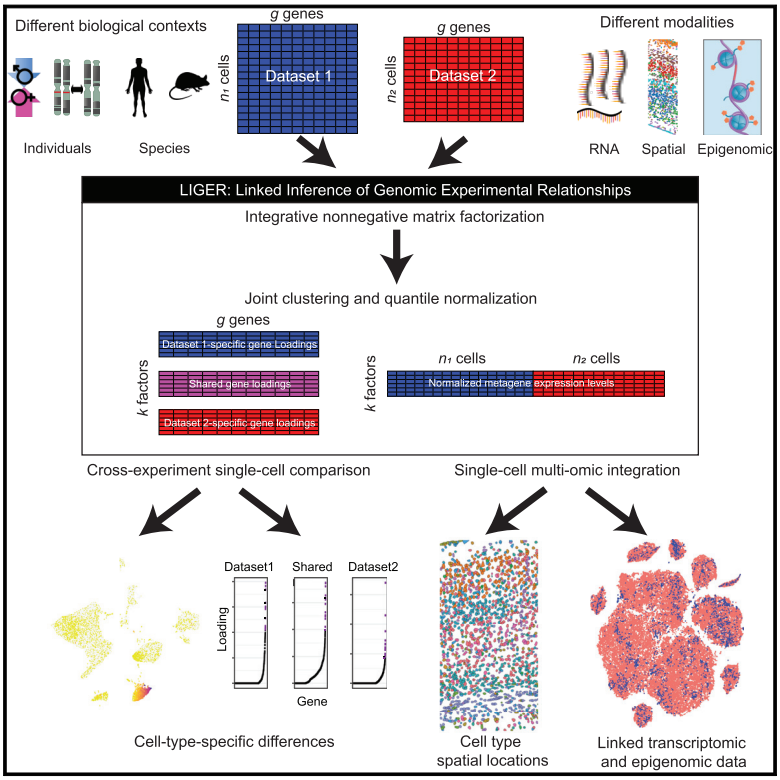
Resource

Cell

Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity

Graphical Abstract



Highlights

- Shared and dataset-specific metagene factors enable single-cell data integration
- LIGER reveals inter-individual differences in bed nucleus and substantia nigra cells
- Integration of *in situ* and dissociated scRNA-seq maps cell types in space
- Joint definition of cortical cell types from single-cell RNA and epigenome profiles

Authors

Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, Evan Z. Macosko

Correspondence

jwelch@broadinstitute.org (J.D.W.), emacosko@broadinstitute.org (E.Z.M.)

In Brief

A platform called LIGER allows for the integration of gene expression, epigenetic regulation, and spatial relationships across single-cell datasets.

Welch et al., 2019, Cell *177*, 1873–1887
June 13, 2019
https://doi.org/10.1016/j.cell.2019.05.006

CellPress

Resource

Cell

Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity

Joshua D. Welch,[1,3,*] Velina Kozareva,[1] Ashley Ferreira,[1] Charles Vanderburg,[1] Carly Martin,[1] and Evan Z. Macosko[1,2,4,*]
[1]Broad Institute of Harvard and MIT, Stanley Center for Psychiatric Research, 450 Main Street, Cambridge, MA, USA
[2]Massachusetts General Hospital, Department of Psychiatry, 55 Fruit Street, Boston, MA, USA
[3]Present address: University of Michigan, Department of Computational Medicine and Bioinformatics, 100 Washtenaw Avenue, Ann Arbor, MI, USA
[4]Lead Contact
*Correspondence: jwelch@broadinstitute.org (J.D.W.), emacosko@broadinstitute.org (E.Z.M.)
https://doi.org/10.1016/j.cell.2019.05.006

SUMMARY

Defining cell types requires integrating diverse single-cell measurements from multiple experiments and biological contexts. To flexibly model single-cell datasets, we developed LIGER, an algorithm that delineates shared and dataset-specific features of cell identity. We applied it to four diverse and challenging analyses of human and mouse brain cells. First, we defined region-specific and sexually dimorphic gene expression in the mouse bed nucleus of the stria terminalis. Second, we analyzed expression in the human substantia nigra, comparing cell states in specific donors and relating cell types to those in the mouse. Third, we integrated *in situ* and single-cell expression data to spatially locate fine subtypes of cells present in the mouse frontal cortex. Finally, we jointly defined mouse cortical cell types using single-cell RNA-seq and DNA methylation profiles, revealing putative mechanisms of cell-type-specific epigenomic regulation. Integrative analyses using LIGER promise to accelerate investigations of cell-type definition, gene regulation, and disease states.

INTRODUCTION

The function of the mammalian brain is dependent upon the coordinated activity of highly specialized cell types. Advances in high-throughput single-cell RNA sequencing (scRNA-seq) analysis (Klein et al., 2015; Macosko et al., 2015; Rosenberg et al., 2018; Zheng et al., 2017) have provided an unprecedented opportunity to systematically identify these cellular specializations, across multiple regions (Saunders et al., 2018; Tasic et al., 2016; Zeisel et al., 2018), in the context of perturbations (Hrvatin et al., 2018), and in related species (Hodge et al., 2018; Lake et al., 2016; Tosches et al., 2018). Furthermore, new technologies can now measure DNA methylation (Luo et al., 2017; Mulqueen et al., 2018), chromatin accessibility (Cusanovich et al., 2018), and *in situ* expression (Coskun and Cai, 2016; Moffitt and Zhuang, 2016; Wang et al., 2018), in thousands to millions of cells. Each of these experimental contexts and measurement modalities provides a different glimpse into cellular identity.

Integrative computational tools that can flexibly combine individual single-cell datasets into a unified, shared analysis offer many exciting biological opportunities. The major challenge of integrative analysis lies in reconciling the immense heterogeneity observed across individual datasets. Within one modality of measurement—like scRNA-seq—datasets may differ by many orders of magnitude in the number of cells sampled, or in the depth of sequencing allocated to each cell. Across modalities, datasets may vary widely in dynamic range (gene expression versus chromatin accessibility), direction of relationship (RNA-seq versus DNA methylation), or in the number of genes measured (targeted quantification versus unbiased approaches). To date, the most widely used data alignment approaches (Butler et al., 2018; Haghverdi et al., 2018; Johnson et al., 2007; Risso et al., 2014) implicitly assume that the differences between datasets arise entirely from technical variation and attempt to eliminate them or map datasets into a completely shared latent space using dimensions of maximum correlation (Butler et al., 2018). However, in many kinds of analysis, both dataset similarities and differences are biologically important, such as when we seek to compare and contrast scRNA-seq data from healthy and disease-affected individuals.

To address these challenges, we developed a new computational method called LIGER (linked inference of genomic experimental relationships). We show here that LIGER enables the identification of shared cell types across individuals, species, and multiple modalities (gene expression, epigenetic, or spatial data), as well as dataset-specific features, offering a unified analysis of heterogeneous single-cell datasets.

RESULTS

Comparing and Contrasting Single-Cell Datasets with Shared and Dataset-Specific Factors

LIGER takes as input multiple single-cell datasets, which may be scRNA-seq experiments from different individuals, time points, or species—or measurements from different molecular modalities, such as single-cell epigenome data or spatial gene expression data (Figure 1A). LIGER then employs integrative non-negative matrix factorization (iNMF) (Yang and Michailidis, 2016) to learn a low-dimensional space in which each cell is defined by one set of dataset-specific factors, or metagenes, and another set of

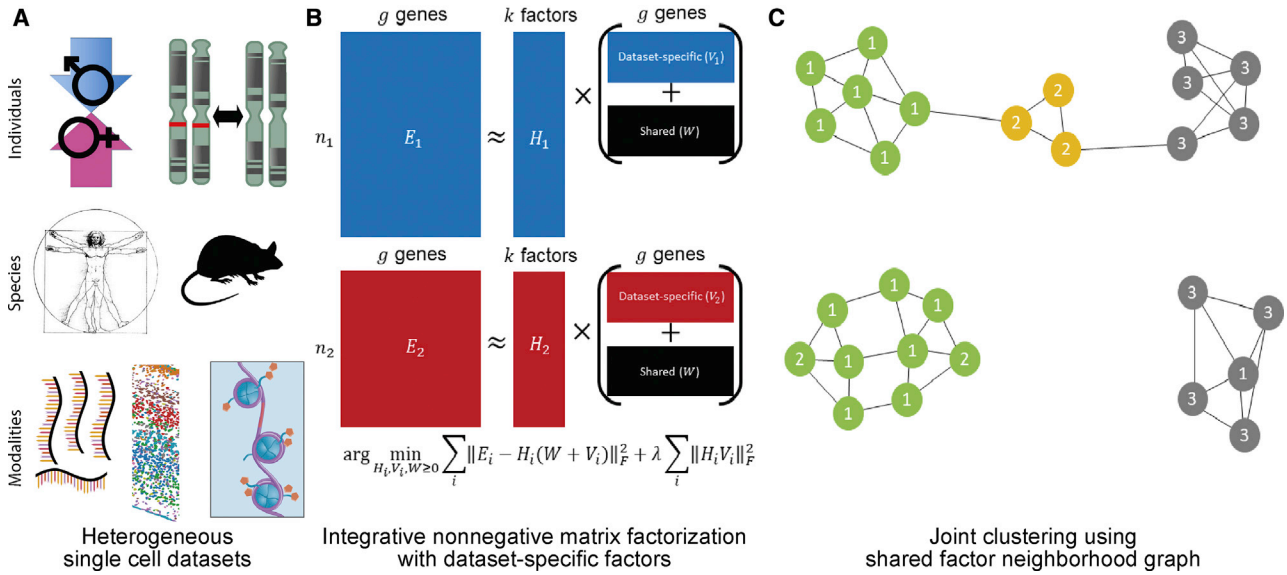


Figure 1. LIGER Approach to Integration of Highly Heterogeneous Single-Cell Datasets

(A) LIGER takes as input two or more datasets, which may come from different individuals, species, or modalities, that share corresponding gene-level features.

(B) Integrative nonnegative matrix factorization (Yang and Michailidis, 2016) identifies shared and dataset-specific metagenes across datasets.

(C) Building a graph in the resulting factor space, based on comparing neighborhoods of maximum factor loadings (STAR Methods). Each cell is numbered by its maximum factor loading and connected to its nearest neighbors within each dataset. The shared factor neighborhood graph leverages the factor loading values of neighboring cells to prevent the spurious integration of divergent cell types across datasets (such as the yellow cells shown).

See also Figure S1.

shared metagenes (Figure 1B). Each factor often corresponds to a biologically interpretable signal—like the genes that define a particular cell type. A tuning parameter, λ, allows adjusting the size of dataset-specific effects to reflect the divergence of the datasets being analyzed. We found that iNMF performs comparably to both NMF and principal-component analysis (PCA) in reconstructing the original data (Figures S1A and S1B). After performing iNMF, we use a novel strategy that increases robustness of joint clustering. We first assign each cell a label based on the maximum factor loading and then build a shared factor neighborhood graph (Figure 1C), in which we connect cells that have similar factor loading patterns (STAR Methods).

We derived a novel algorithm for iNMF optimization, which scales well with the size of large single-cell datasets (Figures S1C and S1D; STAR Methods). To aid selection of the key parameters—the number of factors k and the tuning parameter λ—we developed heuristics based on factor entropy and dataset alignment (STAR Methods). Overall, these heuristics performed well across different analyses (Figures S1I and S1J), though we have observed that manual tuning can sometimes improve the results. Additionally, we derived novel algorithms for rapidly updating the factorization to incorporate new data or change parameters (STAR Methods; Figures S1E–S1H). We anticipate that this capability will be useful for leveraging a rapidly growing corpus of single-cell data.

Liger Shows Robust Performance on Highly Divergent Datasets

We assessed the performance of LIGER through the use of two metrics: alignment and agreement. Alignment (Butler et al., 2018) measures the uniformity of mixing for two or more samples in the aligned latent space. This metric should be high when datasets share underlying cell types, and low when datasets do not share cognate populations. The second metric, agreement, quantifies the similarity of each cell's neighborhood when a dataset is analyzed separately versus jointly with other datasets. High agreement indicates that cell-type relationships are preserved with minimal distortion in the joint analysis.

We calculated alignment and agreement metrics using published datasets (Baron et al., 2016; Gierahn et al., 2017; Saunders et al., 2018), comparing the LIGER analyses to those generated by the Seurat package (Butler et al., 2018). We first ran our analyses on a pair of scRNA-seq datasets from human blood cells that show primarily technical differences (Gierahn et al., 2017) and should thus yield a high degree of alignment. Indeed, LIGER and Seurat show similarly high alignment statistics (Figures 2A–2C), and LIGER's joint clusters match the published cluster assignments for the individual datasets. LIGER and Seurat also performed similarly when integrating human and mouse pancreatic data, with LIGER showing slightly higher alignment (Figure 2C).

In both analyses, LIGER produced considerably higher agreement than Seurat (Figure 2D), suggesting better preservation of the underlying cell-type architectures in the integrated space. We expected this advantage should be especially beneficial when analyzing very divergent datasets that share few or no common cell populations. To confirm this, we jointly analyzed profiles of hippocampal oligodendrocytes and interneurons (Saunders et al., 2018), two cell classes with very different developmental origins. LIGER generated minimal false alignment

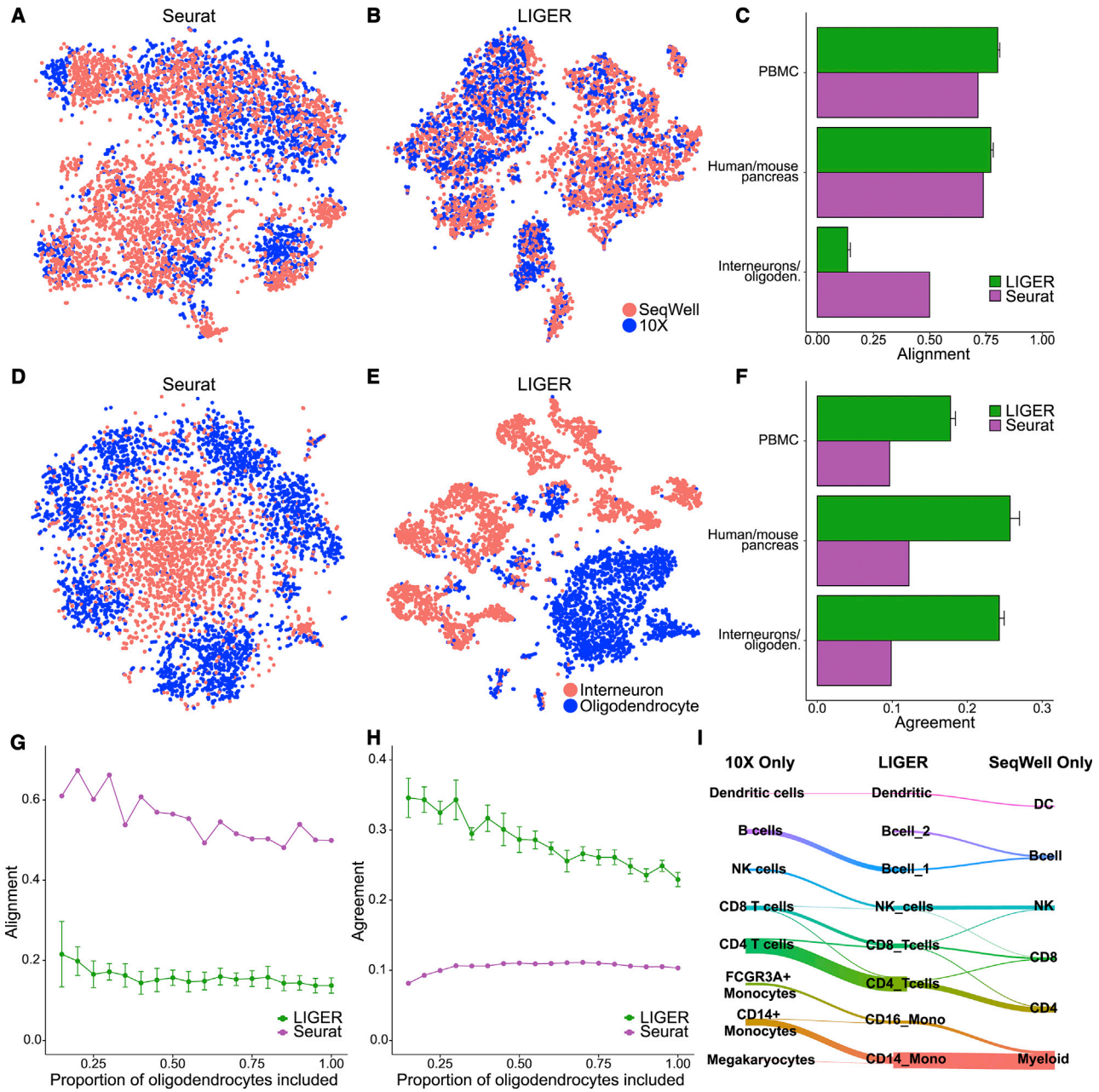


Figure 2. Benchmarking LIGER Performance

(A and B) t-SNE visualizations of Seurat (Butler et al., 2018) (A) and LIGER (B) analyses of two scRNA-seq datasets prepared from human blood cells.

(C) Alignment metrics for the Seurat and LIGER analyses of the human blood cell datasets, human and mouse pancreas datasets, and hippocampal interneuron and oligodendrocyte datasets. Error bars on the LIGER data points represent 95% confidence intervals across 20 random iNMF initializations.

(D and E) t-SNE visualizations of Seurat (D) and LIGER (E) analyses of 3,212 hippocampal interneurons and 2,524 oligodendrocytes. Note the small shared population of doublets in the middle of the t-SNE, highlighting LIGER's ability to identify rare populations.

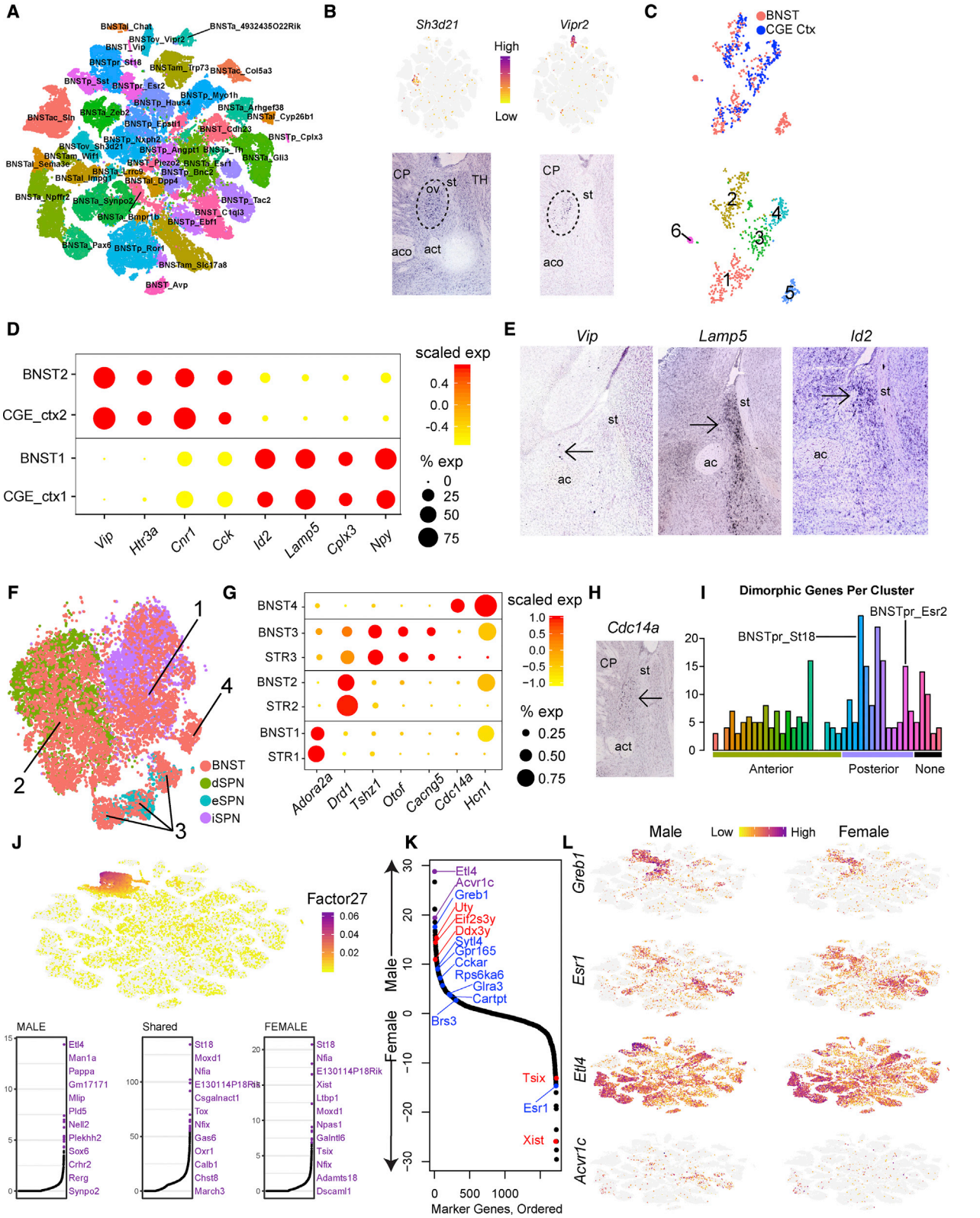
(F) Agreement metrics for Seurat and LIGER analyses of the datasets listed in (C).

(G and H) Alignment (G) and agreement (H) for varying proportions of oligodendrocytes mixed with a fixed number of interneurons.

(I) Riverplot comparing the previously published clustering results for each blood cell dataset with the LIGER joint clustering assignments.

See also Figure S2.

between these classes and demonstrated a good preservation of complex internal substructure (Figures 2D–2F and S2A–S2C), even across considerable changes in dataset proportion (Figures 2G and 2H). In each of the three analyses described above, the LIGER joint clustering result closely matched the published cluster assignments for the individual datasets (Figures 2I,



(legend on next page)

S2D, and S2E). Together, these analyses indicate that LIGER sensitively detects common populations without spurious alignment and preserves complex substructure, even when applied across divergent datasets.

Interpretable Factors Unravel Complex and Dimorphic Expression Patterns in the Bed Nucleus

An important application of integrative single-cell analysis in neuroscience is to quantify cell-type variation across different brain regions and different members of the same species. To examine LIGER's performance in these tasks, we analyzed the bed nucleus of the stria terminalis (BNST), a subcortical region composed of multiple subnuclei (Dong and Swanson, 2004) implicated in social, stress-related, and reward behaviors (Bayless and Shah, 2016). To date, scRNA-seq has not been performed on BNST, providing an opportunity to clarify how cell types are shared between this structure and datasets generated from related tissues.

We isolated, sequenced, and analyzed 204,737 nuclei enriched for the BNST region (Figure S3A; STAR Methods). Initial clustering identified 106,728 neurons, of which 70.2% were localized to BNST by examination of marker expression in the Allen Brain Atlas (ABA) (Lein et al., 2007) (Figure S3B). Clustering analysis revealed 41 transcriptionally distinct populations of BNST-localized neurons (Figure 3A). In agreement with previous estimates (Kudo et al., 2012), 85.9% of the cells were inhibitory (expressing *Gad1* and *Gad2*), while the remaining 14% were excitatory (expressing *Slc17a6* [9.4%] or *Slc17a8* [4.7%]) (Figure S3C). Examination of cluster markers in the ABA showed that many cell types localized to specific BNST substructures, including the principal, oval, and anterior commissure nuclei (Figures S3C and S3D). For example, we identified two molecularly distinct subpopulations in the oval nucleus of the anterior BNST (ovBNST) (Figure 3B), a structure known to regulate anxiety (Kim et al., 2013) and to manifest a robust circadian rhythm of expression of *Per2* (Amir et al., 2004), similar to the superchiasmatic nucleus (SCN) of the hypothalamus.

Two clusters, BNST_Vip and BNSTp_Cplx3, expressed markers of caudal ganglionic eminence (CGE)-derived interneurons found in cortex and hippocampus. Part of the BNST has embryonic origins in the CGE (Nery et al., 2002), suggesting that this structure may harbor such cell types. To examine this possibility, we integrated the 352 nuclei from the BNST_Vip and BNST_Cplx3 clusters with 330 CGE interneuron cell profiles sampled from our recent adult mouse frontal cortex dataset (Saunders et al., 2018). Four clusters in the LIGER analysis showed meaningful alignment between BNST nuclei and cortical CGE cells (Figure 3C). One population (cluster 1), which was *Vip*-negative (Figure 3D) and likely localized to the posterior BNST (Figure 3E), expressed Id2, Lamp5, Cplx3, and Npy, all markers known to be present in cortical neurogliaform (NG) cells (Tasic et al., 2016). A second population (cluster 2) expressed *Vip*, *Htr3a*, *Cck*, and *Cnr1*, likely corresponding to VIP+ basket cells. (Rudy et al., 2011) (Figures 3D and 3E). Although, to our knowledge, NG cells have not been described in the BNST before, cells with NG-like anatomy and physiology have been observed within the amygdala (Mańko et al., 2012), a structure with related functional roles.

Spiny projection neurons (SPNs) are the principal cell type of the striatum, a structure just lateral to the BNST, but cells expressing the canonical SPN marker *Ppp1r1b* have also been documented in multiple anterior BNST nuclei (Gustafson and Greengard, 1990). The molecular relationship between striatal SPNs and these BNST cells is not known. We identified three *Ppp1r1b*$^{+}$ populations—one specifically BNST-localized and two without BNST-specific localization (8,200 nuclei; Figures S3B and S3D). To relate these putative SPNs to striatal SPNs, we used LIGER to integrate these three clusters with 10,643 published striatal SPN profiles (Saunders et al., 2018). Many of the nuclei from our dataset aligned to clusters 1 and 2 (Figure 3F)

Figure 3. LIGER Reveals Region-Specific and Sex-Specific Cellular Specialization in the Bed Nucleus of the Stria Terminalis

(A) t-SNE visualization of 74,910 bed nucleus neurons analyzed by LIGER, colored by cluster, and labeled by an exclusive marker.

(B) Top, feature plots showing expression of *Sh3d21* and *Vipr2*, in the LIGER BNST analysis. Bottom, sagittal ABA images of *Sh3d21* and *Vipr2*, showing restricted expression to the BNST oval nucleus.

(C) t-SNE visualization of a LIGER analysis of 352 BNST nuclei in clusters BNST_Vip and BNSTp_Cplx3 and 330 CGE-derived cortical interneurons (Saunders et al., 2018) are colored by dataset (top) and LIGER cluster (bottom).

(D) Dot plot showing the relative expression of genes, by dataset, in clusters 1 and 2 of the analysis shown in (C). Each dataset is scaled separately to reconcile differences in sampling between whole cells and nuclei (STAR Methods).

(E) Sagittal ABA images of *Vip*, *Lamp5*, and *Id2* expression; arrows highlight the signal present in the BNST.

(F) t-SNE visualization of a LIGER analysis of 8,200 nuclei, drawn from three clusters positive for the SPN marker *Ppp1r1b* (Figure S3), and 10,643 striatal SPNs (Saunders et al., 2018). The striatal SPNs are colored according to their published clustering into three major transcriptional categories (direct, indirect, and eccentric).

(G) Dot plots showing expression of canonical SPN genes in the clusters defined in (F). Markers include those of iSPN identity (*Adora2a*), dSPN identity (*Drd1*), and the recently described eSPN identity (*Tshz1*, *Otof*, and *Cacng5*), as well as two markers of the BNST-specific cluster 4 (*Cdc14a* and *Hcn1*).

(H) Coronal ABA image of *Cdc14a*, showing exclusive expression in the rhomboid nucleus of the anterolateral BNST.

(I) Bar plot quantifying dimorphically expressed genes per BNST neuron cluster, identified by a bootstrap analysis (STAR Methods). Note that the two BNSTpr clusters (BNSTpr_St18 and BNSTpr_Esr2) show high numbers of dimorphic genes.

(J) Cell factor loading values (top) and gene loading plots (bottom) of top loading dataset-specific and shared genes (bottom) for factor 27, which loads primarily on one of the BNSTpr clusters.

(K) Genes ranked by degree of dimorphism (STAR Methods); positive values indicate increased expression in males, while negative values indicate increased female expression. Positions of previously validated dimorphic genes and X and Y chromosome genes are indicated in blue and red, respectively. The two novel genes shown in (L) are indicated in purple.

(L) Feature plots showing expression patterns of known (*Greb1* and *Esr1)* and novel (*Elt4* and *Acvr1c*) dimorphic genes across BNST neurons.

Abbreviations in in situ hybridization (ISH) images: ac, anterior commissure; act, anterior commissure, temporal limb; ov, oval nucleus; st, stria terminalis; CP, caudate putamen; TH, thalamus. See also Figure S3 and Data S1.

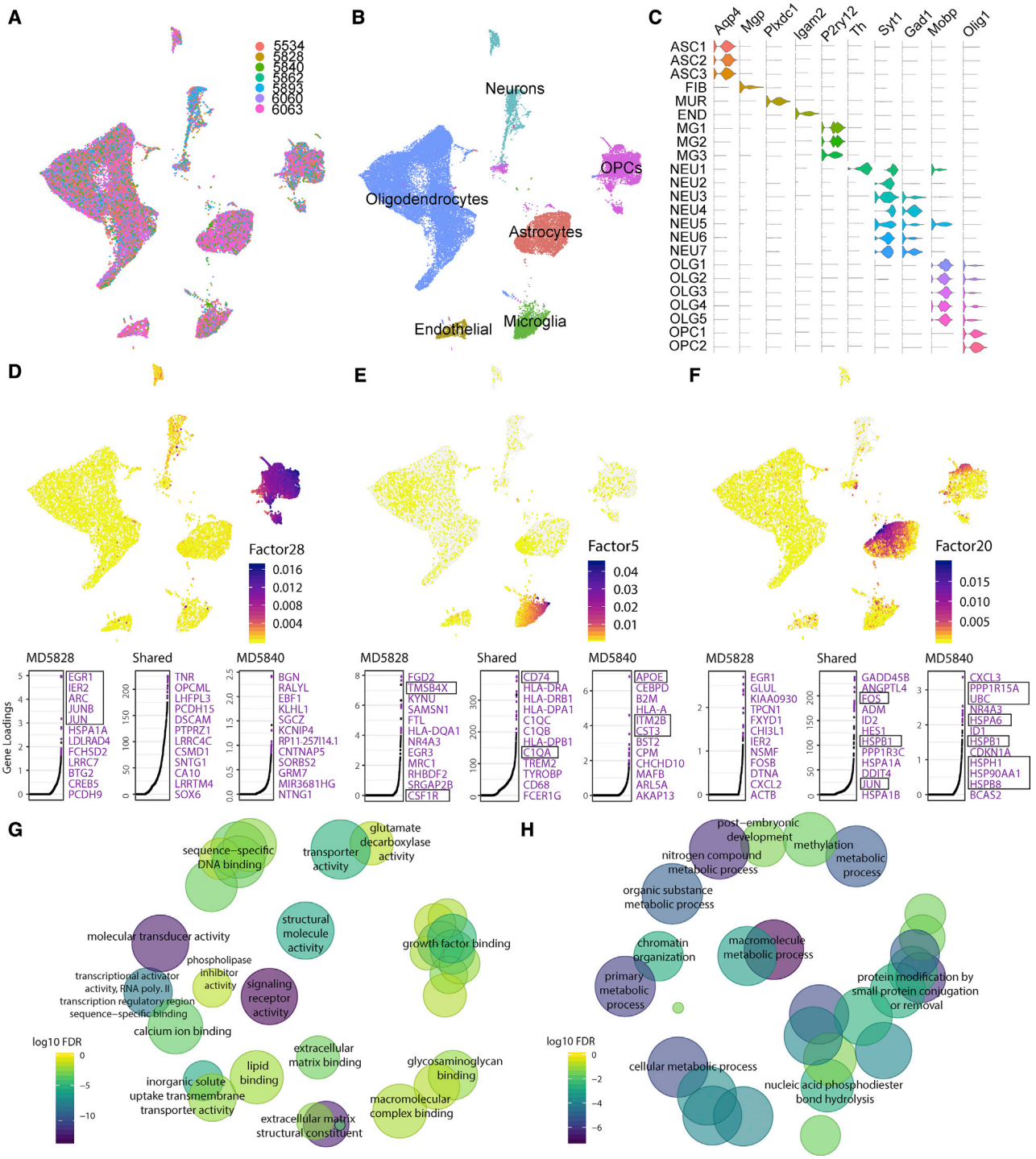


Figure 4. LIGER Allows Analysis of Substantia Nigra across Individuals and Species

(A and B) Uniform manifold approximation and projection (UMAP) plots of a LIGER analysis of 44,274 nuclei derived from the SN of 7 human donors, colored by donor (A) and major cell class (B).

(C) Violin plots showing expression of marker genes across the 25 human SN populations identified by two rounds of LIGER analysis.

(D–F) UMAP plots showing cell factor loading values (top) and gene loading plots (bottom) for factors corresponding to an acutely activated polydendrocyte state (D), an activated microglia state (E), and a reactive astrocyte state (F). In gene loading plots, gene names are sorted in decreasing order of magnitude of their factor loading contribution and correspond to colored points in scatterplots. Plots are organized to show the metagene specific to tissue donors MD5828 and MD5840 and the shared metagene common to all datasets. Genes mentioned in the text are boxed.

(legend continued on next page)

corresponding to canonical striatal SPNs of the indirect spiny projection neuron (iSPN) and direct spiny projection neuron (dSPN) types, respectively. A second population of our nuclei aligned to cluster 3, containing the striatal eccentric spiny projection neurons (eSPNs) we recently described (Saunders et al., 2018). A fourth population, cluster 4, expressed markers localizing it exclusively to the rhomboid nucleus of BNST (Figures 3G and 3H). These results suggest that the BNST contains a combination of SPN-like neurons with high homology to striatal SPNs, while also harboring at least one *Ppp1r1b*$^{+}$ population with tissue-specific specializations.

In addition to its high molecular and anatomical diversity, BNST also displays significant sexual dimorphism, both in size (Allen and Gorski, 1990; Hines et al., 1992) and gene expression (Xu et al., 2012). To identify cell-type-specific BNST dimorphism, we used LIGER to identify sex-specific metagene factors. X and Y chromosome genes such as *Xist*, *Tsix*, *Eif2s3y*, *Ddx3y*, and *Uty* showed high loading values on dataset-specific factors, reinforcing that these factors captured dimorphic gene expression. We then used the dataset-specific factor loadings to quantify the number of cell-type-specific dimorphic genes for each cluster (STAR Methods).

Our analysis revealed a complex pattern of dimorphic expression involving differences across many individual cell types. Clusters BNSTpr_St18 and BNSTpr_Esr2 from the BNST principal nucleus (BNSTpr) showed some of the highest numbers of dimorphic genes (Figure 3I), consistent with previous reports that BNSTpr is particularly dimorphic (Hines et al., 1992; Xu et al., 2012). To illustrate the interpretability of the factorization and the complexity of the dimorphism patterns it reveals, we plotted the loading pattern and cell-type-specific dimorphic genes derived from one particular factor (factor 27) that loads strongly on the BNSTpr_St18 cluster (Figure 3J). Among the top dimorphic genes for this factor were *Xist*, *Tsix*, and *Etl4*. We devised a metric from the LIGER analysis to rank genes by their cell-type-specific dimorphism (Figure 3K; STAR Methods), flagging genes expressed at higher levels in male or female within a specific population. Among 12 genes previously confirmed to be dimorphic in BNST (Xu et al., 2012), we found that most had high cell-type-specific expression metrics. We also identified new dimorphic genes, often with complex cell-type-specific dimorphisms across the many BNST subpopulations (Figure 3L; Data S1).

Integration of Substantia Nigra Profiles across Different Human Postmortem Donors and Species

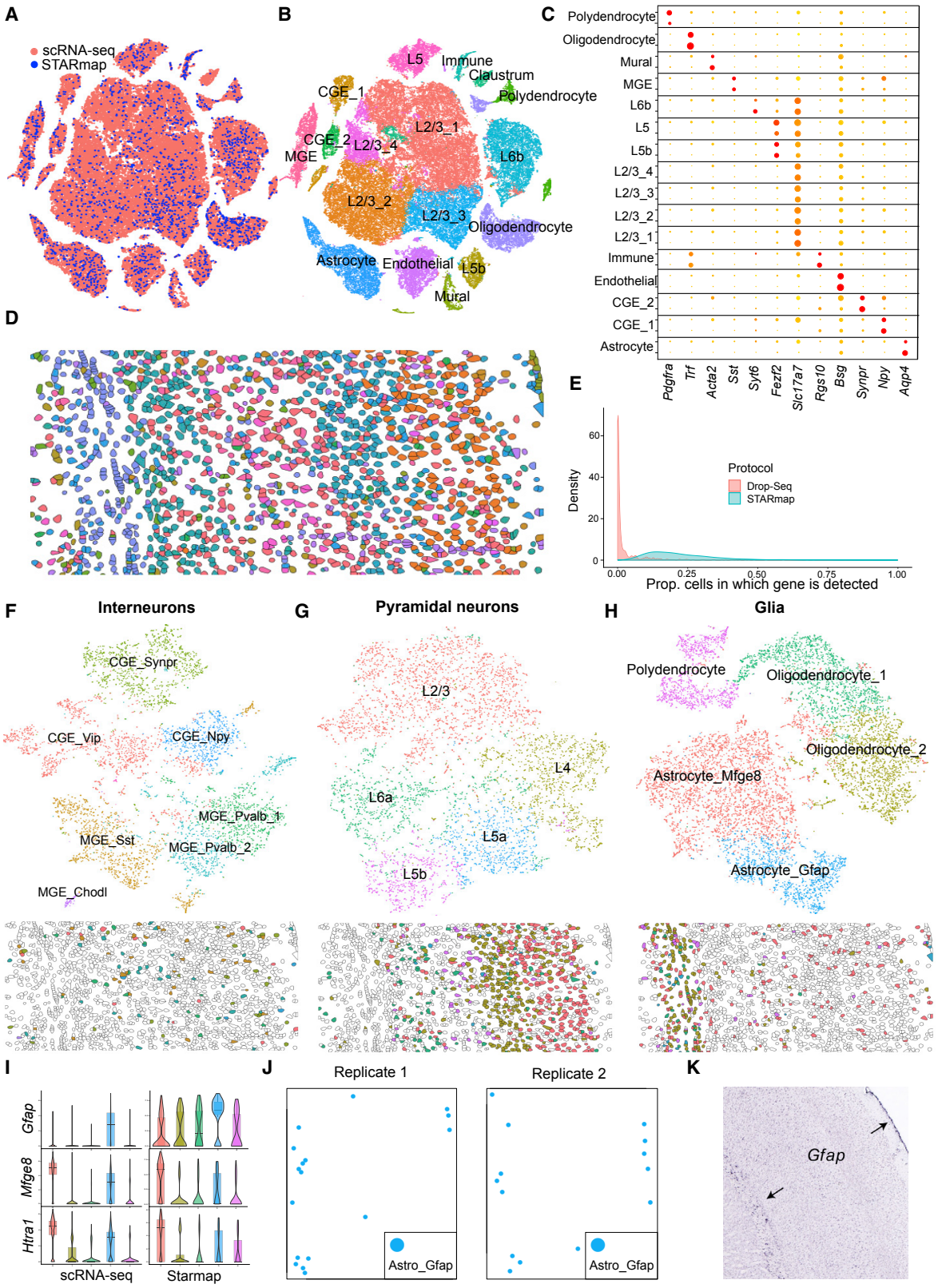
Profiling of individual nuclei from archival postmortem human brain samples (Habib et al., 2017; Lake et al., 2016) provides an exciting opportunity to comprehensively characterize transcriptional heterogeneity across the human brain. However, many ante- and postmortem variables create complex technical variation in gene expression, complicating efforts to identify biological variation in cell state. To explore how well LIGER can integrate individual human postmortem samples, we isolated and sequenced 44,274 nuclei derived from the substantia nigra (SN) of seven individuals designated as neurotypical controls (STAR Methods). The SN is a subcortical structure that functions in reward and movement execution and degenerates in Parkinson's disease. Despite considerable inter-individual variation (Figure S4A), LIGER accurately integrated each of the cell-type substituents of the SN across datasets (Figure 4A). Specifically, we identified 24 clusters spanning all known resident cell classes: astrocytes, fibroblasts, mural cells, microglia, neurons (including TH^{+} dopaminergic neurons and multiple inhibitory types), oligodendrocytes, and oligodendrocyte progenitor cells (polydendrocytes) (Figures 4B and 4C).

Glial activation is an important hallmark and driver of many brain diseases, including neurodegeneration and traumatic brain injury (TBI). To uncover datasets with atypical glial expression patterns, we examined the dataset-specific metagenes of glial cell types. The dataset-specific component of factor 28 showed that subject MD5828 had high expression of immediate early genes within polydendrocytes (Figure 4D), consistent with an acute injury (Dimou et al., 2008). Although this subject was coded as a control, the cause of death strongly suggested brain trauma (STAR Methods). In addition, the MD5828-specific metagene for factor 5, which was microglia specific, had high loadings of *TMSB4X* and *CSF1R*, both of which play important roles in the acute response to TBI (Luo et al., 2013; Xiong et al., 2012). By contrast, in subject 5840, the dataset-specific loadings on the microglial factor 5 included genes upregulated in response to amyloid deposition (Figure 4E). Review of this subject's postmortem report revealed a histological diagnosis of cerebral amyloid angiopathy (CAA), in which amyloid deposits within the walls of CNS vasculature. Intriguingly, two of the three genes known to cause hereditary CAA (Biffi and Greenberg, 2011), *CST3* and *ITM2B*, were also strong contributors to MD5840-specific factor 5. In an astrocyte-specific factor (factor 20), subject MD5840 showed remarkable upregulation of multiple genes involved in protein misfolding response (Figure 4F) (Tsaytler et al., 2011), several of which are known to be amyloid-responsive (Bruinsma et al., 2011).

A deeper understanding of cell types often arises from comparisons across species. We therefore used LIGER to compare our newly generated human SN data with a recently published dataset from the mouse SN (Saunders et al., 2018). The joint analysis identified both corresponding broad cell classes across species and subtler cell types within each class after a second round of analysis (Figures S4B–S4F). In our subanalysis of the neurons, LIGER avoided false-positive alignments of human profiles to mouse cell types outside the dissection zone of the human tissue (Figures S4G and S4H). Overall, we observed strong concordance between mouse and human cell clusters, consistent with a recent analysis of mouse and human cortex (Hodge et al., 2018).

Understanding how expression of homologous genes within the SN differs across species could reveal differences in how

(G) GO terms enriched in homologous genes with strong expression correlation across SN clusters in the LIGER comparative analysis of human and mouse.
(H) GO terms enriched in homologous genes with weak expression correlation. Colors indicate false discovery rate, while size of the circles indicates the number of genes associated with each GO term in (G) and (H).
See also Figure S4.



(legend on next page)

these genes function within the tissue. We performed gene ontology (GO) term enrichment analysis to evaluate whether genes with the highest and lowest correlation across species share any functional relationships. Homologous gene pairs with high expression correlation were enriched for GO terms related to brain cell identity and basic molecular functions, including ion channels, transcription factors, transmembrane receptors, and extracellular matrix structural components (Figure 4G). In contrast, the least correlated homologous gene pairs were enriched for basic metabolic processes, such as macromolecule metabolism and DNA repair (Figure 4H). Intriguingly, genes involved in chromatin remodeling (sharing the GO function "chromatin organization") also showed less expression conservation, hinting at species differences in epigenetic regulation.

Integrating scRNA-Seq and *In Situ* Transcriptomic Data Locates Frontal Cortex Cell Types in Space

Spatial context is an important aspect of cellular identity, but most studies have used scRNA-seq from dissociated cells to define cell types. Integrating these data types using LIGER could offer two potential advantages compared to separate analyses: (1) assigning spatial locations to cell clusters observed in data from dissociated cells; and (2) increasing the resolution for detecting cell clusters from the *in situ* data.

We jointly analyzed frontal cortex scRNA-seq profiles (Saunders et al., 2018) and *in situ* spatial transcriptomic data from the same tissue generated by STARmap (Wang et al., 2018). These two datasets differ widely in number of cells (71,000 scRNA-seq versus 2,500 STARmap) and genes measured per cell (scRNA-seq is unbiased, while STARmap is targeted). Nevertheless, LIGER correctly defined joint cell populations across the datasets (Figures 5A and 5B), with expression of key marker genes confirming the correspondence of cells across these different modalities (Figure 5C). Only one population in the scRNA-seq data was dataset specific, corresponding to cells from the claustrum, an anatomical structure not included in the STARmap field of view (Figure S5A). Our integrated analysis spatially located each of the jointly defined populations (Figure 5D) and reflected the known spatial features of the mouse cortex, including meninges and sparse layer 1 interneurons at the surface, excitatory neurons organized in layers 2–6, and oligodendrocyte-rich white matter below the cortex (Figure 5D). One replicate of the STARmap data also showed a chain of endothelial cells running through the cortex, likely a contiguous segment of vasculature (Figure 5D). The success of this integrative analysis is especially noteworthy given the very different global distributions of gene expression values in the scRNA-seq data compared to the STARmap data (Figure 5E).

Incorporating the scRNA-seq data also identified cell populations from STARmap with greater resolution than the published clustering. Specifically, we identified 7 interneuron clusters and 5 glial clusters compared to 4 and 2 clusters, respectively, in the initial STARmap analysis. These additional populations accorded well with cell-type distinctions defined in the original scRNA-seq analysis. The 5 glial clusters we identified included two astrocyte clusters, polydendrocytes, and two clusters of oligodendrocytes (Wang et al., 2018). The two astrocytic subpopulations expressed patterns of marker genes consistently between both the scRNA-seq and STARmap datasets (Figure 5F). The larger population expressed high levels of *Mfge8* and *Htra1*, while the second population showed high expression of *Gfap* (Figure 5F). The *Gfap*-expressing astrocyte population is located outside the cortical gray matter, in both the meningeal lining and the white matter below layer 6 (Figures 5G and 5H), consistent with a more fibrous identity. In contrast, the larger second population of astrocytes was spread uniformly throughout the cortical layers, consistent with a protoplasmic phenotype. Identifying the localization of the *Gfap*-expressing astrocyte population also clarified our human-mouse SN analysis (Figure S4E), suggesting that this same *Gfap*-expressing population is likely missing from the human data because of dissection differences. These results show the power of jointly leveraging large-scale scRNA-seq and *in situ* gene expression data for defining cell types in the brain.

We also investigated whether it is possible to predict the spatial patterns of genes not assayed by STARmap. To do this, we assigned each STARmap cell to the average of its nearest scRNA-seq neighbors in the aligned factor space (STAR Methods). Comparison of predicted gene patterns with the ABA showed that LIGER is able to reveal even complex spatial expression patterns across many individual genes (Figures S5B–S5D). Most high-error genes either showed technical differences in measurement between STARmap and scRNA-seq (e.g., *Aldoc*, *Tsnax*, *Hlf*) or possessed no obvious spatial pattern (e.g., *Elmo1*, *Glul*, *Scg2*) (Figures S5E–S5G).

LIGER Defines Cell Types Using Both Single-Cell Transcriptome and Single-Cell DNA Methylation Profiles

Linking single-cell epigenomic data with scRNA-seq would open exciting avenues for investigation. First, it is unknown whether clusters defined from gene expression reflect epigenetic distinctions and vice versa. Second, integrating single-cell epigenomic and transcriptomic data provides an opportunity to study the

Figure 5. Locating Cortical Cell Types in Space Using scRNA-Seq and STARmap

(A and B) t-SNE plots of a LIGER analysis of 71,000 frontal cortex scRNA-seq profiles (Saunders et al., 2018) and 2,500 cells profiled by STARmap (Wang et al., 2018) colored by technology (A) and LIGER cluster assignment (B). Labels in (B) derive from the published annotations of the Drop-seq dataset.

(C) Dot plot showing marker expression for STARmap cells (top line of each gene) and Drop-seq cells (bottom line) across LIGER joint clusters.

(D) Spatial locations of STARmap cells colored by LIGER cluster assignments.

(E) Density plot showing proportion of cells in which each gene is detected for the scRNA-seq (red) and STARmap (blue) datasets.

(F–H) t-SNE plots and spatial locations for LIGER subclustering analyses of interneurons (F), pyramidal neurons, (G), and glia (H).

(I) Violin plots of marker genes for two astrocyte populations identified in subclustering analysis of glia.

(J) Spatial coordinates for *Gfap*-expressing astrocyte populations (two STARmap replicates shown).

(K) *Gfap* staining data from the Allen Brain Atlas showing localization of *Gfap* to both meninges and white matter layer below cortex.

See also Figure S5.

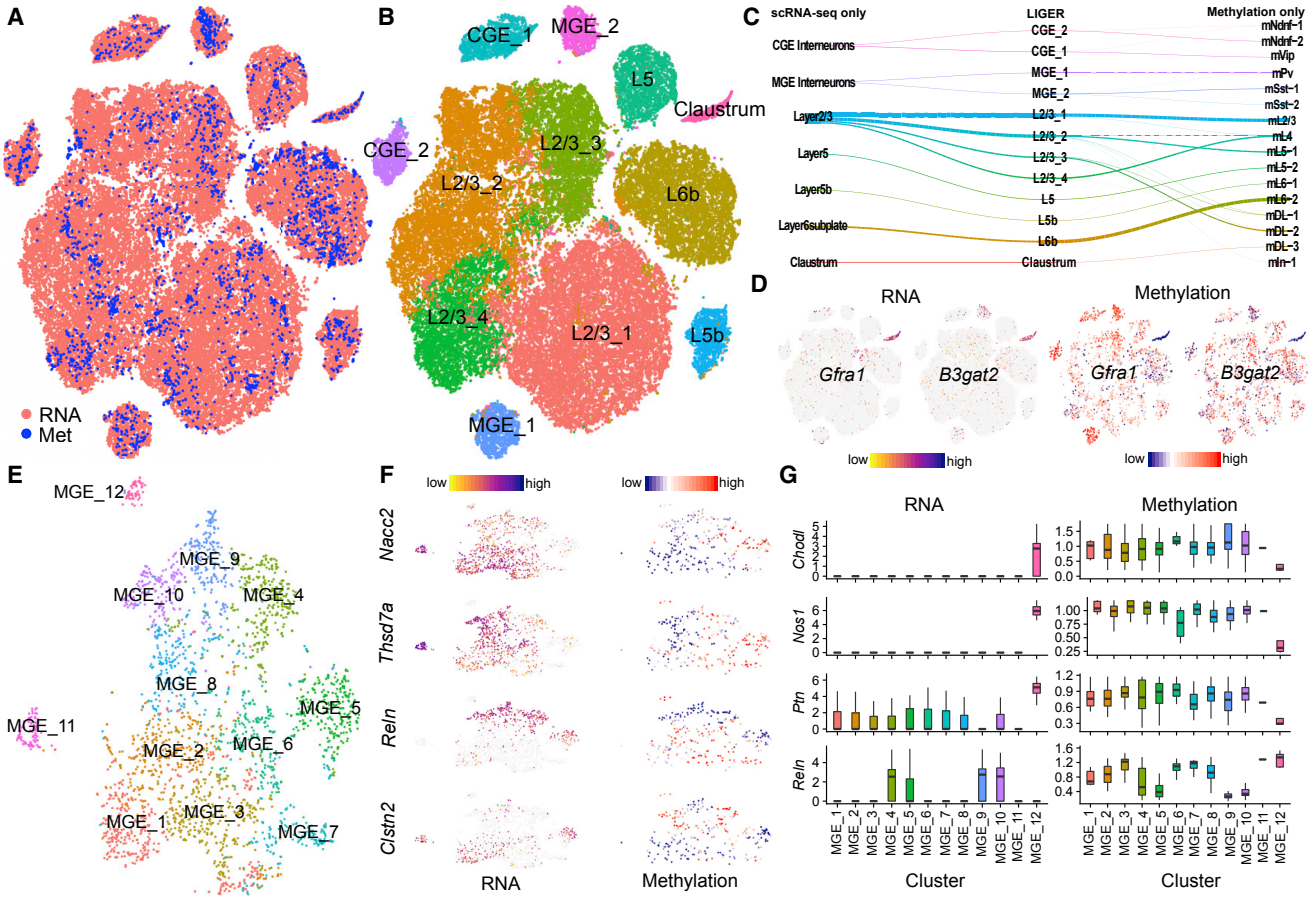


Figure 6. Defining Cortical Cell Types Using Both scRNA-Seq and DNA Methylation

(A and B) t-SNE visualization of LIGER analysis of scRNA-seq data (Saunders et al., 2018) and methylation data (Luo et al., 2017) from mouse frontal cortex, colored by modality (A) and LIGER cluster assignment (B).

(C) Riverplot showing relationship between published cluster assignments of RNA and methylation data and LIGER joint clusters.

(D) Expression and methylation of two claustrum markers.

(E) t-SNE representation of the LIGER subcluster analysis of MGE interneurons.

(F) Expression and methylation of 4 marker genes for different MGE subpopulations.

(G) Boxplots of expression and methylation markers for *Sst-Chodl* cells (cluster MGE_12).

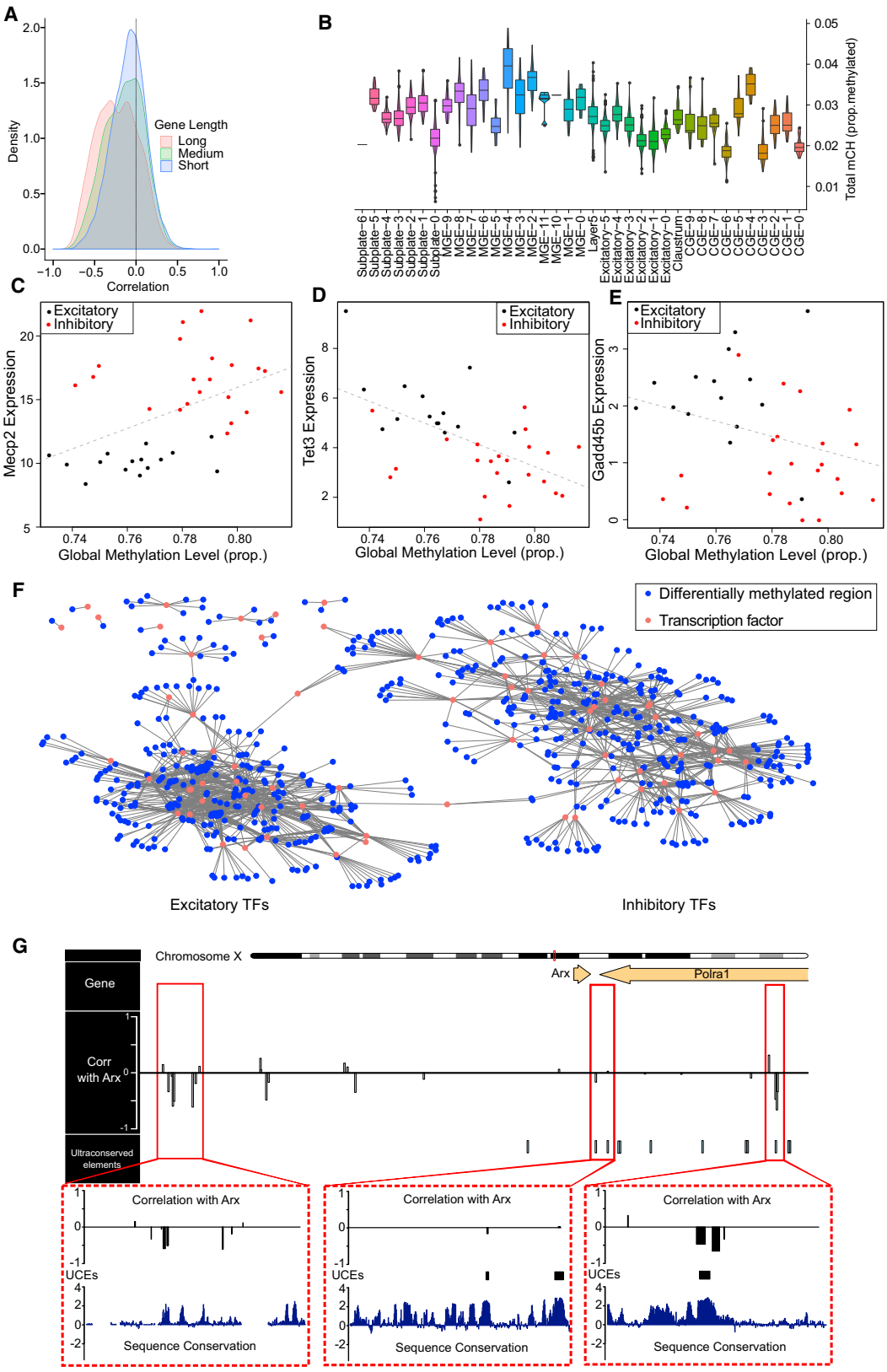
See also Figure S6.

mechanisms by which epigenomic information regulates gene expression to determine cell identity. Finally, such integration may improve sensitivity and interpretability compared to analyzing the epigenomic data in isolation, since scRNA-seq technology can offer greater throughput and capture more information per cell.

To investigate these possibilities, we performed an integrated analysis of two single-cell datasets prepared from mouse frontal cortical neurons: one of gene expression (55,803 cells) (Saunders et al., 2018) and another of genome-wide DNA methylation (3,378 cells) (Luo et al., 2017). We reasoned that, because non-CpG (mCH) gene body methylation is generally anticorrelated with gene expression in neurons (Mo et al., 2015), reversing the direction of the methylation signal would allow joint analysis. Indeed, LIGER successfully integrated the datasets, jointly identifying the neuronal cell types of the frontal cortex and according well with the published analyses of each dataset (Figures 6A–6C).

Our joint analysis clarified the identities of some methylation clusters. We found that a cluster annotated as "deep-layer cluster 3" aligned uniquely to an RNA-seq cluster that we previously had annotated as claustrum (Saunders et al., 2018) (Figures 6C and 6D). In addition, a cluster annotated as "layer 6 cluster 1" aligned with a cluster that we identified as layer 5b. The canonical marker genes have relatively low overall methylation levels, making it challenging to assign the identity of this cell type from methylation alone. However, the expression of several specific layer 5b marker genes, most notably *Slc17a8* (Sorensen et al., 2015), and their corresponding low methylation pattern in the aligned cluster mL6-1 cells, enabled us to confirm this assignment (Figure S6A).

We performed four sub-analyses of the broad cell classes in the frontal cortex: CGE-derived interneurons, MGE-derived interneurons, superficial excitatory neurons, and deep-layer excitatory neurons (Figures S5C–S5E), identifying a total of 37



(legend on next page)

clusters. Joint analysis of MGE interneurons revealed 11 populations, considerably more than was possible using the methylation data alone (Figure 6E). Examining expression and methylation of marker genes confirmed that these populations are real and not simply forced alignment (Figure 6F). We were further able to credibly identify 25 methylation profiles corresponding to an interneuron population expressing *Pvalb* and *Th* (Figure S6B), as well as 5 profiles aligning to the cluster expressing *Sst* and *Chodl* (Figure 6G). Together, these results indicate that epigenomic and expression data produce meaningful joint neural cell-type definitions, and even the finest distinctions among neural cell types defined from gene expression can be reflected by DNA methylation differences.

Our joint cluster definitions offered an opportunity to investigate the regulatory relationship between expression and methylation at cell-type-specific resolution. We first aggregated the gene expression and methylation values within each cluster and then calculated correlation between the expression of each gene and its gene body methylation levels across the set of clusters. We confirmed the well-established overall negative relationship between methylation and expression (Figure 7A). We also leveraged this inverse relationship to predict spatial methylation patterns (Figure S7A; STAR Methods). Consistent with previous work (Luo et al., 2017; Mo et al., 2015), we found that non-CpG methylation within the gene body, rather than CpG methylation (mCG), was more anticorrelated with expression (Figure S7B), and anticorrelation was weaker in mCH deserts (Figure S7C), megabase-scale regions with very low mCH relative to mCG (Lister et al., 2013). We also found that using mCG resulted in poorer cluster separation compared with mCH (Figures S7D and S7E). Longer genes showed stronger negative correlation with gene expression than shorter genes (Figure 7A), consistent with a known mechanism of gene repression by DNA methylation, in which the MECP2 protein binds methylated nucleotides (Fasolino and Zhou, 2017). The degree of MECP2 repression has been shown to be proportional to the number of methylated nucleotides, which is strongly related to gene length (Kinde et al., 2016). Since gene length also affects the amount of measured methylation signal in these sparse profiles, we cannot completely rule out the influence of technical factors in this observed relationship.

We observed a wide range of global methylation levels across our set of clusters (Figure 7B), providing an opportunity to investigate the basic molecular machinery involved in regulating methylation. We correlated the expression of several key genes with the global methylation level of each cell. We found that expression of *Mecp2* correlated strongly ($\rho = 0.46$, $p = 0.0039$) with global methylation level (Figure 7C), supporting a model in which MECP2 represses gene expression by specifically binding to methylated nucleotides (Kinde et al., 2016), creating a stoichiometric requirement for increased *Mecp2* expression in cells with higher overall methylation levels. In addition, we found that *Tet3*, which converts 5mC to 5hmC, strongly anticorrelated ($\rho = -0.57$, $p = 0.0002$) with global methylation (Figure 7D). Intriguingly, the other TET genes were not anticorrelated with global methylation despite similar overall expression levels (Figures S7E and S7F), suggesting that TET3 could be the dominant TET protein regulating global methylation in mature neurons. *Gadd45b*, a gene with a well-established role in demethylating neuronal DNA (Bayraktar and Kreutz, 2018), also showed a strong negative relationship ($\rho = -0.30$, $p = 0.0685$) with global methylation. Consistent with our analysis, *Gadd45b* is thought to regulate DNA demethylation by recruiting TETs (Bayraktar and Kreutz, 2018). By contrast, none of the DNA methyltransferase enzymes (DNMTs) were strongly related to overall methylation level (Figures S7G–S7I). These analyses show the value of an integrated analysis to formulate hypotheses about the mechanisms by which expression and methylation are regulated.

Our integrated analysis could also enable the identification of intergenic elements regulating cell-type-specific gene expression. We defined a set of stringent criteria that combined intergenic methylation status, transcription factor expression, and transcription factor sequence specificity, to identify such intergenic regions—and the transcription factors that may bind them—in specific cell types (STAR Methods) (Figure 7F). These represent strong candidates for cell-type-specific transcriptional regulatory elements, as they harbor unmethylated transcription factor binding motifs in cell types with high expression of the corresponding transcription factors.

Finally, our integrated definition of cell types from methylation and expression allowed us to examine the relationship between intergenic methylation and the expression of nearby genes. The *Arx* locus harbors 8 ultraconserved elements (UCEs)—long stretches of sequence showing perfect conservation among human, mouse, and rat (Bejerano et al., 2004; Colasante et al., 2008). Several distal regulatory elements, including some located within neighboring UCEs, have recently been demonstrated to regulate *Arx* expression (Colasante et al., 2008; Dickel et al., 2018). To nominate putative elements regulating *Arx*, we correlated *Arx* expression and methylation of nearby differentially methylated regions (DMRs) across our joint clusters (Figure 7G). We observed several clusters of DMRs whose methylation is anticorrelated with *Arx* expression, a pattern expected if hypo-methylation within certain cell types makes available a

Figure 7. Investigating the Connection between DNA Methylation and Gene Expression

(A) Density plot of the correlation between gene body methylation and expression for short, medium, and long genes.

(B) Violin plots showing the wide range of global methylation levels across neural cell types defined by the LIGER analysis in Figure 6.

(C–E) Scatterplots of global methylation and aggregate expression for (C) *Mecp2*, (D) *Tet3*, and (E) *Gadd45b* across our joint neural cell clusters.

(F) Network of predicted interactions between transcription factors (TFs; red) and differentially methylated regions (DMRs; blue). An edge connects a TF to a DMR if the region contains a binding motif for the TF and has methylation anticorrelated with the expression of the transcription factor (Pearson $\rho < -0.45$). The network segregates into two largely disconnected components, which are enriched for TFs expressed in excitatory (left) and inhibitory (right) neurons.

(G) Genome browser view showing locations of differentially methylated regions near the *Arx* locus and their correlation with the expression of *Arx*. The bars indicate sign and magnitude of the correlation. The 3 bottom panels show zoomed-in views of three clusters of DMRs.

See also Figure S7.

regulatory element that enhances *Arx* expression. One of these anticorrelated DMRs is a validated *Arx* enhancer (Dickel et al., 2018) just downstream of the end of the *Arx* gene (Figure 7G, middle). Another pair of DMRs strongly correlated with *Arx* expression overlap a UCE further downstream of *Arx* (Figure 7G, right). A third group of DMRs upstream of the *Arx* site lies in a region of very high conservation (though not a UCE), with three clear spikes in conservation that aligned precisely with the locations of the DMRs (Figure 7H, left). In summary, these DMRs represent strong candidates for putative elements regulating *Arx* expression, highlighting the value of our integrative approach for investigating gene regulatory mechanisms.

DISCUSSION

A credible definition of cell type requires distinguishing the invariant properties of cell identity from the dispensable across a myriad of settings and measurements. LIGER promises to be a broadly useful analytical tool for such efforts because of several key technical advantages. First, the nonnegativity constraint of NMF yields interpretable factors, such that each factor generally corresponds to a biologically meaningful signal. Second, the inclusion of dataset-specific terms allows us to identify dataset differences, rather than attempting to force highly divergent datasets into a completely shared latent space. Finally, LIGER's inference of both shared and dataset-specific factors enables a more transparent and nuanced definition of *how* cells correspond across datasets. In cases where complete correspondence is not necessarily expected—such as connecting fully differentiated cells to progenitors or relating pathological cells to healthy counterparts—a characterization of the metagenes that both unite and separate such populations is crucial. We note that one limitation of existing methods for multi-omic integration, including LIGER, is their inability to incorporate different types of features, such as gene expression and intergenic methylation, in the definition of cell types. Future studies may investigate how best to incorporate such information.

Another integration algorithm, described in this issue by Stuart et al., (2019), uses canonical correlation analysis (CCA) to identify a completely shared subspace of maximum correlation and then uses these shared components to identify anchor points across heterogeneous datasets. CCA solves a convex optimization problem and thus guarantees a deterministic, globally optimal solution. In contrast, LIGER uses integrative nonnegative matrix factorization, which solves a non-convex optimization problem and thus produces a different factorization depending on the initialization used. LIGER infers interpretable shared and dataset-specific factors, which often correspond to important biological signals, including signals that are not orthogonal to cell type, or technical signals, enabling their removal from downstream analysis.

We envision LIGER serving several important needs in neurobiology, beyond its capacity to better define cell types. First, a key opportunity in single-cell analysis is the identification of cell-type-specific gene expression patterns associated with disease risk, onset, and progression in human tissue samples. Early efforts at such investigation have yielded some exciting results (Keren-Shaul et al., 2017), but increased discovery is likely possible with robust integrative analysis of many tissue donors. Such analyses may also help localize genetic risk loci for neuropsychiatric diseases to specific human cell types. Second, the integration of data from epigenomic and transcriptomic datasets provides a path toward nominating functional genomic elements important in cell-type-specific gene regulation. Such elements are compelling candidates for cell-type-specific enhancers to drive expression of genetic tools in specific subsets of brain cells and may also help narrow the search for causative alleles at specific genetic risk loci. Finally, as *in vitro* models of complex brain tissues become more sophisticated (Birey et al., 2017; Quadrato et al., 2017), single-cell gene expression measurements, together with an integrative analysis like LIGER, will help provide systematic, information-rich comparisons of such models with their *in vivo* counterparts. To facilitate adoption of the tool in the community, we have developed an R package that supports analysis of large-scale datasets and includes ancillary functions for tuning algorithmic parameters, visualizing results, and quantifying integrative performance. We hope its widespread deployment opens many exciting new avenues in single-cell biology.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Animals
  - Human postmortem tissue
- METHOD DETAILS
  - Liger Workflow
  - Performing iNMF Using Block Coordinate Descent
  - Efficient Updating for New K, New Data, and New λ
  - Algorithm 1: Updating the factorization with a new k
  - Algorithm 2: Adding new data to the factorization
  - Heuristics to guide choice of k and λ
  - Joint clustering and factor normalization
  - Calculating alignment and agreement metrics
  - Generation of BNST nuclei profiles
  - Generation of profiles from human substantia nigra
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Analysis of bed nucleus data
  - Analysis of substantia nigra data
  - Analysis of STARmap data
  - Analysis of methylation data
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.05.006.

ACKNOWLEDGMENTS

We thank Aleks Goeva for advice on computational algorithms. This work was supported by the Stanley Center for Psychiatric Research, the Chan

Zuckerberg Initiative (grant numbers 2017-175259 to E.Z.M. and 2018-183155 to J.D.W.), and NIH/NIMH BRAIN Grant 1U19MH114821 to E.Z.M.

AUTHOR CONTRIBUTIONS

J.D.W. and E.Z.M. designed the study and wrote the paper. J.D.W. derived and implemented the computational algorithms, with help from V.K. J.D.W., V.K., and E.Z.M. performed the analyses. J.D.W. and V.K. wrote the R package. A.F., C.M., and C.V. performed the dissections, nuclear isolations, and library preparations from mouse bed nucleus and human SN.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 2, 2018
Revised: February 21, 2019
Accepted: April 30, 2019
Published: June 6, 2019

REFERENCES

Allen, L.S., and Gorski, R.A. (1990). Sex difference in the bed nucleus of the stria terminalis of the human brain. J. Comp. Neurol. *302*, 697–706.

Amir, S., Lamont, E.W., Robinson, B., and Stewart, J. (2004). A circadian rhythm in the expression of PERIOD2 protein reveals a novel SCN-controlled oscillator in the oval nucleus of the bed nucleus of the stria terminalis. J. Neurosci. *24*, 781–790.

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., et al. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst *3*, 346–360.

Bayless, D.W., and Shah, N.M. (2016). Genetic dissection of neural circuits underlying sexually dimorphic social behaviours. Philos. Trans. R. Soc. Lond. B Biol. Sci. *371*, 20150109.

Bayraktar, G., and Kreutz, M.R. (2018). The Role of Activity-Dependent DNA Demethylation in the Adult Brain and in Neurological Disorders. Front. Mol. Neurosci. *11*, 169.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325.

Bertsekas, D. (1999). Nonlinear Programming (Athena Scientific).

Biffi, A., and Greenberg, S.M. (2011). Cerebral amyloid angiopathy: a systematic review. J. Clin. Neurol. *7*, 1–9.

Birey, F., Andersen, J., Makinson, C.D., Islam, S., Wei, W., Huber, N., Fan, H.C., Metzler, K.R.C., Panagiotakos, G., Thom, N., et al. (2017). Assembly of functionally integrated human forebrain spheroids. Nature *545*, 54–59.

Bruinsma, I.B., de Jager, M., Carrano, A., Versleijen, A.A., Veerhuis, R., Boelens, W., Rozemuller, A.J., de Waal, R.M., and Verbeek, M.M. (2011). Small heat shock proteins induce a cerebral inflammatory reaction. J. Neurosci. *31*, 11992–12000.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Colasante, G., Collombat, P., Raimondi, V., Bonanomi, D., Ferrai, C., Maira, M., Yoshikawa, K., Mansouri, A., Valtorta, F., Rubenstein, J.L., and Broccoli, V. (2008). Arx is a direct target of Dlx2 and thereby contributes to the tangential migration of GABAergic interneurons. J. Neurosci. *28*, 10674–10686.

Coskun, A.F., and Cai, L. (2016). Dense transcript profiling in single cells by image correlation decoding. Nat. Methods *13*, 657–660.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309–1324.

Dickel, D.E., Ypsilanti, A.R., Pla, R., Zhu, Y., Barozzi, I., Mannion, B.J., Khin, Y.S., Fukuda-Yuzawa, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Ultraconserved Enhancers Are Required for Normal Development. Cell *172*, 491–499.

Dimou, L., Simon, C., Kirchhoff, F., Takebayashi, H., and Götz, M. (2008). Progeny of Olig2-expressing progenitors in the gray and white matter of the adult mouse cerebral cortex. J. Neurosci. *28*, 10434–10442.

Ding, S.L., Royall, J.J., Sunkin, S.M., Ng, L., Facer, B.A., Lesnar, P., Guillozet-Bongaarts, A., McMurray, B., Szafer, A., Dolbeare, T.A., et al. (2016). Comprehensive cellular-resolution atlas of the adult human brain. J. Comp. Neurol *524*, 3127–3481.

Dong, H.W., and Swanson, L.W. (2004). Projections from bed nuclei of the stria terminalis, posterior division: implications for cerebral hemisphere regulation of defensive and reproductive behaviors. J. Comp. Neurol. *471*, 396–433.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Fasolino, M., and Zhou, Z. (2017). The Crucial Role of DNA Methylation and MeCP2 in Neuronal Function. Genes (Basel) *8*. Published online May 13, 2017. 10.3390/genes8050141.

Gierahn, T.M., Wadsworth, M.H., 2nd, Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat. Methods *14*, 395–398.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Gustafson, E.L., and Greengard, P. (1990). Localization of DARPP-32 immunoreactive neurons in the bed nucleus of the stria terminalis and central nucleus of the amygdala: co-distribution with axons containing tyrosine hydroxylase, vasoactive intestinal polypeptide, and calcitonin gene-related peptide. Exp. Brain Res. *79*, 447–458.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat. Methods *14*, 955–958.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. *36*, 421–427.

Hines, M., Allen, L.S., and Gorski, R.A. (1992). Sex differences in subregions of the medial nucleus of the amygdala and the bed nucleus of the stria terminalis of the rat. Brain Res. *579*, 321–326.

Hodge, R.D., Bakken, T.E., Miller, J.A., Smith, K.A., Barkan, E.R., Graybuck, L.T., Close, J.L., Long, B., Penn, O., Yao, Z., et al. (2018). Conserved cell types with divergent features between human and mouse cortex. bioRxiv. https://doi.org/10.1101/384826.

Hrvatin, S., Hochbaum, D.R., Nagy, M.A., Cicconet, M., Robertson, K., Cheadle, L., Zilionis, R., Ratner, A., Borges-Monroy, R., Klein, A.M., et al. (2018). Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. Nat. Neurosci. *21*, 120–129.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T.K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., et al. (2017). A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. Cell *169*, 1276–1290.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. *46* (D1), D260–D266.

Kim, J.H., Yunlong, H., and Park, H. (2014). Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. J. Glob. Optim. *58*, 285–319.

Kim, S.Y., Adhikari, A., Lee, S.Y., Marshel, J.H., Kim, C.K., Mallory, C.S., Lo, M., Pak, S., Mattis, J., Lim, B.K., et al. (2013). Diverging neural pathways assemble a behavioural state from separable features in anxiety. Nature *496*, 219–223.

Kinde, B., Wu, D.Y., Greenberg, M.E., and Gabel, H.W. (2016). DNA methylation in the gene body influences MeCP2-mediated gene repression. Proc. Natl. Acad. Sci. USA *113*, 15114–15119.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Kudo, T., Uchigashima, M., Miyazaki, T., Konno, K., Yamasaki, M., Yanagawa, Y., Minami, M., and Watanabe, M. (2012). Three types of neurochemical projection from the bed nucleus of the stria terminalis to the ventral tegmental area in adult mice. J. Neurosci. *32*, 18035–18046.

Lake, B.B., Ai, R., Kaeser, G.E., Salathia, N.S., Yung, Y.C., Liu, R., Wildberg, A., Gao, D., Fung, H.L., Chen, S., et al. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science *352*, 1586–1590.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global epigenomic reconfiguration during mammalian brain development. Science *341*, 1237905.

Luo, J., Elwood, F., Britschgi, M., Villeda, S., Zhang, H., Ding, Z., Zhu, L., Alabsi, H., Getachew, R., Narasimhan, R., et al. (2013). Colony-stimulating factor 1 receptor (CSF1R) signaling in injured neurons facilitates protection and survival. J. Exp. Med. *210*, 157–172.

Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600–604.

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.

Mańko, M., Bienvenu, T.C., Dalezios, Y., and Capogna, M. (2012). Neurogliaform cells of amygdala: a source of slow phasic inhibition in the basolateral complex. J. Physiol. *590*, 5611–5627.

Mo, A., Mukamel, E.A., Davis, F.P., Luo, C., Henry, G.L., Picard, S., Urich, M.A., Nery, J.R., Sejnowski, T.J., Lister, R., et al. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron *86*, 1369–1384.

Moffitt, J.R., and Zhuang, X. (2016). RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). Methods Enzymol. *572*, 1–49.

Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D., Sinnamon, J.R., Shendure, J., Trapnell, C., O'Roak, B.J., et al. (2018). Highly scalable generation of DNA methylation profiles in single cells. Nat. Biotechnol. *36*, 428–431.

Nery, S., Fishell, G., and Corbin, J.G. (2002). The caudal ganglionic eminence is a source of distinct cortical and subcortical cell populations. Nat. Neurosci. *5*, 1279–1287.

Quadrato, G., Nguyen, T., Macosko, E.Z., Sherwood, J.L., Min Yang, S., Berger, D.R., Maria, N., Scholvin, J., Goldman, M., Kinney, J.P., et al. (2017). Cell diversity and network dynamics in photosensitive human brain organoids. Nature *545*, 48–53.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science *360*, 176–182.

Rudy, B., Fishell, G., Lee, S., and Hjerling-Leffler, J. (2011). Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. Dev. Neurobiol. *71*, 45–61.

Saunders, A., Macosko, E.Z., Wysoker, A., Goldman, M., Krienen, F.M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., et al. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell *174*, 1015–1030.

Sorensen, S.A., Bernard, A., Menon, V., Royall, J.J., Glattfelder, K.J., Desta, T., Hirokawa, K., Mortrud, M., Miller, J.A., Zeng, H., et al. (2015). Correlated gene expression and target specificity demonstrate excitatory projection neuron diversity. Cereb. Cortex *25*, 433–449.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., III, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, this issue, 1888–1902.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE *6*, e21800.

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.

Tosches, M.A., Yamawaki, T.M., Naumann, R.K., Jacobi, A.A., Tushev, G., and Laurent, G. (2018). Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. Science *360*, 881–888.

Tripodi, R.V., Sebastiano, V., and Pelillo, M. (2016). Context aware nonnegative matrix factorization clustering. arXiv, arXiv:1609.04628. https://arxiv.org/abs/1609.04628.

Tsaytler, P., Harding, H.P., Ron, D., and Bertolotti, A. (2011). Selective inhibition of a regulatory subunit of protein phosphatase 1 restores proteostasis. Science *332*, 91–94.

Waltman, L., and van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. The European Physical Journal B *86*, 471.

Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science *361*. Published online July 27, 2018. https://doi.org/10.1126/science.aat5691.

Xiong, Y., Mahmood, A., Meng, Y., Zhang, Y., Zhang, Z.G., Morris, D.C., and Chopp, M. (2012). Neuroprotective and neurorestorative effects of thymosin β4 treatment following experimental traumatic brain injury. Ann. N Y Acad. Sci *1270*, 51–58.

Xu, W., Xin, L., and Gong, Y. (2003). Document Clustering Based On Non-negative Matrix Factorization. In SIGIR '03 Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM New York), pp. 267–273.

Xu, X., Coats, J.K., Yang, C.F., Wang, A., Ahmed, O.M., Alvarado, M., Izumi, T., and Shah, N.M. (2012). Modular genetic control of sexually dimorphic behaviors. Cell *148*, 596–607.

Yang, Z., and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics *32*, 1–8.

Zeisel, A., Hochgerner, H., Lonnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Haring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular Architecture of the Mouse Nervous System. Cell *174*, 999–1014.

Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Healthy adult substantia nigra brain tissue	University of Maryland Brain & Tissue Bank; http://www.medschool.umaryland.edu/btbank/	UMBN6060, UMBN6063, UMBN5534, UMBN5893, UMBN5828, UMBN5840, UMBN5862
C57BL/6J Mouse bed nucleus brain tissue	The Jackson Laboratory	JAX:000664
Critical Commercial Assays		
Chromium Single Cell 3′ Library & Gel Bead Kit v3	10X Genomics	Cat#1000075
Chromium Single Cell 3′ Library & Gel Bead Kit v2	10X Genomics	Cat#120237
Chromium i7 Multiplex Kit	10X Genomics	Cat#120262
Deposited Data		
Single nucleus RNA-seq from mouse bed nucleus	This paper	GEO: GSE126836
Single nucleus RNA-seq from human substantia nigra	This paper	GEO: GSE126836
Software and Algorithms		
CellRanger v. 3.0.2	10X Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation
GOrilla	Eden et al., 2009	http://cbl-gorilla.cs.technion.ac.il/
reviGO	Supek et al., 2011	http://revigo.irb.hr/index.jsp
methylpy	Luo et al., 2017	https://github.com/yupenghe/methylpy
Annotate v. 1.60.0	Bioconductor	https://www.bioconductor.org/packages/release/bioc/html/annotate.html
FIMO	Grant et al., 2011	http://meme-suite.org/doc/fimo.html
LIGER	This paper	https://github.com/MacoskoLab/liger

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Evan Macosko (emacosko@broadinstitute.org).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Animals

Nuclei suspensions for the BNST experiments were generated from 8 adult male and 7 adult female mice (60-70 days old; C57BL/6J, Jackson Labs). All experiments were approved by and in accordance with Broad Institute IACUC protocol number 0120-09-16.

Human postmortem tissue

Nuclei suspensions for the human substantia nigra experiments were generated from seven tissue donors (five males, three females, age range 18-75), supplied by the University of Maryland Brain Bank, through the NIH NeuroBioBank. Although all seven donors were coded as neurotypical controls, information provided with the tissue revealed that donor 5828 suffered a traumatic brain injury at the time of death, while donor 5840 was diagnosed at autopsy with cerebral amyloid angiopathy. This work was determined by the Office of Research Subjects Protection at the Broad Institute not to meet the definition of human subjects research (project ID NHSR-4235).

METHOD DETAILS

Liger Workflow

A typical workflow for Integrating multiple single-cell datasets using the LIGER software package consists of the following steps:

- Dataset preprocessing to produce a raw digital gene expression (DGE) matrix.
- Variable gene selection (Saunders et al., 2018), normalization by number of UMIs, and scaling of individual genes. We scale but do not center gene expression because NMF requires non-negative values.
- Identifying shared and dataset-specific factors through integrative non-negative matrix factorization (iNMF). We have derived and implemented a novel coordinate descent algorithm for efficiently performing the factorization.
- Jointly clustering cells and normalizing factor loadings.
- Visualization using t-SNE or UMAP and analysis of shared and dataset-specific marker genes.

The important components of each step are described more formally and in detail below, while vignettes which describe specific commands for analyses included in Figure 2 are available online at https://macoskolab.github.io/liger/.

Performing iNMF Using Block Coordinate Descent

We developed a novel block coordinate descent algorithm for performing integrative non-negative matrix factorization (Yang and Michailidis, 2016). This approach learns a set of latent metagene factors, each with both shared and dataset-specific components, to approximate the original datasets. To estimate these matrix factors, we minimize the following objective:

$$\arg\min_{H\geq 0,\ W\geq 0, V\geq 0} \sum_i^d \|E_i - H_i(W+V_i)\|_F^2 + \lambda \sum_i^d \|H_i V_i\|_F^2 \quad (1)$$

This approach attempts to reconstruct each of the original datasets E_i (of dimension $n_i \times m$) using lower dimensional matrices H_i, W, and V_i, such that $Ei \approx H_i(W + V_i)$, where all factor matrices are constrained to be non-negative. Note that W is shared across all datasets $i = 1 \ldots d$, while the H_i and V_i matrices are unique to each dataset. The inner dimension k of these factors can be interpreted as the number of "metagenes" (or conversely "metacells") used to represent the datasets.

We divide the variables into $2d + 1$ blocks (corresponding to H and V for each dataset, as well as W) and perform block coordinate descent, iteratively minimizing the objective with respect to each block, holding the others fixed. We iterate:

1. $$W = \arg\min_{W\geq 0} \left\| \begin{pmatrix} H_i \\ \vdots \\ H_d \end{pmatrix} W - \begin{pmatrix} E_1 - H_1 V_1 \\ \vdots \\ E_d - H_d V_d \end{pmatrix} \right\|_F^2$$

2. $$H_i = \arg\min_{H_i\geq 0} \left\| \begin{pmatrix} W^T + V_i^T \\ \sqrt{\lambda} V_i^T \end{pmatrix} H_i^T - \begin{pmatrix} E_i^T 1 \\ 0_{g\times n_i} \end{pmatrix} \right\|_F^2$$

3. $$V_i = \arg\min_{V_i\geq 0} \left\| \begin{pmatrix} H_i \\ \sqrt{\lambda} H_i \end{pmatrix} V_i - \begin{pmatrix} E_i - H_i W \\ 0_{n_i\times g} \end{pmatrix} \right\|_F^2 (0_{c\times d} \text{ is the zero matrix of dimension } c \times d).$$

until convergence. Each of the optimization subproblems above requires solving a nonnegative least-squares problem; we use the fast block principal pivoting algorithm developed by Kim et al. (Kim.et al, 2014) to solve each of these subproblems exactly. As described in Kim et al., our block coordinate descent algorithm satisfies the requirements of the theorem of Bertsekas (Bertsekas, 1999), because each of the subproblems is convex with respect to the block of variables being optimized. Thus, the algorithm is guaranteed to converge to a fixed point (local minimum). In contrast, the multiplicative updates often used for NMF-like optimization problems do not have a convergence guarantee. Additionally, because we solve each subproblem exactly at each iteration, the algorithm converges very quickly; previous empirical benchmarks (Kim.et al, 2014) and our own have shown that block coordinate descent algorithms for NMF generally converge in many fewer iterations than multiplicative updates (Figures S1A and S1B). We developed an efficient implementation of the algorithm using the Rcpp package in R.

Efficient Updating for New K, New Data, and New λ

We adapted a method, previously developed for regular NMF (Kim.et al, 2014), for rapidly updating a factorization given new data or new values of k or λ. Suppose we have optimized (1) with k_1 factors to give matrices $H_i^{(n_i\times k_1)}$, $V_i^{(k_1\times g)}$, and $W^{(k_1\times g)}$. To efficiently compute a factorization with k_2 factors, we consider two cases: $k_2 > k_1$ and $k_2 < k_1$. If $k_2 > k_1$, we initialize the $k_2 - k_1$ new factors to factorize the residual from the previous solution, then solve using alternating nonnegative least-squares as before. For $k_2 < k_1$, we pick the k_2 factors that make the largest contributions to the factorization and solve as before. The following algorithm formalizes this approach.

Algorithm 1: Updating the factorization with a new k

1. Initialize W^{new}, V_i^{new}, H_i^{new} with random nonnegative values.
2. If $k_2 > k_1$:

$$W^{new} = \arg\min_{W \geq 0} \left|\left| \begin{pmatrix} H_i^{new} \\ \vdots \\ H_d^{new} \end{pmatrix} W - \begin{pmatrix} E_1 - H_1(W + V_1) - H_1^{new} V_1^{new} \\ \vdots \\ E_d - H_d(W + V_d) - H_d^{new} V_d^{new} \end{pmatrix} \right|\right|_F^2$$

$$H_i^{new} = \arg\min_{H_i \geq 0} \left|\left| \begin{pmatrix} (W^{new})^T + (V_i^{new})^T \\ \sqrt{\lambda}(V_i^{new})^T \end{pmatrix} H_i^T - \begin{pmatrix} E_i^T - (W^T + V_i^T) H_i^T \\ 0_{g \times n_i} \end{pmatrix} \right|\right|_F^2$$

$$V_i^{new} = \arg\min_{V_i \geq 0} \left|\left| \begin{pmatrix} H_i^{new} \\ \sqrt{\lambda} H_i^{new} \end{pmatrix} V_i - \begin{pmatrix} E_i - H_i(W + V_1) - H_i^{new} W^{new} \\ 0_{n_i \times g} \end{pmatrix} \right|\right|_F^2$$

Set $H_i = \begin{pmatrix} H_i & H_i^{new} \end{pmatrix}$, $V_i = \begin{pmatrix} V_i \\ V_i^{new} \end{pmatrix}$, $W = \begin{pmatrix} W \\ W^{new} \end{pmatrix}$

3. If $k_2 < k_1$: Let $\delta_k = \|H_{i,k}(W_k + V_{i,k})\|_F^2$. Choose the factors with the largest k_2 values of δ_k. Call the corresponding elements of W, H_i, and V_i

Set $H_i = \begin{pmatrix} H_i^{new} \end{pmatrix}$, $V_i = \begin{pmatrix} V_i^{new} \end{pmatrix}$, $W = \begin{pmatrix} W_i^{new} \end{pmatrix}$

4. Perform block coordinate descent optimization using alternating nonnegative least-squares until convergence.

Similarly, we can efficiently compute a factorization with a new λ value. Assume a previous optimization with k factors and λ_1, and with matrices $H_i^{(n_i \times k)}$, $V_i^{(k \times g)}$, and $W^{(k \times g)}$. To compute a new factorization with $\lambda_2 > \lambda_1$, we can use the matrices H_i and W and simply reoptimize with the new λ value. Empirical benchmarks for updating factorizations with new k and λ values are shown in Figures S1C and S1D.

Suppose we have optimized (1) with k factors to give matrices $H_i^{(n_i \times k)}$, $V_i^{(k \times g)}$, and $W^{(k \times g)}$. To efficiently compute a factorization incorporating new data from the same condition as an existing dataset, we use the previous W and V matrices as initial values and find the optimal H given this starting point. Given a new dataset altogether, we initialize the W and V matrices using the values from the dataset that we expect *a priori* to be the most similar and find the optimal H given these values. To re-run on a subset of the data, we use the W and V matrices as a starting point and simply drop the H rows corresponding to the omitted data. For each case, we subsequently perform optimization using the same block coordinate descent strategy as described above. Algorithm 2 summarizes this approach. Empirical benchmarks for updating factorizations with subsets of data and new data are shown in Figures S1E and S1F.

Algorithm 2: Adding new data to the factorization

1. Given new data from one of the datasets already factorized or a completely new dataset, initialize W and V_i to their previously found values (if a completely new dataset, choose V_i from the dataset that is expected to be most similar), then solve:

$$H_i^{new} = \arg\min_{H_i \geq 0} \left|\left| \begin{pmatrix} W^T + V_i^T \\ \sqrt{\lambda} V_i^T \end{pmatrix} H_i^T - \begin{pmatrix} (E_i^{new})^T \\ 0_{g \times n_i} \end{pmatrix} \right|\right|_F^2$$

Set $H_i = \begin{pmatrix} H_i \\ H_i^{new} \end{pmatrix}$.

2. To update the factorization for only a subset of the data, initialize H_i by simply dropping the H_i rows corresponding to the omitted data.
3. Perform block coordinate descent optimization using alternating nonnegative least-squares until convergence.

Heuristics to guide choice of k and λ

We devised novel heuristics to aid in selecting the number of factors k and the tuning parameter λ. To choose k, we calculate the Kullback-Leibler divergence (compared to a uniform distribution) of the factor loadings for each cell and plot the median across cells as a function of k. We then try to observe a saturation point in the curve, corresponding to the point at which more factors do not significantly change the sparsity of the factor loadings or correspondingly increase the median KL divergence. The intuition behind this heuristic is that, if the number of factors is too low, factors will encode combinations of clusters, and thus cells will load on multiple factors, with the distribution of factor loadings for a particular cell approaching a uniform distribution. As the number of factors approaches the "true" number of clusters in the dataset, each cell will generally be expected to load on only a few factors. To select an appropriate lambda, we plot the alignment metric (after performing factorization and basic alignment with default parameters) as a

function of lambda and again choose the minimum value at which the alignment metric saturates. KL divergence curves and alignment-lambda curves for two benchmark datasets are shown in Figures S1I and S1J.

Joint clustering and factor normalization

After optimizing the iNMF objective function, we use the factor space to identify corresponding cell types across datasets (Figure 1C). Because of the parts-based nature of the factorization, cells can be clustered by simply assigning each cell to the factors on which they have the highest loading. This is a common way of performing clustering using NMF representations (Xu et al, 2003). If we perform such simple assignment, we get clusters that correspond across datasets, because the factors represent the same signal in each dataset. However, we noticed that simple maximum factor assignments sometimes produced spurious alignments in highly divergent datasets. Therefore, we developed a novel clustering strategy that better leverages the dataset-specific information in the factorization to increase the robustness of the joint clustering results. We build a *shared factor neighborhood* graph in which we connect cells across datasets that have similar factor loading patterns, then identify joint clusters by performing community detection on this graph.

More specifically, we build the shared factor neighborhood (SFN) graph as follows:

1. Build k-nearest neighbor graphs separately for each dataset using the H factor loadings
2. Annotate each cell i with the H factor on which the cell has the highest loading; call this $F(i)$. We scale each factor to unit variance before assigning $F(i)$ values, as is standard with NMF (Xu et al, 2003).
3. For each cell i, collect the "factor neighborhood" vector $FN(i)$ by computing a histogram of $F(i)$ for each of its k nearest neighbors.
4. Calculate Manhattan distance between pairs of cells (i,j) across (and within) datasets using the factor neighborhood vectors $FN(i)$ and $FN(j)$
5. Connect pairs of cells with low distance; these represent cells with shared factor neighborhoods.

We then perform Louvain community detection on the graph to jointly identify cell clusters across datasets (Waltman and van Eck, 2013).

One intuition behind this construction is that, *within* an individual dataset, it is much more likely for an individual cell to have a spurious maximum factor loading than for an entire neighborhood of cells to be incorrectly assigned. Thus, leveraging the neighborhood of a cell in assigning it to a cluster increases the robustness of the assignment. This approach is synergistic with iNMF, which reconstructs each dataset separately, accurately preserving the structure of individual datasets. A recent paper used a related idea to refine cluster assignments from NMF by taking into account the neighborhood of each data point (Tripodi et al, 2016). Additionally, our graph construction greatly reduces the chances of spurious matches *across* datasets, because even if a cell type spuriously loads on the same factor as a different cell type in another dataset, they are unlikely to have the same factor neighborhoods.

After performing SFN clustering, we choose a reference dataset (by default the dataset with the largest number of cells) and normalize the quantiles of the factor loadings for each joint cluster in the other datasets to match the quantiles of the reference dataset for that joint cluster. We require that each dataset have a minimum number of cells assigned to a particular cluster (default: 2 cells); cells not satisfying this requirement are not normalized.

Calculating alignment and agreement metrics

We calculated the alignment metric as defined in Butler et al. (Butler et al., 2018). To quantify how well the integrated factor space respects the geometry of each individual dataset, we calculate an "agreement metric" as follows. We first apply a dimensionality reduction technique to each dataset separately. Using these low-dimensional representations, we build a k-nearest neighbor graph for each dataset. We also build k-nearest neighbor graphs for each dataset using the joint, integrated space—the normalized H factor loadings from LIGER or the aligned canonical components from Seurat. We then count how many of each cell's nearest neighbors in the graphs built from the separate low-dimensional representations are also nearest neighbors in the graphs built from the integrated low-dimensional representations. We found that PCA and NMF produced significantly different graphs for different joint dimensionality reduction and alignment techniques, so to ensure a fair comparison, we compared the graphs built from the aligned Seurat space (CCA-based) with graphs built from PCA, and compared the graphs built from the aligned LIGER factor loadings (iNMF-based) with NMF.

For comparisons with Seurat, we followed all documented procedures for running the method, including choosing the number of canonical components. For the PBMC and pancreas datasets, we reproduced the analyses as described by Butler et al. and subsequent Seurat tutorials (Butler et al., 2018).

Generation of BNST nuclei profiles

Tissue was sliced from the anterior until reaching the target region of interest (Figure S3A). To dissect a tissue segment, the block face was punched with a 1 mm biopsy punch, and a 400-micron slice was made to liberate the circular tissue segments from the remaining brain. Nuclei were extracted from frozen tissue using gentle, detergent-based dissociation, according to a protocol (available at

protocols.io) adapted from one generously provided by the McCarroll lab, and loaded into the 10x Chromium V3 system. Reverse transcription and library generation were performed according to the manufacturer's protocol.

Generation of profiles from human substantia nigra

Frozen tissue samples were sectioned on a cryostat to visualize pigmentation of the pars compacta, and manually dissected to remove as much surrounding tissue as possible. Nuclei from the dissected tissue segments were then immediately isolated (Saunders et al., 2018) and profiled on the 10x Chromium V2 system, according to the manufacturer's protocol.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of bed nucleus data

We performed a first round of LIGER analysis on BNST nuclei to identify neurons, then restricted subsequent analyses to neuronal cells. We next assigned neuronal populations to anatomical locations using *in situ* staining data from the Allen Brain Atlas (Figure S3), then performed a second round of clustering on neurons we could assign to the bed nucleus. Clustering of bed nucleus neurons identified 46 populations, three of which were removed because of artifactual signatures (high expression of mitochondrial genes, activity-related genes, or doublets), and two small clusters that expressed markers localizing them outside the boundaries of BNST (Islands of Calleja and lateral septum, respectively). We found it was not necessary to quantile normalize the factor spaces for the bed nucleus analyses, as iNMF alone provides sufficient batch effect correction in this case.

To identify genes with statistically significant dimorphic expression, we performed a bootstrapping procedure designed to control for biological variation within each sex. To do this, we sampled between 1 and 8 replicates uniformly at random from the male mice, sampled between 1 and 7 replicates uniformly at random from the female mice, and computed the male/female fold change for each gene in each cluster for each bootstrap sample. We performed 1000 bootstrap samples. This analysis yielded empirical distributions for the fold changes of each gene in each cluster under the biological variation within each sex. We report these values in Data S1. To further identify a high-confidence set of dimorphic genes shown in Figure 3I, we retained only genes for which (1) log2(fold change) was greater than 0.5 with 95% confidence; (2) the gene was expressed in at least 10% of the cells in the sex with the higher expression; and (3) the gene was expressed in no more than 25% of the cells in the sex with the lower expression. These filters were designed to remove genes with subtle expression shifts and genes with extremely low expression.

We also used the iNMF factorization to identify genes with dimorphic expression by simply taking the genes with nonzero loadings on the dataset-specific factors V. As an additional filtering step, for gene i loading on dataset-specific factor j, we removed i if it had a log fold change of less than 0.75 or fewer than 30 UMIs in the cells with their highest factor loading values on factor j. For the analysis in Figure 3J, we also restricted to only factors with high loadings on the BNSTp and BNSTpr clusters, which showed the highest number of dimorphic genes in our bootstrapping analysis and were previously shown to be the most dimorphic anatomical region of the BNST (Xu et al., 2012). To rank dimorphic genes by their degree of dimorphism (as shown in Figure 3K), we calculated the difference, for each gene, between male and female dataset specific factor loadings for the factors on which the gene is dimorphic.

For the comparison of BNST_Vip and BNSTp_Cplx3 nuclei with CGE-derived cortical interneurons, we randomly sampled 30 cells from each of the 11 published subclusters of frontal cortex global cluster 1 (Saunders et al., 2018), to maintain cell-type diversity in the downsampled set. These 330 CGE neurons were analyzed with the 352 BNST_Vip and BNSTp_Cplx3 profiles with LIGER according to parameters in Table S1. For the comparison of the *Ppp1r1b+* BNST populations with striatal SPNs, we sampled 5,000 dSPNs (published global cluster 10), 5,000 iSPNs (published global cluster 11), and 934 eSPNs (subclusters 1-5 of global cluster 13). These 10,934 SPNs were analyzed with the 8,624 *Ppp1r1b+* nuclei from the BNST analysis using LIGER with the parameters in Table S1, generating nine clusters. Four small clusters were identified as doublets and removed (715 cells total), and the remaining five clusters were merged into four coarser groups (iSPN, dSPN, eSPN, and BNST-unique identities). For the SPN co-clustering, to reduce over-clustering of dominant populations, the quantile-aligned factors were re-clustered using the SLM algorithm (Gierahn et al., 2017), with a resolution parameter setting of 1.2.

Analysis of substantia nigra data

To analyze the human substantia nigra data across donors, we used a separate dataset-specific factor matrix for each human donor. We performed two rounds of clustering with LIGER, first identifying the main cell classes (neurons, endothelial cells, astrocytes, oligodendrocytes, and microglia), then clustering each cell type again to identify additional substructure within these classes (Saunders et al., 2018). For the cross-species analysis, we determined homology relationships using Jackson Laboratories annotations (http://www.informatics.jax.org) and included only genes with one-to-one homologs. We used LIGER to integrate each broad cell class separately, and used two dataset-specific factor matrices (one for each species), rather than treating the data from each human donor as a separate dataset. For the cross-species analysis, we performed variable gene selection on each species separately, then took the union of variable genes across both species.

For the identification of gene ontology categories with conserved patterns of expression across species, we used GOrilla (Eden et al., 2009) in ranked gene list mode, using default settings, to perform gene ontology enrichment analysis and ReviGO (Supek et al., 2011) to summarize and visualize the results.

Analysis of STARmap data

We downloaded the published STARmap expression data and segmented cell boundaries, then normalized and scaled the STARmap gene counts in the same manner as the Drop-seq data of mouse frontal cortex (Saunders et al., 2018). Because the STARmap data assays pre-selected markers, we did not perform variable gene selection, but used all genes that were present in both the Drop-seq and STARmap data. We used the entire frontal cortex dataset from a previously published Drop-seq atlas of mouse brain. We performed two levels of analysis using LIGER, first jointly identifying broad cell classes, then performing a second round of LIGER analysis on excitatory neurons, inhibitory neurons, and glia.

We developed a simple method for predicting the spatial distributions of genes not measured in STARmap data. To do this, we simply compute a cross-dataset *k*-nearest neighbor graph in the aligned factor space. Then we set the value of each missing gene to the unweighted arithmetic mean of its *k* nearest neighbors in the other dataset. We used k = 50 for all analyses in the paper. We assessed the accuracy of the predicted genes by calculating mean absolute deviation (MAD) from the measured STARmap values for genes present in both datasets. As a baseline model, we predicted each gene by setting the value of each gene in each STARmap cell to its average value in the Drop-seq cells within the joint cluster to which the STARmap cell was assigned. We found that our knn prediction strategy resulted in lower MAD than the baseline model for 938 of 989 genes. Visual inspection of genes with clear spatial trends showed that LIGER is able to capture even complex spatial expression patterns (Figure S5). Our predictions agreed strongly with corresponding *in situ* hybridization data of the same genes in the Allen Brain Atlas (Figure S5)

To investigate failure modes of the spatial prediction, we sorted genes present in both STARmap and Drop-seq by their prediction error (defined as MAD between measured and imputed values) and plotted the 10 genes with highest error (Figure S5).

Analysis of methylation data

We downloaded the publicly available gene-level mCH fractions from 3378 frontal cortex profiles (Luo et al., 2017). Because gene expression and gene body mCH are generally anticorrelated, we created gene-level methylation features that correspond to gene expression features by calculating

$$max(X) - X$$

where X is the matrix of mCH values. Because the data from Luo et al. contains only neuronal cells, we used only cells annotated as neurons from the Saunders et al. frontal cortex Drop-seq dataset.

Because the distribution of methylation values is very different from scRNA-seq data, we selected genes by performing a Kruskal-Wallis test on methylation and RNA data separately using the published clusters, then taking the intersection of the top 8000 RNA and methylation markers. We found that, since the methylation data are not sparse (i.e., very few values are exactly 0), the NMF factor loadings need to be centered as well as scaled during the SFN clustering procedure. We also noted that the standard multiplicative update NMF algorithm converges extremely slowly compared to our ALS implementation when running on the non-sparse methylation data. We used the methylpy Python package to calculate methylation levels for the set of differentially methylated regions identified in the original analysis of the methylation data (Luo et al., 2017). We performed an analysis of transcription factor binding using FIMO (Grant et al., 2011) with default settings and the mouse transcription factor binding motifs from the latest version of the non-redundant JASPAR database (Khan et al., 2018). We annotated the DMRs according to the nearest gene using the *annotate* R package.

We identified a set of sequence elements that: (1) showed no overlap with any annotated genes or promoters; (2) showed significant cell-type-specific DNA methylation; (3) contain a conserved sequence element; (4) contain a binding motif for a transcription factor that is expressed in the dataset; and (5) have a methylation profile that is anticorrelated with the expression of the transcription factor whose binding motif occurs in the region. To construct the network of predicted transcription factor/differentially methylated region interactions shown in Figure 7, we identified the transcription factors that showed statistically significant expression differences among neuronal populations (using a Wilcoxon test). Then we identified, for each TF, the top 10 anticorrelated DMRs satisfying the above criteria. Using this set of TFs and DMRs, we connected any TF and DMR for which the correlation between TF expression and DMR methylation across the set of joint LIGER clusters was less than -0.45.

DATA AND SOFTWARE AVAILABILITY

LIGER is freely available as an R package: https://github.com/MacoskoLab/liger

The mouse bed nucleus and human substantia nigra single nucleus RNA-seq data have been deposited in the Gene Expression Omnibus under accession code GEO: GSE126836.

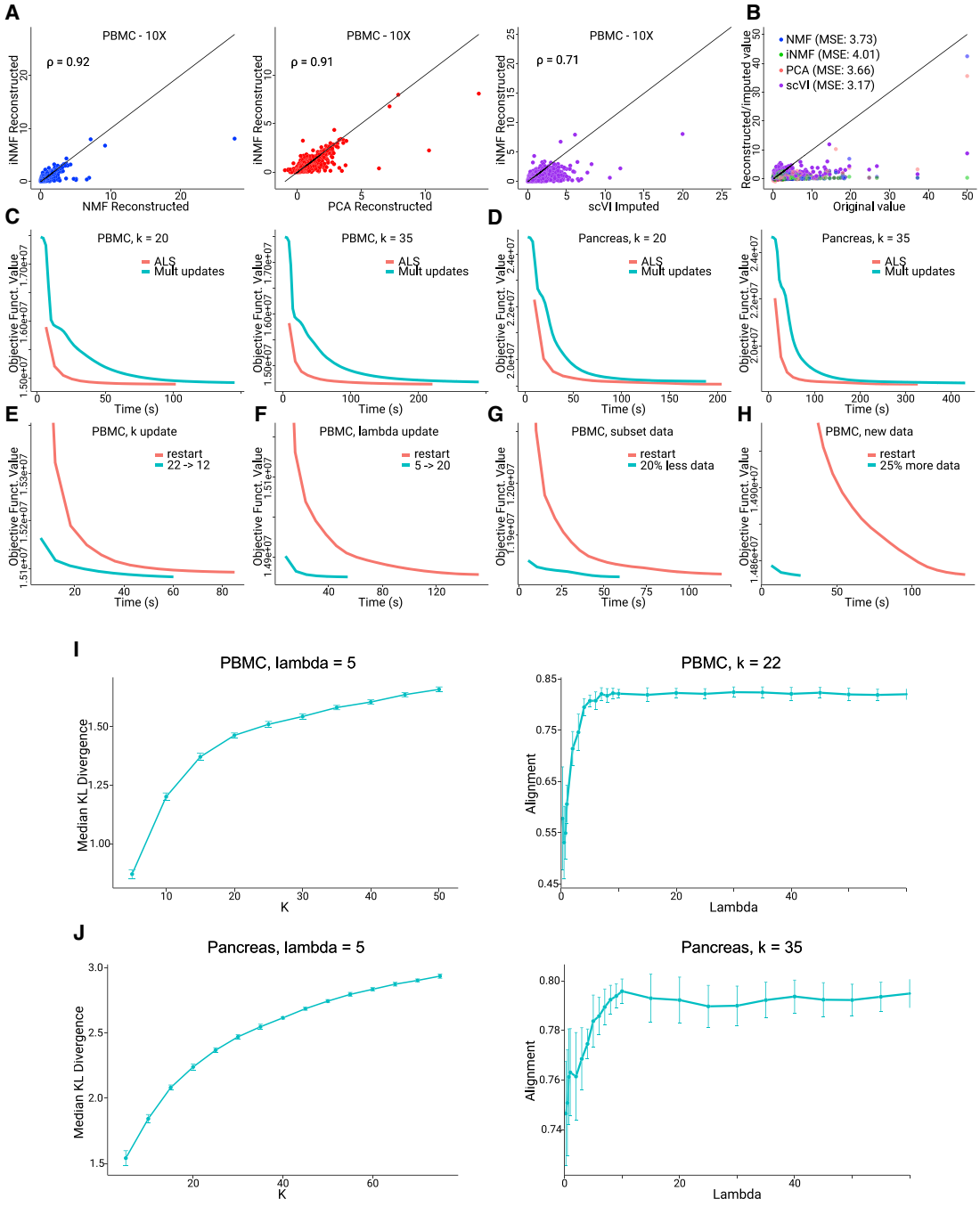


Figure S1. Parameter Selection and Fast Updating, Related to Figure 1

(A) Comparison of reconstructed PBMC 10X gene expression values for iNMF versus NMF, PCA, and scVI (Spearman correlation indicated).

(B) Comparison of reconstructed values and original gene expression values. Comparison calculated for random 1% sample of non-zero values for (A) and (B). Mean squared error values are displayed for each method.

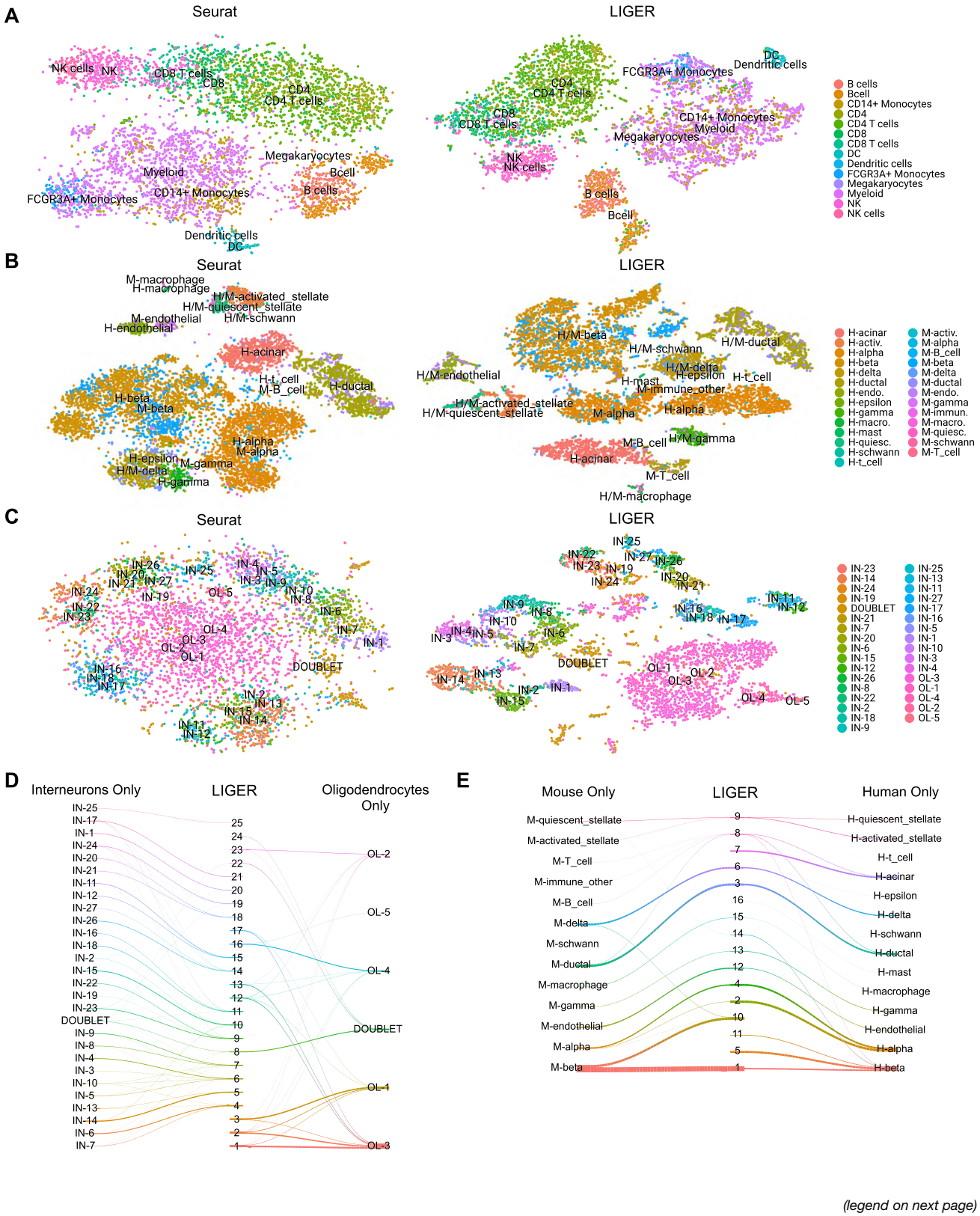
(C and D) Comparison of objective function value as a function of running time for our ALS implementation of iNMF and the original multiplicative update algorithm, for the (C) PBMC datasets and the (D) human-mouse pancreas datasets. Our ALS implementation converged in less time than the original algorithm on these datasets for two different values of *k*.

(E) Comparison of fast updating strategy versus simple recalculation for changing k from 22 to 12.

(F) Comparison of fast updating strategy versus simple recalculation for changing λ from 22 to 12.

(G and H) Fast updating for (G) subsetting the data and (H) adding new data.

(I and J) Parameter selection heuristics for (I) PBMC dataset and (J) pancreas dataset. Error bars represent 95% confidence intervals from twenty random iNMF initializations for KL divergence plots and ten random initializations for lambda-alignment plots.



(legend on next page)

Figure S2. Comparison of LIGER and Seurat for PBMCs and Hippocampal Cells, Related to Figure 2

(A–C) Two-dimensional visualizations of CCA/Seurat (left) and LIGER (right) analyses of (a) PBMC, (b) human and mouse pancreas, and (c) hippocampal interneuron and oligodendrocyte analyses that are shown in Figure 2, colored here by published cluster assignments.

(D and E) River plots comparing clustering in individual and joint LIGER analyses of (D) interneuron and oligodendrocyte cells and (E) mouse and human pancreas data.

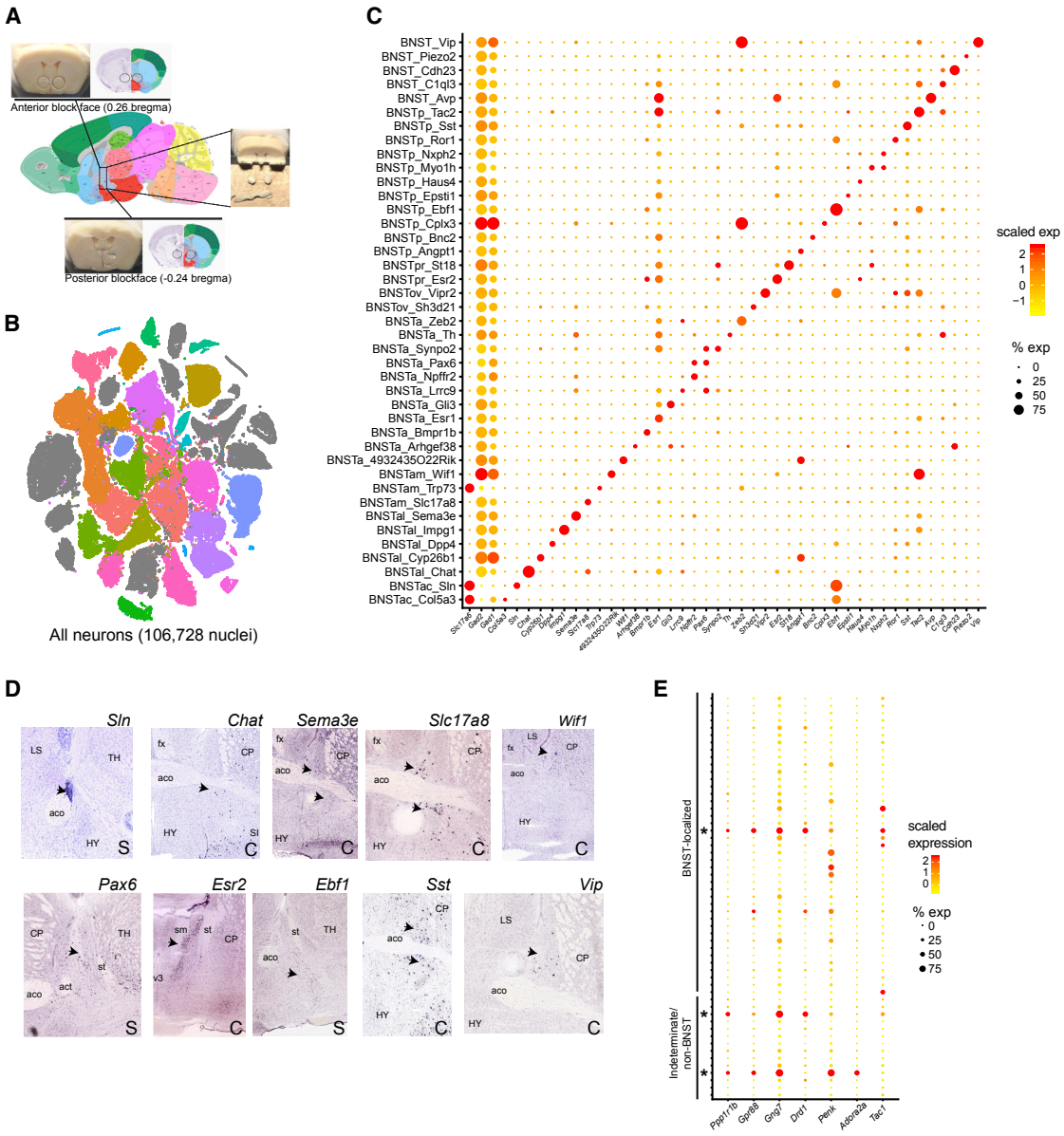


Figure S3. Markers Used to Identify Cell Clusters from the Bed Nucleus, Related to Figure 3

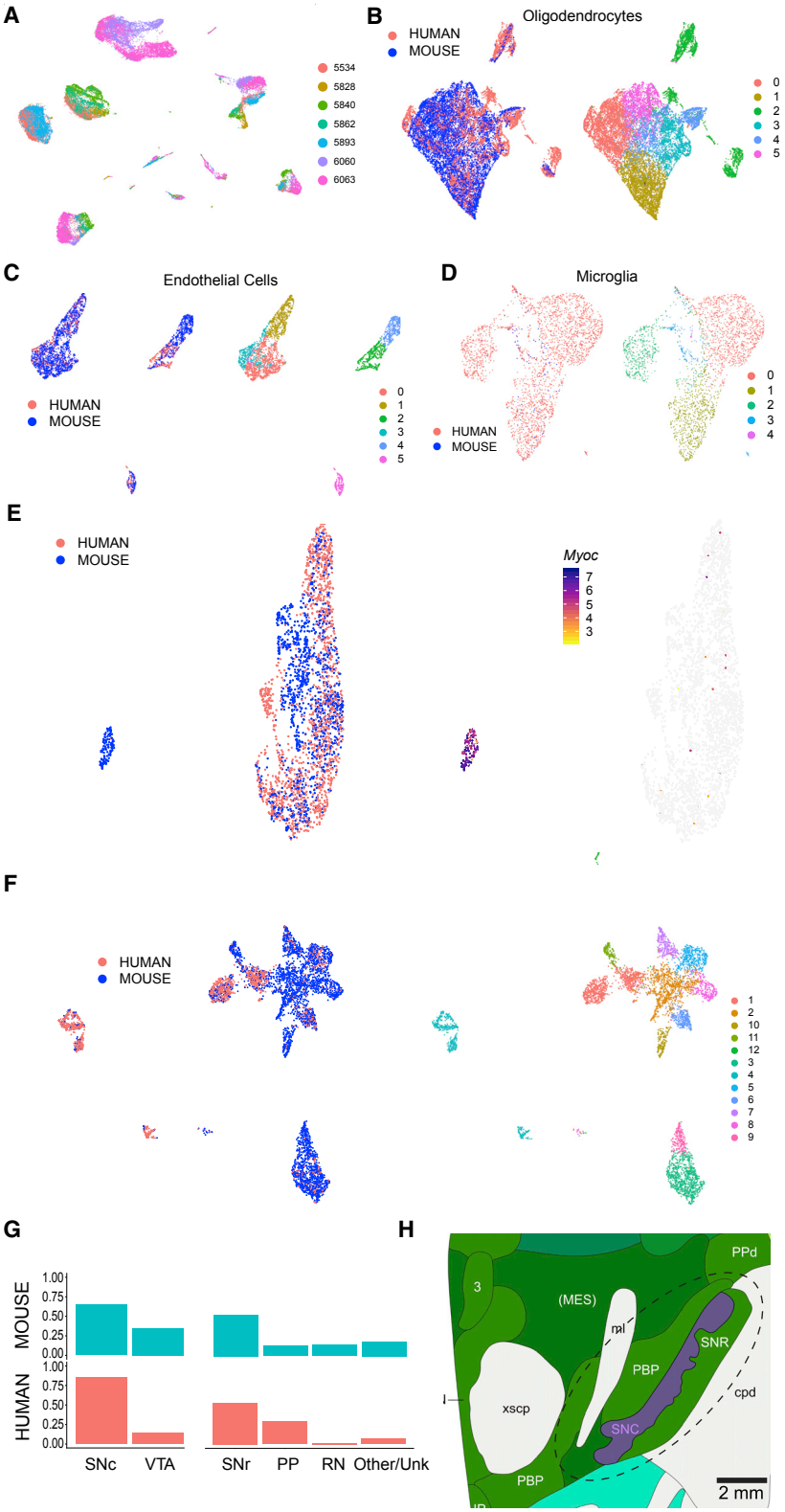
(A) Diagram of dissection strategy for isolating BNST. Arrays denote the dissected tissue fragments for profiling.

(B) t-SNE visualization of LIGER analysis of 106,728 nuclei identified as neurons among the 204,737 nuclei profiled from the tissue dissected in (A). Colored clusters (70.2%) are annotated as localized to BNST; gray clusters were localized to adjacent structures or had an indeterminate localization and were not included in the analysis shown in Figure 3A.

(C) Dot plot of markers of neurochemical identity, and cluster, across the 41 clusters identified in the analysis shown in Figure 3A.

(D) Images showing ISH signals within BNST for marker genes plotted in (C). Arrows indicate signal within BNST region(s). S, sagittal section; C, coronal section; aco, anterior commissure; act, anterior commissure, temporal limb; fx, fornix; st, stria terminalis; v3, third ventricle; CP, caudate putamen; HY, hypothalamus; LS, lateral septum; MS, medial septum; TH, thalamus; SI, substantia innominata.

(E) Dot plot of SPN marker expression across clusters identified in the all-neuron analysis shown in (B). Stars indicate the clusters selected for the SPN analyses in Figures 3F–3H.



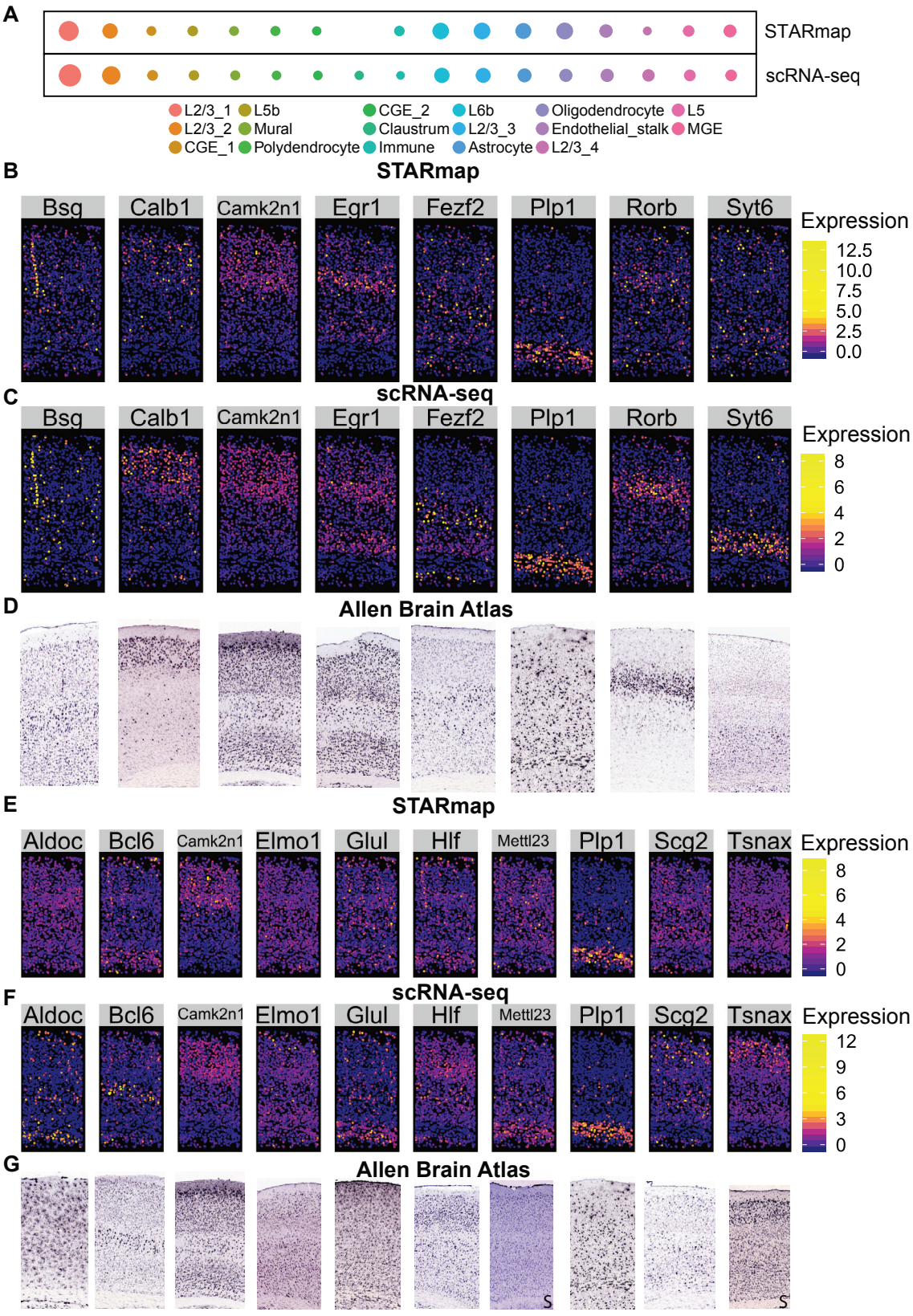
(legend on next page)

Figure S4. Cross-Species LIGER Analyses of Additional Cell Classes Not Displayed in Figure 4, Related to Figure 4

(A) Two-dimensional UMAP representation of a standard Seurat analysis (all datasets merged and scaled together) of the human substantia nigra dataset, colored by donor.

(B–F) Two-dimensional UMAP representations of LIGER human-mouse subanalyses of individual SN cell classes, including (B) oligodendrocytes, (C) endothelial cells, (D) microglia, (E) astrocytes, and (F) neurons. (G) Proportional representation of human nuclei and mouse cells in clusters annotated in the neuron analysis. SNc, substantia nigra pars compacta; VTA, ventral tegmental area; SNr, substantia nigra pars reticulata; PP, peripeduncular nucleus; RN, red nucleus.

(H) Schematic representation of the dissected region of human substantia nigra (dotted oval), sparing several surrounding structures, including the red nucleus. Atlas image is reproduced from the Allen Human Brain Atlas (Ding et al., 2016).



(legend on next page)

Figure S5. Integration of STARmap and scRNA-Seq and Prediction of Spatial Gene Expression Trends, Related to Figure 5

(A) Dot plot showing relative proportions of each joint cluster in STARmap and scRNA-seq datasets. The color of each dot indicates the cluster and the size indicates the proportion of cells belonging to that cluster.

(B–D) Measured STARmap (B), scRNA-seq predicted (C), and Allen Brain Atlas (D) expression levels for selected genes present in both STARmap and scRNA-seq datasets.

(E–G) Measured STARmap (E), scRNA-seq predicted (F), and Allen Brain Atlas (G) expression levels for genes present in both STARmap and scRNA-seq datasets with the 10 highest mean absolute deviation (MAD) values between measured and predicted values (two genes displayed in (B)-(D) for their spatial distribution, *Camk2n1* and *Plp1*, also happen to have high MAD).

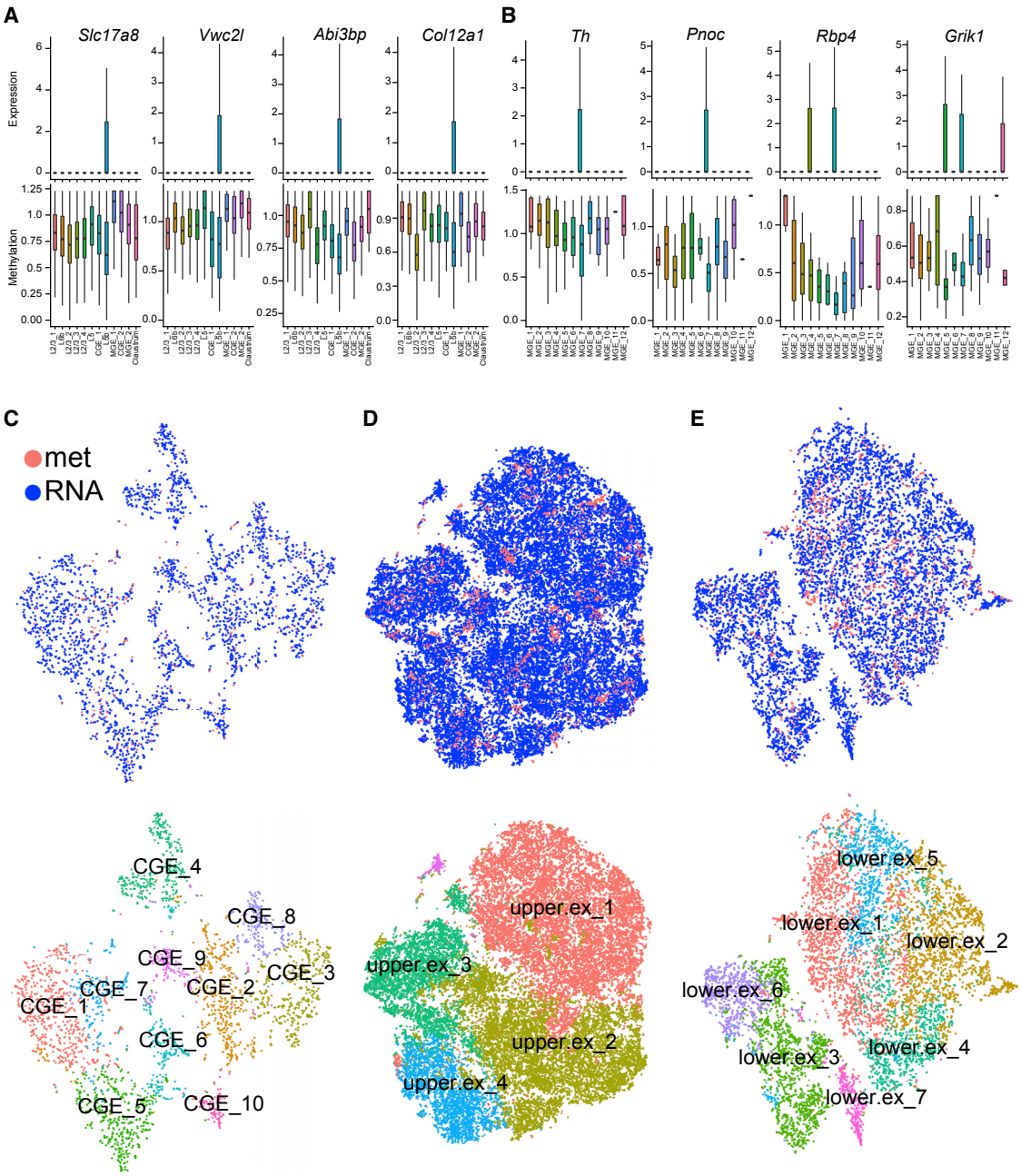
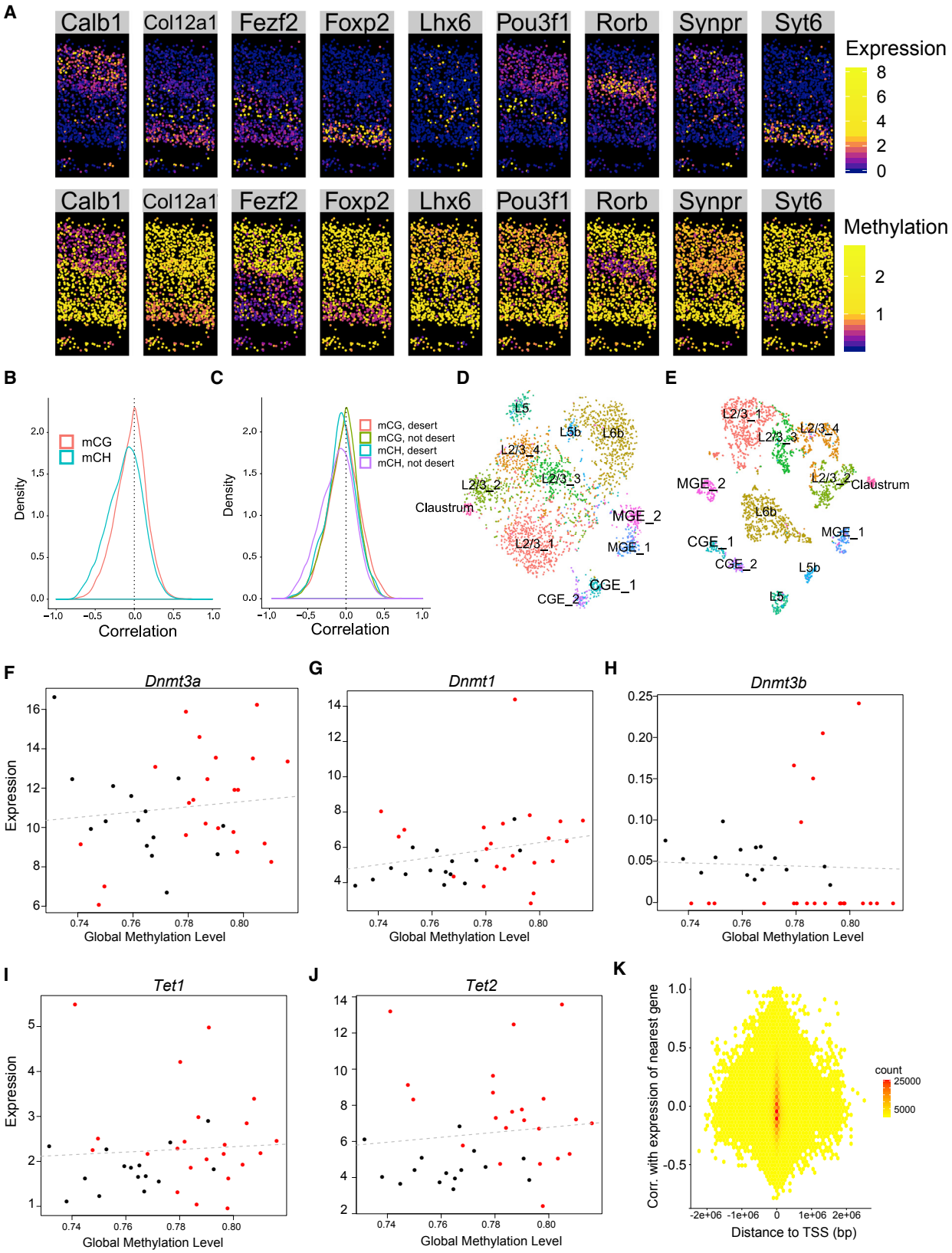


Figure S6. Marker Gene Plots and Additional LIGER Sub-analyses of Joint RNA and Methylation, Related to Figure 6

(A and B) Boxplots are shown highlighting gene expression (top) and methylation (bottom), for selected marker genes of (A) cluster L5b and (B) cluster MGE_7 (Th+).

(C–E) Two-dimensional tSNE representation of LIGER RNA-methylation sub-analyses of (C) CGE interneurons, (D) upper layer excitatory neurons, and (E) lower layer excitatory neurons, colored by measurement modality (top) and cluster identity (bottom).



(legend on next page)

Figure S7. Relationship between Global Methylation and Expression of Methylation Machinery, Related to Figure 7

(A) Predicted spatial expression and methylation patterns. The top row contains expression values predicted by mapping scRNA-seq into STARmap coordinates and the bottom row contains methylation values mapped to scRNA-seq, then to STARmap. Note that there are fewer cells in these plots than in Figure S5 because methylation values are available only for neuronal cells.

(B and C) Density plots comparing the correlation between gene body mCH and gene body mCG with gene expression. The correlations were computed across the LIGER joint methylation and expression clusters. Panel (C) shows both mCH and mCG correlations for all genes and genes within mCH deserts.

(D and E) Plots showing the difference in cluster separation when using gene body mCG (D) or mCH (E).

(F–J) Relationship of expression versus global methylation for selected methylation machinery components. Red dots are inhibitory neuron clusters; black are excitatory neuron clusters, defined by the joint LIGER analysis.

(K) Correlation of methylation and expression of nearest gene as a function of distance to transcriptional start site.