An atlas of gene regulatory elements in adult mouse cerebrum

Yang Eric Li$^{12}$, Sebastian Preissl$^{2,12}$, Xiaomeng Hou$^1$, Ziyang Zhang$^1$, Kai Zhang$^1$, Yunjiang Qiu$^1$, Olivier B. Poirion$^2$, Bin Li$^3$, Joshua Chiu$^{2,4}$, Hanqing Liu$^1$, Antonio Pinto-Duarte$^6$, Naoki Kubo$^1$, Xiaoyu Yang$^5$, Rongxin Fang$^1$, Xinxin Wang$^1$, Jane YX Han$^1$, Jacinta Lucero$^9$, Yiming Yan$^1$, Michael Miller$^6$, Samantha Kuan$^1$, David Gorkin$^2$, Kyle J. Gaulton$^6$, Yin Shen$^{19}$, Michael Nunn$^5$, Eran A. Mukamel$^9$, M. Margarita Behrens$^6$, Joseph R. Ecker$^{5,10}$ & Bing Ren$^{1,2,11}$

The mammalian cerebrum performs high-level sensory perception, motor control and cognitive functions through highly specialized cortical and subcortical structures$^1$. Recent surveys of mouse and human brains with single-cell transcriptomics$^2$–$^6$ and high-throughput imaging technologies$^7,8$ have uncovered hundreds of neural cell types distributed in different brain regions, but the transcriptional regulatory programs that are responsible for the unique identity and function of each cell type remain unknown. Here we probe the accessible chromatin in more than 800,000 individual nuclei from 45 regions that span the adult mouse isocortex, olfactory bulb, hippocampus and cerebral nuclei, and use the resulting data to map the state of 491,818 candidate cis-regulatory DNA elements in 160 distinct cell types. We find high specificity of spatial distribution for not only excitatory neurons, but also most classes of inhibitory neurons and a subset of glial cell types. We characterize the gene regulatory sequences associated with the regional specificity within these cell types. We further link a considerable fraction of the cis-regulatory elements to putative target genes expressed in diverse cerebral cell types and predict transcriptional regulators that are involved in a broad spectrum of molecular and cellular pathways in different neuronal and glial cell populations. Our results provide a foundation for comprehensive analysis of gene regulatory programs of the mammalian brain and assist in the interpretation of noncoding risk variants associated with various neurological diseases and traits in humans.

The cerebral cortex and cerebral nuclei or basal ganglia in adult mice are made up of tens of millions of neurons and glial cells$^7$. The neurons are classified into many different types of excitatory projection neurons and inhibitory interneurons, defined by their distinct morphology, neurotransmitters, and synaptic connections$^{10,12}$. The glial cells include astrocytes, oligodendrocytes, oligodendrocyte precursor cells, microglia and other less abundant non-neuronal cell types. Understanding how the identity and function of each neural cell type is established during development and modified by experience is one of the fundamental challenges in brain research. Despite recent advances in analysis of gene expression patterns using single-cell transcriptomics and spatial transcriptomics assays$^2,8$, we still lack comprehensive maps of the transcriptional regulatory elements in each cell type. Transcriptional regulatory elements recruit DNA-binding transcription factors to exert control of target gene expression in cis in a cell-type-dependent manner$^{13,14}$. Activation of these elements is accompanied by open chromatin, specific histone modifications and DNA hypomethylation$^{13,14}$. To exploit these epigenetic features, candidate cis-regulatory elements (cCREs) have been mapped by techniques such as DNase I hypersensitive sites sequencing (DNase-seq), assay for transposase-accessible chromatin using sequencing (ATAC-seq), chromatin immunoprecipitation followed by sequencing (ChIP–seq) and whole-genome bisulfite sequencing$^{13,16}$. Conventional assays performed using bulk tissue samples are unable to resolve the cCREs in individual neural cell types owing to the extreme cellular heterogeneity of the brain. To overcome this limitation, single-cell genomic technologies have been developed to enable detection of open chromatin in individual cells$^{17,18}$ and identify cell-type-specific transcriptional regulatory sequences in several mouse brain regions$^{19,21,22}$. As part of the BRAIN Initiative Cell Census Network (BICCN), we performed single-nucleus (sn)ATAC-seq assays using single-cell combinatorial indexing (sci)ATAC-seq$^{17,23}$ for more than 800,000 cells from...
Article

45 dissected regions in the adult mouse brain to produce comprehensive maps of cCREs in distinct cerebral cell types. We also integrated the chromatin accessibility data with brain single-cell RNA sequencing (scRNA-seq) data to characterize the gene regulatory programs of different brain cell types, and used the cCREs to interpret genetic risk variants that are associated with neurological diseases and traits.

**snATAC-seq analysis of mouse brain cells**

We dissected 45 brain regions from the isocortex, olfactory bulb (OLF), hippocampus (HIPP) and cerebral nuclei (including the striatum and pallidum) in 8-week-old male mice (Fig. 1a, Extended Data Fig. 1, Supplementary Table 1). Each dissection was made from 600-μm-thick coronal brain slices starting at the frontal pole according to the Allen Brain Reference Atlas28 (Extended Data Fig. 1). For each region, we performed snATAC-seq with two independent biological replicates (Fig. 1a, Extended Data Fig. 2a–f). A total of 81,739 nuclei with a median number of 4,929 fragments per nucleus passed rigorous quality control measures (Supplementary Table 2, Extended Data Fig. 2g–k, Supplementary Note, Methods). Among them, 381,471 nuclei were from the isocortex, 123,434 were from the olfactory area, and 161,556 were from the hippocampus (Fig. 1a, Extended Data Fig. 2l).

We performed iterative clustering with the software package SnapATAC26 and classified the 81,739 nuclei into distinct cell groups (Fig. 1b–e). We classified the cells into glutamatergic neurons (387,060 nuclei, 47.6%), GABAergic (γ-aminobutyric acid-producing) neurons (167,181 nuclei, 20.5%) and non-neuronal cells (259,588 nuclei, 31.9%) (Fig. 1b–d). Next, the three cell classes were further divided into 43 subclasses (also referred to as major types in the accompanying paper28) (Fig. 1b, d) and annotated these on the basis of chromatin accessibility at promoters and the gene bodies of at least three marker genes of known neural cell types22–24 (Fig. 1e, Extended Data Fig. 6, Supplementary Table 4). Finally, we performed another round of clustering for each subclass and identified a total of 160 cell types at optimal resolutions (also referred to as subtypes in the accompanying paper28) (Extended Data Figs. 3–5, Supplementary Tables 2, 3, Supplementary Note, Methods). Among them, 381,471 nuclei were from the isocortex, 123,434 were from the olfactory area, and 161,556 were from the hippocampus (Fig. 1a, Extended Data Fig. 2l).

**Regional specificity of brain cell types**

Taking advantage of our high-resolution brain dissections, we examined the regional specificity of each brain cell type (Extended Data Figs. 11, 12). We calculated a regional specificity score for each subclass and cell type based on the contribution from different brain regions and showed that this score is highly consistent between biological replicates (Fig. 1f, g, Extended Data Fig. 12d, Methods). Overall, we found good agreement between the regional specificity of most neuronal cell types defined using snATAC-seq datasets and the normalized in situ hybridization (ISH) signals of marker genes in each cell type (Extended Data Fig. 13, Supplementary Table 6, Methods). As expected, most glial cell types were ubiquitously distributed throughout the different brain dissections and showed very low regional specificity (Fig. 1f), except for neuronal intermediate progenitor cells (NIPCs) and radial glia-like cells in the dentate gyrus or subventricular zone (labelled as subclass RGL in Fig. 1f). In contrast to the glial cell types, most GABAergic and glutamatergic neurons showed notable regional specificity (Fig. 1f, g). We found a marked separation on the basis of brain sub-regions for distinct neuron types such as granule cells in the dentate gyrus and matrix DI neurons (MDI) in the pallidum. Glutamatergic neurons showed slightly higher regional specificity than GABAergic neurons, consistent with previous single-cell transcriptomic analysis4 (Fig. 1g, bottom). We also observed distinct type-specific functions of Purkinje neuron (PVGA), intra-telencephalic-projection neuron, and hippocampal cornu ammonis (CAI) neuron (CAIGL) that were highly restricted to individual brain regions or dissections (Extended Data Fig. 14, Supplementary Note).

**Mapping of cCREs in mouse brain cells**

To delineate the gene regulatory programs that underlie the identity and function of each brain cell type, we next identified cCREs in each of the 160 brain cell types from the accessible chromatin landscapes. To account for different sequencing depth and/or the number of nuclei in individual clusters, we identified reproducible peaks based on a corrected integer score calculated by model-based analysis of ChIP–seq data (MACS2)30 (Extended Data Fig. 15a–c, Supplementary Note, Methods). We further selected the elements that were determined as open chromatin regions in a significant fraction of cells in each subtype (false discovery rate (FDR) < 0.01, zero-inflated beta model) (Extended Data Fig. 15d, Supplementary Note, Methods), resulting in a union of 491,568 open chromatin regions. These cCREs together made up 14.8% of the mouse genome (Supplementary Tables 7, 8). Of these cCREs, 96.3% were located at least 1 kb away from annotated promoter regions of protein-coding and long noncoding RNA genes (Fig. 2a, Extended Data Fig. 15e). Several lines of evidence support the authenticity of the identified cCREs. First, they strongly overlapped with the DNase hypersensitive sites (DHSs) that were previously mapped in a broad spectrum of bulk mouse tissues and developmental stages (Fig. 2b, Extended Data Fig. 15f). Second, they generally showed higher levels of sequence conservation than random shuffled genomic regions with similar GC content (Fig. 2c, Extended Data Fig. 15g). Third, they were enriched for active chromatin states or potential insulator protein-binding sites that were previously mapped by bulk analysis of mouse brain tissues31–33 (Fig. 2d, Extended Data Fig. 15h).

To define the cell type specificity of the cCREs, we first plotted the median levels of chromatin accessibility against the range of variation for each element (Fig. 2e). We found that most cCREs exhibited highly variable chromatin accessibility across the brain cell types identified, except for 8,188 regions that showed accessible chromatin in almost all cell clusters (Fig. 2e). The invariant cCREs were highly enriched for promoters (81%), with the remainder including CCCTC-binding factor (CTCF)-binding sites (9%) and strong enhancers (Fig. 2f). To characterize the cell type specificity of the cCREs more explicitly, we used non-negative matrix factorization to group them into 42 modules, with elements in each module sharing similar cell type specificity profiles. Aside from the first module (MI) that included mostly cell type invariant cCREs, the remaining 41 modules showed high cell-type-restricted accessibility (Fig. 2g, Supplementary Tables 9, 10). These cell-type-restricted modules were enriched for distinct sets of motifs recognized by known transcriptional regulators (Supplementary Table 11), laying a foundation for investigating the gene regulatory programs in different brain cell types and regions.
Fig. 1 | Single-cell analysis of chromatin accessibility in the adult mouse cerebrum. a, Schematic of single-cell dissection strategy. A detailed list of cell types and associated subtypes is in Supplementary Table 1. b, Uniform manifold approximation and projection (UMAP) embedding and clustering analysis (Supplementary Table 2). A UMAP subspace is Visualized with a hierarchical binary tree. Cell annotations are in Supplementary Table 4. c, Bar chart representing aggregate chromatin accessibility scores of the different cell types. d, Aggregate chromatin accessibility profiles of cluster A and non-neurons showing the regional specificity scores of subclasses and cell types.
Differential chromatin states at cCREs

Because most neuronal types showed highly restricted distribution in the mouse cerebral cortex and basal ganglia, we hypothesized that the regional specificity of different cell types is accompanied by differences in chromatin accessibility at the cCREs, which drive cell-specific gene expression patterns. We performed integrative analysis to delineate these differentially accessible cCREs among different neuronal and glial cell types. We compared the open chromatin landscapes among different cell types using a likelihood ratio test (Methods), and identified a median number of 11,683 cCREs that exhibited differential accessibility (range: 360–31,608) (Extended Data Fig. 16a, b). We characterized the most diverse GABAergic neuron types in the medial septal nucleus (MSGA) (Extended Data Fig. 17, Supplementary Table 12, Supplementary Note). For SSTGA, we detected a total of 50,079 cCREs, 98% of which were promoter distal, that exhibited cell-type-restricted accessibility within the subclass (Figs. 1, 3a, b, Extended Data Fig. 17a–c, Supplementary Table 13). We found a strong motif enrichment of the zinc-finger transcription factor family KLF in cCREs in SSTGA10 cells (also known as Sst-Chodl cells) compared with other SST neurons (Fig. 3c). This observation, coupled with the finding that KLFs, a member of the KLF family, was expressed in Chodl cells, implicates KLFs in the transcriptional control of Chodl cells (Fig. 3d).

We also identified three astrocyte cell types and performed differential chromatin accessibility analysis for cCREs between these (Fig. 3e, Extended Data Fig. 18a). Two cell types were predominantly found in the cortex and hippocampus, whereas the third cell type (ASCN) was expressed in the pallidum (Fig. 3f, Extended Data Fig. 18b–d). Consistent with the previous findings that astrocytes were organized into distinct lineage-associated laminae36, we detected a spatial gradient in ASCG based on chromatin accessibility at several gene loci in ASCG (Extended Data Fig. 18e). We further performed motif analysis for differentially accessible regions in the ASCN cell type, finding enrichment of the binding motif for the GLI family of zinc-finger transcription factors (Fig. 3g, h, Extended Data Fig. 18f, Supplementary Table 14), which mediate the sonic hedgehog (Shh) signalling pathway that maintains neural stem-cell and astrocyte functions36. Notably, we found a cCRE that contained the GLI motif upstream of the Olig2 promoter. This is consistent with a potential role for Shh signalling in regulating Olig2 expression (Fig. 3h). For example, Ntnh3, Scl6a11 and Agt were predominantly expressed in the pallidum, and cCREs at the gene locus were specifically accessible in ASCNs (Fig. 3k, l, Extended Data Fig. 18h–j). We also found enrichment of distinct transcription factor motifs from regional-specific cCREs in ASCGs sampled from different brain regions (Extended Data Fig. 18k, l, Supplementary Note, Supplementary Table 15).

Integrative analysis of gene regulation

To investigate the transcriptional regulatory programs that are responsible for cell-type-specific gene expression patterns in the mouse cerebrum, we carried out integrative analysis that combines the snATAC-seq data collected in the current study with previously published scRNA-seq data from matched brain regions. We first connected 261,204 distal cCREs to 12,722 putative target genes by measuring the co-accessibility using Cicero37,38 (Fig. 4a, Methods). This analysis identified a total of 813,638 gene–cCRE pairs within 500 kb of each other (Supplementary Table 16). Next, we identified the subset of cCREs that might increase the
expression of putative target genes and therefore function as putative enhancers in neuronal or non-neuronal types. To this end, we first identified distal cCREs for which chromatin accessibility was correlated with transcriptional variation of the linked genes in the RNA–ATAC joint cell clusters as defined above (Fig. 4a, b, Extended Data Fig. 10). This analysis revealed a total of 129,404 pairs of positively correlated cCRE (putative enhancers) and genes at an empirically defined significance threshold of FDR < 0.01 (Supplementary Table 16). These included 86,850 putative enhancers and 10,604 genes (Fig. 4b, Extended Data Fig. 19, Supplementary Note). To investigate how the putative enhancers may direct cell-type-specific gene expression, we further classified them into 38 modules, by applying non-negative matrix factorization to the matrix of normalized chromatin accessibility across the RNA–ATAC joint cell clusters (Supplementary Table 17). The putative enhancers in each module had a similar pattern of chromatin accessibility across cell clusters to the expression of putative target genes (Fig. 4c, Supplementary Table 18). This analysis revealed a large group of 12,740 putative enhancers that were linked to 6,373 genes expressed at a higher level in all neuronal cell clusters than in all non-neuronal cell types (module M1) (Fig. 4c). It also uncovered modules of enhancer–gene pairs that were active in a more restricted manner (modules M2–M38) (Fig. 4c, Extended Data Fig. 19, Supplementary Tables 19, 20, Supplementary Note).

Genes associated with module M1 are preferentially expressed in both glutamatergic and GABAergic neurons, but not in glial cell types (Fig. 4c). De novo motif enrichment analysis of the 12,740 cCREs or putative enhancers in this module showed strong enrichment of sequence motifs recognized by the transcription factors CTCF, RFX and MEF2 (Supplementary Table 21, Extended Data Fig. 19d), as well as many known motifs for other transcription factors (Fig. 4d, Supplementary Table 20). CTCF is a ubiquitously expressed DNA-binding protein with a well-established role in transcriptional insulation and chromatin organization. CTCF has also been shown to promote neurogenesis by binding to promoters and enhancers of the proto-cadherin alpha gene cluster and facilitating enhancer–promoter contacts. We found putative enhancers with one or more CTCF-binding motifs linked to 2,601 genes that were broadly expressed in both inhibitory and excitatory neurons (Fig. 4d, Extended Data Fig. 19e), the gene products of which are involved in several neural processes including axon guidance and synaptic transmission (Extended Data Figs. 19f, 20, Supplementary Tables 22, 23, Supplementary Note).

**Neurogenesis in the adult mouse brain**

Neurogenesis in the adult mouse brain is spatially restricted to the subgranular zone (SGZ) in the dentate gyrus of the hippocampus where excitatory neurons are generated, and the subventricular zone (SVZ) of the lateral ventricles that give rise to GABAergic neurons. The NIPCs that are involved in adult neurogenesis could be identified as the cells lined up in trajectories between radial glia-like cells and neuroblasts in both brain regions (Fig. 4e) and their presence in the respective dissections was supported by ISH data from the Allen Brain Atlas for several marker genes (Extended Data Fig. 21a, b). We predicted potential transcription factors that contribute to NIPCs as well as other cell types by integrating RNA expression and motif enrichment analysis using the Taiji pipeline (Fig. 4f, Supplementary Table 24, Methods). Consistent with previous reports, NR2E1 was predicted to be a master regulator in both NIPC populations, and SOX2 was a regulator of the NIPCs from the SGZ, whereas E2F1 contributed to NIPCs from the SVZ. Although chromatin landscapes in the NIPCs from both regions were very similar (Fig. 4g), we identified 200 differentially accessible regions in the NIPC population between the SGZ and the SVZ (Fig. 4h). Several cCREs at Neurog2, which encodes a protein that is crucial for glutamatergic granule neuron specification in the SGZ, were found to be accessible selectively in the SGZ but not in the SVZ (Fig. 4i). By contrast, several cCREs with chromatin accessibility in SVZ NIPCs were located at the Dlx2 locus—a gene that is important for the specification of GABAergic neurons (Extended Data Fig. 21c). An active enhancer previously validated by mouse transgenics was predicted to target the nearby Trappc9 gene, which encodes a protein that is involved in nerve growth factor-induced neuronal differentiation. These observations suggest that NIPCs in the SGZ and SVZ give rise to distinct neuronal cell types by engaging different cCREs involved in controlling region-specific gene expression of key regulator genes.

**Interpreting noncoding risk variants**

Genome-wide association studies (GWASs) have identified genetic variants that are associated with many neurological diseases and traits (Supplementary Table 25), but interpreting the results has been challenging because most variants are located in noncoding parts of the genome that often lack functional annotations. To test whether our maps of
**Fig. 4 | Integrative analysis of gene regulatory programs in different cerebral cell types.** a, Schematic overview of the computational strategy used to identify cCREs that are positively correlated with transcription of target genes. b, In total, 129,404 pairs of positively correlated cCRE and genes (highlighted in red) were identified (FDR < 0.01). Grey filled curve shows distribution of Pearson’s correlation coefficient (PCC) for randomly shuffled cCRE–gene pairs. c, Heat map showing chromatin accessibility of distal cCREs. d, RNA expression of putative target genes. Genes are shown for cCRE–gene pairs. e, Compute PCC between cCRE accessibility and gene expression. f, Candidate driver TFs in adult neurogenesis at SGZ. g, Candidate driver TFs in adult neurogenesis at SVZ. h, VISTA [mm1679] forebrain (4/5 embryos). i, genome browser tracks showing representative differential cCREs of NIPCs in the SGZ (top) and SVZ (bottom). j, Representative images of transgenic mouse embryos showing LacZ reporter gene expression under the control of the indicated enhancers that overlapped the differential cCRE in i (dotted line). Images were downloaded from the VISTA database (https://enhancer.lbl.gov)\(^4\).

---

**ARTICLE**

**Nature** Vol 598 | 7 October 2021

---

**ENCODE**

**Article**

---

**METHODS**

---

**RESULTS**

---

**DISCUSSION**

---

**ACKNOWLEDGEMENTS**

---

**References**

---

**SUPPLEMENTARY INFORMATION**

---

**Correspondence**

---

**AUTHOR CONTRIBUTIONS**

---

**DATA AVAILABILITY**

---
cCREs in different mouse brain cell types could assist the interpretation of noncoding risk variants of neurological diseases, we identified orthologues of the mouse cCREs in the human genome by performing reciprocal homology searches \(^5\) (Methods). For this analysis, we found that for 69.2% of the mouse brain cCREs, human genome sequences with high similarity could be identified (more than 50% of bases lifted over to the human genome) (Extended Data Fig. 22a). Supporting the function of these human orthologues of the mouse brain cCREs, 83.0% of them overlapped with representative DNase hypersensitivity sites in the human genome\(^6\).

We performed linkage disequilibrium score regression\(^7\) analysis and found significant associations between 20 neurological traits (Supplementary Table 25) and the open chromatin landscapes in one or more subclasses of the brain cells we identified (Fig. 3, Methods, Supplementary Note). In particular, we observed widespread and strong enrichment of genetic variants linked to psychiatric and cognitive traits such as major depressive disorder, bipolar disorder and schizophrenia (SCZ) within accessible cCREs across various neuronal cell types (Fig. 5). Other neurological traits—such as attention deficit hyperactivity disorder and autism spectrum disorder—were associated with specific neuronal cell types in cerebral nuclei and the hippocampus (Fig. 5). Risk variants for schizophrenia were not only enriched in cCREs in all excitatory neurons, but also enriched in certain inhibitory neuron cell types\(^8\) (Fig. 5). We also found that more than 25% of homologous sequences of SCZ causal variants reside in the mouse cCREs defined in this study (Extended Data Fig. 22b, Supplementary Note). The strongest enrichment of heritability for bipolar disorder was found in cCREs that mapped in the excitatory neurons from the isocortex (Fig. 5). Risk variants of tobacco use disorder showed significant enrichment in the cell types from the striatum—a cerebral nucleus previously implicated in addiction\(^9\) (Fig. 5).

Understanding the cellular and molecular basis of brain circuits is one of the grand challenges of the twenty-first century\(^10\). In-depth knowledge of the transcriptional regulatory program in brain cells would not only improve our understanding of the molecular inner workings of neurons and non-neuronal cells, but could also shed light on the pathogenesis of a spectrum of neuropsychiatric diseases\(^11\). Here, we report a comprehensive profiling of chromatin accessibility at single-cell subclasses of mouse cerebral cortex. The chromatin accessibility maps of 491,818 cCREs, probed in 813,799 nuclei and 160 cell types, span several cerebral cortical areas and subcortical structures. Taking advantage of our high-resolution brain dissections, we examined the regional specificity in chromatin accessibility of cell types in the mouse cerebrum and showed that most brain cell types exhibit strong regional specificity. The described cCRE atlas (http://catlas.org/mousebrain) represents a rich resource for the neuroscience community to understand the molecular patterns that underlie diversification of brain cell types in complementation to other molecular and anatomical data.

### Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03604-1.
14. Moore, J. E. et al. Expanded encyclopedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
15. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods **10**, 1213–1218 (2013).
16. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).
17. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
18. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
19. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat. Biotechnol. **37**, 916–924 (2019).
20. Satapathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. **37**, 925–936 (2019).
21. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. **36**, 70–80 (2018).
22. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324 (2018).
23. Sinnammon, J. R. et al. The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Res.* **29**, 857–869 (2019).
24. Thornton, C. A. et al. Spatially mapped single-cell chromatin accessibility. Nat. Commun. **12**, 1274 (2021).
25. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat. Neurosci. **21**, 432–439 (2018).
26. Fang, R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat. Commun. **12**, 1337 (2021).
27. Graybuck, L. T. et al. Enhancer viruses for combinatorial cell-subclass-specific labeling. *Neuron* **109**, 1449–1464.e13 (2021).
28. Wang, Q. et al. The Allen Mouse Brain common coordinate framework: a 3D reference atlas. *Cell* **181**, 906–953.e20 (2020).
29. Liu, H. et al. DNA methylation atlas of the mouse brain at single-cell resolution. Nature https://doi.org/10.1038/s41586-018-01382-8 (2020).
30. Zhang, Y. et al. Model-based analysis of ChIP-seq data with Cicero. *Genome Biol.* **9**, R137 (2008).
31. Shen, Y. et al. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
32. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
33. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods **9**, 215–216 (2012).
34. Tatsunami, K. et al. Olig2-lineage astrocytes: a distinct subtype of astrocytes that differs from GFAP astrocytes. *Front. Neuroanat.* **12**, 8 (2018).
35. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell in situ transcriptomic map. Nat. Neurosci. **23**, 500–509 (2020).
36. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
37. Harrow, J. et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7**, S4 (2006).
38. Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. **15**, 234–246 (2014).
39. Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N. & Yagi, T. CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. Cell Rep. **2**, 345–357 (2012).
40. Guo, Y. et al. CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. *Proc. Natl Acad. Sci. USA* **109**, 21081–21086 (2012).
41. Ming, G. L. & Song, H. Adult neurogenesis in the mammalian brain: significant answers and significant questions. *Neuron* **70**, 687–702 (2020).
42. Hsieh, J. Orchestrating transcriptional control of adult neurogenesis. *Genes Dev.* **26**, 1010–1021 (2012).
43. Lein, E. S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
44. Zhang, K., Wang, M., Zhao, Y. & Wang, W. Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development. *Sci. Adv.* **5**, eaaq3262 (2019).
45. Sobhan, P. K. & Funa, K. TLX—its emerging role for neurogenesis in health and disease. Mol. Neurobiol. **54**, 272–280 (2017).
46. Cooper-Kuhn, C. M. et al. Impaired adult neurogenesis in mice lacking the transcription factor E2F1. Mol. Cell. Neurosci. **21**, 312–333 (2002).
47. Silve, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. **35**, D88–D92 (2007).
48. Hu, Y. H. et al. NRPB: a novel NIK and IKKβ-binding protein that enhances NFκB activation. *J. Biol. Chem.* **280**, 29233–29241 (2005).
49. Auffray, C. et al. A brief history of human disease genetics. *Nature* **577**, 179–180 (2020).
50. Youn, C. et al. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. **45** (D1), D626–D634 (2017).
51. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. **47**, 291–295 (2015).
52. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. Nat. Genet. **50**, 825–833 (2018).
53. Volkow, N. D. & Morales, M. The brain on drugs: from reward to addiction. Cell **162**, 712–725 (2015).
54. The BRAIN Initiative Cell Census Consortium. The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. *Neuron* **96**, 542–557 (2017).
55. NIH The BRAIN Initiative. BRAIN 2025 Report https://braininitiative.nih.gov/strategic-planning/brain-2025-report (2014).
56. Sullivan, P. F. & Geschwind, D. H. Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. Cell **177**, 162–183 (2019).
57. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
58. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
59. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. **29**, 24–26 (2011).
60. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021
Methods

Tissue preparation and nuclei isolation
All experimental procedures using live animals were approved by the SALK Institute Animal Care and Use Committee under protocol number 18-00006. Adult C57BL/6J male mice were purchased from Jackson Laboratories. Brains were extracted from 56–63-day-old mice and sectioned into 600-μm coronal sections along the anterior–posterior axis in ice-cold dissection media, by using a brain slicer from CNS muSlice LLC. Specific brain regions were dissected according to the Allen Brain Reference Atlas (Extended Data Fig. 1) and nuclei isolated as previously described. Regions were pooled from 3–31 of the same sex to obtain enough nuclei for snATAC-seq for each biological replica, and two biological replicas were processed for each dissection region.

snATAC-seq using combinatorial indexing
snATAC-seq was performed as described with steps optimized for automation. A step-by-step protocol for library preparation is available at: https://www.protocols.io/view/sequencing-open-chromatin-of-single-cell-nuclei-sn-pjudknw/abstract.

Brain nuclei were pelleted with a swinging bucket centrifuge (500g, 5 min, 4 °C; 5920R, Eppendorf). Nuclei pellets were resuspended in 1 ml nuclei permeabilization buffer (5% BSA, 0.2% IGEPA-CAL-CA630, 1 mM DTT and cOmpleteTM, EDTA-free protease inhibitor cocktail (Roche) in PBS) and pelleted again (500g, 5 min, 4 °C; 5920R, Eppendorf). Nuclei were resuspended in 500 μl high salt tagmentation buffer (36.3 mM Tris-acetate (pH 7.8), 72.6 mM potassium-acetate, 11 mM Mg acetate, 17.6% dimethylformamide) and counted using a haemocytometer. Concentration was adjusted to 1,000–4,500 nuclei per 9 μl, and 1,000–4,500 nuclei were dispensed into each well of a 96-well plate. For tagmentation, 1 μl barcoded Trn transposomes were added using a BenchSmart 96 (Mettler Toledo, RRID:SCR_018093) (Supplementary Table 26), mixed five times, and incubated for 60 min at 37 °C with shaking (500 rpm). To inhibit the Trn reaction, 10 μl of 40 mM EDTA was added to each well with a BenchSmart 96 (Mettler Toledo, RRID:SCR_018093) and the plate was incubated at 37 °C for 15 min with shaking (500 rpm). Next, 20 μl 2× sort buffer (2% BSA, 2 mM EDTA in PBS) were added using a BenchSmart 96 (Mettler Toledo, RRID:SCR_018093). All wells were combined into a FACs tube and stained with 3 μM Draq7 (Cell Signaling) (Extended Data Fig. 23). Using a SH800 (Sony), 20 nuclei were sorted per well into eight 96-well plates (total of 768 wells, 30,720 nuclei total, 15,360 nuclei per sample) containing 10.5 μl EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma). If processing two samples per day, tagmentation was performed with different sets of barcodes in separate 96 well plates. After tagmentation nuclei from individual plates were pooled together. Preparation of sort plates and all downstream pipetting steps were performed on a Biomek i7 Automated Workstation (Beckman Coulter, RRID:SCR_018094). After the addition of 1μl 0.2% SDS, samples were incubated at 55 °C for 5 min with shaking (500 rpm). Then, 1μl 12.5% Triton X-100 was added to each well to quench the SDS. Next, 12.5 μl NEBNext High-Fidelity 2 × PCR Master Mix (NEB) were added and samples were amplified by PCR (72 °C 5 min, 98 °C 30 s, 98 °C 10 s, 63 °C 30 s, 72 °C 60 s) × 12 cycles, held at 12 °C. After PCR, all wells were combined. Libraries were purified according to the MinElute PCR Purification Kit manual (Qiagen) using a vacuum manifold (QIAvac 24 plus, Qiagen) and size selection was performed with SPRI Beads (Beckmann Coulter, 0.5×x and 1.5×x). Libraries were purified one more time with SPRI Beads (Beckmann Coulter, 1.5×x). Libraries were quantified using a Qubit fluorimeter (Life Technologies, RRID:SCR_018095) and the nucleosomal pattern was verified using a Tapestation (High Sensitivity D1000, Agilent). Libraries generated with indexing version 1 (Supplementary Table 1) were sequenced on a HiSeq4000 (RRID:SCR_016386, Illumina) using custom sequencing primers with following read lengths: 50 + 10 + 12 + 50 (Read1 + Index1 + Index2 + Read2). Indexing primers and sequencing primers are in Supplementary Table 26. The nuclei indexing version (v1 or v2) used for each library is indicated in Supplementary Table 26.

Nuclei indexing scheme
To generate snATAC-seq libraries we used initially an indexing scheme as previously described (version 1). Here, 16 p5 and 24 p7 indexes were combined to generate an array of 384 indexes for tagmentation and 16 i5 as well as 48 i7 indexes were combined for an array of 768 PCR indexes. Owing to this library design, it is required to sequence all four indexes to assign a read to a specific nucleus with long reads and a constant base sequence for both indexes reads between i and p barcodes. Therefore, the resulting libraries were sequenced with 25% spike-in library on a HiSeq2500 (RRID:SCR_016383) and these read lengths: 50 + 43 + 37 + 50 (ref. 25).

To generate libraries compatible with other sequencers and not requiring spike-in libraries or custom sequencing recipes, we modified the library scheme (Version 2). For this, we used 384 individual indexes for i7 and combined with one T5 with a universal index sequence for tagmentation (for a total of 384 tagmentation indexes). For PCR, we used 768 different i5 indexes and combined with a universal i7 primer index sequence. Tagmentation indexes were 10 bp and PCR indexes 12 bp long. We made sure, that the hamming distance between every two barcodes was ≥ 4, the GC content between 37.5–62.5%, and the number of repeats ≤ 3. The resulting libraries were sequenced on a HiSeq4000 with custom primers and these read lengths: 50 + 10 + 12 + 50 (Supplementary Table 26).

snATAC-seq data using the 10x Chromium platform
Brain nuclei were pelleted with a swinging bucket centrifuge (500g, 5 min, 4 °C; 5920R, Eppendorf). Nuclei pellets were resuspended in 1 ml nuclei permeabilization buffer (5% BSA, 0.2% IGEPA-CAL-CA630, 1 mM DTT and cOmpleteTM, EDTA-free protease inhibitor cocktail (Roche) in PBS) and pelleted again (500g, 5 min, 4 °C; 5920R, Eppendorf) and washed with wash buffer (10 mM Tris-HCl (pH 7.5), 10mM NaCl, 3 mM MgCl2, 0.1% Tween-20, and 1% BSA (Proliant 7500804) in molecular biology-grade water). Nuclei were pelleted (500g, 5 min, 4 °C; 5920R, Eppendorf) and resuspended in 30 μl of 1× nuclei Buffer (10x Genomics). Nuclei were counted using a haemocytometer, and 15,360 nuclei were used for tagmentation. Single-cell ATAC-seq libraries were generated using the Chromium Single Cell ATAC Library & Gel Bead Kit (10x Genomics, 1000110), Chromium Chip E Single Cell ATAC kit (10x Genomics, 1000155) and Chromium IV Multiplex Kit N, Set A (10x Genomics, 1000084) following manufacturer instructions. Final libraries were quantified using a Qubit fluorimeter (Life Technologies) and the nucleosomal pattern was verified using a Tapestation (High Sensitivity D1000, Agilent). Libraries were sequenced on NextSeq 500 and NovaSeq 6000 sequencers (Illumina) with following read lengths: 50 + 8 + 16 + 50 (Read1 + Index1 + Index2 + Read2). After demultiplexing, the Index2 (cell index) was transferred to the read name, in order to keep the same fastq format for downstream processing.

Processing and alignment of sequencing reads
Paired-end sequencing reads were demultiplexed and the cell index transferred to the read name. Sequencing reads were aligned to mm10 reference genome using bwa. After alignment, we used the R package ATACseqQC (v1.0.2) to check for fragment length contribution which is characteristic for ATAC-seq libraries. Next, we combined the sequencing reads to fragments, and for each fragment we performed the following quality control: (1) keep only fragments quality score MAPQ > 30; (2) keep only the properly paired fragments with length < 1,000 bp; (3) PCR duplicates were further removed with SnapTools.
Article

(https://github.com/r3fang/SnapTools, RRID:SCR_018097)\(^6\). Reads were sorted on the basis of the cell barcode in the read name.

**TSS enrichment calculation**

Enrichment of ATAC-seq accessibility at TSSs was used to quantify data quality without the need for a defined peak set. The method for calculating enrichment at TSS was adapted from previously described. TSS positions were obtained from the GENCODE database v16 (RRID:SCR_014966)\(^3\). In brief, Tn5 corrected insertions (reads aligned to the positive strand were shifted +4 bp and reads aligned to the negative strand were shifted -5 bp) were aggregated ±2,000 bp relative (TSS strand-corrected) to each unique TSS genome-wide. Then this profile was normalized to the mean accessibility ±1,900-2,000 bp from the TSS and smoothed every 11 bp. The max of the smoothed profile was taken as the TSS enrichment.

**Doublet removal**

We used a modified version of Scrublet (RRID:SCR_018098)\(^6\) to remove potential doublets for every dataset independently. Peaks were called using MACS2 scores for aggregate accessibility profiles on each sample. Next, cell-by-peak count matrices were calculated and used as input, with default parameters. Doublet scores were calculated for both observed nuclei \(\{x_i\}\) and simulated doublets \(\{y_i\}\) using Scrublet (RRID:SCR_018098)\(^6\). Next, a threshold \(\theta\) is selected based on the distribution of \(\{y_i\}\), and observed nuclei with doublet score larger than \(\theta\) are predicted as doublets. To determine \(\theta\), we fit a two-component mixture distribution by using function normalmixEM from R package mixtools\(^6\). The lower component contained most embedded doublet types, and the other component contained majority of neo-typic doublets (collision between nuclei from different clusters. We selected the threshold \(\theta\) where the \(p_{\text{pdf}}(x, \mu, \sigma) = p_{\text{pdf}}(x, \mu, \sigma)\) in which \(p\) denotes probability; pdf denotes probability density function. This value suggested that the nuclei have same chance of belonging to both classes.

**Cell clustering**

We used an iterative clustering strategy using the snapATAC\(^26\) package (RRID:SCR_018097) with slight modifications as detailed below. For round 1 clustering, we clustered and finally merged single nuclei to three main cell classes: non-neurons, GABAergic neurons, and glutamatergic neurons. For each main cell class, we performed another round of clustering to identify major cell subclasses. Last, for each subclass, we performed a third round of clustering to find cell types.

Detailed description for every step is as follows. (1) Nuclei filtering. Nuclei with ≥1,000 uniquely mapped fragments and TSS enrichment > 10 were filtered for individual dataset. Second, potential barcode collisions were also removed for individual datasets. (2) Feature bin selection. First, we calculated a cell-by-bin matrix at 5-kb resolution for every dataset independently and subsequently merged the matrices. Second, we converted the cell-by-bin count matrix to a binary matrix. Third, we filtered out any bins overlapping with the ENCODE blacklist (mm10, http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/mm10.blacklist.bed.gz). Fourth, we focused on bins on chromosomes 1–19, X and Y. Last, we removed the top 5% bins with the highest read coverage from the count matrix. (3) Dimensionality reduction. SnapATAC applies a nonlinear dimensionality reduction method called diffusion maps, which is highly robust to noise and perturbation\(^6\). However, the computational time of the diffusion maps algorithm scales exponentially with the increase of number of cells. To overcome this limitation, we combined the Nyström method (a sampling technique)\(^6\) and diffusion maps to present Nyström landmark diffusion map to generate the low-dimensional embedding for large-scale dataset.

A Nyström landmark diffusion maps algorithm includes three major steps:

(1) Sampling. We sampled a subset of \(K (K < N)\) cells from \(N\) total cells as ‘landmarks’.

(2) Embedding. We computed a diffusion map embedding for \(K\) landmarks.

(3) Extension. We projected the remaining \((N - K)\) cells onto the low-dimensional embedding as learned from the landmarks to create a joint embedding space for all cells. Having more than 800,000 single nuclei at the beginning, we decided to apply this strategy on round 1 and 2 clustering. A total of 10,000 cells were sampled as landmarks and the remaining query cells were projected onto the diffusion maps embedding of landmarks. Later for the round 3 clustering, diffusion map embeddings were directly calculated from all nuclei.

(4) Eigenvector selection. To determine the number of eigenvectors of diffusion operator to include for downstream analysis, we generated an ‘elbow plot’, to rank all eigenvectors on the basis of the percentage of variance explained by each one. For each round of clustering, we selected the top 10–20 eigenvectors that captured most of the variance.

(5) Graph-based clustering. Using the selected significant eigenvectors, we next construct a \(k\)-nearest neighbour graph. Each cell is a node and the \(k\)-nearest neighbours of each cell were identified according to the Euclidian distance and edges were drawn between neighbours in the graph. Next, we applied the Leiden algorithm on the \(k\)-nearest neighbour graph using Python package leidenalg (https://github.com/vtraag/leidenalg)\(^6\).

(6) Optimization on cluster resolution. We tested different 'resolution parameter' parameters (step between 0 and 1 by 0.1) to determine the optimal resolution for different cell populations. For each resolution value, we tested whether there was clear separation between nuclei. To do so, we generated a cell-by-cell consensus matrix in which each element represents the fraction of observations two nuclei are part of the same cluster. A perfectly stable matrix would consist entirely of zeros and ones, meaning that two nuclei either cluster together or not in every iteration. The relative stability of the consensus matrices can be used to infer the optimal resolution. To this end, we generated a consensus matrix based on 300 rounds of Leiden clustering with randomized starting seed s. Let \(M^s\) denote the \(N \times N\) connectivity matrix resulting from applying Leiden algorithm to the dataset \(D^s\) with different seeds. The entries of \(M^s\) are defined as follows:

\[
M^s(i, j) = \begin{cases} 
1, & \text{if single nucleus } i \text{ and } j \text{ belong to the same cluster} \\
0, & \text{otherwise} 
\end{cases}
\]

Let \(F\) be the \(N \times N\) identicator matrix in which the \((i, j)\)-th entry is equal to 1 if nucleus \(i\) and \(j\) are in the same perturbed dataset \(D^s\), and 0 otherwise. Then, the consensus matrix \(C\) is defined as the normalized sum of all connectivity matrices of all the perturbed \(D^s\).

\[
C(i, j) = \frac{\sum_{s=1}^{S} M^s(i, j)}{\sum_{s=1}^{S} F(i, j)}
\]

The entry \((i, j)\) in the consensus matrix is the number of times single nucleus \(i\) and \(j\) were clustered together divided by the total number of times they were selected together. The matrix is symmetric, and each element is defined within the range \([0, 1]\). We examined the cumulative distribution function (CDF) curve and calculated proportion of ambiguous clustering (PAC) score to quantify stability at each resolution. The resolution with a local minimum of the PAC scores denotes the parameters for the optimal clusters. In the case these were multiple local minimal PACs, we picked the one with higher resolution. Another measurement is dispersion coefficient, which reflects the dispersion (ranges from 0 to 1) of the consensus matrix \(M\) from the value 0.5. The closer to 1 is the dispersion coefficient, the more perfect is consensus matrix, and thus the more stable is the clustering. In a perfect
Finally, for every cluster, we tested whether we could identify differential features compared to all other nuclei (background) and the nearest nuclei (local background) using the function ‘findDAR’.

(7) Visualization. For visualization, we applied UMAP38. Using the cell embedding, we applied both k-nearest neighbor batch effect test (kBET) and local inverse Simpson’s index (LISI) analysis to test the robustness of the clustering results to variation of sequencing depth, signal-to-noise ratios, and batches.

Clustering for adult neurogenesis lineages
We performed separated cell clustering following above strategy for two lineages:
1. Adult neurogenesis in the SGZ: we extracted 83,583 nuclei from 8 brain dissections at or surrounding the SGZ, including CA-1, CA-2, CA-3, CA-4, DG-1, DG-2, DG-3, DG-4 (Supplementary Table 1). Then, we performed cell clustering on 83,583 nuclei for 6 cell types: astrocytes (ASC); dentate gyrus radial glia-like cells; NPCs; granule neuroblasts (DGNBL1/2); and granule neurons.
2. Adult neurogenesis in the SVZ: we extracted 25,923 nuclei from 8 brain dissections at or surrounding the SVZ, including MOB, AON, ACB-1, ACB-2 CP-1, CP-2, LSX-1 and LSX-2 (Supplementary Table 1). Then, we performed cell clustering on 25,923 nuclei for 5 cell types: astrocytes; subventricular zone radial glia-like cells; neuronal intermediate progenitor cells; neuroblasts (OBGNBL1); and inhibitory neurons in olfactory (OBGA1).

Dendrogram construction for mouse brain cell types
First, we calculated for cCRE the median accessibility per cluster and used this value as cluster centroid. Next, we calculated the coefficient of variant for the cluster centroid of each element across major cell types. Finally, we only kept variable elements with coefficient of variants that were larger than 1.5 for dendrogram construction.

We used the set of variable features defined above to calculate a correlation-based distance matrix. Next, we performed linkage hierarchical clustering using the R package pvclust (v.2.0)39 with parameters method.dist = “cor” and method.hclust = “ward.D2”. The confidence for each branch of the tree was estimated by the bootstrap resampling approach with 1,000 rounds.

Regional specificity of cell types
The specificity score is defined as Jensen–Shannon divergence, which measures the similarity between two probability distributions. For each cell type, the contribution of different brain regions was first calculated. Then, we compared this distribution with the contribution of brain cell type, the contribution of different brain regions was first calculated. Then, we fitted a zero-inflated beta model, and empirically identified a significance threshold of FDR < 0.01 to filter potential false positive peaks. Peak regions with FDR < 0.01 in at least one of the clusters were included in downstream analysis.

Computing chromatin accessibility scores
Accessibility of cCREs in individual clusters was quantified by counting the fragments in individual clusters normalized by read depth (CPM). For each gene, we summed up the counts within the gene body +2 kb upstream to calculate the gene activity score. The gene activity score were used for integrative analysis with scRNA-seq. For better visualization, we smoothed the gene activity score to 50 nearest neighbourhood nuclei using Markov affinity-based graph imputation of cells (MAGIC)39.

Integrative analysis of snATAC-seq and scRNA-seq datasets
For integrative analysis, we downloaded level 5 clustering data from the Mouse Brain Atlas website (http://mousebrain.org)3. First, we filtered brain regions that matched samples profiled in this study using these attributes for ‘region’: ‘CNS’, ‘cortex’, ‘hippocampus’, ‘hippocampal cortex’, ‘olfactory bulb’, ‘striatum dorsal’, ‘striatum ventral’, ‘dentate gyrus’, ‘stratum dorsal, stratum ventral’, ‘stratum dorsal, ventral, dentate gyrus’, ‘palidium’, ‘striatum dorsal, striatum ventral, amygdala’, ‘striatum dorsal, striatum ventral’, ‘telencephalon’, ‘brain’ and ‘sub ventricular zone, dentate gyrus’.

Second, we manually subset cell types into three groups by checking the attribute in ‘taxonomy_group’: non-neurons; ‘vascular and leptomeningeal cells’, ‘astrocytes’, ‘oligodendrocytes’, ‘ependymal cells’, ‘microglia’, ‘oligodendrocyte precursor cells’, ‘olfactory ensheathing cells’, ‘pericytes’, ‘vascular smooth muscle cells’, ‘perivascular macrophages’, ‘dentate gyrus radial glia-like cells’, ‘subventricular zone radial glia-like cells’, ‘vascular smooth muscle cells’, ‘vascular endothelial cells’, ‘vascular and leptomeningeal cells’, ‘gamma-aminobutyric acidergic neurons: ‘non-glutamatergic neuroblasts’, ‘telencephalon projecting inhibitory neurons’, ‘olfactory inhibitory neurons’, ‘glutamatergic neuroblasts’, ‘cholinergic and monoaminergic neurons’, ‘di- and mesencephalon inhibitory neurons’, ‘telencephalon inhibitory interneurons’, ‘peptidergic neurons’, ‘glutamatergic neurons: ‘dentate gyrus granule neurons’, ‘di- and mesencephalon excitatory neurons’, ‘telencephalon projecting excitatory neurons’.
To directly compare our single-nucleus chromatin accessibility-derived cell clusters with the single-cell transcriptomics defined taxonomy of the mouse brain, we first used the snATAC-seq data to impute RNA expression levels (gene activity scores) according to the chromatin accessibility of gene promoter and gene body as previously described. We performed integrative analysis with scRNA-seq using Seurat 3.0 (RRID:SCR_016341) to compare cell annotation between different modalities. We randomly selected 200 nuclei (and used all nuclei for cell cluster with fewer than 200 nuclei) from each cell cluster for integrative analysis. We first generated a Seurat object in R by using previously calculated gene activity scores, diffusion map embeddings and cell cluster labels from snATAC-seq. Then, variable genes were identified from scRNA-seq and used for identifying anchors between these two modalities. Finally, to visualize all the cells together, we co-embended the scRNA-seq and snATAC-seq profiles in the same low dimensional space.

To quantify the similarity between cell clusters from two modalities, we calculated an overlapping score as the sum of the minimum proportion of cells or nuclei in each cluster that overlapped within each co-embedding cluster. Cluster overlaps varied from 0 to 1 and were visualized as a heat map with snATAC-seq clusters in rows and scRNA-seq clusters in columns. We found strong correspondence between the two modalities, which was evidenced by co-embedding of both transcriptomic (T-type) and chromatin accessibility (A-type) cells in the same RNA–ATAC joint clusters (Extended Data Fig. 10a, Supplementary Table 5). For this analysis, we examined GABAergic neurons, glutamatergic neurons and non-neuronal cell classes separately (Extended Data Fig. 10, Supplementary Table 5).

**Identification of cis-regulatory modules**

We used non-negative matrix factorization to group cCREs into cis-regulatory modules on the basis of their relative accessibility across major clusters. We adapted non-negative matrix factorization (Python package: sklearn) to decompose the cell-by-cCRE matrix $V (N \times M, N$ rows: cCRE, M columns: cell clusters) into a coefficient matrix $H (R \times M, R$ rows: number of modules) and a basis matrix $W (V \times R)$, with a given rank $R$:

$$V = WH$$

The basis matrix defines module related accessible cCREs, and the coefficient matrix defines the cell cluster components and their weights in each module. The key issue to decompose the occupancy profile matrix was to find a reasonable value for the rank $R$ (that is, the number of modules). Several criteria have been proposed to decide whether a given rank $R$ decomposes the occupancy profile matrix into meaningful clusters. Here we applied two measurements: ‘sparseness’ and ‘entropy’ to evaluate the clustering result. Average values were calculated from 100 times for non-negative matrix factorization runs at each given rank with random seed, which will ensure the measurements are stable.

Next, we used the coefficient matrix to associate modules with distinct cell clusters. In the coefficient matrix, each row represents a module and each column represents a cell cluster. The values in the matrix indicate the weights of clusters in their corresponding module. The coefficient matrix was then scaled by column (cluster) from 0 to 1. Subsequently, we used a coefficient $> 0.1$ (approximately the 95th percentile of the whole matrix) as threshold to associate a cluster with a module.

In addition, we associated each module with accessible elements using the basis matrix. For each element and each module, we derived a basis coefficient score, which represents the accessible signal contributed by all cluster in the defined module. In addition, we also implemented and calculated a basis-specificity score called ‘feature score’ for each accessible element using the ‘kim’ method. The feature score ranges from 0 to 1. A high feature score means that a distinct element is specifically associated with a specific module. Only features that fulfill both of the following criteria were retained as module specific elements: (1) feature score greater than median + 3 standard deviations; (2) the maximum contribution to a basis component is greater than the median of all contributions (that is, of all elements of $W$).

**Identification of differentially accessible regions and definition of specificity score**

To identify cCREs that were differentially accessible either in subtypes or in brain regions, we constructed a logistic regression model predicting cluster/region membership based on each cCRE individually and compares this to a null model with a likelihood ratio test. We used two functions ‘fit_models’ and ‘compare_models’ in R package Monocle3 (v.0.2.2) to perform the differential test. We designed the full model as

$$\logit(P_{ij}) = a_j + m_i + r_{ij} + \varepsilon_j$$

and a reduced model as

$$\logit(P_{ij}) = a_j + r_{ij} + \varepsilon_j$$

in which $P_{ij}$ represents the probability of $i$th site is accessible in the $j$th cell, $a$ is the log$_{10}$-transformed total number of sites observed as accessible for the $i$th cell, $m$ is membership of the $j$th cell in either cluster or region being tested, $r$ is the replicate label for $j$th cell and $\varepsilon$ is an error term.

For each set of testing, between subtypes or between regions in cell type, we kept only cCREs that overlapped with peaks identified in corresponding cell types. A likelihood ratio test was then used to determine whether the full model (including cell cluster or region membership) provided a significantly better fit of the data than the reduced model. After correcting $P$-values using Benjamini–Hochberg method, we set an FDR cut-off as 0.001 to filter out significant differential cCREs.

The log$_{2}$-transformed fold change is used for two-group comparison, for multiple groups, we calculated a Jensen–Shannon divergence-based specificity score previously described to better assign differential cCREs to cell type or brain region. The fraction of accessibility of each cCRE was then converted to probability by scaling by groups. Then, we calculated Jensen–Shannon divergence between two probability distributions. For example, the probability distribution for the first cCRE as $d_1$, to test whether this cCRE is specific in group 1, we assumed another probability distribution:

$$d_2 = \begin{cases} 1, & \text{group 1} \\ 0, & \text{otherwise} \end{cases}$$

Function ‘JSD’ in R package ‘philentropy’ was used to calculate Jensen–Shannon divergence between these two probability distributions, and Jensen–Shannon-based specificity (JSS) scores was defined as:

$$\text{JSS} = 1 - \sqrt{\text{Jensen Shannon divergence}}$$

For each group, we calculated the JSS score for every cCRE. To find a reasonable cut-off to determine restricted or general cCREs, we consider JSS scores from all cCREs that are not identified as differential accessible (from likelihood ratio test) as a background distribution, and JSS scores from cCREs that passed our likelihood ratio test threshold...
and had positive values to be true positives. We set an empirical FDR cut-off where the type I error was no more than 5%.

Finally, the differential cCREs could be aligned to several cell types or brain regions based on the JSS score, we named the one can be assigned to only one type or region as region-specific or cell-type-specific cCREs.

Comparison between the regional specificity of cell types defined by snATAC-seq data and the spatial ISH signals of cell-type-specific genes

To validate the regional specificity of cell types, we took advantage of the spatially mapped quantified ISH expression from ABA in five matched major brain structures, isocortex, olfactory areas, hippocampal formation (HPF), striatum (STR), pallidum. We used the ‘differential search’ function to identify 10,269 genes with increased expression in these five brain regions compared to all brain regions ‘grey matter’ (expression level >1 and fold change > 1). We also identified cell-type-specific genes using Seurat with default parameters for each joint cluster from integrative analysis (fold change > 1 and FDR < 0.05) (Extended Data Fig. 10). 505 cell-type-specific genes (range from 1 to 53, 15 on average) overlapped with the list of genes with increased expression in the five brain regions. For each cell type, we calculated the regional specificity score (see previous section: regional specificity of cell types) on the basis of the relative contribution from five brain regions estimated from snATAC-seq datasets, and also a coefficient of variation based on averaged normalized ISH signals of cell-type-specific marker genes. For each cell-type-specific gene, we calculated the PCC between cell composition in five brain structures and spatial expression levels across the five brain structures derived from ISH.

Because the astrocyte subtypes identified in our study were not resolved in scRNA-seq studies, we identified subtype-specific genes for astrocyte subtypes using chromatin accessibility from snATAC-seq using a likelihood ratio test. The cell-type-specific genes were filtered by FDR less than 0.001 from the likelihood ratio test and empirical FDR cut-off of no more than 5% for JSS scores. Then, we calculated the fraction of overlap between spatially mapped ISH genes from different brain structures and genes with astrocyte subtype-specific accessibility.

Predicting enhancer–promoter interactions

First, co-accessible regions were identified for all open regions in each cell cluster (randomly selected 200 nuclei, and using all nuclei for cell cluster with <200 nuclei) separately, using Cicero with following parameters: aggregation k = 10, window size = 500 kb, distance constraint = 250 kb. To find an optimal co-accessibility threshold for each cluster, we generated a random shuffled cCRE-by-cell matrix as background and identified co-accessible regions from this shuffled matrix. We fitted the distribution of co-accessibility scores from random shuffled background into a normal distribution model by using the R package fitdistplus. Next, we tested every co-accessibility pairs and set the cut-off at co-accessibility score with an empirically defined significance threshold of FDR < 0.01.

CCRE outside of ±1 kb of the TSS in GENCODE mm10 (v.16, RRID:SCR_014966) were considered distal. Next, we assigned co-accessibility pairs to three groups: proximal-to-proximal, distal-to-distal, and distal-to-proximal. In this study, we focus only on distal-to-proximal pairs. We further used RNA expression from matched T-types to filter out pairs that were linked to non-expressed genes (normalized UMI < 5).

We calculated PCC between gene expression and cCRE accessibility across joint RNA-ATAC clusters to examine the relationship between co-accessibility pairs. To do so, we first aggregated all nuclei or cells from scRNA-seq and snATAC-seq for every joint cluster to calculate accessibility scores (log2(CPM)) and relative expression levels (log2(normalized UMI)). Then, PCC was calculated for every cCRE–gene pair within a 1-Mb window centred on the TSS for every gene. We also generated a set of background pairs by randomly selecting regions from different chromosomes and shuffling of cluster labels. Finally, we fit a normal distribution model and defined a cut-off at PCC score with empirically defined significance threshold of FDR < 0.01, to select significant positively correlated cCRE–gene pairs.

Identification of candidate driver transcription factors

We used the Taiji pipeline to identify candidate driver transcription factors in cell clusters. In brief, for each cell type cluster, we constructed the transcription factor regulatory network by scanning transcription factor motifs at the accessible chromatin regions and linking them to the nearest genes. The network is directed with edges from transcription factors to target genes. The weights of the genes in the network were determined on the basis of the RNA expression level (gene activity score for SGZ NPCs only, because there is no corresponding T-type) of corresponding T-types. The weights of the edges were calculated by the relative accessibility of the promoters of the source transcription factors. We then used the personalized PageRank algorithm to rank the transcription factors in the network. The output of Taiji pipeline is transcription-factor-by-cell type matrix with PageRank scores. From the output matrix, we calculated coefficient of variation across cell types. To identify driver transcription factors, we used following criteria: (1) transcription factors have FDR less than 0.001; (2) transcription factors have coefficient of variant larger than the mean of coefficient of variant; (3) PageRank score should be ranked in the top 25% of all transcription factors for each cell type; (4) RNA expression level (CPM) is larger than 20, which we consider as an expressed transcription factor in corresponding cell type.

Motif enrichment

We performed both de novo and known motif enrichment analysis using Homer (v.4.11, RRID:SCR_010881) with default parameters. Gene Ontology annotation of cCREs was performed using GREAT (version 4.0.4, RRID:SCR_005807) with default parameters. Gene Ontology biological process was used for annotations.

Gene Ontology enrichment

We perform Gene Ontology enrichment analysis using the R package Enrichr (RRID:SCR_001575). The gene set library ‘GO_Biological_Process_4.0.4’ was used with default parameters. The combined score is defined as the P-value computed using the Fisher’s exact test multiplied with the z-score of the deviation from the expected rank.

GWAS enrichment

To enable comparison to GWASs of human phenotypes, we used liftOver with settings ‘-minMatch=0.5’ to convert accessible elements from mm10 to hg19 genomic coordinates. Next, we reciprocal lifted the elements back to mm10 and only kept the regions that mapped to original loci. We further removed converted regions with lengths greater than 1 kb.

We obtained GWAS summary statistics for quantitative traits related to neurological disease and control traits (Supplementary Table 25): age first birth and number of children born, tiredness, Crohn's disease, attention deficit hyperactivity disorder, allergy, birth weight, bipolar disorder, insomnia, sleep duration, neuroticism, coronary artery disease, rheumatoid arthritis, educational attainment, schizophrenia, age at menarche, tobacco use disorder.
We prepared summary statistics to the standard format for linkage disequilibrium score regression. We used homologous sequences for each major cell type as a binary annotation, and the superset of all candidate regulatory peaks as the background control. For each trait, we used cell type-specific linkage disequilibrium score regression (https://github.com/bulk/kdsc) to estimate the enrichment coefficient of each annotation jointly with the background control.62

Statistics
No statistical methods were used to predetermine sample sizes. There was no randomization of the samples, and investigators were not blinded to the specimens being investigated. However, the clustering of single nuclei on the basis of chromatin accessibility was performed in an unbiased manner, and cell types were assigned after clustering. Low-quality nuclei and potential barcode collisions were excluded from downstream analysis as outlined above.

Data availability
Demultiplexed data can be accessed via the NEMO archive (NEMO, RRID:SCR_016152) at: https://assets.nemoarchive.org/dat-wywv153. Additional data are available on our web portal and can be explored here: http://catlas.org/mousebrain. Additional data are available in the NCBI Gene Expression Omnibus (GEO) under accession number GEO173650 and upon request.

Code availability
Custom code and scripts used for analysis can be accessed at: https://github.com/yal054/snATACutils and https://github.com/r3fang/SnapATAC.

62. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science 367, 600–604 (2019).
63. Yao, Z. et al. A transcriptomic and epigenomic atlas of the mouse primary motor cortex. Nature https://doi.org/10.1038/s41586-021-03500-8 (2021).
Acknowledgements We thank J. Huang and E. Lein for critical reading of the manuscript; R. Raviram, Y. Zhang, G. Li and J. Hocker for discussions; and all other members of the Ren laboratory for valuable inputs. We extend our gratitude to the Q2B3 Macrolab at UC Berkeley for purification of the Tn5 transposase. We thank CNS mSlice LLC for making the brain slicer used in this study. This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health (NIH) SIG grant (S10 OD026929). This study was supported by NIH grant U19MH11483. J.R.E is an Investigator of the Howard Hughes Medical Institute. Work at the Center for Epigenomics was also supported by the UC San Diego School of Medicine.

Author contributions Study supervision: B.R. Contribution to data analysis: Y.E.L., K.Z., R.F., Y.Q., O.P., Y.Y. Contribution to data generation: S.P., X.H., J.Y.H., X.W., D.G., S.K., J.L., A.P.-D., M.M.B., X.Y., N.K., M.M., Y.S. Contribution to web portal: Y.E.L., Z.Z., B.L. Contribution to data interpretation: Y.E.L., S.P., B.R., J.R.E., M.M.B., E.A.M., J.C., K.J.G., M.N. Contribution to writing the manuscript: Y.E.L., S.P., B.R. All authors edited and approved the manuscript.

Competing interests B.R. is a co-founder and consultant of Arima Genomics Inc. and co-founder of Epigenome Technologies. K.J.G. is a consultant of Genentech and shareholder in Vertex Pharmaceuticals. A. J.R.E is on the scientific advisory board of Zymo Research, Inc.

Additional information Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03604-1.
Correspondence and requests for materials should be addressed to B.R.
Peer review information Nature thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Anatomic maps of the 45 dissections in the adult mouse cerebrum. **a**, Schematic of brain tissue dissection strategy. Mouse brains were cut into 600-μm-thick coronal slices. **b**, Brain regions dissected from each coronal slice are marked according to the Allen Brain Reference Atlas. The frontal view of each slice from slices 1–11 is shown, with the dissected regions alphabetically labelled on the left, and the anatomic labelling listed on the right. A detailed list of the dissected regions and the full anatomic labelling can be found in Supplementary Table 1.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Quality control metrics of the snATAC-seq datasets. a, Box plots showing the distribution of mapping ratios (the fraction of the mapped sequencing reads) in replicates (rep) 1 and 2 of snATAC-seq experiments from each brain dissection. b, Box plots showing the distribution of the number of proper read pairs (reads are correctly oriented) in replicates 1 and 2 of snATAC-seq experiments. c, Box plots showing the distribution of numbers of unique chromatin fragments detected in replicates 1 and 2 of snATAC-seq experiments. d, Box plots showing the distribution of number of unique barcodes captured in replicates 1 and 2 of snATAC-seq experiments. e, Frequency distribution plot showing the fragment size distribution of each snATAC-seq dataset. f, Heat map showing the pairwise Spearman correlation coefficients between snATAC-seq datasets. The column and row names consist of two parts: brain region name and replicate label. g, Dot plot illustrating fragments per nucleus and individual TSS enrichment. Nuclei in top right quadrant were selected for analysis (TSS enrichment > 10 and > 1,000 fragments per nucleus). h, Fraction of cell collision estimated from species mix samples. Inset shows the fraction of potential barcode collisions detected in snATAC-seq libraries using a modified version of Scrublet™. i, Number of nuclei retained after each step of quality control. j, Distribution of TSS enrichment. k, The number of uniquely mapped fragments per nucleus for individual libraries. l, The number of nuclei passing quality control for subregions. All box plots are stylized as in Fig. 2c.
Extended Data Fig. 3 | Cell clustering based on snATAC-seq data.

a, Schematic diagram of the cell clustering pipeline. b, UMAP embedding of a representative cell clustering at different resolutions from 0.1 to 1.0 using the Leiden algorithm. c, Consensus matrix from 300 iterative clustering runs with different resolutions. d, Cluster stability at different resolutions was assessed using a CDF of consensus matrices. High values illustrate nuclei that clustered together in most cases. e, The PAC and dispersion coefficient at different resolutions. A low PAC and high dispersion coefficient indicates the best and most stable clustering. f, Optimal cell clustering result for a representative subclass (resolution = 0.5). g, The CDF curve of consensus matrix at an optimal resolution for every subclass.
Extended Data Fig. 4 | Summary statistics of snATAC-seq datasets in the current study. a, Violin plots showing the log-transformed number of unique fragments per nucleus in each cell subclass identified. b, Violin plots showing the TSS enrichment in each nucleus of each subclass. c, Acceptance rate from $k$-nearest neighbour batch effect test (kBET) for each subclass of cerebral cells. d, Distribution of the local inverse Simpson’s index (LISI) scores for cells in each subclass.
Extended Data Fig. 5 | Reproducibility of the cell type composition of each brain region estimated from single-cell chromatin accessibility profiles.

a, Bar plot showing the fraction of nuclei from two biological replicates for each of the 43 subclasses of mouse cerebral cells discerned from snATAC-seq data.

b, CDF plot showing the consistency of the estimated fractions of each subclass of cerebral cells between the two biological replicates. Kolmogorov–Smirnov test shows no significant difference between the biological replicates.

c, Box plots of the $P$ values of Kolmogorov–Smirnov tests illustrate consistent results between the two biological replicates for each subclass of cerebral cells across major brain regions, sub-regions and brain dissections tested.

d, Heat map showing the pairwise Spearman correlation coefficients of cell type composition between each replicate of brain dissections. The column and row names consist of two parts: brain region name and replicate label. For example, MOp-1.1 represents the replicate 1 of the first brain dissection of the primary motor cortex (MOp-1). The embedded box plot shows the distribution of Spearman correlation coefficients between two biological replicates, replicates from intra-major brain regions and inter-major brain regions.

*** $P < 0.001$, Wilcoxon rank-sum test. Box plots are stylized as in Fig. 2c.
Extended Data Fig. 6 | Maker genes used for annotation of different subclasses of cerebral cells. 

- **a**, Gene activity scores for marker genes used for subclass annotation. The UMAP embedding from Fig. 1b is shown for reference.

- **b**, Heat map showing the gene activity scores of the marker genes (columns) used for subclass annotation across the 43 subclasses (rows). The colour gradient bar is at the top right.
Extended Data Fig. 7 Iterative clustering identifies cell types for the subclasses of cerebral cells. a, UMAP** embedding of the three classes of cerebral cells, namely non-neurons (left), GABAergic neurons (middle), and glutamatergic neurons (right). b–d, The subclasses of cerebral cells that can be further divided into cell types are shown for the non-neurons (b), GABAergic neurons (c) and glutamatergic neurons (d). UMAP embedding** is shown for each subclass, with the subclass label shown on the top left, and cell-type labels or cluster numbers on each cell type.
Extended Data Fig. 8 | Comparison of cell type compositions in the mouse primary motor cortex determined by snATAC-seq using combinatorial indexing and droplet-based snATAC-seq (10x Genomics) platforms.

a, Individual clustering of snATAC-seq data generated using the single-cell combinatorial indexing (sci) and the droplet-based barcoding (10x) for the primary motor cortex (dissection: 3C). b–c, Co-embedding and joint clustering of snATAC-seq data generated from sci and 10x platforms. b, Dots are coloured by cell clusters. c, Dots are coloured by the experimental platforms (sci or 10x). d, Heat map illustrating the overlap between cell cluster annotations from both platforms. Rows show cell types from combinatorial barcoding; columns show cell types from the droplet-based platform. The overlap between the original clusters and the joint cluster was calculated (overlap score) and plotted on the heat map. e, CDF plot showing the fraction of nuclei in individual cell types for each platform.
Extended Data Fig. 9 | Hierarchical tree depicting the relationships between different subclasses of cerebral cells. Dendrogram tree was constructed using 1,000 rounds of bootstrapping for the subclasses using R package pvclust. Nodes are labelled in grey, approximately unbiased P-values (in red) and bootstrap probability values (in green) are labelled at the shoulder of the nodes, respectively. A full list and a description of cell cluster labels are in Supplementary Table 3.
Extended Data Fig. 10 | The chromatin-accessibility-based cell clustering matches transcriptomics-based cell taxonomy.  

**a**, Heat map showing the similarity between A-type (accessibility) and T-type (transcriptomics) based cell cluster annotations. Each row represents an A-type cluster (a total of 160) and each column represents a T-type cluster (a total of 100). Similarity between original clusters and the joint cluster was calculated as the overlap score, which defined as the sum of minimal proportion of cells/nuclei in each cluster that overlapped within each co-embedding cluster. The overlap score varied from 0 to 1 and was plotted on the heatmap. Joint clusters with an overlap score of >0.5 are highlighted using black dashed line and labelled with RNA–ATAC joint cluster ID. For a full list of cell type labels and descriptions see Supplementary Table 4. 

**b, c**, Bar plots indicating the number of clusters that matched (dark grey) or did not match (light grey) clusters from the other modality. **b**, 155 out of 160 A-types had a matching T-type cluster. **c**, 84 out of 100 T-types had a matching A-type cluster.
Extended Data Fig. 11 | Cellular composition of brain regions, subregions and dissections. a, Cluster dendrogram based on chromatin accessibility as in EDF8. b–d, Normalized percentages (pct) of each subclass in the four major regions (b), the subregions (c), and the dissected regions (d) are shown as different sized dots. The sizes of dots correspond to the percentage, and the colours of the dots indicate the major brain regions, subregions or dissections. Bar plots to the right show the total number of nuclei sampled for each region, subregion, or dissection.
Extended Data Fig. 12 | Distribution of each cerebral cell type across different brain regions, subregions and anatomical dissections. 

a–c. Normalized percentages of cells from each major brain region (a), subregion (b) and dissection (c) are shown as dots in cell types, with the sizes of the dots reflecting the percentages. 

d. Scatter plot showing the reproducibility of regional specificity of each cell type between two biological replicates of brain dissections.
The regional specificity of cerebral cell types defined on the basis of snATAC-seq datasets is consistent with ISH patterns of cell-type-specific genes in the mouse brain. 

**a.** Bar plots showing regional specificity of each cell type determined from relative contribution from five different brain regions, including isocortex, olfactory areas, hippocampal formation, striatum and pallidum. Dot plots showing percentages of each major brain region in the subclass of cerebral cells. The size of each dot reflects the relative contribution of the brain region as indicated by the legend to the right of the panel, and the colour of the dots indicates the brain region. 

**b.** Bar plots showing the coefficients of variation calculated with the average normalized ISH signals of cell-type-specific marker genes for the corresponding cell subclasses across five brain regions. The heat map shows the average normalized ISH signals of cell subclass-specific marker genes. The cell subclass-specific marker genes were identified for joint cell subclasses from RNA-ATAC integrative analysis (Extended Data Fig. 9). For a full list of ISH data of cell subclass-specific genes, see Supplementary Table 6. 

**c.** Scatter plot shows the correlation between the coefficients of variation of marker gene ISH signals and the regional specificity score calculated based on snATAC-seq data for 32 joint cell subclasses from RNA-ATAC integration analysis (Pearson correlation coefficients (PCC) = 0.55). 

**d.** Box plots show the Pearson correlation coefficients (PCC) calculated between cell composition across brain regions based on snATAC-seq data and the spatial distribution ISH signals of cell subclass-specific genes across the five main brain regions for each major brain cell subclass. Box plots are stylized as in Fig. 2c.
Extended Data Fig. 14 | See next page for caption.
Extended Data Fig. 14 | Brain region specificity of different subclasses and cell types. a, UMAP58 embedding of the subclasses of GABAergic neurons. b, Sub-region composition of dopaminergic neurons, D1MSN and D2MSN, and MXD illustrates most MXD neurons were detected in the pallidum. c, Gene activity score for gene *Isl1* in GABAergic neurons predicts expression in MXD. d, RNAISH and bar plot illustration of expression levels show the highest abundance of *Isl1* in the predicted region, pallidum (blue). Data and images were downloaded from © 2004 Allen Institute for Brain Science. Allen Mouse Brain Atlas. Available from: atlas.brain-map.org. e, UMAP58 embedding of PVGA cell types. f, Sub-region composition for each PVGA cell type illustrates that majority of PVGA-7 were detected in nucleus accumbens and caudoputamen. g, Gene activity scores for *Kit* and *Pde3a* predict expression in one PVGA cell type (PVGA-7). h, RNAISH and bar plot illustration of expression levels show expression of *Kit* and *Pde3a* in predicted regions, nucleus accumbens and caudoputamen. Predicted region in blue, other sampled regions in grey and non-sampled regions in white. i, UMAP58 embedding of CA1 glutamatergic neurons (CA1GL) cell types. j, Dissection composition in each CA1GL cell type. k, Gene activity of gene *Dcn* in CA1GL cell types. l, Density of expression level of *Dcn* viewed in BrainExplorer (https://mouse.brain-map.org/static/brainexplorer) shows expression cornu ammonis field 1 (CA1). m, RNAISH and bar plot show highest expression of *Dcn* CA1 and hippocampal formation. The predicted region is coloured in blue, other sampled regions in grey and non-sampled regions in white.
Extended Data Fig. 15 | Statistics of peak calling from snATAC-seq data from each cell type. **a**, Schematic diagram of peak calling and filtering pipeline. **b**, Number of peaks retained after each peak calling and filtering step. **c**, Scatter plots showing the relationship between the number of nuclei in each cluster and the number of peaks before score-per-million (SPM) correction (left) and after SPM correction (right). **d**, Density distribution plot showing the fraction of cells per cell type in which a peak was accessible and a corresponding background for each cell types. For each cell type, the background is defined as same number of non-DHS and non-peak regions randomly picked from genome. **e**, Fraction of different peak sets that overlap with annotated transcriptional start sites, introns, exons, transcriptional termination sites (TTS), and intergenic regions in the mouse genome. **f**, Fraction of different peak sets that overlap with DHSs. **g**, Box plot showing sequence conservation in different peak sets and the control set. **h**, Enrichment analysis of different peak sets with a 15-state ChromHMM model in the mouse brain chromatin. All box plots are stylized as in Fig. 2c.
Extended Data Fig. 16 | Differentially accessible cCREs across different cell types and brain regions.

**a**, Distribution plot of the Jensen–Shannon specificity scores of the differentially accessible cCREs and invariable cCREs. The vertical line shows the cutoff at FDR of 0.05.

**b**, Stacked bar charts showing the number of differential cCREs between cell types for subclasses.

**c**, A graph shows the numbers of differential cCREs in each major brain region.

**d**, A graph shows the numbers of differential cCREs across subregions. The sizes of dots correspond to the number of differential cCREs (log10-transformed) found in each cell type between brain regions, and the colours of the dots indicate the major brain regions in c, subregions in d.
Extended Data Fig. 17 | See next page for caption.
Extended Data Fig. 17 | Variation of chromatin accessibility across different cell types within the subclasses of cerebral neurons. a, UMAP<sup>16</sup> embedding and gene activity scores of Chodl show specificity for one SSTGA cell type. b, Bar chart showing the number of differentially accessible cCREs shared by 1, 2, 3 or 4 cell types of SSTGA. c, Bar chart showing the number of cell-type-specific cCREs, categorized by proximal (black) and distal regions (light grey). d, UMAP<sup>16</sup> embedding of MSGA cell types. e, Bar chart showing the number of differentially accessible cCREs shared by 1, 2, 3, 4 and 5 subtypes of MSGA. f, Number of specific accessible proximal and distal cCREs for each MSGA cell type. g, Heat map showing the normalized accessibility of the cell-type-specific cCREs across different MSGA cell types. h, Heat map showing the promoter accessibility at selected marker genes in each MSGA cell type. i, ISH data (https://portal.brain-map.org<sup>44</sup>) of the marker genes of each MSGA cell type. j, Motif enrichment of uniquely accessible cCREs and expression level of transcription factors in cholinergic neurons (MSGA10). The RNA expression of cholinergic neurons was downloaded from mouse brain atlas (http://mousebrain.org).
Extended Data Fig. 18 | See next page for caption.
Extended Data Fig. 18 | Astrocyte cell types exhibit regional specificity and differential chromatin accessibility. a, UMAP embedding of astrocytes, coloured by cell type (left), fragment depth (middle), and replicate (right). b, Heat map illustrating the overlap between A-type and T-type cell cluster annotations. Each row represents a snATAC-seq subtype and each column represents scRNA-seq cluster (http://mousebrain.org). The overlap between original clusters and the joint cluster was calculated (overlap score) and plotted on the heat map. c, Heat map showing overlap score with cell type defined by integration of multiple transcriptomic and epigenomic modalities. d, Smoothed gene activity scores of marker genes for astrocytes from white and grey matter. e, Smoothed gene activity scores of representative cortical layer-specific marker genes for astrocyte. f, Upset plot showing intersections of differentially accessible cCREs between cell types. cCREs only presented in one cell type are defined as cell-type-specific cCREs. g, Gene activity score for GLI transcription factor family members in astrocyte cell types. h, Smoothed gene activity scores of Oligo2, Itih3, Agt and Slc6a11 in astrocytes show stronger activity in ASCN. i, Genome browser tracks of aggregate chromatin accessibility profiles for each astrocyte cell type at gene Itih3, Slc6a11 and Agt locus. j, Views of ISH experiments from Allen Brain Atlas (atlas.brain-map.org) showing predominant expression of Slc6a11 and Agt in the pallidum. Bar plot showing expression values from ISH experiments in brain structures. Data and images were downloaded from © 2004 Allen Institute for Brain Science. k, UMAP embedding of astrocyte coloured by brain subregions. l, Motif enrichment in cCREs in ASCG with regional specificity.
Extended Data Fig. 19 | Characterization of predicted cCRE–target gene pairs. **a**, Histogram illustrating the 1-D genomic distances between positively correlated distal cCRE and putative target gene promoters. **b**, Box plot showing that genes were linked with a median of 7 putative enhancers. Box plot is stylized as in Fig. 2c. **c**, Accessibility at promoter regions across RNA–ATAC joint cell types; order is as in Fig. 4c. **d**, Enrichment of sequence motifs for CTCF, MEF2 and RFX from de novo motif search in the putative enhancers of module M1 using HOMER. For a full list, see Supplementary Table 21. **e**, Venn diagram illustrating the overlap of putative target genes of cCREs containing binding sites for MEF2, RFX and CTCF, respectively. **f**, Gene Ontology (GO) analysis of the putative target genes of each factor in module M1 was performed using Enrichr. The combined score is the product of the computed P value using the Fisher exact test and the z-score of the deviation from the expected rank. **g, h**, Examples of distal cCRE overlapping a RFX motif (g) or a MEF2 motif (h) and positively correlated putative target genes. Motifs were identified using de novo motif search in HOMER. Genome browser tracks showing chromatin accessibility, mCG methylation levels (see accompanying manuscript) and positively correlated cCRE and gene pairs.
Extended Data Fig. 20 | CTCF-associated cCRE–gene pairs. **a**, Enrichment of CTCF binding sites from 8-week-old mouse forebrain in cCRE modules. **b**, t-SNE of sn-m3C-seq cells (n = 5,142) coloured by clusters. This panel was replicated from the accompanying paper for illustration purposes. **c**, Percentage of cCRE from module M1 with and without CTCF binding overlapping with contact loops identified from snm3C-seq in hippocampus. Shuffled loops are generated by randomly flipping one of the anchors. **d**, Genome browser track view at the Nsg2 locus as an example for CTCF dependent loops. Displayed are chromatin accessibility profiles for several neuronal and non-neuronal clusters, positively correlated cCRE–gene pairs, H3K27ac in postnatal day 0 mouse forebrain and CTCF from cortex of 8-week-old mouse. The triangle heat map (top) shows chromatin contacts in DG neurons and MGC derived from snm3C-seq data (see accompanying manuscript). DG-specific loops are highlighted by boxes.
Extended Data Fig. 21 | Predicted transcription factors involved in adult neurogenesis lineages. 

**a**, Normalized accessibility at marker genes in NIPCs. 

**b**, RNA ISH shows expression levels of marker genes of NIPCs in the SGZ and SVZ. Images were downloaded from © 2004 Allen Institute for Brain Science 44. Allen Mouse Brain Atlas. Available from: atlas.brain-map.org. 

**c**, Genome browser tracks showing representative differentially accessible cCREs in NIPCs from SVZ.
Extended Data Fig. 22 | Mouse cerebral cCREs maps help to interpret potential casual risk variants of neurological diseases. **a**, Bar plot showing the percentage of cCREs identified in the current study with homologous sequences in the human genome (using reciprocal homology search) (Methods). **b**, Chromatin accessibility profiles for several neuronal and non-neuronal cell types, and posterior probabilities of association (PPA) for potential casual variants surrounding the homologous region in the mouse genome to a schizophrenia-associated locus. Grey bars highlight cCREs overlapping potential causal variants. rsID or hg19 coordinates of overlapped variants are labelled: ‘D’ denotes that the alternative allele is a deletion; ‘I’ denotes an insertion. Predicted positively correlated cCRE–gene pairs are shown in red arcs.
Extended Data Fig. 23 | Nuclei gating strategy. After tagmentation, nuclei were pooled and stained with DRAQ7. First, potential nuclei were identified using forward scatter (FSC) area and back scatter (BSC) area (left dot plot). Next, potential doublets were removed based on BSC and FSC signal width (two middle dot plots). Finally, 20 diploid nuclei (2n) were sorted into each well of eight 96-wells plates (right dot plot).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers.

We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Demultiplexed data can be accessed via the NEMO archive at: https://assets.nemoarchive.org/dat-wywv153
Processed data are available on our web portal and can be explored here: http://catlas.org/mousebrain
Additional data are available in the NCBI Gene Expression Omnibus (GEO) under accession number GEO173650 and upon request.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No statistical methods were used to predetermine sample sizes. |
|-------------|---------------------------------------------------------------|
| Data exclusions | No samples were excluded. For analysis only nuclei with > 1,000 reads/nucleus and transcriptional start site enrichment > 10 were selected. Potential barcode collisions were excluded from analysis. |
| Replication | Experiments were performed for 2 biological replicates for each dissected region |
| Randomization | There was no randomization of the samples |
| Blinding | Investigators were not blinded to the specimens being investigated. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒ Antibodies |
| ☒ Eukaryotic cell lines |
| ☒ Palaeontology |
| ☒ Animals and other organisms |
| ☒ Human research participants |
| ☒ Clinical data |

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒ ChIP-seq |
| ☒ Flow cytometry |
| ☒ MRI-based neuroimaging |

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

| Laboratory animals | Adult C57BL/6J male mice were purchased from Jackson Laboratories. |
|--------------------|------------------------------------------------------------------|
| Wild animals | No wild animals were used in this study |
| Field-collected samples | No field-collected samples were used in this study |
| Ethics oversight | All experimental procedures using live animals were approved by the SALK Institute Animal Care and Use Committee under protocol number 18-00006. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.
### Flow Cytometry

#### Plots

Confirm that:
- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a ‘group’ is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

#### Methodology

| Sample preparation | Nuclei were stained with DRAQ7 (#7406, Cell Signaling) |
|--------------------|--------------------------------------------------------|
| Instrument         | Sony SH800                                             |
| Software           | SH800S software                                        |
| Cell population abundance | NA                                                  |
| Gating strategy   | Potential nuclei were first identified using FSC-Area and BSC-Area. Next doublets were removed based on BSC and FSC signal width. DRAQ7 positive nuclei with 2n count were sorted |

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.