Supplementary Materials for

Single Cell Methylomes Identify Neuronal Subtypes and Regulatory Elements in Mammalian Cortex

Chongyuan Luo, Christopher L. Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R. Nery, Justin P. Sandoval, Brian Bui, Terrence J. Sejnowski, Timothy T. Harkins, Eran A. Mukamel, M. Margarita Behrens, Joseph R. Ecker

correspondence to: ecker@salk.edu, mbehrens@salk.edu, emukamel@ucsd.edu

This PDF file includes:

Materials and Methods
Supplementary texts
Figs. S1 to S17
Tables S1 to S9 captions

Materials and Methods

Animal samples

For the production of single neuron methylomes from layer dissected mouse frontal cortex tissue, eight week old C57BL/6J male mice were purchased from Jackson Laboratories, Bar Harbor ME, and allowed a week of acclimation in our animal facility with 12 h light/dark cycles and food ad libitum before sacrificing and dissecting.

Nuclei were also isolated from frontal cortex of the CLSun1-G35-Cre line with no layer dissection. This line was produced by crossing the transgenic line B6;129-Gt(ROSA)26Sortm5(CAG-Sun1/sfGFP)Nat/J (described in (*5*), but backcrossed into a C57BL/6J background for 9 generations), with the G35-Cre line (*26*).

Nuclei of SST+ inhibitory neuron population was isolated from frontal cortex of CLSun1-SST-Cre line. This line was generated by crossing B6;129-Gt(ROSA)26Sortm5(CAG-Sun1/sfGFP)Nat/J backcrossed into C57BL/6J with SST-Cre line (Jackson Labs).

All protocols were approved by the Salk Institute's Institutional Animal Care and Use Committee (IACUC).

Human samples

The human brain specimen was obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD. The frozen middle frontal gyrus tissue belonged to a 25-year-old Caucasian male (UMB#4540) with a PMI of 23 h.

Mouse tissue dissections

To produce the frontal cortex tissue, mouse brains were sectioned coronally at Bregma 2.5 and 0.5 with a razor blade in dissection media [20 mM Sucrose, 28 mM D-Glucose (Dextrose), 0.42 mM $NaHCO_3$, in HBSS]. For cortical layer dissection, the tissue block (devoid of non-cortical tissue) was then dissected under a microscope (SZX16, Olympus). The cortical region was divided parallel to the meninges into three sections of approximately equal width such that “superficial layers” contained layers 1-3 with part of layer 4, and “deep layers” contained mainly layers 6 and part of 5.

Nuclear isolation

Nuclear isolation from mouse and human cortical tissues was performed as described in (*7*) with the following modifications: Proteinase inhibitor (11836153001, Roche) and RNAse inhibitor (30 U/ml, PRN2611 from Promega) were added to the lysis buffer and sucrose gradients. After centrifugation, nuclei were resuspended in 0.5% BSA (AM2616, Ambion) and PBS (Ca^{2+} and Mg^{2+} free, 14190-144 from Life Technologies) with protein and RNAse inhibitors.

Flow cytometry based nuclei sorting

Isolated nuclei from mouse and human tissues were labeled by incubation with 1:1000 dilution of AlexaFluor488 conjugated anti-NeuN antibody (MAB377X, Millipore) at 4°C for 1 hour. Nuclei isolated from CLSun1-G35-Cre line were incubated with AlexaFluor647 conjugated anti-NeuN antibody (anti-NeuN antibody MAB377 labeled using Apex Alexa Fluor 647. A10475, Life Technologies) and AlexaFluor488 conjugated anti-GFP antibody (A21311, Life Technologies). Fluorescence-activated nuclei sorting (FANS) of single nuclei was performed using a BD Influx sorter with an 85 µm nozzle at 22.5 PSI sheath pressure. Single nuclei were sorted into each well of a 384-well plate preloaded with 2 µl of Proteinase K digestion buffer (1 µl M-Digestion Buffer, 0.1 µl 20 µg/µl Proteinase K and 0.9 µl H_2O). The alignment of the receiving 384-well plate was performed by sorting sheath flow into wells of an empty plate and making adjustments based on the liquid drop position. Single cell (1 drop single) mode was selected to ensure the stringency of sorting.

Preparation of single nucleus methylome library

Steps of library preparation prior to SPRI purification were performed in a horizontal laminar flow hood to minimize environmental DNA contamination. Bisulfite conversion of single nuclei was carried out using Zymo EZ-96 DNA Methylation-Direct^{TM} Kit (Deep Well Format, cat. #D5023) following the product manual with reduced reaction volume. 384-well plates (ThermoFisher Armadillo PCR Plate cat. # AB2384) containing FACS isolated single nuclei were heated at 50°C for 20 min. 25 µl CT Conversion Reagent was added to each well, followed by pipetting up and down to mix. Plates were treated with the following program using a thermocycler: 98°C for 8 min, 64°C for 3.5 hours and 4°C forever.

Each well of Zymo-Spin™ I-96 Binding Plates was preloaded with 150 µl M-binding buffer. Bisulfite conversion reactions were transferred from 384-well plates to I-96 Binding Plates followed by pipetting up and down to mix. I-96 Binding Plates were centrifuged at 5,000g for 5 min. Wells were washed with 400 µl of M-Wash Buffer, followed by centrifugation at 5,000g for 5 min. 200 µl of M-Desulphonation Buffer were added to each well and incubated for 15 min at room temperature before removed by centrifugation at 5,000g for 5 min. Each well was then washed with 400 µl of M-Wash Buffer twice. 12 µl of M-Elution Buffer were added to each well and incubated for 5 min at room temperature. I-96 Binding Plate was placed above a 96-well PCR plate (Applied Biosystems MicroAmp® EnduraPlate^{TM} cat. # 4483348) and was centrifuged at 5,000g for 3 min. 9 µl of eluted DNA were commonly collected in each well of the PCR plate.

Each of the four indexed random primers (P5L-AD002-N9, P5L-AD006-N9, P5L-AD008-N9 and P5L-AD010-N9) was used for indexing a 96-well plate containing bisulfite converted single nuclei. The four plates would be combined during a later SPRI step. 1 µl of 5 µM indexed random primer was added to each well of 96-well plate, followed by mixing with vortexing. All DNA oligos were purchased from Integrated DNA Technologies (IDT).

P5L-AD002-N9
/5SpC3/TTCCCTACACGACGCTCTTCCGATCTCGATGT(N1:25252525)(N1)(N1) (N1)(N1)(N1) (N1)(N1)(N1)
P5L-AD006-N9
/5SpC3/TTCCCTACACGACGCTCTTCCGATCTGCCAAT(N1:25252525)(N1)(N1) (N1)(N1)(N1) (N1)(N1)(N1)
P5L-AD008-N9
/5SpC3/TTCCCTACACGACGCTCTTCCGATCTACTTGA(N1:25252525)(N1)(N1) (N1)(N1)(N1) (N1)(N1)(N1)
P5L-AD010-N9
/5SpC3/TTCCCTACACGACGCTCTTCCGATCTTAGCTT(N1:25252525)(N1)(N1) (N1)(N1)(N1) (N1)(N1)(N1)

96-well plates were heated at 95°C using a thermocycler for 3 min to denature sample and were immediately chilled on ice for 2 min. 10 µl enzyme mix containing 2 µl of Blue Buffer (Enzymatics cat. # B0110), 1 µl of 10mM dNTP (NEB cat. # N0447L), 1 µl of Klenow exo- (50U/µl, Enzymatics cat. # P7010-HC-L) and 6 µl H_2O, was added to well. After mixing with vortexing, plate was treated with the following program using a thermocycler: 4°C for 5 min, ramp up to 25°C at 0.1°C/sec, 25°C for 5 min, ramp up to 37°C at 0.1°C/sec, 37°C for 60 min, 4°C. 2 µl of Exonuclease 1 (20U/µl, Enzymatics cat. # X8010L) were added to each well, followed by mixing with vortexing. Plate was incubated at 37°C for 30 min and then 4°C forever using a thermocycler.

17.6 µl of home-made SPRI beads were added to each well. Sample/bead mixture from four plates, indexed using distinct indexed random primers, was combined and followed by pipetting up and down to mix. Sample/bead mixture was incubated at room temperature for 5 min before being placed on a 96-well magnetic separator (DynaMag™-96 Side Magnet, ThermoFisher Cat. # 12331D. DynaMag™-96 Side Skirted Magnet, ThermoFisher Cat. # 12027). Supernatant was removed from each well, followed by three rounds of washing with 180 µl of 80% ethanol. After air drying beads at room temperature, 10 µl M-Elution buffer were added to each well to fully resuspend the beads. Eluted samples were transferred to a new 96-well PCR plate.

PCR plate was heated at 95°C for 3 min using a thermocycler to denature sample and was immediately chilled on ice for more than 2 min. 10.5 ul Adaptase master mix (2 ul Buffer G1, 2 ul Regent G2, 1.25 ul Reagent G3, 0.5 ul Enzyme G4, 0.5 ul Enzyme G5 and 4.25 ul M-Elution buffer; Accel-NGS Adaptase Module for Single Cell Methyl-Seq Library Preparation, Swift Biosciences, cat. # 33096) was added into each well, followed by mixing with vortexing. Plates were incubated at 37°C at 30 min and then 4°C using a thermocycler. 30 µl PCR mix (25 µl KAPA HiFi HotStart ReadyMix, KAPA BIOSYSTEMS, cat. # KK2602, 1 µl 30 µM P5 indexing primer and 5 µl 10 µM P7 indexing primer) were added into each well, followed by mixing with vortexing.

P5 Indexing primers:

P5L_D501
AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGACGCTCT
P5L_D502
AATGATACGGCGACCACCGAGATCTACACATAGAGGCACACTCTTTCCCTACACGACGCTCT
P5L_D503
AATGATACGGCGACCACCGAGATCTACACCCTATCCTACACTCTTTCCCTACACGACGCTCT
P5L_D504
AATGATACGGCGACCACCGAGATCTACACGGCTCTGAACACTCTTTCCCTACACGACGCTCT
P5L_D505
AATGATACGGCGACCACCGAGATCTACACAGGCGAAGACACTCTTTCCCTACACGACGCTCT
P5L_D506
AATGATACGGCGACCACCGAGATCTACACTAATCTTAACACTCTTTCCCTACACGACGCTCT
P5L_D507
AATGATACGGCGACCACCGAGATCTACACCAGGACGTACACTCTTTCCCTACACGACGCTCT
P5L_D508

AATGATACGGCGACCACCGAGATCTACACGTACTGACACACTCTTTCCCTACACGACGCTC
T

P7 indexing primers:

P7L_D701
CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D702
CAAGCAGAAGACGGCATACGAGATTCTCCGGAGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D703
CAAGCAGAAGACGGCATACGAGATAATGAGCGGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D704
CAAGCAGAAGACGGCATACGAGATGGAATCTCGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D705
CAAGCAGAAGACGGCATACGAGATTTCTGAATGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D706
CAAGCAGAAGACGGCATACGAGATACGAATTCGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D707
CAAGCAGAAGACGGCATACGAGATAGCTTCAGGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D708
CAAGCAGAAGACGGCATACGAGATGCGCATTAGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D709
CAAGCAGAAGACGGCATACGAGATCATAGCCGGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D710
CAAGCAGAAGACGGCATACGAGATTTCGCGGAGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D711
CAAGCAGAAGACGGCATACGAGATGCGCGAGAGTGACTGGAGTTCAGACGTGTGCTCTT
P7L_D712
CAAGCAGAAGACGGCATACGAGATCTATCGCTGTGACTGGAGTTCAGACGTGTGCTCTT

PCR plate was treated with the following program using a thermocycler: 95°C for 2 min, 98°C for 30 sec, 17 cycles of (98°C for 15 sec, 64°C for 30 sec, 72°C for 2min), 72°C for 5 min and then 4°C. PCR products were cleaned up using 0.8x SPRI beads and were combined into one tube for each 96-well plate. Pooled PCR product was resolved on 2% agarose gel, smear between 400 bp and 2 Kb were excised and purified using QIAquick Gel Extraction Kit (Qiagen cat. # 28706). Library concentration was determined using Qubit® dsDNA HS (High Sensitivity) Assay Kit (Invitrogen cat. # Q32851).

Sequencing of single nucleus methylome library

Pooled library concentration was adjusted to 700 - 800 pM for cluster generation and was sequenced on Illumina HiSeq 4000 instrument using RTA 2.7.7.

Single cell methylome mapping and data analysis

Sequencing reads were first trimmed to remove sequencing adaptors using Cutadapt 1.11 (*27*) with the following parameters in paired-end mode: -f fastq -q 20 -m 62 -a AGATCGGAAGAGCACACGTCTGAAC -A AGATCGGAAGAGCGTCGTGTAGGGA. For singleplex samples, -m parameter was set to 40. For multiplexed samples, 16 bp were further trimmed from both 5'- and 3'- ends of R1 and R2 reads to remove random primer index sequence and C/T tail introduced by Adaptase, with the following parameters: -f fastq -u 16 -u -16 -m 30. Trimmed reads for mouse and human single nuclei were mapped to mm10 and hg19 reference genomes, respectively. R1 and R2 reads were mapped separately as single-end reads using Bismark v0.15.0 with parameter --bowtie2. --pbat option was activated for mapping R1 reads (*28*). Resulting bam files were sorted using SAMtools 1.3 sort (*29*), followed by removal of duplicate reads using Picard 1.141 MarkDuplicates with the option REMOVE_DUPLICATES=true (https://broadinstitute.github.io/picard/). Non-clonal reads were further filtered for minimal mapping quality (MAPQ ≥ 30) using samtools view with option -q30. To prevent regional mCH level estimation from being skewed by rare reads that failed to be bisulfite converted, reads with read-level mCH level greater than 0.7 were excluded.

The calling of unmethylated and methylated base calls was performed by call_methylated_sites of Methylpy (https://github.com/yupenghe/methylpy/) (*5*, *7*, *23*).

MethylC-seq of mouse SST+ inhibitory neurons

MethylC-seq library was constructed following the protocol described in detail in (*30*).

Genomic sequencing of the human sample

Genomic DNA was extracted from the human specimen using DNeasy Blood & Tissue Kit (Qiagen cat. # 69504). Genomic sequencing library was constructed using the same procedure as MethylC-seq library except bisulfite conversion was not performed.

Calling sequence variants for the human sample

Adaptor sequence was trimmed from sequencing reads using Cutadapt 1.11 with the following options: -f fastq -q 20 -m 50 -a AGATCGGAAGAGCACACGTCTGAAC -A AGATCGGAAGAGCGTCGTGTAGGGA. Trimmed reads were mapped to human hg19 reference genome using Bowtie2 2.2.5 with option -X 2000. Mapped reads were filtered for minimal mapping quality (MAPQ ≥ 20) using samtools view with option -q20. For calling sequence variants, a filtered bam file was processed using samtools mpileup with option -ug with the outputs piped into bcftools called with option -vm0 v (*31*).

Data cleaning

Data were cleaned by excluding low-quality cells using the following set of conservative criteria, ultimately yielding 3376 cells in mouse and 2784 cells in human for analysis. First, non-conversion rate was required to be low (≤1% in mouse and ≤2% in human). We set a minimum on the number of non-clonal mapped reads to eliminate contaminated samples (400K in mouse; 500K in human). We also set an upper limit on coverage to protect against wells with multiple cells (<= 15% of cytosines).

In order to minimize contamination in the human data from exogenous human DNA fragments, we identified potentially contaminated samples using a genomic sequence variant (SNP) matching process. Single nucleotide variants identified from genomic sequencing (*g*) of the human sample were compared to variants observed in each single human nucleus methylome (*m*), with hg19 serving as the reference genome for mapping both data types. SNP compatibility between *g* and *m* was scored for all homozygous variant sites identified in *g* that were covered by methylome reads in *m*. A compatible site between *g* and *m* required identical genotype between *g* and *m* at all sites where the variant sequence was A or T in *g*. For a site with variant sequence C in *g*, sequence = C or T in *m* was considered compatible. For a site with variant sequence G in *g*, sequence = G or A in *m* was considered compatible. For each single human nucleus methylome, compatible SNP rate was defined as the fraction of all scored sites showing compatible genotype between *g* and *m*, and we only retained cells with >0.99 compatible SNP rate.

Clustering analysis

CH methylation data were grouped into non-overlapping 100 kb bins across the whole genome for each cell. Due to the sparsity of the snmC-seq data, few bins had sufficient coverage (>100 base calls) across all cells to be retained in the analysis. We therefore imputed data at bins with coverage in 99.5% or more of cells, replacing missing values with the average methylation across all cells for that bin. This allowed us to include 76.2% of bins in the mouse genome and 63.4% of bins in the human genome in our analysis.

To cluster cells, we adapted an iterative, hierarchical and unsupervised clustering method called BackSPIN that had been previously applied to single cell transcriptome data (*2*). At each iteration, the top 2,000 bins with the greatest variance across cells were selected. The SPIN algorithm was then used to arrange cells in a linear order, with similar cells located near each other (*32*). Next, cells were split into two new clusters at the optimal cut point, where the average correlation within the two new clusters was highest. To retain the split, at least one of the two new subclusters must have >15% increase in the average correlation value over the average of all cells in the original cluster. This procedure was applied recursively to each new cluster, and terminated when no clusters met the splitting criterion. To avoid producing clustering with too few cells for us to confidently analyze, we prevented further splitting of clusters with 50 or fewer cells.

Because BackSPIN can produce different results depending on the initial order of cells, we ran the algorithm with 160 random initializations of the cell order. We selected the clustering that had the highest Dunn Index. The result had 23 clusters for mouse and 40 clusters for human. Initial inspection of the cluster results using tSNE revealed that one of the human clusters, which comprised 44 cells, was highly dissimilar from the other clusters. Cells in this cluster had little detectable mCH (global median: 0.0104), significantly lower than the cluster with the next lowest mCH level (0.0201) and the median across all cells (0.0438). We surmised that this cluster may correspond to non-neuronal cells, and we therefore excluded these cells from subsequent analysis.

To conservatively define neuronal cell types based on robust and biologically interpretable differences in DNA methylation, we next merged clusters with highly similar mCH patterns. Our heuristic choices of criteria for merging clusters does impact the final number and configuration of clusters. Rather than try to accurately determine how many cell types exist in each species, our emphasis was on using consistent clustering parameters and methods in both human and mouse to allow a rigorous cross-species comparison.

To do this we defined a set of mCH marker genes. For this analysis we profiled the mCH level across all gene bodies for each cell, requiring coverage of at least 100 CH bases. We retained genes that were covered in ≥20% of cells in each of the clusters, and in ≥50% of cells in at least one cluster. We further required coverage in at least 10 cells for each cluster. We then combined reads from all cells in each cluster to estimate the mCH level for each cluster at each gene; these mCH levels were then normalized by the average over all cells at each gene. Marker genes for each pair of clusters were defined as those which were strongly hypomethylated (mCH in the bottom 2nd percentile) in one cluster and hypermethylated (mCH above the 80th percentile) for the other cluster. The top 10-20 marker genes with largest methylation difference were identified for each pair of clusters. We then tested the statistical significance of the difference in normalized methylation between the two clusters (2-sample t-test, one-sided, $p<0.05$). If any pair of clusters was separated by fewer than 7 significant marker genes, we merged the pair of clusters with the fewest markers and repeated the procedure (define marker gene, test significance). This process was continued until all cluster pairs had at least 7 marker genes with significantly different mCH.

For visualization purposes, we performed dimensionality reduction using t-Stochastic Neighbor Embedding (tSNE) (*14*), reducing all cells to a point in 2D space. TSNE requires a perplexity parameter that is analogous to how many nearest neighbors to consider in manifold learning algorithms. We examined results using a range of perplexity values (10-1000) and found largely consistent patterns for all perplexity values >50 (Fig. S3G), and all tSNE visualizations shown in this study used perplexity = 150. Importantly, our tSNE results were only for visualization purposes and did not affect the clustering of neurons, although it strongly agreed with the clusters we identified using BackSPIN adapted for DNA methylation data. To illustrate how mCH levels of marker genes vary across individual cells and clusters, we computed gene body methylation as the average mCH level of annotated genic region (from TSS to TES) (Fig. 2E,F) and normalized across cells by dividing by the mean of all cells ((Fig. S6-7).

Validation of clustering

We examined the robustness of our neuronal clusters with respect to several experimental and analytic parameters (Fig. S3). First, we downsampled the number of reads per cell to 10%, 20% and 40% of the full dataset using `samtools view -s,` followed by tSNE (Fig. S3A). To examine whether CG methylation could be used to determine cell types consistent with those estimated using mCH, we summarized CG methylation into 100kb bins followed by tSNE (Fig. S3F). Because CG sites are more sparse than non-CG sites, we lowered the coverage cutoff to >20 base calls for this analysis.

Because backSPIN can produce different clustering outputs given different input order of cells, we compared 200 backSPIN results with independently randomized initializations against our identified neuronal clusters. We did not perform marker-gene based merging on shuffled data as performed to obtain our original clustering, and consequently, we would expect some level of difference between our clusters and the shuffled runs. Using the adjusted Rand index (*33*) and adjusted mutual information (*34*) to quantify similarity, we found clusterings produced from shuffled inputs were highly consistent with our final clustering for mouse and human (Fig. S3H-I). The adjusted Rand index was more variable in human, likely because we had to merge more clusters in the original backSPIN output to obtain our final human clustering.

To quantify how read downsampling affected the presence of our neuronal clusters, we downsampled the number of reads per cell. We quantified the presence of our neuronal clusters in the downsampled results using the inverse of the Davies-Bouldin index (*35*), mean Silhouette coefficient (*36*), and Calinski-Harabasz metrics (*37*) (all implemented in the MATLAB function, `evalclusters`; Fig. S3J-L, left). These metrics reflect the separation between clusters, relative to the variability within each cluster. All three measures showed that cluster quality remains consistently high even with 20-40% of the full reads, corresponding to an average of 280,000-560,000 mapped reads per cell. Cluster quality declines upon further downsampling to 10%. We also used these metrics to quantify how well CG methylation can recapitulate our CH-defined clusters (Fig. S3J-L, right). The quality of clusters is similar when using CG or CH methylation, and in both cases it is significantly greater compared to a shuffled control in which cells were randomly re-assigned to a different cluster. Finally, we applied density-based clustering (DBSCAN, (*16*)) to the data using the tSNE coordinates as input, and found generally consistent, though not identical, results compared with backSPIN (Fig. S3M-N). For DBSCAN, we chose parameters that produced clusters most consistent with the visual separation of cells in the tSNE space (epsilon of 0.6 in mouse and 0.8 in human; minimum points of 5 for mouse and 10 for human).

Next, to examine how many cells are required to identify neuronal clusters, we ran tSNE on a random subsample of 500 or 1,000 cells (Fig. S3B). Even with as few as 500 cells, the cell type structure is clearly present in the tSNE output and there is little mixing of different cell types. As expected, reducing the number of cells has the greatest impact on the least numerous cell types. We also examined how reducing (10kb) or increasing (1Mbp) the bin size, and thus changing the scale of corresponding genome features, affected the clustering results (Fig. S3C). Although tSNE results are altered at these two binning levels, the overall cell type structure is still present.

Furthermore, we examined whether mCH information from intra- or inter-genic regions is sufficient to estimate neuronal cell types. After including only reads from within gene bodies (intragenic) or which fall at least 10kb away from the nearest gene body (intergenic), we repeated the binning and tSNE procedure and found similar results (Fig. S3D). There is therefore sufficient information in both genic and intergenic compartments for cell type classification, although we did find that the ratio of inter-cluster variance to intra-cluster variance for individual genomic bins is generally larger for genic regions (Fig. S3E).

Finally, we examined the relationship between experimental factors (e.g. batches, random primer index) and our clusters to identify any potential experimental confounds. We used a chi-squared test for categorical variables and an ANOVA with scalar variables. Clustering was not significantly associated with experimental factors (adjusted p-value > 0.1, Fig. S5).

Processing of single cell RNA-seq and single nucleus RNA-seq datasets

Single cell RNA-seq dataset of mouse somatosensory cortex was downloaded from NCBI GEO accession GSE60361 (*2*). Single cell RNA-seq dataset of mouse visual cortex was downloaded from NCBI GEO accession GSE71585 (*3*). Processed data (transcripts per million table) of single nucleus RNA-seq of human cortex was downloaded from http://genome-tech.ucsd.edu/public/Lake_Science_2016/ (*4*). Mouse single cell RNA-seq datasets were mapped to gencode VM10 reference followed by computing TPM (transcripts per million) for each annotated genes using RSEM 1.2.3 `rsem-calculate-expression` (*38*).

Cross-species comparison of single neuron clusters

For comparing a given human neuron cluster to mouse clusters, we computed cross-specific spearman correlation, for mCH level of marker genes showing homology between the two species (*19*). Correlations were computed between gene mCH level of each individual human neuron and median gene mCH level of each mouse cluster (e.g. mL2/3). Marker genes were identified as described above - marker genes for each pair of clusters were defined as those which were strongly hypomethylated (mCH in the bottom 2nd percentile) in one cluster and hypermethylated (mCH above the 80th percentile) for the other cluster. The homologous mouse neuron type for a single human neuron was defined by the mouse cluster showing strongest correlation of gene mCH level with the single human neuron. The process effectively assigns each human single neuron to a most likely mouse homologous cluster. Comparison of mouse neuron clusters to human neuron clusters were performed similarly.

Comparison of single cell clusters defined by single cell/nucleus RNA-seq and single cell methylome

For comparing a given neuron cluster defined by DNA methylation to clusters defined by RNA-seq, we computed the Spearman correlation between marker gene mCH level (average mCH across annotated genic region) of each individual neuron and the median gene expression level (TPM) for each cluster defined by RNA-seq. Each single neuron was assigned to the cluster defined by RNA-seq showing minimum correlation coefficient since gene body mCH and transcripts abundance are generally inversely correlated.

In situ hybridization (ISH) and image analysis

Wild type 8wk old C57BL/6J male mice (Jackson Laboratory) were anesthetized with isoflurane and brains were removed. Mouse brains were fixed in 10% neutral buffered formalin for 16 hours at room temperature, and were subjected to paraffin embedding at the UCSD Moores Histology & Sanford Consortium Histology Core lab. The sections were cut by at 5µm thickness and mounted onto Superfrost Plus Slides (Thermo Fisher) and baked at 60^0C for 1 hour. Double ISH was performed using RNAscope® technology by Advanced Cell Diagnostics Pharma Assay Services. ISH slides were imaged with Olympus VS120® Virtual Microscopy Slide Scanning System using a 20x objective. Images in TIFF format were extracted using ImageJ BIOP VSI-Reader plugin (*39*). Images were analyzed using a custom Matlab script. Pixel intensities were first centered around zero by subtracting the average pixel intensity from the entire image. Since RNAscope® assay labels individual RNA molecules, the overlap between co-stained probes was computed at the cell body level. In order to identify neuronal cell bodies, a 30 x 30 pixel sliding window (equivalent to 9.7 x 9.7 µm),similar to the size of a neuronal cell body, was used to scan the image with a step size of 10 pixels. Average pixel intensity was quantified for each sliding position and was standardized by converting to z-score. Probe specific z-score thresholds (Sulf1 - 3, Tle4 - 3, Adgra3 - 7, Pvalb - 7) were defined for the selection of sliding window positions that overlapped with a cell body and showed fluorescent intensity greater than the threshold. Connected sliding window positions were merged to create a list of regions of interests (ROIs), each corresponding to a cell body. The overlap between ROIs of the two imaging channels were counted to determine the co-expression of probed genes.

Identification of CG-DMR and superenhancer-like large CG-DMRs

Files containing unmethylated and methylated cytosine base calls for each cytosine position (allc files) were merged across single cells within each cluster to generate aggregate methylation data for each neuronal cluster. For each CpG site, base calls for the two cytosines located on opposite strands were combined to increase the power of DMR calling. CG-DMRs were then called using Methylpy `DMRfind` with false discovery rate cutoff = 0.01. Differentially methylated sites (DMSs) located within 250 bp of one another were combined into differentially methylated regions (DMRs). DMRs containing at least two DMSs were retained for subsequent analyses.

For the identification of large CG-DMRs, CG-DMRs were first merged allowing 1kb distance between each other using `bedtools merge -d 1000`. Merged CG-DMRs with size greater than 5kb were considered large CG-DMRs. Large CG-DMRs were ranked by their size in Tables S7-8.

Comparison of single cell methylome methods

scBS-seq dataset was downloaded from NCBI GEO accession GSE56879 (*10*), scM&T-seq dataset was downloaded from NCBI GEO accession GSE74535 (*12*). Since sc-WGBS data deposited to NCBI SRA contains non-redundant mapped reads from (*11*), we were not able to determine mapping rate and library complexity using our processing pipeline. scWGBS libraries were generated in-house from single mouse cortical nuclei using Illumina

Truseq Methylation kit as described in (*11*). Sequencing reads were first trimmed to remove sequencing adaptors using Cutadapt 1.11 (*27*) with the following parameters in paired-end mode: -f fastq -q 20 -m 62 -a AGATCGGAAGAGCACACGTCTGAAC -A AGATCGGAAGAGCGTCGTGTAGGGA. For singleplex samples, -m parameter was set to 40. 10 bp were further trimmed from both 5'- and 3'- ends of R1 and R2 reads to remove random primer index sequence with the following parameters: -f fastq -u 16 -u -16 -m 30. R1 and R2 reads were mapped separately as single-end reads using Bismark v0.15.0 with parameter --bowtie2. For mapping scBS-seq data, --pbat option was activated for mapping R1 reads (*28*). For mapping sc-WGBS data, --pbat option was activated for mapping R2 reads. Library complexity was estimated using R1 reads with Preseq gc_extrap function with options -e 5e+09 -s 1e+07 (*40*).

To determine the enrichment of CpG islands (CGI) in single cell methylome data, the fraction of CGI on mouse chromosome 1 covered by a single cell methylome was compared to shuffled regions with matching sizes. The shuffling was carried out using `bedtools shuffle` and was repeated five times and the average fraction of regions covered by reads was used. Bulk MethylC-seq data was downsampled to 1 million non-clonal reads for this anlsysis. For computing the amount of genomic regions covered by reads at different sequencing coverage, 1kb and 10kb bins were generated using `bedtools makewindows` across mouse genome. The bins were intersected with bulk MethylC-seq and single cell methylomes downsampled to 100,000 to 1 million reads.

Transcription factor (TF) binding motif enrichment analysis

TF binding motif enrichment analysis was performed as described in (*5*, *41*) with the following modifications. The analysis of TF binding motif enrichment in mouse and human CG-DMRs only considered TFs with median TPM >= 10 in any clusters defined by single cell/nucleus RNA-seq of mouse visual cortex and human cortex, respectively (*3*, *4*). To summarize enriched or depleted TF binding motifs to TF classes, classification of TFs was downloaded from http://tfclass.bioinf.med.uni-goettingen.de/tfclass (*42*). The folds of enrichment or depletion for TF classes were defined as the strongest enrichment or depletion shown by TF class members.

Prediction of putative enhancers

The enhancers of three major brain cell types (excitatory neurons, PV neurons and VIP neurons) were predicted using Regulatory Element Prediction based on Tissue-specific Local Epigenetic marks (REPTILE) (*20*). REPTILE integrates DNA methylation and chromatin accessibility data to delineate the location of enhancers. REPTILE formulates enhancer prediction as a supervised learning task – it learns the chromatin signatures of enhancers (i.e. enhancer model) using known enhancers and then makes predictions across the whole genome in various cell types and tissues. A unique feature of RETPILE is that it is able to incorporate the data of cells and tissues besides the target sample (as outgroup) and utilize the variation of epigenomic data to improve prediction accuracy.

To generate the putative enhancers of the three brain cell types, we first downloaded the bulk WGBS and ATAC-seq data of excitatory neurons (EXC), PV neurons (PV) and VIP neurons (VIP) from Gene Expression Omnibus (GEO). The accessions of ATAC-seq data are: EXC (GSM1541964, GSM1541965), PV (GSM1541966, GSM1541967) and VIP (GSM1541968, GSM1541969). The accessions of WGBS are EXC (GSM1541958, GSM1541959), PV (GSM1541960, GSM1541961) and VIP (GSM1541962, GSM1541963). In order to train a mouse enhancer model, we also obtained the EP300 ChIP-seq data (GSM723018) and its corresponding input (GSM723020) in mouse embryonic stem cells (mESCs) as well as the bulk ATAC-seq data (GSM2156965) and bulk WGBS (GSM1162043 and GSM1162044) of mESCs.

Next, The WGBS and ChIP-seq data were processed as previously described (*20*). ATAC-seq data were processed in the same way as (*5*). Data of replicates were combined. EP300 binding sites were identified using MACS2 (*43*) similar to (*20*). DMRs were called across the methylomes of mESCs and three brain cell types as previously stated (*20*).

Then, we trained a mouse enhancer model in mESCs. The construction of training dataset as described in (*20*) - the EP300 binding sites were treated as positive instances, whereas promoters and randomly chosen genomic bins were used as negative instances. During the training process, the data of three brain cell types were used as outgroup. After training, we obtained a mouse enhancer model, which is able to distinguish enhancers from genomic background based on the mCG and open chromatin signatures.

Lastly, we applied this model to generate enhancer predictions for three brain cell types. During this process, when REPTILE made enhancer predictions for one brain cell type, mESCs and the other two brain cell types were used as outgroup.

Prediction of excitatory neuron super-enhancers

Excitatory neuron super-enhancers were identified with ROSE (http://younglab.wi.mit.edu/super_enhancer_code.html (*22*)) using the list of H3K27ac peaks identified from cortical excitatory neurons with parameters -s 12500 -t 2500. Excitatory neuron H3K27ac ChIP-seq data and peaks were reported in (*5*).

Comparative analysis of regulatory elements

CG-DMRs were categorized based on the conservation of sequences and methylation states between human and mouse. First, UCSC liftOver (*44*) was used to project CG-DMRs between species based on sequence conservation (minimum ratio of remapped bases = 0.1). CG-DMRs that could be mapped to the other species and mapped back were referred to as mappable CG-DMRs. All other CG-DMRs were called unmapped DMRs. Next, we further divided mappable CG-DMRs into two categories: CG-DMRs located within 1kb of a CG-DMR in the other species (shared CG-DMRs), and CG-DMRs located further than 1 kb away from any DMR in the other species after liftover (specific CG-DMRs).

We then used a hypergeometric test to examine whether mappable CG-DMRs from one species preferentially overlap with CG-DMRs in the homologous cell type from the other species. Specifically, to calculate the expected number of shared DMRs between human cluster *i* and mouse cluster *j*, we mapped (via liftover) human cluster *i* DMRs to the mouse genome and found how many were shared with any mouse DMR (i.e. overlap within 1kb of merged mouse DMRs, N_{ij}). Then this number was divided by the total number of merged human DMRs (N_h) and multiplied by the number of DMRs in human cluster *i* (N_{hi}), to get the expected number of shared DMRs: $E_{ij} = \frac{N_{ij}}{N_h} N_{hi}$. This was compared with the observed number of shared DMRs between human cluster *i* and mouse cluster *j*, N_{ij} .

To quantify the regulatory conservation of CG-DMRs between human and mouse, we computed the correlation of methylation levels at mappable DMRs for each homologous cluster pair. The higher cross-species correlation of inhibitory clusters suggests that inhibitory neurons have greater regulatory conservation between the two species (Fig. 4E, Fig. S17C, p<0.001, Wilcoxon rank-sum test). This finding was further corroborated by examining cross-species enrichment of shared CG-DMRs represented by the fold-change between observation and expectation, which again showed stronger overlap between the two species in inhibitory than excitatory neuron clusters (Fig. S17D-E).

To measure sequence conservation of CG-DMRs, we computed 100-way PhastCons score for human CG-DMRs and 60-way PhastCons score for mouse CG-DMRs by taking the average PhastCons score across all bases in each DMR. Missing values were skipped rather than treated as zero. We observed higher PhastCons scores at CG-DMRs in inhibitory neuron clusters than in excitatory (Fig. 4E, p<0.001, Wilcoxon rank-sum test), suggesting greater sequence conservation of inhibitory neuron regulatory elements across species. We then performed the same analysis at putative enhancers identified by REPTILE from purified neuronal populations (*20*), and also found greater sequence conservation in inhibitory than excitatory neuron putative enhancers (Fig. S17F).

We calculated average PhastCons score across the region surrounding (+/- 10kb) of excitatory or inhibitory neuron CG-DMRs with 100 bp resolution. The conservation of inhibitory neuron CG-DMRs is greater than in excitatory only within 500bp around the CG-DMR (Fig. S17G). Then we investigated whether genes preferentially expressed in inhibitory neurons are more conserved at their gene body. We used nuclear RNA-seq data (*5*) to find genes with at least 2 fold over-expression in one cell type against the other two cell types. Excitatory neuron specific genes showed greater sequence conservation than PV and VIP specific genes at their TSS and similar conservation in flanking regions (Fig. S17H). This result indicates that the higher conservation of inhibitory cells may be restricted to the regulatory elements.

We further divided mappable CG-DMRs into two categories: those proximal to a TSS (within 25kb) and those distal to a TSS, computing the PhastCons score for each. Results showed greater conservation of inhibitory neuron CG-DMRs, with the difference being more pronounced for distal CG-DMRs (Fig. S17I).

Finally, we examined whether excitatory and inhibitory neuron CG-DMRs associated with the same gene (within 25kb of TSS) show different conservation. For each gene, we compared the PhastCons score between DMRs in excitatory cells and inhibitory cells that associated with the genes, and again we observed a significant higher conservation in inhibitory DMRs (Fig. S14J, p<1e-6, Wilcoxon signed-rank test).

Supplementary texts

snmC-seq shows reliable sample multiplexing and high reads mapping rate

Our snmC-seq protocol starts from separating single nuclei using fluorescence-activated cell sorting (FACS) and dispense into wells of 384- well PCR plates followed by proteolytic digestion and bisulfite conversion. snmC-seq is compatible with both fresh and frozen tissues. We incorporated 5'- sequencing adaptors through indexed random primer-initiated DNA synthesis. (Fig. 1A) After pooling four indexed random priming reactions, 3'- sequencing adaptors were incorporated using Adaptase^{TM} technology. The majority of multiplexed pools were constructed by combining two mouse and two human nuclei. Mapping of sequencing reads to both mouse and human reference genomes showed negligible cross-species mapping (Fig. S2A), confirming the fidelity of the multiplexing strategy.

We have compared snmC-seq with published methods for single cell methylome including scBS-seq (*10*), scWGBS (*11*), scM&T-seq (*12*). We specifically examined fraction of reads retained after adaptor trimming, unique mapping rate and library complexity (Fig. S2B-D). We generated scWGBS libraries from single mouse cortical nuclei using Illumina Truseq Methylation kit as described in (*11*). snmC-seq shows significantly greater mapping rate (median = 52.7%) compared to scM&T-seq (median = 19.8%, (*12*)) and scBS-seq (median = 22.5%, Fig. S2C (*10*)). The mapping rate of snmC-seq is comparable to scWGBS libraries (median = 55.4%). However, snmC-seq library contains approximately four times more unique molecules than scWGBS libraries (Fig. S2D).

Reads coverage pattern was compared between downsampled traditional bulk MethylC-seq data, snmC-seq, scBS-seq (*10*) and sc-WGBS (*11*). It was previously shown that the coverage of CpG islands (CGI) is enriched in single cell methylome (*10*, *11*). CGI enrichment was determined relative to shuffled genomic regions with matching sizes (Fig. S2E). snmC-seq showed similar enrichment of CGI (mean fold change = 1.65x) as sc-WGBS (mean fold change = 1.54x), while scBS-seq showed moderately higher CGI enrichment (mean fold change = 2.02x). Traditional MethylC-seq showed depletion of CGI with a mean fold change of 0.52x.

To quantify the evenness of single cell methylome coverage, the fraction of non-overlapping 1kb and 10kb genomic bins covered by sequencing reads were plotted as a function of sequencing depth for each method (Fig. S2F). With a given number of sequencing reads, traditional MethylC-seq data always covers most genomic bins, suggesting less coverage bias than single methylome methods. Single cell methylome methods show moderate difference between their coverage evenness, with snmC-seq showing intermediate evenness between sc-WGBS and scBS-seq measured with 1kb bin coverage, and near identical evenness with sc-WGBS measured with 10kb bin coverage.

hPv-2 is a potentially human specific PV+ inhibitory neuron population

hPv-2 represented a potential human-specific inhibitory population. The strong hypomethylation of *GAD1* and *LHX6* genes in hPv2 suggests these are inhibitory neurons derived from the medial ganglionic eminence (MGE) (Fig. S12B); however, hPv-2 was the only inhibitory neuron cluster in either mouse or human showing hypermethylation of *GAD2*. Notably, hPv-2 cells have low mCH at *CCK* and high methylation at *GRIK3*, similar to caudal ganglionic eminence (CGE) derived interneurons, such as VIP cells, and distinct from MGE-derived inhibitory cells.

Unique large CG-DMRs were also found in hPv-2 (Fig. S16E,J). Gene bodies of *NACC2*, *UNC5B*, *FAM20C* and *FAM222A* were demethylated in hPv-2 but not in any other human or mouse inhibitory neuron clusters (Fig. S16E, J). Thus, these observations suggest hPv-2 is a unique human PV neuron population defined by both marker gene mCH and super-enhancer mCG signatures.

Inhibitory neurons show layer-specific DNA methylation signatures

We found that global mCH level differed among inhibitory neurons within a clusters but located in different cortical layers (Fig. S14A). For example, PV+ interneurons located in superficial layers had significantly less global mCH than middle and deep layer PV+ neurons ($p<1\times10^{-5}$, Wilcoxon rank sum test). Significant global mCH layer differences were also found between superficial and middle layer SST+ neurons ($p<1\times10^{-3}$, Wilcoxon rank sum test). Moreover, we identified 406 genes with layer specific mCH in PV+ neurons (Fig. S14B, one way ANOVA q-value < 0.01; Table S4). The vast majority (358) of these genes were hypomethylated in superficial layer PV+ neurons. In addition, MGE-derived inhibitory populations, including PV+ and SST+ but not CGE-derived VIP+ neurons, shared a significant fraction of genes showing hypomethylation in superficial layers (hypergeometric p-value=8.7×10^{-59}, Fig. S14E), suggesting that layer-specific gene regulation in mature inhibitory neurons may be defined by their progenitor zones. Genes with low mCH in superficial layer PV+ neurons are enriched in functional annotations including neurogenesis, axon guidance functions and synapse part (Fig. S14F-H), suggesting layer-specific epigenetic regulation of synaptic functions in inhibitory neurons.

We identified human PV+ and SST+ neurons that putatively located in different layers by comparing to mouse superficial or deep layer neurons (Fig. S14I). Human PV+ and SST+ neurons putatively located in different layers were separated by tSNE (Fig. S14J and K). Superficial and deep layer human SST+ neurons were also separated by clustering, with hSst-2 correlated with superficial layer mSst-1 whereas hSst-1 and hSst-3 correlated with deep layer mSst-1 (Fig. S14I). Superficial layer human PV+ and SST+ neurons had less global mCH compared with neurons of the same type located in deep layers (Fig. S14M). A group of genes, including *Cux2*, *Nlgn1*, *Grin2a* and *Shank2*, showed similar layer specific mCH patterns between mouse and human (Fig. S14N-O).

Large CG-DMR is a reliable marker for superenhancer

We tested the specificity of superenhancer prediction with large DMR using CG-DMRs found in mL2/3. Putative excitatory neuron superenhancers were predicted using enhancer histone mark H3K27ac profile of purified Camk2a+ excitatory neurons with software ROSE (*5*, *22*). mL2/3 has a high-coverage aggregated methylome from 690 single neurons, which allowed sensitive CG-DMR calling for this cluster. We first merged mL2/3 CG-DMRs located within 1kb from one another and then ranked merged CG-DMRs by their size. We found that the enrichment of H3K27ac over merged CG-DMRs increases along with the size of CG-DMRs (Fig. S16A), suggesting large CG-DMRs are associated with strong regulatory activity. A greater portion of large DMRs (e.g. > 15kb) were overlapped with putative superenhancers, compared to DMRs with smaller sizes (Fig. S16B). For example, 90.3% of merged DMRs larger than 15kb, whereas only 24.8% of merged DMRs with size greater than 2kb were overlapped with putative superenhancers.

Supplementary Figures

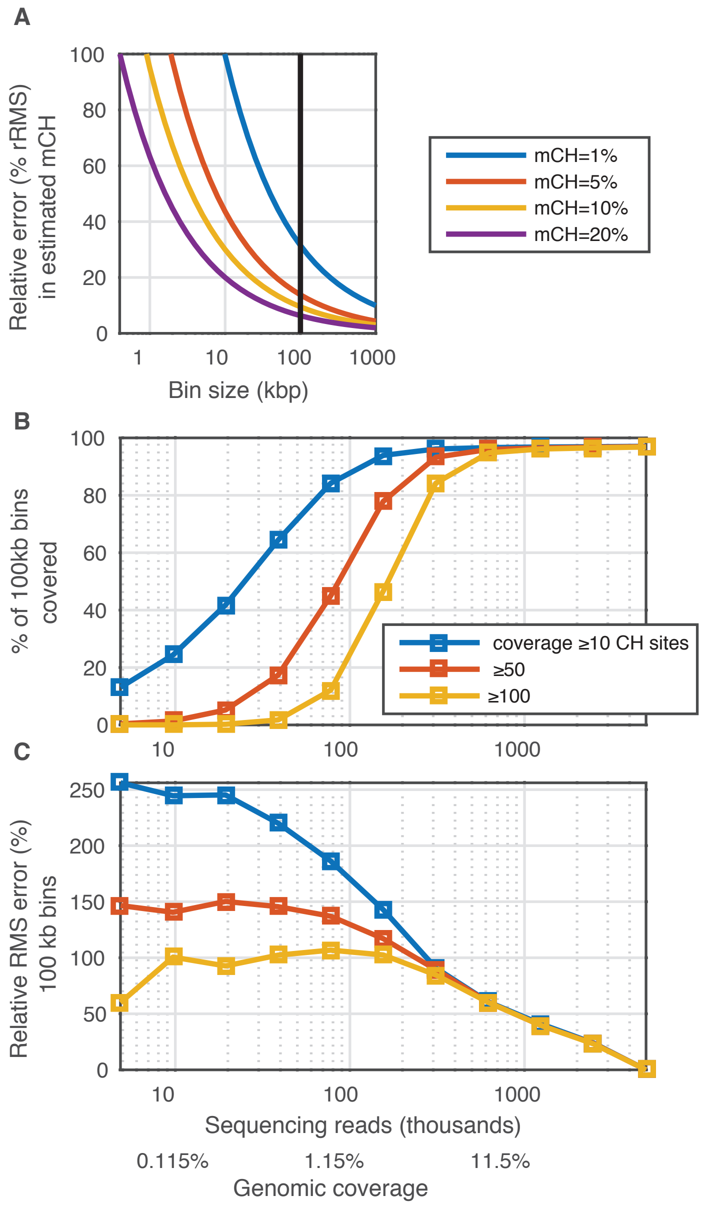
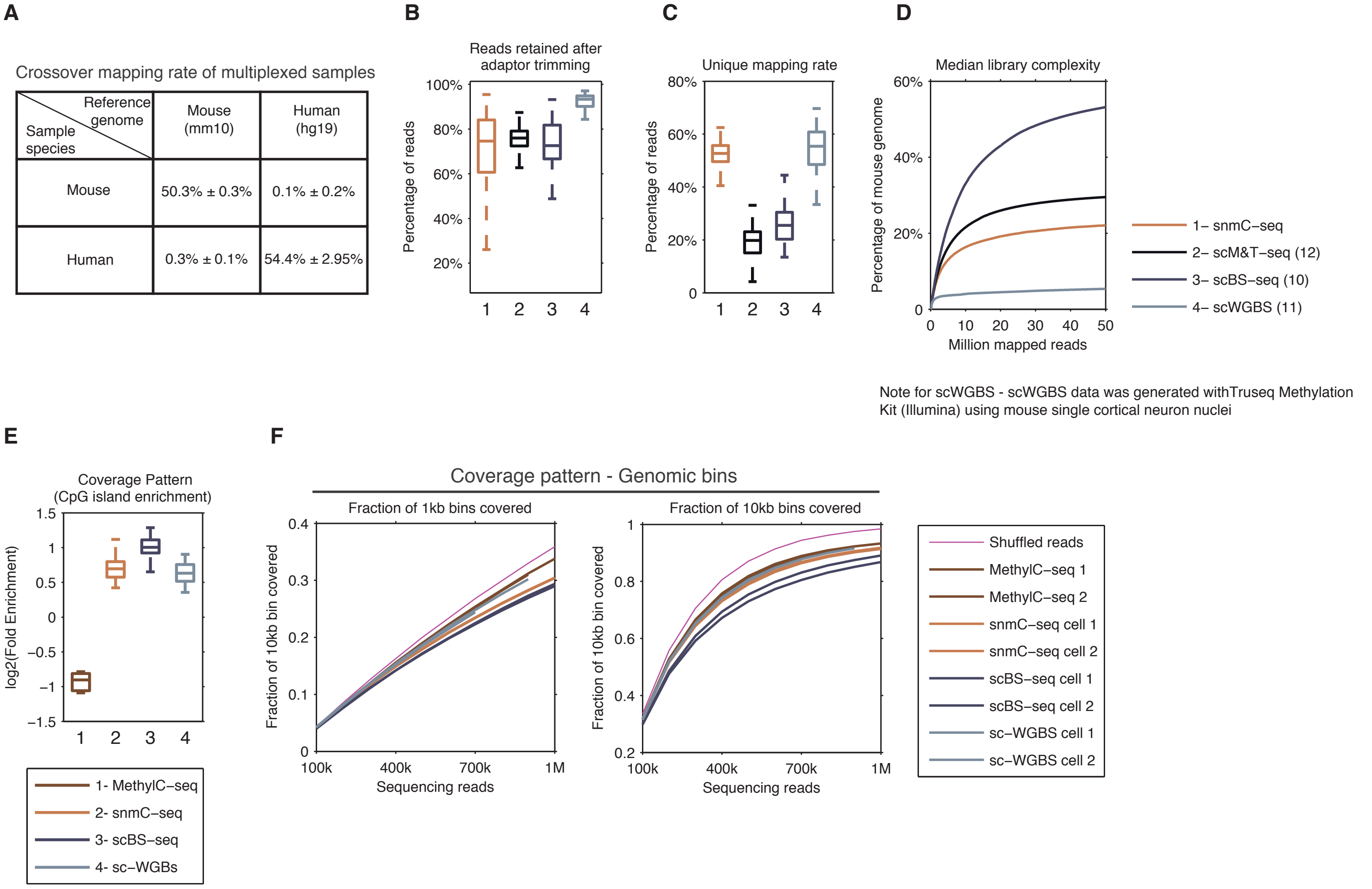


Fig. S1. mCH can be accurately estimated within 100kb bins using sparse snmC-seq data. (A) Theoretical model estimates the relative RMS error in mCH in genomic bins e = sqrt((1-p)/(prcb)), where p ≈ 0-0.2 is the true methylation level, r ≈ 0.05 is the genomic coverage, c ≈ 0.2 is the fraction of CH positions in the genome, and b is the genomic feature size. (B) Downsampling a deeply sequenced single cell methylome shows that >95% of the genome can be covered with ≥100 CH basecalls per 100 kb bin, assuming ~500,000 reads or ~5% genomic coverage. (C) The rRMS error is estimated by comparing the mCH estimated using the full coverage data with downsampled data.

Fig. S1



Crossover mapping rate of multiplexed samples

Sample species \ Reference genome	Mouse (mm10)	Human (hg19)
Mouse	50.3% ± 0.3%	0.1% ± 0.2%
Human	0.3% ± 0.1%	54.4% ± 2.95%

Note for scWGBS - scWGBS data was generated withTruseq Methylation Kit (Illumina) using mouse single cortical neuron nuclei

Fig. S2. snmC-seq is compatible with multiplexing and demonstrates efficient read mappability. (A) Mapping of 100 randomly selected multiplexed snmC-seq samples to both human and mouse reference genomes showed no species crossover between pooled indexed random priming reactions. (B) Percentage of sequencing reads retained after trimming of generic Illumina adaptors, random primer index and low complexity tail introduced by AdaptaseTM for snmC-seq. (C) Percentage of trimmed sequencing reads that were uniquely mapped. (D) Complexity of single cell methylome libraries estimated using R1 reads. (E) Enrichment of CpG islands in DNA methylome generated by traditional MethylC-seq, snmC-seq, scBS-seq and sc-WGBS. (F) Fraction of 1kb and 10kb non-overlapping bins covered by single cell methylome data as a function of the number of sequencing reads.

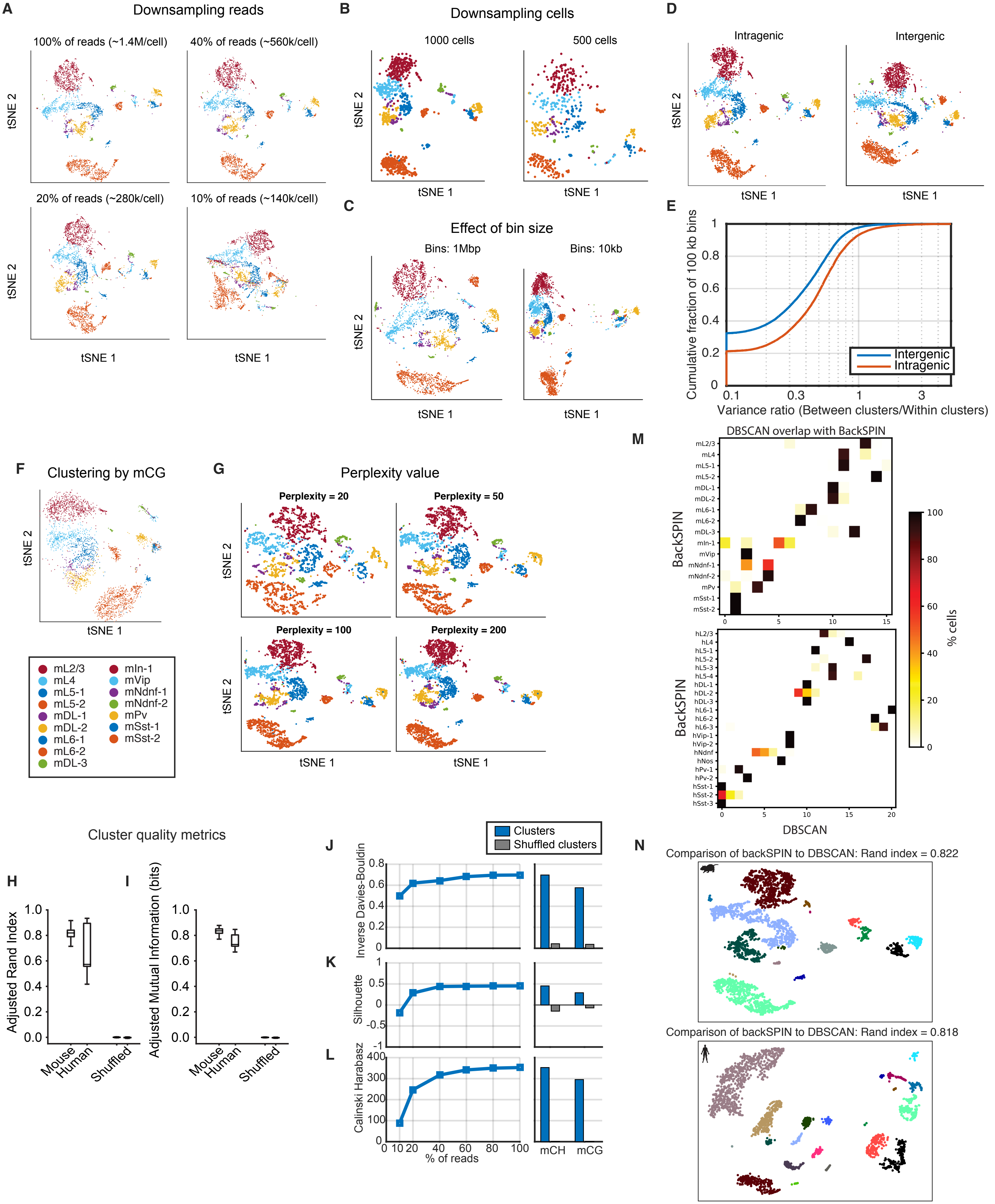

A
Downsampling reads
100% of reads (~1.4M/cell)
40% of reads (~560k/cell)
20% of reads (~280k/cell)
10% of reads (~140k/cell)
tSNE 2
tSNE 1
B
Downsampling cells
1000 cells
500 cells
C
Effect of bin size
Bins: 1Mbp
Bins: 10kb
D
Intragenic
Intergenic
E
Cumulative fraction of 100 kb bins
0
0.2
0.4
0.6
0.8
1
0.1
0.3
1
3
Variance ratio (Between clusters/Within clusters)
Intergenic
Intragenic
F
Clustering by mCG
G
Perplexity value
Perplexity = 20
Perplexity = 50
Perplexity = 100
Perplexity = 200
mL2/3
mL4
mL5-1
mL5-2
mDL-1
mDL-2
mL6-1
mL6-2
mDL-3
mIn-1
mVip
mNdnf-1
mNdnf-2
mPv
mSst-1
mSst-2
M
DBSCAN overlap with BackSPIN
BackSPIN
mL2/3
mL4
mL5-1
mL5-2
mDL-1
mDL-2
mL6-1
mL6-2
mDL-3
mIn-1
mVip
mNdnf-1
mNdnf-2
mPv
mSst-1
mSst-2
0
5
10
15
hL2/3
hL4
hL5-1
hL5-2
hL5-3
hL5-4
hDL-1
hDL-2
hDL-3
hL6-1
hL6-2
hL6-3
hVip-1
hVip-2
hNdnf
hNos
hPv-1
hPv-2
hSst-1
hSst-2
hSst-3
0
5
10
15
20
DBSCAN
100
80
60
40
20
0
% cells
Cluster quality metrics
H
Adjusted Rand Index
0.0
0.2
0.4
0.6
0.8
1.0
Mouse
Human
Shuffled
I
Adjusted Mutual Information (bits)
0.0
0.2
0.4
0.6
0.8
1.0
Mouse
Human
Shuffled
J
Clusters
Shuffled clusters
Inverse Davies-Bouldin
0
0.2
0.4
0.6
0.8
K
Silhouette
-1
-0.5
0
0.5
1
L
Calinski Harabasz
0
100
200
300
400
0 10 20
40
60
80
100
% of reads
mCH
mCG
N
Comparison of backSPIN to DBSCAN: Rand index = 0.822
Comparison of backSPIN to DBSCAN: Rand index = 0.818

Fig. S3

Fig. S3. Single nuclei are consistently clustered by cell type using multiple methylome features across a wide range of genomic length scales. (A) Cell types can be clearly separated in tSNE space despite downsampling reads to 20% of the full mouse dataset (~280,000 uniquely mapped reads per cell); the quality of clustering begins to break down at 10% downsampling (140,000 reads). tSNE analysis was performed using mCH levels in 100kb bins (minimum coverage, 100 base calls), and each cell was colored according to the cluster identity assigned in our analysis of the full dataset. (B) Major cell types are well separated by tSNE analysis using as few as 500 or 1,000 cells; increasing the number of cells increases the representation of minority cell types. (C) Cell type clusters can be identified by tSNE analysis using mCH in bins as small as 10kb or as large as 1Mb. (D) Comparison of tSNE representation of mouse clusters based on mCH in intragenic regions (including all bases between each TSS and TES) vs. intergenic regions (>10 kb away from any gene body). (E) The cumulative distribution over all 100 kb bins of the ratio of between-cluster variance (i.e. the variance of the mean mCH for each cluster) vs. the within-cluster variance (i.e. the variance of all cells, after subtracting the cluster mean) for inter- and intra-genic reads. (F) Cell type clusters can be identified by tSNE analysis using mCG in 100 kb bins. (G) tSNE representation of single mouse neuron methylome with different perplexity values shows consistent patterns. (H-L) The quality of backSPIN clustering is shown for mouse and human using the adjusted Rand index (H), adjusted mutual information (I), inverse of the Davies-Bouldin index (J), mean silhouette index (K) and Calinski-Harabasz index (L). For indices shown in (H-K), a value close to 1 indicates that clusters are well separated relative to the variability within each cluster, while a value close to 0 indicates poor cluster separation. Box plots show the distribution of each index over 200 clustering runs with random initialization, and they are compared with the results for randomly shuffled cluster assignments. (M-N) Comparison of BackSPIN clustering results to clusters generated from tSNE and DBSCAN for mouse and human.

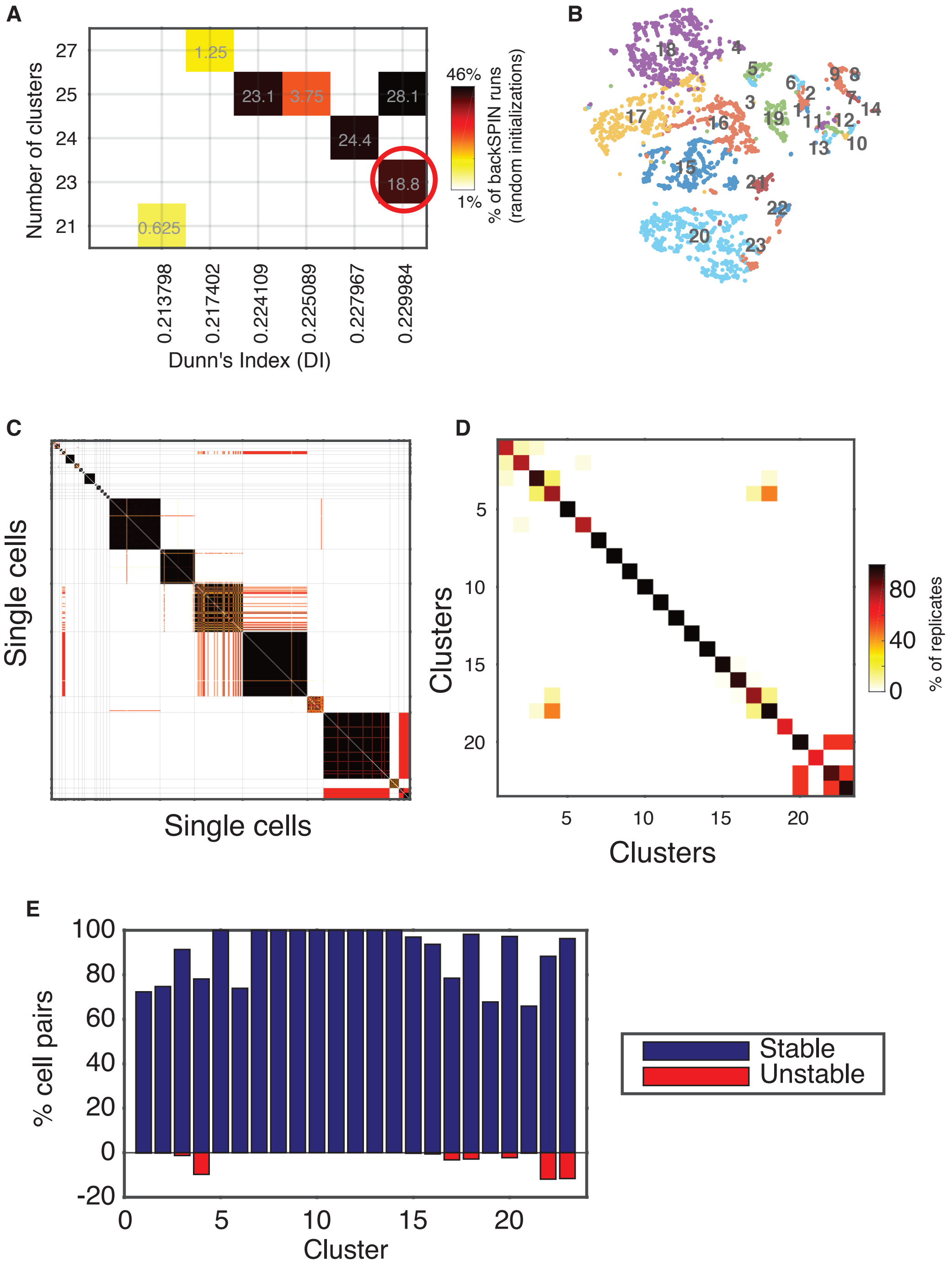


Fig. S4. Cluster robustness. (A) 160 independent clusterings were generated with backSPIN using random initialization. Each backSPIN run converged to one of 7 different clusterings; we selected the clustering with the highest Dunn's Index as a reference clustering (red circle). (B) tSNE plot showing the reference clustering. (C) For each pair of cells, the color shows the fraction of backSPIN runs in which those two cells were co-clustered. (D) Average co-clustering for all cell pairs in two different clusters. (E) For every cell pair in each cluster, we plot the % stability, i.e. the % of runs in which the cells are co-clustered. We also plot the % unstable, i.e. the % of cell pairs that are not in the same cluster which are co-clustered.

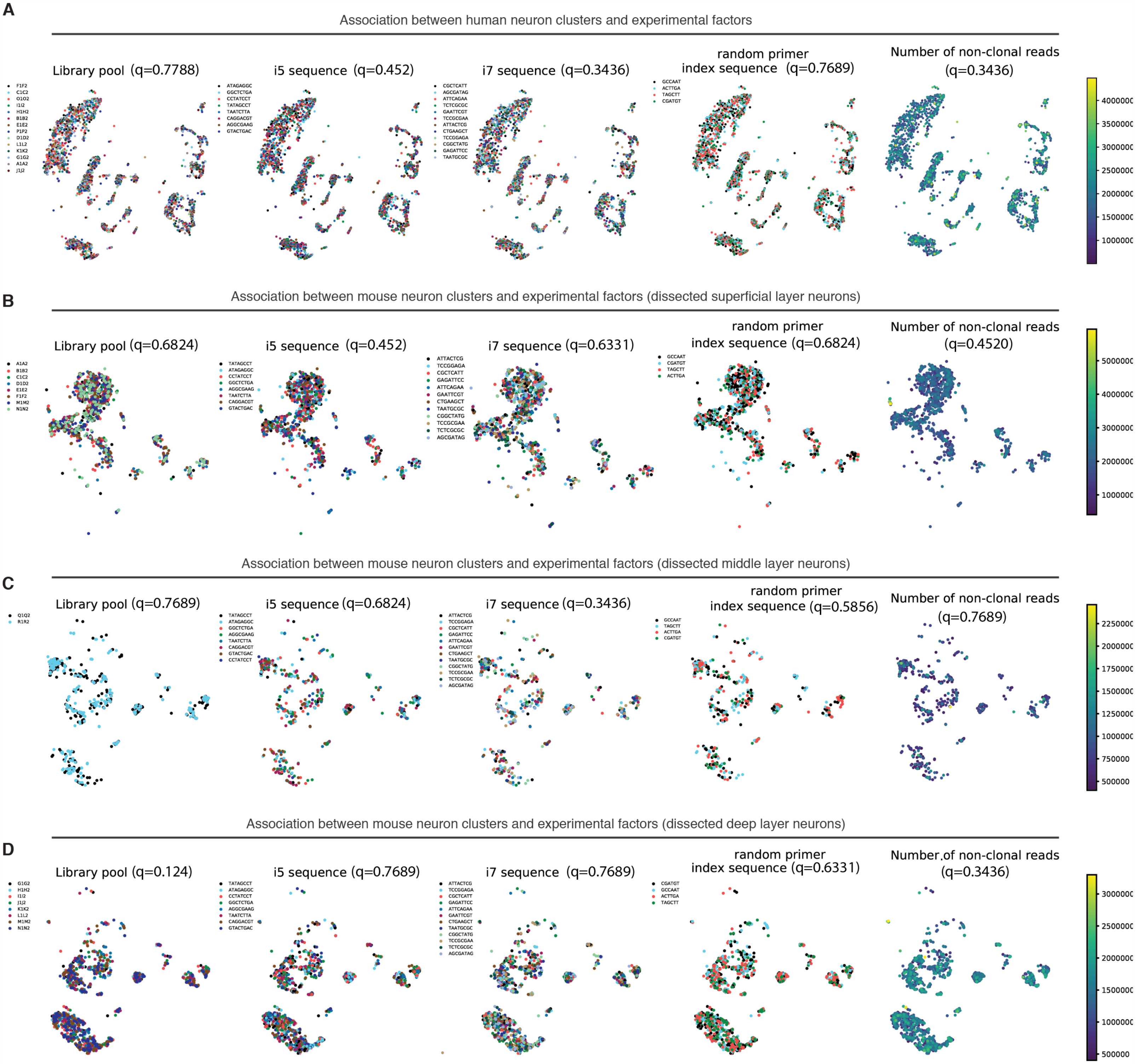


Fig. S5. Absence of strong association between neuron clusters and experimental factors. Statistical comparison of clustering to sequencing experimental factors for human neurons (A), dissected mouse superficial layer (B), middle layer (C) and deep layer (D) neurons.

Fig. S5

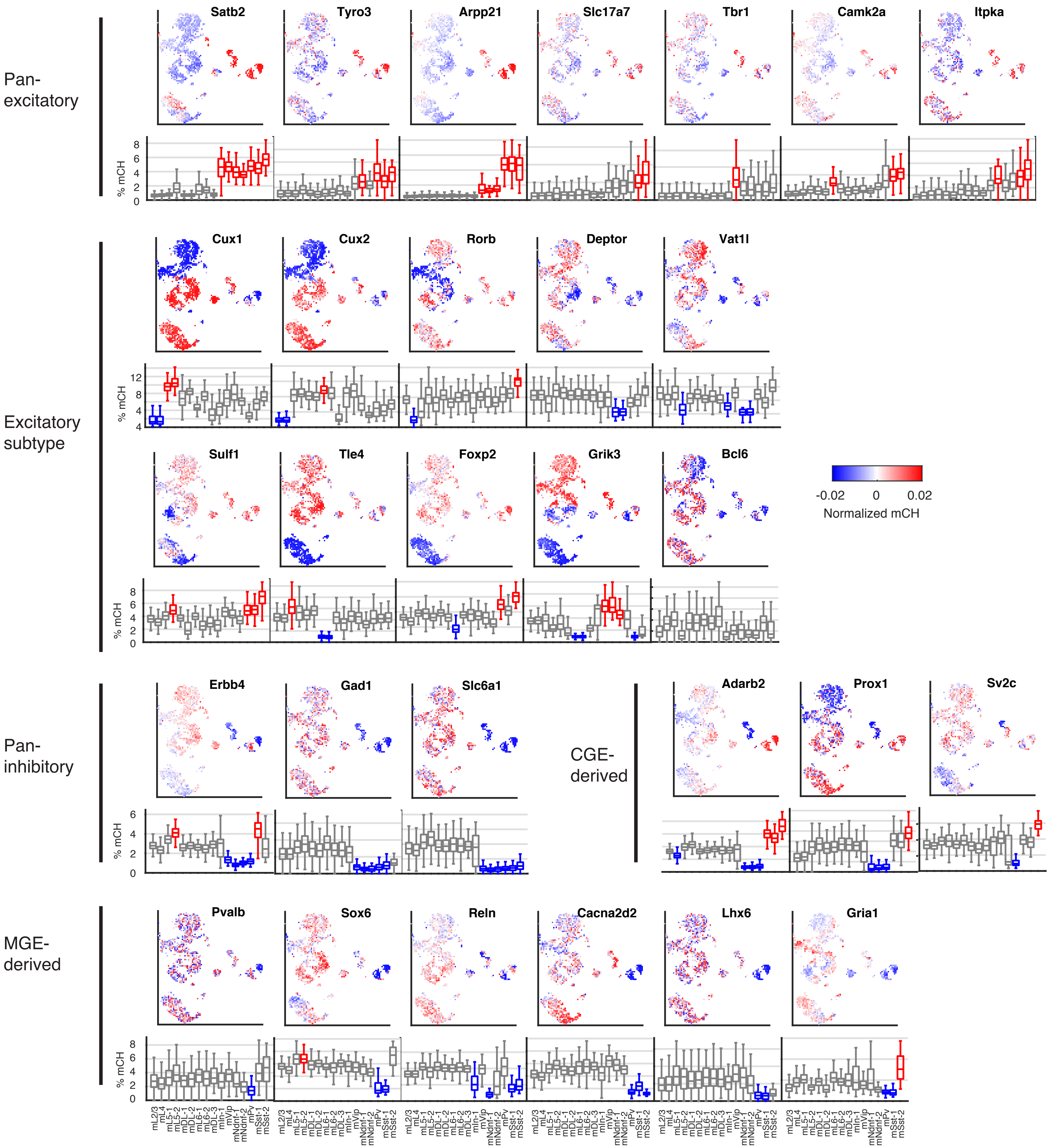


Fig. S6. Mouse marker genes. For each gene, single cells are shown in tSNE representation colored according to the normalized mCH level. Box plots below each tSNE show the distribution of absolute (not normalized) mCH level across all cells within each cluster. For each gene's box plot, we highlight clusters that are significantly hypermethylated (red) or hypomethylated (blue). Hypermethylated (hypomethylated) clusters are defined to be clusters for which at least 75% of cells have higher (lower) methylation than the top (bottom) 25% of cells in all other clusters.

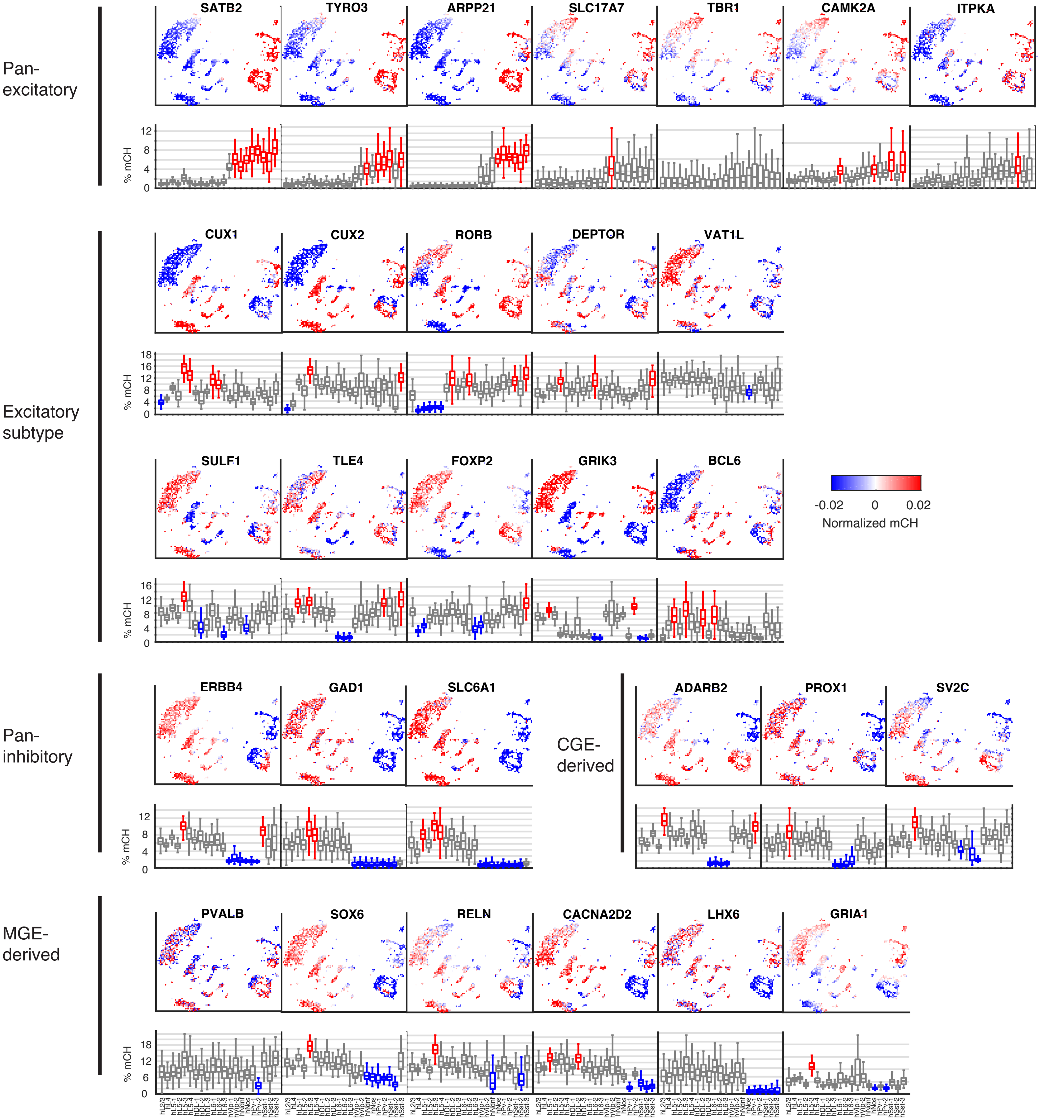


Fig. S7. Human marker genes. For each gene, single cells are shown in tSNE representation colored according to the normalized mCH level. Box plots below each tSNE show the distribution of absolute (not normalized) mCH level across all cells within each cluster. For each gene's box plot, we highlight clusters that are significantly hypermethylated (red) or hypomethylated (blue). Hypermethylated (hypomethylated) clusters are defined to be clusters for which at least 75% of cells have higher (lower) methylation than the top (bottom) 25% of cells in all other clusters.

Fig. S7

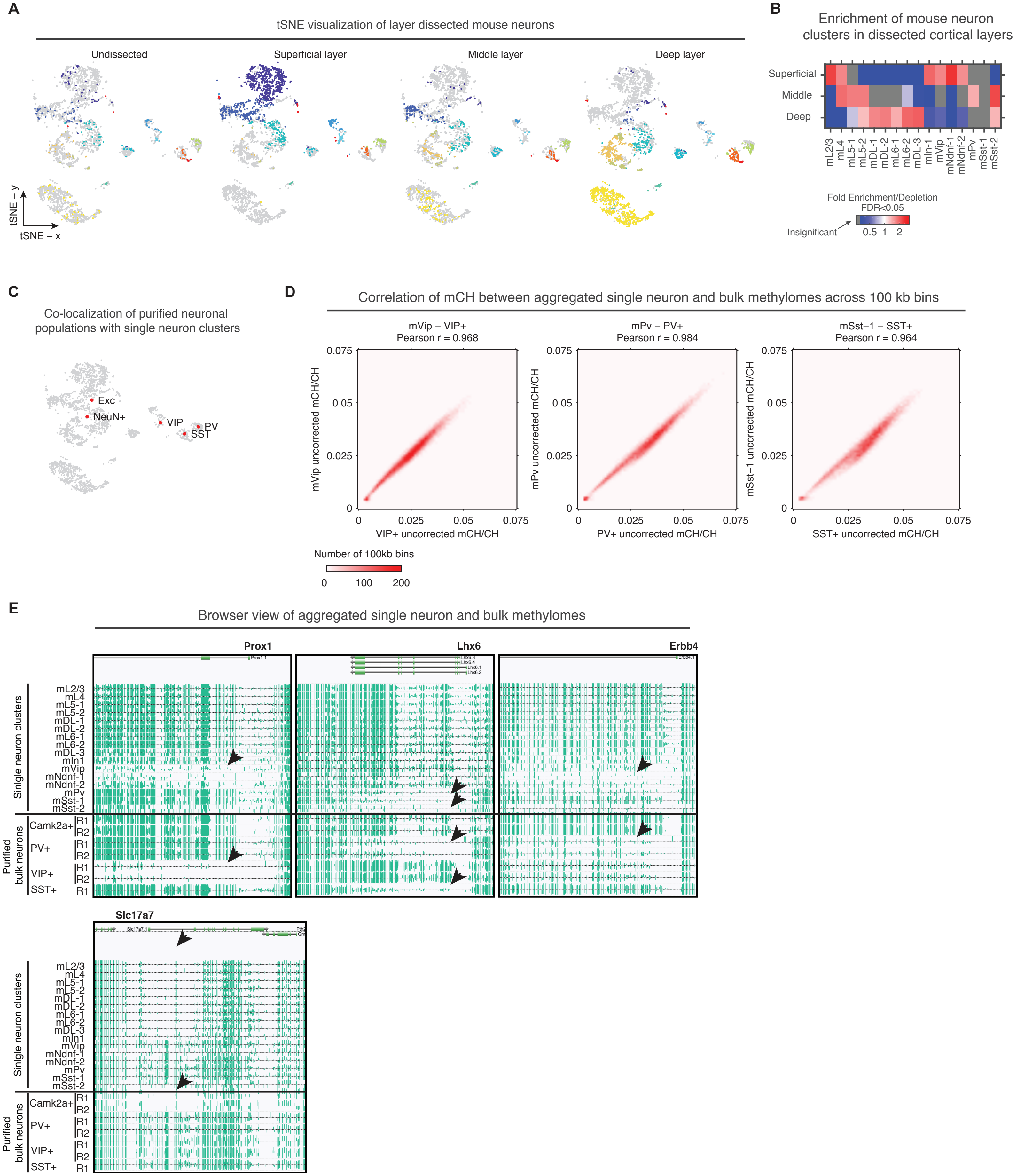

A
tSNE visualization of layer dissected mouse neurons
Undissected
Superficial layer
Middle layer
Deep layer
tSNE – y
tSNE – x
B
Enrichment of mouse neuron clusters in dissected cortical layers
Superficial
Middle
Deep
mL2/3
mL4
mL5-1
mL5-2
mDL-1
mDL-2
mL6-1
mL6-2
mDL-3
mIn-1
mVip
mNdnf-1
mNdnf-2
mPv
mSst-1
mSst-2
Fold Enrichment/Depletion
FDR<0.05
Insignificant
0.5 1 2
C
Co-localization of purified neuronal populations with single neuron clusters
Exc
NeuN+
VIP
PV
SST
D
Correlation of mCH between aggregated single neuron and bulk methylomes across 100 kb bins
mVip – VIP+
Pearson r = 0.968
mPv – PV+
Pearson r = 0.984
mSst–1 – SST+
Pearson r = 0.964
mVip uncorrected mCH/CH
mPv uncorrected mCH/CH
mSst–1 uncorrected mCH/CH
VIP+ uncorrected mCH/CH
PV+ uncorrected mCH/CH
SST+ uncorrected mCH/CH
0
0.025
0.05
0.075
Number of 100kb bins
0
100
200
E
Browser view of aggregated single neuron and bulk methylomes
Prox1
Lhx6
Erbb4
Slc17a7
Prox1.1
Lhx6.3
Lhx6.4
Lhx6.1
Lhx6.2
Erbb4.1
Slc17a7.1
Sinlge neuron clusters
mL2/3
mL4
mL5-1
mL5-2
mDL-1
mDL-2
mL6-1
mL6-2
mDL-3
mIn1
mVip
mNdnf-1
mNdnf-2
mPv
mSst-1
mSst-2
Purified bulk neurons
Camk2a+
PV+
VIP+
SST+
R1
R2
R1
R2
R1
R2
R1

Fig. S8

Fig. S8. Single neuron clusters are correlated with layer dissection and bulk methylome generated from purified neuron populations. (A) Single neurons isolated from undissected and dissected superficial, middle and deep layer frontal cortex tissue are separately visualized using tSNE. (B) Enrichment/depletion of cells from mouse dissected superficial, middle and deep cortical layers in neuron clusters. (C) Immunologically (NeuN+) and genetically labeled (Exc - Camk2a+, PV - PV+, VIP - VIP+, SST - SST+) neuron populations are co-clustered and visualized together with mouse single neurons using tSNE. (D) Consistent mCH profiles between bulk methylome and aggregated single neuron methylomes for nonoverlapping 100 kb bins across the mouse genome. Bulk methylome generated from purified mouse neuronal populations (VIP+, PV+ and SST+) were compared with aggregated single neurons methylomes of corresponding clusters (mVip, mPv and mSst-1). (E) Browser tracks showing concordance of snmC-seq data pooled from neuronal cell type clusters (top tracks) with bulk DNA methylation profiling of purified neuronal cell types. Arrows indicate corresponding single cell clusters and bulk cell types with low methylation levels at these cell-type specific loci.

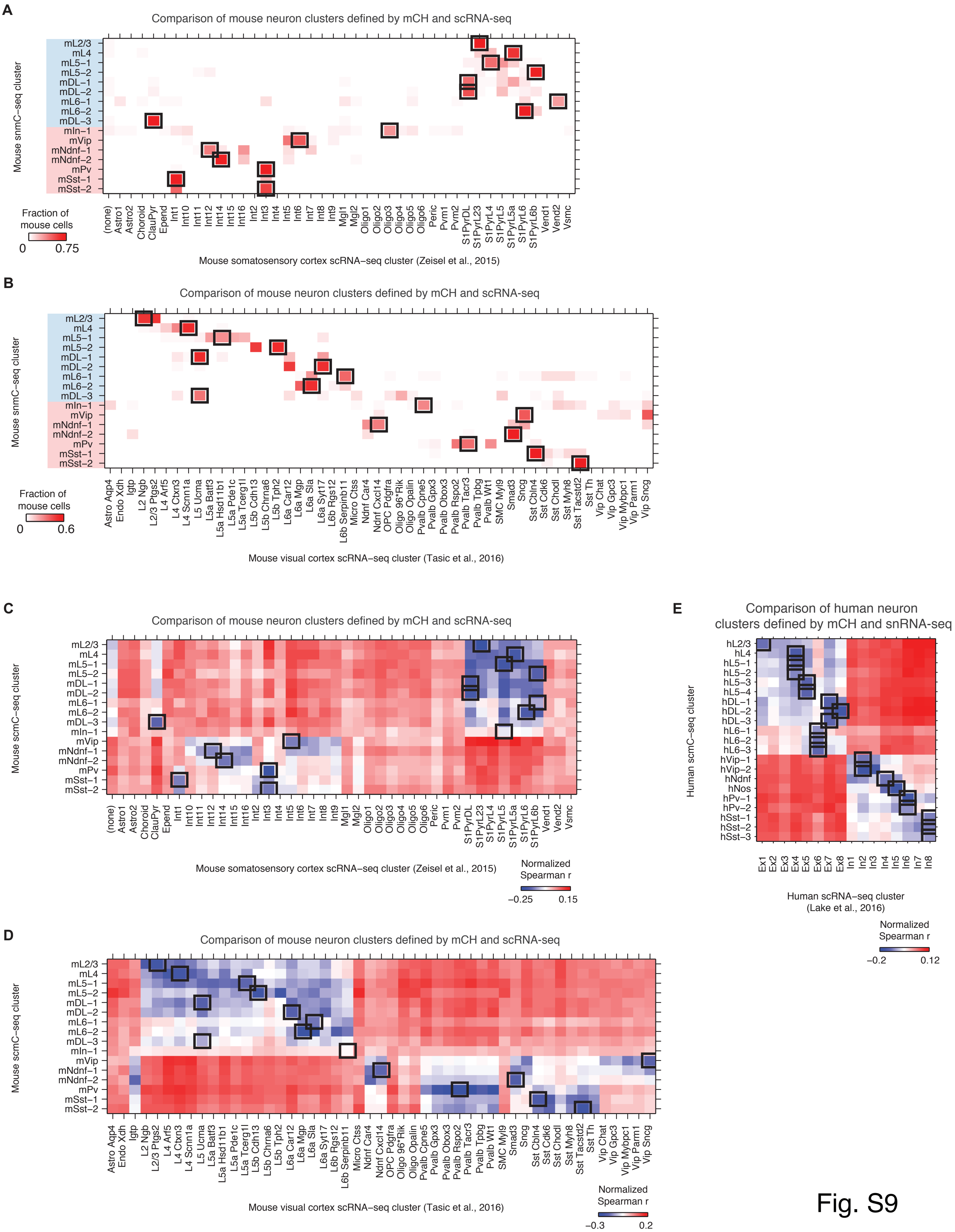

A
Comparison of mouse neuron clusters defined by mCH and scRNA-seq
Mouse snmC-seq cluster
mL2/3
mL4
mL5−1
mL5−2
mDL−1
mDL−2
mL6−1
mL6−2
mDL−3
mIn−1
mVip
mNdnf−1
mNdnf−2
mPv
mSst−1
mSst−2
(none)
Astro1
Astro2
Choroid
ClauPyr
Epend
Int1
Int10
Int11
Int12
Int14
Int15
Int16
Int2
Int3
Int4
Int5
Int6
Int7
Int8
Int9
Mgl1
Mgl2
Oligo1
Oligo2
Oligo3
Oligo4
Oligo5
Oligo6
Peric
Pvm1
Pvm2
S1PyrDL
S1PyrL23
S1PyrL4
S1PyrL5
S1PyrL5a
S1PyrL6
S1PyrL6b
Vend1
Vend2
Vsmc
Fraction of mouse cells
0
0.75
Mouse somatosensory cortex scRNA−seq cluster (Zeisel et al., 2015)
B
Comparison of mouse neuron clusters defined by mCH and scRNA-seq
Mouse snmC−seq cluster
mL2/3
mL4
mL5−1
mL5−2
mDL−1
mDL−2
mL6−1
mL6−2
mDL−3
mIn−1
mVip
mNdnf−1
mNdnf−2
mPv
mSst−1
mSst−2
Astro Aqp4
Endo Xdh
Igtp
L2 Ngb
L2/3 Ptgs2
L4 Arf5
L4 Ctxn3
L4 Scnn1a
L5 Ucma
L5a Batf3
L5a Hsd11b1
L5a Pde1c
L5a Tcerg1l
L5b Cdh13
L5b Chrna6
L5b Tph2
L6a Car12
L6a Mgp
L6a Sla
L6a Syt17
L6b Rgs12
L6b Serpinb11
Micro Ctss
Ndnf Car4
Ndnf Cxcl14
OPC Pdgfra
Oligo 96*Rik
Oligo Opalin
Pvalb Cpne5
Pvalb Gpx3
Pvalb Obox3
Pvalb Rspo2
Pvalb Tacr3
Pvalb Tpbg
Pvalb Wt1
SMC Myl9
Smad3
Sncg
Sst Cbln4
Sst Cdk6
Sst Chodl
Sst Myh8
Sst Tacstd2
Sst Th
Vip Chat
Vip Gpc3
Vip Mybpc1
Vip Parm1
Vip Sncg
Fraction of mouse cells
0
0.6
Mouse visual cortex scRNA−seq cluster (Tasic et al., 2016)
C
Comparison of mouse neuron clusters defined by mCH and scRNA-seq
Mouse scmC−seq cluster
mL2/3
mL4
mL5−1
mL5−2
mDL−1
mDL−2
mL6−1
mL6−2
mDL−3
mIn−1
mVip
mNdnf−1
mNdnf−2
mPv
mSst−1
mSst−2
(none)
Astro1
Astro2
Choroid
ClauPyr
Epend
Int1
Int10
Int11
Int12
Int14
Int15
Int16
Int2
Int3
Int4
Int5
Int6
Int7
Int8
Int9
Mgl1
Mgl2
Oligo1
Oligo2
Oligo3
Oligo4
Oligo5
Oligo6
Peric
Pvm1
Pvm2
S1PyrDL
S1PyrL23
S1PyrL4
S1PyrL5
S1PyrL5a
S1PyrL6
S1PyrL6b
Vend1
Vend2
Vsmc
Mouse somatosensory cortex scRNA−seq cluster (Zeisel et al., 2015)
Normalized Spearman r
−0.25
0.15
D
Comparison of mouse neuron clusters defined by mCH and scRNA-seq
Mouse scmC−seq cluster
mL2/3
mL4
mL5−1
mL5−2
mDL−1
mDL−2
mL6−1
mL6−2
mDL−3
mIn−1
mVip
mNdnf−1
mNdnf−2
mPv
mSst−1
mSst−2
Astro Aqp4
Endo Xdh
Igtp
L2 Ngb
L2/3 Ptgs2
L4 Arf5
L4 Ctxn3
L4 Scnn1a
L5 Ucma
L5a Batf3
L5a Hsd11b1
L5a Pde1c
L5a Tcerg1l
L5b Cdh13
L5b Chrna6
L5b Tph2
L6a Car12
L6a Mgp
L6a Sla
L6a Syt17
L6b Rgs12
L6b Serpinb11
Micro Ctss
Ndnf Car4
Ndnf Cxcl14
OPC Pdgfra
Oligo 96*Rik
Oligo Opalin
Pvalb Cpne5
Pvalb Gpx3
Pvalb Obox3
Pvalb Rspo2
Pvalb Tacr3
Pvalb Tpbg
Pvalb Wt1
SMC Myl9
Smad3
Sncg
Sst Cbln4
Sst Cdk6
Sst Chodl
Sst Myh8
Sst Tacstd2
Sst Th
Vip Chat
Vip Gpc3
Vip Mybpc1
Vip Parm1
Vip Sncg
Mouse visual cortex scRNA−seq cluster (Tasic et al., 2016)
Normalized Spearman r
−0.3
0.2
E
Comparison of human neuron clusters defined by mCH and snRNA-seq
Human scmC−seq cluster
hL2/3
hL4
hL5−1
hL5−2
hL5−3
hL5−4
hDL−1
hDL−2
hDL−3
hL6−1
hL6−2
hL6−3
hVip−1
hVip−2
hNdnf
hNos
hPv−1
hPv−2
hSst−1
hSst−2
hSst−3
Ex1
Ex2
Ex3
Ex4
Ex5
Ex6
Ex7
Ex8
In1
In2
In3
In4
In5
In6
In7
In8
Human scRNA−seq cluster (Lake et al., 2016)
Normalized Spearman r
−0.2
0.12

Fig. S9

Fig. S9. Correlation between single neuron clusters defined by snmC-seq and single cell/nucleus RNA-seq. Comparison of mouse neuron clusters to mouse somatosensory cortex single cell clusters defined by scRNA-seq (A, (2)) and mouse visual cortex single cell clusters defined by scRNA-seq (B, (3)). For (A) and (B), color represents the fraction of cells in each snmC-seq cluster having the best match (strongest inverse correlation) for an RNA-seq cluster. The best RNA-seq cluster match for each of snmC-seq clusters was highlighted with a black rectangle. (C-E) Normalized Spearman correlation coefficients between gene body mCH level in snmC-seq clusters and median transcript abundance (TPM) of RNA-seq clusters. Mouse snmC-seq clusters were compared to mouse somatosensory cortex single cell clusters defined by scRNA-seq (C, (2)) and mouse visual cortex single cell clusters defined by scRNA-seq (D, (3)). (E) Human snmC-seq clusters were compared with human cortical neuron clusters defined by snRNA-seq (4). Spearman correlation coefficients were normalized by subtracting the mean value of each row in the matrix.

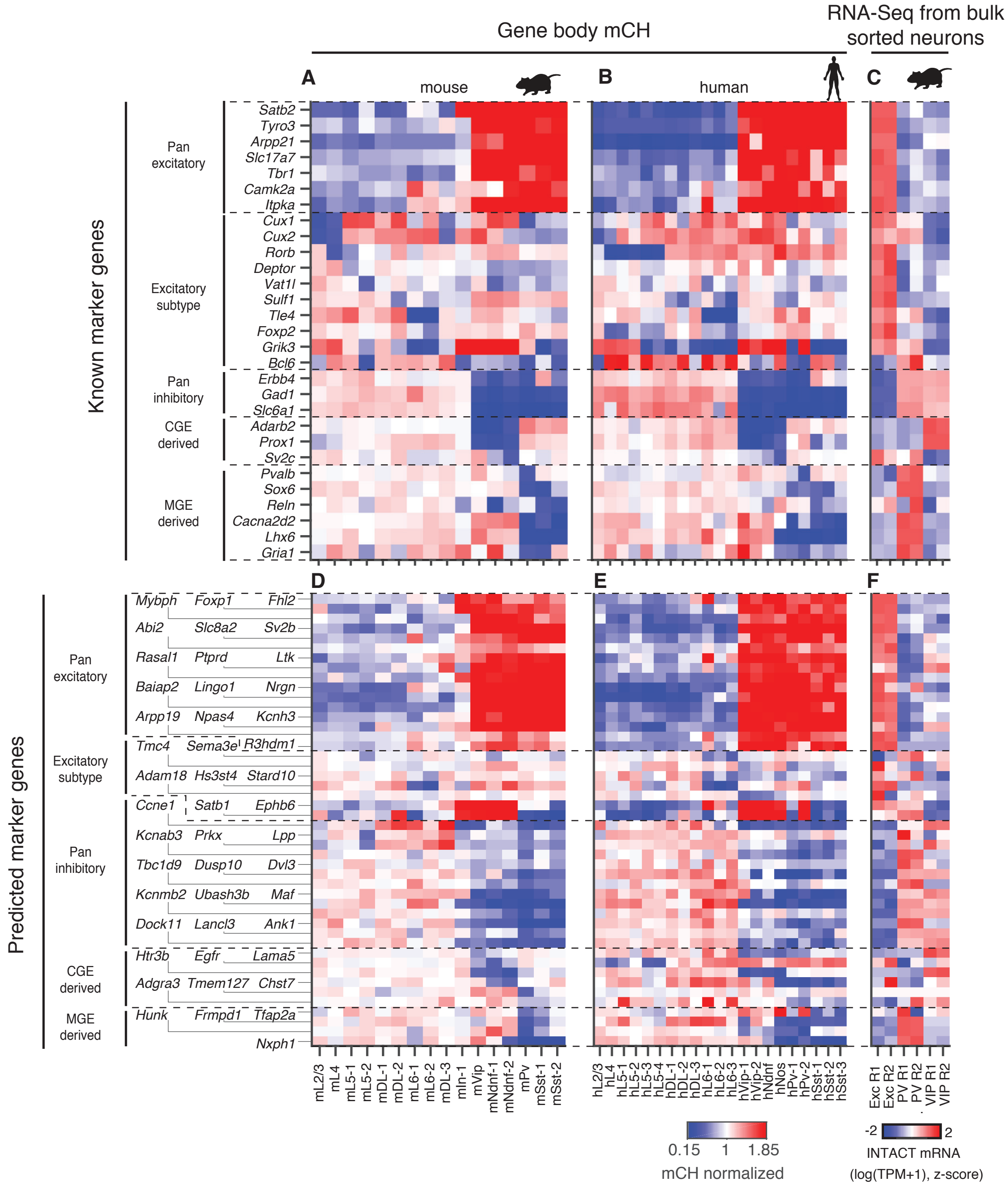


Fig. S10. Prediction of neuron type marker genes with single cell methylomes. (A-B) mCH level of known markers for mouse (A) and human (B) neuron clusters. (C) Gene expression of known markers for Camk2a+ (Exc), PV+ and VIP+ neuron populations. (D-E) mCH level of newly predicted markers for mouse (D) and human (E) neuron clusters. (F) Gene expression of newly predicted markers for Camk2a+, PV+ and VIP+ neuron populations.

Fig. S10

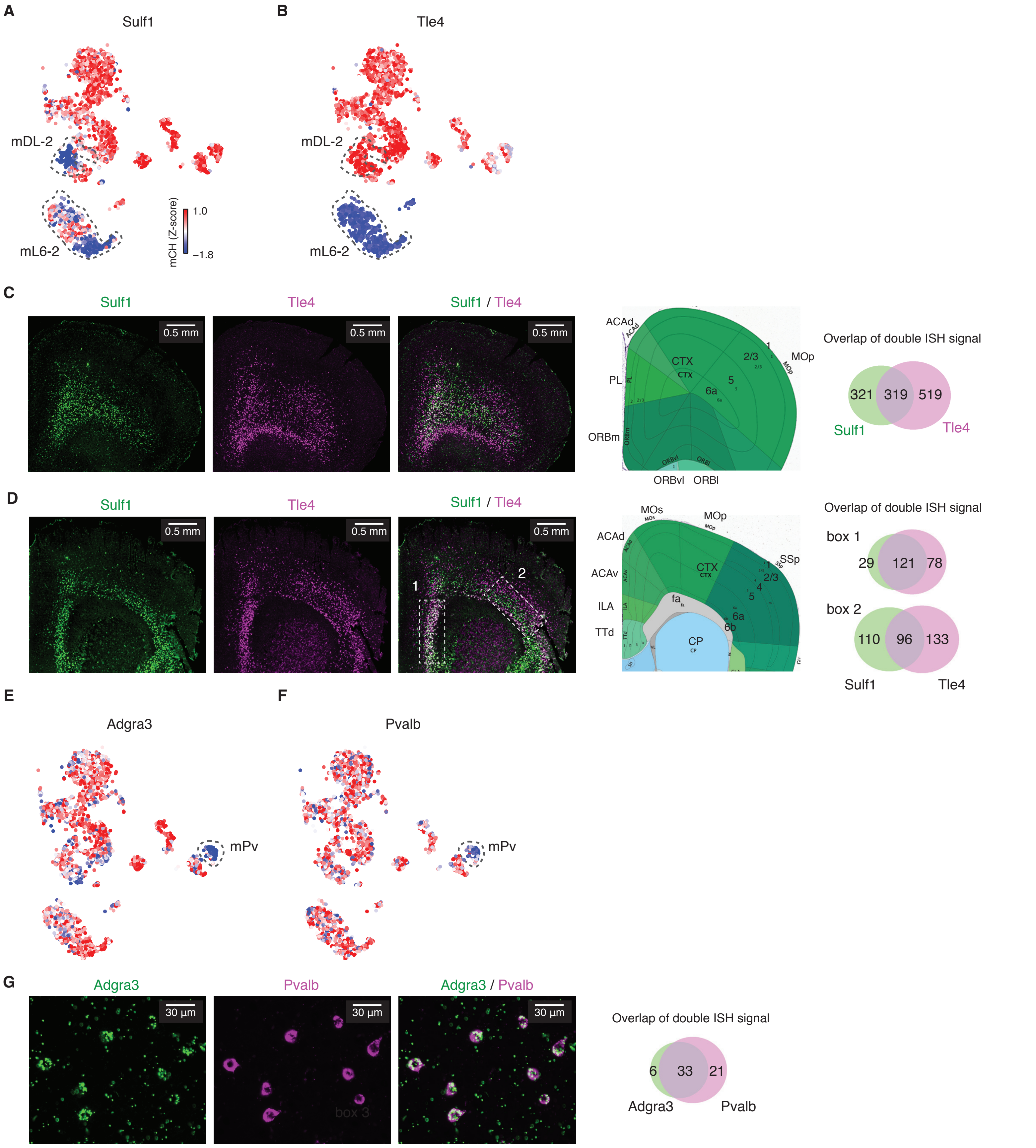


Fig. S11. Double ISH experiments validate novel markers predicted by mCH. (A-B) Relative mCH level (mCH Z-score) of Sulf1 and Tle4. The z-score is defined as the mCH value minus its mean over all cells, divided by the standard deviation across cells. (C-D) Double in situ RNA hybridization results using probes for Sulf1 and Tle4 in mouse FC. (C) and (D) show two coronal sections both in mouse FC with (C) located more rostral than (D). (E-F) Relative mCH level (mCH Z-score) of Adgra3 and Pvalb. (G) Double in situ RNA hybridization results using probes for Adgra3 and Pvalb.

Fig. S11

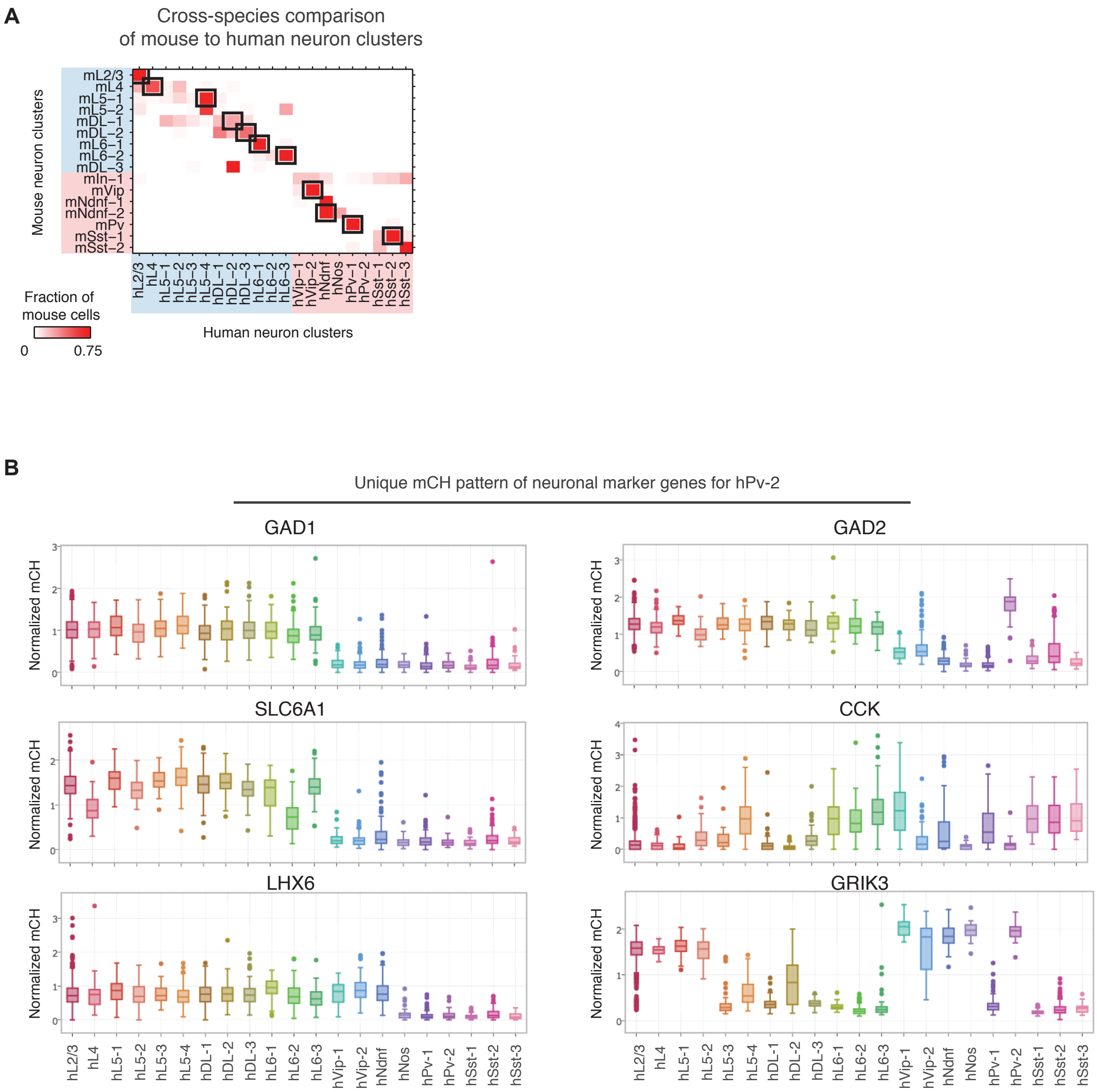


Fig. S12. Expanded neuronal diversity in human FC. (A) Cross-species cluster similarity computed by comparing mouse to human clusters. Color indicates the fraction of neurons in mouse cluster showing strongest correlation (Spearman correlation at homologous gene bodies) with each human cluster. Human and mouse cluster pairs that are mutual best matches are highlighted with black rectangles. (B) hPv-2 shows unique mCH pattern of neuronal marker genes. Boxplots show the distribution of gene body mCH of individual single human neurons, normalized by dividing gene body mCH by global mCH for each cell.

Fig. S12

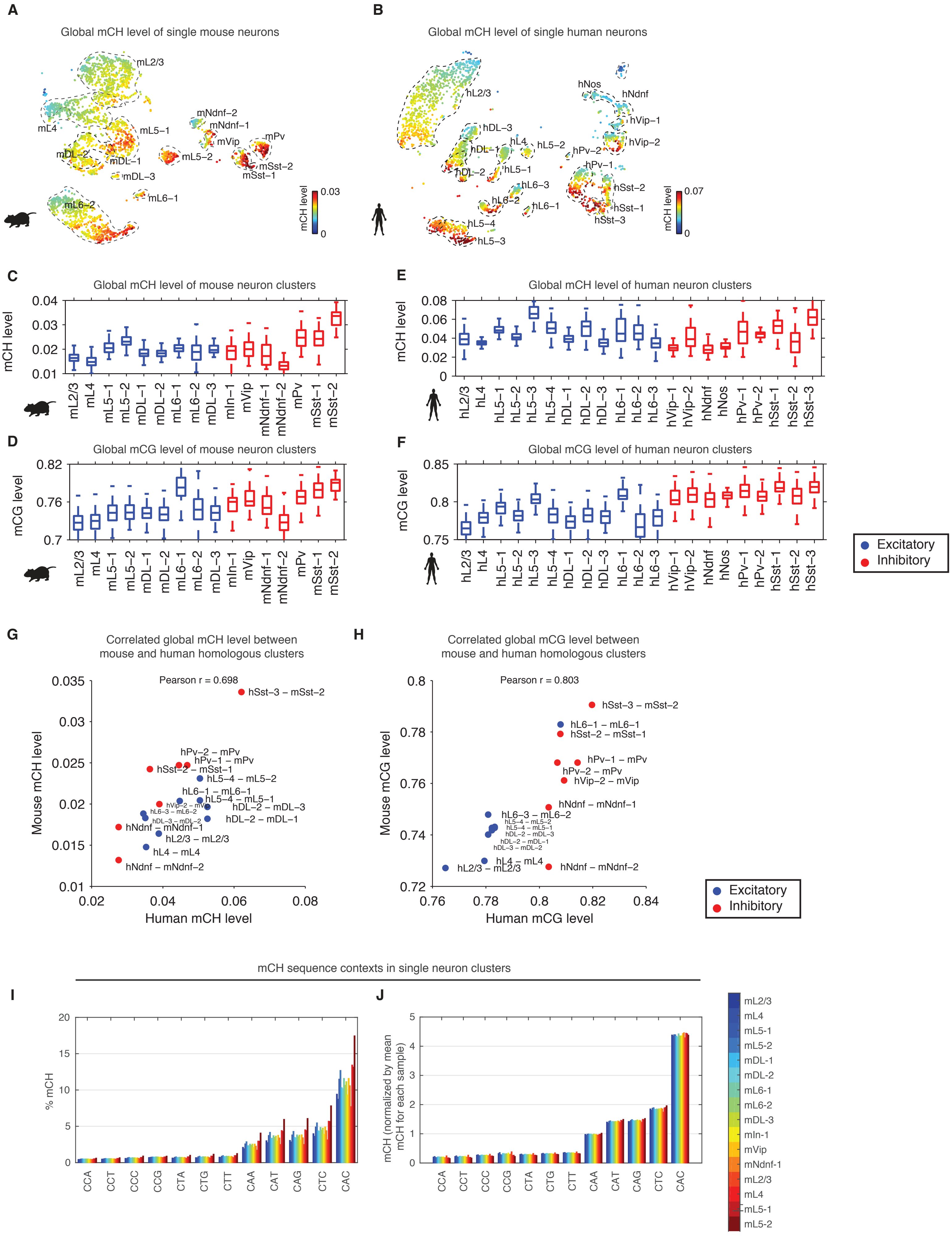

A
Global mCH level of single mouse neurons
mL2/3
mL4
mL5−1
mDL−2
mDL−1
mDL−3
mL6−1
mL6−2
mNdnf−2
mNdnf−1
mVip
mL5−2
mPv
mSst−2
mSst−1
mCH level
0.03
0
B
Global mCH level of single human neurons
hL2/3
hNos
hNdnf
hVip−1
hVip−2
hDL−3
hDL−1
hL4
hL5−2
hPv−2
hPv−1
hDL−2
hL5−1
hL6−3
hL6−2
hL6−1
hSst−2
hSst−1
hSst−3
hL5−4
hL5−3
mCH level
0.07
0
C
Global mCH level of mouse neuron clusters
mCH level
0.04
0.03
0.02
0.01
mL2/3
mL4
mL5−1
mL5−2
mDL−1
mDL−2
mL6−1
mL6−2
mDL−3
mIn−1
mVip
mNdnf−1
mNdnf−2
mPv
mSst−1
mSst−2
E
Global mCH level of human neuron clusters
mCH level
0.08
0.06
0.04
0.02
0
hL2/3
hL4
hL5−1
hL5−2
hL5−3
hL5−4
hDL−1
hDL−2
hDL−3
hL6−1
hL6−2
hL6−3
hVip−1
hVip−2
hNdnf
hNos
hPv−1
hPv−2
hSst−1
hSst−2
hSst−3
D
Global mCG level of mouse neuron clusters
mCG level
0.82
0.76
0.7
mL2/3
mL4
mL5−1
mL5−2
mDL−1
mDL−2
mL6−1
mL6−2
mDL−3
mIn−1
mVip
mNdnf−1
mNdnf−2
mPv
mSst−1
mSst−2
F
Global mCG level of human neuron clusters
mCG level
0.85
0.8
0.75
hL2/3
hL4
hL5−1
hL5−2
hL5−3
hL5−4
hDL−1
hDL−2
hDL−3
hL6−1
hL6−2
hL6−3
hVip−1
hVip−2
hNdnf
hNos
hPv−1
hPv−2
hSst−1
hSst−2
hSst−3
● Excitatory
● Inhibitory
G
Correlated global mCH level between
mouse and human homologous clusters
Pearson r = 0.698
hSst−3 − mSst−2
hPv−2 − mPv
hPv−1 − mPv
hSst−2 − mSst−1
hL5−4 − mL5−2
hL6−1 − mL6−1
hL5−4 − mL5−1
hVip−2 − mVip
hL6−3 − mL6−2
hDL−2 − mDL−3
hDL−3 − mDL−2
hDL−2 − mDL−1
hNdnf − mNdnf−1
hL2/3 − mL2/3
hL4 − mL4
hNdnf − mNdnf−2
Mouse mCH level
0.035
0.03
0.025
0.02
0.015
0.01
0.02
0.04
0.06
0.08
Human mCH level
H
Correlated global mCG level between
mouse and human homologous clusters
Pearson r = 0.803
hSst−3 − mSst−2
hL6−1 − mL6−1
hSst−2 − mSst−1
hPv−1 − mPv
hPv−2 − mPv
hVip−2 − mVip
hNdnf − mNdnf−1
hL6−3 − mL6−2
hL5−4 − mL5−2
hL5−4 − mL5−1
hDL−2 − mDL−3
hDL−2 − mDL−1
hDL−3 − mDL−2
hL4 − mL4
hL2/3 − mL2/3
hNdnf − mNdnf−2
Mouse mCG level
0.8
0.78
0.76
0.74
0.72
0.76
0.78
0.8
0.82
0.84
Human mCG level
● Excitatory
● Inhibitory
mCH sequence contexts in single neuron clusters
I
% mCH
20
15
10
5
0
CCA
CCT
CCC
CCG
CTA
CTG
CTT
CAA
CAT
CAG
CTC
CAC
J
mCH (normalized by mean
mCH for each sample)
5
4
3
2
1
0
CCA
CCT
CCC
CCG
CTA
CTG
CTT
CAA
CAT
CAG
CTC
CAC
mL2/3
mL4
mL5-1
mL5-2
mDL-1
mDL-2
mL6-1
mL6-2
mDL-3
mIn-1
mVip
mNdnf-1
mL2/3
mL4
mL5-1
mL5-2

Fig.S13

Fig. S13. Global mC levels are conserved between mouse and human neuron types. (A and B) Global mCH level for single mouse (A) and human (B) neurons. (C-D) Genome-wide mCH (C) and mCG (D) levels for mouse neuron clusters. (E-F) Genome-wide mCH (E) and mCG (F) levels for human neuron clusters. (G-H) Cross-species comparison of genome-wide mCH and mCG level between homologous clusters. (I) Percentage of mCH basecalls located in each trinucleotide context for mouse neuron clusters. (J) Normalized mCH level of each trinucleotide context for mouse neuron clusters.

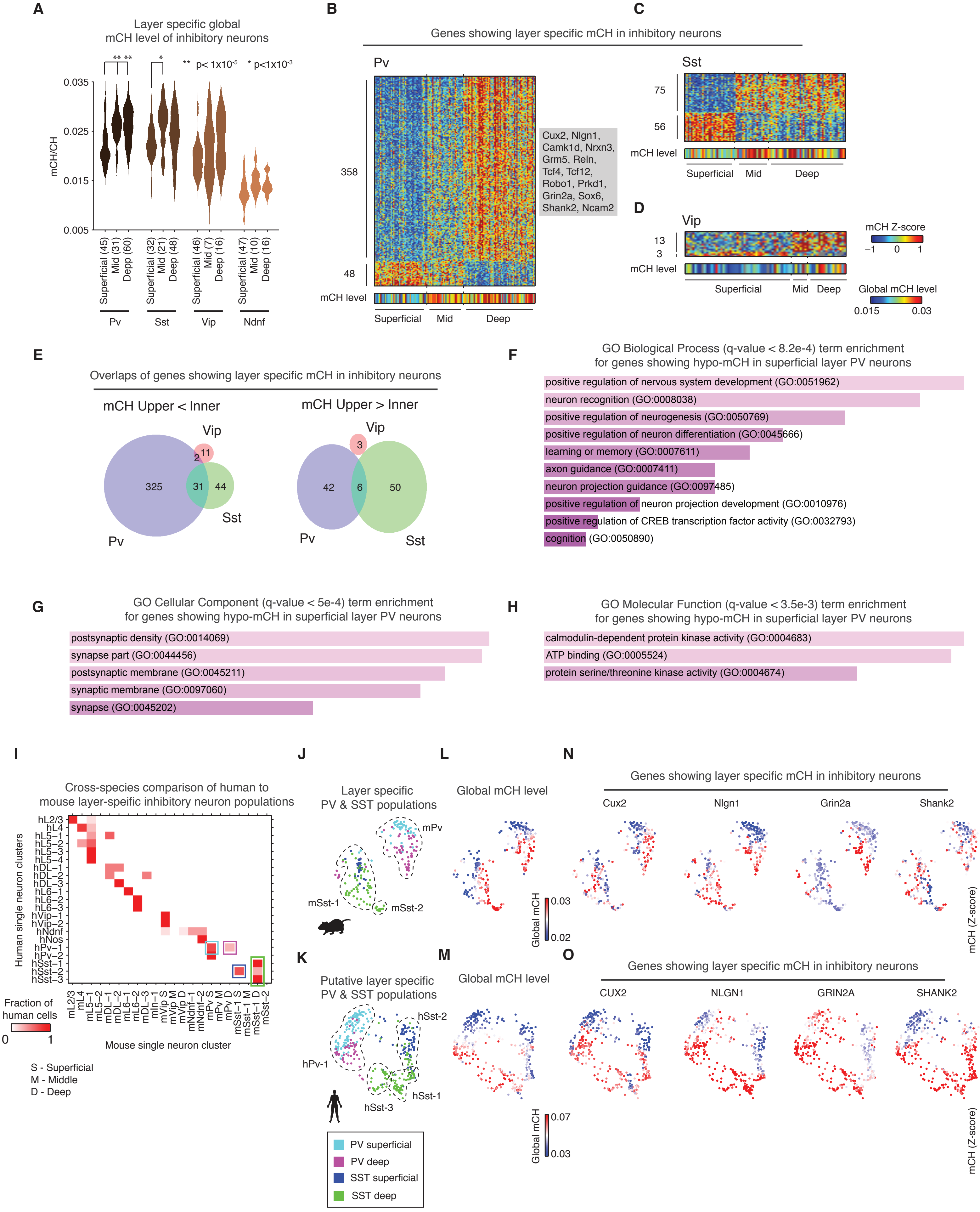
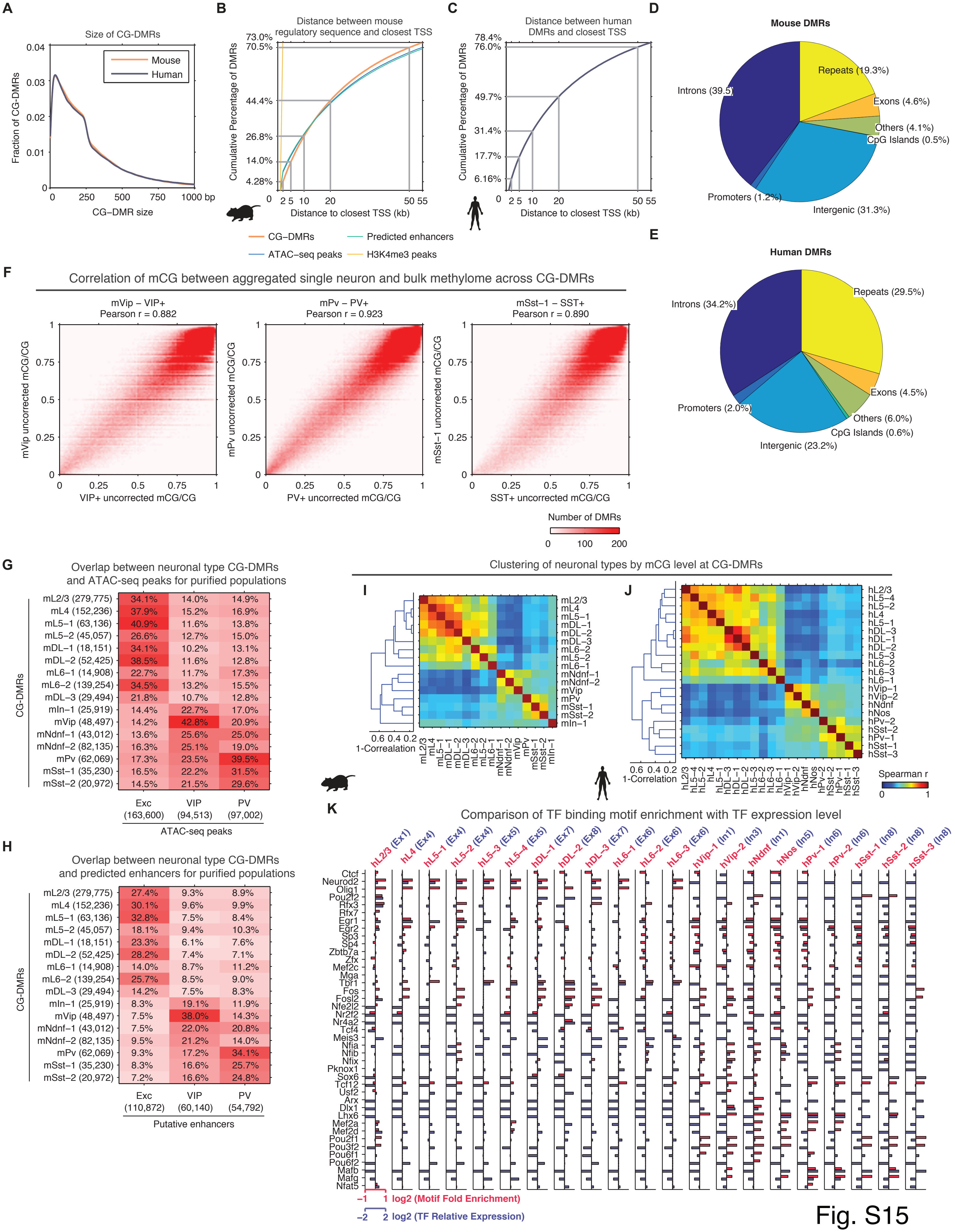

A
Layer specific global
mCH level of inhibitory neurons
** **
*
** p< 1x10^-5 * p<1x10^-3
0.035
0.025
0.015
0.005
mCH/CH
Superficial (45)
Mid (31)
Depp (60)
Superficial (32)
Mid (21)
Deep (48)
Superficial (46)
Mid (7)
Deep (16)
Superficial (47)
Mid (10)
Deep (16)
Pv
Sst
Vip
Ndnf
B
C
Genes showing layer specific mCH in inhibitory neurons
Pv
358
48
mCH level
Superficial
Mid
Deep
Cux2, Nlgn1,
Camk1d, Nrxn3,
Grm5, Reln,
Tcf4, Tcf12,
Robo1, Prkd1,
Grin2a, Sox6,
Shank2, Ncam2
Sst
75
56
mCH level
Superficial
Mid
Deep
D
Vip
13
3
mCH level
Superficial
Mid
Deep
mCH Z-score
−1
0
1
Global mCH level
0.015
0.03
E
Overlaps of genes showing layer specific mCH in inhibitory neurons
mCH Upper < Inner
Vip
2
11
325
31
44
Sst
Pv
mCH Upper > Inner
Vip
3
42
6
50
Pv
Sst
F
GO Biological Process (q-value < 8.2e-4) term enrichment
for genes showing hypo-mCH in superficial layer PV neurons
positive regulation of nervous system development (GO:0051962)
neuron recognition (GO:0008038)
positive regulation of neurogenesis (GO:0050769)
positive regulation of neuron differentiation (GO:0045666)
learning or memory (GO:0007611)
axon guidance (GO:0007411)
neuron projection guidance (GO:0097485)
positive regulation of neuron projection development (GO:0010976)
positive regulation of CREB transcription factor activity (GO:0032793)
cognition (GO:0050890)
G
GO Cellular Component (q-value < 5e-4) term enrichment
for genes showing hypo-mCH in superficial layer PV neurons
postsynaptic density (GO:0014069)
synapse part (GO:0044456)
postsynaptic membrane (GO:0045211)
synaptic membrane (GO:0097060)
synapse (GO:0045202)
H
GO Molecular Function (q-value < 3.5e-3) term enrichment
for genes showing hypo-mCH in superficial layer PV neurons
calmodulin-dependent protein kinase activity (GO:0004683)
ATP binding (GO:0005524)
protein serine/threonine kinase activity (GO:0004674)
I
Cross-species comparison of human to
mouse layer-speific inhibitory neuron populations
hL2/3
hL4
hL5−1
hL5−2
hL5−3
hL5−4
hDL−1
hDL−2
hDL−3
hL6−1
hL6−2
hL6−3
hVip−1
hVip−2
hNdnf
hNos
hPv−1
hPv−2
hSst−1
hSst−2
hSst−3
Human single neuron clusters
mL2/3
mL4
mL5−1
mL5−2
mDL−1
mDL−2
mL6−1
mL6−2
mDL−3
mIn−1
mVip S
mVip M
mVip D
mNdnf−1
mNdnf−2
mPv S
mPv M
mPv D
mSst−1 S
mSst−1 M
mSst−1 D
mSst−2
Mouse single neuron cluster
Fraction of
human cells
0
1
S - Superficial
M - Middle
D - Deep
J
Layer specific
PV & SST populations
mPv
mSst-1
mSst-2
K
Putative layer specific
PV & SST populations
hSst-2
hPv-1
hSst-1
hSst-3
PV superficial
PV deep
SST superficial
SST deep
L
Global mCH level
Global mCH
0.03
0.02
M
Global mCH level
Global mCH
0.07
0.03
N
Genes showing layer specific mCH in inhibitory neurons
Cux2
Nlgn1
Grin2a
Shank2
mCH (Z-score)
1.4
-1.4
O
Genes showing layer specific mCH in inhibitory neurons
CUX2
NLGN1
GRIN2A
SHANK2
mCH (Z-score)
1.2
-1.2

Fig. S14

Fig. S14. Inhibitory neurons possess global and gene level cortical-layer-specific mCH signatures. (A) Global mCH level shows layer-specific differences in PV and SST neurons. (B-D) A subset of genes show layer-specific gene body mCH in inhibitory neurons. (E) Overlap of genes showing layer-specific mCH in PV and SST neurons. (F-H) Gene ontology term enrichment for genes showing hypo-mCH in PV neurons located in superficial layer. (I) Cross species similarity computed by comparing human to mouse clusters, with mouse inhibitory clusters divided into sub-clusters containing neurons located in dissected superficial, middle and deep cortical layers. Cyan rectangle indicates human neuron showing strongest correlation to mouse superficial layer PV neurons, magenta rectangle indicates correlation to mouse deep layer PV neurons, blue rectangle indicates correlation to mouse superficial SST neurons, and green rectangle indicates correlation to mouse deep SST neurons. (J) tSNE visualization of mouse PV and SST neurons located in superficial or deep layers. Colors indicate the cell type and layer for each cell based on layer-specific dissections. (K) tSNE visualization of human PV and SST neurons that were putatively located in superficial or deep layers. (L,M) Global mCH of mouse (L) and human (M) PV and SST neurons. (N,O) mCH level of mouse (N) and human (O) genes showing layer-specific mCH in PV and SST neurons. The z-score is defined as the mCH value minus its mean over all cells, divided by the standard deviation across cells.



CG-DMRs	Exc (163,600)	VIP (94,513)	PV (97,002)
mL2/3 (279,775)	34.1%	14.0%	14.9%
mL4 (152,236)	37.9%	15.2%	16.9%
mL5−1 (63,136)	40.9%	11.6%	13.8%
mL5−2 (45,057)	26.6%	12.7%	15.0%
mDL−1 (18,151)	34.1%	10.2%	13.1%
mDL−2 (52,425)	38.5%	11.6%	12.8%
mL6−1 (14,908)	22.7%	11.7%	17.3%
mL6−2 (139,254)	34.5%	13.2%	15.5%
mDL−3 (29,494)	21.8%	10.7%	12.8%
mIn−1 (25,919)	14.4%	22.7%	17.0%
mVip (48,497)	14.2%	42.8%	20.9%
mNdnf−1 (43,012)	13.6%	25.6%	25.0%
mNdnf−2 (82,135)	16.3%	25.1%	19.0%
mPv (62,069)	17.3%	23.5%	39.5%
mSst−1 (35,230)	16.5%	22.2%	31.5%
mSst−2 (20,972)	14.5%	21.5%	29.6%

CG-DMRs	Exc (110,872)	VIP (60,140)	PV (54,792)
mL2/3 (279,775)	27.4%	9.3%	8.9%
mL4 (152,236)	30.1%	9.6%	9.9%
mL5−1 (63,136)	32.8%	7.5%	8.4%
mL5−2 (45,057)	18.1%	9.4%	10.3%
mDL−1 (18,151)	23.3%	6.1%	7.6%
mDL−2 (52,425)	28.2%	7.4%	7.1%
mL6−1 (14,908)	14.0%	8.7%	11.2%
mL6−2 (139,254)	25.7%	8.5%	9.0%
mDL−3 (29,494)	14.2%	7.5%	8.3%
mIn−1 (25,919)	8.3%	19.1%	11.9%
mVip (48,497)	7.5%	38.0%	14.3%
mNdnf−1 (43,012)	7.5%	22.0%	20.8%
mNdnf−2 (82,135)	9.5%	21.2%	14.0%
mPv (62,069)	9.3%	17.2%	34.1%
mSst−1 (35,230)	8.3%	16.6%	25.7%
mSst−2 (20,972)	7.2%	16.6%	24.8%

Fig. S15

Fig. S15. Neuron-type-specific CG-DMRs reveal regulatory diversity in human and mouse brains. (A) Distribution of CG-DMR size. Note that the DMR calling software (methylpy) merges CG positions spaced closer than 250 bp to call DMRs, which accounts for the drop in the frequency of DMRs around 250 bp. (B) Distance between closest TSS and mouse regulatory sequences defined by CG-DMRs, enhancers predicted by Regulatory Element Prediction based on Tissue-specific Local Epigenetic marks (REPTILE, (20)), ATAC-seq peaks and H3K4me3 peaks. The curve shows the cumulative percentage of DMRs within a certain distance to the closest TSS. (C) Distance between human CG-DMRs and closest TSS. (D-E) Distribution of mouse (D) and human (E) CG-DMRs in genomic compartments. (F) Consistent mCG across CG-DMRs between bulk methylome and aggregated single neuron methylomes. Bulk methylome generated from purified mouse neuronal populations (VIP+, PV+ and SST+) were compared with aggregated single neurons methylomes of corresponding clusters (mVip, mPv and mSst-1). (G) Overlap between neuron-type-specific CG-DMRs and ATAC-seq peaks identified from purified neuronal populations. (H) Overlap between neuron-type-specific CG-DMRs and putative enhancers predicted from purified neuronal populations. For (G) and (H), the percentage of row features overlapping with column features was shown. (I-J) Hierarchical clustering of neuron clusters by mCG level at CG-DMRs. (K) Comparison of TF binding motif enrichment with TF expression level across human neuron clusters. Median TF expression of the best matching snRNA-seq cluster (indicated on the top) identified in (4) for each snmC-seq clusters was shown.

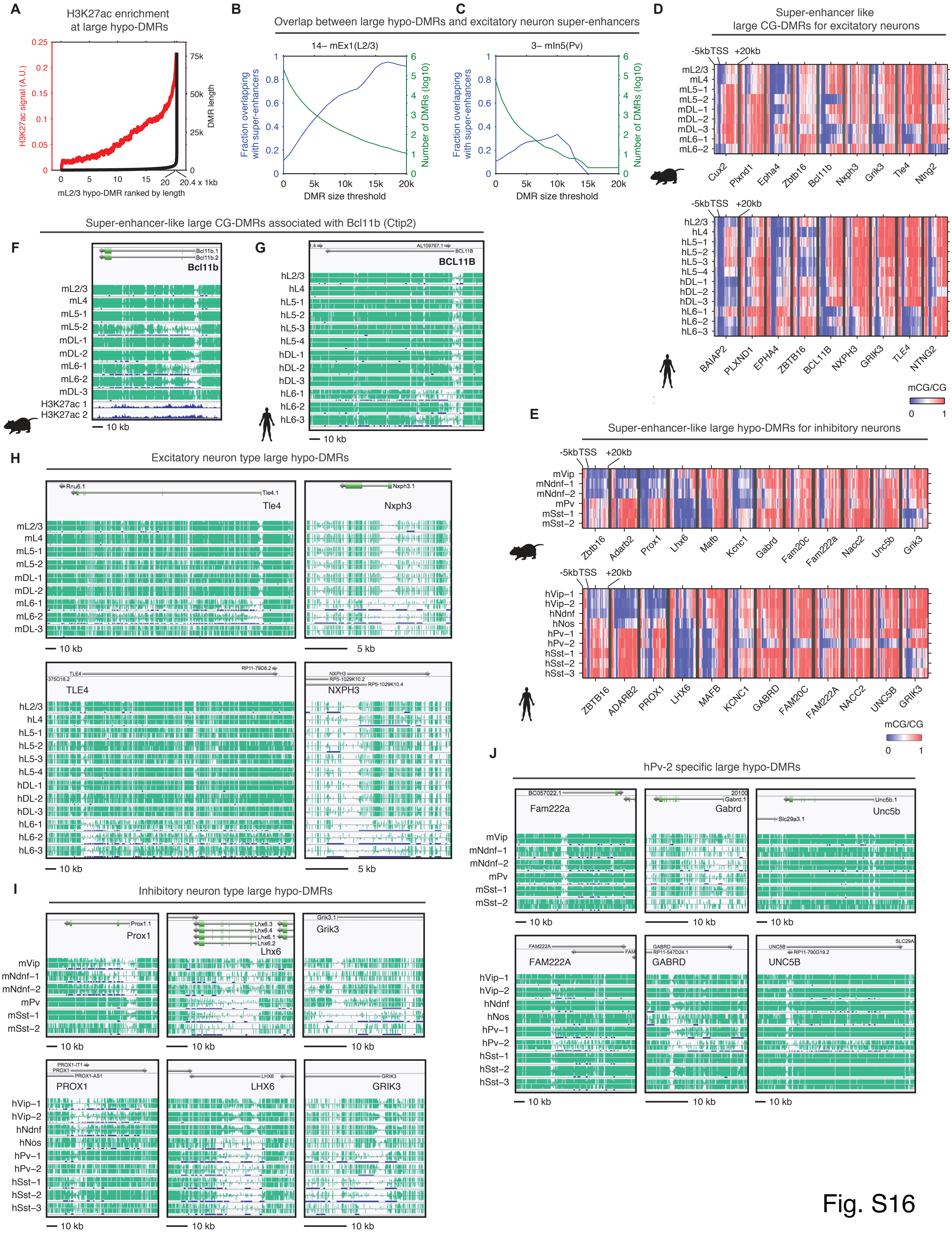

A
H3K27ac enrichment at large hypo-DMRs
H3K27ac signal (A.U.)
0.25
0.2
0.15
0.1
0.05
0
DMR length
75k
50k
25k
0
0
5
10
15
20
20.4 x 1kb
mL2/3 hypo-DMR ranked by length
B
C
Overlap between large hypo-DMRs and excitatory neuron super-enhancers
14– mEx1(L2/3)
3– mIn5(Pv)
Fraction overlapping with super-enhancers
Number of DMRs (log10)
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5
6
0
5k
10k
15k
20k
DMR size threshold
D
Super-enhancer like large CG-DMRs for excitatory neurons
-5kbTSS
+20kb
mL2/3
mL4
mL5–1
mL5–2
mDL–1
mDL–2
mDL–3
mL6–1
mL6–2
Cux2
Plxnd1
Epha4
Zbtb16
Bcl11b
Nxph3
Grik3
Tle4
Ntng2
hL2/3
hL4
hL5–1
hL5–2
hL5–3
hL5–4
hDL–1
hDL–2
hDL–3
hL6–1
hL6–2
hL6–3
BAIAP2
PLXND1
EPHA4
ZBTB16
BCL11B
NXPH3
GRIK3
TLE4
NTNG2
mCG/CG
0
1
Super-enhancer-like large CG-DMRs associated with Bcl11b (Ctip2)
F
Bcl11b.1
Bcl11b.2
Bcl11b
mL2/3
mL4
mL5-1
mL5-2
mDL-1
mDL-2
mL6-1
mL6-2
mDL-3
H3K27ac 1
H3K27ac 2
10 kb
G
AL109767.1
BCL11B
hL2/3
hL4
hL5-1
hL5-2
hL5-3
hL5-4
hDL-1
hDL-2
hDL-3
hL6-1
hL6-2
hL6-3
10 kb
E
Super-enhancer-like large hypo-DMRs for inhibitory neurons
-5kbTSS
+20kb
mVip
mNdnf–1
mNdnf–2
mPv
mSst–1
mSst–2
Zbtb16
Adarb2
Prox1
Lhx6
Mafb
Kcnc1
Gabrd
Fam20c
Fam222a
Nacc2
Unc5b
Grik3
hVip–1
hVip–2
hNdnf
hNos
hPv–1
hPv–2
hSst–1
hSst–2
hSst–3
ZBTB16
ADARB2
PROX1
LHX6
MAFB
KCNC1
GABRD
FAM20C
FAM222A
NACC2
UNC5B
GRIK3
mCG/CG
0
1
H
Excitatory neuron type large hypo-DMRs
Rnu6.1
Tle4.1
Tle4
Nxph3.1
Nxph3
mL2/3
mL4
mL5-1
mL5-2
mDL-1
mDL-2
mL6-1
mL6-2
mDL-3
10 kb
5 kb
TLE4
RP11-79D8.2
375O18.2
NXPH3
RP5-1029K10.2
RP5-1029K10.4
hL2/3
hL4
hL5-1
hL5-2
hL5-3
hL5-4
hDL-1
hDL-2
hDL-3
hL6-1
hL6-2
hL6-3
10 kb
5 kb
I
Inhibitory neuron type large hypo-DMRs
Prox1.1
Prox1
Lhx6.3
Lhx6.4
Lhx6.1
Lhx6.2
Lhx6
Grik3.1
Grik3
mVip
mNdnf–1
mNdnf–2
mPv
mSst–1
mSst–2
10 kb
10 kb
10 kb
PROX1-IT1
PROX1
PROX1-AS1
PROX1
LHX6
GRIK3
hVip–1
hVip–2
hNdnf
hNos
hPv–1
hPv–2
hSst–1
hSst–2
hSst–3
10 kb
10 kb
10 kb
J
hPv-2 specific large hypo-DMRs
BC057022.1
Fam222a
20100
Gabrd.1
Gabrd
Unc5b.1
Unc5b
Slc29a3.1
mVip
mNdnf–1
mNdnf–2
mPv
mSst–1
mSst–2
10 kb
10 kb
10 kb
FAM222A
FAM
FAM222A
GABRD
RP11-547D24.1
GABRD
UNC5B
RP11-790G19.2
SLC29A
UNC5B
hVip–1
hVip–2
hNdnf
hNos
hPv–1
hPv–2
hSst–1
hSst–2
hSst–3
10 kb
10 kb
10 kb

Fig. S16

Fig. S16. Identification of neuron type specific large CG-DMRs with super-enhancer like properties. (A) Average H3K27ac signal was plotted as a function of CG-DMR size. (B and C) The fraction of large CG-DMRs overlapping with putative super-enhancers was examined for different DMR size thresholds for identifying large CG-DMRs (blue line). Green line indicates the number of large CG-DMRs found with each DMR size threshold. The overlap between excitatory neuron (Camk2a+) super-enhancers and Layer 2/3 excitatory neuron and PV+ inhibitory neuron large CG-DMRs was shown in (B) and (C), respectively. (D and E) mCG levels near TSS for super-enhancer-like DMRs in excitatory (D) and inhibitory (E) neurons in mouse and human. (F,G) Large gene body CG-DMRs and H3K27ac ChIP-seq signal from mouse excitatory neurons at Bcl11b (Ctip2) locus in mouse (F) and human (G). The height of green ticks represents mC level at CG dinucleotides. (H-J) Browser view of large CG-DMRs for excitatory neurons (H), inhibitory neurons (I) and hPv-2 (J). For F-J, the height of green ticks represents mC level at CG dinucleotides.

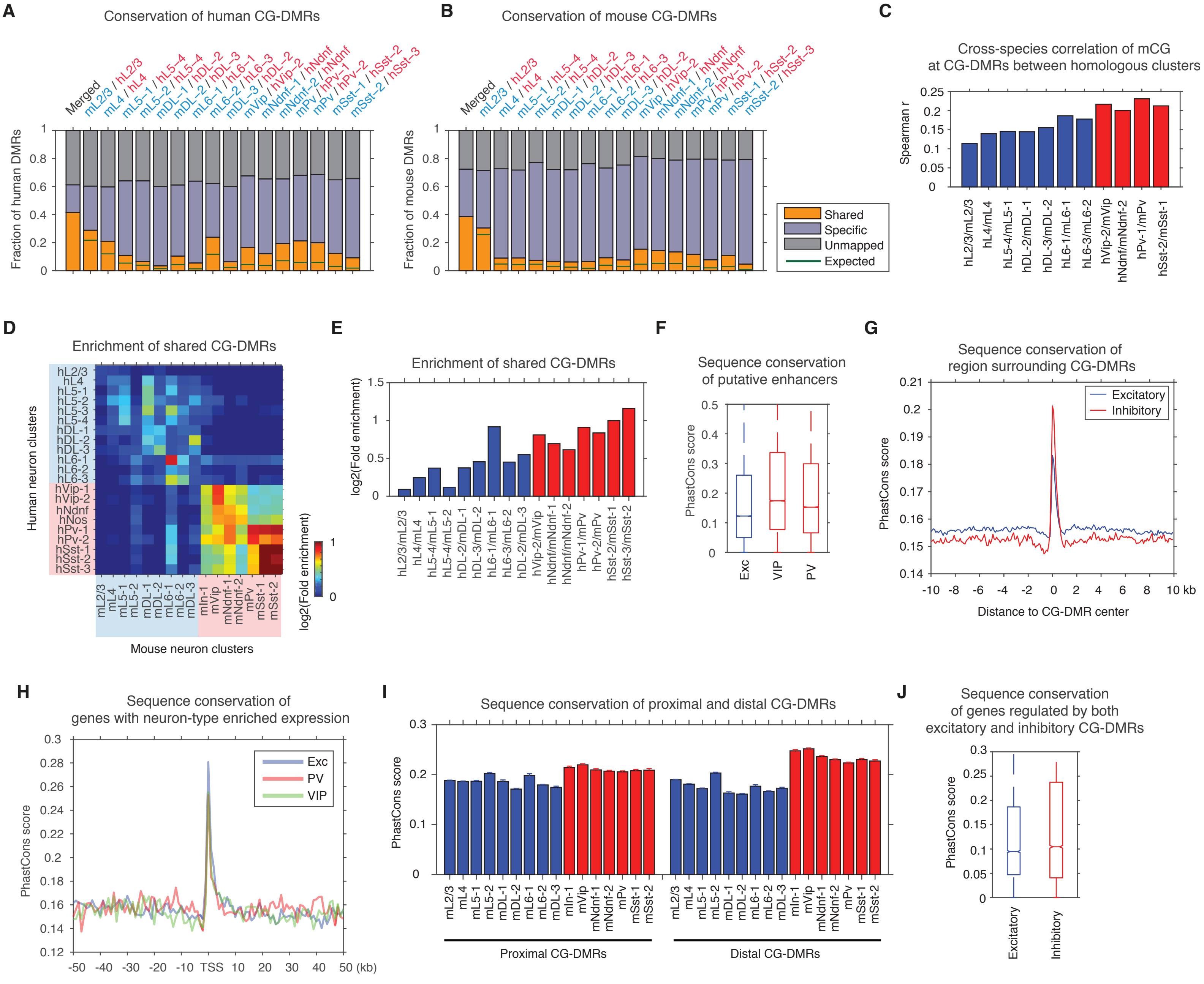


Fig. S17. Regulatory conservation of neuron types. (A and B) Fractions of human (A) and mouse (B) CG-DMRs that overlapped with CG-DMR of the homologous cluster in the other species (shared) had no overlap with CG-DMRs of the homologous cluster in other species (specific), or had no sequence homology in other species (unmapped). (C) Cross-species Spearman correlation of mCG at CG-DMRs between homologous clusters. (D-E) Enrichment of shared DMRs between all human and mouse clusters (D) and for homologous clusters (E). (F-J) Sequence conservation at putative enhancers predicted from purified neuronal populations (F), regions surrounding CG-DMRs identified in excitatory and inhibitory neuron clusters (G), genes with preferential expression in purified neuronal populations (H), proximal and distal mouse CG-DMRs (I), and excitatory and inhibitory CG-DMRs associated with the same set of genes (J).

Supplementary Tables

Table S1. Metadata of mouse single neuron methylomes (separate file)

Table S2. Metadata of human single neuron methylomes (separate file)

Table S3. Marker genes for human and mouse clusters (separate file)

Table S4. Genes showing layer-specific mCH in inhibitory neuron populations (separate file)

Table S5. List of mouse CG-DMRs (separate file)

Table S6. List of human CG-DMRs (separate file)

Table S7. List of mouse large CG-DMRs (separate file)

Table S8. List of human large CG-DMRs (separate file)

Table S9: Outline of bioinformatic analysis procedures (separate file)