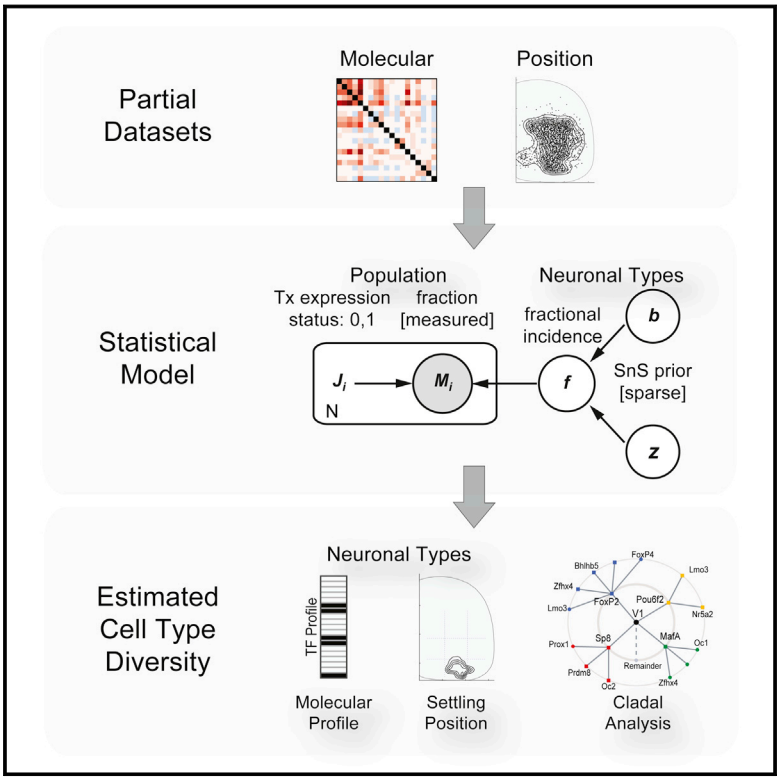
Resource

Cell

Bayesian Sparse Regression Analysis Documents the Diversity of Spinal Inhibitory Interneurons

Graphical Abstract



Authors

Mariano I. Gabitto, Ari Pakman, Jay B. Bikoff, L.F. Abbott, Thomas M. Jessell, Liam Paninski

Correspondence

mig2118@columbia.edu (M.I.G.), liam@stat.columbia.edu (L.P.)

In Brief

A statistical approach based on limited protein expression data and neuronal position provides estimates of cell-type diversity, revealing some 50 discrete types of spinal V1 inhibitory interneurons.

Highlights

- Bayesian framework identifies cell types using a variety of information sources
- Spinal V1 inhibitory interneurons are genetically diverse
- Spinal V1 cell types cluster into localized spatial domains
- Experimental validation confirms model predictions

Gabitto et al., 2016, Cell *165*, 220–233
March 24, 2016
http://dx.doi.org/10.1016/j.cell.2016.01.026

CellPress

Cell

Resource

Bayesian Sparse Regression Analysis Documents the Diversity of Spinal Inhibitory Interneurons

Mariano I. Gabitto,[1,4,5,*] Ari Pakman,[2,5] Jay B. Bikoff,[1,4,5] L.F. Abbott,[1,3] Thomas M. Jessell,[1,4] and Liam Paninski[1,2,*]
[1]Department of Neuroscience, Columbia University, New York, NY 10032, USA
[2]Department of Statistics and Grossman Center for the Statistics of Mind, Columbia University, New York, NY 10027, USA
[3]Department of Physiology and Cellular Biophysics, Columbia University, New York, NY 10032, USA
[4]Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Kavli Institute for Brain Science, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10032, USA
[5]Co-first author
*Correspondence: mig2118@columbia.edu (M.I.G.), liam@stat.columbia.edu (L.P.)
http://dx.doi.org/10.1016/j.cell.2016.01.026

SUMMARY

Documenting the extent of cellular diversity is a critical step in defining the functional organization of tissues and organs. To infer cell-type diversity from partial or incomplete transcription factor expression data, we devised a sparse Bayesian framework that is able to handle estimation uncertainty and can incorporate diverse cellular characteristics to optimize experimental design. Focusing on spinal V1 inhibitory interneurons, for which the spatial expression of 19 transcription factors has been mapped, we infer the existence of ~50 candidate V1 neuronal types, many of which localize in compact spatial domains in the ventral spinal cord. We have validated the existence of inferred cell types by direct experimental measurement, establishing this Bayesian framework as an effective platform for cell-type characterization in the nervous system and elsewhere.

INTRODUCTION

Tissues and organs are comprised of diverse cell types, possessing characteristic morphology and specialized function. The diversification of cell types attains prominence in the nervous system, where neuronal distinctions depend on the activities of transcription factors (TFs) and their downstream effectors (Kohwi and Doe, 2013). Attempts to define the link between transcriptional identity and neuronal diversity have benefitted from the analysis of long-distance projection neurons, for which distinctions in target innervation provide a clear correlate of functional divergence (Molyneaux et al., 2007; Sanes and Masland, 2015). In the retina and cerebral cortex, functional subclasses of ganglion and pyramidal neurons have been delineated through their transcriptional identities (Siegert et al., 2009; Greig et al., 2013). Similarly, the hierarchical ordering of motor neuron subtypes in the spinal cord has its origins in discrete profiles of transcription factor expression (Dasen et al., 2005; Dasen and Jessell, 2009). Yet, local interneurons represent by far the most prevalent neurons within the mammalian CNS, collectively shaping the output of long-range-projection neurons (Isaacson and Scanziani, 2011). The local confinement of interneuron axons, however, has made it difficult to obtain objective measures of identity and diversity.

Genome-wide mRNA expression profiles have been informative in distinguishing neuronal cell types (Usoskin et al., 2015; Zeisel et al., 2015; Macosko et al., 2015; Tasic et al., 2016). Nevertheless, documented dissociations between mRNA and protein expression (Gygi et al., 1999; Vogel and Marcotte, 2012) emphasize the merits of analysis of protein expression at the level of individual neurons (Sharma et al., 2015). But if many genes are involved in defining individual subpopulations, then the validation of protein co-expression will be constrained by the limited repertoire of primary and secondary antibodies.

This practical limitation could be overcome through the development of a statistical method that is able to resolve the extent of neuronal diversity from sparsely sampled transcriptional datasets. Such a method should provide: (1) an objective measure of confidence in the existence of cell types and their prevalence within a parental population, (2) improvement in estimation accuracy upon integrating independent cellular characteristics with molecular phenotype, and (3) informative predictions to guide further experiments. To meet these goals, we developed a sparse Bayesian framework that models co-expression data based on incomplete combinations of TFs. Our focus on TF expression was governed by the well-established role of DNA-binding proteins in defining neuronal identity (Dalla Torre di Sanguinetto et al., 2008; Amamoto and Arlotta, 2014).

We used this Bayesian approach to assess the diversity of V1 interneurons in the spinal cord, a major inhibitory interneuron population implicated in motor control (Zhang et al., 2014). V1 interneurons are defined by developmental expression of the homeodomain transcription factor En1 (Saueressig et al., 1999) and include Renshaw cells and Group Ia reciprocal interneurons, which mediate recurrent and reciprocal inhibition, respectively (Sapir et al., 2004; Zhang et al., 2014). Yet these two physiologically defined subtypes represent only a small fraction of the parental V1 population (Alvarez et al., 2005), implying a greater diversity of V1 neurons. Indeed, the V1 population has recently been subdivided on the basis of the expression of 19 TFs (Bikoff et al., 2016 [this issue of *Cell*]).

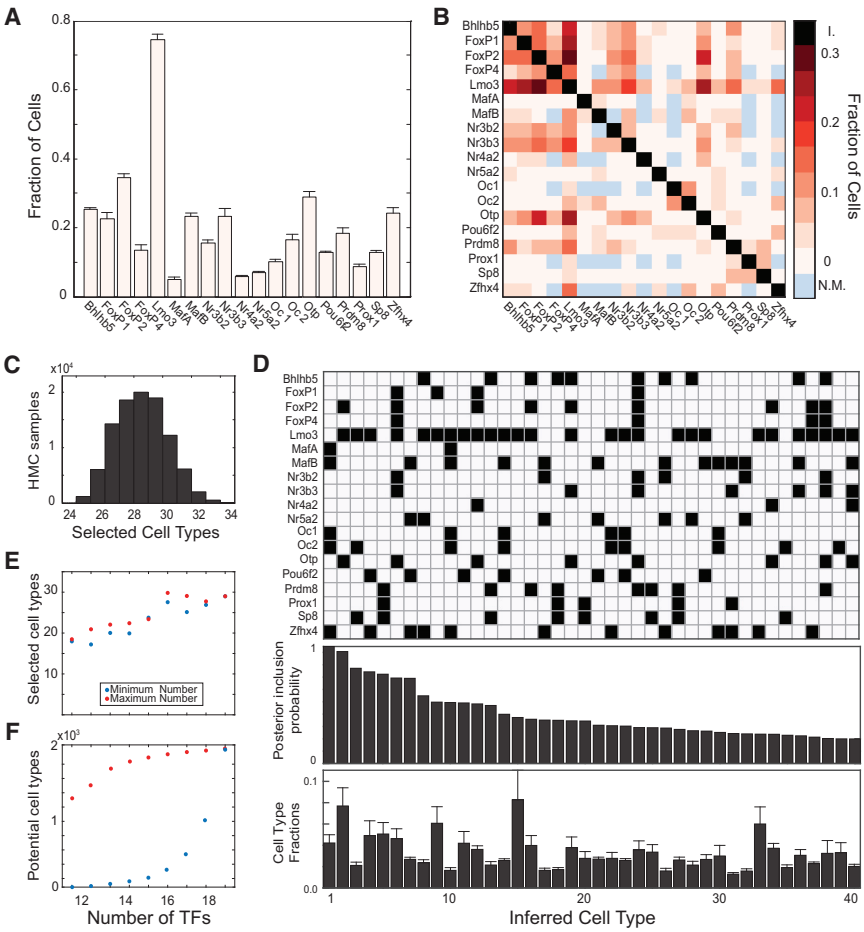


Figure 1. Cell Type Discovery Using Transcription Factor Expression Information

(A) Fraction of V1 interneurons labeled by each of the 19 individual TFs, in p0 lumbar spinal cord. Mean ± SEM, n ≥ 3 animals.

(B) Fraction of V1 interneurons labeled by pairs of TFs. (N.M., not measured). Diagonal values represent identity. Mean ± SEM, n ≥ 3 animals.

(C) Number of cell types selected per HMC iteration (for which the fraction f_k was nonzero).

(D) Transcriptional profiles of top 40 inferred cell types. Cell types (top) are arranged by descending posterior inclusion probability (middle). Black indicates TF expression, white indicates absence of expression. Bottom: fraction of each cell type in the parental V1 population (mean ± SD of all nonzero sampled values).

(E) Number of selected cell types remains close to 29 when varying the set of observed TFs. Red and blue curves denote the maximum and minimum number for different TF sets.

(F) Number of potential cell types. Red and blue curves denote maximum and minimum numbers after reduction by measured TF pairs that exhibit no co-expression.

See also Figure S1.

Here, we apply sparse Bayesian approach to assess the diversity of V1 neuronal subtypes marked by these TFs. Our analysis has generated three major findings: (1) an estimate of the number of V1 interneuron types, on the order of 50 subtypes; (2) an estimate of the expression profiles of the 19 TFs across these inferred types; and (3) a cladistic description of V1 diversity that guides experiments aimed at monitoring and manipulating distinct V1 subpopulations. In several instances, the predicted assignment of V1 neuronal type has been validated through single-cell qRT-PCR, immunohistochemistry, and assessment of spatial distribution. Finally, we demonstrate that this sparse Bayesian analysis serves as an effective platform for identifying cell types within other diverse populations.

RESULTS

A Sparse Bayesian Approach for Uncovering Neuronal Diversity

The companion paper (Bikoff et al., 2016) used comparative microarray screening to identify and map the expression of 19 TFs in V1 interneurons. We used three sets of data to infer V1 neuronal diversity: (1) the fraction of neurons within the parental V1 population that express each of the 19 TFs (Figure 1A), (2) the fractions of neurons co-expressing various pairs of TFs (Figure 1B), and (3) the position of V1 interneurons expressing each of the 19 TFs (Figure 3A). Complete analysis of all TF pairs is hindered by the fact that primary antibodies generated in the same host species cannot be distinguished easily by fluorescently tagged secondary antibodies. We therefore developed an approach that permits statistical inference on the basis of incomplete data.

In this statistical analysis, a cell type is defined by the expression pattern of the 19 TFs under consideration. We characterize TFs as either expressed or not expressed, and thus each expression pattern is specified by a vector of 19 binary numbers, $J_{k,a}$, for pattern k, with a ranging from 1 to 19. $J_{k,a}$ is set to 1 if TF a is expressed in expression pattern k and to 0 if it is not. This results in 2^{19} possible binary expression patterns for the 19 TFs. This large number was reduced by eliminating combinations that include pairs of factors not co-expressed within the same neuron. Analysis of the pairwise expression revealed that 67 out of 148 measured TF pairs fail to co-express (Figure 1B), thereby reducing the possible diversity to 1,978 potential expression patterns (see Statistical Model in the Supplemental Information). Thus, for the variables $J_{k,a}$ specifying the expression patterns, k runs from 1 to 1,978.

We designate the fraction of cells with expression pattern k, the cell-type fraction, denoted by f_k, with k again ranging across all the potential expression patterns (1 to 1,978). Cell-type fractions must be positive ($f_k \geq 0$) and sum to 1 ($\Sigma_k f_k = 1$), indicating that the entire V1 population is accounted for. The fraction of V1 neurons expressing TF a (the data in Figure 1A) is $\Sigma_k f_k J_{k,a}$, and the fraction co-expressing factors a and b (the data in Figure 1B)

is $\Sigma_k f_k J_{k,a} J_{k,b}$ (Supplemental Information). Fitting data within this framework amounts to choosing a set of cell-type fractions that provide a good match to the expression and co-expression data and that satisfy non-negativity and sum-to-one constraints (by the definition of f_k). Under these conditions, the number of non-zero inferred cell-type fractions determines the inferred number of cell types, and the variables $J_{k,a}$ for a = 1, ... 19 and for k values with $f_k \neq 0$, provide candidate expression patterns of these selected cell types.

In principle, the model could be fit to observed data by minimizing the summed squared difference between the measurements and the predictions generated by the inferred fractions. This amounts to a non-negative constrained least-squares (NNCLS) minimization problem (see Experimental Procedures) (Wang et al., 2006; Abbas et al., 2009; Gong et al., 2011; Grange, et al., 2014). But the NNCLS approach fails in this case because, despite the constraint of non-negativity, it generates an infinite number of equally valid solutions. Indeed, for any single presumed cell type it is possible to find alternative solutions that exclude this cell type while maintaining an optimal summed squared difference.

We therefore resorted to a Bayesian approach in which unknown cell-type fractions are modeled as random variables, allowing their uncertainty to be characterized by probability distributions. The use of a prior distribution enables previous knowledge and expectations to be incorporated into the model, and a likelihood function reflects the probability that the observed data were generated by the model. As a biologically plausible prior distribution over cell-type fractions, we chose a constrained "spike-and-slab" (SnS) distribution (Ishwaran and Rao, 2005). This prior incorporates the biologically reasonable assumption that only a small fraction of the 1,978 potential cell types actually exist within the parental V1 population. The SnS prior favors configurations in which only a small subset of the coefficients f_k are non-zero (Supplemental Information).

The use of Bayes' rule to combine prior and data likelihoods results in a posterior distribution from which estimates of confidence about the existence and identity of cell types can be determined. In our case, the posterior distribution cannot be computed directly, necessitating the use of a Monte Carlo sampling method (Gelman et al., 2013). In particular, we adapted a Hamiltonian Monte Carlo (HMC) algorithm to draw random samples from the posterior distribution. This Monte Carlo procedure is specialized for constrained SnS posteriors and permits efficient sampling from our posterior distributions (Pakman and Paninski, 2013, 2014) (Figure S1).

Each iteration of the sampling algorithm generates a set of cell-type fractions that satisfy the constraints and provide a good fit to the data. The number of selected cell types and their expression patterns vary across iterations. Combining samples across a large number of iterations allows us to infer the properties of the posterior probability distribution. For example, the proportion of Monte Carlo samples for which a particular expression pattern is selected determines that type's posterior inclusion probability and serves as a confidence measure of its necessity to explain the data. We also computed the distribution of the number of cell types selected in each iteration, which provides an estimate of the total number of distinct cell types required to explain the observed data. Repeated sampling also enables us to compute cross-correlations between cell-type fractions that are used to construct a list of candidate expression profiles along with the probabilities of their correspondence to actual cell types. As with an expression pattern, a candidate expression profile is a 19-component vector, with each component representing a different TF, but now these components are allowed to be real numbers between 0 and 1. In this scheme, component *a* of the candidate expression profile represents the probability that TF *a* is expressed. Finally, the quantification of uncertainty in the Bayesian approach provides a tool that can enhance experimental design by selecting, in a principled way, the measurement that is expected to maximally reduce uncertainty (see Experimental Design in the Supplemental Information).

We validated the ability of the Bayesian approach to accurately infer cellular diversity by performing computational cross-validation experiments, as well as experiments on simulated datasets, for which the underlying cell types and corresponding cell-type fractions are known (see Computational Validation of the Bayesian Model in the Supplemental Information). This approach provides convincing evidence that meaningful and accurate estimates of cellular diversity can be extracted, encouraging us to apply the Bayesian approach to V1 datasets.

V1 Diversity Extracted Solely from Transcription Factor Expression Data

We first applied this Bayesian framework to TF expression without including spatial information (Figures 1A and 1B, but not 3A). As discussed, each iteration of the HMC sampling algorithm generates a possible set of cell types, but their number and identity vary across HMC iterations. Over the course of the full HMC run, the number of types selected (those with non-zero cell-type fractions) ranged from 25 to 33 with a mean ± SD of 29 ± 2 (Figure 1C). The identity of the selected cell types varied across different HMC iterations.

Computing the posterior inclusion probability of each expression pattern across many samples led to a rank-ordered list of candidate expression patterns. The 40 candidate patterns with the highest inclusion probabilities and their inferred cell-type fractions are shown in Figure 1D. The expression pattern with the highest inclusion probability corresponds to the Renshaw interneuron, a defined V1 neuronal type that mediates recurrent inhibition of motor neurons (Renshaw, 1946) and co-expresses the TFs Oc1, Oc2, and MafB (Stam et al., 2012). This analysis also infers the existence of $MafA^+$ and $MafA^-$ subsets of Renshaw interneurons (patterns 1 and 30 in Figure 1D), a molecular diversity that may correspond to their known morphological heterogeneity (Fyffe, 1990).

We examined the sensitivity of these results to the number of TFs used in the analysis, selecting 11 to 18 of the 19 measured TFs. The average number of selected cell types, 29 for the full 19 factors, decreases only gradually when smaller numbers of TFs are analyzed. Moreover, when 16 to 19 TFs are incorporated, the number of selected cell types remains relatively constant, close to 29 (Figure 1E). In contrast, the number of potential cell types (1,978 for the case of 19 factors) depends much more strongly on the chosen TF subset (Figure 1F). These findings

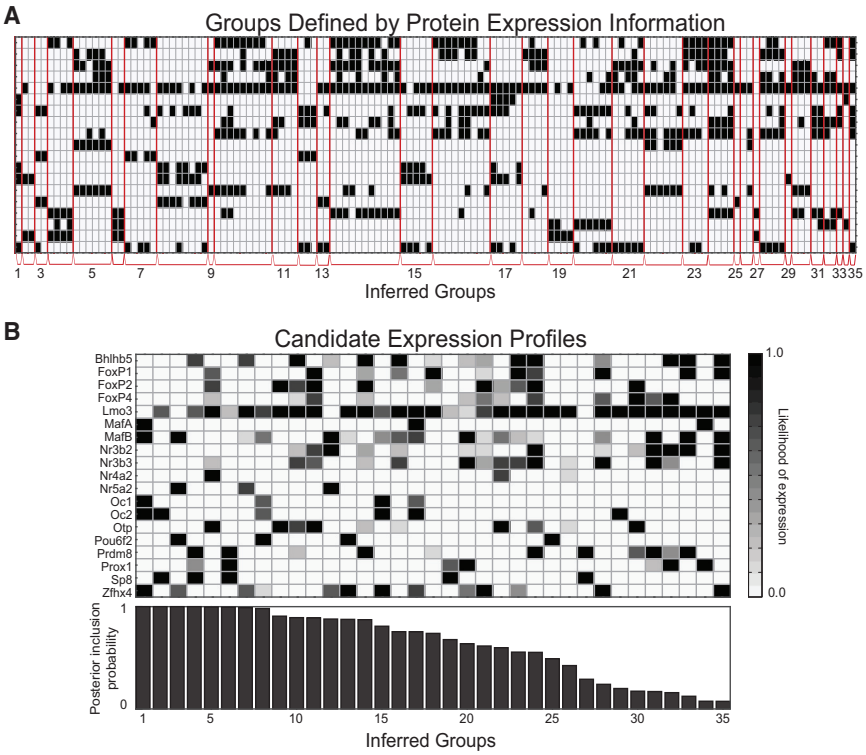


Figure 2. Clustering Algorithm Arranging Cell Types into Correlated Groups

(A) TFs expressed by cell types with a posterior inclusion probability >5%. Inferred group members are arranged in between red lines.

(B) Representation of inferred groups. Top: candidate expression profiles derived from the 35 V1 groups. Gray scale indicates the likelihood of each TF expressed within the group. Bottom: posterior inclusion probability for each V1 group.

See also Figure S2.

suggest that Bayesian calculations of cellular diversity based solely on the available TF data may be close to saturating with the use of a majority subset of the 19 TFs examined here.

What is the diversity of potential TF expression patterns within the V1 population? We detected 131 different expression patterns with posterior inclusion probabilities >0.05 (i.e., appearing in more than 5% of the HMC samples). We constructed candidate expression profiles by clustering the 131 most likely expression patterns into "groups" (Figure 2A; see Clustering Cell Types into Groups in the Supplemental Information). A group is defined as a set of expression patterns that satisfies two conditions: (1) the members of a group express similar sets of TFs (Figure S2A), and (2) in all or almost all of the HMC samples, only a single member of a group is selected (i.e., has a non-zero cell-type fraction), although different members may be selected in different samples (Figure S2B). The second condition causes the members of a group to be negatively correlated with each other (Figure S2A). These conditions permit the interpretation of a group as a single cell type with an uncertain expression pattern.

We developed a recursive algorithm for constructing these groups. All candidate expression profiles with inclusion probabilities >5% were assigned, with most groups having only a single member selected across all of the HMC samples and no group having more than one member selected in >3% of the samples (Figure S2D). To examine the robustness of the inferred groups, we varied systematically the threshold for selecting the list of candidate expression patterns from which groups were constructed. As this threshold is lowered, the number of groups first increases linearly because each high-ranked expression pattern spawns its own group (Figure S2C). However, this growth slows as lower-ranked patterns join existing groups, resulting in a weak dependence on the inclusion threshold. With an inclusion threshold of 5%, the clustering algorithm identifies 35 groups (Figure 2A).

Each group gives rise to a single candidate expression profile (Figure 2B), and for each profile, we assign an expression probability to each TF, weighting the binary expression patterns of each member of the group by the frequency with which it appears in the HMC samples (Figure 2B, top). In addition, we compute a posterior inclusion probability (Figure 2B, bottom) for each candidate expression profile. These are much higher than the inclusion probabilities of the corresponding candidate expression patterns from which they are constructed (Figure 1D). Nevertheless, there is still considerable uncertainty in the identity of the cell types predicted by TF expression data alone (Figure 2B). Furthermore, the existence of 131 candidates for only ~30 cell types (the average number per sample; Figure 1C), and the fact that few of the top expression patterns in Figure 1D have posterior inclusion probabilities near one both indicate that expression-only data is insufficient for specifying cell-type identity (Figure S3).

Incorporating Spatial Information Reveals Further V1 Interneuron Diversity

Certain classes of spinal interneurons are known to localize in discrete spatial domains (Thomas and Wilson, 1965; Hultborn et al., 1971), prompting us to ask whether the incorporation of spatial information could refine estimates of V1 group diversity. For spatial analysis, we divided the ventral spinal cord into 196 bins, mapped spatial expression data into specific bins, and defined cell-type fractions for each bin (Figure 3A). We then applied an appropriately generalized Bayesian analysis to the spatially resolved data (see Supplemental Information, Sections 2.1 and 3.1).

Incorporating spatial information into the Bayesian analysis increased the number of candidate cell types that the HMC sampler selected per iteration from 29 to 50 ± 2 (mean ± SD); and the degree of confidence in the inferred expression profiles also increased (Figure 3B). With the addition of spatial

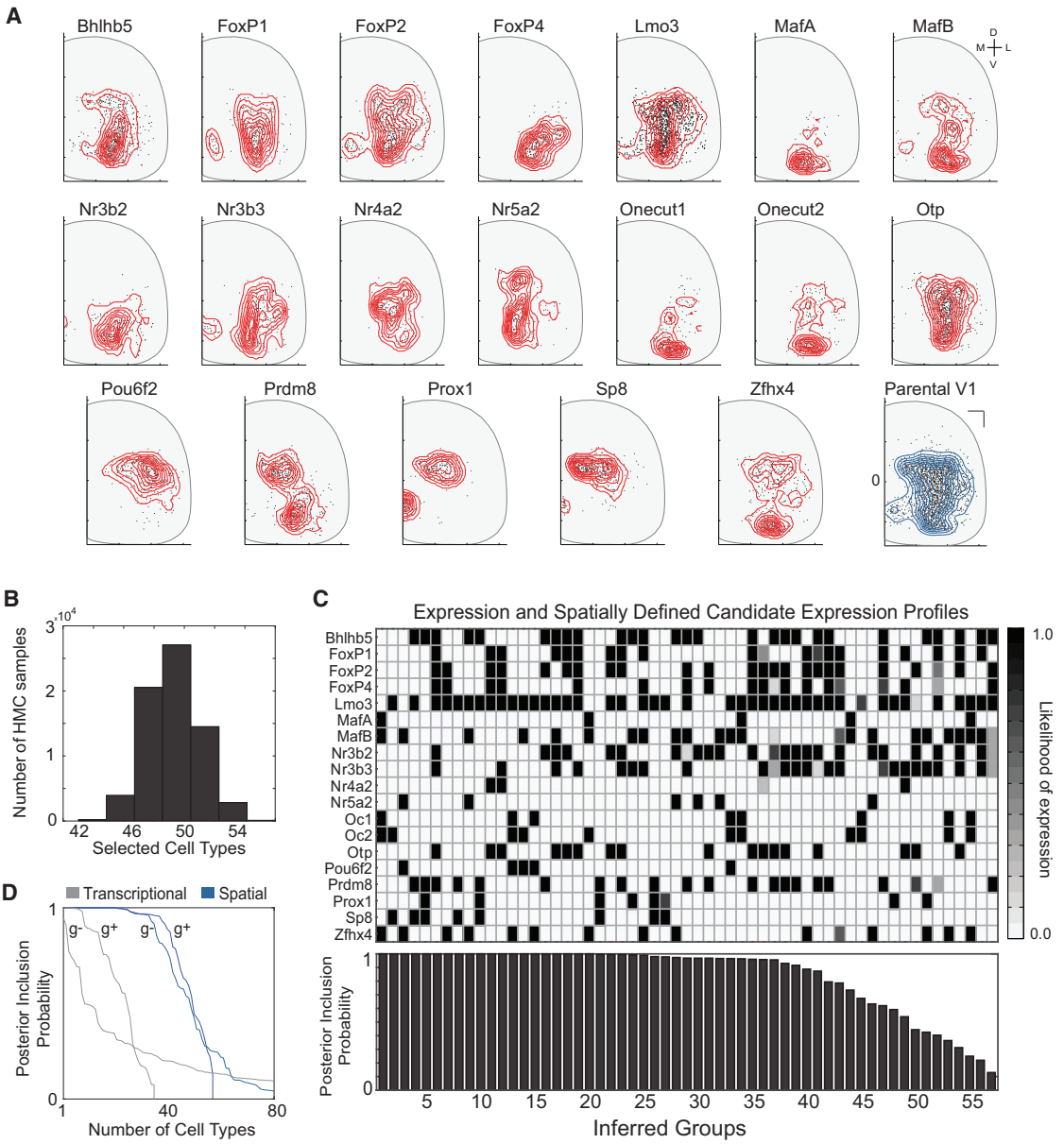


Figure 3. Cell Types Revealed by Incorporating Transcription Factor Spatial Information

(A) Spatial distributions for each of the 19 TF V1 subpopulations. Black dots represent cell positions. Red contours represent kernel density estimates, based on a 14 × 14 grid (a.u.). Scale bar, 100 μm.

(B) Number of selected cell types in each HMC iteration (cf. Figure 1C).

(C) Condensed representation of candidate expression profiles of 57 V1 groups. Gray scale indicates the likelihood that each TF is expressed within the group (as in Figure 2B). Bottom: posterior inclusion probability for each V1 group.

(D) Posterior inclusion probability for expression-inferred cell types and groups (gray) and expression-and-spatially inferred cell types and groups (blue); "g^+" indicates groups, and "g^-" indicates cell types.

See also Figure S3.

information, only 75 total expression patterns are assigned posterior inclusion probabilities >0.05 (compared to 131 for the non-spatial analysis), and many of their inclusion probabilities are close to one (Figure S2F; cf. Figure 1D). We repeated the grouping procedure for these 75 total cell types and uncovered 57 candidate expression profiles, most of which are identified with high inclusion probabilities (Figure 3C), permitting a more confident assignment of expression patterns (Figure 3D).

An additional benefit of this spatial analysis is that it provides estimates of how each inferred cell type localizes in the ventral spinal cord (Figure 4). Although the method does not impose continuity on cell-type distributions, we find that many of the

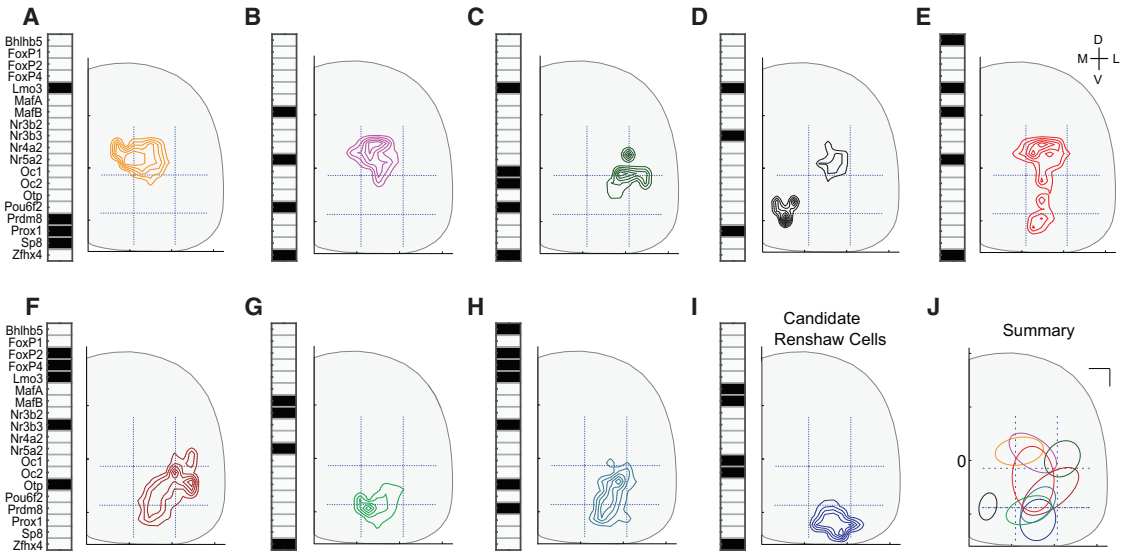


Figure 4. Inferred Cell Type Spatial Distributions Segregate V1 Interneurons into Compact Domains

(A–I) Positional distributions of inferred V1 cell types. These populations are confined to compact spatial domains. (I) Spatial distribution of an inferred cell type corresponding to candidate Renshaw interneurons, defined by expression of known Renshaw markers (MafB, Oc1, and Oc2) and localization in an extreme ventral position.

(J) Spatial distributions from cell types in (A)–(I) aggregated in a single plot. Each cell type is represented by its confidence ellipse under a Gaussian approximation to the posterior spatial distribution of each cell type (66% confidence ellipse). Scale bar, 100 μm.

See also Figure S4.

inferred cell types are localized in relatively compact, contiguous domains, covering the full positional spectrum of the parental V1 interneuron distribution along both the dorsoventral and mediolateral axes. Notably, one inferred cell type with the expression profile of Renshaw interneurons (expressing MafB, Oc1, and Oc2) is predicted to be confined to the most ventral region within the parental V1 population (Figure 4I), in agreement with their known settling position (Alvarez and Fyffe, 2007; Stam et al., 2012). Other inferred cell types, characterized by FoxP2, FoxP4, Nr3b3, and/or Nr4a2 expression, showed clustered distributions ventral to the central canal and dorsomedial to motor neurons (Figure S4). Such distributions are similar to the proposed location of group Ia reciprocal interneurons (Hultborn et al., 1971), a subset of which are known to reside within the parental V1 population (Zhang et al., 2014). Taken together, these findings document novel molecular and spatial diversity in the V1 interneuron population.

A Cladistic Analysis of Transcription Factor Expression

To characterize the minimal number of TFs needed to provide selective access to an individual cell type, we developed a classification scheme that relies on a recursive algorithm to sequentially subdivide the parental population and arrange every cell type along a clade diagram (see Supplemental Information; Figure S5). In this representation, the central node of the diagram corresponds to the full V1 population, with branches representing TFs expressed in a mutually exclusive fashion covering the highest fraction of the parental population. This process is repeated until the candidate cell types from which the analysis is constructed are revealed at the extremities of the plot. Our analysis reveals that 64% of the V1 parental population can be divided into four main clades on the basis of the mutually exclusive expression of FoxP2, Pou6f2, Sp8, and MafA (Figure 5A). Each clade contains from 4 to 19 cell types, with a total of 36 cell types falling within the four main clades (Figure 5B). The minimum number of TFs needed to target each cell type ranges from two (in the case of Pou6f2, Nr5a2 neurons) to six (as in the final leaves of the FoxP2 clade) with a mean number of 4 ± 1 (mean ± SD).

Analysis of clade settling positions shows that the MafA, Sp8, and Pou6f2 clades exhibited little spatial overlap, whereas the FoxP2 clade displays a broad spatial distribution with significant overlap with the other three clades (Figure 5C). At the second tier of our clade assignment, V1 subclasses become more spatially restricted (Figure 5D), with additional restrictions for progressively higher tiers (not shown). Cell types within the Pou6f2 clade exhibit medio-lateral gradations in their spatial distributions, determined by the expression of the TFs Nr5a2 and Lmo3, respectively. In certain instances, however, cell types within a single clade show overlapping spatial distributions, best exemplified by the FoxP2 clade, characterized by numerous intermingled cell types with no statistically significant difference in their centroid coordinates (Figure S4). In summary, this clade analysis provides predictive insight into the relative contributions of individual TFs as well as spatial information of use in delineating the hierarchy of V1 interneuron candidate cell types.

Validation of Bayesian Model Predictions

The merits of our computational analysis depend critically on the ability to infer cellular diversity accurately (see Computational Validation of the Bayesian Model in the Supplemental

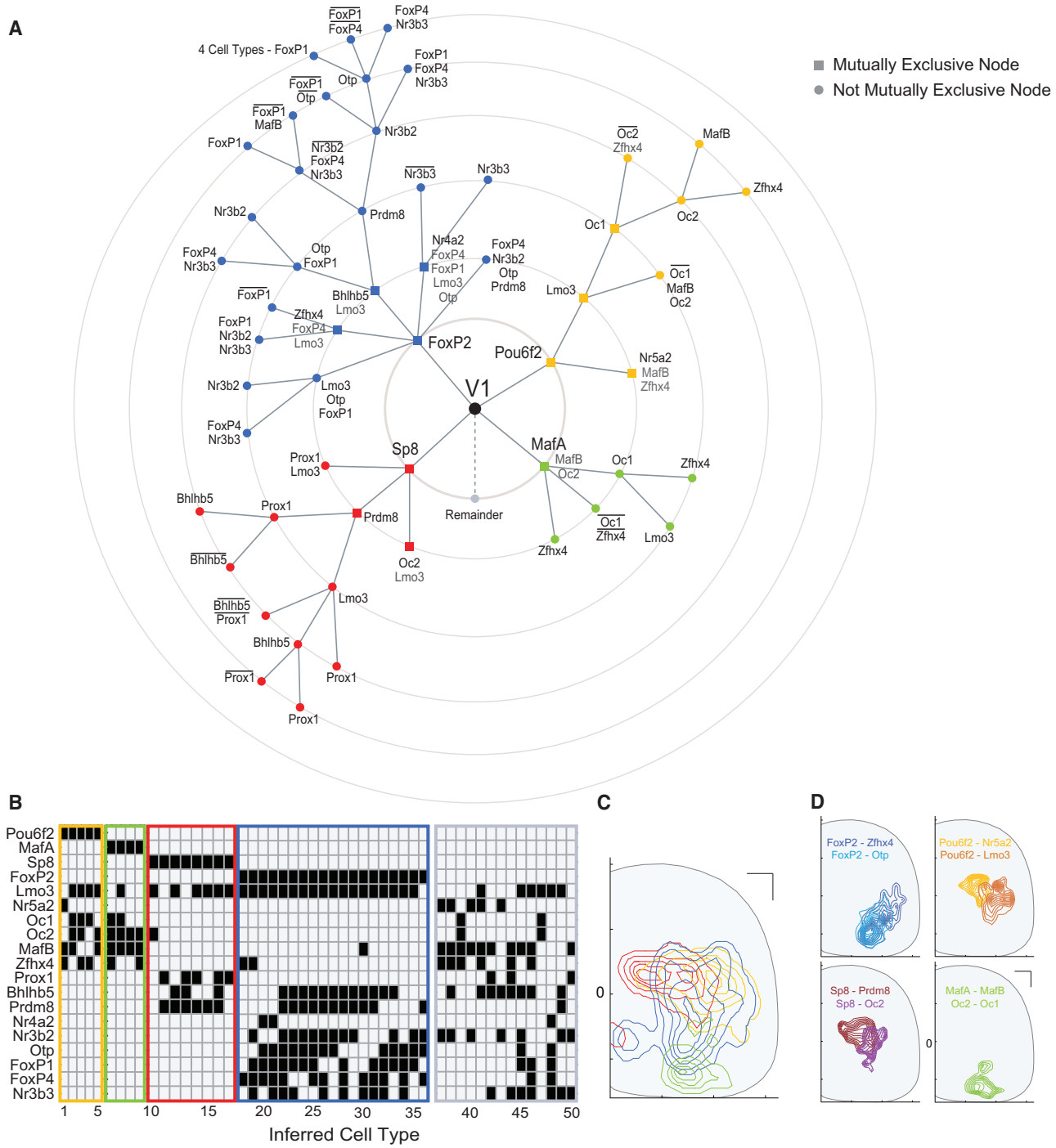


Figure 5. Mutually Exclusive Cell Types Divide the V1 Parental Population into Four Clades

(A) Clade diagram constructed from the set of 50 cell types corresponding to the collection that occurs most frequently among the samples (mode of the posterior). Each terminal node corresponds to a cell type, with its TF profile obtained by traversing the diagram from the center to the outermost levels. Diagram is portrayed up to level 6 in the hierarchy and contains 15 cell types out of the 19 belonging to the FoxP2 clade. Bar above a TF name denotes lack of expression.

(B) Expression profiles of inferred cell types contained within each of the four clades. Gray box contains remaining cell types not expressed within $V1^{FoxP2}$, $V1^{MafA}$, $V1^{Pou6f2}$, or $V1^{Sp8}$ clades.

(C) $V1^{MafA}$, $V1^{Pou6f2}$, and $V1^{Sp8}$ clades represent mutually exclusive subsets but overlap spatially with the $V1^{FoxP2}$ clade. Scale bar, 100 μm.

(D) V1 spatial distributions corresponding to subpopulations at the second tier of the clade diagram. Blue corresponds to cell types within the $V1^{FoxP2}$ clade, green correspond to $V1^{MafA}$, yellow corresponds to $V1^{Pou6f2}$, and red corresponds to $V1^{Sp8}$. Scale bar, 100 μm.

See also Figure S5.

Information; Figure S6). We sought to assess the biological accuracy of the Bayesian model's predictions, comparing first experimental findings and inferred results from Bayesian analysis with single-cell qRT-PCR data (Figure 6). We focused on 15 TFs for which reliable qRT-PCR probes could be identified and analyzed TF expression within $En1^+$ neurons, isolated at random from p0 lumbar spinal segments of *En1::Cre; RCE::lsl.GFP* mice. We assessed whether qRT-PCR and immunocytochemistry gave comparable co-expression values. Appropriate thresholds for each gene were set, relative to the expression of the ubiquitously expressed gene β-actin, with the aim of comparing the measured qRT-PCR patterns against our immunohistochemical measurements. Thresholds were chosen by minimizing the distance between qRT-PCR generated expression values and immunohistochemical measurements (Experimental Procedures, Single Cell qRT-PCR Characterization). After applying appropriate thresholds, TFs were classified either as "expressed" or "not expressed" within individual $En1^+$ interneurons. The patterns of gene expression emerging from qPCR exhibited high correlation with immunohistochemical data for both individual and paired TF measurements (Figures 6A–6C, correlation coefficient = 0.88).

We then asked whether qRT-PCR transcriptional patterns correlate with the clade results of our Bayesian analysis, seeking to validate the general organization of our candidate expression patterns, instead of individual cell types. Individual V1 interneuron gene expression profiles can be segregated along the four major inferred clade populations, indicating that our computational predictions accurately reflect gene expression relationships in vivo (Figure 6D). Importantly, qRT-PCR identified closely related expression patterns predicted by the model. These results validate the general organization of the Bayesian cell-type predictions.

We next tested predictions involving triple labeling of combinations of TFs not previously measured, guided by experimental design considerations (Figure S7). We focused on $V1^{Sp8}$ interneurons, noting that predicted fractions for the seven potential combinations of Sp8, Prox1, and Prdm8 are in good agreement with their measured values (Figure 7A; Table S1). Predictions arising from spatial analyses tend to be more constrained, with smaller SDs and are generally more accurate than the non-spatial predictions. Moreover, we validated the predicted absence of the combination $Prdm8^+$, $Prox1^+$, $Sp8^-$ (Figure 7B). Taken together, these results indicate that the Bayesian approach accurately infers the TF expression profile of neuronal types within the parental V1 population.

Finally, our results also enabled us to test predictions about unmeasured spatial distributions of neurons expressing pairs of TFs, on the basis of measured spatial information for single TFs. We focused on cases in which a combination of two TFs confined an inferred V1 neuronal type to a highly restricted region of the parental V1 distribution. We found that our predictions faithfully co-localize with the actual distributions assessed in p0 caudal lumbar spinal segments (Figures 7C–7E). These results indicate that the Bayesian approach, by virtue of incorporating dual cellular sources of information, correctly predicts the spatial distribution of novel gene combinations.

Generalization of Bayesian Diversity Estimates

To establish the general applicability of our Bayesian approach, we evaluated its ability to identify cell types in systems where an estimate of cellular diversity had been extracted by other analysis procedures.

We first focused on the zebrafish embryo, for which single cells have been transcriptionally profiled by RNA sequencing (RNA-seq) and mapped to their location of origin (Satija et al., 2015). Although the delineation of the entire cellular repertoire was not attempted in that work, the analysis of single cell cluster profiles across the marginal region of the embryo is consistent with seven cell types (Figure 7G). We sought to determine whether our sparse Bayesian methods are able to achieve this result given simulated data generated by randomly subsampling the dataset from Satija et al. (2015) (see Ground Truth Data Generation from RNA-Seq Measurements in the Supplemental Information). In the absence of spatial information, the sparse Bayesian algorithm estimated 5 ± 1 cell types. The transcriptional profile of each inferred cell type corresponds to one of the ground-truth candidates, but our procedure underestimated total cell-type number (Figure 7I). Introducing spatial information into the analysis increases the number of correctly inferred cell types to 6 ± 1 (Figure 7J), close to 7, the ground-truth number. Thus as in the V1 study, inference is improved by incorporating additional cellular characteristics—in this case location. The similar shape of some spatial distributions, together with the random selection of the subset of genes incorporated in our analysis (see Supplemental Information for details), are likely reasons that the algorithm slightly underestimates cell-type diversity. Nevertheless, the results obtained by the Bayesian approach are generally in good agreement with those obtained by clustering the original RNA-seq data.

We next analyzed cortical interneuron diversity, where 16 interneuronal cell types have been identified on the basis of RNA-seq data in mouse somatosensory cortex and hippocampal CA1 neurons (Zeisel et al., 2015). From this dataset, single and pairwise measurements were created, with errors assigned to each measurement (Supplemental Information). We used this dataset to construct Bayesian estimates of neocortical diversity in the absence of spatial information, varying the number of genes used in the analysis, the noise in the measurements and the amount of missing data (representing antibody incompatibility). From this, we inferred 12.7 ± 0.3 cell types over a range of 13 to 16 genes used, all 12 corresponding to correctly inferred expression profiles (Figure 7K). In every example, the sparse Bayesian analysis outperformed the NNCLS approach, which overestimated the number of cell types by nearly 100%. We next used all selected genes to estimate sensitivity to noise and missing data and observed a larger effect for noise (Figure 7L). As the noise level and amount of missing data tend to zero, we correctly infer the total number of cell types and their expression profiles (Figure 7M).

These analyses establish the sparse Bayesian approach as an effective means of estimating neuronal type diversity and provide further insight into the benefits of incorporating spatial information when obtaining accurate estimates.

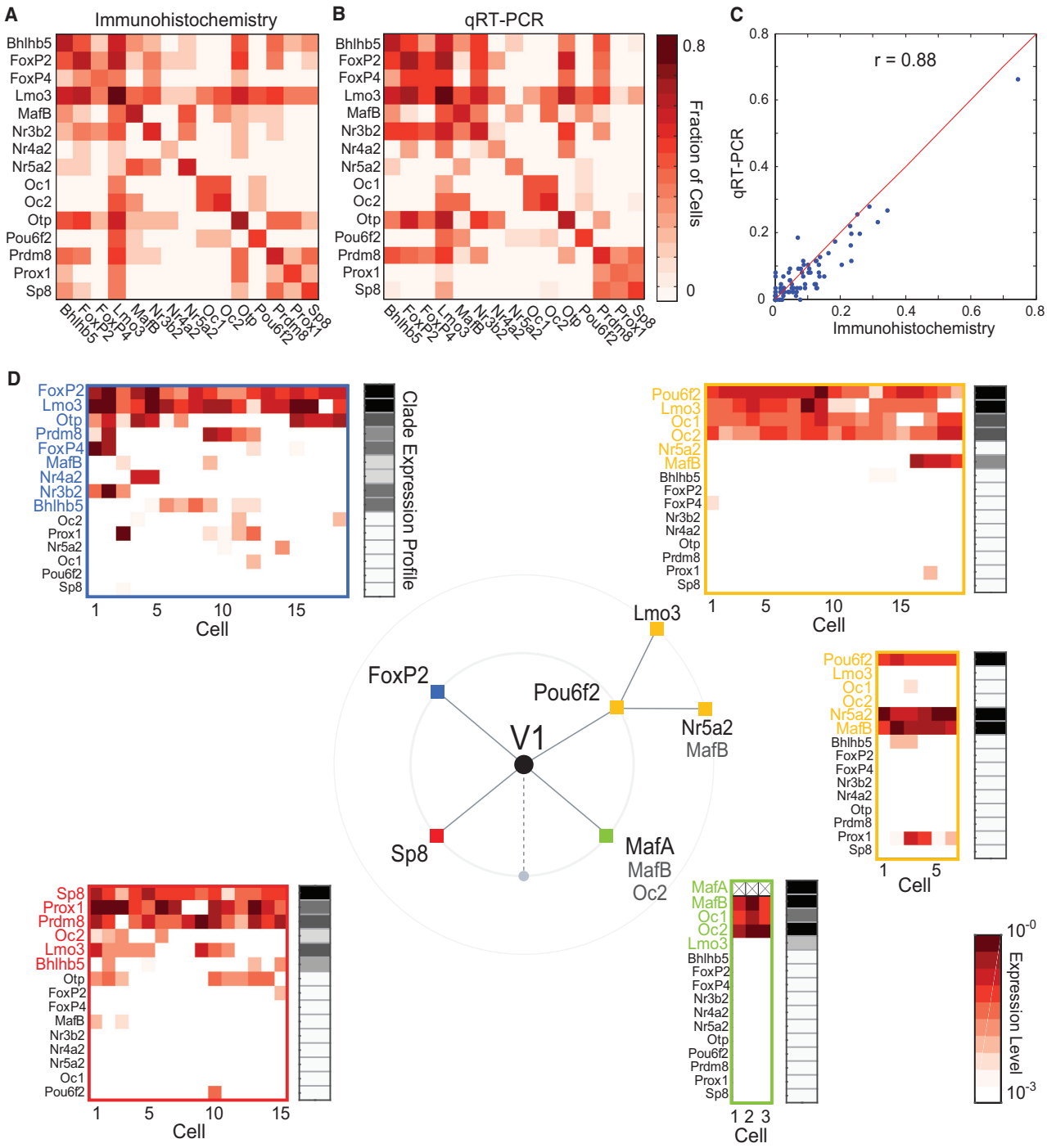


Figure 6. Single-Cell qRT-PCR Confirms Antibody Measurements and Validates Candidate Clade Expression Profiles

(A) Matrix of V1 interneurons representing fraction of cells expressing single and paired TFs (same as in Figures 1A and 1B, reproduced to compare against B here). Fractional values of single TFs represented as diagonal elements.

(B) Co-expression matrix calculated using single-cell qRT-PCR information (plotted as in A); n = 86 cells.

(C) Immunohistochemistry versus qRT-PCR values. (A) versus (B) show a correlation value of 0.88.

(D) Single-cell expression profiles can be arranged according to cladistic analysis; second tier predictions are corroborated in the Pou6f2 clade. The clade expression profile (C.E.P.) was computed by averaging expression profiles of inferred cell types belonging to each clade, weighted by their posterior inclusion probability. These profiles, computed solely from immunohistochemistry data, match the clustered qRT-PCR measurements. Twenty-five of 86 total cells span the remainder (i.e., were not assigned to the clusters shown here; data not shown), consistent with the ratio of remainder cell types shown in Figure 5B.

See also Figure S6 and Table S2.

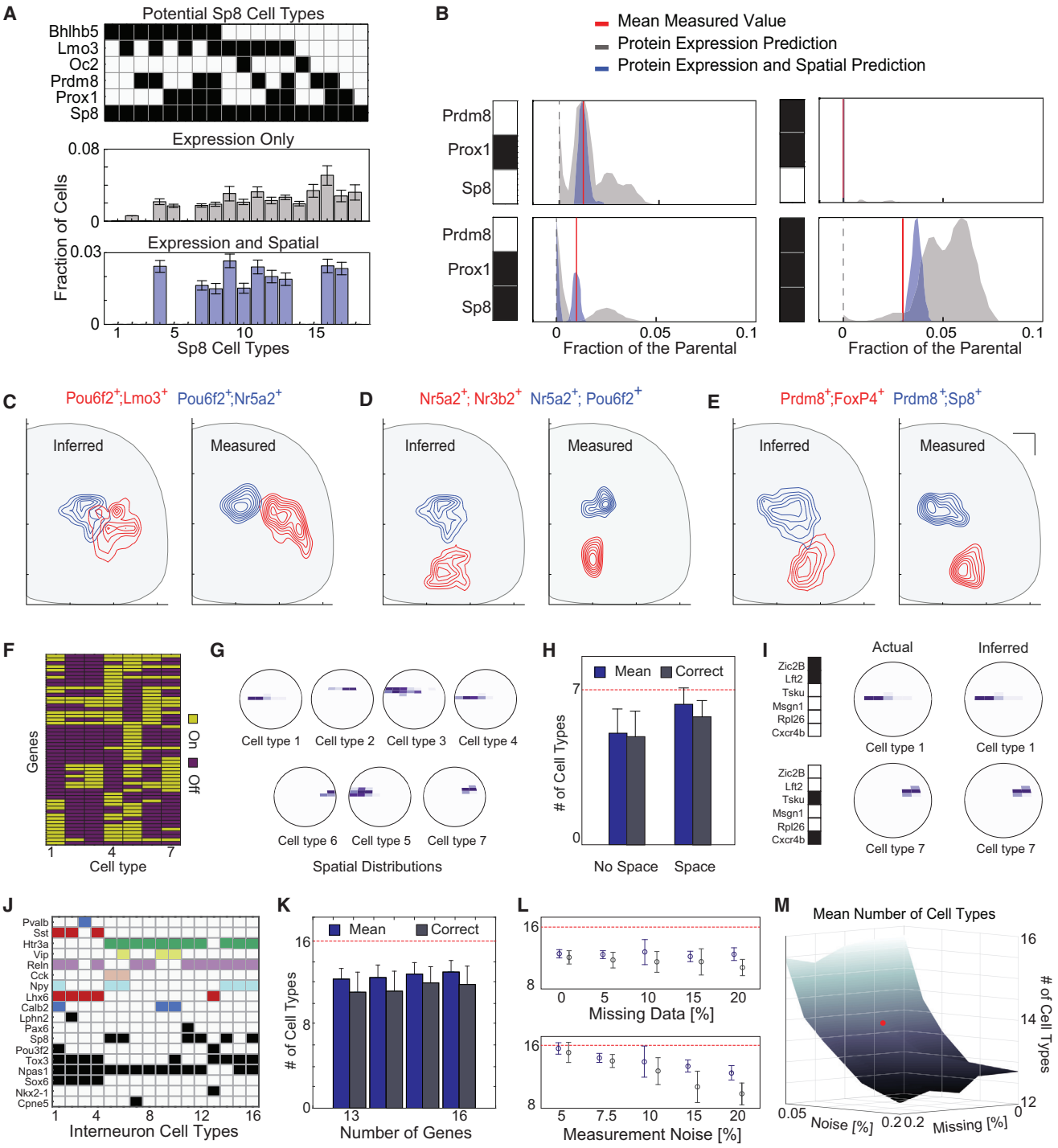


Figure 7. Biological Validation of Bayesian Predictions

(A) Eighteen potential cell types expressing Sp8 TF (top), along with their inferred fractions within the parental population (mean ± SD, as in Figure 2A, middle and bottom), calculated using solely TF expression information (middle) or both expression and spatial information (bottom).

(B) Predicted prevalence for measured triplet antibody combinations. Mean measured value is depicted as a red line. Predicted values are indicated in gray or blue (computed using protein expression information only, or with spatial information, respectively). See also Table S1.

(C–E) Spatial distributions for dual transcription factor-gated V1 subsets can be predicted accurately. Left: indicates prediction. Right: measured distributions. Scale bar, 100 μm.

(F) Expression profiles of zebrafish cell types identified by Satija et al. (2015).

(G) Spatial distribution of each cell type in (F).

(legend continued on next page)

DISCUSSION

Objective assessment of the extent of mammalian cellular diversity has remained challenging. The sparse Bayesian framework presented here provides a general method for characterizing cellular heterogeneity on the basis of sparsely sampled biological information. We have used this framework to study spinal V1 interneuron diversity. By analyzing the spatial expression densities of individual TFs as well as their patterns of pairwise expression, candidate expression profiles for ~50 inferred V1 cell types are provided. The integration of distinct phenotypic aspects of cellular heterogeneity, in this instance TF expression and settling position, markedly enhances the confidence in assignment of predicted V1 interneuron cell types. We note that this approach provides a general method for delineating the heterogeneity of cell types in any mixed tissue.

Bayesian and Other Approaches to Cell-Type Diversity

Estimates of cell-type diversity can be obtained through computational approaches that employ hierarchical clustering (see Armañanzas and Ascoli, 2015 for a recent review). But these methods have drawbacks—it is challenging to use hierarchical clustering to determine the number of cell types automatically and their inferences can be sensitive to the choice of similarity measures (Augen, 2005).

A different set of computational approaches based on deconvolution algorithms have been used to characterize cellular diversity from information about gene expression profiles (Shen-Orr and Gaujoux, 2013). These can be divided into two major methodologies. Regression approaches can be applied when the expression profile of cell types of interest is known a priori (Wang et al., 2006; Abbas et al., 2009; Gong et al., 2011; Zuk et al., 2013; Grange, et al., 2014). In contrast, matrix-factorization approaches become relevant when cell-type expression profiles are not known (Repsilber et al., 2010; Erkkilä et al., 2010; Bazot et al., 2013; Zhong et al., 2013; Liebner et al., 2014). The latter approaches suffer from similar limitations as the NNCLS method: matrix factorizations are inherently non-unique, and estimation of confidence levels can be difficult. Past attempts to overcome these challenges have relied on the premise that particular genes are expressed in only a single cell type, which does not represent the general biological case (Gaujoux and Seoighe, 2012). A more recent approach, conceptually akin to the methods developed here, uses SnS priors to reduce the mathematical ambiguities inherent in matrix factorization and has recently been applied to the identification of genetic disease pathways (Shen et al., 2015), albeit without addressing cell-type diversity.

Our approach sidesteps previous limitations by focusing on the fractional prevalence of individual and paired TFs within a parental population, with well-defined values between zero and one, in contrast to expression levels that may be scaled arbitrarily (see Supplemental Information for further discussion). Mathematically, our model resembles previous regression models (Wang et al., 2006; Abbas et al., 2009; Gong et al., 2011; Zuk et al., 2013; Grange et al., 2014), but has the virtue of incorporating the SnS prior and a binary set of regressors that span all possible cell types. For the data considered here, the resulting regression problem is highly ill-posed (Figure S3), necessitating a fully Bayesian approach to capture uncertainty in resultant estimates.

A distinctive feature of our approach is that we start by considering all of the possible expression patterns for the genes being considered, 2^{19} in our case and, in general, 2^N for a study involving N genes. We note that this factor 2^N may be prohibitive for applications of the method to RNA-seq data, where a large number of genes are typically tracked. Although applications in which N is several thousand would appear impractical, it may be possible to identify a subset of genes expected to be particularly informative about cell type and restrict the analysis to this subset. Even with a reduced N, 2^N may be dauntingly large, but it is important to recall that in our analysis a preliminary screening reduced the number of expression patterns by a factor of ~265, from 2^{19} (524,288) to 1,978. Greater N values may yield even larger reductions.

The Extent of Transcriptional Diversity

Our analysis has identified extensive transcriptional diversity within V1 interneurons on the basis of the expression of 19 TFs. The first issue this raises is whether further diversity will follow inevitably with the inclusion of additional V1 TFs. We analyzed the impact of varying the number of TFs in our analysis and found only a weak dependence of the number of cell types on the number of TFs (Figures 1E and 1F). Thus, 19 TFs appear sufficient to uncover a substantial fraction of the total underlying transcriptional heterogeneity.

A second issue is the impact of incorporating spatial information on cell type. In our analysis, the inclusion of spatial data increased inferred cell-type number by ~70% and markedly enhanced confidence in the inferred expression profiles. Spatial

(H) In blue, posterior mean ± SD number of cell types per sample. In gray, mean ± SD number of correctly identified cell types.
(I) Examples of two correctly inferred spatial distributions.
(J) Interneuronal cell types identified by Zeisel et al. (2015). Commonly used markers are color coded as in Zeisel et al. (2015) and additional markers are colored in black.
(K) Sparse Bayesian regression underestimates total cell type. We randomly selected 13 to 16 genes from the list of markers defined in (J). In blue, posterior mean ± SD number of cell types per sample. In gray, mean ± SD number of correctly identified cell types. The expression profile of the first 16 patterns is compared to the true patterns. The red dashed line indicates the ground-truth value of 16.
(L) Impact of missing data and error in the measurement dataset. Top: fixing the measurement error at 10% and using all the genes described in (J), the performance of the algorithm remains constant when varying the amount of data removed. Bottom: fixing the amount of missing data at 10% and using all the genes described in (J), the performance of the algorithm decreases as the measurement noise approaches 20%. The red dashed line indicates the ground-truth value of 16.
(M) Landscape representing the mean number of inferred cell types when varying the amount of missing data and the noise in the measurements. Red Dot indicates a level of missing data and noise similar to V1 interneurons.
See also Figure S7.

information has a much stronger impact on cell-type assignment than variation in the number of TFs, indicating that settling position carries significant cell-type information which is independent from the information carried by the expression patterns of the 19 TFs examined here.

A third issue is the impact of V1 cell-type spatial segregation on synaptic input specificity (Bikoff et al., 2016). An inordinate diversity in spatially restricted cell types may be necessary to satisfy highly diverse spinal circuit computations. Our data indicate that spatial segregation only partially accounts for spinal V1 diversity. A large number of inferred cell types, marked by expression of FoxP2, localize in the same ventral region (Figure S4). Additional heterogeneity might be realized by the expression of surface molecules that impose and constrain connectivity with different synaptic partners.

We note that transcriptional diversity could, in some instances, reflect variation in functional cell state, rather than indicating a distinct neuronal subtype. However, consistent with the genetic specification of cell-type identity, we find that the position of transcriptionally distinct V1 subsets are segregated, stereotyped from animal to animal, and stable across development (see Bikoff et al., 2016) (Figures 3H, S4A, and S4B). The spatial segregation argues strongly against the "cell state" possibility. Nevertheless, activity-shaped differences in Er81 expression in fast-spiking cortical interneurons have been shown to mark delay-type or non-delayed firing states (Dehorter et al., 2015). In addition, activity-dependent induction of Npas4 expression has been described for cortical neurons (Lin et al., 2008), with implications for homeostatic regulation of sensitivity to inhibitory transmitters. Further studies will therefore be needed to dissect the functional consequences of V1 diversity, to resolve whether certain state-dependent functional properties are reflected in the diversity of V1 transcriptional profiles.

Broader Implications of a Bayesian Analysis of Cellular Diversity

Our statistical approach has relevance well beyond a focus on spinal V1 interneurons and could prove useful in further delineating neuronal cell types elsewhere in the nervous system. Cortical projection neurons fractionate into a few broad classes based on patterns of target innervation and distinctions in gene expression, yet the extent to which any single broad class of pyramidal neurons is itself heterogeneous remains unclear (Greig et al., 2013). The classification of interneuron cell types in the brain has proven particularly challenging (Ascoli, 2008; DeFelipe et al., 2013; Kepecs and Fishell, 2014), although studies of hippocampal interneuron diversity suggest a degree of heterogeneity that approaches that found for spinal V1 interneurons. Within CA1 hippocampus, over 20 inhibitory interneuron subtypes have been identified, based on anatomical, molecular, or electrophysiological distinctions (Krook-Magnuson et al., 2012). Single-cell transcriptome analysis of primary somatosensory cortex or CA1 hippocampus interneurons has identified 16 molecularly distinct interneuron cell types, which likely represents a lower bound on diversity (Zeisel et al., 2015). Thus, insight into interneuronal diversity in the spinal cord may inform studies to address heterogeneity throughout the brain.

Our analysis also has implications for genetic strategies aimed at manipulating circuit elements throughout the nervous system. The minimal number of TFs needed to define a single V1 cell type uniquely has been identified on the basis of clade profiles and is, on average, 4 ± 1. This indicates that individual TFs are generally not sufficient to isolate V1 neuronal types, consistent with findings in other neuronal systems (Sanes and Masland, 2015). The difficulty in identifying single TFs that uniquely define a cell type may reflect the prevalence of combinatorial TF codes (Philippidou and Dasen, 2015) and could explain the difficulty in delineating individual motor neuron pools (De Marco Garcia and Jessell, 2008).

EXPERIMENTAL PROCEDURES

Immunohistochemistry

Immunohistochemistry was performed as in Bikoff et al. (2016). Briefly, p0 *En1::cre; Tau.lsl.nLacZ* mice were transcardially perfused with 4% paraformaldehyde in 0.1 M phosphate buffer, followed by a 2-hr postfixation. Tissue was then washed, cryoprotected by equilibration in 30% sucrose in 0.1 M phosphate buffer, embedded in OCT, and cryostat-sectioned in the transverse plane at 20 μm. Immunohistochemistry was performed on tissue through sequential exposure to primary antibodies (overnight at 4°C) and fluorophore-conjugated secondary antibodies (1 hr at room temperature). Sections were mounted using Fluoromount-G (SouthernBiotech) and coverslipped for imaging. Confocal images were obtained on an LSM 710 Meta Confocal microscope (Carl Zeiss) at 1,024 × 1,024 resolution, using a Plan-Apochromat 20×/0.8 M27 objective. See Bikoff et al. (2016) for a description of antibodies used.

Transcription Factor Co-expression

Confocal images of transcription factor co-expression were analyzed in Imaris (Bitplane) using the "Colocalization" and "Spots" functions with thresholds set to exclude nonspecific background reactivity, followed by manual validation to confirm co-expression within V1 interneurons. For each transcription factor combination, at least two independent sections from three or more animals were analyzed, totaling >580 lumbar sections and >1,100 lumbar hemisections in this dataset.

Spatial Analysis

Interneuron spatial distributions are described in Bikoff et al. (2016). Sections were normalized to a standardized spinal cord hemisection (distance from central canal to lateral boundary: 650 μm; distance from central canal to bottom-most boundary: 400 μm). Coordinates were exported from Imaris and plotted using the contour function in MATLAB.

Bayesian Sparse Regression Model

The Supplemental Information includes a detailed description of the Bayesian model and Hamiltonian Monte Carlo algorithm for sampling from the posterior distribution of the fractional values $\boldsymbol{f}$ given the observed data, under the SnS prior.

Single-Cell qRT-PCR Characterization

GFP^+ cells from lumbar spinal cords of p0 *En1::Cre; RCE.lsl.GFP* mice were dissociated using the Papain Dissociation Kit (Worthington), and single cells were isolated by fluorescence-activated cell sorting using a Beckman Coulter MoFlo Astrios cell sorter. Cells were directly deposited into 96-well plates containing lysis buffer (Ambion Single Cell-to-CT Kit; Life Technologies). Reverse transcription and pre-amplification were performed according to the manufacturer's protocol (Ambion, Life Technologies). Pre-amplified products were loaded on Biomark 48.48 Dynamic Arrays and run on the Biomark HD microfluidic multiplex qRT-PCR platform (Fluidigm). Gene expression levels were assayed using TaqMan probes (Life Technologies) directed against housekeeping genes and 15 of the 19 transcription factor-encoding genes, excluding FoxP1, MafA, Nr3b3, and Zfhx4 for which reliable probes could not be identified (see Table S2 for a description of TaqMan probes

and Supplemental Experimental Procedures). Ct values were measured by Biomark software, where relative transcript levels were determined by 2-Ct normalization to Actb transcript levels. The co-expression matrix for single cell qRT-PCR data was calculated using an optimization approach. Briefly, a threshold for each gene was computed by minimizing the distance between a co-expression matrix calculated by qRT-PCR and immunohistochemistry. The fraction of cells expressing gene X was computed simply by calculating the number of cells expressing gene X divided by the total number of cells. All experiments and procedures were performed according to NIH guidelines and approved by the Institutional Animal Care and Use Committee of Columbia University.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.01.026.

AUTHOR CONTRIBUTIONS

M.I.G., A.P., L.A., and L.P. designed computational analysis. J.B.B. performed biological experiments upon which statistical analysis was based. M.I.G. and J.B.B. performed and analyzed qPCR experiments. M.I.G., J.B.B., A.P., L.A., L.P., and T.M.J. prepared the manuscript.

ACKNOWLEDGMENTS

We thank H. Lee for sharing advice on qPCR experiments. R. Axel, W. Fischler, A. Miri, D. Gutnisky, and J.C. Tapia provided valuable discussions and comments on the manuscript. C. Zuker suggested validation experiments. We thank A. Karpova and S. Druckman for quantitative insight. M.I.G. was supported by the National Science Foundation through a GRFP fellowship. T.M.J. was supported by NIH grant NS033245, the Harold and Leila Y. Mathers Foundation, the Brain Research Foundation, and Project A.L.S. and is an investigator of the Howard Hughes Medical Institute. L.F.A. was supported by NIH grant MH093338 and by the Gatsby, Swartz, and Mathers Foundations. L.P. and A.P. were supported by ONR grant N00014-14-1-0243 and ARO MURI grant W911NF-12-1-0594.

Received: July 27, 2015
Revised: November 30, 2015
Accepted: January 15, 2016
Published: March 3, 2016

REFERENCES

Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H.F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS ONE *4*, e6098.

Alvarez, F.J., and Fyffe, R.E. (2007). The continuing case for the Renshaw cell. J. Physiol. *584*, 31–45.

Alvarez, F.J., Jonas, P.C., Sapir, T., Hartley, R., Berrocal, M.C., Geiman, E.J., Todd, A.J., and Goulding, M. (2005). Postnatal phenotype and localization of spinal cord V1 derived interneurons. J. Comp. Neurol. *493*, 177–192.

Amamoto, R., and Arlotta, P. (2014). Development-inspired reprogramming of the mammalian central nervous system. Science *343*, 1239882.

Armañanzas, R., and Ascoli, G.A. (2015). Towards the automatic classification of neurons. Trends Neurosci. *38*, 307–318.

Ascoli, G.A. (2008). Neuroinformatics grand challenges. Neuroinformatics *6*, 1–3.

Augen, J. (2005). Bioinformatics in the Post-Genomic Era: Genome, Transcriptome, Proteome, and Information-Based Medicine (Addison-Wesley Press).

Bazot, C., Dobigeon, N., Tourneret, J.Y., Zaas, A.K., Ginsburg, G.S., and Hero, A.O., 3rd. (2013). Unsupervised Bayesian linear unmixing of gene expression microarrays. BMC Bioinformatics *14*, 99.

Benito-Gonzalez, A., and Alvarez, F.J. (2012). Renshaw cells and Ia inhibitory interneurons are generated at different times from p1 progenitors and differentiate shortly after exiting the cell cycle. J. Neurosci. *32*, 1156–1170.

Bikoff, J.B., Gabitto, M.I., Rivard, A.F., Drobac, E., Machado, T.A., Miri, A., Brenner-Morton, S., Famojure, E., Diaz, C., Alvarez, F.J., et al. (2016). Spinal inhibitory interneuron diversity delineates variant motor microcircuits. Cell *165*, this issue, 207–219.

Dalla Torre di Sanguinetto, S.A., Dasen, J.S., and Arber, S. (2008). Transcriptional mechanisms controlling motor neuron diversity and connectivity. Curr. Opin. Neurobiol. *18*, 36–43.

Dasen, J.S., and Jessell, T.M. (2009). Hox networks and the origins of motor neuron diversity. Curr. Top. Dev. Biol. *88*, 169–200.

Dasen, J.S., Tice, B.C., Brenner-Morton, S., and Jessell, T.M. (2005). A Hox regulatory network establishes motor neuron pool identity and target-muscle connectivity. Cell *123*, 477–491.

De Marco Garcia, N.V., and Jessell, T.M. (2008). Early motor neuron pool identity and muscle nerve trajectory defined by postmitotic restrictions in Nkx6.1 activity. Neuron *57*, 217–231.

DeFelipe, J., López-Cruz, P.L., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., Burkhalter, A., Cauli, B., Fairén, A., Feldmeyer, D., et al. (2013). New insights into the classification and nomenclature of cortical GABAergic interneurons. Nat. Rev. Neurosci. *14*, 202–216.

Dehorter, N., Ciceri, G., Bartolini, G., Lim, L., del Pino, I., and Marín, O. (2015). Tuning of fast-spiking interneuron properties by an activity-dependent transcriptional switch. Science *349*, 1216–1220.

Erkkilä, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I., and Lähdesmäki, H. (2010). Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics *26*, 2571–2577.

Fyffe, R.E. (1990). Evidence for separate morphological classes of Renshaw cells in the cat's spinal cord. Brain Res. *536*, 301–304.

Gaujoux, R., and Seoighe, C. (2012). Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. Infect. Genet. Evol. *12*, 913–921.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., and Vehtari, A. (2013). Bayesian Data Analysis (Chapman & Hall/CRC).

Gong, T., Hartmann, N., Kohane, I.S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., and Szustakowski, J.D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS ONE *6*, e27156.

Grange, P., Bohland, J.W., Okaty, B.W., Sugino, K., Bokil, H., Nelson, S.B., Ng, L., Hawrylycz, M., and Mitra, P.P. (2014). Cell-type-based model explaining coexpression patterns of genes in the brain. Proc. Natl. Acad. Sci. USA *111*, 5397–5402.

Greig, L.C., Woodworth, M.B., Galazo, M.J., Padmanabhan, H., and Macklis, J.D. (2013). Molecular logic of neocortical projection neuron specification, development and diversity. Nat. Rev. Neurosci. *14*, 755–769.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. *19*, 1720–1730.

Hultborn, H., Jankowska, E., and Lindström, S. (1971). Recurrent inhibition of interneurones monosynaptically activated from group Ia afferents. J. Physiol. *215*, 613–636.

Isaacson, J.S., and Scanziani, M. (2011). How inhibition shapes cortical activity. Neuron *72*, 231–243.

Ishwaran, H., and Rao, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat. *33*, 730–773.

Kepecs, A., and Fishell, G. (2014). Interneuron cell types are fit to function. Nature *505*, 318–326.

Kohwi, M., and Doe, C.Q. (2013). Temporal fate specification and neural progenitor competence during development. Nat. Rev. Neurosci. *14*, 823–838.

Krook-Magnuson, E., Varga, C., Lee, S.H., and Soltesz, I. (2012). New dimensions of interneuronal specialization unmasked by principal cell heterogeneity. Trends Neurosci. *35*, 175–184.

Liebner, D.A., Huang, K., and Parvin, J.D. (2014). MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. Bioinformatics *30*, 682–689.

Lin, Y., Bloodgood, B.L., Hauser, J.L., Lapan, A.D., Koon, A.C., Kim, T.K., Hu, L.S., Malik, A.N., and Greenberg, M.E. (2008). Activity-dependent regulation of inhibitory synapse development by Npas4. Nature *455*, 1198–1204.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Molyneaux, B.J., Arlotta, P., and Macklis, J.D. (2007). Molecular development of corticospinal motor neuron circuitry. Novartis Found. Symp. *288*, 3–15, discussion 15–20, 96–98.

Pakman, A., and Paninski, L. (2013). Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions. Adv. Neural Inf. Process. Syst. *26*, 2490–2498.

Pakman, A., and Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussian. J. Comput. Graph. Stat. *23*, 2.

Philippidou, P., and Dasen, J.S. (2015). Sensory-motor circuits: Hox genes get in touch. Neuron *88*, 437–440.

Renshaw, B. (1946). Central effects of centripetal impulses in axons of spinal ventral roots. J. Neurophysiol. *9*, 191–204.

Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G.F., Selbig, J., Parida, S.K., Kaufmann, S.H., and Jacobsen, M. (2010). Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC Bioinformatics *11*, 27.

Sanes, J.R., and Masland, R.H. (2015). The types of retinal ganglion cells: current status and implications for neuronal classification. Annu. Rev. Neurosci. *38*, 221–246.

Sapir, T., Geiman, E.J., Wang, Z., Velasquez, T., Mitsui, S., Yoshihara, Y., Frank, E., Alvarez, F.J., and Goulding, M. (2004). Pax6 and engrailed 1 regulate two distinct aspects of renshaw cell development. J. Neurosci. *24*, 1255–1264.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Saueressig, H., Burrill, J., and Goulding, M. (1999). Engrailed-1 and netrin-1 regulate axon pathfinding by association interneurons that project to motor neurons. Development *126*, 4201–4212.

Sharma, K., Schmitt, S., Bergner, C.G., Tyanova, S., Kannaiyan, N., Manrique-Hoyos, N., Kongi, K., Cantuti, L., Hanisch, U.K., Philips, M.A., et al. (2015). Cell type- and brain region-resolved mouse brain proteome. Nat. Neurosci. *18*, 1819–1831.

Shen, Y., Rahman, M., Piccolo, S.R., Gusenleitner, D., El-Chaar, N.N., Cheng, L., Monti, S., Bild, A.H., and Johnson, W.E. (2015). ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways. Bioinformatics *31*, 1745–1753.

Shen-Orr, S.S., and Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr. Opin. Immunol. *25*, 571–578.

Siegert, S., Scherf, B.G., Del Punta, K., Didkovsky, N., Heintz, N., and Roska, B. (2009). Genetic address book for retinal cell types. Nat. Neurosci. *12*, 1197–1204.

Stam, F.J., Hendricks, T.J., Zhang, J., Geiman, E.J., Francius, C., Labosky, P.A., Clotman, F., and Goulding, M. (2012). Renshaw cell interneuron specialization is controlled by a temporally restricted transcription factor program. Development *139*, 179–190.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. *19*, 335–346.

Thomas, R.C., and Wilson, V.J. (1965). Precise localization of Renshaw cells with a new marking technique. Nature *206*, 211–213.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P.V., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat. Neurosci. *18*, 145–153.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Wang, M., Master, S.R., and Chodosh, L.A. (2006). Computational expression deconvolution in a complex mammalian organ. BMC Bioinformatics *7*, 328.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

Zhang, J., Lanuza, G.M., Britz, O., Wang, Z., Siembab, V.C., Zhang, Y., Velasquez, T., Alvarez, F.J., Frank, E., and Goulding, M. (2014). V1 and V2b interneurons secure the alternating flexor-extensor motor activity mice require for limbed locomotion. Neuron *82*, 138–150.

Zhong, Y., Wan, Y.W., Pang, K., Chow, L.M., and Liu, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics *14*, 89.

Zuk, O., Amir, A., Zeisel, A., Shamir, O., and Shental, N. (2013). Accurate profiling of microbial communities from massively parallel sequencing using convex optimization. In String Processing and Information Retrieval-Lecture Notes in Computer, O. Kurland, M. Lewenstein, and E. Porat, eds. (Springer International Publishing), pp. 279–297.