Title

Data-driven approaches for improving the interpretability of patch-seq data

Authors

Shreejoy J. Tripathy[1,2,*], Lilah Toker[1,2], Claire Bomkamp[2], B. Ogan Mancarci[1,2], Manuel Belmadani[1,2], Paul Pavlidis[1,2*]

[1] Michael Smith Laboratories, University of British Columbia, Vancouver BC, Canada

[1] Department of Psychiatry, University of British Columbia, Vancouver BC, Canada

[*] Corresponding Author
Email: stripathy@msl.ubc.ca
Email: paul@msl.ubc.ca

Abstract

Patch-seq, enabling simultaneous measurement of a transcriptomic, electrophysiological, and morphological features, has recently emerged as a powerful tool for neuronal characterization. However, we show the method is susceptible to technical artifacts, including the presence of mRNA contaminants from multiple cells, that limit the interpretability of the data. We present a straightforward marker gene-based approach for controlling for these artifacts and show that our method improves the correspondence between gene expression and electrophysiological features.

Manuscript body

The patch-seq methodology combines recent advances in single cell RNAseq with standard patch-clamp electrophysiological methods [1–5]. Patch-seq mRNA sample collection differs from standard single-cell or single-nucleus RNAseq in two major ways. First, as opposed to relying on cellular dissociation, the micropipette used for electrical recording is used for mRNA extraction via aspiration. While guiding the patch pipette to the soma of a cell of interest, the pipette often must travel through the processes of other cells, presenting an opportunity for contamination. Second, the effectiveness of cell content aspiration is difficult to control, so the amount of mRNA extracted tends to vary from cell to cell.

To quantitatively assess the influence of these technical confounds, we performed a re-analysis of three published patch-seq datasets obtained from mouse acute brain slices (Supplementary Table 1) [1,3,5]. First, we wanted to know if the single cell RNAseq samples may have been contaminated by mRNA from other cells. We made use of the fact that the broad identities of the recorded cells could be ascertained from morphological and electrophysiological features without relying on the transcriptomic data (see Methods). We inspected the expression of various classes of marker genes for each scRNAseq sample, with markers based on cell type-specific expression profiles from dissociated cells (Supplementary Table 4) [6]. We observed single cell samples that strongly expressed markers of multiple distinct cell types (Figure 1, Supplement Figure 1). For example, some of the elongated neurogliaform cells (eNGCs) characterized in the Cadwell dataset appeared to also express marker genes specific to

pyramidal cells. Similarly, many of the cells identified as regular spiking GABAergic interneurons in the Földy dataset also expressed high levels of microglial and pyramidal cell markers.

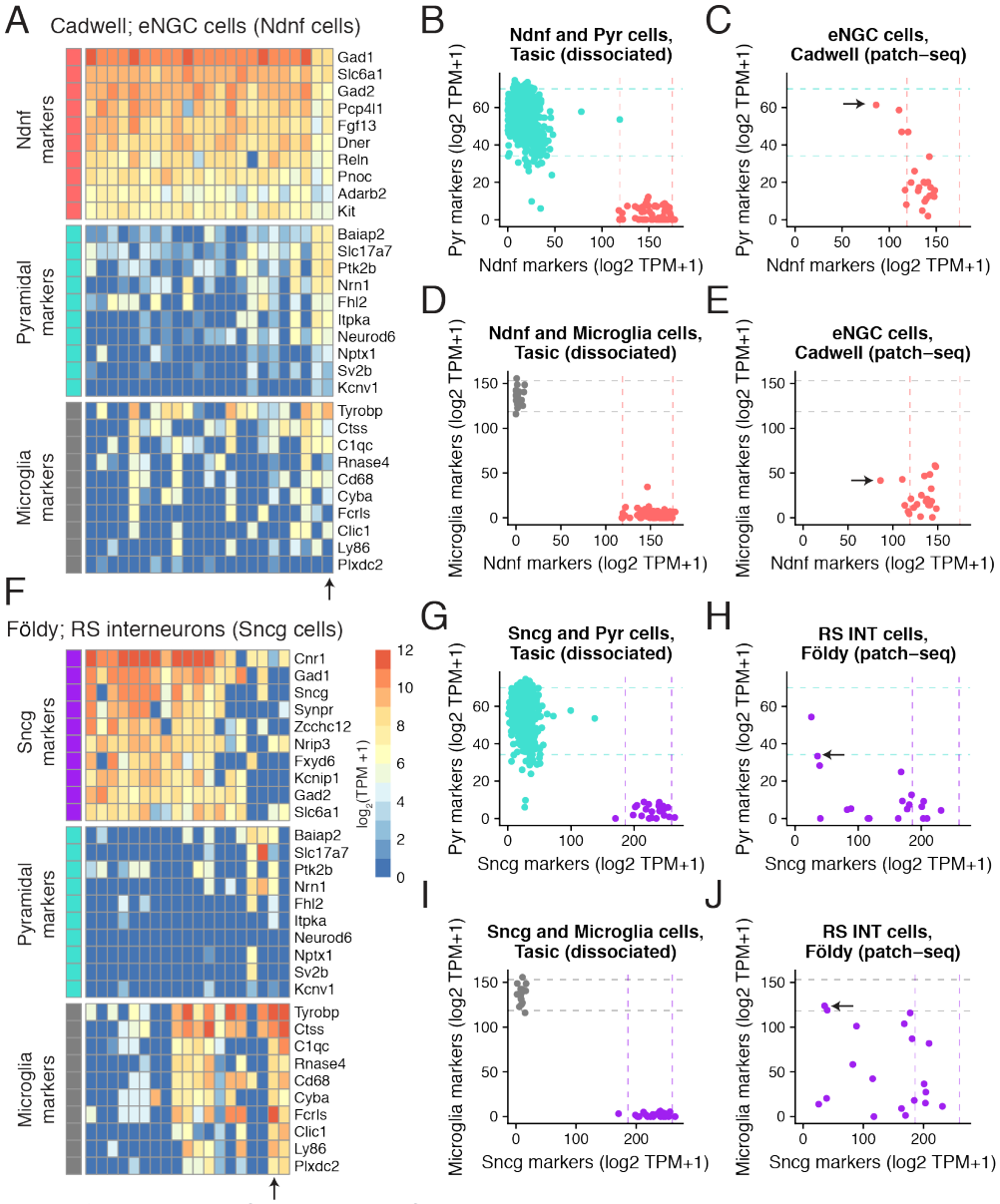


Figure 1: Expression of cell type-specific marker genes in single-cell samples collected using patch-seq. A) Gene expression profiles for GABAergic elongated neurogliaform cells (eNGCs, similar to layer 1 Ndnf cellular subtype) for various cell type-specific markers. Each column reflects a single-cell sample. B) Summed expression of cell type-specific marker genes for Pyramidal cell (y-axis) and Layer 1 Ndnf cell (x-axis) markers. Dots reflect Pyramidal (turquoise) and Ndnf (red) single cells collected in Tasic dataset, based on dissociated scRNAseq. Dashed lines reflect 95% intervals of marker expression for each cell type. C) Same as B, but showing summed marker expression for eNGC cells shown in A based on patch-seq data. Arrow shows single-cell marked in A. D,E) Same as B and C, but for microglial cell markers. F-J) Same as A-E, but for hippocampal GABAergic regular spiking interneurons (RS INT cells, similar to Sncg cells from in Tasic) characterized in Földy dataset.

To further quantify the extent of apparent cellular contamination, we directly compared the patch-seq-based expression profiles to cellular dissociation-based transcriptomes from two recent surveys of mouse cortical diversity from Tasic et al. and Zeisel et al. [7,8]. After matching cell type identities across studies (shown in Supplementary Table 2), we found that compared to

dissociated cells, patch-seq-based samples expressed markers for multiple cell types at considerably higher levels (Figure 1. C, H, J; Supplement Figure 2A, B). Importantly, patch-seq-based samples that appeared more contaminated expressed markers of their own cell type at lower levels relative to less contaminated cells (Supplement Figure 3). We saw less off-cell-type marker expression in the Fuzik dataset (Supplement Figure 2), suggesting either less contamination in these cells or that the lower gene detection rate in this dataset (Figure 2A) obscures our ability to use expression profiles to identify cellular contamination.

Next, we explored if there are identifiable technical factors that can help explain the large variance in genes detected per cell, from 8500 +/- 1200 genes / cell in Cadwell to 2000 +/- 1000 genes / cell in Fuzik (mean +/- s.d.; Figure 2A). Since the Cadwell dataset (and some cells in the Földy dataset) employed spike-in ERCC mRNAs, where the same amount of synthetic mRNAs was added to each sample prior to library preparation, we used these to estimate the relative amounts of extracted cellular mRNA [9,10]. Single-cell samples with lower initial amounts of cellular mRNA (i.e., with a higher ratio of spike-in / total reads) had fewer detected genes (Figure 2B). More contaminated samples, defined by total off cell-type marker expression (see Methods), had greater numbers of detected genes (Figure 2C), consistent with previous reports [9,11]. We next asked how these factors contribute to the numbers of genes detected per cell, using a sequential analysis of variance approach (see Methods). Relative to the Ndnf cell subset of the Tasic dataset (with 7000 +/- 650 genes/cell), we found that even after accounting for differences in library sizes (i.e., numbers of sequenced reads per cell), the spike-in ratio and cellular contamination explain considerably more variance in the Cadwell patch-seq data relative to Tasic (Figure 2D). Our analysis of the subset of the Földy dataset containing spike-ins further suggests that sufficient mRNA extraction is critical for detecting cellular contamination (when present) and the expression of expected markers (*Supplement Figure 4*).

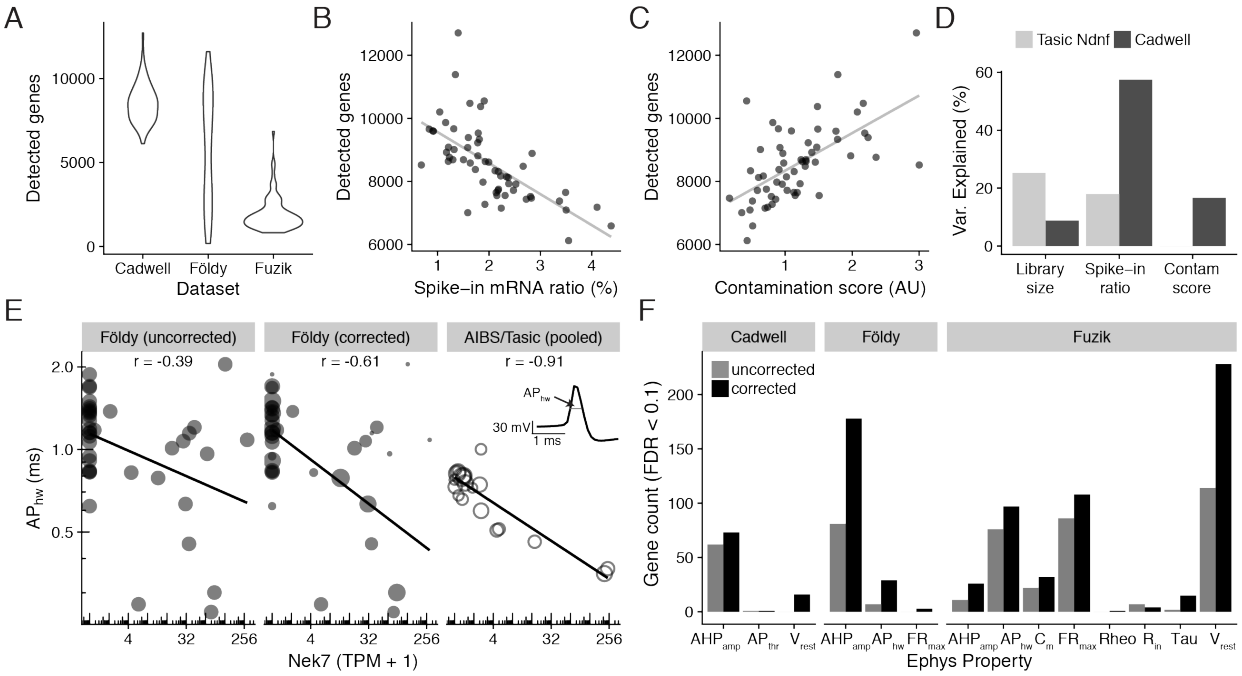


Figure 2. Patch-seq experimental confounds affect numbers of genes detected and correspondence with electrophysiological measures. A) Violin plots showing numbers of genes detected per cell across patch-seq datasets. B) Spike-ins as a fraction of all sequenced reads per single-cell patch-seq sample (dots) in the Cadwell dataset. Samples with lower cellular mRNA content (indicated by higher spike-in ratios) have lower gene counts. C) Same as B, but x-axis shows cellular contamination score, quantified by summing normalized contamination values across tested cell types (arbitrary units). D) Factors associated with numbers of genes detected per cell in dissociated

scRNAseq (Tasic, Ndnf cells) and patch-seq (Cadwell). Bars indicate percent variance in detected gene counts explained by library size, spike-in ratio, and cellular contamination scores. E) Comparison of gene expression (Nek7; x-axis) with electrophysiological features (action potential half-width; AP_{hw}; y-axis). Left panel shows single-cell samples (circles) from the Földy dataset. Middle panel shows same data as left, but size of circles proportional to each sample's quality score, defined as the similarity of marker expression to dissociated cell-based reference data. Right panel shows cell type-level analysis based on pooled cell type data from Allen Institute cell types database (AIBS/Tasic), where scRNAseq and electrophysiology were performed on different cells from same type [12]. Each open circle reflects one cell type and circle size is proportional to the number of cells representing each cell type. Inset illustrates calculation of action potential half-width (schematic). F) Count of genes significantly correlated (FDR < 0.1) with various electrophysiological properties before (grey) and after (black) correcting for contamination.

Lastly, we performed an integrated analysis of gene expression and electrophysiological features, reasoning that more contaminated patch-seq samples would be overall less informative of relationships between cellular electrophysiology and gene expression [12]. We first calculated a quality score for each patch-seq sample, based on the similarity of its marker expression to dissociated cells of its same type (see Methods; Supplement Figure 3). After statistically down-weighting lower quality cells, we observed a modest improvement in the correspondence between gene expression and electrophysiology, as evidenced by an increase in the number of genes significantly correlated with electrophysiological features (FDR < 0.1, Figure 2E-F). In addition, after correction, we found more genes overlapping with those identified in our previous gene-electrophysiology correlation analysis based on pooled cell types [12] (Figure 2E; Supplement Figure 5). While the biological implications of these correlations require further investigation, this analysis suggests that controlling for these technical factors can help improve the interpretability of patch-seq data.

To summarize, though patch-seq provides a powerful method for multi-modal neuronal characterization [2–5], it is susceptible to a number of methodology-specific technical artifacts, such as an increased likelihood of mRNA contamination from adjacent cells. These artifacts strongly bias traditional scRNAseq quality metrics such as the numbers of genes detected per cell. By leveraging high-quality reference atlases of single-cell transcriptomic diversity [7,8], we argue that inspection of cell type-specific marker expression should be an essential patch-seq quality control step prior to downstream analyses.

Acknowledgements

We thank Cathryn Cadwell, Janos Fuzik, and Csaba Földy for sharing data and Jim Berg, Cedric Bardy, Dmitry Kobak, and Philipp Berens for helpful discussions. This work was supported by Kids Brain Health Network, a Canadian Institute for Health Research Post-doctoral Fellowship, and NIH grant MH111099.

Author contributions

SJT and PP conceived the project. SJT implemented the methodology and generated the results with assistance from LT, OBM, CB and MB. All authors contributed to interpreting the results. SJT and PP wrote the paper with assistance from all authors.

Competing interests

The authors declare no competing financial interests.

References

1. Cadwell, C. R. *et al.* Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3445
2. Cadwell, C. R. *et al.* Multimodal profiling of single-cell morphology, electrophysiology, and gene expression using Patch-seq. *Nat. Protoc.* **12,** nprot.2017.120 (2017).
3. Földy, C. *et al.* Single-cell RNAseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. *Proc. Natl. Acad. Sci.* **113,** E5222–E5231 (2016).
4. Bardy, C. *et al.* Predicting the functional states of human iPSC-derived neurons with single-cell RNA-seq and electrophysiology. *Mol. Psychiatry* (2016). doi:10.1038/mp.2016.158
5. Fuzik, J. *et al.* Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nat. Biotechnol.* **34,** 175–183 (2016).
6. Mancarci, B. O. *et al.* Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. *eNeuro* ENEURO.0212-17.2017 (2017). doi:10.1523/ENEURO.0212-17.2017
7. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19,** 335–346 (2016).
8. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347,** 1138–1142 (2015).
9. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* (2017). doi:10.1038/nmeth.4292
10. Lun, A. T. L., Calero-Nieto, F. J., Haim-Vilmovsky, L., Göttgens, B. & Marioni, J. C. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* (2017). doi:10.1101/gr.222877.117
11. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17,** 29 (2016).

12. Tripathy, S. J. *et al.* Transcriptomic correlates of neuron electrophysiological diversity. *PLOS Comput. Biol.* **13,** e1005814 (2017).

13. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

14. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29,** 15–21 (2013).

15. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9,** 884 (2018).

16. Teeter, C. *et al.* Generalized leaky integrate-and-fire models classify multiple neuron types. *Nat. Commun.* **9,** 709 (2018).

Methods

Dataset overview

We made use of 3 previously published patch-seq datasets (Cadwell, Földy, Fuzik) [1,3,5] and 2 cellular dissociation-based single-cell RNAseq datasets (Tasic, Zeisel) [7,8]. We downloaded single-cell transcriptomic data from each study from accessions provided in Supplementary Table 1). We obtained patch-seq-based electrophysiological data for the Cadwell and Fuzik datasets from the authors. For all patch-seq datasets, electrophysiological data were provided as a spreadsheet containing a set of summarized electrophysiological features per cell (e.g., input resistance, resting membrane potential, etc.). Electrophysiological data from the Allen Institute Cell Types database (celltypes.brain-map.org) were obtained and processed as described previously [12].

Transcriptome data pre-processing

We reprocessed transcriptomic data for the Cadwell, Földy, and Tasic datasets directly from Gene Expression Omnibus (GEO) or Array Express. Data from GEO was downloaded using fastq-dump version 2.8.2 from the Sequence Read Archive Toolkit. Technical reads such as barcodes and primers were filtered out during extraction. Adapter sequences were clipped from the raw reads. The list of option used is as follows: '--gzip --skip-technical --readids --dumpbase --split-files --clip'. Data from ArrayExpress was downloaded and used directly as prepared by the European Bioinformatics Institute.

The reference mouse transcriptome was produced using the 'rsem-prepare-reference' script provided by the RSEM RNA-Seq transcript quantifier [13]. The assembly version used was Ensembl GRCm38, packaged by Illumina for the iGenomes collection. Alignment was

performed using STAR [14] version 2.4.0h, provided as the aligner to RSEM v1.2.31. Default parameters were used (with the exception of parallel processing and logging related options). Transcript definitions used to detect ERCC spike-ins were obtained from the ERCC92 version fasta and GTF files. Spike-ins were concatenated to the GRCm38 assembly before applying rsem-prepare-reference, and independently to create a standalone ERCC assembly. Both the concatenated and standalone spike-ins assemblies showed highly comparable proportions of spike-in expression. For the Fuzik and Zeisel datasets, we made use of the quantified summarized unique molecule counts (UMIs) made available at GEO.

Mapping of patch-seq cell types onto taxonomies derived from dissociated cells

Using descriptions for cellular identities provided in the original patch-seq publications, we manually mapped each of the cell types represented across the three patch-seq datasets onto transcriptomically-defined cellular clusters reported in the two dissociated cell datasets (shown in Supplementary Table 2). For example, given that the elongated neurogliaform cells and single bouquet cells characterized in Cadwell are both cortical layer 1 cells, we manually mapped these to the layer 1 cells defined in Tasic as Ndnf cells. Similarly, we mapped the hippocampal regular-spiking interneurons characterized in Foldy to the Sncg cluster from Tasic (personal communication with Csaba Földy). To align cell subtype clusters between Tasic and Zeisel, we used mappings provided by MetaNeighbor [15] (shown in Supplementary Table 2). The mappings between broad cell types in Tasic with Zeisel are provided in Supplementary Table 3. As with our previous work mapping cells and cell types across datasets [6,12], we note that these cross-dataset mappings are approximate and ideally would be guided by the use methods for unambiguously aligning cell types across experiments (e.g., transgenic mouse lines with specific cell types labeled by fluorescent proteins).

Identification of cell type-specific marker genes

For this study, we defined two classes of marker genes, termed “on” and “off” markers. The first class, “on” markers, are genes that are highly and ubiquitously expressed in the cell type of interest with enriched expression relative to other cell types. The second class, “off” markers, are expected to be expressed at low levels in a given patch-seq cell type. These are genes that are specifically expressed in a single cell type (e.g., microglia) and, if expressed, are an indicator of possible cellular contamination. To identify marker genes, we employed two recent surveys of mouse cortical diversity from Tasic et al. and Zeisel et al. [7,8].

To identify “on” marker genes, we initially used the Tasic dataset, and selected genes whose average expression in the chosen cell type was >10 times relative all other cell types in the dataset, with an average expression in the cell type of >100 TPM. From this initial gene list, we next filtered these genes to only include those that were expressed >10 TPM/cell in >75% of all cells of that type in Tasic, and >1 UMI/cell in >50% of all cells of that type in Zeisel. Using the Tasic nomenclature, we defined “on” markers for Ndnf, Sncg, Pvalb, and Pyramidal cell types.

To identify “off” marker genes for broad cell types (shown in Supplementary Table 3), as an initial listing we used the set of cell type-specific marker genes for broad cell classes in the mouse cortex, defined in our previous work using the NeuroExpresso database [7]. Specifically, we used the set of cortical markers derived from single-cell RNA-seq for astrocytes, endothelial cells, microglia, oligodendrocytes, oligodendrocyte precursor cells, and pyramidal cells. From this list, we first filtered out lowly expressed genes that were expressed <10 TPM/cell in >50% of all cells of that type in Tasic, and <1 UMI/cell in >50% of all cells of that type in Zeisel. Next, we filtered genes too broadly expressed in our patch-seq cell types of interest by assessing the

expression of these genes in the Ndnf, Sncg, Pvalb, and Pyramidal cell types, removing genes that were expressed at a level greater than >10 TPM/cell in >33% of all cells of that type in Tasic, and >2 UMI/cell in >33% of all cells of that type in Zeisel.

When defining on and off marker genes for inhibitory cell subtypes (e.g., the Ndnf cell type), we did not compare these cells to other GABAergic cells. For example, when defining "on" markers for Ndnf cells, we did not compare these cells' expression to Pvalb or Sst cells. We note that this choice limits our ability to identify inhibitory-to-inhibitory cell contamination, for example, an Ndnf cell contaminated by Sst-cell specific markers. To define an initial set of "off" markers for GABAergic inhibitory cells, we first obtained a list of genes based on Tasic where in GABAergic cells had average expression >10 times all other non-GABAergic cells in the dataset and with an average expression of at least 100 TPM.

The final list of filtered cell type specific marker genes used in this study are provided in Supplementary Table 4.

Summarizing cell type-specific marker expression

When directly comparing expression values from patch-seq data to dissociated cell data, we compared the Cadwell and Földy datasets to Tasic, as these all were quantified using TPM, and the Fuzik dataset to Zeisel, as these both were quantified using unique molecule identifiers (UMIs), normalized as UMI counts per million. We summarized a single-cell sample's expression of multiple cell type-specific markers using the sum of the $\log_2$ normalized expression values. Given a patch-seq sample of cell type identity A (e.g., a pyramidal cell) and wanting to quantify its normalized expression of "off" markers for cell type B (e.g., microglial markers), we used the dissociated cell data to estimate the median expression of cell type B's "markers in cells of type A (e.g., median expression level of microglial markers in pyramidal cells) and the median expression of cell type B's markers in cells of type B (e.g., median expression level of microglial markers in microglia cells). Specifically, we normalized expression to a value of approximately 0 to 1, as follows:

*(PatchSeqCellTypeA*_markersB – median(*DissocCellTypeA*_markersB))/
(median(*DissocCellTypeB*_markersB) - median(*DissocCellTypeA*_markersB))

where we set all negative values to 0. Next, to obtain a single contamination score per single cell, we summed all contamination scores for all broad cell types, excluding the patch-seq cell's assigned broad cell type.

Lastly, to obtain a scalar quality score for transcriptomic data from patch-seq samples (e.g., for analysis of electrophysiological data), we used the Spearman correlation of each patch-seq sample's expression of "on" and "off" marker genes to the average expression profile of dissociated cells of the same cell type (shown in Supplement Figure 3). For example, for an Ndnf patch-seq sample from Cadwell, we first calculated the average expression profile of Ndnf cells from Tasic across the set of all "on" and "off" marker genes (i.e., Ndnf markers, pyramidal cell markers, astrocyte markers, etc.), and then calculated the correlation between the patch-seq cell's marker expression to the mean dissociated cell expression profile. Since these correlations could potentially be negative, we set quality scores to a minimum of 0.1. A convenient feature of this quality score is that it yields low correlations for samples with relatively high contamination as well as those where contamination is largely undetected but expression of endogenous "on" markers is also low (Supplement Figure 4F, G).

Analysis of factors influencing the numbers of genes detected per cell

We analyzed how the following factors influenced the numbers of genes detected per cell: library size, defined as the total numbers of reads sequenced per cell; spike-in ratio, defined as the number of reads mapping to ERCC spike-ins divided by total sequenced reads; and cellular contamination scores, as define in the previous section. For the Cadwell and Tasic datasets, where each single-cell sample also contained ERCC spike-ins, we fit a linear model (implemented using the 'lm' function in R) for numbers of detected genes per each cell as follows:

num_genes ~ log10(library_size) + log10(spike-in ratio) + contam_score

We used a sequential analysis of variance approach (implemented using the 'anova' function in R, applied to the model fits), where we asked how much variance in the numbers of genes detected per cell could be explained by each subsequent term in the prior equation (i.e., first, library size is used to explain variance in number of genes detected, and subsequently, spike-in ratio, and lastly, contamination score). Because of the sequential nature of the analysis, if two variables are correlated (e.g., in Cadwell, spike-in ratio and contamination score have a correlation of r = -0.53), then the first variable will get assigned more overall variance explained.

Combined analysis of transcriptomic and electrophysiological features

We analyzed correlations between transcriptomic and electrophysiological features using an approach similar to our previous work [12]. For each patch-seq dataset, we first filtered for genes whose average expression was > 30th percentile relative to all genes in the dataset. We analyzed electrophysiological features overlapping with our previous analysis, specifically, input resistance (Rin), resting membrane potential (Vrest , action potential threshold (APthr) , action potential amplitude (APamp) , action potential half-width (APhw), membrane time constant (Tau), after-hyperpolarization amplitude (AHPamp), rheobase (Rheo), maximum firing rate (FRmax), and capacitance (Cm). We calculated Pearson correlations between the set of electrophysiology features and gene expression values, both without weighting cells by their overall quality scores (based on correlation of markers to dissociated cell samples), and after weighting cells using their quality scores.

We performed an analogous analysis for comparison of pooled-cell correlations based on the AIBS/Tasic dataset, where we computationally merged different groups of cells characterized using dissociated cell scRNAseq (based on Tasic et al, [7]) with cells characterized using patch-clamp electrophysiology [16] based on the overlap of same mouse transgenic lines and coarse cortical layers (i.e., upper vs lower mouse visual cortex). For example, we merged 14 QC-passing scRNAseq samples from the Sst-IRES-cre mouse line from visual cortex dissections specific to lower layers with 89 patch-clamp samples from the same mouse line from cortical layers 4 through 6b. After merging single-cells into cell types, we averaged expression and electrophysiological values; since cell types tended to be represented by differing numbers of cells, in our gene-electrophysiology correlation analyses we weighted cell types based on the numbers of cells available using the square root of the harmonic mean of the number of cells characterized by electrophysiology and electrophysiology.

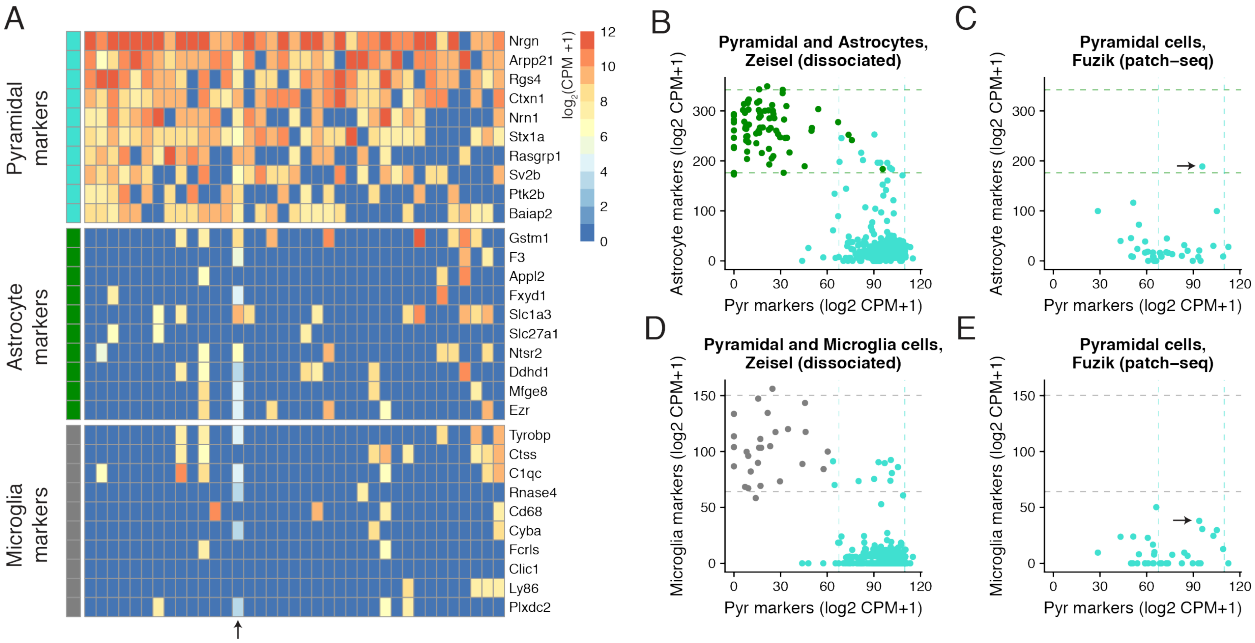
Statistical information

We used the R weights toolbox (v0.85) to calculate weighted Pearson correlations and raw p-values. We used the Benjamini-Hochberg False Discovery Rate (FDR) to account for analysis of multiple correlations.

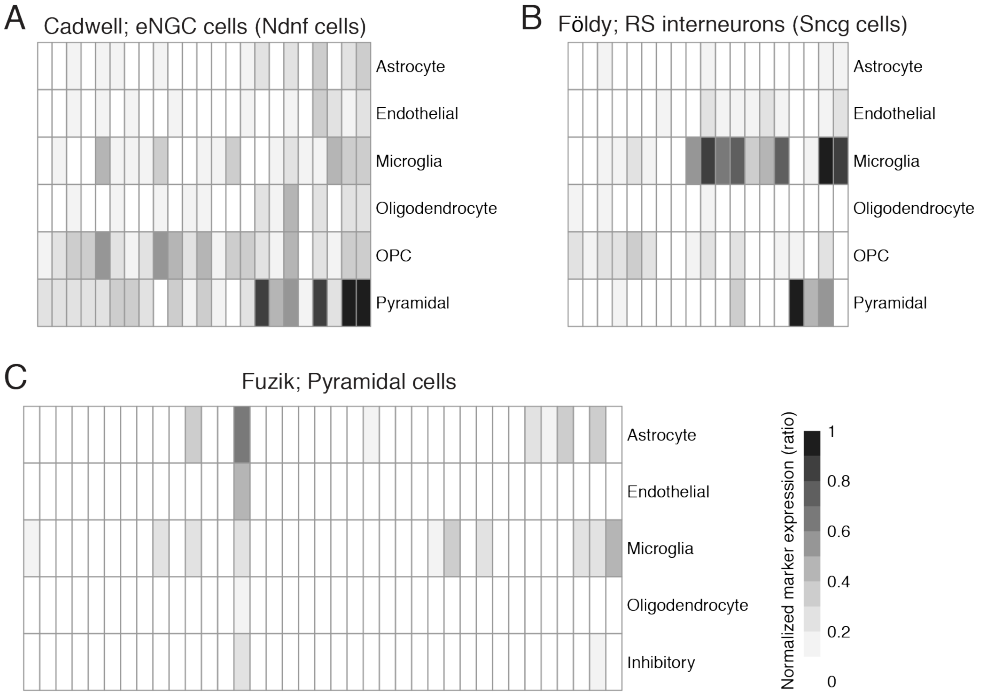
Computer code and data availability

All computational code and associated data has been made accessible at https://github.com/PavlidisLab/patchSeqQC and code for the RNAseq pipeline is accessible at https://github.com/PavlidisLab/rnaseq-pipeline.

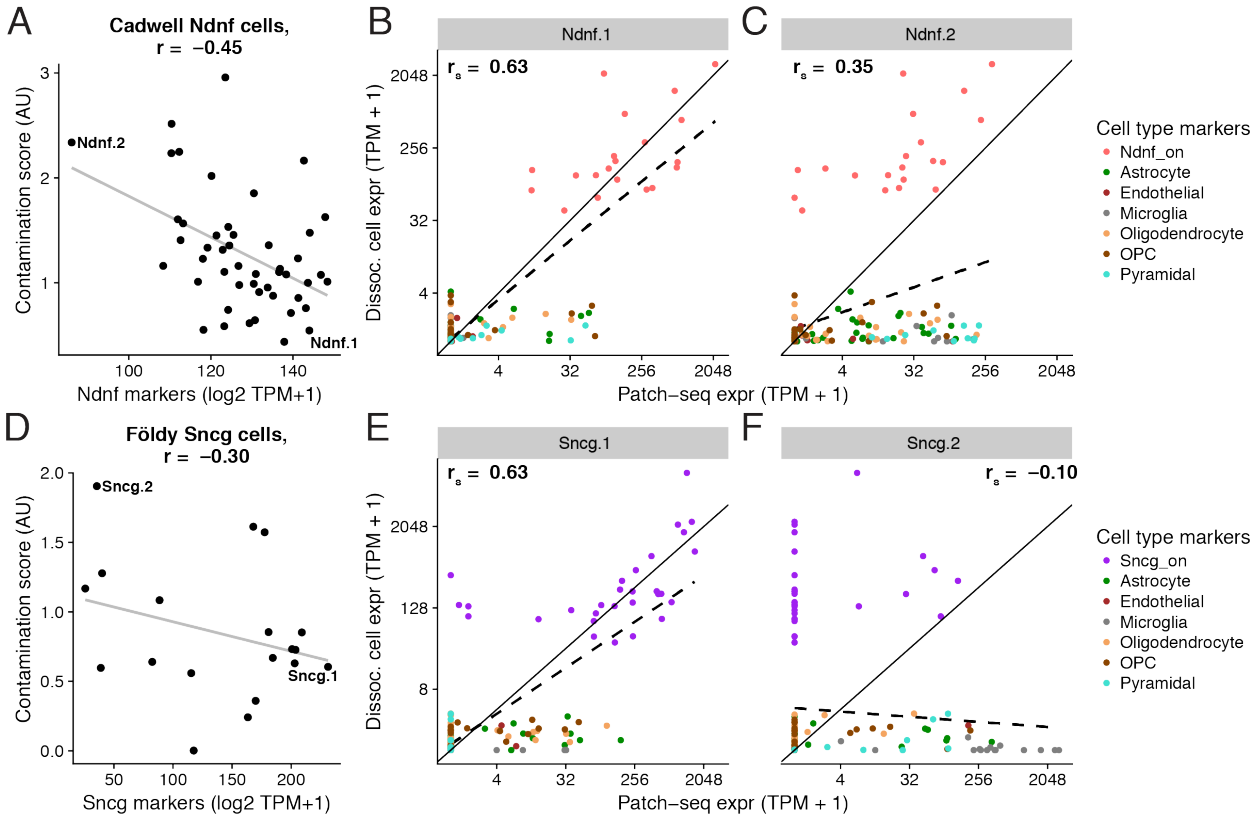
Supplemental Figures



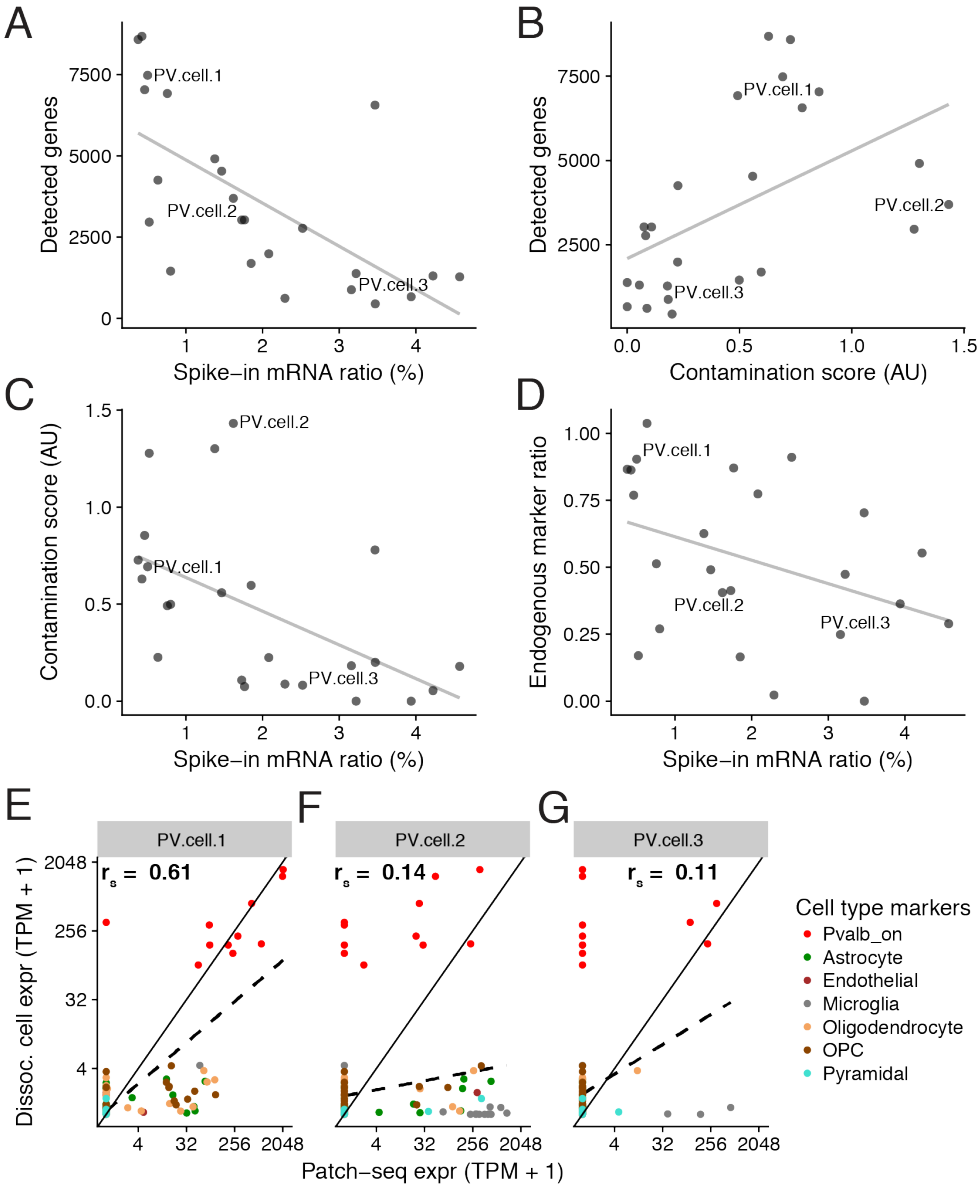
Supplement Figure 1. Expression of cell type-specific marker genes in patch-seq samples from Fuzik. A) Gene expression profiles for sampled pyramidal cells for various cell type-specific markers. B) Summed expression of cell type-specific marker genes for Pyramidal cell (x-axis) and Astrocyte (y-axis) markers. Dots reflect cortical Pyramidal cell (turquoise) and Astrocyte (green) single cells collected in Zeisel dataset, based on dissociated scRNAseq. Dashed lines reflect 95% intervals of marker expression for each cell type. C) Same as B, but showing summed marker expression for Pyramidal cells shown in A based on patch-seq data. Arrow denotes same single-cell highlighted in A. D,E) Same as B and C, but showing comparison of microglial marker expression.



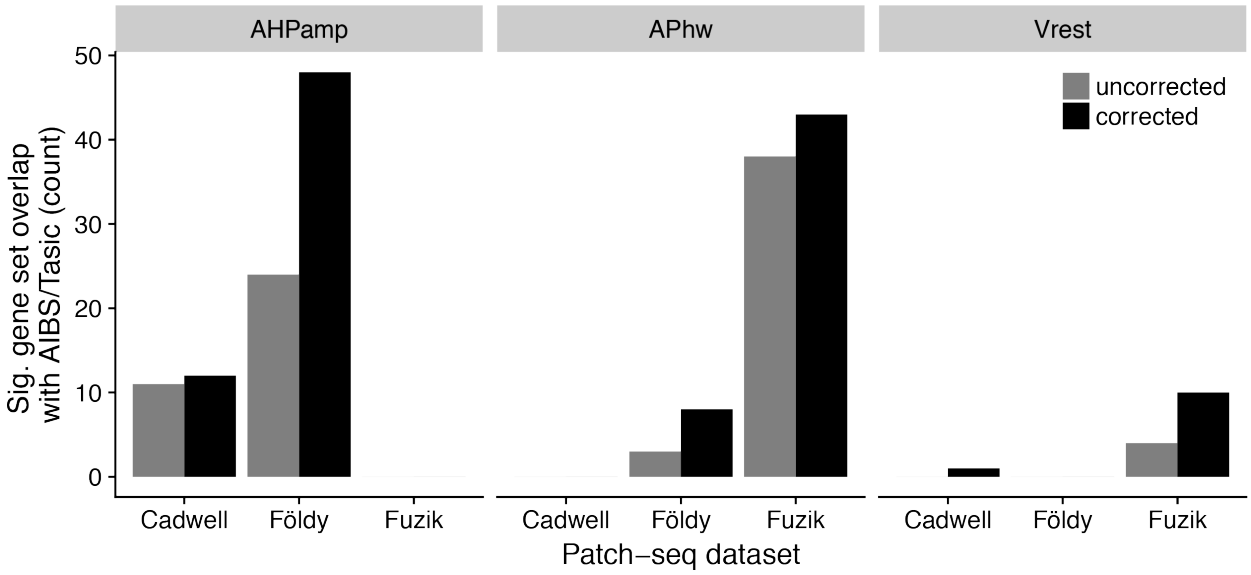
Supplement Figure 2. Summarized marker gene expression in patch-seq samples for broad cell classes. A) Cortical Layer 1 elongated neurogliaform cells (eNGCs) from Cadwell; B) Hippocampus regular spiking (RS) GABAergic interneurons from Földy; C) Cortical Pyramidal cells from Fuzik. Each column reflects a single-cell sample and columns are sorted as in Figure 1 and Supplement Figure 1. Heatmap colors show cell type-specific marker expression, normalized to expected expression based on dissociated cell reference datasets (Tasic, A, B; Zeisel, C). 0 indicates little-to-no detected off cell-type marker contamination (relative to dissociated cells) and 1 indicates strong expression of off-cell-type markers. Oligodendrocyte precursor cells not available in C because this cell type was not explicitly annotated in the Zeisel dataset.



Supplement Figure 3. Relationship between inferred contamination and endogenous marker expression. A) Summed expression of endogenous "on"-cell type cellular markers (x-axis) versus normalized contamination scores (y-axis, summing across normalized contamination values across broad cell types) for individual Ndnf cells from the Cadwell dataset (dots). B, C) Examples of "on"- and "off"-cell type marker expression for two single-cell patch-seq samples indicated in A. X-axis shows expression of marker genes (dots) in an individual patch-seq sampled cell and y-axis shows the average expression of the same markers in Ndnf-type dissociated cells from Tasic. Solid line is unity line, dashed line shows best linear fit, and r_s denotes Spearman correlation between patch-seq and mean dissociated cell marker expression. Cell Ndnf.1 (shown in B) illustrates a patch-seq sample with high expression of "on"-type endogenous markers and relatively little "off"-cell type marker expression whereas cell Ndnf.2 (shown in C) expresses endogenous markers less strongly (relative to dissociated cells of same type) and higher levels "off"-cell type marker expression. D-F) Same as A-C, but for hippocampal GABAergic regular spiking interneurons (i.e., Sncg cells) characterized in Földy dataset.



Supplement Figure 4. Numbers of detected genes, off-cell type contamination, and endogenous marker expression are limited by amounts of extracted mRNA. A) Spike-ins as a fraction of all sequenced reads per patch-seq sample (dots) in Földy dataset. Samples with lower mRNA content (indicated by higher spike-in ratios) show fewer detected genes. Note that ERCC spike-ins were added to only a subset of cells in this dataset (n = 24 cells). B) Same as A, but x-axis shows normalized contamination scores; as with the Cadwell dataset, cells with higher contamination scores have more genes detected. C) Contamination scores as a function of spike-in ratios. Note that less contamination is detected when less mRNA is extracted (i.e., higher spike-in ratios). D) Endogenous marker ratio (defined as the normalized sum of "on"-type marker expression, relative to dissociated cells of same type) versus spike-in ratios. Less "on"-cell type marker expression (endogenous markers) are detected when less mRNA is extracted. E-G) Examples of individual single-cell patch-seq samples highlighted in A-D. Solid line is unity line, dashed line shows best linear fit, and r_s denotes Spearman correlation between patch-seq and dissociated cell marker expression. E shows a cell (PV.cell.1) with a relatively high amount of extracted mRNA, high endogenous marker expression, and low-to-moderate detected contamination. F shows a cell (PV.cell.2) with a moderate amount of extracted mRNA, relatively high amount of contamination, and low-to-moderate endogenous marker expression. G shows a cell (PV.cell.3) with a relatively low amount of extracted mRNA and few detected genes, leading to both low detected contamination and low detected amounts of endogenous marker expression.



Supplement Figure 5. Comparison of genes significantly correlated (BH FDR < 0.1) with electrophysiological features based on patch-seq data with analogous correlations based on AIBS/Tasic dataset, pooled to the level of cell types based on cre-lines. Bars indicate count of overlapping genes between patch-seq and AIBS/Tasic pooled-cell data without correcting for contamination (uncorrected) and with correction (corrected).

Supplementary Tables

Dataset	Experiment type	Description	Accession	Number of cells
Cadwell	patch-seq	Cortical layer 1 interneurons	E-MTAB-4092	58
Földy	patch-seq	Hippocampal CA1 and Subiculum pyramidal cells and regular- and fast-spiking interneurons	GSE75386	93
Fuzik	patch-seq	Cortical layer 1/2 interneurons and pyramidal cells	GSE70844	80
Tasic	dissociated cell scRNAseq	Visual cortex neurons and glia	GSE71585	1366
Zeisel	dissociated cell scRNAseq	Somatosensory cortex and hippocampus neurons and glia	GSE60361	3005
Allen Institute Cell Types database	patch-clamp electrophysiology	Visual cortex neurons	celltypes.brain-map.org	952

Supplementary Table 1: Description of datasets used.

Patch-seq dataset	Cell type (patch-seq)	Matched cell type (dissociated cell; Tasic)	Matched cell type (dissociated cell; Zeisel)
Cadwell	Cortex Layer 1 elongated neuragliaform cell (eNGC)	Ndnf cluster (Ndnf Car4, Ndnf Cxcl14)	Int12, Int15
Cadwell	Cortex Layer 1 single bouquet cells (SBC)	Ndnf cluster (Ndnf Car4, Ndnf Cxcl14)	Int12, Int15
Földy	Hippocampus regular-spiking (RS) interneurons	Sncg	Int 5
Földy	Hippocampus CA1 and Subiculum Pyramidal cells	Pyramidal cluster	Pyramidal cluster (excluding CA1PyrInt)
Földy	Hippocampus fast-spiking (FS) interneurons	Pvalb cluster (Pvalb Gpx3, Pvalb Wt1, Pvalb Tacr3, Pvalb Tpbg, Pvalb Cpne5, Pvalb Rspo2, Pvalb Obox3)	Int 3
Fuzik	Cortex Layer 1 and 2 interneurons	Ndnf cluster (Ndnf Car4, Ndnf Cxcl14)	Int12, Int15
Fuzik	Cortex Pyramidal cells	Pyramidal cluster	Pyramidal cluster (excluding CA1PyrInt)

Supplementary Table 2: Matching of patch-seq cell types to dissociated cell reference atlases.

Broad cell type	Tasic subtypes	Zeisel subtypes
Astrocyte	Astro Gja1	Astro2, Astro1
Endothelial	Endo Myl9, Endo Tbc1d4	Vsmc
Inhibitory	Vip Chat, Vip Parm1, Vip Mybpc1, Vip Gpc3, Pvalb Gpx3, Ndnf Cxcl14, Vip Sncg, Ndnf Car4, Sst Myh8, Sst Th, Sst Chodl, Sst Tacstd2, Sst Cdk6, Pvalb Wt1, Sncg, Sst Cbln4, Pvalb Tacr3, Igtp, Smad3, Pvalb Tpbg, Pvalb Cpne5, Pvalb Rspo2, Pvalb Obox3	Int10, Int6, Int9, Int2, Int4, Int1, Int3, Int13, Int16, Int14, Int11, Int5, Int7, Int8, Int12, Int15
Microglia	Micro Ctss	Mgl1, Mgl2
Oligodendrocyte	Oligo Opalin, Oligo 96_Rik	Oligo1, Oligo3, Oligo4, Oligo2, Oligo6, Oligo5
OPC	OPC Pdgfra	*
Pyramidal	L2/3 Ptgs2, L2 Ngb, L4 Ctxn3, L4 Scnn1a, L5a Batf3, L5a Pde1c, L6a Mgp, L6b Serpinb11, L6b Rgs12, L5a Hsd11b1, L4 Arf5, L5a Tcerg1l, L6a Sla, L6a Syt17, L6a Car12, L5b Cdh13, L5 Ucma, L5b Tph2, L5 Chrna6	S1PyrL4, ClauPyr, S1PyrL5, S1PyrL23, S1PyrDL, S1PyrL5a, SubPyr, CA1Pyr1, S1PyrL6b, S1PyrL6, CA1Pyr2, CA2Pyr2

*Supplementary Table 3. Mapping of broad cell types between Tasic and Zeisel dissociated cell reference datasets. * denotes oligodendrocyte precursor cell type not being explicitly labelled in Zeisel.*

Supplementary Table 4: List of cell type-specific markers based on re-analysis of published dissociated cell-based scRNAseq experiments from mouse brain.