Title
Matrix Inversion and Subset Selection (MISS): A pipeline for mapping of diverse cell types across the murine brain

Journal
Proceedings of the National Academy of Sciences of the United States of America, 119(14)

ISSN
0027-8424

Authors
Mezias, Christopher
Torok, Justin
Maia, Pedro D
et al.

Publication Date
2022-04-05

DOI
10.1073/pnas.2111786119

Peer reviewed

PNAS RESEARCH ARTICLE | NEUROSCIENCE

Matrix Inversion and Subset Selection (MISS): A pipeline for mapping of diverse cell types across the murine brain

Christopher Mezias[a,1], Justin Torok[a,b,1], Pedro D. Maia[c], Eric Markley[d], and Ashish Raj[a,b,2]

Edited by Joseph Takahashi, The University of Texas Southwestern Medical Center, Dallas, TX; received June 25, 2021; accepted December 14, 2021

The advent of increasingly sophisticated imaging platforms has allowed for the visualization of the murine nervous system at single-cell resolution. However, current experimental approaches have not yet produced whole-brain maps of a comprehensive set of neuronal and nonneuronal types that approaches the cellular diversity of the mammalian cortex. Here, we aim to fill in this gap in knowledge with an open-source computational pipeline, Matrix Inversion and Subset Selection (MISS), that can infer quantitatively validated distributions of diverse collections of neural cell types at 200-μm resolution using a combination of single-cell RNA sequencing (RNAseq) and in situ hybridization datasets. We rigorously demonstrate the accuracy of MISS against literature expectations. Importantly, we show that gene subset selection, a procedure by which we filter out low-information genes prior to performing deconvolution, is a critical preprocessing step that distinguishes MISS from its predecessors and facilitates the production of cell-type maps with significantly higher accuracy. We also show that MISS is generalizable by generating high-quality cell-type maps from a second independently curated single-cell RNAseq dataset. Together, our results illustrate the viability of computational approaches for determining the spatial distributions of a wide variety of cell types from genetic data alone.

transcriptomics | neuroanatomy | cell-type maps | deconvolution

Significance

The current state-of-the-art mappings of cell types fall short regarding finely resolved subtypes of neural cells, especially γ-aminobutyric acidergic and glutamatergic subtypes. Most such maps compromise on either the number or specificity of unique cell types quantified in each study. Others only use qualitative validation for their maps and fail to address whether gene subset selection is necessary for optimal maps. The Matrix Inversion and Subset Selection pipeline uses publicly available in situ hybridization and single-cell RNA sequencing gene expression data to infer cell-type distributions to map diverse cell types across the murine brain. Most importantly, we demonstrate that data-driven feature selection is necessary to arrive at quantitatively optimal cell-type maps using inversion-, deconvolution-, and correlation-based mapping approaches.

Characterizing whole-brain distributions of neural cell types is a topic of keen interest in modern neuroanatomy, with many applications to both basic and clinical neuroscience research (1–6). Advances in molecular methods for quantifying gene expression and data analytic cell clustering techniques based on morphologic or genetic profiles are enabling the mapping of meso- and microscale neuronal and nonneuronal cell–type architecture at a whole-brain level (7–14). Mammalian whole-brain cell-type mapping has historically focused on neuromodulatory systems, largely because the identification of catecholamine-producing subpopulations using molecular markers is rather straightforward (15–18). More recently, serial two-photon tomography imaging of cells expressing individual cell-type markers genetically tagged to green fluorescent protein successfully mapped three important subpopulations of inhibitory γ-aminobutyric acidergic (GABAergic) interneurons (7) and cholinergic neurons (19) across the murine brain. Although the above animal laboratory techniques cover the entire brain, they are expensive and time consuming to apply to mice, impractical to apply to larger-brained model organisms, and impossible to apply to human subjects.

Recent computational work demonstrates the feasibility of using existing datasets of cell-type gene expression or cell markers for mapping cells in space across the vertebrate brain. Recent work produced whole-brain maps of broad classes of neuronal and glial subpopulations at single-cell resolution using purely computational methods (11), with the limitation that the mapped cells were classed into large metagroups, such as GABAergic vs. glutamatergic neurons, rather than more specific cell types. Pioneering work mapping single-cell RNA sequencing (scRNAseq) data from aquatic flatworms and zebrafish onto in situ hybridization (ISH) expression provided a plausible route to mapping highly specified cell types in mammalian nervous systems (20, 21). Others have inferred the spatial distribution of murine cell types by deconvolving type-specific microarray expression profiles from the ISH-based Allen Gene Expression Atlas (AGEA) (22, 23). In theory, using a matrix-inversion–based approach provides a better estimate of cell density than correlation-based mapping because cell types with highly similar gene expression profiles will necessarily have highly correlated spatial profiles. However, the original work pioneering matrix inversion for cell-type mapping provided no external quantitative validation of their maps. We hypothesized that, with several modifications, the approach of deconvolving ISH data into cell-type densities using recently available scRNAseq data could create voxel-level cell-type maps that could be externally validated both qualitatively and quantitatively.

Author contributions: C.M., J.T., P.D.M., and A.R. designed research; C.M., J.T., and E.M. performed research; C.M., J.T., and E.M. analyzed data; and C.M., J.T., and A.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]C.M. and J.T. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: ashish.raj@ucsf.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2111786119/-/DCSupplemental.

Published April 1, 2022.

We present an information-theoretic computational pipeline that can supply per-voxel estimates of cell densities of distinct and highly specified cell types across the whole murine brain, which is the best characterized mammalian nervous system at a molecular level. After deriving matrices of type-specific expression profiles from scRNAseq data (8, 12, 24) and the matrix of spatial gene expression information from the AGEA (23), we then solve a linear system of equations for cell-type density per voxel using a nonnegative least-squares algorithm, like prior approaches (22). Two key methodological innovations make the present maps possible. First, we hypothesize that mapping accuracy will improve dramatically if we filter out all "low-information" genes prior to matrix inversion. Although using only a subset of the genes runs counter to similar prior approaches (22, 25), we suspected that the inclusion of genes that are lowly expressed or poorly differential across cell types would deteriorate the quality of the resulting maps. Therefore, we developed an information-theoretic algorithm, Minimal Redundancy–Maximum Relevance–Minimum Residual (MRx3; *Algorithm*), selecting an informative subset of genes to capture cell densities across the mouse brain (*Materials and Methods* has details) and used it to perform subset selection prior to matrix inversion. Second, we formulated several objective metrics to assess the biological accuracy of cell-type maps, relying in part upon regional quantification of individual cell types where available (7). We further evaluated the generalizability of our approach by mapping a larger, more widely sampled, and independently collected scRNAseq dataset (8).

Overall, we demonstrate that both MRx3-based subset selection and matrix inversion are necessary to achieve superior cell-type maps as compared with correlation-based methods and deconvolution/inversion methods without subsetting (22). All cell-type–specific maps across both scRNAseq datasets are available for download along with this methodological pipeline, which we call Matrix Inversion and Subset Selection (MISS). The MISS pipeline is designed to project any arbitrary single-cell expression data onto any arbitrary spatial expression atlas and can be applied to other brains, such as those from macaques and humans. The aim of the current paper is to demonstrate the accuracy of the present maps and the importance of gene subset selection with data-driven approaches, such as our MRx3 algorithm (*Algorithm*) using mouse data where quality assessments are tractable.

Results

Overview of MISS. A schematic of the MISS pipeline is displayed in Fig. 1. We extracted publicly available ISH (23) and scRNAseq data (8, 12, 24) and collected all overlapping genes between scRNAseq and ISH datasets (Fig. 1, step 1). Since nonspecific genes can introduce noise and other artifacts, we developed an information-theoretic algorithm (*Algorithm*) for reordering the genes, referred to herein as "MRx3" (Fig. 1, step 2). After this reordering, we find the elbow of the residual curve, defined as the point closest to the origin, and exclude all genes past this point (Fig. 1, step 3). We then use the solution to the nonnegative matrix inversion using only the genes included in the subset after elbow selection to yield densities per voxel (Fig. 1, steps 4 and 5). Notably, all three-dimensional (3D) whole-brain illustrations of all cell types in this paper are not at individual cell resolution as our method does not produce individual cell locations. Rather, they are voxel-level point cloud illustrations (*Materials and Methods* has details) with the density of points per voxel controlled by the density of that type of cell in that voxel. MISS-inferred densities for all cell types using

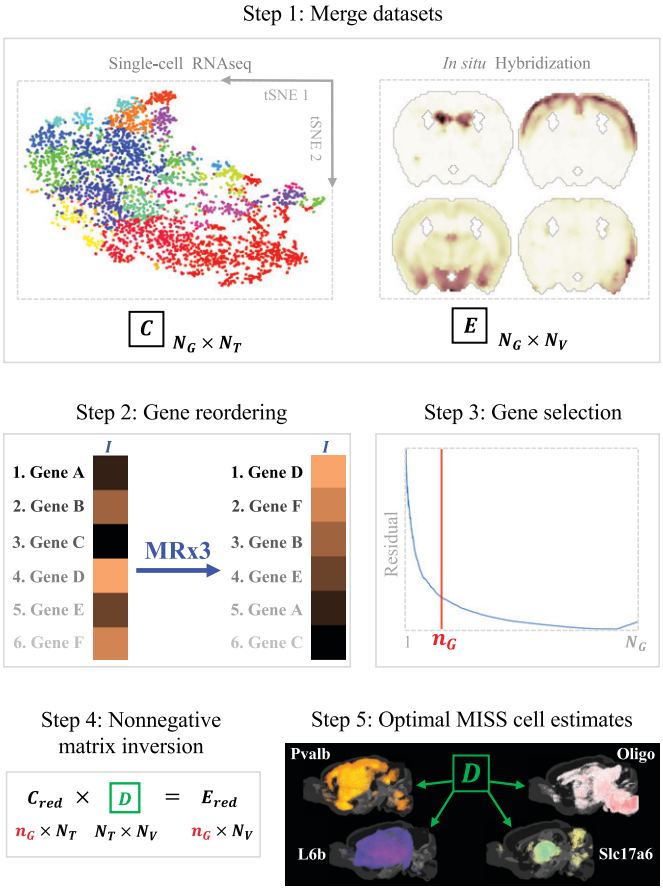

Fig. 1. A visual outline of the MISS pipeline for mapping cell-type clusters. Step 1 consists of finding appropriate scRNAseq cell cluster expression data and combining the data with spatial gene expression data, such as the AIBS gene expression atlas (23) used here. In step 2, the MRx3 algorithm (*Algorithm*) posed here is used to reorder the genes according to information content relevant to the mapping problem at hand. Step 3 chooses a cutoff point in the reordered gene list, and only genes ranked at or above the selected index value, n_G, are used for inversion. This subset selection is accomplished by plotting subset size vs. the residual and then choosing an elbow, defined as the point on the curve closest to the origin. The inversion using only the chosen genes and the MISS-inferred maps produced are steps 4 and 5, respectively.

scRNAseq data from the Allen Institute for Brain Science (AIBS) (12, 24), including all neurons and glia, can be found in Dataset S3. This dataset, which we refer to throughout the paper as "the Tasic et al. (12) dataset," pools data collected from the visual and motor cortices as well as the lateral geniculate complex (LGd) of the thalamus. *Materials and Methods* and *SI Appendix* have a full description of the MISS pipeline and the subset selection procedure. Further methodological details, such as the hierarchical clustering levels at which we map the Tasic et al. (12) scRNAseq dataset and the elbow curves for mapping both scRNAseq datasets, can be found in *SI Appendix*, Fig. 1.

MISS Produces Quantitatively Superior Maps. Fig. 2*A* shows whole-brain illustrations of the Tasic et al. (12) MISS results using the selected MRx3 ordered gene set for *Pvalb*+, *Sst*+, and *Vip*+ interneurons at $n_G = 606$ (*SI Appendix*, Fig. 1*B* shows the elbow curve). We achieve significant quantitative agreement with interneuron densities reported in prior work (7) across the neocortex, with Pearson's $R = 0.84$ and Spearman's $\rho = 0.85$ for *Pvalb*+ cells, $R = 0.52$ and $\rho = 0.59$ for *Sst*+ cells, and $R = 0.63$ and $\rho = 0.66$ for *Vip*+ cells (all $P < 0.001$) (Fig. 2*B*). Although using inversion without performing MRx3 subset selection

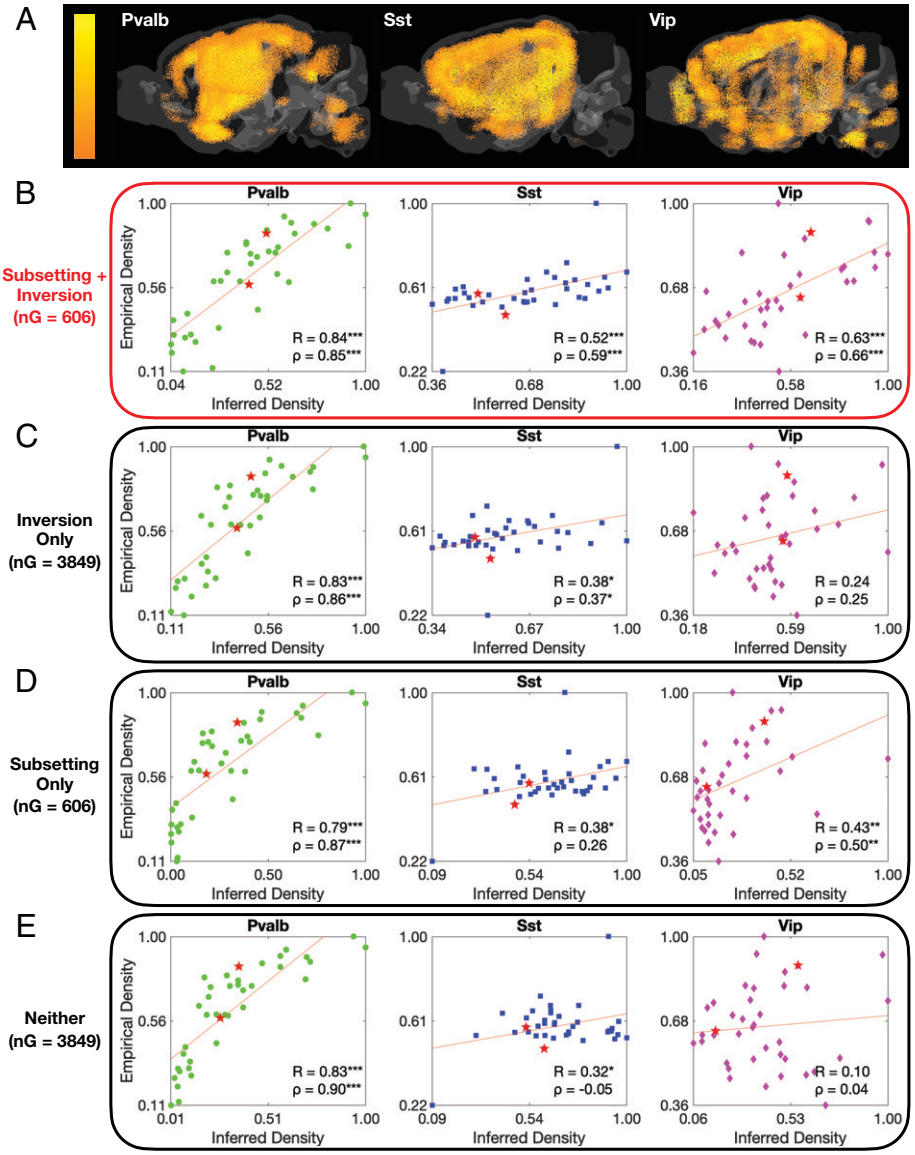

Fig. 2. Matrix inversion after MRx3 gene subset selection produces *Pvalb*+, *Sst*+, and *Vip*+ generally outperforms inversion without subset selection and correlation-based mapping. (*A*) Sagittal axis views of whole-brain MISS maps of *Pvalb*+, *Sst*+, and *Vip*+ interneurons in the mouse CCF (26). Scatterplots depicting correlations between empirical measurements of *Pvalb*+, *Sst*+, and *Vip*+ interneuron densities across neocortical regions (7) and (*B*) MISS estimates, (*C*) matrix inversion without gene subsetting, (*D*) correlation-based mapping using the chosen MRx3 gene subset, and (*E*) correlation-based mapping using all the genes. Red asterisks indicate sampled regions in the scRNAseq dataset (12). *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

yielded maps of similar quality for *Pvalb*+ interneurons ($R = 0.83$, $\rho = 0.86$, $P < 0.001$), it resulted in significantly worse *Sst*+ ($R = 0.38$, $\rho = 0.37$, $P < 0.05$) and *Vip*+ interneuron maps ($R = 0.24$, $\rho = 0.25$, $P > 0.05$) (Fig. 2*C*). The correlation-based mapping procedure also generally performed worse than MISS at recreating *Sst*+ and *Vip*+ interneuron distributions, but performance improved when we excluded genes that failed to satisfy the MRx3 algorithm (*Algorithm* and Fig. 2 *D* and *E*). Overall, our hypothesis that data-driven subset selection with matrix inversion would be important for the quality of cell-type or class maps is confirmed by these interneuron results, as we can only achieve high-accuracy results for all three cell types with MISS.

MISS Layer-Specific Cell-Type Distributions Reproduce Neocortical Laminar Architecture. We next used a metric based on Kendall's τ, τ_{adj} (*Materials and Methods* and *SI Appendix* have details), to compare the ordering of laminar glutamatergic projection neurons in our maps vs. their expected order given identity and sampling location (12). Using our proposed processing pipeline at $n_G = 606$ (*SI Appendix*, Fig. 1*B*) with matrix inversion yields $\tau_{adj} = 0.75$, while matrix inversion without subset selection, $\tau_{adj} = 0.57$, and correlation-based mapping, $\tau_{adj} = 0.56$, both perform worse (Fig. 3*A*). Qualitative assessment finds that layer 2/3 (L2/3) neurons inferred by MISS are most enriched in a band barely inside of the cortical surface. In contrast, L6 neuron enrichment forms a band that traces the interior border between the neocortex and white matter tracts, demarcated by ventricles. L4 and L5 neurons show enrichment in bands that are intermediary between L2/3 and L6 neurons in the expected order (Fig. 3*A*). Notably, maps produced with matrix inversion without subset selection appear to be worse than our MISS maps because they contain more nonneocortical and therefore, off-target estimated cell density for these types (Fig. 3*A*). However, while correlation-based mapping using the subset-selected gene set has very little off-target expression, akin with the MISS maps, it does not produce clearly defined bands for each expected cortical layer, likely explaining its lower τ_{adj} compared with MISS maps (Fig. 3*A*).

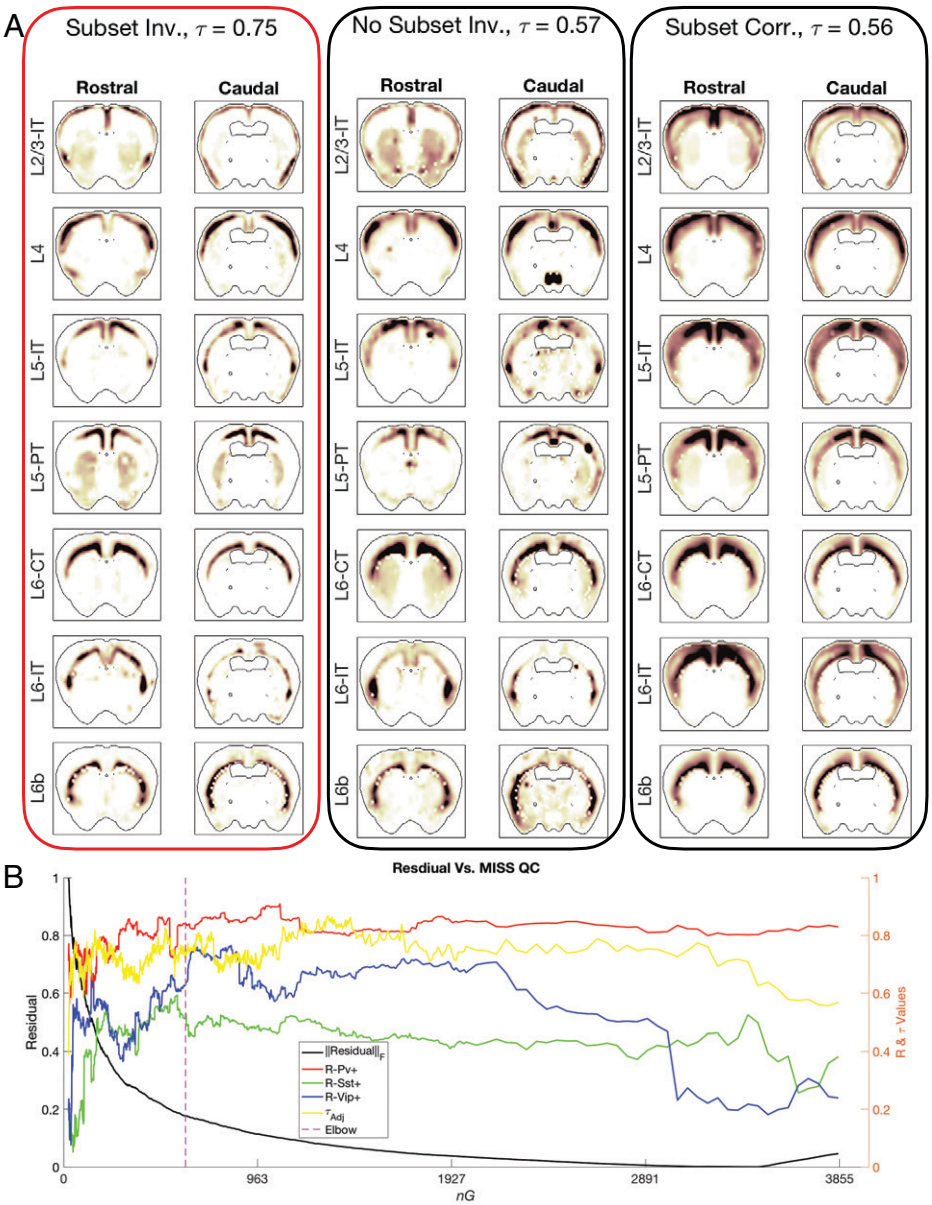

Fig. 3. MISS with MRx3 produces laminar glutamatergic projection neuron maps that generally outperform those produced with inversion without gene subset selection and correlation-based maps. (*A*) The τ_{adj} value for MISS projection neuron maps (*Left*) is better than inversion using all available genes (*Center*) as well as correlation-based mapping using the MRx3-based subset (*Right*). In general, the MISS maps produce clear bands for each projection neuron class in the appropriate cortical layer, which are less clear the correlation-based maps, while there is more significant off-target signal when there is no prior gene subset selection. (*B*) Within a range of about 100 genes on either side of the chosen elbow, both the interneuron *R* values and the laminar excitatory neuron ordering metric τ_{adj} jointly achieve peak or close to peak performance, indicating that the elbow of the residual curve is a suitable ground-truth–independent metric on which we choose gene subset size, n_G.

MISS Provides Stable, Accurate Cell-Type Maps without Notable Overfitting. We note that our methodological decision to choose an elbow n_G based on residual error, calculated as $E_{red} - C_{red} D_F$ and is, therefore, ground truth independent, does not produce a uniquely high-performing gene subset; Fig. 3*B* shows that, relative to the elbow curve, any n_G subset size between ∼520 and ∼675 will outperform using the entire set of genes. This range of values also produces maps that correlate with those using the elbow subset (*SI Appendix*, Fig. 2*C*), indicating that our results are not a product of overfitting. Our maps show no bias toward scRNAseq sampled regions (i.e., the distribution of absolute error [in all voxels] for scRNAseq-sampled and nonsampled regions did not differ) (*SI Appendix*, Fig. 2*B*). Similarly, a spatial map of each region's average per-voxel residual did not highlight sampled regions in a whole-brain 3D illustration (*SI Appendix*, Fig. 2*A*). These results indicate that, at the level of cell-type specificity mapped, types from the scRNAseq dataset were general enough to map across many brain regions without bias toward sampled regions. We further demonstrate the robustness of MRx3 gene reordering, elbow selection, and the validation metrics with bootstrapping (*SI Appendix*, *MISS pipeline: MRx3 with bootstrapping* and Fig. 3).

MISS Maps of Cell Types with Previously Uncharacterized Spatial Distributions Can Help Shed Light on Cell Types' Roles. We also use MISS to produce maps of cell types without prior spatial characterization and that lack informative functional annotations. First, we map all neuronal cell types without prior spatial characterization from the combined Tasic et al. (12) dataset sampled from the thalamus (Fig. 4*A* and *SI Appendix*, Fig. 4) and neocortex (Fig. 4*D* and *SI Appendix*, Fig. 5).

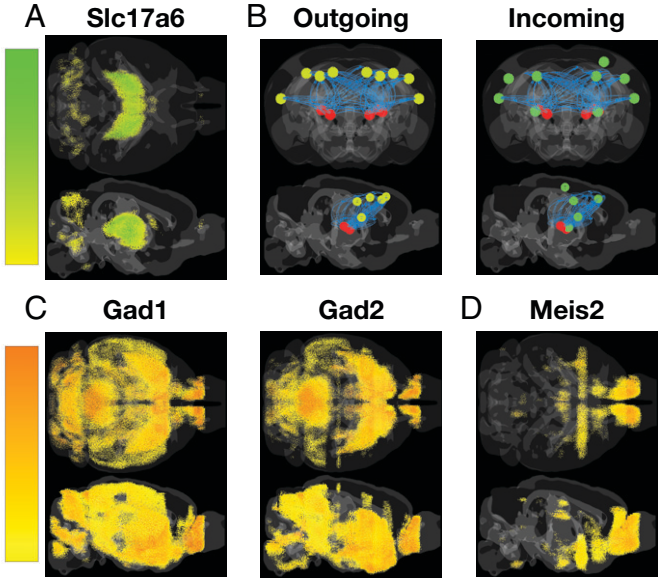

Fig. 4. MISS facilitates the spatial characterization of cell types with unclear functional annotations. (*A*) Axial and sagittal views of the LGd-derived glutamatergic neuron, Slc17a6, with pronounced and specific localization in caudal thalamic nuclei. (*B*) Regions most enriched in Slc17a6 neurons (red spheres) connect most strongly with neocortical regions in both outgoing (yellow spheres) and incoming (green spheres) directions, suggesting that its functionality is likely in constructing or maintaining thalamocortical loops. (*C*) Axial and sagittal views of *Gad1* and *Gad2* expression from the AGEA (23), showing that GABAergic neurons are widely distributed throughout the mouse brain. (*D*) The GABAergic neuron, Meis2, in contrast to *Gad1* and *Gad2* is almost exclusively limited to olfactory areas, parts of the pallidum and amygdala, and several neocortical regions.

Glutamatergic cell type Slc17a6 has a particularly interesting distribution, with expression almost entirely limited to thalamic nuclei (Fig. 4*A*). To probe its potential function, we picked the top five regions of Slc17a6 enrichment and used the Mouse Brain Connectivity Atlas (26) to examine the top two regions with the strongest incoming and outgoing connectivity with Slc17a6-enriched regions. We found that regions with the strongest incoming and outgoing connectivity with Slc17a6-enriched regions were almost entirely neocortical, with no clear bias toward visual areas as opposed to other neocortical regions (red spheres in Fig. 4*B*). We posit that Slc17a6 is likely a thalamic-end neuron in thalamocortical loops. GABAergic neurons from the Tasic et al. (12) dataset also exhibit unique, spatially localized, and bilaterally symmetric enrichment patterns; Meis2 neurons are a particularly pronounced example of this (Fig. 4*D*). These GABAergic neuron maps, and in particular, the Meis2 maps, highlight that inhibitory neuronal subtypes have greater spatial diversity than can be gleaned from paninhibitory markers, such as *Gad1* and *Gad2* (27) (Fig. 4 *C* and *D* and *SI Appendix*, *SI Text* and Figs. 4 and 5). We also mapped the four nonneuronal cell types in the Tasic et al. (12) dataset and demonstrate that their distributions agree with our current understanding of their biological functions (*SI Appendix*, *SI Text* and Fig. 6).

MISS Successfully Maps Cell Types in an Independent scRNAseq Dataset. Despite our successful mapping of the cell types from the Tasic et al. (12) scRNAseq dataset, it remained unclear if the MISS pipeline would generalize to other datasets, particularly those with more numerous and more finely specified cell types. Therefore, we applied MISS to the Zeisel et al. (8) scRNAseq dataset, which contains 200 cell types sampled throughout the entire mouse brain. For comparison, we recreated the maps presented by the original authors using the combined set of differential genes they identified across cell types, which we then correlate at the voxel level with the output of the MISS pipeline using the MRx3-chosen gene set (elbow $n_G = 1{,}360$) (*SI Appendix*, Fig. 1*C*).

We first examined the maps from four individual cell types: telencephalic glutamatergic neuron type 12 (TEGLU12; a telencephalic glutamatergic neuron), CBPC (a cerebellar Purkinje neuron), midbrain dopaminergic neuron type 2 (MBDOP2; a midbrain dopaminergic neuron), and MOL3-enriched oligodendrocytes (MOL3; a type of oligodendrocyte) (Fig. 5 *A*–*D* and *F*). These cell types were selected for two reasons. 1) They represent a sampling of different general cell classes in the brain (excitatory, inhibitory, and modulatory neurons as well as a glial cell), and 2) these cells were sampled from independent regions (i.e., 4 of 12 sampled regions are represented here). Our MISS-derived maps exhibit strong agreement at a per-voxel level with the correlation-based mapping procedure and with expectations of their spatial distributions based on where the cell types were originally sampled (8). Of particular note is the inferred distribution of CBPC neurons, which confirms that they are confined to the cerebellum, as is expected (Fig. 5*F*, second panel from the top). A prior attempt to map Purkinje neurons exhibited significant off-target signal (22), which may have been a limitation of the microarray-based gene expression assay used for that study (28). Generally, we found that the two sets of maps were highly correlated across all types within the several major classes of cell types contained in this dataset (Fig. 5*E*), with overall median and mean R values of 0.56 and 0.54, respectively, at the per-voxel density level. Taken together, despite significant differences in the protocols for mapping, these results demonstrate that MISS faithfully reproduces expected cell-type distributions.

Discussion

Summary of Key Results. We provide a method to accurately infer the per-voxel density of a diverse range of neuronal and nonneuronal cell types from gene expression data at a submillimeter scale at both whole-neocortical and whole-brain levels of coverage. We are able to obtain and evaluate the accuracy of our maps for two key reasons. First, MISS incorporates gene subset selection as an essential preprocessing step, distinguishing it from previous deconvolution approaches for the purpose of mapping cell types (22, 25). We proposed and thoroughly evaluated a subset selection algorithm, MRx3 (*Algorithm*), which outperformed conventional approaches that utilized all available genes. Second, we created evaluation metrics for cell-type maps and show that our inferred maps give strong quantitative agreement with independent literature-derived regional estimates of GABAergic interneurons (7) (Fig. 2) and faithfully reproduce the laminar architecture of the neocortex (Fig. 3). We also demonstrate that MISS can be applied to larger scRNAseq datasets with larger numbers of more finely specified cell types (Fig. 5), generalizing our methodology to other gene expression datasets.

Why Does MRx3-Based Subset Selection Provide the Best Results? Our results depend critically on the quality of gene subset selection. This was a combination of methodologies, as prior subset selection approaches focused on differential expression or using literature-derived marker genes (8, 11, 15–18), and prior mapping attempts using deconvolution or matrix inversion did not employ feature selection (22, 25). Although previous work has suggested that using all available genes

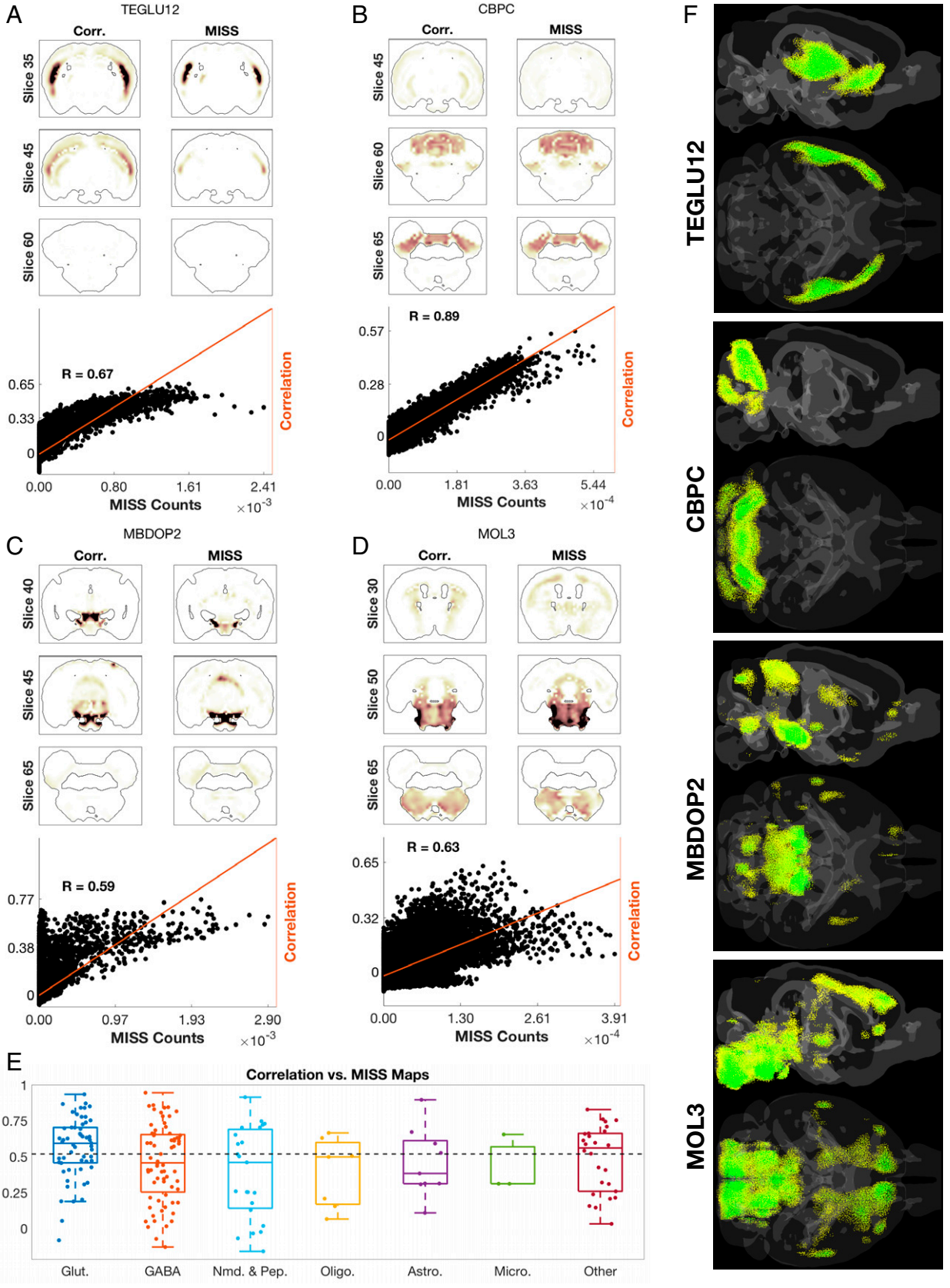

Fig. 5. MISS maps of cell classes using data from a more widely sampled scRNAseq dataset with 200 cell classes sampled from a more comprehensive set of brain regions (8) produce maps that agree with those produced by the original authors. We show comparisons between the two approaches for four distinct cell types: (*A*) TEGLU12, (*B*) CBPC (cerebellar Purkinje neurons), (*C*) MBDOP2, and (*D*) MOL3. There is strong visual agreement between the approach taken by Zeisel et al. (8) and MISS throughout the brain (*Left* and *Right*, respectively, in *A*–*D*) as well as strong quantitative agreement at a per-voxel level (scatterplots). Further, the neuronal cell types exhibit enrichment within the regions from which they were sampled. (*E*) Box plots of correlations between the two approaches per cell type grouped by major cell class. The overall mean and median *R* values across all 200 cell classes from Zeisel et al. (8) were 0.54 and 0.56, respectively.(*F*) Axial and sagittal views of the MISS maps of each of the cell types in *B*–*E*. Glut. - glutamatergic; GABA - GABAergic; Nmd. & Pep. - neuromodulatory and peptidergic; Oligo. - oligodendrocytes; Astro. - astrocytes; Micro. - microglia.

provides good mapping results (22, 25), we find that without filtering of low-information genes, the resulting maps become qualitatively and quantitatively inaccurate in places (Figs. 2 and 3), with significant diffuse abundance patterns that are biologically implausible (Fig. 3*A*, second column). The issue is exacerbated by unsampled cell types in the AIBS data used here (12, 24), since a matrix with missing cell types in Eq. **2** may potentially lead to error in the least-squares solution, particularly in regions anatomically dissimilar to those sampled. Our subset selection step helps ensure that unsampled cell types do not appreciably contaminate the inference of sampled ones, as it specifically selects only the genes most relevant to cells in the scRNAseq datasets. Additionally, many genes neither are specific to the central nervous system, nor do they show appreciable gradients across the brain. The inclusion of such genes can lead to diffuse effects in inferred maps without contributing any useful signal. However, we note that driving feature selection to an extreme is also suboptimal. Small numbers of genes do not give a good performance in R or τ_{adj} values (Fig. 3*B*), and there is apparent strong cross–cell-type expression of some genes in the scRNAseq data (*SI Appendix*, Fig. 7; gene names are in Dataset S2), indicating that the best performance range is achieved at an intermediate subset of genes, whose identification is not trivial.

There are several reasons why MRx3 in particular may yield strong results. While feature selection approaches generally rely upon a measure of differential expression as a criterion for selecting high-information genes, there are other important considerations for determining an optimal gene set for the purposes of matrix-inversion–based inference of cell counts. High-information genes are not homogeneously distributed between cell types, meaning that a simple filtering that yields one or a small number of genes per cell type does not perform well (Fig. 3*B*). The minimum redundancy criterion contained within the Maximum Relevance-Minimum Redundancy (mRMR) (29) and MRx3 algorithms (*Algorithm*) circumvents this issue by preventing genes from being added if their expression profiles between cell types are too similar to other genes already selected. MRx3 goes a step further than mRMR by also including a minimum residual criterion that prevents genes that, if added, would result in high degrees of error when reconstructing the spatial gene expression matrix from the scRNAseq data and inferred cell-type densities. Such genes may be highly noisy in the ISH expression atlas even if they have high information content in the scRNAseq dataset and will, therefore, lead to unstable or inaccurate results after matrix inversion (note the sharp increase in residual as those high-noise genes are added back into the inversion near the end of the curve in Fig. 3*B*). The net effect of removing them, as MRx3 does, is to produce high-information gene sets specifically for the purpose of generating quantitatively validated cell-type maps.

Further Uses and Potential Applications of the MISS Pipeline. Key advantages of MISS are its flexibility and low input cost to generate results. First, MISS is computationally inexpensive and fast, as it performs linear inference on metadata rather than a time- and labor-intensive microscopy and image processing pipeline. This allows us to run the entire pipeline from start to finish to generate our cell-type maps, including visualization, in hours on a laptop with a 3.1-GHz processor and 16 GB of random access memory (RAM). The method achieves strong fits to empirical data despite the fact that choosing the elbow n_G is dependent only on the input datasets; furthermore, this elbow falls in a range of possible n_G values that yield quantitatively strong results. We, therefore, anticipate that as more scRNAseq data becomes available, users will be able to implement the MISS pipeline directly to generate whole-brain maps of yet more cell types; we make all code publicly available on GitHub. Future work includes utilizing the MISS pipeline to infer cell densities in the human brain using human scRNAseq datasets mapped onto the AIBS human gene expression atlas (30). A computational approach for cell-type mapping is particularly appealing in humans, where brain size and tissue accessibility make experimental techniques pioneered in mice prohibitively expensive and time consuming.

Inferred cell-type maps from MISS can also help address the extent to which cellular identities of brain regions govern the formation of synaptic connections (31) and interregional neural connectivity (32, 33). Questions surrounding whether certain behavioral, cognitive, or sensory processing abilities are correlated with certain cell types, their spatial distribution, or their location within connectivity networks (34–37) can also benefit from our maps. Clinically, MISS could further understanding of the selective vulnerability of brain regions, such as the entorhinal cortex to early τ-inclusions in Alzheimer's disease (38) or the substantia nigra pars compacta to early synuclein inclusions in Parkinson's disease (39). The spatially varying abundances of cell types considered selectively vulnerable to τ- or α-synuclein inclusions can be mapped using MISS, and their correspondence with the spatial pattern of protein pathologies can be tested; for instance, recent experiments suggest cell-type selectivity of τ-pathology (1, 2, 6). Cellular vulnerability in other neurological conditions can also be interrogated using MISS, including psychiatric diseases, such as schizophrenia (4), and traumatic brain injury, which was hypothesized to preferentially involve certain types of cells in both injury and recovery phases (3, 5).

Methodological Limitations of the MISS Pipeline. The most significant limitation is that the scales of cell densities presented in Datasets S3 and S4 are reliable across voxels but are not fully so across cell types; it is possible that a per–cell-type scaling factor could address this issue. In future work, we plan to utilize Nissl and 4′,6-diamidino-2-phenylindole (DAPI) stains coregistered to the mouse common coordinate framework (CCF) using the connectivity atlas parcellation (26) to infer actual counts per voxel. There is also a risk of off-target predictions of cell types due to two confounding factors. First, our method cannot differentiate between gene expression signal in the ISH atlas coming from somatodendritic compartments of cells as opposed to their axon terminals. Second, as not all cell types will be contained in any single scRNAseq dataset, cells not included in a dataset but which have similar gene expression profiles to an included cell could be erroneously mapped. Finally, there is a risk of both overfitting and of choosing a suboptimal number of genes given our selection procedure. However, our selection procedure chooses a value in the middle of a range of values that produce high correlation values between proposed interneuron maps and empirical density data in the neocortex (Fig. 2) as well as faithful recreations of the laminar patterns of neocortical projection neurons (Fig. 3). Furthermore, we find no bias toward lower residuals (that is, $E_{red} - C_{red} D_F$) in sampled vs. unsampled regions, especially within the neocortex (*SI Appendix*, Fig. 2*A*). The strength of our results indicates that, despite these limitations, we can reproduce cell densities at per-voxel resolution using MISS with superior accuracy.

Conclusions. We propose a computational pipeline for high-accuracy, per-voxel cell-type density inference using ISH and scRNAseq data across the entire mouse brain. Our results demonstrate that verifiable mapping of neuronal and glial subpopulations

with well-differentiated glutamatergic and GABAergic subpopulations can be obtained using relatively small numbers of cell types and sampled brain regions. Most importantly, we demonstrate that data-driven gene subset selection prior to cell-type mapping, which we accomplish with our MRx3 algorithm (*Algorithm*), is vital for producing more accurate maps. Furthermore, we find that matrix inversion is superior to correlation-based mapping procedures for yielding accurate cell-type distributions but that subset selection was independently important for improving cell-type map accuracy regardless of mapping method. We also show that MISS can be applied to other mouse scRNAseq datasets, demonstrating the generalizability of the pipeline. The presented maps and computational pipeline can be used as an inexpensive alternative to single-cell counting for generating density distributions of more cell types than current whole-brain approaches can readily accommodate.

Materials and Methods

Input Datasets. Cell-type–specific gene expression data came from two publicly available sources of scRNAseq data: 1) the combined Tasic et al. (12) dataset, which contains cell types sampled from the anterior lateral motor cortex, the primary visual cortex (12), and the dorsal part of the LGd (24), and 2) the Zeisel et al. (8) dataset, which contains cell types sampled from 12 locations around the whole mouse brain. Spatial gene expression data came from the coronal AGEA (23). For connectivity analyses, we use the AIBS mesoscale mouse connectome (26). A complete description of our data preprocessing pipeline is in *SI Appendix*.

MISS Pipeline. Grange et al. (22) first introduced a mathematical framework for inferring voxel-wise cell-type densities, positing that ISH voxel energy for a given gene is proportional to each cell type's expression value for that gene multiplied by its density in that voxel, summed across all cell types within that voxel. Similar recent work used scRNAseq profiles to infer cell densities within spatial RNA-seq sampled areas (25).

We can generalize the relationship between expression energy per voxel and cell-type–specific gene expression across all voxels and genes in matrix form as follows:

$$CD = E, \qquad [1]$$

where E is the row-normalized genes by voxels ($N_G \times N_V$) expression matrix extracted from the ISH data, C is the row-normalized genes by cell types ($N_G \times N_T$) expression matrix extracted from the scRNAseq data, and D is a cell types by voxels ($N_T \times N_V$) matrix of densities. We find the solution to Eq. **1** in a least-squares sense, $\tilde{D}$, using the nonnegativity constrained lsqnonneg function in MATLAB:

$$\tilde{D} = \underset{D,\ D_{ij} \geq 0\ \forall i,j}{\arg\min} \|E - CD\|_F. \qquad [2]$$

MRx3-Based Gene Subset Selection. Previous deconvolution-based mapping utilized all available genes, which could introduce noise into the predicted cell-type densities. Here, we introduce a subset selection procedure, MRx3. The first two components of MRx3 are derived from the mRMR algorithm (29), while the third prevents genes that contribute most to reconstruction residual error, $\|E - C \cdot D_F\|$, from being added to the subset. What follows is a brief description of the MRx3 algorithm (*Algorithm*); *SI Appendix* has a more in-depth explanation.

Prior to performing MRx3, we make a cell size correction by dividing each column of C by its mean, yielding the column-normalized gene expression matrix, C_{col}. We then define the maximum relevance criterion, F_i, for any candidate gene $i \in$ gene set G using the following formula:

$$F_i = \sum_{j=1}^{N_T} \left(\frac{\left(C_{\text{col}}(i,j) - \overline{C_{\text{col}}(i,:)} \right)^2}{N_T - 1} \right), \qquad [3]$$

where $C_{\text{col}}(i,j)$ is the column-normalized expression of gene i of cell-type j, N_T is the number of cell types (columns) within C, and $\overline{C_{\text{col}}(i,:)}$ is the mean gene expression of gene i across all cell types following the column normalization of C.

The minimum redundancy component is given by the mean of the absolute value of the Pearson's correlation between the candidate gene i and the already-selected genes within set S:

$$\text{Redund}(i|S) = \frac{1}{|S|} \sum_{j \in S} |R(i,j)|, \qquad [4]$$

where $|S|$, the cardinality of set S, is the number of genes already selected and $R(i,j)$ is the Pearson's correlation between genes i and j in S.

For the minimum residual component, we approximate per-gene residual error, ε_i, using the Sherman–Morrison rank-1 update rule (40):

$$\varepsilon_i = \|e_i - C_{\sim i}(C_{\sim i}^T C_{\sim i})^{-1} C_{\sim i}^T e_i\|_2^2 + \|e_i\|_2^2 \cdot \|(C_{\sim i}^T C_{\sim i})^{-1} c_i\|_2^2, \qquad [5]$$

where $c_i = C(g_i, :)$ and $e_i = E(g_i, :)$.

We denote the following algorithm to perform MRx3 gene reordering (Fig. 1, step 2).

Algorithm: MRx3

Result: S, an n_G-element set of MRx3-selected genes
for each gene i of full gene set G, **do**
  Let
$\varepsilon_i = \|e_i - C_{\sim i}(C_{\sim i}^T C_{\sim i})^{-1} C_{\sim i}^T e_i\|_2^2 + \|e_i\|_2^2 \cdot \|(C_{\sim i}^T C_{\sim i})^{-1} c_i\|_2^2$
end
Let G_{90} be the subset of G at or above the 90$^{\text{th}}$ percentile of ε_i;
Let $\hat{G} = G \setminus G_{90}$;
Initialize $S_0 = \emptyset$ and $k = 1$;
while $|S| \leq n_G$, **do**
  Let $g_k = \text{argmax}_{i \in \hat{G} \setminus S_{k-1}} \frac{F_i}{Redund(i|S_{k-1})}$;
  Let $S_k = \{S_{k-1} \bigcup g_k\}$;
  Let $k = k + 1$;
end

After reordering the genes in the intersection set between the scRNAseq datasets and the AGEA, we then choose a gene subset size, n_G, that balances minimizing reconstruction residual error and number of included genes (Fig. 1, step 3). This allows us to create reduced matrices C_{red} and E_{red}, which contain only MRx3-selected genes, and then, solve the nonnegative least-squares problem posed by Eq. **2**, substituting C_{red} and E_{red} for C and E, respectively (Fig. 1, step 4).

Method Validation. We validate MISS-derived cell-type maps using quantitative comparisons. The ground truth *Pvalb*+, *Sst*+, and *Vip*+ interneuron distributions in the neocortex come from prior work directly imaging cells positive for these markers via a *Cre*-based expression system (7). The calculation of the layer-ordering metric, τ_{adj}, is described fully in *SI Appendix*. Briefly, we skeletonize maps of Tasic et al. (12) layer-specific glutamatergic neuronal types (per coronal slice containing neocortex) and rank them by mean distance between their bands and the cortical surface. These are then correlated to the expected ordering based on the Tasic et al. (12) ontology using Kendall's τ.

Data Availability. All code and data used for running the MISS pipeline are hosted on GitHub (https://github.com/Raj-Lab-UCSF/MISS-Pipeline). Previously published data were also used for this work (https://portal.brain-map.org/atlases-and-data/rnaseq and http://mousebrain.org).

ACKNOWLEDGMENTS. We acknowledge Dr. Pablo Damasceno for his help with subset selection algorithms and Dr. Chuying Xia for her prior work in the laboratory organizing and formatting the raw Allen Institute Mouse Gene Expression Atlas ISH data. NIH Grants R01NS092802, RF1AG062196, R56AG064873, and R01AG072753 provided funding support for this project.

Author affiliations: [a]Department of Radiology, University of California, San Francisco, CA 94143; [b]Department of Radiology, Weill Cornell Medicine of Cornell University, New York, NY 10065; [c]Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019; and [d]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

1. H. Fu, J. Hardy, K. E. Duff, Selective vulnerability in neurodegenerative diseases. *Nat. Neurosci.* **21**, 1350–1358 (2018).
2. H. Fu *et al.*, A tau homeostasis signature is linked with the cellular and regional vulnerability of excitatory neurons to tau pathology. *Nat. Neurosci.* **22**, 47–56 (2019).
3. D. Arneson *et al.*, Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nat. Commun.* **9**, 3894 (2018).
4. N. G. Skene *et al.*, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
5. K. V. Rama Rao *et al.*, A single primary blast-induced traumatic brain injury in a rodent model causes cell-type dependent increase in nicotinamide adenine dinucleotide phosphate oxidase isoforms in vulnerable brain regions. *J. Neurotrauma* **35**, 2077–2090 (2018).
6. C. R. Muratore *et al.*, Cell-type dependent Alzheimer's disease phenotypes: Probing the biology of selective neuronal vulnerability. *Stem Cell Reports* **9**, 1868–1884 (2017).
7. Y. Kim *et al.*, Brain-wide maps reveal stereotyped cell-type-based cortical architecture and subcortical sexual dimorphism. *Cell* **171**, 456–469.e22 (2017).
8. A. Zeisel *et al.*, Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018).
9. T. C. Murakami *et al.*, A three-dimensional single-cell-resolution whole-brain atlas using CUBIC-X expansion microscopy and tissue clearing. *Nat. Neurosci.* **21**, 625–637 (2018).
10. X. Wang *et al.*, Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
11. C. Erö, M.-O. Gewaltig, D. Keller, H. Markram, A cell atlas for the mouse brain. *Front. Neuroinform.* **12**, 84 (2018).
12. B. Tasic *et al.*, Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
13. S. Codeluppi *et al.*, Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
14. J. R. Moffitt *et al.*, Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
15. S. R. Vincent, H. Kimura, Histochemical mapping of nitric oxide synthase in the rat brain. *Neuroscience* **46**, 755–784 (1992).
16. A. Björklund, S. B. Dunnett, Dopamine neuron systems in the brain: An update. *Trends Neurosci.* **30**, 194–202 (2007).
17. A. Pazos, R. Cortés, J. M. Palacios, Quantitative autoradiographic mapping of serotonin receptors in the rat brain. II. Serotonin-2 receptors. *Brain Res.* **346**, 231–249 (1985).
18. K. T. Beier *et al.*, Circuit architecture of VTA dopamine neurons revealed by systematic input-output mapping. *Cell* **162**, 622–634 (2015).
19. X. Li *et al.*, Generation of a whole-brain atlas for the cholinergic system and mesoscopic projectome analysis of basal forebrain cholinergic neurons. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 415–420 (2018).
20. K. Achim *et al.*, High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
21. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
22. P. Grange *et al.*, Cell-type-based model explaining coexpression patterns of genes in the brain. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5397–5402 (2014).
23. E. S. Lein *et al.*, Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
24. Allen Cell Types Database, "Allen Cell Types Database Technical White Paper: Transcriptomics" (Allen Cell Types Database, 2018), v.7, pp. 1–14, https://help.brain-map.org/display/celltypes/Documentation. Accessed 24 March 2022.
25. A. Andersson *et al.*, Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 565 (2020).
26. S. W. Oh *et al.*, A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
27. M. G. Erlander, N. J. Tillakaratne, S. Feldblum, N. Patel, A. J. Tobin, Two genes encode distinct glutamate decarboxylases. *Neuron* **7**, 91–100 (1991).
28. B. W. Okaty, K. Sugino, S. B. Nelson, A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. *PLoS One* **6**, e16493 (2011).
29. H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005).
30. M. J. Hawrylycz *et al.*, An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
31. H. S. Meyer *et al.*, Cell type-specific thalamic innervation in a column of rat vibrissal cortex. *Cereb. Cortex* **20**, 2287–2303 (2010).
32. E. G. Jones, P. Rakic, Radial columns in cortical architecture: It is the composition that counts. *Cereb. Cortex* **20**, 2261–2264 (2010).
33. J. Szentágothai, The 'module-concept' in cerebral cortex architecture. *Brain Res.* **95**, 475–496 (1975).
34. L. Pinto, Y. Dan, Cell-type-specific activity in prefrontal cortex during goal-directed behavior. *Neuron* **87**, 437–450 (2015).
35. T. K. Roseberry *et al.*, Cell-type-specific control of brainstem locomotor circuits by basal ganglia. *Cell* **164**, 526–537 (2016).
36. Y. Senzai, G. Buzsáki, Physiological properties and behavioral correlates of hippocampal granule cells and mossy cells. *Neuron* **93**, 691–704.e5 (2017).
37. T. Sippy, D. Lapray, S. Crochet, C. C. H. Petersen, Cell-type-specific sensorimotor processing in striatal projection neurons during goal-directed behavior. *Neuron* **88**, 298–305 (2015).
38. H. Braak, E. Braak, Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259 (1991).
39. H. Braak *et al.*, Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol. Aging* **24**, 197–211 (2003).
40. J. Sherman, W. J. Morrison, Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Stat.* **21**, 124–127 (1950).