www.sciencemag.org/cgi/content/full/science.aaa1934/DC1

Supplementary Materials for

Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq

Amit Zeisel, Ana B. Muñoz Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler,* Sten Linnarsson*

*Corresponding author. E-mail: sten.linnarsson@ki.se (S.L.); jens.hjerling-leffler@ki.se (J.H.-L.)

Published 19 February 2015 on *Science* Express
DOI: 10.1126/science.aaa1934

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S11
References

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.aaa1934/DC1)

Tables S1 and S2 (as Excel files)

Materials and Methods

Animals

Wild type (CD-1) and transgenic mice (5HT3EGFP on a CD-1 background) between postnatal 21 and 31 days of both sexes were used. In the transgenic mouse line EGFP is expressed under the control of the Htr3a promoter (GENSAT project, Rockefeller University, NY). All experimental procedures performed followed the guidelines and recommendations of local animal protection legislation and were approved by the local committee for ethical experiments on laboratory animals (Stockholms Norra Djurförsöksetiska nämnd, Sweden)

Tissue dissociation

Somatosensory cortex (Bregma, AP: 1.54 to -1.82 mm), or CA1 hippocampal (Bregma, AP: -2.06 to -3.80 mm) brain regions (Fig. S11) were dissociated into a single cell suspension. Mice were deeply anesthetized with a mixture of ketamine/xylazine (80mg/kg; 10mg/kg), and the brain was quickly dissected and transferred to ice-cold oxygenated cutting solution (87 mM NaCl, 2.5 mM KCl, 1.25 mM NaH_2PO_4, 26 mM $NaHCO_3$, 75 mM sucrose, 20 mM glucose, 1 mM $CaCl_2$, and 2 mM $MgSO_4$) and kept in the same solution during sectioning on a vibratome (VT1200 S, Leica) in 300 μm thick slices. The area of interest was dissected from each slice, and the tissue was dissociated using the Papain dissociation system (Worthington) following the manufacturer's instructions. All the solutions were oxygenated for at least 10 minutes with a mixture of 5% CO_2 in O_2 (Labline). Oxygenation and a short time of dissection were crucial to keep a high rate of survival in the cell suspension. After this, the cell suspension obtained was filtered with 20 μm filter (Partec) and kept in cold HBSS solution (SIGMA) with 0.2% BSA and 0.3% glucose. Then the cells were immediately loaded in the C1 chip, or FACS sorted and then loaded in the chip.

FACS sorting

After dissociation of the somatosensory cortex in the $5HT3a^{EGFP}$ mouse, the $EGFP^{+}$ cells were FACS sorted using a BD FACSAria" III Cell Sorter B5/R3/V3 system. In the cortex of this transgenic line all the EGFP labeled cells are GABAergic interneurons derived from the caudal ganglionic eminence (*27*). A total of 91 (passed-QC) FACS sorted cortical interneurons were obtained in this way.

Cell capture and imaging

A cell suspension of either somatosensory S1 cortex or CA1 hippocampus with 600-1000 cells/μL was used. C1 Suspension Reagent was added (all 'C1' reagents were from Fluidigm, Inc.) in a ratio of 4 μL to every 7 μL cell suspension. 11 μL of the cell suspension mix was loaded on a C1 Single-Cell AutoPrep IFC microfluidic chip designed for 10- to 17-μm cells, and the chip was then processed on a Fluidigm C1 instrument using the '*mRNA Seq: Cell Load (1772x/1773x)*' script (30 min at 4°C). The plate was then transferred to an automated microscope (Nikon TE2000E), and a brightfield and EGFP fluorescence image (20× magnification) was acquired for each capture site using μManager (http://micro-manager.org/ (*28*)), which took <15 min.

Lysis, reverse transcription and PCR

The plate was returned to the lab where Lysis mix, RT mix and PCR mix (previously described (*11*)) were added to the chip. The plate was then placed in the

Fluidigm C1 instrument and the '*mRNA Seq: RT + Amp (1772x/1773x)*' script was executed, which took ~8.5 h and included lysis, reverse transcription and 21 cycles of PCR. When the run had finished, the amplified cDNA was harvested in a total of 13 µL C1 Harvesting Reagent and the quality of cDNA was assessed on an Agilent BioAnalyzer. The typical yield was 1 ng/µL.

Quality control and selection

After each capture experiment, the individual cell images were used to determine cell diameter by an automated image analysis algorithm in MATLAB. Manual inspection of each capture site was used to identify empty chambers, unhealthy cells or chambers with more than a single-cell. Only chambers containing a single healthy-looking cell were further processed. In preliminary experiments, we had confirmed that cells judged unhealthy rarely yielded useable cDNA.

Tagmentation and isolation of 5' fragments

Amplified cDNA was simultaneously fragmented and barcoded by tagmentation, i.e. using Tn5 DNA transposase to transfer adaptors to the target DNA as previously described (*11*). 100 µl Dynabeads MyOne Streptavidin C1 beads (Invitrogen) were washed in 2× BWT (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 2 M NaCl, 0.02% Tween-20) then resuspended in 2 ml 2× BWT. Twenty microliters of beads were added to each well and incubated at room temperature for 15 min. All fractions were pooled, the beads were immobilized and the supernatant removed (thus removing all internal fragments and retaining only the 5′- and 3′-most fragments). The beads were then resuspended in 100 µL TNT (20 mM Tris, pH 7.5, 50 mM NaCl, 0.02% Tween), washed in 100 µL Qiagen Qiaquick PB, then washed twice in 100 µL TNT. The beads were then resuspended in 100 µL restriction mix (1× NEB CutSmart, 0.4 U/µL PvuI-HF enzyme), designed to cleave 3′ fragments carrying the PvuI recognition site. The mix was incubated for 1 h at 37 °C, then washed three times in TNT. Finally, to elute DNA, beads were resuspend in 30 µL ddH_2O and incubated 10 minutes at 70°C. Beads were then immediately bound to magnet and the supernatant was collected. To remove short fragments, Ampure beads (Beckman Coulter) were used at 1.8× volume and eluted in 30 µL.

Illumina high-throughput sequencing

The library molar concentration was quantified by qPCR using KAPA Library Quant (Kapa Biosystems) and library fragment length was estimated using Bioanalyzer of a reamplified (12 cycles) library. Sequencing was performed on an Illumina HiSeq 2000 instrument using C1-P1-PCR-2 as the read 1 primer, and C1-TN5-U as the index read primer (*11*). Reads of 50 bp were generated along with 8 bp index reads corresponding to the cell-specific barcode. Each read was expected to start with a 6 bp unique molecular identifier (UMI), followed by 3-5 guanines, followed by the 5' end of the transcript.

Processing sequencing reads to molecule counts

Read processing was performed as described (*11*), except that we removed any RNA molecule (i.e. UMI) supported only by a single read. This considerably reduced the number of true molecules detected (we estimate that as many as 30% of all molecules were detected as singletons), but also removed a large number of likely 'ghost' molecules. Such artefacts can arise by sequencing error, PCR-induced mutations or translocations and cross-contamination.

To estimate cross-contamination, we exploited the fact that choroid plexus epithelial cells are rare, but express *Ttr* at exceptionally high levels. In ten plates that did not contain a choroid plexus cell, we observed zero *Ttr* reads in every well. In contrast, in a plate with a single choroid plexus epithelial cell, we found 546,996 reads mapped to *Ttr* in the correct well, and an average of 25 reads per well in the other wells. The estimated cross-contamination rate was thus 0.0045%. There may be several reasons for this cross-contamination. For example, actual leakage between channels in the C1 chip, leakage of RNA from choroid plexus cells in solution prior to single-cell isolation, or cross-contamination during sample preparation.

Single-molecule fluorescence RNA *in situ* hybridization (smFISH)

Postnatal day 21 wild type CD1 or 5HTR3-GFP mice were perfused with 4% cold PFA. The brains were then collected, embedded in Tissue-Tek OCT (Sakura, Alphen aan den Rijn, The Netherlands), frozen on dry ice and stored at -80°C until used.

smFISH was carried out as previously described (*29*) with minor modifications. The tissue sections were permeabilized using PBS-TritonX 0.5% for 1 hour at room temperature followed by 24 hours of hybridization with 250 nM fluorescent label probes (Biosearchtech, Petaluma, CA, USA; Table S2) at 37°C and counterstained with DAPI (Life Technologies). The sections were mounted with pro-long gold (Life Technologies) and image stacks (0.3 μm distance) were acquired using a customized automated scanning microscope controlled by μManager (*28*). Images were analyzed in Matlab (Mathworks, Natick, MA, USA) using a custom script. After blind deconvolution the cell nuclei were isolated by marker-controlled segmentation of the DAPI signal. A Laplacian-of-Gaussian followed by a Wiener filter were used to reduce the smFISH background and to enhance the RNA dots that were then identified by computing the extended maxima transform of the flattened image (*30*). Each mRNA dot was assigned to the closest nucleus if the distance between the mRNA and nucleus centroids was below or equal twice the geometric average of the major and minor axis of the closest nucleus mask.

Electrophysiological recordings and immunohistochemistry

Whole-cell patch-clamp electrophysiological recordings and immunohistochemistry was performed as described in (*31*). For electrophysiological recordings we targeted layer 1 cells in acute brain slices prepared from P18 wildtype mice. Biocytin was included in the recording pipette and the recorded neurons were stained using Alexa555-streptavidin and rabbit anti Pax6 (1:500; Millipore) in combination with secondary antibody goat anti rabbit Alexa488 (1:1000; Invitrogen) to reveal morphology and Pax6 expression. Immunohistochemistry was performed on neocortex from p21-p28 5HT3a^{EGFP} mice and antibodies used were: chicken anti-EGFP (1:2000; Abcam), rabbit anti-PAX6 (1:500, Millipore) and Aldolase C (1:100 Santa Cruz). Secondary antibodies conjugated with Alexa Fluor dyes 488, 594 and 647 (1:1000; Invitrogen) were used to visualize the signals. Images were acquired on Lsm720 or Lsm760 confocal microscopes (Zeiss).

Supplementary Text

Data analysis and clustering (level 1)

Molecule counts data from all sequencing runs were merged into one single database which included metadata (such as age, sex, diameter etc.) about each cell.

Cells were selected to be valid for analysis if they passed manual inspection of images (described above) and had more than 2500 total detected RNA molecules (excluding repeats RNA and mitochondrial RNA). This resulted in 3315 valid cells. To select genes for clustering we used the following filters:

1. Remove all genes that have less than 25 molecules in total over all cells (resulted in 15310 genes)
2. Calculate the correlation matrix over the genes and define a threshold as 90^{th} percentile of this matrix ($\rho = 0.2091$). Remove genes which have less than 5 other genes which correlate more than this threshold (resulted in 12114 genes)
3. For the remaining genes calculate for each gene the mean and coefficient of variation (CV; standard deviation divided by the mean). We fit a simple noise model as $\log_2(CV) = \log_2(mean^{\alpha} + k)$, where the best fit was found to be $\alpha = -0.55$, $k = 0.64$ see Figure S2J Next we ranked all genes by their distance from the fit line and select the top 5000 genes as informative for further clustering

Next, we developed BackSPIN, an iterative and automated version of the SPIN algorithm (*32*) as a two-way unsupervised clustering approach (*33*). The algorithm is described in detail below. About 310 cells were excluded from further analysis as they clustered separately from everything else and did not express any distinct markers. Based on their expression profiles, these cells were probably neurons of low quality. When applying the BackSPIN method to the remaining 3005 cells, the algorithm identified 77 groups, stopping after 12 splits along the deepest branch. For initial analysis, we limited each branch to 5 splits, where we merged the neuronal groups into three main groups (each containing four subgroups) and oligodendrocytes into one group (containing 3 subgroups). This resulted in the nine main cell classes shown in Figure 1B.

The nine major clusters were visualized using t-Stochastic Neighbor Embedding (tSNE), using the Barnes-Hut algorithm and implementation (http://homepage.tudelft.nl/19j49/t-SNE.html). Each cluster was interpreted using known markers. For example, interneurons were identified based on their shared expression of *Gad1*, *Gad2* (the glutamic acid decarboxylases), *Slc32a1* (previously known as Vgat; the vesicular GABA transporter) and *Slc6a1* (previously known as Gat1; the transmembrane GABA transporter responsible for removing GABA from the synaptic cleft).

Subclass analysis (level 2)

To further analyze each of the nine main groups we started again with all genes and made a statistical t-test comparing the expression of each gene between any two possible groups. Each gene was assigned to the group with the lowest value of the maximum *p* value. Having the relevant genes for each group is important to achieve optimal clustering within groups. Thus the aim of the t test was not to discover genes that were highly specific to e.g. interneurons, but to exclude genes that were likely to be more specific to other groups. This was particularly important for the vascular cells, where we found that a subset of endothelial cells carried a pericyte signature, and vice versa. We interpreted this as incomplete cell dissociation, e.g. that some endothelial cells had part of a pericyte stuck onto it. If pericyte-specific genes were

included when clustering endothelial cells, this would create artefactual subtypes of endothelial cells.

These gene exclusions were only used during clustering. Once clustering had been achieved, we used a full Bayesian regression model to determine each gene's expression pattern across all subclasses (see below). While in principle it would have been desirable to use the regression model – with its more realistic noise model – during clustering, in practice this was not computationally feasible. Our Bayesian regression model was computed by numerical Monte Carlo sampling, and it took about 48 hours to run a single regression for every gene, on all 32 cores of a high-end server.

Since the three neuronal groups shared many relevant genes, we merged the genes assigned to interneurons, S1 pyramidal and CA1 pyramidal neurons. Next, each one of the nine groups and its set of relevant genes were subjected to the BackSPIN algorithm with manually defined number of splits (according to group variability).

BackSPIN algorithm

The BackSPIN algorithm is based on the SPIN algorithm (*32*) – which essentially sorts the expression matrix by cell-cell or gene-gene similarity. In contrast to the SPIN algorithm which does not identify clusters, here the aim was to identify groups of cells/genes in an unsupervised manner. SPIN is a powerful method to sort a distance/correlation matrix without reducing dimensionality, and it converges to a 1D order of the features.

Step 1 – sort the expression matrix. We used SPIN to sort the expression matrix by both cells and genes (Fig. S3), running the algorithm several iterations (~10) for each width parameter, starting with a width of 40% of the matrix dimension and slowly going down to 1. The width parameter controls the normal distribution used for the weight matrix in the SPIN algorithm. This process resulted in a two-way sorted matrix.

Step 2 – split the matrix. A splitting step over the cell dimension comprises:

1. Find the best splitting point over the correlation matrix, where R is the correlation matrix and N is the matrix dimension.

$$S_{left} = \frac{\sum_{j,h=1}^{I_{split}} R_{j,h}}{(I_{split})^2} \Big/ \frac{\sum_{j,h=1}^{N} R_{j,h}}{N^2} \qquad \text{Eq. 1}$$

$$S_{right} = \frac{\sum_{j,h=1+I_{split}}^{N} R_{j,h}}{(N - I_{split})^2} \Big/ \frac{\sum_{j,h=1}^{N} R_{j,h}}{N^2} \qquad \text{Eq. 2}$$

$$\forall i = 1,2,3...N \; \frac{\sum_{j,h=1}^{i} R_{j,h} + \sum_{j,h=i+1}^{N} R_{j,h}}{i^2 + (N-i)^2}, split = \text{argmax}_i \, S(i) \qquad \text{Eq. 3}$$

2. If $max(S_{left}, S_{right}) > 1.15$, use the splitting point to break data to left and right groups, otherwise do not split.

3. Assign genes to each group by calculating the average expression in each group and assign each gene to the group with the highest expression.

Step 3 – recursion on each sub-matrix. The two halves (right/left) of the matrix are resorted and resplit separately, by restarting at **Step** 1 but including only the genes and cells assigned to each half.

The algorithm gets as input the full matrix after sorting (1D order) and the max number of split cycles allowed. For every cycle, the number of groups may increase up to two-fold. To avoid splitting in case that the data is very homogenous the algorithm has a stopping condition (step 2). The output of the algorithm is the updated order of cells/genes and the group assignment in each splitting step for each gene/cell. Because genes are always assigned to either right or left halves, both genes and cells are clustered simultaneously. This works well up to the point where genes are truly expressed in multiple subsets of cells, which cannot be consistently split.

Validation of cluster robustness

Applying BackSPIN for S1 pyramidal cells class resulted in 13 clusters after allowing maximum 4 splits as shown in Figure S5A. To test the stability of clusters generated by BackSPIN we used a resampling approach. The idea is to test the stability of the clusters by eliminating part of the data and running the clustering algorithm again. This approach is free from the assumptions required when testing robustness using artificial noise or creating synthetic data (*34*) . Resampling was performed as follows:

1. Select at random 80% of the cells.
2. Run BackSPIN with 4 splits using the same set of genes as in FigS5A.
3. For every one of the original clusters (1-13) count the fraction of cells that stay together in the largest corresponding cluster in the resampled data.
4. Repeat steps 1-3.
5. As a negative control (null model) perform the same analysis on a random permutation of the original cluster order (which controls for differences in clusters size).

The results of 100 resampling realizations (Fig. S4F) confirmed the stability of the clusters, i.e. that the same cells consistently tended to form the same clusters even when leaving out a random 20% of the cells.

Comparison with hierarchical clustering

We applied standard hierarchical clustering (*Mathematica* version 10; Wolfram Research Inc.) to the log-transformed raw data with correlation distance and Ward's linkage. The result was similar to BackSPIN on the 9 major classes (Fig. S4A) although some oligodendrocytes were erroneously lumped together with S1 pyramidal cells. On the lower levels, clusters tended to be more fragmented. For example, within S1 pyramidal cells, several clusters were in near-perfect agreement with BackSPIN, while others were diffuse, even though BackSPIN clustering was well supported by in situ hybridization (Fig. S4B).

We determined that one reason for the fragmentation was the fact that standard clustering methods are based on a global cell-cell distance measure, which is computed from all genes. However, when clustering, say, oligodendrocytes, most

genes are enriched elsewhere. They are thus not informative, and contribute at best only noise. For example, consider *Rasgrf2* (Fig. S4C). This gene is a strong layer 2-3 marker in S1 pyramidal cells, but was also detected in interneurons, and in a seemingly random subset of oligodendrocytes and other cell types. Thus this gene will tend to align some oligodendrocytes with layer 2-3 neurons, even though this conflicts with oligodendrocyte-specific genes such as *Mog*. Since there are thousands of genes that are more highly expressed outside oligodendrocytes than within them, this contributes a kind of 'centrifugal force' drawing these cells out of their proper placement.

Biclustering solves this problem because at each level, genes enriched outside of the cells under consideration are excluded. So, for example, *Rasgrf2* would be excluded when clustering oligodendrocytes because it is enriched in neurons, and conversely, *Mog* would be excluded when clustering neurons.

Comparison with affinity propagation

Another recently developed clustering algorithm, which performs well on a variety of datasets, is affinity propagation (*35*). This algorithm has the advantage that the exact number of clusters does not need to be predefined (in contrast to K-means clustering). It achieves clustering by an iterative procedure that gradually identifies exemplars to which other nodes (cells in this case) are assigned.

On level 1, corresponding to Fig. S3, affinity propagation found ten clusters in total (Fig. S4D; using damping = 0.95 and preference = -14). It found two clusters of S1 pyramidal cells, two of mainly CA1 pyramidals and two of oligodendrocytes. However, it lumped together mural, endothelial and interneurons in two clusters, microglia, oligodendrocytes and interneurons in one cluster, and ependymal cells with astrocytes.

Applying affinity propagation to S1 Pyramidal cells (Fig. S4E), in contrast, resulted in near-perfect agreement with BackSPIN (and with validation by *in situ* hybridization).

Residual variance and merging of clusters

We used a heuristic stopping rule (described above). To determine if there was residual variance in the final subclasses – which would suggest that they could be further split – we performed the following test. We computed the variance explained by each principal component (i.e. the eigenvalues of the standardized correlation matrix). We then used the 'broken stick' criterion (*36*) to determine if the first several principal components explained more variance than expected by chance. Each of the 9 major classes had several significant principal components. In contrast, 46 of the 47 subclasses had only a single significant principal component (exception: Oligo6, which had two). Although this shows that in principle, it may be possible to further split some of the clusters, we have preferred to be conservative when calling cell classes. For clarity, we show all the raw clusters in supplementary figures (black rectangles in Figs. S6, S8, S9 and S10), where we have also indicated how some clusters were merged.

In brief: we merged two clusters to form S1PyrL23 (layer II/III pyramidal neurons; Fig. S6A), although a few genes such as *Bdnf, Penk* and *Inhba* suggested possible subtypes. Two clusters were merged to form S1PyrDL (deep-layer pyramidal neurons; Fig. S6A). CA1 type 1 and type 2 neurons (CA1Pyr1 and CA1Pyr2) were merged from three and two clusters, respectively. As noted in the main text,

hippocampal pyramidal neurons varied along at least two dimensions, with genes such as *Dcn* and *Grp* suggesting subgroups within both type 1 and type 2 neurons.

Similarly, we report a single Oligo5 subclass from two clusters (possibly representing sub-stages of oligodendrocyte differentiation: Fig S9B). For astrocytes, we report two subclasses, Astro1 and Astro2 from four clusters (Fig. S10A). These were confirmed by immunohistochemistry (Fig. 3B), and we suspect that the further splits represent intermediate forms also indicated by the immunostaining. For example, *Gfap* and *Mfge8* stained distinct populations residing in the superficial glia limitans and in the parenchyme, respectively, but there were cells in the vicinity of the surface that were apparently positive for both markers.

Finally, in the case of mural end endothelial cells, we merged two clusters to form a single vascular smooth muscle subclass (Vsmc; Fig. S10B) and three clusters forming type 2 endothelial cells (Vend2). In these cases, specifically, we wanted to avoid the risk of over-splitting due to cross-contamination, as mural cells and endothelial cells both reside under the basement membrane and we were not confident that we could always physically separate them. Any contamination of endothelial cells by (fragments of) pericytes, for example, would potentially show up as an artefactual endothelial subtype.

Negative binomial generalized linear regression

Clustering by BackSPIN was used to discover classes and subclasses of cells. However, BackSPIN always assigns each gene to a single cluster, when in reality genes may be expressed in many cell types, and may have any combination of cell type-specific, basal, sex-specific or age-regulated expression patterns. To rigorously assign expression to defined classes of cells, we therefore developed a Bayesian generalized linear regression model (GLM), using a negative binomial noise model.

In a regression model, the outcomes (i.e. the measured molecule counts in individual cells) are viewed as being sampled from some distribution whose mean is given by a linear combination of *predictors*. A predictor can be a scalar (e.g. cell size) or a 0/1 indicator (e.g. 'this cell belongs to the class of microglia'). With K predictors x_i there will be K coefficients β_i. For each cell, the outcome and predictors are known, and we wish to determine the values of the coefficients.

Note that since molecule counts are always non-negative, we constrain the coefficients and the predictors to also be non-negative. This gives a straightforward interpretation to the coefficients; they represent the additional number of molecules expressed when $x_i = 1$.

We model the following predictors:

Variable	Type	Implementation
Sex	Categorical	x_{male}, x_{female}
Cell type	Categorical	$x_{S1PyrL23}$, x_{Mgl}, x_{Astro2}, …
Age	Categorical	x_{young}, x_{old}
Basal	Scalar	x_{basal}

The *Basal* predictor is meant to capture constitutive expression, also known as housekeeping genes. We regarded each gene as having a basal expression proportional to the total transcriptome size, on top of which is added molecules due to the age, type and sex of the cell. The *Basal* variable is set proportional to the total

molecule count of the cell, normalized to the mean count in all cells. Hence, a neuron twice as large as the average cell will have twice the basal expression.

We set up a regression model that explains the inferred expression level μ as a linear combination of the predictors. We introduce coefficients β_k, one for each explanatory variable (thus $k \in [1, K]$); they tell us about the effect, if any, of the various predictors on the gene expression level. For a binary explanatory variable x_k, its coefficient β_k is the number of additional molecules that are expressed in cells for which $x_k = 1$. Therefore,

$$\mu = \sum_{k=1}^{K} \beta_k x_k \qquad \text{Eq. 4}$$

Note that there is no constant (intercept); instead basal expression is assumed to be proportional to cell size (total molecule count) and is modeled by its own explanatory variable.

Next, we note that real count data is typically *overdispersed* compared to an ideal Poisson distribution. To model this additional source of variation, we use the negative binomial distribution, which can be equivalently viewed as a gamma mixture of Poisson distributions. That is, if y is the observed count, the model is:

$$y \mid \lambda \sim Poisson(\lambda) \qquad \text{Eq. 5}$$

$$\lambda \mid a, b \sim Gamma(a, b) \qquad \text{Eq. 6}$$

The mean and standard deviation are then

$$\mu = ab \qquad \text{Eq. 7}$$

$$\sigma = \sqrt{\frac{ab}{1+b}}\left(1+b\right) \qquad \text{Eq. 8}$$

However, we also need to take into account the way in which the overdispersion scales with the mean. For example, the naïve approach of simply modelling noise using an extra predictor for the standard deviation fails, as it essentially amounts to saying that the standard deviation is the same at all levels of expression (which is very far from true, see Fig. S1J).

Instead, we make use of the fact that the observed standard deviations scale roughly as the square root of the mean, with a constant offset (Fig. S2J). We introduce an overdispersion factor r, defined by

$$\sigma = r\sqrt{\mu} \qquad \text{Eq. 9}$$

and hence

$$a = \frac{\mu}{r^2 - 1} \qquad \text{Eq. 10}$$

$$b = r^2 - 1 \qquad \text{Eq. 11}$$

To make a full Bayesian model, we attach prior distributions to the coefficients as well as the overdispersion factor. For β_k, we used a Pareto distribution $y \sim Pareto(0, 1.5)$, which was a good fit to the actual distribution of gene expression over all genes (not shown)[1].

For r, we use a Cauchy distribution, leading to the complete Bayesian negative binomial regression model (next page).

[1] When implementing this model, we actually used Pareto(1, 1.5) which shifts the coefficiencts by 1.0. We then subtract 1.0 when calculating the mean. This is necessary because the Pareto distribution is undefined at zero. We also add a small constant (0.001) to the mean to avoid numerical instability when the mean is very close to zero.

The complete regression model

$$\mu = \sum_{k=1}^{K} \beta_k x_k \qquad \text{(Eq. 4 repeated for clarity)}$$

$$y \mid \lambda \sim Poisson(\lambda) \qquad \text{(Eq. 5 repeated for clarity)}$$

$$\lambda \mid \mu, r \sim Gamma\left(\frac{\mu}{r^2 - 1}, r^2 - 1\right) \qquad \text{Eq. 12}$$

$$r \sim Cauchy(0,1) \qquad \text{Eq. 13}$$

$$\beta_k \sim Pareto(0,1.5) \qquad \text{Eq. 14}$$

The model was implemented in Stan (`http://mc-stan.org`):

```
data {
   int<lower=0> N;           # number of outcomes
   int<lower=0> K;           # number of predictors
   matrix<lower=0>[N,K] x;   # predictor matrix
   int y[N];                 # outcomes
}

parameters {
   vector<lower=1>[K] beta;  # coefficients
   real<lower=0.001> r;      # overdispersion
}

model {
   vector<lower=0.001>[N] mu;
   vector<lower=1.001>[N] rv;

   # priors
   r ~ cauchy(0, 1);
   beta ~ pareto(1, 1.5);

   # vectorize the overdispersion
   for (n in 1:N) {
      rv[n] <- square(r + 1) - 1;
  }

   # regression
   mu <- x * (beta - 1) + 0.001;
   y ~ neg_binomial(mu ./ rv, 1 / rv[1]);
}
```

Gene expression enrichment analysis

We used negative binomial generalized linear regression (previous section) to obtain posterior probability distributions for the subclass-specific (as well as basal and sex-specific) contributions to each gene's expression. In order to find genes expressed

preferentially in a particular subclass (or set of subclasses), we compared these posterior probability distributions in two different ways, as follows.

For transcription factors (Fig. 4A), we used a procedure designed to find all-or-none expression patterns. Given a test set (i.e. a set of named subclasses, such as the two astrocyte subclasses), we defined a control set by taking interneurons, S1 Pyramidal cells, CA1 pyramidal cells, oligodendrocytes, microglia, astrocytes, ependymal cells, mural cells and vascular endothelial cells, removing from the control set any subclass included in the test set. We then calculated the posterior probability that expression in the test set was greater than the expression in each member of the control set (using the numerical samples obtained by Markov chain Monte Carlo).

We systematically analyzed enrichment in every subtree of the hirearchical clustering (Fig. 1), using a cutoff of 90% posterior probability, leading to a false discovery rate <5%. When a gene was enriched on several nested subtrees, we assigned it to the largest subtree. When it was enriched on several non-overlapping subtrees, we indicated this fact by an asterisk in Fig. 4A.

All enriched transcription factors were then manually examined, rejecting those with substantial (but not statistically significant) background expression in some or most subclasses. Fig. 4A shows the resulting enriched transcription factors.

Enrichment of genes expressed in ependymal cells (Fig. 4B) was calculated by comparing expression in each of the nine major classes individually against the basal expression given by the linear regression. We retained only genes that were enriched with 99.9% posterior probability in ependymal cells, and not enriched in any other class. The same analysis was applied to all nine major classes of cells, and the results are given in Table S1.

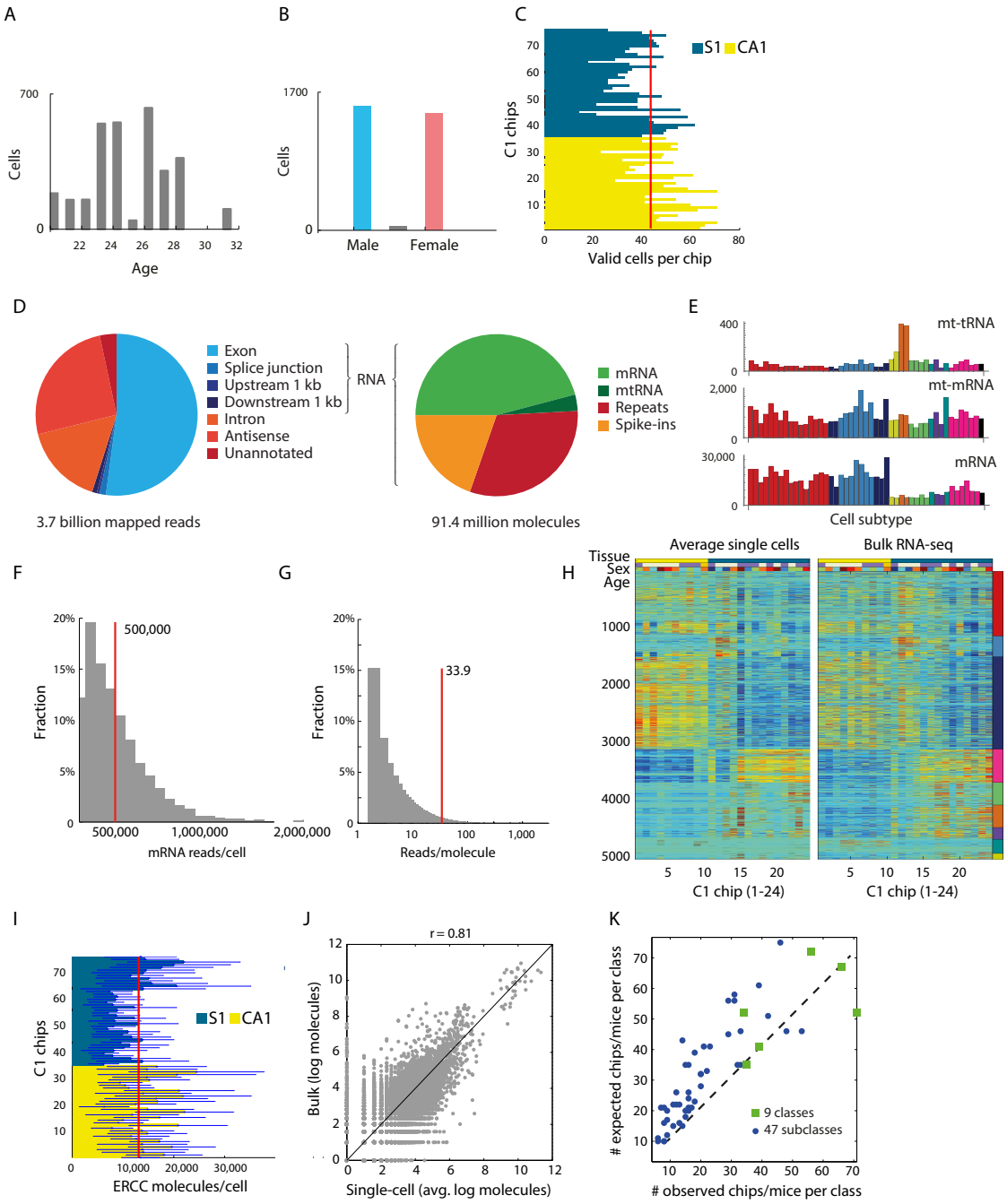

Fig. S1. Technical performance.

(A) Total cell counts by mouse age. **(B)** Total cell counts by sex (gray bar, not determined). **(C)** Number of cells passing QC from each Fluidigm C1 run. **(D)** Breakdown of all reads in our data according their alignment in the genome (pie chart, left) and breakdown of all RNA molecules by genomic feature class (right). mRNA, polyadenylated RNAs, including non-coding RNAs, but excluding expressed repeats. **(E)** Average total molecules per cell in each subclass of cells, for mRNA, mt-RNA and mt-tRNA. Colored and ordered as in Fig. S3A (the two brown bars showing high mt-tRNA expression are endothelial cells, Vend1 and Vend2). **(F)** Histogram of number of reads per cell. **(G)** Histogram of number of reads per molecule. **(H)** Average and standard-deviation of number of molecules per cell for every Fluidigm

C1 chip. **(I)** Scatter plot showing correlation between average single-cell and bulk control data over 24 Fluidigm C1 chips. **(J)** Gene expression heatmap comparing average single-cell RNA-seq with bulk RNA-seq. Gene order is the same as in Fig. S3A and chips are ordered by tissue (S1 or CA1, indicated at top). **(K)** Representation of inferred cell classes and subclasses across chips (and hence, mice; with one exception each chip was processed from a different mouse), compared to that expected by chance. Red vertical lines in all panels show means.

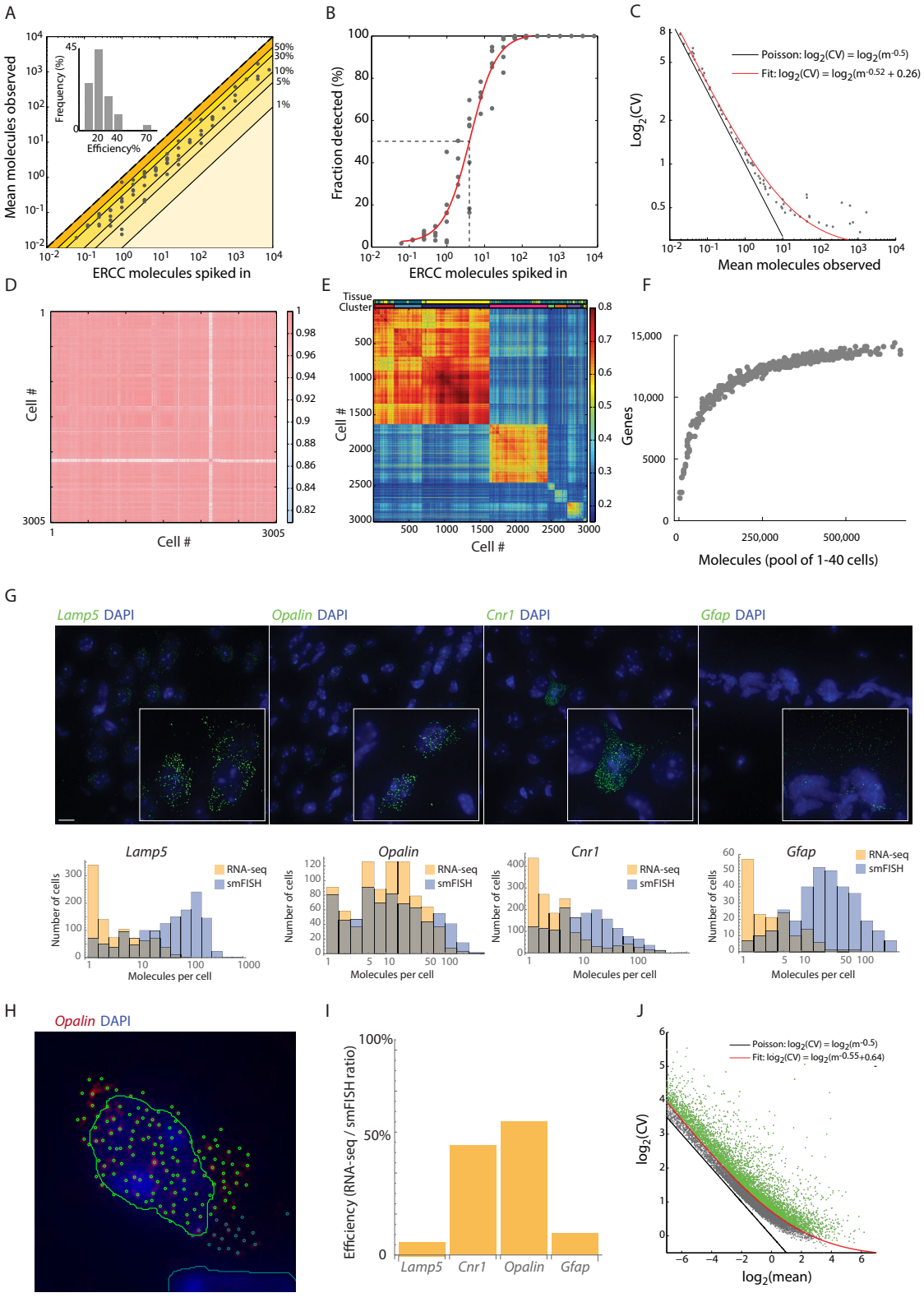

Fig. S2. Quantitative sensitivity and accuracy

(A) Efficiency of detecting ERCC spike-in RNA. Scatter plot shows number of spiked-in molecules versus the average number of observed molecules. Each dot represents one ERCC RNA species averaged over all 3005 cells in the dataset. Diagonal fields in the background represent equal efficiency of capturing RNA

molecules. Inset shows the distribution of capture efficiency across all ERCC spike-in transcripts (average was 22%). **(B)** Detection limit: the probability to detect (at least one molecule) of an ERCC transcript as function of the average actual number of molecules spiked-in. Dashed line indicates a 50% probability of detection, at 4 molecules per cell. **(C)** Coefficient of variance (CV; i.e. standard deviation divided by the mean) as a function of the mean observed number of ERCC spike-in molecules, shows properties of the technical noise. For low abundance transcripts, noise approached the limit given by the Poisson distribution, while at high abundance it was greater. Red curve shows empirical fit to the data. **(D)** Cell-cell correlation matrix based on ERCC transcripts over all cells in the data, revealing that reaction conditions were highly uniform across several months of experiments. **(E)** Cell-cell correlation matrix based on endogenous genes, showing biological variability in excess of technical noise. Cells are sorted in the same order as in Figure 1B. **(F)** Number of genes detected as a function of total number of molecules when sampling randomly 1-40 cells from the data. Each dot represents one sampling where the total number of molecules and the total number of detected genes was calculated. **(G)** Single molecule RNA-FISH (smFISH) targeting *Lamp5* (pyramidal and interneurons), *Opalin* (oligodendrocytes), *Cnr1* (interneurons) and *Gfap* (astrocytes). Histograms below show comparison between smFISH and single-cell RNA-seq counts for the same genes. **(H)** Example of output from automated counting of smFISH. Nucleus represented by DAPI signal is surrounded by a contour and smFISH spots assigned to the same cell appear in the color of the contour. **(I)** Estimated efficiency of single-cell RNA-seq, measured as ratio of average single-cell RNA-seq to smFISH (molecule counts). **(J)** Scatter plot of CV versus the mean for all detected genes over all cells in the dataset. Each dot represents one gene, measured 3,005 times. Black line show expected noise from Poisson distribution and red line show fit to a model with additive constant component. Genes were selected for clustering analysis based on their distance from the fit line as a measure of variability.

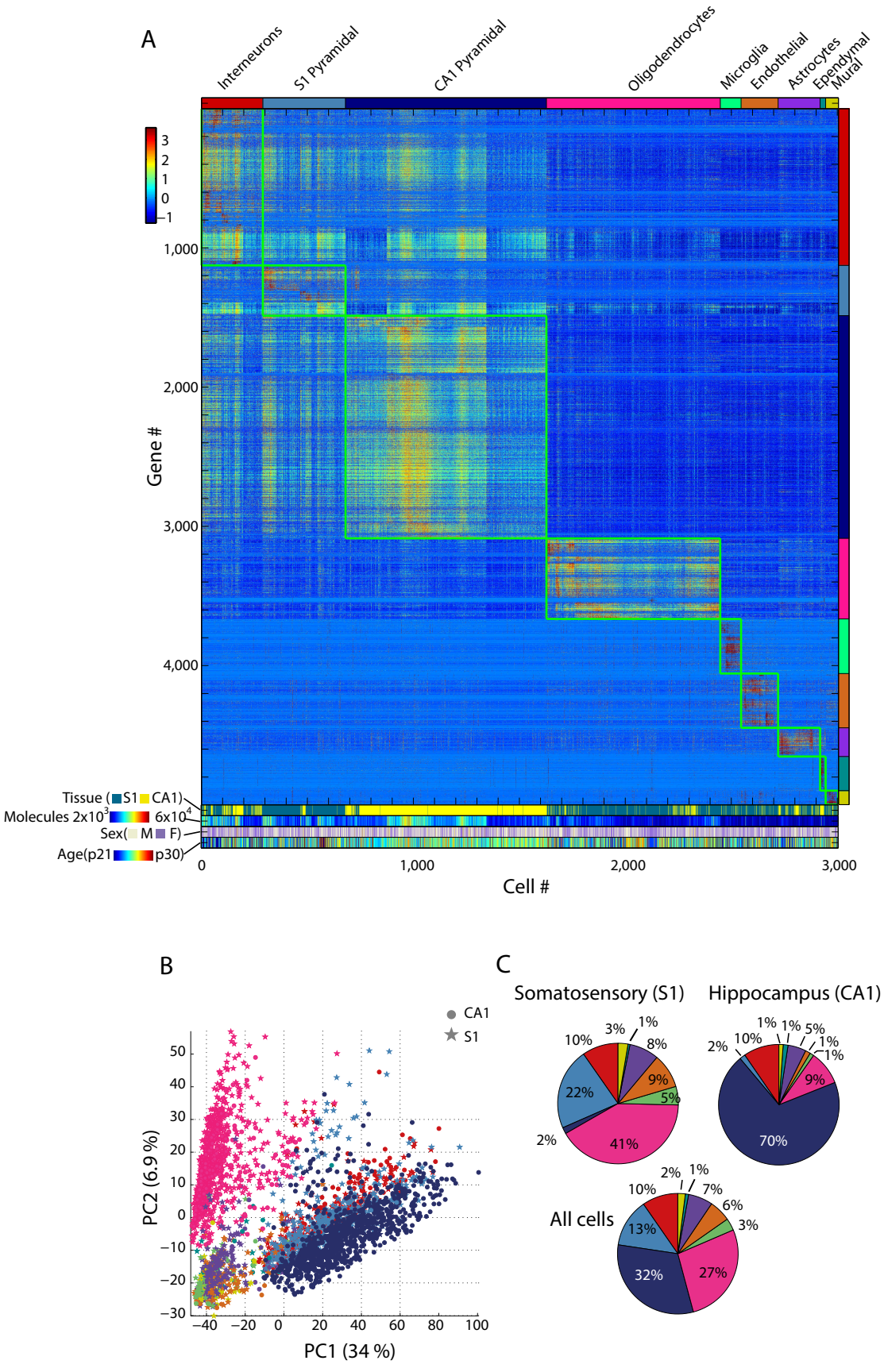
A
Interneurons
S1 Pyramidal
CA1 Pyramidal
Oligodendrocytes
Microglia
Endothelial
Astrocytes
Ependymal
Mural
3
2
1
0
−1
1,000
2,000
3,000
4,000
Gene #
Tissue (S1 CA1)
Molecules 2x10^3 6x10^4
Sex(M F)
Age(p21 p30)
0
1,000
2,000
3,000
Cell #
B
CA1
S1
50
40
30
20
10
0
−10
−20
−30
PC2 (6.9 %)
−40
−20
0
20
40
60
80
100
PC1 (34 %)
C
Somatosensory (S1)
Hippocampus (CA1)
All cells
10%
3%
1%
8%
9%
5%
22%
2%
41%
2%
10%
1%
1%
5%
1%
1%
9%
70%
2%
1%
10%
7%
6%
3%
13%
32%
27%

Fig. S3. Molecular census of somatosensory S1 cortex and hippocampus CA1 by unbiased sampling and single-cell RNA-seq

(A) Gene expression heatmap of 3,005 single-cells using 5,000 selected highly variable genes. Data was clustered using BackSpin resulting in nine main clusters, indicated at top, identified based on known cell type-specific genes. **(B)** Principal components analysis on the same data showing the similarity within and between cell-types. Colors indicate clusters of cells in (A), top row. **(C)** Frequency of major cell classes in S1 cortex, CA1 hippocampus and all cells shown as pie charts. Colored as in (A) and (B).

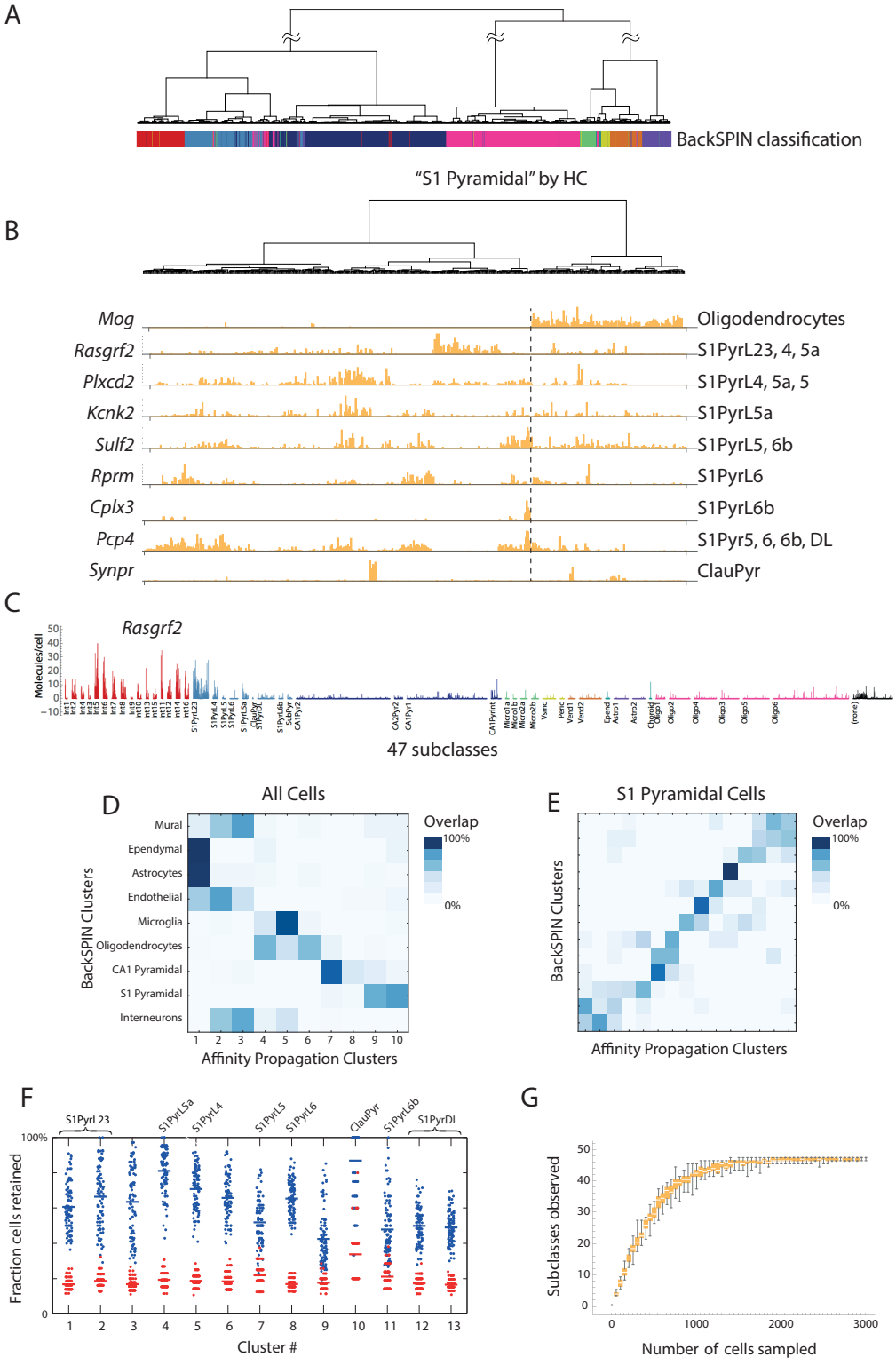
A
BackSPIN classification
"S1 Pyramidal" by HC
B
Mog
Rasgrf2
Plxcd2
Kcnk2
Sulf2
Rprm
Cplx3
Pcp4
Synpr
Oligodendrocytes
S1PyrL23, 4, 5a
S1PyrL4, 5a, 5
S1PyrL5a
S1PyrL5, 6b
S1PyrL6
S1PyrL6b
S1Pyr5, 6, 6b, DL
ClauPyr
C
Rasgrf2
Molecules/cell
47 subclasses
D
All Cells
Mural
Ependymal
Astrocytes
Endothelial
Microglia
Oligodendrocytes
CA1 Pyramidal
S1 Pyramidal
Interneurons
BackSPIN Clusters
Affinity Propagation Clusters
Overlap
100%
0%
E
S1 Pyramidal Cells
BackSPIN Clusters
Affinity Propagation Clusters
Overlap
100%
0%
F
S1PyrL23
S1PyrL5a
S1PyrL4
S1PyrL5
S1PyrL6
ClauPyr
S1PyrL6b
S1PyrDL
Fraction cells retained
Cluster #
G
Subclasses observed
Number of cells sampled

Fig. S4. Validation of clustering algorithm

(A) Comparison with hierarchicical clustering, with colors as in Fig. S3A showing the corresponding cluster assignment for each cell in BackSPIN. **(B)** Hierarchical clustering of the S1 pyramidal cells, and selected markers indicative of layer-specific cell classes, and oligodendrocytes. **(C)** Expression of *Rasgrf2* across the full set of 3,005 cells. **(D)** Comparison with affinity propagation (all cells). **(E)** Comparison with affinity propagation (S1 pyramidal cells). **(F)** Cluster robustness by resampling in S1 pyramidal cells. Blue, resampled BackSPIN clusters. Red, resampled BackSPIN clusters after random permutation of cluster labels (null model). **(G)** Saturation plot showing the total number of subclasses observed, as a function of number of cells sampled, with the requirement that at least four cells be observed in each cluster.

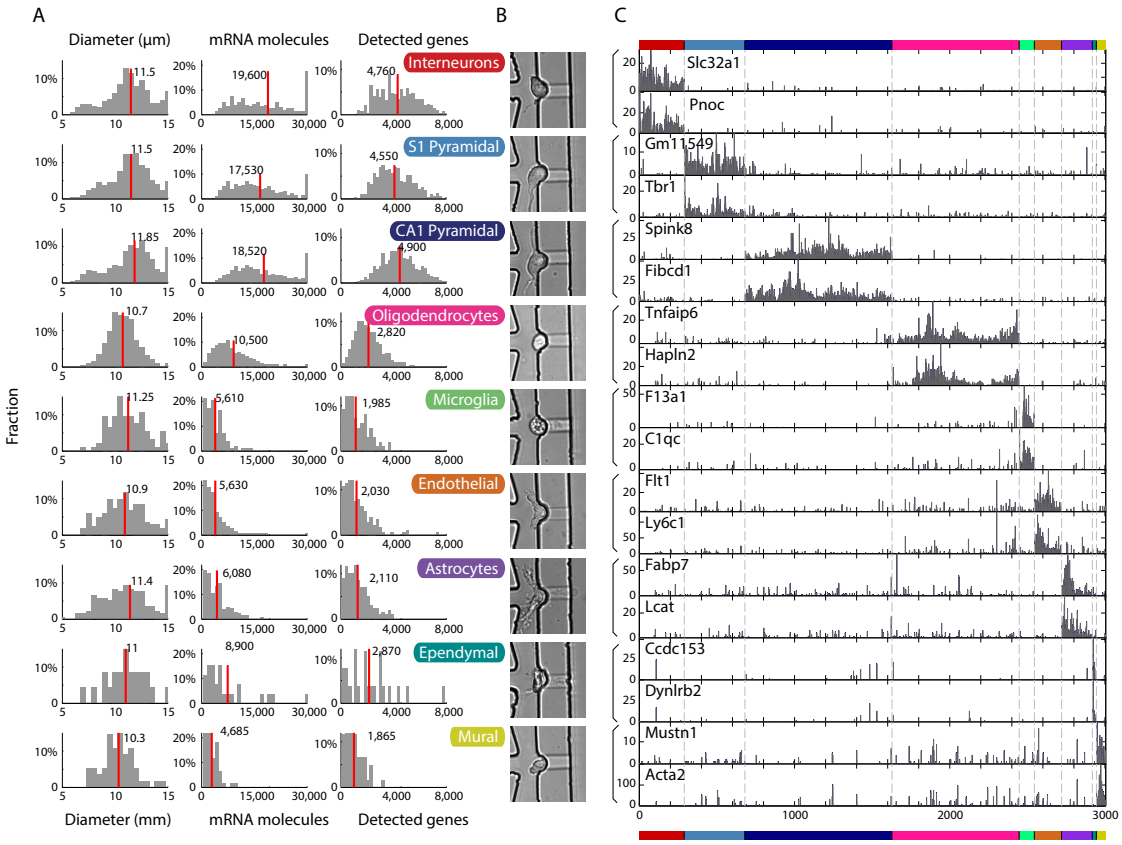

Fig. S5. General properties of major classes of cortical cells

(A) Histograms of cell diameter, total number of molecules and number of detected genes for each one of the nine classes. **(B)** Representative image of cell from each class as they appear during capture on the C1 chip. **(C)** Examples of class-specific genes, two per class. Barplots show raw detected molecule counts over all cells (ordered as in Fig. S3) without any normalization.

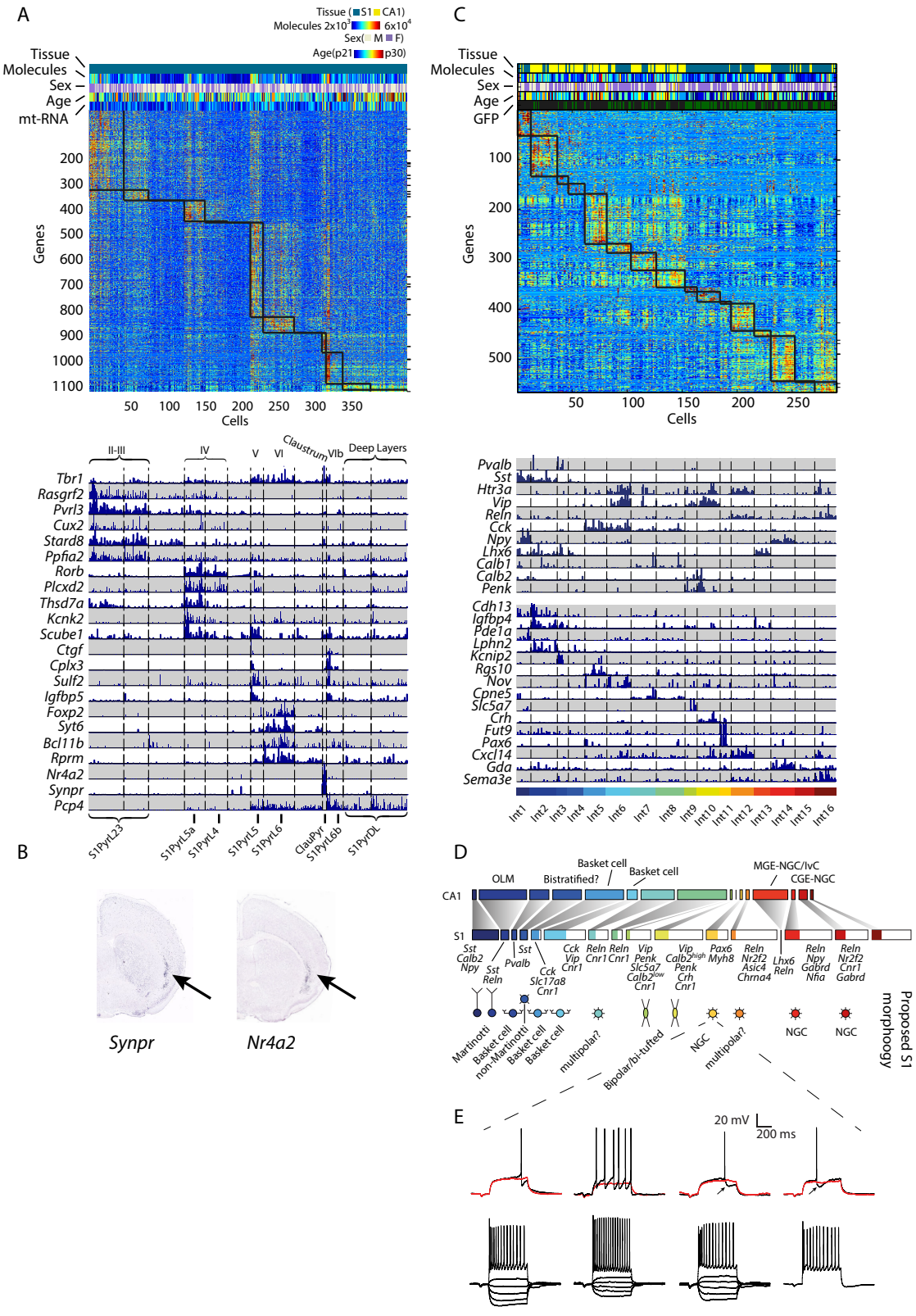

Fig. S6. Neuron subclasses in the somatosensory cortex

(A) Identification of subclasses in the somatosensory cortex S1 pyramidal group using BackSPIN. Heatmap shows expression of all genes enriched in these neurons, with rectangles indicating BackSPIN clusters. Barplots below show expression of selected

known and novel markers. Subclasses were named after their layer-specificity as determined using marker expression in the Allen Brain Atlas. S1PyrL23, layer II-III; S1PyrL4, layer IV; S1PyrL5a, layer Va; S1PyrL5, layer V; S1PyrL6, layer VI; S1PyrL6b, layer VIb; S1PyrDL, deep layers. **(B)** *In situ* hybridization (Allen Brain Atlas). ClauPyr was identified as cells mainly located in Claustrum (arrows), based on *Synpr* and *Nr4a2* expression, but scattered cells expressing these markers were observed scattered throughout S1 and in layer 6b respectively. **(C)** Identification of 16 interneurons subclasses using BackSPIN. Heatmap shows the gene expression matrix and barplots show selected known and novel markers. Cells with detected EGFP signal are marked in green on the GFP bar above heatmap. Fraction of S1/CA1 cells is depicted at bottom: blue, S1; yellow, CA1; white, flow sorted Htr3a^{+} cells from S1. **(D)** Proportion of interneurons subgroups in S1 or CA1 depicted by the length of colored blocks. White blocks represent FACS sorted cells, marked to facilitate a fair comparison of subclass abundances. Names of selected known markers expressed in each subgroup are shown together with the proposed morphology in S1. **(E)** Example traces additional PAX6+ neurons recorded in layer 1 showing a late-spiking firing. Two of the cells had a slight bi-phasic after hyperpolarization (black arrows) not seen in the other PAX6+ cells.

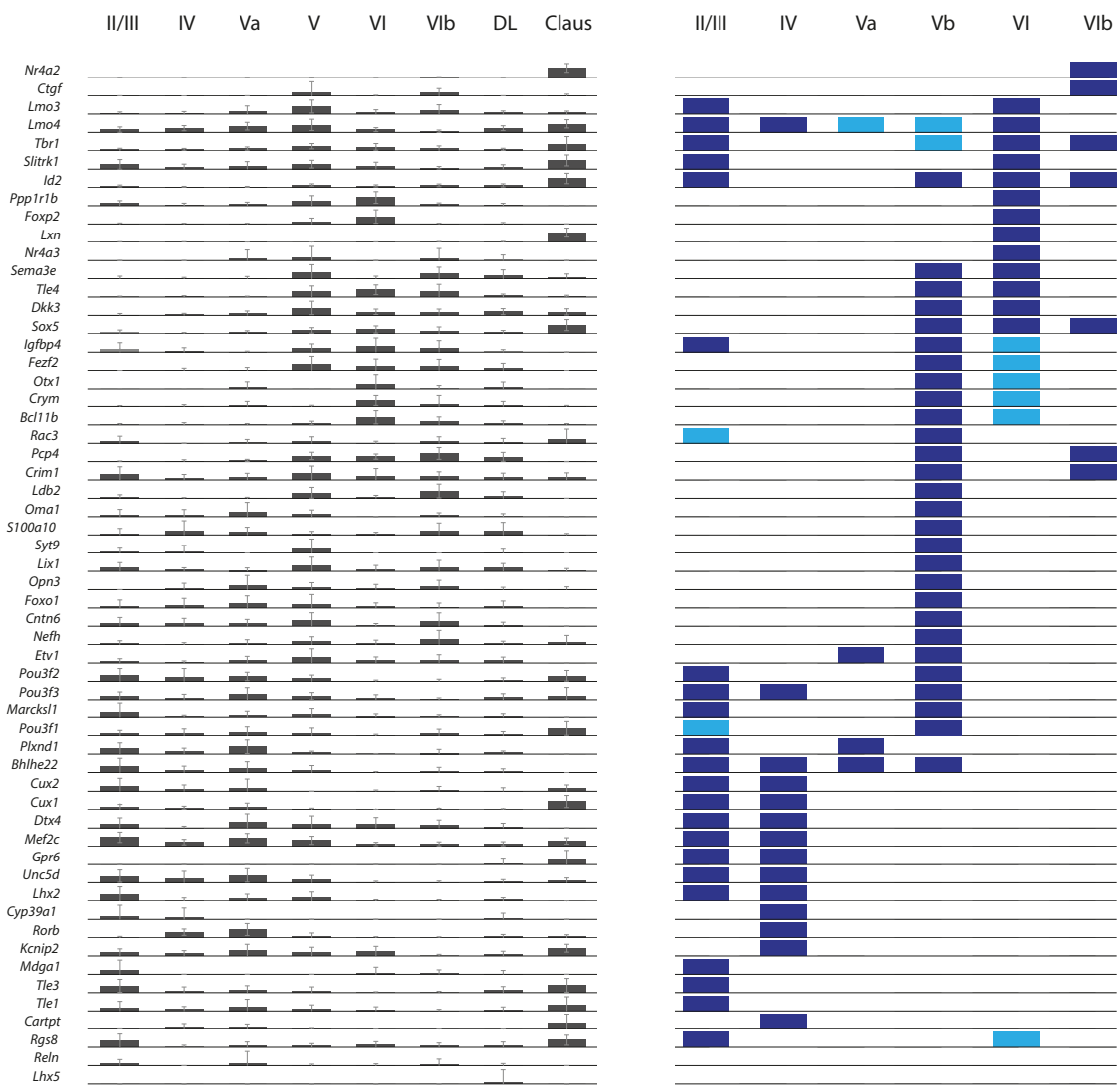

Fig. S7. Previously described layer-specific markers

Barchart showing the average (± standard deviation) expression of previously described layer-specific markers, in subclasses defined in this paper (left) and previously (right, from ref. (*1*)).

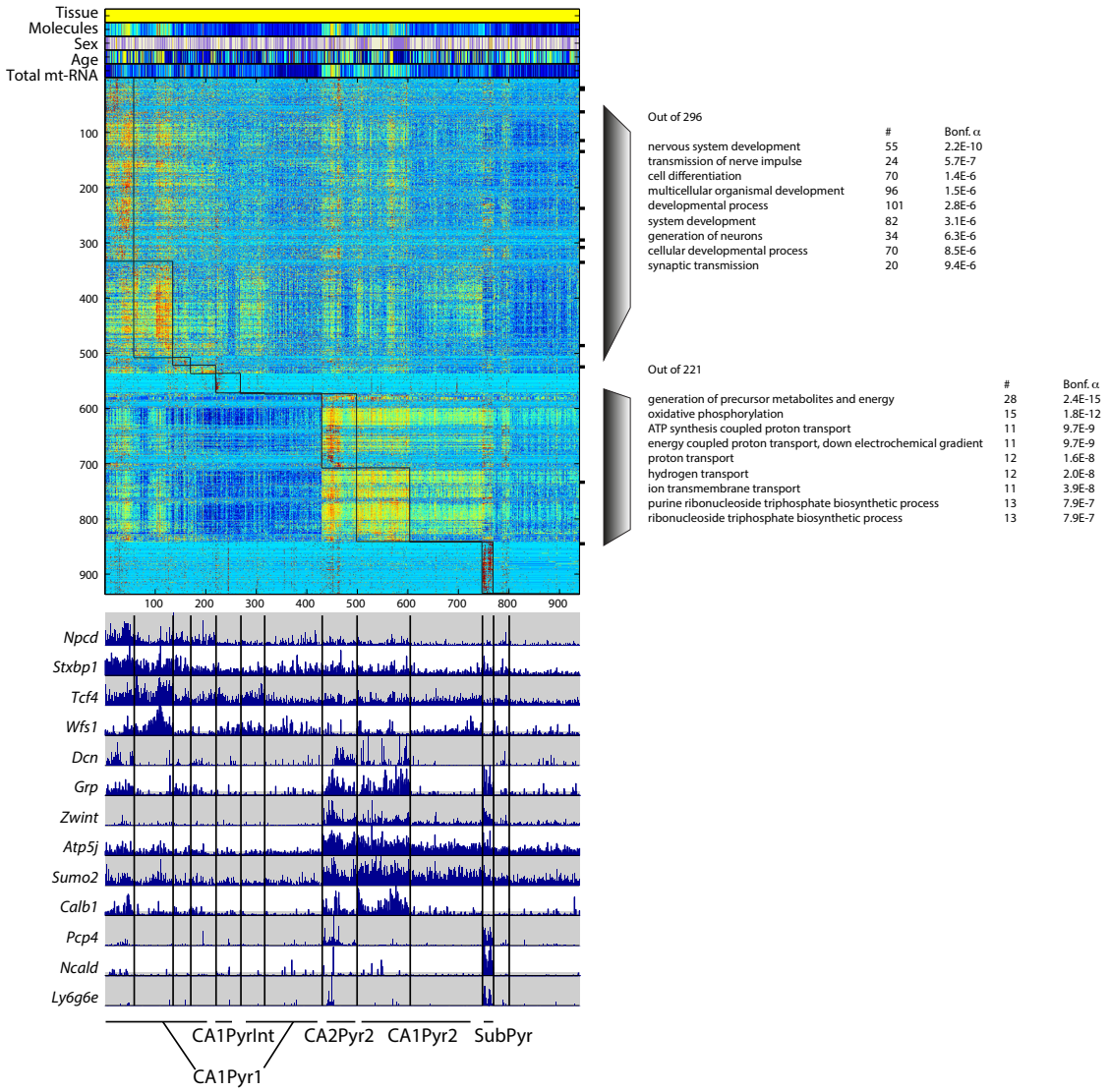

Fig. S8. Hippocampus CA1 pyramidal cells

Clustering using BackSPIN of pyramidal CA1 neurons. We classified the resulting clusters into three CA1 pyramidal neurons subclasses (CA1Pyr1, CA1Pyr2, CA1PyrInt) and two originating from adjacent tissues (CA2Pyr2, hippocampus CA2 pyramidal neuron; and SubPyr, subiculum pyramidal neuron). Gene enrichment analysis of the two main blocks of genes is shown at right. α values are *P* values Bonferroni-corrected for multiple testing.

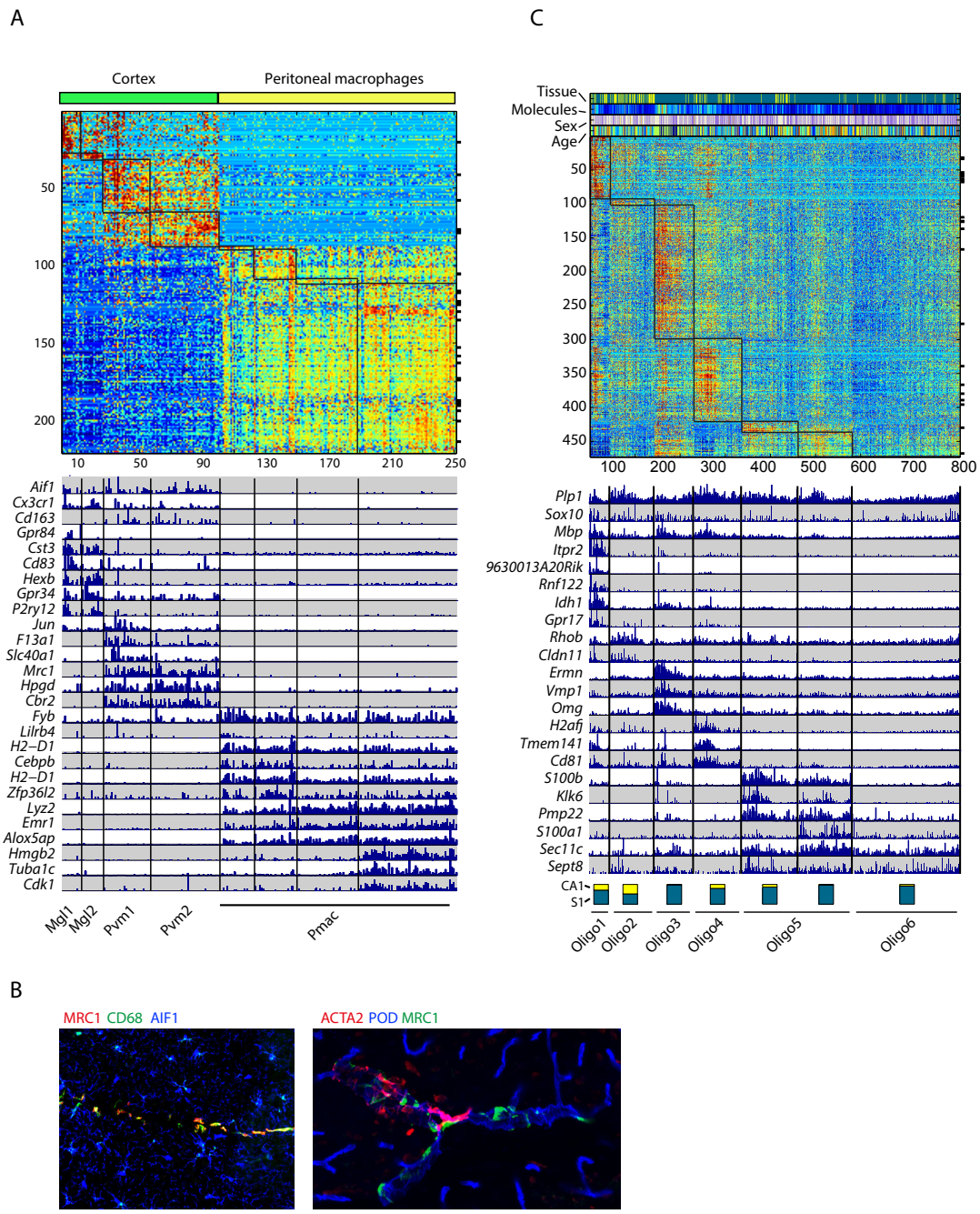

Fig. S9. Microglia and oligodendrocytes.

(A) Cluster analysis of microglia, brain perivascular and peritoneal macrophages. **(B)** Additional immunofluorescence stainings for microglia (Mgl) and perivascular macrophages (Pvm). MRC1 (Pvm), CD68 (Pvm), AIF1 (also known as Iba-1, Mgl + Pvm), ACTA2 (also known as ASMA, vascular smooth muscle cells), MRC1 (Pvm) and POD (Podocalyxin, vessels/endothelial cells). **(C)** Cluster analysis of oligodendrocytes, and barplots for selected known and novel markers.

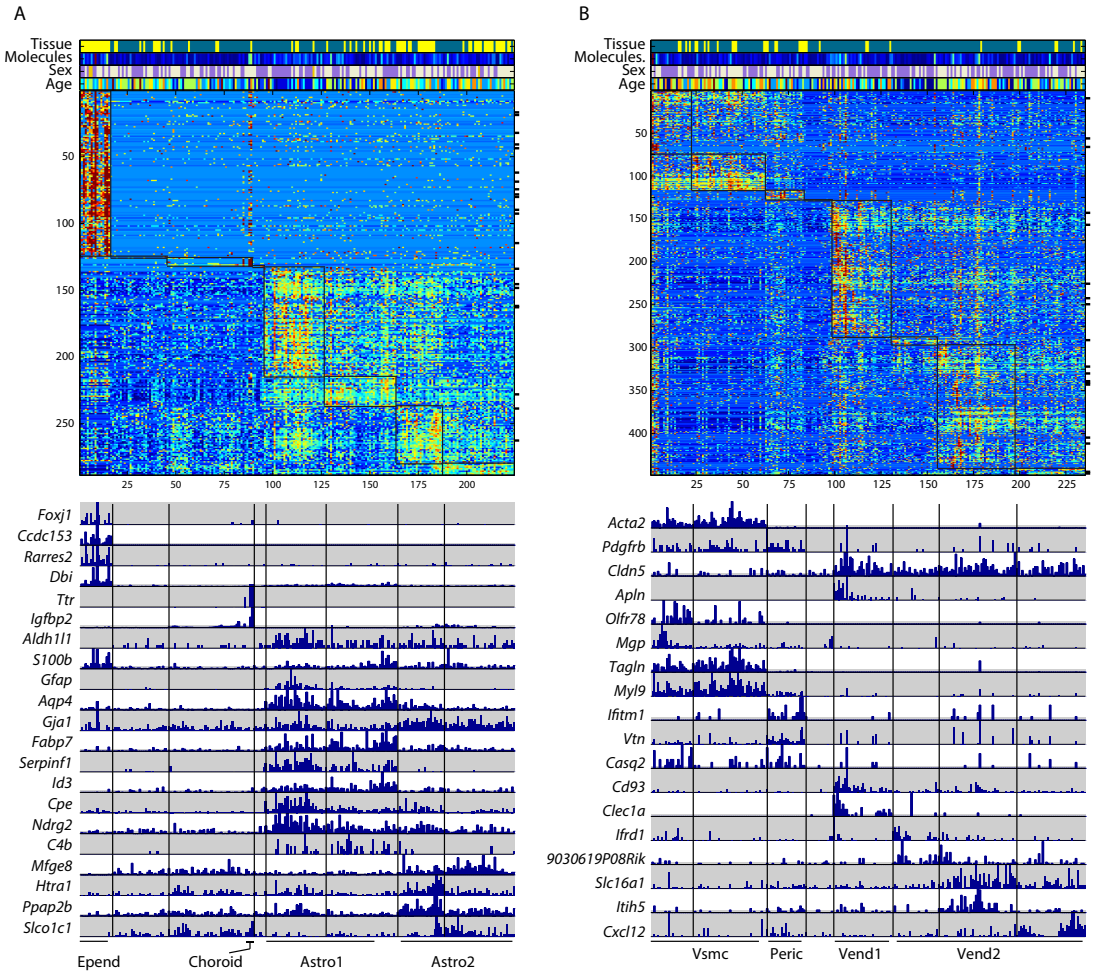

Fig. S10. Classification of astrocytes and vascular-related cells

(A) Clustering of ependymal, choroid plexus epithelial and astrocytes cells, using BackSPIN. Single ependymal (Epend) and choroid plexus epithelial cell (Choroid) classes were found, but analysis suggested two main subclasses of astrocytes (Astro1 and Astro2). Barplots show representative markers, ordered left-to-right as in the heatmap. **(B)** Clustering of vascular related cells, using BackSPIN. Vascular smooth muscle cells (Vsmc) and pericytes (Peric) were identified, as well as two distinct classes of endothelial cells (Vend1 and Vend2). Bar plots show representative markers.

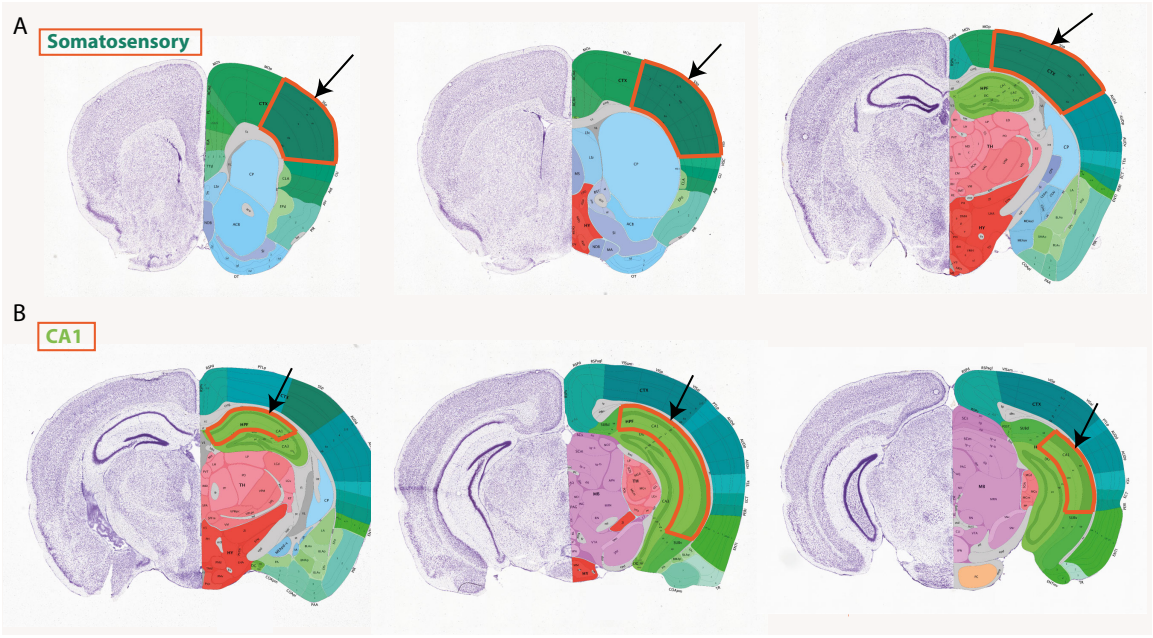

Fig. S11. Dissection scheme

(A) Dissected brain regions in the somatosensory cortex shown in orange contour on three coronal sections from different levels. **(B)** Dissected brain regions in the CA1 hippocampus shown in orange contour on three coronal sections.

References

1. B. J. Molyneaux, P. Arlotta, J. R. Menezes, J. D. Macklis, Neuronal subtype specification in the cerebral cortex. *Nat. Rev. Neurosci.* **8**, 427–437 (2007). Medline doi:10.1038/nrn2151

2. T. Klausberger, P. Somogyi, Neuronal diversity and temporal dynamics: The unity of hippocampal circuit operations. *Science* **321**, 53–57 (2008). Medline doi:10.1126/science.1149381

3. J. DeFelipe, P. L. López-Cruz, R. Benavides-Piccione, C. Bielza, P. Larrañaga, S. Anderson, A. Burkhalter, B. Cauli, A. Fairén, D. Feldmeyer, G. Fishell, D. Fitzpatrick, T. F. Freund, G. González-Burgos, S. Hestrin, S. Hill, P. R. Hof, J. Huang, E. G. Jones, Y. Kawaguchi, Z. Kisvárday, Y. Kubota, D. A. Lewis, O. Marín, H. Markram, C. J. McBain, H. S. Meyer, H. Monyer, S. B. Nelson, K. Rockland, J. Rossier, J. L. Rubenstein, B. Rudy, M. Scanziani, G. M. Shepherd, C. C. Sherwood, J. F. Staiger, G. Tamás, A. Thomson, Y. Wang, R. Yuste, G. A. Ascoli, New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nat. Rev. Neurosci.* **14**, 202–216 (2013). Medline

4. K. Sugino, C. M. Hempel, M. N. Miller, A. M. Hattox, P. Shapiro, C. Wu, Z. J. Huang, S. B. Nelson, Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nat. Neurosci.* **9**, 99–107 (2006). Medline doi:10.1038/nn1618

5. G. Fishell, B. Rudy, Mechanisms of inhibition within the telencephalon: “Where the wild things are”. *Annu. Rev. Neurosci.* **34**, 535–567 (2011). Medline doi:10.1146/annurev-neuro-061010-113717

6. E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T. M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H. W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K. R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, A. R. Jones, Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007). Medline doi:10.1038/nature05453

7. A. Kepecs, G. Fishell, Interneuron cell types are fit to function. *Nature* **505**, 318–326 (2014). Medline doi:10.1038/nature12983

8. D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, I. Amit, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014). Medline

9. B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, S. R. Quake, Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014). Medline doi:10.1038/nature13173

10. A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D. W. Kemp 2nd, M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein, J. A. West, Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014). Medline doi:10.1038/nbt.2967

11. S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, S. Linnarsson, Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014). Medline doi:10.1038/nmeth.2772

12. T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011). Medline doi:10.1038/nmeth.1778

13. D. Tsafrir, I. Tsafrir, L. Ein-Dor, O. Zuk, D. A. Notterman, E. Domany, Sorting points into neighborhoods (SPIN): Data analysis and visualization by ordering distance matrices. *Bioinformatics* **21**, 2301–2308 (2005). Medline doi:10.1093/bioinformatics/bti329

14. C. Shi, E. G. Pamer, Monocyte recruitment during infection and inflammation. *Nat. Rev. Immunol.* **11**, 762–774 (2011). Medline doi:10.1038/nri3070

15. O. Kann, C. Huchzermeyer, R. Kovács, S. Wirtz, M. Schuelke, Gamma oscillations in the hippocampus require high complex I gene expression and strong functional performance of mitochondria. *Brain* **134**, 345–358 (2011). Medline doi:10.1093/brain/awq333

16. H. W. Dong, L. W. Swanson, L. Chen, M. S. Fanselow, A. W. Toga, Genomic-anatomic evidence for distinct functional domains in hippocampal field CA1. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11794–11799 (2009). Medline doi:10.1073/pnas.0812608106

17. K. Mizuseki, K. Diba, E. Pastalkova, G. Buzsáki, Hippocampal CA1 pyramidal cells form functionally distinct sublayers. *Nat. Neurosci.* **14**, 1174–1181 (2011). Medline doi:10.1038/nn.2894

18. L. Tricoire, K. A. Pelkey, M. I. Daw, V. H. Sousa, G. Miyoshi, B. Jeffries, B. Cauli, G. Fishell, C. J. McBain, Common origins of hippocampal Ivy and nitric oxide synthase expressing neurogliaform cells. *J. Neurosci.* **30**, 2165–2176 (2010). Medline doi:10.1523/JNEUROSCI.5123-09.2010

19. M. Prinz, J. Priller, Microglia and brain macrophages in the molecular age: From origin to neuropsychiatric disease. *Nat. Rev. Neurosci.* **15**, 300–312 (2014). Medline doi:10.1038/nrn3722

20. I. Galea, K. Palin, T. A. Newman, N. Van Rooijen, V. H. Perry, D. Boche, Mannose receptor expression specifically reveals perivascular macrophages in normal, injured, and diseased mouse brain. *Glia* **49**, 375–384 (2005). Medline doi:10.1002/glia.20124

21. Y. Zhang, C. Pak, Y. Han, H. Ahlenius, Z. Zhang, S. Chanda, S. Marro, C. Patzke, C. Acuna, J. Covy, W. Xu, N. Yang, T. Danko, L. Chen, M. Wernig, T. C. Südhof, Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* **78**, 785–798 (2013). Medline doi:10.1016/j.neuron.2013.05.029

22. M. Kohyama, W. Ise, B. T. Edelson, P. R. Wilker, K. Hildner, C. Mejia, W. A. Frazier, T. L. Murphy, K. M. Murphy, Role for Spi-C in the development of red pulp macrophages and splenic iron homeostasis. *Nature* **457**, 318–321 (2009). Medline doi:10.1038/nature07472

23. J. Thomas, L. Morlé, F. Soulavie, A. Laurençon, S. Sagnol, B. Durand, Transcriptional control of genes involved in ciliogenesis: A first step in making cilia. *Biol. Cell* **102**, 499–513 (2010). Medline doi:10.1042/BC20100035

24. E. R. Brooks, J. B. Wallingford, Multiciliated cells. *Curr. Biol.* **24**, R973–R982 (2014). Medline doi:10.1016/j.cub.2014.08.047

25. M. A. Zariwala, H. Y. Gee, M. Kurkowiak, D. A. Al-Mutairi, M. W. Leigh, T. W. Hurd, R. Hjeij, S. D. Dell, M. Chaki, G. W. Dougherty, M. Adan, P. C. Spear, J. Esteve-Rudd, N. T. Loges, M. Rosenfeld, K. A. Diaz, H. Olbrich, W. E. Wolf, E. Sheridan, T. F. Batten, J. Halbritter, J. D. Porath, S. Kohl, S. Lovric, D. Y. Hwang, J. E. Pittman, K. A. Burns, T. W. Ferkol, S. D. Sagel, K. N. Olivier, L. C. Morgan, C. Werner, J. Raidt, P. Pennekamp, Z. Sun, W. Zhou, R. Airik, S. Natarajan, S. J. Allen, I. Amirav, D. Wieczorek, K. Landwehr, K. Nielsen, N. Schwerk, J. Sertic, G. Köhler, J. Washburn, S. Levy, S. Fan, C. Koerner-Rettberg, S. Amselem, D. S. Williams, B. J. Mitchell, I. A. Drummond, E. A. Otto, H. Omran, M. R. Knowles, F. Hildebrandt, ZMYND10 is mutated in primary ciliary dyskinesia and interacts with LRRC6. *Am. J. Hum. Genet.* **93**, 336–345 (2013). Medline

26. M. I. Chung, T. Kwon, F. Tu, E. R. Brooks, R. Gupta, M. Meyer, J. C. Baker, E. M. Marcotte, J. B. Wallingford, Coordinated genomic control of ciliogenesis and cell movement by RFX2. *eLife* **3**, e01439 (2014). Medline doi:10.7554/eLife.01439

27. S. Lee, J. Hjerling-Leffler, E. Zagha, G. Fishell, B. Rudy, The largest group of superficial neocortical GABAergic interneurons expresses ionotropic serotonin receptors. *J. Neurosci.* **30**, 16796–16808 (2010). Medline doi:10.1523/JNEUROSCI.1869-10.2010

28. A. Edelstein, N. Amodaj, K. Hoover, R. Vale, N. Stuurman, Computer control of microscopes using μManager. *Curr. Protoc. Mol. Biol.* **92**, 14.20.1–14.20.17 (2010). Medline doi:10.1002/0471142727.mb1420s92

29. A. Lyubimova, S. Itzkovitz, J. P. Junker, Z. P. Fan, X. Wu, A. van Oudenaarden, Single-molecule mRNA detection and counting in mammalian tissue. *Nat. Protoc.* **8**, 1743–1758 (2013). Medline doi:10.1038/nprot.2013.109

30. A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, S. Tyagi, Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008). Medline doi:10.1038/nmeth.1253

31. A. B. Muñoz-Manchado, C. Foldi, S. Szydlowski, L. Sjulson, M. Farries, C. Wilson, G. Silberberg, J. Hjerling-Leffler, Novel striatal GABAergic interneuron populations labeled in the 5HT3aEGFP mouse. *Cereb. Cortex* (2014). Medline doi:10.1093/cercor/bhu179

32. D. Tsafrir, I. Tsafrir, L. Ein-Dor, O. Zuk, D. A. Notterman, E. Domany, Sorting points into neighborhoods (SPIN): Data analysis and visualization by ordering distance matrices. *Bioinformatics* **21**, 2301–2308 (2005). Medline doi:10.1093/bioinformatics/bti329

33. G. Getz, E. Levine, E. Domany, Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 12079–12084 (2000). Medline doi:10.1073/pnas.210134797

34. E. Levine, E. Domany, Resampling method for unsupervised estimation of cluster validity. *Neural Comput.* **13**, 2573–2593 (2001). Medline doi:10.1162/089976601753196030

35. B. J. Frey, D. Dueck, Clustering by passing messages between data points. *Science* **315**, 972–976 (2007). Medline doi:10.1126/science.1136800

36. R. Cangelosi, A. Goriely, Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* **2**, 2 (2007). Medline doi:10.1186/1745-6150-2-2