PLOS BIOLOGY

METHODS AND RESOURCES

# Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics

**Kenneth D. Harris[1,2]*, Hannah Hochgerner[3], Nathan G. Skene[1,3], Lorenza Magno[4], Linda Katona[5], Carolina Bengtsson Gonzales[3], Peter Somogyi[5], Nicoletta Kessaris[4], Sten Linnarsson[3], Jens Hjerling-Leffler[3]**

**1** University College London Institute of Neurology, London, United Kingdom, **2** University College London Department of Neuroscience, Physiology and Pharmacology, London, United Kingdom, **3** Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden, **4** University College London Wolfson Institute for Biomedical Research, London, United Kingdom, **5** Department of Pharmacology, University of Oxford, Oxford, United Kingdom

* kenneth.harris@ucl.ac.uk





**Citation:** Harris KD, Hochgerner H, Skene NG, Magno L, Katona L, Bengtsson Gonzales C, et al. (2018) Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. PLoS Biol 16(6): e2006387. https://doi.org/10.1371/journal.pbio.2006387

**Academic Editor:** Peter Jonas, Institute of Science and Technology Austria, Austria

**Received:** April 18, 2018

**Accepted:** May 22, 2018

**Published:** June 18, 2018



**Data Availability Statement:** Raw data are available on GEO under accession number GSE99888. Processed data and analysis results are available at https://figshare.com/articles/Transcriptomic_analysis_of_CA1_inhibitory_interneurons/6198656. For any additional questions, please contact kenneth.harris@ucl.ac.uk.

**Funding:** Wellcome Trust (grant number 108726). Received by KDH, PS, NK, SL, JH-L. The funder had no role in study design, data collection and

## Abstract

Understanding any brain circuit will require a categorization of its constituent neurons. In hippocampal area CA1, at least 23 classes of GABAergic neuron have been proposed to date. However, this list may be incomplete; additionally, it is unclear whether discrete classes are sufficient to describe the diversity of cortical inhibitory neurons or whether continuous modes of variability are also required. We studied the transcriptomes of 3,663 CA1 inhibitory cells, revealing 10 major GABAergic groups that divided into 49 fine-scale clusters. All previously described and several novel cell classes were identified, with three previously described classes unexpectedly found to be identical. A division into discrete classes, however, was not sufficient to describe the diversity of these cells, as continuous variation also occurred between and within classes. Latent factor analysis revealed that a single continuous variable could predict the expression levels of several genes, which correlated similarly with it across multiple cell types. Analysis of the genes correlating with this variable suggested it reflects a range from metabolically highly active faster-spiking cells that proximally target pyramidal cells to slower-spiking cells targeting distal dendrites or interneurons. These results elucidate the complexity of inhibitory neurons in one of the simplest cortical structures and show that characterizing these cells requires continuous modes of variation as well as discrete cell classes.

## Author summary

Single-cell RNA sequencing allows scientists to count the number of copies of each gene expressed in multiple individually isolated cells. Because different cell types express genes in different amounts, "clusters" of cells with similar expression patterns are likely to correspond to different cell types. As well as discrete classes, however, cells also show continuous variation in gene expression. To study the relationship between cell classes and

analysis, decision to publish, or preparation of the manuscript. Medical Research Council (grant number MC_UU_12024/4). Received by PS. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. European Research Council (grant number 261063). Received by SL. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Swedish Research Council (grant number STARGET). Received by SL. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Swedish Research Council (grant number 2014-3863). Received by J-HL. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Knut and Alice Wallenberg Foundation (grant number). Received by SL. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. European Union FP7/Marie Curie Actions (grant number 322304). Received by JH-L. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Chan-Zuckerberg Initiative (grant number 2018-182811). Received by KDH. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. StratNeuro (grant number). Received by JH-L. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Swedish Brain Foundation (Hjärnfonden) (grant number). Received by JH-L. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

**Abbreviations:** aCSF, artificial cerebrospinal fluid; AP, alkaline phosphatase; BIC, Bayesian Information Criterion; CA, *Cornu Ammonis*; CGE, caudal ganglionic eminence; EM, Expectation-Maximization; FACS, fluorescence-activated cell sorting; GABA, gamma-Aminobutyric acid; GO, gene ontology; I-S, interneuron-selective; nbtSNE, negative binomial tSNE; O-Bi, Oriens-Bistratified; OCT, optimal cutting temperature; O-LM, oriens/lacunosum-moleculare; ProMMT, Probabilistic Mixture Modeling for Transcriptomics; PVALB, parvalbumin; SBC, single-bouquet cell; scRNA-seq, single-cell RNA-sequencing; slm, stratum lacunosum-moleculare; so, stratum oriens; sp, stratum pyramidale; sr, stratum radiatum; tSNE, t-stochastic neighbor embedding; UMI, unique

continua in a well-understood brain system, we applied new analysis methods to a dataset of inhibitory interneurons from area CA1 of the mouse hippocampus. Thanks to decades of intensive work, at least 23 classes of CA1 interneurons have been previously defined. We were able to identify them all with our transcriptomic clusters but unexpectedly found three to be identical. Because the connectivity of these cells has already been established, we were also able to identify the primary mode of continuous variation in these cells, which related to their axon target location. This in-depth understanding of the relatively simple cortical circuit of CA1 not only clarifies the cellular composition of this important brain structure but also will form a solid foundation for understanding more complex structures, such as the isocortex.

## Introduction

Cortical circuits are composed of highly diverse neurons, and a clear definition of cortical cell types is essential for the explanation of their contribution to network activity patterns and behavior. Cortical neuronal diversity is strongest amongst GABAergic neurons. In hippocampal area CA1—one of the architecturally simplest cortical structures—GABAergic neurons have been divided so far into at least 23 classes of distinct connectivity, firing patterns, and molecular content [1–6]. A complete categorization of CA1 inhibitory neurons would provide not only essential information to understand the computational mechanisms of the hippocampus but also a canonical example to inform studies of more complex structures, such as six-layered isocortex.

CA1 GABAergic neurons have been divided into six major groups based on connectivity and expression patterns of currently used molecular markers. Parvalbumin (PVALB)-positive neurons (including basket, bistratified, and axo-axonic cells) target pyramidal cells' somata, proximal dendrites, or axon initial segments, firing fast spikes that lead to strong and rapid suppression of activity [7,8]. Somatostatin (SST)-positive oriens/lacunosum-moleculare (O-LM) cells target pyramidal cell distal dendrites and exhibit slower firing patterns [9]. GABAergic long-range projection cells send information to distal targets and comprise many subtypes, including SST-positive hippocamposeptal cells; NOS1-positive backprojection cells targeting dentate gyrus and CA3; and several classes of hippocamposubicular cells, including trilaminar, radiatum-retrohippocampal, and PENK-positive neurons [10–14]. Cholecystokinin (CCK)-positive interneurons are a diverse class characterized by asynchronous neurotransmitter release [15,16] that have been divided into at least five subtypes targeting different points along the somadendritic axis of pyramidal cells [17–21]. Neurogliaform and Ivy cells release GABA diffusely from dense local axons and can mediate volume transmission as well as conventional synapses [22,23]. Interneuron-selective (I-S) interneurons comprise at least three subtypes specifically targeting other inhibitory neurons and expressing one or both of Vasoactive intestinal polypeptite (VIP) and calretinin (CALB2) [2,24–26]. Finally, additional rare types, such as large SST/NOS1 cells [27], have been described at a molecular level, but their axonal targets and relationship to other subtypes is unclear.

This already complex picture likely underestimates the intricacy of CA1 inhibitory neurons. Currently defined classes likely divide into several further subtypes, and additional neuronal classes likely remain to be found (e.g., [28]). Furthermore, it is unclear whether a categorization into discrete classes is even sufficient to describe the diversity of cortical inhibitory neurons [29,30]. For example, several CCK interneuron classes have been described, targeting pyramidal cells at multiple locations ranging from their somata to distal dendrites, and the

molecular identifier; V1, primary visual cortex; YFP, yellow fluorescent protein.

molecular profile and spiking phenotype of these cells correlates with their synaptic target location, with fast-spiking cells more likely to target proximal segments of pyramidal neurons [18,19,21]. Do such cells represent discrete classes with sharp interclass boundaries, or do they represent points along a continuum? Finally, while a cell's large-scale axonal and dendritic structure likely remains fixed throughout life, both gene expression and electrophysiological properties can be modified by factors such as neuronal activity [31–34]. To what extent is the observed molecular diversity of interneurons consistent with activity-dependent modulation of gene expression?

Single-cell RNA sequencing (scRNA-seq)—which can read out the expression levels of all genes in large numbers of individual cells—provides a powerful opportunity to address these questions. This method has successfully identified the major cell classes in several brain regions [35–46]. Nevertheless, identifying fine cortical cell classes has not been straightforward because of both incomplete prior information on the underlying cell types and complicating factors, such as potential continuous variability within these classes. The large body of prior work on CA1 interneurons provides a valuable opportunity to identify transcriptomic clusters with known cell types in an important cortical circuit, enabling confident identification of known and novel classes and investigation of questions such as continuous variability.

Here, we describe a transcriptomic analysis of 3,663 inhibitory neurons from mouse CA1. This analysis revealed 49 clusters, of which we could identify 41 with previously described cell types, with the remaining 8 representing putative novel cell types. All previously described CA1 GABAergic classes could be identified in our database, but our results unexpectedly suggest that three of them are identical. The larger number of clusters occurring in our transcriptomic analysis reflected several previously unappreciated subtypes of existing classes and tiling of continua by multiple clusters. Our data suggest a common genetic continuum exists between and within classes, from faster-firing cells targeting principal cell somata and proximal dendrites, to slower-firing cells targeting distal dendrites or interneurons. Several classes previously described as discrete represent ranges along this continuum of gene expression.

## Results

### Data collection and identification of inhibitory cells

We collected cells from six *Slc32a1-Cre;R26R-tdTomato* mice, three of age p60 and three of age p27. Cells were procured using enzymatic digestion and manual dissociation [46], and data were analyzed using the 10X Genomics "cellranger" pipeline, which uses unique molecular identifiers (UMIs) to produce an absolute integer quantification of each gene in each cell. The great majority of cells (4,572/6,971 cells total; 3,283/3,663 high-quality interneurons) came from the older animals. Because we observed no major difference in interneuron classes between ages, data were pooled between them (S1 Fig). Fluorescence-activated cell sorting (FACS) yielded an enriched but not completely pure population of GABAergic neurons. A first-round clustering (using the method described below) was therefore run on the 5,940 cells passing quality control, identifying 3,663 GABAergic neurons (as judged by the expression of genes *Gad1* and *Slc32a1*).

### Cluster analysis

We analyzed the data using a suite of four novel algorithms derived from a probabilistic model of RNA distributions. All four methods were based on the observation that RNA counts within a homogeneous population can be approximated using a negative binomial distribution (see Methods, [47,48]). The negative binomial distribution accurately models the high variance of transcriptomic read counts (S2A and S2B Fig). As a consequence, algorithms based on this

distribution weight the presence or absence of a gene more than its numerical expression level—for example, this distribution treats read counts of 0 and 10 as more dissimilar than read counts of 500 and 1,000 (S2C Fig).

The algorithm we used for clustering was termed ProMMT (Probabilistic Mixture Modeling for Transcriptomics). This algorithm fits gene expression in each cluster $k$ by a multivariate negative binomial distribution with cluster-specific mean $\boldsymbol{\mu}_k$. The mean expression levels of only a small subset of genes are allowed to vary between clusters (150 for the current analysis; S3 Fig); these genes are selected automatically by the algorithm by maximum likelihood methods. The use of such "sparse" methods is essential for probabilistic classification of high-dimensional data [49], and the genes selected represent those most informative for cluster differentiation. The number of clusters was chosen automatically using the Bayesian Information Criterion (BIC) [50]. The ProMMT algorithm also provides a natural measure of the distinctness of each cluster, which we term the isolation metric (see Methods).

The ProMMT algorithm divided CA1 interneurons into 49 clusters (Fig 1). We named the clusters using a multilevel scheme after genes that are strongly expressed at different hierarchical levels; for example, the cluster *Calb2.Vip.Nos1* belongs to a first-level group characterized by strong expression of *Calb2* (indicating I-S interneurons); a second-level group *Calb2.Vip*; and a third-level group distinguished from other *Calb2.Vip* cells by stronger expression of *Nos1*. This naming scheme was based on the results of hierarchical cluster analysis of cluster means, using a distance metric based on the negative binomial model (Methods; Fig 1).

## Data visualization

To visualize cell classes in two dimensions, we modified the t-stochastic neighbor embedding (tSNE) algorithm [51] for data with negative binomial variability, terming this approach nbtSNE (negative binomial tSNE). In conventional tSNE, the similarity between data points is defined by their Euclidean distance, which corresponds to conditional probabilities under a Gaussian distribution. We obtained greater separation of clusters and a closer correspondence to known cell types by replacing the Gaussian distribution with the same negative binomial distribution used in our clustering algorithm (see Methods; S4 Fig).

The nbtSNE maps revealed that cells were arranged in 10 major "continents" (Fig 2). The way expression of a single gene differed between classes could be conveniently visualized on these maps by adjusting the symbol size for each cell according to that gene's expression level. Consistent with previous transcriptomic analyses, we found that classes were rarely, if ever, identified by single genes but rather by combinatorial expression patterns. Thanks to the extensive literature on CA1 interneurons, 25 genes together sufficed to identify the main continents with known cell classes (Fig 3), and it was also possible to identify nearly all the finer subclasses using additional genes specific to each class (S1 Text).

## Identification of cell types

Previous work has extensively characterized the connectivity, physiology, and firing patterns of CA1 inhibitory neurons, and these cellular properties have been related to the expression of large numbers of marker genes. We next sought to identify our transcriptomic clusters with previously defined cell types, taking advantage of the "Rosetta stone" provided by this extensive prior research. Explaining how the identifications were made requires an extensive discussion of the previous literature, which is presented in full as S1 Text, online. Here, we briefly summarize the major subtypes identified (summarized in Fig 4).

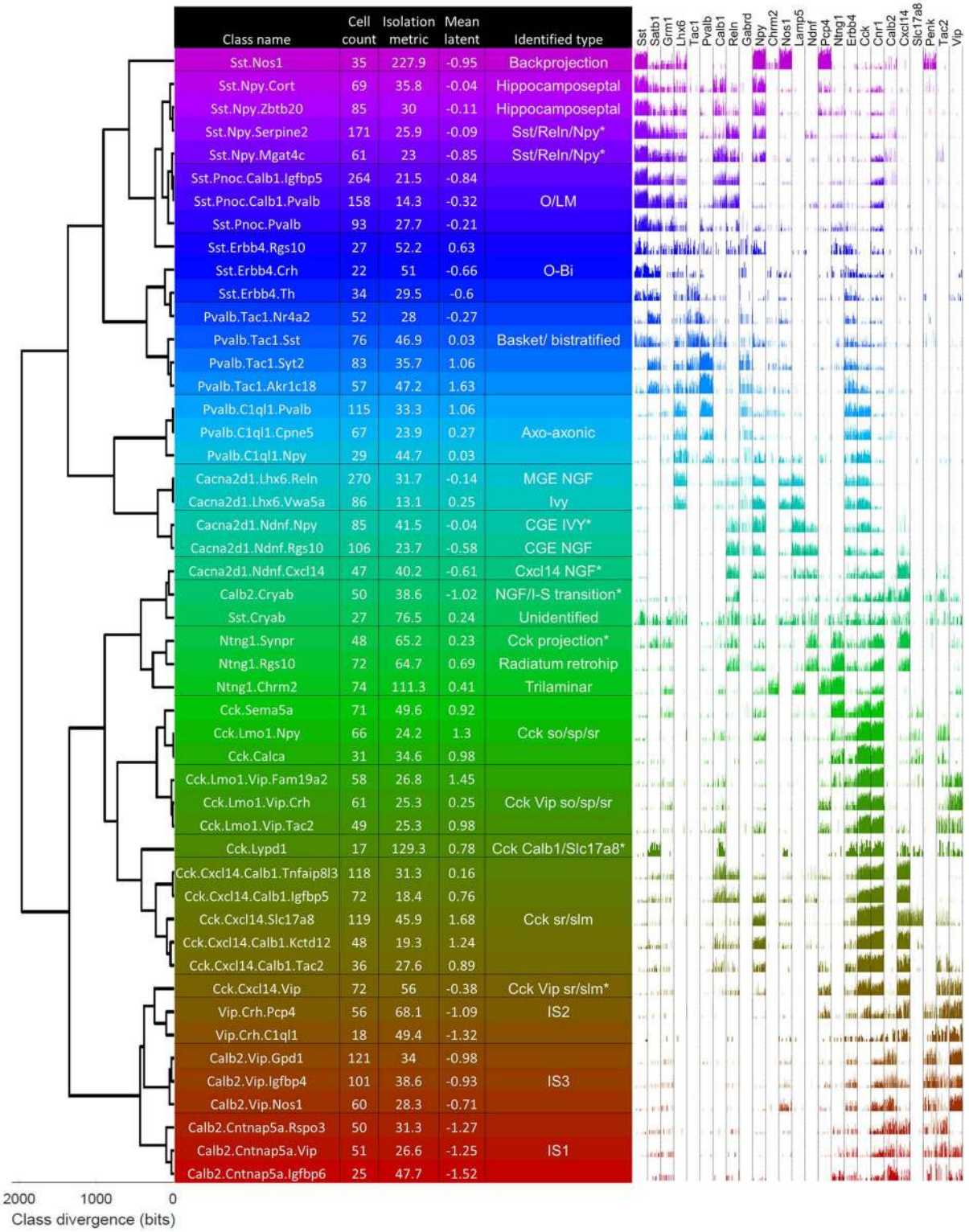


**Fig 1. The ProMMT algorithm split CA1 GABAergic neurons into 49 clusters.** Dendrogram (left) shows a hierarchical cluster analysis of these classes. Table shows class names (chosen hierarchically according to strongly expressed genes), number of cells per class, isolation metric of each class (higher for distinct classes), the mean value of latent variable analysis for cells in this class, and the biological cell type identified from its gene expression pattern. Asterisks indicate hypothesized novel classes. Right, bar chart showing log expression of 25 selected genes for all cells in the class. Note the expression pattern of *Lhx6*, which suggests a developmental origin in medial ganglionic eminence for clusters

*Cacna2d1.Lhx6.Vwa5a* and above. CGE, caudal ganglionic eminence; I-S/IS, interneuron-selective; MGE, medial ganglionic eminence; NGF, neurogliaform cell; O-Bi, Oriens-Bistratified; O-LM, oriens/lacunosum-moleculare; ProMMT, Probabilistic Mixture Modeling for Transcriptomics; slm, stratum lacunosum-moleculare; so, stratum oriens; sp, stratum pyramidale; sr, stratum radiatum.

https://doi.org/10.1371/journal.pbio.2006387.g001

Continent 1 was identified with the *Sst*-positive hippocamposeptal and O-LM cells of stratum oriens (so). These cells all expressed *Sst* and *Grm1*, and were further divided into two *Npy+/Ngf+* clusters identified as hippocamposeptal neurons [52] and three *Pnoc+/Reln+/Npy–* clusters identified with O-LM cells [9]. In addition, continent 1 contains a previously undescribed subclass positive for *Sst*, *Npy*, and *Reln*.

Continent 2 was identified as basket and bistratified cells. These were all positive for *Tac1* (the precursor to the neuropeptide Substance P), as well as *Satb1* and *Erbb4*, but were negative for *Grm1*. They were divided into two *Pvalb+/Sst–* clusters identified with basket cells, two *Pvalb+/Sst+/Npy+* clusters identified with bistratified cells (Klausberger et al., 2004), and three *Pvalb–* clusters identified with Oriens-Bistratified (O-Bi) cells [53].

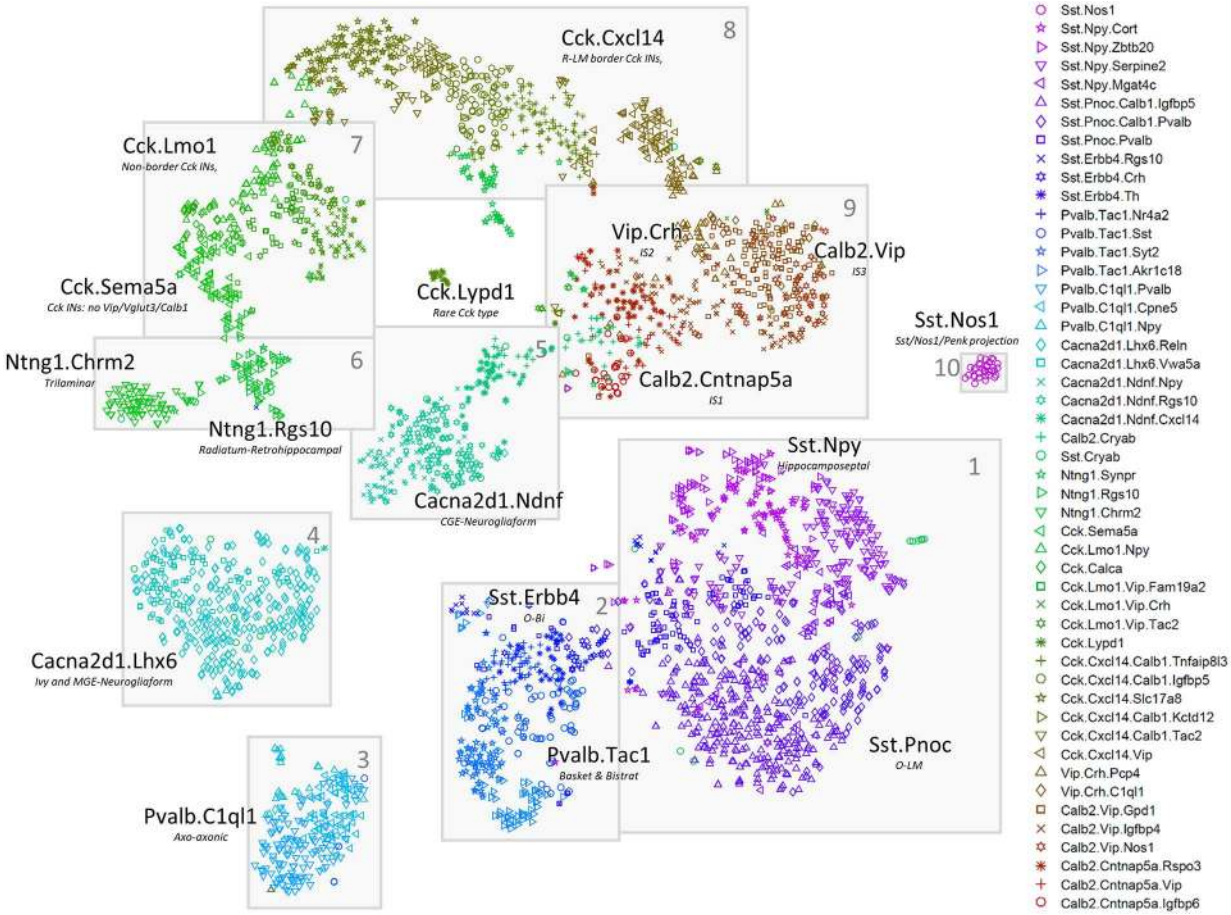


**Fig 2. Two-dimensional visualization of expression patterns using nbtSNE algorithm, which places cells of similar expression close together.** Each symbol represents a cell, with different color/glyph combinations representing different cell classes (legend, right). Grey boxes and numbers refer to the "continents" mentioned in the text and subsequent figures. nbtSNE, negative binomial t-stochastic neighbor embedding; O-LM, oriens/lacunosum-moleculare.

https://doi.org/10.1371/journal.pbio.2006387.g002

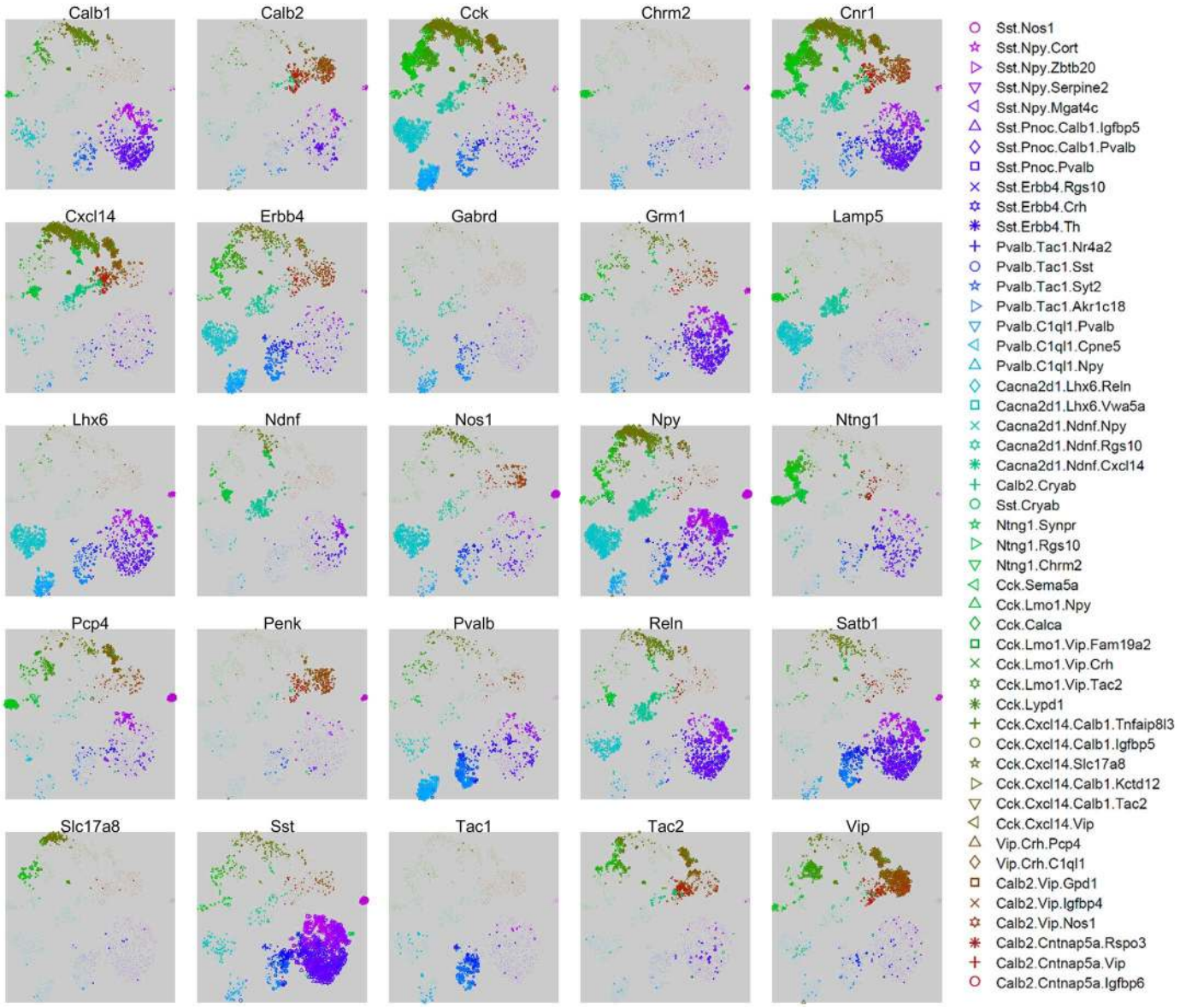


**Fig 3. Expression levels of 25 selected genes that together allow identification of major cell classes.** Each subplot shows an nbtSNE map of all cells, with marker size indicating log-expression level of the gene named above the plot. Similar maps for all genes can be found online at http://linnarssonlab.org/ca1/.

https://doi.org/10.1371/journal.pbio.2006387.g003

Continent 3 was identified as axo-axonic cells because of their expression of *Pvalb* but not *Satb1* [54]. This continent's three clusters were *Tac1* negative but positive for other markers, including *Snca*, *Pthlh*, and *C1ql1*, which have also been associated with axo-axonic cells in isocortex [43,44]. We note that this dichotomy of *Pvalb* interneurons into *Tac1*-positive and -negative subclasses is likely homologous to previous observations in isocortex [55].

Continent 4 was identified as Ivy cells and medial ganglionic eminence (MGE)-derived neurogliaform cells. These cells expressed *Cacna2d1*, which we propose as a unique identifier of hippocampal neurogliaform/ivy cells, as well as *Lhx6* and *Nos1* [56]. They were divided into a *Reln*+ cluster identified with MGE-derived neurogliaform cells and a *Reln*−/*Vwa5a*+ cluster identified with Ivy cells [23]. This continent is homologous to the isocortical *Igtp* class defined

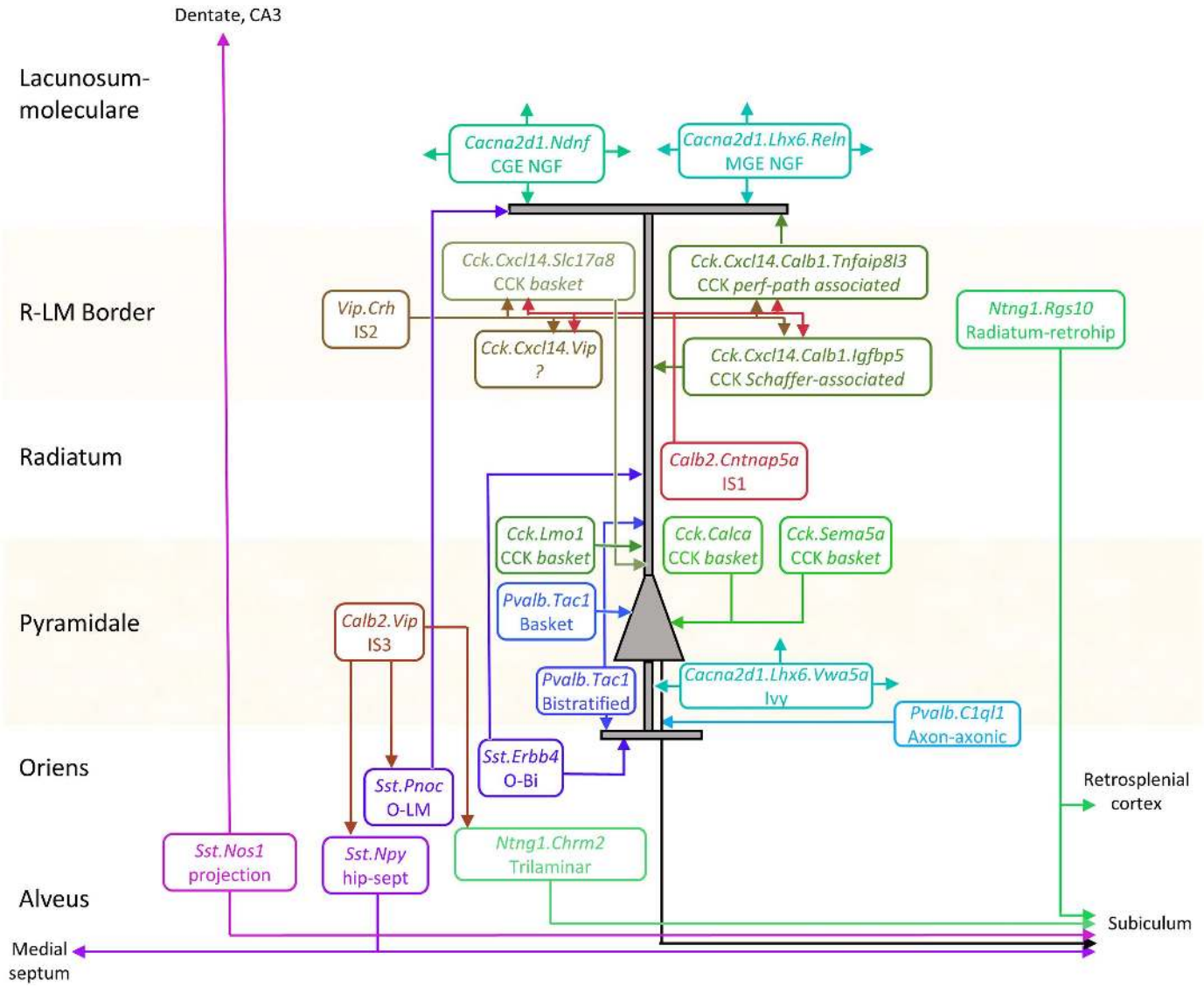


**Fig 4. Inferred circuit diagram of identified GABAergic cell types.** The identification of transcriptomic clusters with known cell classes is described in full in S1 Text. Laminar locations and connections between each class are derived from previous literature. CGE, caudal ganglionic eminence; IS, interneuron-selective MGE, medial ganglionic eminence; NGF, neurogliaform cell; O-Bi, Oriens-Bistratified; O-LM, oriens/lacunosum-moleculare; R-LM, Radiatum/lacunosum-moleculare.

https://doi.org/10.1371/journal.pbio.2006387.g004

by Tasic and colleagues [44], which we hypothesize may represent isocortical neurogliaform cells of MGE origin; this hypothesis could be confirmed using fate mapping.

Continent 5 was identified as caudal ganglionic eminence (CGE)-derived neurogliaform cells. Its three clusters contained *Cacna2d1* and many other genes in common with those of continent 4, but lacked *Lhx6* and *Nos1* [56]. Similar to isocortical putative neurogliaform cells, this continent expressed *Ndnf* and contained a distinct subtype positive for *Cxcl14* [44]. As with continent 4, continent 5 mainly expressed *Reln* but also contained a small *Reln*-negative cluster, which we suggest forms a rare and novel class of CGE-derived ivy cell.

Continent 6 was identified with *Sst*-negative long-range projection interneurons. It divided into two distinct clusters, both of which were strongly positive for *Ntng1*. The first strongly

expressed *Chrm2* but lacked *Sst* and *Pvalb*, identifying them as trilaminar cells [12,57]. The second subgroup lacked most classical molecular markers; this fact, together with their inferred laminar location at the stratum radiatum / stratum lacunosum-moleculare (sr/slm) border, identified them as putative radiatum-retrohippocampal neurons that project to the retrosplenial cortex [12,58].

Continents 7 and 8 were identified as what are traditionally called *Cck* interneurons. This term is somewhat unfortunate: while these cells indeed strongly express *Cck*, many other inhibitory classes express *Cck* at lower levels, including even *Pvalb*+ basket cells [59]. Continents 7 and 8 cells comprised 13 highly diverse clusters but shared strong expression of *Cnr1*, *Sncg*, *Trp53i11*, and several other novel genes. Continent 8 is distinguished by expression of *Cxcl14*, which localizes these cells to the sr/slm border. This continent comprised a continuum ranging from soma-targeting basket cells, identified by their *Slc17a8* (vGlut3) expression, to dendrite-targeting cells, identified by expression of *Calb1* or *Reln* [19,21]. Continent 7, lacking *Cxcl14*, was identified as *Cck* cells of other layers and contained multiple subtypes characterized by the familiar markers *Calb1*, *Vip*, and *Slc17a8* [21] as well novel markers such as *Sema5a* and *Calca*. Associated with continent 8 were several apparently novel subtypes: a rare and distinct group positive for both *Scl17a8* and *Calb1* and marked nearly exclusively by *Lypd1*; a *Ntng1*+/*Ndnf*+ subgroup related to cells of continent 6; and a group strongly expressing both *Vip* and *Cxcl14*, which therefore likely corresponds to a novel *Vip*+/*Cck*+ interneuron at the sr/slm border.

Continent 9 was identified as I-S interneurons. Its eight clusters fell into three groups: *Calb2*+/*Vip*− neurons identified as IS-1 cells, *Calb2*−/*Vip*+ neurons identified as IS-2 cells, and *Calb2*+/*Vip*+ neurons identified as IS-3 cells [2,24,26,60]. All expressed *Penk* [61]. These cells contained at least two novel subgroups: an IS-3 subtype positive for *Nos1* and *Myl1*, homologous to the *Vip Mybpc2* class defined in isocortex [44], and a rare subclass of IS-1 cells positive for *Igfbp6*.

Continent 10 contained a single highly distinct cluster located in an "island" off continent 1. It contained cells strongly positive for *Sst* and *Nos1* [27], whose expression pattern is consistent with that of both backprojection cells [13] and PENK-positive projection cells [10], suggesting that these three previously identified classes reflect a single cell type.

## Comparison with isocortical classes

Our finding of 49 clusters in a sample of 3,663 CA1 cells contrasts with a previous study of isocortical area V1 (primary visual cortex), which found 23 clusters from a sample of 761 inhibitory neurons [44]. One can imagine three reasons for the greater number of clusters found in the present study: the larger sample size used here may have resulted in our resolving more clusters, the use of a different clustering algorithm may have allowed the current study to reveal finer cell types, or area CA1 might genuinely contain more diverse inhibitory neurons than isocortex. To address these questions, we performed two analyses. First, we applied our clustering algorithm to the data of Tasic and colleagues (2016), and second we reanalyzed subsamples of the data of both the current study and of Tasic and colleagues (2016) to see how the number of clusters found varies with cell count and with sequencing depth.

Applying the ProMMT algorithm to the Tasic dataset yielded 30 clusters (S5A and S5B Fig). The cluster assignments almost completely overlapped as far as top-level groupings but showed some more subtle distinctions in finer-level clusters (S5C Fig). We examined three of these novel classes' differences in more depth, to ask whether the finer distinctions found by the ProMMT algorithm could correspond to genuine biological cell classes. The most

notable of these was cluster 11, which contained neurons that had previously been assigned to the neurogliaform clusters Ndnf Cxcl14, Ndnf Car4, but lacked common neurogliaform markers such as *Lamp5* and *Gabrd*. Instead, cells in these clusters expressed *Calb2* and *Penk* but not *Vip*, suggesting cells homologous to hippocampal IS-1 cells and potentially matching the *Vip*-negative I-S layer 1 "single-bouquet cells" (SBCs) described by Jiang and colleagues [62,63]. To test whether cluster 11 indeed corresponds to SBCs, we took advantage of a Patch-seq study [64] that contrasted gene expression in anatomically identified layer 1 SBCs and neurogliaform cells (S5D Fig). We found that the genes that Cadwell and colleagues had reported as distinguishing SBCs from neurogliaform cells indeed occurred in almost entirely nonoverlapping populations of cells; furthermore, these populations closely matched the ProMMT clusters identified with SBCs and neurogliaform cells. Examination of two further subdivisions found by the ProMMT algorithm again revealed genes uniquely expressed in nonoverlapping subpopulations of the Sst Cbln4 and Vip Parm1 clusters (S6 Fig). We conclude that the larger number of clusters identified by the ProMMT algorithm at least in part results from its ability to distinguish subtle variations in gene expression between related cell types.

To ask whether the greater number of clusters found in the current study might in part arise from its larger sample size, we reran the cluster analysis on randomly selected subsets of cells from our dataset. We found a strong linear increase in the number of clusters found with increasing sample size (S7A Fig). To investigate what effect sequencing depth may have had, we resampled our dataset to simulate lower read counts for the same cells, and again found an approximately linear increase in the number of identified clusters with read count (S7B and S7C Fig). We performed similar analyses on Tasic and colleagues' data and obtained similar results (S7D and S7E Fig).

We therefore conclude that the larger number of clusters found by the current study is more likely to reflect a combination of larger sample size and more sensitive clustering algorithms than a greater number of biological cell types in CA1 than in V1. Furthermore, we expect that an even larger sample size or greater sequencing depth would have revealed yet more, finely distinguished cell types.

## Continuous variation between and within cell classes

Although the major continents of the expression map were clearly separated, clusters within these continents often appeared to blend into each other continuously. This suggests continuous gradation in gene expression patterns: while our probabilistic mixture model will group cells from a common negative binomial distribution into a single cluster, it will tile cells of continuously graded mean expression into multiple clusters.

Although visualization methods such as nbtSNE can suggest whether classes are discrete or continuously separated, they are not sufficient to confirm the suggestion. Such methods exhibit local optima, raising the possibility that apparent continuity only occurs for particular initialization conditions. Furthermore, as nbtSNE is based on a subset of genes, it is conceivable that discrete/continuous patterns occur only for this subset.

To confirm the apparent continuity or discreteness of these groups, we therefore employed a novel method of negative binomial discriminant analysis that is independent of nbtSNE and considers all genes. Given a pair of cell classes, this method compares how close each cell's whole-genome expression pattern is to each class, using a cross-validated likelihood ratio statistic. For two classes identified as basket and axo-axonic cells, the histogram of likelihood ratios was clearly bimodal (Fig 5A, top), indicating that every cell exhibited a much stronger fit

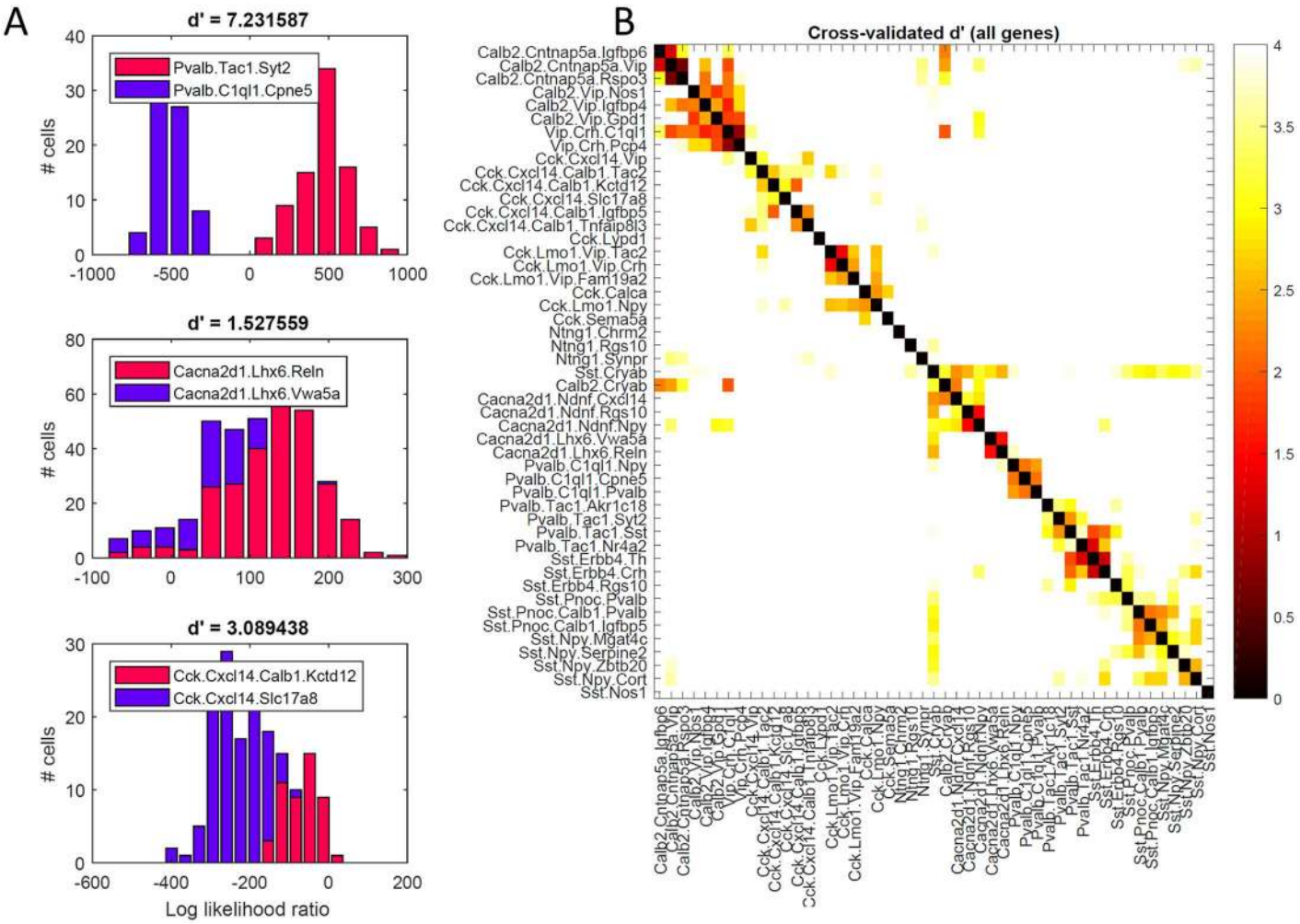


**Fig 5. Analysis of discrete versus continuous variation by negative binomial discriminant analysis.** (**A**) Histogram of log-likelihood ratios for three example cluster pairs, measuring how much better each cell's whole-genome expression pattern is explained by one or the other cluster. The top histogram (basket versus axo-axonic cells) is clearly bimodal, indicating discrete separation. The bottom two histograms (ivy versus MGE-neurogliaform cells; two subclasses of Cck/Cxcl14 cells) show substantial overlap, indicating continuous variation between clusters. The degree of bimodality is captured by the d' statistic above each plot. (**B**) Pseudocolor matrix showing continuity of each pair of clusters, as assessed by d' statistic. White means strongly bimodal; darker colors indicate continuity.

https://doi.org/10.1371/journal.pbio.2006387.g005

to its own class than to the other, and confirming the discrete separation of these classes. A second example of clusters identified with Ivy and MGE-neurogliaform cells, however, showed different behavior (Fig 5A, middle): a unimodal likelihood ratio histogram indicated that the two clusters ran smoothly into each other, tiling a continuum of gene expression patterns. The bimodality of the likelihood ratio can be captured by a d' statistic, which for these two examples was 7.2 and 1.5, respectively. Perhaps ironically, the degree to which two neighboring classes are discrete or continuous was itself a continuous variable. For example, *Slc17a8*-expressing *Cxcl14/Cck* neurons showed largely continuous overlap with their neighboring *Cck/Cxcl14* cells, but with some small indication of bimodality, characterized by a d' of 3.1 (Fig 5A, bottom). We conclude that while truly discrete cluster separations do exist, the dataset is not fully described as a set of discrete classes, and that many clusters tile continuous dimensions (Fig 5B).

## Latent factor analysis reveals a common mode of variation across all cell types that correlates with axon target location

The existence of continuous variation in gene expression suggests that cluster analysis is not giving a complete picture of neuronal gene expression patterns. To further study the biological significance of continuously varying gene expression, we therefore applied a complementary method, latent factor analysis. Cluster analysis can be viewed as an attempt to summarize the expression of all genes using only a single discrete label per cell (the cell's cluster identity); the value this label takes for each cell is not directly observed but is "latent" and inferred from the data. Latent factor analysis also attempts to predict the expression of all genes using only a single variable (the "latent factor"), but now with a continuous rather than discrete distribution. As with cluster analysis, the latent factor is not directly observed but is inferred for each cell. Latent factor analysis operates without knowledge of cluster identity and therefore requires that the same rules be used to predict gene expression from the latent factor for cells of all types. Clearly, one should expect neither method to precisely predict the expression of all genes from a single variable, but the rules of cellular organization they reveal may provide important biological information.

As expected, latent factor analysis produced a complementary view to cluster analysis (Fig 6A). Knowing a cell's cluster identity did not suffice to predict its latent factor value, and vice versa. For example, the ranges of latent factor values for cells in the clusters identified with *Cck* and *Pvalb* basket cells overlapped. Nevertheless, the range of possible latent factor values was not identical between clusters, and the mean latent factor value of each cluster differed in a manner that had a clear biological interpretation.

The mean latent factor value of each cluster correlated with the axon target location of the corresponding cell type (Fig 6A). The clusters showing largest mean latent factor values were identified with soma-targeting basket cells (both *Pvalb* and *Cck* expressing) and with axo-axonic cells. Lower values of the latent factor were found in clusters identified with dendrite-targeting *Cck* cells and with bistratified, Ivy, and hippocamposeptal cells, which target pyramidal cells' proximal dendrites [14,23]. Still lower values of the latent factor were found in clusters identified with neurogliaform and O-LM cells, which target pyramidal distal dendrites. The lowest values of all were found in clusters identified with cells synapsing on inhibitory interneurons: the IS cells of continent 9 and the *Sst/Penk/Nos1* cells of continent 10, whose local targets are *Pvalb* cells [10].

While mean values of the latent factor differed between continents, there was also substantial variability within cells of a single continent. For example, a gradient of latent factor values was seen within continent 8 (identified with *Cck*-positive neurons at the sr/slm border), with larger values in the west smoothly transitioning to smaller values in the east (Fig 6B). Comparison of gene expression patterns in continent 8 to previous work again suggested that this gradient in latent factor values correlates with axon target location. Indeed, immunohistochemistry has demonstrated that CCK-positive cells expressing SLC17A8 (expressed in western continent 8) project to the pyramidal layer [21], while those expressing CALB1 (expressed in the east) target pyramidal cell dendrites [3,18,65]. The cannabinoid receptor *Cnr1*, which is more strongly expressed in soma-targeting neurons [66,67], was also more strongly expressed in western cells with larger latent factor values.

As expected, the expression levels of many individual genes correlated with the latent factor; furthermore, the directions of these correlations were consistent, even within distantly related cell types. We investigated the relationships of genes to the latent factor by focusing initially on the *Pvalb*- and *Cck*-expressing cells of continents 2 and 8 (Fig 6C). Most genes correlated similarly with the latent factor in both classes. For example, the $Na^+/K^+$ pump *Atp1b1* and the

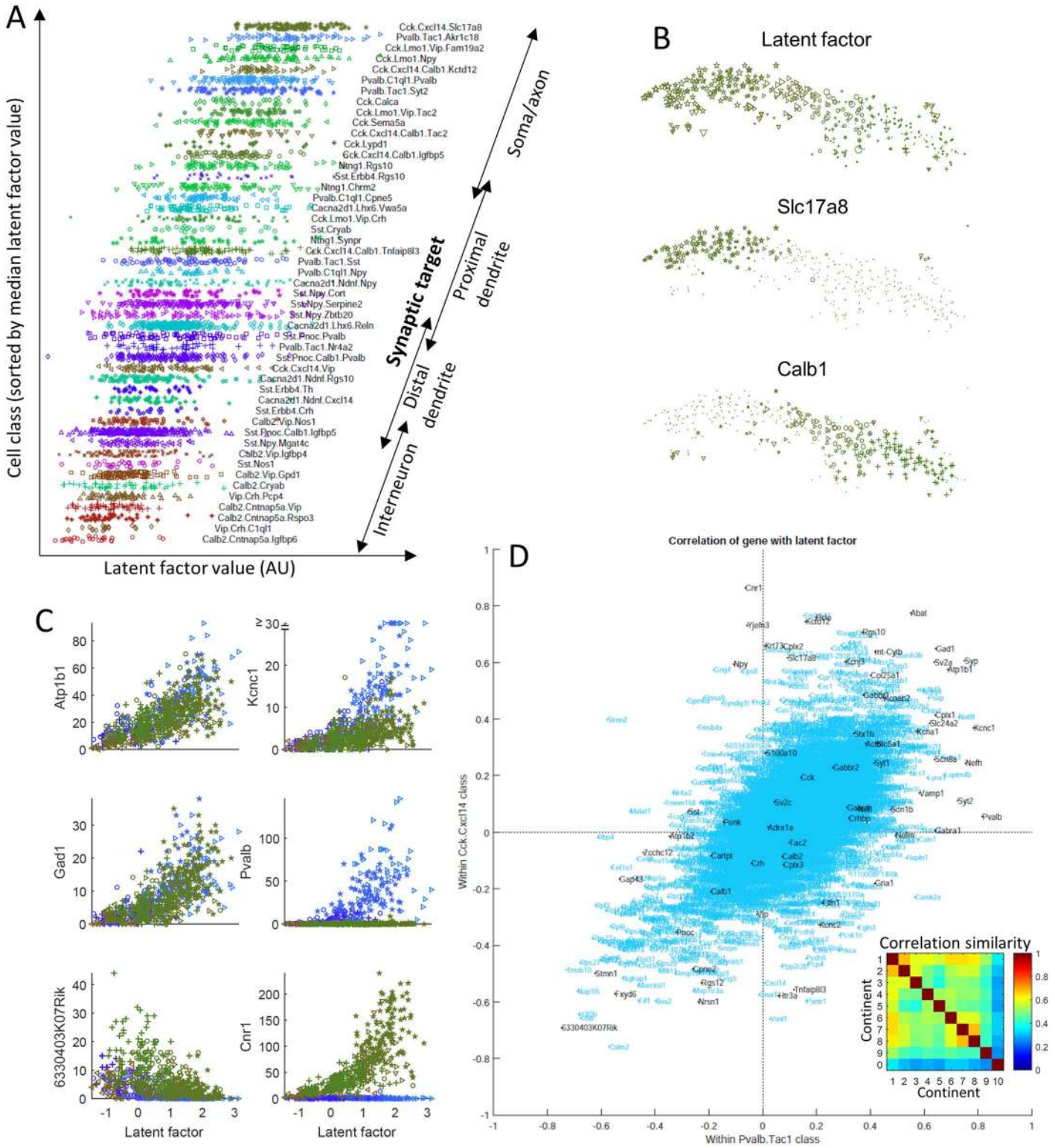


**Fig 6. Latent factor analysis reveals a common mode of continuous variation that is consistent across cell types and correlates with axon target location.** **(A)** Latent factor analysis assigns a single number to each cell via a search for the factor values that best predict expression of multiple genes. Mean factor values within each cluster differ systematically in a way that correlates with the identified cell class's axon target location. Each point represents a cell, with the x-coordinate showing latent factor value and y-coordinate showing cluster, sorted by mean latent factor value. **(B)** Continuous gradient of latent factor values across continent 8 (top; symbol size denotes latent factor value). Largest values are found in western *Slc17a8*-expressing neurons identified with soma-targeting *Cck* basket cells; smallest values are found in eastern *Calb1*-expressing neurons identified with dendrite-targeting *Cck* cells. **(C)** Correlation of latent factor values with expression of six example genes.

Symbols as in (A); blue, continent 2 (basket/bistratified), green, continent 8 (*Cck/Cxcl14*). **(D)** Correlations of genes with the latent factor are preserved across cell classes. X-axis: correlation of gene with latent factor in continent 2; y-axis, correlation in continent 8; Spearman's $\rho = 0.58$, $p < 10^{-100}$. Inset, Spearman $\rho$ values for all pairs of continents, $p < 10^{-100}$ in each case. AU, arbitrary units.

https://doi.org/10.1371/journal.pbio.2006387.g006

GABA synthesis enzyme *Gad1* correlated positively with the latent factor for multiple cell types, while *6330403K07Rik*, a gene of unknown function, correlated negatively. Some genes' expression levels depended on both cell type and latent factor value. For example, the ion channel *Kcnc1* (which enables rapid action potential repolarization in fast-spiking cells) correlated positively with the latent factor in both *Pvalb* and *Cck* cells, but its expression was stronger in *Pvalb* cells, even for the same latent factor value. Other genes showed correlations with the latent factor, but only within the specific classes that expressed them. For example, expression of *Pvalb* correlated with the latent factor within cells of continent 2, but the gene was essentially absent from cells of continent 8; conversely, *Cnr1* expression correlated with the latent factor in continent 8 but was essentially absent in cells of continent 2. Thus, the latent factor value is not alone sufficient to predict a cell's gene expression pattern but provides a summary of continuous gradation in the expression of multiple genes in multiple cell types.

The relationship of genes to latent factor values was statistically similar across cell types. To demonstrate this, we computed the Spearman correlation of each gene's expression level with the latent factor, separately, within cells of each continent (S1 Table). As expected from the scatterplots (Fig 6C), the correlation coefficients for *Atp1b1*, *Gad1*, and *6330403K07Rik* were similar between continents 2 and 8 (Fig 6D). Also as expected, *Pvalb* and *Cnr1* showed strong positive correlations with the latent factor within the continent where these genes were expressed, but correlations close to zero within the continent where they were barely expressed. In general, the correlation coefficients of genes with the latent factor were preserved between continents 2 and 8 (Fig 6D; Spearman rank correlation $\rho = 0.58$, $p < 10^{-100}$). A similar relationship was found across all continents (Fig 6D, inset; $p < 10^{-100}$ in each case), although cells of continents 9 and 10 showed less similarity than continents 1–8. Furthermore, similar results were obtained when analyzing isocortical data, most notably in isocortical *Pvalb* cells (S8 Fig).

In summary, the expression of many genes correlates with a single continuous variable, the latent factor value assigned to each cell. While this latent factor does not provide a complete summary of a cell's gene expression pattern, the direction and strength of the correlation of individual genes to the latent factor is largely preserved across cell types. Furthermore, while a cell's latent factor value was not simply a function of its cell class, mean latent factor values differed between clusters, being largest for clusters identified with cell types whose axons target pyramidal somata or axon initial segments and smallest for clusters identified with cell types targeting pyramidal distal dendrites or interneurons.

## Biological significance of genes correlating with the latent factor

The above results suggest that the expression of a large set of genes is modulated in a largely consistent way across multiple cell types in a manner that correlate with their axonal targets. What biological functions might these genes serve? While one might certainly expect structural genes be differentially expressed between soma- and dendrite-targeting interneurons, these cells also differ in their physiology. Indeed, *Pvalb*-expressing basket cells are known for their fast-spiking phenotype, which produces rapid, powerful perisomatic inhibition and is mediated by a set of rapidly acting ion channels and synaptic proteins, including *Kcnc1*, *Kcna1*, *Scn1a*, *Scn8a*, and *Syt2* [8]. Although most other interneurons show regular-spiking

phenotype, CCK-expressing basket cells with a fast-spiking phenotype have also been reported [18,20]. We therefore hypothesized that genes responsible for the fast-spiking phenotype might be positively correlated with the latent factor, because of increased expression in soma-targeting cells of all classes.

Consistent with this hypothesis, genes associated with fast-spiking phenotype (*Kcnc1*, *Kcna1*, *Scn1a*, *Scn8a*, *Syt2*) were amongst the genes most positively correlated with the latent factor in both *Pvalb* and *Cck* basket cells (Fig 6D). However, this positive correlation was not restricted to these cell types: in an ordering of the correlations of all genes with the latent factor (taking into account cells of all types), these genes ranked in the 99.9th, 98.3rd, 99.5th, 98.9th, and 95th percentiles, respectively (S1 Table).

Other gene families positively correlated with the latent factor included genes associated with mitochondria (e.g., *mt-Cytb*), ion exchange and metabolism (e.g., *Atp1b1; Slc24a2*), GABA synthesis and transport (e.g., *Gad1*, *Slc6a1*), vesicular release (e.g., *Syp*, *Sv2a*, *Cplx2*, *Vamp1*), and fast ionotropic glutamate and GABA receptors (e.g., *Gria1*, *Gabra1*) as well as $GABA_B$ receptors (e.g., *Gabbr1*, *Gabbr2*, *Kcnj3*, *Kctd12*). The genes correlating negatively with the latent factor were less familiar but included *Atp1b2*, a second isoform of the $Na^+/K^+$ pump; *Fxyd6*, which modulates its activity; *Nrsn1*, whose translation is suppressed after learning [68], as well as many neuropeptides (e.g., *Sst*, *Vip*, *Cartpt*, *Tac2*, *Penk*, *Crh*; exceptional neuropeptides such as *Cck* showed positive correlation). Genes associated with neurofilaments and intermediate filaments (e.g., *Nefh*, *Nefl*, *Krt73*) tended to show positive weights, while genes associated with actin processing (e.g., *Gap43*, *Stmn1*, *Tmsb10*) tended to show negative weights. Many other genes of as yet unknown function correlated positively and negatively with the latent factor (for example, *6330403K07Rik*). Relating the latent factor correlations of each gene to their gene ontology (GO) annotations (which are not granular enough to list annotations such as fast-spiking physiology) suggested that negatively correlated genes tended to be associated with translation and ribosomes, while positively correlated genes were associated with diverse functions, including transcription, signal transduction, ion transport, and vesicular function and with cellular compartments, including mitochondria, axons, and dendrites (S2 Table).

We therefore suggest that cells with large values of the latent factor not only target more proximal components of pyramidal cells but also express genes enabling a faster spiking firing pattern, more synaptic vesicles, and larger amounts of GABA release; receipt of stronger excitatory and inhibitory inputs; and faster metabolism. These are all characteristics of *Pvalb*-expressing fast-spiking interneurons [8]; however, a similar continuum was observed within all cell types, suggesting that these genes are commonly regulated in all CA1 interneurons.

The fact that the latent factor differs systematically between cells with different axonal targets suggests that this property is in good measure fixed, as it seems unlikely that neurons would make major changes to their axonal targets in adulthood. Nevertheless, interneuronal gene expression can be modulated by activity, and some of the genes that were most strongly correlated with the latent factor (*Pvalb*, *Kcna1*) are amongst those with activity-dependent modulation [31–34,69].

To investigate whether the genes correlated with the latent factor might also be partially modulated by neuronal activity, we correlated each gene's latent factor score with that gene's modulation by in vivo light exposure after dark housing, using data from three classes of visual cortical interneurons (made available by Mardinly and colleagues [33]). We observed a moderate relationship of latent factor weighting to activity modulation in *Sst* neurons ($r = 0.26$; $p < 10^{-12}$; S9 Fig), suggesting that activity-dependent modulation of *Sst* cells may cause them to move along the continuum of latent factor values. A weaker but still significant correlation was observed for *Pvalb* neurons ($r = 0.11$; $p < 0.002$), whereas no significant relationship was

found for *Vip* neurons ($p = 0.17$). These data therefore suggest that a portion of the continuous variability of gene expression observed in CA1 interneurons may arise from activity-dependent modulation but that such modulation is unlikely to be a full explanation for the genetic continua revealed by latent factor analysis.

## Histological confirmation of transcriptomic predictions

The transcriptomic classification we derived makes a large number of predictions for the combinatorial expression patterns of familiar and novel molecular markers in distinct CA1 interneuron types. To verify our transcriptomic classification, we set out to test some of these predictions using traditional methods of molecular histology.

Our first tests regarded the very distinct *Sst.Nos1* cluster of continent 10. This cluster's expression pattern matched three previously reported rare hippocampal inhibitory cell types: large SST-immunopositive cells that are intensely immunoreactive for NOS1 throughout the cytoplasm, revealing their full dendrites [27]; PENK-positive projection cells [10]; and strongly NADPH diaphorase-labeled (i.e., NOS1-positive) backprojection cells [13]. We therefore hypothesized that these cell types, previously regarded as separate, may in fact be identical. To test this hypothesis, we performed a series of triple and quadruple immunoreactions, focusing on the intensely NOS1-positive neurons ($n = 3$ mice, $n = 70$ cells: 39% in so/alveus; 10% in stratum pyramidale (sp); 27% in sr; 24% at the sr/slm border). Similar to previously reported PENK-projection, backprojection, and SST/NOS1 cells [10,13,27]—but unlike SST-positive O-LM cells [9]—these neurons all showed spiny or sparsely spiny dendrites. As expected from the *Sst.Nos1* cluster, we found that they were all SST/NPY double positive ($n = 20/20$) and were virtually all weakly positive for CHRM2 ($n = 36/38$) and GRM1 ($n = 17/17$) in the somatodendritic plasma membrane, strongly positive for PCP4 ($n = 19/21$) in the cytoplasm and nucleus, and strongly positive for PENK ($n = 35/42$) in the Golgi apparatus and granules in the soma and proximal dendrites (Fig 7). By contrast, the more numerous moderately NOS1-positive cells (which include many interneuron types such as ivy, MGE-neurogliaform, and a subset of IS-3 neurons) were mostly immunonegative for CHRM2, PCP4, and PENK, although some were positive for GRM1. Our results are therefore consistent with the hypothesis that all three previously reported classes correspond to the *Sst1.Nos1* cluster.

A second prediction of our classification was the expression of *Npy* in multiple subclasses of *Cck* cell, most notably the *Slc17a8*- and *Calb1*-expressing clusters of continent 8. This was unexpected, as NPY (at least at the protein level) has instead been traditionally associated with SST-expressing neurons and ivy/neurogliaform cells (Fuentealba et al., 2008a; Katona et al., 2014). Nevertheless, no studies to our knowledge have yet examined immunohistochemically whether the neuropeptides NPY and CCK can be colocalized in the same interneurons. We therefore tested this by double immunohistochemistry in sr and slm (Fig 8A, $n = 3$ mice). Consistent with our predictions, 119 out of 162 (74% ± 6%) of the cells immunopositive for pro-CCK were also positive for NPY (an additional 73 cells were positive for NPY only, which, according to our identifications, should represent neurogliaform and radiatum-retrohippocampal cells). A subset (176 cells) of NPY and/or pro-CCK immunopositive neurons were further tested for CALB1 in triple immunoreactions. As expected, nearly all CALB1-positive neurons were pro-CCK positive (89% ± 2%), and CALB1 immunoreactivity was seen in a subset of the cells containing both pro-CCK and NPY (27% ± 3%). Additional triple immunohistochemistry for NPY, pro-CCK, and SLC17A8 (VGLUT3) revealed triple positive cells in sr and particularly at the sr/slm border, as predicted by the class *Cck.Cxcl14.Slc17a8* (Fig 8B). Because of the low level of somatic immunoreactivity for SLC17A8 (which, as a vesicular transporter, is primarily trafficked to axon terminals), we could not count these cells reliably;

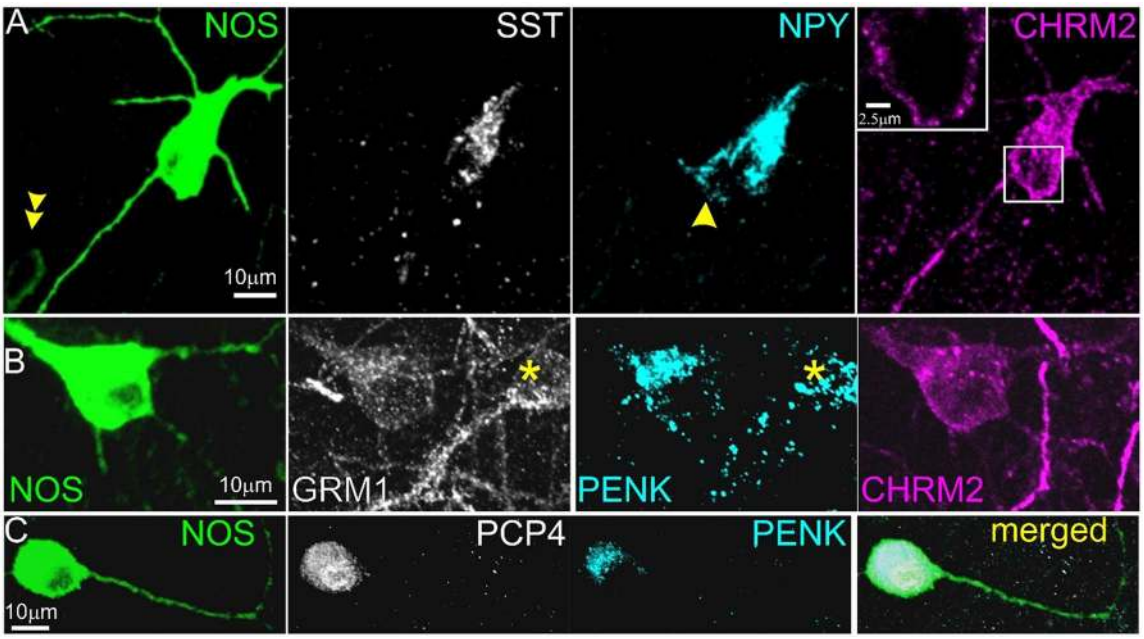


**Fig 7. Immunohistochemical characterization of intensely NOS1-positive neurons. (A)** A large multipolar neuron in stratum pyramidale is strongly SST and NPY positive in the somatic Golgi apparatus and weakly positive for CHRM2 in the somatodendritic plasma membrane (maximum intensity projection, z stack, height 11 μm; inset, maximum intensity projection of three optical slices, z stack height 2 μm). A smaller, more weakly NOS1-positive cell (double arrow) in the lower left is immunonegative for the other molecules; a second NPY-positive cell (arrow) adjoining the NOS1+ neuron is immunonegative for the other three molecules. (**B**) A NOS1-positive cell and another NOS1-immunonegative cell (asterisk) at the border of stratum radiatum and lacunosum-moleculare are both positive for GRM1 in the plasma membrane and PENK in the Golgi apparatus and in granules, but only the NOS1+ cell is immunopositive for CHRM2 (maximum intensity projection, z stack, height 10 μm). (**C**) An intensely NOS1-positive cell in stratum radiatum is also positive for PCP4 in the cytoplasm and nucleus and for PENK in the Golgi apparatus and in granules (maximum intensity projection, z stack, height 15 μm).

https://doi.org/10.1371/journal.pbio.2006387.g007

however, of the cells that were unambiguously immunopositive for SLC17A8, in a majority we detected NPY. Additional analysis combining double in situ hybridization for *Slc17a8* and *Npy* with immunohistochemistry for pro-CCK (Fig 8C, $n = 3$ mice) confirmed that the great majority of *Slc17a8*-expressing cells were also positive for *Npy* and pro-CCK (84% ± 3%). As predicted by our identifications, the converse was not true: a substantial population of *Npy*/pro-CCK double-positive cells (57% ± 7% of the total) did not show detectable *Slc17a8*, which we identify with dendrite-targeting neurons in the east of continent 8.

Several cell types in our classification expressed *Cxcl14*, a gene whose expression pattern in the Allen Brain Atlas shows localization largely at the sr/slm border. The *Cxcl14*-positive population includes all clusters of continent 8, which express *Cck* and contain subclusters expressing *Npy*, *Calb1*, *Reln*, and *Vip*; a subtype of CGE-derived neurogliaform cell that expresses *Reln* and *Npy* but lacks *Nos1* and expresses *Kit* at most weakly; as well as IS-1, IS-2, and radiatum-retrohippocampal cells. However, as all *Cxcl14*-positive clusters lacked *Lhx6*, we conclude they should be distinct from all MGE-derived neurons, including MGE-derived neurogliaform cells.

To test these predictions, we performed in situ hybridization for *Cxcl14* simultaneously with in situ hybridization or immunohistochemistry to detect *Reln*, *Npy*, CALB1, CCK, PVALB, *Sst*, *Nos1*, and *Kit* ($n = 3$ mice; Fig 9). In addition, we combined fluorescent in situ hybridization for *Cxcl14* with immunohistochemistry for yellow fluorescent protein (YFP) in *Lhx6-Cre/R26R-YFP* mice, which allows identification of developmental origin by marking

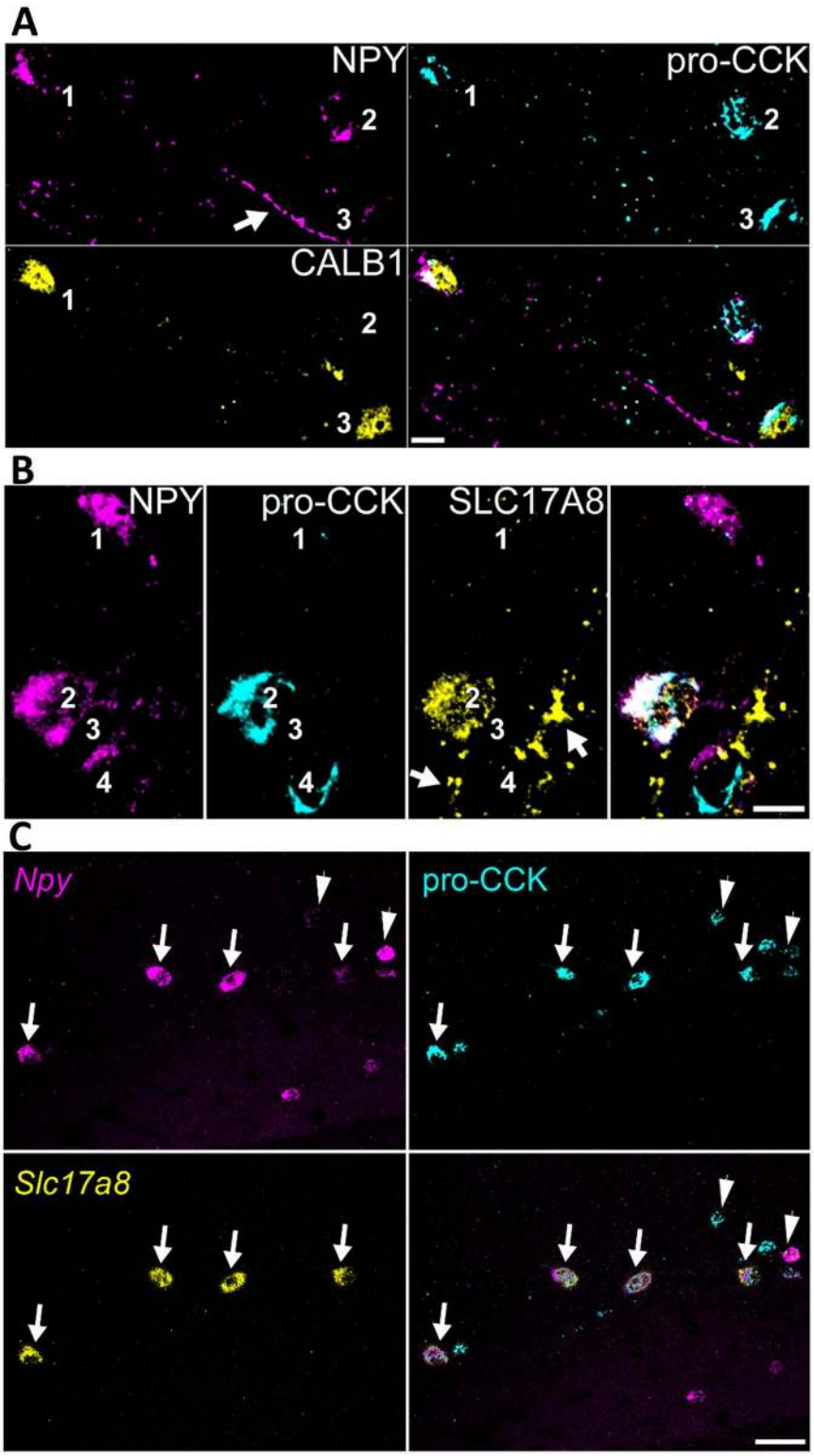


**Fig 8. Confirmation of predicted colocalization of NPY and pro-CCK.** (A) Interneurons at the sr/slm border immunopositive for both NPY and pro-CCK (cells 1 and 2), one of which (cell 1) is also immunopositive for CALB1. A third neuron is positive only for pro-CCK and CALB1 (cell 3). (B) Interneurons at the sr/slm border immunopositive for NPY (cells 1–3), pro-CCK (cells 2 and 4), and SLC17A8 (VGLUT3, cell 2). Note SLC17A8-positive terminals targeting unlabeled cells (arrows). (A, B) Both NPY and pro-CCK are detected in the Golgi apparatus and endoplasmic

reticulum surrounding cell nuclei; in addition, some axons are also immunopositive for NPY (see arrow in (A); average intensity projections, z stacks, height 6.3 μm and 10.4 μm, respectively). (C) Combined double in situ hybridization and immunohistochemistry shows that nearly all Slc17a8-expressing cells also express Npy and are immunopositive for pro-CCK (arrows), but some Npy/pro-CCK cells do not express Slc17a8 (arrowheads). Scale bars: 10 μm (A, B), 50 μm (C). sr/slm, stratum radiatum and stratum lacunosum-moleculare.

https://doi.org/10.1371/journal.pbio.2006387.g008

MGE-derived interneurons (Fogarty et al., 2007). The results of these experiments were consistent with our hypotheses. We found that within CA1, *Cxcl14*-expressing cells were primarily located at the sr/slm border (71% ± 3%), although a subpopulation of cells were also found in other layers. We found no overlap of *Cxcl14* with YFP in the *Lhx6-Cre/R26R-YFP* mouse, confirming the CGE origin of *Cxcl14*-expressing neurons (Fig 9A). The majority of *Cxcl14*-positive cells expressed *Reln* (72% ± 4%), accounting for 42% ± 9% of Reln-expressing neurons (substantial populations of *Reln*+/*Cxcl14*– cells located in so and slm likely represent O-LM and MGE-neurogliaform cells, respectively (Fig 9B). Indeed, although less than half of *Reln* cells were located at the R-LM border (44% ± 1%), the great majority of *Reln*+/*Cxcl14*+ cells were

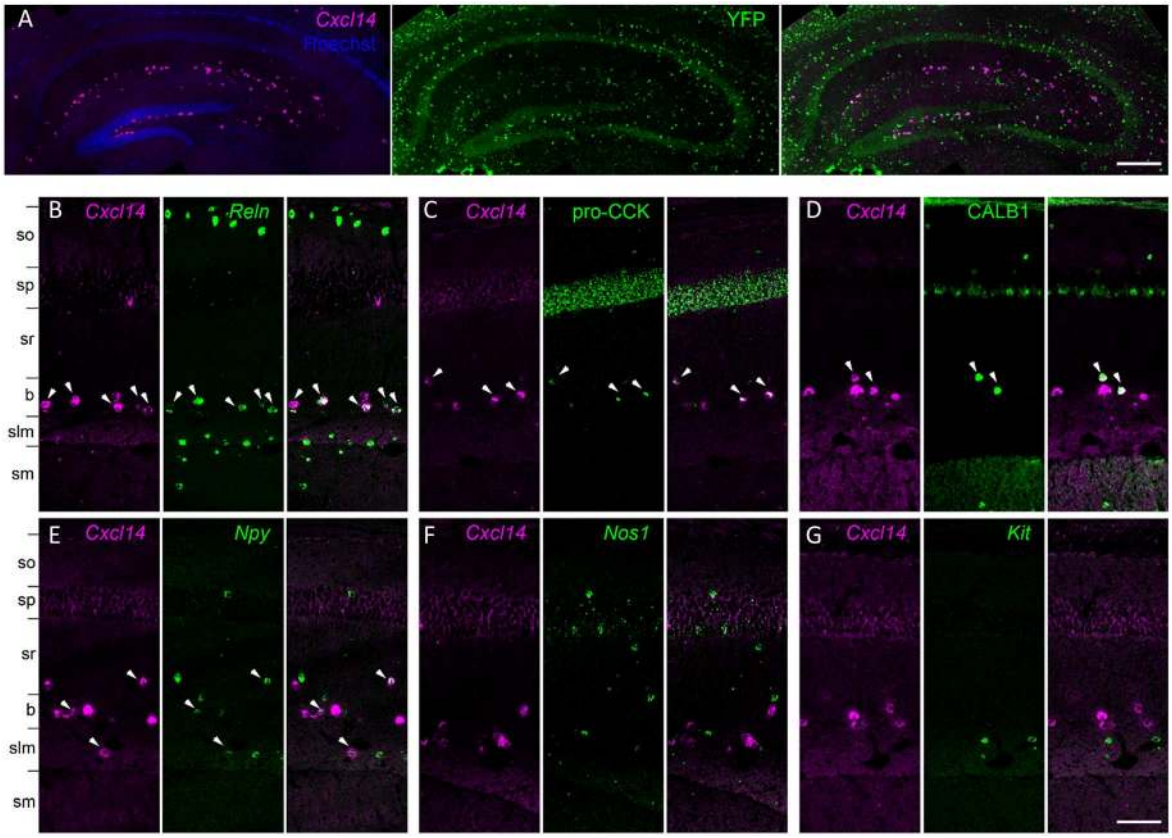


**Fig 9. Analysis of *Cxcl14* co-expression patterns confirms predicted properties *Cck.Cxcl14* cells. (A)** *Cxcl14*-expressing cells are CGE derived: in situ hybridization for *Cxcl14* combined with immunohistochemistry for YFP in the *Lhx6-Cre/R26R-YFP* mouse yields no double labeling. **(B)** Double in situ hybridization for *Cxcl14* and *Reln* marks a population of neurons located primarily at the sr/slm border. Note *Reln* expression without *Cxcl14* in so and slm, likely reflecting O-LM and neurogliaform cells. **(C–E)** Subsets of the *Cxcl14*-positive neurons are positive for pro-CCK or CALB1 (in situ hybridization plus immunohistochemistry), or *Npy* (double in situ hybridization). **(F, G)** No overlap was seen of *Cxcl14* with *Nos1* or *Kit*. In all panels, arrowheads indicate double-expressing neurons. Scale bars: 200 μm (A), 100 μm (B–G). b, sr/slm border region; O-LM, oriens/lacunosum-moleculare; slm, stratum lacunosum-moleculare; sm, stratum moleculare of the dentate gyrus; so, stratum oriens; sp, stratum pyramidale; sr, stratum radiatum; sr/slm, stratum radiatum and stratum lacunosum-moleculare; YFP, yellow fluorescent protein.

https://doi.org/10.1371/journal.pbio.2006387.g009

found there (88% ± 6%). Consistent with the expected properties of continent 8 cells, a large fraction of the *Cxcl14* population were immunoreactive for pro-CCK (62% ± 6%; Fig 9C), while substantial minorities were positive for CALB1 (29% ± 2%; Fig 9D) or *Npy* (25% ± 5%; Fig 9E). However, as expected from the lack of *Cxcl14* in MGE-derived neurogliaform and IS-3 cells, we observed no overlap of *Cxcl14* with *Nos1* (0 out of 209 cells; Fig 9F) and very weak overlap with *Kit*, which is primarily expressed in clusters *Cacna2d1.Ndnf.Npy* and *Cacna2d1.Ndnf.Rgs10*, associated with the *Cxcl14*-negative CGE-neurogliaform population (1 of 264 cells, respectively, from all mice; Fig 9G).

The cluster *Cck.Cxcl14.Vip* presented a puzzle, because *Cxcl14* is located primarily at the sr/slm border, whereas immunohistochemistry in rat has localized CCK/VIP basket cells to sp [24]. Because *Cxcl14* expression can sometimes also be found in sp, we tested whether this cluster reflects sp cells by combining in situ hybridization for *Cxcl14* with immunohistochemistry against VIP in mouse CA1 (Fig 10). This revealed frequent co-expression at the sr/slm border (8% ± 1% *Cxcl14* cells positive for *Vip*; 23% ± 1% *Vip* cells positive for *Cxcl14*) but very few *Cxcl14* cells in sp, and essentially no double labeling (1 of 147 *Vip* cells in sp was weakly labeled

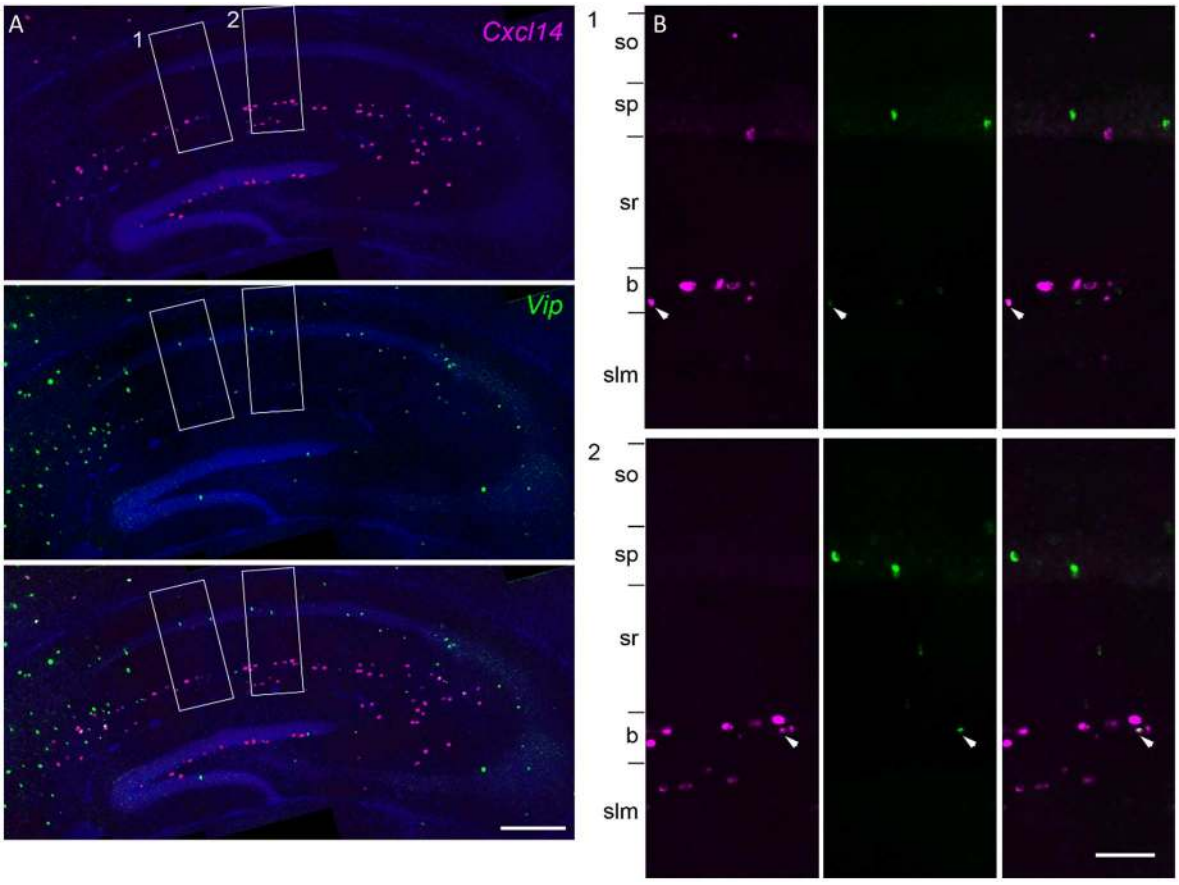


**Fig 10. Overlap of *Cxcl14* and *Vip*.** The class *Cck.Cxcl14.Vip* represented a puzzle: *Vip/Cck* cells had previously been reported in sp, but *Cxcl14* is detected primarily at the sr/slm border, although exceptional cells can be detected in sp also. **(A)** Double fluorescent in situ hybridization images reveal that the vast majority of cells co-expressing *Cxcl14* and *Vip* were found at the sr/slm border, confirming the location of this novel class. **(B)** Zoom into rectangles 1 and 2. Arrowheads: double-expressing cells. b, sr/slm border region; slm, stratum lacunosum-moleculare; so, stratum oriens; sp, stratum pyramidale; sr, stratum radiatum; sr/slm, stratum radiatum and stratum lacunosum-moleculare.

https://doi.org/10.1371/journal.pbio.2006387.g010

for *Cxcl14)*. We therefore conclude that this cluster indeed represents a novel cell type located at the sr/slm border, expressing *Cck*, *Vip*, and *Cxcl14*.

## Discussion

The molecular architecture of CA1 interneurons has been intensively studied over the last decades, leading to the identification of 23 inhibitory classes. Our transcriptomic data showed a remarkable correspondence to this previous work, with all previously described classes identified in our database. Our analysis also revealed a continuous mode of variability common across multiple cell types, eight hypothesized novel classes, as well as additional molecular subdivisions of previously described cell types.

Surprisingly, these data suggest that three previously described CA1 cell groups in fact represent a single cell class, a fact previously overlooked because of the limited combinations of molecules tested in prior work. The *Sst.Nos1* class is strongly positive for *Nos1* and also expresses *Sst*, *Npy*, *Chrm2*, *Pcp4*, and *Penk*, but unlike *Penk*-positive I-S cells of continent 9, it lacks *Vip*. This class is homologous to the "Int1" and "Sst Chodl" classes defined in isocortex, which have been identified with long-range projecting sleep-active neurons [44,46,70,71]. The three previously described classes identified with *Sst.Nos1* are PENK-immunopositive neurons with projections to subiculum, which were shown to be VIP negative, but not tested for SST or NOS1 (Fuentealba et al., 2008b); the NADPH diaphorase-labeled (i.e., strongly NOS1-positive) axons reported by Sik and colleagues (1994) as projecting to CA3 and dentate, but not tested for SST or PENK; and the SST/NOS1 cells identified by Jinno and Kosaka (2004) in mouse, which were not tested for long-range projections or for PENK. While it remains possible that a larger transcriptomic sample of these rare neurons would reveal subclasses, our present data suggest that *Sst.Nos1* cells are a homogeneous population: the nbtSNE algorithm, BIC criterion, and further manual exploration failed to reveal any finer distinctions. We therefore suggest that they constitute a class of inhibitory neurons with diverse long-range projection targets. Interestingly, the targets of PENK-positive projection cells are most commonly PVALB-positive interneurons, unlike conventional IS cells, which preferentially target SST cells [10]. As these cells are identified as sleep active, this fact may provide an important clue to the mechanisms underlying sleep in cortical circuits.

The match between our transcriptomic analysis and previous immunohistochemical work (primarily in rat) is so close that it is simpler to describe the few areas of disagreement than the many areas of agreement. First, ACTN2 has been used as a neurogliaform marker in rat [72] but was almost completely absent from any cell type of our database. We suggest this reflects a species difference, as previous attempts with multiple ACTN2 antibodies have been unsuccessful in mouse (J. H.-L., unpublished observations), and *Actn2* labeling is not detectable in the Allen atlas [73]. Second, we observed *Calb2* in a subset of putative O-LM cells; these *Calb2*-expressing neurons typically also expressed *Calb1*. Such O-LM cells have not been described in rat [9], but CALB2/SST neurons have been observed in mouse isocortex [44,74]. A third inconsistency regards NCALD, which in rat was reported not to overlap with PVALB, SST, or NPY [75], but did so in our data. Finally, it has previously been reported that a subset of O-LM cells show *Htr3a* expression [76]. In our data, we observed at most weak expression of *Htr3a* in *Sst* cells, and the cells showing it belonged to clusters identified as hippocamposeptal rather than O-LM cells.

Our analysis revealed several rare and presumably novel cell groups, although we cannot exclude that some of these were inadvertently included from neighboring areas such as subiculum (S10 Fig). *Sst.Npy.Serpine2* and *Sst.Npy.Mgat4c*, which simultaneously expressed *Sst*, *Npy*, and *Reln*, fit the expected expression pattern of neither O-LM nor hippocamposeptal cells; *Sst*.

*Erbb4.Rgs10* is a distinct group related to *Pvalb* basket and bistratified cells; *Cck.Lypd1* formed a rare and highly distinct class expressing *Cck*, *Slc17a8*, and *Calb1*; *Ntng1.Synpr* showed an expression pattern with features of both sr/slm *Cck* neurons and projection cells; and *Cck.Cxcl14.Vip* represents a cell class strongly positive for both *Cck* and *Vip* located at the sr/slm border that appears to be a pyramidal- rather than interneuron-targeting class. The analysis also revealed subdivisions of known types, such as the division of IS-3 cells into *Nos1*-positive and -negative groups, and the division of CGE-NGF cells into *Car4*- and *Cxcl14*-expressing subtypes. Finally, our data suggested that with more cells or deeper sequencing, even rarer types are likely to be found, as subsetting analysis showed a linear increase in the number of clusters with cell count and read depth, with little sign of saturation as yet. The data appeared to contain several novel cell types not containing enough cells to overcome the algorithm's parsimony penalty, such as a small group of cells with features of both basket and axo-axonic cells located off the coast of continent 3; such cells have indeed been rarely encountered by quantitative electron microscopic analysis of synaptic targets in the rat (P. S., unpublished observations).

Latent factor analysis revealed a common continuum of gene expression across the database, suggesting a large "module" of genes that are coregulated in multiple types of hippocampal interneuron. The latent factor differed between clusters, and clusters with larger latent factor values were identified with interneuron types targeting pyramidal cell somas or proximal dendrites (such as *Pvalb* or *Cck/Slc17a8* expressing basket cells), while those with low mean values were identified with interneurons targeting pyramidal distal dendrites (such as *Sst* or *Cck/Calb1* expressing dendrite-targeting cells) or targeting other interneurons. Subtler differences in latent factor were found within clusters, suggesting that a similar continuum exists within cells of a single type. Genes positively correlated with the latent factor are associated with fast-spiking phenotype, presynaptic function, GABA release, and metabolism. Consistent with this expression pattern, perisomatic inhibitory cells show fast-spiking phenotypes and deliver powerful, accurately timed inhibition [8], but interneurons targeting distal dendrites show slower-spiking patterns; presumably because distal inputs are subject to passive dendritic filtering, their presynaptic vesicle release does not need to be so accurately timed. I-S cells had the lowest mean values of the latent factor, consistent with their small axonal trees and metabolic machinery [77]. The stronger expression of many neuropeptides in cells of low latent factor suggests that these slower, distal-targeting interneurons may also rely more heavily on neuropeptide signaling, for which slow firing rates support outputs transduced by slower G-protein–coupled receptors. Interestingly, a study conducted independently of the present work identified enriched expression of a gene module similar to our latent factor in isocortical *Pvalb* neurons [43] and suggested it is controlled by the transcription factor PGC-1α [78,79]. Our results suggest that *Cck*-expressing basket cells have a similar expression pattern and that, more generally, expression of this module correlates with a neuron's axonal target location.

Several novel genes correlating with the factor appear to be interesting candidates for future research, such as *Trp53i11*, *Yjefn3*, and *Rgs10*, associated with faster-spiking *Cck* cells; *Zcchc12* and *6330403K07Rik*, both associated with slower-firing cells of all classes; and *Fxyd6*, associated with slow spiking, which may modulate ion exchange. Intriguingly, genes for neurofilaments and other intermediate filaments were positively correlated with the latent factor, while genes involved in actin processing were negatively correlated; we hypothesize that this might reflect a different cytoskeletal organization required for somatic- and dendritic-targeting neurons.

The question of how many cell classes a given neural circuit contains is often asked of transcriptomic analyses, but we argue this question will not have a clearly defined answer. For example, our data indicate no sharp dividing line between ivy cells and MGE-derived neurogliaform cells. Yet cells at the two ends of the continuum are clearly different: not only do their

gene expression patterns differ substantially, but their different axonal targets indicate different roles in circuit function [23]. In statistics, multiple criteria can be used to define how many clusters should be assigned to a dataset; a common approach (which is used by the ProMMT algorithm) is to consider a cluster indivisible if within-cluster fluctuations cannot be distinguished from random noise. Using this criterion, the number of clusters of CA1 interneurons increased with the number of cells and read depth analyzed, showing no sign of saturation in the current dataset. Furthermore, we observed several apparent rare classes that were too small to be assigned their own clusters at present, together with further subtle gradations within currently assigned clusters. The fact that we observed more clusters in CA1 than the 23 previously identified in isocortex [44] should therefore not be taken as implying that CA1 is a more complex circuit but simply that our larger sample size and different clustering algorithm were able to detect finer distinctions. Indeed, our data suggest that while the divisions between the 10 major "continents" are unambiguous, the organization of gene expression within these continents is complex and subtle, and likely far more detailed than characterized by our present 49 clusters. An understanding of this multiscale variability in gene expression in CA1 interneurons will be a key tool to understanding the function of this circuit.

## Methods

### scRNA-seq

**Animals.** *Slc32a1* (vesicular GABA transporter)-*Cre* BAC transgenic mice [80] were crossed with a tdTomato reporter line to generate mice with fluorescently labeled inhibitory neurons. Both the *Slc32a1-Cre* and tdTomato mouse lines were of mixed B6 and CD1 backgrounds. Three of these mice were used for both the p28 and p63 cohorts; both males and females were used at each age. All experimental procedures followed the guidelines and recommendations of Swedish animal protection legislation and were approved by the local ethical committee for experiments on laboratory animals (Stockholms Norra Djurförsöksetiska nämnd, Sweden; N282/14). The raw data are available on GEO under accession number GSE99888, and a gene expression matrix together with results of clustering, latent factor analysis, and nbtSNE are available at https://figshare.com/articles/Transcriptomic_analysis_of_CA1_inhibitory_interneurons/6198656. An online viewer for expression maps (cf. Fig 3) is available at http://linnarssonlab.org/ca1/.

**Single-cell suspension and FACS.** Dissection and single cell dissociation were carried out as described before (Marques et al. 2016), with slight alterations for p63 animals, for which NMDG-HEPES–based solution was used in all steps to enable better recovery of the aged cells (Tanaka, 2008). The NMDG-HEPES–based cutting solution contained 93 mM NMDG, 2.5 mM KCl, 1.2 mM $NaH_2PO_4$, 30 mM $NaHCO_3$, 20 mM HEPES, 25 mM glucose, 5 mM sodium ascorbate, 2 mM thiourea, 3 mM sodium pyruvate, 10 mM $MgSO_4{}^*7H_2O$, 0.5 mM $CACl_2{}^*2H_2O$, and 12 mM N-acetyl-L-cysteine; it was adjusted to pH 7.4 with 10N HCl. Mice were humanely killed by an overdose of Isoflurane and Ketamine/Xylazine, followed by transcardial perfusion through the left ventricle with artificial cerebrospinal fluid (aCSF) equilibrated in 95% $O_2$ 5% $CO_2$ before use. The brains were removed and CA1 was microdissected from 300 μm vibratome sections. Single-cell suspensions were prepared using Papain (Worthington) with 30 min enzymatic digestion, followed by manual trituration with fire-polished Pasteur pipettes. The albumin density gradient was only performed for the p63 samples. On a BD FACSAria II, tdTomato-positive cells were sorted into oxygenated aCSF at 4 ˚C, concentrated, inspected for viability, and counted.

To assess the accuracy of our dissection, we studied the gene expression patterns of simultaneously collected pyramidal cells using previously published genetic criteria (S10 Fig). No cells

exhibited expression patterns consistent with CA2 or CA3 [35], but a fraction of these cells (62 of 357 total excitatory neurons) expressed genes seen in a region stretching from the dorsomedial lip of CA1 to the subiculum. Although this result is consistent with dissection of only CA1 interneurons, we also cannot rule out the presence of a small number of atypical interneuron classes located at the dorsomedial lip, or of inclusion of some subicular interneurons.

**10X Chromium mRNA-seq.** Sorted suspensions were added to 10X Chromium RT mix, aiming at 2,500 cells recovered per experiment. Downstream cDNA synthesis (14 PCR cycles) and library preparation were carried out as instructed by the manufacturer (10X Genomics Chromium Single Cell Kit Version 1). Libraries were sequenced on the Illumina HiSeq2500 to an average depth of 112,000 reads per cell (raw), yielding on average 3,600 distinct molecules and 1,700 genes per cell. Demultiplexed samples were aligned to the reference genome and converted to mRNA molecule counts using the "cellranger" pipeline version 1.1, provided by the manufacturer.

**Normalization.** Prior to many analyses (clustering, latent factor analysis, and nbtSNE), the expression vectors for each cell were normalized, so that each cell's total RNA expression became equal to the total cellular RNA count averaged over all cells in the database. However, scatterplots of expression (Fig 6C) show unnormalized values.

**Quality control.** Cells showing abnormally high values of nuclear noncoding RNAs (*Meg3*, *Malat1*, *Snhg11*) or mitochondrial genes were discarded, as this can signify cell lysis. Cells were discarded if the summed normalized expression of these genes exceeded a threshold of 600.

## Immunohistochemistry (Oxford)

Six adult (20 wk old) male C57BL/6J mice (Charles River, Oxford, UK) were perfusion fixed following anesthesia and tissue preparation for immunofluorescence (Katona et al., 2014) and analysis using wide-field epifluorescence microscopy [21] was performed as described. The following primary antibodies were used: anti-calbindin (goat, Fronteir Inst, Af104); anti-pro-CCK (rabbit, 1:2,000, Somogyi et al., 2004); anti-metabotropic glutamate receptor 1a (GRM1, rabbit, 1:1,000; guinea pig, 1:500; gifts from Prof. M. Watanabe, Frontier Institute); anti-muscarinic acetylcholine receptor 2 (CHRM2, rat, 1:400, EMD Millipore Corporation, MAB367); anti-NOS1 (rabbit, 1:1,000, EMD Millipore Corporation, AB5380; mouse, 1:1,000, Sigma-Aldrich, N2280); anti-NPY (mouse, 1:5,000, Abcam, #ab112473); anti-Purkinje cell protein 4 (PCP4, rabbit, 1:500, Santa Cruz Biotechnology, sc-74816); anti-pre-pro-enkephalin (PENK, guinea pig, 1:1,000, gift from Takahiro Furuta, Kyoto University, Japan; rabbit, 1:5,000, Life-Span Biosciences, LS-C23084); anti-SST (sheep, 1:500, Fitzgerald Industries International, CR2056SP); anti-VGLUT3 (guinea pig, Somogyi et al 2004). Secondary antibodies were raised in donkey against immunoglobulin G of the species of origin of the primary antibodies and conjugated to Violet 421 (1:250); DyLight405 (1:250); Alexa 488 (1:1,000); cyanine 3 (1:400); Alexa 647 (1:250); cyanine 5 (Cy5, 1:250). With the exception of donkey-antimouse-Alexa488 purchased from Invitrogen, all secondary antibodies were purchased from Stratech.

For cell counting, image stacks (212 × 212 μm area; 512 × 512 pixels; z stack height on average 12 μm) were acquired using LSM 710/AxioImager.Z1 (Carl Zeiss) laser scanning confocal microscope equipped with Plan-Apochromat 40×/1.3 Oil DIC M27 objective and controlled using ZEN (2008 v5.0 Black, Carl Zeiss). In a second set of sections, images were taken using Leitz DM RB (Leica) epifluorescence microscope equipped with PL Fluotar 40×/0.5 objective. Counting was performed either using ImageJ (v1.50b, Cell Counter plugin) on the confocal image stacks or OPENLAB software for the epifluorescence documentation. For the CCK counts, numbers were pooled from two separate reactions testing for a given combination of

primary antibodies ($n$ = 3 mice each reaction, 2–3 sections each mouse) and reported as average values ± standard deviation. For the testing of intensely nNOS-positive neurons, cells were selected using Leitz DM RB (Leica) epifluorescence microscope equipped with PL Fluotar 40×/0.5 objective. Cells were pooled from three separate reactions testing for a given combination of primary antibodies ($n$ = 3 mice each reaction, 2 sections each mouse) and reported as pooled data. Image processing was performed using ZEN (2012 Blue, Carl Zeiss), ImageJ (v1.51m, open source), Inkscape (0.92, open source), and Photoshop (CS5, Adobe).

## In situ hybridization (UCL)

Wild-type (C57BL/6/CBA) male and female adult (p30) mice and *Lhx6-Cre*$^{Tg}$ transgenic mice were perfusion-fixed, as previously described (Rubin et al., 2010), followed by immersion fixation overnight in 4% paraformaldehyde. Fixed samples were cryoprotected by overnight immersion in 20% sucrose, embedded in optimal cutting temperature (OCT) compound (Tissue Tek, Raymond Lamb Ltd Medical Supplies, Eastbourne, UK), and frozen on dry ice. 30 $\mu$m cryosections were collected in DEPC-treated PBS and double in situ hybridization was carried out as described (Rubin et al., 2010). Probes used included either a *Cxcl14*-(digoxgenin)DIG RNA probe in combination with *Reln*-(fluorescein)FITC; *Npy*-FITC, *Sst*-FITC, or *Vip*-FITC probes; or a *Cxcl14*-FITC probe with *Nos1*-DIG, *Kit*-DIG, *Scl17a8*-DIG, or *Pvalb*-DIG probes. DIG-labeled probes were detected with an anti-DIG-alkaline phosphatase (AP)-conjugated antibody followed by application of a Fast Red (Sigma) substrate. The first reaction was stopped by washing 3 × 10 min in PBS, and the sections were incubated with an anti-FITC-Peroxidase (POD)-conjugated antibody (1:1,500—Roche) overnight. The POD signal was developed by incubating the sections with Tyramide-FITC:amplification buffer (1:100, TSA-Plus, Perkin Elmer) for 10 min at room temperature. For immunohistochemistry after in situ hybridization, the following antibodies were used: anti-Calbindin (rabbit, 1:1,000, Swant, Bellinzona, Switzerland); anti-pro-CCK (rabbit, 1:2,000, Somogyi et al., 2004); anti-GFP (chicken, 1:500, Aves Labs). All sections were counterstained with Hoechst 33258 dye (Sigma, 1,000-fold dilution) and mounted with Dako Fluorescence Mounting Medium (DAKO).

For cell counts, images (at least two sections per mouse) were acquired on an epifluorescence microscope (Zeiss) with a 10× objective. Several images spanning the entire hippocampal CA1 were stitched using Microsoft Image Composite Editor. Cells were counted manually in the CA1 area, including sr and slm, and in a subregion spanning 100 $\mu$m across the border between sr and slm, where most *Cxcl14*-positive cells are located. Confocal images (z stack height on average 25 μm, 2 $\mu$m spacing) were taken on a Leica confocal microscope under a 10× objective and processed for contrast and brightness enhancement with Photoshop (CS5, Adobe). A final composite was generated in Adobe Illustrator (CS5, Adobe).

## Cluster analysis

Code for cluster analysis and all other algorithms can be found at https://github.com/cortex-lab/Transcriptomics.

**Sparse mixture model.** The ProMMT algorithm performs cluster analysis by modeling molecular counts by a mixture of sparse multivariate negative binomial distributions. Specifically, let $\mathbf{x}$ represent the $N_{genes}$-dimensional vector summarizing the expression of all genes in a single cell. We model the probability distribution of $\mathbf{x}$ with a mixture model:

$$\Pr(\mathbf{x}) = \sum_k \Pr(\mathbf{x}|k)\pi_k \quad (1)$$

Here, $k$ denotes a cell class, $p(\mathbf{x}|k)$ denotes the probability that a cell in this class will have expression vector $\mathbf{x}$, and the "class prior" $\pi_k$ represents the fraction of cells belonging to this class. To model $p(\mathbf{x}|k)$, we use the following distribution family:

$$\Pr(\boldsymbol{x}|k) = \prod_g \begin{cases} \Pr(x_g|\mu_{g,0}) & g \notin S \\ \Pr(x_g|\mu_{g,k}) & g \in S \end{cases} \quad (2)$$

In this family, the distribution of all genes is modelled as conditionally independent within a class. The within-class distribution of each gene g depends on a single parameter $\mu_{g,k}$ (the mean level of the gene in that class). Furthermore, the distributions of only a subset $S$ of genes are allowed to vary between classes, while the remainder are constrained to have a class-independent distribution with mean $\mu_{g,0}$. Taking $S$ to have a fixed and small size $N_S$ ensures a "sparse model," which can be fit robustly in high dimensions from only a small number of cells. Note that while the set S could in principle vary between classes, we have found that using a single set $S$ for all classes provides good results.

**Negative binomial distribution.** To model the variability of each gene within a class, we use a negative binomial distribution. The negative binomial distribution is a model of count data with greater variance than the Poisson distribution and is frequently used as a model for gene expression levels [47,48]. The negative binomial is specified by two parameters, $r$ and $p$, and has distribution

$$Pr(x; r, p) = \binom{x + r - 1}{x} p^x (1 - p)^r \quad (3)$$

This distribution has mean $\mu = \frac{rp}{1-p}$, and for fixed $r$, the maximum likelihood estimate of parameter $p$ is $\frac{\mu}{r+\mu}$, where $\mu$ is the sample mean. For fixed $r$, the standard deviation of this distribution scales asymptotically linearly with its mean: $\sigma = \sqrt{\frac{\mu^2}{r} + \mu}$. In contrast, the Poisson distribution has a smaller standard deviation, which scales with the square root of the mean.

We verified that a negative binomial with fixed $r$ is appropriate for scRNA-seq data by considering a relatively homogeneous class (CA1 pyramidal cells; S2A Fig; data from Zeisel et al [46]). This analysis confirmed that the negative binomial with $r = 2$ accurately modelled the relationship of standard deviation to mean in these data. The "wide" shape of the negative binomial distribution (S2B Fig) has a consequence that the absolute expression levels of a gene matters much less than whether the gene is expressed at all. Indeed, examining the symmetrized Kullback-Leibler divergence of negative binomials with different means (S2C Fig)—an indication of the penalty paid for misestimating the mean expression level—indicates that a much smaller penalty is paid for fitting a mean of 500 to a distribution whose actual mean is 1,000, than for fitting a mean of 10 to a distribution whose actual mean is 0.

**EM algorithm.** To fit the model, we fix $r = 2$ and fit the parameters $S$, $\mu$, and $\pi$ by maximum likelihood. Because maximum likelihood fitting involves a sum over (unknown) class assignments, we use a standard Expectation-Maximization (EM) algorithm [81,82]. We define $z_{c,k}$ to be the expected value of an indicator variable taking the value 1 if cell $c$ belongs to class $k$:

$$z_{c,k} = \Pr(k|\mathbf{x}_c; S, \mu).$$

The algorithm alternates between an E step, where $z_{c,k}$ is computed using the current values of the parameters $S$ and $\mu$, and an M step, where $S$ and $\mu$ are optimized according to the current values of $z_{c,k}$.

**E-step.** The E step is straightforward. Observe that

$$\log \Pr(\mathbf{x}_c|k) = const + \log(\pi_k) + \sum_{g \in S} x_{c,g} \log(p_{g,k}) + r \log(1 - p_{g,k})$$

The constant term includes the contributions of all genes not in $S$, as well as the binomial coefficient from Eq 3, none of which depend on the value of $k$ and therefore do not affect the result.

One can compute $z_{c,k}$ from this using Bayes' theorem; in practice, however, we found that when the set $S$ contains a reasonable number of genes (approximately 100 or more), all values of $z_{c,k}$ are close to 0 or 1, so there is little to lose by employing a much faster "hard EM" algorithm, in which, for all cells $c$, only a single winning $k_c$ has $z_{c,k_c} = 1$, with all others 0.

**M-step.** In the M-step, we are given $z_{c,k}$ and must find the set $S$ of genes that are allowed to differ between classes, and their class means $\mu_{g,k}$, by maximum likelihood. Although one might expect finding $S$ to pose an intractable combinatorial optimization problem, it can in fact be solved quickly and exactly. The derivation below is for a hard EM algorithm; the soft case can be derived easily but requires substantially more computation time, without a noticeable increase in performance.

We first define a quantity $L_0$ to be the log likelihood of the data under a model in which $S = \emptyset$, so all the expression of each gene $g$ is determined by its grand mean $\mu_{g,0}$, independent of cluster assignments. Observe that

$$L = L_0 + \sum_{g \in S} Y_g$$

where

$$Y_g = \sum_c x_{g,c} [\log(p_{g,k_c}) - \log(p_{g,0})] + r[\log(1 - p_{g,k_c}) - \log(1 - p_{g,0})]$$

represents the gain in log likelihood obtained when the distribution of gene $g$ is allowed to vary between classes. To compute the optimal value of the set $S$, we note that the values of $Y_g$ are independent of each other. Thus, the optimal set $S$ is simply the $N_s$ genes with the largest values of $Y_g$.

The maximum likelihood estimates of the negative binomial parameters $p_{g,k}$ are given by $\frac{\mu_{g,k}}{r+\mu_{g,k}}$, where $\mu_{g,k}$ denotes the average expression of genes $g$ for the cells currently assigned to cluster $k$, and $\mu_{g,0}$ is the mean expression of gene $g$ for all cells in the database. Because the negative binomial distribution can give zero likelihoods if any $\mu_{g,k} = 0$, we use a regularized mean estimate:

$$\mu_{g,k} = \frac{A + \sum_{c \in k} x_{g,c}}{B + N_k}$$

where $N_k$ denotes the number of cells in cluster $k$, and the regularization parameters take the values $A = 10^{-4}$, $B = 1$.

Finally, we compute the priors $\pi_k$ as the fraction of cells $c$ with $k_c = k$, as is standard in EM.

**BIC penalty.** To automatically choose the number of clusters, we employed the BIC method [50], which for our model takes the form of a penalty $\frac{|S| \log(N_c)}{2}$ per cluster added to the log likelihood.

**Cluster splitting.** As is typical for cluster analysis, the likelihood function has multiple local maxima, and steps must be taken to ensure the algorithm does not become trapped in a suboptimal position. To do this, we use a heuristic that splits clusters that are poorly fit by a

negative binomial distribution. The full clustering method consists of a divisive approach that alternates such splits with EM runs that then re-optimize the parameters.

For each cluster $k$, the splitting heuristic searches for genes $g$ whose likelihood would be substantially increased if the cluster was split in two, according to whether the expression of gene $g$ is above a threshold $\Theta_g$. Note that after splitting, the amount by which the log likelihood gain $Y_g$ changes can be written as

$$\begin{aligned}\Delta Y_{g,\Theta} &= \sum_{c:x_{g,c}<\Theta} x_{g,c}[\log(p_g^{<}) - \log(p_g)] + r[\log(1-p_g^{<}) - \log(1-p_g)] \\ &+ \sum_{c:x_{g,c}\geq\Theta} x_{g,c}[\log(p_g^{\geq}) - \log(p_g)] + r[\log(1-p_g^{\geq}) - \log(1-p_g)]\end{aligned}$$

Here, $p_g$ represents the maximum-likelihood parameter for gene $g$ in the cluster under consideration, $p_g^{<}$ represents this parameter computed only for cells with $x_g < \Theta_g$, and $p_g^{\geq}$ represents this parameter for cells with $x_g \geq \Theta_g$. The only values of $\Theta$ for which a split need be considered correspond to the expression levels of cells in the cluster, and $\Delta Y_{g,\Theta}$ can therefore be rapidly computed for all $g$ and $\Theta$ using cumulative summation, with computational cost linear in the size of the expression matrix.

**Full algorithm.** The full algorithm consists of repeatedly alternating the EM algorithm with cluster splitting and merging operations to escape from local maxima. The algorithm is initialized by assigning all cells to a single cluster.

On each iteration, all clusters are first split in two using the splitting heuristic. Specifically, for each cluster, $\Delta Y_{g,\Theta}$ is computed for all $g$ and $\Theta$, and the optimal split points $\Theta_g$ are found for each gene. The 10 genes giving top values of $\Delta Y_{g,\Theta_g}$ are found. For each of them, the cluster is split, an EM algorithm run to convergence on the resulting cluster pair, and the split providing the highest increase in likelihood is kept. Once all clusters have been split, they are combined to produce a dataset with twice the original number of clusters, and the EM algorithm is run, to allow points to be reassigned between the split clusters.

The iteration ends with a round of cluster pruning. For each cluster, we compute the deletion loss: the decrease in log likelihood that would occur if all points in the cluster were reassigned to their second-best matching cluster. If this loss does not outweigh the BIC penalty, the cluster's points are so reassigned, and EM is run on the full dataset. This process continues until no cluster's deletion loss is smaller than the BIC penalty.

The algorithm is run for a set number of iterations (50 in the current case) and the final result corresponds to the clustering that gave the highest score.

## Isolation metric

To measure how well separated each cluster is from its neighbors, we define an isolation metric equal to the deletion loss (described in the previous section), divided by $N_k \log(2)$, where $N_k$ is the number of cells assigned to cluster $k$. This has an information-theoretic interpretation, as the number of additional bits that would be required to communicate the gene expression pattern of a cell in cluster $k$, using a code defined by the probability model if cluster $k$ were deleted.

## Hierarchical cluster clustering

Each cluster produced by the EM algorithm is specified by a mean expression vector. To understand the relationship between these cluster means, we applied a clustering method to

the clusters themselves. This was achieved using Ward's method, with a distance matrix given by the K-L divergence between cluster means, weighted by the number of cells per cluster.

## nbtSNE algorithm

To visualize the locations of the cells, we derived a variant of the tSNE algorithm [51] appropriate for data following a negative binomial distribution.

Stochastic neighbor embedding algorithms such as tSNE start by converting Euclidean distances between pairs of high-dimensional vectors $x_i$ into conditional probabilities according to a Gaussian distribution: $p_{j|i} = N(x_j; x_i, \sigma_i^2)/\sum_{k\neq i} N(x_k; x_i, \sigma_i^2)$. The tSNE algorithm then adjusts the locations of low-dimensional representation $y_i$ in order to minimize the K-L divergence of a symmetrized $p_{j|i}$, with a t-distribution on the $y_i$.

The Gaussian distribution, however, is not the most appropriate choice for transcriptomic data. We found that we obtained better results using the same negative binomial distribution as in the ProMMT algorithm:

$$p_{j|i} = NB(x_j; x_i, r)/\sum_{k\neq i} NB(x_k; x_i, r)$$

where

$$NB\left(x_j; x_i, r\right) = \exp\left[\sum_{g\in S} x_{gj}\log\left(\frac{x_{gi}}{x_{gi}+r}\right) + r\log\left(\frac{r}{x_{gi}+r}\right)\right]$$

excluding a binomial coefficient that cancels when computing $p_{j|i}$. The sum runs over the set of genes $g$ that were chosen by the ProMMT algorithm.

In the original tSNE algorithm, variations in distance between the points $x_i$ are overcome by adjusting the variance $\sigma_i^2$ for each point $i$ to achieve constant perplexity of the symmetrized conditional distributions. We took the same approach, finding a scale factor $\lambda_i$ for each cell $i$ to ensure that the scaled symmetrized distribution

$$p_{ji} = \exp\lambda_i(log(p_{j|i}) + log(p_{i|j}))$$

had a fixed perplexity of 15. This computation and the implementation K-L minimization was achieved using Laurens van der Maaten's drtoolbox (https://lvdmaaten.github.io/drtoolbox/). The algorithm was initialized by placing all points on a unit circle, with angular position determined by their parent cluster, linearly ordered by the hierarchical cluster clustering.

For comparison, we ran four other methods of tSNE analysis (S4 Fig) using either all genes or the 150 genes found by ProMMT, and either a Euclidean metric or a Euclidean metric after log(1+x) transformation. Perplexity of 15 was again used and initialization was the same as before. Using all genes gave results that were difficult to interpret, particularly for log-transformed data, which we ascribe the noise arising from the large number of weakly expressed genes in the database. Using the gene subset provided more interpretable results, and combining the gene subset with log(1+x) transformation yielded results similar to nbtSNE, while a Euclidean metric yielded a less clear distinction of isolated classes such as *Cck.Lypd1* and *Sst. Nos1*. We conclude that the alignment of nbtSNE to the probability distribution of RNA counts allows the algorithm to take into account differences between weakly expressed genes, and that a log(1+x) transformation approximates this probability distribution. We also conclude that gene subsetting prevents noise from the large number of genes that do not differ between classes from swamping the signal, and that this is particularly important with algorithms sensitive to changes in weakly expressed genes. We suggest that nbtSNE provides a

principled probabilistic method for choosing the transformation and gene subset required for informative tSNE analysis.

## Negative binomial discriminant analysis

To investigate whether a pair of clusters were discretely separated or tiled a continuum we developed a method of cross-validated negative binomial discriminant analysis. This analysis assesses the separation of two clusters $k_1$ and $k_2$ by computing the log likelihood ratio for each cell to belong to the two clusters. It is simple to show that this ratio for a cell $c$ is given by

$$\Delta_c = \sum_{g \in G} x_{c,g}(\log p_{g,k_1} - \log p_{g,k_2}) + r(\log(1 - p_{g,k_1}) - \log(1 - p_{g,k_2}))$$

The sum runs over all genes $g$ in the database, not just the set $S$ found by the ProMMT algorithm.

The degree to which clusters $k_1$ and $k_2$ are discrete is visible by the bimodality of the histogram of $\Delta_c$, which can be quantified using a d' statistic, $\frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}$, where $\mu_i$ and $\sigma_i$ represent the mean and standard deviation of $\Delta_c$ for cells arising in cluster $i$. In this analysis, it is essential that the ratios $\Delta_c$ are computed on a separate "test set" of cells to the "training set" used to estimate $p_{g,k}$, otherwise even a random division of a single homogeneous cluster would give an apparently bimodal histogram because of overfitting.

## Latent factor analysis

To model continuous variation between cells, we used a negative binomial latent factor model. The model is parametrized by two matrices, **W** and **F** of size $N_{genes} \times N_{factors}$ and $N_{factors} \times N_{cells}$. The distribution of each cell follows a negative binomial distribution with mean $\mu_{gc} = r \exp(\Sigma_f W_{gf} F_{fc})$:

$$\Pr(x_{gc}; \mathbf{W}, \mathbf{F}) = NB\left(x_{gc}; r\exp\left(\sum_f W_{gf} F_{fc}\right), r\right)$$

This corresponds to the natural parameterization of the negative binomial, $p = 1 / (1 + \exp(\Sigma_f W_{gf} F_{fc}))$. As usual, we take a fixed value of $r = 2$. For the analysis described in this study, we use only a single latent factor, but add a second column to **F** of all ones to allow the mean expression level to vary between genes.

Given a dataset $x_{gc}$, we fit the matrices **W** and **F** by maximum likelihood. As the negative binomial distribution with fixed $r$ belongs to the exponential family, we can use the simple alternating method of Collins and colleagues [83]. Note that we do not require a sparse algorithm because (unlike in clustering), the number of parameters is fixed. However, to avoid instability, only genes that have reasonable expression levels in the database are kept (genes are included if at least 10 cells express at least 5 copies of the RNA), and a quadratic regularization penalty $-50[|\mathbf{W}|^2 + |\boldsymbol{F}|^2]$ is added to the log likelihood.

To relate the correlations of each gene with the latent factor to their GO categories (S2 Table), we used the MGI mouse GO database (downloaded 2 April 2018), accessed via MATLAB's bioinformatics toolbox. An enrichment score was computed for each GO term by summing the Spearman rank correlations of gene expression with the latent factor, over all genes annotated with that term.

## Supporting information

**S1 Fig. No major difference was observed in cell type distribution between the four samples collected.** Each plot shows the location of all cells collected in a sample on the nbtSNE plots. Top row: mice of age p60; bottom row, age p27. Statistical testing of homogeneity of cluster IDs between mice yielded a weakly significant overall result ($p = 3 \times 10^{-5}$; $\chi^2$ test), but post hoc analysis (individual 2×2 contingency tables, Bonferroni corrected) did not identify any clusters with individually significant differences between mice. We suspect that the weakly significant difference might reflect slightly differential dissection of layers in the four samples. nbtSNE, negative binomial t-stochastic neighbor embedding.
(TIF)

**S2 Fig. Detected RNA counts can be fit by a negative binomial distribution.** (**A**) Standard deviation versus mean expression in a population of CA1 pyramidal cells (Zeisel et al. 2015). Each point represents a single gene. Red curve shows prediction of negative binomial distribution with $r = 2$. (**B**) Example negative binomial probability distribution. Low probability is assigned to very small counts, but for larger counts, there is no strong dependence on precise count value. (**C**) Symmetrized Kullback-Liebler divergence between two negative binomial distributions, as a function of their mean values. High values (indicating poor fit) are obtained when one mean is close to zero, and the other is large; when both means are far from zero, their precise values do not have a major effect.
(TIF)

**S3 Fig. Pseudocolor representation showing mean expression levels of the 150 genes selected by the ProMMT algorithm in each cluster (unnormalized, log scale).** ProMMT, Probabilistic Mixture Modeling for Transcriptomics.
(TIF)

**S4 Fig. Four alternative methods of tSNE analysis.** Left columns show results when all genes are used; right columns show results with the 150 genes selected by ProMMT. Top row shows a Euclidean metric; bottom row shows Euclidean metric after log(1+x) transformation. All methods were initialized from the same starting point as nbtSNE. Of the four methods, only the log(1+x) transformed data with gene subset gave comparable results to nbtSNE. This indicates that the primary effect of the negative binomial distribution is to downweight differences in expression between strongly expressed genes, similarly to the log(1+x) transformation, and that gene subsetting produces more interpretable results whether or not transformation is used. nbtSNE, negative binomial t-stochastic neighbor embedding; ProMMT, Probabilistic Mixture Modeling for Transcriptomics; tSNE, t-stochastic neighbor embedding.
(TIF)

**S5 Fig. Analysis of isocortical interneurons.** (**A**) nbtSNE algorithm applied to 761 interneurons of mouse V1, from Tasic and colleagues (2016). Symbols indicate 23 clusters assigned by Tasic and colleagues. (**B**) Same data, with symbols representing 30 clusters assigned by ProMMT algorithm. (**C**) Confusion matrix relating cluster assignments made by the two algorithms. Right; cell classes identified with ProMMT clusters (black), and genes used to make the identification (red: expressed, blue: not expressed). (**D**) Reprint of figure from Cadwell and colleagues (2016) showing expression of selected genes in layer 1 SBCs and eNGCs. (**E**) Scatterplot matrix showing expression of these genes in Tasic and colleagues' data support the identification of SBCs made by ProMMT algorithm. eNGC, elongated neurogliaform cell; nbtSNE, negative binomial t-stochastic neighbor embedding; ProMMT, Probabilistic Mixture

Modeling for Transcriptomics; SBC, single-bouquet cell.
(TIF)

**S6 Fig. Additional cluster divisions found by the ProMMT algorithm in the data of Tasic and colleagues (2016).** Left and right panels show scatterplot matrices for sets of genes with near-exclusive expression in further subdivisions of the Vip Parm1 and Sst Cbln4 clusters. Red and green points indicate which subcluster the cell was placed in by the ProMMT algorithm. ProMMT, Probabilistic Mixture Modeling for Transcriptomics.
(TIF)

**S7 Fig. Number of detected clusters increases linearly with cell count and read depth.** (**A**) To investigate how the number of detected clusters might change with the number of cells analyzed, we reclustered random subsets of different numbers of cells. The number of clusters identified increased with cell count. (**B**) To investigate how the number of detected clusters might change with read depth, we resampled reads independently for each cell and gene, following a binomial distribution with probability between 0 and 1. Again, cluster count increased linearly with read depth; although a marginally sublinear trend was potentially visible, this was not statistically significant ($p > 0.05$, power-law regression). (**C**) Expected gene count (i.e., mean number of genes with expression >0, averaged over cells in a class), computed as a function of the binomial probability. Color scheme indicated below. (**D,E**) Similar analysis as (A,B) for the data of Tasic and colleagues (2016).
(TIF)

**S8 Fig. Latent factor analysis of isocortical interneurons yields similar results to analysis in CA1.** (**A**) Mean latent factor values differ between cell classes (cf. Fig 6A). Each point represents a cell; x-axis shows latent factor value; y-axis shows original cluster assignments. (**B**) Correlations of genes with the latent factor for isocortical *Pvalb* cells (y-axis) are similar to those of their CA1 counterparts (x-axis; cf. Fig 6D).
(TIF)

**S9 Fig. Activity-dependent modulation of gene expression partially matches variation along genetic latent factor.** Each panel represents a cell type analyzed by Mardinly and colleagues (2016). Within each panel, every point represents a gene, and the y-axis value shows the log ratio of its expression level after 7.5 h of light exposure, compared to dark housing, in the corresponding subtype of visual cortical interneurons. The x-axis value shows that gene's latent factor weighting as determined from our CA1 data. Blue line shows linear regression fit, which was strongest for *Sst* neurons ($r = 0.25$; $p < 10^{-12}$), weaker but significant for *Pvalb* neurons ($r = 0.11$; $p < 0.002$), and insignificant for *Vip* neurons ($r = 0.05$, $p = 0.17$). Only genes of strong mean expression were analyzed (>5,000 normalized reads). A small number of example genes of particular interest are highlighted with red text.
(TIF)

**S10 Fig. Analysis of simultaneously collected pyramidal cells.** The main group of cells exhibited a continuous gradation of gene expression, consistent with previous reports of graded expression of genes such as Wfs1 and Dcn between dorsal and ventral CA1. No cells were identified as CA2 or CA3, as we did not detect populations consistently expressing genes such as Cacng5, Sostdc1, S100b, Ccdc3, Iyd, or Coch. However, three small clusters together containing 62 of the total 357 excitatory neurons were identified as either occurring at the dorsomedial lip of stratum oriens or subiculum, because of their expression of genes such as Tshz2, Tle4, Lxn, and Car10, whose Allen atlas expression patterns are shown at bottom.
(TIF)

**S1 Table. Contains correlation coefficients and significance *p*-values of a set of example genes correlated with the latent factor.**
(XLSX)

**S2 Table. Contains a list of gene ontology terms most correlated with the latent factor.**
(XLSX)

**S1 Text. Contains the in-depth reasoning behind the identifications of our transcriptomic clusters with known cell classes.**
(PDF)

## Acknowledgments

We thank T. Viney, A. Joshi, G. Unal, and B. Bekkouche for discussions and comments on the manuscript.

## Author Contributions

**Conceptualization:** Kenneth D. Harris, Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, Jens Hjerling-Leffler.

**Data curation:** Kenneth D. Harris.

**Formal analysis:** Kenneth D. Harris.

**Funding acquisition:** Kenneth D. Harris, Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, Jens Hjerling-Leffler.

**Investigation:** Kenneth D. Harris, Hannah Hochgerner, Nathan G. Skene, Lorenza Magno, Linda Katona, Carolina Bengtsson Gonzales, Peter Somogyi.

**Methodology:** Kenneth D. Harris.

**Project administration:** Kenneth D. Harris.

**Resources:** Peter Somogyi, Sten Linnarsson.

**Software:** Kenneth D. Harris.

**Supervision:** Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, Jens Hjerling-Leffler.

**Visualization:** Kenneth D. Harris.

**Writing – original draft:** Kenneth D. Harris.

**Writing – review & editing:** Kenneth D. Harris, Hannah Hochgerner, Nathan G. Skene, Lorenza Magno, Linda Katona, Carolina Bengtsson Gonzales, Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, Jens Hjerling-Leffler.

## References

1. Bezaire MJ, Soltesz I. Quantitative assessment of CA1 local circuits: knowledge base for interneuron-pyramidal cell connectivity. Hippocampus. 2013; 23: 751–85. https://doi.org/10.1002/hipo.22141 PMID: 23674373
2. Freund TF, Buzsaki G. Interneurons of the hippocampus. Hippocampus. 1996; 6: 347–470. PMID: 8915675
3. Klausberger T, Somogyi P. Neuronal diversity and temporal dynamics: the unity of hippocampal circuit operations. Science. 2008; 321: 53–7. https://doi.org/10.1126/science.1149381 PMID: 18599766

4. Pelkey KA, Chittajallu R, Craig MT, Tricoire L, Wester JC, McBain CJ. Hippocampal GABAergic Inhibitory Interneurons. Physiol Rev. 2017; 97: 1619–1747. https://doi.org/10.1152/physrev.00007.2017 PMID: 28954853
5. Somogyi P. Hippocampus: intrinsic organization. In: Shepherd GM, Grillner S, editors. Handbook of Brain Microcircuits. New York: Oxford University Press; 2010.
6. Wheeler DW, White CM, Rees CL, Komendantov AO, Hamilton DJ, Ascoli GA. Hippocampome.org: a knowledge base of neuron types in the rodent hippocampus. Elife. 2015; 4. https://doi.org/10.7554/eLife.09960 PMID: 26402459
7. Buhl EH, Halasy K, Somogyi P. Diverse sources of hippocampal unitary inhibitory postsynaptic potentials and the number of synaptic release sites [see comments] [published erratum appears in Nature 1997 May 1;387(6628):106]. Nature. 1994; 368: 823–8. https://doi.org/10.1038/368823a0 PMID: 8159242
8. Hu H, Gan J, Jonas P. Interneurons. Fast-spiking, parvalbumin$^{+}$ GABAergic interneurons: from cellular design to microcircuit function. Science. 2014; 345: 1255263. https://doi.org/10.1126/science.1255263 PMID: 25082707
9. Katona L, Lapray D, Viney TJ, Oulhaj A, Borhegyi Z, Micklem BR, et al. Sleep and movement differentiates actions of two types of somatostatin-expressing GABAergic interneuron in rat hippocampus. Neuron. 2014; 82: 872–86. https://doi.org/10.1016/j.neuron.2014.04.007 PMID: 24794095
10. Fuentealba P, Tomioka R, Dalezios Y, Marton LF, Studer M, Rockland K, et al. Rhythmically active enkephalin-expressing GABAergic cells in the CA1 area of the hippocampus project to the subiculum and preferentially innervate interneurons. J Neurosci. 2008; 28: 10017–22. https://doi.org/10.1523/JNEUROSCI.2052-08.2008 PMID: 18829959
11. Jinno S. Structural organization of long-range GABAergic projection system of the hippocampus. Front Neuroanat. 2009; 3: 13. https://doi.org/10.3389/neuro.05.013.2009 PMID: 19649167
12. Jinno S, Klausberger T, Marton LF, Dalezios Y, Roberts JD, Fuentealba P, et al. Neuronal diversity in GABAergic long-range projections from the hippocampus. J Neurosci. 2007; 27: 8790–804. https://doi.org/10.1523/JNEUROSCI.1847-07.2007 PMID: 17699661
13. Sik A, Ylinen A, Penttonen M, Buzsaki G. Inhibitory CA1-CA3-hilar region feedback in the hippocampus. Science. 1994; 265: 1722–4. PMID: 8085161
14. Takács VT, Freund TF, Gulyás AI. Types and synaptic connections of hippocampal inhibitory neurons reciprocally connected with the medial septum. Eur J Neurosci. 2008; 28: 148–164. https://doi.org/10.1111/j.1460-9568.2008.06319.x PMID: 18662340
15. Daw MI, Tricoire L, Erdelyi F, Szabo G, McBain CJ. Asynchronous transmitter release from cholecystokinin-containing inhibitory interneurons is widespread and target-cell independent. J Neurosci. 2009; 29: 11112–22. https://doi.org/10.1523/JNEUROSCI.5760-08.2009 PMID: 19741117
16. Hefft S, Jonas P. Asynchronous GABA release generates long-lasting inhibition at a hippocampal interneuron–principal neuron synapse. Nat Neurosci. 2005; 8: 1319–1328. https://doi.org/10.1038/nn1542 PMID: 16158066
17. Armstrong C, Soltesz I. Basket cell dichotomy in microcircuit function: Basket cells as dichotomous microcircuit modulators. J Physiol. 2012; 590: 683–694. https://doi.org/10.1113/jphysiol.2011.223669 PMID: 22199164
18. Cope DW, Maccaferri G, Marton LF, Roberts JD, Cobden PM, Somogyi P. Cholecystokinin-immunopositive basket and Schaffer collateral-associated interneurones target different domains of pyramidal cells in the CA1 area of the rat hippocampus. Neuroscience. 2002; 109: 63–80. PMID: 11784700
19. Klausberger T, Marton LF, O'Neill J, Huck JH, Dalezios Y, Fuentealba P, et al. Complementary roles of cholecystokinin- and parvalbumin-expressing GABAergic neurons in hippocampal network oscillations. J Neurosci. 2005; 25: 9782–93. https://doi.org/10.1523/JNEUROSCI.3269-05.2005 PMID: 16237182
20. Pawelzik H, Hughes DI, Thomson AM. Physiological and morphological diversity of immunocytochemically defined parvalbumin- and cholecystokinin-positive interneurones in CA1 of the adult rat hippocampus. J Comp Neurol. 2002; 443: 346–367. https://doi.org/10.1002/cne.10118 PMID: 11807843
21. Somogyi J, Baude A, Omori Y, Shimizu H, El Mestikawy S, Fukaya M, et al. GABAergic basket cells expressing cholecystokinin contain vesicular glutamate transporter type 3 (VGLUT3) in their synaptic terminals in hippocampus and isocortex of the rat. Eur J Neurosci. 2004; 19: 552–69. PMID: 14984406
22. Armstrong C, Krook-Magnuson E, Soltesz I. Neurogliaform and Ivy Cells: A Major Family of nNOS Expressing GABAergic Neurons. Front Neural Circuits. 2012; 6. https://doi.org/10.3389/fncir.2012.00023 PMID: 22623913
23. Fuentealba P, Begum R, Capogna M, Jinno S, Marton LF, Csicsvari J, et al. Ivy cells: a population of nitric-oxide-producing, slow-spiking GABAergic neurons and their involvement in hippocampal network activity. Neuron. 2008; 57: 917–29. https://doi.org/10.1016/j.neuron.2008.01.034 PMID: 18367092

24. Acsady L, Arabadzisz D, Freund TF. Correlated morphological and neurochemical features identify different subsets of vasoactive intestinal polypeptide-immunoreactive interneurons in rat hippocampus. Neuroscience. 1996; 73: 299–315. PMID: 8783251

25. Acsady L, Gorcs TJ, Freund TF. Different populations of vasoactive intestinal polypeptide-immunoreactive interneurons are specialized to control pyramidal cells or interneurons in the hippocampus. Neuroscience. 1996; 73: 317–34. PMID: 8783252

26. Gulyás AI, Hájos N, Freund TF. Interneurons containing calretinin are specialized to control other interneurons in the rat hippocampus. J Neurosci Off J Soc Neurosci. 1996; 16: 3397–3411.

27. Jinno S, Kosaka T. Patterns of colocalization of neuronal nitric oxide synthase and somatostatin-like immunoreactivity in the mouse hippocampus: quantitative analysis with optical disector. Neuroscience. 2004; 124: 797–808. https://doi.org/10.1016/j.neuroscience.2004.01.027 PMID: 15026120

28. Katona L, Micklem B, Borhegyi Z, Swiejkowski DA, Valenti O, Viney TJ, et al. Behavior-dependent activity patterns of GABAergic long-range projecting neurons in the rat hippocampus. Hippocampus. 2017; 27: 359–377. https://doi.org/10.1002/hipo.22696 PMID: 27997999

29. Markram H, Toledo-Rodriguez M, Wang Y, Gupta A, Silberberg G, Wu C. Interneurons of the neocortical inhibitory system. NatRevNeurosci. 2004; 5: 793–807.

30. Parra P, Gulyas AI, Miles R. How many subtypes of inhibitory cells in the hippocampus? Neuron. 1998; 20: 983–93. PMID: 9620702

31. Dehorter N, Ciceri G, Bartolini G, Lim L, del Pino I, Marin O. Tuning of fast-spiking interneuron properties by an activity-dependent transcriptional switch. Science. 2015; 349: 1216–20. https://doi.org/10.1126/science.aab3415 PMID: 26359400

32. Donato F, Rompani SB, Caroni P. Parvalbumin-expressing basket-cell network plasticity induced by experience regulates adult learning. Nature. 2013; 504: 272–6. https://doi.org/10.1038/nature12866 PMID: 24336286

33. Mardinly AR, Spiegel I, Patrizi A, Centofante E, Bazinet JE, Tzeng CP, et al. Sensory experience regulates cortical inhibition by inducing IGF1 in VIP neurons. Nature. 2016; 531: 371–375. https://doi.org/10.1038/nature17187 PMID: 26958833

34. Spiegel I, Mardinly AR, Gabel HW, Bazinet JE, Couch CH, Tzeng CP, et al. Npas4 regulates excitatory-inhibitory balance within neural circuits through cell-type-specific gene programs. Cell. 2014; 157: 1216–1229. https://doi.org/10.1016/j.cell.2014.03.058 PMID: 24855953

35. Cembrowski MS, Wang L, Sugino K, Shields BC, Spruston N. Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. eLife. 2016; 5: e14997. https://doi.org/10.7554/eLife.14997 PMID: 27113915

36. Cembrowski MS, Bachman JL, Wang L, Sugino K, Shields BC, Spruston N. Spatial Gene-Expression Gradients Underlie Prominent Heterogeneity of CA1 Pyramidal Neurons. Neuron. 2016; 89: 351–368. https://doi.org/10.1016/j.neuron.2015.12.013 PMID: 26777276

37. Chevée M, Robertson JD, Cannon GH, Brown SP, Goff LA. Variation in neuronal activity state, axonal projection target, and position principally define the transcriptional identity of individual neocortical projection neurons. bioRxiv. 2017; 157149. https://doi.org/10.1101/157149

38. Ecker JR, Geschwind DH, Kriegstein AR, Ngai J, Osten P, Polioudakis D, et al. The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. Neuron. 2017; 96: 542–557. https://doi.org/10.1016/j.neuron.2017.10.007 PMID: 29096072

39. Frazer S, Prados J, Niquille M, Cadilhac C, Markopoulos F, Gomez L, et al. Transcriptomic and anatomic parcellation of 5-$HT_{3A}$R expressing cortical interneuron subtypes revealed by single-cell RNA sequencing. Nat Commun. 2017; 8: 14219. https://doi.org/10.1038/ncomms14219 PMID: 28134272

40. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta JJ, et al. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science. 2016; 353: 925–928. https://doi.org/10.1126/science.aad7038 PMID: 27471252

41. Habib N, Basu A, Avraham-Davidi I, Burks T, Choudhury SR, Aguet F, et al. DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. bioRxiv. 2017; https://doi.org/10.1101/115196

42. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015; 161: 1202–14. https://doi.org/10.1016/j.cell.2015.05.002 PMID: 26000488

43. Paul A, Crow M, Raudales R, Gillis J, Huang ZJ. Transcriptional Architecture of Synaptic Communication Delineates Cortical GABAergic Neuron Identity. bioRxiv. 2017; 180034. https://doi.org/10.1101/180034

44. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016; 19: 335–346. https://doi.org/10.1038/nn.4216 PMID: 26727548

45. Usoskin D, Furlan A, Islam S, Abdo H, Lonnerberg P, Lou D, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 2015; 18: 145–53. https://doi.org/10.1038/nn.3881 PMID: 25420068

46. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015; 347: 1138–42. https://doi.org/10.1126/science.aaa1934 PMID: 25700174

47. Lu J, Tomfohr JK, Kepler TB. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. BMC Bioinformatics. 2005; 6: 165. https://doi.org/10.1186/1471-2105-6-165 PMID: 15987513

48. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. 2008; 9: 321–332. https://doi.org/10.1093/biostatistics/kxm030 PMID: 17728317

49. Bouveyron C, Brunet-Saumard C. Model-based clustering of high-dimensional data: A review. Comput Stat Data Anal. 2014; 71: 52–78. https://doi.org/10.1016/j.csda.2012.12.008

50. Schwarz G. Estimating the Dimension of a Model. Ann Stat. 1978; 6: 461–464. https://doi.org/10.1214/aos/1176344136

51. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008; 9: 2579–2605.

52. Acsády L, Pascual M, Rocamora N, Soriano E, Freund TF. Nerve growth factor but not neurotrophin-3 is synthesized by hippocampal GABAergic neurons that project to the medial septum. Neuroscience. 2000; 98: 23–31. PMID: 10858608

53. Losonczy A, Zhang L, Shigemoto R, Somogyi P, Nusser Z. Cell type dependence and variability in the short-term plasticity of EPSCs in identified mouse hippocampal interneurones. J Physiol. 2002; 542: 193–210. https://doi.org/10.1113/jphysiol.2002.020024 PMID: 12096061

54. Viney TJ, Lasztoczi B, Katona L, Crump MG, Tukker JJ, Klausberger T, et al. Network state-dependent inhibition of identified hippocampal CA3 axo-axonic cells in vivo. Nat Neurosci. 2013; 16: 1802–1811. https://doi.org/10.1038/nn.3550 PMID: 24141313

55. Vruwink M, Schmidt HH, Weinberg RJ, Burette A. Substance P and nitric oxide signaling in cerebral cortex: anatomical evidence for reciprocal signaling between two classes of interneurons. J Comp Neurol. 2001; 441: 288–301. PMID: 11745651

56. Tricoire L, Pelkey KA, Daw MI, Sousa VH, Miyoshi G, Jeffries B, et al. Common origins of hippocampal Ivy and nitric oxide synthase expressing neurogliaform cells. J Neurosci. 2010; 30: 2165–76. https://doi.org/10.1523/JNEUROSCI.5123-09.2010 PMID: 20147544

57. Ferraguti F, Klausberger T, Cobden P, Baude A, Roberts JD, Szucs P, et al. Metabotropic glutamate receptor 8-expressing nerve terminals target subsets of GABAergic neurons in the hippocampus. J Neurosci. 2005; 25: 10520–36. https://doi.org/10.1523/JNEUROSCI.2547-05.2005 PMID: 16280590

58. Miyashita T, Rockland KS. GABAergic projections from the hippocampus to the retrosplenial cortex in the rat. Eur J Neurosci. 2007; 26: 1193–204. https://doi.org/10.1111/j.1460-9568.2007.05745.x PMID: 17767498

59. Tricoire L, Pelkey KA, Erkkila BE, Jeffries BW, Yuan X, McBain CJ. A blueprint for the spatiotemporal origins of mouse hippocampal interneuron diversity. J Neurosci. 2011; 31: 10948–70. https://doi.org/10.1523/JNEUROSCI.0323-11.2011 PMID: 21795545

60. Tyan L, Chamberland S, Magnin E, Camire O, Francavilla R, David LS, et al. Dendritic inhibition provided by interneuron-specific cells controls the firing rate and timing of the hippocampal feedback inhibitory circuitry. J Neurosci. 2014; 34: 4534–47. https://doi.org/10.1523/JNEUROSCI.3813-13.2014 PMID: 24671999

61. Blasco-Ibanez JM, Martinez-Guijarro FJ, Freund TF. Enkephalin-containing interneurons are specialized to innervate other interneurons in the hippocampal CA1 region of the rat and guinea-pig. Eur J Neurosci. 1998; 10: 1784–95. PMID: 9751150

62. Jiang X, Wang G, Lee AJ, Stornetta RL, Zhu JJ. The organization of two new cortical interneuronal circuits. Nat Neurosci. 2013; 16: 210–8. https://doi.org/10.1038/nn.3305 PMID: 23313910

63. Jiang X, Shen S, Cadwell CR, Berens P, Sinz F, Ecker AS, et al. Principles of connectivity among morphologically defined cell types in adult neocortex. Science. 2015; 350: aac9462. https://doi.org/10.1126/science.aac9462 PMID: 26612957

64. Cadwell CR, Palasantza A, Jiang X, Berens P, Deng Q, Yilmaz M, et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. Nat Biotechnol. 2016; 34: 199–203. https://doi.org/10.1038/nbt.3445 PMID: 26689543

65. Gulyás AI, Freund TF. Pyramidal cell dendrites are the primary targets of calbindin D28k-immunoreactive interneurons in the hippocampus. Hippocampus. 1996; 6: 525–534. PMID: 8953305

66. Dudok B, Barna L, Ledri M, Szabó SI, Szabadits E, Pintér B, et al. Cell-specific STORM super-resolution imaging reveals nanoscale organization of cannabinoid signaling. Nat Neurosci. 2015; 18: 75–86. https://doi.org/10.1038/nn.3892 PMID: 25485758

67. Lee S-H, Földy C, Soltesz I. Distinct endocannabinoid control of GABA release at perisomatic and dendritic synapses in the hippocampus. J Neurosci Off J Soc Neurosci. 2010; 30: 7993–8000. https://doi.org/10.1523/JNEUROSCI.6238-09.2010 PMID: 20534847

68. Cho J, Yu N-K, Choi J-H, Sim S-E, Kang SJ, Kwak C, et al. Multiple repressive mechanisms in the hippocampus during memory formation. Science. 2015; 350: 82–87. https://doi.org/10.1126/science.aac7368 PMID: 26430118

69. Cohen SM, Ma H, Kuchibhotla KV, Watson BO, Buzsáki G, Froemke RC, et al. Excitation-Transcription Coupling in Parvalbumin-Positive Interneurons Employs a Novel CaM Kinase-Dependent Pathway Distinct from Excitatory Neurons. Neuron. 2016; 90: 292–307. https://doi.org/10.1016/j.neuron.2016.03.001 PMID: 27041500

70. Gerashchenko D, Wisor JP, Burns D, Reh RK, Shiromani PJ, Sakurai T, et al. Identification of a population of sleep-active cerebral cortex neurons. Proc Natl Acad Sci U S A. 2008; 105: 10227–10232. https://doi.org/10.1073/pnas.0803125105 PMID: 18645184

71. Magno L, Oliveira MG, Mucha M, Rubin AN, Kessaris N. Multiple embryonic origins of nitric oxide synthase-expressing GABAergic neurons of the neocortex. Front Neural Circuits. 2012; 6: 65. https://doi.org/10.3389/fncir.2012.00065 PMID: 23015780

72. Price CJ, Cauli B, Kovacs ER, Kulik A, Lambolez B, Shigemoto R, et al. Neurogliaform neurons form a novel inhibitory network in the hippocampal CA1 area. J Neurosci Off J Soc Neurosci. 2005; 25: 6775–6786. https://doi.org/10.1523/JNEUROSCI.1135-05.2005 PMID: 16033887

73. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007; 445: 168–76. https://doi.org/10.1038/nature05453 PMID: 17151600

74. Xu X, Roby KD, Callaway EM. Mouse cortical inhibitory neuron type that coexpresses somatostatin and calretinin. J Comp Neurol. 2006; 499: 144–160. https://doi.org/10.1002/cne.21101 PMID: 16958092

75. Martínez-Guijarro FJ, Briñón JG, Blasco-Ibáñez JM, Okazaki K, Hidaka H, Alonso JR. Neurocalcin-immunoreactive cells in the rat hippocampus are GABAergic interneurons. Hippocampus. 1998; 8: 2–23. PMID: 9580316

76. Chittajallu R, Craig MT, McFarland A, Yuan X, Gerfen S, Tricoire L, et al. Dual origins of functionally distinct O-LM interneurons revealed by differential 5-HT(3A)R expression. Nat Neurosci. 2013; 16: 1598–607. https://doi.org/10.1038/nn.3538 PMID: 24097043

77. Gulyás AI, Buzsáki G, Freund TF, Hirase H. Populations of hippocampal inhibitory neurons express different levels of cytochrome c. Eur J Neurosci. 2006; 23: 2581–2594. https://doi.org/10.1111/j.1460-9568.2006.04814.x PMID: 16817861

78. Lucas EK, Markwardt SJ, Gupta S, Meador-Woodruff JH, Lin JD, Overstreet-Wadiche L, et al. Parvalbumin Deficiency and GABAergic Dysfunction in Mice Lacking PGC-1α. J Neurosci. 2010; 30: 7227–7235. https://doi.org/10.1523/JNEUROSCI.0698-10.2010 PMID: 20505089

79. Lucas EK, Dougherty SE, McMeekin LJ, Reid CS, Dobrunz LE, West AB, et al. PGC-1α Provides a Transcriptional Framework for Synchronous Neurotransmitter Release from Parvalbumin-Positive Interneurons. J Neurosci. 2014; 34: 14375–14387. https://doi.org/10.1523/JNEUROSCI.1222-14.2014 PMID: 25339750

80. Ogiwara I, Iwasato T, Miyamoto H, Iwata R, Yamagata T, Mazaki E, et al. Nav1.1 haploinsufficiency in excitatory neurons ameliorates seizure-associated sudden death in a mouse model of Dravet syndrome. Hum Mol Genet. 2013; 22: 4784–4804. https://doi.org/10.1093/hmg/ddt331 PMID: 23922229

81. Bishop CM. Pattern Recognition and Machine Learning | Christopher Bishop | Springer verlag; 2006.

82. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B Methodol. 1977; 39: 1–38.

83. Collins M, Dasgupta S, Schapire RE. A generalization of principal component analysis to the exponential family. Advances in Neural Information Processing Systems. MIT Press; 2001.