries due to invitable differences in lysis efficiency, priming rate at RT, amplification efficiency during the initial PCR, the equalization for tagmentation, and ratios in the final sequencing pooling [36]. Although several experimental methods have been developed to mitigate biases, including, for example, adding spike-in or using unique molecular identifiers, we note that these methods would only help to reduce, at best, the amount of bias introduced after the initial PCR step, however significant amount of bias occurs before that step. We assume that cells of the same type should highly express a set of genes that are tightly regulated and exhibit small "real" intercelluar variability. An example of such a gene set includes ribosomal and cytoskeleton genes in stem cells or housekeeping genes in dentritic cells that was previously used to normalize single cell sequencing data [37]. However, there is no concensus housekeeping gene set for brain cells that consist of both mature neurons, immature neurons, and glia cells. To normalized cells, we developed a computational normalization procedure based on Bland-Altman (MA) plot and density estimation (fig. S4). For a pair of cells, our procedure normalizes one cell with respect to another so that genes belonging to this gene set are not differentially expressed on average. Using only a small set of highly expressed and lowly variable genes, as opposed to using all genes [36] or genes within the middle quantile, provides robustness against noise, because measurements of highly expressed genes are resistant to sampling noise, and lowly variable measurements unlikely to have been corrupted by large noise. In addition, small intercellular variance enables simple statistical models, such as Gaussian model, to yield good estimates. Similar reasoning underlies previously described normalized methods such as TMM [38], DESeq [39]. However, these methods are designed for population RNA-Seq data, and we empirically found that they not compatible with single cell data. A modified DESeq normalization which takes into account of massive false negatives common to single cell data did give comparable performance to our procedure.

We first discuss the case of two cells, and later we show how to generalize to a set of arbitrary size. To identify the set of genes for normalization, we first calculate differences and averages of log transformed expression level of each gene between a given pair of cells, and plot the distribution of differences by averages on an MA plot. Then, gene density in this distribution is estimated [40] and genes within the most densely plotted regions are selected. We calculate a scaling factor as the average of the log expression differences of selected genes. The second cell is normalized with respect to the first cell by dividing gene expressions of the second cell by the scaling factor.

Specifically, the log expression difference of gene j between two cells is given by

$$r_{12\_j} = \log(e_{2j}) - \log(e_{1j}) \ ,$$

and the average of log expression of gene j is given by

$$a_{12\_j} = [\,\log(e_{1j}) + \log(e_{2j})\,]/2 \ .$$

where e_{ij} denotes the the expression level of gene j in cell i. Gene j is selected into the gene set $\mathbb{S}_J$, if r_{12_j} and a_{12_j}, coordinates of gene j, are within the region having density above the top 70 percentile in the MA plot.

The scaling factor is obtained by

$$s = \sum_{j,\, j \in \mathbb{S}_J} r_{12\_j} / |\mathbb{S}_J| \ ,$$

where $|S|$ indicates the cadinality of a set S. Then the second cell is normalized as

$$e'_{2j} = e_{2j}/s \ .$$

To normalize single cells of different types, cells are first clustered into separate groups, each of which contains cells of a similar type. This step ensures that normalization complies with our assumption that cells are of the same type. Then normalization is performed for each group separately. Within each group, scaling factors are estimated for each cell with respect to multiple reference cells, which are chosen based on the number of genes detected, for example, cells having number of genes detected around the 80 percentile.

Although any particular reference cell could be affected by erroneous measurements to various degrees, using multiple reference cells reduces the effect of these errors in the normalization.

Specifically, for a given group of cells $\{\, i \mid i \in \mathbb{C}_g \text{ and } g \in \mathbb{G}\,\}$, a set of cells that have number of genes detected above 80 percentile are selected as reference cells $\{\, r \mid r \in \mathbb{C}_{gr} \text{ and } \mathbb{C}_{gr} \subset \mathbb{C}_g\,\}$. The scaling factor s_{ir} for each cell i with respect to each reference cell r is calculated. To relate s_{ir} obtained with different reference cells, we solve the optimization problem

$$\{\hat{a}_r \mid r \in \mathbb{C}_{gr}\,\} = \operatorname*{arg\,max}_{a_r,\, r \in \mathbb{C}_{gr}} \sum_{i \in \mathbb{C}_g} \operatorname*{Var}_{r \in \mathbb{C}_{gr}} \left[\log(s_{ir}) - \log(a_r)\right] \ ,$$

and scaling factors are estimated as

$$s_i = \operatorname*{median}_{r \in \mathbb{C}_{gr}} \left(\frac{s_{ir}}{\hat{a}_r}\right) \ .$$

To normalize cells from different groups, we use group scaling factors estimated for each group aggregates, which are obtained by averaging all cells within a same group. Cells from a same group are normalized using their group scaling factor.

Specifically, for each group $g \in \mathbb{G}$, the group aggregate is calculated as

$$e_{gj} = \sum_{i \in \mathbb{C}_g} e'_{ij} \ ,$$

where e_{gj} denotes the expression level of gene j in group g, and e'_{ij} is the normalized expression level of gene g in cell i. Multiple reference group aggregates are selected for the estimation of group scaling factors.

Comparison of our normalization method with TMM and DESeq

We consider a model for observed expression level e_{ij} given true expression level x_{ij},

$$e_{ij} = s_i \cdot \epsilon_{ij} \cdot x_{ij} \ .$$

where s_i represents the scaling factor of cell i, ϵ_{ij} represents the technical noise of gene j measured in cell i, and x_{ij} represents the true expression level of gene j measured in cell i. Rewrite e_{ij} on log scale,

$$\log(e_{ij}) = \log(s_i) + \log(\epsilon_{ij}) + \log(x_{ij}) \ .$$

In our normalization, the normalization factor is obtained by averaging, between cell $i1$ and $i2$, the differences in the expression of selected subset of genes $\mathbb{S}_J$.

$$\begin{aligned}\sum_{j \ \in \ \mathbb{S}_J} \big(\log(e_{1j}) - \log(e_{2j})\big)/|\mathbb{S}_J| \ &= \ \log(s_1) - \log(s_2) + \\ &\sum_{j \ \in \ \mathbb{S}_J} \big(\log(\epsilon_{1j}) - \log(\epsilon_{2j})\big)/|\mathbb{S}_J| + \\ &\sum_{j \ \in \ \mathbb{S}_J} \big(\log(x_{1j}) - \log(x_{2j})\big)/|\mathbb{S}_J| \ .\end{aligned}$$

As ϵ_{ij} for $j \ \in \ \mathbb{S}_J$ is assumed to be lognormally distributed with zero mean (modeling PCR and sampling noise), and genes within $\mathbb{S}_J$ are not differentially expressed on average, it follows that

$$\log(s_1) - \log(s_2) = \sum_{j \ \in \ \mathbb{S}_J} \big(\log(e_{1j}) - \log(e_{2j})\big)/|\mathbb{S}_J|.$$

In TMM normalization, the $\mathbb{S}_J$ is replaced by $\mathbb{S}_Q = \{j \mid e_{ij} \ \in \ [e_{qa}, e_{qb}]\}$, where e_{qa} and e_{qb} are a^{th} and b^{th} quantiles of e_{ij}. We find the assumption that $\sum_{j \ \in \ \mathbb{S}_Q}[\log(x_{1j}) - \log(x_{2j})] = 0$ might not hold true for single cell RNA-Seq data.

In DESeq normalization, e_{ij} is first normalized by its geometric mean across all cells,

$$\begin{aligned}\log(e_{ij}) - \sum_i \log(e_{ij})/|I| \ &= \ \log(s_i) - \sum_i \log(s_i)/|I| + \\ &\quad \log(\epsilon_{ij}) - \sum_i \log(\epsilon_{ij})/|I| + \\ &\quad \log(x_{ij}) - \sum_i \log(x_{ij})/|I| \ .\end{aligned}$$

Then median is taken over all genes,

$$\begin{aligned}\operatorname*{median}_j \Big(\log(e_{ij}) - \sum_i \log(e_{ij})/|I| \Big) \ &= \ \log(s_i) - \sum_i \log(s_i)/|I| + \\ &\quad \operatorname*{median}_j \Big(\log(\epsilon_{ij}) - \sum_i \log(\epsilon_{ij})/|I| \Big) + \\ &\quad \operatorname*{median}_j \Big(\log(x_{ij}) - \sum_i \log(x_{ij})/|I| \Big) \ .\end{aligned}$$

Assume that the median of ϵ_{ij} can be replaced by the mean of ϵ_{ij},

$$\begin{aligned}\operatorname*{median}_j \Big(\log(\epsilon_{ij}) - \sum_i \log(\epsilon_{ij})/|I| \Big) &= \sum_j \Big(\log(\epsilon_{ij}) - \sum_i \log(\epsilon_{ij}) \Big) /|I||J| = \\ &\quad \sum_j \log(\epsilon_{ij})/|J| - \sum_i \frac{1}{|I|} \sum_j \log(\epsilon_{ij})/|J|.\end{aligned}$$

It shows that the median of normalized e_{ij} is a good estimator for the scaling factor s_i only if

$$\sum_j \log(\epsilon_{ij}) = 0 \quad \text{and} \quad \operatorname*{median}_j \Big(\log(x_{ij}) - \sum_i \log(x_{ij})/|I| \Big) = 0.$$

However, because single cell RNA-Seq data contains substantial amount of false negative measurements, as discussed in the next section, these conditions might not hold true generally. We propose a modified DESeq normalization, which gives comparable performance to our normalization method when applied to synthetic test data.

In the modified DESeq normalization, the geometric mean and median are taken over only genes whose measured expression level $e_{ij} > 0$. This leads to

$$\begin{aligned}\operatorname*{median}_{j,\ e_{ij} \neq 0} \Big(\log(e_{ij}) - \sum_i \log(e_{ij})/|I| \Big) \ &= \ \log(s_i) - \sum_i \log(s_i)/|I| + \\ &\quad \operatorname*{median}_{j,\ e_{ij} \neq 0} \Big(\log(\epsilon_{ij}) - \sum_i \log(\epsilon_{ij})/|I| \Big) + \\ &\quad \operatorname*{median}_{j,\ e_{ij} \neq 0} \Big(\log(x_{ij}) - \sum_i \log(x_{ij})/|I| \Big) \ .\end{aligned}$$

In this formulation, the expression level ϵ_{ij} for $\{j \mid e_{ij} > 0\}$ is not subjected to false negative, and is assumed to be lognormally distributed with zero mean. Therefore, the median of ϵ_{ij} for $\{j \mid e_{ij} > 0\}$ is

$$\underset{j,\ e_{ij}\neq 0}{\text{median}} \Big(\log(\epsilon_{ij}) - \sum_i \log(\epsilon_{ij})/|I| \Big) = \sum_{j,\ e_{ij}\neq 0} \log(\epsilon_{ij})/|J| - \sum_i \frac{1}{|I|} \sum_{j,\ e_{ij}\neq 0} \log(\epsilon_{ij})/|J| = 0 \ .$$

And further assume that there exist some genes that are not differentially expressed among all cells, then the median is a robust measure to find one such gene,

$$\underset{j,\ e_{ij}>0}{\text{median}} \Big(\log(x_{ij}) - \sum_i \log(x_{ij})/|I| \Big) = 0 \ .$$

Therefore, we can obtain the scaling factor by

$$\log(s_i) - \sum_i \log(s_i) \ = \ \underset{j,\ e_{ij}>0}{\text{median}} \Big(\log(e_{ij}) - \sum_i \log(e_{ij})/|I| \Big) \ .$$

Estimation of missed detection probability

Single nuclei transcriptome libraries are amplified from extremely small input materials. As such, we expect that some transcripts that are lowly expressed will not be detected (false negatives). The probabilty of such missed detection increases for lowly expressed transcripts and lower quality libraries. Such false negatives are detrimental to various analyses. For example, they invalidate the normal distribution assumption underlying typically used Student's t-test, leaving the statistical test unjustified. In addition, false negatives confound the identification of bimodally expressed genes, such as cell type specific markers. Previous studies accounted for such false negatives by combining estimation of cell quality and gene expression [36, 41]. These methods were based on parametric estimation of gene expression distribution. However, distribution of gene expression cannot be readily fitted by a single parametric function. In contrast to these methods, we developed a Bayesian method to estimate the likelihood of an observed zero measurement being a missed detection. Our approach is based on a non-parametric estimation for gene expression distribution.

Our method is based on two observations: a) Detection rates depend on expression level. The higher a gene is expressed, the more likely it can be detected. b) Detection rates depend on library quality. Genes are more likely to be detected in libraries of high quality. We model these two observations as

- prior distributions: distributions of expression levels for each gene in cells of the same type

- sampling probabilites: detection probabilities at different expression levels for each cell

For each observed $e_{ij} = 0$ of gene j in cell i, we then estimate the posterior distribution for two mutually exclusive hypotheses that e_{ij} is a missed detection or that gene j is not expressed in cell i.

Specifically, the distribution of expression level of gene j is calculated as mixture of two distributions. The first one is the probability that gene j is not expressed

$$p_j(x = 0) = \frac{\sum_{i \in \{e_{ij}=0\}} 1}{\sum_i 1} \ ,$$

where x denotes the true expression level. The second one is a conditional distribution of expression levels of gene j given that gene j is expressed. This distribution is estimated using a KDE (kernel density estimation) based method [40] using gene expression levels e_{ij} from cells i, $\{\, i \mid e_{ij} > 0 \,\}$. Combining two parts yields

$$p_j(x) = p_j(x = 0) + [\, 1 - p_j(x = 0)\,] p_{j\_KDE}(x) \ ,$$

where x denotes the expression level.

The detection probability (1 - dropout probability) for a cell i is modeled using a geometric distribution parameterized by $\boldsymbol{\beta_i}$, as it captures the Poisson sampling process, mechanism underlying detection stochasticity

$$\begin{aligned} \Lambda(x, \boldsymbol{\beta_i}) &= 1 - e^{-t}, t = \boldsymbol{\beta_i} \begin{bmatrix} 1 \\ x \end{bmatrix} = \beta_{i0} + \beta_{i1} x \\ 0 \le \Lambda(x, \boldsymbol{\beta_i}) &\le 1 \ , \end{aligned}$$

where x denotes expression level.

Given observed data e_{ij}, the expected value of the log likehood function is given by

$$\mathrm{E}[L] = \sum_{j \in \{e_{ij} > 0\}} \log(1 \cdot \Lambda(e_{ij}, \boldsymbol{\beta_i})) + \sum_{j \in \{e_{ij} = 0\}} \sum_x p_j(x) \log(p_j(x)(1 - \Lambda(x, \boldsymbol{\beta_i}))) \ .$$

In each iteration, the log likehood function is maximized using gradient descent.

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \underset{\boldsymbol{\beta_i}}{\operatorname{argmax}} \ \mathrm{E}[L]$$

$$\frac{\partial\, \mathrm{E}[L]}{\partial \boldsymbol{\beta_i}} = \sum_{j \in \{e_{ij} > 0\}} \frac{1}{\Lambda(e_{ij}, \boldsymbol{\beta_i})} \frac{\partial \Lambda(e_{ij}, \boldsymbol{\beta_i})}{\partial \boldsymbol{\beta_i}} + \sum_{j \in \{e_{ij} = 0\}} \sum_x p_j(x) \frac{1}{1 - \Lambda(x, \boldsymbol{\beta_i})} (-1) \frac{\partial \Lambda(x, \boldsymbol{\beta_i})}{\partial \boldsymbol{\beta_i}} \ .$$

Because $\Lambda(x, \boldsymbol{\beta})$ is constrainted to be non-negative, its derivative is modified with a rectifier so that $\Lambda(x, \boldsymbol{\beta})$ is differentiable for any x,

$$\begin{aligned}
h(x) &= \frac{\log(\exp(x \cdot N) + 1)}{N}, \text{where } N \text{ is a large number} \\
\frac{\partial \Lambda}{\partial \boldsymbol{\beta}} &\approx \frac{\partial h}{\partial \Lambda} \frac{\partial \Lambda}{\partial \boldsymbol{\beta}} = \frac{1}{1 + \exp(-\Lambda(x, \boldsymbol{\beta}) \cdot N)} \cdot e^{-t} \begin{bmatrix} 1 \\ x \end{bmatrix} \ .
\end{aligned}$$

Then the distribution of expression levels are updated by

$$\begin{aligned}
p\left(e_{ij} = 0 | x_{ij} > 0\right) &= \sum_x p_j(x)(1 - \Lambda(x, \hat{\boldsymbol{\beta}}_{\boldsymbol{i}})) \\
p\left(x_{ij} = 0 | e_{ij} = 0\right) &= \frac{p_j(x_{ij} = 0)}{p_j(x_{ij} = 0) + p(e_{ij} = 0 | x_{ij} > 0)} \\
p_j(x = 0) &= \frac{\sum_{j \in \{e_{ij} = 0\}} p(x_{ij} = 0 | e_{ij} = 0)}{\sum_{j \in \{e_{ij} = 0\}} p(x_{ij} = 0 | e_{ij} = 0) + \sum_{j \in \{e_{ij} > 0\}} 1} \\
p_j(x) &= p_j(x = 0) + [1 - p_j(x = 0)] p_{j\_KDE}(x) \ ,
\end{aligned}$$

where $p\left(x_{ij} = 0 | e_{ij} = 0\right)$ denotes the probability that gene j is not expressed in cell i.

We implemented an expectation-maximization (EM) algorithm that alternates between performing an expectation step for L, and a maximization step for searching the maximizer $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ of $\mathrm{E}[L]$.

The probability $p\left(x_{ij} = 0 | e_{ij} = 0\right)$ is incorporated in calculations of summary statistics and distances to weight zero measurements. The higher the probability, the more likely that an observed zero represents a truly unexpressed gene in a cell, and the more we weight the contribution of the zero. Conversely, the lower the probability, the higher the chance that it is false negative, and the lower we weight its contribution in an analysis.

Specifically, we weight summary statistics, Euclidean distance, Pearson correlation coefficient, and cosine similarity in the following ways.

I. the weighted gene expression mean:

$$u_j = \sum_i e_{ij} w_{ij} / \sum_i w_{ij} \ ,$$

where

$$w_{ij} = \begin{cases} p\left(x_{ij} = 0 | e_{ij} = 0\right) & \text{if } e_{ij} = 0 \\ 1 & \text{if } e_{ij} > 0 \ . \end{cases}$$

II. the weighted Euclidean distance between two cells x, y:

$$\begin{aligned} w_j &= w_{xj} w_{yj} \\ d_{xy} &= \frac{\sum_j (e_{xj} - e_{yj})^2 w_j}{\sum_j w_j} \; . \end{aligned}$$

III. the weigthed Pearson correlation coefficient between two cells x, y:

$$\begin{aligned} \hat{\mathbf{e}}_{\mathbf{x}} &= \mathbf{e}_{\mathbf{x}} - u_x, & \hat{\mathbf{e}}_{\mathbf{y}} &= \mathbf{e}_{\mathbf{y}} - u_y & & \\ S_{xy} &= \sum_j \hat{e}_{xj} \hat{e}_{yj} w_j, & S_{xx} &= \sum_j \hat{e}_{xj}^2 w_j, & S_{yy} &= \sum_j \hat{e}_{yj}^2 w_j \\ \rho_{xy} &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \; . & & & & \end{aligned}$$

IV. the weighted cosine similarity is calculated in a similar way except no data centering.

V. the weighted Euclidean distance between two cells x, y under a linear transformation of linear combinations of genes, $Y = XA$, where X is an $i \times j$ matrix, and A is a $j \times k$ transformation matrix, is given by

$$\begin{aligned} w_j &= w_{xj} w_{yj} \\ d_{xy} &= \sum_k \big(\frac{\sum_j a_{jk} (e_{xj} - e_{yj}) w_j}{\sum_j w_j} \big)^2 \; . \end{aligned}$$

VI. the weighted Pearson correlation coefficient between two cells x, y under a linear transformation of linear combinations of genes as above is given by

$$\begin{aligned} u_x &= \frac{1}{|K|} \sum_k \frac{\sum_j a_{jk} e_{xj} w_j}{\sum_j w_j}, & u_y &= \frac{1}{|K|} \sum_k \frac{\sum_j a_{jk} e_{yj} w_j}{\sum_j w_j} & & \\ \hat{e}_{xk} &= \frac{\sum_j a_{jk} e_{xj} w_j}{\sum_j w_j} - u_x, & \hat{e}_{yk} &= \frac{\sum_j a_{jk} e_{yj} w_j}{\sum_j w_j} - u_y & & \\ S_{xy} &= \sum_k \hat{e}_{xk} \hat{e}_{yk}, & S_{xx} &= \sum_k \hat{e}_{xk}^2, & S_{yy} &= \sum_k \hat{e}_{yk}^2 \\ \rho_{xy} &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \; . & & & & \end{aligned}$$

VII. the weighted cosine similarity is calculated similarly as the weighted correlation coefficient except no data centering.

VIII. the weighted covariance between two genes under a linear transformation of linear combinations of genes as above is given by

$$u_{xk} = \frac{\sum_i \sum_j a_{jk}\hat{e}_{ij}w_{ij}}{\sum_i \sum_j w_{ij}} = \frac{\sum_j a_{jk} \sum_i \hat{e}_{ij}w_{ij}}{\sum_i \sum_j w_{ij}} = 0,\ \hat{e} \text{ is centered along } i$$

$$cov(k, k') = \frac{\sum_i(\sum_j a_{jk}\hat{e}_{ij}w_{ij})(\sum_j a_{jk'}\hat{e}_{ij}w_{ij})}{\sum_i(\sum_j w_{ij})^2}\ .$$

PCA and tSNE

To project cells to two dimensional space, we first perform principal component analysis (PCA) to project the original data to reduced linear dimensions, where most significant variance of the data is preserved as determined based on the largest eigenvalue gap. We then calculate the cosine distance of cells on the PCA reduced dimensional space. Finally, we use t-distributed Stochastic Neighbor Embedding (tSNE) [34, 42, 43] with the cosine distance to further map cells to two dimensions, where Euclidean distances of closely projected cells represent their cosine distances.

The cosine distance depends on the angle between two vectors defined by gene expressions in the high dimensional space. It is preferred in our analysis over Euclidean distance and correlation distance, because it is more robust to noise than Euclidean distance and it is invariant under rotational transformations, such as PCA.

I. weighted PCA

The PCA analysis is performed commonly using singular value decomposition (SVD) or eigenvalue decomposition (EVD) on the covariance matrix, which scales quadratically with the number of genes. Given the large number of genes, more than 25,000, in our data, it is computational costly to directly perform SVD or EVD on the large covariance matrix. In order to get principal components, or the transformation matrix A, while accounting for weights, we first center the original data matrix E across genes to get $\hat{E}$, where e_{ij} is the expression level of gene j in cell i. Next, we perform SVD on centered data matrix $\hat{E}$ to get A^*. We calculate the weighted covariance matrix C_w on $\hat{E}$ under the linear transformation defined by the matrix A^*. We then perform SVD or EVD on C_w to get A.

II. tSNE with cosine distance

We modified the original tSNE to allow dimensionality reduction based on a weighted cosine similarity. The original tSNE technique projects data in a non-linear way to low dimensional space, such that Euclidean distances between neighboring data points in the low dimensional space overall represent distances between these neighboring data points, or local distances, in the high dimensional space. The input to tSNE is a distance matrix, describing all pairwise distances in the high dimensional space. In order to apply tSNE, we first transform the weighted cosine similarity to cosine distance by exploring relationships between the two measures on the closest data points.

Specifically, given a cell and its gene expression measurements denoted by a n dimensional vector $\mathbf{x}$, the measurements of its neighbor $\mathbf{y}$ is modeled as

$$\mathbf{y} = k\,(\mathbf{x} + \mathbf{d}) \ ,$$

where k is a scaling factor and $\mathbf{d}$ denotes the distance between $\mathbf{x}$ and $\mathbf{y}$. Under the null hypothesis that $\mathbf{x}$ and $\mathbf{y}$ are measured from two cells of the same type, $\mathbf{d}$ is drawn from a Gaussian distribution with zero mean and variance σ. Our goal is to estimate the distance magnitude $|\mathbf{d}|$, given the measured angle ϕ between $\mathbf{x}$ and $\mathbf{y}$.

Geometrically, the vector $\mathbf{d}$ lies on a hypersphere defined by radius $|\mathbf{d}|$. The volume and surface area of a hypersphere of dimension n (n-sphere) has the following properties

$$\begin{aligned} S_n &= (n+1)V_{n+1} \\ \mathrm{d}S_n &= (n+1)\mathrm{d}V_{n+1} \end{aligned}$$

the volume element is

$$\begin{aligned} \mathrm{d}V_{n+1} &= |\mathbf{d}|^n sin^{n-1}(\phi_1) sin^{n-2}(\phi_2) \cdots sin(\phi_{n-1})\,\mathrm{d}|\mathbf{d}|\,\mathrm{d}\phi_1\,\mathrm{d}\phi_2\,\cdots\,\mathrm{d}\phi_n \\ &= sin^{n-1}(\phi_1)\,\mathrm{d}\phi_1\ \cdot g(|\mathbf{d}|, \phi_2, \ldots, \phi_n) \\ \mathrm{d}S_n &= sin^{n-1}(\phi_1)\,\mathrm{d}\phi_1\ \cdot (n+1) g(|\mathbf{d}|, \phi_2, \ldots, \phi_n). \end{aligned}$$

The probability of drawing $\mathbf{d}$ in a n-sphere of radius $|\mathbf{d}|$ with an angle ϕ from $\mathbf{x}$ scales as $sin^{n-1}(\phi)$.

When n is large, most of $\mathbf{d}$ lie perpendicular to $\mathbf{x}$, thus there exists a unique mapping between $|\mathbf{d}|$ and ϕ.

$$cos(\phi) = \frac{1}{\sqrt{(1+|\mathbf{d}|)^2+1}}$$
$$|\mathbf{d}| = \sqrt{\frac{1}{\cos^2(\phi)} - 1}$$

Differential gene expression and pathway analysis

We use an adjusted Welch's t-test for identifying differentially expressed genes. We applied weights in the calculation of summary statistics, such as sample mean, sample variance, and effective degrees of freedom, used in Welch's t-test.

Specifically, to find the significance level of gene j between cells in group $\mathbb{X}$ and cells in group $\mathbb{Y}$,

$$\begin{aligned} n_{xj} &= \sum_{i,\, i\in\mathbb{X}} w_{ij} \ , & n_{yj} &= \sum_{i,\, i\in\mathbb{Y}} w_{ij} \ , \\ u_{xj} &= \sum_{i,\, i\in\mathbb{X}} e_{ij} w_{ij}/n_{xj} \ , & u_{yj} &= \sum_{i,\, i\in\mathbb{Y}} e_{ij} w_{ij}/n_{yj} \ , \\ S_{xj} &= \sum_{i,\, i\in\mathbb{X}} (e_{ij}-u_{xj})^2 w_{ij}/(n_{xj}-1) \ , & S_{yj} &= \sum_{i,\, i\in\mathbb{Y}} (e_{ij}-u_{yj})^2 w_{ij}/(n_{yj}-1) \ , \end{aligned}$$

$$\begin{aligned} \text{t statistic} \qquad t_j &= \frac{u_{xj}-u_{yj}}{\sqrt{\frac{S_{xj}}{n_{xj}} + \frac{S_{yj}}{n_{yj}}}} \ , \\ \text{degrees of freedom} \qquad v_j &\approx \frac{(S_{xj}/n_{xj} + S_{yj}/n_{yj})^2}{S_{xj}^2/[n_{xj}^2(n_{xj}-1)] + S_{yj}^2/[n_{yj}^2(n_{yj}-1)]} . \end{aligned}$$

The false discovery rate (FDR) is calculated for each differentially expressed gene in multiple hypothesis testing using the Benjamini and Hochberg procedure [44].

Density clustering and selection of the number of clusters

We used a density based clustering method [35] to partition cells embedded in the 2-D space. The method searches cluster centers that are characterized by two quantities: (1) high local density ρ_i and (2) large distance δ_i from points of higher density, which are centers of other clusters. We unify the two quantities into a single metric by taking the product of the two quantities, $s_i = \rho_i \cdot \delta_i$.

To select cluster centers, we rank each data points by their s_i in descending order. For a given n, the number of desired clusters, we select the top ranked n cluster centers, and perform the cluster assignment as described previously [35]. To evaluate the quality of the clustering, we calculate the Dunn index for each n with $d(i, j)$ and $d'(k)$ defined as local distances. The calculation of the Dunn index can be operated in $O(N^3)$, where N is the number of total data points.

Algorithm: Identification of maximum steps on shortest paths (MaxStep)

```
Input: pairwise distance of data points ($D$)
Output: the pairwise shortest link ($D'$)
$D' \leftarrow D$
n← # of data points
for $k \leftarrow 1$ to $n$ do
    for $i \leftarrow 1$ to $n-1$ do
        for $j \leftarrow i+1$ to $n$ do
            $D'(i,j) \leftarrow \min(D'(i,j), \max(D'(i,k), D'(k,j)))$
        end
    end
end
return $D'$
```

Algorithm: Calculation of the Dunn index defined on local distances (DunnLocal)

```
Input: pairwise distance of data points in the 2-D embedding (D), clustering assignment (Cl)
Output: the Dunn index ($\theta$)
cl_uiq← unique(Cl)
n← # of cl_uiq
$d'_k \leftarrow$ empty array with a length of n
$d_{ij} \leftarrow$ empty matrix with a size of (n, n)
for $i \leftarrow 1$ to $n$ do
    ii← index of data whose clustering assignment is cl_uiq(i)
    $d'_k(i) \leftarrow$ max(MaxStep(D(ii,ii)))
end
for $i \leftarrow 1$ to $n-1$ do
    for $j \leftarrow i+1$ to $n$ do
        ii← index of data whose clustering assignment is either cl_uiq(i) or cl_uiq(j)
        $d_{ij}(i,j) \leftarrow$ max(MaxStep(D(ii,ii)))
    end
end
$\theta \leftarrow \min(d_{ij}) / \max(d'_k)$
return $\theta$
```

Large scale comparison between RNA-Seq data and ISH data

We selected genes differentially expressed between any bipartition of DG, CA1, CA2, CA3 clusters in RNA-Seq data. For example, a gene is selected if it is differentially expressed between cells in a combined DG and CA2 cluster, and cells in a combined CA1 and CA3 cluster.

Specifically, the differential expression was tested using the adjusted t-test between cells $\in \mathbb{C}_1$, $\mathbb{C}_1 \subset \{\text{DG, CA1, CA2, CA3}\}$ and cells $\in \mathbb{C}_2$, $\mathbb{C}_2 = \{\text{DG, CA1, CA2, CA3}\} \setminus \mathbb{C}_1$. Gene j is selected if

$$\begin{aligned} \text{difference in mean} \quad & m_{\mathbb{C}_1 j} - m_{\mathbb{C}_2 j} > 1 \\ \text{mean of cells} \in \mathbb{C}_1 \quad & m_{\mathbb{C}_1 j} > 20 \text{ TPM} \\ \text{mean of cells} \in \mathbb{C}_2 \quad & m_{\mathbb{C}_2 j} < 5 \text{ TPM} \\ p \text{ value of t-test} \quad & p_j < 0.01 \ . \end{aligned}$$

The quantified ISH data [10] with 200µm resolution was downloaded from Allen Brain Atlas (Website: 2015 Allen Institute for Brain Science. Allen Mouse Brain Atlas [Internet]. Available from: http://mouse.brain-map.org.) Mean expression level of ISH data was calculated as averaged energy level for each of the DG, CA1, CA2, CA3 regions. Specifically, averaged energy level $e_{\mathbb{G}}$ for grids in a region $\mathbb{G}$ is given by

$$e_{\mathbb{G}} = \sum_{g, g \in \mathbb{G}} d_g \cdot i_g / |\mathbb{G}| \ ,$$

where d_g is the quantified expression density for grid g, and i_g is the quantified expression intensity for grid g. The Indices for DG, CA1, CA2, CA3 regions are 726, 382, 423, 463. We obtained two vectors $\mathbf{e} \in \mathbb{R}^4$ comprising averaged expression levels of DG, CA1, CA2, CA3 regions for each gene, one from RNA-Seq data, and another from ISH data. Pearson correlation coefficient was calculated between these two vectors for each selected gene.

BiSNE

Cells positioned in proximity in the tSNE mapping coexpress a set of genes that are not expressed by distal cells. These set of genes could be used to distinguish different cell subpopulations. These genes are coexpressed in the cells grouped in proximity, and therefore they have localized expression patterns in the tSNE mapping.

Statistics for scoring expression patterns

Motivated by this observation, we use two different statistics to identify genes with significantly localized expression patterns in the tSNE mapping and then perform PCA-tSNE using the union of these identified genes to cluster cells.

I. Moran's I

Moran's I [45] scores correlation between a measurement on a set of mapping positions and pairwise distances of these mapping positions. Given tSNE coordinates, the Moran's I for gene k is given by

$$I(k) = \frac{\sum_i \sum_j Q_{ij}(e_{ik} - u_k)(e_{jk} - u_k) w_{ik} w_{jk} / \sum_i \sum_j Q_{ij} w_{ik} w_{jk}}{\sum_i (e_{ik} - u_k)^2 w_{ik} / \sum_i w_{ik}} \ ,$$

where Q_{ij} denotes the pairwise similarity transformed from d_{ij}, the Euclidean distances between cell i and j in the tSNE mapping. We obtain Q_{ij} from d_{ij} using the Gaussian function,

$$Q_{ij} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \ .$$

We choose σ to set the minimal size of localized expressed pattern, as $d_{ij} \approx \sigma$ weights around 60% and $d_{ij} \approx 2\sigma$ weights around 13.5%.

The statistical significance of the pattern of gene k is tested by converting $I(k)$ to a z score,

$$\begin{aligned}
&E[I] = -1/(N-1), \qquad \text{where } N \text{ is the length of } \mathbf{e_k} \\
&V[I] = \frac{1}{S_0^2(N^2-1)(N^2 S_1 - N S_2 + 3S_0^2)} - E[I]^2 \\
&S_0 = 2\sum_i \sum_j Q_{ij}, \qquad S_1 = 2\sum_i \sum_j Q_{ij}^2, \qquad S_2 = 4\sum_i (\sum_j Q_{ij})^2 \\
&z = -\frac{I - E[I]}{\sqrt{V[I]}} \ .
\end{aligned}$$

Moran's I uses gene expression levels in its calculation. When identifying marker genes, only the information about whether a gene is expressed or not is necessary. We use a modified Moran's I on binarized gene expression levels.

Specifically, we binarize gene expression level by a threshold,

$$\hat{e}_{ij} = \begin{cases} 1 & \text{if } e_{ij} > 3\,\text{TPM} \\ 0 & \text{if } e_{ij} \leq 3\,\text{TPM} \ . \end{cases}$$

We then calculate the modified Moran's I by,

$$I(k) = \frac{\sum_i \sum_j Q_{ij}(\hat{e}_{ik} - \hat{u}_k)(\hat{e}_{jk} - \hat{u}_k)w_{ik}w_{jk}}{\sum_i \sum_j Q_{ij}w_{ik}w_{jk}} \ .$$

Moran's I is a global measure. It has biases towards genes that are widely expressed. To reduce false positives, we filtered out genes expressed in more than 80% of cells.

II. Manhattan distance and order statistics

The manhattan distance is an alternative to the Euclidean distance in quantifying proximity. The advantage of using mahattan distance is that x and y coordinates can be tested independently using order statistics. Assume a given set of cells that express gene j and their positions $\mathbf{z}$ on a coordinate z, $\bar{\mathbf{z}}$ is defined as the normalized $\mathbf{z}$ such that $\bar{z}_i = (z_i - min(\mathbf{z}))/(max(\mathbf{z}) - min(\mathbf{z}))$, $i \in \{i \mid e_{ij} > \text{TPM } 3\}$, and $\hat{\mathbf{z}}$ is defined as the ordered list of $\bar{\mathbf{z}}$, such that $\hat{z}_i < \hat{z}_{i+1}$. The range w_z is defined as $w_z = \hat{z}_n - \hat{z}_1$. Assume that $\hat{z}$ is a vector of i.i.d. samples from a uniform distribution, the significance level p of w_z can be found using order statistics

$$\begin{array}{ll} \text{PDF of } w & f_w(w) = n(n-1)\displaystyle\int_{-\infty}^{+\infty} [F(x+w) - F(w)]^{(n-2)} f(x) f(x+w)\mathrm{d}x \\ & \text{where } f(x) \text{ and } F(x) \text{ are PDF and CDF of z} \\ \text{CDF of } w & F_w(w) = w^{n-1}[n - (n-1)w], \qquad \text{under null hypothesis} \\ & p_z = F_w(w_z) \ , \end{array}$$

where PDF is the probability density function and CDF is the cumulative density function. To robustly estimate w in the presence of outliers, the distribution of $\mathbf{z}$ is fitted using the Gaussian distribution with robust estimators of mean and variance [46].

$$\begin{aligned} u_z &= \text{median}(\mathbf{z}) \\ S_z &= 1.1926 \ \underset{i}{\text{median}}(\underset{j}{\text{median}}(|z_i - z_j|)) \\ p_i &= \Phi(-|\frac{z_i - u_z}{S_z}|) \ , \end{aligned}$$

where Φ denotes the CDF of the standard normal distribution. Samples with $p_i < \epsilon$, a predefined threshold, are considered outliers and are excluded from the estimation of w.

A single p value is calculated for each gene by taking the product of p_x and p_y, the p values obtained for x and y coordinates, respectively. It measures the overall significance level of each gene in both coordinates.

Selection of significant genes

For each statistic, we rank genes based on their significance. Genes ranked high are likely to be informative for clustering cells, whereas genes ranked low are more likely to be noises that suppress clustering separation. We use a cut off rank to select informative genes, chosen based on the statistic of eigenvalues of random matrices [47], which states that inclusion of a noisy row (gene) in a data matrix would lead to a reduction in the maximum eigenvalue gap of the matrix. Conversely, inclusion of an informative row(gene) would lead to an increase in the maximum eigenvalue gap, as the variance it introduces aligns with variances of some other genes. Therefore, the change in the maximum eigenvalue gap measures the extent to which a gene is informative. After genes are ranked, we start with a data matrix containing the top ranked genes, and add subsequent genes with lower rank incrementally. For each addition, we calculate the change in the maximum eigenvalue gap before and after adding the gene. Additionally, we randomly permute measurements of this gene across cells and calculate the change in the maximum eigenvalue gap induced by adding this permuted gene. We then select a cut off rank, below which there is no difference in the change of the maximum eigenvalue gap between adding a gene or its permuted counterpart. The selection cut-off can also be formally tested using minimum hypergeometric test [48].

Specifically, for a data matrix $E_{1,j-1}$ and a gene j, We form a new matrix

$$E_{1,j} = \begin{bmatrix} E_{1,j-1} \\ \mathbf{e_j} \ , \end{bmatrix}$$

and we obtain the eigenvalues of $E_{1,j-1}E_{1,j-1}^T$ using weighted SVD. The eigenvalues are normalized and sorted in order

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n \text{ , and } \sum_i \lambda_i = 1 \ .$$

The distribution density (Marchenko-Pastur distribution) of higher order eigenvalues can be approximated by a linear function [47], and its cumulative distribution can be approximated by a quadratic polynomial. The sorted eigenvalues follow the inverse function of the cumulative distribution, and are fitted by

$$\hat{\lambda}_i = f(i) = \alpha_0 + \alpha_1 \sqrt{\frac{i}{n}} \text{ , } \alpha_0 \text{ and } \alpha_1 \in \mathbb{R} \ .$$

The eigenvalue gap is approximated as

$$\Delta_j = \sum_{i=1}^{n} (\lambda_i - \hat{\lambda}_i) \ .$$

For permutation comparision, expression of gene j is permuted,

$$\begin{aligned} &\tilde{\mathbf{e}}_\mathbf{j} : \tilde{e}_{ij} = e_{i'j} \text{ , } i' \text{ is drawn without replacement from } [1, n] \\ &\tilde{E}_{1,j} = \begin{bmatrix} E_{1,j-1} \\ \tilde{\mathbf{e}}_\mathbf{j} \end{bmatrix} \ , \end{aligned}$$

where i' denotes randomly permuted cell index. The eigenvalue gap $\tilde{\Delta}_j$ is obtained for the permuted matrix $\tilde{E}_{1,j}$. A cut off rank is chosen at k, if the change in the eigenvalue gap $\Delta' - \tilde{\Delta}'$ is not significant for genes ranked below k. To combine top genes, we take the union of genes selected by different statistics.

Clustering of gene signatures using cross correlation

To cluster genes into gene signatures while taking into account of the similarity between cells expressing these genes, we compute cross correlations between high scoring genes while taking account of the proximity of cells expressing these genes, convert the correlation coefficient to distances, and cluster these genes using t-SNE and density clustering.

Specifcally, spatial cross correlation between gene k and k' is given by

$$I(k, k') = \frac{\sum_i \sum_j Q_{ij}(e_{ik} - u_k)(e_{jk'} - u_k) w_{ik} w_{jk'} / \sum_i \sum_j Q_{ij} w_{ik} w_{jk'}}{\sqrt{(\sum_i (e_{ik} - u_{k'})^2 w_{ik} / \sum_i w_{ik})(\sum_i (e_{ik'} - u_{k'})^2 w_{ik'} / \sum_i w_{ik'})}} \ .$$

It has been noted that the range of I is not $[-1, 1]$, unlike Pearson's correlation coefficient. We empirically found that I is positively biased in the tSNE mapping. The positive bias may underestimate the strength of anti-correlation genes having complementary patterns. A scalar transformation of I that has the exact range $[-1, 1]$ has been proposed [49].

$$\tilde{W} = (n\bar{w})^{-1} H^T W H, \qquad \text{where } \tilde{W} \text{ is a } (n-1) \times (n-1) \text{ matrix, and } \bar{w} = \sum_{i,j=1}^{n} w_i j / n^2$$

$$H = (\mathbf{h}_1, \dots, \mathbf{h}_{n-1}) \text{ is defined based on Helmert orthogonal matrix}$$

$$\mathbf{h}_i^T = (\mathbf{1}_i^T, -i, \mathbf{0}_{n-i-1}^T) / \sqrt{i(i+1)}, \qquad \text{for } i = 1, \dots, n-1 \ .$$

The scalar transformation of Moran's I is given by

$$I_M = \begin{cases} [(n-1)I + 1]/[|(n-1)\lambda_{(1)} + 1|] & \text{if } (n-1)I + 1 < 0 \\ [(n-1)I + 1]/[(n-1)\lambda_{(n-1)} + 1] & \text{if } (n-1)I + 1 \geq 0 \ , \end{cases}$$

where $\lambda_{(1)}$ and $\lambda_{(n-1)}$ are the smallest and largest eigenvalues of the matrix $\tilde{W}$.

The calculation of spatial cross correlation has a computational complexity that scales quadratically with the number of gene and cells as of $O(N^2M^2)$, where N is the number of cells and M is the number of genes. When the number of cells and the number of genes are large, it becomes inpractical to calculate the spatial cross correlation. However, for clustering genes using tSNE [50], only the information about k nearest neighor (knn) data points is necessary, requiring a linear complexity as of $O(N^2MK)$. The data with knn defined on a metric space can be organized using structures such as vantage point (VP) tree [51] for efficient computation. We develop a conversion between spatial correlation coefficient and a metric.

Theorem *For a given similarity $I(k, k')$, $I(k, k') \in [-B, B]$, $B \in \mathbb{R}$ and $B > 0$, define $g(k, k')$*

$$g(k,k') = \begin{cases} 0 & \textit{if } k = k' \\ \sqrt{a - I(k,k')} & \textit{if } k \neq k' \end{cases}$$

with $a > \frac{5}{3}B$, and $g(I(k, k'))$ is a metric.

Proof: For $k = k'$, the proof is trivial. For $k \neq k'$,

1. non-negativity, $g(k, k') = \sqrt{a - I(k, k')} > \sqrt{\frac{2}{3}B} > 0$
2. coincidence, $g(k, k') = \sqrt{a - I(k, k')} > 0$
3. symmetry, $g(k, k') = \sqrt{a - I(k, k')} = \sqrt{a - I(k', k)} = g(k', k)$
4. triangle inequality, $g(k, k'') + g(k'', k') \geq 2\sqrt{a - B} > \sqrt{a - (-B)} > g(k, k')$

□

Selection of principal components

We choose top principal components (PCs) based on the largest Eigen value gap. We used top 15 PCs for all cells, top 11 PCs for glial cells, top 13 PCs for DG granule cells, top 7 PCs and top 4 PCs before and after biSNE feature selection for GABAergic cells, top 3 PCs and top 4 PCs before and after biSNE feature selection for CA1 pyramidal cells, top 2 PCs and top 5 PCs before and after biSNE feature selection for CA3 pyramidal cells, and top 2 PCs for immature neuronal cells.

Comparison of biSNE and generalized linear model

We used an in-house implemented generalized linear model (GLM) [52, 53] to select highly variable genes in the GABAergic nuclei data. Three different set of genes were chosen based on three significance levels. PCA-tSNE embeddings were performed on the nuclei data using each of the chosen sets of genes. The cluster assignments were obtained on the PCA-tSNE embedding that corresponds to the most stringent significance level. We used biSNE to select three sets of correlated highly variable genes in the same nuclei data. Each set contains the same number of genes as that in the corresponding set selected by GLM. PCA-tSNE embedding and the cluster assignments were performed using each set of genes.

Validations of glia sub-types expression signatures

Differentially expressed marker genes were found for each of the glia sub-clusters and for the neuronal clusters. Differential genes were averaged across each glia cluster and averaged across all neuronal clusters combined. Spearman correlation was calculated between these average expression patterns and cell type specific bulk RNA-Seq performed in the cerebral cortex [11]. The published dataset was log transformed.

Identification of nuclei identity based on a single marker gene

We performed *in silico* cell sorting based on *Pvalb* expression, and found that the sorted cells constitute a subset of the identified *Pvalb* interneurons. This demonstrates that cell type identification based on the expression level of a single marker gene can suffer from false negatives, if only because of “drop outs” in single cell RNA-Seq or Nuc-Seq. Fortunately, the *Pvalb* expressing interneurons also share similarity in the expression of many other genes, enabling the recovery of genes commonly expressed by Pvalb interneurons, providing a robust way to determine cell types.

Localization of subclusters to anatomical regions

Localizing subclusters requires a spatial reference map of a few landmark genes [24] and the expression level of these landmark genes in each subcluster. We first created a spatial reference map by dividing an anatomical region into a grid. We manually scored the expression levels of known landmark genes [10] in this grid as not expressed, weakly, or highly expressed in these grids. Next, we generated for each subcluster a "landmark profile" by the percentage of cells expressing each landmark in this subcluster. We developed an approach similar to Seurat [24] to infer whether a given landmark gene is expressed in each cell by exploiting information from all non-landmark genes. The technique leverages the fact that many genes that are co-regulated with the landmark genes are measured in Nuc-Seq and that their expression pattern contains information about landmark genes [24]. Our anatomical alignment method is similar to Seurat in concept. Unlike Seurat, however, our method can accommodate situations when far fewer landmark genes are available (a common situation in many system unlike the heavily-studied zebrafish embryo, on which Seurat was demonstrated). We calcualted the percentage of inferred expressing cells in each subcluster. To relate the subclusters to the reference map, we evaluted the correlation between each subcluster's landmark profile and the profile of landmark genes in each part of the reference map. we positioned each of the subclusters to the highly correlated parts of the map. The accuracy of this spatial mapping is dependent on the quality of ISH images of landmark genes from the Allan brain atlas.

The selected landmark genes for CA1 region are *Nov*, *Ndst4*, *Dcn*, *Gpc3*, *Zbtb20*, *Calb1*, *Prss12*, *Wfs1*, *Col5a1*, *Grp*, *Gpr101*. The selected landmark genes for CA3 region are *Kcnq5*, *Kctd4*, *Ttn*, *Rph3a*, *Mas1*, *Plagl1*, *Col6a1*, *Prkcd*, *Loxl1*, *Grp*, *Ptgs2*, *Dkk3*, *St18*, *Mylk*.

We used a supervised machine learning algorithm to fit and binarize expression of marker genes. To obtain a traning data set for a given marker gene j, we ranked subclusters by weighted mean expression of the marker gene, and select cells expressing the marker gene above TPM 8 in the top ranked three subclusters as positive training samples. We selected cells not expressing or lowly (less than TPM 3) expressing the marker gene in the bottom ranked three subsclusters as negative training samples.

Specifically, we use all genes except marker genes as feature data $\mathbf{z}$ in an L1-regularized L2-loss support vector machine

$$z_{ik} = \begin{cases} 1 - p\left(x_{ik} = 0 | e_{ik} = 0\right) & \text{if } e_{ik} = 0 \\ 1 & \text{if } e_{ik} > 0 \end{cases}$$

$$y_{ij} = \begin{cases} 0 & \text{if } i \in \text{negative training samples} \\ 1 & \text{if } i \in \text{positive training samples} \end{cases},$$

where $k \notin$ markers, and $i \in$ training cells. We solved the unconstrained optimization problem using liblinear package [54]

$$\min_{w_j} \|\mathbf{w_j}\|_1 + C\sum_{i=1}^{l}(\max(0, 1 - y_{ij}\mathbf{w_j}^T\mathbf{z_i}^T))^2$$

where C denotes the penalty parameter. We performed coarse search followed by fine search using 5 fold cross validation for parameter C that yielded the best accuracy for the training data.

To predict whether the marker gene is expressed in cells not included in the training samples, we used the decision function

$$\hat{y}_{ij} = sgn(\mathbf{w_j}^T\mathbf{z_i}^T) \ .$$

The fraction of cells expressing marker gene j in a subcluster $\mathbb{C}$ is given by

$$f_{\mathbb{C}j} = \sum_{i,\ i\in\mathbb{C}} \hat{y}_{ij}/|\mathbb{C}| \ .$$

We predicted expression of all the marker genes in this way and calculate Pearson correlation coefficient between subclusters and subregions using $\mathbf{f}_{\mathbb{C}}$ and manually quantified expression intensity.

To test whether the subclusters were driven by the selected landmark genes, we excluded the landmark genes from PCA-tSNE and biSNE steps, and repeated the clustering. We consistently obtained the same clustering.

Indexing cells along a trajectory on projected continuum

To obtain the ranking of cells along a given trajectory, we treat the indexing as a traveling salesman problem (TSP). Cells at the start and the end points of a given trajectory are manually selected. The Euclidean distances between cells on the projected space are calculated, and normalized to integers

$$\hat{d} = \lceil 10\ d/\min(d)\rceil$$

The distance between start and points is set to 0. The normalized distances are used in Lin-Kernighan heuristic (LKH) solver [55, 56] for TSP. The obtained ordering of cells is shifted, so that the manually selected start cell is numbered as the first.

Single molecule *in situ* hybridization tissue assay

For double fluorescent *in situ* hybridization (dFISH) assay, Mice were sacrificed by a lethal dose of Ketamine/Xylazine 2 weeks post EdU injection, and transcardially perfused with PBS. Brain samples were immediately frozen in tissue freezing medium (O.C.T.) and kept in -80°C overnight. Coronal sections were cut at 15µm at -15°C. dFISH assay on O.C.T. embedded sections was performed according to Affymetrix provided protocol for O.C.T. samples, which combines QuantiGene ViewRNA ISH Tissue 2-plex Assay Kit (Affymetrix, #QVT0012) and ViewRNA ISH Cell Assay Kit (Affymetrix, #QVCM0001). Proprietary probes designed for *Calb2*, *Htr3a*, *Vip*, *Pvalb*, *Penk*, and *Oprd1* were purchased from the vendor (Affymetrix) and used.

Images were taken using fluorescent microscopy (Zeiss microscope and Hamamatsu camera C11440-22CU) and were processed in Matlab. Image background due to non-uniform illumination was removed using Matlab function strel('disk',25). The image brightness and contrast were adjusted to obtain the maximum dynamic range.

EdU labeling for staining

Labeling of proliferating cells for staining in mice was performed by intraperitoneal (i.p.) injection of EdU (5-ethynyl-2′-deoxyuridine) (Thermo Fisher Scientific, #A10044) at a dose of 100mg/kg. Mice were sacrificed by a lethal dose of Ketamine/Xylazine 2 weeks post EdU injection, and transcardially perfused with PBS followed by 4% PFA. Brain coronal sections of 30 µm were cut using vibratome (Leica, VT1000S). Sections were washed twice in PBST with 3% BSA, permeabilized in PBS with 0.5% Triton X-100 for 20 min, and washed three times in PBST with 3% BSA. EdU staining was performed using Click-iT Edu Imaging Kit (Thermo Fisher Scientific, #C10086) according to the manufacturer's protocol. Briefly, Click-iT reaction mix was prepared as follows: 100µl Click-iT reaction buffer, 800µl $CuSO_4$, 100µl 1X Click-iT reaction buffer additive, and Alexa Fluor 488 azide. Sections were incubated with 0.5ml reaction mix in 6 well plate for 30 min at room temperature covered in dark. Sections were washed twice in PBS 3% BSA post reaction, followed by mounting and imaging.

Div-Seq

Labeling proliferating cells in mice for Div-Seq was performed by intraperitoneal (i.p.) injection of EdU at a dose of 100 mg/kg. Fresh tissue was microdissected into RNA-later as described above. 24 hours after dissection nuclei were isolated as described above and resuspended in 100 µl resuspension buffer (with RNAse inhibitor), filtered and transferred to a 15 ml tube. EdU staining was performed immediately using Click-iT Edu Flow Cytometry assay Kit (Thermo Fisher Scientific, #C10086), 500 µl reaction buffer was added directly to the resuspension buffer (mix is made following the manufacturer's protocol), mixed well and left in RT for 30min; 3-5ml of 1% BSA PBS wash solution was added to the resuspended nuclei and mixed well, then nuclei were spun down for 5min at 500g in 4°C, buffer was removed and nuclei were resuspended in 400-700 µl resuspension buffer with ruby-dye (1:800) and FACS sorted immediately.

For DG and SC samples, mice were injected with a single dose of EdU, or three doses of EdU with a 12 hour interval, and then sacrificed at 2,4,6,7 and 14 days post EdU injection). For OB samples, mice were injected with three doses of EdU with 12 hour intervals, and then sacrificed at 7 days post EdU injection.

Clustering of adult newborn cells and reconstructing pseudotime along the maturation trajectory

To place the DG neurons on the maturation trajectory, nuclei were clustered iteratively. EdU labeled nuclei were clustered together with non-EdU labeled nuclei to identify nuclei of the neuronal lineage. The PCA-tSNE followed by density clustering [35] (as described above) assigned the majority of the EdU labeled nuclei together with a few non-EdU labeled nuclei into a distinct cluster. Then, a second iteration of clustering was performed using nuclei only from this cluster, and outlier nuclei were removed (which are low quality cells, doublets or potentially other cell types). biSNE was used to find the trajectory of the remaining nuclei: nuclei were (1) clustered by PCA-tSNE (2); scored by all genes (as described in the selection of significant genes, combining top scoring genes from the Moran's I, Moran's I on binarized expression data, and Manhattan scores as described above); (3) clustered using the top high scoring genes and additional known stage specific marker genes. The top 3 PCs were used to place these nuclei on a trajectory.

The intercellular Euclidean distance on the tSNE embedding reflects the intercellular transcriptional divergence. The embedding of EdU labeled cells forms a trajectory-like distribution. The Euclidean distances along the trajectory reflect transcriptional changes along the underlying biological process. Positions of each cell on that trajectory should indicate how far the cell has progressed along the process. Thus, the position of each cell along the trajectory is correlated with the pseudotime of a cell in the biological process.

There is also a considerable cell distribution that makes up the width of the trajectory. The Euclidean distances orthogonal to the longitudinal axis of the trajectory reflect transcriptional divergence due to other cellular variabilities or noises. In order to find the position of each cell along the trajectory, we need to distinguish the distances along the trajectory from the distances orthogonal to the trajectory.

Previous methods find the cell positions using minimal spanning tree [57], or shortest possible route (travelling salesman problem) [2], neither of which take into account of the noise or other cellular variabilities. An improved method [58] uses randomization heuristics to mitigate the effect of noises.

In contrast to these methods, we model the noise explicitly and find a shortest spanning curve along the trajectory (Occam's razor). We then project cells onto this spanning curve, and find their projected positions.

Specifically, we find a curve that minimizes the following objective function,

$$f = \sum_i (\boldsymbol{x_i} - SP(\hat{q}_i, \boldsymbol{cp}))^2 + \lambda \int_0^1 \| \frac{\partial}{\partial t} SP(t, \boldsymbol{cp}) \| \, \mathrm{d}t \ ,$$

where the first term reflects Gaussian noises that model the orthogonal distances, the second term is the total length of the spanning curve, and $\boldsymbol{x_i}$ are the coordinates of the tSNE embedding of the cell i. The λ reflects the prior knowledge on the relative amount of noises and the transcriptional changes that align with the trajectory. The $SP(\hat{q}_i, \boldsymbol{cp})$ are the coordinates of the projected positions of cell i on the curve, and $\hat{q}_i$ is the pseudotime of the cell i along the trajectory.
The $\hat{q}_i$ is given by

$$\hat{q}_i = \underset{0 \leq t \leq 1}{\operatorname{argmax}} \ (\boldsymbol{x_i} - SP(t, \boldsymbol{cp}))^2 \ .$$

The $SP(x, \boldsymbol{cp})$ is the b-spline function [59] given by

$$SP(x, \boldsymbol{cp}) = \sum_i B_{i,n}(x)\ \boldsymbol{cp}_i\ ,$$

where $\boldsymbol{cp}$ are control points, and $B_{i,n}(x)$ is the b-spline basis function of degree n given by the following recursion formula

$$\begin{aligned} B_{i,1}(x) :&= \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1} \\ 0 & \text{otherwise} \end{cases} \ , \\ B_{i,k}(x) :&= \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(x)\ , \end{aligned}$$

where $\boldsymbol{t}$ is a knot vector, and k is the degree of the b-spline. We used a knot vector uniformly spaced between 0 and 1, and a third order b-spline.

The spanning curve is found by searching for control points that minimize the objective function f,

$$\hat{\boldsymbol{cp}} = \operatorname{argmax} f\ .$$

To initialize the curve, we calculated a smoothed shortest path (using Dijkstra's algorithm) that follows the trajectory. The smoothed shortest path contains 16 points spanning from the progenitor cells to immature neurons. These points were used as the initial control points. We then searched for the optimal control points using gradient descent,

$$\begin{aligned} \frac{\partial f}{\partial \boldsymbol{cp_i}} =& 2\sum_i (\boldsymbol{x_i} - SP(\hat{q}_i, \boldsymbol{cp}))(-SP(\hat{q}_i, \boldsymbol{e_i})) + \\ & \lambda \int_0^1 \frac{1}{2}\Big(\|\frac{\partial}{\partial t} SP(t, \boldsymbol{cp})\|\Big)^{-1} \Big(2\frac{\partial}{\partial t} SP(t, \boldsymbol{cp}) \frac{\partial}{\partial t} SP(t, \boldsymbol{e_i})\Big) \mathrm{d}t\ , \end{aligned}$$

where $\boldsymbol{e_i}$ is a matrix that has the same size as $\boldsymbol{cp}$, and entries of $\boldsymbol{e_i}$ are equal to 1 at column i corresponding to the control point i and zero elsewhere.

To quantify the expression of genes along the trajectory, running averages of gene expressions along the smoothed shortest path that follows the trajectory were calculated. Next, the running averages were subtracted by their mean to obtain normalized expressions. To find the dynamically expressed genes along the trajectory biSNE was applied, and the top scoring genes were clustered by their normalized expression patterns. A total of 5000 iterations of Kmeans clustering were performed, and the top consensus clusters were chosen. The consensus clusters were found by hierarchically clustering the frequency of pairwise co-assignment of genes within the same cluster across all Kmeans iterations (Hamming distance of the cluster assignments matrix).

Pathway and regulator analysis of adult newborn cells

Differentially expressed genes between immature neurons and adult neurons were found using the adjusted t-test. Enriched pathways in dynamic gene clusters and differentially expressed signatures were found (Hypergeometric p-value < 0.01) using the MsigDB/GSEA resource (combining Hallmark pathways, REACTOME, KEGG, GO and BIOCARTA) [60]. Dynamically regulated TFs were defined as genes within the genes clusters that are annotated by GO category [61] to be involved in transcription regulation, DNA binding or chromatin remodeling and modification. The gene list for the semaphoring signaling pathway was taken from KEGG mouse axon guidance pathway (mmu04360) and the IPA Semaphoring signaling pathway. We defined a maturation signature as the linear combination of centered expression levels of the set of up-regulated and down-regulated genes in mature granule cells compared to the immature granule cells in adult mice. The average relative expression of the up-regulated genes minus the average relative expression of the down-regulated genes was used to define a maturation score for each granule DG nuclei, in adult (3 months), adolescent (1 month) and old (2 year) mice.

Single molecule *in situ* hybridization tissue assay and EdU co-staining

Labeling of proliferating cells for staining in mice was performed by intraperitoneal (i.p.) injection of EdU (5-ethynyl-2′-deoxyuridine) (Thermo Fisher Scientific, #A10044) every 12 hr for 3 injections at a dose of 100mg/kg. Mice were sacrificed by a lethal dose of Ketamine/Xylazine 2 weeks post EdU injection, and transcardially perfused with PBS. Brain samples were immediately frozen in tissue freezing medium (O.C.T.) and kept in -80°C overnight. Coronal sections were cut at 15µm at -15°C. dFISH assay on O.C.T. embedded sections was performed according to Affymetrix provided protocol for O.C.T. samples, which combines QuantiGene ViewRNA ISH Tissue 2-plex Assay Kit (Affymetrix, #QVT0012) and ViewRNA ISH Cell Assay Kit (Affymetrix, #QVCM0001). Proprietary probes designed for *Eomes*, *Draxin*, and *Rrm2* were purchased from the vendor (Affymetrix) and used. Immediately following FISH protocol, while sections were still hydrated, EdU staining was performed using Click-iT Edu Imaging Kit (Thermo Fisher Scientific, #C10340) according to the manufacturer's protocol. After the protocol was completed, sections were washed twice in 1X wash buffer from the ViewRNA ISH Cell Assay Kit, followed by mounting and imaging.

The FISH images were acquired using a Nikon Ti-E microscope (Nikon, NY) equipped with a spinning disk confocal system (Andor, Belfast UK, WD system) and EMCCD camera (Andor, iXon Ultra 888). Excitation laser lines are 405 nm (DAPI), 488 nm (probe type 4, Alex Fluor 488), 561 nm (probe type 1, Alexa Fluor 568), 640 nm (EdU, Alexa Fluor 647). The maximum projection of the Z sections of confocal images were calculated and the image contrast of the resulting projection images were adjusted for better visualization of the FISH punta.

Differential isoforms

Gene isoform expression levels (TPM) and percent of mapped reads (compared to other all other isoforms of the same gene) were quantified using RSEM (as described in "Sequencing reads initial processing"). We restricted the analysis to highly expressed isoforms only, *e.g.* genes that have at least two isoforms with expression level of log(TPM+1) $>$ 4 in at least 10% of the analyzed nuclei. Analysis of differentially expressed isoforms between immature and mature granule neurons was done using t-test on the isoform percentage. A pair of isoforms are considered differentially expressed if both are significant in the t-test (FDR $<$ 0.01, log-ratio $>$ 1) and one is upregulated in immature neuron and the other is down regulated in the immature neuron.

Spinal cord analysis

All the 7 day EdU labeled and unlabeled cells from the spinal cord and DG were clustered by PCA-tSNE and density clustering as described in "Analysis of nuclei clusters". The identities of each clusters were determined based on differentially expressed genes and known marker genes. Immature and mature neurons were clustered by biSNE (top 2 PCs, 2,522 high scoring genes with $p < 5.4e\text{-}5$). Differentially expressed genes between immature neurons in the DG and spinal cord were calculated using student's t-test, with FDR $<$ 0.05, log(ratio) $>$ 1, and the average expression across samples in one region to be log(TPM+1) $>$ 2 and in the other region log(TPM+1) $<$ 3.

Div-Seq applied to the spinal cord and olfactory bulb

The 1-7 day EdU labeled and unlabeled cells from the spinal cord and DG were clustered by PCA-tSNE and density clustering as described in “Analysis of nuclei clusters”. The identities of each cluster was determined based on differentially expressed genes and known cell type marker genes. Each nucleus was assigned a cell type based on its cluster assignment. The first PC in the unbiased clustering of all the nuclei separated neuronal nuclei from glia nuclei (mainly oligodendrocytes, oligodendrocyte precursor cells, and astrocytes). The 50 highest scoring genes were defined as the “neuronal signature genes” and the 50 lowest scoring genes were defined as the “glia signature genes”. These signatures were used to calculate the glia-neuron score, which was defined as the difference in the total centered expression of the neuronal signature genes and the glia signature genes. The centered expression for each gene was defined as the log(TPM+1) expression of the gene subtracted by its mean expression across all nuclei. To place the SC neurons on the maturation trajectory the neuronal lineage nuclei were clustered by biSNE similarly as done for the DG with one exception. Here, the top 670 differential genes and markers from the SC trajectory were used, and the top 3 PCs were used in the clustering, which embedded the nuclei on a branched trajectory.

Differentially expressed genes between immature neurons in the DG and spinal cord were calculated using student’s t-test, with $FDR < 0.05$, $\log(ratio) > 1$, and the log transformed average expression in at least one region to be $\log(TPM+1) > 2$. For the olfactory bulb comparison, the immature neurons in the olfactory bulb (6-7 days post EdU labeling) were identified by clustering all EdU labeled and non-labeled cells in the SC and OB (similarly as done for the SC) and identifying the immature neuronal cluster by marker genes. The differentially expressed genes between the SC and DG immature neurons were clustered by hierarchical clustering based on their expression in the immature neurons from the OB, SC and DG. The cluster of genes up-regulated in both the OB and SC but not in the DG was chosen for further analysis.

BrdU labeling and immunohistochemistry

C57Bl6 adult mice (male, 6-8 weeks) were injected intraperitoneally with 200 mg/kg BrdU (Thermo Fisher Scientific, #B23151) every 12 h for 2 days. 8 days after the last injection mice were deeply anesthetized with isoflurane, and transcardially perfused with 4% PFA. Brain and spinal cord have been dissected in ice-cold PBS and postfixed for 24 h in 4% PFA. After dehydration in a graded ethanol series, Xylene (Sigma Aldrich) incubation for 15 min, and paraffin embedding, 8 µm sections were cut (Leica, Jung Multicut 2045). Sections were rehydrated in H_2O and incubated in 20 mM citric acid, 60 mM disodium phosphate, and 1.5% (vol/vol) H_2O_2 at room temperature for 15 min. After boiling in 40 mM Tris and 1 mM EDTA (pH 9.0), cooling to room temperature (1 h), and washing in PBS, sections were blocked with 3% Normal Goat Serum (NGS), 2% Donkey Serum (DS) in PBS and 0.1% Triton X-100 (PBST) for 1 h. Primary antibodies were added in 1.5% Normal Goat Serum (NGS), 1% Donkey Serum (DS) in PBST: rat anti-BrdU (1:50; Abcam, MA, #ab6326), rabbit anti-Pbx3 (1:100; Abcam, #ab56239) and mouse anti-NeuN (1:200; Millipore, MA, #MAB377). After washing in PBST, the sections were incubated with Alexa Fluor 488 goat anti-rat, Alexa Fluor 555 goat anti-rabbit and Alexa Fluor 647 goat anti-mouse including (all 1:1000; Thermo Fisher Scientific). The sections were finally washed in PBST and mounted in Vectashield mounting medium with DAPI (Vectorlabs, #H-1500). Confocal images were taken by confocal laser-scanning microscopy (Zeiss, LSM510) and assembled using Adobe Photoshop (Adobe Systems).

SUPPLEMENTARY TEXT

Comparison of nuclei and tissue RNA

We tested whether the type and complexity of nuclear mRNA can be effectively used for sensitive classification of cell types and states in the CNS on a large scale. Given the relative low total amount and non-uniform distribution of RNA within neuronal subcellular compartments (nuclei, soma, axons, and dendrites), analysis of nuclear RNA may introduce biases. We thus compared RNA profiles of bulk tissue and populations of nuclei from the hippocampus dentate gyrus (DG) and found they showed remarkable agreement, with similar RNA complexity and profiles (fig. S2A), and that nuclear RNA enriches for long non-coding RNAs (fig. S2B).

Cell sub-Type classification from sNuc-Seq data

Analysis of the glia nuclei (cluster 7 in **Fig. 1B**) by PCA-tSNE recovered five known glial cells sub-types [11] (fig. S6A-C), despite the low number of nuclei in each sub-cluster. Moreover, averaged expressions across each sub-cluster correlated well with published population RNA-Seq data [11] (fig. S6D and table S1).

BiSNE analysis of the GABAergic neurons nuclei (cluster 5 in **Fig. 1B**) partitioned the neurons into eight sub-clusters (fig. S8A-B), each with unique expression of individual or pairs of canonical interneuron marker genes, such as *Pvalb*, *Vib* and *Htr3a* (fig. S8C, validated by double fluorescent RNA *in situ* hybridization, fig. S9). We further characterized the sub-clusters by differential gene expression analysis (table S2), revealing for example that the calcium channel *Cacna1i* is specifically expressed in *Pvalb* or *Sst* positive GABAergic neurons (fig. S8D and table S2).

BiSNE analysis of the glutamatergic neurons from CA1 (cluster 4 in **Fig. 1B**), CA3 (cluster 2), and DG (clusters) into 8, 6, and 2 sub-clusters, respectively (**Fig. 2A** and fig. S10A). Analysis of sub-cluster specific gene expression highlighted several known landmark genes that exhibit spatially restricted expression patterns in sub-regions of the hippocampus, indicating a correspondence between hippocampal sub-regions and sub-clusters of glutamatergic nuclei. We then used the spatial expression patterns [10] of these landmark genes to map sub-clusters in CA1, CA3, and DG to distinct spatial sub-regions (**Fig. 2B** and fig. S11, S12, S13). Previous studies using single-neuron RNA-Seq in CA1 reported two cell clusters that did not match spatial position [1] (fig. S15), whereas our spatial mapping of Nuc-Seq data corresponds to continuous transcriptional transitions within the CA1 and CA3 regions, in coordination with recent observations [12, 62].

Comparison of adult neurogenesis dynamics in the SC and DG

Neuronal lineage nuclei from the full Div-Seq time course, showed a continuous trajectory in both the DG (**Fig. 3C**), and the SC (**Fig. 4C**), and broadly comparable between the two regions (**Fig. 3D-E** and fig. S20C). However, there were also key distinctions between the processes in the DG and SC. First, while gene expression patterns are similar at early and later stages, expression levels along intermediate time points change more sharply in the SC compared to the DG (**Fig. 3D, 4D** and fig S20B). Second, in contrast to the mostly unbranched path we observed in DG neurogenesis, the SC trajectory has several branches (**Fig. 4C**). Interestingly, the gene expression profiles of nuclei at side branches resemble a glia expression pattern more than a neuronal pattern (**Fig. 4E**).

SUPPLEMENTARY FIGURES

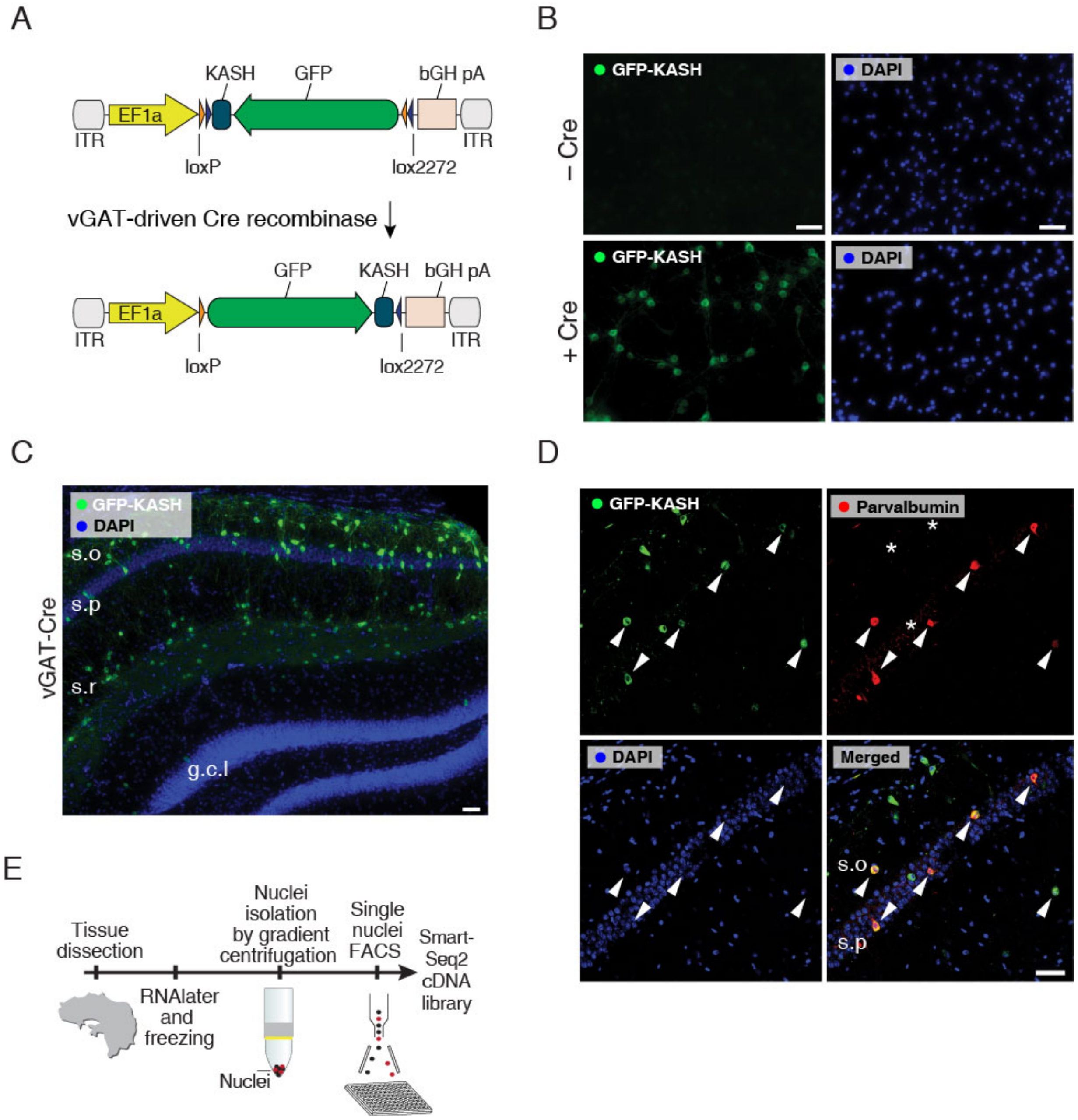

Figure S1: Nuc-Seq is compatible with genetic labeling for enrichment of rare cells. (**A**) Genetic labeling of GABAergic interneurons using AAV expression vectors. Cre-mediated recombination of inverted transgenic cassette flanked by oppositely oriented loxP and lox2272 (Double-floxed Inverted Orientation, DIO) sites drives expression of GFP-KASH. Top: before recombination. Bottom: after cre-driven recombination. (**B**) Primary cortical neurons infected with pAAV-EF1a-DIO-GFP-KASH-bGH-polyA alone (top) or co-infected with pAAV-EF1a-Cre-WPRE-bGH-polyA (bottom). (**C**) Expression of GFP-KASH in hippocampus of vGAT-Cre mice 14 d after viral delivery of pAAV-EF1a-DIO-GFP-KASH-bGH-polyA into CA1/CA2 stratum pyramidale (s.p.). (**D**) GFP-KASH labeled parvalbumin positive (arrowheads) and negative (asterisk) interneurons in hippocampus of vGAT-Cre mice shown in C. (ITR – inverted terminal repeat; GFP – green fluorescent protein; KASH – Klarsicht, ANC1, Syne Homology nuclear transmembrane domain; hGH pA – human growth hormone polyadenylation signal; WPRE – Woodchuck Hepatitis virus posttranscriptional regulatory element, s.o. – stratum oriens, s.p. – stratum pyramidale, s.r. – stratum radiatum, g.c.l. – granule cell layer). Scale bars: 50µm (**E**) Nuc-Seq method overview. Dissected tissue is fixed in RNA-later for 24 hours at 4°C (and can be subsequently stored in -80°C or further processed); nuclei are isolated using a gradient centrifugation method [8] (samples kept at 4°C or on ice), resuspended, and sorted using FACS to a single nucleus per well in plates. Plates are processed using a modified Smart-Seq2 RNA-Seq protocol [63, 21].

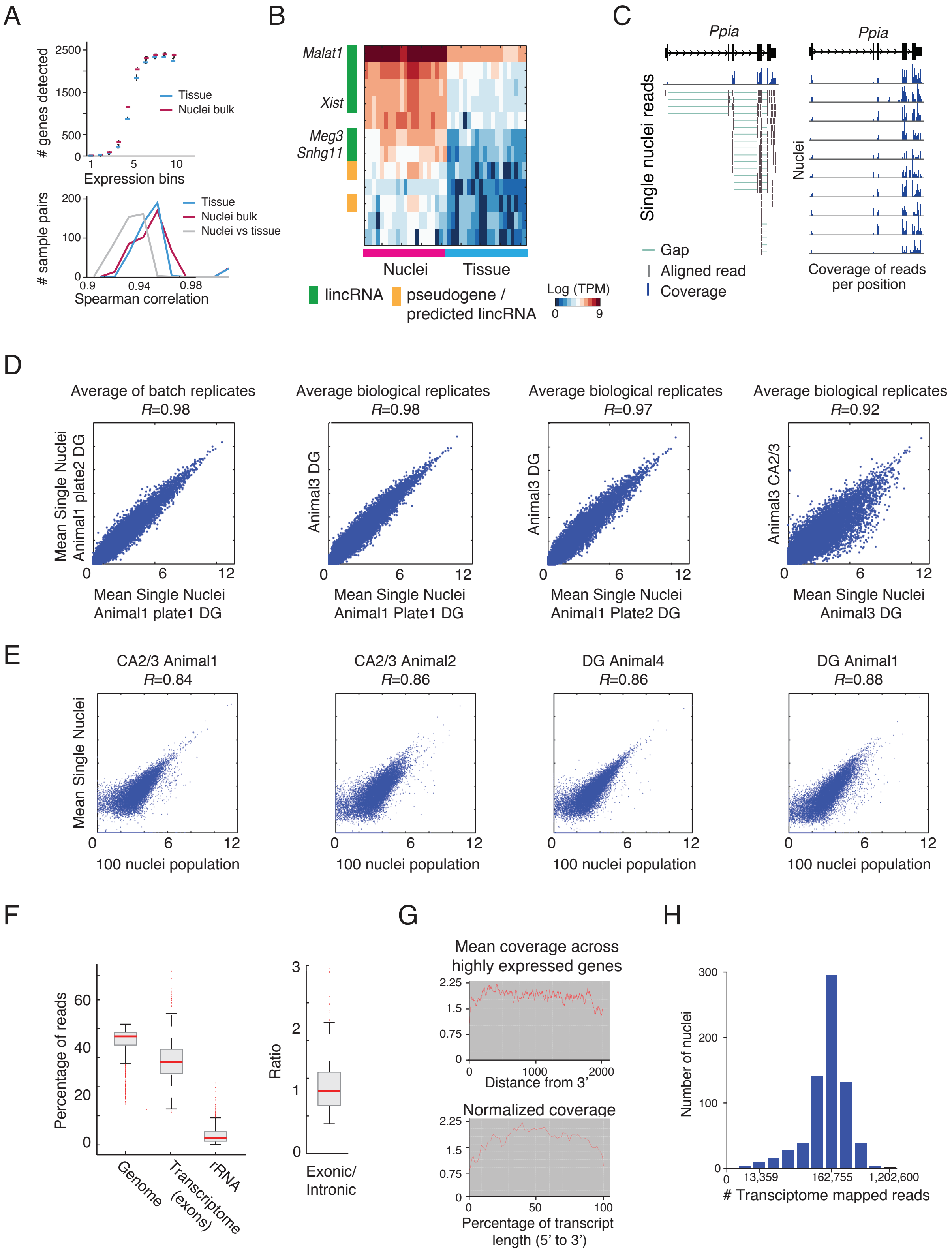
A
# genes detected
2500
1500
500
0
1
5
10
Expression bins
Tissue
Nuclei bulk
# sample pairs
200
100
0
0.9
0.94
0.98
Spearman correlation
Tissue
Nuclei bulk
Nuclei vs tissue
B
Malat1
Xist
Meg3
Snhg11
Nuclei
Tissue
lincRNA
pseudogene / predicted lincRNA
Log (TPM)
0
9
C
Ppia
Ppia
Single nuclei reads
Nuclei
Gap
Aligned read
Coverage
Coverage of reads per position
D
Average of batch replicates
R=0.98
Mean Single Nuclei Animal1 plate2 DG
Mean Single Nuclei Animal1 plate1 DG
Average biological replicates
R=0.98
Animal3 DG
Mean Single Nuclei Animal1 Plate1 DG
Average biological replicates
R=0.97
Animal3 DG
Mean Single Nuclei Animal1 Plate2 DG
Average biological replicates
R=0.92
Animal3 CA2/3
Mean Single Nuclei Animal3 DG
0
6
12
E
CA2/3 Animal1
R=0.84
CA2/3 Animal2
R=0.86
DG Animal4
R=0.86
DG Animal1
R=0.88
Mean Single Nuclei
100 nuclei population
0
6
12
F
Percentage of reads
60
40
20
0
Genome
Transcriptome (exons)
rRNA
Ratio
3
2
1
0
Exonic/ Intronic
G
Mean coverage across highly expressed genes
2.25
1.5
0.75
0
0
1000
2000
Distance from 3'
Normalized coverage
0
50
100
Percentage of transcript length (5' to 3')
H
Number of nuclei
300
200
100
0
0
13,359
162,755
1,202,600
# Transciptome mapped reads

Figure S2: Quality measurements of Nuc-Seq libraries. (**A**) Nuc-Seq faithfully captures tissue RNA. Comparing Nuc-Seq on populations of nuclei and RNA-seq on tissue samples from the DG brain region. Shown are number of genes detected (TPM $>$ 3) per expression quantile (top) and distribution of pairwise spearman correlations across samples (bottom). (**B**) Nuc-Seq libraries are enriched for long non-coding RNAs (lincRNAs). Heatmap showing differentially expressed genes between Nuc-Seq on population of nuclei (columns, pink) and tissue RNA-Seq (columns, blue). T-test FDR q-value $<$ 0.05 with log-ratio $>$ 1, mean log(TPM+1) $>$ 2 in at least one condition, 21 samples per condition. Left: colorbar showing the classification of genes as lincRNAs (green) and pseudogene/predicted lincRNA (orange). Names of known nuclear localized lincRNAs are marked (left). (**C**) Nuc-Seq detects full-length, spliced transcripts. Showing RNA-Seq read coverage at the *Ppia* genomic locus. Top track: exons/introns, thick and thin lines. Left: Alignments of individual spliced reads from one single nucleus at the locus. Grey bar: individual read, green line: gapped alignment. Right: RNA-Seq read coverage at the locus in ten individual nuclei (rows). (**D**) Average of dentate gyrus (DG) and CA2/3 Nuc-Seq data correlates between replicates. Scatter plots showing comparison between average of single nuclei across technical and biological replicates. Data is shown in log(TPM+1). Spearman correlation between replicates (R), top. (**E**) Average of Nuc-Seq data correlates with population samples. Scatter plots showing comparison between average of single nuclei (Y axis) to populations of 100 nuclei (X axis). (**F**) Mapping rates of Nuc-Seq data. Left: Box plots showing the mapping rates to the genome, transcriptome (exons only) and rRNA. In box plots, the median (red), 75% and 25% quantile (box), error bars (dashed lines), and outliers (red dots). Right: Box plots showing the percentage of reads mapped to introns and exons in Nuc-Seq libraries. (**G**) 3' and 5' bias. Top: Mean read coverage across highly expressed genes per distance from the 3' of the gene. Showing constant coverage with a decrease around 2000 bp from the 3' end. Bottom: Mean read coverage throughout the transcripts, averaged per percentage of the transcript length (3' to 5'). (**H**) Distribution of number of reads mapped to the genome per Nuc-Seq library.

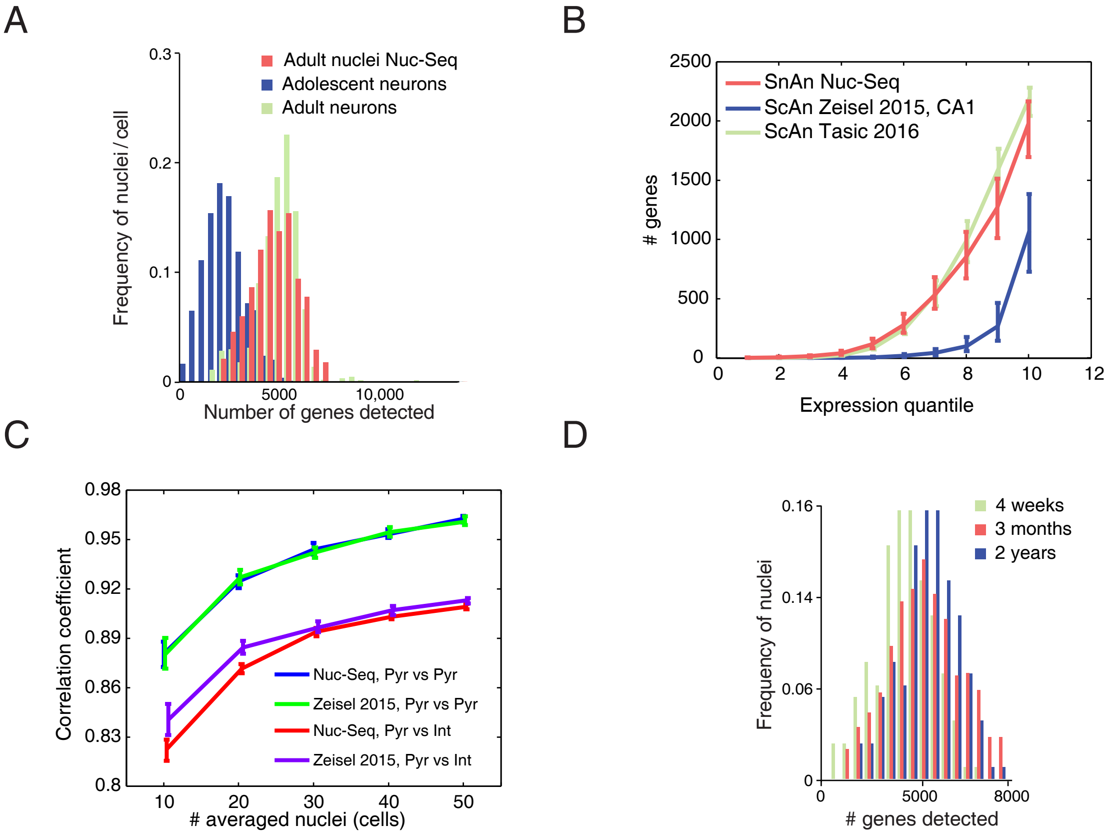

Figure S3: Comparison of Nuc-Seq and single cell RNA-Seq. (**A**) Nuc-Seq detects consistently higher number of genes (TPM/FPKM $>$ 3 or UMI $>=$ 1.1) compared to published single neuron RNA-seq in adolescent or adult mice. (**B**) Nuc-Seq detects comparable or more genes than single cell RNA-Seq across wide expression range. Shown is the distribution of number of genes detected (Y axis, log(TPM+1) $>$ 1.1) per nucleus for Nuc-Seq and per cell for Zeisel 2015 (only CA1 neurons) [1], and Tasic 2016 [3] across expression quantiles (X axis). A different threshold (X axis, TPM $>$ 1.1) was used in the calculation of number of genes detected for Zeisel 2015, which used unique molecular identifier (UMI) counts. SnAn: single nuclei sequencing of adult neuron; ScAn: single cell sequencing of adult/adolescent neuron; Error bar: 75% and 25% quantile. (**C**) Transcriptional profiles between different cell types are more distinct in NucSeq than in single cell RNA-Seq [1]. Plots showing Spearman correlation coefficients (Y axis) between two subsets of averaged pyramidal neurons (Pyr) or between subsets of averaged pyramidal neurons and

averaged GABAergic interneurons (Int). A subset of neurons are first randomly sampled from Nuc-Seq and single-cell RNA-Seq. Then Spearman correlation is calculated between the averages of the subsets (see Materials and Methods). (**D**) Nuc-Seq detects similar number of genes across animal ages, 4 weeks, 3 months and 2 years old (detected gene defined as log(TPM+1) $>$ 1.1).

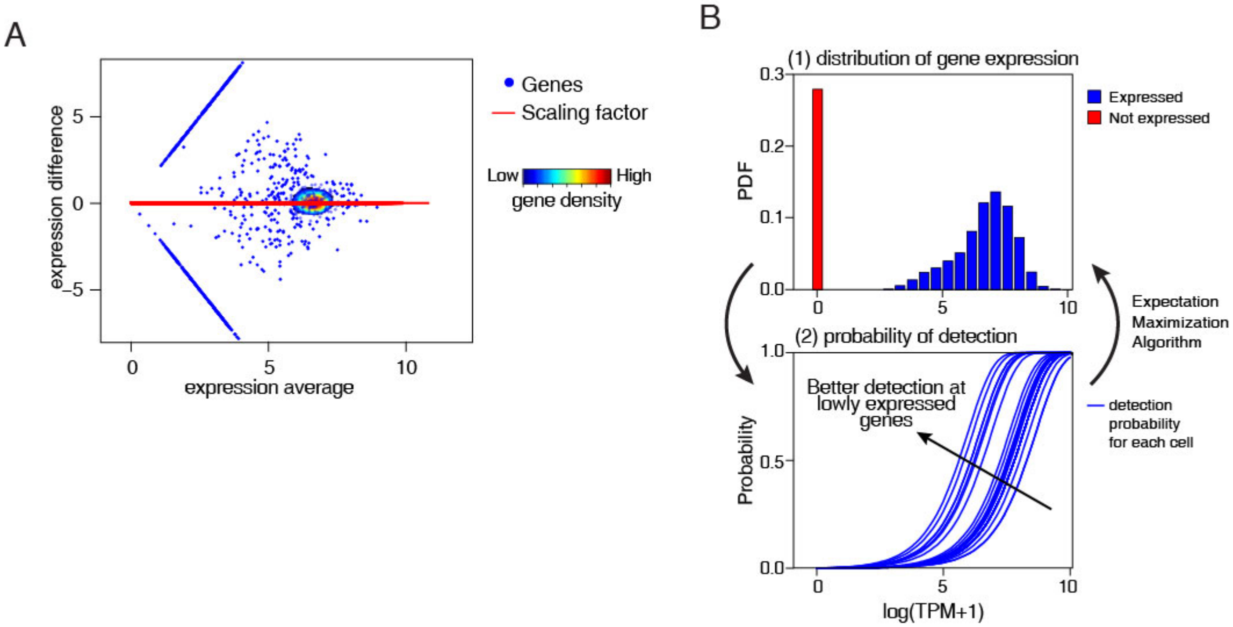

Figure S4: Computational methods. (**A**) Density MA plot normalization method. MA plot showing the average log (X axis) versus the log-ratio (Y axis) of TPM expression of all genes between two single nuclei. High density region marked by a color scale. Genes within the colored density region are used to calculate the scaling factor between libraries for normalization. (**B**) Illustration of false negative estimation method. An expectation maximization algorithm alternates between estimation of gene expression distribution per gene (top) and the probability of detection (bottom) per cell. Top: histogram shows estimated distribution of expression of an example gene, PDF: probability density function. Bottom: each blue curve represent the probability of successfully detecting genes expressed at different levels in each cell.

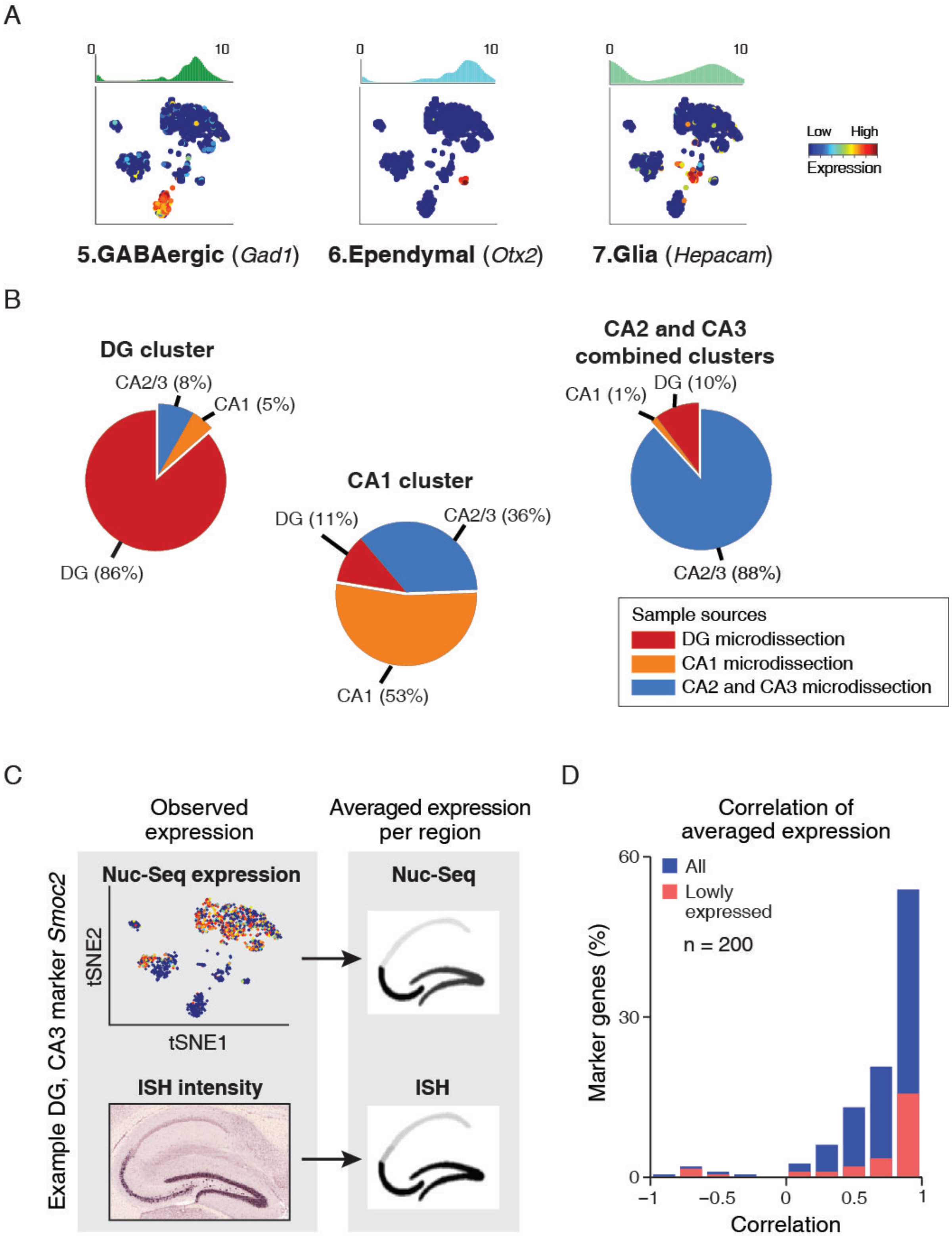
A
0
10
0
10
0
10
Low High
Expression
5.GABAergic (*Gad1*)
6.Ependymal (*Otx2*)
7.Glia (*Hepacam*)
B
DG cluster
CA2/3 (8%)
CA1 (5%)
DG (86%)
CA1 cluster
DG (11%)
CA2/3 (36%)
CA1 (53%)
CA2 and CA3 combined clusters
CA1 (1%)
DG (10%)
CA2/3 (88%)
Sample sources
DG microdissection
CA1 microdissection
CA2 and CA3 microdissection
C
Observed expression
Averaged expression per region
Example DG, CA3 marker *Smoc2*
Nuc-Seq expression
tSNE2
tSNE1
Nuc-Seq
ISH intensity
ISH
D
Correlation of averaged expression
60
30
0
All
Lowly expressed
n = 200
Marker genes (%)
−1
−0.5
0
0.5
1
Correlation

Figure S5: Validation of cell type classification based on Nuc-Seq data. (**A**) Identification of GABAergic, ependymal and glial clusters. For each cluster, marker gene expression is shown in two ways: **1**, histogram quantifying expression level of the marker gene across all nuclei in the relevant cluster (top); and **2**, 2-D embedding of nuclei (as in **Fig. 1B**) showing relative expression level of the marker gene across all nuclei. (**B**) Nuc-Seq clusters agree with the anticipated cell types based on the microdissected anatomical regions. Shown are the distributions of nuclei from each microdissection source (DG, CA1, CA2 and CA3) within each of the nuclei clusters identified as DG, CA1, and CA2 and CA3 combined. (**C**) Computational pipeline for the validation of expression patterns using ISH. An example of comparison of the expression pattern of the *Smoc2* gene across CA1, CA2, CA3, and DG Nuc-Seq clusters to its expression in the corresponding regions in ISH data. Top left: scatter plot of 2-D embedding of all nuclei (as in **Fig. 1B**) colored by the expression of *Smoc2* across all nuclei (Nuc-Seq data). Top right: the average Nuc-Seq expression levels in the CA1, CA2, CA3, and DG clusters presented in a schematics of the hippocampus (gray scale, high expression in dark grey and low expression in light grey). Bottom left: the expression pattern of *Smoc2* in Allen ISH [10] image. Bottom right: the average expression levels in the CA1, CA2, CA3, and DG regions presented in a schematics of the hippocampus (gray scale). (**D**) Distribution of correlation coefficients of average RNA-Seq expressions and ISH [10] intensities per gene, across all differentially expressed genes between the CA1, CA2, CA3 and DG regions. Shown are all genes (blue) and lowly expressed (red) defined as averaged expression in all regions within bottom 25%. quantile.

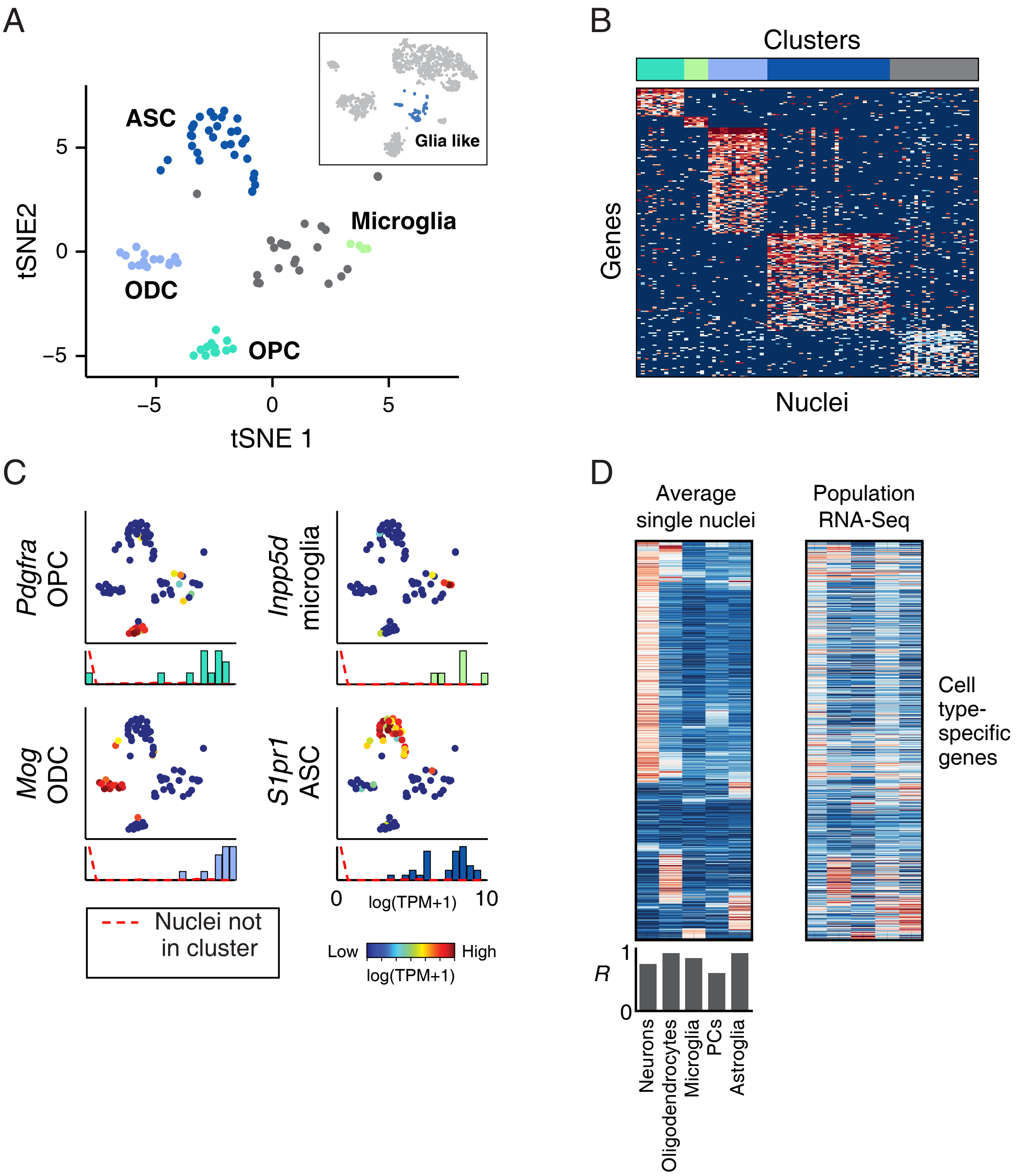

Figure S6: Nuc-Seq identifies glial cell types. (**A**) Clustering of glial nuclei. Top insert: the glial cluster (blue) within all other nuclei from **Fig. 1B**. The glial nuclei are divided to five clusters by PCA-tSNE: oligodendrocytes (ODC), astroglia (ASC), oligodendrocyte precursor cells (OPC), microglia, and

a sparse cluster of diverse cells (grey). (**B**) Marker genes. Heatmap shows the expression of marker genes (rows, t-test FDR q-value < 0.05 with log-ratio > 1 across all pairwise comparisons between sub-clusters) specific for each of the five clusters in (A) (color bar, top, matches cluster color in A) across the single nuclei (columns). (**C**) Identification of each glial sub-cluster by marker genes. For each cluster, a marker gene expression is shown in two ways: Top: 2-D embedding of nuclei (as in A) showing relative expression level of the marker gene across all nuclei. Bottom: histogram quantifying expression level of the marker gene nuclei in the relevant cluster (colored bars) and the distribution across all other nuclei (dashed red line). (**D**) Single nuclei transcriptional profiles match population RNA-Seq. Heat map showing the expression of top marker genes in the average of single nuclei (left) and in population RNA-Seq [11]. Bottom: Bar plot of the Pearson correlation (R) of each expression signature to the relevant population.

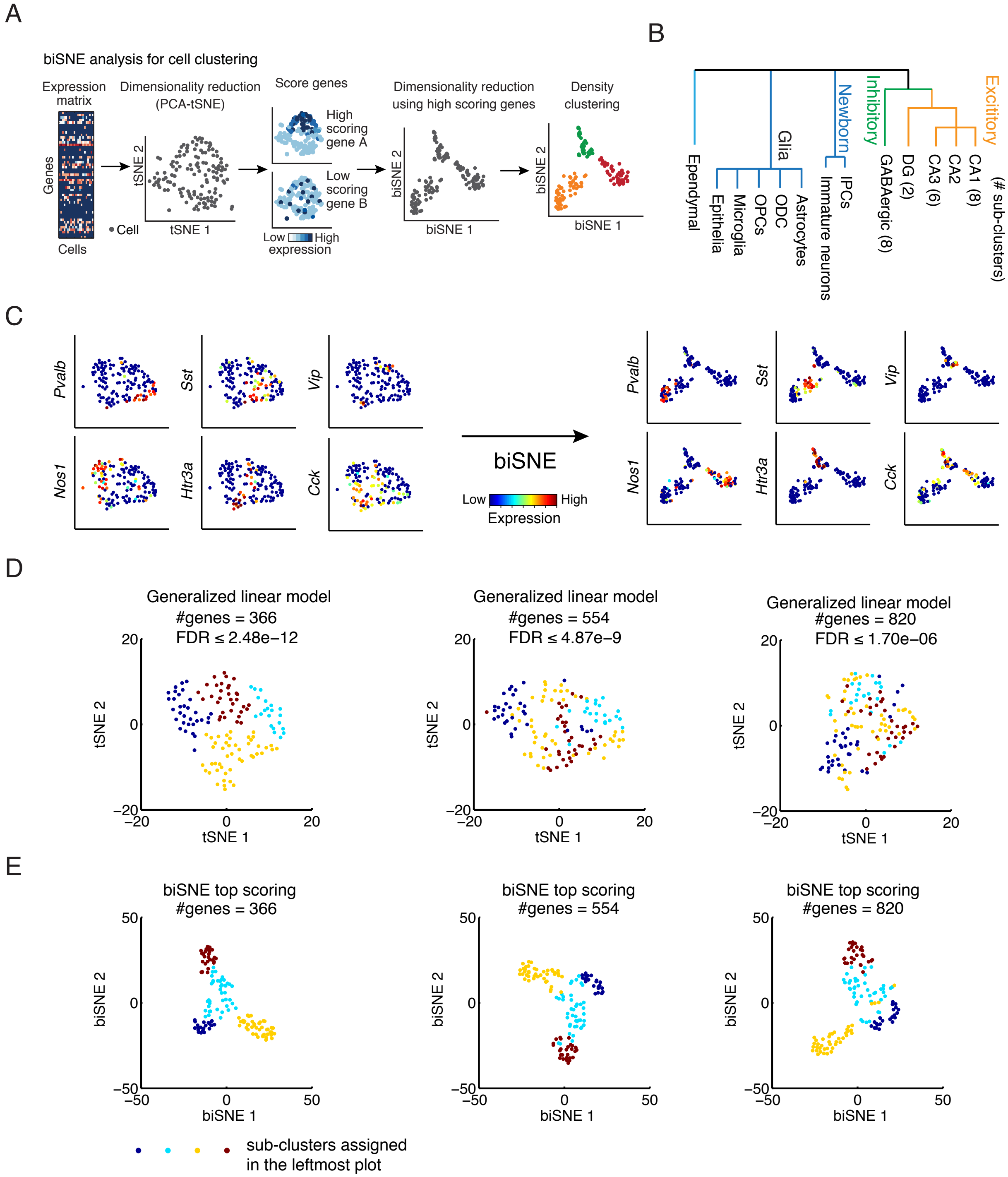

Figure S7: BiSNE algorithm. (**A**) BiSNE algorithm. Top row, left to right: BiSNE takes as input an expression matrix of genes (rows) across nuclei (or cells, columns). It generates a 2-D plot of nuclei

by dimensionality reduction using PCA followed by tSNE non-linear embedding, and then scores each gene by their expression across the 2-D plot, such that genes expressed in nuclei in proximity on the 2-D plot (dark blue points, top) are high scoring, whereas those expressed in nuclei scattered across the plot (dark blue points, bottom) are low scoring. Next, it takes an expression matrix of only high scoring genes (heatmap, genes (rows) across all nuclei (columns)), and repeats the dimensionality reduction. BiSNE is followed by density clustering (colored, bottom left). (**B**) BiSNE sub-clustering. Dendrogram of all nuclei clusters along with number of sub-clusters found by biSNE. NPC: neuronal precursor cells, ODC: oligodendrocytes, ASC: astroglia, OPC: oligodendrocyte precursor cell. (**C**) Expression of marker genes across 2-D embedded nuclei before and after biSNE. Shown is a panel of the same tSNE 2-D embedding of the GABAergic nuclei (from **Fig. 1B**), with each panel colored by the expression of a marker genes (denoted on the left). Left: using PCA-tSNE only. Right: using biSNE. (**D**) 2-D embedding of cells using genes selected by generalized linear model (GLM) with different thresholds (Top). Cells are grouped in leftmost 2-D embedding and denoted by group colors. GLM with less stringent thresholds selects more genes (from left to right), and results in different 2-D embedding without preserving cell grouping (from left to right). (**E**) 2-D embedding of cells using genes selected by biSNE with different thresholds (Top). biSNE with different thresholds results in similar 2-D embedding preserving cell grouping.

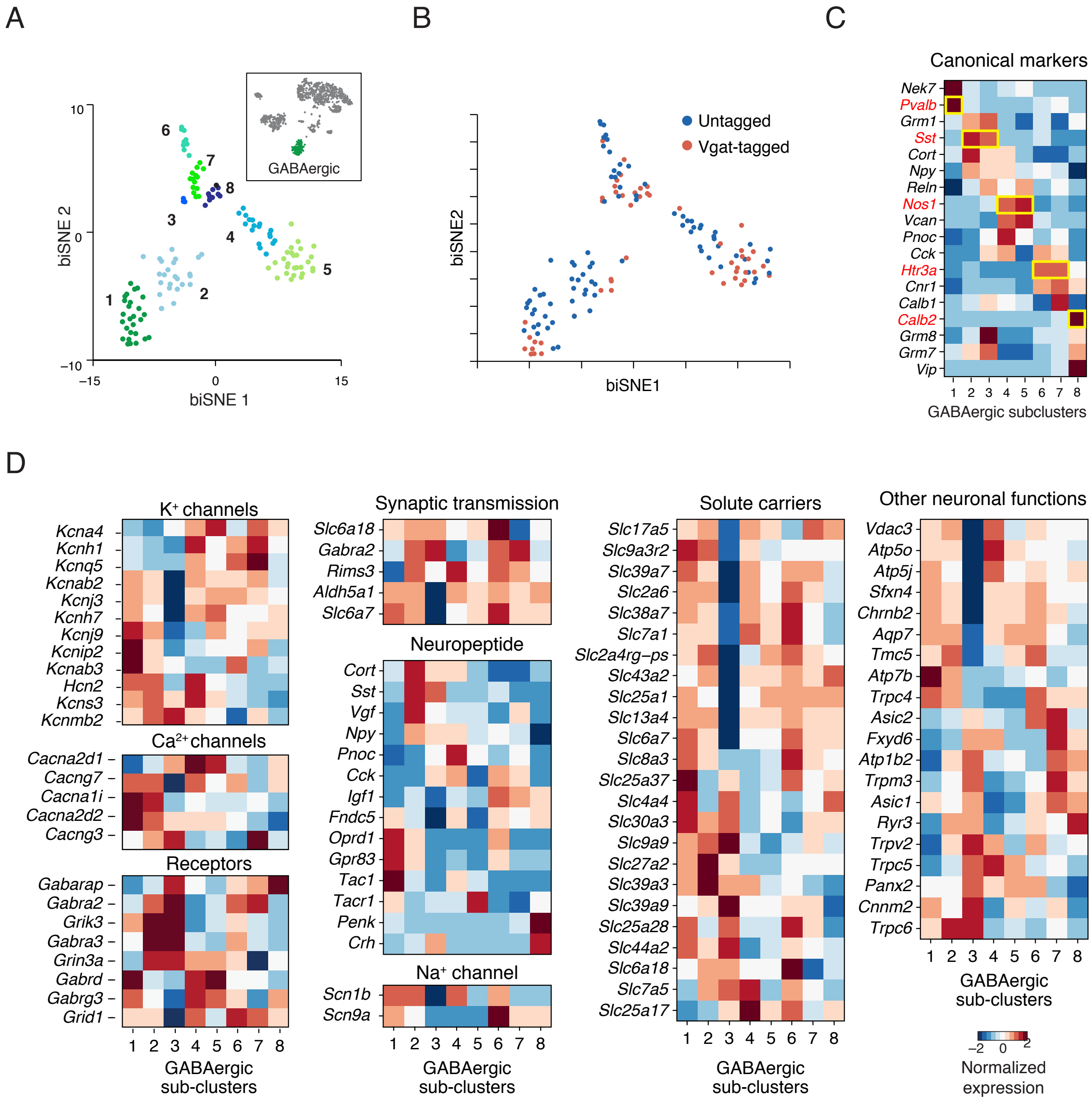

Figure S8: Transcriptional profiles of GABAergic interneurons. (**A**) Sub-clusters of GABAergic interneurons identified by biSNE. Shown is a biSNE 2-D embedding of GABAergic nuclei with 8 sub-clusters. Top insert: the GABAergic cluster within all other nuclei from Fig. 1B. (**B**) biSNE clustering of GABAergic interneurons is independent of AAV infection or expression of transgene. Showing tSNE 2-D embedding of the GABAergic nuclei clustered with biSNE (from A) displaying untagged nuclei

(blue) and Vgat-tagged nuclei (red) from Vgat-cre mice (fig. S1). (**C**) GABAergic sub-clusters (from A) are characterized by a combination of canonical marker genes. Heat map with averaged expression of canonical neuron markers (rows) across GABAergic sub-clusters (columns). (**D**) Differentially expressed neuronal functional genes across GABAergic sub-clusters (from A). Average centered expression of differentially expressed (t-test FDR q-value < 0.05 with log-ratio > 1 in at least one pairwise comparison between sub-cluster) K^+ channels, ${Ca_2}^+$ channels, receptors, synaptic transmission, neuropeptides, Na^+ channels, solute carriers, and other neuronal function across GABAergic sub-clusters (columns, from A).

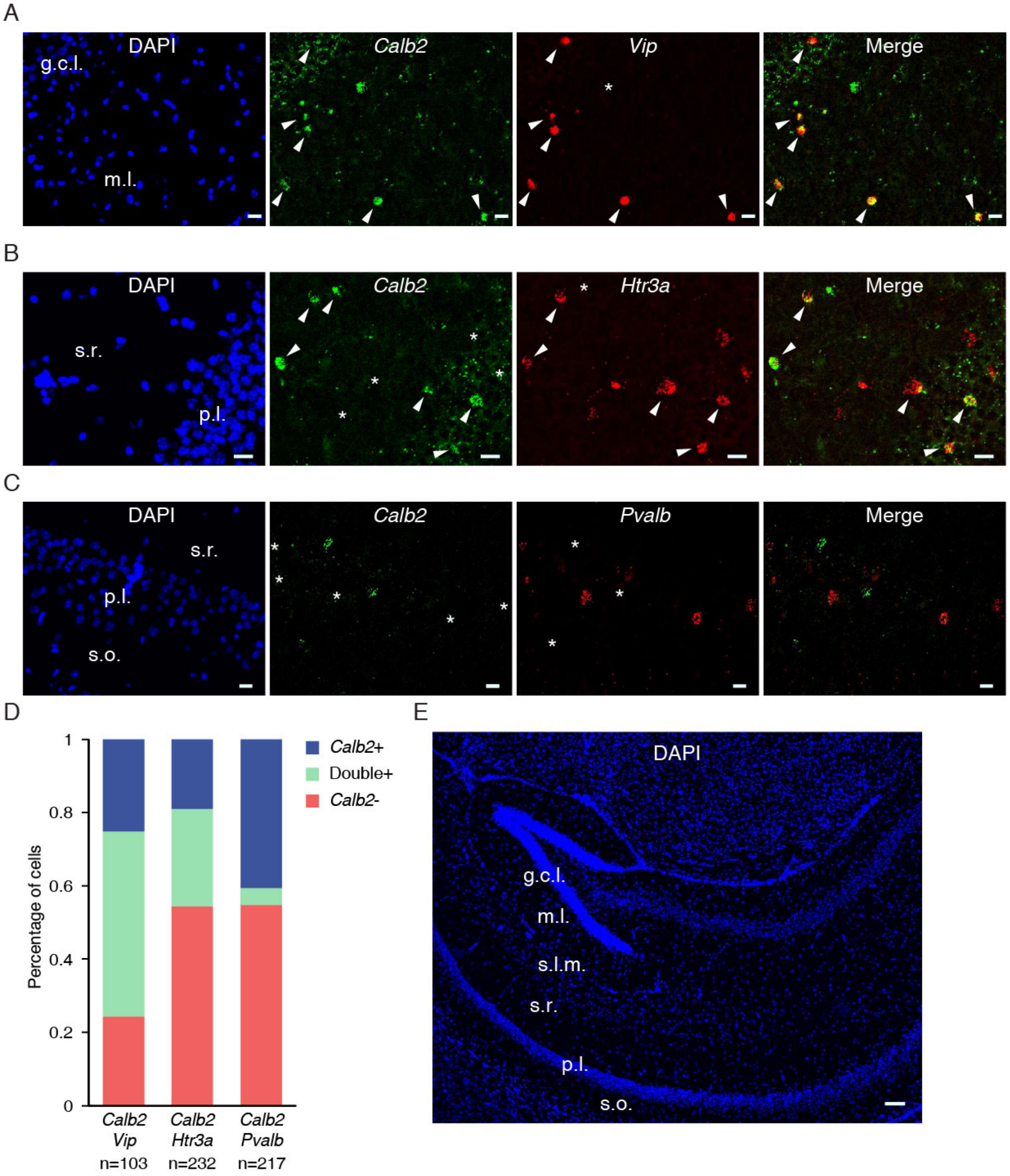
A
DAPI
g.c.l.
m.l.
Calb2
Vip
Merge
B
DAPI
s.r.
p.l.
Calb2
Htr3a
Merge
C
DAPI
s.r.
p.l.
s.o.
Calb2
Pvalb
Merge
D
1
0.8
0.6
0.4
0.2
0
Percentage of cells
Calb2+
Double+
Calb2-
Calb2 Vip n=103
Calb2 Htr3a n=232
Calb2 Pvalb n=217
E
DAPI
g.c.l.
m.l.
s.l.m.
s.r.
p.l.
s.o.

Figure S9: Validation of GABAergic interneuron subtypes. (**A**) double fluorescent RNA in situ hybridization (dFISH) of *Calb2* (green) and *Vip* (red). Expressions of *Calb2* and *Vip* are largely overlapped (arrowheads). Scale bar: 20μm. (**B**) dFISH of *Calb2* (green) and *Htr3a* (red). Expressions of *Calb2* and *Htr3a* are partially overlapped (arrowheads). Scale bar: 20μm. (**C**) dFISH of *Calb2* (green) and *Pvalb* (red). Expressions of *Calb2* and *Pvalb* are not overlapped (asterisks). Scale bar: 20μm. (**D**) Quantification of dFISH images. Bar plots showing the percent of single and double labeled cells in FISH images for each pair of genes. (**E**) DAPI image showing the entire view of hippocampus. Scale bar 100μm. g.c.l. – granule cell layer; m.l. – molecular layer; s.l.m. – stratum locunosum-moleculare; s.r. – stratum radiatum; p.l. – pyramidal layer; s.o. – stratum oriens.

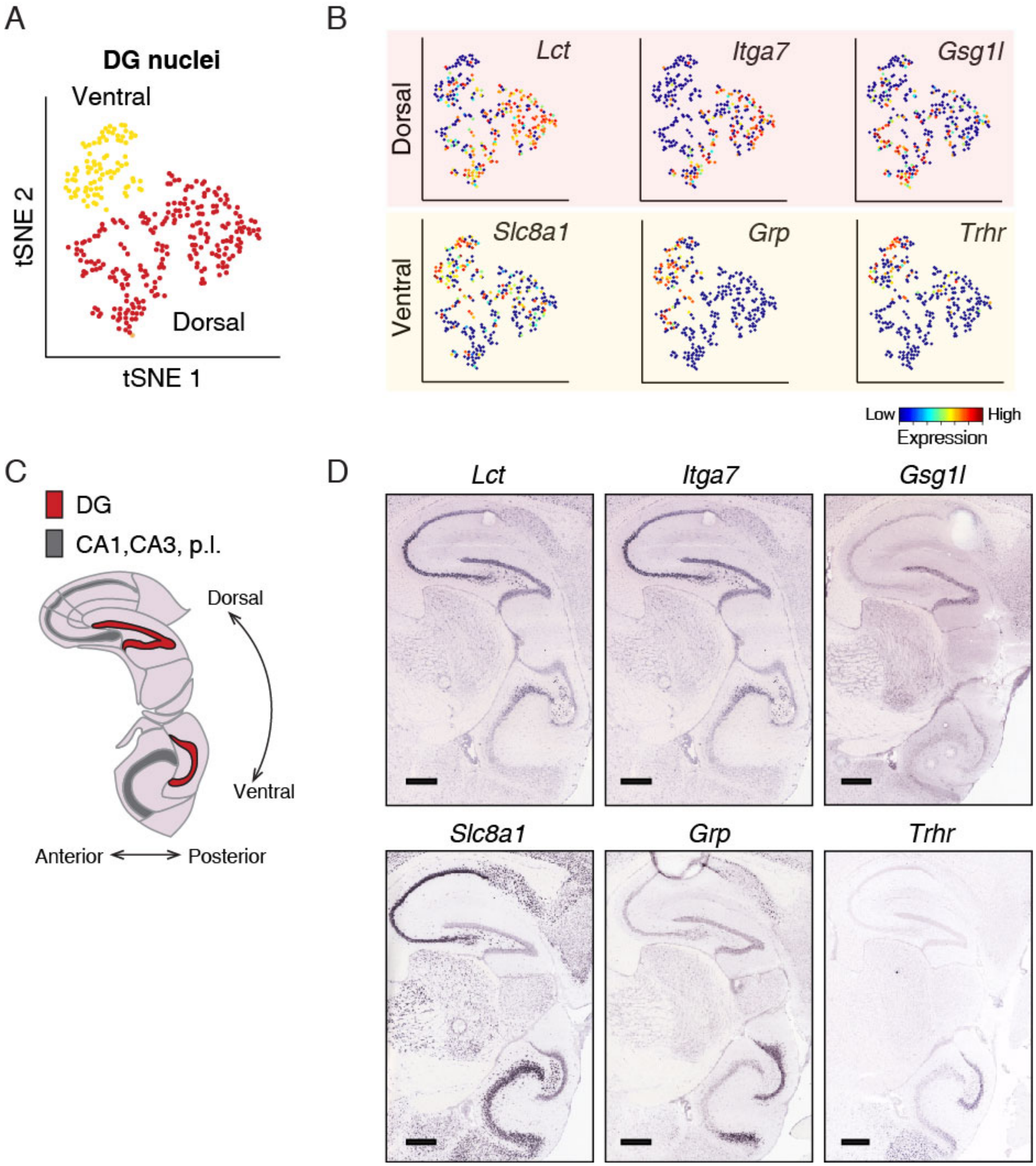
A
DG nuclei
Ventral
Dorsal
tSNE 2
tSNE 1
B
Dorsal
Lct
Itga7
Gsg1l
Ventral
Slc8a1
Grp
Trhr
Low
High
Expression
C
🟥 DG
⬛ CA1,CA3, p.l.
Dorsal
Ventral
Anterior
Posterior
D
Lct
Itga7
Gsg1l
Slc8a1
Grp
Trhr

Figure S10: Spatial pattern of DG granule cells. (**A**) DG granule cells sub-clusters. Shown is a biSNE 2-D embedding of the DG granule nuclei with 3 sub-clusters denoted by colors. (**B**) Differential genes between clusters that have a distinct spatial pattern. 2-D embedding of nuclei (as in A), each showing the relative expression level of a gene expressed in the dorsal DG (top) or the ventral DG (bottom). (**C**) Schematics of hippocampal anatomy in a sagittal plane. DG marked in red. p.l. – pyramidal layer. (**D**) Spatial pattern of genes in the DG. ISH [10] sagittal image of the genes in (B). Top: Dorsal expression pattern. Bottom: Ventral expression. Scale bar: 400µm.

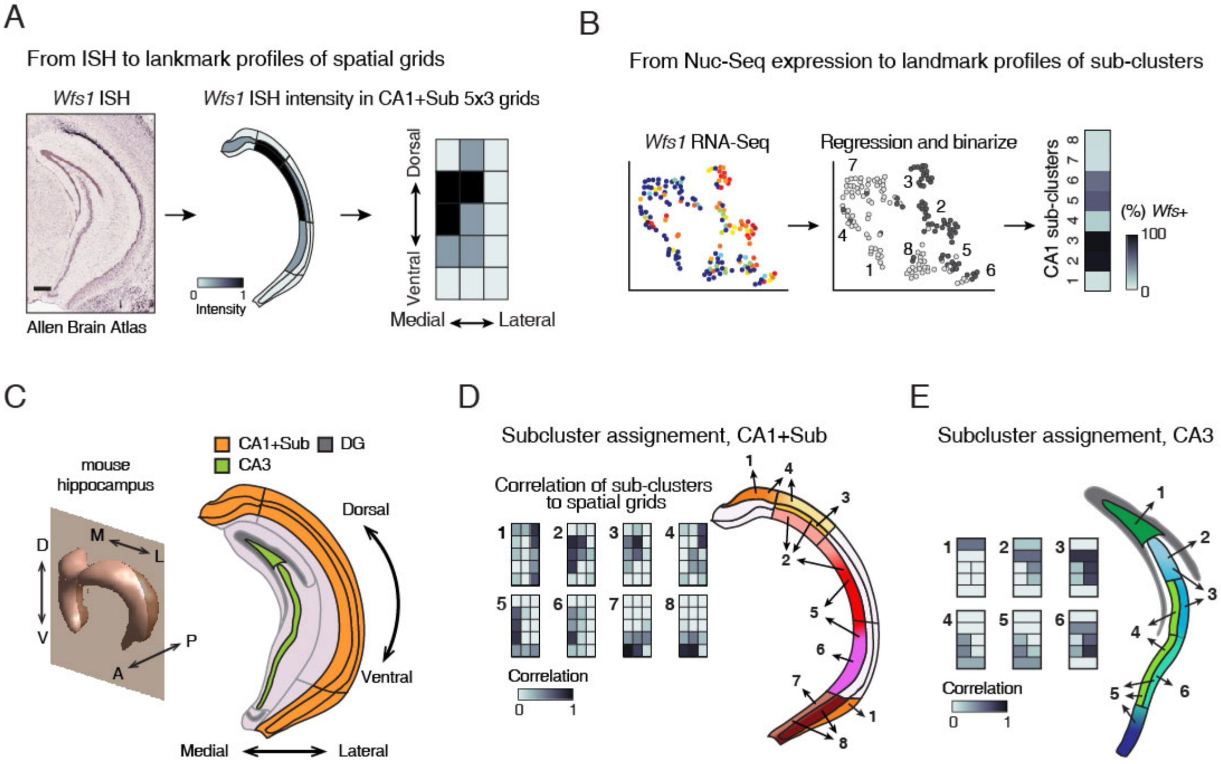
A
From ISH to lankmark profiles of spatial grids
Wfs1 ISH
Wfs1 ISH intensity in CA1+Sub 5x3 grids
Allen Brain Atlas
0
1
Intensity
Dorsal
Ventral
Medial
Lateral
B
From Nuc-Seq expression to landmark profiles of sub-clusters
Wfs1 RNA-Seq
Regression and binarize
CA1 sub-clusters
(%) Wfs+
100
0
C
mouse hippocampus
CA1+Sub
DG
CA3
D
V
M
L
A
P
Dorsal
Ventral
Medial
Lateral
D
Subcluster assignement, CA1+Sub
Correlation of sub-clusters to spatial grids
Correlation
0
1
E
Subcluster assignement, CA3
Correlation
0
1

Figure S11: Spatial assignment method (**A**) Spatial assignment method using landmark gene expression. Nuclei sub-clusters are assigned to brain regions, by comparing a spatial map of landmark gene expression from ISH data to the expression of the landmark genes in each of the sub-clusters. An example using the landmark gene *Wfs1*. Left to right: creating a spatial landmark expression map from ISH data - *Wfs1* Allen ISH [10] image data is quantified for its intensity in 15 bin grid (dividing the CA1 region into five grid bins along the dorsal-ventral axis and three grid bins along medial-lateral axis). (**B**) Left to right: Sub-cluster expression map - *Wfs1* RNA-Seq expression across CA1 pyramidal nuclei (left) is fitted with regression and binarized (middle) to generate a profile (right) of the percentage of *Wfs1* expressing nuclei (greyscale) in each of the CA1 pyramidal sub-clusters. (**C**) Hippocampus spatial anatomy in a coronal sections. Left: The mouse hippocampus 3-D structure and the coronal section (brown plane) used in this analysis. Right: Schematics of the coronal section shown on the left. CA1 (including subiculum): orange; CA3 (including the hilus): green; DG: dark grey. M: medial, L: lateral, D: dorsal, V: ventral. Sub: Subiculum. (**D**) Registration of CA1 pyramidal sub-clusters to CA1 sub-regions. Left: correlation of each sub-cluster to CA1 sub-regions using landmark genes. Right: sub-cluster assignments (numbered arrows and color code). (**E**) Registration of CA3 pyramidal sub-clusters to CA3 sub-regions. Shown as

in (D). Dentate gyrus in gray is included in the schematic for spatial reference.

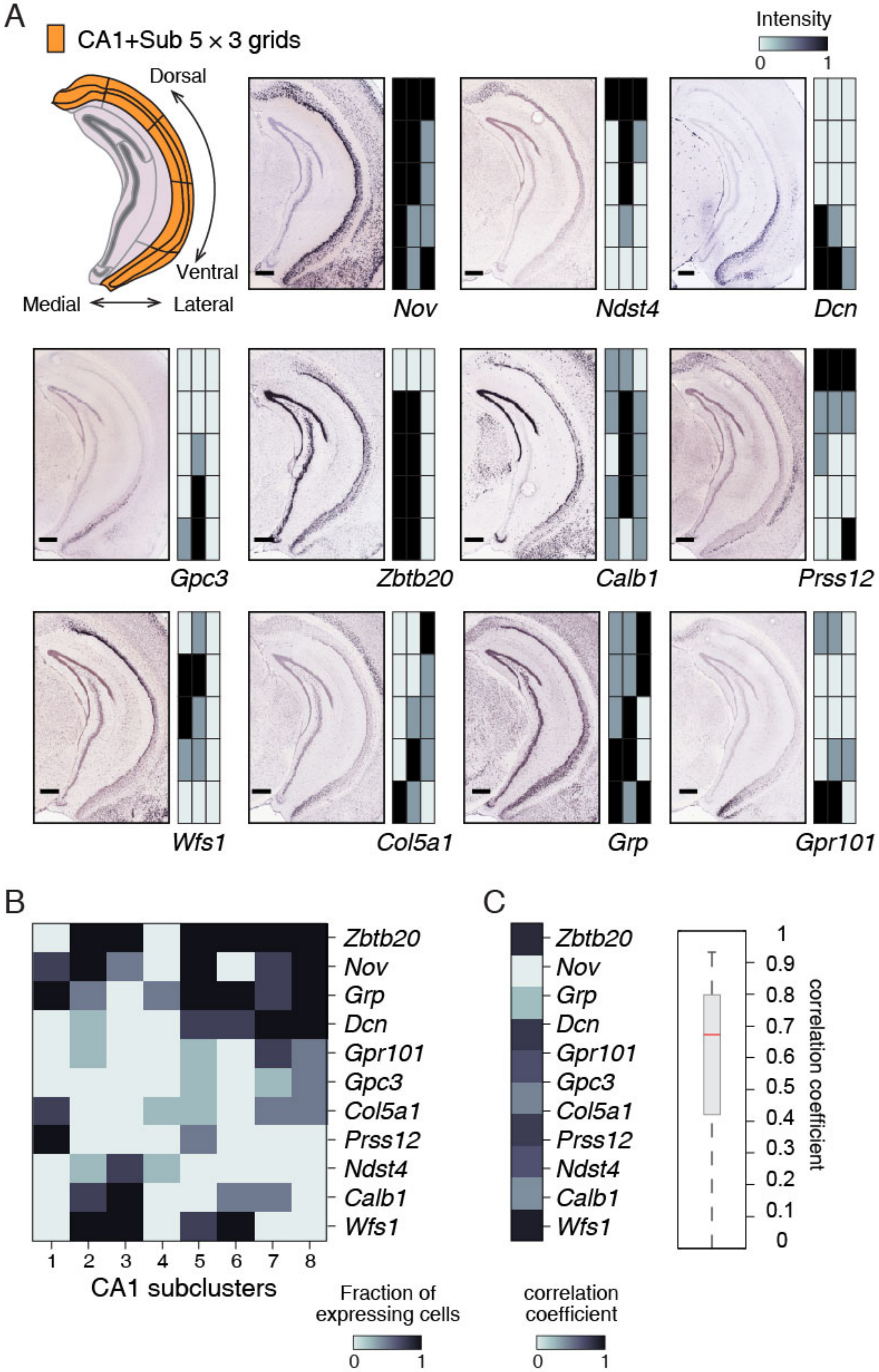
A
CA1+Sub 5 × 3 grids
Intensity
0
1
Dorsal
Ventral
Medial
Lateral
Nov
Ndst4
Dcn
Gpc3
Zbtb20
Calb1
Prss12
Wfs1
Col5a1
Grp
Gpr101
B
Zbtb20
Nov
Grp
Dcn
Gpr101
Gpc3
Col5a1
Prss12
Ndst4
Calb1
Wfs1
1 2 3 4 5 6 7 8
CA1 subclusters
Fraction of expressing cells
0
1
C
Zbtb20
Nov
Grp
Dcn
Gpr101
Gpc3
Col5a1
Prss12
Ndst4
Calb1
Wfs1
correlation coefficient
0
1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
correlation coefficient

Figure S12: Spatial landmark genes in CA1. (**A**) Spatial landmark genes in the CA1. Top left: Schematics of the hippocampus marking the CA1 and subiculum grid (orange). Displaying for each landmark gene an ISH [10] image showing its expression pattern in CA1 (right) and a heatmap showing the quantification of ISH intensities across the grid (left). Scale bar: 400μm. (**B**) Expression of landmark genes across the CA1 pyramidal sub-clusters. Heatmap showing the fractions of nuclei expressing each landmark genes in each biSNE sub-cluster. (**C**) Expression intensity of landmark genes in ISH [10] correlates with expression intensity predicted using Nuc-Seq data. Displayed in heat map (left) and box plot (right). In box plot, the median (red), 75% and 25% quantile (box), error bars (dashed lines).

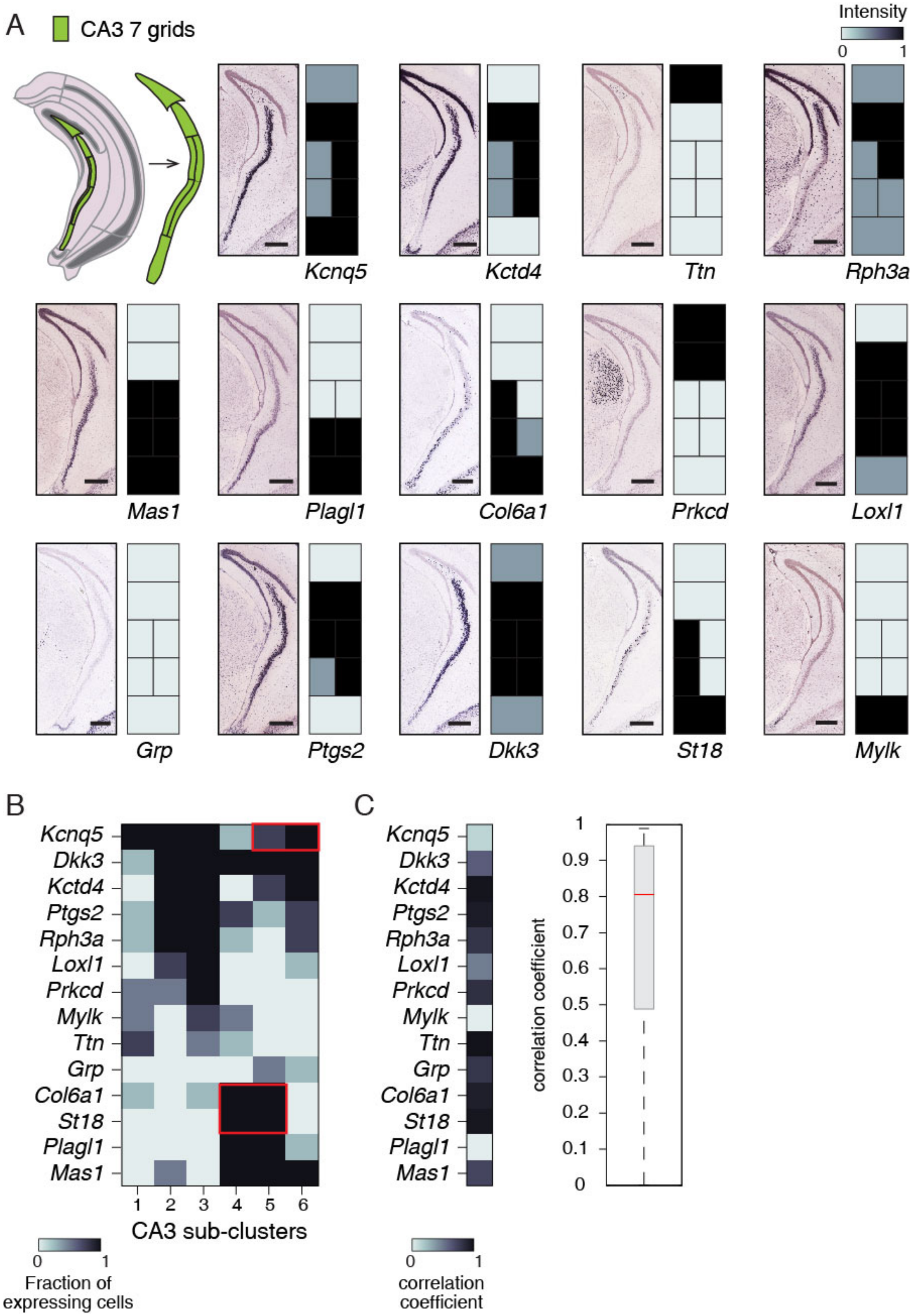
A
CA3 7 grids
Intensity
0
1
Kcnq5
Kctd4
Ttn
Rph3a
Mas1
Plagl1
Col6a1
Prkcd
Loxl1
Grp
Ptgs2
Dkk3
St18
Mylk
B
Kcnq5
Dkk3
Kctd4
Ptgs2
Rph3a
Loxl1
Prkcd
Mylk
Ttn
Grp
Col6a1
St18
Plagl1
Mas1
1 2 3 4 5 6
CA3 sub-clusters
0
1
Fraction of expressing cells
C
Kcnq5
Dkk3
Kctd4
Ptgs2
Rph3a
Loxl1
Prkcd
Mylk
Ttn
Grp
Col6a1
St18
Plagl1
Mas1
0
1
correlation coefficient
correlation coefficient
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Figure S13: Spatial landmark genes in CA3. (**A**) Spatial landmark genes in the CA3. Top left: Schematics of the hippocampus marking the CA3 grid (green). Displaying for each landmark gene an ISH [10] image showing its expression pattern in CA3 (right) and a heatmap showing the quantification of ISH intensities across the grid (left). Scale bar: 400µm. (**B**) Expression of landmark genes across the CA3 pyramidal sub-clusters. Heatmap showing the fractions of nuclei expressing each landmark genes in each biSNE sub-cluster. Marking the differentially expressed landmark genes in CA3.4, 5, 6 sub-clusters (red box). (**C**) Expression intensity of landmark genes in ISH [10] correlates with expression intensity predicted using Nuc-Seq data. Displayed in heat map (left) and box plot (right). In box plot, the median (red), 75% and 25% quantile (box), error bars (dashed lines).

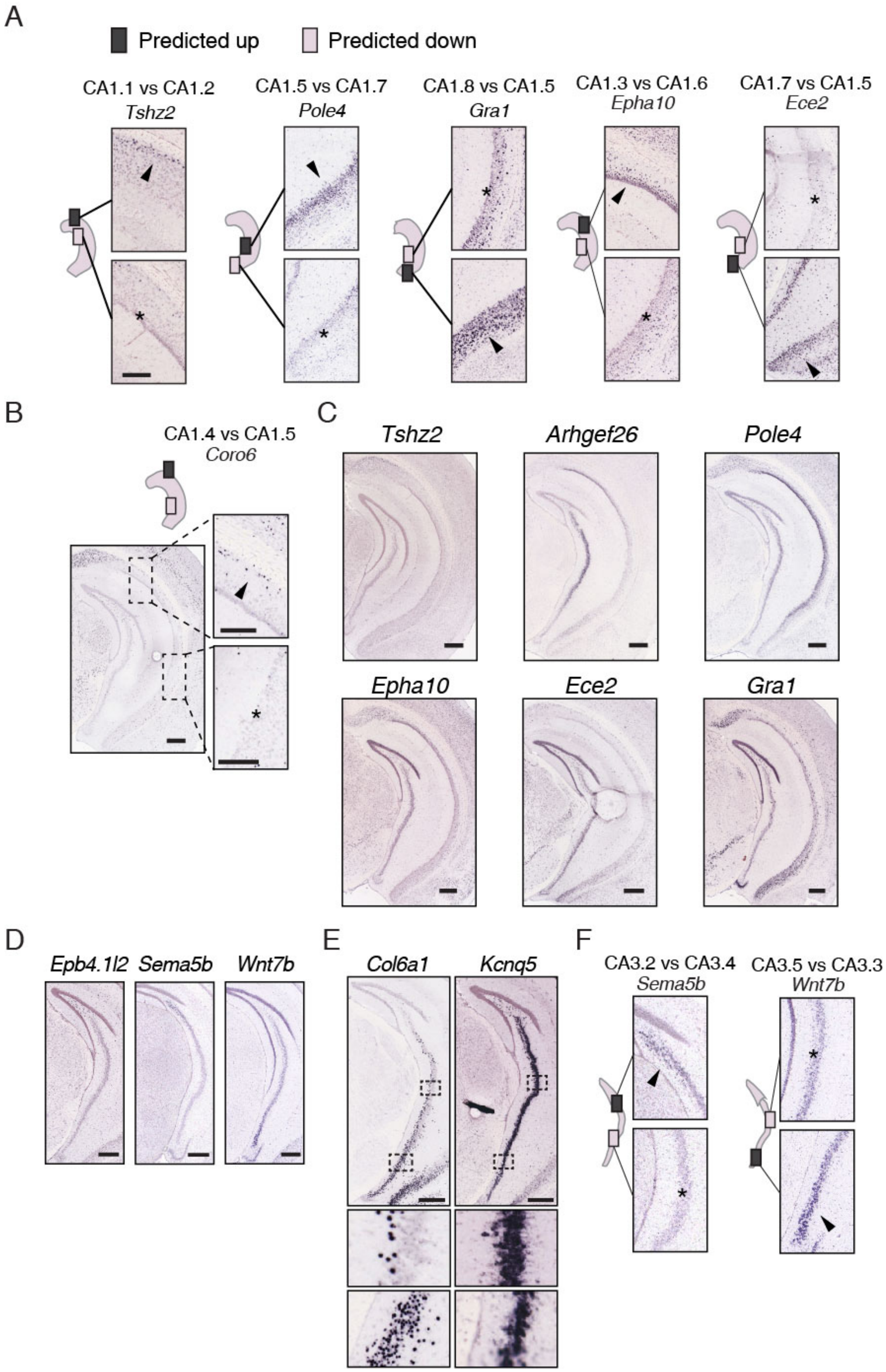
A
■ Predicted up
□ Predicted down
CA1.1 vs CA1.2
Tshz2
CA1.5 vs CA1.7
Pole4
CA1.8 vs CA1.5
Gra1
CA1.3 vs CA1.6
Epha10
CA1.7 vs CA1.5
Ece2
B
CA1.4 vs CA1.5
Coro6
C
Tshz2
Arhgef26
Pole4
Epha10
Ece2
Gra1
D
Epb4.1l2
Sema5b
Wnt7b
E
Col6a1
Kcnq5
F
CA3.2 vs CA3.4
Sema5b
CA3.5 vs CA3.3
Wnt7b

Figure S14: Examples of CA1 and CA3 predicted spatial expression. (**A-B**) Validation of spatial assignments in the CA1 pyramidal sub-clusters (denoted as CA1.1,...,CA1.8). Predictions (left illustrations; dark and light boxes showing predicted differential expression regions) match well with Allen ISH [10] images (right; arrowhead: high expression; asterisk: depletion) in pairwise comparison of genes differentially expressed between two sub-clusters (genes and clusters labeled on top). (**C**) ISH image [10] of the example genes showing the entire view of dorsal-ventral CA1. Scale bar: 400µm. (**D**) ISH image [10] of the example genes showing the entire view of dorsal-ventral CA3. Scale bar: 400µm. (**E**) Restricted spatial expression pattern in ventral CA3 of *Col6a1* and *Kcnq5*. Showing ISH [10] images. Top: entire view of CA3. Middle: view of region marked by the upper dashed box. Bottom: view of the region marked by the lower dashed box. (**F**) Validation of spatial assignments in the CA3 pyramidal sub-clusters. Shown as in (A).

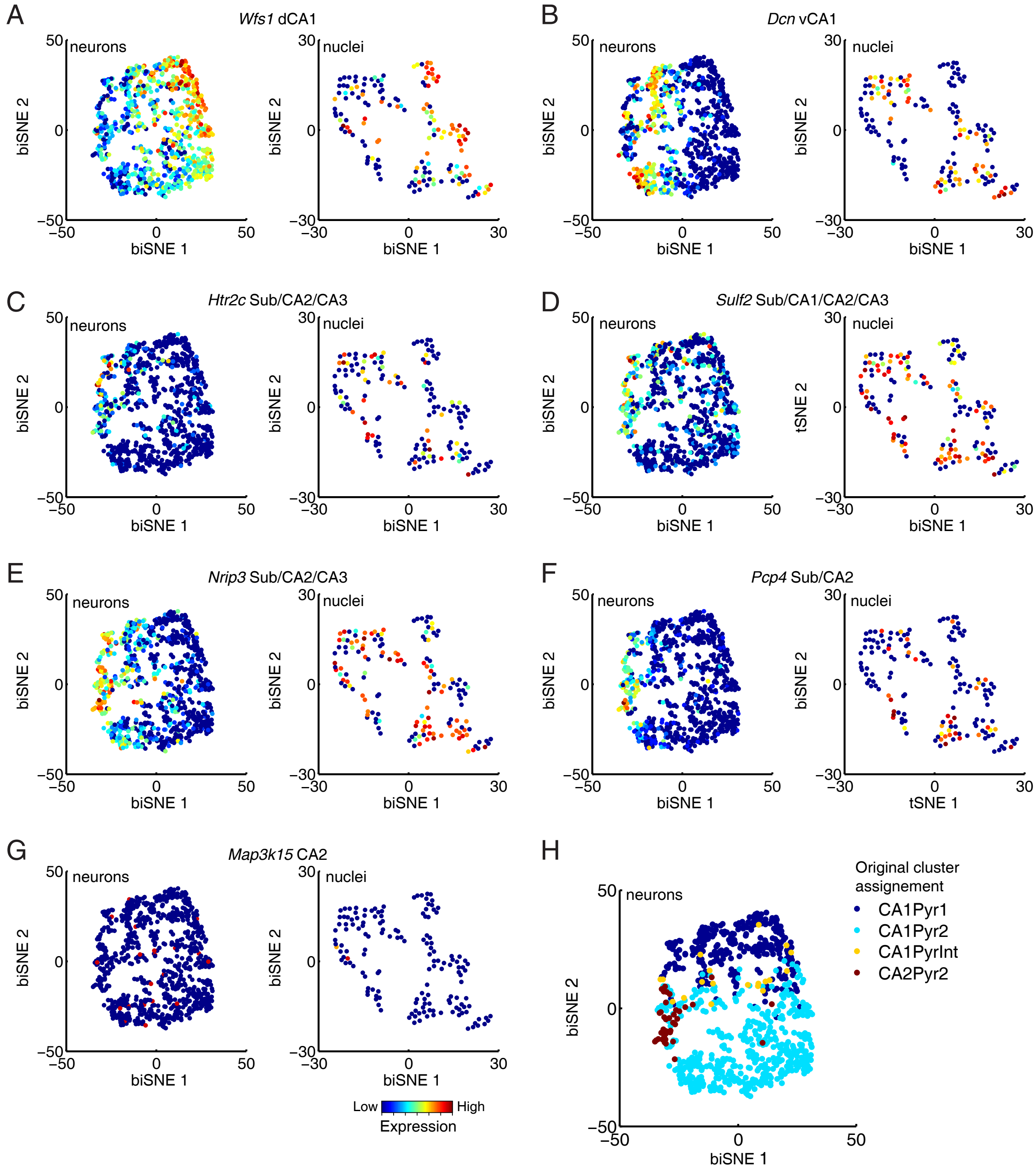

Figure S15: Clustering of CA1 pyramidal neurons from published single cell RNA-Seq data. (**A-F**) Nuc-Seq and biSNE improve cell sub-type classification of CA1 pyramidal neurons compared to

single neuron RNA-seq. Pairwise comparison of the expression levels of spatial landmark genes across 2-D biSNE embedding of of CA1 pyramidal neurons (left, data from single neurons RNA-seq [1]) and Nuc-Seq (right), showing the relative expression level of the gene (color scale). The expression of each gene is not restricted to any sub-cluster in the single neuron data [1], but is restricted to distinct sub-clusters in Nuc-Seq data. biSNE identified differential genes that have localized expression pattern the 2-D embedding of the single neuron RNA-Seq data. On top of each pair of plots, the anatomical region where the expression pattern of this gene is restricted to (identified in ISH [10]) is marked on the left, and the gene name on the right. dCA1: dorsal CA1; vCA1: ventral CA1; Sub: Subiculum. (**G**) A CA2 landmark gene *Map3k15* is not selected by biSNE and does not have localized expression pattern in the 2-D embedding of the CA1 pyramidal neurons from the single cell RNA-Seq data. 2-D embedding of CA1 pyramidal neurons (Left: data from [10]) and nuclei (Right: data from Nuc-Seq) showing the relative expression level of the gene (as in A). (**H**) 2-D embedding of the CA1 neurons showing the original assignment to 4 sub-clusters identified in [10] and denoted by colors. CA1Pyr1: CA1 pyramidal neuron type 1; CA2Pyr2: CA1 pyramidal neuron type 2; CA1PyrInt: CA1 pyramidal intermediate; CA2Pyr2: CA2 pyramidal neuron.

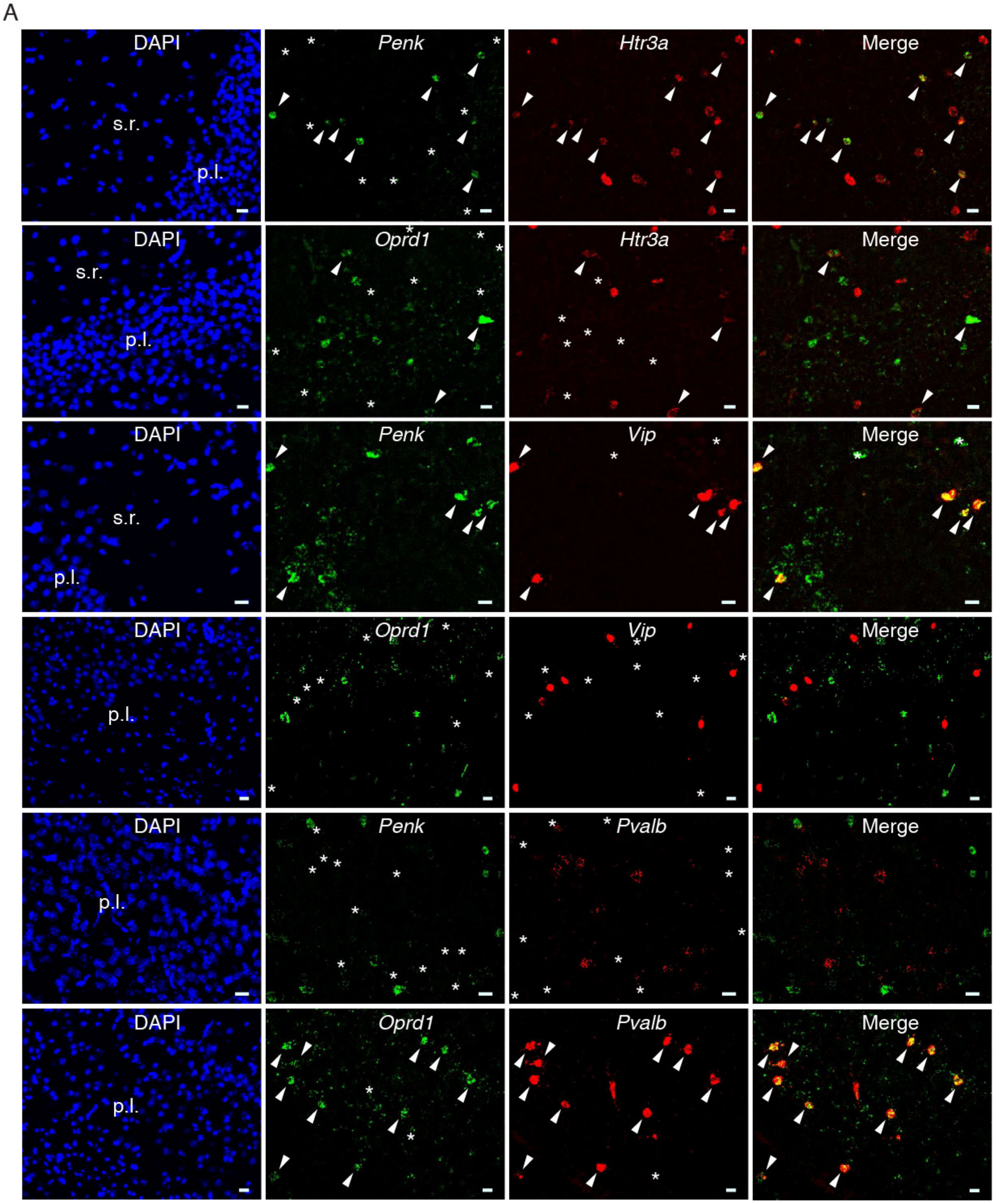

Figure S16: Double FISH validates mutual exclusive expression of *Penk* and *Oprd1*.

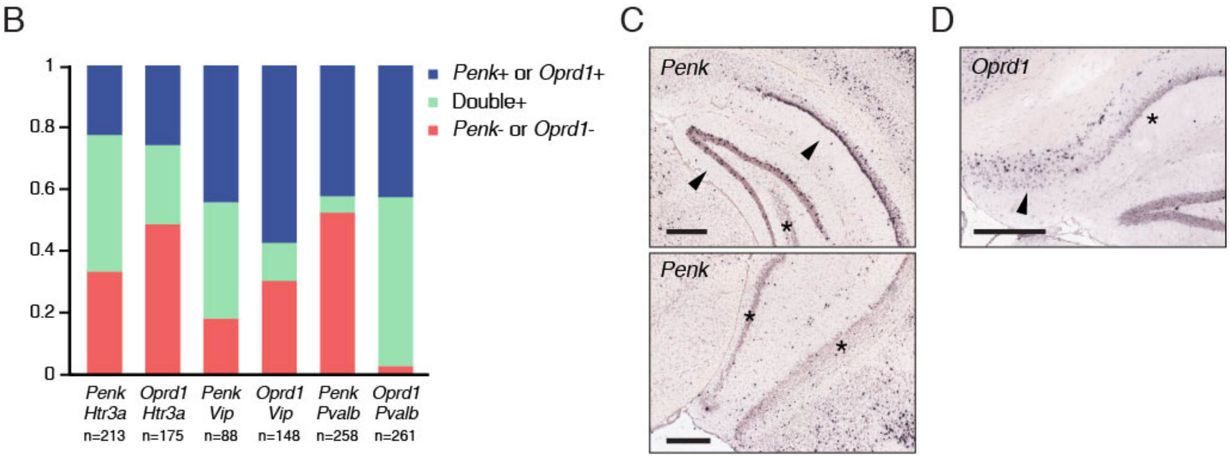

Figure S16: ***Continued*** **Double FISH validates mutual exclusive expression of *Penk* and *Oprd1*.** (**A**) From Top to Bottom: dFISH *Penk*/*Oprd1* (green) and *Htr3a*/*Vip*/*Pvalb* (red) showing expressions of *Penk* and *Htr3a* are partially overlapped; expressions of *Oprd1* and *Htr3a* are mostly not overlapped; expressions of *Penk* and *Vip* are largely overlapped; expressions of *Oprd1* and *Vip* are not overlapped; expressions of *Penk* and *Pvalb* are not overlapped; expressions of *Oprd1* and *Pvalb* are largely overlapped. Scale bar: 20µm. s.r. – stratum radiatum; p.l. – pyramidal layer. (**B**) Quantification of dFISH images. Bar plots showing the percent of single and double labeled cells in FISH images for each pair of genes. (**C**) Allen ISH [10] image of *Penk* gene with view of the upper DG (top) and the lower DG (bottom) shows its expression pattern in the dorsal CA1 and DG (arrowheads) and its depletion in ventral CA1. Scale bar: 400µm. (**D**) Allen ISH [10] image of *Oprd1* gene (as in B) shows its expression in the subiculum (arrowheads) and its depletion in the dorsal CA1 and DG regions (asterisks). Scale bar: 400µm.

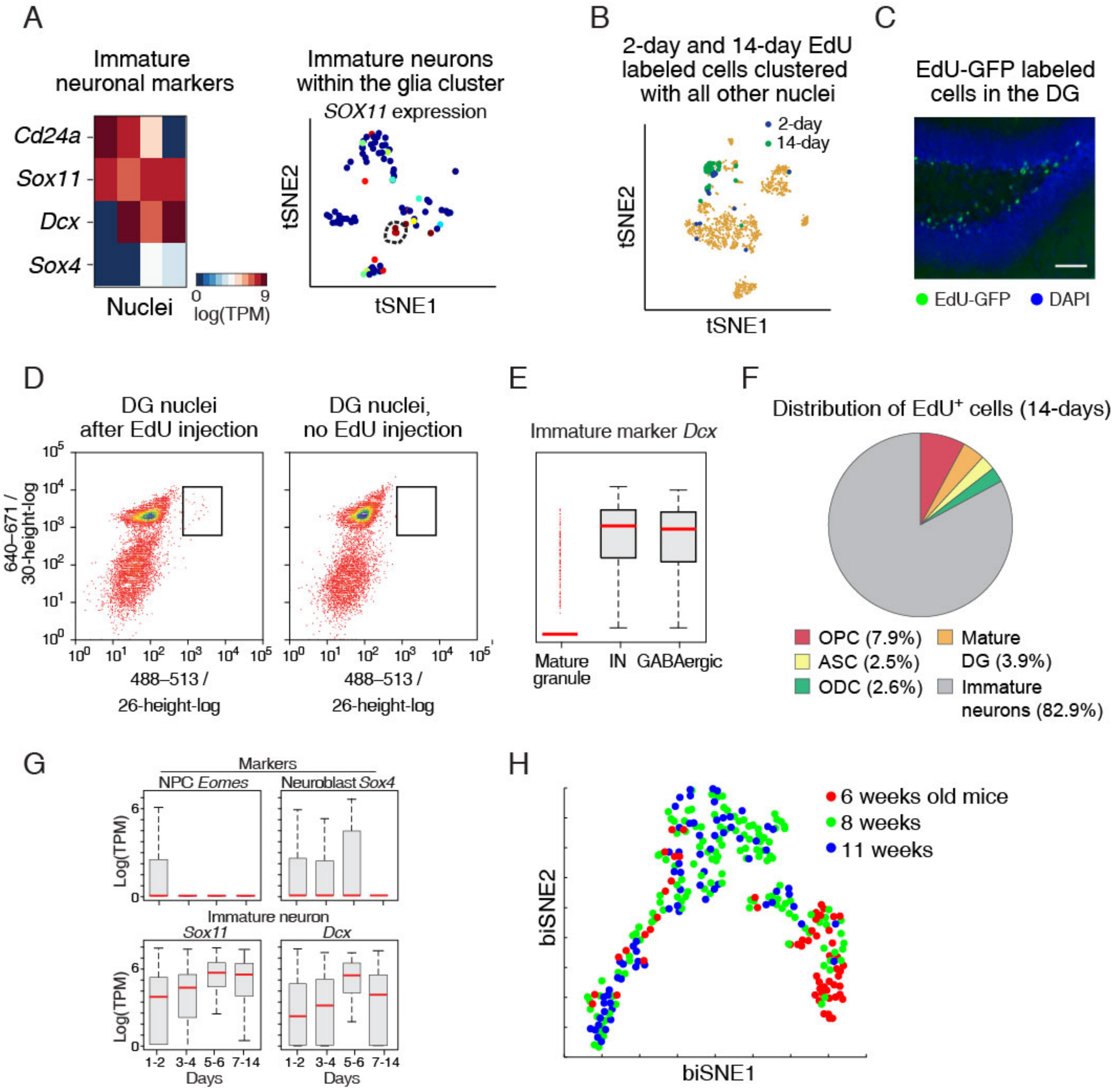
A
Immature neuronal markers
Cd24a
Sox11
Dcx
Sox4
Nuclei
0
9
log(TPM)
Immature neurons within the glia cluster
SOX11 expression
tSNE2
tSNE1
B
2-day and 14-day EdU labeled cells clustered with all other nuclei
2-day
14-day
tSNE2
tSNE1
C
EdU-GFP labeled cells in the DG
EdU-GFP
DAPI
D
DG nuclei after EdU injection
DG nuclei, no EdU injection
640–671 / 30-height-log
488–513 / 26-height-log
10^0
10^1
10^2
10^3
10^4
10^5
E
Immature marker Dcx
Mature granule
IN
GABAergic
F
Distribution of EdU$^+$ cells (14-days)
OPC (7.9%)
ASC (2.5%)
ODC (2.6%)
Mature DG (3.9%)
Immature neurons (82.9%)
G
Markers
NPC Eomes
Neuroblast Sox4
Immature neuron
Sox11
Dcx
Log(TPM)
0
6
1-2
3-4
5-6
7-14
Days
H
6 weeks old mice
8 weeks
11 weeks
biSNE2
biSNE1

Figure S17: Nuc-seq combined with labeling of dividing cells (Div-Seq) profiles adult newborn precursors and neurons. (**A**) Cells expressing immature neuronal markers with EdU tagging. Left: heatmap showing 4 nuclei expressing immature neuronal marker genes: *Sox4*, *Dcx*, *Sox11*, and *Cd24a*. Right: 2-D embedding of the glial cluster of nuclei (from **Fig. 1B**), clustered as in fig. S6A colored by the expression level of Sox11 gene. These nuclei are marked in the 2-D embedding of glial like cells as in fig. S6A (black dashed circle) (**B**) EdU labeled cells cluster separately from other cells. Shown is a biSNE 2-D embedding of all nuclei including the EdU labeled nuclei extracted after 2-day and 14-day post labeling. Most labeled nuclei form a distinct cluster. (**C**) EdU labeling tagged cells in the subgranular zone (SGZ) region. Shown are EdU staining (GFP click chemistry) and DAPI staining (blue) of tissue slice two weeks post EdU injections. (**D**) FACS sorting of EdU labeled nuclei. Shown is a scatter plot of log GFP intensity (X axis) and the log ruby-dye intensity (Y axis) from FACS of nuclei isolated two days after EdU injection (left) and with no EdU injections (right). Both samples were treated with click chemistry as in B. (**E**) *Dcx*, a commonly used marker for immature neurons, was expressed in all mature GABAergic neurons in the hippocampus, highlighting the limits of using single marker genes to identify cell types. Box plots showing expression levels of the *Dcx* gene across mature granule neurons, immature neurons (EdU labeled) and GABAergic neurons. In box plots, the median (red), 75% and 25% quantile (box), error bars (dashed lines), and outliers (red dots). (**F**) Most of the 14 days EdU labeled nuclei are immature neurons. Shown is the distribution of 14 days EdU labeled nuclei across cell types, assigned by clustering (as in B) and marker gene expression: Oligodendrocyte precursor cells, OPC; dentate gyrus granule cells, DG; Astrocytes, ASC; Oligodendrocytes, ODC. (**G**) Div-Seq captured cells expressing known markers of neuronal precursors, neuroblasts and immature neurons. Box plots for the 1-14 days EdU labeled nuclei (excluding nuclei classified as non-neuronal). (**H**) Newborn neurons cluster along a continuous trajectory independent of animal age. Data includes nuclei from 6, 8 and 11 weeks old mice. Showing 2-D embedding of 1-14 days EdU labeled nuclei colored by animal age.

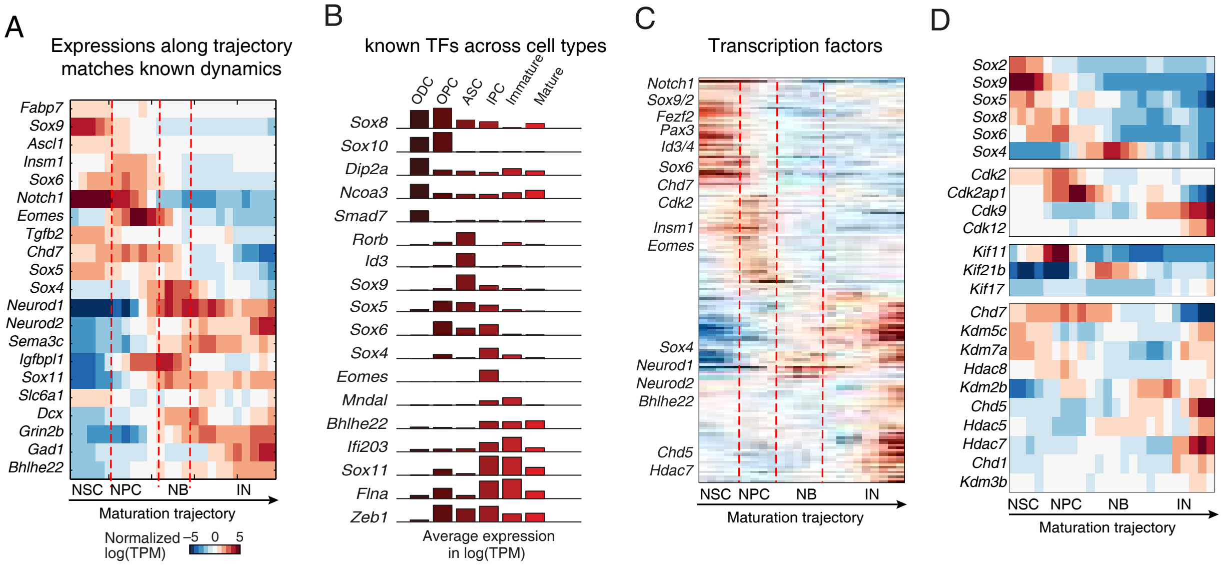

Figure S18: Transcriptional and epigenetic switch during adult neurogenesis and neuronal maturation in the DG. (**A**) Expression of known marker genes along the trajectory matches the expected dynamics. Left: Heatmap of the expression of the markers and related genes (rows), sorted by their expected pattern, along the neurogenesis trajectory (columns, running average along the trajectory). Data in log(TPM+1). (**B**) Expression level of known transcription factors (TF) across cell types, showing known regulators of each cell type. Shown are the relative average expression levels (bars) across cells. (**C**) Dynamically regulated TFs and chromatin regulator. Heatmap of the running average expression (log(TPM+1)) of the regulators (rows) along the trajectory (columns). Genes are sorted based on their cluster identities (as in **Fig. 3D**). Red lines mark the transition from neuronal precursor cells (NPCs) to neuroblast (NB) and from NB to immature neurons. (**D**) Examples of dynamic expression patterns of families of regulators. Top: Sox family genes. Middle: Cyclin (Cdk) genes. Bottom: kinesin superfamily.

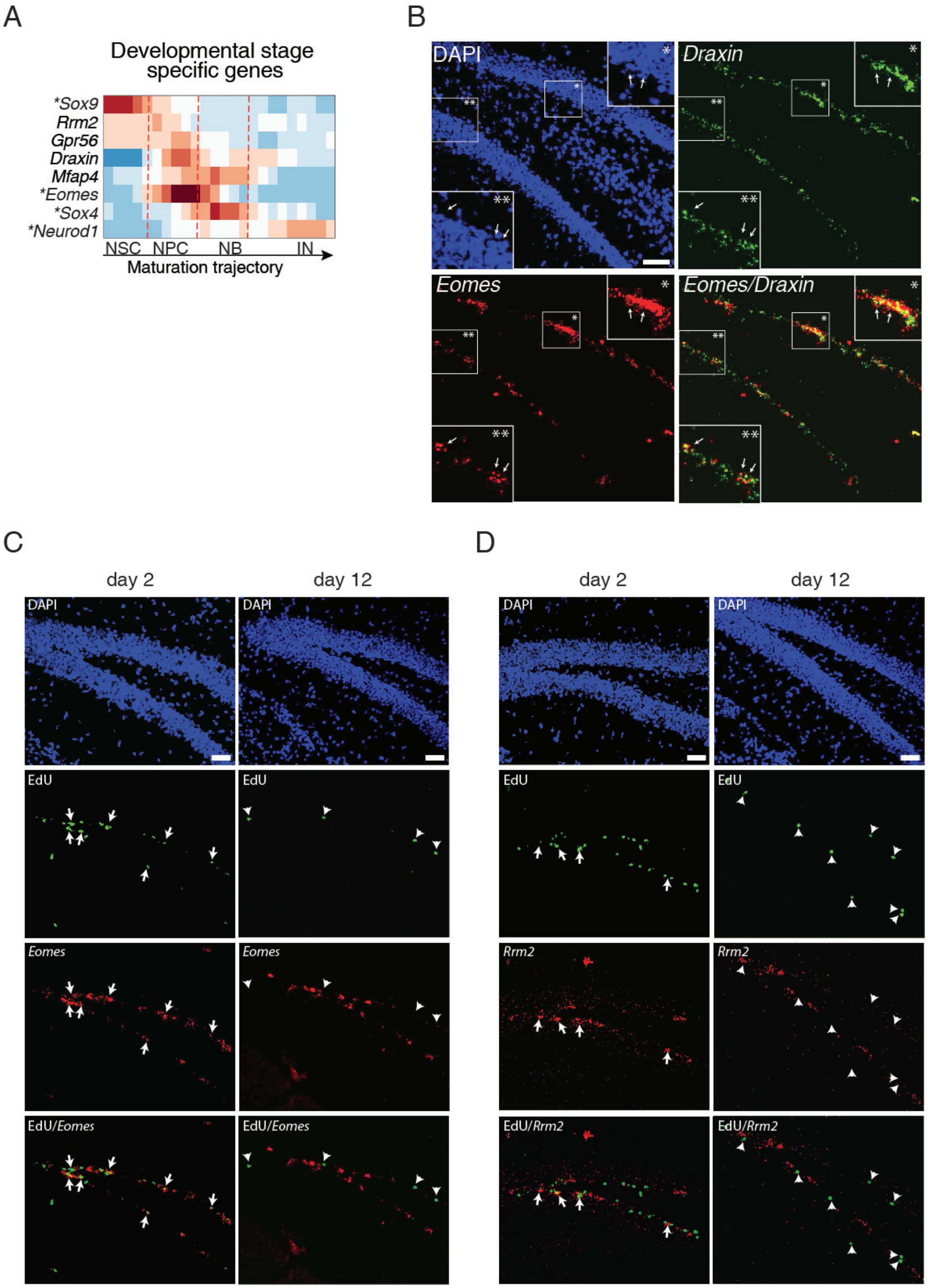
A
Developmental stage specific genes
*Sox9
Rrm2
Gpr56
Draxin
Mfap4
*Eomes
*Sox4
*Neurod1
NSC
NPC
NB
IN
Maturation trajectory
B
DAPI
Draxin
Eomes
Eomes/Draxin
*
**
C
day 2
day 12
DAPI
EdU
Eomes
EdU/Eomes
D
day 2
day 12
DAPI
EdU
Rrm2
EdU/Rrm2

Figure S19: Tissue validation of markers of immature neurons in hippocampus. (**A**) Heatmap of running average expression of genes along the DG maturation trajectory, showing known and novel stage specific gene expressions. Known marker genes are marked by asterisks. (**B**) Coronal sections of adult mouse dentate gyrus stained with co-FISH of *Draxin* (green) and *Eomes* (red). Cell nuclei were labeled with DAPI. Insets show higher magnifications of the boxed areas indicated with asterisks. Overlaps of *Draxin* and *Eomes* are indicated with arrowheads. (**C**) and (**D**) Coronal sections of the adult mouse dentate gyrus stained with EdU labeling (green) and FISH (red) of *Eomes* (C) or *Rrm2* (D) at 2 and 12 days after intraperitoneal (i.p.) EdU injection. Cell nuclei were labeled with DAPI. Overlaps of *Eomes* and *Rrm2* with EdU are found at 2 days (arrows) but not at 12 days (arrowhead) post i.p. EdU injection. Scale bars: 50 µm.

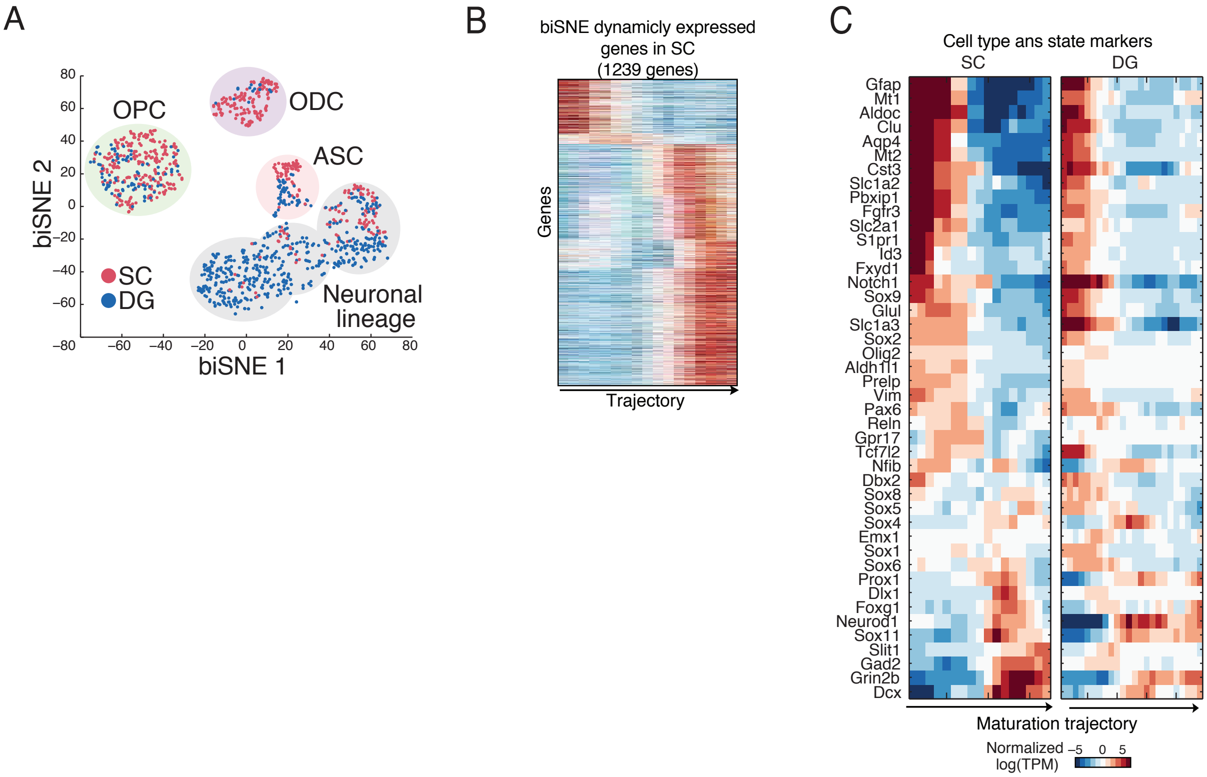

Figure S20: Dynamic expression of genes during the SC adult neurogenesis. (**A**) Nuclei cluster primarily by cell type and maturation state and secondarily by region. Shown is biSNE 2-D embedding of cells from SC and the DG. Nuclei are colored by tissue. (**B**) Dynamic gene expression clusters along the SC newborn neuronal maturation trajectory. Four clusters are shown from top to bottom, presented as a heatmap of running average expression of all genes along the trajectory (n = 1,239 genes). (**C**) Heatmap of running average expression of known cell stage and cell type marker genes along the trajectory in the SC (left) and DG (right).

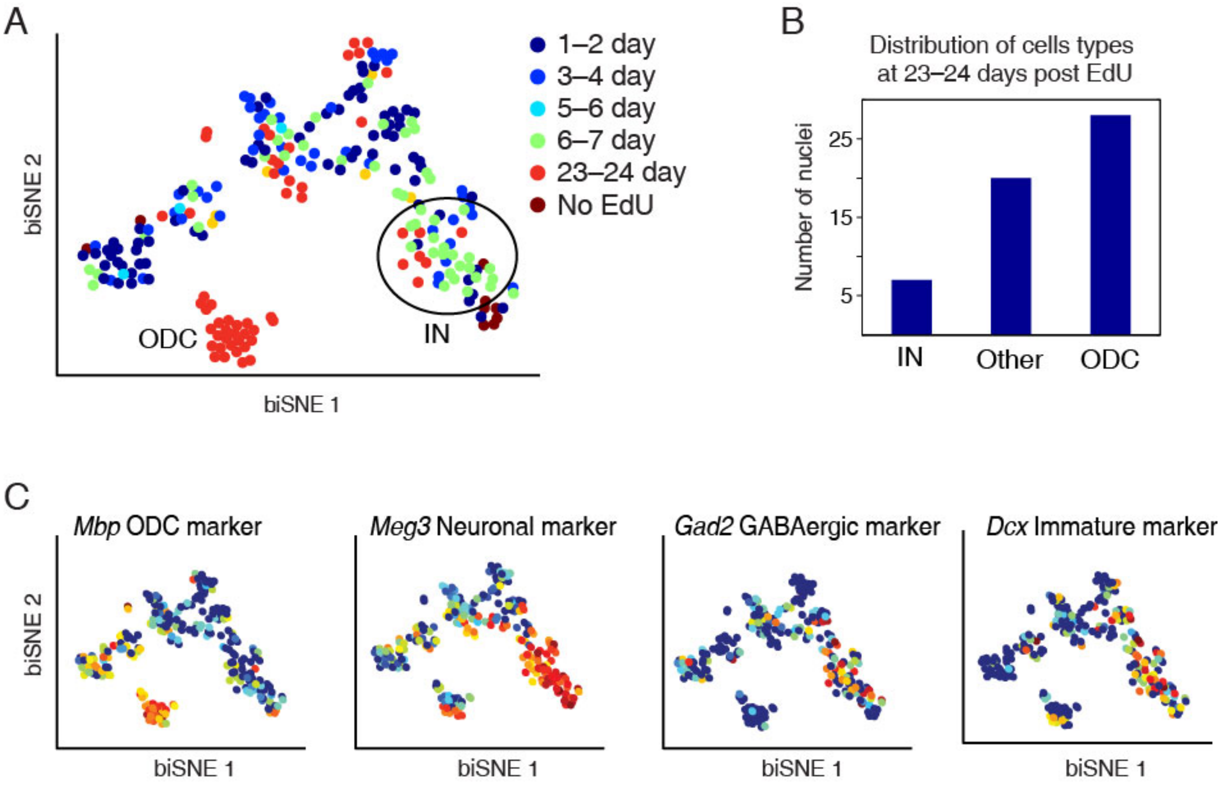

Figure S21: Survival of newborn neurons in the SC. (**A**) 23-24 days post EdU nuclei embedded into the 2-D clustering of neuronal lineage genes (from **Fig. 4C**). Showing a set (10%) of nuclei that cluster with the immature neuronal nuclei along the trajectory. (**B**) Bar plot showing the number of nuclei classified as oligodendrocytes (ODC), immature/young neurons (IN) or other cells types. (**C**) Marker genes expressed along the combined neuronal and 23-24 days EdU labeled nuclei trajectory. From left to right: *Mbp* oligodendrocyte marker, *Meg3* neuronal marker, *Gad2* GABAergic marker, and *Dcx* immature neuronal marker.

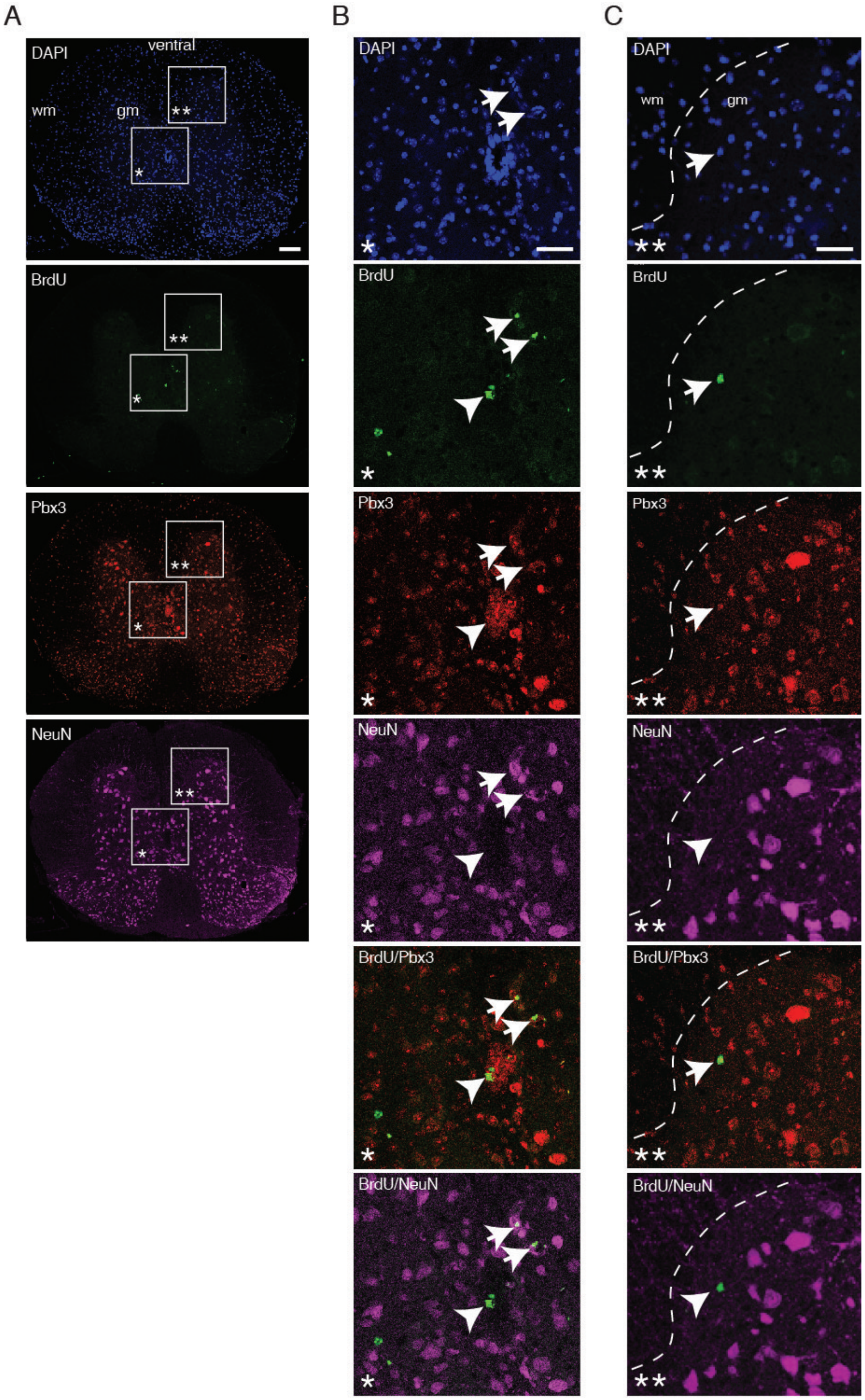
A
B
C
DAPI
ventral
wm
gm
**
*
BrdU
**
*
Pbx3
**
*
NeuN
**
*
DAPI
*
BrdU
*
Pbx3
*
NeuN
*
BrdU/Pbx3
*
BrdU/NeuN
*
DAPI
wm
gm
**
BrdU
**
Pbx3
**
NeuN
**
BrdU/Pbx3
**
BrdU/NeuN
**

Figure S22: Tissue validation of Pbx3 expression in newborn cells of the spinal cord. Cross section of adult mouse spinal cord 8 days after intraperitoneal BrdU injection stained with anti-BrdU (green), Pbx3 (red) and NeuN (magenta) antibodies. Cell nuclei were labeled with DAPI. (**A**) Overview of spinal cord shows spare BrdU labeling in grey matter (gm) and white matter (wm). B and C) Higher magnifications of insets shown in (A) as indicated by asterisks. (**B**) Overlap of BrdU, Pbx3 and NeuN in newborn cells proximate to the central canal (arrows). Overlap of BrdU and Pbx3 but not NeuN in a newborn cell within the central canal ependymal cell layer (arrowhead). (**C**) Overlap of BrdU and Pbx3 (arrow) but not NeuN (arrowhead) in a newborn cell at the border between gm and wm (indicated by dotted line). Scale bars: 50 µm.

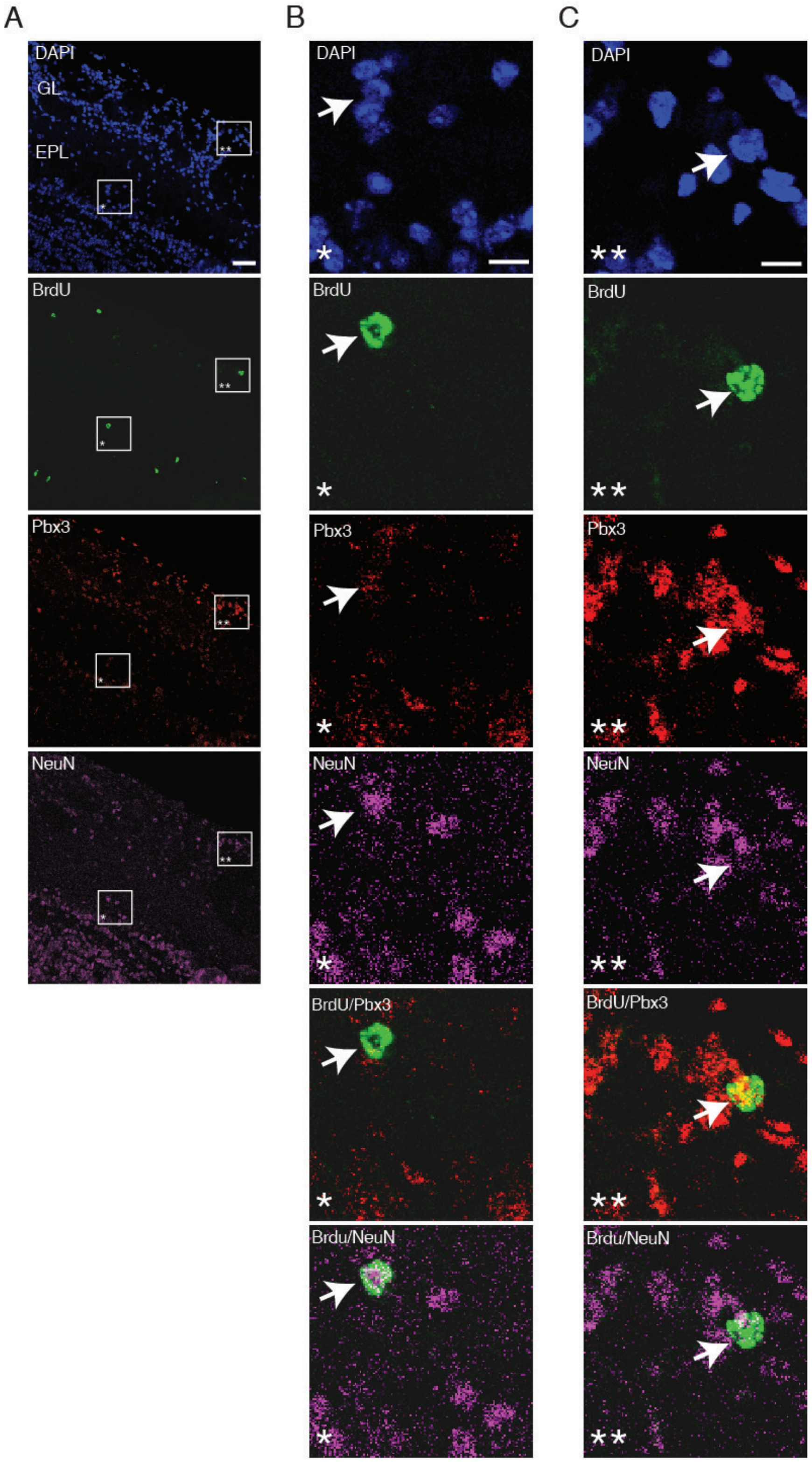
A
B
C
DAPI
GL
EPL
**
*
BrdU
Pbx3
NeuN
BrdU/Pbx3
Brdu/NeuN

Figure S23: Tissue validation of Pbx3 expression in newborn cells of the olfactory bulb. Sagittal section of adult mouse olfactory bulb 8 days after intraperitoneal BrdU injection stained with anti-BrdU (green), Pbx3 (red) and NeuN (magenta) antibodies. Cell nuclei were labeled with DAPI. (**A**) Overview of olfactory bulb shows spare BrdU labeling in the glomerular layer (GL) and external plexiform layer (EPL). (**B** and **C**) Higher magnifications of insets shown in (A) as indicated by asterisks. Overlap of BrdU, Pbx3 and NeuN in newborn cells are shown (arrows). Scale bars: 40 µm (A) and 10 µm (B,C).

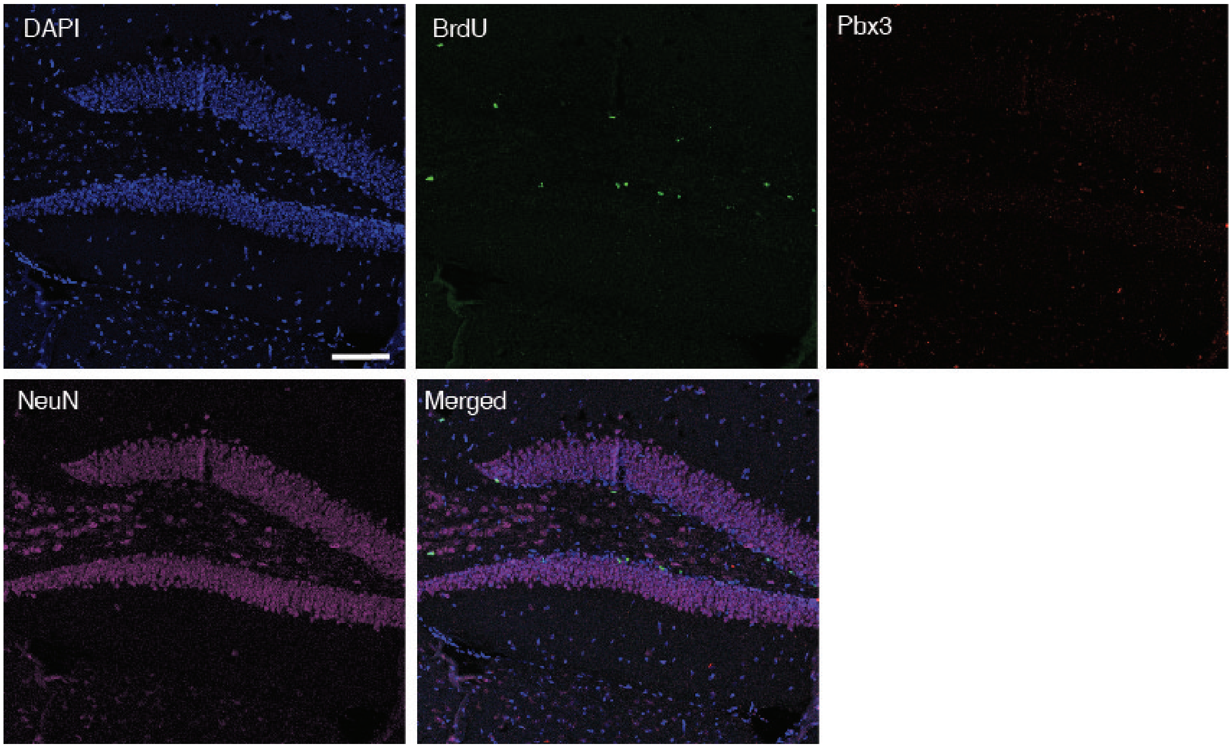

Figure S24: Tissue validation of Pbx3 expression in newborn cells of the dentate gyrus. Sagittal section of adult mouse hippocampus 8 days after intraperitoneal BrdU injection stained with anti-BrdU (green), Pbx3 (red) and NeuN (magenta) antibodies. No detectable Pbx3 expression levels in the dentate gyrus and no overlap of Pbx3 and NeuN with newborn cells. Same confocal microscope settings have been used as in fig. S22 and fig. S23. Scale bar: 100µm.

SUPPLEMENTARY TABLES

The supplementary tables are provided online as separate Excel files.

Table S1: Marker genes for the major cell types in the adult hippocampus.

Table S2: Marker genes for the GABAergic sub-clusters.

Table S3: Differential genes between CA pyramidal neuron sub-clasters.

Table S4: Information of tissue samples and animal treatments.

Table S5: Dynamically regulated genes along the neurogensis trajectory.

Table S6: Differential gene splice isoforms between immature and mature neurons in the DG.

Table S7: Differential genes between immature neurons from the SC, OB, and DG.

SUPPLEMENTARY REFERENCES

[1] A. Zeisel, *et al.*, Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq, *Science* **347**, 1138 (2015).

[2] J. Shin, *et al.*, Single-cell rna-seq with waterfall reveals molecular cascades underlying adult neurogenesis, *Cell stem cell* **17**, 360 (2015).

[3] B. Tasic, *et al.*, Adult mouse cortical cell taxonomy revealed by single cell transcriptomics, *Nature neuroscience* (2016).

[4] G. li Ming, H. Song, Adult neurogenesis in the mammalian brain: Significant answers and significant questions, *Neuron* **70**, 687 (2011).

[5] D. L. Moore, G. A. Pilz, M. J. Arauzo-Bravo, Y. Barral, S. Jessberger, A mechanism for the segregation of age in mammalian neural stem cells, *Science* **349**, 1334 (2015).

[6] Materials and methods are available as supplementary materials on science online.

[7] B. Lacar, *et al.*, Nuclear RNA-seq of single neurons reveals molecular signatures of activation, *Nature Communications* **7**, 11022 (2016).

[8] L. Swiech, *et al.*, In vivo interrogation of gene function in the mammalian brain using crispr-cas9, *Nature biotechnology* **33**, 102 (2015).

[9] H. Hu, J. Gan, P. Jonas, Fast-spiking, parvalbumin(+) GABAergic interneurons: From cellular design to microcircuit function, *Science* **345**, 1255263 (2014).

[10] E. S. Lein, *et al.*, Genome-wide atlas of gene expression in the adult mouse brain, *Nature* **445**, 168 (2007).

[11] Y. Zhang, *et al.*, An rna-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex, *The Journal of Neuroscience* **34**, 11929 (2014).

[12] M. S. Cembrowski, *et al.*, Spatial gene-expression gradients underlie prominent heterogeneity of CA1 pyramidal neurons, *Neuron* **89**, 351 (2016).

[13] B. P. Roques, M.-C. Fournié-Zaluski, M. Wurm, Inhibiting the breakdown of endogenous opioids and cannabinoids to alleviate pain, *Nature Reviews Drug Discovery* **11**, 292 (2012).

[14] E. Llorens-Bobadilla, *et al.*, Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury, *Cell Stem Cell* **17**, 329 (2015).

[15] M. Schouten, M. R. Buijink, P. J. Lucassen, C. P. Fitzsimons, New neurons in aging brains: Molecular control by small non-coding RNAs, *Front. Neurosci.* **6** (2012).

[16] D. M. Feliciano, A. Bordey, L. Bonfanti, Noncanonical sites of adult neurogenesis in the mammalian brain, *Cold Spring Harbor Perspectives in Biology* **7**, a018846 (2015).

[17] P. J. Horner, *et al.*, Proliferation and differentiation of progenitor cells throughout the intact adult rat spinal cord, *The Journal of Neuroscience* **20**, 2218 (2000).

[18] R. Shechter, Y. Ziv, M. Schwartz, New GABAergic interneurons supported by myelin-specific t cells are formed in intact adult spinal cord, *STEM CELLS* **25**, 2277 (2007).

[19] Z. Agoston, *et al.*, Meis2 is a pax6 co-factor in neurogenesis and dopaminergic periglomerular fate specification in the adult olfactory bulb, *Development* **141**, 28 (2014).

[20] C. A. Rottkamp, K. J. Lobur, C. L. Wladyka, A. K. Lucky, S. O'Gorman, Pbx3 is required for normal locomotion and dorsal horn development, *Developmental Biology* **314**, 23 (2008).

[21] S. Picelli, *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in single cells, *Nature methods* **10**, 1096 (2013).

[22] C. A. Paul, B. Beltz, J. Berger-Sweeney, Sectioning of brain tissues, *Cold Spring Harbor Protocols* **2008**, pdb (2008).

[23] H. Hideo, T. Keiko, Y. Nobuyuki, M. Tsuyoshi, Dissection of hippocampal dentate gyrus from adult mouse, *Journal of Visualized Experiments* (2009).

[24] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data, *Nature biotechnology* **33**, 495 (2015).

[25] C. M. Hempel, K. Sugino, S. B. Nelson, A manual method for the purification of fluorescently labeled neurons from the mammalian brain, *Nature protocols* **2**, 2924 (2007).

[26] C. Trapnell, L. Pachter, S. L. Salzberg, Tophat: discovering splice junctions with rna-seq, *Bioinformatics* **25**, 1105 (2009).

[27] C. Trapnell, *et al.*, Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nature biotechnology* **28**, 511 (2010).

[28] J. T. Robinson, *et al.*, Integrative genomics viewer, *Nature biotechnology* **29**, 24 (2011).

[29] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, C. N. Dewey, Rna-seq gene expression estimation with read mapping uncertainty, *Bioinformatics* **26**, 493 (2010).

[30] B. Langmead, S. L. Salzberg, Fast gapped-read alignment with bowtie 2, *Nature methods* **9**, 357 (2012).

[31] D. S. DeLuca, *et al.*, Rna-seqc: Rna-seq metrics for quality control and process optimization, *Bioinformatics* **28**, 1530 (2012).

[32] J. C. Dunn, Well-separated clusters and optimal fuzzy partitions, *Journal of cybernetics* **4**, 95 (1974).

[33] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**, 2323 (2000).

[34] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, *Journal of Machine Learning Research* **9**, 85 (2008).

[35] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, *Science* **344**, 1492 (2014).

[36] A. K. Shalek, *et al.*, Single cell rna seq reveals dynamic paracrine control of cellular variation, *Nature* **510**, 363 (2014).

[37] A. K. Shalek, *et al.*, Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells, *Nature* **498**, 236 (2013).

[38] M. D. Robinson, D. J. McCarthy, G. K. Smyth, edger: a bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics* **26**, 139 (2010).

[39] M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for rna-seq data with deseq2, *Genome Biol* **15**, 550 (2014).

[40] Z. I. Botev, J. F. Grotowski, D. P. Kroese, *et al.*, Kernel density estimation via diffusion, *The Annals of Statistics* **38**, 2916 (2010).

[41] P. V. Kharchenko, L. Silberstein, D. T. Scadden, Bayesian approach to single-cell differential expression analysis, *Nature methods* **11**, 740 (2014).

[42] E.-a. D. Amir, *et al.*, visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia, *Nature biotechnology* **31**, 545 (2013).

[43] E. Z. Macosko, *et al.*, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, *Cell* **161**, 1202 (2015).

[44] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300 (1995).

[45] P. A. Moran, Notes on continuous stochastic phenomena, *Biometrika* pp. 17–23 (1950).

[46] P. J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, *Journal of the American Statistical association* **88**, 1273 (1993).

[47] A. Edelman, N. R. Rao, Random matrix theory, *Acta Numerica* **14**, 233 (2005).

[48] E. Eden, D. Lipson, S. Yogev, Z. Yakhini, Discovering motifs in ranked lists of dna sequences, *PLoS Comput Biol* **3**, e39 (2007).

[49] Y. Maruyama, An alternative to moran's i for spatial autocorrelation, *arXiv preprint arXiv:1501.06260* (2015).

[50] L. Van Der Maaten, Accelerating t-sne using tree-based algorithms, *The Journal of Machine Learning Research* **15**, 3221 (2014).

[51] P. N. Yianilos, *SODA* (1993), vol. 93, pp. 311–321.

[52] S. Anders, W. Huber, Differential expression analysis for sequence count data, *Genome Biology* **11**, R106 (2010).

[53] P. Brennecke, *et al.*, Accounting for technical noise in single-cell rna-seq experiments, *Nature methods* **10**, 1093 (2013).

[54] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: A library for large linear classification, *The Journal of Machine Learning Research* **9**, 1871 (2008).

[55] K. Helsgaun, An effective implementation of the lin–kernighan traveling salesman heuristic, *European Journal of Operational Research* **126**, 106 (2000).

[56] K. Helsgaun, General k-opt submoves for the lin–kernighan tsp heuristic, *Mathematical Programming Computation* **1**, 119 (2009).

[57] C. Trapnell, *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, *Nature biotechnology* **32**, 381 (2014).

[58] S. C. Bendall, *et al.*, Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development, *Cell* **157**, 714 (2014).

[59] C. De Boor, A practical guide to splines, *Mathematics of Computation* (1978).

[60] A. Subramanian, *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545 (2005).

[61] M. Ashburner, *et al.*, Gene ontology: tool for the unification of biology, *Nature genetics* **25**, 25 (2000).

[62] B. A. Strange, M. P. Witter, E. S. Lein, E. I. Moser, Functional organization of the hippocampal longitudinal axis, *Nature Reviews Neuroscience* **15**, 655 (2014).

[63] J. J. Trombetta, *et al.*, Preparation of single-cell rna-seq libraries for next generation sequencing, *Current Protocols in Molecular Biology* pp. 4–22.