

Title: A comparative atlas of single-cell chromatin accessibility in the human brain

Authors: Yang Eric Li¹, Sebastian Preissl², Michael Miller², Nicholas D. Johnson³, Zihan Wang⁴, Henry Jiao², Chenxu Zhu¹, Zhaoning Wang¹, Yang Xie¹, Olivier Poirion², Colin Kern², Antonio Pinto-Duarte³, Wei Tian³, Kimberly Siletti⁵, Nora Emerson³, Julia Osteen³, Jacinta Lucero³, Lin Lin², Qian Yang², Quan Zhu², Sarah Espinoza², Anna Marie Yanny⁶, Julie Nyhus⁶, Nick Dee⁶, Tamara Casper⁶, Nadiya Shapovalova⁶, Daniel Hirschstein⁶, Rebecca D. Hodge⁶, Sten Linnarsson⁵, Trygve Bakken⁶, Boaz Levi⁶, C. Dirk Keene⁷, Jingbo Shang⁴, Ed S. Lein⁶, Allen Wang², M. Margarita Behrens³, Joseph R. Ecker^{3,8}, and Bing Ren^{1,2,9*}.

Affiliations:

¹ Department of Cellular and Molecular Medicine, University of California, San Diego, San Diego, CA, USA

² Center for Epigenomics, University of California San Diego, School of Medicine, La Jolla, CA, USA.

³ The Salk Institute for Biological Studies, La Jolla, CA, USA

⁴ Department of Computer Science and Engineering, University of California San Diego, CA, USA

⁵ Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Stockholm, Sweden

⁶ Allen Institute for Brain Science, Seattle, WA, USA

⁷ Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA

⁸ Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA, USA

⁹ Ludwig Institute for Cancer Research, La Jolla, CA, USA

*Corresponding author. Email: biren@health.ucsd.edu

Abstract:

The human brain contains an extraordinarily diverse set of neuronal and glial cell types. Recent advances in single cell transcriptomics have begun to delineate the cellular heterogeneity in different brain regions, but the transcriptional regulatory programs responsible for the identity and function of each brain cell type remain to be defined. Here, we carried out single nucleus ATAC-seq analysis to probe the open chromatin landscape from over 1.1 million cells in 42 brain regions of three neurotypical adult donors. Integrative analysis of the resulting data identified 107 distinct cell types and revealed the cell-type-specific usage of 544,735 candidate cis-regulatory DNA elements (cCREs) in the human genome. Nearly 1/3 of them displayed sequence conservation as well as chromatin accessibility in the mouse brain. On the other hand, nearly 40% cCREs were human specific, with chromatin accessibility associated with species-restricted gene expression. Interestingly, these human specific cCREs were enriched for distinct families of retrotransposable elements, which displayed cell-type-specific chromatin accessibility. We uncovered strong associations between specific brain cell types and neuropsychiatric disorders. We further developed deep learning models to predict regulatory function of non-coding disease risk variants.

Main Text:

Neurological disorders and mental illnesses are the leading cause of disease burdens in the United States(1). Tens of thousands of sequence variants in the human genome have been linked to the etiology of neuropsychiatric disorders(2, 3). However, interpreting the mode of action of the identified risk variants remains a daunting challenge since the vast majority of them are non-protein-coding (4, 5). It is increasingly suspected that a large fraction of the non-coding risk variants contribute to disease etiology by perturbing transcriptional regulatory elements and target gene expression in the disease-relevant cell types(6–8). However, a lack of the maps and tools to explore gene activities and their transcriptional regulatory sequences at high cellular and anatomical resolution in the brain prevents a clearer mechanistic understanding of the broad spectrum of neuropsychiatric disorders.

The human brain is made up of hundreds of billions of neurons, which through trillions of synapses form a complex neurocircuitry to carry out diverse neurocognitive functions. The functionality of the neural circuitry is supported and maintained by an even greater number of glial cells including astrocytes, oligodendrocytes, oligodendrocyte precursor cells, and microglia, among others. Single-cell RNA-seq and high throughput imaging experiments have produced detailed cell taxonomies for mouse brains and a few cortical regions in human (9–17), leading to a comprehensive view of cell types and their molecular signatures in several brain regions(18–20). Analysis of gene expression patterns using single-cell transcriptomics and spatial transcriptomics assays (9, 11, 12, 16, 21–24), have further advanced our understanding of the transcriptional landscapes in different brain cell types. In contrast to the large body of transcriptomic analyses, analysis of the regulatory elements that drive the cell-type specific expression of genes is lagging. Current catalogs of candidate regulatory sequences in the human genome, most notably those generated by ENCODE and Epigenome Roadmap consortia (6–8, 25, 26), still lack the information about cell-type-specific activities of each element especially those identified from brain tissues, because conventional assays performed using bulk tissue samples, unfortunately, fail to resolve cCREs in individual cell types comprising the heterogeneous tissues. To address this problem, recent technological advances have enabled the analysis of open chromatin at single cell resolution (27–32) in adult mouse tissues and several brain regions (27, 29, 33–35), generating cell-

type-specific maps of gene regulatory elements for a limited number of human brain cell types and brain regions (36–38).

As part of the BRAIN Initiative Cell Census Network (BICCN), we have carried out single-cell profiling of transcriptome, chromatin accessibility, and DNA methylome across >40 regions in the human brain from multiple neurotypical adult donors. Here, we describe a single-cell chromatin accessibility atlas comprising ~1.1 million human brain cells and results from integrative analysis of this genomic resource. We used this chromatin atlas to define 107 distinct cell types, and uncover the state of chromatin accessibility at 544,735 cCREs in one or more of these brain cell types. We integrate our chromatin atlas with the companion single cell transcriptome and DNA methylome atlases to link cCREs to putative target genes. We compared with a previous catalog of cCREs from mouse cerebrum, finding strong evolutionary conservation in both sequence and chromatin accessibility for nearly 1/3 of human brain cCREs. We further predict disease relevant cell types for 19 neurological traits and diseases. Finally, we developed machine learning models to predict the regulatory function of disease risk variants. We created an interactive web atlas to disseminate this resource (cis-element ATLAS [CATLAS]; <http://catlas.org>).

Result

A single-cell chromatin accessibility atlas of human brains

We dissected 42 brain regions from the human cortex (CTX), hippocampus (HIP), basal nuclei (BN), midbrain (MB), thalamus (THM), cerebellum, and pons from three neurotypical male donors (D1, D2, and D4) at the age of 29, 42, and 58, respectively (**Fig. 1A, Table S1**), according to the Allen Brain Reference Atlas (39). For each brain sample, we performed snATAC-seq using a protocol described previously (40) (**Fig. 1A, Fig. S1A-D, Table S2**). The reliability of the data was evidenced by sequencing reads showing nucleosome-like periodicity (**Fig. S1E**), excellent correlation between datasets from similar brain regions across the three donors (**Fig. S1F**), high transcription start site (TSS) enrichments as well as other quality control metrics (see **Methods**). A total of 1,290,974 nuclei passed

stringent quality control criteria (**Fig. S1G**, see **Methods**). After removing an additional 156,614 snATAC-seq profiles that likely resulted from potential barcode collision or doublets (**Fig. S1H-J**, see **Methods**), a total of 1,134,360 nuclei were retained. Among them, 595,713 were from cortex, 72,190 from HIP, 317,480 from BN, 23,114 from MB, 50,768 from THM, 51,775 cerebellum and 25,459 from pons (**Table S3**). On average 4,970 chromatin fragments were detected in each nucleus (**Table S3**, **Fig. S1K-M**, see **Methods**).

We next carried out iterative clustering with the above snATAC-seq profiles and classified them into three major classes: glutamatergic (vGlut+) or excitatory neurons (11.8%), GABAergic (GABA+) or inhibitory neurons (6.8%) and non-neuronal cells (81.4%) (**Fig. 1B, D, F**, **Fig. S2**, **Fig. S3A-B**). We further classified the three cell classes into 14 sub-classes of glutamatergic neurons, 2 sub-classes of granule cell types, 1 sub-class of cholinergic neurons, 4 sub-classes of dopaminergic neurons, 2 sub-classes of thalamic and midbrain derived neurons, 11 sub-classes of cortical GABAergic neurons, and 8 sub-classes of non-neuronal cells (**Fig 1B, D, F**). Each sub-class was annotated based on chromatin accessibility at promoters and gene bodies of at least three marker genes of known brain cell types, together with the brain region (**Fig. 1C, E, G**, **Fig. S3C**, **Table S4**, and **S5**). For each sub-class, we also conducted a third round of clustering and identified a total of 107 distinct brain cell types (**Fig. 1H**, **Fig. S4**, **Table S3**, see **Methods**). To determine the optimal number of cell types within each sub-class, we evaluated the relative stability from a consensus matrix based on 100 rounds of clustering with randomized starting seeds (see **Methods**). We then calculated the proportion of ambiguous clustering (PAC) score and dispersion coefficient (DC) to find the optimal resolution (local minimum and maximum) for cell type clustering (**Fig. S2B-E**). For example, *VIP* positive (*VIP*⁺) neurons (*VIP*) were further divided into multiple cell types with distinct chromatin accessibility at multiple gene loci (**Fig. 1I**, **Fig. S2B-E**, **Fig. S4**). We found that the clustering result of snATAC-seq was robust to variation of sequencing depth, and signal-to-noise ratios, and most cell sub-classes showed no batch effect from at least two donors using local inverse Simpson's index (LISI) analysis, with the exception of two cell sub-classes (SUB, granule cells from hippocampus, SMC, vascular smooth muscle cells) that were mostly captured from one donor (**Fig. S5**). To capture the relative similarity in chromatin landscapes

among the 42 sub-classes we constructed a robust hierarchical dendrogram (**Fig. 1H, Fig. S7**). This dendrogram shows known organizing principles of human brain cells: the non-neuronal class is separated from the neuronal class, which are further separated based on neurotransmitter types (GABAergic, dopaminergic, cholinergic, and glutamatergic) and developmental origins (**Fig. 1H, Fig. S6**, see **Methods**).

All neuronal and some glial cell types are distributed in the human brain in a non-uniform fashion (**Fig. 1J**). We defined a regional specificity score for each sub-class based on the contribution from different brain regions. While the majority of glial cell types are ubiquitously distributed throughout the brain, showing very low regional specificity (**Fig. 1J**, right), there are exceptions. For example, the Bergmann glia (ACBGM) is specifically found in the cerebellum. On the other hand, all neuron types are characterized by moderate to high regional specificity (**Fig. 1J**, right). We found a stark separation based on brain sub-regions for distinct neuron types including the granular cells in the cerebellum (CBGRC) and medium spiny neurons (MSN) in the basal ganglia. For glutamatergic neurons, we also observed distinct types of intra-telencephalic (IT) cortical neurons in the primary visual cortex (V1C) (IT-V1C), and excitatory neurons from the amygdala (AMY) to be highly restricted to specific brain regions or dissections.

We also compared our single-nucleus chromatin accessibility derived cell clusters with the cell taxonomy defined from other modalities. We performed the integrative analysis with single-cell RNA-seq (scRNA-seq) and single-nucleus methylome sequencing (snmC-seq) (companion papers, Siletti, et al. 2022 and Tian, et al. 2022, see **Methods**). For 38 of 42 sub-classes defined based on chromatin accessibility (A-Type), we could identify a corresponding sub-class defined based on transcriptomics (T-Type; overlap score cut-off >0.2 , **Fig. S7**). In addition, for 39 out of 42 A-types we identified one, or in some cases more, corresponding sub-class defined based on DNA methylation (M-Type, **Fig. S7**). Of note, a few cell clusters can hardly be aligned with other modalities, possibly due to differences in brain dissections and/or under sampling of neurons in our datasets. For example, the A-type granule cells from cerebellum (CBGRC) were not identified in T-types, but aligned well with M-type. The

cholinergic neurons were not captured in M-types, but aligned to a mixture class of T-types called Glycinergic, cholinergic, monoaminergic, or peptidergic neurons.

Mapping and characterization of human brain cCREs

As a first step towards defining the gene regulatory programs that underlie the identity and function of each brain cell type, we identified the open chromatin and candidate *cis-regulatory* DNA elements (cCREs) in each of the 107 brain cell types. We aggregated the chromatin accessibility profiles from the nuclei comprising each cell cluster/type and identified the open chromatin regions with MACS2(41) (Fig. S8A). We filtered the resulting accessible chromatin regions based on whether they were called in at least two donors, or in two pseudo-replicates (Fig. S8A-B). From our previous study, we found that read depth or cluster size can affect MACS2 peak calling scores. We therefore used “score per million” (SPM)(42) to correct this bias (Fig. S8C, see Methods). About 1000 nuclei in a cluster were sufficient to identify over 80% of the accessible regions in a cell type, consistent with our previous finding(40) (Fig. S8C). We iteratively merged the open chromatin regions identified from every cell type, and kept the summits with the highest MACS2 score for overlapped regions. On average, we detected 62,045 open chromatin regions per cell type (500 bp in length), and a union of 544,735 open chromatin regions across all 107 cell types (Fig. S8D, Table S6, and S7, see Methods). These cCREs together made up 8.8% of the human genome (hg38) (Table S8). Of these cCREs, 95.3% were located at least 2 kbp away from annotated promoter regions of protein-coding and lncRNA genes (Fig. 2A, Table S8). The promoter-distal cCREs were distributed in introns (34.8%), intergenic (27.8%), and other genomic regions. Interestingly, 22% of them overlap with endogenous retrotransposable elements, including long terminal repeats class (LTR, 6.8%), LINEs (long interspersed nuclear elements, 11.3%), SINEs (short interspersed nuclear elements, 3.9%). Several lines of evidence support the authenticity of the identified cCREs. First, both proximal and distal cCREs showed higher levels of sequence conservation than random genomic regions with similar GC content (Fig. 2B). Second, 89.6% of cCREs overlapped with DNase hypersensitive sites (DHS) previously mapped in a broad spectrum of bulk human tissues and cell types including fetal and adult brains (43). This list further expands candidate

CREs previously annotated in the human genome by the ENCODE(26) and a recent survey of chromatin accessibility in single nuclei across fetal and adult human tissues (44)(**Fig. 2C**).

To define the cell type specificity of the above cCREs, we first plotted the median levels of chromatin accessibility against the maximum variation for each element (**Fig. 2D**). We found that the majority of cCREs displayed highly variable chromatin accessibility across the brain cell types identified in the current study, except for a small proportion of invariable cCREs (2.0 %) that showed accessibility in virtually all cell clusters, of which 87% were at proximal regions to TSS (**Fig. 2D**). To characterize the cell type specificity of the cCREs more explicitly, we used non-negative matrix factorization to group them into 37 modules, with elements in each module sharing similar cell type specificity profiles. Except for the first module (M1) which included mostly cell-type invariant cCREs, the remaining 36 modules displayed highly cell-type restricted accessibility (**Fig. 2E, Table S9, and S10**). These cell type restricted cCRE modules were enriched for distinct sets of motifs recognized by known transcriptional regulators (**Table S11**). For example, the NEUROG2 and ASCL1 enriched in module M4 for intratelencephalic (IT) neurons at cortical layer 2/3 (IT-L2/3) have been reported to be proneural genes and critical for cortical development (**Table S11**)(45). The SOX family factors in module M35 for oligodendrocytes (OGC) are pivotal regulators of a variety of developmental processes (**Table S11**)(46). These results lay a foundation for dissecting the gene regulatory programs in different brain cell types and regions.

Linking distal cCREs to target genes

To investigate the transcriptional regulatory programs that are responsible for cell-type-specific gene expression patterns in the human brain, we carried out an integrative analysis that combines the snATAC-seq data collected in the current study with scRNA-seq data generated by a companion paper (Siletti, et al. 2022) from matched brain regions (**Fig. S7**). We first connected 255,828 distal cCREs to 14,861 putative target genes by measuring the co-accessibility across single nuclei in every cell subclass using Cicero(47) (**Fig. 2F**, upper, see **Methods**), which resulted in a total of 1,661,975 gene–

cCRE pairs within 500 kb of each other. Next, we identified the subset of cCREs the accessibility of which positively correlates with the expression of putative target genes and therefore function as putative enhancers in neuronal or non-neuronal types (**Fig. 2F**, bottom). This analysis was restricted to distal cCREs and expressed genes captured from 27 matched cell sub-classes defined above from integrative analysis between snATAC-seq and scRNA-seq (**Fig. S7**). We revealed a total of 265,049 pairs of positively correlated cCRE (putative enhancers) and genes at an empirically defined significance threshold of $FDR < 0.05$ (**Table S12**). These included 114,877 putative enhancers and 13,094 genes (**Fig. 2G, Fig. S9, Table S12**). The median distance between the putative enhancers and the target promoters was 176,345 bp (**Fig. S9A**). Each promoter region was assigned to an average of 7 putative enhancers, and each putative enhancer was assigned to two genes on average (**Fig. S9B-C**).

We further classified these putative enhancers into 27 modules by using non-negative matrix factorization (**Table S13, and S14**), to investigate how cell-type-specific gene expression is regulated. The putative enhancers in each module displayed a similar pattern of chromatin accessibility across cell sub-classes (**Fig. 2H**), and the expression of putative target genes showed a correlated pattern (**Fig. 2I**). This analysis also revealed a large group of 5,113 putative enhancers that were linked to 4,775 genes expressed at a higher level across all neuronal cell clusters than in non-neuronal cell types (module M1) (**Fig. 2H, I, Table S13, and S14**). These putative enhancers are strongly enriched for CTCF, and RFX binding sites (**Table S15**), which is consistent with what we previously found in the mouse cerebrum(40).

We also uncovered modules of enhancer–gene pairs that were active in a more restricted manner (modules M2–M27) (**Fig. 2H-J**). For example, we identified modules (M2-M7) associated with several cortical glutamatergic neurons (IT-L2/3, IT-L4, IT-L5, IT-L6-1, IT-L6-2), in which the putative enhancers were enriched for sequence motifs recognized by the bHLH factors NEUROD1 (**Fig. 2J, Table S13-S15**). Another example was module M15 associated with medium spiny neurons (MSN), in which putative enhancers were enriched for motifs recognized by MEIS factors, which play an important role in establishing the striatal inhibitory neurons (**Fig. 2J, Table S13-S15**). Module M25

was associated with microglia (MGC). Genes linked to putative enhancers in this module were related to immune genes and the putative enhancers were enriched for the binding motif for ETS-factor PU.1, a known master transcriptional regulator of this cell lineage (**Fig. 2J**, **Table S13-S15**). This observation is consistent with the paradigm that cell-type-specific gene expression patterns are largely established by distal enhancer elements.

Regional specificity of glial and neuronal cell cCREs

The single-cell atlas of chromatin accessibility generated in this study provides a unique opportunity to characterize the heterogeneity of the gene regulatory programs that might underlie the specialized functions of glial and neuronal cells in each brain region. While most non-neuronal cell types, including oligodendrocytes (OGCs), oligodendrocyte precursor cells (OPCs), microglia (MGC), telencephalon astrocytes (ASCTs), non-telencephalon (ASCNTs), and various vascular cells were ubiquitously distributed throughout the different brain dissections (**Fig. 1J**), molecular diversity has been recently reported in these cells in juvenile and adult vertebrates(48–51). We therefore leverage the largest collection of >900,000 single nuclei of non-neuronal cells and the high-resolution brain dissections in the present study (**Fig. S1N**), to more thoroughly characterize the regulatory diversity of the non-neuronal populations.

The UMAP embeddings in the brain regional spaces showed a gradient among cell types of OGCs, OPCs, MGCs, and ASCTs (**Fig. 3A-B, E-F, I-J, and M-N, Fig. S10**). We hypothesized that these gradients may reflect heterogeneity in cCRE usage in these glial cells across brain regions. We first calculated the averaged chromatin accessibility and coefficient of variation (CV) across 42 brain regions for every cCRE identified in OGCs, OPCs, MGCs, ASCTs, respectively (**Fig. 3C, G, K, and O, Fig. S10**). To our surprise, a large number of cCREs displayed highly variable chromatin accessibility across the brain regions (**Fig. 3C, G, K, and O, Fig. S10**). In total 55,304 or 40.1% of total cCREs identified in OGCs, 43,574 or 33.0% of total cCREs identified in OPCs, 37,962 or 34.5% of total cCREs identified in MGC, and 46,979 or 33.1% of total cCREs identified in ASCTs, were found to show variability in

chromatin accessibility across major brain regions. These variable cCREs show similar regional specificities across three donors. Next, applying k-means clustering analyses to the variable cCREs identified in the above glial cell populations (**Fig. 3D, H, L, and P**), we revealed distinct open chromatin patterns in OGCs, OPCs, and ASCTs from the cerebellum (CB). Interestingly, a large fraction of these variable cCREs showed higher chromatin accessibility in the cerebellum (**Fig. 3D, H, and P**). We also observed loss of chromatin accessibility in a large number of cCREs in distinct brain structures (**Fig. 3D, H, L, and P**).

We found a diverse population of both telencephalic (ASCT) and non-telencephalic astrocytes (ASCNT) in different major brain structures (**Fig. 3M, and N**). We identified three ASCNT cell types from sub-clustering of astrocytes, and one cell population restricted to the cerebellum that was annotated as Bergmann glial cell (ASBGM) (**Fig. 3Q, and R, Table S5**). One cell type (ASCNT-1) was detected mostly in the thalamus, midbrain, and pons, whereas the other two ASCNT cell types were predominantly found in the cortex, hippocampus, and cerebral nuclei (CN). To characterize the dynamic epigenome, we compared the open chromatin landscapes among different cell types using a likelihood ratio test (**Fig. 3S, Table S16, see Methods**), and identified a total number of 8,790 cCREs that exhibited cell-type-restricted accessibility (range: 100–3,787) (**Fig. 3S**). A human enhancer, specifically accessible in the ASCNT-1 type, was previously validated by mouse transgenics to be active in the midbrain (**Fig. 3T**). We further performed motif analysis for differentially accessible regions in these cell types, finding enrichment of both shared and specific TF binding motifs. For example, we found CCAAT box-binding transcription factor (CTF) NF1 enriched in differential regions identified from both ASCT-1 and ASBGM, whereas TF motifs from nuclear receptors (NRs) and zinc-finger families are specifically enriched in different types (**Fig. 3U, Table S17**).

Medium spiny neurons (MSN) have two primary characteristic types: D1-type and D2-type MSNs defined by the expression of two dopamine receptors DRD1 and DRD2. Although these two sub-classes showed significantly different gene expression patterns and were well separated from each other in transcriptomic and methylation datasets (companion paper, Siletti, et al. 2022 and Wei, et al. 2022), we

found that the cell populations of MSNs were well separated by the sub-regions in basal ganglia, rather than D1- and D2-types. (**Fig. S11A-C**). We identified and annotated several cell types as D1-/D2-MSN types based on the chromatin accessibility at gene *DRD1* and *DRD2* loci and sub-regions (**Fig. S11C**). Specifically, we found both D1- and D2-MSN populations from putamen (Pu) (**Fig. S11A-C**). Similarly, D1- and D2-types were also found in the body of caudate (CaB) (**Fig. S11A-C**). We identified one cell type restricted in nucleus accumbens (NAC), although the nuclei also show a distinct pattern of chromatin accessibility at *DRD1* and *DRD2* loci (**Fig. S11A-C**). We hypothesized that the regional specificity of D1-/D2-MSN types is accompanied by differential chromatin accessibility at the cCREs, which likely underlies cell-specific gene expression patterns. We found a total of 35,203 differential accessible cCREs (**Fig. S11D, Table S18**). Higher chromatin accessibility variation has been found in cell populations from different brain sub-regions, compared to the difference found between D1- and D2-MSN types (**Fig. S11D**). Consistent with above observations, MSN-types were better separated by brain sub-regions, other than separated by D1- and D2- types (**Fig. S11A-C**). The motif analysis suggested the gene expression in these cell types can be regulated by both brain region specific, and D1-/D2- cell-type specific TFs. For example, we found *DLX* and *LHX* from the Homeobox TF family were specifically enriched in D1/2-NAC, whereas *MEF2* and *RORA/G* were enriched in both D1- and D2-MSN types from putamen but not in other brain sub-regions (**Fig. S11E, Table S19**). In addition, we found *EGR1* and *WT1* from zinc-finger TF family were specifically enriched in D2-MSN type but not in D1-MSN types from caudate (**Fig. S11E, Table S19**).

Conservation of chromatin accessibility at cCREs between mouse and human brain cells

To determine how conserved the gene regulatory landscapes are between human and mouse brains, we compared the human brain cCREs defined in this study with our previously published map of mouse cerebrum cCREs(40). We first performed joint clustering of 18 neuronal and glial cell sub-classes (each with >1,000 single nuclei) (**Fig. 4A, Fig. S12**)(40). We used multiple molecular features, including “gene activity scores” at homologous genes, chromatin accessibility at homologous cCREs, and

transcription factor motif enrichment scores (**Fig. S12A-G**, see **Methods**). We note that clustering based on gene activity scores alone does not align brain cell types between the two species (**Fig. S12A, D, and E**), because of a lack of general conservation of gene expression patterns, as reported previously (52, 53). We also identified orthologues of the human cCREs in the mouse genome by performing reciprocal homology searches, and overlapped with cCREs identified in the mouse cerebrum (**Table S20, Fig. S12H-I**, see **Methods**). Clustering using the chromatin accessibility at homologous cCREs alone also failed to align corresponding cell types, likely due to significant levels of CRE turnovers (**Fig. S12B, and F**). Instead, we found that clustering based on TF motif enrichment allows for a reasonable alignment of brain sub-classes between species (**Fig. S12C, and G**). We observed that sequence motif enrichment scores were the most conserved molecular features that reliably align similar cell sub-classes between human and mouse brains (**Fig. 4B, Fig. S12G**). These analyses suggested that although the open chromatin landscapes were largely divergent between species, the gene regulatory programs of similar cell types may share a similar “vocabulary”, which is the specific combination of transcription factors in each cell lineage.

We found that for ~60% of the human cCREs, mouse genome sequences with high similarity could be identified (more than 50% of bases lifted over to the human genome) (**Fig. 4C, Fig. S12H-I, Table S20**). Among these orthologues’ genome sequences, only half of them (32.8% of total human cCREs) were identified as open chromatin regions in any cell sub-classes from the mouse cerebrum, whereas the other half were not. We thus defined the 32.8% of human cCREs with both DNA sequence similarity and open chromatin conservation as chromatin accessibility (CA) conserved cCREs, and 26.8% of human cCREs with only DNA sequence similarity as chromatin accessibility (CA) divergent cCREs. In addition, we found 40.4% of human cCREs to be without orthologous genome sequences in the mouse genome, and refer to them as human-specific cCREs (**Fig. 4C, left, Table S21**). This general pattern was consistent with what has been reported in other cell types between human and mouse(54). Next, we further performed the same analyses on the cCREs within the corresponding cell sub-classes from both species. We observed a similar pattern and the proportion of different categories of cCREs was relatively consistent between various cell sub-classes (**Fig. 4C, right, Fig. S12J, Table S21**).

We next characterized the genomic distribution of different categories of cCREs. Not surprisingly, we observed that a large proportion of CA-conserved cCREs were located at or near the promoter-TSS regions in the human genome. In addition, the human-specific cCREs were enriched for transposable elements (TE), such as LINEs, SINEs, and LTRs (**Fig. 4D, Table S21**). Most TE families show invariable chromatin accessibility across sub-classes (**Fig. 4E, and F**), and they are dominated by LINE and LTR superfamilies (**Fig. 4F**). Interestingly, 178 TE families display cell-type-specific patterns of chromatin accessibility (**Fig. 4G**). Previous reports suggest that certain transposable elements are active in mammalian brains, and could hypothetically contribute to vulnerability to disease(55, 56). Our findings support the hypothesis that distinct TE families might be activated in specific brain cell types. For example, LTRs, including but not limited to LTR13A, LTR71A, and HERV9-int, display chromatin accessibility in microglia but not in other sub-classes of brain cells (**Fig. 4G**). The LTR13A has been reported to act as cellular gene enhancers(57, 58). We observed that the LTR13A also shows variable accessibility in microglia populations in different brain regions. For example, we found higher accessibility in brain regions such as posterior parahippocampal gyrus (TH-TL), primary visual cortex (A1C) and primary motor cortex (M1C) from the cortex, substantia innominata and nearby nuclei (SI) and corticomedial nuclear group (CMN) from cerebral nuclei, while lower accessibility in midbrain and cerebellum (**Fig. 4H**). Furthermore, we observed that chromatin accessibilities at LTR13A in microglia from TH-TL, CMN and SI vary considerably among the donors, (**Fig. 4I**), but not in other brain regions (**Fig. 4J**). The chromatin accessibility of LTR13A was associated with the activation indicated by RNA expression signals (**Fig. 4K**), though the biological significance of this observation requires further investigation.

Taking advantage of the integration of multi-modal datasets, we can better delineate the gene regulatory programs that underlie the identity and function of each brain cell type (**Fig. 5A, Fig. S13A-E**). The data integration was anchored on shared genomic features. Here, the RNA expression levels of a common set of genes were profiled from snRNA-seq and Paired-Tag assays. The chromatin accessibility from snATAC-seq at both 2kb upstream and gene body, defined as “gene activity score”,

showed a strong correlation with the expression level, which was used for integration(47). Given the negative correlation of DNA methylation levels, especially mCG at gene bodies with RNA expression, the expression levels of genes were imputed from both snmC and snm3C-seq data (**Fig. 5A**, see **Methods**). Through the integration of those single cell genomic datasets, we were able to better interrogate the gene regulatory programs in cell types with information from gene expression, open chromatin, histone modification, DNA methylation, and chromosome contact (**Fig. 5A**). We collected single-cell genomic datasets profiled from the human primary cortex (M1C) and middle temporal gyrus (MTG), including 149,891 cells from scRNA-seq, 27,383 cells from Paired-Tag (only from M1C), 55,974 cells from snATAC-seq, 10,604 cells from snmC-seq, and 16,257 cells from snm3C-seq (**Fig. 5B-C**). We performed co-embedding cell clustering analysis on these datasets. We first normalized technical confounders to recover sharper biological heterogeneity using SCTransform(59). Then, dimensionality reduction (i.e., PCA) was conducted by using a common gene set with Seurat(60). Later, the PCA embeddings were iteratively corrected between assays using the Harmony algorithm(61). Finally, the cell clusters were identified from the k-nearest neighbor space calculated from the corrected PCA and visualized on UMAP embedding (see **Methods**). The different single-cell assays showed excellent agreement in the same co-embedding space, which indicates the high quality of common variable features and the success of the integration strategy (**Fig. 5C**).

We characterized the gene program in VIP-4, one cell type of VIP⁺ neurons, which showed distinct chromatin accessibility at the *CHRNA2* gene locus (**Fig. 5D**). *CHRNA2* encodes a subunit of nicotinic cholinergic receptor, which is involved in fast synaptic transmission. The co-localization of marker gene *BTBD11* for VIP⁺ neurons and *CHRNA2* from RNAscope(62) fluorescence *in situ* hybridization (FISH) experiment first validated the existence of VIP-4 type in the human cortex (**Fig. 5E, Fig. S13F**). We also noticed that the expression of *CHRNA2* was restricted in human VIP⁺ cell types, but not in any VIP⁺ cell types identified from mouse brain (**Fig. 5F**) (9, 63, 64). We explored whether the human-specific expression of *CHRNA2* was regulated by specific cCREs in the VIP-4 type. We identified a total of 40,086 differential cCREs between 7 VIP⁺ cell types (**Figure 5G, Table S22**). One differential cCRE located downstream of the gene *CHRNA2* showed higher accessibility in VIP-4, than in

Oligodendrocyte precursor cells (OPC) (**Fig. 5H**, ATAC tracks for 7 VIP⁺ cell type and aggregated signals for VIP and OPC). This cCRE was also characterized as a human-specific cCRE in VIP⁺ neurons (**Fig. 4C**, **Table S20**). The specific accessibility of this cCRE and promoter of *CHRNA2* in VIP⁺ neurons were supported by mCG signals from snmC-seq (**Fig. 5H**, mCG tracks). This cCRE was predicted as one putative enhancer that regulates the expression of *CHRNA2* (**Fig. 5H**, red arcs). The potential active function of this CRE was supported by H3K27ac modification from Paired-Tag (**Fig. 5H**, K27ac tracks). We further confirmed the chromatin interactions between this CRE and the promoter of *CHRNA2* (**Fig. 5I**). Taken together, the above data suggested that this human-specific CRE could be an enhancer that regulates the distinct expression of *CHRNA2* in the VIP-4 type from the human brain.

Although the cCREs from human and mouse brains shared considerable evolutionary conservation in DNA sequences, the orthologous elements also displayed a significant amount of divergence in chromatin accessibility in each cell type. We categorized CA-conserved and CA-divergent cCREs in matched cell sub-classes identified from human and mouse brains (**Fig. 4C**). First, we unsurprisingly found ~40% of the CA-conserved cCREs from every cell sub-class were at TSS proximal genomic regions, compared to ~10% of CA-divergent cCREs distributed in TSS proximal genomic regions (**Fig. 6A**). Second, we observed that a significantly larger amount of distal CA-conserved cCREs are marked by either H3K27ac or H3K27me3, than CA-divergent cCREs (Wilcoxon signed-rank test, p -value < 0.001, **Fig. 6B**, and **C**, **Table S23**, see **Methods**). Third, the CA-conserved cCREs were largely shared between diverse cell sub-classes, while the CA-divergent cCREs show less overlap between sub-classes and higher cell type-specificity except a few cell sub-classes with close relationship (OPC and OGC, IT neurons, **Fig. 6D**). Fourth, on average, 45% of distal CA-conserved cCREs have been predicted to regulate putative target genes (**Fig. 2F-H**), which is significantly higher than the proportion found in CA-divergent cCREs (**Fig. 6E**). However, the distribution of 1-D genomic distance from putative enhancers to promoters of their putative target genes showed no difference between these two categories of cCREs (**Fig. 6F**).

Sequence changes underlie epigenetic divergence in cCREs in distinct brain cell types

We hypothesize that the epigenetic divergence of cCREs is partly due to evolutionary changes in DNA sequences. To test this hypothesis, we picked IT-L2/3 neurons, LAMP5⁺ interneurons, and MGC as representative cell sub-classes, and trained gapped-kmer SVM classifiers (gkmSVM) (65) from the DNA sequences in cCREs (**Fig. 6G-H**, see **Methods**). These models achieved excellent performance (area under receiver operating characteristic curve (AUROC) ranging from 0.856 to 0.928, and area under precision-recall curve (AUPRC) ranging from 0.850 to 0.912) in the prediction of open chromatin regions within the corresponding species (**Fig. 6G-H**). Next, we predicted different categories of mouse cCREs using gkmSVM models trained with human DNA sequence at cCREs in corresponding cell sub-classes (**Fig. 6I**). Importantly, these models also achieved high accuracy (ranging from 0.83 to 0.91) in the prediction of CA-conserved mouse cCREs, and slightly lower accuracy in CA-divergent (ranging from 0.79 to 0.82) and mouse-species cCREs (ranging from 0.71 to 0.80) (**Fig. 6I**). Similarly, the gkmSVM models trained with mouse DNA sequences archived high accuracy in predicting human CA-conserved cCREs (ranging from 0.89 to 0.93), and slightly lower accuracy for human CA-divergent (ranging from 0.75 to 0.86) and human-specific cCREs (ranging from 0.77 to 0.84) (**Fig. 6J**). The human CA-divergent cCREs that failed to be predicted from mouse gkmSVM models have a potential function in regulating genes involved in specific biological processes, including glutamate receptor signaling pathway (GO:0007215), synaptic transmission - GABAergic (GO: 0051932), and various fatty acid elongation (GO: 0019367, GO:0019368, GO:0034625) in IT-L2/3 neurons, LAMP5⁺ interneurons, and MGC, respectively (**Fig. S14, Table S24**). These results suggested that the regulatory divergence is at least in part due to evolution of DNA sequences. However, we cannot exclude a role of non-sequence factors in this process.

Predicting disease relevant cell types for neuropsychiatric disorders

Genome-wide association studies (GWASs) have identified genetic variants that are associated with many mental diseases and traits (**Table S25**), but >90% of variants are located in non-protein-coding

regions of the genome (4, 5). Previous studies have shown that non-coding risk variants are enriched in cCREs active in disease relevant cell types (6–8). Leveraging the newly annotated cell-type-resolved human brain cCREs, we predict the cell types relevant to the different neuropsychiatric disorders and traits. We performed linkage disequilibrium score regression (LDSC) analysis to determine if the genetic heritability of DNA variants associated with neuropsychiatric traits are significantly enriched within cCREs showing chromatin accessibility in the major brain cell types in the present study (**Table S25**, see **Methods**). We found significant associations between 19 neuropsychiatric disorders and traits (**Table S25**, and **S26**) with the open chromatin landscapes in one or more cell types we identified (**Fig. 7A**, see **Methods**). In particular, we observed widespread and strong enrichment of genetic variants linked to psychiatric and cognitive traits such as schizophrenia (SCZ), intelligence, bipolar disorder, and educational attainment within accessible cCREs across various neuronal cell types (**Fig. 7A**, **Table S26**). Other neuropsychiatric traits, such as tobacco use disorder and alcohol usage were associated with specific neuronal cell types in basal ganglia, which were previously implicated in addiction(66). Another example is neuroticism, which was restricted and associated with IT neurons from the cortex (**Fig. 7A**, **Table S26**). In addition, as reported before(67, 68), the risk variants from Alzheimer’s disease were significantly enriched in the cCREs found in microglia, but not in other cell types (**Fig. 7A**, **Table S26**).

We further provided breakout reports of LDSC analysis by using three categories of cCREs defined above (**Fig. 4C**). We found most significant associations between cell sub-classes and GWAS traits were found in the analysis of epigenetic conserved elements (**Fig. 7B**, **Table S27**). Fewer significant associations were observed in the analysis of epigenetic divergent elements, and most of GWAS traits showed no significant associations when human-specific elements were used for LDSC (**Fig. 7B**). For example, the risk variants in schizophrenia showed the most significant enrichment in epigenetic conserved elements. We obtained 4,356 likely SCZ causal variants with a posterior probability of association (PPA) score greater than 1% based on Bayesian fine-mapping determined by the Psychiatric Genomics Consortium (<https://www.med.unc.edu/pgc/>). By overlapping with different categories of elements, we found one schizophrenia-associated locus harboring a CA-conserved element, and another

CA-divergent element containing two likely causal variants, rs324017 and rs324015, respectively (**Fig. 7C**). The CA-conserved cCRE, displayed accessibility in multiple neuronal cell types, was predicted to regulate the expression of the SCZ-associated gene *NAB2* (**Fig. 7C**) (69). These observations provide new hypotheses regarding the potential functions of noncoding SCZ risk variants.

Interestingly, LDSC analysis using human-specific elements revealed a significant association between Alzheimer's disease (AD) and microglia (**Fig. 7B, Table S27**). This raises the possibility that the AD-related risk variants could reside in human-specific regulatory elements, and contribute to human-specific gene regulation programs in microglia(70). This observation suggests potential limitations of animal models of AD in revealing disease pathology in humans (71). One AD risk locus contains multiple microglia-specific cCREs, which cannot find any homologous sequence in the mouse genome (**Fig. 7D**). One of these cCREs harboring AD-risk variants rs6733839 has been predicted to be a microglia-specific enhancer that can regulate the expression of *BIN1* gene, and its function was supported by both H3K27ac modification and previous validation experiment (**Fig. 7D**) (68).

Deep learning models predict the influence of risk variants on gene regulation

To further understand how risk variants contribute to the function of regulatory elements, we used deep learning (DL) models to predict chromatin accessibility from DNA sequences (**Fig. 7E, Figure S15A-D, see Methods**). The deep learning model architecture was inspired by Enformer (72), which adapts attention-based architecture, Transformer, that could better capture syntactic forms (e.g., order and combination of words in a sentence) and outperforms most existing models in natural language processing tasks (73). We trained the novel DL model called Epiformer on the normalized pseudo bulk ATAC-seq profiles in human microglia. The resulting model successfully predict the cell type-specific accessibility of cCREs at the *BIN1* locus with a Pearson correlation coefficient of 0.74 (**Fig. 7F**). By contrast, DL models trained using ATAC-seq profiles from other cell types the accessibilities of these cCREs failed to predict the chromatin accessibility profiles from microglia (**Fig. S15F**). To predict the

regulatory effects of the risk variant, we further performed *in silico* mutagenesis on the above microglia-specific enhancer near *BINI* and compared the changes of accessibility predicted from reference and altered DNA sequences. Every nucleotide within this 500 bp enhancer was mutated *in silico*, and the influence on accessibility was measured by assessing the difference between the predicted accessibility for the reference and altered sequences (**Fig. 7G, Table S28**, see **Methods**). Among these *in silico* single nucleotide mutations, most in the flanking regions did not affect the predicted accessibility, while a few nucleotide substitutions significantly increased or decreased the predicted accessibility (**Fig. 7G, Table S28**). The DNA sequence most negatively associated with the predicted accessibility was predicted to contain binding motifs for transcription factors SPI1, SPIB, and SPIC, which bind to the PU-box and play a key role in the differentiation of macrophages, myeloid and B-lymphoid cells(74, 75). This result indicates that only specific sequence motifs within the 500bp region are critical for determining the function of CREs. Lastly, the model also predicts that the nucleotide substitutions (C > T / C > G) of AD-risk variants (rs6733839) would increase the accessibility of human microglia specific enhancer (**Fig. 7H**, upper, in dark red, **Table S28**). This risk variant likely creates a new binding motif for MEF2A or other TFs in microglia (**Fig. 7H**, bottom, **Table S28**). Taken together, we hypothesize that this risk variant (rs6733839) will increase the binding and combined effect of TFs SPI1 and MEF2A, which may result in a higher expression level of AD-risk gene *BINI* in the microglia. These results support a model where the nucleotides within regulatory elements are not equally contributing to function, and capturing the regulatory code of sequence motifs is key to understanding how risk variants influence the potential function of regulatory elements.

Discussion

Understanding the cellular and molecular basis of brain circuits is one of the grand challenges of the twenty-first century. 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 disorders. Here, we

report a comprehensive profiling of chromatin accessibility at single-cell resolution in 42 human brain regions. The chromatin accessibility maps of 544,735 cCREs, were probed in >1.1 million nuclei. Taking advantage of our high-resolution brain dissections, we examined the regional specificity in chromatin accessibility of cell types in the human brain and showed that most brain cell types exhibit strong regional specificity. The described cCRE atlas (<http://catlas.org>) represents a rich resource for the neuroscience community to understand the molecular patterns that underlie the diversification of brain cell types in complementation to other molecular and anatomical data.

Genome-wide association studies (GWAS) have been used to enhance our understanding of polygenic human traits and reveal clinically relevant therapeutic targets for neuropsychiatric disorders. However, our ability to interpret the risk variants has been hampered by a lack of annotation of the non-coding genome. By leveraging both epigenetic conserved, divergent, and human-specific cCRE identified from various human cell types, we prioritize likely causal variants in linkage disequilibrium (LD), link distal cCREs to putative target genes, and predict motifs altered by risk variants using cutting-edge deep learning methods. We revealed hundreds of cell-type trait associations and created a framework to systematically interpret noncoding risk variants.

References and Notes

1. C. J. L. Murray, C. Atkinson, K. Bhalla, G. Birbeck, R. Burstein, D. Chou, R. Dellavalle, G. Danaei, M. Ezzati, A. Fahimi, D. Flaxman, Foreman, S. Gabriel, E. Gakidou, N. Kassebaum, S. Khatibzadeh, S. Lim, S. E. Lipshultz, S. London, Lopez, M. F. MacIntyre, A. H. Mokdad, A. Moran, A. E. Moran, D. Mozaffarian, T. Murphy, M. Naghavi, C. Pope, T. Roberts, J. Salomon, D. C. Schwebel, S. Shahrzaz, D. A. Sleet, Murray, J. Abraham, M. K. Ali, C. Atkinson, D. H. Bartels, K. Bhalla, G. Birbeck, R. Burstein, H. Chen, M. H. Criqui, Dahodwala, Jarlais, E. L. Ding, E. R. Dorsey, B. E. Ebel, M. Ezzati, Fahami, S. Flaxman, A. D. Flaxman, D. Gonzalez-Medina, B. Grant, H. Hagan, H. Hoffman, N. Kassebaum, S. Khatibzadeh, J. L. Leasher, J. Lin, S. E. Lipshultz, R. Lozano, Y. Lu, L. Mallinger, M. M. McDermott, R. Micha, T. R. Miller, A. A. Mokdad, A. H. Mokdad, D. Mozaffarian, M. Naghavi, K. M. V. Narayan, S. B. Omer, P. M. Pelizzari, D. Phillips, D. Ranganathan, F. P. Rivara, T. Roberts, U. Sampson, E. Sanman, A. Sapkota, D. C. Schwebel, S. Sharaz, R. Shivakoti, G. M. Singh, D. Singh, M. Tavakkoli, J. A. Towbin, J. D. Wilkinson, A. Zabetian, Murray, J. Abraham, M. K. Ali, M. Alvarado, C. Atkinson, L. M. Baddour, E. J. Benjamin, K. Bhalla, G. Birbeck, I. Bolliger, R. Burstein, E. Carnahan, D. Chou, S. S. Chugh, A. Cohen, K. E. Colson, L. T. Cooper, W. Couser, M. H. Criqui, K. C. Dabhadkar, R. P. Dellavalle, Jarlais, D. Dicker, E. R. Dorsey, H. Duber, B. E. Ebel, R. E. Engell, M. Ezzati, D. T. Felson, M. M. Finucane, S. Flaxman, A. D. Flaxman, T. Fleming, Foreman, M. H. Forouzanfar, G. Freedman, M. K. Freeman, E. Gakidou, R. F. Gillum, D. Gonzalez-Medina, R. Gosselin, H. R. Gutierrez, H. Hagan, R. Havmoeller, H. Hoffman, K. H. Jacobsen, S. L. James, R. Jasrasaria, S. Jayarman, N. Johns, N. Kassebaum, S. Khatibzadeh, Q. Lan, J. L. Leasher, S. Lim, S. E. Lipshultz, S. London, Lopez, R. Lozano, Y. Lu, L. Mallinger, M. Meltzer, G. A. Mensah, C. Michaud, T. R. Miller, C. Mock, T. E. Moffitt, A. A. Mokdad, A. H. Mokdad, A. Moran, M. Naghavi, K. M. V. Narayan, R. G. Nelson, C. Olives, S. B. Omer, K. Ortblad, B. Ostro, P. M. Pelizzari, D. Phillips, M. Raju, H. Razavi, B. Ritz, T. Roberts, R. L. Sacco, J. Salomon, U. Sampson, D. C. Schwebel, S. Shahrzaz, K. Shibuya, D. Silberberg, J. A. Singh, K. Steenland, J. A. Taylor, G. D. Thurston, M. S. Vavilala, T. Vos, G. R. Wagner, M. A. Weinstock, M. G. Weisskopf, S. Wulf, Murray, U. S. B. of D. Collaborators, The State of US Health, 1990-2010: Burden of Diseases, Injuries, and Risk Factors. *Jama*. 310, 591–606 (2013).
2. A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousseau, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, H. Parkinson, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 47, D1005–D1012 (2019).
3. M. Claussnitzer, J. H. Cho, R. Collins, N. J. Cox, E. T. Dermitzakis, M. E. Hurles, S. Kathiresan, E. Kenny, C. M. Lindgren, D. G. MacArthur, K. N. North, S. E. Plon, H. L. Rehm, N. Risch, C. N. Rotimi, J. Shendure, N. Soranzo, M. I. McCarthy, A brief history of human disease genetics. *Nature*. 577, 179–189 (2020).
4. A. Frydas, E. Wauters, J. van der Zee, C. V. Broeckhoven, Uncovering the impact of noncoding variants in neurodegenerative brain diseases. *Trends Genet*. 38, 258–272 (2021).
5. B. Liu, S. B. Montgomery, Identifying causal variants and genes using functional genomics in specialized cell types and contexts. *Hum Genet*. 139, 95–102 (2020).
6. I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Fritze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B.-K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shores, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. J. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D.

1 Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E.
2 Birney, I. Dunham, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. L. Dillon, L.
3 B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E.
4 Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C.
5 Giddings, T. R. Gingeras, E. D. Green, R. Guigó, R. C. Hardison, T. J. Hubbard, M. Kellis, W. J.
6 Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A.
7 Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P.
8 Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B.
9 Epstein, N. Shores, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X.
10 Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler,
11 M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A.
12 Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C.
13 Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T.
14 Baer, P. Batut, I. Bell, K. Bell, S. Chakraborty, X. Chen, J. Chrast, J. Curado, T. Derrien, J.
15 Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M.
16 J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P.
17 Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A.
18 Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L.-H. See, A. Shahab, J. Skancke,
19 A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y.
20 Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon,
21 M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, T. R. Gingeras, K. R. Rosenbloom, C. A.
22 Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G.
23 Heitner, D. Karolchik, W. J. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney,
24 T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B.-K. Lee, A. Battenhouse, N. C. Sheffield, J. M.
25 Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z.
26 Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M.
27 J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E.
28 Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan,
29 Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter,
30 E. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M.
31 Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K.
32 McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian,
33 N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. S. Newberry, S. E. Levy, D. M. Absher, A.
34 Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H.
35 Margulies, S. C. J. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry,
36 A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M.
37 Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix,
38 C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland,
39 Z. Lu, D. Manthavadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez,
40 G. Saunders, A. Sboner, S. Searle, C. Sis, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress,
41 M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler,
42 M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guigó, J. Harrow, T. J. Hubbard, S. G. Landt, S.
43 Fietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K.
44 Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C.
45 Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S.
46 Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J.
47 Leng, J. Lian, M. Lindahl-Alten, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen,
48 Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Shi, C.
49 Sis, T. Slifer, H. Witt, L. Wu, X. Xu, K.-K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M.
50 Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenenbaum, L. O. Penalva, F. Doyle, S.
51 Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil,
52 T. Slifer, A. Vectorsen, X. Yang, M. Snyder, K. P. White, T. Auer, L. Centanin, M. Eichenlaub, F.
53 Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt,
54 L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P.
55 J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R.

- 1 Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A.
- 2 K. Ebersol, T. Frum, K. Garg, E. Gist, R. S. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A.
- 3 K. Johnson, E. M. Johnson, T. V. Kuttyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J.
- 4 Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez,
- 5 R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J.
- 6 Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O.
- 7 Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A.
- 8 Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M.
- 9 Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N.
- 10 Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C.
- 11 Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B.
- 12 Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. D. Sahu,
- 13 P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J.
- 14 Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A.
- 15 Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K.-K. Yan, K. Y. Yip, E. Birney, An
- 16 Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*. 489, 57–74 (2012).
- 17 7. A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z.
- 18 Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G.
- 19 Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. Pfenning, X. Wang, M. C. Liu, C. Coarfa, R. A.
- 20 Harris, N. Shores, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong,
- 21 P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J.
- 22 Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni,
- 23 D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R.
- 24 C. Sallari, K. T. Siebenthal, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang,
- 25 X. Zhou, N. Abdennur, M. Adli, M. Akerman, L. Barrera, J. Antosiewicz-Bourget, T. Ballinger, M. J.
- 26 Barnes, D. Bates, R. J. A. Bell, D. A. Bennett, K. Bianco, C. Bock, P. Boyle, J. Brinchmann, P.
- 27 Caballero-Campo, R. Camahort, M. J. Carrasco-Alfonso, T. Charnecki, H. Chen, Z. Chen, J. B.
- 28 Cheng, S. Cho, A. Chu, W.-Y. Chung, C. Cowan, Q. A. Deng, V. Deshpande, M. Diegel, B. Ding, T.
- 29 Durham, L. Echipare, L. Edsall, D. Flowers, O. Genbacev-Krtolica, C. Gifford, S. Gillespie, E. Giste,
- 30 I. A. Glass, A. Gnirke, M. Gormley, H. Gu, J. Gu, D. A. Hafler, M. J. Hangauer, M. Hariharan, M.
- 31 Hatan, E. Haugen, Y. He, S. Heimfeld, S. Herlofsen, Z. Hou, R. Humbert, R. Issner, A. R. Jackson, H.
- 32 Jia, P. Jiang, A. K. Johnson, T. Kadlec, B. Kamoh, M. Kapidzic, J. Kent, A. Kim, M.
- 33 Kleinewietfeld, S. Klugman, J. Krishnan, S. Kuan, T. Kuttyavin, A.-Y. Lee, K. Lee, J. Li, N. Li, Y. Li,
- 34 K. L. Ligon, S. Lin, Y. Lin, J. Liu, Y. Liu, C. J. Luckey, Y. P. Ma, C. Maire, A. Marson, J. S. Mattick,
- 35 M. Mayo, M. McMaster, H. Metsky, T. Mikkelsen, D. Miller, M. Miri, E. Mukame, R. P. Nagarajan,
- 36 F. Neri, J. Nery, T. Nguyen, H. O'Geen, S. Paithankar, T. Papayannopoulou, M. Pelizzola, P. Plettner,
- 37 N. E. Propson, S. Raghuraman, B. J. Raney, A. Raubitschek, A. P. Reynolds, H. Richards, K. Riehle,
- 38 P. Rinaudo, J. F. Robinson, N. B. Rockweiler, E. Rosen, E. Rynes, J. Schein, R. Sears, T. Sejnowski,
- 39 A. Shafer, L. Shen, R. Shoemaker, M. Sigaroudinia, I. Slukvin, S. Stehling-Sun, R. Stewart, S. L.
- 40 Subramanian, K. Sukuntha, S. Swanson, S. Tian, H. Tilden, L. Tsai, M. Urich, I. Vaughn, J.
- 41 Vierstra, S. Vong, U. Wagner, H. Wang, T. Wang, Y. Wang, A. Weiss, H. Whitton, A. Wildberg, H.
- 42 Witt, K.-J. Won, M. Xie, X. Xing, I. Xu, Z. Xuan, Z. Ye, C. Yen, P. Yu, X. Zhang, X. Zhang, J. Zhao,
- 43 Y. Zhou, J. Zhu, Y. Zhu, S. Ziegler, A. E. Beaudet, L. A. Boyer, P. L. D. Jager, P. J. Farnham, S. J.
- 44 Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson,
- 45 T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J.
- 46 F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos,
- 47 T. Wang, M. Kellis, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi,
- 48 P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward,
- 49 A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M.
- 50 Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shores, C. B. Epstein, E. Gjoneska, D. Leung, W.
- 51 Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T.
- 52 K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S.
- 53 Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V.
- 54 Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthal, N. A. Sinnott-Armstrong,

- 1 M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. D. Jager, P.
2 J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S.
3 Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H.
4 Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B.
5 Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human
6 epigenomes. *Nature*. 518, 317–330 (2015).
- 7 8. M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R.
8 Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutayvin, S. Stehling-Sun, A. K. Johnson,
9 T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A.
10 Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, J. A.
11 Stamatoyannopoulos, Systematic Localization of Common Disease-Associated Variation in
12 Regulatory DNA. *Science*. 337, 1190–1195 (2012).
- 13 9. R. D. Hodge, T. E. Bakken, J. A. Miller, K. A. Smith, E. R. Barkan, L. T. Graybuck, J. L. Close, B.
14 Long, N. Johansen, O. Penn, Z. Yao, J. Eggermont, T. Höllt, B. P. Levi, S. I. Shehata, B. Aevertmann,
15 A. Beller, D. Bertagnolli, K. Brouner, T. Casper, C. Cobbs, R. Dalley, N. Dee, S.-L. Ding, R. G.
16 Ellenbogen, O. Fong, E. Garren, J. Goldy, R. P. Gwinn, D. Hirschstein, C. D. Keene, M. Keshk, A. L.
17 Ko, K. Lathia, A. Mahfouz, Z. Maltzer, M. McGraw, T. N. Nguyen, J. Nyhus, J. G. Ojemann, A.
18 Oldre, S. Parry, S. Reynolds, C. Rimorin, N. V. Shapovalova, S. Somasundaram, A. Szafer, E. R.
19 Thomsen, M. Tieu, G. Quon, R. H. Scheuermann, R. Yuste, S. M. Sunkin, B. Lelieveldt, D. Feng, L.
20 Ng, A. Bernard, M. Hawrylycz, J. W. Phillips, B. Tasic, H. Zeng, A. R. Jones, C. Koch, E. S. Lein,
21 Conserved cell types with divergent features in human versus mouse cortex. *Nature*. 573, 61–68
22 (2019).
- 23 10. C. Luo, H. Liu, F. Xie, E. J. Armand, K. Siletti, T. E. Bakken, R. Fang, W. I. Doyle, R. D. Hodge,
24 L. Hu, B.-A. Wang, Z. Zhang, S. Preissl, D.-S. Lee, J. Zhou, S.-Y. Niu, R. Castanon, A. Bartlett, A.
25 Rivkin, X. Wang, J. Lucero, J. R. Nery, D. A. Davis, D. C. Mash, J. R. Dixon, S. Linnarsson, E. Lein,
26 M. M. Behrens, B. Ren, E. A. Mukamel, J. R. Ecker, *Biorxiv*, in press,
27 doi:10.1101/2019.12.11.873398.
- 28 11. B. Tasic, Z. Yao, L. T. Graybuck, K. A. Smith, T. N. Nguyen, D. Bertagnolli, J. Goldy, E. Garren,
29 M. N. Economo, S. Viswanathan, O. Penn, T. Bakken, V. Menon, J. Miller, O. Fong, K. E. Hirokawa,
30 K. Lathia, C. Rimorin, M. Tieu, R. Larsen, T. Casper, E. Barkan, M. Kroll, S. Parry, N. V.
31 Shapovalova, D. Hirschstein, J. Pendergraft, H. A. Sullivan, T. K. Kim, A. Szafer, N. Dee, P.
32 Groblewski, I. Wickersham, A. Cetin, J. A. Harris, B. P. Levi, S. M. Sunkin, L. Madisen, T. L.
33 Daigle, L. Looger, A. Bernard, J. Phillips, E. Lein, M. Hawrylycz, K. Svoboda, A. R. Jones, C. Koch,
34 H. Zeng, Shared and distinct transcriptomic cell types across neocortical areas. *Nature*. 563, 72–78
35 (2018).
- 36 12. J. R. Moffitt, D. Bambach-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D.
37 Rubinstein, J. Hao, A. Regev, C. Dulac, X. Zhuang, Molecular, spatial, and functional single-cell
38 profiling of the hypothalamic preoptic region. *Science*. 362 (2018), doi:10.1126/science.aau5324.
- 39 13. J. R. Ecker, D. H. Geschwind, A. R. Kriegstein, J. Ngai, P. Osten, D. Polioudakis, A. Regev, N.
40 Sestan, I. R. Wickersham, H. Zeng, The BRAIN Initiative Cell Census Consortium: Lessons Learned
41 toward Generating a Comprehensive Brain Cell Atlas. *Neuron*. 96, 542–557 (2017).
- 42 14. D.-S. Lee, C. Luo, J. Zhou, S. Chandran, A. Rivkin, A. Bartlett, J. R. Nery, C. Fitzpatrick, C.
43 O'Connor, J. R. Dixon, J. R. Ecker, Simultaneous profiling of 3D genome structure and DNA
44 methylation in single human cells. *Nat Methods*. 16, 999–1006 (2019).
- 45 15. C. Luo, C. L. Keown, L. Kurihara, J. Zhou, Y. He, J. Li, R. Castanon, J. Lucero, J. R. Nery, J. P.
46 Sandoval, B. Bui, T. J. Sejnowski, T. T. Harkins, E. A. Mukamel, M. M. Behrens, J. R. Ecker, Single-

- cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*. 357, 600–604 (2017).
16. C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulana, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, L. Cai, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*. 568, 235–239 (2019).
17. Z. Yao, H. Liu, F. Xie, S. Fischer, R. S. Adkins, A. I. Aldridge, S. A. Ament, A. Bartlett, M. M. Behrens, K. V. den Berge, D. Bertagnolli, H. R. de Bézieux, T. Biancalani, A. S. Boeshaghi, H. C. Bravo, T. Casper, C. Colantuoni, J. Crabtree, H. Creasy, K. Crichton, M. Crow, N. Dee, E. L. Dougherty, W. I. Doyle, S. Dudoit, R. Fang, V. Felix, O. Fong, M. Giglio, J. Goldy, M. Hawrylycz, B. R. Herb, R. Hertzano, X. Hou, Q. Hu, J. Kancharla, M. Kroll, K. Lathia, Y. E. Li, J. D. Lucero, C. Luo, A. Mahurkar, D. McMillen, N. M. Nadaf, J. R. Nery, T. N. Nguyen, S.-Y. Niu, V. Ntranos, J. Orvis, J. K. Osteen, T. Pham, A. Pinto-Duarte, O. Poirion, S. Preissl, E. Purdom, C. Rimorin, D. Risso, A. C. Rivkin, K. Smith, K. Street, J. Sulc, V. Svensson, M. Tieu, A. Torkelson, H. Tung, E. D. Vaishnav, C. R. Vanderburg, C. van Velthoven, X. Wang, O. R. White, Z. J. Huang, P. V. Kharchenko, L. Pachter, J. Ngai, A. Regev, B. Tasic, J. D. Welch, J. Gillis, E. Z. Macosko, B. Ren, J. R. Ecker, H. Zeng, E. A. Mukamel, A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*. 598, 103–110 (2021).
18. A. Kepecs, G. Fishell, Interneuron cell types are fit to function. *Nature*. 505, 318–326 (2014).
19. S. Hrvatin, D. R. Hochbaum, M. A. Nagy, M. Cicconet, K. Robertson, L. Cheadle, R. Zilionis, A. Ratner, R. Borges-Monroy, A. M. Klein, B. L. Sabatini, M. E. Greenberg, Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat Neurosci*. 21, 120–129 (2018).
20. E.-L. Yap, M. E. Greenberg, Activity-Regulated Transcription: Bridging the Gap between Neural Activity and Behavior. *Neuron*. 100, 330–348 (2018).
21. A. Zeisel, H. Hochgerner, P. Lönnerberg, A. Johnsson, F. Memic, J. van der Zwan, M. Häring, E. Braun, L. E. Borm, G. L. Manno, S. Codeluppi, A. Furlan, K. Lee, N. Skene, K. D. Harris, J. Hjerling-Leffler, E. Arenas, P. Ernfors, U. Marklund, S. Linnarsson, Molecular Architecture of the Mouse Nervous System. *Cell*. 174, 999–1014.e22 (2018).
22. A. Saunders, E. Z. Macosko, A. Wysoker, M. Goldman, F. M. Krienen, H. de Rivera, E. Bien, M. Baum, L. Bortolin, S. Wang, A. Goeva, J. Nemesh, N. Kamitaki, S. Brumbaugh, D. Kulp, S. A. McCarroll, Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell*. 174, 1015–1030.e16 (2018).
23. B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, D. Bertagnolli, J. Goldy, N. Shapovalova, S. Parry, C. Lee, K. Smith, A. Bernard, L. Madisen, S. M. Sunkin, M. Hawrylycz, C. Koch, H. Zeng, Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci*. 19, 335–346 (2016).
24. R. Fang, C. Xia, J. L. Close, M. Zhang, J. He, Z. Huang, A. R. Halpern, B. Long, J. A. Miller, E. S. Lein, X. Zhuang, Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH. *Science*. 377, 56–62 (2022).
25. Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, B. Ren, A map of the cis-regulatory sequences in the mouse genome. *Nature*. 488, 116–120 (2012).

- 1 26. F. Abascal, R. Acosta, N. J. Addleman, J. Adrian, V. Afzal, R. Ai, B. Aken, J. A. Akiyama, O. A.
- 2 Jammal, H. Amrhein, S. M. Anderson, G. R. Andrews, I. Antoshechkin, K. G. Ardlie, J. Armstrong,
- 3 M. Astley, B. Banerjee, A. A. Barkal, I. H. A. Barnes, I. Barozzi, D. Barrell, G. Barson, D. Bates, U.
- 4 K. Baymuradov, C. Bazile, M. A. Beer, S. Beik, M. A. Bender, R. Bennett, L. P. B. Bouvrette, B. E.
- 5 Bernstein, A. Berry, A. Bhaskar, A. Bignell, S. M. Blue, D. M. Bodine, C. Boix, N. Boley, T.
- 6 Borrmann, B. Borsari, A. P. Boyle, L. A. Brandsmeier, A. Breschi, E. H. Bresnick, J. A. Brooks, M.
- 7 Buckley, C. B. Burge, R. Byron, E. Cahill, L. Cai, L. Cao, M. Carty, R. G. Castanon, A. Castillo, H.
- 8 Chaib, E. T. Chan, D. R. Chee, S. Chee, H. Chen, H. Chen, J.-Y. Chen, S. Chen, J. M. Cherry, S. B.
- 9 Chhetri, J. S. Choudhary, J. Chrast, D. Chung, D. Clarke, N. A. L. Cody, C. J. Coppola, J. Coursen,
- 10 A. M. D'Ippolito, S. Dalton, C. Danyko, C. Davidson, J. Davila-Velderrain, C. A. Davis, J. Dekker,
- 11 A. Deran, G. DeSalvo, G. Despacio-Reyes, C. N. Dewey, D. E. Dickel, M. Diegel, M. Diekhans, V.
- 12 Dileep, B. Ding, S. Djebali, A. Dobin, D. Dominguez, S. Donaldson, J. Drenkow, T. R. Dreszer, Y.
- 13 Drier, M. O. Duff, D. Dunn, C. Eastman, J. R. Ecker, M. D. Edwards, N. El-Ali, S. I. Elhajjajy, K.
- 14 Elkins, A. Emili, C. B. Epstein, R. C. Evans, I. Ezkurdia, K. Fan, P. J. Farnham, N. P. Farrell, E. A.
- 15 Feingold, A.-M. Ferreira, K. Fisher-Aylor, S. Fitzgerald, P. Flicek, C. S. Foo, K. Fortier, A. Frankish,
- 16 P. Freese, S. Fu, X.-D. Fu, Y. Fu, Y. Fukuda-Yuzawa, M. Fulciniti, A. P. W. Funnell, I. Gabdank, T.
- 17 Galeev, M. Gao, C. G. Giron, T. H. Garvin, C. A. Gelboin-Burkhart, G. Georgolopoulos, M. B.
- 18 Gerstein, B. M. Giardine, D. K. Gifford, D. M. Gilbert, D. A. Gilchrist, S. Gillespie, T. R. Gingeras,
- 19 P. Gong, A. Gonzalez, J. M. Gonzalez, P. Good, A. Goren, D. U. Gorkin, B. R. Graveley, M. Gray, J.
- 20 F. Greenblatt, E. Griffiths, M. T. Groudine, F. Grubert, M. Gu, R. Guigó, H. Guo, Y. Guo, Y. Guo, G.
- 21 Gursoy, M. Gutierrez-Arcelus, J. Halow, R. C. Hardison, M. Hardy, M. Hariharan, A. Harmanci, A.
- 22 Harrington, J. L. Harrow, T. B. Hashimoto, R. D. Hasz, M. Hatan, E. Haugen, J. E. Hayes, P. He, Y.
- 23 He, N. Heidari, D. Hendrickson, E. F. Heuston, J. A. Hilton, B. C. Hitz, A. Hochman, C. Holgren, L.
- 24 Hou, S. Hou, Y.-H. E. Hsiao, S. Hsu, H. Huang, T. J. Hubbard, J. Huey, T. R. Hughes, T. Hunt, S.
- 25 Ibarrientos, R. Issner, M. Iwata, O. Izuogu, T. Jaakkola, N. Jameel, C. Jansen, L. Jiang, P. Jiang, A.
- 26 Johnson, R. Johnson, I. Jungreis, M. Kadaba, M. Kasowski, M. Kasparian, M. Kato, R. Kaul, T.
- 27 Kawli, M. Kay, J. C. Keen, S. Keles, C. A. Keller, D. Kelley, M. Kellis, P. Kheradpour, D. S. Kim, A.
- 28 Kirilusha, R. J. Klein, B. Knoechel, S. Kuan, M. J. Kulik, S. Kumar, A. Kundaje, T. Kutayavin, J.
- 29 Lagarde, B. R. Lajoie, N. J. Lambert, J. Lazar, A. Y. Lee, D. Lee, E. Lee, J. W. Lee, K. Lee, C. S.
- 30 Leslie, S. Levy, B. Li, H. Li, N. Li, S. Li, X. Li, Y. I. Li, Y. Li, Y. Li, Y. Li, J. Lian, M. W. Libbrecht,
- 31 S. Lin, Y. Lin, D. Liu, J. Liu, P. Liu, T. Liu, X. S. Liu, Y. Liu, Y. Liu, M. Long, S. Lou, J. Loveland,
- 32 A. Lu, Y. Lu, E. Lécuyer, L. Ma, M. Mackiewicz, B. J. Mannion, M. Mannstadt, D. Manthavadi, G.
- 33 K. Marinov, F. J. Martin, E. Mattei, K. McCue, M. McEown, G. McVicker, S. K. Meadows, A.
- 34 Meissner, E. M. Mendenhall, C. L. Messer, W. Meuleman, C. Meyer, S. Miller, M. G. Milton, T.
- 35 Mishra, D. E. Moore, H. M. Moore, J. E. Moore, S. H. Moore, J. Moran, A. Mortazavi, J. M. Mudge,
- 36 N. Munshi, R. Murad, R. M. Myers, V. Nandakumar, P. Nandi, A. M. Narasimha, A. K. Narayanan,
- 37 H. Naughton, F. C. P. Navarro, P. Navas, J. Nazarovs, J. Nelson, S. Neph, F. J. Neri, J. R. Nery, A. R.
- 38 Nesmith, J. S. Newberry, K. M. Newberry, V. Ngo, R. Nguyen, T. B. Nguyen, T. Nguyen, A.
- 39 Nishida, W. S. Noble, C. S. Novak, E. M. Novoa, B. Nuñez, C. W. O'Donnell, S. Olson, K. C. Onate,
- 40 E. Otterman, H. Ozadam, M. Pagan, T. Palden, X. Pan, Y. Park, E. C. Partridge, B. Paten, F. Pauli-
- 41 Behn, M. J. Pazin, B. Pei, L. A. Pennacchio, A. R. Perez, E. H. Perry, D. D. Pervouchine, N. N.
- 42 Phalke, Q. Pham, D. H. Phanstiel, I. Plajzer-Frick, G. A. Pratt, H. E. Pratt, S. Preissl, J. K. Pritchard,
- 43 Y. Pritykin, M. J. Purcaro, Q. Qin, G. Quinones-Valdez, I. Rabano, E. Radovani, A. Raj, N.
- 44 Rajagopal, O. Ram, L. Ramirez, R. N. Ramirez, D. Rausch, S. Raychaudhuri, J. Raymond, R. Razavi,
- 45 T. E. Reddy, T. M. Reimann, B. Ren, A. Reymond, A. Reynolds, S. K. Rhie, J. Rinn, M. Rivera, J. C.
- 46 Rivera-Mulia, B. S. Roberts, J. M. Rodriguez, J. Rozowsky, R. Ryan, E. Rynes, D. N. Salins, R.
- 47 Sandstrom, T. Sasaki, S. Sathe, D. Savic, A. Scavelli, J. Scheiman, C. Schlaffner, J. A. Schloss, F. W.
- 48 Schmitges, L. H. See, A. Sethi, M. Setty, A. Shafer, S. Shan, E. Sharon, Q. Shen, Y. Shen, R. I.
- 49 Sherwood, M. Shi, S. Shin, N. Shores, K. Siebenthal, C. Sis, T. Slifer, C. A. Sloan, A. Smith, V.
- 50 Snetkova, M. P. Snyder, D. V. Spacek, S. Srinivasan, R. Srivas, G. Stamatoyannopoulos, J. A.
- 51 Stamatoyannopoulos, R. Stanton, D. Steffan, S. Stehling-Sun, J. S. Strattan, A. Su, B. Sundararaman,
- 52 M.-M. Suner, T. Syed, M. Szynkarek, F. Y. Tanaka, D. Tenen, M. Teng, J. A. Thomas, D. Toffey, M.
- 53 L. Tress, D. E. Trout, G. Trynka, J. Tsuji, S. A. Upchurch, O. Ursu, B. Uszczynska-Ratajczak, M. C.
- 54 Uziel, A. Valencia, B. V. Biber, A. G. van der Velde, E. L. V. Nostrand, Y. Vaydylevich, J. Vazquez,
- 55 A. Victorson, J. Vielmetter, J. Vierstra, A. Visel, A. Vlasova, C. M. Vockley, S. Volpi, S. Vong, H.

- 1 Wang, M. Wang, Q. Wang, R. Wang, T. Wang, W. Wang, X. Wang, Y. Wang, N. K. Watson, X.
2 Wei, Z. Wei, H. Weisser, S. M. Weissman, R. Welch, R. E. Welikson, Z. Weng, H.-J. Westra, J. W.
3 Whitaker, C. White, K. P. White, A. Wildberg, B. A. Williams, D. Wine, H. N. Witt, B. Wold, M.
4 Wolf, J. Wright, R. Xiao, X. Xiao, J. Xu, J. Xu, K.-K. Yan, Y. Yan, H. Yang, X. Yang, Y.-W. Yang,
5 G. G. Yardımcı, B. A. Yee, G. W. Yeo, T. Young, T. Yu, F. Yue, C. Zaleski, C. Zang, H. Zeng, W.
6 Zeng, D. R. Zerbino, J. Zhai, L. Zhan, Y. Zhan, B. Zhang, J. Zhang, J. Zhang, K. Zhang, L. Zhang, P.
7 Zhang, Q. Zhang, X.-O. Zhang, Y. Zhang, Z. Zhang, Y. Zhao, Y. Zheng, G. Zhong, X.-Q. Zhou, Y.
8 Zhu, J. Zimmerman, J. E. Moore, M. J. Purcaro, H. E. Pratt, C. B. Epstein, N. Shores, J. Adrian, T.
9 Kawli, C. A. Davis, A. Dobin, R. Kaul, J. Halow, E. L. V. Nostrand, P. Freese, D. U. Gorkin, Y.
10 Shen, Y. He, M. Mackiewicz, F. Pauli-Behn, B. A. Williams, A. Mortazavi, C. A. Keller, X.-O.
11 Zhang, S. I. Elhajjajy, J. Huey, D. E. Dickel, V. Snetkova, X. Wei, X. Wang, J. C. Rivera-Mulia, J.
12 Rozowsky, J. Zhang, S. B. Chhetri, J. Zhang, A. Victorson, K. P. White, A. Visel, G. W. Yeo, C. B.
13 Burge, E. Lécuyer, D. M. Gilbert, J. Dekker, J. Rinn, E. M. Mendenhall, J. R. Ecker, M. Kellis, R. J.
14 Klein, W. S. Noble, A. Kundaje, R. Guigó, P. J. Farnham, J. M. Cherry, R. M. Myers, B. Ren, B. R.
15 Graveley, M. B. Gerstein, L. A. Pennacchio, M. P. Snyder, B. E. Bernstein, B. Wold, R. C. Hardison,
16 T. R. Gingeras, J. A. Stamatoyannopoulos, Z. Weng, Expanded encyclopaedias of DNA elements in
17 the human and mouse genomes. *Nature*. 583, 699–710 (2020).
- 18 27. D. A. Cusanovich, A. J. Hill, D. Aghamirzaie, R. M. Daza, H. A. Pliner, J. B. Berletch, G. N.
19 Filippova, X. Huang, L. Christiansen, W. S. DeWitt, C. Lee, S. G. Regalado, D. F. Read, F. J.
20 Steemers, C. M. Distech, C. Trapnell, J. Shendure, A Single-Cell Atlas of In Vivo Mammalian
21 Chromatin Accessibility. *Cell*. 174, 1309-1324.e18 (2018).
- 22 28. J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y.
23 Chang, W. J. Greenleaf, Single-cell chromatin accessibility reveals principles of regulatory variation.
24 *Nature*. 523, 486–490 (2015).
- 25 29. C. A. Lareau, F. M. Duarte, J. G. Chew, V. K. Kartha, Z. D. Burkett, A. S. Kohlway, D.
26 Pokholok, M. J. Aryee, F. J. Steemers, R. Lebofsky, J. D. Buenrostro, Droplet-based combinatorial
27 indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol*. 37, 916–924 (2019).
- 28 30. A. Mezger, S. Klemm, I. Mann, K. Brower, A. Mir, M. Bostick, A. Farmer, P. Fordyce, S.
29 Linnarsson, W. Greenleaf, High-throughput chromatin accessibility profiling at single-cell resolution.
30 *Nat Commun*. 9, 3647 (2018).
- 31 31. X. Chen, R. J. Miragaia, K. N. Natarajan, S. A. Teichmann, A rapid and robust method for single
32 cell chromatin accessibility profiling. *Nat Commun*. 9, 5345 (2018).
- 33 32. B. B. Lake, S. Chen, B. C. Sos, J. Fan, G. E. Kaeser, Y. C. Yung, T. E. Duong, D. Gao, J. Chun,
34 P. V. Kharchenko, K. Zhang, Integrative single-cell analysis of transcriptional and epigenetic states in
35 the human adult brain. *Nat Biotechnol*. 36, 70–80 (2018).
- 36 33. J. R. Sinnamoni, K. A. Torkenczy, M. W. Linhoff, S. A. Vitak, R. M. Mulqueen, H. A. Pliner, C.
37 Trapnell, F. J. Steemers, G. Mandel, A. C. Adey, The accessible chromatin landscape of the murine
38 hippocampus at single-cell resolution. *Genome Res*. 29, 857–869 (2019).
- 39 34. S. Preissl, R. Fang, H. Huang, Y. Zhao, R. Raviram, D. U. Gorkin, Y. Zhang, B. C. Sos, V. Afzal,
40 D. E. Dickel, S. Kuan, A. Visel, L. A. Pennacchio, K. Zhang, B. Ren, Single-nucleus analysis of
41 accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional
42 regulation. *Nat Neurosci*. 21, 432–439 (2018).
- 43 35. L. T. Graybuck, T. L. Daigle, A. E. Sedeño-Cortés, M. Walker, B. Kalmbach, G. H. Lenz, E.
44 Morin, T. N. Nguyen, E. Garren, J. L. Bendrick, T. K. Kim, T. Zhou, M. Mortrud, S. Yao, L. A.
45 Siverts, R. Larsen, B. B. Gore, E. R. Szelenyi, C. Trader, P. Balaram, C. T. J. van Velthoven, M.

- 1 Chiang, J. K. Mich, N. Dee, J. Goldy, A. H. Cetin, K. Smith, S. W. Way, L. Esposito, Z. Yao, V.
- 2 Gradinaru, S. M. Sunkin, E. Lein, B. P. Levi, J. T. Ting, H. Zeng, B. Tasic, Enhancer viruses for
- 3 combinatorial cell-subclass-specific labeling. *Neuron*. 109, 1449-1464.e13 (2021).
- 4 36. M. R. Corces, A. Shcherbina, S. Kundu, M. J. Gloudemans, L. Frésard, J. M. Granja, B. H. Louie,
- 5 T. Eulalio, S. Shams, S. T. Bagdatli, M. R. Mumbach, B. Liu, K. S. Montine, W. J. Greenleaf, A.
- 6 Kundaje, S. B. Montgomery, H. Y. Chang, T. J. Montine, Single-cell epigenomic analyses implicate
- 7 candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat Genet*.
- 8 52, 1158–1168 (2020).
- 9 37. S. Domcke, A. J. Hill, R. M. Daza, J. Cao, D. R. O'Day, H. A. Pliner, K. A. Aldinger, D.
- 10 Pokholok, F. Zhang, J. H. Milbank, M. A. Zager, I. A. Glass, F. J. Steemers, D. Doherty, C. Trapnell,
- 11 D. A. Cusanovich, J. Shendure, A human cell atlas of fetal chromatin accessibility. *Science*. 370
- 12 (2020), doi:10.1126/science.aba7612.
- 13 38. B. I. C. C. N. (BICCN), B. I. C. C. N. (BICCN) C. authors, E. M. Callaway, H.-W. Dong, J. R.
- 14 Ecker, M. J. Hawrylycz, Z. J. Huang, E. S. Lein, J. Ngai, P. Osten, B. Ren, A. S. Tolias, O. White, H.
- 15 Zeng, X. Zhuang, B. contributing principal investigators, G. A. Ascoli, M. M. Behrens, J. Chun, G.
- 16 Feng, J. C. Gee, S. S. Ghosh, Y. O. Halchenko, R. Hertzano, B. K. Lim, M. E. Martone, L. Ng, L.
- 17 Pachter, A. J. Ropelewski, T. L. Tickle, X. W. Yang, K. Zhang, P. manuscript editors, M. writing and
- 18 figure generation, T. E. Bakken, P. Berens, T. L. Daigle, J. A. Harris, N. L. Jorstad, B. E. Kalmbach,
- 19 D. Kobak, Y. E. Li, H. Liu, K. S. Matho, E. A. Mukamel, M. Naeemi, F. Scala, P. Tan, J. T. Ting, F.
- 20 Xie, M. Zhang, Z. Zhang, J. Zhou, B. Zingg, A. coordination, I. data analysis, E. Armand, Z. Yao, D.
- 21 Bertagnolli, T. Casper, K. Crichton, N. Dee, D. Diep, S.-L. Ding, W. Dong, E. L. Dougherty, O.
- 22 Fong, M. Goldman, J. Goldy, R. D. Hodge, L. Hu, C. D. Keene, F. M. Krienen, M. Kroll, B. B. Lake,
- 23 K. Lathia, S. Linnarsson, C. S. Liu, E. Z. Macosko, S. A. McCarroll, D. McMillen, N. M. Nadaf, T.
- 24 N. Nguyen, C. R. Palmer, T. Pham, N. Plongthongkum, N. M. Reed, A. Regev, C. Rimorin, W. J.
- 25 Romanow, S. Savoia, K. Siletti, K. Smith, J. Sulc, B. Tasic, M. Tieu, A. Torkelson, H. Tung, C. T. J.
- 26 van Velthoven, C. R. Vanderburg, A. M. Yanny, A. data generation and processing, R. Fang, X. Hou,
- 27 J. D. Lucero, J. K. Osteen, A. Pinto-Duarte, O. Poirion, S. Preissl, X. Wang, M. data production and
- 28 analysis, A. I. Aldridge, A. Bartlett, L. Boggeman, C. O'Connor, R. G. Castanon, H. Chen, C.
- 29 Fitzpatrick, C. Luo, J. R. Nery, M. Nunn, A. C. Rivkin, W. Tian, E. data generation and processing,
- 30 B. Dominguez, T. Ito-Cole, M. Jacobs, X. Jin, C.-T. Lee, K.-F. Lee, P. A. Miyazaki, Y. Pang, M.
- 31 Rashid, J. B. Smith, M. Vu, E. Williams, T. Biancalani, A. S. Boeshaghi, M. Crow, S. Dudoit, S.
- 32 Fischer, J. Gillis, Q. Hu, P. V. Kharchenko, S.-Y. Niu, V. Ntranos, E. Purdom, D. Risso, H. R. de
- 33 Bézieux, S. Somasundaram, K. Street, V. Svensson, E. D. Vaishnav, K. V. den Berge, J. D. Welch, T.
- 34 and connectivity data generation, X. An, H. S. Bateup, I. Bowman, R. K. Chance, N. N. Foster, W.
- 35 Galbavy, H. Gong, L. Gou, J. T. Hatfield, H. Hintiryan, K. E. Hirokawa, G. Kim, D. J. Kramer, A. Li,
- 36 X. Li, Q. Luo, R. Muñoz-Castañeda, D. A. Stafford, M. data generation and reconstruction, Z. Feng,
- 37 X. Jia, S. Jiang, T. Jiang, X. Kuang, R. Larsen, P. Lesnar, Y. Li, Y. Li, L. Liu, H. Peng, L. Qu, M.
- 38 Ren, Z. Ruan, E. Shen, Y. Song, W. Wakeman, P. Wang, Y. Wang, Y. Wang, L. Yin, J. Yuan, S.
- 39 Zhao, X. Zhao, O. and other data generation, A. Narasimhan, R. Palaniswamy, M. analysis
- 40 connectivity and imaging, S. Banerjee, L. Ding, D. Huilgol, B. Huo, H.-C. Kuo, S. Laturus, X. Li, P.
- 41 P. Mitra, J. Mizrahi, Q. Wang, P. Xie, F. Xiong, Y. Yu, S. resolved single-cell transcriptomics
- 42 (MERFISH), S. W. Eichhorn, M. profiling (Patch-seq), J. Berg, M. Bernabucci, Y. Bernaerts, C. R.
- 43 Cadwell, J. R. Castro, R. Dalley, L. Hartmanis, G. D. Horwitz, X. Jiang, A. L. Ko, E. Miranda, S.
- 44 Mulherkar, P. R. Nicovich, S. F. Owen, R. Sandberg, S. A. Sorensen, Z. H. Tan, T. tools, S. Allen, D.
- 45 Hockemeyer, A. Y. Lee, M. B. Veldman, N. archive and analytics, R. S. Adkins, S. A. Ament, H. C.
- 46 Bravo, R. Carter, A. Chatterjee, C. Colantuoni, J. Crabtree, H. Creasy, V. Felix, M. Giglio, B. R.
- 47 Herb, J. Kanherla, A. Mahurkar, C. McCracken, L. Nickel, D. Olley, J. Orvis, M. Schor, B. I. L.
- 48 (BIL) archive, G. Hood, D. archive, B. Dichter, M. Grauer, B. Helba, B. C. D. C. (BCDC), A.
- 49 Bandrowski, N. Barkas, B. Carlin, F. D. D'Orazi, K. Degatano, T. H. Gillespie, F. Khajouei, K.
- 50 Konwar, C. Thompson, P. management, K. Kelly, S. Mok, S. Sunkin, A multimodal cell census and
- 51 atlas of the mammalian primary motor cortex. *Nature*. 598, 86–102 (2021).

- 1 39. S.-L. Ding, J. J. Royall, S. M. Sunkin, L. Ng, B. A. C. Facer, P. Lesnar, A. Guillozet-Bongaarts,
2 B. McMurray, A. Szafer, T. A. Dolbeare, A. Stevens, L. Tirrell, T. Benner, S. Caldejon, R. A. Dalley,
3 N. Dee, C. Lau, J. Nyhus, M. Reding, Z. L. Riley, D. Sandman, E. Shen, A. van der Kouwe, A.
4 Varjabedian, M. Write, L. Zollei, C. Dang, J. A. Knowles, C. Koch, J. W. Phillips, N. Sestan, P.
5 Wohnoutka, H. R. Zielke, J. G. Hohmann, A. R. Jones, A. Bernard, M. J. Hawrylycz, P. R. Hof, B.
6 Fischl, E. S. Lein, Comprehensive cellular-resolution atlas of the adult human brain: Adult human
7 brain atlas. *J Comp Neurol.* 524, Spc1–Spc1 (2016).
- 8 40. Y. E. Li, S. Preissl, X. Hou, Z. Zhang, K. Zhang, Y. Qiu, O. B. Poirion, B. Li, J. Chiou, H. Liu, A.
9 Pinto-Duarte, N. Kubo, X. Yang, R. Fang, X. Wang, J. Y. Han, J. Lucero, Y. Yan, M. Miller, S.
10 Kuan, D. Gorkin, K. J. Gaulton, Y. Shen, M. Nunn, E. A. Mukamel, M. M. Behrens, J. R. Ecker, B.
11 Ren, An atlas of gene regulatory elements in adult mouse cerebrum. *Nature.* 598, 129–136 (2021).
- 12 41. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoutte, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M.
13 Myers, M. Brown, W. Li, X. S. Liu, Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* 9,
14 R137 (2008).
- 15 42. M. R. Corces, J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, C.
16 Groeneveld, C. K. Wong, S. W. Cho, A. T. Satpathy, M. R. Mumbach, K. A. Hoadley, A. G.
17 Robertson, N. C. Sheffield, I. Felau, M. A. A. Castro, B. P. Berman, L. M. Staudt, J. C. Zenklusen, P.
18 W. Laird, C. Curtis, T. C. G. A. A. Network, W. J. Greenleaf, H. Y. Chang, R. Akbani, C. C. Benz, E.
19 A. Boyle, B. M. Broom, A. D. Cherniack, B. Craft, J. A. Demchok, A. S. Doane, O. Elemento, M. L.
20 Ferguson, M. J. Goldman, D. N. Hayes, J. He, T. Hinoue, M. Imielinski, S. J. M. Jones, A. Kemal, T.
21 A. Knijnenburg, A. Korkut, D.-C. Lin, Y. Liu, M. K. A. Mensah, G. B. Mills, V. P. Reuter, A.
22 Schultz, H. Shen, J. P. Smith, R. Tarnuzzer, S. Trefflich, Z. Wang, J. N. Weinstein, L. C. Westlake, J.
23 Xu, L. Yang, C. Yau, Y. Zhao, J. Zhu, The chromatin accessibility landscape of primary human
24 cancers. *Science.* 362 (2018), doi:10.1126/science.aav1898.
- 25 43. W. Meuleman, A. Muratov, E. Rynes, J. Halow, K. Lee, D. Bates, M. Diegel, D. Dunn, F. Neri,
26 A. Teodosiadis, A. Reynolds, E. Haugen, J. Nelson, A. Johnson, M. Frerker, M. Buckley, R.
27 Sandstrom, J. Vierstra, R. Kaul, J. Stamatoyannopoulos, Index and biological spectrum of human
28 DNase I hypersensitive sites. *Nature.* 584, 244–251 (2020).
- 29 44. K. Zhang, J. D. Hocker, M. Miller, X. Hou, J. Chiou, O. B. Poirion, Y. Qiu, Y. E. Li, K. J.
30 Gaulton, A. Wang, S. Preissl, B. Ren, A single-cell atlas of chromatin accessibility in the human
31 genome. *Cell.* 184, 5985-6001.e19 (2021).
- 32 45. A.-M. Oproescu, S. Han, C. Schuurmans, New Insights Into the Intricacies of Proneural Gene
33 Regulation in the Embryonic and Adult Cerebral Cortex. *Front Mol Neurosci.* 14, 642016 (2021).
- 34 46. J. Wittstatt, S. Reiprich, M. Küspert, Crazy Little Thing Called Sox—New Insights in
35 Oligodendroglial Sox Protein Function. *Int J Mol Sci.* 20, 2713 (2019).
- 36 47. H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa, D. A. Cusanovich, R. M. Daza, D.
37 Aghamirzaie, S. Srivatsan, X. Qiu, D. Jackson, A. Minkina, A. C. Adey, F. J. Steemers, J. Shendure,
38 C. Trapnell, Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin
39 Accessibility Data. *Mol Cell.* 71, 858-871.e8 (2018).
- 40 48. M. Y. Batiuk, A. Martirosyan, J. Wahis, F. de Vin, C. Marneffe, C. Kusserow, J. Koeppen, J. F.
41 Viana, J. F. Oliveira, T. Voet, C. P. Ponting, T. G. Belgard, M. G. Holt, Identification of region-
42 specific astrocyte subtypes at single cell resolution. *Nat Commun.* 11, 1220 (2020).
- 43 49. S. Marques, A. Zeisel, S. Codeluppi, D. van Bruggen, A. M. Falcão, L. Xiao, H. Li, M. Häring, H.
44 Hochgerner, R. A. Romanov, D. Gyllborg, A. B. Muñoz-Manchado, G. L. Manno, P. Lönnerberg, E.

- 1 M. Floriddia, F. Rezayee, P. Ernfors, E. Arenas, J. Hjerling-Leffler, T. Harkany, W. D. Richardson, S.
2 Linnarsson, G. Castelo-Branco, Oligodendrocyte heterogeneity in the mouse juvenile and adult
3 central nervous system. *Science*. 352, 1326–1329 (2016).
- 4 50. Y.-L. Tan, Y. Yuan, L. Tian, Microglial regional heterogeneity and its role in the brain. *Mol*
5 *Psychiatr*. 25, 351–367 (2020).
- 6 51. A. C. Yang, R. T. Vest, F. Kern, D. P. Lee, M. Agam, C. A. Maat, P. M. Losada, M. B. Chen, N.
7 Schaum, N. Khoury, A. Toland, K. Calcuttawala, H. Shin, R. Pálovics, A. Shin, E. Y. Wang, J. Luo,
8 D. Gate, W. J. Schulz-Schaeffer, P. Chu, J. A. Siegenthaler, M. W. McEnerney, A. Keller, T. Wyss-
9 Coray, A human brain vascular atlas reveals diverse mediators of Alzheimer’s risk. *Nature*. 603, 885–
10 892 (2022).
- 11 52. T. E. Bakken, N. L. Jorstad, Q. Hu, B. B. Lake, W. Tian, B. E. Kalmbach, M. Crow, R. D. Hodge,
12 F. M. Krienen, S. A. Sorensen, J. Eggermont, Z. Yao, B. D. Aevertmann, A. I. Aldridge, A. Bartlett,
13 D. Bertagnolli, T. Casper, R. G. Castanon, K. Crichton, T. L. Daigle, R. Dalley, N. Dee, N. Dembrow,
14 D. Diep, S.-L. Ding, W. Dong, R. Fang, S. Fischer, M. Goldman, J. Goldy, L. T. Graybuck, B. R.
15 Herb, X. Hou, J. Kanherla, M. Kroll, K. Lathia, B. van Lew, Y. E. Li, C. S. Liu, H. Liu, J. D. Lucero,
16 A. Mahurkar, D. McMillen, J. A. Miller, M. Moussa, J. R. Nery, P. R. Nicovich, S.-Y. Niu, J. Orvis,
17 J. K. Osteen, S. Owen, C. R. Palmer, T. Pham, N. Plongthongkum, O. Poirion, N. M. Reed, C.
18 Rimorin, A. Rivkin, W. J. Romanow, A. E. Sedeño-Cortés, K. Siletti, S. Somasundaram, J. Sulc, M.
19 Tieu, A. Torkelson, H. Tung, X. Wang, F. Xie, A. M. Yanny, R. Zhang, S. A. Ament, M. M. Behrens,
20 H. C. Bravo, J. Chun, A. Dobin, J. Gillis, R. Hertzano, P. R. Hof, T. Höllt, G. D. Horwitz, C. D.
21 Keene, P. V. Kharchenko, A. L. Ko, B. P. Lelieveldt, C. Luo, E. A. Mukamel, A. Pinto-Duarte, S.
22 Preissl, A. Regev, B. Ren, R. H. Scheuermann, K. Smith, W. J. Spain, O. R. White, C. Koch, M.
23 Hawrylycz, B. Tasic, E. Z. Macosko, S. A. McCarroll, J. T. Ting, H. Zeng, K. Zhang, G. Feng, J. R.
24 Ecker, S. Linnarsson, E. S. Lein, Comparative cellular analysis of motor cortex in human, marmoset
25 and mouse. *Nature*. 598, 111–119 (2021).
- 26 53. T. M. E. Consortium, F. Yue, Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom,
27 Z. Ma, C. Davis, B. D. Pope, Y. Shen, D. D. Pervouchine, S. Djebali, R. E. Thurman, R. Kaul, E.
28 Rynes, A. Kirilusha, G. K. Marinov, B. A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I.
29 Antoshechkin, G. DeSalvo, L.-H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J.
30 Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. Li, M. A. Bender, M. Zhang, R. Byron, M. T.
31 Groudine, D. McCleary, L. Pham, Z. Ye, S. Kuan, L. Edsall, Y.-C. Wu, M. D. Rasmussen, M. S.
32 Bansal, M. Kellis, C. A. Keller, C. S. Morrissey, T. Mishra, D. Jain, N. Dogan, R. S. Harris, P.
33 Cayting, T. Kawli, A. P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V. S. Malladi,
34 M. S. Cline, D. T. Erickson, V. M. Kirkup, K. Learned, C. A. Sloan, K. R. Rosenbloom, B. L. de
35 Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W. J. Kent, M. R. Santos, J.
36 Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P. J.
37 Sabo, M. S. Wilken, T. A. Reh, E. Giste, A. Shafer, T. Kuttyavin, E. Haugen, D. Dunn, A. P.
38 Reynolds, S. Neph, R. Humbert, R. S. Hansen, M. D. Bruijn, L. Selleri, A. Rudensky, S. Josefowicz,
39 R. Samstein, E. E. Eichler, S. H. Orkin, D. Levasseur, T. Papayannopoulou, K.-H. Chang, A.
40 Skoultschi, S. Gosh, C. Disteché, P. Treuting, Y. Wang, M. J. Weiss, G. A. Blobel, X. Cao, S. Zhong,
41 T. Wang, P. J. Good, R. F. Lowdon, L. B. Adams, X.-Q. Zhou, M. J. Pazin, E. A. Feingold, B. Wold,
42 J. Taylor, A. Mortazavi, S. M. Weissman, J. A. Stamatoyannopoulos, M. P. Snyder, R. Guigo, T. R.
43 Gingeras, D. M. Gilbert, R. C. Hardison, M. A. Beer, B. Ren, A comparative encyclopedia of DNA
44 elements in the mouse genome. *Nature*. 515, 355–364 (2014).
- 45 54. L. Minnoye, I. I. Taskiran, D. Mauduit, M. Fazio, L. V. Aerschot, G. Hulselmans, V. Christiaens,
46 S. Makhzami, M. Seltenhammer, P. Karras, A. Primot, E. Cadieu, E. van Rooijen, J.-C. Marine, G.
47 Egidy, G. E. Ghanem, L. Zon, J. Wouters, S. Aerts, *Genome Res*, in press,
48 doi:10.1101/gr.260844.120.

55. M. T. Reilly, G. J. Faulkner, J. Dubnau, I. Ponomarev, F. H. Gage, The Role of Transposable Elements in Health and Diseases of the Central Nervous System. *J Neurosci.* 33, 17577–17586 (2013).
56. A. Ahmadi, I. D. Toma, N. Vilor-Tejedor, M. R. E. Ghamsari, I. Sadeghi, Transposable elements in brain health and disease. *Ageing Res Rev.* 64, 101153 (2020).
57. Ö. Deniz, M. Ahmed, C. D. Todd, A. Rio-Machin, M. A. Dawson, M. R. Branco, Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nat Commun.* 11, 3506 (2020).
58. J. Ito, R. Sugimoto, H. Nakaoka, S. Yamada, T. Kimura, T. Hayano, I. Inoue, Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *Plos Genet.* 13, e1006883 (2017).
59. C. Hafemeister, R. Satija, Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 20, 296 (2019).
60. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. *Cell.* 184, 3573-3587.e29 (2021).
61. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods.* 16, 1289–1296 (2019).
62. F. Wang, J. Flanagan, N. Su, L.-C. Wang, S. Bui, A. Nielson, X. Wu, H.-T. Vo, X.-J. Ma, Y. Luo, RNAscope A Novel in Situ RNA Analysis Platform for Formalin-Fixed, Paraffin-Embedded Tissues. *J Mol Diagnostics.* 14, 22–29 (2012).
63. T. E. Bakken, N. L. Jorstad, Q. Hu, B. B. Lake, W. Tian, B. E. Kalmbach, M. Crow, R. D. Hodge, F. M. Krienen, S. A. Sorensen, J. Eggermont, Z. Yao, B. D. Aeversmann, A. I. Aldridge, A. Bartlett, D. Bertagnolli, T. Casper, R. G. Castanon, K. Crichton, T. L. Daigle, R. Dalley, N. Dee, N. Dembrow, D. Diep, S.-L. Ding, W. Dong, R. Fang, S. Fischer, M. Goldman, J. Goldy, L. T. Graybuck, B. R. Herb, X. Hou, J. Kancharla, M. Kroll, K. Lathia, B. van Lew, Y. E. Li, C. S. Liu, H. Liu, J. D. Lucero, A. Mahurkar, D. McMillen, J. A. Miller, M. Moussa, J. R. Nery, P. R. Nicovich, S.-Y. Niu, J. Orvis, J. K. Osteen, S. Owen, C. R. Palmer, T. Pham, N. Plongthongkum, O. Poirion, N. M. Reed, C. Rimorin, A. Rivkin, W. J. Romanow, A. E. Sedeño-Cortés, K. Siletti, S. Somasundaram, J. Sulc, M. Tieu, A. Torkelson, H. Tung, X. Wang, F. Xie, A. M. Yanny, R. Zhang, S. A. Ament, M. M. Behrens, H. C. Bravo, J. Chun, A. Dobin, J. Gillis, R. Hertzano, P. R. Hof, T. Höllt, G. D. Horwitz, C. D. Keene, P. V. Kharchenko, A. L. Ko, B. P. Lelieveldt, C. Luo, E. A. Mukamel, A. Pinto-Duarte, S. Preissl, A. Regev, B. Ren, R. H. Scheuermann, K. Smith, W. J. Spain, O. R. White, C. Koch, M. Hawrylycz, B. Tasic, E. Z. Macosko, S. A. McCarroll, J. T. Ting, H. Zeng, K. Zhang, G. Feng, J. R. Ecker, S. Linnarsson, E. S. Lein, Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature.* 598, 111–119 (2021).
64. Z. Yao, C. T. J. van Velthoven, T. N. Nguyen, J. Goldy, A. E. Seden-Cortes, F. Baftizadeh, D. Bertagnolli, T. Casper, M. Chiang, K. Crichton, S.-L. Ding, O. Fong, E. Garren, A. Glandon, N. W. Gouwens, J. Gray, L. T. Graybuck, M. J. Hawrylycz, D. Hirschstein, M. Kroll, K. Lathia, C. Lee, B. Levi, D. McMillen, S. Mok, T. Pham, Q. Ren, C. Rimorin, N. Shapovalova, J. Sulc, S. M. Sunkin, M. Tieu, A. Torkelson, H. Tung, K. Ward, N. Dee, K. A. Smith, B. Tasic, H. Zeng, A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell.* 184, 3222-3241.e26 (2021).

- 1 65. M. Ghandi, M. Mohammad-Noori, N. Ghareghani, D. Lee, L. Garraway, M. A. Beer, gkmSVM:
2 an R package for gapped-kmer SVM. *Bioinform Oxf Engl.* 32, 2205–7 (2015).
- 3 66. J. Nelson, R. Bundoc-Baronia, G. Comiskey, T. F. McGovern, Facing Addiction in America: The
4 Surgeon General’s Report on Alcohol, Drugs, and Health: A Commentary. *Alcohol Treat Q.* 35, 445–
5 454 (2017).
- 6 67. Y. Li, M.-S. Tan, T. Jiang, L. Tan, Microglia in Alzheimer’s Disease. *Biomed Res Int.* 2014,
7 437483 (2014).
- 8 68. A. Nott, I. R. Holtman, N. G. Coufal, J. C. M. Schlachetzki, M. Yu, R. Hu, C. Z. Han, M. Pena, J.
9 Xiao, Y. Wu, Z. Keulen, M. P. Pasillas, C. O’Connor, C. K. Nickl, S. T. Schafer, Z. Shen, R. A.
10 Rissman, J. B. Brewer, D. Gosselin, D. D. Gonda, M. L. Levy, M. G. Rosenfeld, G. McVicker, F. H.
11 Gage, B. Ren, C. K. Glass, Brain cell type–specific enhancer–promoter interactome maps and
12 disease-risk association. *Science.* 366, 1134–1139 (2019).
- 13 69. K. K. Marballi, A. L. Gallitano, Immediate Early Genes Anchor a Biological Pathway of Proteins
14 Required for Memory Formation, Long-Term Depression and Risk for Schizophrenia. *Front Behav*
15 *Neurosci.* 12, 23 (2018).
- 16 70. R. Kosoy, J. F. Fullard, B. Zeng, J. Bendl, P. Dong, S. Rahman, S. P. Kleopoulos, Z. Shao, K.
17 Girdhar, J. Humphrey, K. de P. Lopes, A. W. Charney, B. H. Kopell, T. Raj, D. Bennett, C. P.
18 Kellner, V. Haroutunian, G. E. Hoffman, P. Roussos, Genetics of the human microglia regulome
19 refines Alzheimer’s disease risk loci. *Nat Genet.* 54, 1145–1154 (2022).
- 20 71. E. Drummond, T. Wisniewski, Alzheimer’s disease: experimental models and reality. *Acta*
21 *Neuropathol.* 133, 155–175 (2017).
- 22 72. Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael,
23 J. Jumper, P. Kohli, D. R. Kelley, Effective gene expression prediction from sequence by integrating
24 long-range interactions. *Nat Methods.* 18, 1196–1203 (2021).
- 25 73. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin,
26 Attention Is All You Need. *Arxiv* (2017), doi:10.48550/arxiv.1706.03762.
- 27 74. A. Zakrzewska, C. Cui, O. W. Stockhammer, E. L. Benard, H. P. Spaink, A. H. Meijer,
28 Macrophage-specific gene functions in Spi1-directed innate immunity. *Blood.* 116, e1–e11 (2010).
- 29 75. R. Dahl, D. L. Ramirez-Bergeron, S. Rao, M. C. Simon, Spi-B can functionally replace PU.1 in
30 myeloid but not lymphoid development. *Embo J.* 21, 2220–2230 (2002).
- 31 76. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform.
32 *Bioinformatics.* 25, 1754–1760 (2009).
- 33 77. J. Ou, H. Liu, J. Yu, M. A. Kelliher, L. H. Castilla, N. D. Lawson, L. J. Zhu, ATACseqQC: a
34 Bioconductor package for post-alignment quality assessment of ATAC-seq data. *Bmc Genomics.* 19,
35 169 (2018).
- 36 78. R. Fang, S. Preissl, Y. Li, X. Hou, J. Lucero, X. Wang, A. Motamedi, A. K. Shiau, X. Zhou, F.
37 Xie, E. A. Mukamel, K. Zhang, Y. Zhang, M. M. Behrens, J. R. Ecker, B. Ren, Comprehensive
38 analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun.* 12, 1337 (2021).

- 1 79. A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C.
2 Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T.
3 Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T.
4 Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martínez, S. Mohanan, P.
5 Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E.
6 Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A.
7 Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B.
8 Paten, M. L. Tress, P. Flicek, GENCODE 2021. *Nucleic Acids Res.* 49, gkaa1087- (2020).
- 9 80. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational Identification of Cell Doublets in
10 Single-Cell Transcriptomic Data. *Cell Syst.* 8, 281-291.e9 (2019).
- 11 81. T. Benaglia, D. Chauveau, D. R. Hunter, D. Young, mixtools : An R Package for Analyzing Finite
12 Mixture Models. *J Stat Softw.* 32 (2009), doi:10.18637/jss.v032.i06.
- 13 82. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic
14 Regions of the Genome. *Sci Rep-uk.* 9, 9354 (2019).
- 15 83. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected
16 communities. *Sci Rep-uk.* 9, 5233 (2019).
- 17 84. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and
18 Projection. *J Open Source Softw.* 3, 861 (2018).
- 19 85. R. Suzuki, H. Shimodaira, Pvcust: an R package for assessing the uncertainty in hierarchical
20 clustering. *Bioinformatics.* 22, 1540–1542 (2006).
- 21 86. H.-G. Drost, Philentropy: Information Theory and Distance Quantification with R. *J Open Source*
22 *Softw.* 3, 765 (2018).
- 23 87. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoutte, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M.
24 Myers, M. Brown, W. Li, X. S. Liu, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9,
25 R137 (2008).
- 26 88. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features.
27 *Bioinformatics.* 26, 841–842 (2010).
- 28 89. D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L.
29 Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe’er, Recovering
30 Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell.* 174, 716-729.e27 (2018).
- 31 90. R. D. Hodge, T. E. Bakken, J. A. Miller, K. A. Smith, E. R. Barkan, L. T. Graybuck, J. L. Close,
32 B. Long, N. Johansen, O. Penn, Z. Yao, J. Eggermont, T. Höllt, B. P. Levi, S. I. Shehata, B.
33 Aevertmann, A. Beller, D. Bertagnolli, K. Brouner, T. Casper, C. Cobbs, R. Dalley, N. Dee, S.-L.
34 Ding, R. G. Ellenbogen, O. Fong, E. Garren, J. Goldy, R. P. Gwinn, D. Hirschstein, C. D. Keene, M.
35 Keshk, A. L. Ko, K. Lathia, A. Mahfouz, Z. Maltzer, M. McGraw, T. N. Nguyen, J. Nyhus, J. G.
36 Ojemann, A. Oldre, S. Parry, S. Reynolds, C. Rimorin, N. V. Shapovalova, S. Somasundaram, A.
37 Szafer, E. R. Thomsen, M. Tieu, G. Quon, R. H. Scheuermann, R. Yuste, S. M. Sunkin, B. Lelieveldt,
38 D. Feng, L. Ng, A. Bernard, M. Hawrylycz, J. W. Phillips, B. Tasic, H. Zeng, A. R. Jones, C. Koch,
39 E. S. Lein, Conserved cell types with divergent features in human versus mouse cortex. *Nature.* 573,
40 61–68 (2019).

91. Y. E. Li, M. Xiao, B. Shi, Y.-C. T. Yang, D. Wang, F. Wang, M. Marcia, Z. J. Lu, Identification of high-confidence RNA regulatory elements by combinatorial classification of RNA–protein binding sites. *Genome Biol.* 18, 169 (2017).
92. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python. *Arxiv* (2012), doi:10.48550/arxiv.1201.0490.
93. P. O. Hoyer, Non-negative matrix factorization with sparseness constraints. *Arxiv* (2004), doi:10.48550/arxiv.cs/0408058.
94. H. Kim, H. Park, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics.* 23, 1495–1502 (2007).
95. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol.* 32, 381–386 (2014).
96. M. L. Delignette-Muller, C. Dutang, fitdistrplus : An R Package for Fitting Distributions. *J Stat Softw.* 64 (2015), doi:10.18637/jss.v064.i04.
97. A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A. Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress, P. Flicek, GENCODE 2021. *Nucleic Acids Res.* 49, gkaa1087- (2020).
98. C. Zhu, Y. Zhang, Y. E. Li, J. Lucero, M. M. Behrens, B. Ren, Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat Methods.* 18, 283–292 (2021).
99. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25 (2009).
100. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9, 357–359 (2012).
101. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods.* 14, 975–978 (2017).
102. F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44, W160-5 (2016).
103. P. Shaw, J. Uszkoreit, A. Vaswani, Self-Attention with Relative Position Representations. *Arxiv* (2018), doi:10.48550/arxiv.1803.02155.
104. N. Shazeer, GLU Variants Improve Transformer. *Arxiv* (2020), doi:10.48550/arxiv.2002.05202.

- 1 105. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H.
2 Singh, C. K. Glass, Simple Combinations of Lineage-Determining Transcription Factors Prime cis-
3 Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell*. 38, 576–589 (2010).
- 4 106. E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, A. Ma'ayan,
5 Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *Bmc Bioinformatics*.
6 14, 128 (2013).
- 7 107. B. T. Lee, G. P. Barber, A. Benet-Pagès, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N.
8 Gonzalez, A. S. Hinrichs, C. M. Lee, P. Muthuraman, L. R. Nassar, B. Nguy, T. Pereira, G. Perez, B.
9 J. Raney, K. R. Rosenbloom, D. Schmelter, M. L. Speir, B. D. Wick, A. S. Zweig, D. Haussler, R. M.
10 Kuhn, M. Haeussler, W. J. Kent, The UCSC Genome Browser database: 2022 update. *Nucleic Acids*
11 *Res*. 50, D1115–D1122 (2021).
- 12 108. G. Schumann, C. Liu, P. O'Reilly, H. Gao, P. Song, B. Xu, B. Ruggeri, N. Amin, T. Jia, S. Preis,
13 M. S. Lepe, S. Akira, C. Barbieri, S. Baumeister, S. Cauchi, T.-K. Clarke, S. Enroth, K. Fischer, J.
14 Hällfors, S. E. Harris, S. Hieber, E. Hofer, J.-J. Hottenga, Å. Johansson, P. K. Joshi, N. Kaartinen, J.
15 Laitinen, R. Lemaitre, A. Loukola, J. Luan, L.-P. Lyytikäinen, M. Mangino, A. Manichaikul, H.
16 Mbarek, Y. Milanese, A. Moayyeri, K. Mukamal, C. Nelson, J. Nettleton, E. Partinen, R. Rawal, A.
17 Robino, L. Rose, C. Sala, T. Satoh, R. Schmidt, K. Schraut, R. Scott, A. V. Smith, J. M. Starr, A.
18 Teumer, S. Trompet, A. G. Uitterlinden, C. Venturini, A.-C. Vergnaud, N. Verweij, V. Vitart, D.
19 Vuckovic, J. Wedenoja, L. Yengo, B. Yu, W. Zhang, J. H. Zhao, D. I. Boomsma, J. Chambers, D. I.
20 Chasman, T. Daniela, E. de Geus, I. Deary, J. G. Eriksson, T. Esko, V. Eulenburg, O. H. Franco, P.
21 Froguel, C. Gieger, H. J. Grabe, V. Gudnason, U. Gyllensten, T. B. Harris, A.-L. Hartikainen, A. C.
22 Heath, L. Hocking, A. Hofman, C. Huth, M.-R. Jarvelin, J. W. Jukema, J. Kaprio, J. S. Kooner, Z.
23 Kutalik, J. Lahti, C. Langenberg, T. Lehtimäki, Y. Liu, P. A. F. Madden, N. Martin, A. Morrison, B.
24 Penninx, N. Pirastu, B. Psaty, O. Raitakari, P. Ridker, R. Rose, J. I. Rotter, N. J. Samani, H. Schmidt,
25 T. D. Spector, D. Stott, D. Strachan, I. Tzoulaki, P. van der Harst, C. M. van Duijn, P. Marques-Vidal,
26 P. Vollenweider, N. J. Wareham, J. B. Whitfield, J. Wilson, B. Wolffenbuttel, G. Bakalkin, E.
27 Evangelou, Y. Liu, K. M. Rice, S. Desrivieres, S. A. Kliewer, D. J. Mangelsdorf, C. P. Müller, D.
28 Levy, P. Elliott, KLB is associated with alcohol drinking, and its gene product β -Klotho is necessary
29 for FGF21 regulation of alcohol preference. *Proc National Acad Sci*. 113, 14372–14377 (2016).
- 30 109. I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K.
31 Karlsson, S. Hägg, L. Athanasiu, N. Voyle, P. Proitsi, A. Witoelar, S. Stringer, D. Aarsland, I. S.
32 Almdahl, F. Andersen, S. Bergh, F. Bettella, S. Björnsson, A. Brækhus, G. Bräthen, C. de Leeuw, R.
33 S. Desikan, S. Djurovic, L. Dumitrescu, T. Fladby, T. J. Hohman, P. V. Jonsson, S. J. Kiddle, A.
34 Rongve, I. Saltvedt, S. B. Sando, G. Selbæk, M. Shoai, N. G. Skene, J. Snaedal, E. Stordal, I. D.
35 Ulstein, Y. Wang, L. R. White, J. Hardy, J. Hjerling-Leffler, P. F. Sullivan, W. M. van der Flier, R.
36 Dobson, L. K. Davis, H. Stefansson, K. Stefansson, N. L. Pedersen, S. Ripke, O. A. Andreassen, D.
37 Posthuma, Genome-wide meta-analysis identifies new loci and functional pathways influencing
38 Alzheimer's disease risk. *Nat Genet*. 51, 404–413 (2019).
- 39 110. A. N. G. Initiative, E. D. W. G. of the P. G. Consortium, H. J. Watson, Z. Yilmaz, L. M.
40 Thornton, C. Hübel, J. R. I. Coleman, H. A. Gaspar, J. Bryois, A. Hinney, V. M. Leppä, M.
41 Mattheisen, S. E. Medland, S. Ripke, S. Yao, P. Giusti-Rodríguez, K. B. Hanscombe, K. L. Purves, R.
42 A. H. Adan, L. Alfredsson, T. Ando, O. A. Andreassen, J. H. Baker, W. H. Berrettini, I. Boehm, C.
43 Boni, V. B. Perica, K. Buehren, R. Burghardt, M. Cassina, S. Cichon, M. Clementi, R. D. Cone, P.
44 Courtet, S. Crow, J. J. Crowley, U. N. Danner, O. S. P. Davis, M. de Zwaan, G. Dedoussis, D.
45 Degortes, J. E. DeSocio, D. M. Dick, D. Dikeos, C. Dina, M. Dmitrzak-Weglarz, E. Docampo, L. E.
46 Duncan, K. Egberts, S. Ehrlich, G. Escaramis, T. Esko, X. Estivill, A. Farmer, A. Favaro, F.
47 Fernández-Aranda, M. M. Fichter, K. Fischer, M. Föcker, L. Foretova, A. J. Forstner, M. Forzan, C.
48 S. Franklin, S. Gallinger, I. Giegling, J. Giuranna, F. Gonidakis, P. Gorwood, M. G. Mayora, S.
49 Guillaume, Y. Guo, H. Hakonarson, K. Hatzikotoulas, J. Hauser, J. Hebebrand, S. G. Helder, S.

- 1 Herms, B. Herpertz-Dahlmann, W. Herzog, L. M. Huckins, J. I. Hudson, H. Imgart, H. Inoko, V.
- 2 Janout, S. Jiménez-Murcia, A. Julià, G. Kalsi, D. Kaminská, J. Kaprio, L. Karhunen, A. Karwautz, M.
- 3 J. H. Kas, J. L. Kennedy, A. Keski-Rahkonen, K. Kiezebrink, Y.-R. Kim, L. Klareskog, K. L. Klump,
- 4 G. P. S. Knudsen, M. C. L. Via, S. L. Hellard, R. D. Levitan, D. Li, L. Lilenfeld, B. D. Lin, J.
- 5 Lissowska, J. Luykx, P. J. Magistretti, M. Maj, K. Mannik, S. Marsal, C. R. Marshall, M.
- 6 Mattingdsdal, S. McDevitt, P. McGuffin, A. Metspalu, I. Meulenbelt, N. Micali, K. Mitchell, A. M.
- 7 Monteleone, P. Monteleone, M. A. Munn-Chernoff, B. Nacmias, M. Navratilova, I. Ntalla, J. K.
- 8 O'Toole, R. A. Ophoff, L. Padyukov, A. Palotie, J. Pantel, H. Papezova, D. Pinto, R. Rabionet, A.
- 9 Raevuori, N. Ramoz, T. Reichborn-Kjennerud, V. Ricca, S. Ripatti, F. Ritschel, M. Roberts, A.
- 10 Rotondo, D. Rujescu, F. Rybakowski, P. Santonastaso, A. Scherag, S. W. Scherer, U. Schmidt, N. J.
- 11 Schork, A. Schosser, J. Seitz, L. Slachtova, P. E. Slagboom, M. C. T. S.-O. 't Landt, A. Slopian, S.
- 12 Sorbi, B. Świątkowska, J. P. Szatkiewicz, I. Tachmazidou, E. Tenconi, A. Tortorella, F. Tozzi, J.
- 13 Treasure, A. Tsitsika, M. Tyszkiewicz-Nwafor, K. Tziouvas, A. A. van Elburg, E. F. van Furth, G.
- 14 Wagner, E. Walton, E. Widen, E. Zeggini, S. Zerwas, S. Zipfel, A. W. Bergen, J. M. Boden, H.
- 15 Brandt, S. Crawford, K. A. Halmi, L. J. Horwood, C. Johnson, A. S. Kaplan, W. H. Kaye, J. E.
- 16 Mitchell, C. M. Olsen, J. F. Pearson, N. L. Pedersen, M. Strober, T. Werge, D. C. Whiteman, D. B.
- 17 Woodside, G. D. Stuber, S. Gordon, J. Grove, A. K. Henders, A. Juréus, K. M. Kirk, J. T. Larsen, R.
- 18 Parker, L. Petersen, J. Jordan, M. Kennedy, G. W. Montgomery, T. D. Wade, A. Birgegård, P.
- 19 Lichtenstein, C. Norring, M. Landén, N. G. Martin, P. B. Mortensen, P. F. Sullivan, G. Breen, C. M.
- 20 Bulik, Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric
- 21 origins for anorexia nervosa. *Nat Genet.* 51, 1207–1214 (2019).
- 22 111. D. Demontis, R. K. Walters, J. Martin, M. Mattheisen, T. D. Als, E. Agerbo, G. Baldursson, R.
- 23 Belliveau, J. Bybjerg-Grauholm, M. Bækvad-Hansen, F. Cerrato, K. Chambert, C. Churchhouse, A.
- 24 Dumont, N. Eriksson, M. Gandal, J. I. Goldstein, K. L. Grasby, J. Grove, O. O. Gudmundsson, C. S.
- 25 Hansen, M. E. Hauberg, M. V. Hollegaard, D. P. Howrigan, H. Huang, J. B. Maller, A. R. Martin, N.
- 26 G. Martin, J. Moran, J. Pallesen, D. S. Palmer, C. B. Pedersen, M. G. Pedersen, T. Poterba, J. B.
- 27 Poulsen, S. Ripke, E. B. Robinson, F. K. Satterstrom, H. Stefansson, C. Stevens, P. Turley, G. B.
- 28 Walters, H. Won, M. J. Wright, Ö. Albayrak, R. J. L. Anney, M. J. Arranz, T. J. Banaschewski, C.
- 29 Bau, J. Biederman, J. K. Buitelaar, M. Casas, A. Charach, J. Crosbie, A. Dempfle, A. E. Doyle, R. P.
- 30 Ebstein, J. Elia, C. Freitag, M. Föcker, M. Gill, E. Grevet, Z. Hawi, J. Hebebrand, B. Herpertz-
- 31 Dahlmann, A. Hervas, A. Hinney, S. Hohmann, P. Holmans, M. Hutz, A. Ickowitz, S. Johansson, L.
- 32 Kent, S. Kittel-Schneider, N. Lambregts-Rommelse, G. Lehmkuhl, S. K. Loo, J. J. McGough, J.
- 33 Meyer, E. Mick, F. Middleton, A. Miranda, N. R. Mota, F. Mulas, A. Mulligan, F. Nelson, T. T.
- 34 Nguyen, R. D. Oades, M. C. O'Donovan, M. J. Owen, H. Palmason, J. A. Ramos-Quiroga, T. J.
- 35 Renner, M. Ribasés, M. Rietschel, O. Rivero, J. Romanos, M. Romanos, A. Rothenberger, H. Royers,
- 36 C. Sánchez-Mora, A. Scherag, B. G. Schimmelmann, H. Schäfer, J. Sergeant, J. Sinzig, S. L. Smalley,
- 37 H.-C. Steinhausen, M. Thompson, A. Todorov, A. A. Vasquez, S. Walitza, Y. Wang, A. Warnke, N.
- 38 Williams, S. H. Witt, L. Yang, T. Zayats, Y. Zhang-James, G. D. Smith, G. E. Davies, E. A. Ehli, D.
- 39 M. Evans, I. O. Fedko, C. U. Greven, M. M. Groen-Blokhuys, M. Guxens, A. R. Hammerschlag, C. A.
- 40 Hartman, J. Heinrich, J. J. Hottenga, J. Hudziak, A. Jugessur, J. P. Kemp, E. Krapohl, M. Murcia, R.
- 41 Myhre, I. M. Nolte, D. R. Nyholt, J. Ormel, K. G. Ouwens, I. Pappa, C. E. Pennell, R. Plomin, S.
- 42 Ring, M. Standl, E. Stergiakouli, B. S. Pourcain, C. Stoltenberg, J. Sunyer, E. Thiering, H. Tiemeier,
- 43 C. M. T. Tiesler, N. J. Timpson, M. Trzaskowski, P. J. van der Most, N. Vilor-Tejedor, C. A. Wang,
- 44 A. J. O. Whitehouse, H. Zhao, M. Agee, B. Alipanahi, A. Auton, R. K. Bell, K. Bryc, S. L. Elson, P.
- 45 Fontanillas, N. A. Furlotte, D. A. Hinds, B. S. Hromatka, K. E. Huber, A. Kleinman, N. K. Litterman,
- 46 M. H. McIntyre, J. L. Mountain, C. A. M. Northover, S. J. Pitts, J. F. Sathirapongsasuti, O. V.
- 47 Sazonova, J. F. Shelton, S. Shringarpure, C. Tian, V. Vacic, C. H. Wilson, O. A. Andreassen, P.
- 48 Asherson, C. L. Burton, D. I. Boomsma, B. Cormand, S. Dalsgaard, B. Franke, J. Gelernter, D.
- 49 Geschwind, H. Hakonarson, J. Haavik, H. R. Kranzler, J. Kuntsi, K. Langley, K.-P. Lesch, C.
- 50 Middeldorp, A. Reif, L. A. Rohde, P. Roussos, R. Schachar, P. Sklar, E. J. S. Sonuga-Barke, P. F.
- 51 Sullivan, A. Thapar, J. Y. Tung, I. D. Waldman, S. E. Medland, K. Stefansson, M. Nordentoft, D. M.
- 52 Hougaard, T. Werge, O. Mors, P. B. Mortensen, M. J. Daly, S. V. Faraone, A. D. Børghlum, B. M.
- 53 Neale, Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity
- 54 disorder. *Nat Genet.* 51, 63–75 (2019).

- 1 112. J. Grove, S. Ripke, T. D. Als, M. Mattheisen, R. K. Walters, H. Won, J. Pallesen, E. Agerbo, O.
2 A. Andreassen, R. Anney, S. Awashti, R. Belliveau, F. Bettella, J. D. Buxbaum, J. Bybjerg-
3 Grauholm, M. Bækvad-Hansen, F. Cerrato, K. Chambert, J. H. Christensen, C. Churchhouse, K.
4 Dellenvall, D. Demontis, S. D. Rubeis, B. Devlin, S. Djurovic, A. L. Dumont, J. I. Goldstein, C. S.
5 Hansen, M. E. Hauberg, M. V. Hollegaard, S. Hope, D. P. Howrigan, H. Huang, C. M. Hultman, L.
6 Klei, J. Maller, J. Martin, A. R. Martin, J. L. Moran, M. Nyegaard, T. Nærland, D. S. Palmer, A.
7 Palotie, C. B. Pedersen, M. G. Pedersen, T. dPoterba, J. B. Poulsen, B. S. Pourcain, P. Qvist, K.
8 Rehnström, A. Reichenberg, J. Reichert, E. B. Robinson, K. Roeder, P. Roussos, E. Saemundsen, S.
9 Sandin, F. K. Satterstrom, G. D. Smith, H. Stefansson, S. Steinberg, C. R. Stevens, P. F. Sullivan, P.
10 Turley, G. B. Walters, X. Xu, N. R. Wray, M. Trzaskowski, E. M. Byrne, A. Abdellaoui, M. J.
11 Adams, T. M. Air, T. F. M. Andlauer, S.-A. Bacanu, A. T. F. Beekman, T. B. Bigdeli, E. B. Binder,
12 D. H. R. Blackwood, J. Bryois, H. N. Buttenschön, N. Cai, E. Castelao, T.-K. Clarke, J. R. I.
13 Coleman, L. Colodro-Conde, B. Couvy-Duchesne, N. Craddock, G. E. Crawford, G. Davies, I. J.
14 Deary, F. Degenhardt, E. M. Derks, N. Direk, C. V. Dolan, E. C. Dunn, T. C. Eley, V. Escott-Price, F.
15 F. H. Kiadeh, H. K. Finucane, A. J. Forstner, J. Frank, H. A. Gaspar, M. Gill, F. S. Goes, S. D.
16 Gordon, L. S. Hall, T. F. Hansen, S. Herms, I. B. Hickie, P. Hoffmann, G. Homuth, C. Horn, J.-J.
17 Hottenga, M. Ising, R. Jansen, E. Jorgenson, J. A. Knowles, I. S. Kohane, J. Kraft, W. W.
18 Kretschmar, J. Krogh, Z. Kutalik, Y. Li, P. A. Lind, D. J. MacIntyre, D. F. MacKinnon, R. M. Maier,
19 W. Maier, J. Marchini, H. Mbarek, P. McGrath, P. McGuffin, S. E. Medland, D. Mehta, C. M.
20 Middeldorp, E. Mihailov, Y. Milaneschi, L. Milani, F. M. Mondimore, G. W. Montgomery, S.
21 Mostafavi, N. Mullins, M. Nauck, B. Ng, M. G. Nivard, D. R. Nyholt, P. F. O'Reilly, H. Oskarsson,
22 M. J. Owen, J. N. Painter, R. E. Peterson, E. Pettersson, W. J. Peyrot, G. Pistis, D. Posthuma, J. A.
23 Quiroz, J. P. Rice, B. P. Riley, M. Rivera, S. S. Mirza, R. Schoevers, E. C. Schulte, L. Shen, J. Shi, S.
24 I. Shyn, E. Sigurdsson, G. C. B. Sinnamon, J. H. Smit, D. J. Smith, F. Streit, J. Strohmaier, K. E.
25 Tansey, H. Teismann, A. Teumer, W. Thompson, P. A. Thomson, T. E. Thorgeirsson, M. Traylor, J.
26 Treutlein, V. Trubetskoy, A. G. Uitterlinden, D. Umbrecht, S. V. der Auwera, A. M. van Hemert, A.
27 Viktorin, P. M. Visscher, Y. Wang, B. T. Webb, S. M. Weinsheimer, J. Wellmann, G. Willemsen, S.
28 H. Witt, Y. Wu, H. S. Xi, J. Yang, F. Zhang, V. Arolt, B. T. Baune, K. Berger, D. I. Boomsma, S.
29 Cichon, U. Dannlowski, E. J. C. de Geus, J. R. DePaulo, E. Domenici, K. Domschke, T. Esko, H. J.
30 Grabe, S. P. Hamilton, C. Hayward, A. C. Heath, K. S. Kendler, S. Kloiber, G. Lewis, Q. S. Li, S.
31 Lucae, P. A. F. Madden, P. K. Magnusson, N. G. Martin, A. M. McIntosh, A. Metspalu, B. Müller-
32 Myhsok, M. M. Nöthen, M. C. O'Donovan, S. A. Paciga, N. L. Pedersen, B. W. J. H. Penninx, R. H.
33 Perlis, D. J. Porteous, J. B. Potash, M. Preisig, M. Rietschel, C. Schaefer, T. G. Schulze, J. W.
34 Smoller, H. Tiemeier, R. Uher, H. Völzke, M. M. Weissman, C. M. Lewis, D. F. Levinson, G. Breen,
35 M. Agee, B. Alipanahi, A. Auton, R. K. Bell, K. Bryc, S. L. Elson, P. Fontanillas, N. A. Furlotte, B.
36 S. Hromatka, K. E. Huber, A. Kleinman, N. K. Litterman, M. H. McIntyre, J. L. Mountain, E. S.
37 Noblin, C. A. M. Northover, S. J. Pitts, J. F. Sathirapongsasuti, O. V. Sazonova, J. F. Shelton, S.
38 Shringarpure, J. Y. Tung, V. Vacic, C. H. Wilson, K. Stefansson, D. H. Geschwind, M. Nordentoft,
39 D. M. Hougaard, T. Werge, O. Mors, P. B. Mortensen, B. M. Neale, M. J. Daly, A. D. Børglum,
40 Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet.* 51, 431–444
41 (2019).
- 42 113. L. Hou, S. E. Bergen, N. Akula, J. Song, C. M. Hultman, M. Landén, M. Adli, M. Alda, R.
43 Arda, B. Arias, J.-M. Aubry, L. Backlund, J. A. Badner, T. B. Barrett, M. Bauer, B. T. Baune, F.
44 Bellivier, A. Benabarro, S. Bengesser, W. H. Berrettini, A. K. Bhattacharjee, J. M. Biernacka, A.
45 Birner, C. S. Bloss, C. Brichant-Petitjean, E. T. Bui, W. Byerley, P. Cervantes, C. Chillotti, S. Cichon,
46 F. Colom, W. Coryell, D. W. Craig, C. Cruceanu, P. M. Czerski, T. Davis, A. Dayer, F. Degenhardt,
47 M. D. Zompo, J. R. DePaulo, H. J. Edenberg, B. Étain, P. Falkai, T. Foroud, A. J. Forstner, L. Frisé,
48 M. A. Frye, J. M. Fullerton, S. Gard, J. S. Garnham, E. S. Gershon, F. S. Goes, T. A. Greenwood, M.
49 Grigoriou-Serbanescu, J. Hauser, U. Heilbronner, S. Heilmann-Heimbach, S. Herms, M. Hipolito, S.
50 Hitturlingappa, P. Hoffmann, A. Hofmann, S. Jamain, E. Jiménez, J.-P. Kahn, L. Kassem, J. R.
51 Kelsoe, S. Kittel-Schneider, S. Kliwicki, D. L. Koller, B. König, N. Lackner, G. Laje, M. Lang, C.
52 Lavebratt, W. B. Lawson, M. Leboyer, S. G. Leckband, C. Liu, A. Maaser, P. B. Mahon, W. Maier,
53 M. Maj, M. Manchia, L. Martinsson, M. J. McCarthy, S. L. McElroy, M. G. McInnis, R. McKinney,
54 P. B. Mitchell, M. Mitjans, F. M. Mondimore, P. Monteleone, T. W. Mühleisen, C. M. Nievergelt, M.

- 1 M. Nöthen, T. Novák, J. I. Nurnberger, E. A. Nwulia, U. Ösby, A. Pfennig, J. B. Potash, P. Propping,
2 A. Reif, E. Reininghaus, J. Rice, M. Rietschel, G. A. Rouleau, J. K. Rybakowski, M. Schalling, W. A.
3 Scheftner, P. R. Schofield, N. J. Schork, T. G. Schulze, J. Schumacher, B. W. Schweizer, G. Severino,
4 T. Shekhtman, P. D. Shilling, C. Simhandl, C. M. Slaney, E. N. Smith, A. Squassina, T. Stamm, P.
5 Stopkova, F. Streit, J. Strohmaier, S. Szelinger, S. K. Tighe, A. Tortorella, G. Turecki, E. Vieta, J.
6 Volkert, S. H. Witt, A. Wright, P. P. Zandi, P. Zhang, S. Zollner, F. J. McMahon, Genome-wide
7 association study of 40,000 individuals identifies two novel loci associated with bipolar disorder.
8 *Hum Mol Genet.* 25, 3383–3394 (2016).
- 9 114. S. E. Jones, J. M. Lane, A. R. Wood, V. T. van Hees, J. Tyrrell, R. N. Beaumont, A. R. Jeffries,
10 H. S. Dashti, M. Hillsdon, K. S. Ruth, M. A. Tuke, H. Yaghootkar, S. A. Sharp, Y. Jie, W. D.
11 Thompson, J. W. Harrison, A. Dawes, E. M. Byrne, H. Tiemeier, K. V. Allebrandt, J. Bowden, D. W.
12 Ray, R. M. Freathy, A. Murray, D. R. Mazzotti, P. R. Gehrman, D. A. Lawlor, T. M. Frayling, M. K.
13 Rutter, D. A. Hinds, R. Saxena, M. N. Weedon, Genome-wide association analyses of chronotype in
14 697,828 individuals provides insights into circadian rhythms. *Nat Commun.* 10, 343 (2019).
- 15 115. A. Okbay, J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G.-B.
16 Chen, V. Emilsson, S. F. W. Meddens, S. Oskarsson, J. K. Pickrell, K. Thom, P. Timshel, R. de
17 Vlaming, A. Abdellaoui, T. S. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, J. H. Brandsma,
18 M. P. Concas, J. Derringer, N. A. Furlotte, T. E. Galesloot, G. Girotto, R. Gupta, L. M. Hall, S. E.
19 Harris, E. Hofer, M. Horikoshi, J. E. Huffman, K. Kaasik, I. P. Kalafati, R. Karlsson, A. Kong, J.
20 Lahti, S. J. van der Lee, C. deLeeuw, P. A. Lind, K.-O. Lindgren, T. Liu, M. Mangino, J. Marten, E.
21 Mihailov, M. B. Miller, P. J. van der Most, C. Oldmeadow, A. Payton, N. Pervjakova, W. J. Peyrot,
22 Y. Qian, O. Raitakari, R. Rueedi, E. Salvi, B. Schmidt, K. E. Schraut, J. Shi, A. V. Smith, R. A. Poot,
23 B. S. Pourcain, A. Teumer, G. Thorleifsson, N. Verweij, D. Vuckovic, J. Wellmann, H.-J. Westra, J.
24 Yang, W. Zhao, Z. Zhu, B. Z. Alizadeh, N. Amin, A. Bakshi, S. E. Baumeister, G. Biino, K.
25 Bønnelykke, P. A. Boyle, H. Campbell, F. P. Cappuccio, G. Davies, J.-E. D. Neve, P. Deloukas, I.
26 Demuth, J. Ding, P. Eibich, L. Eisele, N. Eklund, D. M. Evans, J. D. Faul, M. F. Feitosa, A. J.
27 Forstner, I. Gandin, B. Gunnarsson, B. V. Halldórsson, T. B. Harris, A. C. Heath, L. J. Hocking, E. G.
28 Holliday, G. Homuth, M. A. Horan, J.-J. Hottenga, P. L. de Jager, P. K. Joshi, A. Jugessur, M. A.
29 Kaakinen, M. Kähönen, S. Kanoni, L. Keltigangas-Järvinen, L. A. L. M. Kiemeny, I. Kolcic, S.
30 Koskinen, A. T. Kraja, M. Kroh, Z. Kutalik, A. Latvala, L. J. Launer, M. P. Lebreton, D. F. Levinson,
31 P. Lichtenstein, P. Lichtner, D. C. M. Liewald, L. C. Study, A. Loukola, P. A. Madden, R. Mägi, T.
32 Mäki-Opas, R. E. Marioni, P. Marques-Vidal, G. A. Meddens, G. McMahon, C. Meisinger, T.
33 Meitinger, Y. Milaneschi, L. Milani, G. W. Montgomery, R. Myhre, C. P. Nelson, D. R. Nyholt, W.
34 E. R. Ollier, A. Palotie, L. Paternoster, N. L. Pedersen, K. E. Petrovic, D. J. Porteous, K. Räikkönen,
35 S. M. Ring, A. Robino, O. Rostapshova, I. Rudan, A. Rustichini, V. Salomaa, A. R. Sanders, A.-P.
36 Sarin, H. Schmidt, R. J. Scott, B. H. Smith, J. A. Smith, J. A. Staessen, E. Steinhagen-Thiessen, K.
37 Strauch, A. Terracciano, M. D. Tobin, S. Ulivi, S. Vaccargiu, L. Quaye, F. J. A. van Rooij, C.
38 Venturini, A. A. E. Vinkhuyzen, U. Völker, H. Völzke, J. M. Vonk, D. Vozzi, J. Waage, E. B. Ware,
39 G. Willemsen, J. R. Attia, D. A. Bennett, K. Berger, L. Bertram, H. Bisgaard, D. I. Boomsma, I. B.
40 Borecki, U. Bültmann, C. F. Chabris, F. Cucca, D. Cusi, I. J. Deary, G. V. Dedoussis, C. M. van
41 Duijn, J. G. Eriksson, B. Franke, L. Franke, P. Gasparini, P. V. Gejman, C. Gieger, H.-J. Grabe, J.
42 Gratten, P. J. F. Groenen, V. Gudnason, P. van der Harst, C. Hayward, D. A. Hinds, W. Hoffmann, E.
43 Hyppönen, W. G. Iacono, B. Jacobsson, M.-R. Järvelin, K.-H. Jöckel, J. Kaprio, S. L. R. Kardia, T.
44 Lehtimäki, S. F. Lehrer, P. K. E. Magnusson, N. G. Martin, M. McGue, A. Metspalu, N. Pendleton,
45 B. W. J. H. Penninx, M. Perola, N. Pirastu, M. Pirastu, O. Polasek, D. Posthuma, C. Power, M. A.
46 Province, N. J. Samani, D. Schlessinger, R. Schmidt, T. I. A. Sørensen, T. D. Spector, K. Stefansson,
47 U. Thorsteinsdottir, A. R. Thurik, N. J. Timpson, H. Tiemeier, J. Y. Tung, A. G. Uitterlinden, V.
48 Vitart, P. Vollenweider, D. R. Weir, J. F. Wilson, A. F. Wright, D. C. Conley, R. F. Krueger, G. D.
49 Smith, A. Hofman, D. I. Laibson, S. E. Medland, M. N. Meyer, J. Yang, M. Johannesson, P. M.
50 Visscher, T. Esko, P. D. Koellinger, D. Cesarini, D. J. Benjamin, Genome-wide association study
51 identifies 74 loci associated with educational attainment. *Nature.* 533, 539–542 (2016).

- 1 116. G. Scotland, S. M. van den Berg, M. H. M. de Moor, K. J. H. Verweij, R. F. Krueger, M.
2 Luciano, A. A. Vasquez, L. K. Matteson, J. Derringer, T. Esko, N. Amin, S. D. Gordon, N. K.
3 Hansell, A. B. Hart, I. Seppälä, J. E. Huffman, B. Konte, J. Lahti, M. Lee, M. Miller, T. Nutile, T.
4 Tanaka, A. Teumer, A. Viktorin, J. Wedenoja, A. Abdellaoui, G. R. Abecasis, D. E. Adkins, A.
5 Agrawal, J. Allik, K. Appel, T. B. Bigdeli, F. Busonero, H. Campbell, P. T. Costa, G. D. Smith, G.
6 Davies, H. de Wit, J. Ding, B. E. Engelhardt, J. G. Eriksson, I. O. Fedko, L. Ferrucci, B. Franke, I.
7 Giegling, R. Grucza, A. M. Hartmann, A. C. Heath, K. Heinonen, A. K. Henders, G. Homuth, J.-J.
8 Hottenga, W. G. Iacono, J. Janzing, M. Jokela, R. Karlsson, J. P. Kemp, M. G. Kirkpatrick, A.
9 Latvala, T. Lehtimäki, D. C. Liewald, P. A. F. Madden, C. Magri, P. K. E. Magnusson, J. Marten, A.
10 Maschio, H. Mbarek, S. E. Medland, E. Mihailov, Y. Milaneschi, G. W. Montgomery, M. Nauck, M.
11 G. Nivard, K. G. Ouwens, A. Palotie, E. Pettersson, O. Polasek, Y. Qian, L. Pulkki-Råback, O. T.
12 Raitakari, A. Realo, R. J. Rose, D. Ruggiero, C. O. Schmidt, W. S. Slutske, R. Sorice, J. M. Starr, B.
13 S. Pourcain, A. R. Sutin, N. J. Timpson, H. Trochet, S. Vermeulen, E. Vuoksimaa, E. Widen, J.
14 Wouda, M. J. Wright, L. Zgaga, D. Porteous, A. Minelli, A. A. Palmer, D. Rujescu, M. Ciullo, C.
15 Hayward, I. Rudan, A. Metspalu, J. Kaprio, I. J. Deary, K. Räikkönen, J. F. Wilson, L. Keltikangas-
16 Järvinen, L. J. Bierut, J. M. Hettema, H. J. Grabe, B. W. J. H. Penninx, C. M. van Duijn, D. M. Evans,
17 D. Schlessinger, N. L. Pedersen, A. Terracciano, M. McGue, N. G. Martin, D. I. Boomsma, Meta-
18 analysis of Genome-Wide Association Studies for Extraversion: Findings from the Genetics of
19 Personality Consortium. *Behav Genet.* 46, 170–182 (2016).
- 20 117. P. R. Jansen, K. Watanabe, S. Stringer, N. Skene, J. Bryois, A. R. Hammerschlag, C. A. de
21 Leeuw, J. S. Benjamins, A. B. Muñoz-Manchado, M. Nagel, J. E. Savage, H. Tiemeier, T. White, M.
22 Agee, B. Alipanahi, A. Auton, R. K. Bell, K. Bryc, S. L. Elson, P. Fontanillas, N. A. Furlotte, D. A.
23 Hinds, K. E. Huber, A. Kleinman, N. K. Litterman, J. C. McCreight, M. H. McIntyre, J. L. Mountain,
24 E. S. Noblin, C. A. M. Northover, S. J. Pitts, J. F. Sathirapongsasuti, O. V. Sazonova, J. F. Shelton, S.
25 Shringarpure, C. Tian, C. H. Wilson, J. Y. Tung, D. A. Hinds, V. Vacic, X. Wang, P. F. Sullivan, S.
26 van der Sluis, T. J. C. Polderman, A. B. Smit, J. Hjerling-Leffler, E. J. W. V. Someren, D. Posthuma,
27 Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional
28 pathways. *Nat Genet.* 51, 394–403 (2019).
- 29 118. J. E. Savage, P. R. Jansen, S. Stringer, K. Watanabe, J. Bryois, C. A. de Leeuw, M. Nagel, S.
30 Awasthi, P. B. Barr, J. R. I. Coleman, K. L. Grasby, A. R. Hammerschlag, J. A. Kaminski, R.
31 Karlsson, E. Krapohl, M. Lam, M. Nygaard, C. A. Reynolds, J. W. Trampush, H. Young, D. Zabaneh,
32 S. Hägg, N. K. Hansell, I. K. Karlsson, S. Linnarsson, G. W. Montgomery, A. B. Muñoz-Manchado,
33 E. B. Quinlan, G. Schumann, N. G. Skene, B. T. Webb, T. White, D. E. Arking, D. Avramopoulos, R.
34 M. Bilder, P. Bitsios, K. E. Burdick, T. D. Cannon, O. Chiba-Falek, A. Christoforou, E. T. Cirulli, E.
35 Congdon, A. Corvin, G. Davies, I. J. Deary, P. DeRosse, D. Dickinson, S. Djurovic, G. Donohoe, E.
36 D. Conley, J. G. Eriksson, T. Espeseth, N. A. Freimer, S. Giakoumaki, I. Giegling, M. Gill, D. C.
37 Glahn, A. R. Hariri, A. Hatzimanolis, M. C. Keller, E. Knowles, D. Koltai, B. Konte, J. Lahti, S. L.
38 Hellard, T. Lencz, D. C. Liewald, E. London, A. J. Lundervold, A. K. Malhotra, I. Melle, D. Morris,
39 A. C. Need, W. Ollier, A. Palotie, A. Payton, N. Pendleton, R. A. Poldrack, K. Räikkönen, I.
40 Reinvang, P. Roussos, D. Rujescu, F. W. Sabb, M. A. Scult, O. B. Smeland, N. Smyrnis, J. M. Starr,
41 V. M. Steen, N. C. Stefanis, R. E. Straub, K. Sundet, H. Tiemeier, A. N. Voineskos, D. R.
42 Weinberger, E. Widen, J. Yu, G. Abecasis, O. A. Andreassen, G. Breen, L. Christiansen, B.
43 Debrabant, D. M. Dick, A. Heinz, J. Hjerling-Leffler, M. A. Ikram, K. S. Kendler, N. G. Martin, S. E.
44 Medland, N. L. Pedersen, R. Plomin, T. J. C. Polderman, S. Ripke, S. van der Sluis, P. F. Sullivan, S.
45 I. Vrieze, M. J. Wright, D. Posthuma, Genome-wide association meta-analysis in 269,867 individuals
46 identifies new genetic and functional links to intelligence. *Nat Genet.* 50, 912–919 (2018).
- 47 119. N. R. Wray, S. Ripke, M. Mattheisen, M. Trzaskowski, E. M. Byrne, A. Abdellaoui, M. J.
48 Adams, E. Agerbo, T. M. Air, T. M. F. Andlauer, S.-A. Bacanu, M. Bækvad-Hansen, A. F. T.
49 Beekman, T. B. Bigdeli, E. B. Binder, D. R. H. Blackwood, J. Bryois, H. N. Buttenschøn, J. Bybjerg-
50 Grauholm, N. Cai, E. Castelao, J. H. Christensen, T.-K. Clarke, J. I. R. Coleman, L. Colodro-Conde,
51 B. Couvy-Duchesne, N. Craddock, G. E. Crawford, C. A. Crowley, H. S. Dashti, G. Davies, I. J.

- 1 Deary, F. Degenhardt, E. M. Derks, N. Direk, C. V. Dolan, E. C. Dunn, T. C. Eley, N. Eriksson, V.
- 2 Escott-Price, F. H. F. Kiadeh, H. K. Finucane, A. J. Forstner, J. Frank, H. A. Gaspar, M. Gill, P.
- 3 Giusti-Rodríguez, F. S. Goes, S. D. Gordon, J. Grove, L. S. Hall, E. Hannon, C. S. Hansen, T. F.
- 4 Hansen, S. Herms, I. B. Hickie, P. Hoffmann, G. Homuth, C. Horn, J.-J. Hottenga, D. M. Hougaard,
- 5 M. Hu, C. L. Hyde, M. Ising, R. Jansen, F. Jin, E. Jorgenson, J. A. Knowles, I. S. Kohane, J. Kraft,
- 6 W. W. Kretschmar, J. Krogh, Z. Kutalik, J. M. Lane, Y. Li, Y. Li, P. A. Lind, X. Liu, L. Lu, D. J.
- 7 MacIntyre, D. F. MacKinnon, R. M. Maier, W. Maier, J. Marchini, H. Mbarek, P. McGrath, P.
- 8 McGuffin, S. E. Medland, D. Mehta, C. M. Middeldorp, E. Mihailov, Y. Milaneschi, L. Milani, J.
- 9 Mill, F. M. Mondimore, G. W. Montgomery, S. Mostafavi, N. Mullins, M. Nauck, B. Ng, M. G.
- 10 Nivard, D. R. Nyholt, P. F. O'Reilly, H. Oskarsson, M. J. Owen, J. N. Painter, C. B. Pedersen, M. G.
- 11 Pedersen, R. E. Peterson, E. Pettersson, W. J. Peyrot, G. Pistis, D. Posthuma, S. M. Purcell, J. A.
- 12 Quiroz, P. Qvist, J. P. Rice, B. P. Riley, M. Rivera, S. S. Mirza, R. Saxena, R. Schoevers, E. C.
- 13 Schulte, L. Shen, J. Shi, S. I. Shyn, E. Sigurdsson, G. B. C. Sinnamon, J. H. Smit, D. J. Smith, H.
- 14 Stefansson, S. Steinberg, C. A. Stockmeier, F. Streit, J. Strohmaier, K. E. Tansey, H. Teismann, A.
- 15 Teumer, W. Thompson, P. A. Thomson, T. E. Thorgeirsson, C. Tian, M. Traylor, J. Treutlein, V.
- 16 Trubetskoy, A. G. Uitterlinden, D. Umbricht, S. V. der Auwera, A. M. van Hemert, A. Viktorin, P. M.
- 17 Visscher, Y. Wang, B. T. Webb, S. M. Weinsheimer, J. Wellmann, G. Willemsen, S. H. Witt, Y. Wu,
- 18 H. S. Xi, J. Yang, F. Zhang, eQTLGen, 23andMe, V. Arolt, B. T. Baune, K. Berger, D. I. Boomsma,
- 19 S. Cichon, U. Dannlowski, E. C. J. de Geus, J. R. DePaulo, E. Domenici, K. Domschke, T. Esko, H. J.
- 20 Grabe, S. P. Hamilton, C. Hayward, A. C. Heath, D. A. Hinds, K. S. Kendler, S. Kloiber, G. Lewis,
- 21 Q. S. Li, S. Lucae, P. F. A. Madden, P. K. Magnusson, N. G. Martin, A. M. McIntosh, A. Metspalu,
- 22 O. Mors, P. B. Mortensen, B. Müller-Myhsok, M. Nordentoft, M. M. Nöthen, M. C. O'Donovan, S.
- 23 A. Paciga, N. L. Pedersen, B. W. J. H. Penninx, R. H. Perlis, D. J. Porteous, J. B. Potash, M. Preisig,
- 24 M. Rietschel, C. Schaefer, T. G. Schulze, J. W. Smoller, K. Stefansson, H. Tiemeier, R. Uher, H.
- 25 Völzke, M. M. Weissman, T. Werge, A. R. Winslow, C. M. Lewis, D. F. Levinson, G. Breen, A. D.
- 26 Børglum, P. F. Sullivan, M. D. D. W. G. of the P. G. Consortium, Genome-wide association analyses
- 27 identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet.* 50, 668–
- 28 681 (2018).
- 29 120. I. M. S. G. Consortium, N. A. Patsopoulos, S. E. Baranzini, A. Santaniello, P. Shoostari, C.
- 30 Cotsapas, G. Wong, A. H. Beecham, T. James, J. Replogle, I. S. Vlachos, C. McCabe, T. H. Pers, A.
- 31 Brandes, C. White, B. Keenan, M. Cimpean, P. Winn, I.-P. Panteliadis, A. Robbins, T. F. M.
- 32 Andlauer, O. Zarzycki, B. Dubois, A. Goris, H. B. Søndergaard, F. Sellebjerg, P. S. Sorensen, H.
- 33 Ullum, L. W. Thørner, J. Saarela, I. Couru-Rebeix, V. Damotte, B. Fontaine, L. Guillot-Noel, M.
- 34 Lathrop, S. Vukusic, A. Berthele, V. Pongratz, D. Buck, C. Gasperi, C. Graetz, V. Grummel, B.
- 35 Hemmer, M. Hoshi, B. Knier, T. Korn, C. M. Lill, F. Luessi, M. Mühlau, F. Zipp, E. Dardiotis, C.
- 36 Agliardi, A. Amoroso, N. Barizzzone, M. D. Benedetti, L. Bernardinelli, P. Cavalla, F. Clarelli, G.
- 37 Comi, D. Cusi, F. Esposito, L. Ferrè, D. Galimberti, C. Guaschino, M. A. Leone, V. Martinelli, L.
- 38 Moiola, M. Salvetti, M. Sorosina, D. Vecchio, A. Zauli, S. Santoro, N. Mancini, M. Zuccalà, J.
- 39 Mescheriakova, C. van Duijn, S. D. Bos, E. G. Celius, A. Spurkland, M. Comabella, X. Montalban, L.
- 40 Alfredsson, I. L. Bomfim, D. Gomez-Cabrero, J. Hillert, M. Jagodic, M. Lindén, F. Piehl, I. Jelčić, R.
- 41 Martin, M. Sospedra, A. Baker, M. Ban, C. Hawkins, P. Hysi, S. Kalra, F. Karpe, J. Khadake, G.
- 42 Lachance, P. Molyneux, M. Neville, J. Thorpe, E. Bradshaw, S. J. Caillier, P. Calabresi, B. A. C.
- 43 Cree, A. Cross, M. Davis, P. W. I. de Bakker, S. Delgado, M. Dembele, K. Edwards, K. Fitzgerald, I.
- 44 Y. Frohlich, P.-A. Gourraud, J. L. Haines, H. Hakonarson, D. Kimbrough, N. Isobe, I. Konidari, E.
- 45 Lathi, M. H. Lee, T. Li, D. An, A. Zimmer, L. Madireddy, C. P. Manrique, M. Mitrovic, M. Olah, E.
- 46 Patrick, M. A. Pericak-Vance, L. Piccio, C. Schaefer, H. Weiner, K. Lage, ANZgene, IIBDGC,
- 47 WTCCC2, A. Compston, D. Hafler, H. F. Harbo, S. L. Hauser, G. Stewart, S. D'Alfonso, G.
- 48 Hadjigeorgiou, B. Taylor, L. F. Barcellos, D. Booth, R. Hintzen, I. Kockum, F. Martinelli-Boneschi,
- 49 J. L. McCauley, J. R. Oksenberg, A. Oturai, S. Sawcer, A. J. Ivinson, T. Olsson, P. L. D. Jager,
- 50 Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility.
- 51 *Science.* 365 (2019), doi:10.1126/science.aav7188.

121. M. Luciano, S. P. Hagenaars, G. Davies, W. D. Hill, T.-K. Clarke, M. Shirali, S. E. Harris, R. E. Marioni, D. C. Liewald, C. Fawns-Ritchie, M. J. Adams, D. M. Howard, C. M. Lewis, C. R. Gale, A. M. McIntosh, I. J. Deary, Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat Genet.* 50, 6–11 (2018).
122. M. W. Logue, A. B. Amstadter, D. G. Baker, L. Duncan, K. C. Koenen, I. Liberzon, M. W. Miller, R. A. Morey, C. M. Nievergelt, K. J. Ressler, A. K. Smith, J. W. Smoller, M. B. Stein, J. A. Sumner, M. Uddin, The Psychiatric Genomics Consortium Posttraumatic Stress Disorder Workgroup: Posttraumatic Stress Disorder Enters the Age of Large-Scale Genomic Collaboration. *Neuropsychopharmacol.* 40, 2287–2297 (2015).
123. S. Ripke, B. M. Neale, A. Corvin, J. T. R. Walters, K.-H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. B. Jr, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. K. Chan, R. Y. L. Chen, E. Y. H. Chen, W. Cheng, E. F. C. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. D. Favero, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, I. Giegling, P. Giusti-Rodríguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julià, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskis, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kähler, C. Laurent, J. L. C. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K.-Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lönngqvist, M. M. Jr, P. K. E. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Meshulam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Müller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O’Callaghan, C. O’Dushlaine, F. A. O’Neill, S.-Y. Oh, A. Olincy, L. Olsen, J. V. Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietiläinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H.-C. So, ChrisC. A. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Söderman, S. Thirumalai, D. Toncheva, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. M. Wong, B. K. Wormley, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, W. T. C.-C. Consortium, R. Adolfsson, O. A. Andreassen, D. H. R. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jönsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nöthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. S. Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, M. C. O’Donovan, Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 511, 421–427 (2014).

124. S. E. Jones, J. Tyrrell, A. R. Wood, R. N. Beaumont, K. S. Ruth, M. A. Tuke, H. Yaghootkar, Y. Hu, M. Teder-Laving, C. Hayward, T. Roenneberg, J. F. Wilson, F. D. Greco, A. A. Hicks, C. Shin, C.-H. Yun, S. K. Lee, A. Metspalu, E. M. Byrne, P. R. Gehrman, H. Tiemeier, K. V. Allebrandt, R. M. Freathy, A. Murray, D. A. Hinds, T. M. Frayling, M. N. Weedon, Genome-Wide Association Analyses in 128,266 Individuals Identifies New Morningness and Sleep Duration Loci. *Plos Genet.* 12, e1006125 (2016).
125. I. C. for B. P. GWAS, C. C. A. and L. Group, C. C. I. Group, V. Deary, S. P. Hagenaars, S. E. Harris, W. D. Hill, G. Davies, D. C. M. Liewald, A. M. McIntosh, C. R. Gale, I. J. Deary, Genetic contributions to self-reported tiredness. *Mol Psychiatr.* 23, 609–620 (2018).
126. W. Zhou, J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. Gifford, L. A. Bastarache, W.-Q. Wei, J. C. Denny, M. Lin, K. Hveem, H. M. Kang, G. R. Abecasis, C. J. Willer, S. Lee, Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet.* 50, 1335–1341 (2018).
127. S. W. G. of the P. G. Consortium, B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 47, 291–295 (2015).

Acknowledgments: We would like to thank Dr. Erica Melief, Aimee Schantz, Katie Kern, Amanda Keen, and Lisa Keene of the University of Washington BioRepository and Integrated Neuropathology (BRaIN) Laboratory for outstanding technical and administrative support for tissue collection, preservation, and characterization, and the brain donors and their loved ones, without whom this research would be impossible. We would like to thank the QB3 Macrolab at UC Berkeley for purification of the Tn5 transposase. 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 SIG grant (#S10 OD026929). We thank Ethan Armand for discussion and support, Qiurui Zeng for data management and transferring.

Funding: This work was supported by grants from NIMH U01MH121282 to B.R., M.M.B and J.R.E., UM1 MH130994 to J.R.E., M.M.B. and B.R., NIMH U01MH114812 to E.L and S.L, and in part by the Nancy and Buster Alvord Endowment to C.D.K. The development of deep learning models was sponsored in part by NSF Convergence Accelerator under award OIA-2040727, NIH Bridge2AI Center Program under award U54HG012510, as well as generous gifts from Google, Adobe, and Teradata. Dr. Z.W. is a DDBrown Awardee of the Life Sciences Research Foundation. J.R.E is an investigator of the Howard Hughes Medical Institute.

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

Competing interests: J.R.E is a member of the scientific advisor for Zymo Research and Ionis. B.R. is a co-founder and consultant of Arima Genomics Inc. and co-founder of Epigenome Technologies.

Data and materials availability: Demultiplexed data can be accessed via the NEMO archive here:

<https://assets.nemoarchive.org/dat-d6r90fb>. Processed data will be available on our web portal

CATLAS: <http://catlas.org>. Additional data are available upon request.

Custom code and scripts used for analysis can be accessed here:

<https://github.com/yal054/snATACutils> and <https://github.com/r3fang/SnapATAC>.

The deep learning model and pre-trained model can be download from

<https://github.com/yal054/epiformer>.

Supplementary materials

Materials and Methods

Figs. S1 to S15

Tables S1 to S29

Figure 1

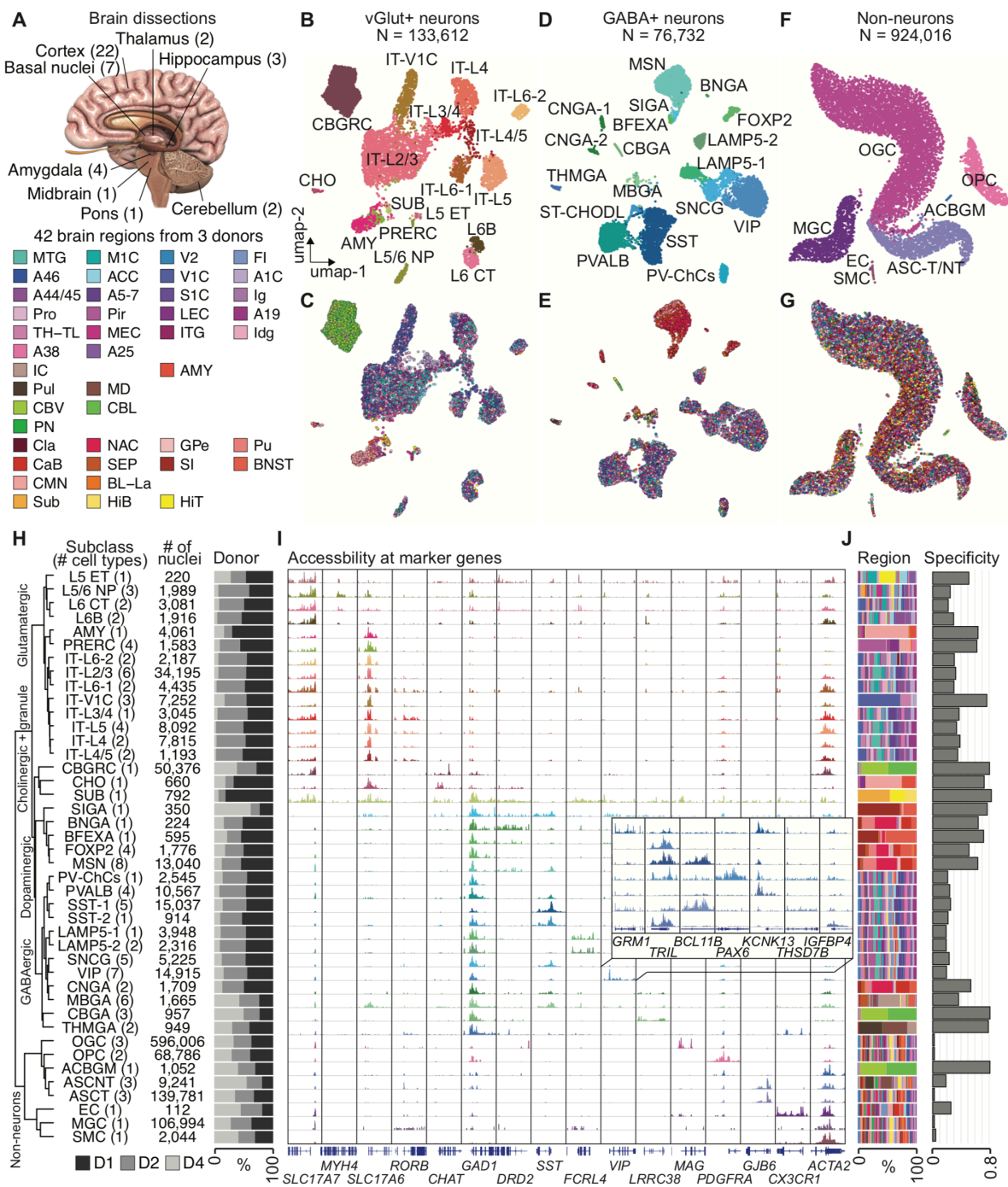
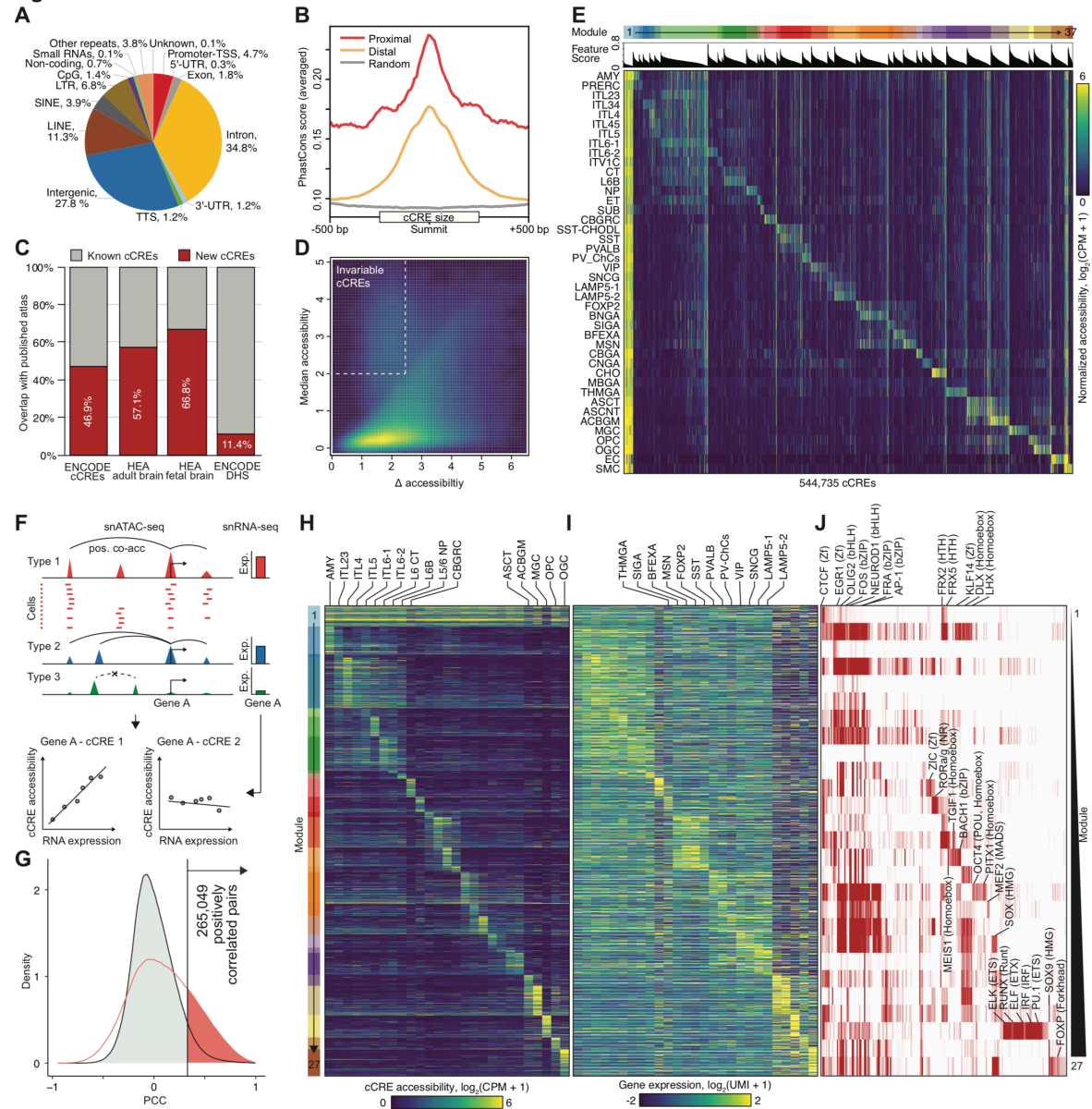


Fig. 1. Single-cell analysis of chromatin accessibility in the human brain

(A) Schematic of sampling strategy and brain dissections. A detailed list of regions is provided in **Table S1**. (B) Uniform manifold approximation and projection (UMAP) embedding and clustering analysis of glutamatergic neurons from snATAC-seq data. Individual nuclei are colored by cell subclasses. A full list and description of cluster labels are provided in **Tables S3**. (C) UMAP embedding of glutamatergic neurons, colored by brain regions. (D) UMAP embedding and clustering analysis of GABAergic neurons, colored by cell subclasses. (E) UMAP embedding of GABAergic neurons, colored by brain regions. (F) UMAP embedding and clustering analysis of non- neurons, colored by cell subclasses. (G) UMAP embedding of non-neurons, colored by brain regions. (H) Left, hierarchical organization of 42 cell subclasses on chromatin accessibility. Middle, the number of nuclei in each subclass. Right, Bar chart representing the relative contribution of 3 donors to each cell subclass. (I) Genome browser tracks of aggregate chromatin accessibility profiles for each subclass at selected marker gene loci that were used for cell cluster annotation. A full list and description of subclass annotations are in **Table S4**. (J) Left, bar chart representing the relative contribution of brain regions to cell subclasses. Right, regional specificity scores of cell subclasses.

Figure 2



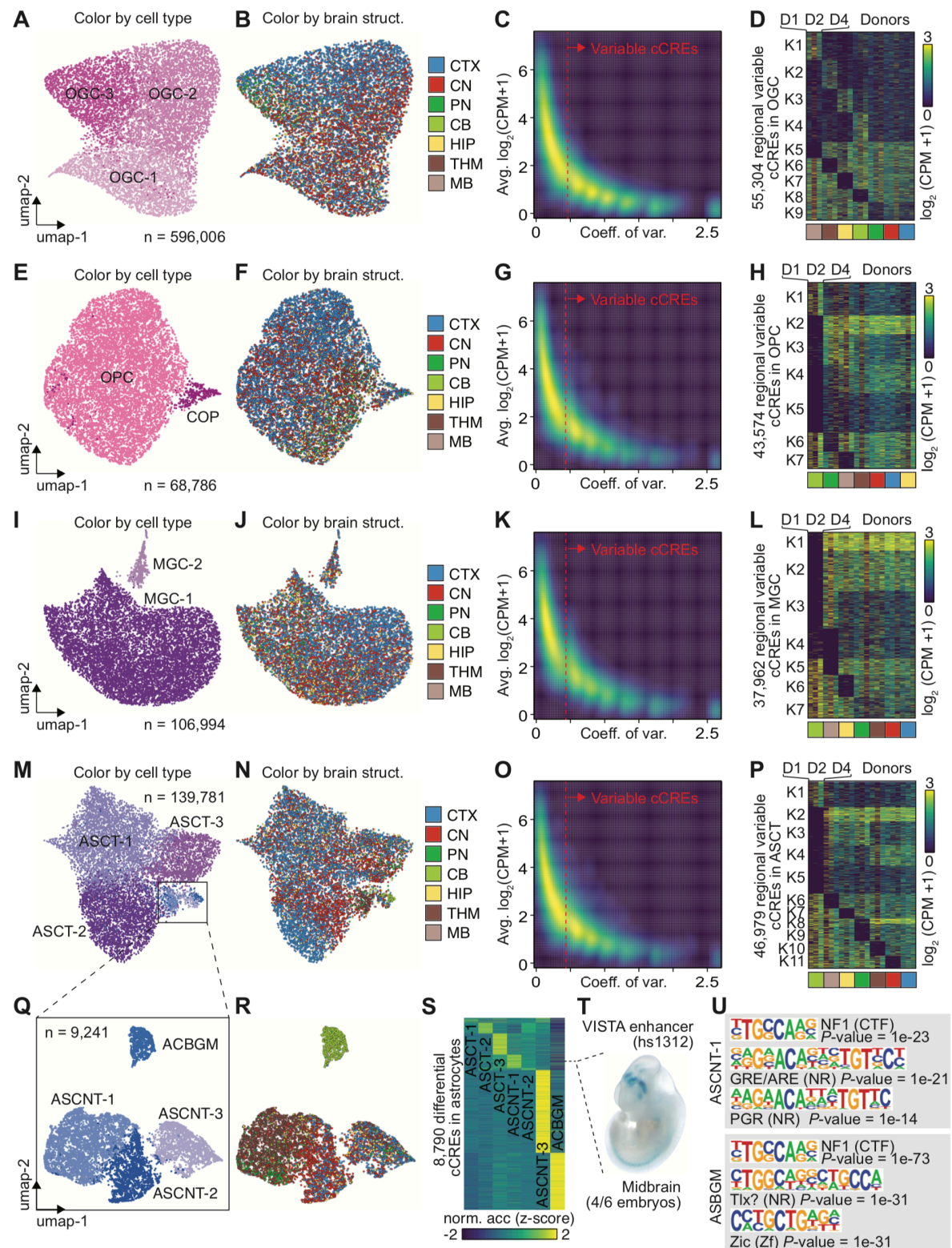
1

2

Fig. 2: Identification and characterization of candidate CREs (cCREs) across human brain cell types

(A) Pie chart showing the fraction of cCREs that overlaps with different classes of annotated sequences in the human genome. TSS, transcription start site; TTS, transcription termination site. UTR, untranslated region. LINE, long interspersed nuclear element. SINE, short interspersed nuclear element. LTR, long terminal repeats. (B) Average phastCons conservation scores of proximal (in red) and distal cCREs (in yellow), and random genomic background is indicated in gray. (C) Stacked bar plot showing the percentage of new cCREs defined in this study (in red) and percentage of cCREs that overlapped with public recourse (in grey), including the cCREs and DHSs in the SCREEN database, cCREs identify in human enhancer atlas (HEA) fetal and adult brain. (D) Density map comparing the median and maximum variation of chromatin accessibility at each cCRE across cell types. Each dot represents a cCRE. (E) Heat map showing association of the 42 subclasses (rows) with 37 cis-regulatory modules (top, from left to right). Columns represent cCREs. A full list of subclass or module associations is in **Table S9**, and the association of cCREs to modules is in **Table S10**. CPM, counts per million. (F) Schematic overview of the computational strategy used to identify cCREs that are positively correlated with transcription of target genes. (G) In total, 265,049 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. (H) Heat map showing chromatin accessibility of putative enhancers and (I) the expression of linked genes (right). Genes are shown for each putative enhancer separately. UMI, unique molecular identifier. (J) Enrichment of HOMER known transcription factor (TF) motifs in distinct enhancer–gene modules.

Figure 3



1

2

Fig. 3: Regional specificity of cell types correlates with chromatin accessibility.

(A) UMAP embedding of cell types of oligodendrocytes (OGCs). (B) UMAP embedding of oligodendrocytes, colored by major brain structures. (C) Density scatter plot comparing the averaged accessibility and coefficient of variation across brain structures at each cCRE. Variable cCREs for OGCs are defined on the right side of dash line. (D) Heat map showing the normalized accessibility of variable cCREs in OGCs. CPM, counts per million. (E) UMAP embedding of cell types of oligodendrocyte precursor cells (OPCs). (F) UMAP embedding of OPCs, colored by major brain structures. (G) Density scatter plot comparing the averaged accessibility and coefficient of variation across brain structures at each cCRE. Variable cCREs for OPCs are defined on the right side of dash line. (H) Heat map showing the normalized accessibility of variable cCREs in OPCs. (I) UMAP embedding of cell types of microglia (MGC). (J) UMAP embedding of MGC, colored by major brain structures. (K) Density scatter plot comparing the averaged accessibility and coefficient of variation across brain structures at each cCRE. Variable cCREs for MGC are defined on the right side of dash line. (L) Heat map showing the normalized accessibility of variable cCREs in MGC. (M) UMAP embedding of cell types of astrocytes (ASCs). (N) UMAP embedding of ASCs, colored by major brain structures. (O) Density scatter plot comparing the averaged accessibility and coefficient of variation across brain structures at each cCRE. Variable cCREs for ASCs are defined on the right side of dash line. (P) Heat map showing the normalized accessibility of variable cCREs in ASCs. (Q) UMAP embedding of non-telencephalon ASCs (ASCNTs). (R) UMAP of ASCNTs, colored by major brain structures. (S) Normalized chromatin accessibility of 8,790 cell-type-specific cCREs. (T) Representative images of transgenic mouse embryos showing LacZ reporter gene expression under the control of the indicated enhancers that overlapped the differential cCRE in S (dotted line). Images were downloaded from the VISTA database (<https://enhancer.lbl.gov>). (U) Top enriched known motifs for astrocyte cell-type-specific cCREs.

Figure 4

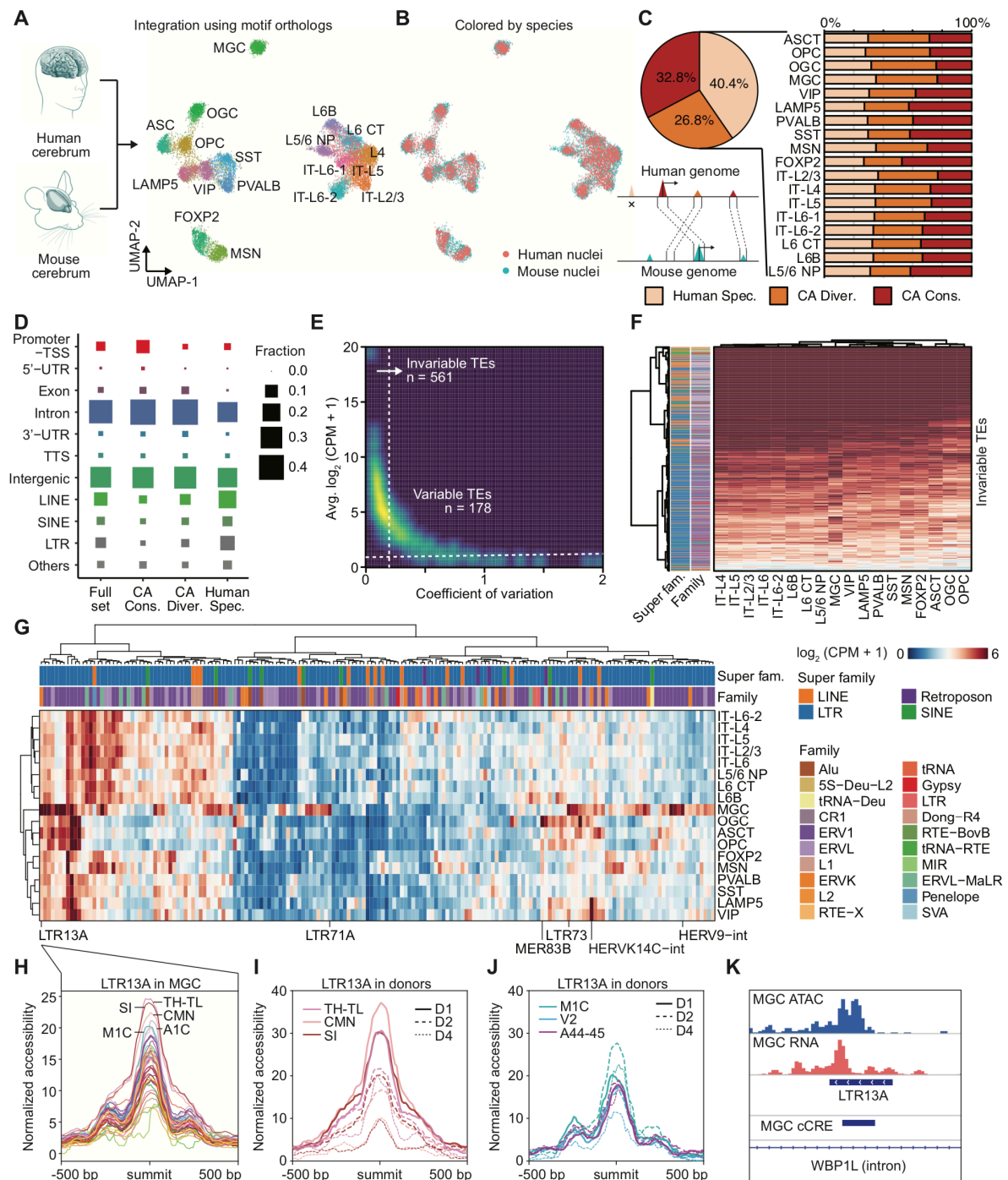
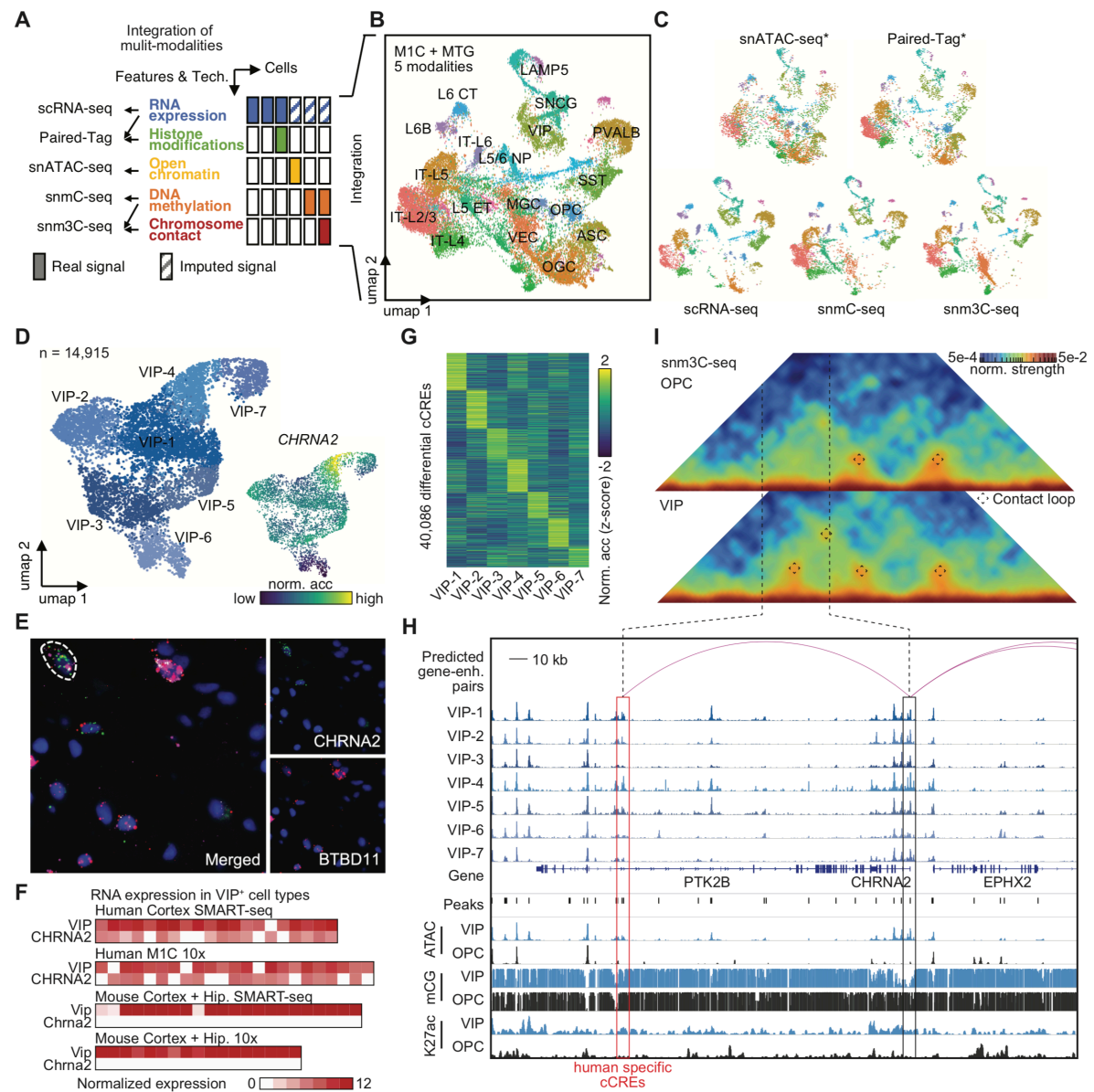


Fig. 4: Comparative analyses of chromatin accessibility between human and mouse cerebrum.

(A) UMAP co-embedding of 18 cell subclasses from both human and mouse cerebrum. (B) UMAP co-embedding of single nuclei colored by human and mouse. (C) Left, pie chart showing fraction of three categories of cCREs, including human specific, CA-divergent and CA-conserved cCREs. The CA-conserved cCREs are both DNA sequence conserved across species and have open chromatin in orthologous regions. The CA divergent cCREs are sequence conserved to orthologous regions but have not been identified as open chromatin regions in other species. Human specific cCREs are not able to find orthologous regions in the mouse genome. Right, bar plot showing three categories of cCREs in corresponding cell subclasses from human and mouse. (D) Dot plot showing fraction of genomic distribution of three categories of cCREs. (E) Density scatter plot comparing the averaged accessibility and coefficient of variation across cell subclasses at each transposon elements (TEs). Variable TEs are defined on the upper right side of dash lines, invariable TEs are defined on the upper left of dash lines. (F) Normalized accessibility at invariable TEs in different cell subclasses. CPM, counts per million. (G) Normalized accessibility at variable TEs in different cell subclasses. (H) Average chromatin accessibility of LTR13A in microglia across different brain regions. (I) Variable chromatin accessibility of LTR13A across donors in microglia from different brain regions. (J) Invariable chromatin accessibility of LTR13A across donors in microglia from different brain regions. (K) Representative genomic locus showing chromatin accessibility and expression levels on LTR13A in microglia.

Figure 5



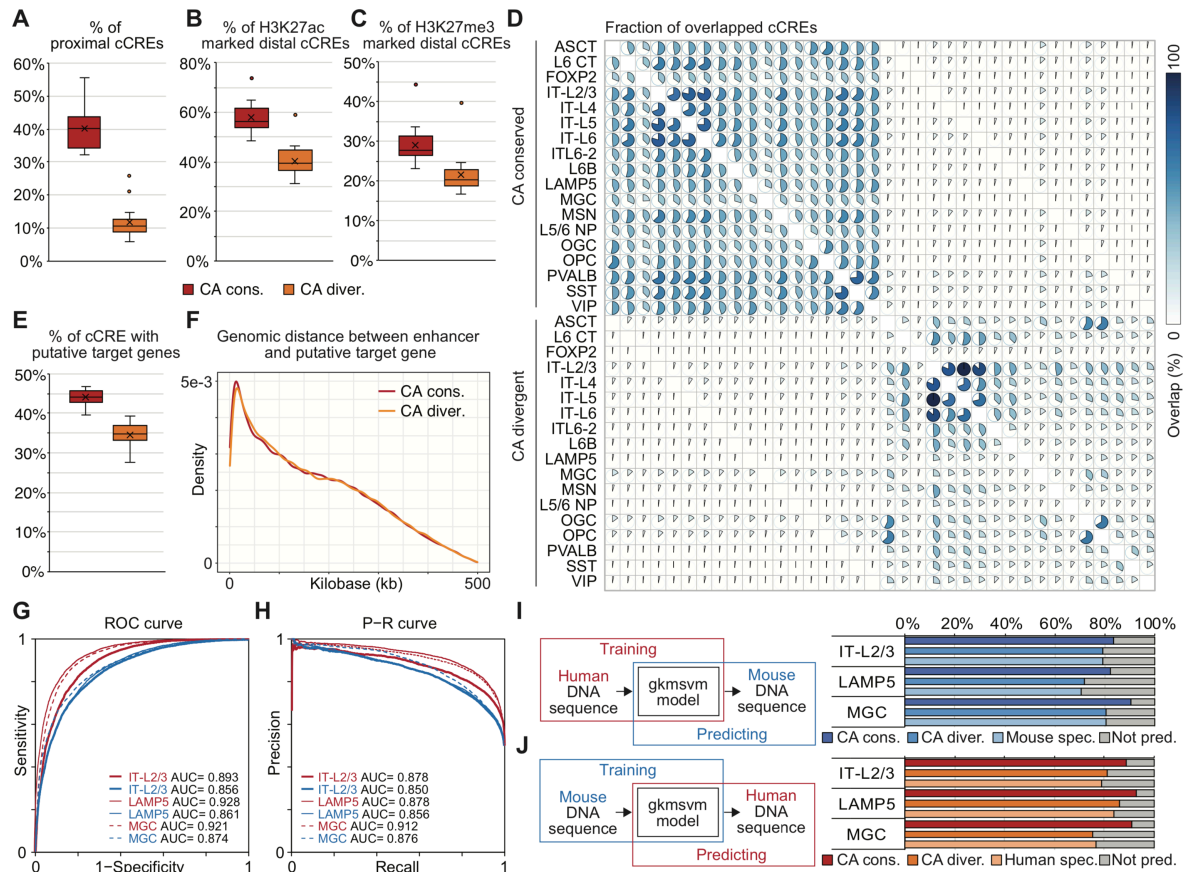
1

2

Fig. 5: Integration of multi-modal single cell datasets of cortical cells.

(A) Summary of single cell technologies and multi-model data integration strategies. (B) UMAP embedding and integrative clustering analysis of 18 major cell types. (C) Co-embedding of multi-model single cell datasets showing excellent agreement. * for assays using unbiased sampling strategy. (D) Left, UMAP embedding of VIP positive (VIP⁺) GABAergic cell types. Upper right, colored by donors. Bottom right, normalized accessibility at gene *CHRNA2*. (E) RNAscope validation of *CHRNA2*-lineage of VIP⁺ GABAergic neurons. (F) Expression of gene *VIP* and *CHRNA2* in human VIP⁺ cell types, and expression of gene *Vip* and *Chrna2* in mouse VIP⁺ cell types from Allen Cell Types Database: RNA-Seq Data. (G) Normalized chromatin accessibility of 40,086 VIP⁺ cell-type-specific cCREs. (H) Genome browser track view at the *CHRNA2* locus as an example for candidate enhancers predicted from single cell multi-model datasets. Displayed chromatin accessibility profiles from snATAC-seq; DNA methylation signals (mCG) from snm3C-seq; and histone modification signals (H3K27ac) from Paired-Tag for several VIP⁺ neurons and oligodendrocytes precursor cells (OPC). Red Arcs represent the predicted enhancer for gene *CHRNA2*. (I) Triangle heat map show chromatin contacts in VIP⁺ neurons and OPC derived from snm3C-seq data at gene *CHRNA2* locus.

Figure 6



1

2

3

Fig. 6: Epigenetic conservation and divergence of human orthologous cCREs

(A) The boxplots showing percentage of proximal cCREs in CA-conserved and CA-divergent cCREs from various cell subclasses. (B) The boxplots showing percentage of H3K27ac marked distal cCREs in CA-conserved and CA-divergent cCREs from various cell subclasses. (C) The boxplots showing percentage of H3K27me3 marked distal cCREs in CA-conserved and CA-divergent cCREs from various cell subclasses. (D) Fraction of overlap between CA-conserved and CA-divergent cCREs defined in various cell subclasses. (E) The boxplots showing percentage of distal cCREs from various cell subclasses are predicted to regulate putative target genes in CA-conserved and CA-divergent cCREs (Fig. 2). (F) The distribution of 1-D genomic distance between epigenetic conserved/divergent enhancers and putative target genes. (G) Receiver operating characteristic (ROC) curve and area under curve (AUC) from gkmsvm models trained for representative human and mouse cell types. (H) Precision-recall curve (PRC) curve and area under curve (AUC) from gkmsvm models trained for representative human and mouse cell types. (I) Prediction for mouse epigenetic conserved, mouse CA-divergent, and mouse specific cCREs from gkmsvm models trained in corresponding human cell subclasses. (J) Prediction for human epigenetic conserved, human CA-divergent, and human specific cCREs from gkmsvm models trained in corresponding mouse cell subclasses.

Figure 7

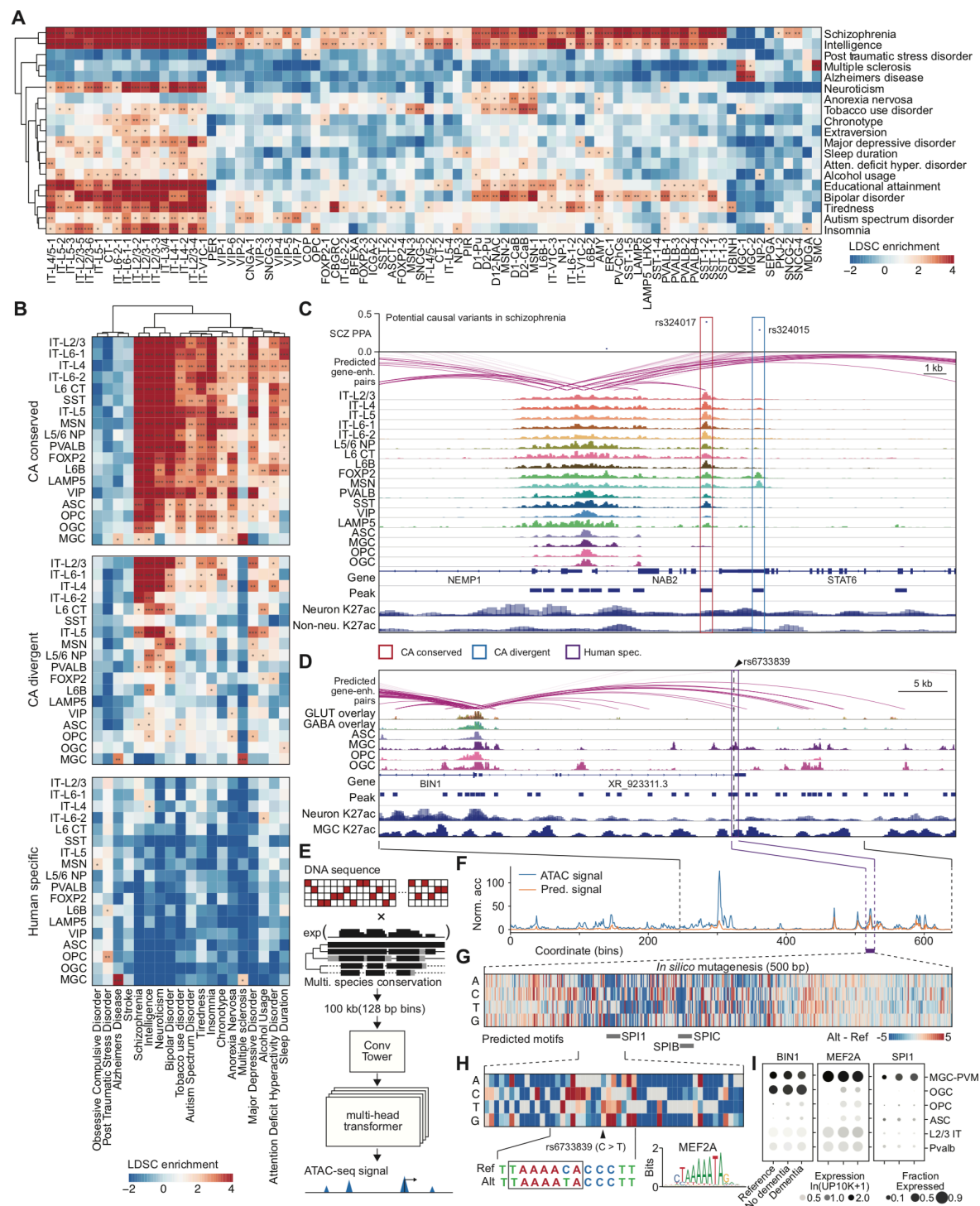


Fig. 7: Interpreting noncoding risk variant of neurological disorder and traits.

(A) Heat map showing enrichment of risk variants associated with neurological disorder and traits from genome wide association studies in human cell type-resolved cCREs. Cell-type specific linkage disequilibrium score regression (LDSC) analysis was performed using GWAS summary statistics. Total cCREs identified independently from each human cell type were used as input for analysis. *P*-values were corrected using the Benjamini Hochberg procedure for multiple tests. FDRs of LDSC coefficient are displayed. *, FDR < 0.05; **, FDR < 0.01; ***, FDR<0.001. Detailed results are reported in **Table S26**. (B) Heat map showing enrichment of risk variants associated with mental disorder and traits in three categories of cCREs. Detailed results are reported in **Table S27**. (C) Fine mapping and molecular characterization of schizophrenia (SCZ) risk variants in different categories of cCREs from multiple neuronal types. Genome browser tracks (GRCh38) display chromatin accessibility profiles from snATAC-seq; histone modification signals (H3K27ac) from Paired-Tag, and red arcs represent the predicted enhancer for gene *NAB2*. (D) Molecular characterization of Alzheimer's disease (AD) risk variants in human specific enhancer in microglia. Genome browser tracks (GRCh38) display chromatin accessibility profiles from snATAC-seq; histone modification signals (H3K27ac) from Paired-Tag, and red arcs represent the predicted enhancer for gene *BIN1*. PPA: Posterior probability of association. (E) Schematic diagram of deep learning model for predicting chromatin accessibility. (F) Chromatin accessibility at *BIN1* enhancer loci predicted in human microglia. (G) *In silico* nucleotide mutagenesis influenced the prediction of accessibility. Larger signals (in dark red) represent a higher accessibility prediction on altered sequence, and lower signals (in dark blue) represent lower accessibility on altered sequence. (H) Higher accessibility predicted on *BIN1* enhancer with risk variant rs6733839 C>T. The risk variant rs6733839 creates a TF motif for MEF2 family. (I) Gene expression level of *BIN1*, *MEF2A* and *SP11* in patients with Alzheimer's disease collected from Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD).

Materials and Methods

Tissue preparation and nuclei isolation

Tissues from 3 human donors (males aged 29, 42, and 58, R hemisphere) with post-mortem interval <12 hours and RIN score >7.5 were obtained from the Allen Institute. Tissue included 42 regions spanning diverse brain structures, dissected using standard architectonic landmarks and guided by the Allen Brain Human Brain Atlas(39). Tissue blocks were separated into 100-150mg chunks in RNase-free glass dishes on dry ice and stored at -80°C until homogenized.

Frozen tissue was homogenized in 4.5 mL Lysis Buffer (0.32 M sucrose, 0.1% Triton X-100, 5 mM CaCl₂, 3 mM Mg(Ace)₂, 0.1 mM EDTA pH 8.0, 10 mM Tris-HCl pH 8.0, 1 mM DTT, Roche cOmplete Mini Protease Inhibitor, 24 U/mL Promega RNasin Plus) using glass Dounce Homogenizers. For iodixanol-gradient purification of nuclei, tissue was Dounce homogenized in 5 mL NIMT Buffer (0.25 M sucrose, 0.1% Triton X-100, 25 mM KCl, 5 mM MgCl₂, 10 mM Tris-Cl pH 8.0, 1 mM DTT, 20 U/mL SUPERase IN, 40 U/mL RNase OUT), mixed with 50% iodixanol buffer (50% iodixanol (Optiprep), 25 mM KCl, 5 mM MgCl₂, 20 mM Tris-HCl pH 8.0) to a concentration of 20% iodixanol, and layered onto 25% iodixanol buffer (25% iodixanol (Optiprep), 0.125 M sucrose, 12.5 mM KCl, 2.5 mM MgCl₂, 5 mM Tris-Cl pH 8.0). Samples were centrifuged at 10,000 x g for 20 min at 4°C in a Beckman Optima-80K Ultracentrifuge with SW-41-Ti swinging bucket rotor. The supernatant was carefully aspirated before nuclei rehydration. All steps were performed on ice using nuclease-free equipment.

Single nucleus ATAC-seq using combinatorial indexing

Single nucleus ATAC-seq was performed as described with steps optimized for automation and modifications in permeabilization and sort buffers (34). A step-by-step-protocol for library preparation is available here: <https://www.protocols.io/view/sequencing-open-chromatin-of-single-cell-nuclei-sn-pjudknw/abstract>.

Brain nuclei were pelleted with a swinging bucket centrifuge (500 x g, 5 min, 4°C; 5920R, Eppendorf). Nuclei pellets were resuspended in 1 ml nuclei permeabilization buffer (10mM Tris-HCL (pH 7.5), 10mM NaCl, 3mM MgCl₂, 0.1% Tween-20 (Sigma), 0.1% IGEPAL-CA630 (Sigma), 0.01% Digitonin (Promega) and 1% BSA (Proliant 7500804) in molecular biology-grade water) and pelleted again (500 x g, 5 min, 4°C; 5920R, Eppendorf). Nuclei were resuspended in 1 mL high salt tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM potassium-acetate, 11 mM Mg-acetate, 17.6% DMF) and counted using a hemocytometer. Concentration was adjusted to 2,000 nuclei/9 µl, and 2,000 nuclei were dispensed into each well of a 96-well plate. For tagmentation, 1 µL barcoded Tn5 transposomes²⁶ 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 Tn5 reaction, 10 µL of 40 mM EDTA was added to each well with a BenchSmart™ 96 (Mettler Toledo, RRID:SCR_018093), mixed 10 times, and the plate was incubated at 37 °C for 15 min with shaking (500 rpm). Next, 10 µL 3 x sort buffer (3 % BSA, 3 mM EDTA in PBS) was added using a BenchSmart™ 96 (Mettler Toledo, RRID:SCR_018093). If processing two or more 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, added to a FACS tube and stained with 3 µM Draq7 (Cell Signaling). Using a SH800 (Sony), 20 nuclei per sample were sorted per well into eight 96-well plates (total of 768 wells, 15,360 nuclei per sample) containing 10.5 µL EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma). Preparation of sort plates and all

downstream pipetting steps were performed on a Biomek i7 Automated Workstation (Beckman Coulter, RRID:SCR_018094). After addition of 1 μ L 0.2% SDS, samples were incubated at 55 °C for 7 min with shaking (500 rpm). 1 μ L 12.5% Triton-X was added to each well to quench the SDS. Next, 12.5 μ L NEBNext High-Fidelity 2 $\times$ PCR Master Mix (NEB) were added and samples were PCR-amplified (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72 °C 60 s) $\times$ 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, 0.55x and 1.5x). Libraries were purified one more time with SPRI Beads (Beckmann Coulter, 1.5x). 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 were sequenced on a NextSeq500 (RRID:SCR_014983, Illumina), HiSeq4000 (RRID:SCR_016386, Illumina) or NovaSeq6000 (RRID:SCR_016387, Illumina) using custom sequencing primers with following read lengths: 50 + 10 + 12 + 50 (Read1 + Index1 + Index2 + Read2). Indexing primers and sequencing primers are in **Table S29**.

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(76). After alignment, we used the R package ATACseqQC (1.10.2)(77) 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 <1000 bp. 3) PCR duplicates were further removed with SnapTools (<https://github.com/r3fang/SnapTools>)(78). Reads were sorted based on 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 v29 (hg38)(79). Briefly, 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 $\pm 2,000$ bp relative (TSS strand-corrected) to each unique TSS genome-wide. Then this profile was normalized to the mean accessibility $\pm 1,900$ -2,000 bp from the TSS and smoothed every 11bp. The max of the smoothed profile was taken as the TSS enrichment.

Doublet removal

We used a modified version of Scrublet(80) to remove potential doublets for every dataset independently. First, we count the reads on gene body and 2 kb upstream of TSS for every single nucleus that passed quality control. Next, cell-by-gene count matrices were used as input, with default parameters, the doublet scores were calculated for both observed nuclei $\{x_i\}$ and simulated doublets $\{y_i\}$ using Scrublet(80). Then, a threshold θ is selected based on the distribution of $\{y_i\}$, and observed nuclei with doublet score larger than θ are predicted as doublets. To determine θ , we fit a two-component mixture distribution by using function normalmixEM from R package mixtools(81). The lower component contained majority of embedded doublet types, and the other component contained majority of neo-typic doublets (collision

between nuclei from different clusters. We selected the threshold θ where the $p_1 \cdot pdf(x, \mu_1, \sigma_1) = p_2 \cdot pdf(x, \mu_2, \sigma_2)$. This value suggested that the nuclei have same chance of belonging to both classes.

Clustering and cluster annotation

We used an iterative clustering strategy using the SnapATAC package(78) with 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 types. Last, for each major cell types, we performed a third round of clustering to find sub-types.

Detailed description for every step is listed below:

1) Nuclei filtering

Nuclei with $\geq 500/1,000$ uniquely mapped fragments and TSS (transcription start site) enrichment $> 5/7$ were filtered for individual dataset.

2) Doublet removal:

The potential barcode collisions were also removed for individual datasets.

3) Feature bin selection

First, we calculated a cell-by-bin matrix at 500 kb resolution for every sample 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 (hg38, <https://github.com/Boyle-Lab/Blacklist>)(82). Fourth, we focused on bins on chromosomes 1-22, X, and Y. Last, we removed the top 5% bins with the highest read coverage from the count matrix.

4) Dimensionality reduction

SnapATAC applies a nonlinear dimensionality reduction method called diffusion maps, which is highly robust to noise and perturbation. However, the computational time of the diffusion maps algorithm scales exponentially with the increase of number of cells. To overcome this limitation, we used adjacency spectral embedding (<https://github.com/yixuan/RSPSpectra>) instead of diffusion maps, which can easily scale up to tens of thousands of nuclei without memory issue. We combined the Nyström method (a sampling technique)⁶⁸ to generate the low-dimensional embedding for large-scale dataset.

A Nyström landmark diffusion maps algorithm includes three major steps:

1. sampling: sample a subset of K ($K \ll N$) cells from N total cells as “landmarks”.
2. embedding: compute a diffusion map embedding for K landmarks;
3. extension: project 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 1.1 million single nuclei at the beginning, we decided to apply this strategy on level 1 and 2 clustering. 35,000 cells were sampled as landmarks and the remaining query cells were projected onto the diffusion maps embedding of landmarks. Later for the level III clustering, diffusion map embeddings were directly calculated from all nuclei.

5) Batch correction

To correct the potential bias from donors and technical replicates, we applied batch correction on these variables using R package “Harmony”(61).

6) 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 based on the percentage of variance explained by each

one. For each round of clustering, we selected the top 10-20 eigenvectors that captured the majority of the variance.

7) Generate K Nearest Neighbor (KNN) Graph

Using the selected significant eigenvectors, we next construct a K Nearest Neighbor (KNN) Graph. Each cell is a node and the k-nearest neighbors of each cell were identified according to the Euclidian distance and edges were drawn between neighbors in the graph. To overcome the limitation on CPU memory and speed on the process, we applied Annoy (Approximate Nearest Neighbors Oh Yeah) instead of the original KNN function(60).

8) Graph-based clustering

We applied the Leiden algorithm on the KNN graph using python package leidenalg (<https://github.com/vtraag/leidenalg>)(83).

9) 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 if 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 100 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:

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

Let I^s be the $N \times N$ indicator matrix where 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 .

$$C(i, j) = \frac{\left(\sum_{s=1}^S M^s(i, j) \right)}{\left(\sum_{s=1}^S I^s(i, j) \right)}$$

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 (DC), 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 DC, the more perfect is consensus matrix, and thus the more stable is the clustering. In a perfect consensus matrix, all entries are 0 or 1, meaning that all connectivity matrices are identical. The DC is defined as:

$$\rho = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n 4 * \left(M(i, j) - \frac{1}{2} \right)^2.$$

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’.

9) Visualization

For visualization, we applied Uniform Manifold Approximation and Projection (UMAP)(84).

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 variation (CV) for the cluster centroid of each element across major cell types. Finally, we only kept variable elements with CV larger than 10% quantile and smaller than 90% quantile 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 `pvcust` (v.2.0) (85) 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.

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 sub-class, the contribution of different brain regions is first calculated. Then, we compared this distribution with the contribution of brain regions calculated from all sampled cells. We used function ‘JSD’ from R package `philentropy` for this analysis(86).

Identification of reproducible peak sets in each cell cluster

We performed peak calling according to the ENCODE ATAC-seq pipeline (<https://www.encodeproject.org/atac-seq/>). For every cell cluster, we combined all properly paired reads to generate a pseudo-bulk ATAC-seq dataset for individual biological replicates. In addition, we generated two pseudo-replicates which comprise half of the reads from each biological replicate. We called peak for each of the four datasets and a pool of both replicates independently. Peak calling was performed on the Tn5-corrected single-base insertions using the MACS2(87) with these parameters: `--shift -75 --extsize 150 --nomodel --call-summits --SPMR -q 0.01`. Finally, we extended peak summits by 250 bp on either side to a final width of 501 bp for merging and downstream analysis. To generate a list of reproducible peaks, we kept peaks that 1) were detected in the pooled dataset and overlapped $\geq 50\%$ of peak length with a peak in both individual replicates or 2) were detected in the pooled dataset and overlapped $\geq 50\%$ of peak length with a peak in both pseudo-replicates.

We found that when cell population varied in read depth or number of nuclei, the MACS2 score varied proportionally due to the nature of the Poisson distribution test in MACS2(87). Ideally, we should perform a reads-in-peaks normalization, but in practice, this type of normalization is not possible because we don’t know how many peaks we will get. To account for differences in performance of MACS2(87) based on read depth and/or number of nuclei in individual clusters, we converted MACS2 peak scores ($-\log_{10}(q\text{-value})$) to “score per million”⁷². We filtered reproducible peaks by choosing a “score per million” cut-off of 2 was used to filter reproducible peaks.

We only kept reproducible peaks on chromosome 1-19 and both sex chromosomes, and filtered ENCODE blacklist (hg38, <https://github.com/Boyle-Lab/Blacklist>)(82). A union peak list for the whole dataset was obtained by merging peak sets from all cell clusters using BEDtools(88).

Computing chromatin accessibility scores

Accessibility of cCREs in individual clusters was quantified by counting the fragments in individual clusters normalized by read depth (counts per million: CPM). For each gene, we summed up the counts within the gene body + 2kb upstream to calculate “gene activity score (GAS)” GAS was used for integrative analysis with other single cell modalities. For better visualization, we smoothed GAS to 50 nearest neighbour nuclei using Markov Affinity-based Graph Imputation of Cells (MAGIC)(89).

Integrative analysis of single nucleus ATAC-seq and single cell datasets

For integrative analysis, we first filtered brain regions that matched samples profiled in this study (same brain dissections were used in companion papers). Second, we manually subset cell types into two groups, neurons and non-neurons, by checking both signal on marker genes and taxonomy labels from scRNA-seq, snmC-seq, and snm3C-seq. To directly compare our single nucleus chromatin accessibility derived cell clusters with the single cell transcriptomics defined taxonomy of the human brain, we first used the snATAC-seq data to impute RNA expression levels according to the chromatin accessibility of gene promoter and gene body as described previously. For single cell methylome, snmC-seq and snm3C-seq, the signal on gene body was negatively correlated with the gene expression. We imputed the RNA expression levels from methylation signals by subtracting it from 1.

Next, we used a 3-step method analogous to Seurat v3 to project three modalities onto the same space: 1) Using canonical correlation analysis (CCA) to capture the shared variance across cells between datasets; 2) finding anchors as 50 mutual nearest neighbors (MNN) between the two datasets; 3) pulling the three modalities into the same space. Detailed description can be found in companion paper (Wei, et al. 2022). To quantify the similarity between cell clusters from two modalities, we calculated an overlapping score as the sum of the minimum proportion of cells/nuclei in each cluster that overlapped within each co-embedding cluster(90). 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, for example scRNA-seq and snATAC-seq, which was evidenced by co-embedding of both transcriptomic (T-type) and chromatin accessibility (A-type) cells in the same joint clusters, as well as co-embed with methylation (M-type) cells (**Fig. S6**). For this analysis, we examined neurons and non-neuronal cell classes separately, due to the CG methylation was more informative for non-neurons comparing to CH methylation.

Identification of *cis* regulatory modules

We used Nonnegative Matrix Factorization (NMF)(91) to group cCREs into *cis* regulatory modules based on their relative accessibility across major clusters. We adapted NMF (Python package: sklearn(92) 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 ($N \times R$), with a given rank R :

$$V \approx WH,$$

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 (i.e., 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”(93) and “Entropy”(94) to evaluate the clustering result. Average values were calculated from 100 times for NMF 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 (~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(94). 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 following criteria were retained as module specific elements:

1. feature score greater than median + 2 standard deviations;
2. the maximum contribution to a basis component is greater than the median of all contributions (i.e. of all elements of W).

Identification of differentially accessible regions and definition of specificity score

To identify cCRE that differentially accessible in either in cell cluster or brain region, 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 (v0.2.2)(95) to perform the differential test. We designed the full model as

$$\text{logit}(P_{ij}) = a_j + m_j + d_j + \varepsilon_j$$

and a reduced mode as

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

Where 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 j^{th} cell, m is membership of the j^{th} cell in either cluster or region being tested, d is the donor label for j^{th} cell and ε is an error term.

For each set of testing, between cell type or between regions in every cell sub-class, we only kept cCREs that overlapped with peaks identified in corresponding cell types. A likelihood ratio test is then used to determine if the full model (including cell cluster or region membership) provided a significantly better fit of the data than reduced model. After correcting p -value using Benjamini-Hochberg method, we set an FDR cutoff as 0.001 to filter out significant differential cCREs.

$\log_2$ fold change is used for two-group comparison, for multiple groups, we calculated a Jensen Shannon divergence-based specificity score described in previous study,²² in order to better assign differential cCREs to cell type or brain region. The fraction of accessibility of each cluster f was first calculated for each i^{th} site. We normalized these scores by multiplying by corresponding scaling factors, which are considering different overall complexity across groups. To do so, median number of sites accessible c in individual cells for each cluster was calculated and followed with $\log_{10}$ -transforming. Then, we took the ratio of the average of c (across all clusters) over the median accessibility in each cluster as scaling factor. These corrected fraction of accessibility for 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 1st cCRE as *d1*, to test whether this cCRE is specific in group1, we assumed another probability distribution:

$$d2 = \begin{cases} 1, & \text{group 1} \\ 0, & \text{otherwise} \end{cases}$$

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

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

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

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

Predicting enhancer-promoter interactions

First, co-accessible regions are identified for all open regions in each cell cluster (randomly selected 200 nuclei, and used all nuclei for cell cluster with <200 nuclei) separately, using Cicero(47) with following parameters: aggregation *k* = 10, window size = 500 kb, distance constraint = 250 kb. In order 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 R package *fitdistrplus*(96). 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.

The cCREs outside ± 2 kb of transcriptional start sites (TSS) in GENCODE hg38 (v29)(97) 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 calculated Pearson's correlation coefficient (PCC) between gene expression and cCRE accessibility across matched T-type and A-type clusters to examine the relationship between co-accessibility pairs. To do so, we first aggregated all nuclei/cells from scRNA-seq and snATAC-seq for every joint cluster to calculate accessibility scores ($\log_2$ CPM) and relative expression levels ($\log_2$ normalized UMI). Then, PCC was calculated for every gene-cCRE pair within a 1 Mbp window centered 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, in order to select significant positively correlated cCRE-gene pairs.

RNAscope fluorescence *in situ* hybridization

- 1) Tissue preparation: All tissues were cryostat sectioned at 14 μ m onto SuperFrost Plus charged slides and allowed to adhere to the slide. The slides were then stored at -80 C until further use.
- 2) RNAscope in situ hybridization multiplex version 2 was performed as per the protocol established by Advanced Cell Diagnostics (ACD) with minor modifications. In brief, slides were fixed using cold (4°C)

4% PFA for 15 minutes. The samples were then dehydrated in 50% ethanol (5 min), 70% ethanol (5 min) and 100% ethanol (5 min, twice) at room temperature. The slides were air dried for 10 min, after which boundaries were drawn around each section using a hydrophobic pen (ImmEdge PAP pen; Vector Labs). After the hydrophobic boundaries dried, sections were incubated with hydrogen peroxide (ACD, ~5 drops per section) for 10 MIN at RT. Sections were then washed twice with PBS for 2 MIN at RT and, subsequently, incubated with 4-5 drops of RNAscope® Protease IV (15 min) at RT. Slides were washed twice in 1X phosphate buffered saline (PBS, pH 7.4) at room temperature (2 min each). Each slide was then placed in a prewarmed humidity control tray (ACD) containing dampened filter paper, and a mixture of Channel probes (50:1:1 dilution, as per ACD instructions) was pipetted onto each section until fully covered. The humidity control tray was placed in a HybEZ oven (ACD) for 2 hours at 40°C. Following probe incubation, the slides were washed twice in 1X RNAscope wash buffer and stored overnight in 5X SSC (Sigma-Aldrich) at RT. The following day slides were returned to the oven for 30 minutes after submersion in AMP-1 reagent (~5 drops). Two washes and amplification were repeated using AMP reagents with a 30-min, 30-min, and 15-min incubation period, respectively. HRP-C1 (ACD, ~5 drops) was added to the slide and incubated in the HybEZ oven at 40°C for 15 min. Two washes with 1X wash buffer (2 min each) were performed and 150 µL pre-assigned Opal dye added (Opal Dyes; Akoya Biosciences, Opal Dye diluted 1:1500 in TSA buffer). The slides were incubated in the oven for 30 minutes before being washed twice in 1X wash buffer. Slides were then covered in HRP blocker (ACD, ~5 drops) and incubated for 15 minutes before being washed twice. Slides were then washed two times in 0.1M phosphate buffer (PBS, pH=7.4) and incubated with DAPI (ACD, ~5 drops) for 30 seconds before being washed in PBS, air dried, and cover-slipped with Prolong Gold Antifade mounting medium. Slides were stored at 4°C until imaging.

3) Image quantification: Fluorescence signals were acquired using a Zeiss AiryScan confocal laser scanning microscope equipped with a Plan-APOCHROMAT 63× oil-immersion lens (N/A 1.4). Six fields of view (three per layer) of one tissue section were analyzed using Imaris (Bitplane). Prior to quantification, regions containing lipofuscin signals (defined by overlapping non-specific signals appearing in both channels) were subtracted from each image. Then, positive-stained nuclear signals, defined by automated surface detection function, were segmented, and classified into two groups depending on whether they expressed Gad2. Finally, Chrna2 expression within those two groups was quantified.

Paired-Tag experimental procedure

Paired-Tag experiment was performed as previously described(98) with minor modifications detailed below. First, antibodies against H3K27ac (Abcam, ab4729) or H3K27me3 (Abcam, ab195477) were pre-incubated with protein A-Tn5 fusion protein, and then incubated with permeabilized nuclei overnight, to target the binding of protein A-fused Tn5 to chromatin. Tagmentation reaction and reverse transcription (RT) were then sequentially performed. Reactions were carried out in 12 different wells, each well containing roughly 2M nuclei and a well-specific DNA barcode included in the Tn5 transposase adaptors and RT primers, to label different samples or replicates. Next, a ligation-based combinatorial barcoding strategy was used to introduce the 2nd and 3rd rounds of DNA barcodes to the nuclei, by sequentially attaching well-specific DNA barcodes to the 5'-end of both chromatin DNA fragments and cDNA from RT in 4 x 96-well plates in each round. Finally, the barcoded nuclei were divided into sub-libraries of 25,000 nuclei each and lysed with Proteinase K (NEB, #P8102). The chromatin DNA and cDNA were subsequently purified as a whole with SPRI beads (Beckman Coulter), pre-amplified, and splitted into two parts for RNA and DNA libraries. For the RNA part, libraries were constructed using the Nextera XT DNA Library

Preparation Kit (Illumina, FC-131-1024). For the DNA part, an adaptor containing Truseq i5 sequencing primer binding site was ligated to the library fragments, followed by DNA cleanup using SPRI beads and indexing PCR. Final libraries were quantified using qPCR, and sequenced on a NovaSeq 6000 platform (Illumina).

Paired-Tag data processing

A detailed, step-by-step Paired-Tag data processing pipeline can be found at: <https://github.com/cxzhu/Paired-Tag>.

1) Preprocessing: Cellular barcodes from the sequencing reads are first extracted by matching the linker sequences adjacent to the cellular barcodes, which are then mapped to the cellular barcodes reference using bowtie(99), with reads with more than 1 nt mismatch discarded. The adaptor sequences are trimmed from 3' of DNA and RNA libraries, with Poly-dT and random hexamer sequences further trimmed from 3' of RNA libraries. The low-quality reads (L = 30, Q = 30) are excluded from further analysis. DNA reads are mapped to hg38 using bowtie2(100) with default parameters. Reads with MAPQ<10 are removed. PCR duplicates are also removed.

2) Cell clustering: RNA alignment files are converted to a matrix with cells as columns and genes as rows. Cells with less than 200 features in both DNA and RNA matrices are removed. The clustering of single-cells based on RNA-profiles was performed with Seurat package(60). Briefly, cell-to-gene counts are normalized, and the variable genes are selected for dimension reduction by PCA. Batch effects are corrected with Harmony(61), and single cell gene expression profiles are visualized with UMAP and clustered with Louvain algorithm(83).

3) Cell-type-specific histone modification maps: After clustering, the DNA component of Paired-Tag data corresponding to each histone mark in each cell cluster is aggregated to generate profiles of histone modification for the cluster. The aggregated aligned reads were then used to generate genome browser tracks to reveal the landscape of each histone modification in each cell type as defined by the RNA modality from the Paired-Tag data.

4) To annotate each cCRE, we counted sequencing reads on cCREs captured from Paired-Tag experiments on two histone modifications, and fit a Poisson distribution to filter out cCREs that significantly (FDR<0.05) marked by H3K27ac and H3K27me3.

Integration analyses of chromatin accessibility between human and mouse brain cell types

To better compare the open chromatin landscape within a similar cellular context, I extracted 18 major neuronal and glial cell types from the mouse cerebrum (40) for alignment with human snATAC-seq from this study. We first calculated chromatin accessibility at orthologous genes and cCREs, and enrichment of orthologous motifs from pseudo-bulk profiles in each cell sub-classes. Then, we performed multidimensional scaling analyses and project cell classes on two dimensional spaces.

To integrate at single cell level, we randomly selected 1000 nuclei from each cell sub-class in two species. Similarly, we first calculated fragment counts on orthologous genes and cCREs, and variability of orthologous motifs using chromVAR(101). For the integration analyses using orthologous genes and motifs as features, we generated Seurat objects in R. Next, variable features were identified and used for identifying anchors between cells from two species. Finally, to visualize all the cells together, we co-embedded the cell profiles from two species in the same low dimensional space. For the integration analysis

using orthologous cCREs, we performed by using the SnapATAC package(78) as described above. Of note, we did not perform any correction for potential bias from species.

Deep learning model Epiformer

The deep learning model takes both one-hot-encoded DNA sequences ($A = [1, 0, 0, 0]$, $C = [0, 1, 0, 0]$, $G = [0, 0, 1, 0]$, $T = [0, 0, 0, 1]$) and conservation scores (phastCons, range from 0 to 1) calculated from multiple species alignment as inputs. We first divide the human genome (hg38) into segments in length of 98,304 bp. Any genomic segments overlapped with human ENCODE “blacklist” regions, or contain ‘N’ are discarded. The remaining one-hot-encoded DNA sequences are then multiplied by the exponential transformed conservation score:

$$inputs = one\ hot\ encoding\ DNA\ sequence \times e^{phastCons}$$

The ATAC-seq signals on the corresponding genomic segments are treated as the predicting targets. We first generate and normalize the genomic signal tracks by using deepTools(102) with parameter “--normalizeUsing RPKM”. Then, each segment is cut into 768 genomic bins in size of 128-bp, the mean value of ATAC signal across 128 bp is taken and assigned to that bin. To avoid potential bias from outlier signals, we cutoff signal values higher than 128, so that the bin signal is in the range of 0 to 128.

The model architecture is inspired by Enformer(72) with some modifications (see Supplementary **Fig. S15**). It is written based on one open-source machine learning framework, PyTorch, and contains three major parts: [1] 6 residual convolutional blocks with max pooling (Conv Tower). The kernel size (4,15) is used for the 1st residual convolutional block, and kernel size (4, 5) is used for the rest of residual convolutional block with padding, in order to extract informative sequence features at lower and higher resolution, respectively; The batch normalization and Gaussian Error Linear Unit (GELU) activation function are inserted between each convolutional layer. The convolutional blocks with pooling reduce the spatial dimension from 98,304 bp to 768 so that each sequence position vector represents a bin of 128 bp. [2] Transformer blocks including 4 multi-head attention layers is followed by the Conv Tower, to capture the long-range combinations and orders of sequence features. To inject positional information, we add relative positional encodings(103). Dropout rate of 0.4 is used for attention layer to avoid potential overfitting. The SwiGLU(104), a GELU variant, is used as activation function. [3] A pointwise convolution block is included to aggregate the weights from the last transformer layer and eventually output the predicted signals. The GELU and SwiGLU activation functions are used with dropout rate of 0.1.

To train the model, we randomly selected 5,000, 10,000 and full set of inputs together with its targets. We found that the model performance increase with the size of training set (**Fig. S15**). To balance the training time and performance, we used 10,000 inputs together with its target for training. Among 80% of inputs are used as training set, and the rest of 20% are used as validation set. To avoid overfitting problem, we introduce data augmentation during training by randomly shifting the input sequence by up to 3 bp and reverse-complementing the input sequence while reversing the targets every other epoch. We used the Adam optimizer from PyTorch with a learning rate of 0.0001 and default settings for other hyperparameters. The model was trained on 4 GPUs (NVIDIA GeForce RTX 3090 24GB). To overcome the limitation in GPU memory, we finally used a gradient accumulation approach to save the gradients and update network weights every 8 batches, to achieve an effective batch size of 64 (16 per GPU). We implemented a cosine

warmup learning rate scheduler by linearly increase the learning rate from 0 to target value in the first 10 epoches, and then decrease the learning rate to 0 by multiplying the cosine function in the next 90 epoches. The negative log likelihood loss with Poisson distribution of target (PoissonNLLLoss) is used for evaluating how well the model predict the targets. A cropping layer is introduced to trims 64 positions on each side to avoid computing the loss on the far ends because these regions are disadvantaged because they can observe regulatory elements only on one side (toward the sequence center) and not the other (the region beyond the sequence boundaries).

To evaluate the model, we test the model performance during the training on validation set, by calculating both loss and Pearson Correlation Coefficient (PCC) between predicted and target values across every input in batch. The model checkpoints with either lowest loss or highest PCC have been saved. We finally picked with model with lowest loss for downstream analyses. To evaluate the influence of risk variants on predicted chromatin accessibility, we performed *In silico* mutagenesis by altering every nucleotide within one cCRE. The predicted signals were calculated for both reference and alternative sequence for every output genomic position/bin. We then took the sum of predicted signals on output positions/bins that overlapped with cCRE, and measure the difference by subtracting the sum value on reference sequence from the sum value on alternative sequences.

Motif enrichment

We performed both *de novo* and known motif enrichment analysis using Homer (v4.11)(105). For cCREs in the consensus list, we scanned a region of ± 250 bp around the center of the element. Randomly selected background regions are used for motif discovery.

Gene ontology enrichment

We perform gene ontology enrichment analysis using R package Enrichr(106). Gene set library "GO_Biological_Process_2018" was used with default parameters. The combined score is defined as the p-value computed using the Fisher exact test multiplied with the z-score of the deviation from the expected rank.

GWAS enrichment

To enable comparison to GWAS of human phenotypes, we used liftOver with default setting “-minMatch=0.95” to convert accessible elements from hg38 to hg19 genomic coordinates(107). 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 length > 1kb.

We obtained GWAS summary statistics for quantitative traits related to neurological disease and control traits (**Table S25**): Alcohol Usage(108), Alzheimer’s Disease(109), Anorexia Nervosa(110), Attention Deficit Hyperactivity Disorder(111), Autism Spectrum Disorder(112), Bipolar Disorder(113), Chronotype(114), Educational Attainment(115), Extraversion(116), Insomnia(117), Intelligence(118), Major Depressive Disorder(119), Multiple sclerosis(120), Neuroticism(121), Posttraumatic Stress Disorder(122), Schizophrenia(123), Sleep Duration(124), Tiredness(125), and Tobacco use disorder(126). We prepared summary statistics to the standard format for Linkage disequilibrium (LD) score regression. We used homologous sequences for each major cell types as a binary annotation, and the superset of all candidate regulatory peaks as the background control. For each trait, we used cell type specific (CTS) LD

score regression (<https://github.com/bulik/ldsc>) to estimate the enrichment coefficient of each annotation jointly with the background control(127).

Fine mapping

We obtained 99% credible sets for schizophrenia from the Psychiatric Genomics Consortium website (<https://www.med.unc.edu/pgc/>). Potential causal variants with a posterior probabilities of association (PPA) score larger than 1% are used for overlapping with cCREs.

External datasets

We listed published datasets we used in this study for intersection analysis:

rDHS regions for hg38 is obtained from SCREEN database (<https://screen.encodeproject.org>).

The cCREs identified from adult and fetal human brains were downloaded from previously published snATAC-seq dataset.

The PhastCons⁵⁹ conserved elements were download from the UCSC Genome Browser (<http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/>).

The JASPAR motif prediction was downloaded from the UCSC Genome Browser (http://expdata.cmm.ubc.ca/JASPAR/downloads/UCSC_tracks/2022/).

The gene expression level of Vip⁺/VIP⁺ cell type from human and mouse brain were downloaded from Allen Brain Map: Cell Types Database (<https://portal.brain-map.org/atlas-and-data/rnaseq>).

The gene expression level of gene *BINI*, *MEF2A*, and *SPII* in normal humans and patients with Alzheimer's disease were download from Allen Brain Map: Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) (<https://portal.brain-map.org/explore/seattle-alzheimers-disease>).

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, clustering of single nuclei based on 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.

Supplementary Figures

Figure S1

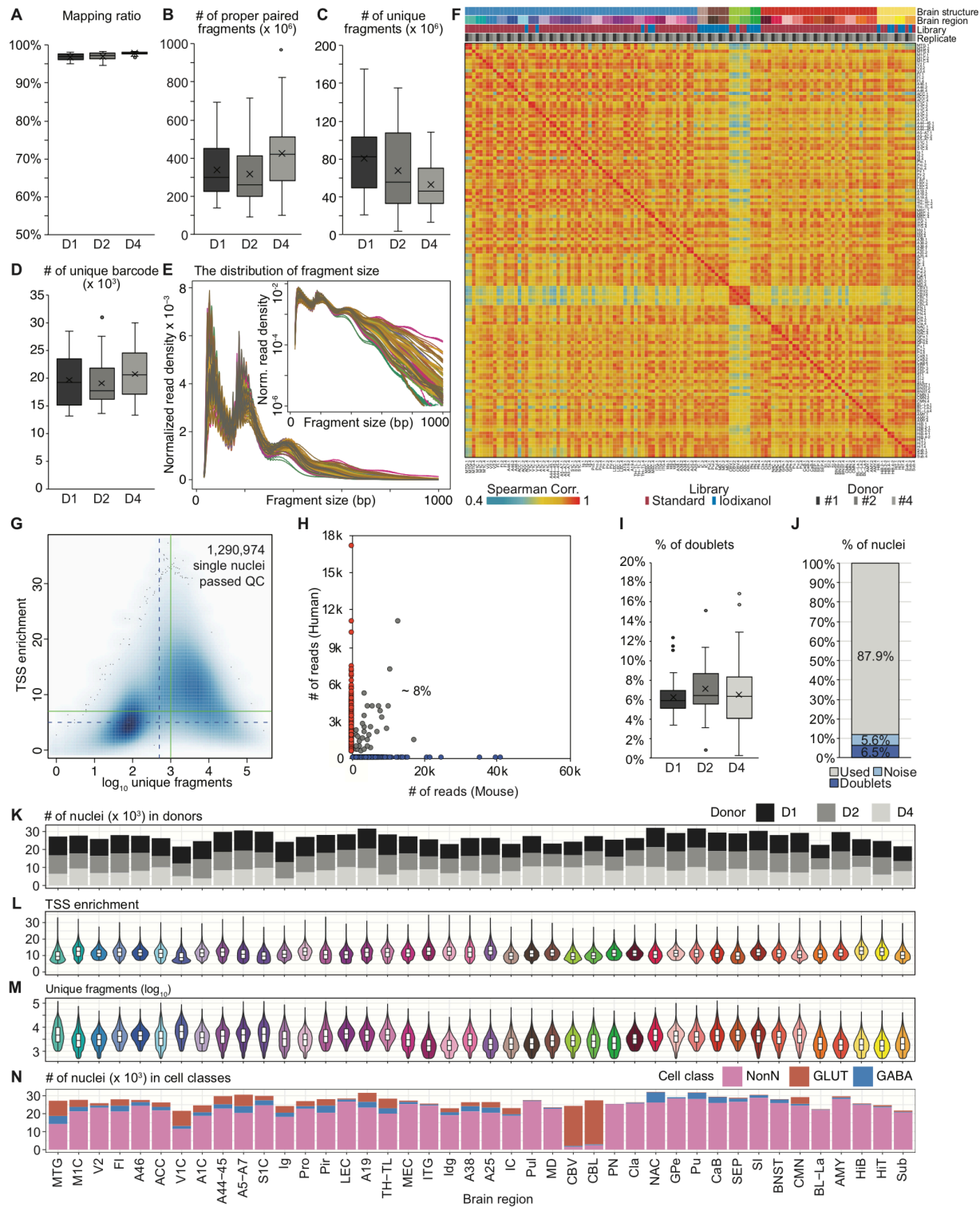


Fig. S1: Quality control metrics of the snATAC-seq datasets.

(A) Box plots showing the distribution of mapping ratios (the fraction of the mapped sequencing reads) in donor (biological replicate) D1, D2, and D4 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 donor D1, D2, and D4 of snATAC-seq experiments. (C) Box plots showing the distribution of numbers of unique chromatin fragments detected in donor D1, D2, and D4 of snATAC-seq experiments. (D) Box plots showing the distribution of number of unique barcodes captured in donor D1, D2, and D4 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 donor label. (G) Density scatter plot illustrating fragments per nucleus and individual TSS enrichment. Nuclei in top right quadrant were selected for analysis (TSS enrichment > 5 as acceptable and >7 as ideal, and > 500 fragments per nucleus as acceptable and >1,000 fragments as ideal). (H) Fraction of cell collision estimated from species mix samples. (I) Box plot shows the fraction of potential barcode collisions detected in snATAC-seq libraries using a modified version of *Scrublet*. (J) Number of nuclei retained after each step of quality control. (K) Number of nuclei passing quality control for each brain region from donors. (L) Distribution of TSS enrichment. (M) The number of uniquely mapped fragments per nucleus for individual libraries. Boxes span the first to third quartiles (Q1 to Q3), horizontal line denotes the median, and whiskers show 1.5× the interquartile range (IQR). (N) Number of nuclei from three cell classes in each brain region.

Figure S2

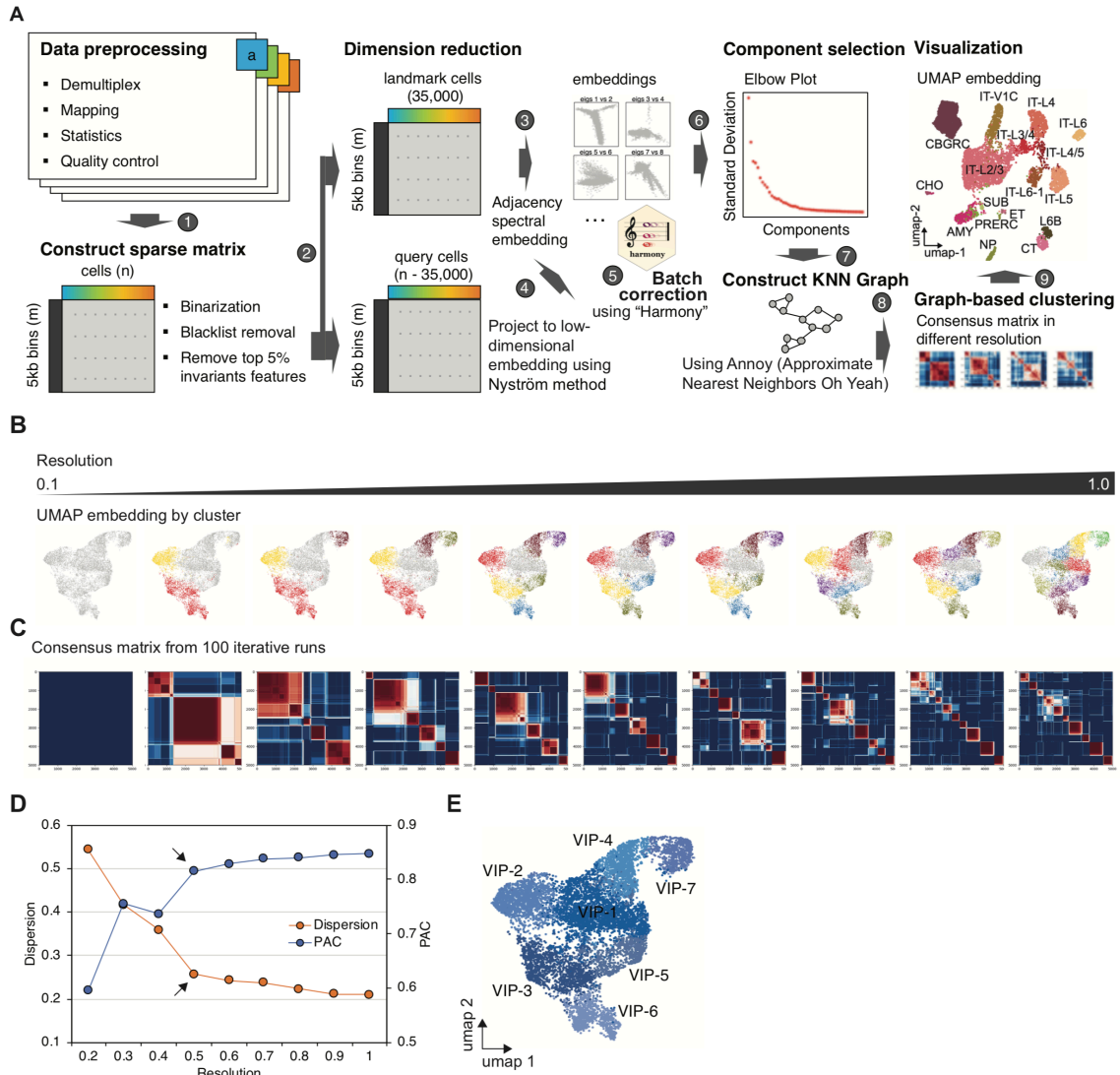


Fig. S2: Cell clustering pipeline

(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 Leiden algorithm. (C), Consensus matrix from 100 iterative clustering runs with different resolutions. (D) The proportion of ambiguous clustering (PAC) score and dispersion coefficient at different resolutions. A low PAC and high dispersion coefficient indicate the best and most stable clustering. (E) Optimal cell clustering result for a representative subclass (resolution = 0.5).

A Colored by subclasses

vGlut+ neurons
N = 133,612

GABA+ neurons
N = 76,732

Non-neurons
N = 924,016

n = 1,134,360

umap-2

umap-1

B SLC17A7 GAD1 GABRA6

SLC17A6 GAD2

Smoothed gene activity score

Low High

C Top accessible gene loci

Normalized gene activity score

Fig. S3: Maker genes used for annotation of different subclasses

(A) UMAP embedding of >1.1 million single nuclei and three major classes. (B) Smoothed gene activity score at marker loci used for annotation of three major classes. (C) Heat map showing normalized gene activity scores at marker gene loci used for subclass annotation.

Figure S4

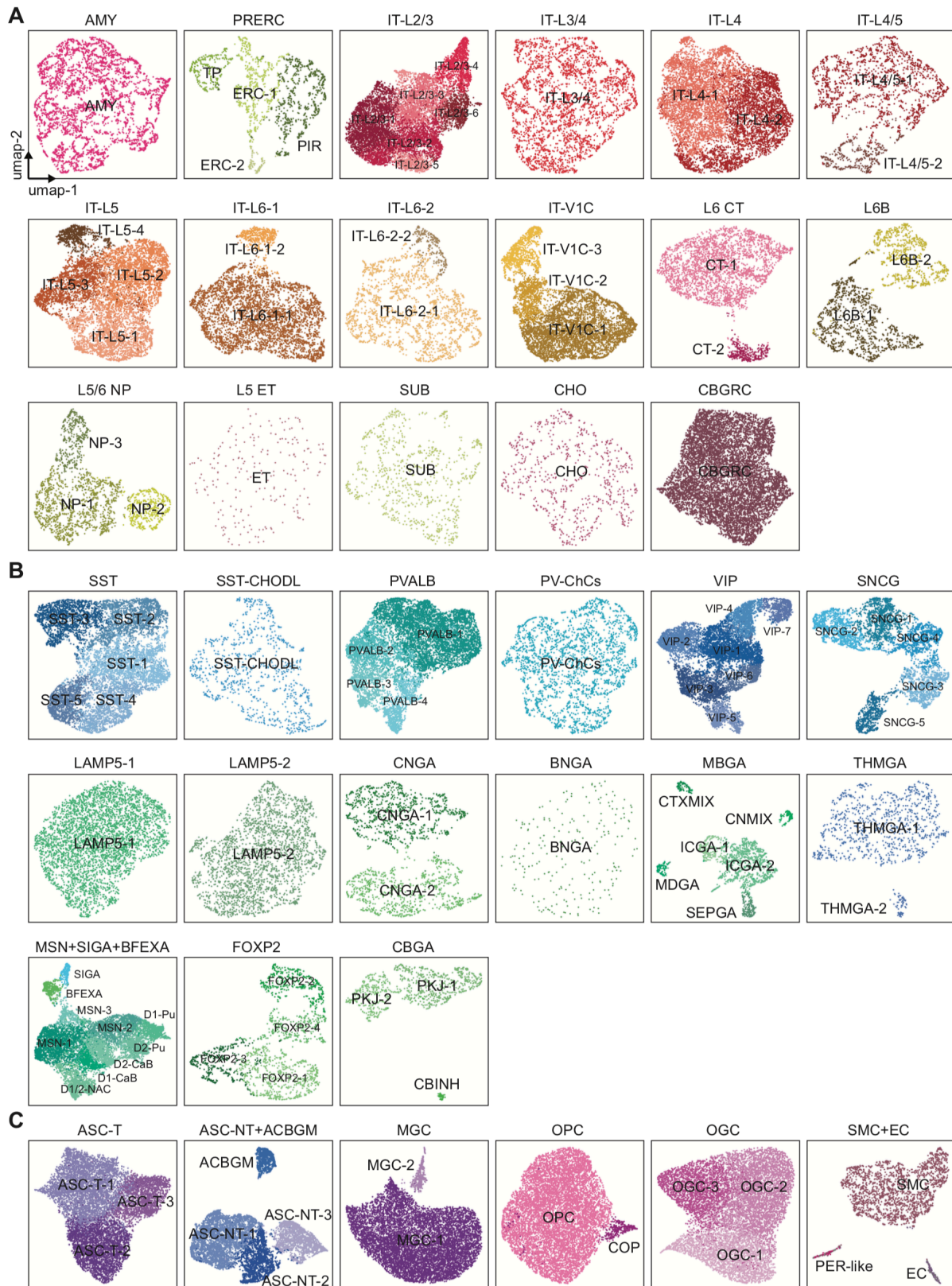


Fig. S4: Iterative clustering identifies cell types for the subclasses

The subclasses of cerebral cells that can be further divided into cell types are shown for the glutamatergic neurons (**A**), GABAergic neurons (**B**), and non-neurons (**C**). UMAP embedding is shown for each subclass, with the subclass label shown on the top, and cell-type labels on each cell type.

Figure S5

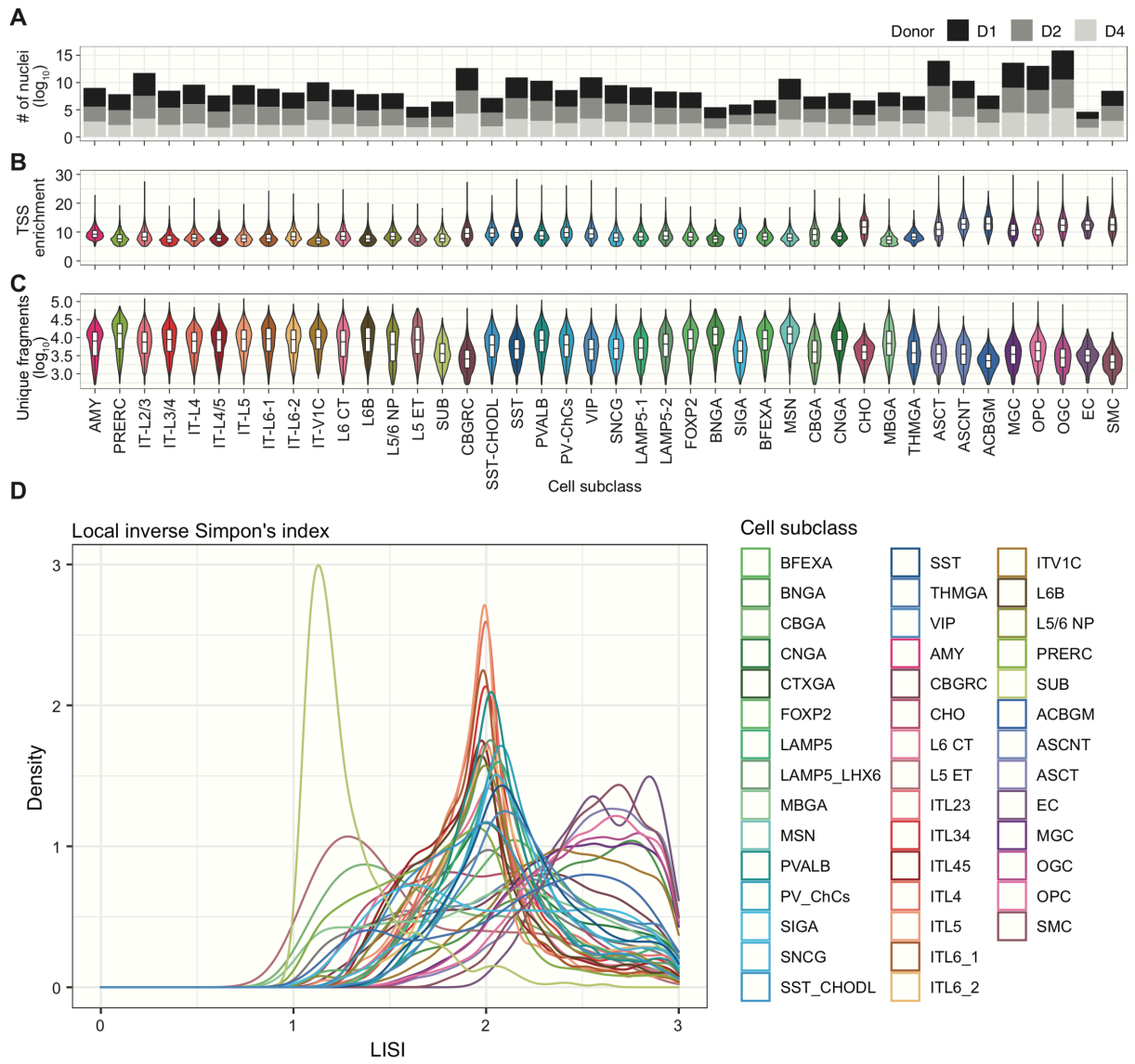


Fig. S5: Summary statistics of cell subclasses

(A) Number of nuclei retained for each cell subclasses from donors. (B) Violin plots showing the TSS enrichment in each nucleus of each subclass. (C) Violin plots showing the log-transformed number of unique fragments per nucleus in each cell subclass identified. (D) Distribution of the local inverse Simpson's index (LISI) scores for every single cell in each subclass.

Figure S6

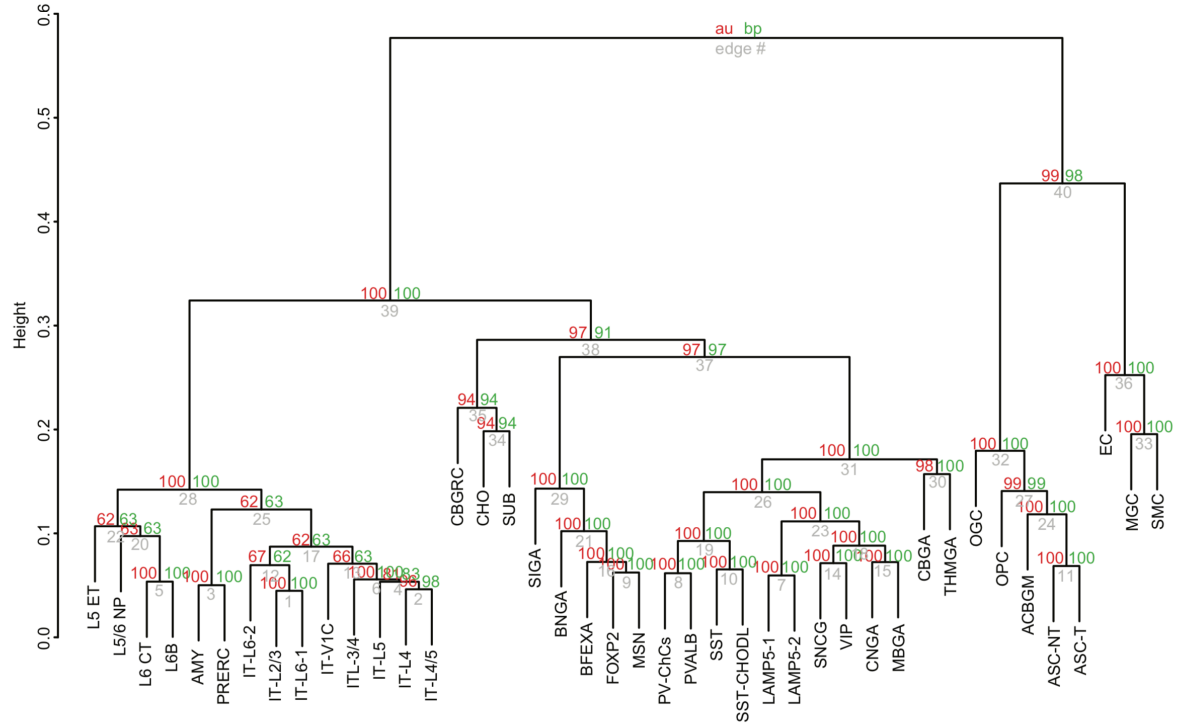


Fig. S6: Hierarchical tree depicting the relationships between different subclasses.

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 Table S4.

Figure S7

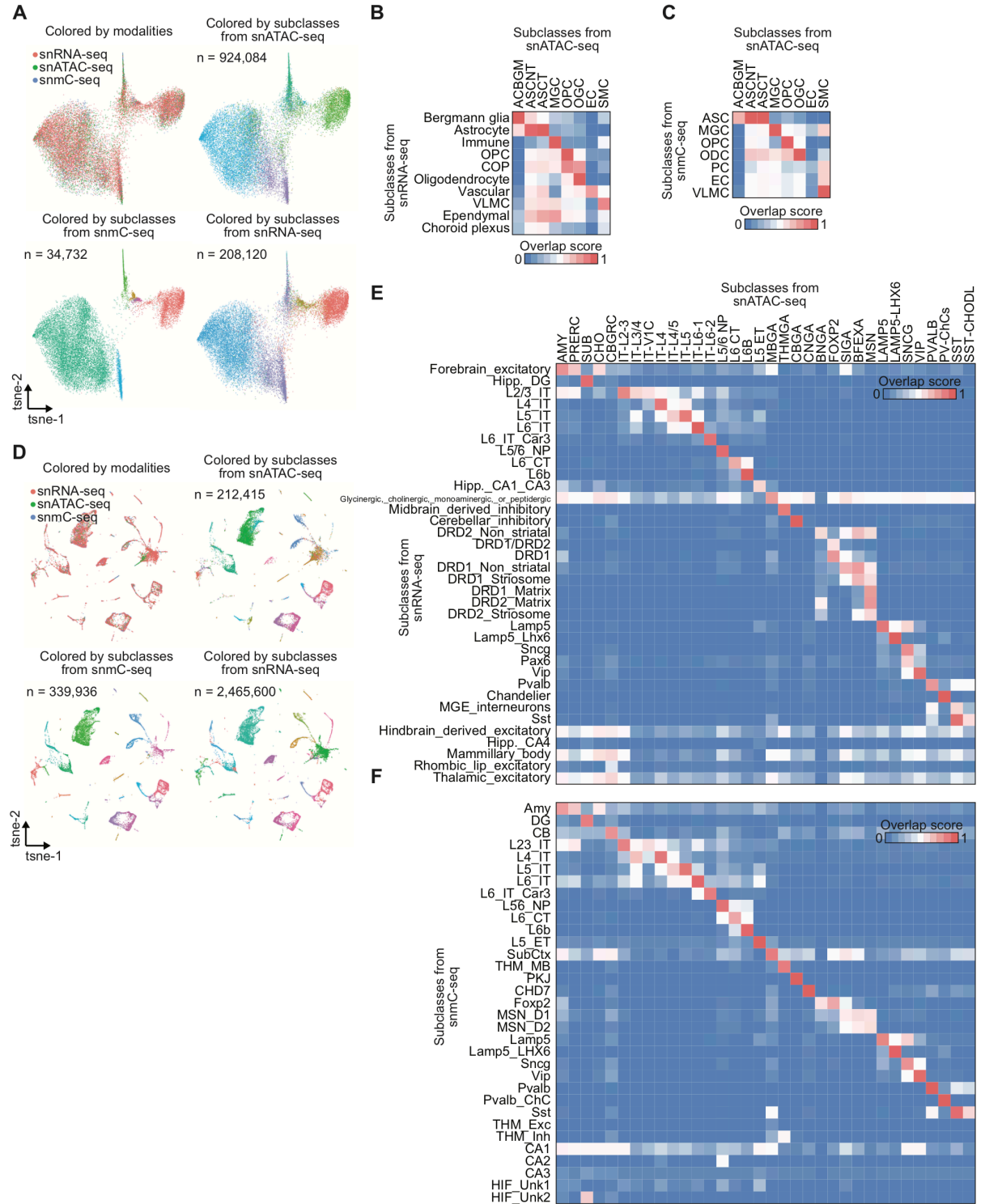


Fig. S7: Integration analysis of snATAC-seq, snRNA-seq and snmC-seq

(A) UMAP embedding of non-neuronal types identified from snATAC-seq, snRNA-seq, and snmC-seq. Single cells are colored by modalities, cell subclasses defined from snATAC-seq, snmC-seq and snRNA-seq clockwise from top left. (B) Heat map showing the similarity between A-type (accessibility) and T-type (transcriptomics) based non-neuronal cell cluster annotations. Each row represents an A-type cluster and each column represents a T-type cluster. 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. For a full list of cell type labels and descriptions see Table S5. (C) Heat map showing the similarity between A-type (accessibility) and M-type (methylomes) based non-neuronal cell cluster annotations. (D) UMAP embedding of neuronal types identified from snATAC-seq, snRNA-seq, and snmC-seq. Single cells are colored by modalities, cell subclasses defined from snATAC-seq, snmC-seq and snRNA-seq clockwise from top left. (E) Heat map showing the similarity between A-type (accessibility) and T-type (transcriptomics) based neuronal cell cluster annotations. (F) Heat map showing the similarity between A-type (accessibility) and M-type (methylomes) based neuronal cell cluster annotations.

Figure S8

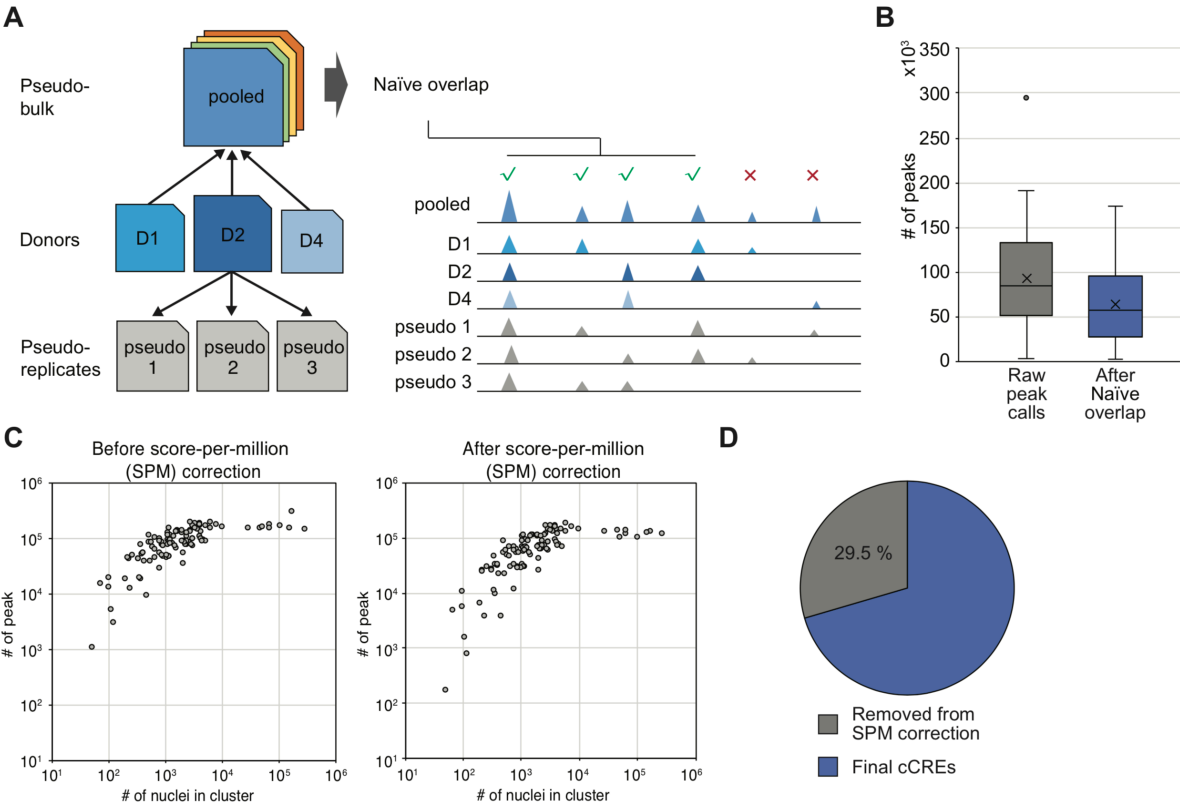


Fig. S8: Statistics of peak calling from snATAC-seq data from each cell type.

(A) Schematic diagram of peak calling and filtering pipeline. (B) Number of peaks retained after peak calling and naïve overlap filtering step for each cell type. (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) Fraction of peaks removed in SPM correction step.

Figure S9

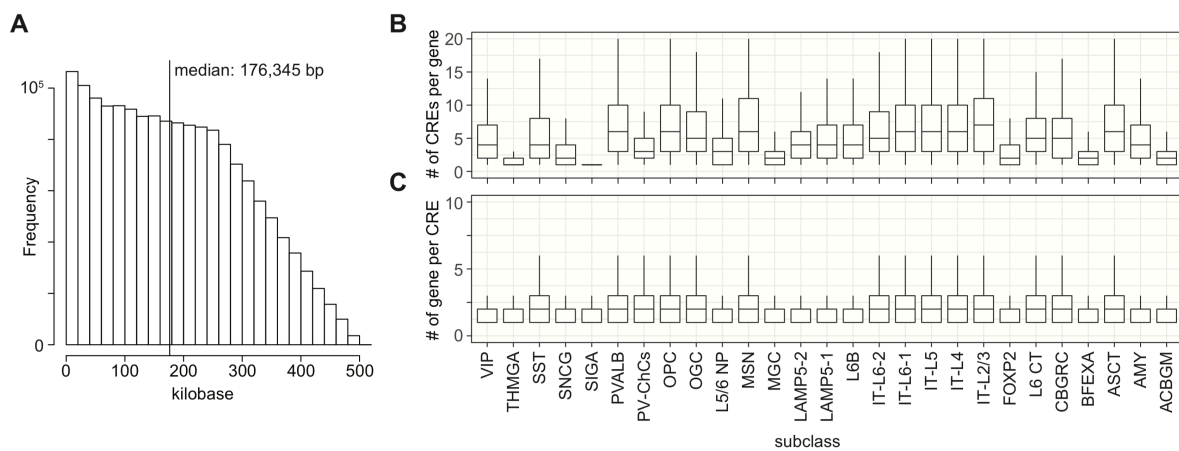


Fig. S9: Statistics of putative enhancer-gene pairs.

(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 across cell subclasses. (C) Box plot showing that putative were linked with a median of 2 genes across cell subclasses.

Figure 10

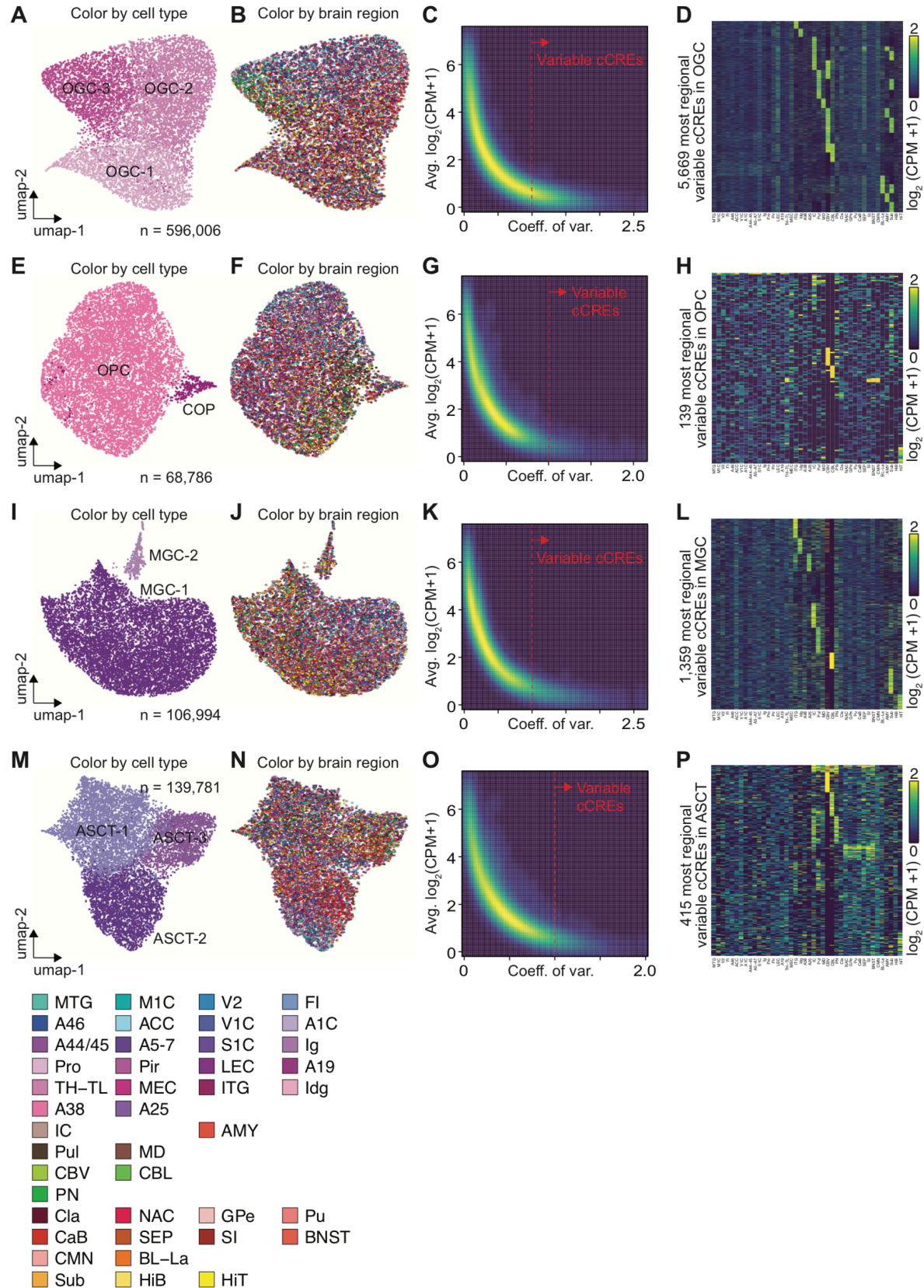


Fig. S10: Regional specificity of glial cell types

(A) UMAP embedding of cell types of oligodendrocytes (OGCs). (B) UMAP embedding of oligodendrocytes, colored by brain regions. (C) Density scatter plot comparing the averaged accessibility and coefficient of variation across brain regions at each cCRE. Most variable cCREs for OGCs are defined on the right side of dash line. (D) Heat map showing the normalized accessibility of most variable cCREs in OGCs. CPM, counts per million. (E) UMAP embedding of cell types of oligodendrocyte precursor cells (OPCs). (F) UMAP embedding of OPCs, colored by major brain regions. (G) Density scatter plot comparing the averaged accessibility and coefficient of variation across brain regions at each cCRE. Most variable cCREs for OPCs are defined on the right side of dash line. (H) Heat map showing the normalized accessibility of most variable cCREs in OPCs. (I) UMAP embedding of cell types of microglia (MGC). (J) UMAP embedding of MGC, colored by major brain regions. (K) Density scatter plot comparing the averaged accessibility and coefficient of variation across brain regions at each cCRE. Most variable cCREs for MGC are defined on the right side of dash line. (L) Heat map showing the normalized accessibility of most variable cCREs in MGC. (M) UMAP embedding of cell types of astrocytes (ASCs). (N) UMAP embedding of ASCs, colored by major brain regions. (O) Density scatter plot comparing the averaged accessibility and coefficient of variation across brain regions at each cCRE. Most variable cCREs for ASCs are defined on the right side of dash line. (P) Heat map showing the normalized accessibility of most variable cCREs in ASCs.

Figure S11

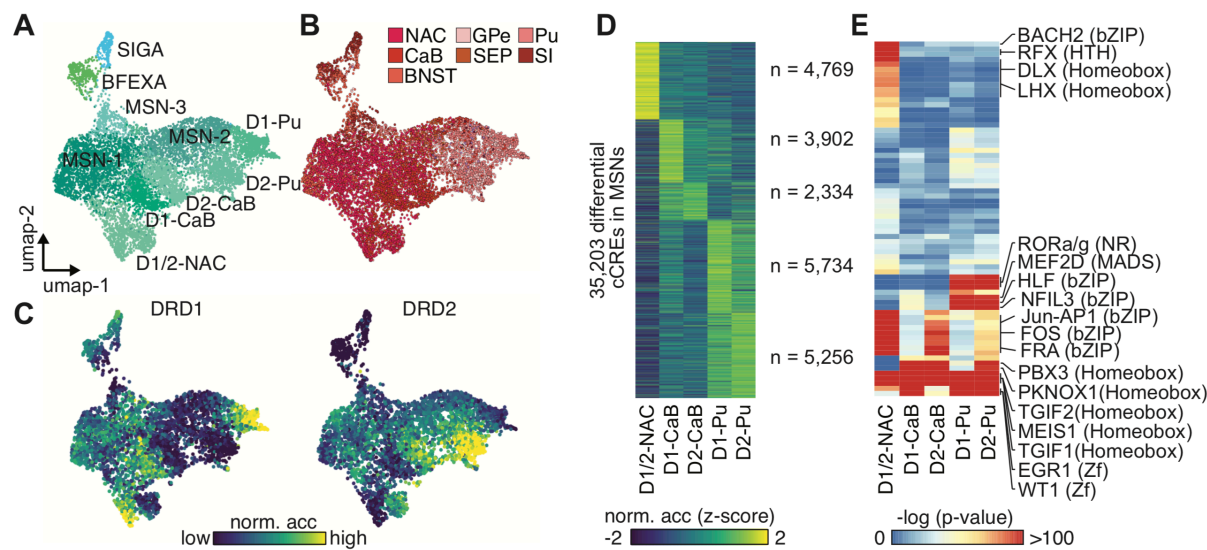


Fig. S11: Regional difference in MSN.

(A) UMAP embedding of cell types of medium spiny neurons (MSNs) and cell types from cerebral nuclei. (B) UMAP embedding of MSNs, colored by brain regions. (C) UMAP embedding of normalized accessibility at gene *DRD1* and *DRD2* in MSNs. (D) Normalized chromatin accessibility of 35,203 MSN cell-type-specific cCREs. (E) Enrichment of known transcription factor (TF) motifs in MSN cell-type-specific cCREs.

Figure S12

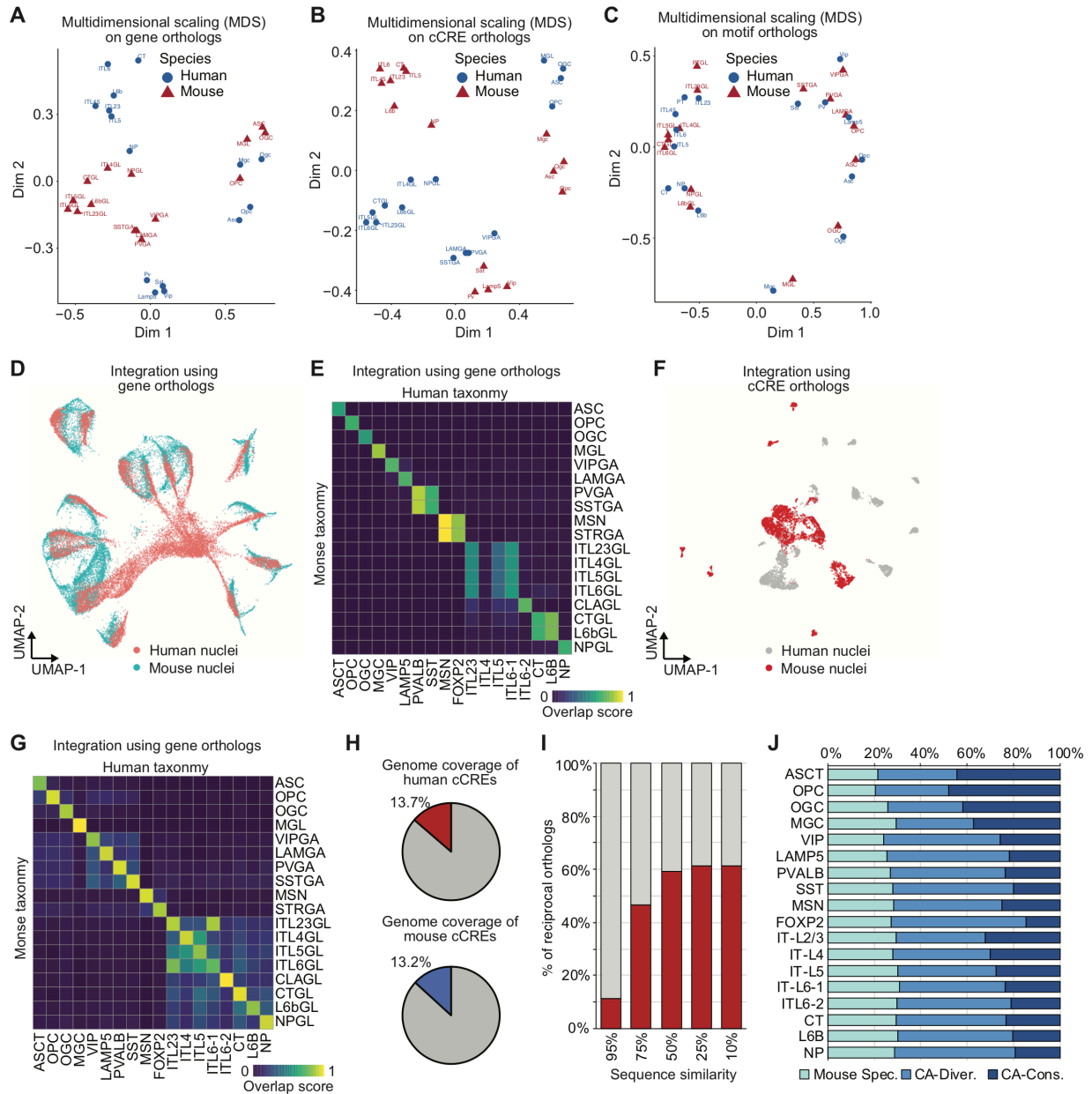


Fig. S12: Integration analysis of snATAC-seq data between human and mouse cerebrum

(A) Multidimensional scaling (MDS) embedding of human and mouse subclasses using gene orthologs. (B) MDS embedding of human and mouse subclasses using cCRE orthologs. (C) MDS embedding of human and mouse subclasses using motif orthologs. (D) Co-embedding analysis of single nuclei from human and mouse subclasses using “gene activity score” from gene orthologs. (E) Heat map illustrating the overlap between cell subclasses from both species from integration analysis using gene orthologs. Rows show cell subclasses from mouse cerebrum; columns show cell subclasses from the human cerebrum. The overlap between the human and mouse cell subclasses was calculated (overlap score) and plotted on the heat map. (F) Co-embedding analysis of single nuclei from human and mouse subclasses using chromatin accessibility from cCRE orthologs. (G) Heat map illustrating the overlap between cell subclasses from both species from integration analysis using motif orthologs. (H) Genome coverage of cCREs from used human and mouse subclasses. (I) Percentage of reciprocal orthologous cCREs identified in human by using different cutoff of sequence similarity. (J) bar plot showing three categories of mouse cCREs in corresponding cell subclasses.

Figure S13

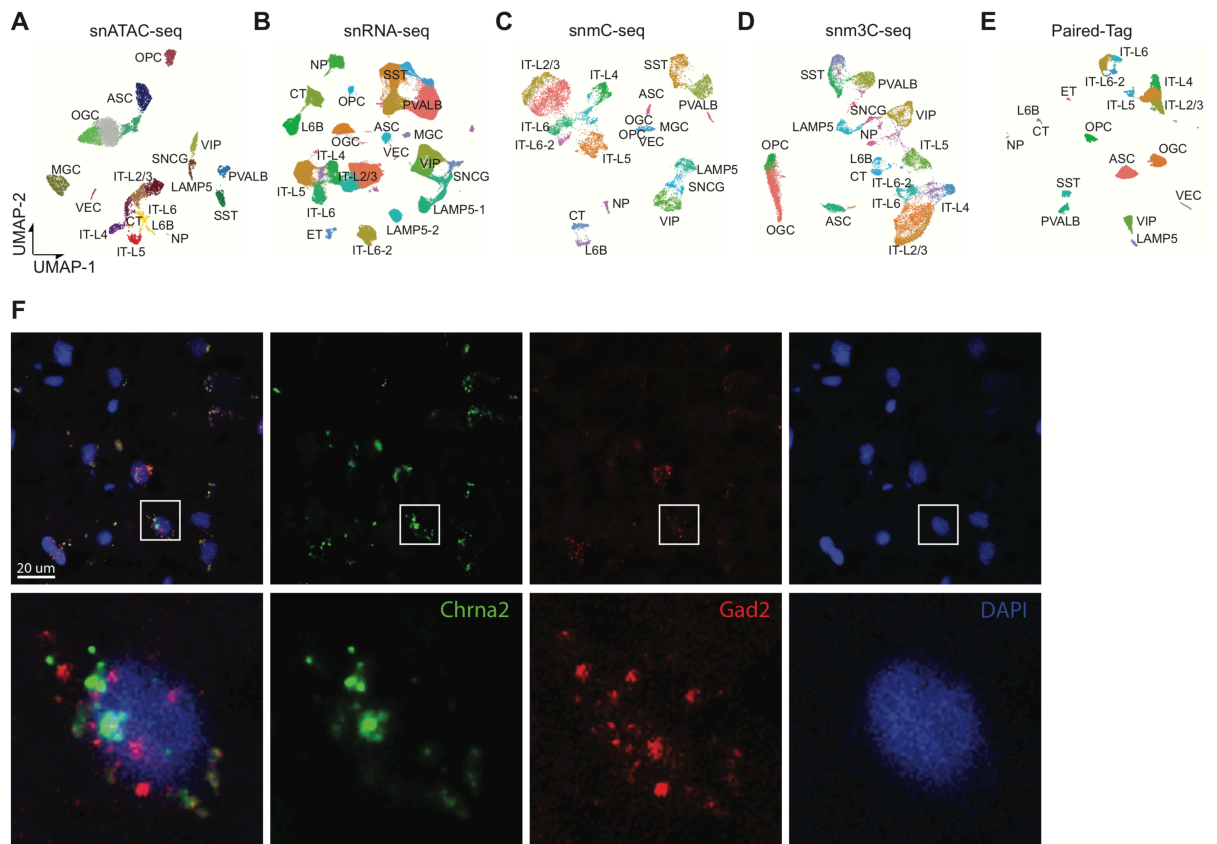


Fig. S13: Integration analysis of multi-modal omics datasets

UMAP embedding of cell clustering and annotation for snATAC-seq (**A**), snRNA-seq (**B**), snmC-seq (**C**), snm3C-seq (**D**), Paired-Tag (**E**) from brain region M1C and MTG. (**F**) RNAscope validation of *CHRNA2* position lineage of VIP⁺ GABAergic neurons.

Figure S14

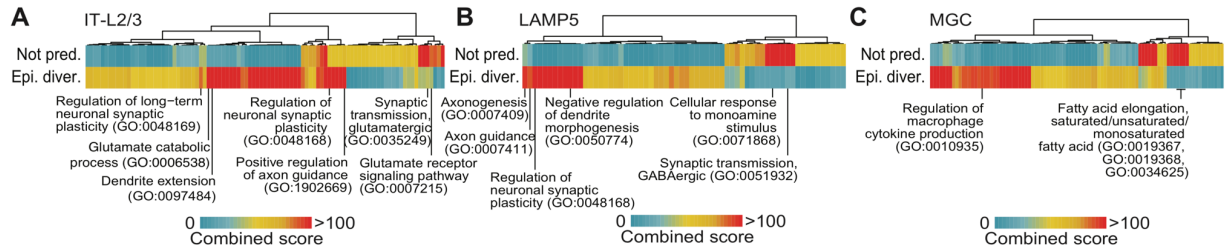


Fig. S14: Gene ontology of putative target genes that linked with human divergent cCREs predicted and not predicted from mouse gkmsvm models

(A) Gene ontology of putative target genes that linked to human divergent cCREs in IT-L2/3, which can be predicted and cannot predicted from mouse gkmsvm model. (B) Gene ontology of putative target genes that linked to human divergent cCREs in IT-L2/3, which can be predicted and cannot predicted from mouse gkmsvm model. (C) Gene ontology of putative target genes that linked to human divergent cCREs in IT-L2/3, which can be predicted and cannot predicted from mouse gkmsvm model. The combined score is defined in R package *Enrichr*, which calculated by multiplying the log-transformed p-value from Fisher's exact test with the z-score computed to assess the deviation from the expected rank.

Figure S15

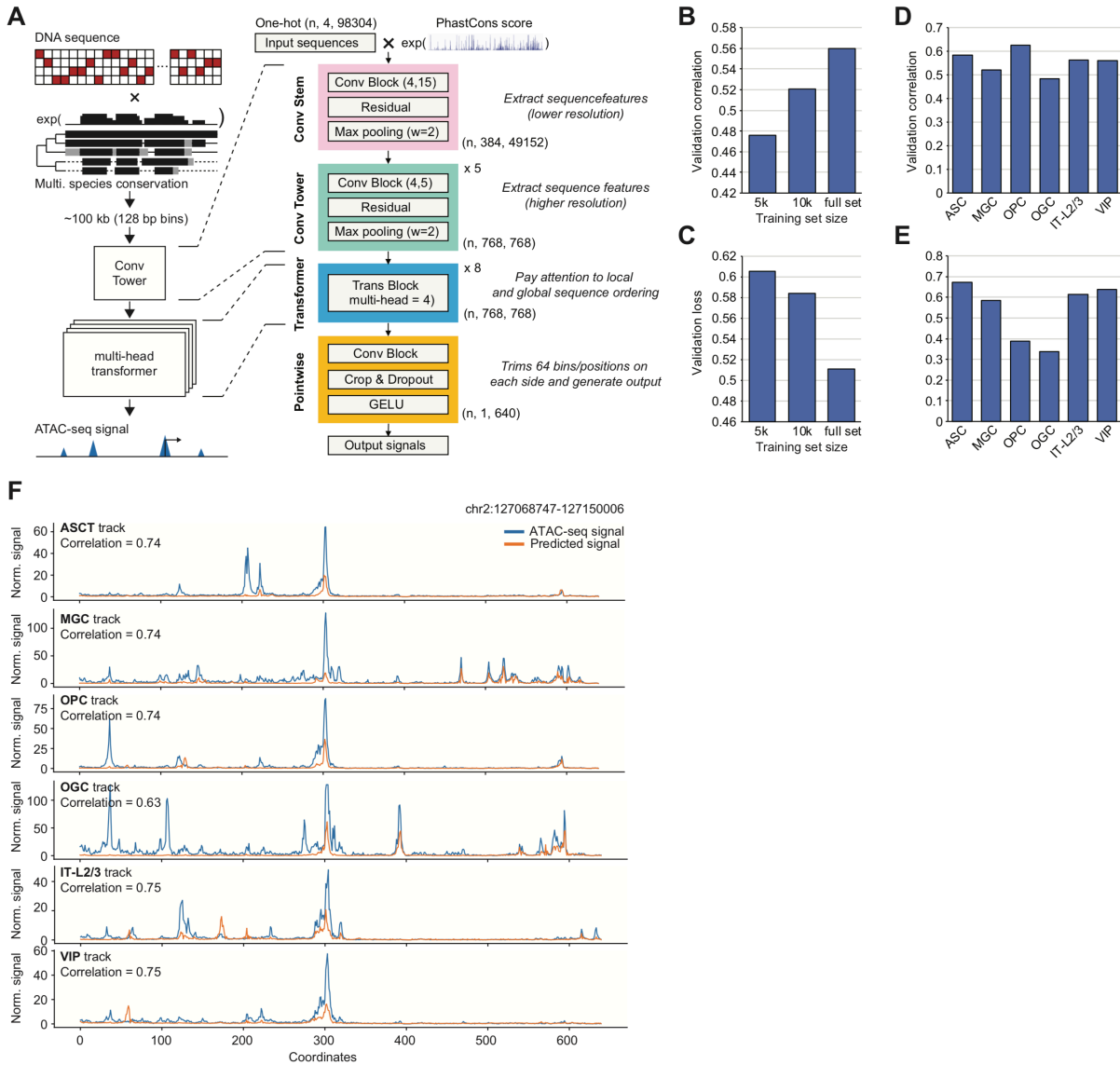


Fig. S15: Deep learning model architecture and statistics

(A) The graphic diagram showing the model architecture. Output shapes (without batch dimensions) are shown as tuples on the right side of the blocks. n, number of inputs in mini-batch, w, width for padding. (B) Benchmark of model performance on validation set using different size of training set. The Pearson correlation coefficient showed was calculated between true ATAC-seq signals and predicted signals. (C) Benchmark of loss on validation set using different size of training set. The loss showed was calculated from function *PoissonNLLLoss*. (D) Bar plots showing model performance on different human cell subclasses. (E) Bar plots showing model loss on different human cell subclasses. (F) Predicted signal centered at BIN1 promoter from models trained from different cell subclasses.

Supplementary Tables

Table S1: Summary of brain samples and dissections.

Table S2: Metadata for the snATAC-seq experiments.

Table S3: Metadata and annotation of single nuclei.

Table S4: Human brain cell taxonomy and annotation.

Table S5: Cell subclass annotation and marker genes.

Table S6: List of cCREs defined in the current study.

Table S7: Cell type assignment of cCREs.

Table S8: Genomic annotation of cCREs defined in the current study.

Table S9: The cis regulatory modules and the cell subclasses

Table S10: The cis regulatory modules and cCREs.

Table S11: Known motif enrichment in the cis regulatory modules.

Table S12: Summary of gene-cCRE correlations.

Table S13: Association modules of enhancer-gene pairs with RNA-ATAC matched cell subclasses.

Table S14: Association modules of enhancer-gene pairs with individual putative enhancers.

Table S15: Known motif enrichment in the putative enhancers from different modules.

Table S16: Differential cCRE in the cell types of astrocyte.

Table S17: Known motif enrichment in the cell types of astrocyte.

Table S18: Differential cCRE in the cell types of D1-/D2-MSN.

Table S19: Known motif enrichment in the cell types of D1-/D2-MSN.

Table S20: List of different categories of cCREs.

Table S21: Genomic annotation of different categories of cCREs.

Table S22: Differential cCRE in the cell types of VIP⁺ neurons.

Table S23: Different categories of cCREs marked by K27ac and K27me3.

Table S24: Gene Ontology analysis of the putative target gene of different categories of cCREs.

Table S25: List and references of genome-wide association studies.

Table S26: LDSC analysis of cCREs for every cell type.

Table S27: LDSC analysis of different categories of cCREs.

Table S28: in silico mutagenesis on microglia-specific enhancer.

Table S29: Primer sequences and nuclei barcodes.