High-throughput mapping of single-cell molecular and projection architecture of neurons by retrograde barcoded labelling

Peibo Xu[1,2,6#], Jian Peng[4,6,7#], Tingli Yuan[1,6#], Zhaoqin Chen[1,6#], Ziyan Wu[1], Ting Li[4], Xiaodong Li[1,2], Luyue Wang[5], Wu Wei[5], Chengyu T. Li[1,3], Zhen-Ge Luo[4*], Yuejun Chen[1,3*]

1Institute of Neuroscience, State Key Laboratory of Neuroscience, Chinese Academy of Sciences, CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai 200031, China

2University of Chinese Academy of Sciences, Beijing 100049, China

3Shanghai Center for Brain Science and Brain-Inspired Intelligence Technology, Shanghai 201210, China

4School of Life Science and Technology, ShanghaiTech University, 393 Huaxia Middle Road, Shanghai 201210, China

5CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Science, Shanghai 200031, China

6These authors contributed equally

7Present address: BGI-ShenZhen, Shenzhen 518103, China

*Correspondence: luozhg@shanghaitech.edu.cn; yuejunchen@ion.ac.cn

Highlights:

- MERGE-seq recovers single-neuron transcriptome and projectome simultaneously
- MERGE-seq partitions projection pattern into neuronal layer and subtype
- MERGE-seq bridges projection pattern to differentially expressed genes
- Machine learning models interpret relationship between digital projectome and gene expression

Abstract

Deciphering mesoscopic connectivity of the mammalian brain is a pivotal step in neuroscience. Most imaging-based conventional neuroanatomical tracing methods identify area-to-area or sparse single neuronal labeling information. Although recently developed barcode-based connectomics has been able to map a large number of single-neuron projections efficiently, there is a missing link in single-cell connectome and transcriptome. Here, combining single-cell RNA sequencing technology, we established a retro-AAV barcode-based multiplexed tracing method called MEGRE-seq (Multiplexed projEction neuRons retroGrade barcodE), which can resolve projectome and transcriptome of source neurons simultaneously. Using the ventromedial prefrontal cortex (vmPFC) as a proof-of-concept neocortical region, we investigated projection patterns of its excitatory neurons targeting five canonical brain regions, as well as corresponding transcriptional profiles. Dedicated, bifurcated or collateral projection patterns were inferred by digital projectome. In combination with simultaneously recovered transcriptome, we find that certain projection pattern has a preferential layer or neuron subtype bias. Further, we fitted single-neuron two-modal data into a machine learning-based model and delineated gene importance by each projection target. In summary, we anticipate that the new multiplexed digital connectome technique is potential to understand the organizing principle of the neural circuit by linking projectome and transcriptome.

Introduction

A wiring diagram of brain can be divided into three levels: the macroscale connectome studying inter-areal connections, the mesoscale connectome describing cellular level connections, and the microscale connectome dissecting synaptic level connections[1]. By studying neural circuit architecture, the mesoscale connectome can be regarded as an intermediate level description of

how distinct information flow among each brain grid[2]. Traditionally, a neuroanatomical tracer is used to portray regional connectivity matrix[3]. To obtain cell-type-specific connectivity, one can use recombinant virus-based tracer in a transgenic model organism or more precisely trace a specific subset of a neural circuit by viral-genetic tracing of the input-output organization[4–6]. However, these methods highly rely on complex recombinant virus design and genetically modified model organism, as well as be obscure or predefined in the function of neuron subtype.

Connectome at single-neuron resolution is believed to transform our understanding of a functional brain unit. Previous studies using high-throughput fluorescence microscopy have been able to reconstruct a full morphology of a single neuron[7,8]. Besides, harnessing the power of next-generation sequencing, the barcode-based circuit mapping technique allows one to study single neuron projection in a high-throughput manner[9–12]. Although these significant progress have revolutionized the view of neuronal diversity and connection complexity, there is still not a comprehensive demonstration of the connectome, considering the unclear role of the neuron subtype in its corresponding input/output information[13–15]. Although the recently developed method can assess the relationship between projection pattern and gene expression of single intratelencephalic neurons, the accessible number of genes is limited by highly multiplexed single-molecule fluorescence in situ hybridization (smFISH) and there are potential difficulties in being jointed with in situ transcriptomics[16].

To address this challenge, we modified the rAAV2-retro virus vector by inserting a 15-bp random barcode as a predetermined index readable for commercial single-cell RNA-seq platform (10X Chromium)[17]. As a proof of concept, we chose to investigate vmPFC, a crucial brain region involving internal information processing such as memory and emotions, and internal goal-directed action[18]. By injecting indexed retrograde AAVs into five canonical projection targets of vmPFC: agranular insular (AI)[19], dorsomedial striatum (DMS)[20],

basolateral amygdala (BLA)[20], lateral hypothalamus (LH)[21,22], and medial dorsal nucleus (MD)[23], we were able to acquire gene expression pattern and projection pattern of excitatory projection neurons, simultaneously. We show that major neuron subtypes of vmPFC are barcoded and then recovered, spanning from layer 2/3, layer 5, layer 6, and lack of layer 4. Collateral projection patterns were easily observed without tedious imaging data registration. More intriguingly, we find that certain projection patterns are preferentially restricted to transcriptionally defined subtypes. Moreover, we implemented a machine learning-based method to establish a mathematical model to describe transcriptional relevance for individual projection target. Bridging the gap of single-neuron projectome and transcriptome, we assume that this method can shed light on the underlying principle of neural circuit assembly and provide a comprehensive view of intricate neural information processing of vmPFC.

Results

Accessing single-neuron transcriptome and projectome simultaneously.

To make transcripts of rAAV2-retro virus infected cells compatible with 10X Chromium scRNA-seq system, we first modified the virus vector by inserting a 15-bp barcode index and polyadenylation signal sequence at the C-terminus of EGFP, driven by short CAG promoter (Fig. 1a,b). Then, five types of the distinct barcoded viruses were injected into selected canonical efferent projection brain regions of P60 mouse vmPFC: AI, DMS, BLA, LH, and MD. One and a half months later, mice were sacrificed and vmPFC, referring to prelimbic area (PrL) and infralimbic area (IL) in this study, were carefully dissected to perform scRNA-seq. Single-cell transcriptional libraries were obtained following library preparation protocol according to 10X Genomics, and projection barcode expression libraries were obtained by using user-defined primers to enrich cDNA fragments of projection barcode index, unique molecular identifiers (UMIs) , and

cell barcode (Fig. 1a). To preserve the intactness of the whole vmPFC transcriptome and projectome, all single-cell suspensions generated were loaded for scRNA-seq. A total of 21,261 single cells passing quality control (QC) from 3 independent Gel Beads-in-Emulsion (GEM) wells were recovered (Fig. 1c). Another 1791 EGFP-sorted cells passing QC were also recovered and most of them (about 95.48% in sorted cells) are excitatory neurons as expected[24] (Supplementary Fig. 1a,b). By transcriptional profiling of all cells, major cell types were recovered as previously reported[25] (Fig. 1d,e).

To determine barcoded cells with confident projectome, we calculated UMI counts of empirical cumulative distribution recovered from each projection target and set a stringent cumulative density threshold of 0.95, where corresponding UMI counts are determined as filtering threshold (see Methods, Supplementary Fig. 1a-f). Across all cell types detected, barcoded cells are mainly distributed among excitatory neurons rather than inhibitory neurons or other non-neurons (3883 barcoded in 10005 excitatory neurons and 26 barcoded in 1425 inhibitory neurons, Fig. 1f,g and Supplementary Fig. 1g,h), which is consistent with the previously observed phenomenon that projection neurons in mPFC are excitatory[24].

Neuronal subtype distribution of projection index barcoded excitatory neurons.

Focusing on the excitatory neuronal population, we extracted the annotated excitatory cluster by examining the Slc17a7 expression, a protein also known as vesicular glutamate transporter Vglut1 (Supplementary Fig. 2a). We examined them roughly by typical markers of layers[14,25] (layer 2/3, Cux2; layer 5, Etv1; layer 6, Sulf1; Supplementary Fig. 2b-g) and further clustered them into 7 subtypes by calculated differentially expressed genes (DEGs) of each cluster (3.8% Calb1, 7.4% Rorb, 4.1% Bcl6, 3.4% Htr2c, 15% S100b, 11.9% Npy and 54.4%

Syt6; Fig. 2a and Supplementary Table 1). Excitatory neuron subtypes are transcriptionally similar to neuron subtypes as previously reported[25], except that we captured more layer 6 neurons than superficial layer neurons (cluster Syt6, Fig. 2b). This collection or capturing bias is possibly due to differences in dissociation protocol. Layer and subtype marker genes are also found expressed specifically in the corresponding layers of vmPFC according to in situ hybridization image data provided by Mouse Brain of Allen Brain Atlas (Supplementary Fig. 2b-g and Supplementary Fig. 3a-g).

Valid retrograde barcoded cells spanned all distinguished layers of vmPFC (layer 2/3, layer 5, and layer 6) and recapitulated all 7 neuron subtypes, which are transcriptionally similar to mPFC excitatory neuronal subtypes of the previous reported dataset (Fig. 2b-c and Supplementary Fig. 3h). Among neuron subtype Htr2c, only 6 neurons were determined as valid barcoded, possibly due to non-preferential projecting to our chosen five targets.

Neurons projecting to DMS were most abundant (n=1961), whereas neurons projecting to BLA were rare (n=329). These statistics are comparable with the number of neurons observed by conventional fluorescence-based retrograde tracing and also comparable to rat[24]. Since our method can acquire transcriptome and projectome simultaneously, it is easy to map two-modal categorical data to discover whether there is any neuron subtype-specific program of projection. We calculated the distribution of all five projection targets in all neuronal subtypes. We found that neurons projecting to LH and MD are mainly distributed in cluster Syt6 (70.00% of all neurons projecting to LH and 81.16% of all neurons projecting to MD), while neurons projecting to AI, DMS, and BLA are mainly distributed in cluster S100b, cluster Npy and cluster Rorb (for cluster S100b, 47.35%, AI; 42.58%, DMS; 53.50%, BLA; for cluster Npy, 26.01%, AI; 19.33%, DMS; 18.20%, BLA; for cluster Rorb, 20.09%, AI; 19.36%, DMS; 12.20%, BLA; Fig. 2e). We further found neurons projecting to different targets have distinct layer distribution preferences as indicated by immunohistochemistry

staining. DMS-projecting and AI-projecting neurons are mainly located in layer 5 and layer 2/3 (for ipsilateral DMS, layer 5, 57%±5.5%; layer 2/3, 21%±6.8%; layer 6, 22%±4.6%; for AI, layer 5, 48%±6.9%; layer 2/3, 33%±7.6%; layer 6, 19%±1.8%; Fig. 2f-g), while MD-projecting neurons are mainly located in layer 5 and layer 6 (layer 5, 44%±4.1%; layer 6, 50%±4.4%; layer 2/3, 6%±3.1%; Fig. 2f-g). And no significant difference of layer distribution was observed between ipsilateral and contralateral side for AI, DMS, and MD (Fig. 2f-g and Supplementary Fig. 3i). The statistics inferred from digital projectome and transcriptome are comparable to immunohistochemistry staining results (Fig. 2e,g), which may suggest layer distribution inherency of neuronal subtypes.

Projection diversity in the mouse vmPFC.

To explore the projection diversity of excitatory neurons in mPFC, we binarized the projection pattern of five types of valid barcoded neurons. One-target projections (dedicated projection) are dominant among total recovered projection patterns (73.83% of 3706 neurons). For collateral projection patterns (greater than one-target projection), about 21.84% of neurons have two-target projections (bifurcated projection), 3.72% have three-target projections, and 0.61% of neurons send axons to more than three targets (Fig. 3a). By calculating the conditional probability of the same neurons projecting to two targets, we found that DMS is more likely to receive vmPFC input from neurons with collateral projections DMS/AI and DMS/BLA (Fig. 3b). A relatively high conditional probability can also be observed with MD/LH projection pattern (Fig. 3b). We used an established statistical analysis method to compare our experimental results to a null model assuming the independent probability of projection between targets, simply as binomial distribution of projection[11,26]. We found that 5 projection patterns are overrepresented and 13 projection patterns are underrepresented after performing adjustment p-values for multiple comparisons (Bonferroni correction, Supplementary Table 3, Fig. 2c-d), and all projection patterns greater than three-target are not significantly over- or under-represented

(Fig. 2c-d). These significantly represented projection patterns are about 58% (18/31) of all, suggesting that a majority of vmPFC neurons project to selected five targets in a non-random manner.

Here, we only focus on the top 10 frequent patterns, including five dedicated projections (DMS, AI, MD, LH, and BLA) and five bifurcated projections (DMS/AI, MD/LH, DMS/MD, DMS/BLA, and DMS/LH). We conducted a principal component analysis (PCA) on projection matrix and mapped binary projection labels on PC embeddings. We found that binary projection clustering results aligned well with clusters at PC1- and PC2-defined embeddings, except scattered distribution of DMS/BLA bifurcated group (Fig. 3e and Supplementary Fig. 4a).

By computing hierarchical clustering on given 10 projection patterns, considering projection strength, there is a clear hierarchical branch of corticothalamic projection (MD-projecting and LH-projecting) and subcortical (DMS-projecting and BLA-projecting) or intracortical (AI-projecting) projection. When we annotated these neurons by binary projection clusters, DMS/LH-bifurcated projection neurons were grouped together with LH- and MD-projecting neurons, including LH-dedicated, MD-dedicated, and LH/MD-bifurcated projection pattern, while DMS-dedicated projection neurons were grouped together with AI- and MD-projecting neurons (Fig. 3f). This may reflect the projection bias to certain targets in bifurcated projection patterns.

Relating distinct projection patterns to differentially expressed genes.

Next, we want to discover whether there is any transcriptional difference between neurons with distinct projection patterns. Taking advantage of MERGE-seq, we explored the differentially expressed genes (DEGs) of each projection pattern. We performed a differential expression analysis using MAST and found several projection-specific gene sets (Fig. 3g). For dedicated projection patterns,

Foxp2 and Crym are specific genes expressed in neurons targeting LH and MD, Nptxr, Cnr1, Cck and Nrn1 are specific genes expressed in neurons targeting AI, DMS and BLA (p-value cutoff=10^{-20}, Log_2 fold change cutoff=1; Fig. 3h-k and Supplementary Fig. 4b).

We asked whether there is any transcriptional difference between dedicated projection neurons and collateral projection neurons. In general, DEGs between pattern A/B versus A or pattern A/B versus B are rare (Supplementary Fig. 5a-i), indicating the transcriptional similarity is dominated by one or the other projection target. For instance, it is interesting to notice that DMS/LH-bifurcated projection neurons are similar to LH-dedicated projection neurons rather than DMS-dedicated projection neurons, according to the dendrogram of projection heatmap (Fig. 3f).

Validating inferred bifurcated projection and projection-specific gene expression.

Taking advantage of MERGE-seq, we mapped top 10 projection patterns with single-neuron transcriptomic clusters (Fig. 4a). We found one distinct collateral projection pattern (DMS/LH) which is transcriptionally distinct to DMS, but not to LH (Fig. 4a-c). Moreover, this projection pattern is mainly contributed by neuron subtype Bcl6 (67.07%, Supplementary Fig. 5j, fourth pie chart).

To validate the bifurcated or collateral projection patterns inferred by digital projectome, we injected two different color-coded retrograde viruses into two selected projection targets separately and analyzed projection patterns by immunohistochemistry. We observed a comparable amount of dual-color labeled neurons in DMS/LH (17.8%±0.11%) by histological validation. Histological validation also confirms that most dual-color labeled neurons are located in layer5 (69.32%±2.14% of total dual-color labeled neurons, Fig. 4d-f). Similar

results were observed also in in AI/DMS with 23.1%±2.03% dual-color labeled neurons, in MD/LH with 27.9%±4.10% dual-color labeled neurons, in DMS/BLA with 6.59%±1.55% dual-color labeled neurons, and in BLA/LH, which is a bifurcated projection pattern undetected by MERGE-seq, with 1.66%±0.92% dual-color labeled neurons.

Moreover, we used RNA fluorescence in situ hybridization (FISH) to examine our predicted DMS/LH-collateral projection-enriched gene, Pou3f1, along with dual-color tracing patterns. First, we found that Pou3f1 is confined within layer5 (10.8%±7.67% of all Fig. 4g), which is consistent with the previous results[25]. Further, Pou3f1^{+} neurons colocalized with DMS/LH collaterally projecting neurons (white arrowheads, Fig. 4h) and LH dedicatedly projecting neurons (blue arrows, Fig. 4h), but not DMS dedicatedly projecting neurons (white arrows, Fig. 4h). This result is consistent with our prediction based on single-neuron digital projectome data (Fig. 4a-c), indicating the presence of a layer 5 neuron subtype (cluster Bcl6, Pou3f1^{+}) that has a distinct non-overlapping pattern between DMS-dedicated versus LH-dedicated or DMS-dedicated versus DMS/LH-collateral projection patterns. A further step toward the functional and physiological relevance of this group of neurons is worth to be examined.

Machine learning-based modeling of projection and gene expression.

To evaluate the potential relationship of transcriptome and projectome, we used optimized distributed gradient boosting algorithm XGBoost[27] to predict binary projection pattern for each projection targets based on the transcript profile. In detail, we applied the XGBoost on top 150 highly variable genes (HVGs) and randomly chosen 150 genes as a control model (see Methods). Unexpectedly, both models (top 150 HVGs and random 150 genes) gave rise to high prediction accuracy, the area under roc curve (AUC), and F1-score (harmonic mean of the precision and recall), but top 150 HVGs-based models did significantly increase

the accuracy, AUC and F1-score across 10-splits test data validation (Fig. 5a-b and Supplementary Fig. 7a-d). Except for that extreme metric values for BLA group, it is possibly due to few cells in BLA group as model training input (n=329 cells in BLA group). Thus, these results suggest that the top 150 HVGs are more informative when predicting and decoding the projection patterns.

To interpret the important genes contributing to certain valid projection patterns, we used a game-theoretic approach to explain the output of HVGs-based XGBoost models[28]. Effects of HVGs were summarized in SHAP (SHapley Additive exPlanations) values for each projection model (Fig. 5c-d, Supplementary Fig. 7e-g). For example, the Nrip3 gene is a top positive feature for LH projection, which reveals that a cell with high mRNA expression of Nrip3 will have a higher probability of projecting LH; Nptxr gene is a top positive feature for AI projection, revealing that a cell with high mRNA expression of Nptxr will have a higher probability projecting AI. By examining top effective genes on PC embeddings of the projection matrix, we found that positive features share similar cell clustering and distribution with projection barcode index expression or binary projection pattern (Fig. 5e-f).

Discussion

Given the complexity of brain circuits, one would dive into their own study in multiple views, including region-to-region brain connectivity, cell type taxonomy by physiological properties or by transcriptional profile, and decoding information processing of a group of neurons in certain behavioral paradigm. It is intricate to understand brain circuits depending on the information derived from only one modality. Recently, there are several examples delineating the brain circuit coordinating multiple views[16,29–31]. In this study, we have developed a multiplexed barcoding method coordinated with scRNA-seq, which injects retrograde AAV at multiple target regions simultaneously to label source region

projection neurons and enables accessible transcriptome and projectome profile digitaly.

First, the method described here is time-saving and effortless to study projectome. To achieve single-neuron resolution of projectome, it requires a high-level imaging facility, skilled technicians, and labor-intensive image registration. Usually, it would take months to acquire reliable mapping results. MERGE-seq, using barcode-based digital connectome technology, can only take 5~7 weeks, routine virus injection and commercial scRNA-seq platform to investigate brain-region-of-interest, which is a good start for an exploratory study.

Second, MERGE-seq is scalable in downstream projection targets for an individual experimental subject. Current common methods studying structural connection rely on consuming numerous experimental subjects. Powered by the infinite barcode indexes of MERGE-seq, one is able to portray projection patterns of plentiful brain-region-of-interest for an individual experimental brain, which is especially promising for studying less understandable non-human primates brain. Utilizing multiplexed readouts of digital projectome, we can also infer bifurcated or collateral projection patterns further validated by traditional image-based tracing technique. For instance, we only sacrificed 3 mice in this study to observe a single-neuron projectome derived from 5 targets region. This advantage could help researchers to escape from the cumbersome combination of all possible collateral projection patterns using the traditional image-based tracing techniques. While we only gave a proof-of-concept for studying projection pattern or efferent connection map in this study, we believe it is also technically available to establish afferent connection map in a multiplexed way[29,32,33].

Finally, we have attempted to reveal and understand the relationship of region-to-neuron connectivity and gene expression, by jointing single-neuron efferent projection with the transcriptome. Due to rising single-cell technology, the

transcriptional composition of a group of projection-defined neurons can be revealed[2,34]. However, it requires the multiplexed method to get the whole picture of transcriptional neuron subtypes and downstream targets, especially for collateralization at single-neuron resolution. Recently, Jabaudon and colleagues developed $Connect_{id}$ linking connectome with transcriptome using MAPseq and scRNA-seq[29]. MERGE-seq is conceptually similar. Except that our method captures two-modality information in the same neuron at the expense of downstream targets spatial information, while $Connect_{id}$ captures transcriptome at source area and connectome at downstream targets area at the expense of mapping rate and recovery rate. We expect that combining with commercial or self-made spatial transcriptomics technology[35–39], MERGE-seq can add spatial information as a new modality.

Our method mapped neuronal subtypes and projection patterns into a shared coordinate system. In terms of vmPFC, we found that neurons projecting to certain brain regions are homogenous in layer location or subtype composition (neurons, projecting to MD or LH, do not tend to project to DMS, AI or BLA; Fig. 2e, Fig. 3f and Fig. 4a). Moreover, neurons giving certainly bifurcated or collateral projections are also made up of homogenous subtype-specific neurons, which can be further validated by a set of marker genes. For instance, neurons projecting to DMS/LH are mainly cluster Bcl6 (67.07%) specifically expressing Bcl6 and Pou3f1. Interestingly, we found a small group of genes can be assigned as important features given a machine learning model for each projection target. This observation resembles the prediction power of a small number (~10) of genes reported by a recent technique called BRICseq[31]. Even though functional connectivity is not examined and discussed here, we anticipate that a shared coordinate system provided by MERGE-seq could be more precise and less variable in predicting functional connectivity based on gene expression.

Cell capturing bias is a potential caveat. In this study, much more deep layer excitatory neurons were captured than superficial layer excitatory neurons (Fig.

2a-b). We anticipated this result may be an inherent caveat of dissociation-based protocol, which in our case may selectively capture more deep layer excitatory neurons with higher viability than the superficial layer.

It also should be noted that some assigned dedicated projection patterns could possibly be included in unexamined projection targets or not be recovered during retrograde labeling, single-cell capturing, and projection index barcode recovery. These scenarios could decrease the frequency of cells that actually belong to bifurcated or collateral projection patterns.

In summary, we anticipate that MERGE-seq can help researchers to perform big-data driven research at low cost, to understand the organizing principle of the neural circuit across modality, and achieve a more comprehensive mesoscale connectome.

Methods

Experimental methods.

AAV vector design

Plasmid pAAV-CAG-tdTomato was first modified by replacing tdTomato with EGFP by T4 DNA Ligase mediated ligation. 15 bp barcode sequence was then inserted after the stop codon of EGFP, linked by EcoRI restriction enzyme recognition site, by T4 DNA Ligase mediated ligation. Barcode 0 sequences represented AI target is: CTGCACCGACGCATT; Barcode 1 sequences represented DMS target is: GAAGGCACAGACTTT; Barcode 2 sequences represented MD target is: GTTGGCTGCAATCCA; Barcode 3 sequences represented BLA target is: AAGACGCCGTCGCAA; Barcode 4 sequences represented LH target is: TATTCGGAGGACGAC.

Virus injection.

Male adult C57BL/6 mice (8 weeks of age) were anesthetized intraperitoneally by pentobarbital sodium (10mg/mL, 120 mg/kg b.w.) and unilaterally injected with rAAV-EGFP-BARCODE virus into five projection targets simultaneously. Coordinates for these injections are as follows. AI at two locations (from Bregma, in mm: 2.0 AP, 2.52 ML, -2.0 DV; 1.6 AP, 2.97 ML, -2.2 DV), DMS at one location (from Bregma, in mm: 0.6 AP, 1.8 ML, -2.2 DV, 8 degree angle), MD at one location (from Bregma, in mm: -1.25 AP, 1.35 ML, -3.55 DV, 20 degree angle), BLA at one location (from Bregma, in mm: -1.5 AP, 3.2 ML, -4.2 DV), LH at one location (from Bregma, in mm: -0.94 AP, 1.2 ML, -4.55 DV). At the end of each injection, the micropipette was left in tissue for 10 min before slowly withdrawn, to prevent virus spilling and backflow. Mice were sacrificed 6 weeks from virus injection. The single-cell suspension was generated as described in (scRNA-seq sample and library preparation).

For dual-color retrograde virus tracing, one location was unilaterally injected rAAV-EGFP (or rAAV-EGFPnls) and rAAV-tdTomato virus simultaneously. For DMS(EGFP)/LH(tdTomato), DMS(EGFP)/BLA(tdTomato) and MD(EGFP)/LH(tdTomato) group, rAAV-EGFPnls was used to avoid dense fiber staining when performing immunohistochemistry. The injection procedure performed and coordinates were as described above. Note that mice used for layer distribution ratio analysis (Fig. 2f,g) of each target were the same as used for dual-color retrograde virus tracing analysis (Fig. 4d-f and Supplementary Fig. 6)

scRNA-seq sample and library preparation.

Virus injected mice were anaesthetized, followed by transcranial perfusion with ice-cold oxygenated self-made dissection buffer (in mM: 92 Choline chloride, 2.5 KCl, 1.2 NaH2PO4, 30 NaHCO3, 20 HEPES, 25 Glucose, 5 Sodium ascorbate, 2 Thiourea, 3 Sodium pyruvate, 10 MgSO4.7H2O, 0.5 CaCl2.2H2O, 12 N-Acetyl-L-Cysteine). The brain was removed, 300um vibratome sections were collected and the regions of PrL and IL areas were microdissected under a

stereo microscope with a cooled platform. Brain slices were incubated in dissection buffer added 10uM AMPA receptor antagonist CNQX (Abcam, Cat #ab120017) and 50uM NMDA receptor antagonist D-AP5 (Abcam, Cat #ab120003) at 33°C for 30 min. The pieces were dissociated first using the ice-cold oxygenated dissection buffer added papain(20 units/ml, Worthington, Cat #LS003126), 0.067mM 2-mercaptoethanol (Sigma, Cat #M6250), 1.1mM EDTA (Invitrogen, Cat #15575020), L-Cysteine hydrochloride monohydrate (Sigma, Cat #C7880) and Deoxyribonuclease I (Sigma, Cat #D4527), with 30-40 min enzymatic digestion at 37°C, then followed by 30 min 1mg/ml protease (Sigma, Cat #P5147) and 1mg/ml dispase (Worthington, Cat #LS02106) enzymatic digestion at 25°C. Supernatant was removed and digestion was terminated using dissection buffer containing 2% Fetal Bovine Serum (FBS, Bioind, Cat #04-002-1A). Single-cell suspension was generated by manual trituration using fire-polishing pasteur pipettes and filtered through a 35um DM-equilibrated cell strainer (Falcon, Cat #352052). Cells were then pelleted at 400 g, 5 min, supernatant carefully removed and resuspended in a 1-2 ml dissection buffer containing 2% FBS. Then, the cell suspension was proceeded debris removal step using Debris Removal Solution according to the manufacturer's recommendations (Miltenyi, Cat #130-109-398). Cell pellets were then resuspended and 48000 cells were loaded in 3 lanes to perform 10x Genomics sequencing. The rest cells were proceeded with cell sorting for enriching EGFP positive retro-AAV labeled cells. 4893 EGFP positive cells were captured and loaded to perform 10x Genomics sequencing. Chromium Single Cell 3' Reagent Kits (v3) were used for library preparation according to the manufacturer's recommendations (10X Genomics). Libraries were sequenced on an Illumina Novaseq 6000 system.

Projection barcode library preparation.

Parallel PCR reactions were performed containing 50 ng of post cDNA amplification reaction cleanup material as a template, 200nM P5-Read1 (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGC

TC) and 200nM P7-index-Read2-EGFP (CAAGCAGAAGACGGCATACGAGATAGGATTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGgCATGGACGAGCTGTACAAG) as primers, and NEBNext Ultra II Q5 Master Mix (NEB, Cat #M0544L). Amplification was performed according to the following PCR protocol: (1) 33°C for 1 min, (2) 98for 10 s, then 65°C for 75 s (20-24 cycles), (3) 75°C for 5 min. Reactions were re-pooled during 1X SPRI selection (Beckman, Cat #B23317), which harvested virus projection barcodes library. 431bp-437bp (with 120bp adaptors) libraries were sequenced using Illumina HiSeq X Ten.

Immunohistochemistry.

Mice were sacrificed after virus injection for 6 weeks. Briefly, mice were transcardially perfused with phosphate-buffered (PB) followed by 4% paraformaldehyde (PFA). Brain samples were extracted and cryoprotected in 20% sucrose/4% PFA fixed at 4°C overnight and then transferred to 30% sucrose/PB for more than 24h. Brain samples were flash-frozen on dry ice and sectioned at 30μm on a cryostat (SM2010R, Leica). For dual-color retrograde virus tracing, brain slices were blocked in 10% donkey serum and 0.3% Triton X-100 at 37°C for 1h. Slices were then incubated with primary antibodies against the green fluorescent protein (GFP, 1:500, Nacalai, Cat #04404-84) and tdTomato (1:500, OriGene, Cat #AB8181-200) at room temperature for 2h, then 4°C overnight. Slices were washed three times using phosphate-buffered saline (PBS) and incubated with Hoechst 33342 (1:1000, Lifetech, Cat #H3570), secondary antibodies donkey anti-rat Alexa Fluor 488 (1:800, Invitrogen, Cat #A21208) donkey anti-goat Alexa Fluor 568 (1:800, Invitrogen, Cat #A11057) at room temperature for 1h. Slices were washed three times using PBS and coverslipped. Stained slices were imaged with a 4X objective with numerical aperture 0.16 as a map, followed by 1.5 um increment z stacks with a 10X objective with numerical aperture 0.4 (FV3000, OLYMPUS). Composite images were automatically stitched in the X-Y plane using the ImageJ/FIJI.

RNA FISH experiment was performed using RNA-Scope reagents and protocols

(ACD Bioscience, CA) following the manufacturer's instructions for fresh-frozen tissue. For Fig 4. g-h, immunohistochemistry was performed following RNA FISH.

Computational methods.

ScRNA-seq data pre-processing.

ScRNA-seq data were aligned with modified mouse reference genome mm10-3.0.0 adding five projection barcodes as separate genes. This was only taken as preliminary validation that projection barcodes can be compatible with 10X Genomics system. Further projection barcode expression was obtained as described in (**Projection barcode library preparation and Projection barcode FASTQ alignment**). ScRNA-seq data was demultiplexed using the default parameters of Cellranger software (10x Genomics, v3.0.2). Obtained filtered transcription count matrix was used for downstream analysis.

Projection barcode FASTQ alignment.

Demultiplexing projection index barcode was performed using deMULTIplex R package (v1.0.2) (https://github.com/chris-mcginnis ucsf/MULTI-seq). Tag parameters in 'MULTIseq.preProcess' function were adjusted according to our user-defined position of index barcode length and position.

ScRNA-seq transcriptional expression analysis.

The filtered count matrix was analyzed and processed using Seurat and Scanpy, including data filtering, normalization, highly variable genes selection, scaling, dimension reduction, and clustering[40–42]. First, scRNA-seq data from 3 samples of unsorted cells and 1 sample of sorted EGFP positive cells were created as Seurat object separately; genes with less than 3 counts were removed and cells with fewer than 200 genes detected were removed. Second, four Seurat objects were merged using the 'merge' function in Seurat. Details of downstream analysis on merged Seurat object are described as following: (1). Data filtering: cells with a mitochondrial gene ratio of greater than 20% were excluded. Then,

cells with more than 500 genes detected, less than 8000 genes detected (cells with more than 8000 genes detected were considered as potential doublets), more than 1000 counts detected, and less than 60000 counts (cells with more than 60000 counts detected are considered as potential doublets) detected were kept. (2). Data normalization: for each cell, the counts were log normalized with the "NormalizeData" function in Seurat, "scale.factor" was set for 50000. (3). Highly variable genes selection: 2000 highly variable genes were calculated using "FindVariableFeatures" function in Seurat. (4). Data scaling: the Seurat object was performed the "ScaleData" function with default parameters. The number of counts, number of genes, mitochondrial gene ratio, and sorting condition were regressed out in "ScaleData". (5). Principal component analysis: highly variable genes were used to calculate principal components in "RunPCA" function. 100 principal components (PCs) were obtained and stored in Seurat object for computing neighborhood graph and uniform manifold approximation and projection (umap) in following part. (6). Leiden clustering: Seurat object was converted into loom file and imported by Scanpy. A neighborhood graph of observations was computed by "scanpy.pp.neighbors" function in Scanpy. Then, leiden algorithm was used to cluster cells by "scanpy.tl.leiden" function in Scanpy. (6). Cluster merge and trimming: top 200 differential expression genes for each cluster were calculated by "scanpy.tl.rank_genes_groups" function in Scanpy using parameters method='wilcoxon' and n_genes=200. Cluster annotation was done manually based on previously reported markers of PFC all cell type, layer, neuron subtypes, and mouse brain atlas[14,25]. For cell clusters with similar marker genes expression, we merged them into one cluster. Complete markers lists for all cell types and all excitatory neuron subtypes calculated using 'FindAllMarkers' function in Seurat were provided (See Supplemental Table 2).

Cell type replicability assessment.

To evaluate whether transcriptional cell types we recovered and annotated were correlated with previously reported[25], we applied MetaNeighbor R package (v1.6.0)[43] to calculate AUROC score as a performance vector on all cell types

and excitatory neuron subtypes. First, we chose 2000 genes as integration features of our Seurat object and public datasets prepared as Seurat object. Then, we integrated two datasets using anchors found by the 'FindIntegrationAnchors' function. Next, we prepared a normalized data matrix of 2000 anchor genes (variable genes) as SummarizedExperiment class using the SummarizedExperiment R package (v1.16.1). Fast, low memory, and unsupervised version of MetaNeighbor was used to calculate the AUROC score (MetaNeighborUS function, fast version was implemented). Cross-dataset mean AUROC scores were plotted in the heatmap.

Binary projection pattern classification.

To determine valid barcoded cells, we hypothesized that EGFP-sorted cells are positive control and non-neuron cells defined by transcriptional profile are negative control which does not send projection or cannot be retrograded labeled. First, we calculated the UMI empirical cumulative distribution function (ECDF) of each projection target using the 'stat_ecdf' function in the ggplot2 R package (v3.3.1). Then, we set a stringent threshold of UMI cumulative density, 0.95, corresponding UMI counts for each projection index barcode is the threshold we chose (Supplementary Figure 1). Finally, we drop UMI counts of determined unbarcoded cells to zero to obtain the index barcode counts matrix used for downstream analysis. Binary projection patterns were calculated by five projection targets set intersections of corresponding barcoded cells. Only the top 10 frequent projection patterns were kept for reliable inference.

The conditional probability of projection pattern and determination of over- and under-represented projection patterns.

In short, we denoted conditional probability $P(B|A)$ as a proportion of neurons projecting to region B (N_{AB}) within a subgroup of neurons projecting to region A (N_A), thus $P(B|A)=\frac{N_{AB}}{N_A}$.

Methods for statistical significance analysis of over- and under-represented projection patterns were described previously[11,26].

Projection pattern-specific differentially expressed genes analysis.

Differentially expressed genes were calculated using the default parameters of the 'FindMarkers' function in Seurat, except the MAST algorithm was used to do DE testing.

For the DEG volcano plot, the chosen cut-off for statistical significance was 10^{-20}, and chosen cut-off for absolute $\log_2$ fold-change was 1. Volcano plots were implemented using the EnhancedVolcano R package (v1.4.0).

For the DEG heatmap in Fig. 3b, top 10 DEGs ordered by average log2 fold-change were chosen from each binary cluster, and heatmap was implemented using 'sc.pl.heatmap' function in Scanpy.

Machine learning implementation on projection and transcription data.

XGBoost was applied to perform a machine learning classification task. For model simplicity and accuracy, we only consider binary projection labels for each projection target (barcoded and unbarcoded). Top 150 highly variable genes and sized match random genes were used to build a model. Cells were consistent with cells belonging to the top 10 frequent projection patterns. A corresponding normalized data matrix was extracted. Training/test data split was 0.7. Hyperparameters of each model were tuned using 10-fold stratified cross-validation by 'tune_model' function in the pycaret python package (v1.0.0). For each tuned model, we random split test data and computed AUC, accuracy, F1 score, kappa value, precision, and recall across 10 trials.

To interpret contribution of important genes for each HVGs-based XGBoost model, we implemented tree explainer of SHAP python package (v0.32.1) to summarize the effects of genes. SHAP values were calculated using 'shap.TreeExplainer(model).shap_values(test_data)' function, and plotted by 'shap.summary_plot()' function.

Statistical tests.

All statistical analyses and plots were done in R. No statistical methods were used to predetermine sample size. The experiments were not randomized and

investigators were not blinded to allocation during experiments and outcome assessment.

Acknowledgments

We thank Dr. Liye Zhang and Pin Wu for generous advice on some of bioinformatic analysis. This study was supported in part by the National Key Research and Development Program of China (2018YFA0108000), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB32030200), the Shanghai Municipal Science and Technology Major Project 2018SHZDZX05), the National Natural Science Foundation of China (31771137), the Thousand Young Talents Program.

Author contributions

Y.C. and Z.L., conceived experiments. J.P. and Z.W. dissected mouse brain and performed scRNA-seq experiments. T.Y. performed immunohistochemistry experiments. Z.C. performed retro-barcode virus injection experiments. J.P. and Q.L. constructed retro-AAV plasmids. P.X. analysed data with considerable input from Z.W. and T.Y.. X.L., L.W., and W.W. helped with some of bioinformatic analysis. C.T.L., Z.L., and Y.C. supervised the project. P.X. and Y.C. wrote the manuscript.

Code availability

The computational code used in the study is available at GitHub (https://github.com/MichaelPeibo/mpfc_projectome).

Reference

1. Zeng, H. Mesoscale connectomics. *Curr. Opin. Neurobiol.* **50**, 154–162 (2018).
2. Oh, S. W. *et al.* A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
3. Cowan, W. M. The emergence of modern neuroanatomy and developmental neurobiology. *Neuron* **20**, 413–426 (1998).
4. Nassi, J. J., Cepko, C. L., Born, R. T. & Beier, K. T. Neuroanatomy goes viral! *Front. Neuroanat.* **9**, 80 (2015).
5. Schwarz, L. A. *et al.* Viral-genetic tracing of the input-output organization of a central noradrenaline circuit. *Nature* **524**, 88–92 (2015).
6. Ghosh, S. *et al.* Sensory maps in the olfactory cortex defined by long-range viral tracing of single neurons. *Nature* **472**, 217–220 (2011).
7. Gong, H. *et al.* High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nat. Commun.* **7**, 12142 (2016).
8. Rompani, S. B. *et al.* Different Modes of Visual Integration in the Lateral Geniculate Nucleus Revealed by Single-Cell-Initiated Transsynaptic Tracing. *Neuron* **93**, 767–776 (2017).
9. Kebschull, J. M. *et al.* High-Throughput Mapping of Single-Neuron Projections by Sequencing of Barcoded RNA. *Neuron* **91**, 975–987 (2016).
10. Zador, A. M. *et al.* Sequencing the connectome. *PLoS Biol.* **10**, e1001411 (2012).
11. Han, Y. *et al.* The logic of single-cell projections from visual cortex. *Nature* **556**, 51–56 (2018).
12. Chevée, M. et al. Variation in activity state, axonal projection, and position define the transcriptional identity of individual neocortical projection neurons. *Cell Rep.* **22**, 441–55 (2018)
13. Fornito, A., Arnatkevičiūtė, A. & Fulcher, B. D. Bridging the Gap between Connectome and Transcriptome. *Trends Cogn. Sci.* **23**, 34–50 (2019).

14. Sorensen, S. A. *et al.* Correlated gene expression and target specificity demonstrate excitatory projection neuron diversity. *Cereb. Cortex* **25**, 433–449 (2015).
15. Klingler, E. *et al.* A Translaminar Genetic Logic for the Circuit Identity of Intracortically Projecting Neurons. *Curr. Biol.* **29**, 332-339.e5 (2019).
16. Chen, X. *et al.* High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing. *Cell* **179**, 772-786.e19 (2019).
17. Tervo, D. G. R. *et al.* A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. *Neuron* **92**, 372–382 (2016).
18. Le Merre, P., Ährlund-Richter, S. & Carlén, M. The mouse prefrontal cortex: Unity in diversity. *Neuron* 1–20 (2021) doi:10.1016/j.neuron.2021.03.035.
19. Zhu, J. *et al.* Transient Delay-Period Activity of Agranular Insular Cortex Controls Working Memory Maintenance in Learning Novel Tasks. *Neuron* **105**, 934-946.e5 (2020).
20. Hunnicutt, B. J. *et al.* A comprehensive excitatory input map of the striatum reveals novel functional organization. *Elife* **5**, e19103 (2016).
21. Reppucci, C. J. & Petrovich, G. D. Organization of connections between the amygdala, medial prefrontal cortex, and lateral hypothalamus: a single and double retrograde tracing study in rats. *Brain Struct. Funct.* **221**, 2937–2962 (2016).
22. Hurley, K. M., Herbert, H., Moga, M. M. & Saper, C. B. Efferent projections of the infralimbic cortex of the rat. *J. Comp. Neurol.* **308**, 249–276 (1991).
23. Vertes, R. P. Differential projections of the infralimbic and prelimbic cortex in the rat. *Synapse* **51**, 32–58 (2004).
24. Gabbott, P. L. A., Warner, T. A., Jays, P. R. L., Salway, P. & Busby, S. J. Prefrontal cortex in the rat: Projections to subcortical autonomic, motor, and limbic centers. *J. Comp. Neurol.* **492**, 145–177 (2005).
25. Bhattacherjee, A. *et al.* Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. *Nat. Commun.* **10**,

1–18 (2019).
26. Gergues, M. M. *et al.* Circuit and molecular architecture of a ventral hippocampal network. *Nat. Neurosci.* **23**, 1444–1452 (2020).
27. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **13-17-Augu**, 785–794 (2016).
28. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
29. Klingler, E. et al. Single-cell molecular connectomics of intracortically-projecting neurons. Preprint at https://www.biorxiv.org/content/10.1101/378760v1 (2018)
30. Lui, J. H. *et al.* Differential encoding in prefrontal cortex projection neuron classes across cognitive tasks. *Cell* **184**, 489-506.e26 (2021).
31. Huang, L. *et al.* BRICseq Bridges Brain-wide Interregional Connectivity to Neural Activity and Gene Expression in Single Animals. *Cell* **182**, 177-188.e27 (2020).
32. Hanchate, N. K. *et al.* Connect-seq to superimpose molecular on anatomical neural circuit maps. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 4375–4384 (2020).
33. Clark, I. C. *et al.* Barcoded viral tracing of single-cell interactions in central nervous system inflammation. *Science* **372**, eabf1230 (2021).
34. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
35. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science.* **361**, eaat5691 (2018).
36. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci.* **116**, 19490 LP – 19499 (2019).
37. Zhang, M. et al. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. Preprint at

https://www.biorxiv.org/content/10.1101/2020.06.04.105700v1 (2020)
38. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science.* **362**, eaau5324 (2018).
39. Chen, A. *et al.* Large field of view-spatially resolved transcriptomics at nanoscale resolution. Preprint at https://www.biorxiv.org/content/10.1101/2021.01.17.427004v2 (2021)
40. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
41. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
42. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
43. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 1–12 (2018).

Figure Legends

Fig. 1

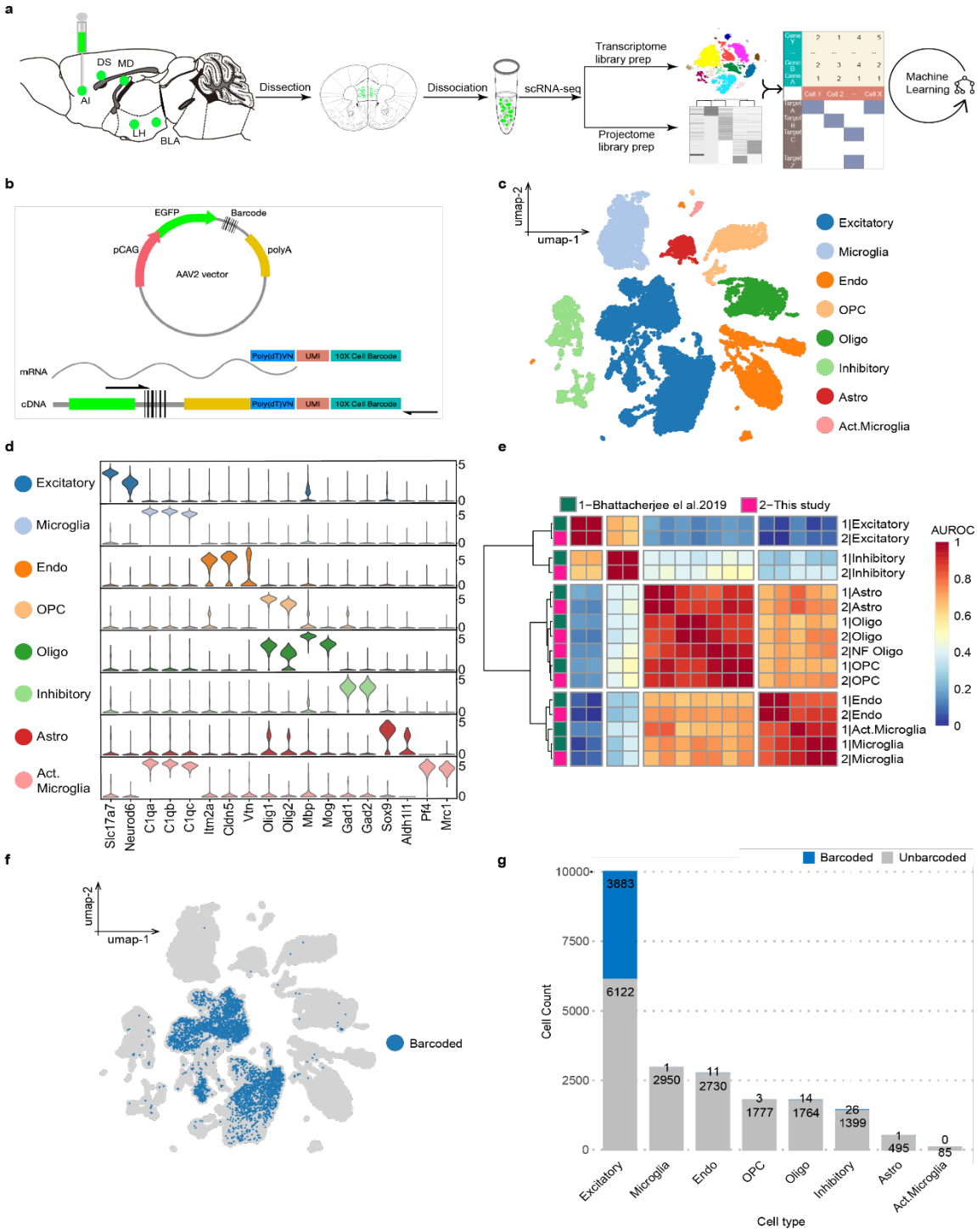


Fig. 1 | MERGE-seq can recover transcriptome and projectome simultaneously. a, Schematic diagram of experimental workflow. **b**, Retro AAV plasmid vector design and projectome library preparation diagram. **c**, Umap embedding of transcriptional clustering results of all vmPFC cells. **d**, Stacked violin plot showing selected markers expression of each cluster. **e**, Heatmap showing correlation between cell types recovered in this study and dataset from Bhattacherjee et al. **f**, Umap embedding of all determined barcoded cells (blue, see Methods). **g**, Bar plot showing frequency of barcoded (blue) and unbarcoded (grey) cells in all recovered cell types.

Fig. 2

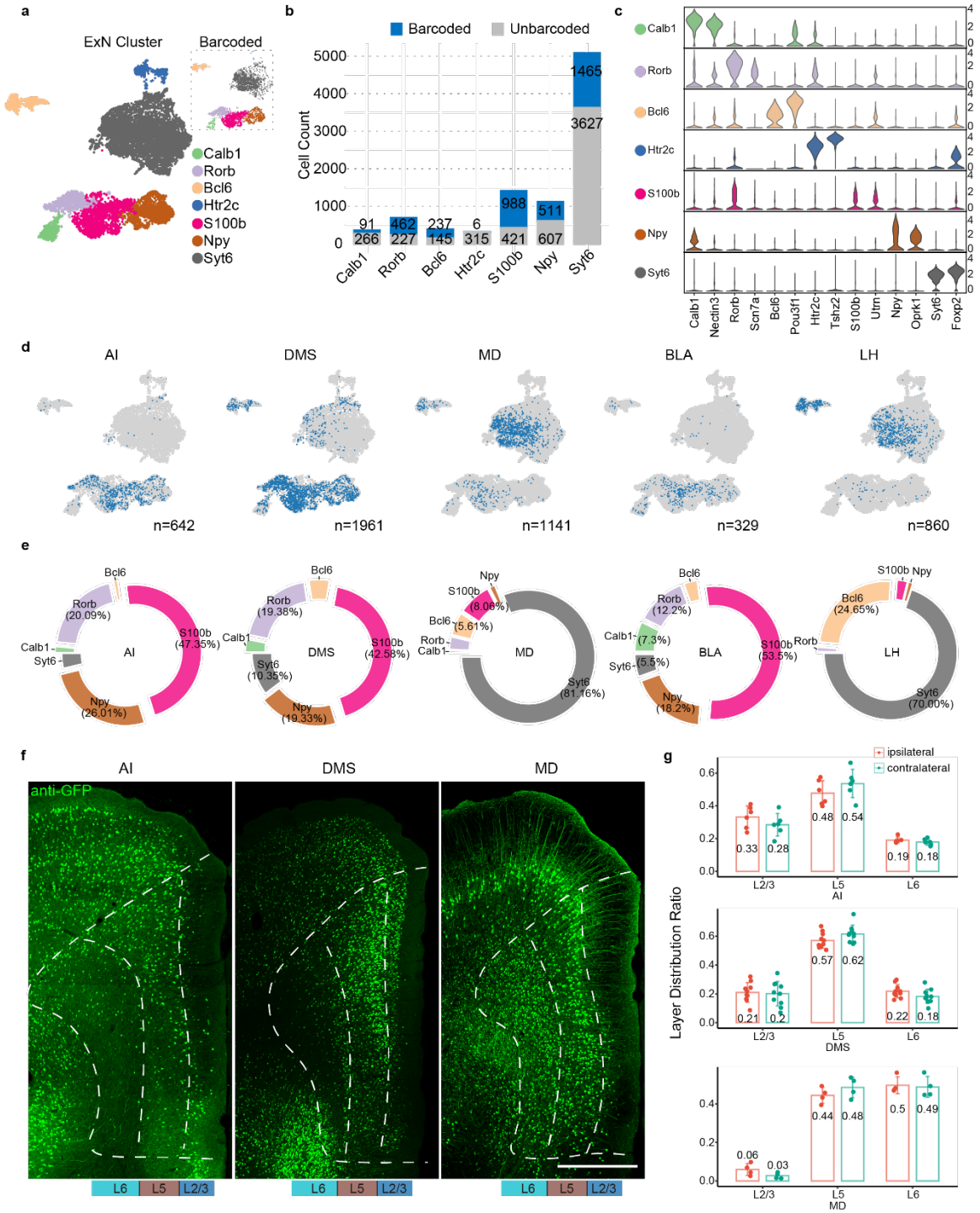

a
ExN Cluster
Barcoded
Calb1
Rorb
Bcl6
Htr2c
S100b
Npy
Syt6
b
Barcoded
Unbarcoded
Cell Count
5000
4000
3000
2000
1000
0
91
266
462
227
237
145
6
315
988
421
511
607
1465
3627
Calb1
Rorb
Bcl6
Htr2c
S100b
Npy
Syt6
c
Calb1
Rorb
Bcl6
Htr2c
S100b
Npy
Syt6
Calb1
Nectin3
Rorb
Scn7a
Bcl6
Pou3f1
Htr2c
Tshz2
S100b
Utm
Npy
Oprk1
Syt6
Foxp2
d
AI
DMS
MD
BLA
LH
n=642
n=1961
n=1141
n=329
n=860
e
AI
Bcl6
Rorb (20.09%)
S100b (47.35%)
Calb1
Syt6
Npy (26.01%)
DMS
Bcl6
Rorb (19.38%)
Calb1
S100b (42.58%)
Syt6 (10.35%)
Npy (19.33%)
MD
Npy
S100b (8.08%)
Bcl6 (5.61%)
Rorb
Calb1
Syt6 (81.16%)
BLA
Bcl6
Rorb (12.2%)
Calb1 (7.3%)
S100b (53.5%)
Syt6 (5.5%)
Npy (18.2%)
LH
S100b Npy
Bcl6 (24.65%)
Rorb
Syt6 (70.00%)
f
AI
DMS
MD
anti-GFP
L6
L5
L2/3
g
ipsilateral
contralateral
Layer Distribution Ratio
0.6
0.4
0.2
0.0
0.33
0.28
0.48
0.54
0.19
0.18
L2/3
L5
L6
AI
0.21
0.2
0.57
0.62
0.22
0.18
DMS
0.06
0.03
0.44
0.48
0.5
0.49
MD

Fig. 2 | Neuronal subtype distribution of retrograde barcoded excitatory neurons. **a**, Umap embedding of excitatory neuron subtype annotation. **b**, Bar plot showing frequency of barcoded (blue) and unbarcoded (grey) neurons in distinct neuron subtypes. **c**, Stacked violin plot showing selected markers expression of each neuronal subtype. **d**, Umap embeddings of barcoded (blue) neurons projecting to each target. **e**, Donut plot describing neuronal subtypes (donut) distribution of barcoded neurons for each projection targets. Threshold to show label as a ratio of total was set to 0.1. Neuronal subtype color codes are same as (**a**). **f**, Typical images showing layer distribution of retrograde AAV labeled neurons projecting AI, DMS and MD from ipsilateral side. Scale bars, 500 um. **g**, Layer distribution ratio of all $EGFP^+$ neurons from both ipsilateral and contralateral sides (AI: n=6, DMS: n=10, MD: n=4). The data are presented as the mean ± SD.

Fig. 3

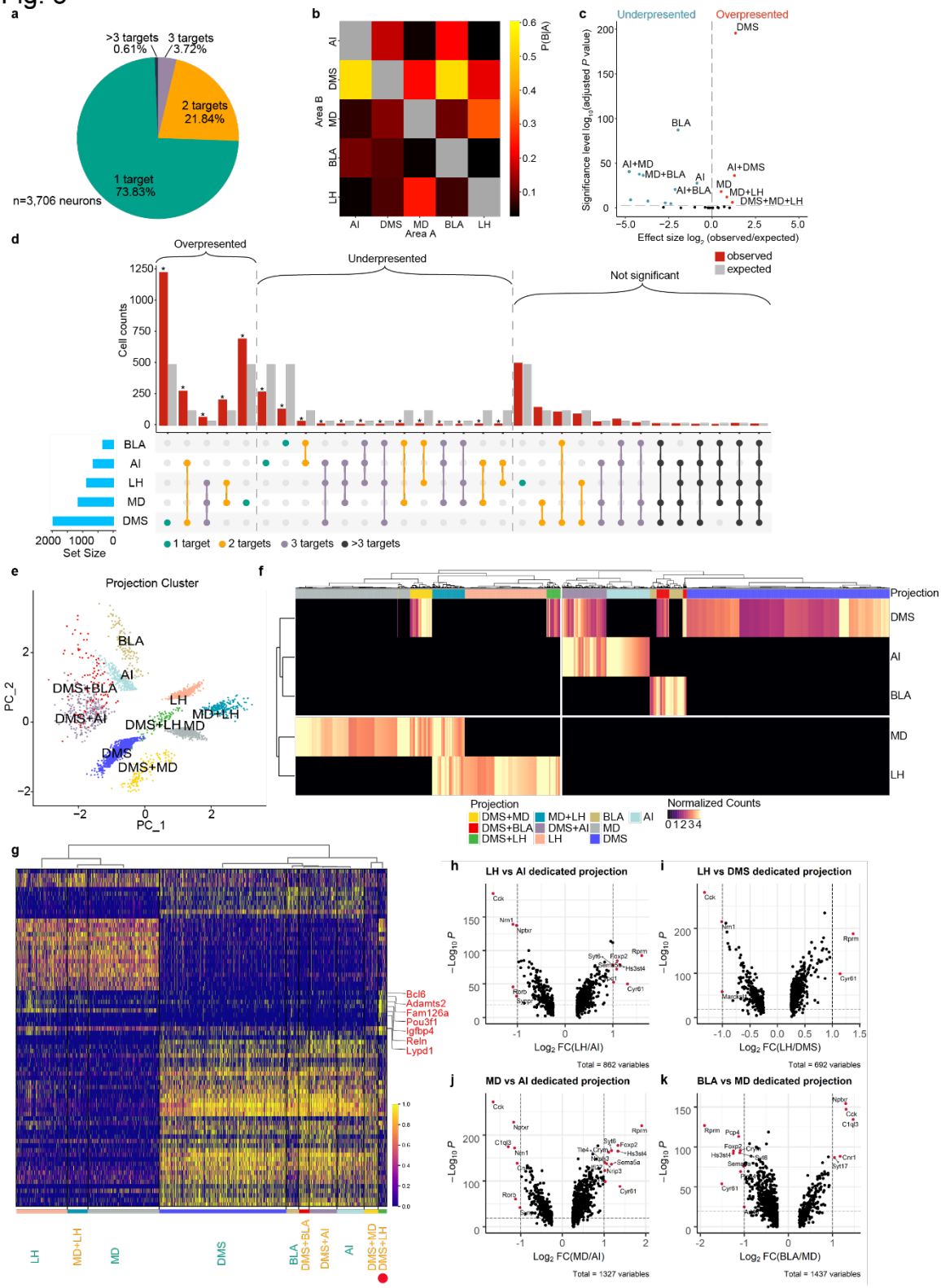

a
>3 targets 0.61%
3 targets 3.72%
2 targets 21.84%
1 target 73.83%
n=3,706 neurons
b
Area B
Area A
AI
DMS
MD
BLA
LH
P(B|A)
c
Underpresented
Overpresented
Significance level log10(adjusted P value)
Effect size log2 (observed/expected)
observed
expected
d
Overpresented
Underpresented
Not significant
Cell counts
Set Size
1 target
2 targets
3 targets
>3 targets
e
Projection Cluster
PC_1
PC_2
f
Projection
Normalized Counts
g
Bcl6
Adamts2
Fam126a
Pou3f1
Igfbp4
Reln
Lypd1
h
LH vs AI dedicated projection
Log2 FC(LH/AI)
Total = 862 variables
i
LH vs DMS dedicated projection
Log2 FC(LH/DMS)
Total = 692 variables
j
MD vs AI dedicated projection
Log2 FC(MD/AI)
Total = 1327 variables
k
BLA vs MD dedicated projection
Log2 FC(BLA/MD)
Total = 1437 variables

Fig. 3 | MERGE-seq can reveal distinct projection patterns of vmPFC and corresponding DEGs. a, Percentage of number of projection targets of barcoded vmPFC neurons recovered by MERGE-seq. **b**, Heatmap showing condition probability of projecting to area A within the same group of neurons projecting to area B. **c**, Volcano plot of adjusted statistical significance level and effect size of over and underrepresented projection patterns (*p-value ≤0.001, adjusted P values and effect sizes are provided in Supplementary Table 3, see Methods). **d**, Upset plot showing intersection of five projection targets. Horizontal blue bar represents frequency of barcoded cells of each projection target. Vertical red and grey bar represents observed and expected neurons of each projection pattern, respectively. Comparison was performed between observed and expected neurons using two-sided binomial tests and p values are adjusted by Bonferroni method. **e,** Binary projection annotation labeled on PC1 and PC2 embeddings. Note that only top 10 frequent binary projection patterns are included. **f**, Heatmap showing normalized projection strength. Rows represent five projection targets and columns represent cells labeled by top 10 binary projection patterns or labeled by transcriptional neuron subtypes. According to clustering results of rows, MD and LH are hierarchically similar, while DMS, AI and BLA are hierarchically similar, at first two hierarchical branch. According to clustering results of columns, projection pattern MD, LH, MD+LH, DMS+MD and DMS+LH are hierarchically similar, while projection pattern DMS, BLA, AI, DMS+BLA and DMS+AI are hierarchically similar, at first two hierarchical branch. **g,** Heatmap showing scaled expression of calculated DEGs based on 10 projection patterns. Top 10

DEGs ordered by average log_2 fold change of each pattern was selected. Genes labeled in red were DEGs of DMS/LH collateral projection neurons. **h-k**, Volcano plots of selected dedicated projection patterns. Assigned DEGs (red dots) were determined using threshold: Log_2 fold change=1, p value cutoff= 10^{-20}.

Fig. 4

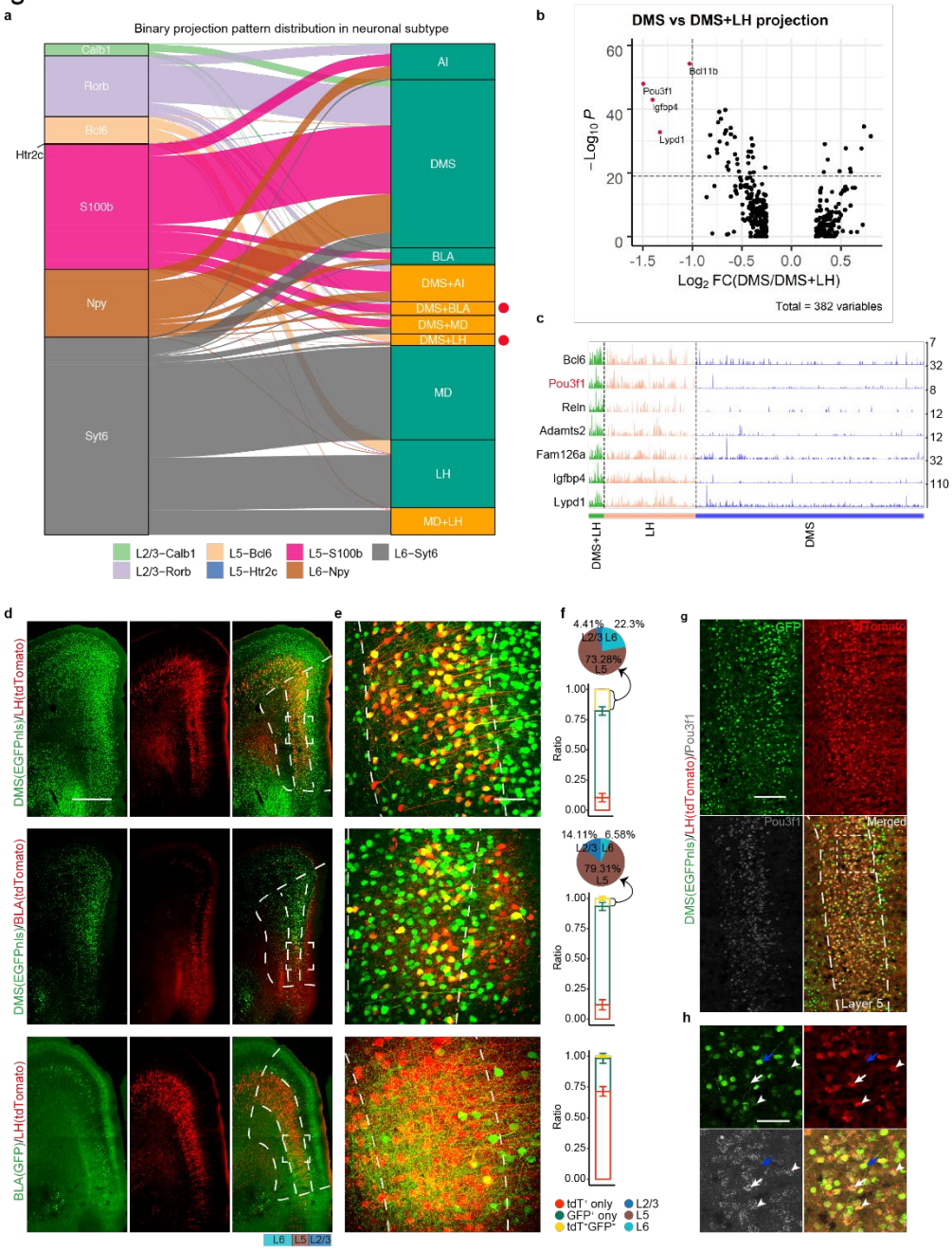

a
Binary projection pattern distribution in neuronal subtype
Calb1
Rorb
Bcl6
Htr2c
S100b
Npy
Syt6
AI
DMS
BLA
DMS+AI
DMS+BLA
DMS+MD
DMS+LH
MD
LH
MD+LH
L2/3-Calb1
L2/3-Rorb
L5-Bcl6
L5-Htr2c
L5-S100b
L5-Npy
L6-Syt6
b
DMS vs DMS+LH projection
Bcl11b
Pou3f1
Igfbp4
Lypd1
$-Log_{10}$ P
Log_2 FC(DMS/DMS+LH)
Total = 382 variables
c
Bcl6
Pou3f1
Reln
Adamts2
Fam126a
Igfbp4
Lypd1
DMS+LH
LH
DMS
d
DMS(EGFPnls)/LH(tdTomato)
DMS(EGFPnls)/BLA(tdTomato)
BLA(GFP)/LH(tdTomato)
L6
L5
L2/3
e
f
4.41%
22.3%
73.28%
14.11%
6.58%
79.31%
Ratio
tdT+ only
GFP+ ony
tdT+GFP+
L2/3
L5
L6
g
GFP
tdTomato
Pou3f1
Merged
DMS(EGFPnls)/LH(tdTomato)/Pou3f1
Layer 5
h

Fig. 4 | MERGE-seq can infer bifurcated projection patterns and projection-specific genes of vmPFC. a, Alluvial plot showing top 10 frequent projection patterns distribution in distinct neuronal subtypes. neuron subtype color code is same as Fig. 2a and projection pattern color code is same as Fig. 3d. **b**, Volcano plot showing genes differentially expressed in DMS/LH group (red spots) calculated between DMS and DMS/LH projection pattern. **c**, Tracks plot showing raw counts of the selected DEGs (red labeled genes in Fig. 3g) in DMS dedicated, LH dedicated and DMS/LH bifurcated projection pattern. Gene labeled in red (Pou3f1) was used to perform RNA FISH in (**g-h**). **d**, Immunostaining of dual-color traced retrograde labeled neurons of selected targets (DMS/LH labeled in red in Fig. 4a, DMS/BLA labeled in red in Fig. 4a, and BLA/LH). Dotted line depicts layer2/3, layer5 and layer6 of vmPFC. Scale bars, 500 um. **e,** A zoom view of dotted box in **(d)**. Scale bars, 100um. **f**, Histogram shows quantitative data for single (red, green) and double (yellow) labeled neurons as mean percentages of total retrograde AAV labeled neurons (n=3). The data are presented as the mean ± SD. Colocalization ratio (yellow bar) from ipsilateral side for DMS/LH, 17.8%±0.11%; for DMS/BLA, 6.59%±1.55%; for BLA/LH, 1.66%±0.92%. Pie chart showing layer distribution of double (yellow) labeled neurons. **g**, Examining Pou3f1 and DMS/LH projection pattern using RNA FISH and immunostaining of dual-color traced retrograde labeled neurons. Virus injection scheme was same as **(d)**. Scale bars, 200um. **h**, A zoom view of dotted box in **(g)**. Arrow heads indicate $Pou3f1^{+}GFP^{+}tdTomato^{+}$ neurons, white arrows indicate $Pou3f1^{+}GFP^{-}tdTomato^{+}$ neurons and blue arrows indicate $Pou3f1^{-}GFP^{+}tdTomato^{-}$ neurons. Scale bars, 100um.

Fig. 5

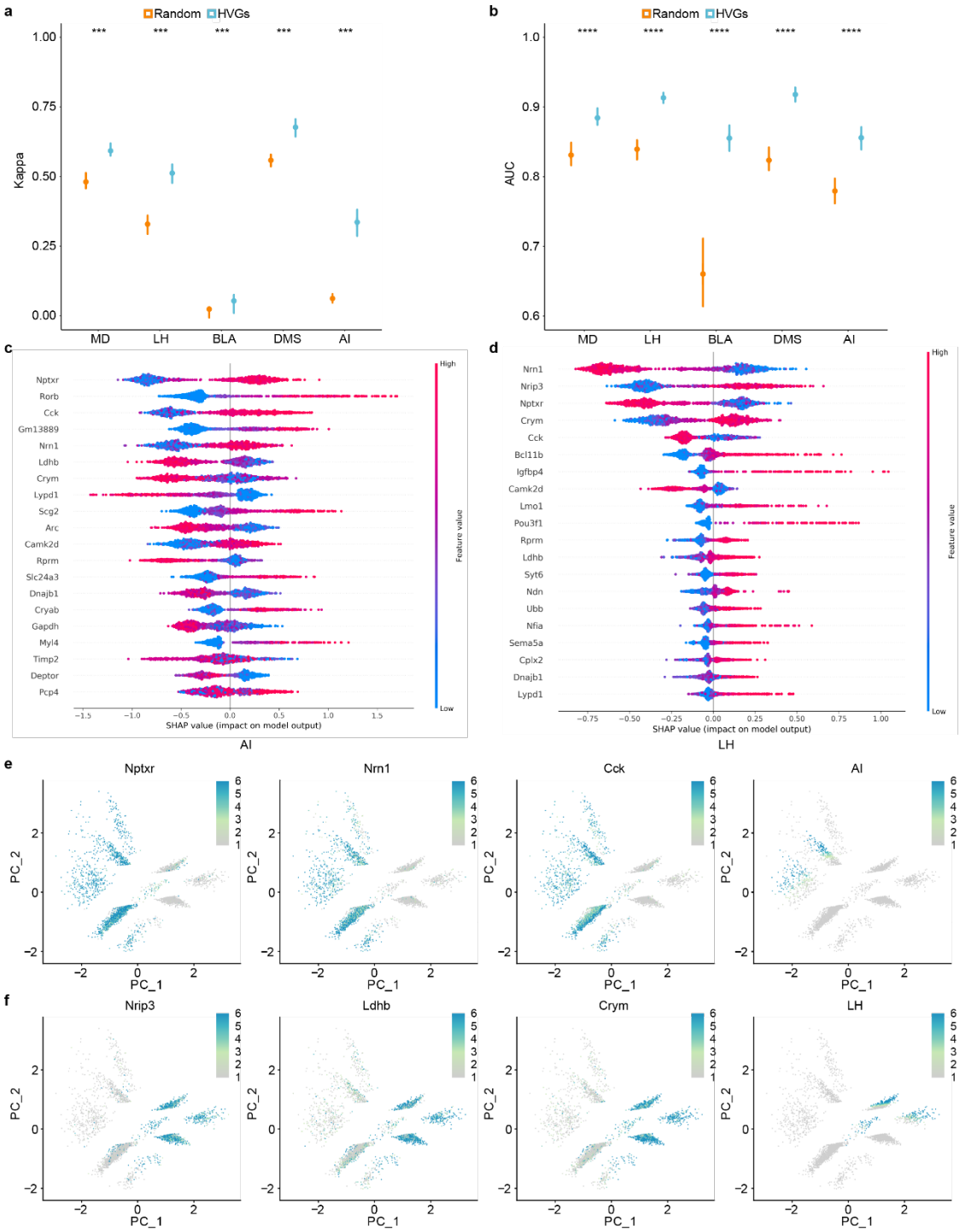
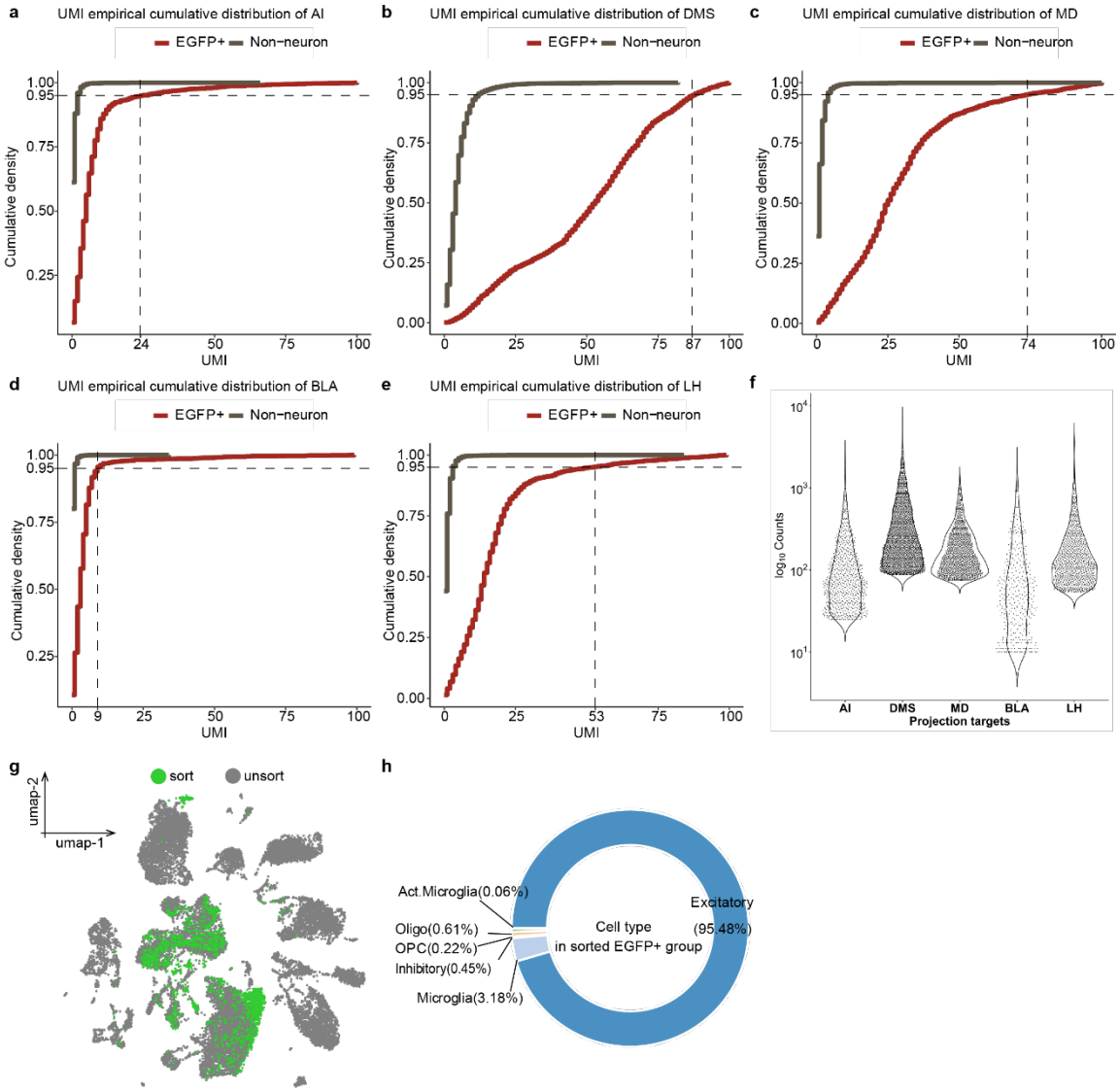

a
Random
HVGs
Kappa
MD
LH
BLA
DMS
AI
b
AUC
c
Nptxr
Rorb
Cck
Gm13889
Nrn1
Ldhb
Crym
Lypd1
Scg2
Arc
Camk2d
Rprm
Slc24a3
Dnajb1
Cryab
Gapdh
Myl4
Timp2
Deptor
Pcp4
SHAP value (impact on model output)
Feature value
High
Low
d
Nrn1
Nrip3
Nptxr
Crym
Cck
Bcl11b
Igfbp4
Camk2d
Lmo1
Pou3f1
Rprm
Ldhb
Syt6
Ndn
Ubb
Nfia
Sema5a
Cplx2
Dnajb1
Lypd1
e
Nptxr
Nrn1
Cck
AI
PC_1
PC_2
f
Nrip3
Ldhb
Crym
LH

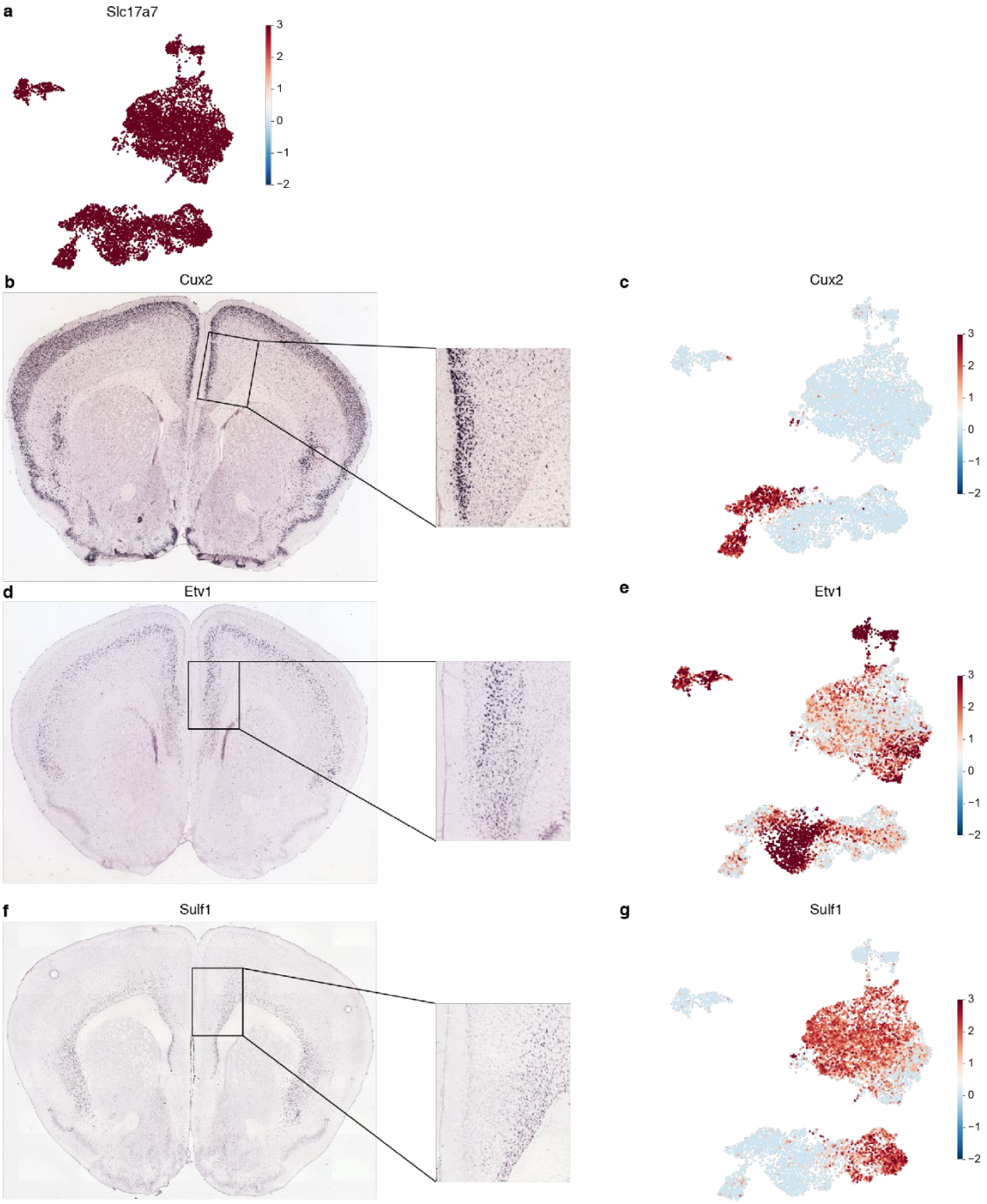
Fig. 5 | Machine-learning modeling partially predicts projection pattern based on gene expression. a-b, 10 trials of prediction results using random split test data and comparison is made between models built by top 150 HVGs and random 150 genes for each projection target. **(a)**, Kappa value; **(b)**, AUC score. The displayed P value is computed using a two-sided Wilcoxon test. The data are presented as the mean ± SD. Degree of significance was represented as follows: * p-value ≤0.05; ** p-value ≤0.01; *** p-value ≤0.001; **** p-value ≤0.0001; n.s., not significant. Error bars indicate standard deviation. **(c-d),** SHAP summary plots of AI and LH showing important features (genes) with feature effects. For each model, unbarcoded cells were set to class 0 and barcoded cells were set to class 1. **e-h**, Normalized expression of top important genes with positive feature effects in XGBoost modeling of AI (**e-g**) and normalized expression of barcode AI (**h**) on PC1 and PC2 embeddings. **i-l**, Normalized expression of top important genes with positive feature effects in XGBoost modeling of LH (**i-k**) and normalized expression of barcode LH (**l**) on PC1 and PC2 embeddings.

Supplementary Figure 1



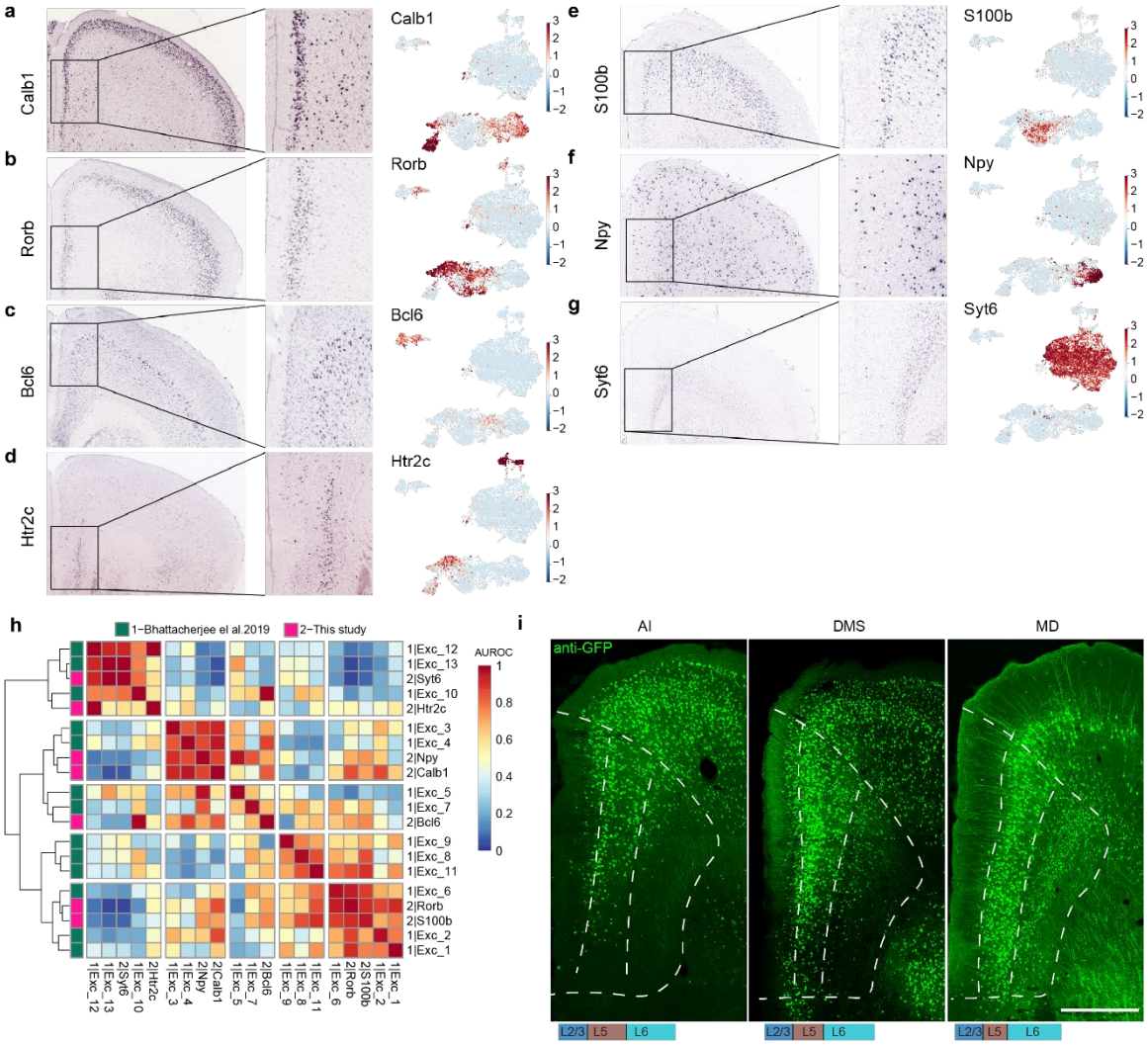
Supplementary Figure 1. Determination of valid barcoded cells. (**a-e**) UMI counts of empirical cumulative distribution for each projection index barcode. A stringent cumulative density threshold was set to 0.95. Grey line represents UMI counts of all non-neuron projection index barcode, red line represents UMI counts of FAC-sorted EGFP+ neuron. UMI counts threshold was chosen at the UMI value where cumulative density of each index barcode UMI counts equals to 0.95 for EGFP+ neuron group (crossing point of horizontal and vertical dashed line). (**f**) Violin plots of Log_{10} normalized projection index barcode counts. Note that, after choosing the UMI counts threshold, UMI counts below threshold were dropped to zero. (**g**) Umap embedding of FACS condition in all cells. Green represents sorted EGFP+ cells and grey represents unsorted cells. (**h**) Donut plot depicting percentage of each cell type in sorted EGFP+ group. 95.48% of all sorted cells are excitatory neurons.

Supplementary Figure 2



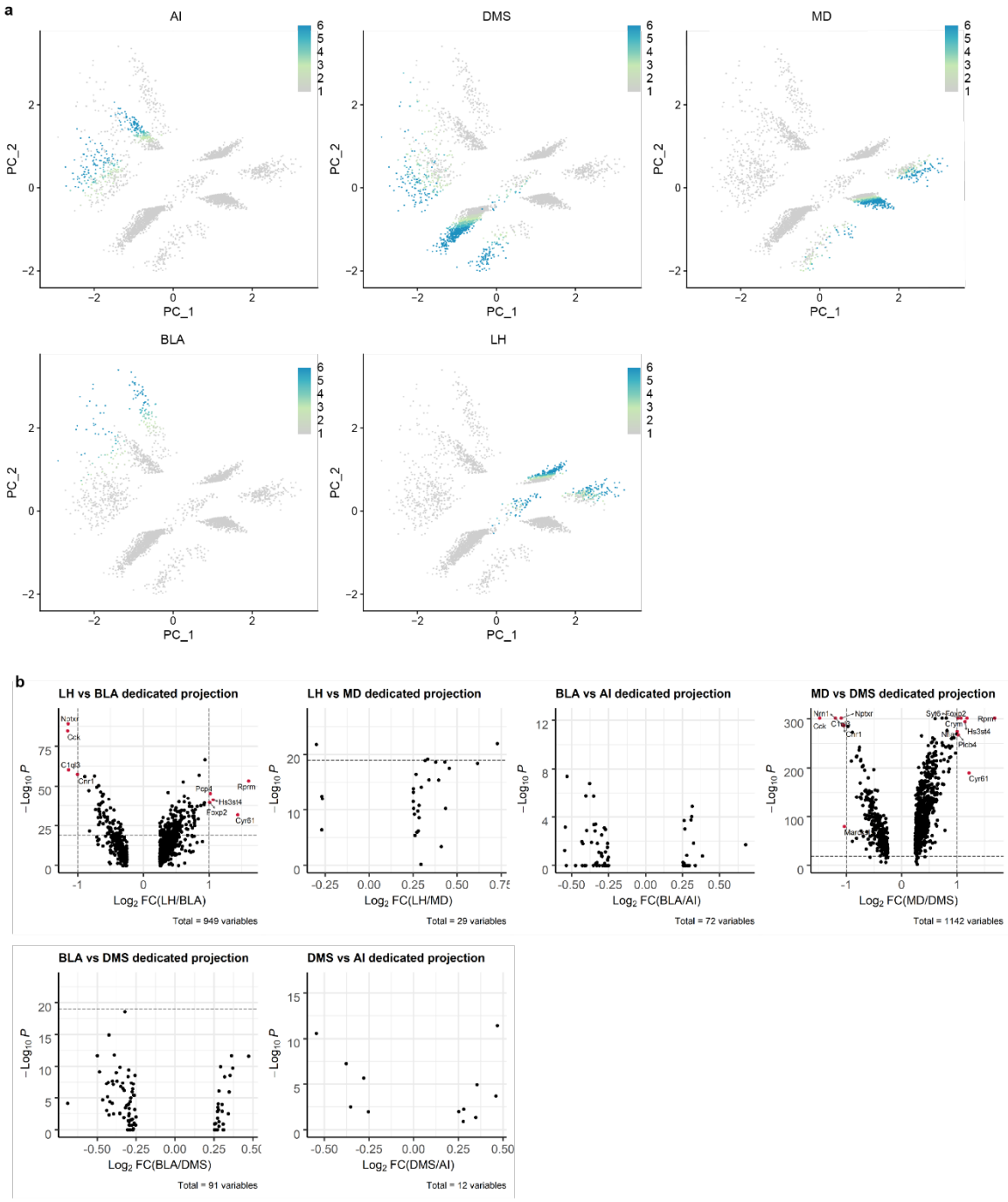
Supplementary Figure 2. Layer annotation using mouse brain atlas. (**a**) Normalized Slc17a7 (vGlut1) expression on all extracted excitatory neurons. (**b, d, f**) In situ hybridization of typical layer specific markers among vmPFC region from Adult Mouse Brain Atlas. Cux2 is layer2/3 specific, Etv1 is layer5 specific and Sulf1 is layer6 specific. **(c, e, g)** Normalized expression of Cux2, Etv1 and Sulf1 at umap embedding, also see **Fig. 2a**.

Supplementary Figure 3



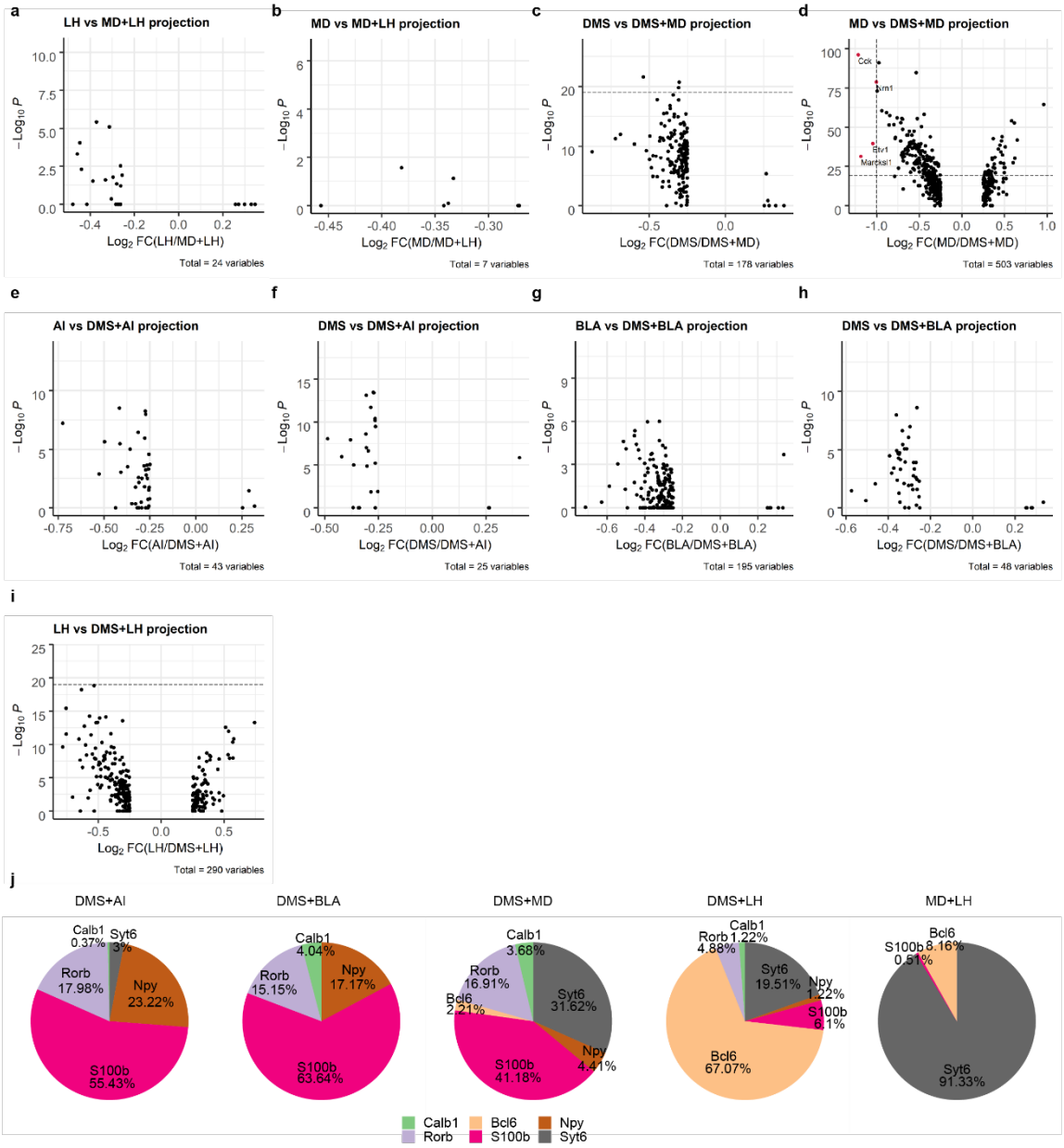
Supplementary Figure 3. Excitatory neuron subtype annotation using mouse brain atlas and scRNA-seq transcriptomic comparison with public dataset. (**a-g**). In situ hybridization of typical neuronal subtype specific markers among vmPFC region from Adult Mouse Brain Atlas. Calb1 and Rorb are layer2/3 specific, Htr2c and S100b are layer5 specific; Bcl6 is around transition of layer2/3 and layer5, based on previous research (Sorensen et al., 2015), we define Blc6+ cluster as layer5 specific; Syt6 and Npy are layer6 specific, though Npy is distributed sporadically. Corresponding normalized gene expression embedded in umap is plotted at right panel. (**h**). Heatmap showing correlation between annotated excitatory neuron subtypes recovered in this project and dataset from Bhattacherjee et al. (**i**). Typical images showing layer distribution of retrograde AAV labeled neurons projecting AI, DMS and MD from contralateral side. Scale bars, 500 um.

Supplementary Figure 4



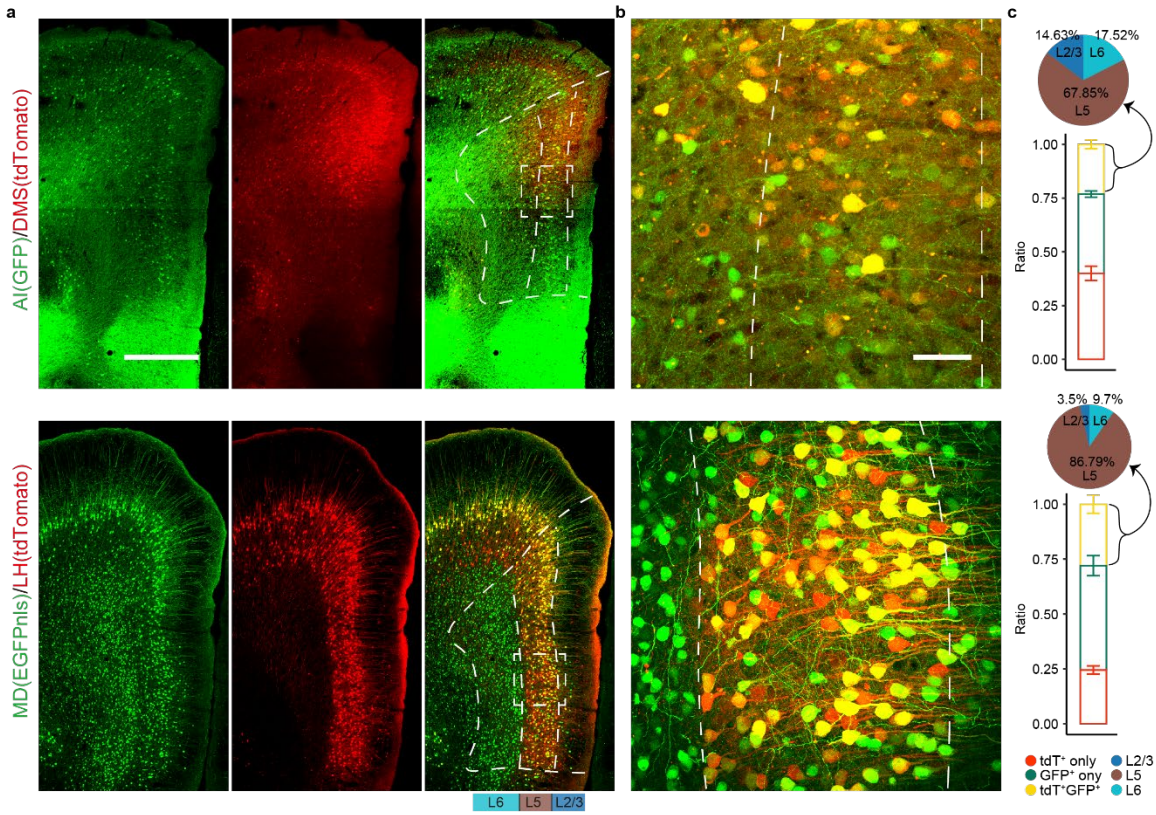
Supplementary Figure 4. Digital projection strength and DEGs calculated between dedicated projection patterns. (**a**) Normalized projection index barcode expression on PC1 and PC2 embeddings. (**b**) Volcano plots of selected dedicated projection patterns. See also **Fig. 3**. Assigned DEGs (red dots) were determined using threshold: Log_2 fold change=1, p value cutoff= 10^{-20}.

Supplementary Figure 5



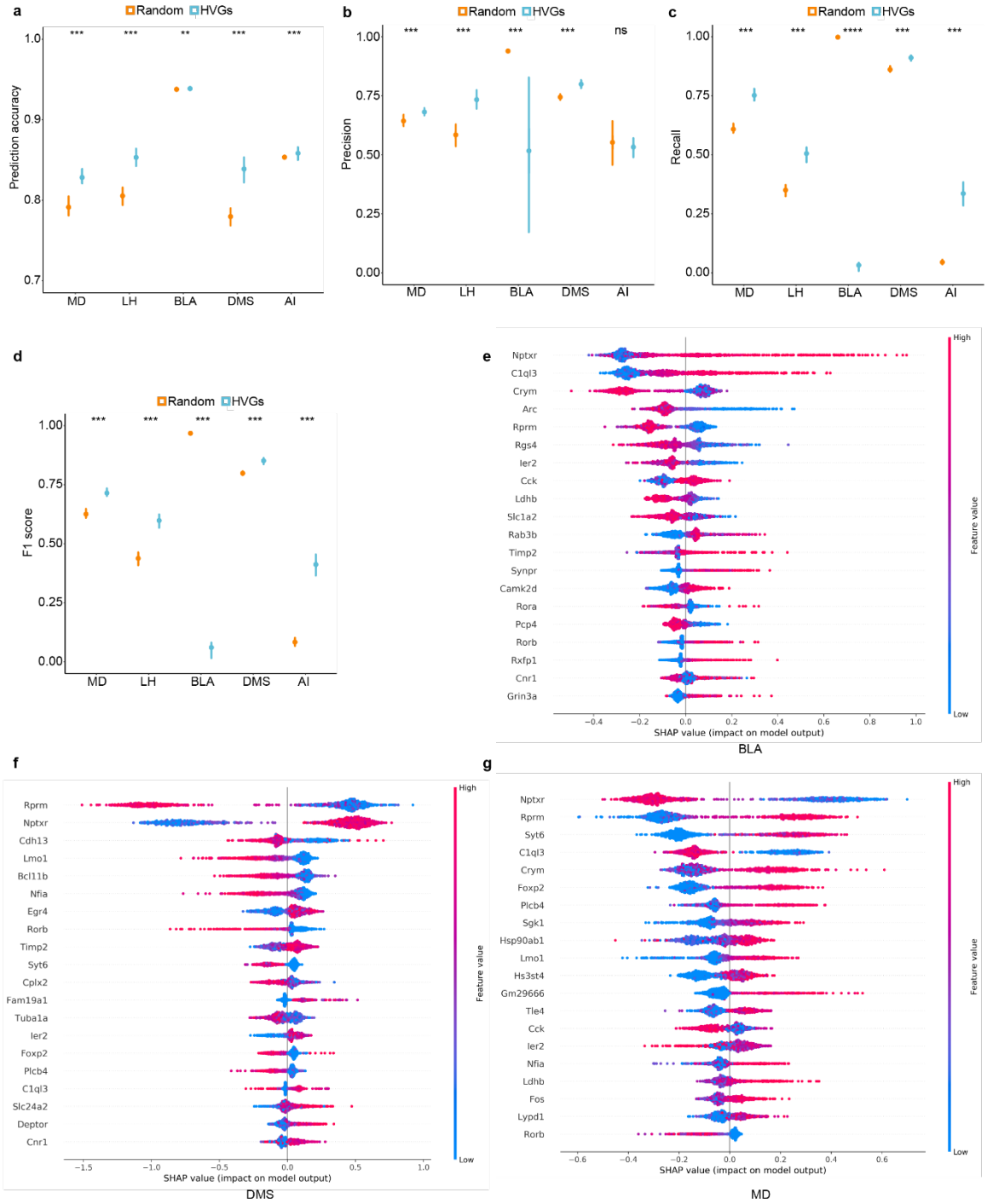
Supplementary Figure 5. DEGs calculated between dedicated and bifurcated projection patterns. (**a-i**) Volcano plots of DESgs calculated between A and A/B projection pattern. See also **Fig. 4**. Assigned DEGs (red dots) were determined using threshold: Log_2 fold change=1, p value cutoff= 10^{-20}. (**j**) Pie charts describing selected top 5 frequent bifurcated projection patterns (DMS+AI, DMS+BLA, DMS+MD, DMS+LH and MD+LH) partitioned in neuronal subtype. For DMS+BLA, neurons projecting DMS+BLA are mostly made up of neuron subtypes S100b (63.64%). For DMS+LH, neurons projecting DMS+LH are mostly made up of neuron subtypes Bcl6 (67.07%)

Supplementary Figure 6



Supplementary Figure 6. Immunostaining of dual-color traced retrograde labeled neurons and statistical analysis. (**a**) Immunostaining of dual-color traced retrograde labeled neurons of AI(GFP)/DMS(tdTomato) and MD(GFP)/LH(tdTomato). Dotted line depicts layer 2/3, layer 5 and layer 6. Scale bars, 500 um. (**b**) A zoom view of dotted box in (**a**). Scale bars, 100 um. (**c**) Histogram shows quantitative data for single (red, green) and double (yellow) labeled neurons as mean percentages of total retrograde AAV labeled neurons (AI(GFP)/DMS(tdTomato): n=3, MD(GFP)/LH(tdTomato): n=4. The data are presented as the mean ± SD. Colocalization ratio (yellow bar) of each group from ipsilateral side: AI/DMS=23.1%±2.03% and MD/LH=27.9%±4.10%. Pie chart showing layer distribution of double (yellow) labeled neurons.

Supplementary Figure 7



Supplementary Figure 7. SHAP summary plots of XGBoost models. **(a-c)** SHAP summary plots of BLA, MD and DMS showing important features (genes) with feature effects. For each model, unbarcoded cells were set to class 0 and barcoded cells were set to class 1. **(d-g)** 10 trials of prediction results using random split test data and comparison is made between models built by top 150 HVGs and random 150 genes for each projection target. **(d),** prediction accuracy; **(e),** precision. **(f),** recall rate. **(g),** F1 score. The displayed P value is computed using a two-sided Wilcoxon test. The data are presented as the mean ± SD. Degree of significance was represented as follows: * p-value ≤0.05; ** p-value ≤0.01; *** p-value ≤0.001; **** p-value ≤0.0001; n.s., not significant. Error bars indicate standard deviation.