

1 **DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution**

2 Hanqing Liu^{1,2*}, Jingtian Zhou^{1,3*}, Wei Tian¹, Chongyuan Luo^{1,4,5}, Anna Bartlett¹, Andrew
3 Aldridge¹, Jacinta Lucero⁶, Julia K. Osteen⁶, Joseph R. Nery¹, Huaming Chen¹, Angeline Rivkin¹,
4 Rosa G Castanon¹, Ben Clock⁷, Yang Eric Li⁸, Xiaomeng Hou⁹, Olivier B. Poirion⁹, Sebastian
5 Preissl⁹, Carolyn O'Connor¹⁰, Lara Boggeman¹⁰, Conor Fitzpatrick¹⁰, Michael Nunn¹, Eran A.
6 Mukamel¹¹, Zhuzhu Zhang¹, Edward M. Callaway¹², Bing Ren^{8,9}, Jesse R. Dixon⁷, M. Margarita
7 Behrens⁶, Joseph R. Ecker^{1,4†}

8 ¹Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037

9 ²Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093

10 ³Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA
11 92093

12 ⁴Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 N. Torrey
13 Pines Road, La Jolla, CA 92037;

14 ⁵Department of Human Genetics, University of California Los Angeles, Los Angeles, CA 90095

15 ⁶Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA,
16 92037

17 ⁷Peptide Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037

18 ⁸Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, CA 92093, USA

19 ⁹Center for Epigenomics, Department of Cellular and Molecular Medicine, Institute of Genomic
20 Medicine, Moores Cancer Center, University of California San Diego, School of Medicine, La
21 Jolla, CA, 92037

22 ¹⁰Flow Cytometry Core Facility, The Salk Institute for Biological Studies, La Jolla, CA 92037

23 ¹¹Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92037

24 ¹²Systems Neurobiology Laboratories, The Salk Institute for Biological Studies, La Jolla, CA
25 92037

26 †Correspondence: ecker@salk.edu

27 *These authors contributed equally

28 **Summary**

29 Mammalian brain cells are remarkably diverse in gene expression, anatomy, and function, yet
30 the regulatory DNA landscape underlying this extensive heterogeneity is poorly understood. We
31 carried out a comprehensive assessment of the epigenomes of mouse brain cell types by
32 applying single nucleus DNA methylation sequencing to profile 110,294 nuclei from 45 regions
33 of the mouse cortex, hippocampus, striatum, pallidum, and olfactory areas. We identified 161
34 cell clusters with distinct spatial locations and projection targets. We constructed taxonomies of
35 these epigenetic types, annotated with signature genes, regulatory elements, and transcription
36 factors. These features indicate the potential regulatory landscape supporting the assignment of
37 putative cell types, and reveal repetitive usage of regulators in excitatory and inhibitory cells for
38 determining subtypes. The DNA methylation landscape of excitatory neurons in the cortex and
39 hippocampus varied continuously along spatial gradients. Using this deep dataset, an artificial
40 neural network model was constructed that precisely predicts single neuron cell-type identity
41 and brain area spatial location. Integration of high-resolution DNA methylomes with
42 single-nucleus chromatin accessibility data allowed prediction of high-confidence
43 enhancer-gene interactions for all identified cell types, which were subsequently validated by
44 cell-type-specific chromatin conformation capture experiments. By combining multi-omic
45 datasets (DNA methylation, chromatin contacts, and open chromatin) from single nuclei and
46 annotating the regulatory genome of hundreds of cell types in the mouse brain, our DNA
47 methylation atlas establishes the epigenetic basis for neuronal diversity and spatial organization
48 throughout the mouse brain.

49 **Introduction**

50 Epigenomic dynamics is associated with cell differentiation and maturation in the mammalian
51 brain, and plays an essential role in regulating neuronal functions and animal behavior¹⁻³.
52 Cytosine DNA methylation (5mC) is a stable covalent modification that persists in post-mitotic
53 cells throughout their lifetime, and is critical for proper gene regulation^{2,4,5}. In mammalian
54 genomes, cytosine methylation predominantly occurs at CpG sites, showing dynamic patterns at
55 regulatory elements with tissue and cell-type specificity^{2,6-8}, modulating binding affinity of
56 transcription factors (TF)^{9,10}, and controlling gene transcription¹¹. As a unique signature of

57 neuronal cells, non-CpG (CH) cytosines are also abundantly methylated in the mouse and
58 human brain², which can directly affect DNA binding of MeCP2 methyl-binding protein
59 responsible for Rett Syndrome¹²⁻¹⁴. mCH levels at gene bodies are anti-correlated with gene
60 expression and show a remarkable diversity across neuronal cell types^{7,8}.

61 Deeper understanding of epigenomic diversity in the mouse brain not only provides a
62 complementary approach to transcriptome-based profiling methods to identify all brain cell types
63 but additionally allows genome-wide prediction of the regulatory DNA sequences and
64 transcriptional networks that underlie this diversity. Our previous studies demonstrated the utility
65 of studying brain cell-type and regulatory diversity using single nucleus methylome sequencing
66 (snmC-seq)^{8,15}. In this study, we use snmC-seq2¹⁶ to conduct thorough methylome profiling with
67 detailed spatial dissection in the adult postnatal day 56 (P56) mouse brain (Fig. 1a). In Li et al.
68 (Companion Manuscript # 11), the same tissue samples were also profiled in parallel using
69 single nucleus ATAC-seq (snATAC-seq) to identify genome-wide accessible chromatin¹⁷,
70 providing complementary epigenomic information to aid in cell-type specific regulatory genome
71 annotation. Moreover, to further study cis-regulatory elements and their potential targeting
72 genes across the whole genome, we applied a multi-omic assay sn-m3C-seq¹⁸ to profile
73 methylome and chromatin conformation in the same cells.

74 The deep epigenomic datasets described here provide a detailed and comprehensive census of
75 the diversity of cell-types across mouse brain regions, allowing prediction of cell-type-specific
76 regulatory elements as well as their candidate target genes and upstream TFs. Highlights of the
77 findings of our study include:

- 78 ● Single nucleus methylome sequencing of 110,294 cells from the 45 dissected regions
79 across cortex, hippocampus, striatum, pallidum, and olfactory areas of the mouse brain
80 provides the largest single-cell base-resolution DNA methylation dataset.
- 81 ● A robust single cell methylome analysis framework identifies 161 predicted mouse brain
82 subtypes. The relationship between subtypes in distinct brain areas provides novel
83 insights into anatomical cell/sub-structure associations.
- 84 ● Comparing subtype-level methylomes allows identification of 3.5M cell-type-specific
85 differentially CG methylated regions (CG-DMR) that cover ~ 50% (1,240 Mb) of the
86 mouse genome.

- 87 ● Differentially methylated TF genes and binding motifs can be associated with branches
88 in subtype phylogeny, allowing the prediction of cell-type gene regulatory programs
89 specific for each developmental lineage.
- 90 ● Integrative analysis with chromatin accessibility based cell clusters validate most
91 methylome-derived subtypes, allowing prediction of 1.6M enhancer-like DMRs (eDMR)
92 across cell subtypes.
- 93 ● Identify cis-regulatory interactions between enhancers and genes with computational
94 prediction and single-cell chromatin conformation profiling (in hippocampus).
- 95 ● A continuous three-dimensional spatial methylation gradient was observed in cortical
96 excitatory and dentate gyrus granule cells, with region-specific TFs and motifs found
97 associated with the gradients.
- 98 ● An artificial neural network model constructed using this deep dataset precisely predicts
99 single neuron cell-type identity and brain area spatial location using only its methylome
100 profile as input.

101 **Results**

102 **A single cell DNA methylome atlas of 45 mouse brain regions**

103 We used single-nucleus methylcytosine sequencing 2 (snmC-seq2)¹⁶ to profile genome-wide
104 cytosine DNA methylation at single-cell resolution across four major brain structures: cortex,
105 hippocampus (HIP), striatum and pallidum (or cerebral nuclei, CNU), and olfactory areas (OLF)
106 (Fig. 1d-g) using adult male C57BL/6 mice (age P56-63). In total we analyzed 45 dissected
107 regions in two replicates (Extended Data Fig. 1, Supplementary Table 2).
108 Fluorescence-activated nuclei sorting (FANS) of antibody-labeled nuclei was applied to capture
109 NeuN positive (NeuN⁺, 92%) neurons, while also sample a smaller number of NeuN negative
110 (NeuN⁻, 8%) non-neuronal cells (Fig. 1a). In total, we profiled the DNA methylomes of 110,294
111 single nuclei yielding, on average, 1.5 million stringently filtered reads/cell (mean \pm SD: 1.5×10^6
112 $\pm 5.8 \times 10^5$), covering $6.2 \pm 2.6\%$ of the mouse genome per cell. In each cell, we reliably
113 quantified DNA methylation for $95 \pm 4\%$ of the mouse genome in 100kb bins and $81 \pm 8\%$ of
114 gene bodies (Fig. 1i). In parallel, we performed single-nucleus ATAC-seq (snATAC-seq)¹⁷ on
115 nuclei from the same samples to identify sites of accessible chromatin (see below and Li et al.,
116 companion manuscript # 11).

117 Based on the DNA methylome profiles in both CpG sites (CG methylation or mCG hereafter)
118 and non-CpG sites (CH methylation or mCH) in 100kb bins throughout the genome, we
119 performed a three-level iterative clustering analysis to categorize the epigenomic cell
120 populations (Fig. 1b, c). After quality control and preprocessing (see Methods), in the first level
121 (cell class), we clustered 103,982 cells into 67,472 (65%) excitatory neurons, 28,343 (27%)
122 inhibitory neurons, and 8,167 (8%) non-neurons. The portion of non-neurons is determined by
123 the FANS of NeuN⁺ nuclei (Supplementary Table 3, Extended Data Fig. 2b). The second round
124 of iterative analysis of each cell class identified 41 major types in total (cluster size range
125 95-11,919, median 1,819 cells), and the third round separated these major types further into 161
126 cell subtypes (cluster size range 12-6,551, median 298 cells). Each subtype can be
127 distinguished from all others based on: 1) a supervised model that reproduces the cluster label
128 with > 0.9 cross-validated accuracy; and 2) at least 5 differentially methylated genes between
129 each pair of subtypes (see Methods, and Extended Data Fig. 2, Supplementary Table 4 for
130 subtype names and signature genes). All subtypes are highly conserved across replicates
131 (Extended Data Fig. 2d), and replicates from the same brain region are co-clustered compared
132 to samples from other brain regions (Extended Data Fig. 2e-g).

133 The spatial distribution of each cell type is assessed based on where the cells were dissected
134 from (Supplementary Table 6). Here we used the Uniform Manifold Approximation and
135 Projection (UMAP)¹⁹ to visualize epigenetic differences based on cell location (Fig. 1b, Extended
136 Data Fig. 3) and major cell type (Extended Data Fig. 2a). Major non-neuronal types have a
137 similar distribution across brain regions (Fig. 1h), with the exception of Adult Neuron Progenitors
138 (ANP). We found two subtypes of ANP presumably corresponding to neuron precursors in the
139 subgranular zone of the dentate gyrus (DG)²⁰ (“ANP anp-dg”, 121 cells) and the rostral
140 migratory stream (RMS)²¹ in CNU and OLF (“ANP anp-olf-cnu”, 210 cells). Major excitatory
141 neuron types from isocortex, OLF, and HIP formed separate subtypes, with some exceptions
142 potentially due to overlaps in dissected regions (see “Potential Overlap” column in
143 Supplementary Table 2). Cells from isocortex were assigned to subtypes based on their
144 projection types^{8,22,23}. The Intratelencephalic (IT) neurons from all cortical regions contain four
145 major types corresponding to the laminar layers (L2/3, L4, L5, and L6), with additional subtypes
146 in each layer that differ across cortical regions. We also identified major types from cortical
147 subplate structures, including the claustrum (CLA) and endopiriform nucleus (EP) from isocortex
148 and OLF dissections.

149 GABAergic inhibitory neurons from isocortex and HIP cluster together into five major types, but
150 most HIP neurons are then separated from isocortical neurons at the subtype level, in
151 agreement with a companion transcriptomic study²⁴. Intriguingly, in one major GABAergic type
152 marked by low methylation at the gene *Unc5c* (suggesting high RNA expression), cells from HIP
153 and isocortex remain to be co-clustered at subtype levels. The isocortex cells in this cluster
154 display hypo-mCH in the gene body of both *Lhx6* and *Adarb2*, corresponding to the previously
155 reported “Lamp5-Lhx6” cluster in transcriptome studies^{22,25,26}. The close relationship between
156 this cell type and HIP interneurons was also reported in a single-cell transcriptomic study in
157 marmoset²⁷. In contrast with excitatory neurons, subtypes of isocortical GABAergic neurons do
158 not separate by cortical region. Interneurons from CNU and OLF group into nine major types
159 and many subtypes which further separate by region (see vignette below), indicating substantial
160 spatial diversity among CNU and OLF interneurons. In the next section, we provide several
161 analytic vignettes to illustrate the unprecedented level of neuronal subtype and spatial diversity
162 observed in their DNA methylomes.

163 **Neuron subtype vignettes**

164 *Methylome similarity between Indusium Griseum and Hippocampal region CA2 neuron subtypes*

165 Nearly all hippocampal excitatory neurons formed their own clusters separately from cells in
166 other brain regions (Fig. 1h), except the IG-CA2 neurons (745 cells, three subtypes; IG:
167 Indusium griseum, CA: Cornu Ammonis) (Fig. 2c). Several markers of the hippocampal region
168 CA2²⁸ (e.g., *Pcp4*²⁹, *Cacng5*³⁰, *Ntf3*³¹) were marked by low gene body mCH (Fig. 2d) in IG-CA2
169 subtypes. However, one subtype “IG-CA2 Xpr1” (subtype 1 in Fig. 2c, 152 cells) was located in
170 the Anterior Cingulate Area (ACA, 72 cells) and Lateral Septal Complex (LSX, 71 cells), which
171 are anatomically distinct from the hippocampus (Extended Data Fig. 1). In-situ hybridization
172 (ISH) data from the Allen Brain Atlas (ABA) of *Ntf3* (Fig. 2d, S4c) indicates that those cells
173 potentially come from the IG region³² (Fig. 2e), which is a thin layer of gray matter lying dorsal to
174 the corpus callosum at the base of the anterior half of the cingulate cortex, included in the ACA
175 and LSX dissection regions (Fig. 2c, 3D model). With the power of single-cell epigenomic
176 profiling, we are able to classify cells from this region without additional dissection. Moreover the
177 similarity of their DNA methylomes indicates the possible functional and/or developmental

178 relationship between the CA2 and IG excitatory neurons. Finally, the hypo-mCG regions
179 surrounding the marker genes like *Ntf3* (Extended Data Fig. 4b) identify candidate
180 cell-type-specific enhancers which may further our understanding of how specific gene
181 expression programs are regulated in these structures.

182 *Methylation signatures of striatal medium spiny neurons located in patch compartments*

183 A major GABAergic inhibitory cell type in the striatum, the *Drd1+* medium spiny neuron
184 (MSN-D1) from the caudoputamen (CP, dorsal) and nucleus accumbens (ACB, ventral), is
185 further separated into four subtypes (Fig. 2f, S4d). Two subtypes “MSN-D1 *Plxnc1*” (subtype 3
186 in Fig.2f) and “MSN-D1 *Ntn1*” (subtype 4 in Fig.2f), mainly (79%) from the ACB, are further
187 separated by location along the anterior-posterior axis (Extended Data Fig. 4e, based on
188 dissection), indicating spatial diversity may exist in addition to the canonical dorsal-ventral
189 gradient of the striatum^{33,34}. “MSN-D1 *Khdrbs3*” and “MSN-D1 *Hrh1*” subtypes are mostly (94%)
190 from CP dissections, one of them marked by gene *Khdrbs3* and its potential regulatory elements
191 (Fig. 2g, h, Extended Data Fig. 4f) corresponds to the neurochemically defined patch³⁵
192 compartments in CP. Here the methylome profiling data provides evidence of previously unseen
193 spatial epigenetic diversity in the striatum D1 neurons, which is also observed in the other major
194 type MSN-D2 (*Drd2+*) of the striatum (Extended Data Fig. 4g, e, i).

195 *Spatial and projection specificity of extra-telencephalic (ET) neuron subtypes*

196 We found that cortical excitatory subtypes have distinct spatial distributions. For example,
197 subtypes of L5-ET consist of cells from different cortical regions (Fig. 2i, k). This is further
198 confirmed by integrating our snmC-seq2 data with data from neurons with defined projections
199 from a parallel study using Epi-Retro-Seq (Companion Manuscript # 10)³⁶ (Fig. 2j).
200 Epi-Retro-Seq uses retrograde viral labeling to select neurons projecting to specific target brain
201 regions followed by methylome analysis of their epigenetic subtypes. To further infer the
202 projection target of L5-ET cells profiled in our study, we performed co-clustering on the
203 integrated datasets and calculated the overlap between unbiased (snmC-seq2) and targeted
204 (Epi-Retro-Seq) profiling experiments (Extended Data Fig. 4m). Interestingly, some subtypes
205 identified from the same cortical area show different projection specificity. For example, in the
206 L5-ET neurons, SSp and MOp neurons were mainly enriched in three subtypes marked by
207 *Kcnh1*, *Tmtc2*, and *Nectin1*, respectively. However, medulla projecting neurons in the MOp and

208 SSp only integrate with the subtype marked with *Kcnh1* (Fig. 2l), suggesting that the subtypes
209 identified in unbiased methylome profiling have distinct projection specificity.

210 *Consensus epigenomic profiles of brain cell subtypes*

211 Integrating single-cell datasets collected using different molecular profiling modalities (e.g.
212 snRNA-seq, snmC-seq2, snATAC-seq) can help to establish a consensus cell type atlas^{23,37}.
213 Using a mutual nearest neighbor based approach^{38,39} (see Methods), we integrated the
214 methylome data with the chromatin accessibility data profiled using snATAC-seq on the same
215 brain samples from a parallel study (Li et al., Companion Manuscript # 11). As visualized by the
216 ensemble UMAP, after integration the two modalities validated each other at the level of subtype
217 cluster assignment (Fig. 2m, n, Extended Data Fig. 5a). We then performed co-clustering of the
218 integrated data and calculated Overlap Scores (OS) between the original methylation subtypes
219 (m-types) and the chromatin accessibility subtypes (a-types) (Fig. 2o, Extended Data Fig. 5b).
220 The OS describes the level of co-clustering of two groups of cells, where a higher value means
221 the two groups are more similarly distributed across co-clusters (see Methods and
222 Supplementary Table 7). The high overlap of CG-DMRs and open chromatin peaks in the
223 hippocampus also confirmed the correct match of cell-type identities (Fig. 2p).

224 **Regulatory hierarchy of neuronal subtypes**

225 Having developed a consensus map of cell types based on their DNA methylomes, we further
226 explored the gene regulatory relationship between neuronal subtypes. We constructed two
227 phylogeny trees for 68 excitatory types (Extended Data Fig. 6a) and 77 inhibitory types
228 (Extended Data Fig. 6b) respectively, based on the gene body mCH level of 2,503 differentially
229 methylated genes (CH-DMGs) between every pair of clusters. The taxonomy tree structures
230 represent the similarities of these discrete subtypes, and may reflect the developmental history
231 of neuronal type specification^{22,25}.

232 A unique advantage of single-neuron methylome profiling is that it captures the information of
233 both cell-type-specific gene expression and predicted regulatory events. Specifically, gene body
234 mCH is predictive of gene expression in neurons^{2,7,8}, while CG-DMRs indicate cell-type specific
235 regulatory elements, and TFs whose motifs enriched in these CG-DMRs predict the crucial
236 regulators of the cell type⁶⁻⁸. We used both CH-DMGs and CG-DMRs to further annotate the

237 tree and explore the features specifying cell subtypes (Exc: Fig. 3a-c, Inh: Extended Data Fig.
238 6c, d).

239 To better understand which genes contribute to cell type specifications in the branches of the
240 tree, we calculated a branch-specific “methylation impact score” for each gene that summarises
241 all of the pairwise comparisons related to that branch (Fig. 3d; see Methods). The impact score
242 ranges from 0 to 1, with a higher score predicting stronger functional relevance to the branch.
243 With impact scores > 0.3, a total of 6,038 unique genes were assigned to branches within the
244 excitatory phylogeny (5,975 in inhibitory), including 406 TF genes (412 in inhibitory).

245 Key TFs determine the neuronal cell-type identity by targeting downstream genes through the
246 binding to regulatory DNA elements during brain development⁴⁰. Timed expression of these TFs
247 and their co-regulators throughout the neuron's lifespan is critical to maintain cell-type
248 identity⁴⁰⁻⁴². To better understand what regulatory elements and TFs contribute to cell-type
249 specification in the tree, we annotated branches of the tree with TF motif usage. We first
250 identified 3.9 million DMRs across all the cell types, with an average size of 624 ± 176 bp (mean
251 \pm SD), covering 50.0% of the genome CpG sites spanning 1,240 Mb in total (see Methods). We
252 next scanned 719 known TF binding motifs from the JASPAR database⁴³ across these DMRs
253 and performed motif enrichment analyses using all pairwise DMRs identified between neuronal
254 subtypes (Exc: Fig. 3c, Inh: Extended Data Fig. 6d). Similar to the analysis of DMGs, the
255 pairwise enriched motifs were assigned to branches on the subtype phylogeny using their
256 impact scores.

257 This analysis revealed a possible regulatory role for several TFs assigned to the upper nodes of
258 the phylogeny. For example, motifs from the ROR(NR1F) family were assigned to the branch
259 that separates superficial layer IT neurons from deeper layer IT neurons (Fig. 3h, node 9),
260 whereas motifs from the CUX family were assigned to the L2/3-IT branch separating from
261 L4/5-IT neurons (Fig. 3h, node 11). Both families contain known members, for example *Cux1*,
262 *Cux2*⁴⁴, *Rorb*⁴⁵ that show laminar expression in the corresponding layers and regulate cortical
263 layer differentiation during development.

264 After impact score assignment, each branch of this phylogeny was associated with multiple TF
265 genes and motifs, which potentially function in combination to shape cell-type identities⁴⁶ (Fig.

266 3g, h). As an example, we focused on two brain structures of interests: the claustrum (CLA) and
267 the Endopiriform Nucleus (EP)^{47,48}. Single excitatory neurons from these two structures were
268 included in the dissection of several cortical and piriform cortex regions (Supplementary Table
269 6). At the major-cell-type level, two distinct clusters are marked by *Npsr1* (EP) or *B3gat2* (CLA),
270 and the known EP/CLA marker TF *Nr4a2*^{47,49} were observed to be significantly hypomethylated
271 in both clusters compared to the others. Accordingly, the NR4A2 motif is also associated with a
272 branch that splits CLA neurons from L6-IT neurons (Fig. 3g, h, node 6). On another branch
273 separating EP from CLA and L6-IT neurons, several TFs including NF-1 family gene *Nfia* and
274 *Nfib*, and RFX family gene *Rfx3*, together with corresponding motifs (Fig. 3g, 3h, node 5) rank
275 near the top. Our findings suggest that these TFs may function together with *Nr4a2*, potentially
276 separating EP neurons from CLA and L6-IT neurons.

277 Beyond identifying specific cell subtype characteristics, we hypothesized that ranking of genes
278 or motifs by methylation variation may provide a route toward understanding their relative
279 importance in cell type diversification and/or function. Thus, for each gene or motif, we
280 calculated a total impact (TI) score to summarize their variation among the subtypes, using all
281 the branch specific impact scores (see Methods). Genes or motifs with higher TI score impact
282 more branches of the phylogeny, thus are predicted to have higher importance in distinguishing
283 multiple subtypes. Intriguingly, by comparing the TI scores of genes and motifs calculated from
284 the inhibitory and excitatory phylogenies, we found more TF genes and motifs having large TI
285 scores in both cell classes than specific to either one (Extended Data Fig. 6i). For instance,
286 *Bcl11b* distinguishes “OLF-Exc” and Isocortex IT neurons in the excitatory lineage, and
287 distinguishes “CGE-Lamp5” and “CGE-Vip” in the inhibitory lineage. Similarly, *Satb1* separates
288 L4-IT from L2/3-IT, and MGE from CGE in excitatory and inhibitory cells, respectively. These
289 findings indicate broad repurposing of TFs for cell type specification among distinct
290 developmental lineages.

291 In contrast, we also find TF genes and motifs only having large TI scores in one cell class
292 (Extended Data Fig. 6i). For example, *Tshz1* gene body mCH shows a striking difference of
293 diversity between excitatory and inhibitory cells (Extended Data Fig. 6h), suggesting that it may
294 function in shaping inhibitory subtypes, but not excitatory subtypes. Similarly, bHLH DNA
295 binding motifs show much higher TI scores for excitatory subtypes compared with inhibitory
296 (Extended Data Fig. 6i). While genes in this TF family such as *Neurod1/2* have long been

297 known to participate in excitatory neuron development, they have not been reported to regulate
298 GABAergic neuron differentiation⁵⁰.

299 **Defining regulatory interaction between enhancers and genes at the cell type level**

300 To systematically identify enhancer-like regions in specific cell types, we predicted
301 enhancer-DMRs (eDMR) by integrating matched DNA methylome and chromatin accessibility
302 profiles of 161 subtypes using the REPTILE algorithm⁵¹ (Fig. 4a). We identified 1,612,198
303 eDMR (34% of total DMRs), 73% of which overlapped with separately identified snATAC peaks
304 (Fig. 4b). Fetal-enhancer DMRs (feDMR, eDMRs identified across developing time points) of
305 forebrain bulk tissues⁶ show high (88%) overlap with eDMRs. Surprisingly, the eDMRs also
306 cover 74% of the feDMRs from other fetal tissues⁶, indicating extensive reuse of enhancer-like
307 regulatory elements across mammalian tissue types (Fig. 4b).

308 Next, we examined the relationship between the cell type signature genes and their potential
309 regulatory elements. After regressing out the global methylation level, we calculated the partial
310 correlation between all DMG-DMR pairs within 1 Mb distance, using methylation levels across
311 145 neuronal subtypes (see Methods). Non-neuronal subtypes were not included in this
312 analysis due to the large difference in the global methylation level compared to neurons. We
313 identified a total of 1,038,853 (64%) eDMR that correlated with at least one gene (correlation >
314 0.3 with empirical P value < $1e-4$, two-sided test, Extended Data Fig. 7a). Notably, for those
315 strongly positive-correlated DMR-DMG pairs (correlation > 0.5), the DMRs are largely (63%)
316 within 100kb to the TSSs of the corresponding genes, but depleted from ± 1 kb (Fig. 4c,
317 Extended Data Fig. 7b), whereas for the negatively correlated DMR-DMG pairs, only 11% of
318 DMRs are found within 100kb of the TSS (Extended Data Fig. 7c).

319 eDMR-DMG correlation analysis predicts regulatory interactions between enhancers and genes
320 at subtype resolution. Using the gene-enhancer interactions predicted by this correlation
321 analysis, we assigned eDMRs to their target genes, and calculated the percentage of overlap
322 with feDMR. We found that these percentages vary dramatically among genes (Fig. 4d). For
323 example, in *Bcl11b*, an early developmental TF gene⁵², 63% of the positively correlated eDMRs
324 overlap with forebrain feDMRs (Fig. 4f, first row). However, for *Tle4*, a gene known to be
325 functional in L6-CT⁵³ and MSN-D1/D2 neurons, the percentage of eDMR that overlapped with

326 feDMRs was only 20% (Fig. 4d). Interestingly, when further classifying *Tle4* correlated eDMR
327 into three subgroups using K-means clustering (k=3, see Methods), we identified subgroups
328 with different cell-type-specificity (Fig. 4f, second row. Extended Data Fig. 7d). One group (G2)
329 of elements that displayed little diversity during development in bulk data showed highly specific
330 mCG and open chromatin signals in MSN-D1/D2 neurons, while another group (G3) is specific
331 to L6-CT neurons. These two groups of DMRs suggest possible alternative regulatory elements
332 usage to regulate the same gene in different cell-types although further experiments are
333 required to validate this hypothesis.

334 Together, these analyses indicate that previously defined developmental regulatory elements
335 from bulk tissues are limited in terms of the cell type resolution. For those broadly effective
336 regulatory elements (e.g., feDMRs correlated with *Bcl11b*) that also have been identified in bulk
337 tissues, single-cell profiling allows us to carefully chart their cell type specificity into subtype
338 level. In addition, we identified many novel regulatory elements (e.g., eDMRs correlated with
339 *Tle4* in MSN-D1/D2) that show more restricted specificity, providing abundant candidates for
340 further pursuing enhancer-driven AAVs⁵⁴⁻⁵⁶ that target fine cell types.

341 **Single-nucleus multi-omic profiling of chromatin contact and DNA methylation validates** 342 **gene-enhancer interactions**

343 Distal enhancers typically regulate gene expression by physical interaction with promoters⁵⁷.
344 Therefore, to examine whether our correlation-based predictions of enhancer-gene associations
345 are supported by physical chromatin contacts, we generated single-nucleus methylation and
346 chromosome conformation capture sequencing (sn-m3C-seq)¹⁸ data for 5,142 single nuclei from
347 the DG and CA (152k contacts per cell on average). We assigned each of these cells to one of
348 the 161 subtypes based on their methylome, using a supervised model trained with >100k
349 snmC-Seq2 cells. By merging the contact matrices of all single cells belonging to the same
350 methylation-defined subtype, we generated contact maps for eight major cell types that include
351 more than 100 cells. In total, 19,151 chromosome loops were identified in at least one of the
352 cell-types at 25kb resolution (range from 1,173 to 12,614).

353 Since sn-m3C-seq data identified structural interactions between genes and regulatory
354 elements, we next asked whether these interactions were consistent with the predictions from

355 eDMR-DMG correlation analysis. Using DG and CA1 as examples (the neuronal types with
356 highest coverage), a notably higher correlation was observed between enhancers and genes at
357 loop anchors compared to random enhancer-gene pairs after controlling for genomic distance
358 (Fig. 4g; P value=5.9e-74 for DG and 3.0e-158 for CA1, two-sided Wilcoxon rank-sum tests).
359 Reciprocally, the enhancer-gene pairs showing the stronger correlation of methylation were
360 more likely to be found linked by chromosome loops or within the same looping region (Fig. 4h,
361 Extended Data Fig. 7f). We also compared the concordance of methylation patterns between
362 genes and enhancers linked by different interaction-based methods, and found the pairs linked
363 by loop anchors or closest genes had the highest correlation of methylation (Extended Data Fig.
364 7g). Together, these analyses validate the physical proximity of co-associated, hypomethylated
365 enhancer-gene pairs predicted by our correlation based method in specific cell types.

366 In addition, we observed significant cell type specificity of 3D genome structures. The major cell
367 types could be distinguished on UMAP embedding based on chromosome interaction at 1 Mb
368 resolution (Fig. 4i), indicating the dynamic nature of genome architecture across cell types.
369 Among the 19,151 chromosome loops at 25 kb resolution, 48.7% showed significantly different
370 contact frequency between cell types (Fig. 4j). eDMRs were highly enriched at these differential
371 loop anchors (Fig. 4k; $p < 0.005$, permutation test). mCG levels at distal cis-elements are typically
372 anti-correlated with enhancer activity⁶. Thus, we hypothesized that enhancers at differential loop
373 anchors might also be hypomethylated in the corresponding cell-type. Indeed, using the loops
374 identified in DG granule cells and CA1 as an example, we observed that in DG, enhancers at
375 the anchors of DG specific loops were hypomethylated compared to enhancers at the anchors
376 of CA1 specific loops (P value=2.9e-103, two-sided Wilcoxon rank-sum test); the opposite
377 scenario was found in CA1 (Fig. 4l; P value=3.5e-5, two-sided Wilcoxon rank-sum test).

378 Many differential loops were observed near the marker genes of the corresponding cell type.
379 For example, *Foxp1*, a highly expressed TF in CA1 but not DG⁵⁸, has chromosome loops
380 surrounding its gene body in CA1 but not DG (Fig. 4m). eDMRs and open chromatin were
381 observed at these loop anchors (Fig. 4m). Intriguingly, three loops in CA1 anchored at the TSS
382 of the same transcript of *Foxp1* (Fig. 4m, box 2-4). Stronger demethylation and chromatin
383 accessibility were also observed at the same transcript compared to the others (Fig. 4m, view
384 E). These epigenetic patterns might suggest a specific transcript of *Foxp1* (*Foxp1*-225) is
385 selectively activated in CA1. In contrast, *Lrrtm4*, a DG marker gene encoding a presynaptic

386 protein that mediates excitatory synapse development in granule cells⁵⁹, shows extensive
387 looping to distal elements in DG but not CA1 (Extended Data Fig. 7h). Notably, 34 genes
388 showed alternative loop usage, 20 of which expressed in both DG and CA1 (CPM > 1). *Grm7* is
389 an example where its TSS interacts with an upstream enhancer in DG but with gene-body
390 enhancers in CA1 (Extended Data Fig. 7i).

391 **Spatial gradients in intra-telencephalic (IT) excitatory neurons**

392 IT neurons from different cortical areas are typically classified into major types corresponding to
393 their laminar locations: L2/3, L4, L5, and L6 (Fig. 5a). In agreement with the correlation between
394 transcription²² and DNA methylation, we found that IT neurons show hypomethylation of the
395 marker genes defined in Tasic et al.²² in the corresponding layers (Extended Data Fig. 8c). For
396 example, the *Rorb*⁺ L4 cells show unique gene body hypomethylation in both somatomotor (MO)
397 and somatosensory areas (SS) (Extended Data Fig. 8c). These findings suggest that although
398 the MO lacks a cytoarchitectural visible L4⁶⁰, there is a population of IT neurons that are
399 epigenetically similar to L4 neurons in SS. The presence of a putative L4 neuron population in
400 MO is also supported by transcriptomics²³ and neuronal connectivity⁶⁰. Furthermore,
401 unsupervised UMAP embedding (Fig. 5a) reveals a continuous gradient of IT neurons
402 resembling the medial-lateral distribution of the cortical areas (Fig. 5b), strongly suggesting that
403 the cortical arealization information is well preserved in the DNA methylome.

404 To systematically explore the spatial gradient of DNA methylation patterns, we remerge the cells
405 into spatial groups based on their cortical layer and anatomical regions. A phylogenetic tree was
406 generated using the mCH level of highly variable 100kb chromosome bins of these spatial
407 groups' centroids (see Methods). The phylogeny split the cells into four laminar layer groups,
408 followed by cortical area separation within each layer (Fig. 5c). This provides a clear structure
409 for calculating the methylation total impact score of layer-related or region-related DMGs and TF
410 binding motifs separately (Fig. 5d, see Methods). In brief, the top layer-related TFs included
411 most of the known laminar marker genes together with their DNA binding motifs (Fig. 5d), while
412 some of them show regional specific methylation differences between layers. For example,
413 *Cux1*, a top-ranked TF marker gene for L2/3 and L4 neurons, is hypomethylated in MO and SS,
414 but is hypermethylated in L2/3 of other regions we sampled, in agreement with patterns from
415 in-situ hybridization⁶¹. *Cux2*, a gene from the same TF family, does not show the same regional

416 specificity (Extended Data Fig. 8c). We also identified many additional TFs having cortical
417 region specificity (Fig. 5e, f). For example, *Etv6* is only hypomethylated in medial dissection
418 regions, while *Zic4* is hypermethylated in those regions across layers. In contrast, *Rora* shows
419 anterior-posterior methylation gradient only in the L4 and L5 cells. Together, these observed
420 methylome spatial gradients demonstrated the value of our dataset for further exploring the
421 cortical arealization with cell-type resolution.

422 **Spatial gradients in dentate gyrus granule cells**

423 Another striking spatial gradient was observed in granule cells from DG. UMAP embedding of
424 granule cells displayed continuous global mCH and mCG level gradients that correlated with
425 their dorsal-ventral location. Granule cells from the ventral DG show higher global CG and CH
426 methylation compared to cells from the dorsal DG (Fig. 5g). To investigate the biological
427 significance of these gradients, we identified 219,498 gradient CG-DMRs by grouping DG cells
428 according to their global mCH level. Among them, 139,387 DMRs are positively correlated with
429 global mCH, and 80,111 are negatively correlated DMRs (Fig. 5h). Positively or negatively
430 correlated DMRs showed significant enrichment in two different sets of genes (245 pos. corr.;
431 183 neg. corr.) which include TF genes such as *Tcf4* and *Rfx3* compared to the genome
432 background (Fig. 5i-j, see Methods). Based on Gene Ontology analysis, the negatively
433 correlated DMRs are enriched in genes related to synaptic functions (Extended Data Fig. 8d),
434 while the positively correlated DMRs are enriched in developmental related genes (Extended
435 Data Fig. 8e). Together, these results indicate a continuous methylation gradient exists in the
436 dorsal-ventral axis of the DG granule layers, supporting an intra-cell-type molecular shift that
437 parallels the known functional and anatomical differences in dorsal-ventral DG^{62,63} that are
438 possibly controlled via gradient DNA methylation⁶⁴.

439 Next, we investigated whether the global dynamic of methylation is correlated with changes in
440 3D genome architecture. By plotting the interaction strength against the genomic distance
441 between the anchors (decay curve), a higher proportion of short-range contacts and smaller
442 proportion of long-range contacts were observed in the groups with higher global mCH (Fig. 5k).
443 This might indicate a more compact nuclear structure in these groups. The genome is organized
444 into specific 3D features with different levels of resolution, including compartments, domains,
445 and loops. We tested the differences of these features across DG cells grouped based on their

446 global mCH levels. Although compartment strengths were not correlated with these methylation
447 changes (Extended Data Fig. 8f), the number of intra-domain contacts was positively correlated
448 with global mCH across single cells (Fig. 5l), which is concordant with the patterns observed in
449 the contact frequency decay curves. Interestingly, after normalizing the effect of decay, we found
450 the insulation scores at domain boundaries were significantly lower in the groups with high
451 global mCH levels (Fig. 5m; all $p < 1e-10$, two-sided Wilcoxon signed-rank test), which suggests
452 that the domain structures might be more condensed over flanking regions in these cell groups.

453 **Deep neural network learning of cell identity and spatial location**

454 To further test the extent that spatial and cell-type information is encoded in a single neuron's
455 DNA methylome, we built a multi-task deep Artificial Neural Network (ANN) using cell-level
456 methylome profiles in this study (Fig. 6a). Specifically, mCH rates of non-overlapping 100kb bins
457 from 94,383 neurons were used to train and test the network with five-fold cross-validation. The
458 ANN was able to predict subtype identity and spatial location simultaneously for each testing
459 cell with 95% and 89% accuracies, respectively (Fig. 6b-d). Importantly, the location prediction
460 accuracy using DNA methylation is significantly higher than only using the spatial distribution
461 information of each subtype (overall accuracy increased by 38%, Extended Data Fig. 9b),
462 suggesting that spatial diversity is well-preserved in the neuronal DNA methylome. We did,
463 however, notice errors in location prediction in a few cell types, most notably in MGE and CGE
464 derived inhibitory neurons (Fig. 6e, Extended Data Fig. 9b). This finding is consistent with
465 previous transcriptome based studies²², suggesting these inhibitory neurons do not display
466 strong cortical region specificity.

467 Next, we selected the features (100kb bins) that capture most spatial information (Fig. 6f, see
468 Methods), and found that genes related to neuronal system development are highly enriched in
469 these bins (Fig. 6g, h). Many cell type marker genes are also located in these bins, such as
470 *Foxp2* (Fig. 6f). In addition to distinguishing L6-CT neurons from other cell types, *Foxp2* shows
471 notably mCH difference among different dissected regions in L6-CT (Fig. 6l), indicating the
472 extensive impact of neuron spatial origin on their methylomes.

473 **Discussion**

474 In this study, we describe the generation and analysis of a comprehensive single-cell DNA
475 methylation atlas for the mouse brain, encompassing over 110,000 methylomes; the largest
476 methylation dataset for any organism to date. By profiling of 45 brain regions, a total of 161
477 subtypes: 68 excitatory, 77 inhibitory, and 16 non-neuronal subtypes were classified using the
478 three-levels of iterative clustering. These subtypes are associated with a total of 24,208
479 CH-DMGs and 3.9M CG-DMRs, covering 50% (1,240 Mb) of the mouse genome. Through
480 integration with snATAC-seq (Li et al. Companion Manuscript # 11), we matched the chromatin
481 accessibility clusters to each of the methylome subtypes and used the combined epigenomic
482 information to predict 1.5M active-enhancer-like eDMRs, including 72% cell-type-specific
483 elements that were missed from previous tissue level bulk studies^{6,65}.

484 To describe cell-type specificity of genes and TF binding motifs in the context of cell type
485 phylogeny, we defined a metric called the methylation impact score. This metric aggregates
486 pairwise DMGs and DMRs between subtypes, and assigns them to branches of a cell type
487 phylogeny. These assignments allow us to describe cell-type specificity at different levels of the
488 phylogeny, potentially relating to different stages of their neuronal developmental lineages. We
489 found that many known TF genes and their corresponding DNA-binding motifs were
490 co-associated with the same branch in the phylogeny, supporting the biological significance of
491 this approach. An important outcome of this analysis is the discovery of many novel TFs with
492 high impact scores, providing a rich source of candidate TFs for future study.

493 To study the associations of eDMRs and their targeting genes, we first established the
494 eDMR-gene landscape via correlation between the mCH rate of genes and mCG rate of DMRs.
495 Furthermore, in the hippocampus, we also generated high-resolution single-cell 3D chromatin
496 conformation and DNA methylome from the same cells using sn-m3C-seq. With this multi-omic
497 dataset, we obtained cell-type-specific loops in eight cell types. The physical loops and
498 eDMR-gene correlation loops together linked candidate enhancers to their target genes in
499 specific cell types.

500 Through the analyses of single-cell methylome from detailed spatial brain region dissections,
501 our brain-wide epigenomic dataset reveals extraordinary spatial diversity encoded in the DNA
502 methylomes of neurons. The iterative clustering analysis separates fine-grained spatial
503 information, which is accurately reproduced by the ANN trained on the single-cell methylome
504 profiles. Moreover, the ANN also reveals additional spatial specifications within most of the
505 subtypes, indicating continuous spatial DNA methylation gradients widely exist among cell
506 types. The spatial gradients observed for excitatory IT neurons within distinct cortical areas are
507 of special interest. During cortex development, glutamatergic neurons are regionalized by a
508 proto map formed from an early developmental gradient of TF expression^{66,67}. Similarly we
509 observed that many TF genes and their corresponding DNA binding motifs showed gradients of
510 DNA methylation in adult IT neurons from distinct cortical regions. Additionally, we found
511 intra-subtype methylation gradients in DG granule cells. These CG-DMR are enriched in
512 essential synaptic genes, suggesting the existence of a methylation gradient in the
513 dorsal-ventral axis in the DG granule layer⁶⁴.

514 Overall, we present the first single-cell resolution DNA methylomic mouse brain atlas with
515 detailed spatial dissection and subtype level classification. Our analysis highlights the power of
516 this dataset for characterizing cell types with both gene activity information in the coding regions
517 and the regulatory elements in the non-coding region. This comprehensive epigenomic dataset
518 provides a valuable resource for answering fundamental questions about gene regulation in
519 specifying cell type spatial diversity and provides the raw material to develop new genetic tools
520 for targeting specific cell types and functional testing of them.

521 **Acknowledgments**

522 We thank Dr. Yupeng He for the advice on the methylpy and REPTILE analysis. We thank Dr.
523 Terrence Sejnowski for the advice on the ANN analysis. This work is supported by NIMH
524 U19MH11483 to J.R.E. and E.M.C. The Flow Cytometry Core Facility of the Salk Institute is
525 supported by funding from NIH-NCI CCSG: P30 014195. J.R.E is an investigator of the Howard
526 Hughes Medical Institute.

527 **Author Contributions**

528 J.R.E., H.L., B.R., M.M.B., C.L., J.R.D. conceived the study. H.L., J.Z., W.T. analyzed the
529 snmC-seq data and drafted the manuscript. J.R.E., C.L., E.A.M., J.R.D., M.M.B. edited the
530 manuscript. J.R.E., H.L., M.M.B., A.B., J.L., S.P. coordinated the research. M.M.B., A.B., A.A.,
531 H.L., J.L., J.R.N., A.R., J.K.O., C.O., L.B., C.F., C.L., J.R.E. generated the snmC-seq2 data.
532 J.R.D., B.C., A.B., J.L., J.Z., A.A., J.K.O., C.L., J.R.N., C.O., L.B., C.F., R.G.C., M.M.B., J.R.E.
533 generated the sn-m3C-seq data. S.P., M.M.B., X.H., J.L., O.B.P., Y.E.L., J.K.O., B.R. generated
534 the snATAC-seq data. Z.Z., J.Z., E.M.C., M.M.B., J.R.E., A.B., A.A., J.R.N., C.O., L.B., C.F.,
535 R.G.C., A.R. generated the Epi-Retro-Seq data. H.L., H.C., E.A.M., M.N., C.L. contributed to
536 data archive/infrastructure. J.R.E. supervised the study.

537 **Competing interests**

538 J.R.E serves on the scientific advisory board of Zymo Research Inc.

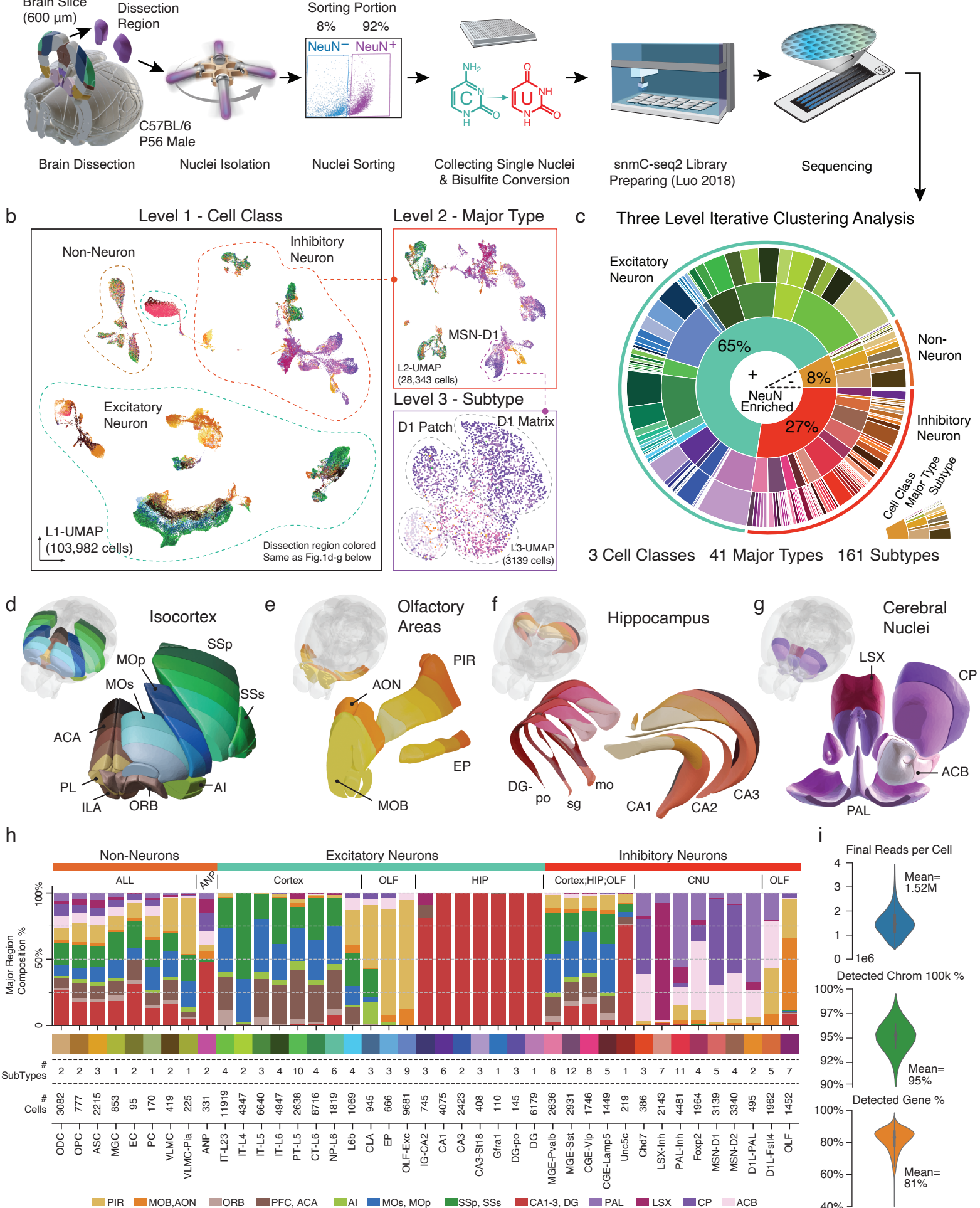


Figure 1

539 **Figure 1. A survey of single-cell DNA methylomes in the mouse brain.**

540 **a**, The workflow of dissection, FANS, and snmC-seq2 library preparation. **b**, Level1 UMAP
541 (L1-UMAP) is colored by the dissection region; the color palette is the same as **(d-g)**. Panels
542 show examples of Level 2 and Level 3 UMAP embeddings from iterative analyses. **c**, Sunburst
543 of three-level iterative clustering. From inner to outer: L1 Cell Classes, L2 Major Types, and L3
544 Subtypes. **d-g**, Brain 3D dissection models of Isocortex (**d**), olfactory areas (**e**), hippocampus
545 (**f**), and cerebral nuclei (**g**) colored by 45 dissection regions (see Extended Data Fig. 1 for the
546 detail labels). **h**, An integrated overview of brain region composition, subtype, and cell numbers
547 of the major types. **i**, The number of final pass QC reads, the percentage of nonoverlapping
548 chromosome 100kb bins detected, and the percentage of GENCODE vm22 genes detected per
549 cell.

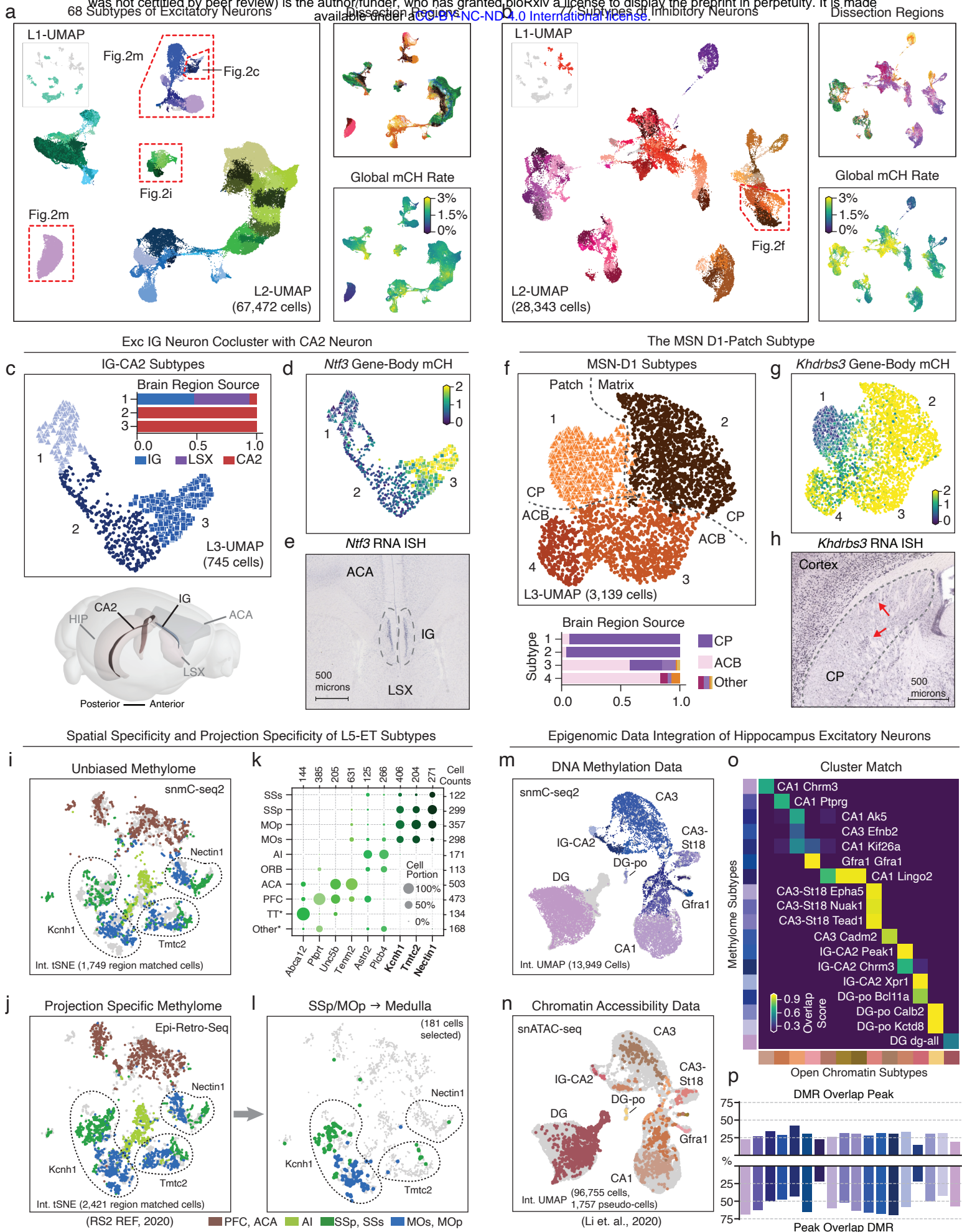


Figure 2

550 **Figure 2. Cellular and spatial epigenomic diversity of neurons.**

551 **a, b**, Level 2 UMAP of excitatory neurons (**a**) and inhibitory neurons (**b**), colored by subtypes,
552 dissection regions, and global mCH rate. **c**, Level 3 UMAP of IG-CA2 neurons colored by
553 subtypes. Barplot showing sub-region composition of the subtypes: 1) “Xpr1”, 2) “Chrm3”, and 3)
554 “Peak1”. The 3D model below illustrates these regions’ spatial relationships. **d, e**, mCH rate and
555 an ISH experiment from the Allen brain atlas (ABA) of the *Ntf3* gene. **f**, Level 3 UMAP of
556 MSN-D1 neurons colored by subtypes. Barplot showing sub-region composition of the subtypes:
557 1) “Khdrbs3”, 2) “Hrh1” 3) “Plxnc1” and 4) “Ntn1”. **g, h**, mCH rate (**g**) and an ISH experiment
558 from ABA (**h**) of the *Khdrbs3* gene. **i, j**, Integration t-SNE on L5-ET cells profiled by snmC-seq2
559 (**i**) and Epi-Retro-Seq (**j**), colored by matched dissection regions in both studies. The positions
560 of three SSp/MOp enriched subtypes are circled and labeled by their marker gene, see
561 Extended Data Fig. 4i for the full subtype labels. **k**, L5-ET subtype spatial composition, each
562 column sum to 100%. **l**, Medulla projecting neurons from SSp/MOp dissections profiled by
563 Epi-Retro-Seq. **m, n**, Integration UMAP of the hippocampus excitatory neurons profiled by
564 snmC-seq2 (**m**) and snATAC-seq (**n**, showing pseudo-cells). **o**, Overlap score matrix matching
565 the open chromatin subtypes to the methylome subtypes. **p**, Overlap of subtype matched DMR
566 and ATAC peaks (see Methods).

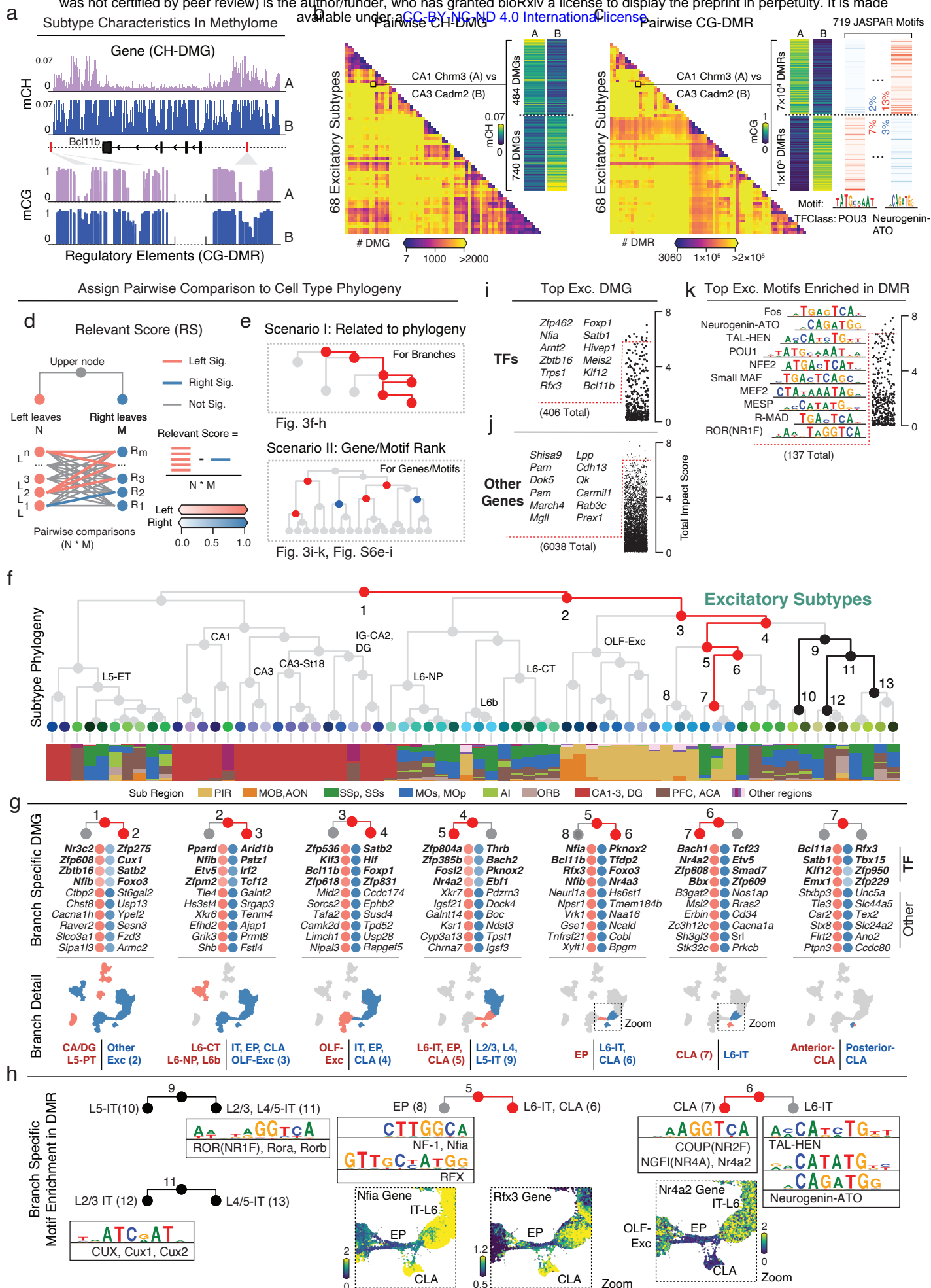


Figure 3

567 **Figure 3. Relating genes and regulatory elements to cell subtype phylogeny.**

568 **a**, Genome browser schematic of the two characteristics contained in the methylome profiles.
569 Genes used here are *Bcl11b*, with mCH and mCG rates from “CA1 Chrm3” (A) and “CA3
570 *Cadm2*” (B). **b**, **c**, Pairwise CH-DMG (**b**) or CG-DMR (**c**) counts heatmap between 68 excitatory
571 subtypes. The detailed example from the same cluster pair as in (**a**). In each DMR set, we
572 further scan the occurrence of 719 JASPAR motifs, two differentially enriched motifs from POU3
573 and Neurogenin-ATO family are shown for the example pair in (**c**). **d-e**, Schematic of impact
574 score calculation (**d**) and two scenarios of discussing impact scores (**e**). **f**, Excitatory subtype
575 phylogeny tree. Leaf nodes are colored by subtypes, and the barplot shows subtype
576 composition. The numbered nodes correspond to the panels below. **g**, Top impact scores of
577 ranked genes for the left and right branches of node 1-7. Top four TF genes are in bold, followed
578 by other protein-coding genes. The scatter plot below shows the L2-UMAP of excitatory cells
579 (from Fig. 2a) colored by cells involved in each branch. **h**, Branch specific TF motif families. The
580 zoomed UMAP plots show particular TF genes in those families, whose differential mCH rate
581 concordant with their motif enrichment. **i-k**, Total impact of TFs (**i**), TF motifs (**k**), and other
582 protein-coding genes (**j**). Top ranked items are listed on the left.

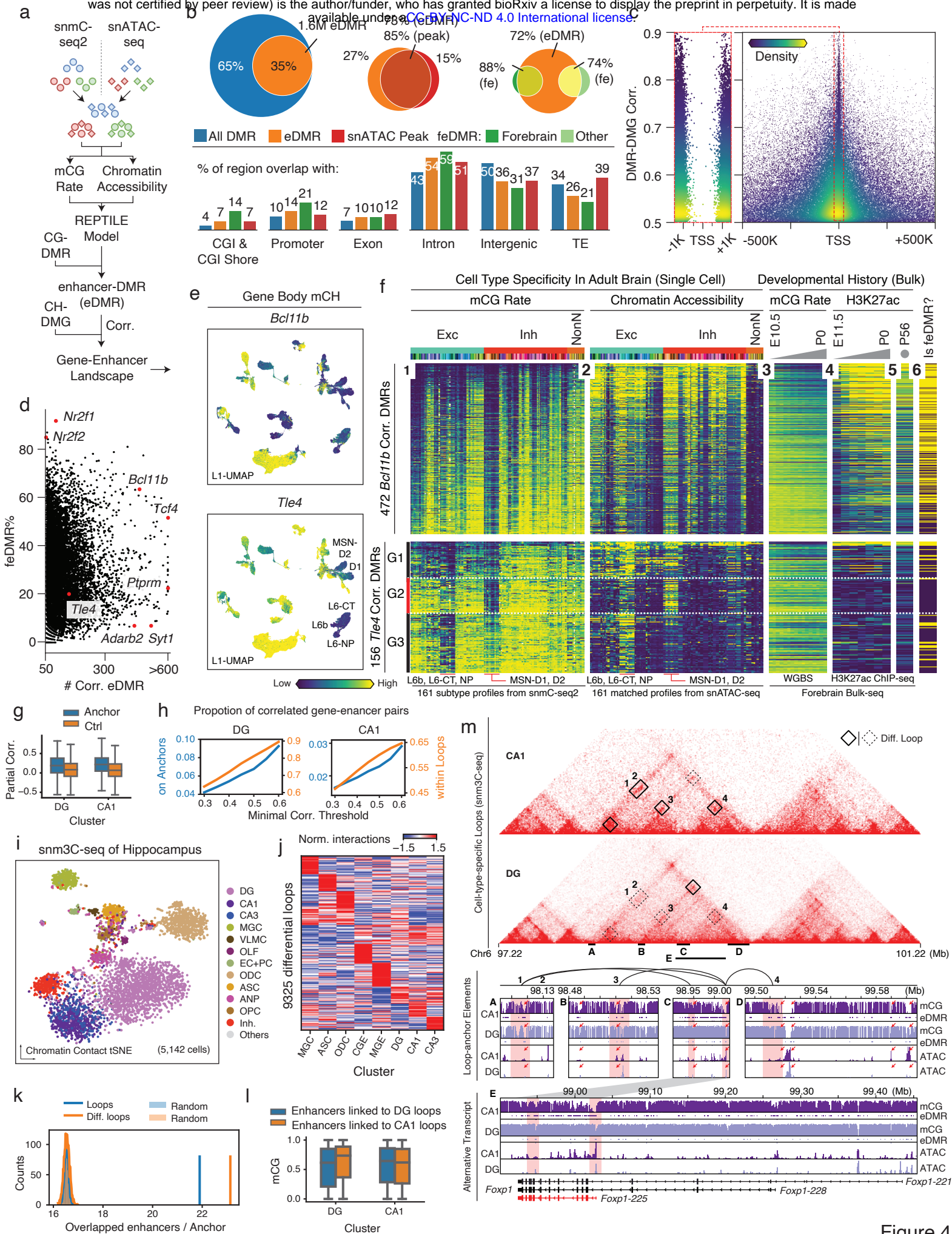


Figure 4

583 **Figure 4. Gene-enhancer landscapes in neuronal subtypes.**

584 **a**, Schematic of enhancer calling using matched DNA methylome and chromatin accessibility
585 subtype profiles. The REPTILE algorithm was deployed followed by building the gene-enhancer
586 landscape through their methylation correlation. **b**, Overlap of regulatory elements identified in
587 this study, and other epigenomic studies (snATAC-seq peaks from Li. et al, Companion
588 Manuscript 11; fetal enhancer DMRs⁶). **c**, DMR-DMG correlation (y-axis) and the distance
589 between DMR center and gene TSS (x-axis), each point is a DMR-DMG pair colored by kernel
590 density. **d**, Percentage of positively correlated eDMR that overlap with forebrain feDMR (from
591 He et al.⁶). Each point is a gene, while x-axis is the number of positively correlated eDMRs to
592 each gene. **e**, The gene body mCH rate of *Bcl11b* (top) and *Tle4* (bottom) gene. **f**, The predicted
593 enhancer landscape of *Bcl11b* (top) and *Tle4* (bottom). Each row is a correlated eDMR to the
594 gene, columns from left to right are: (1) mCG rate and (2) ATAC FPKM in 161 subtypes; (3) bulk
595 developing forebrain tissue mCG rate⁶ and (4) H3K27ac FPKM⁶⁸; (5) adult frontal cortex
596 H3K27ac²; and (6) is a feDMR or not⁶. **g**, Partial correlation between mCG of enhancers and
597 mCH of genes on separated loop anchors of DG (left) and CA1 (right) compared to random
598 anchors with comparable distance (n=4,171, 4,036, 4,326, 5,133). **h**, Proportion of loop
599 supported enhancer-gene pairs among those linked by correlation analyses surpassing different
600 correlation thresholds. **i**, t-SNE of sn-m3C-seq cells (n=5,142) colored by clusters. Cluster labels
601 were predicted using the model trained with snmC-Seq data, and some major cell-types were
602 merged. **j**, Interaction level of 9,325 differential loops in eight clusters at 25kb resolution. Values
603 shown are z-score normalized across rows and columns. **k**, Number of enhancers per loop
604 anchor (blue) or per differential loop anchor (orange) compared to randomly selected 25kb
605 regions across the genome. The permutation was repeated for 2,000 times. **l**, mCG of
606 enhancers linking to DG specific loops (blue, n=13,854) and CA1 specific loops (orange,
607 n=14,373) in DG (left) or CA1 (right). **m**, Comprehensive epigenomic signatures surrounding
608 *Foxp1*. The triangle heatmap shows CA1 and DG chromatin contacts with differential loops
609 labeled with black boxes; the next genome browser section shows detail mCG and ATAC
610 profiles near four CA1-specific loops' anchors, with red rectangles indicate loop anchor and red
611 arrows indicate notable regulatory elements; genome browser image depicts the mCG and
612 ATAC profiles at the *Foxp1* gene. The elements of boxplots in (**g**) and (**l**) are defined as: center
613 line, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range.

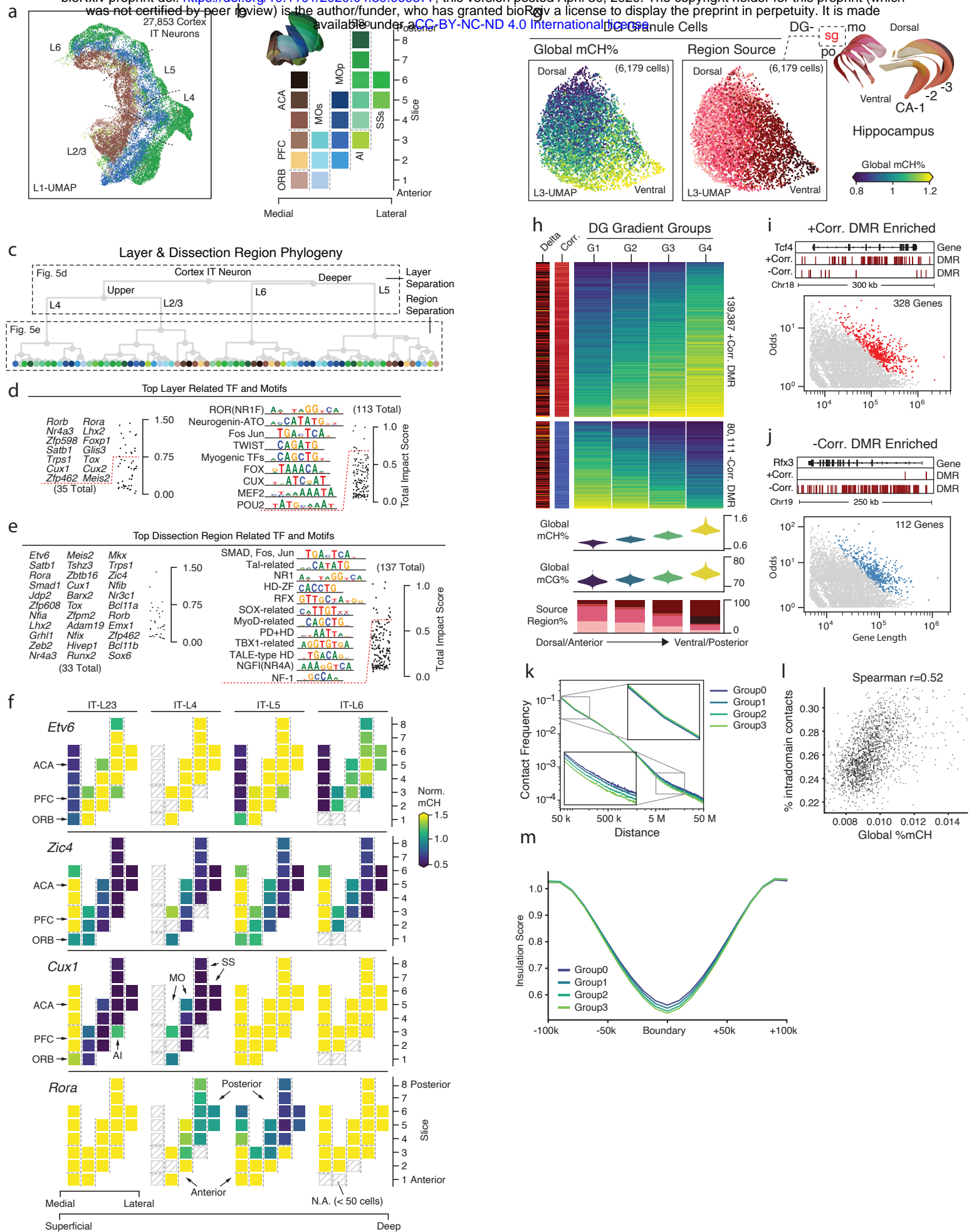


Figure 5

614 **Figure 5. Brain-wide spatial gradients of DNA methylation.**

615 **a**, L1-UMAP for cortex IT neurons colored by dissection regions. **b**, 2-D layout of twenty-one
616 dissection regions profiled in this study, colored by dissection regions. **c**, IT spatial group
617 phylogeny. The top three nodes separate four layers, and downstream nodes separate
618 dissection regions (see Extended Data Fig. 8a for details). **d**, **e**, The top layer (**d**) and dissection
619 region (**e**) related TFs and JASPAR motifs ranked by total impact score (see Methods). **f**, Gene
620 body mCH rate of TF genes in (**e**) using the same layout as (**b**). **g**, L3-UMAPs for DG granule
621 cells colored by cell global mCH rate and dissection regions. **h**, Compound figure showing four
622 DG gradient cell groups and the two groups of gradient DMR separated by their sign of
623 correlation to the cell's global mCH level. **i**, **j**, Gradient DMR enriched genes. Red dots indicate
624 genes with positive (**i**) or negative (**j**) DMR enriched in their gene body. DMRs of *Tcf4* and *Rfx3*
625 gene. Genome browser view of representative positive and negative correlated DMR enriched
626 genes respectively. **k**, Interaction frequency decays with increasing genome distances in
627 different groups. **l**, Correlation between global mCH and proportion of intra-domain contacts
628 across 1,904 DG cells. **m**, Insulation scores of 9,160 domain boundaries and flanking 100kb
629 regions.

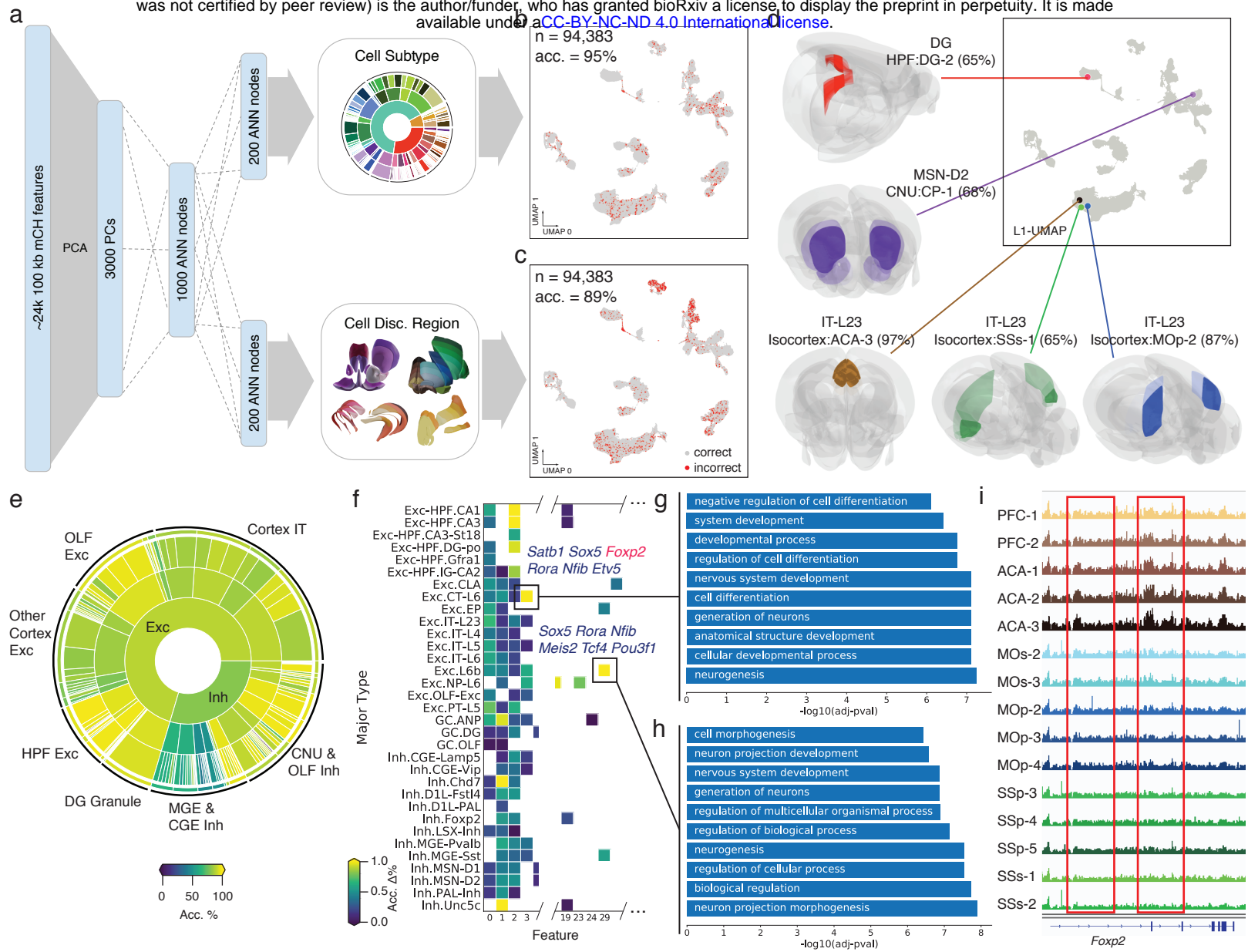


Figure 6

630 **Figure 6. A methylome-based predictive model captures both cellular and spatial**
631 **characteristics of neurons.**

632 **a**, The architecture of the artificial neural network for predicting both cell-type identity and spatial
633 origin. **b, c**, Performance of subtype (**b**) and dissection region (**c**) predictions. incorrect
634 predictions are colored in red on L1 UMAP. **d**, Examples of using the model to predict cell type
635 identity and spatial location simultaneously. The color intensity denotes the probability that the
636 cell came from the dissection region. The highest probabilities are shown in brackets. **e**, The
637 overall predictability of spatial location for each level of neuronal types (see Extended Data Fig.
638 9b for details). **f**, Feature importance evaluation for spatial origin prediction. **g, h**, GO term
639 enrichment of top-loading genes of features that are important for predicting the spatial location
640 of L6-CT (**g**) and L6b (**h**). **i**, Genome browser view of the mCH rate of the *Foxp2* gene in each
641 cortical dissection region.

642 **Methods**

643 **Mouse brain tissues**

644 All experimental procedures using live animals were approved by the Salk Institute Animal Care
645 and Use Committee under protocol number 18-00006. Adult (P56) C57BL/6J male mice were
646 purchased from Jackson Laboratories and maintained in the Salk animal barrier facility on 12 h
647 dark-light cycles with food ad-libitum for a maximum of 10 days. Brains were extracted and
648 sliced coronally at 600 μ m from the frontal pole across the whole brain (for a total of 18 slices) in
649 ice-cold dissection buffer containing 2.5mM KCl, 0.5mM CaCl₂, 7mM MgCl₂, 1.25mM NaH₂PO₄,
650 110mM sucrose, 10mM glucose, and 25mM NaHCO₃. The solution was kept ice-cold and
651 bubbled with 95% O₂ / 5% CO₂ for at least 15 min before starting the slicing procedure. Slices
652 were kept in 12-well plates containing ice-cold dissection buffers (for a maximum of 20 min) until
653 dissection aided by an SZX16 Olympus microscope equipped with an SDF PLAPO 1XPF
654 objective. Each brain region was dissected from slices along the anterior-posterior axis
655 according to the Allen Brain reference Atlas CCFv3⁶⁹ (see Extended Data Fig. 1 for the
656 depiction of a posterior view of each coronal slice). Slices were kept in ice-cold dissection media
657 during dissection and immediately frozen in dry ice for posterior pooling and nuclei production.
658 For nuclei isolation, each dissected region was pooled from 6-30 animals, and two biological
659 replicas were processed for each slice.

660 **Nuclei isolation and Fluorescence Activated Nuclei Sorting (FANS)**

661 Nuclei were isolated as previously described^{2,8}. Isolated nuclei were labeled by incubation with
662 1:1000 dilution of AlexaFluor488-conjugated anti-NeuN antibody (MAB377X, Millipore) and a
663 1:1000 dilution of Hoechst 33342 at 4°C for 1 hour with continuous shaking.
664 Fluorescence-Activated Nuclei Sorting (FANS) of single nuclei was performed using a BD Influx
665 sorter with an 85 μ m nozzle at 22.5 PSI sheath pressure. Single nuclei were sorted into each
666 well of a 384-well plate preloaded with 2 μ l of Proteinase K digestion buffer (1 μ l M-Digestion
667 Buffer, 0.1 μ l 20 μ g/ μ l Proteinase K and 0.9 μ l H₂O). The alignment of the receiving 384-well
668 plate was performed by sorting sheath flow into wells of an empty plate and making adjustments
669 based on the liquid drop position. Single-cell (1 drop single) mode was selected to ensure the
670 stringency of sorting. For each 384-well plate, columns 1-22 were sorted with NeuN+ (488+)

671 gate, and column 23-24 with NeuN- (488-) gate, reaching an 11:1 ratio of NeuN+ to NeuN-
672 nuclei.

673 **Library preparation and Illumina sequencing**

674 Detailed methods for bisulfite conversion and library preparation were previously described for
675 snmC-seq^{8,16}. The snmC-seq2 and sn-m3C-seq (see below) libraries generated from mouse
676 brain tissues were sequenced using an Illumina Novaseq 6000 instrument with S4 flowcells and
677 150 bp paired-end mode.

678 **The sn-m3C-seq specific steps of library preparation**

679 Single-nucleus methyl-3C sequencing (sn-m3C-seq) was performed as previously described¹⁸.
680 In brief, the same batch of dissected tissue samples from the dorsal dentate gyrus (DG-1 and
681 DG-2, see Supplementary Table 2), ventral dentate gyrus (DG-3 and DG-4), dorsal
682 hippocampus (CA-1 and CA-2), and ventral hippocampus (CA-3 and CA-4), were frozen in
683 liquid nitrogen. The samples were then pulverized while frozen using a mortar and pestle, and
684 then immediately fixed with 2% formaldehyde in DPBS for 10 minutes. The samples were
685 quenched with 0.2M Glycine and stored at -80C until ready for further processing. After isolating
686 nuclei as previously described¹⁸, nuclei were digested overnight with NlaIII and ligated for 4
687 hours. Nuclei were then stained with Hoechst 33342 and filtered through a 0.2µM filter and
688 sorted similarly to the snmC-seq2 samples. Libraries were generated using the snmC-seq2
689 method.

690 **Mouse brain region nomenclature**

691 The mouse brain dissection and naming of anatomical structures in this study followed the Allen
692 Mouse Brain Atlas⁶⁹. Based on the hierarchical structure of the Allen CCF, we used a three-level
693 spatial region organization to facilitate description: a. The major region, e.g., isocortex,
694 hippocampus, b. The sub-region, e.g., MOp, SSp, within isocortex, c. The dissection region,
695 e.g., MOp-1, MOp-2, within MOp. Supplementary Table 1 contains full names of all
696 abbreviations used in this study.

697 **Analysis stages**

698 The following method sections were divided into three stages. The first stage describes
699 mapping and generating files in the single-cell methylation-specific data format. The second

700 stage describes clustering, identifying differentially methylated genes (DMG), or integrating
701 other datasets, which all happened at the single-cell level. The third stage describes the
702 identification of putative cell-type-specific regulatory elements using cluster-merged
703 methylomes. Other figure-specific analysis topics may combine results from more than one
704 stage.

705 **STAGE I. MAPPING AND FEATURE GENERATION**

706 **Mapping and feature count pipeline**

707 We implemented a versatile mapping pipeline (cemba-data.rtfid.io) for all the single-cell
708 methylome based technologies developed by our group^{8,15,16}. The main steps of this pipeline
709 included: 1) Demultiplexing FASTQ files into single-cell; 2) Reads level QC; 3) Mapping; 4) BAM
710 file processing and QC; 5) final molecular profile generation. The details of the five steps for
711 snmC-seq2 were previously described¹⁶. We mapped all of the reads to the mouse mm10
712 genome. After mapping, we calculated the methylcytosine counts and total cytosine counts for
713 two sets of genomic regions in each cell. Non-overlapping chromosome 100kb bins of the
714 mm10 genome (generated by “bedtools makewindows -w 100000”), were used for clustering
715 analysis and ANN model training, and the gene body region \pm 2kb defined by the mouse
716 GENCODE vm22, were used for cluster annotation and integration with other modalities.

717 **sn-m3C-seq specific steps or read mapping and chromatin contact analysis**

718 Methylome sequencing reads were mapped following the TAURUS-MH pipeline as previously
719 described¹⁸. Specifically, reads were trimmed for Illumina adaptors and then an additional 10bps
720 was trimmed on both sides. Then R1 and R2 reads were mapped separately to the mm10
721 genome using bismark with bowtie. The unmapped reads were collected and split into shorter
722 reads representing the first 40bps, the last 40bps, and the middle part of the original reads (if
723 read length > 80bp after trimming). The split reads were mapped again using Bismark with the
724 Bowtie. The reads with MAPQ<10 were removed. To generate the methylation data, the filtered
725 bam files from split and unsplit R1 and R2 reads were deduplicated with picard and merged into
726 a single bam file. Methylypy (v1.4.2)⁷⁰ was used to generate an allc file from the bam file for
727 every single cell. To generate the Hi-C contact map, we paired the R1 and R2 bam files where
728 each read pair represents a potential contact. For generating contact files, read pairs where the
729 two ends mapped within 1kbp of each other were removed.

730 **STAGE II. CLUSTERING RELATED**

731 **Single-cell methylome data quality control and preprocessing**

732 **Cell filtering.** We filtered the cells based on these main mapping metrics: 1) mCCC rate < 0.03.
733 mCCC rate reliably estimates the upper bound of bisulfite non-conversion rate⁸, 2) overall mCG
734 rate > 0.5, 3) overall mCH rate < 0.2, 4) total final reads > 500,000, 5) bismark mapping rate >
735 0.5. Other metrics such as genome coverage, PCR duplicates rate, index ratio were also
736 generated and evaluated during filtering. However, after removing outliers with the main metrics
737 1-5, few additional outliers can be found.

738 **Feature filtering.** 100kb genomic bin features were filtered by removing bins with mean total
739 cytosine base calls < 250 or > 3000. Regions that overlap with the ENCODE blacklist⁷¹ were
740 also excluded from further analysis.

741 **Computation and normalization of the methylation rate.** For CG and CH methylation, the
742 computation of methylation rate from the methyl-cytosine and total cytosine matrices contains
743 two steps: 1) prior estimation for the beta-binomial distribution and 2) posterior rate calculation
744 and normalization per cell.

745 Step 1, for each cell we calculated the sample mean, m , and variance, v , of the raw mc rate
746 (mc / cov) for each sequence context (CG, CH). The shape parameters (α , β) of the beta
747 distribution were then estimated using the method of moments:

$$748 \alpha = m(m(1 - m)/v - 1)$$

$$749 \beta = (1 - m)(m(1 - m)/v - 1)$$

750 This approach used different priors for different methylation types for each cell and used weaker
751 prior to cells with more information (higher raw variance).

752 Step 2, We then calculated the posterior: $\hat{m}c = \frac{\alpha + mc}{\alpha + \beta + cov}$., We normalized this rate by the cell's
753 global mean methylation, $m = \alpha / (\alpha + \beta)$. Thus, all the posterior $\hat{m}c$ with 0 cov will be constant 1
754 after normalization. The resulting normalized mc rate matrix contains no NA (not available)
755 value, and features with less cov tend to have a mean value close to 1.

756 **Selection of highly variable features.** Highly variable methylation features were selected
757 based on a modified approach using the scanpy package `scanpy.pp.highly_variable_genes`
758 function⁷². In brief, the `scanpy.pp.highly_variable_genes` function normalized the dispersion of a
759 gene by scaling with the mean and standard deviation of the dispersions for genes falling into a

760 given bin for mean expression of genes. In our modified approach, we reasoned that both the
761 mean methylation level and the mean cov of a feature (100kb bin or gene) could impact *mc* rate
762 dispersion. We grouped features that fall into a combined bin of mean and cov. We then
763 normalized the dispersion within each *mean-cov* group. After dispersion normalization, we
764 selected the top 3000 features based on normalized dispersion for clustering analysis.

765 **Dimension reduction and combination of different mC types.** For each selected feature, *mc*
766 rates were scaled to unit variance and zero mean. PCA was then performed on the scaled *mc*
767 rate matrix. The number of significant PCs was selected by inspecting the variance ratio of each
768 PC using the elbow method. The CH and CG PCs were then concatenated together for further
769 analysis in clustering and manifold learning.

770 **Consensus clustering**

771 **Consensus clustering on concatenated PCs.** We used a consensus clustering approach
772 based on multiple Leiden-clustering⁷³ over K-Nearest Neighbor (KNN) graph to account for the
773 randomness of the Leiden clustering algorithms. After selecting dominant PCs from PCA in both
774 mCH and mCG matrix, we concatenated the PCs together to construct a KNN graph using
775 *scanpy.pp.neighbors* with euclidean distance. Given fixed resolution parameters, we repeated
776 the Leiden clustering 300 times on the KNN graph with different random starts and combined
777 these cluster assignments as a new feature matrix, where each single Leiden result is a feature.
778 We then used the outlier-aware DBSCAN algorithm from the scikit-learn package to perform
779 consensus clustering over the Leiden feature matrix using the hamming distance. Different
780 epsilon parameters of DBSCAN are traversed to generate consensus cluster versions with the
781 number of clusters that range from minimum to the maximum number of clusters observed in
782 the multiple Leiden runs. Each version contained a few outliers that usually fall into three
783 categories: 1. Cells located between two clusters that had gradient differences instead of clear
784 borders, e.g., border of IT layers; 2. Cells with a low number of reads that potentially lack
785 information in essential features to determine the specific cluster. 3. Cells with a high number of
786 reads that were potential doublets. The amount of type 1 and 2 outliers depends on the
787 resolution parameter and is discussed in the choice of the resolution parameter section. The
788 type 3 outliers were very rare after cell filtering. The supervised model evaluation below then
789 determined the final consensus cluster version.

790 **Supervised model evaluation on the clustering assignment.** For each consensus clustering
791 version, we performed a Recursive Feature Elimination with Cross-Validation (RFECV)⁷⁴
792 process from the scikit-learn package to evaluate clustering reproducibility. We first removed the
793 outliers from this process, and then we held out 10% of the cells as the final testing dataset. For
794 the remaining 90% of the cells, we used ten-fold cross-validation to train a multiclass prediction
795 model using the input PCs as features and *sklearn.metrics.balanced_accuracy_score*⁷⁵ as an
796 evaluation score. The multiclass prediction model is based on *BalancedRandomForestClassifier*
797 from the *imblearn* package that accounts for imbalanced classification problems⁷⁶. After training,
798 we used the 10% testing dataset to test the model performance using the
799 *balanced_accuracy_score* score. We kept the best model and corresponding clustering
800 assignments as the final clustering version. Finally, we used this prediction model to predict
801 outliers' cluster assignments, we rescued the outlier with prediction probability > 0.3, otherwise
802 labeling them as outliers.

803 **Manifold learning for visualization.** In each round of clustering analysis, the t-SNE^{77,78} and
804 UMAP¹⁹ embedding were run on the PC matrix the same as the clustering input using the
805 implementation from the scanpy⁷² package. The coordinates from both algorithms were in
806 Supplementary Table 5.

807 **Choice of resolution parameter.** Choosing the resolution parameter of the Leiden algorithm is
808 critical for determining the final number of clusters. We selected the resolution parameter by
809 three criteria: 1. The portion of outliers < 0.05 in the final consensus clustering version. 2. The
810 ultimate prediction model accuracy > 0.9. 3. The average cell per cluster ≥ 30 , which controls
811 the cluster size to reach the minimum coverage required for further epigenome analysis such as
812 DMR calls. All three criteria prevented the over-splitting of the clusters; thus, we selected the
813 maximum resolution parameter under meeting the criteria using a grid search.

814 **Pairwise Differential Methylated Gene (DMG) identification**

815 We used a pairwise strategy to calculate DMGs for each pair of clusters within the same round
816 of analysis. We used the gene body ± 2 kb regions of all the protein-coding and long non-coding
817 RNA genes with evidence level 1 or 2 from the mouse GENCODE vm22. We used the per
818 cluster normalized mCH rate (same as the "Computation and normalization of the methylation
819 rate" in the clustering step above) to calculate markers between all neuronal clusters. We
820 compared non-neuron clusters separately using the normalized mCG rate. For each pairwise
821 comparison, we used the Wilcoxon test to select genes that have a significant decrease

822 (hypo-methylation). Marker gene was chosen based on adjusted P-value $< 1e-3$ with multitest
823 correction using Benjamini-Hochberg procedure, delta normalized methylation level change $<$
824 -0.5 (hypo-methylation), AUROC > 0.8 . We required each cluster to have ≥ 5 DMGs compared
825 to any other cluster. Otherwise, the smallest cluster that did not meet this criterion was merged
826 to the closest cluster based on cluster centroids euclidean distance in the PC matrix that was
827 used for clustering. Then the marker identification process was repeated until all clusters found
828 enough marker genes.

829 **Three-level of iterative clustering analysis**

830 Based on the consensus clustering steps described above, we used an iterative approach to
831 cluster the data into three levels of categories. In the first level termed CellClass, clustering
832 analysis is done using all cells, and then manually merged into three canonical classes:
833 excitatory neurons, inhibitory neurons, and non-neurons based on marker genes. Then within
834 each CellClass, we performed all the preprocessing and clustering steps again to get clusters
835 for the MajorType level using the same stop criteria. And within each MajorType, we obtained
836 clusters for the SubType level. All clusters' annotations and relationships are in Supplementary
837 Table 4.

838 **Subtype phylogeny tree**

839 To build the phylogeny tree of subtypes, we selected the top 50 genes that show the most
840 number of significant changes for each subtypes' pairwise comparisons. We then used the
841 union of these genes from all subtypes and obtained a total of 2503 unique genes. We
842 calculated the median mCH rate level of these genes in each subtype and applied bootstrap
843 resampling based hierarchical clustering with average linkage and the correlation metric using
844 the R package pvclust⁷⁹ (v.2.2).

845 **Impact Score and Total Impact Score**

846 We defined the Impact Score (IS) as a way to summarize pairwise comparisons for two subtype
847 groups, where one group A contains M clusters, the other group B contains N clusters. For each
848 gene or motif, the number of total related pairwise comparisons is $M \times N$, the number of
849 significant comparisons with desired change (hypo-methylation for gene or enrichment for motif)
850 in group A is a , and in group B is b . The IS is then calculated as $IS_A = \frac{a-b}{M \times N}$ and

851 $IS_B = \frac{b-a}{M \times N}$ for the two directions. For either group, IS ranges from -1 to 1 and 0 means no
852 impact, 1 means full impact, -1 means full impact in the other group.

853 We explored two scenarios of using the IS to describe cluster characteristics (Fig. 3e). The first
854 scenario is considering each pair of branches in the subtype phylogeny tree as group A and
855 group B. Thus the IS can quantify and rank genes or motifs to the upper nodes based on the
856 leaves' pairwise comparisons (Fig. 3i-h). The second scenario was a summarization of the total
857 impact for specific genes or motifs regarding the phylogeny tree based on the calculation in the
858 first scenario. In a subtype phylogeny tree with n subtypes, the total non-singleton node was
859 $n-1$, and each node i had a height h_i and associated IS_A for one of the branches ($IS_B = -IS_A$).
860 The node height weighted total IS was then calculated by:

$$861 \quad IS_{total} = \sum_{i=1}^{n-1} h_i \cdot |IS_A|$$

862 The larger total IS indicated a gene or motif show more cell-type-phylogeny related significant
863 changes. The height weight intended to focus on the higher branches (major cell type
864 separations), but the total IS can also be calculated in a sub-tree or any combination of
865 interests, to rank gene and motifs most related to that combination (See Figure 5 related
866 methods about calculating layer and region total IS from the same tree).

867 **Integration with snATAC-seq data**

868 A portion of exact same brain tissue sample used in this study for methylome profiling was also
869 processed with snATAC-seq in a parallel study about chromatin accessibility (Li et al.,
870 Companion Manuscript # 11). The final high-quality snATAC-seq cells were assigned to 160
871 chromatin accessibility clusters (a-clusters). The snATAC specific data analysis steps are
872 described in Li et al. Here, we performed cross-modality data integration and label-transferring
873 to assign the 160 a-clusters to the 161 methylome subtypes in the following steps:

- 874 1. We manually grouped both modalities into five integration groups (e.g., all IT neurons as
875 a group) and only performed the integration of cells within the same group to decrease
876 computation time. These groups were distinct in the clustering steps of both modalities
877 and can be matched with great confidence using known marker genes. Step 2-6 were
878 repeated for each group, see Extended Data Fig. 5 for the group design.
- 879 2. We used a similar approach as described above to identify pairwise differential
880 accessible genes (DAG) between all pairs of snATAC-seq clusters. The cutoff for DAG is
881 adjusted P-value < 1e-3, fold change > 2, AUROC > 0.8.

- 882 3. We then gather DMGs from related subtypes' comparisons in the same group. Both
883 DAGs and DMGs were filtered by recurring in > 5 pairwise comparisons. The
884 intersection of the remaining genes was used as the feature set of integration.
- 885 4. After identifying DAGs using cell level snATAC-seq data, we merged the snATAC-seq
886 cells into pseudo-cells to increase snATAC-seq data coverage. Within each a-cluster, we
887 did a K-means clustering ($K = \#$ of cells in that cluster / 50) on the same PCs that were
888 used in snATAC-seq clustering. We discarded (about 5% of the cells) small K-means
889 clusters with < 10 cells and merged each remaining K-means cluster into a pseudo-cell.
890 Each pseudo-cell had about 50 times more fragments than a single cell.
- 891 5. We then used the MNN based Scanorama³⁹ method with default parameters to integrate
892 the snmC-seq cells and snATAC-seq pseudo-cells using genes from 3. After Scanorama
893 integration, we did co-clustering on the integrated PC matrix using the clustering
894 approaches described above.
- 895 6. We used the intermediate clustering assignment from 5 to calculate the overlap score
896 (see the section below) between the original methylome subtypes and the a-clusters. We
897 used the overlap score > 0.3 to assign snATAC-seq clusters to each methylome subtype.
898 For those subtypes that have no match under this threshold, we assign the top a-cluster
899 ranked by the overlap score.

900 **Overlap Score (OS)**

901 We used the overlap score to match a-cluster and methylome subtypes together. The overlap
902 score range from 0 to 1 was defined as the sum of the minimum proportion of samples in each
903 cluster that overlapped within each co-cluster²⁶. A higher score between one methylome
904 subtype and one a-cluster indicates they consistently co-clustered within one or more
905 co-clusters. Besides matching clusters in integration analysis, the OS was also used in two
906 other cases: 1. To quantify replicates and region overlaps over methylome subtypes (Extended
907 Data Fig. 2e-g); 2. To quantify the overlap of each L5-ET subtype overlapping with “soma
908 location” and “projection target” labels from epi-retro-seq cells (Extended Data Fig. 4k) through
909 integration with the epi-retro-seq dataset.

910 **STAGE III. CELL-TYPE-SPECIFIC REGULATORY ELEMENTS**

911 **Differentially Methylated Region (DMR) analysis**

912 After clustering analysis, we used the subtype cluster assignments to merge single-cell ALLC
913 files into the pseudo-bulk level and then used methylpy⁷⁰ *DMRfind* function to calculate mCG
914 DMRs across all clusters. The base calls of each pair of CpG sites were added before analysis.
915 In brief, the methylpy function used a permutation-based root-mean-square test of
916 goodness-of-fit to identify differentially methylated sites simultaneously across all samples
917 (subtypes in this case), and then merge the DMS within 250bp into DMR. We further excluded
918 DMS calls that have low absolute mCG rate difference by using a robust-mean-based approach.
919 For each DMR that was merged from the DMS, we order all the samples by their mCG rate and
920 calculate the robust mean m using the samples between 25th and 75th percentiles. We then
921 reassign hypo-DMR and hyper-DMR to each sample when a region met two criteria: 1) the
922 sample mCG rate of this DMR is lower than $(m - 0.3)$ for hypo-DMR or $(m + 0.3)$ for hyper-DMR,
923 and 2) the DMR is originally a significant hypo- or hyper-DMR in that sample judged by
924 methylpy. DMRs without any hypo- or hyper-DMR assignment were excluded from further
925 analyses.

926 **Enhancer prediction using DNA methylation and chromatin accessibility**

927 We performed enhancer prediction using the REPTILE⁵¹ algorithm. The REPTILE is a
928 random-forest-based supervised method that incorporates different sources of epigenomic
929 profiles with base-level DNA methylation data to learn and then distinguish the epigenomic
930 signatures of enhancers and genomic background. We trained the model in a similar way as in
931 the previous studies^{6,51} using CG methylation, chromatin accessibility of each subtype, and
932 mouse embryonic stem cells (mESC). The model was first trained on mESC data and then
933 predicted a quantitative score we termed enhancer score for each subtype's DMRs. The
934 positives were 2kb regions centered at the summits of top 5,000 EP300 peaks in mESCs.
935 Negatives include randomly chosen 5,000 promoters and 30,000 2kb genomic bins. The bins
936 have no overlap with any positive region or gene promoter⁶.
937 Methylation and chromatin accessibility profiles in bigwig format for mESC were from the GEO
938 database (GSM723018). The mCG rate bigwig file was generated from subtype-merged ALLC
939 files using the ALLCools package (<https://github.com/lhqing/ALLCools>). For chromatin

940 accessibility of each subtype, we merged all fragments from snATAC-seq cells that were
941 assigned to this subtype in the integration analysis and used “*deeptools bamcoverage*” to
942 generate CPM normalized bigwig files. All bigwig file bin sizes were 50bp.

943 **Motif enrichment analysis**

944 We used 719 motif PWMs from the JASPAR 2020 CORE vertebrates database⁴³, where each
945 motif was able to assign corresponding mouse TF genes. The specific DMR sets used in each
946 motif enrichment analysis are described in figure specific methods below. For each set of
947 DMRs, we standardized the region length to the center \pm 250bp and used the FIMO tool from
948 the MEME suite⁸⁰ to scan the motifs in each enhancer with the log-odds score p-value $< 10^{-6}$ as
949 the threshold. To calculate motif enrichment, we use the adult non-neuronal mouse tissue
950 DMRs⁶⁵ as background regions unless expressly noted. We subtracted enhancers in the region
951 set from the background, and then scanned the motifs in background regions using the same
952 approach. We then used Fisher’s exact test to find motifs enriched in the region set, and the
953 Benjamini-Hochberg procedure to correct multiple tests. We used the TFClass⁸¹ classification to
954 group TFs with similar motifs.

955 **DMR-DMG partial correlation**

956 To calculate DMR-DMG partial correlation, we used the mCG rate of DMRs and the mCH rate of
957 DMGs in each neuronal subtypes. We first used linear regression to regress out variance due to
958 global methylation difference (using `scanpy.pp.regress_out` function), then use the residual
959 matrix to calculate Pearson correlation between DMR and DMG pairs where the DMR center is
960 within 1Mb the DMG’s TSSs. To generate the null distribution, we shuffled the subtype orders in
961 both matrices and recalculated all pairs 100 times.

962 **Identification of loops and differential loops from sn-m3C-seq data**

963 After merging the chromatin contacts from cells belonging to the same type, we generated a .hic
964 file of the cell-type with *Juicer tools pre. HICCUPS*⁸² was used to identify loops in each cell-type.
965 The loops from eight major cell-types were concatenated and deduplicated, and used as the
966 total samples for differential loop calling. A loop-by-cell matrix was generated, where each
967 element represents the number of contacts supporting each loop in each cell. The matrix was
968 used as input of EdgeR to identify differential interactions with ANOVA tests. Loops with FDR $<$
969 $1e-5$ and min-max fold-change > 2 were used as differential loops.

970 **FIGURE-SPECIFIC METHODS**

971 **Figure 1 related**

972 **3D model of dissection regions (Fig. 1d-g).** We created in-silico dissection regions based on
973 the Allen CCFv3⁶⁹ 3D model using blender 2.8 that precisely follows our dissection plan. To
974 ease visualization of all different regions, we modified the layout and removed some of the
975 symmetric structures, but all the actual dissections were applied symmetrically to both
976 hemispheres.

977 **Calculating the genome feature detected ratio (Fig. 1i).** The detected ratio of chromosome
978 100kb bins and gene bodies is calculated as the percentage of bins with > 20 total cytosine
979 coverage. Non-overlapping chromosome 100kb bins generated by “bedtools makewindows -w
980 100000”; gene body definition from the GENCODE vm22 GTF file.

981 **Figure 2 related**

982 **Integration with epi-retro-seq L5-ET cells (Fig. 2i-l, Extended Data Fig. 4h-j).** Epi-retro-seq
983 is an snmC-seq2 based method that combines retrograde AAV labeling (Companion Manuscript
984 # 11)³⁶. The L5-ET cells’ non-overlapping chromosome 100kb bin matrix gathered by the
985 epi-retro-seq dataset was concatenated with all the L5-ET cells from this study to do
986 co-clustering and embedding as described in “STAGE II” above. We then calculated the OS
987 between subtypes in this study and the “soma location” or “projection target” labels of
988 epi-retro-seq cells. The first OS helped quantify how consistent the spatial location is between
989 the two studies, the second OS allowed us to impute the projection targets of subtypes in this
990 study.

991 **Figure 3 related**

992 **Pairwise DMR and motif enrichment analysis (Fig. 3c, h).** The total subtype DMRs were
993 identified as described in “STAGE III” via comparing all subtypes. We then assigned DMRs to
994 each subtype pair if the DMRs were: 1) significantly hypomethylated in only one of the subtypes;
995 and 2) the mCG rate difference between the two subtypes > 0.4. Each subtype pair was
996 associated with two exclusive sets of pairwise DMRs. We carried out motif enrichment analysis
997 as described in “STAGE III” on each DMR set using the other set as background. Motifs

998 enriched in either direction were then used to calculate the impact score and were associated
999 with upper nodes of the phylogeny.

1000 **Figure 4 related**

1001 **Overlapping eDMR with genome regions (Fig. 4b).** The cluster-specific snATAC-seq peaks
1002 were identified in Li et al. (Companion Manuscript # 11). We used “bedtools merge” to
1003 aggregate the total non-overlap peak regions, and “bedtools intersect” to calculate the overlap
1004 between peaks and eDMRs. The developing forebrain and other tissue feDMR were identified in
1005 He et al.⁶ using methylC-seq⁸³ for bulk whole-genome bisulfite sequencing (WGBS-seq). All of
1006 the genome features used in Fig. 4b were defined as in He et al, except using an updated
1007 mm10 CGI region and RepeatMaster transposable elements lists (UCSC table browser
1008 downloaded on 09/10/2019, and the GENCODE vm22 gene annotation).

1009 **Assembly epigenome card (Fig. 4f) of the gene-enhancer landscape.** The eDMRs for each
1010 gene selected by eDMR-Gene correlation > 0.3. Sections of the heatmaps in Fig. 4f were
1011 gathered by: 1) mCG rate of each eDMR in 161 subtypes from this study, 2) snATAC
1012 subtype-level FPKM of each eDMR in the same subtype orders. The subtype snATAC profiles
1013 were merged from integration results as described in “STAGE II”, 3) mCG rate of each eDMR in
1014 forebrain tissue during ten developing time points from E10.5 to P0, data from He et al.⁶, 4)
1015 H3K27ac FPKM of each eDMR in 7 developing time points from E11.5 to P0, data from Gorokin
1016 et al.⁶⁸, 5) H3K27ac FPKM of each eDMR in P56 frontal brain tissue, data from Lister et al.² and
1017 6) eDMR is overlapped with forebrain feDMR using “bedtools intersect”.

1018 **Embedding of cells with chromosome interactions (Fig. 4i).** scHiCluster⁸⁴ was used to
1019 generate the t-SNE embedding of the sn-m3C-seq cells. Specifically, a contact matrix at 1Mb
1020 resolution was generated for each chromosome of each cell. Then the matrices were smoothed
1021 by linear convolution with pad = 1 and random walk with restart probability = 0.5. The top 20th
1022 percentile of strongest interactions on the smoothed map was extracted, binarized, and used for
1023 PCA. The first 20 PCs were used for t-SNE.

1024 **Figure 5 related**

1025 **IT layer-dissection-region group DMG and DMR analysis (Fig. 5a-f).** In order to collect
1026 enough cells for dissection region analysis, we only used the major types (which corresponds to
1027 L2/3, L4, L5, and L6) of IT neurons. We grouped cells into layer-dissection-region groups and
1028 kept groups with > 50 cells in further analysis (Extended Data Fig. 8b). We performed pairwise

1029 DMG, DMR, and motif enrichment analysis the same as the subtype analysis in Fig. 3, but using
1030 the layer-dissection-region group labels. We then built a spatial phylogeny for these groups and
1031 calculated impact scores based on it. To rank layer-related or dissection-region-related genes
1032 and motifs separately, we used two sets of the branches (Extended Data Fig. 8a, upper set for
1033 layers, lower set for regions) in the phylogeny and calculated two total impact scores using
1034 equations in above.

1035 **DG cell group and gradient DMR analysis (Fig. 5h).** DG cells were grouped into four
1036 even-sized groups according to cells' global mCH rates, and cutoff thresholds were 0.45%,
1037 0.55%, and 0.69%. We then randomly chose 400 cells from each group to call gradient-DMRs
1038 using methods described in "STAGE II". To ensure the DMRs identified between intra-DG
1039 groups were not due to stochasticity, we also randomly sampled 15 groups of 400 cells from all
1040 DG cells regardless of their global mCH and called DMR among them as control-DMRs (2,003
1041 using the same filtering condition). Only 0.04% of gradient-DMRs overlapped with the
1042 control-DMRs and thus were removed from further analysis. Pearson correlation coefficients
1043 (ρ 's) of mCG rates of each gradient-DMR were calculated against a linear sequence [1, 2, 3, 4]
1044 to quantify the gradient trend. DMRs with $\rho < -0.75$ or $\rho > 0.75$ were considered to be significantly
1045 correlated. Weakly correlated DMRs (10%) were not included in further analysis.

1046 **DMR/DMS enriched genes (Fig. 5i-j).** To investigate the correlated DMR/DMSs enrichment in
1047 specific gene bodies, we compare the number of DMS and cytosine inside the gene body with
1048 the number of DMS and cytosine in the ± 1 Mb regions through Fisher's exact test. We chose
1049 genes passing both criteria: 1) adjusted P-value < 0.01 with multitest correction using
1050 Benjamini-Hochberg procedure, and 2) overlap with > 20 DMS. Gene ontology analysis of
1051 DMR/DMS enriched genes was carried out using GOATOOLS⁸⁵ all protein coding genes with
1052 gene body length > 5 kb were used as background to prevent gene length bias.

1053 **Compartment strength analysis.** We Z-score normalized the total chromosome contacts in
1054 each 1 Mb bin of DG contact matrix, and the bins with normalized coverage between -1 and 2
1055 were kept for the analysis. After filtering, the PC1 of genome-wide KR normalized contact matrix
1056 was used as the compartment score. The score was divided into 50 categories with equal sizes
1057 from low to high, and bins were assigned into the categories. The intra-chromosomal
1058 observation/expectation (ove) matrices of each group were used to quantify the compartment
1059 strength. We computed the average ove values within each pair of categories to generate the
1060 50x50 saddle matrices. The compartment strength was computed with the average of upper left

1061 and lower right 10x10 matrices divided by the average of upper right and lower left 10x10
1062 matrices⁸⁶.

1063 **Domain analysis.** We identified 4,580 contact domains at 10kb resolution in DG using
1064 *Arrowhead*⁸². For bin i , the insulation score is computed by

$$1065 I_i = \frac{\text{mean}_{i-10 \leq j' < i; i \leq j' < i+10} A_{i'j'}}{\max(\text{mean}_{i-10 \leq j' < i; i-10 \leq j' < i} A_{i'j'}, \text{mean}_{i \leq j' < i+10; i \leq j' < i+10} A_{i'j'})}$$

1066 , where A is the oved of KR normalized matrices. For each group, insulation scores of domain
1067 boundaries and 100kb flanking regions were computed and averaged across all boundaries.

1068 **Figure 6 related**

1069 **Prediction model description.** To reduce the computing complexity, we applied principal
1070 component analysis (PCA) on the dataset of 100kb-bin-mCH features to obtain the first 3,000
1071 principal components (PCs), which retains ~ 61% variance of the original data. These 3,000
1072 PCs were then used to train and test the predicting model. We used an artificial neural network
1073 with two hidden layers to predict cell subtypes and their dissection regions simultaneously (Fig.
1074 6a). The input layer contains 3,000 nodes, followed by a shared layer with 1,000 nodes. The
1075 shared layer is further connected simultaneously to two branch hidden layers of the subtype of
1076 the dissection region, each containing 200 nodes. Branch hidden layers are followed by the
1077 corresponding one-hot encoding output layers. We used 5-fold cross-validation to access the
1078 model performance. During the training, we applied the dropout technique⁸⁷ with a dropout rate
1079 $p=0.5$ on each hidden layer to prevent overfitting. Adam optimization⁸⁸ was used to train the
1080 network with a cross-entropy loss function. The training epoch number and batch size are 10
1081 and 100, respectively. The training and testing processes were conducted via TensorFlow 2.0⁸⁹.

1082 **Model performance.** For each single cell input, the two output layers generate two probabilistic
1083 vectors as the prediction results for cell subtypes and dissection regions, respectively. The
1084 subtype and dissection region label with highest probabilities were used as the prediction results
1085 for each cell to calculate accuracy. When calculating the cell dissection region accuracy (Fig.
1086 6c), we defined two kinds of accuracy with different stringency: 1) the exact accuracy using the
1087 predicted label, and 2) the fuzzy accuracy using predicted labels or its potential overlap
1088 neighbors. The potential overlap neighbors curated based on Allen CCF (Extended Data Fig.
1089 9a, Supplementary Table 2) stood for adjacent regions of a particular dissection region. The
1090 exact accuracy of the ANN model is 69%, and the fuzzy accuracy is 89%. To evaluate how
1091 much of the dissection region accuracy was improved via ANN, we calculated fuzzy accuracy

1092 just based on naive guesses in each subtype based on the dissection region composition (gray
1093 dots in Extended Data Fig. 9b).

1094 **Biological feature importance for dissection region prediction.** To assess what DNA
1095 regions store information of cell spatial origins that is distinguishable using our model, we
1096 evaluated the PC feature importance by examining how permutation of each PC feature across
1097 cells affects prediction accuracy. We did five permutations for each feature and used the
1098 average accuracy decreasing as PC feature importance. For a given cell type, we examined
1099 genes contained in the 100kb bins with the top 1% PCA factor loadings for the most important
1100 PC feature.

1101 **Data availability**

1102 Single cell raw and processed data included in this study were deposited to NCBI GEO/SRA
1103 with accession number GSE132489 and to the NeMO archive: <https://portal.nemoarchive.org/>.
1104 Cluster merged methylome profiles can be visualized at
1105 http://neomorph.salk.edu/mouse_brain.php.

1106 **Code availability**

1107 The mapping pipeline for snmC-seq2 data: <https://cemba-data.readthedocs.io/en/latest/>; The
1108 ALLCools package for post-mapping analysis and snmC-seq2 related data structure:
1109 <https://github.com/lhqing/ALLCools>; The jupyter notebooks for reproducing specific analysis:
1110 https://github.com/lhqing/mouse_brain_2020.

1111 **Supplementary Tables**

1112 Supplementary Table 1. Glossary table for all the abbreviations.

1113 Supplementary Table 2. Metadata of the 45 brain dissection regions.

1114 Supplementary Table 3. Summary of cell numbers.

1115 Supplementary Table 4. Cell type names and annotations.

1116 Supplementary Table 5. Cell metadata and manifold learning coordinates.

1117 Supplementary Table 6. Cell type by dissection region cell counts.

1118 Supplementary Table 7. snATAC-seq clusters matching to subtypes.

1119 Reference

- 1120 1. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right
1121 time. *Science* **361**, 1336–1340 (2018).
- 1122 2. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development.
1123 *Science* **341**, 1237905 (2013).
- 1124 3. Fagiolini, M., Jensen, C. L. & Champagne, F. A. Epigenetic influences on brain
1125 development and plasticity. *Curr. Opin. Neurobiol.* **19**, 207–212 (2009).
- 1126 4. Lavery, L. A. *et al.* Losing Dnmt3a dependent methylation in inhibitory neurons impairs
1127 neural function by a mechanism impacting Rett syndrome. *Elife* **9**, (2020).
- 1128 5. Li, J., Pinto-Duarte, A., Zander, M., Lai, C. Y. & Osteen, J. Polycomb-mediated repression
1129 compensates for loss of postnatal DNA methylation in excitatory neurons. *bioRxiv* (2019).
- 1130 6. He, Y. *et al.* Spatiotemporal DNA Methylome Dynamics of the Developing Mammalian
1131 Fetus. *bioRxiv* 166744 (2017) doi:10.1101/166744.
- 1132 7. Mo, A. *et al.* Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*
1133 **86**, 1369–1384 (2015).
- 1134 8. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in
1135 mammalian cortex. *Science* **357**, 600–604 (2017).
- 1136 9. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human
1137 transcription factors. *Science* **356**, eaaj2239 (2017).
- 1138 10. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA
1139 methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).
- 1140 11. Stricker, S. H., Köferle, A. & Beck, S. From profiles to function in epigenomics. *Nat. Rev.*
1141 *Genet.* **18**, 51–66 (2017).

- 1142 12. Gabel, H. W. *et al.* Disruption of DNA-methylation-dependent long gene repression in Rett
1143 syndrome. *Nature* **522**, 89–93 (2015).
- 1144 13. Chen, L. *et al.* MeCP2 binds to non-CG methylated DNA as neurons mature, influencing
1145 transcription and the timing of onset for Rett syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **112**,
1146 5509–5514 (2015).
- 1147 14. Stroud, H. *et al.* Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic
1148 States. *Cell* **171**, 1151–1164.e16 (2017).
- 1149 15. Luo, C. *et al.* Single nucleus multi-omics links human cortical cell regulatory genome
1150 diversity to disease risk variants. *bioRxiv* 2019.12.11.873398 (2019)
1151 doi:10.1101/2019.12.11.873398.
- 1152 16. Luo, C. *et al.* Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.*
1153 **9**, 3824 (2018).
- 1154 17. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse
1155 forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439
1156 (2018).
- 1157 18. Lee, D.-S. *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in
1158 single human cells. *Nat. Methods* **16**, 999–1006 (2019).
- 1159 19. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection
1160 for Dimension Reduction. *arXiv [stat.ML]* (2018).
- 1161 20. Zhao, C., Deng, W. & Gage, F. H. Mechanisms and functional implications of adult
1162 neurogenesis. *Cell* **132**, 645–660 (2008).
- 1163 21. Carleton, A., Petreanu, L. T., Lansford, R., Alvarez-Buylla, A. & Lledo, P.-M. Becoming a
1164 new neuron in the adult olfactory bulb. *Nat. Neurosci.* **6**, 507–518 (2003).
- 1165 22. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas.

- 1166 *Nature* **563**, 72–78 (2018).
- 1167 23. Yao, Z. *et al.* An integrated transcriptomic and epigenomic atlas of mouse primary motor
1168 cortex cell types. *bioRxiv* 2020.02.29.970558 (2020) doi:10.1101/2020.02.29.970558.
- 1169 24. Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal
1170 formation. *bioRxiv* 2020.03.30.015214 (2020) doi:10.1101/2020.03.30.015214.
- 1171 25. Huang, Z. J. & Paul, A. The diversity of GABAergic neurons and neural communication
1172 elements. *Nat. Rev. Neurosci.* **20**, 563–572 (2019).
- 1173 26. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse
1174 cortex. *Nature* (2019) doi:10.1038/s41586-019-1506-7.
- 1175 27. Krienen, F. M. *et al.* Innovations in Primate Interneuron Repertoire. *bioRxiv* 709501 (2019)
1176 doi:10.1101/709501.
- 1177 28. Dudek, S. M., Alexander, G. M. & Farris, S. Rediscovering area CA2: unique properties and
1178 functions. *Nat. Rev. Neurosci.* **17**, 89–102 (2016).
- 1179 29. San Antonio, A., Liban, K., Ikrar, T., Tsyganovskiy, E. & Xu, X. Distinct physiological and
1180 developmental properties of hippocampal CA2 subfield revealed by using anti-Purkinje cell
1181 protein 4 (PCP4) immunostaining. *J. Comp. Neurol.* **522**, 1333–1354 (2014).
- 1182 30. Fukaya, M., Yamazaki, M., Sakimura, K. & Watanabe, M. Spatial diversity in gene
1183 expression for VDCCgamma subunit family in developing and adult mouse brains.
1184 *Neurosci. Res.* **53**, 376–383 (2005).
- 1185 31. Phillips, H. S., Hains, J. M., Laramée, G. R., Rosenthal, A. & Winslow, J. W. Widespread
1186 expression of BDNF but not NT3 by target areas of basal forebrain cholinergic neurons.
1187 *Science* **250**, 290–294 (1990).
- 1188 32. Adamek, G. D., Shipley, M. T. & Sanders, M. S. The indusium griseum in the mouse:
1189 architecture, Timm's histochemistry and some afferent connections. *Brain Res. Bull.* **12**,

- 1190 657–668 (1984).
- 1191 33. Heimer, L. & Wilson, R. D. The subcortical projections of the allocortex: similarities in the
1192 neural connections of the hippocampus, the piriform cortex and the neocortex. Santini M,
1193 editor. *Perspectives in neurobiology* (1975).
- 1194 34. Voorn, P., Vanderschuren, L. J. M. J., Groenewegen, H. J., Robbins, T. W. & Pennartz, C.
1195 M. A. Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci.* **27**,
1196 468–474 (2004).
- 1197 35. Smith, J. B. *et al.* Genetic-Based Dissection Unveils the Inputs and Outputs of Striatal
1198 Patch and Matrix Compartments. *Neuron* **91**, 1069–1084 (2016).
- 1199 36. Zhang, Z. *et al.* Epigenomic Diversity of Cortical Projection Neurons in the Mouse Brain.
1200 *bioRxiv* 2020.04.01.019612 (2020) doi:10.1101/2020.04.01.019612.
- 1201 37. Mukamel, E. A. & Ngai, J. Perspectives on defining cell types in the brain. *Curr. Opin.*
1202 *Neurobiol.* **56**, 61–68 (2019).
- 1203 38. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell
1204 RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat.*
1205 *Biotechnol.* **36**, 421–427 (2018).
- 1206 39. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell
1207 transcriptomes using Scanorama. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0113-3.
- 1208 40. Deneris, E. S. & Hobert, O. Maintenance of postmitotic neuronal cell identity. *Nat. Neurosci.*
1209 **17**, 899–907 (2014).
- 1210 41. Kepecs, A. & Fishell, G. Interneuron cell types are fit to function. *Nature* **505**, 318–326
1211 (2014).
- 1212 42. Paul, A. *et al.* Transcriptional Architecture of Synaptic Communication Delineates
1213 GABAergic Neuron Identity. *Cell* **171**, 522–539.e20 (2017).

- 1214 43. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor
1215 binding profiles. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz1001.
- 1216 44. Cubelos, B. *et al.* Cux1 and Cux2 regulate dendritic branching, spine morphology, and
1217 synapses of the upper layer neurons of the cortex. *Neuron* **66**, 523–535 (2010).
- 1218 45. Oishi, K., Aramaki, M. & Nakajima, K. Mutually repressive interaction between Brn1/2 and
1219 Rorb contributes to the establishment of neocortical layer 2/3 and layer 4. *Proc. Natl. Acad.*
1220 *Sci. U. S. A.* **113**, 3371–3376 (2016).
- 1221 46. Arendt, D. *et al.* The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757
1222 (2016).
- 1223 47. Smith, J. B. *et al.* The relationship between the claustrum and endopiriform nucleus: A
1224 perspective towards consensus on cross-species homology. *J. Comp. Neurol.* **527**,
1225 476–499 (2019).
- 1226 48. Crick Francis C & Koch Christof. What is the function of the claustrum? *Philos. Trans. R.*
1227 *Soc. Lond. B Biol. Sci.* **360**, 1271–1279 (2005).
- 1228 49. Puelles, L. Chapter 4 - Development and Evolution of the Claustrum. in *The Claustrum*
1229 (eds. Smythies, J. R., Edelman, L. R. & Ramachandran, V. S.) 119–176 (Academic Press,
1230 2014).
- 1231 50. Ross, S. E., Greenberg, M. E. & Stiles, C. D. Basic helix-loop-helix factors in cortical
1232 development. *Neuron* **39**, 13–25 (2003).
- 1233 51. He, Y. *et al.* Improved regulatory element prediction based on tissue-specific local
1234 epigenomic signatures. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1633–E1640 (2017).
- 1235 52. Arlotta, P. *et al.* Neuronal subtype-specific genes that control corticospinal motor neuron
1236 development in vivo. *Neuron* **45**, 207–221 (2005).
- 1237 53. Allen, T. & Lobe, C. G. A comparison of Notch, Hes and Grg expression during murine

- 1238 embryonic and post-natal development. *Cell. Mol. Biol.* **45**, 687–708 (1999).
- 1239 54. Hrvatin, S. *et al.* PESCA: A scalable platform for the development of cell-type-specific viral
1240 drivers. *bioRxiv* 570895 (2019) doi:10.1101/570895.
- 1241 55. Mich, J. K. *et al.* Epigenetic landscape and AAV targeting of human neocortical cell classes.
1242 *bioRxiv* 555318 (2019) doi:10.1101/555318.
- 1243 56. Dimidschstein, J. *et al.* A viral strategy for targeting and manipulating interneurons across
1244 vertebrate species. *Nat. Neurosci.* **19**, 1743–1749 (2016).
- 1245 57. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell*
1246 **164**, 1110–1121 (2016).
- 1247 58. Ferland, R. J., Cherry, T. J., Preware, P. O., Morrissey, E. E. & Walsh, C. A. Characterization
1248 of Foxp2 and Foxp1 mRNA and protein in the developing and mature brain. *J. Comp.*
1249 *Neurol.* **460**, 266–279 (2003).
- 1250 59. Siddiqui, T. J. *et al.* An LRRTM4-HSPG complex mediates excitatory synapse development
1251 on dentate gyrus granule cells. *Neuron* **79**, 680–695 (2013).
- 1252 60. Yamawaki, N., Borges, K., Suter, B. A., Harris, K. D. & Shepherd, G. M. G. A genuine layer
1253 4 in motor cortex with prototypical synaptic circuit connectivity. *Elife* **3**, e05422 (2014).
- 1254 61. Nieto, M. *et al.* Expression of Cux-1 and Cux-2 in the subventricular zone and upper layers
1255 II–IV of the cerebral cortex. *J. Comp. Neurol.* **479**, 168–180 (2004).
- 1256 62. Fanselow, M. S. & Dong, H.-W. Are the dorsal and ventral hippocampus functionally distinct
1257 structures? *Neuron* **65**, 7–19 (2010).
- 1258 63. Cembrowski, M. S., Wang, L., Sugino, K., Shields, B. C. & Spruston, N. Hipposeq: a
1259 comprehensive RNA-seq database of gene expression in hippocampal principal neurons.
1260 *Elife* **5**, e14997 (2016).
- 1261 64. Zhang, T.-Y. *et al.* Environmental enrichment increases transcriptional and epigenetic

- 1262 differentiation between mouse dorsal and ventral dentate gyrus. *Nat. Commun.* **9**, 298
1263 (2018).
- 1264 65. Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation
1265 maps from adult mouse tissues. *Nat. Genet.* **45**, 1198–1206 (2013).
- 1266 66. Sansom, S. N. & Livesey, F. J. Gradients in the brain: the control of the development of
1267 form and function in the cerebral cortex. *Cold Spring Harb. Perspect. Biol.* **1**, a002519
1268 (2009).
- 1269 67. O’Leary, D. D. M., Chou, S.-J. & Sahara, S. Area patterning of the mammalian cortex.
1270 *Neuron* **56**, 252–269 (2007).
- 1271 68. Gorkin, D. U. *et al.* Systematic mapping of chromatin state landscapes during mouse
1272 development. *bioRxiv* 166652 (2017) doi:10.1101/166652.
- 1273 69. Allen Institute for Brain Science. Allen Mouse Brain Reference Atlas CCF v3. *Allen Mouse*
1274 *Brain Reference Atlas CCF v3* <http://atlas.brain-map.org> (2017).
- 1275 70. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation
1276 variation. *Nature* **523**, 212–216 (2015).
- 1277 71. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of
1278 Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
- 1279 72. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data
1280 analysis. *Genome Biol.* **19**, 15 (2018).
- 1281 73. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing
1282 well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
- 1283 74. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification
1284 using Support Vector Machines. *Mach. Learn.* **46**, 389–422 (2002).
- 1285 75. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The Balanced Accuracy

- 1286 and Its Posterior Distribution. in *2010 20th International Conference on Pattern Recognition*
1287 3121–3124 (2010).
- 1288 76. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle
1289 the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **18**, 1–5
1290 (2017).
- 1291 77. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**,
1292 2579–2605 (2008).
- 1293 78. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Fast
1294 interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat.*
1295 *Methods* **16**, 243–245 (2019).
- 1296 79. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in
1297 hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).
- 1298 80. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif.
1299 *Bioinformatics* **27**, 1017–1018 (2011).
- 1300 81. Wingender, E., Schoeps, T., Haubrock, M., Krull, M. & Dönitz, J. TFClass: expanding the
1301 classification of human transcription factors to their mammalian orthologs. *Nucleic Acids*
1302 *Res.* **46**, D343–D347 (2018).
- 1303 82. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles
1304 of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- 1305 83. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylC-seq library
1306 preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10**,
1307 475–483 (2015).
- 1308 84. Zhou, J. *et al.* Robust single-cell Hi-C clustering by convolution- and random-walk-based
1309 imputation. *Proceedings of the National Academy of Sciences* vol. 116 14011–14018

- 1310 (2019).
- 1311 85. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. *Sci.*
1312 *Rep.* **8**, 10872 (2018).
- 1313 86. Zhang, H. *et al.* Chromatin structure dynamics during the mitosis-to-G1 phase transition.
1314 *Nature* **576**, 158–162 (2019).
- 1315 87. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a
1316 simple way to prevent neural networks from overfitting. (2014).
- 1317 88. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).
- 1318 89. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. in *12th {USENIX}*
1319 *Symposium on Operating Systems Design and Implementation ({OSDI} 16)* 265–283
1320 (2016).

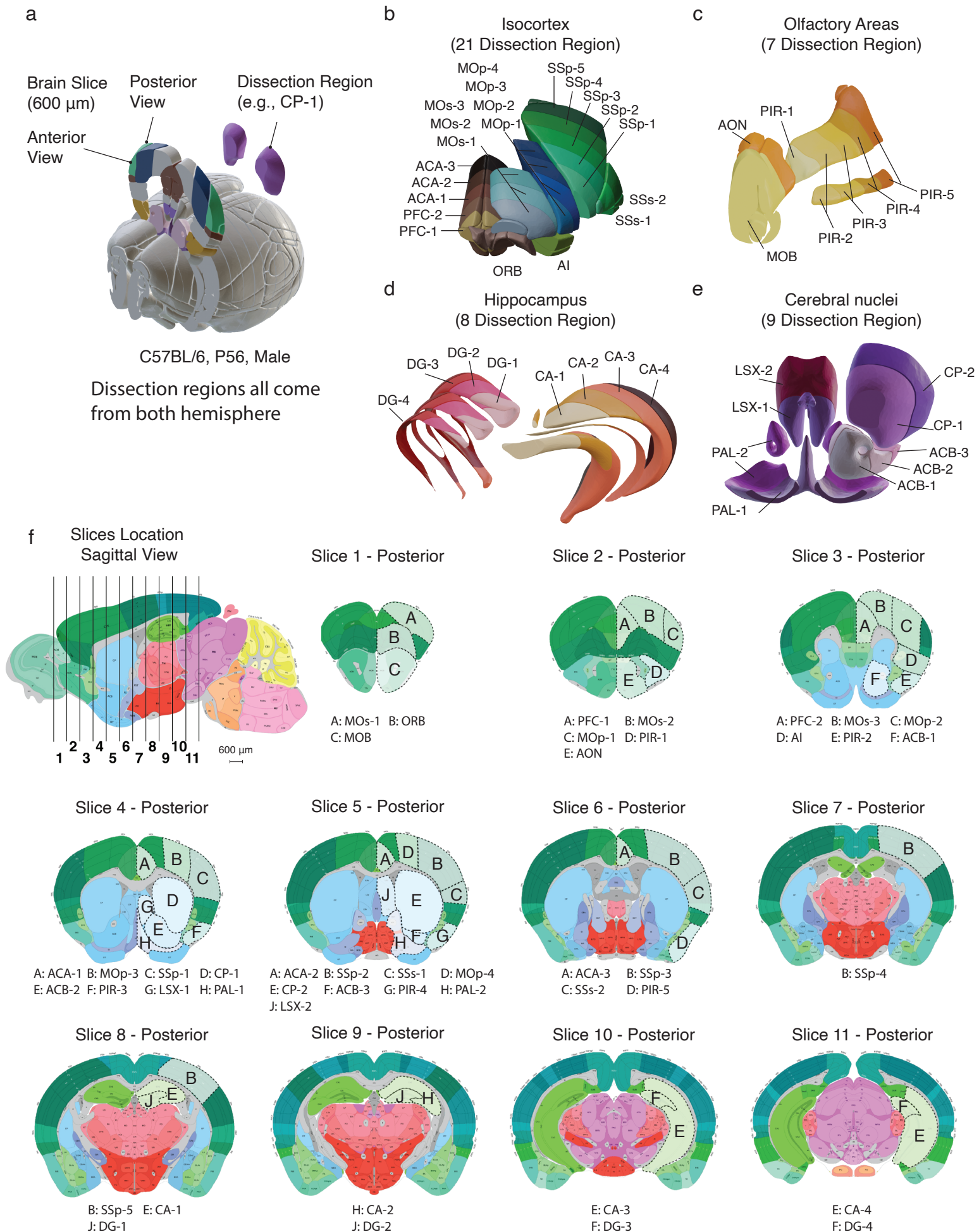


Figure S1

1321 **Extended Data Fig. 1. Brain dissection regions**

1322 **a**, 3D schematic of brain dissection steps. Each P56, male, C57BL/6 mouse brain is dissected
1323 into 600-micron slices. we then dissect brain regions from both hemispheres within a specific
1324 slice. **b-e**, 3D adult P56 mouse brain schematic adapted from Allen CCFv3 to display the four
1325 major brain regions and 45 dissection regions. Each color and the corresponding label
1326 represents a dissection region. **f**, 2D adult P56 mouse brain atlas adapted from Allen Mouse
1327 Brain Reference Atlas, the first sagittal image showing the location of each coronal slice,
1328 followed by 11 posterior view images of all coronal slices, the same 45 dissection regions are
1329 labeled on the corresponding slice. All coronal images follow the same scale as the sagittal
1330 image. The posterior view of each slice is the anterior view of the next slice.

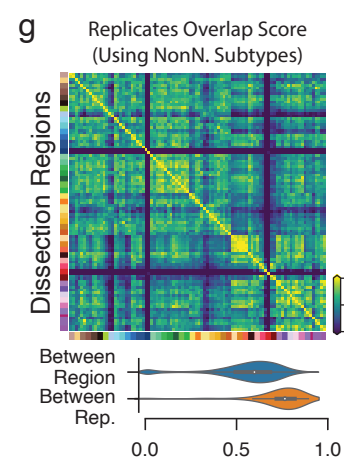
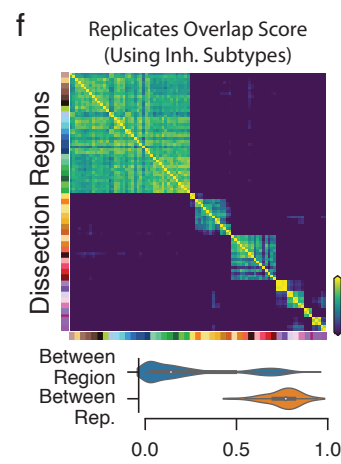
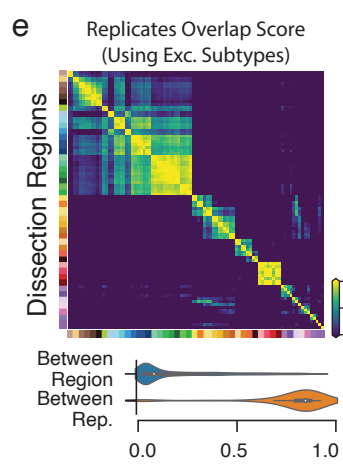
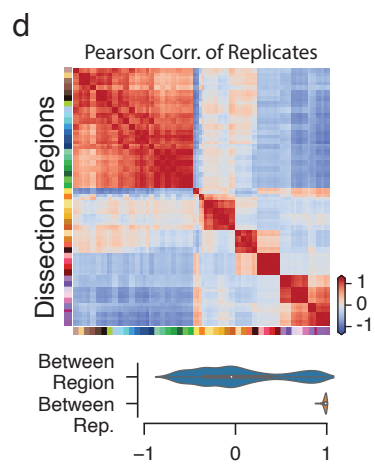
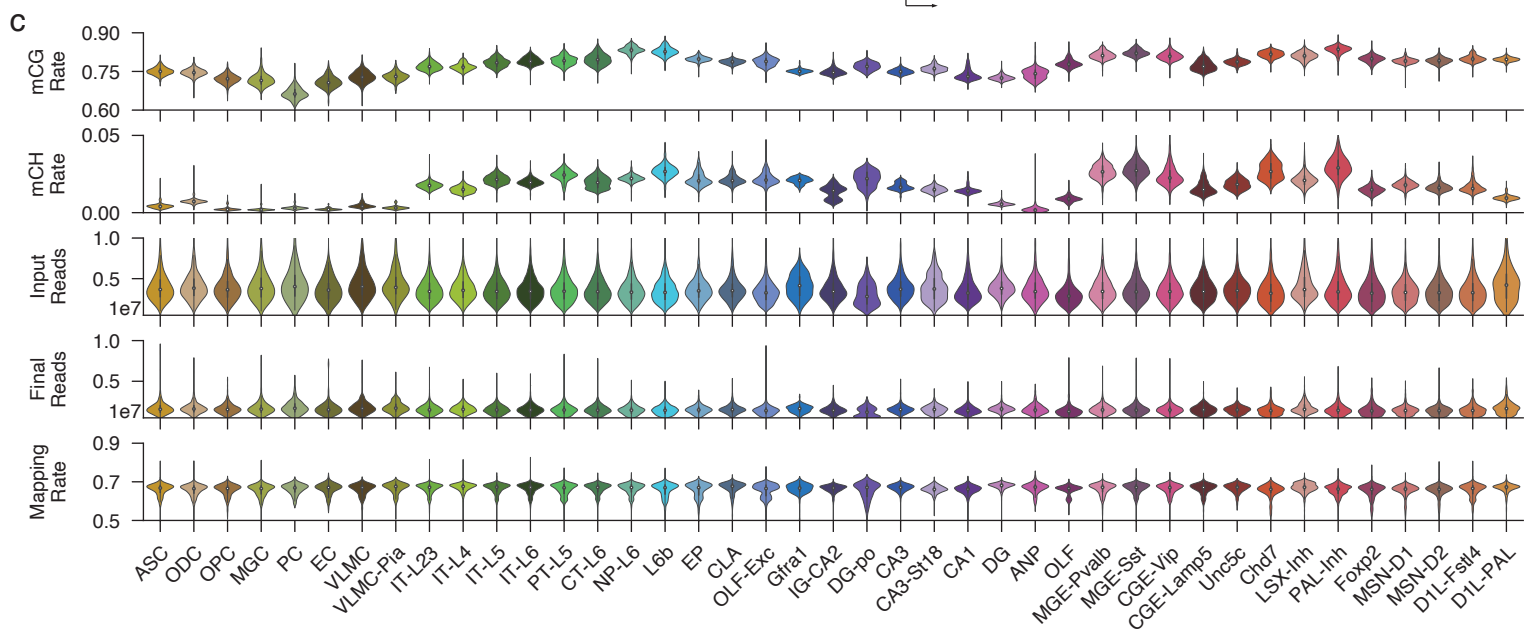
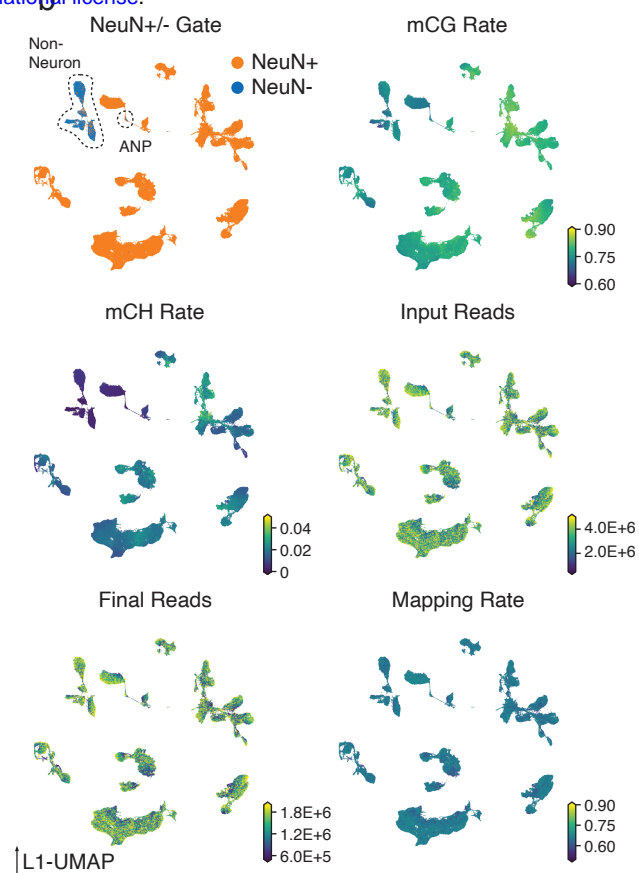
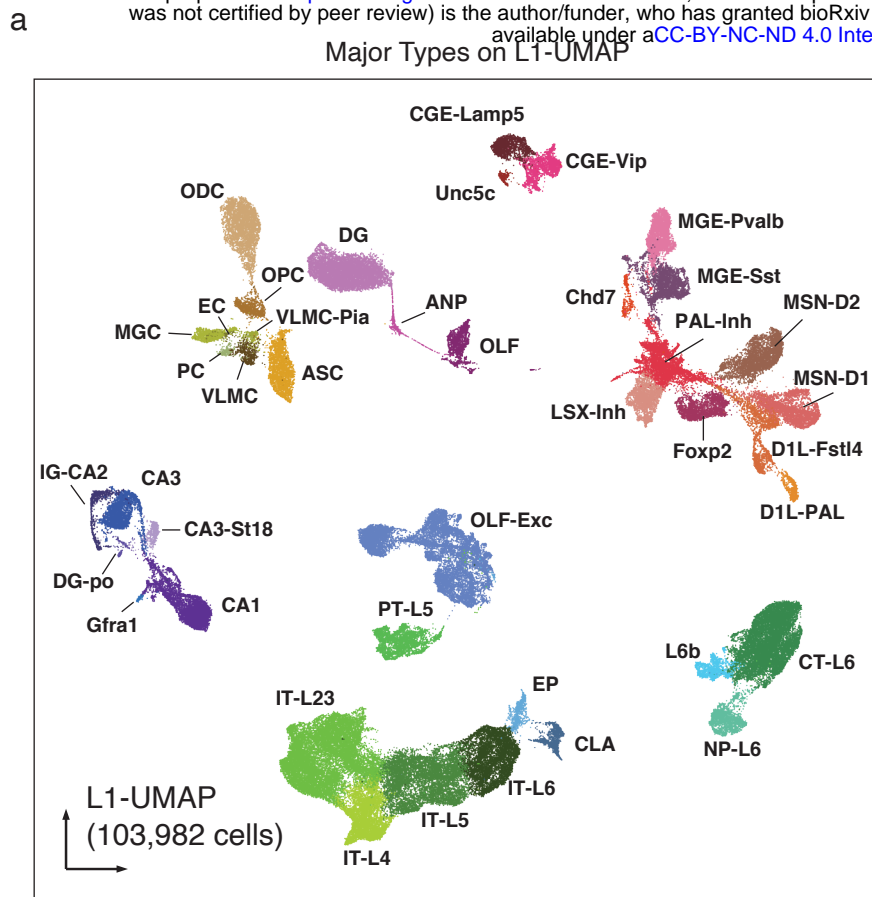


Figure S2

1331 **Extended Data Fig. 2. Major Type labeling and basic mapping metrics of snmC-seq2.**

1332 **a**, L1-UMAP colored and labeled by major cell types. **b**, L1-UMAP colored by NeuN antibody
1333 FACS gates and other snmC-seq2 key read mapping metrics. **c**, Violin plots for all of the key
1334 metrics, group by major types. **d**, Heatmap of Pearson correlation between the average
1335 methylome profiles (mean mCH and mCG rate of all chromosome 100kb bins across all cells
1336 belong to a replicate sample) of the 92 replicates from 45 brain regions. The violin plot below
1337 summarizes the value between replicates within the same brain region, or between different
1338 brain regions. **e-g**, Pairwise overlap score (measuring co-clustering of two replicates) of
1339 excitatory subtypes (**e**), inhibitory subtypes (**f**), and non-neuronal subtypes (**g**). The violin plots
1340 summarize the subtype overlap score between replicates within the same brain region, or
1341 between different brain regions.

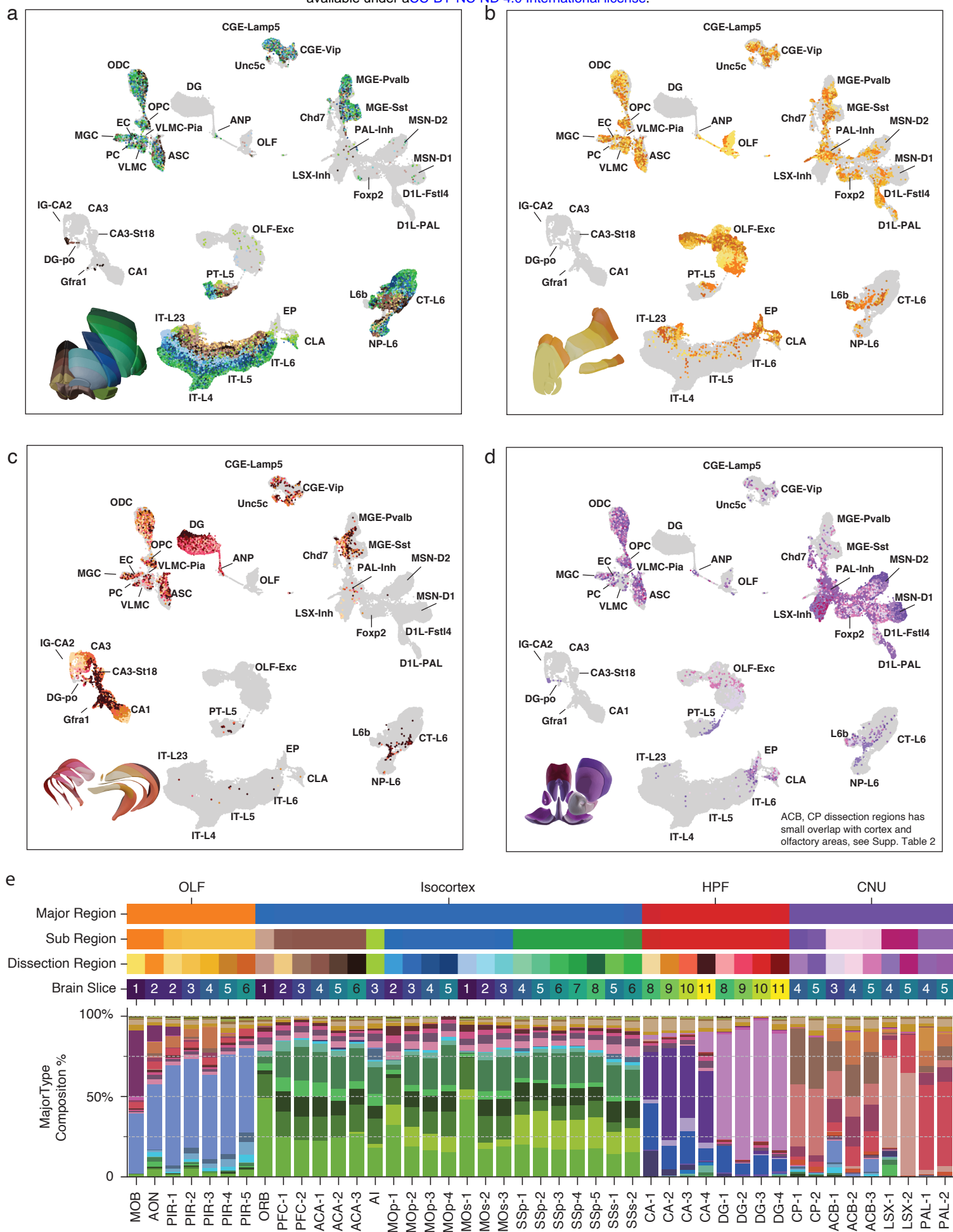


Figure S3

1342 **Extended Data Fig. 3. Cell-type composition of dissection regions**

1343 **a-d**, L1-UMAP labeled by major types and partially colored by dissection regions for cells from
1344 Isocortex (**a**), olfactory areas (**b**), hippocampus (**c**), and cerebral nucleus (**d**). Other cells are
1345 shown in gray as background. **e**, Similar compound bar plot as Fig. 1h, from top to bottom,
1346 showing the organization of dissection regions and the major type composition of each
1347 dissection region.

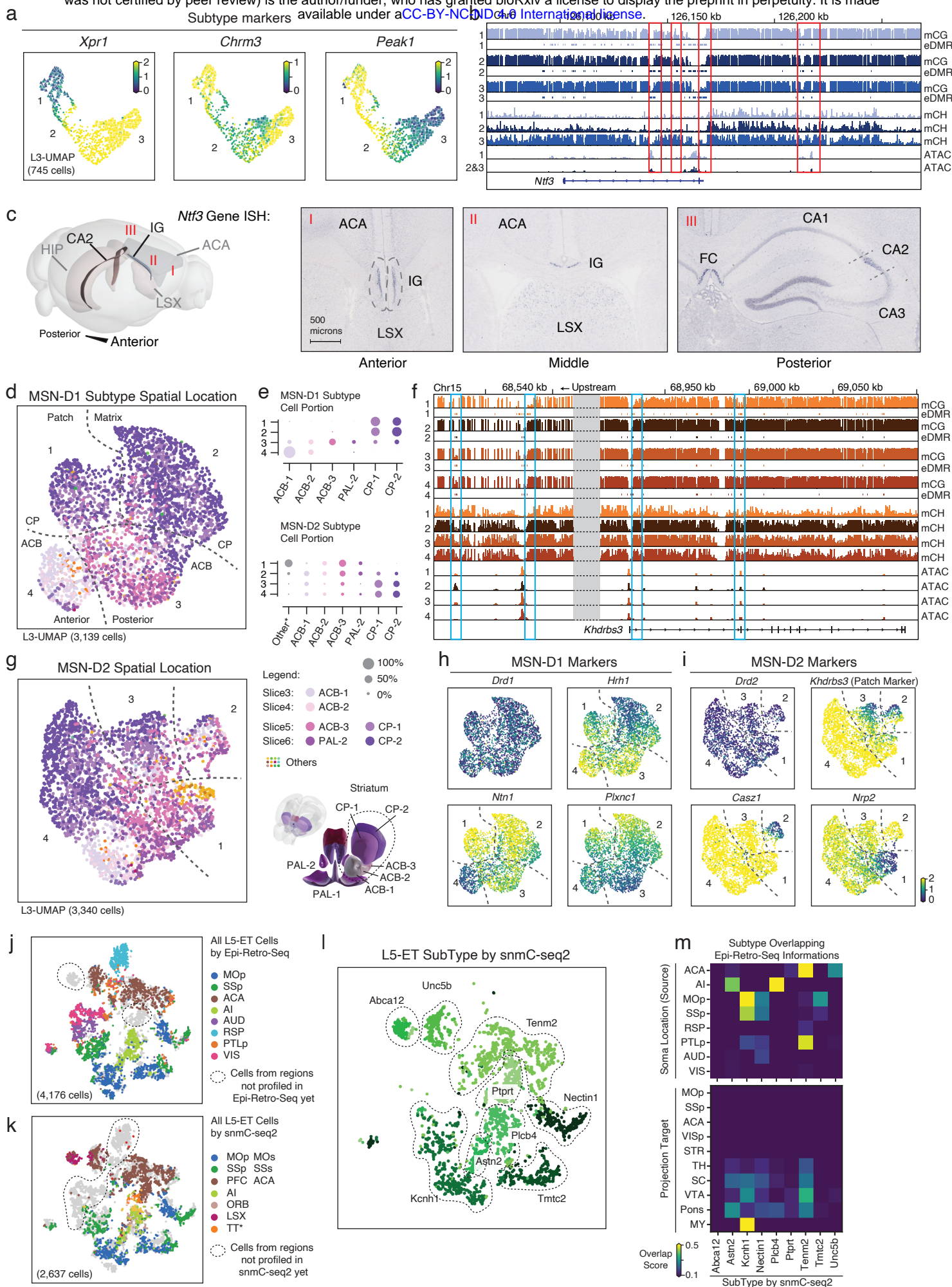
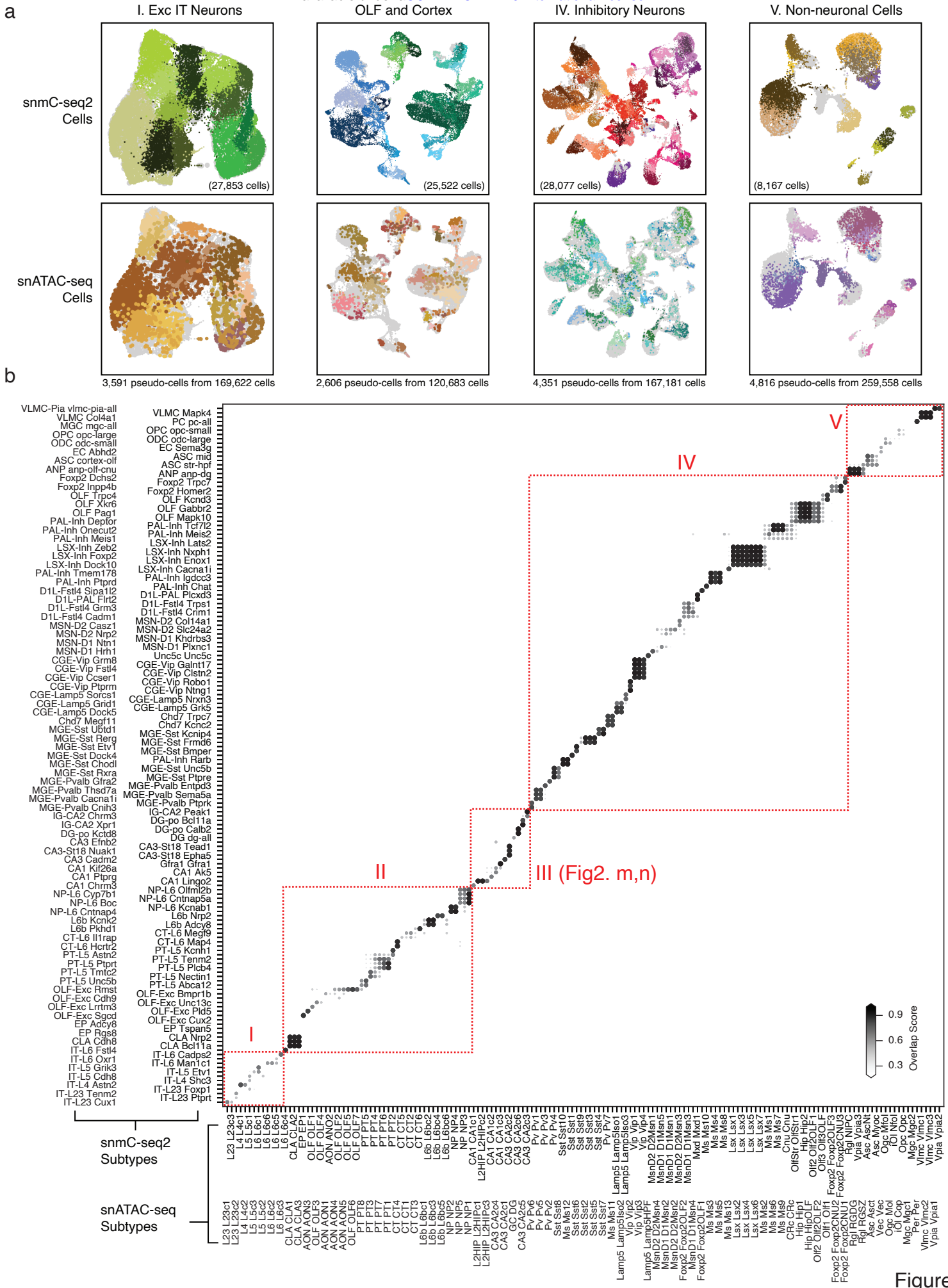


Figure S4

1348 **Extended Data Fig. 4. Supporting details of cellular and spatial diversity of neurons at the**
1349 **subtype level.**

1350 **a**, mCH Rate of marker genes in IG-CA2 cells. **b**, Methylome and chromatin accessibility
1351 genome browser view of *Ntf3* genes and its upstream regions. ATAC and eDMR information are
1352 from Fig.4 analysis. **c**, Three different views of the ISH experiment from ABA, showing the *Ntf3*
1353 gene expressed in both IG and CA2. **d**, L3-UMAP from MSN-D1 cells colored by subregion. The
1354 numbers are subtypes marked by hypomethylation of unique genes: 1) *Khdrbs3*, 2) *Hrh1*, 3)
1355 *Plxnc1*, an 4) *Ntn1*. **e**, The dot plot shows region composition of each subtype of MSN-D1 and
1356 MSN-D2. **f**, Genome browser view of *Khdrbs3* genes similar to (**b**). **g**, MSN-D2 subtypes
1357 marked by hypomethylation of unique genes: 1) *Nrp2*, 2) *Casz1*, 3) *Col14a1*, and 4) *Slc24a2*. **h**,
1358 **i**, mCH Rate of MSN-D1 (**h**) and MSN-D2 (**i**) subtype marker genes. **j**, **k**, Same integration
1359 t-SNE as Fig. 2i-j colored by the dissection regions, but using all cells that have been profiled by
1360 Epi-Retro-Seq (**j**) or snmC-seq2 (**k**), cells from brain regions that have only been profiled via
1361 one of the methods are circled out. **l**, Same t-SNE as (**k**) colored subtypes. **m**, Overlap score
1362 matrix matching the subtypes to the “Soma Location (source)” and “Projection Target”
1363 information labels of Epi-Retro-Seq cells.



1364 **Extended Data Fig. 5. Integration with snATAC-seq**

1365 **a**, Integration UMAP for snmC-seq2 cells (top row) and snATAC-seq pseudo-cells (bottom row).
1366 Each panel is colored by subtypes from the corresponding study, the other dataset shown in
1367 gray in the background. snATAC-seq subtype color token from Li et al., (Companion Manuscript
1368 # 11). Integration is done in five separate groups, including four shown here and one shown in
1369 Fig. 2m-n. **b**, Overlap score matrix matching the 160 snATAC-seq subtypes to the 161
1370 snmC-seq2 subtypes.

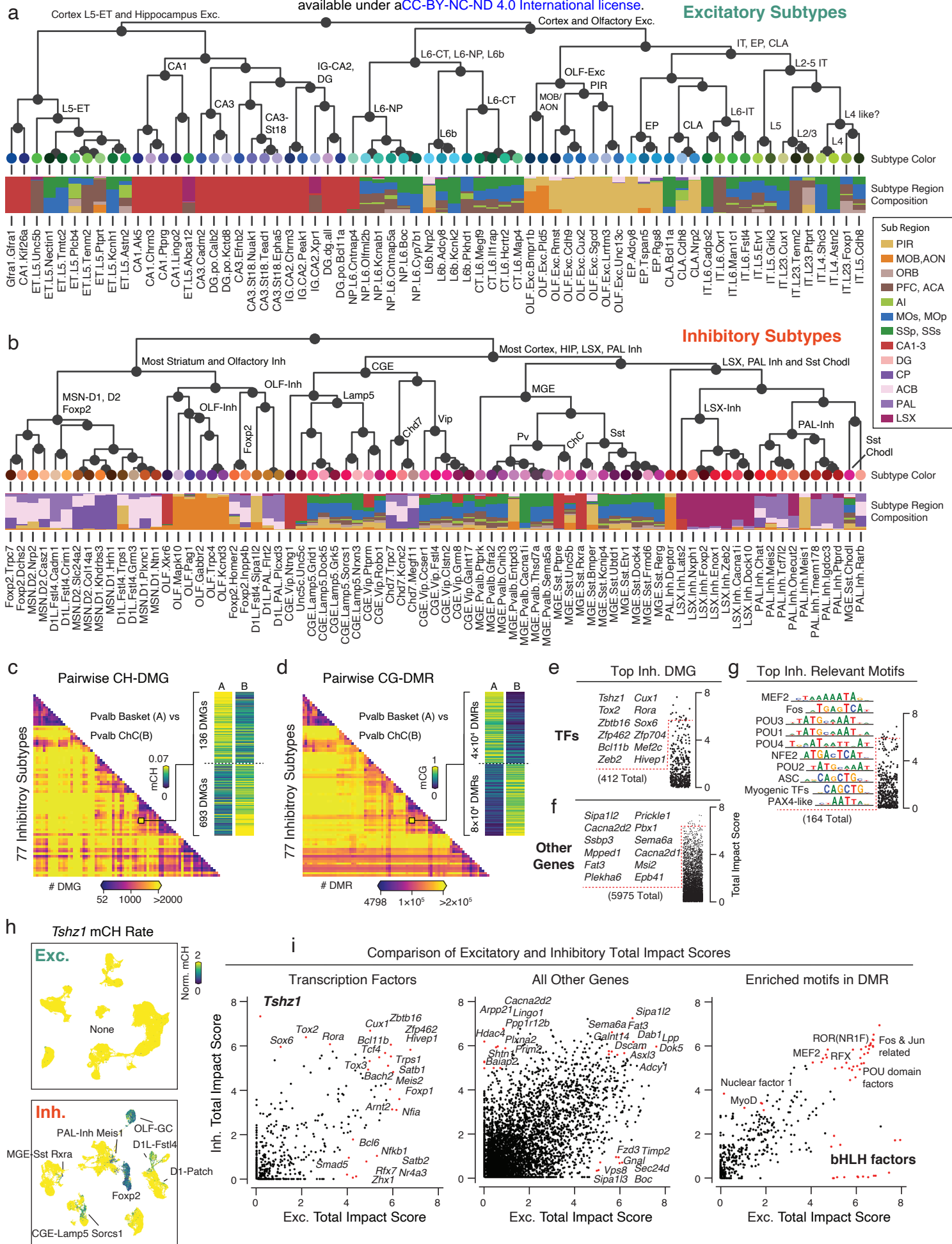
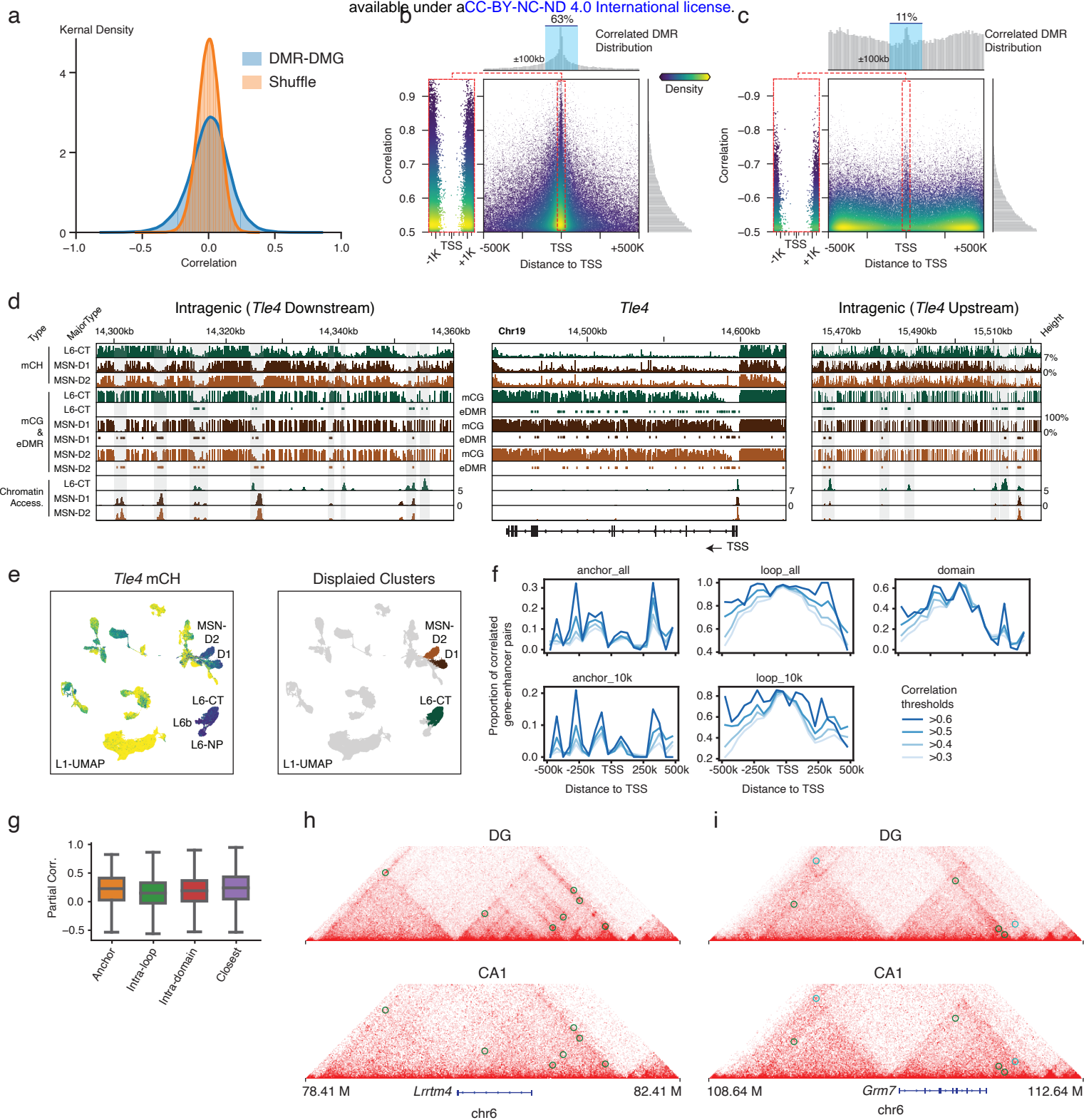


Figure S6

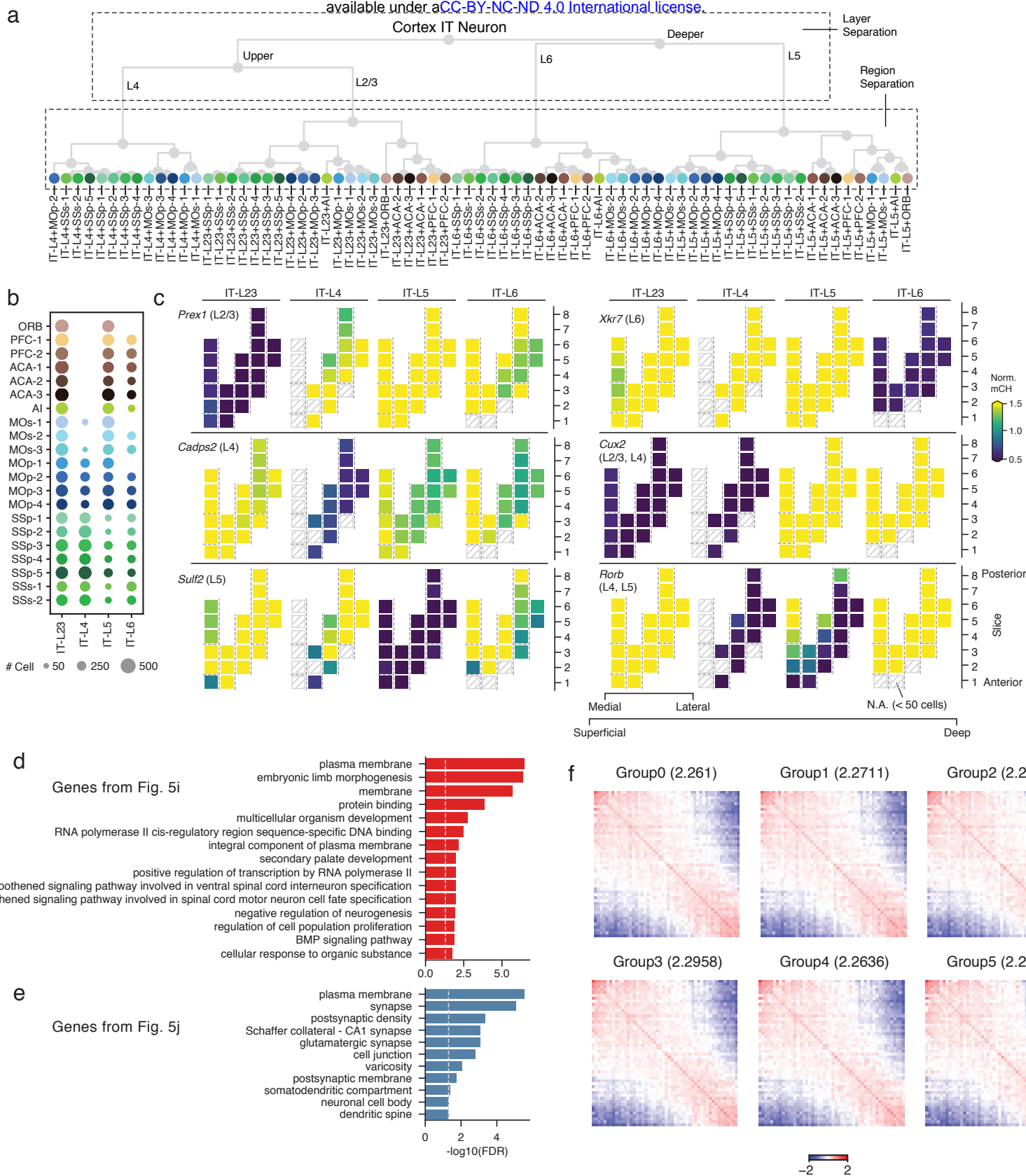
1371 **Extended Data Fig. 6. Subtype phylogeny with related gene and motifs**

1372 **a, b**, Subtype phylogeny of excitatory (**a**) and inhibitory (**b**) neurons. Leaf nodes colored by
1373 subtypes, and the barplot shows subregion composition. **c, d**, Counts heatmap of pairwise
1374 CH-DMG (**c**) and CG-DMR (**d**) between 77 inhibitory subtypes. **e-g**, Top TFs (**e**), other genes
1375 (**f**), and enriched motifs (**g**) rank by total impact score based on the inhibitory subtype
1376 phylogeny. **h**, An example gene *Tshz1* only shows subtype diversity in inhibitory subtypes but
1377 not excitatory subtypes. **i**, Comparison of the total impact scores calculated from either
1378 excitatory subtype phylogeny (x-axis) or inhibitory subtype phylogeny (y-axis) for TFs, other
1379 genes, and enriched motifs.



1380 **Extended Data Fig. 7. Gene-Enhancer landscape related**

1381 **a**, Distribution of actual DMR-DMG partial correlation compared to the shuffled null distribution.
1382 **b, c**, DMR-DMG correlation (y-axis) and the distance between DMR center and gene TSS
1383 (x-axis), each point is a DMR-DMG pair, color represents points kernel density. The positively
1384 **(b)** and negatively **(c)** correlated DMRs are shown separately, due to very different genome
1385 location distributions that are plotted on the top histograms. **d**, detailed view of surrounding
1386 eDMRs that are correlated with *Tle4* gene body mCH. Alternative eDMRs only appear in either
1387 L6-CT or MSN-D1/D2 can be seen on both upstream and downstream of the gene. **e**, *Tle4* gene
1388 body mCH and the major type label on L1-UMAP, paired with **(d)**. **f**, Proportion of loop
1389 supported enhancer-gene pairs among those linked by correlation analyses surpassing different
1390 correlation thresholds at each specific distance. **g**, Partial correlation between mCG of
1391 enhancers and mCH of genes linked by different methods (n=4,171, 127,730, 28,203, 10,058).
1392 The elements of boxplots are defined as: center line, median; box limits, first and third quartiles;
1393 whiskers, 1.5× interquartile range. **h, i**, Interaction maps, mCH, mCG, ATAC, and differential
1394 loops tracks surrounding *Lrrtm4* **(h)** and *Grm7* **(i)**. Circles on the interaction maps represent
1395 differential loops between DG and CA1, where green represents DG loops, and cyan represents
1396 CA1 loops.



1397 **Extended Data Fig. 8. DNA methylation gradient related**

1398 **a**, Layer-dissection-region cell group phylogeny. **b**, Dot plot sized by the number of cells in each
1399 layer-dissection-region combination in excitatory IT neurons. Each group needs at least 50 cells
1400 to be included in the analysis. **c**, Representative marker genes for laminar layers separation.
1401 Using the same dissection region layout as Fig. 5b. **d**, **e**, Top enriched GO terms for positively
1402 (**d**) and negatively (**e**) correlated DMR enriched genes. **f**, Saddle plots of different groups of DG
1403 cells separated by global mCH. Values in the title represent the compartment strengths.

a

bioRxiv preprint doi: <https://doi.org/10.1101/2020.04.30.069377>; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

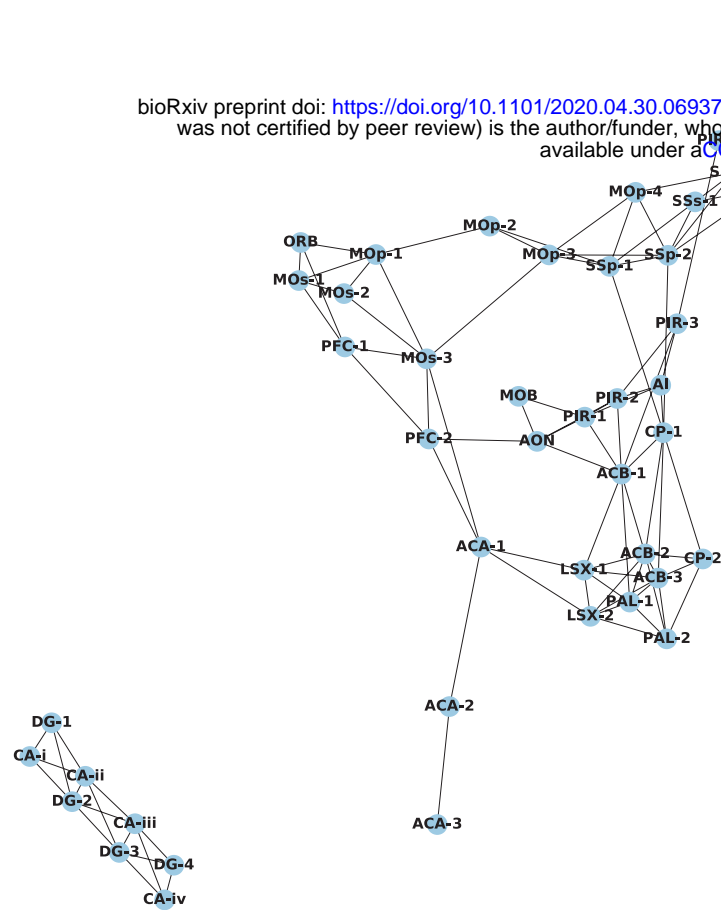
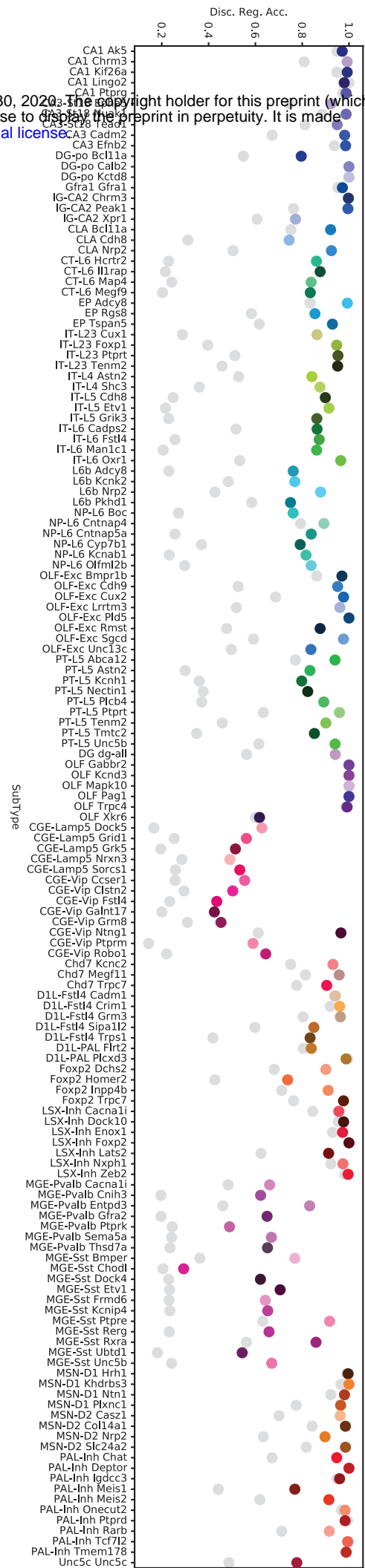
**b**

Figure S9

1404 **Extended Data Fig. 9. Evaluation of the predictive model**

1405 **a**, The neighbor relation among the potential overlapping dissection regions. The network is
1406 constructed based on information of the dissection scheme and the “Potential Overlap” column
1407 in Supplementary Table 2. **b**, Prediction accuracy of dissection region at cell subtype level. The
1408 colored points denote the prediction accuracy of the model, while the grey ones denote the
1409 random guess accuracy when cell subtypes and corresponding spatial distributions are given.