An Atlas of Gene Regulatory Elements in Adult Mouse Cerebrum

Yang Eric Li[1]*, Sebastian Preissl[2*], Xiaomeng Hou[2], Ziyang Zhang[1], Kai Zhang[1], Rongxin Fang[1], Yunjiang Qiu[1], Olivier Poirion[2], Bin Li[1], Hanqing Liu[3], Xinxin Wang[2], Jee Yun Han[2], Jacinta Lucero[4], Yiming Yan[1], Samantha Kuan[1], David Gorkin[2], Michael Nunn[3], Eran A. Mukamel[5], M. Margarita Behrens[4], Joseph Ecker[3,6] and Bing Ren[1,2,7]

*these authors contributed equally

[1]Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, CA 92093, USA
[2]Center for Epigenomics, Department of Cellular and Molecular Medicine, University of California, San Diego, School of Medicine, La Jolla, CA, USA.
[3]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, 92037, USA.
[4]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[5]Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92037, USA.
[6]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA, 92037, USA.
[7]Institute of Genomic Medicine, Moores Cancer Center, University of California San Diego, School of Medicine, La Jolla, CA, USA.

Correspondence: Bing Ren (biren@ucsd.edu)

ABSTRACT

The mammalian cerebrum performs high level sensory, motor control and cognitive functions through highly specialized cortical networks and subcortical nuclei. Recent surveys of mouse and human brains with single cell transcriptomics[1-3] and high-throughput imaging technologies[4,5] have uncovered hundreds of neuronal cell types and a variety of non-neuronal cell types distributed in different brain regions, but the cell-type-specific transcriptional regulatory programs responsible for the unique identity and function of each brain cell type have yet to be elucidated. Here, we probe the accessible chromatin in >800,000 individual nuclei from 45 regions spanning the adult mouse isocortex, olfactory bulb, hippocampus and cerebral nuclei, and use the resulting data to define 491,818 candidate *cis* regulatory DNA elements in 160 distinct sub-types. We link a significant fraction of them to putative target genes expressed in diverse cerebral cell types and uncover transcriptional regulators involved in a broad spectrum of molecular and cellular pathways in different neuronal and glial cell populations. Our results provide a foundation for comprehensive analysis of gene regulatory programs of the mammalian brain and assist in the interpretation of non-coding risk variants associated with various neurological disease and traits in humans. To facilitate the dissemination of information, we have set up a web portal (http://catlas.org/mousebrain).

INTRODUCTION

In mammals, the cerebrum is the largest part of the brain and carries out essential functions such as sensory processing, motor control, emotion, and cognition[6]. It is divided into two hemispheres, each consisting of the cerebral cortex and various cerebral nuclei. The cerebral cortex is further divided into isocortex and allocortex. Isocortex, characterized by six cortical layers, is a phylogenetically more recent structure that has further expanded greatly in primates. It is responsible for sensory motor integration, decision making, volitional motor command and reasoning. The allocortex, by contrast, is phylogenetically the older structure that features three or four cortical layers. It includes the olfactory bulb responsible for processing the sense of smell and the hippocampus involved in learning, memory and spatial navigation.

The cerebral cortex and basal ganglia are made up of a vast number of neurons and glial cells. The neurons can be classified into different types of excitatory projection neurons and inhibitory interneurons, defined by the neural transmitters they produce and their connective patterns with other neurons[7-9]. Understanding how the identity and function of each brain cell type is established during development and modified by experience is one of the fundamental challenges in brain research. Recent single cell RNA-seq and high throughput imaging experiments have produced detailed cell atlases for both mouse and human brains[3-5,10-15], leading to a comprehensive view of gene expression patterns in different brain regions, cell types and physiological states[16-18]. Despite these advances, the gene regulatory programs in most brain cell types have remained to be characterized. A major barrier to the understanding of cell-type specific transcriptional control is the lack of comprehensive maps of the regulatory elements in diverse brain cell types.

Transcriptional regulatory elements recruit transcription factors to exert control of target gene expression in *cis* in a cell-type dependent manner[19]. The regulatory activity of these elements is accompanied by open chromatin, specific histone modifications and DNA hypomethylation[19]. Exploiting these structural features, candidate *cis* regulatory elements (cCREs) have been mapped with the use of tools such as DNase-seq, ATAC-seq, ChIP-

seq and Whole genome bisulfite sequencing[20,21]. Conventional assays, typically performed using bulk tissue samples, unfortunately fail to resolve the cCREs in individual cell types comprising the extremely heterogeneous brain tissues. To overcome this limitation, single cell genomic technologies, such as single cell ATAC-seq, have been developed to enable analysis of open chromatin in individual cells[22-28]. These tools have been used to probe transcriptional regulatory elements in the prefrontal cortex[28,29], cerebellum[29], hippocampus[30], forebrain[31] or the whole brain[24,29], leading to identification of cell-type specific transcriptional regulatory sequences in these brain regions. These initial studies provided proof of principle for the use of single cell chromatin accessibility assays to resolve cell types and cell-type specific regulatory sequences in complex brain tissues, but the number of cells analyzed, and the *cis* regulatory elements identified so far are still limited.

In the present study, as part of the BRAIN Initiative Cell Census Network, we conducted the most comprehensive analysis to date to identify candidate *cis* regulatory elements (cCRE) in the mammalian brain at single cell resolution. Using a semi-automated single nucleus ATAC-seq (snATAC-seq) procedure[22,31], we mapped accessible chromatin in >800,000 cells from the mouse isocortex, hippocampus, olfactory bulb, and cerebral nuclei (including striatum and pallidum). We defined 160 sub-types based on the chromatin landscapes and matched 155 of them to previous cell taxonomy of the mouse brain[1]. We delineated the cell-type specificity for >490,000 cCREs that make up nearly 14.8% of the mouse genome. We also integrated the chromatin accessibility data with available brain single cell RNA-seq data to assess their potential role in cell-type specific gene expression patterns, and gain mechanistic insights into the gene regulatory programs of different brain cell types. We further demonstrated that the human counterparts of the identified mouse brain cCREs are enriched for risk variants associated with neurological disease traits in a cell-type-specific and region-specific manner.

RESULTS

Single cell analysis of chromatin accessibility of the adult mouse brain

We performed snATAC-seq, also known as sci-ATAC-seq[22,31], for 45 brain regions dissected from isocortex, olfactory bulb (OLF), hippocampus (HIP) and cerebral nuclei (CNU) (Fig. 1a, Extended Data Figure 1, Supplementary Table 1, see **Methods**) in 8-week-old male mice. Each dissection was made from 600 μm thick coronal brain slices according to the Allen Brain Reference Atlas (Extended Data Figure 1)[32]. For each region, snATAC-seq libraries from two independent biological replicates were generated with a protocol[31] that had been optimized for automation (Fig. 1a, see **Methods**). The libraries were sequenced, and the reads were deconvoluted based on nucleus-specific barcode combinations. We confirmed that the dataset of each replicate met the quality control metrics (Extended Data Figure 2a-e, see **Methods**). We selected nuclei with at least 1,000 sequenced fragments that displayed high enrichment (>10) in the annotated transcriptional start sites (TSS; Extended Data Figure 2b). We also removed the snATAC-seq profiles likely resulting from potential barcode collision or doublets using a procedure modified from Scrublet[33] (Extended Data Figure. 2c, see **Methods**). Altogether, we obtained chromatin profiles from 813,799 nuclei with a median of 4,929 fragments per nucleus (Supplementary Table 2). Among them, 381,471 were from isocortex, 123,434 from olfactory area, 147,338 from cerebral nuclei and 161,556 from hippocampus (Fig. 1a, Extended Data Figure 2f). Thus, this dataset represents by far the largest number of chromatin accessibility profiles for these brain areas. Excellent correlation between datasets from similar brain regions (0.92-0.99 for isocortex; 0.89-0.98 for OLF; 0.79-0.98 for CNU; 0.88-0.98 for hippocampus) and between biological replicates (0.98 in median, range from 0.95 to 0.99) indicated high reproducibility and robustness of the experiments (Extended Data Figure 2g).

Clustering and annotation of mouse brain cells based on open chromatin landscapes

134 We carried out iterative clustering with the software package SnapATAC[34] to classify the
135 813,799 snATAC-seq profiles into distinct cell groups based on the similarity of chromatin
136 accessibility profiles (Fig. 1b-e, Supplementary Table 2 and 3, see **Methods**)[34].
137 SnapATAC clusters chromatin accessibility profiles using a nonlinear dimensionality
138 reduction method that is highly robust to noise and perturbation[34]. We performed three
139 iterative rounds of clustering, first separating cells into three broad classes, then into
140 major types within each class, and finally into more sub-types. In the first iteration, we
141 grouped cells into glutamatergic neurons (387,060 nuclei, 47.6%), GABAergic neurons
142 (167,181 nuclei, 20.5%) and non-neuronal cells (259,588 nuclei, 31.9%; Fig. 1b-d). For
143 each main cell class, we performed a second round of clustering. We identified a total of
144 43 major types including distinct layer-specific cortical neurons, hippocampal granular
145 cells (GRC) and striatal D1 and D2 medium spiny neurons (D1MSN, D2MSN; Fig 1b, d)
146 which were annotated based on chromatin accessibility in promoters and gene bodies of
147 known marker genes (Fig. 1e)[1,3]. Finally, for each major type we conducted another round
148 of clustering to reveal sub-types. For example, *Lamp5*$^{+}$ neurons (LAMGA) and *Sst*$^{+}$
149 neurons (SSTGA) were further divided into sub-types (Fig. 1d, e, Supplementary Table
150 3)[3,35]. One of the LAMGA subtypes showed accessibility at *Lhx6* and therefore might
151 resemble an unusual transcriptomically defined putative chandelier-like cell type with
152 features from caudal ganglionic and medial ganglionic eminence (Fig. 1b, e)[3]. Similarly,
153 using this third layer clustering we found one SSTGA subpopulation with accessibility at
154 *Chodl* locus which resembles long range projecting GABAergic neurons (Fig. 1b, e)[35].
155 Altogether, we were able to resolve 160 sub-types, with the number of nuclei in each
156 group ranging from 93 to 75,474 and a median number of 5,086 nuclei per cluster
157 (Supplementary Table 3).
158
159 We constructed a hierarchical dendrogram to represent the relative similarity in chromatin
160 landscapes among the 43 major cell groups (Fig. 1d, Extended Data Figure 3). This
161 dendrogram captures known organizing principles of mammalian brain cells: Neurons are
162 separated from non-neuronal types followed by separation of neurons based on
163 neurotransmitter types (GABAergic versus glutamatergic) and finally into more specified
164 cell types which might resemble the developmental origins (Fig. 1d)[3]. Consistent with

previous reports of brain cell types, we found that non-neuronal cells were broadly distributed in all regions while several classes of glutamatergic neurons and GABAergic neurons showed regional specificity (Fig. 1c, f, Extended Data Figure 4)[3]. We also found that glutamatergic neuron types showed more regional specificity than GABAergic types, consistent with transcriptomic analysis (Fig. 1c, f, Extended Data Figure 4)[3].

The chromatin-defined cell types matched well with the previously reported taxonomy based on transcriptomes and DNA methylomes[3,36] (see companion manuscript by Liu, Zhou et al.[37]). To directly compare our single nucleus chromatin-derived cell clusters with the single cell transcriptomics defined taxonomy of the mouse brain[1], 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 (Seurat package[38]). We then performed integrative analysis with scRNA-seq data from matched brain regions of the Mouse Brain Atlas[1]. We found strong correspondence between the two modalities which was evidenced by co-embedding of both transcriptomic (T-type) and chromatin accessibility (A-type) cells in the same joint clusters (Fig. 2a-c, Supplementary Table 4, see **Methods**). For this analysis, we examined GABAergic neurons, glutamatergic neurons and non-neuronal cell classes separately (Fig. 2a-c, Supplementary Table 4, see **Methods**). For 155 of 160 types defined by snATAC-seq (A-Type), we could identify a corresponding cell cluster defined using scRNA-seq data (T-Type, Fig. 2d, e); conversely, for 84 out of 100 T-types we identified one, or in some cases more, corresponding A-types (Fig. 2d, f). Of note, two clusters fell into different classes. The Cajal-Retzius cells (CRC) was part of the GABAergic class in A-type but glutamatergic class in T-type and one small non-neuronal A-type cluster, VPIA3 (Vascular and leptomeningeal like cells) co-clustered with CRC T-type (Fig. 2d). Nevertheless, the general agreement between the open chromatin-based clustering and transcriptomics-based clustering laid the foundation for integrative analysis of cell-type specific gene regulatory programs in the mouse brain using single cell RNA and single nucleus chromatin accessibility assays, as for the mouse primary motor cortex[15].

Identification of cCREs in different mouse brain cell types

196

197 To identify the cCREs in each of the 160 A-types defined from chromatin landscapes, we
198 aggregated the snATAC-seq profiles from the nuclei comprising each cell cluster and
199 determined the open chromatin regions with MACS2[39]. We then selected the genomic
200 regions mapped as accessible chromatin in both biological replicates, finding an average
201 of 93,775 (range from 50,977 to 136,962) sites (500-bp in length) in each sub-type. We
202 further selected the elements that were identified as open chromatin in a significant
203 fraction of the cells in each sub-type (FDR >0.01, zero inflated Beta model, see **Methods**),
204 resulting in a union of 491,818 open chromatin regions. These cCREs occupied 14.8% of
205 the mouse genome (Supplementary Table 5 and 6).

206

207 96.3% of the mapped cCREs were located at least 1 kbp away from annotated promoter
208 regions of protein-coding and lncRNA genes (Gencode V16) (Fig. 3a)[40]. Several lines of
209 evidence support the function of the identified cCREs. First, they largely overlapped with
210 the DNase hypersensitive sites (DHS) previously mapped in a broad spectrum of bulk
211 mouse tissues and developmental stages by the ENCODE consortium (Fig. 3b)[41,42].
212 Second, they generally showed higher levels of sequence conservation than random
213 genomic regions with similar GC content (Fig 3c). Third, they were enriched for active
214 chromatin states or potential insulator protein binding sites previously mapped with bulk
215 analysis of mouse brain tissues (Fig. 3d)[43-45].

216

217 To define the cell-type specificity of the cCREs, we first plotted the median levels of
218 chromatin accessibility against the maximum variation for each element (Fig 3e). We
219 found that the majority of cCREs displayed highly variable levels of chromatin accessibility
220 across the brain cell clusters identified in the current study, with the exception for 8,188
221 regions that showed accessible chromatin in virtually all cell clusters (Fig 3e). The
222 invariant cCREs were highly enriched for promoters (81%), with the remainder including
223 CTC-binding factor (CTCF) binding sites and strong enhancers (Fig 3f). To more explicitly
224 characterize the cell-type specificity of the cCREs, we used non-negative matrix
225 factorization to group them into 42 modules, with elements in each module sharing similar
226 cell-type specificity profiles. Except for the first module (M1) that included mostly cell-type

227 invariant cCREs, the remaining 41 modules displayed highly cell-type restricted
228 accessibility (Fig. 3g, Supplementary Table 7, 8). These cell-type restricted modules were
229 enriched for transcription factor motifs recognized by known transcriptional regulators for
230 such as the SOX family factors for oligodendrocytes OGC (Supplementary Table 9)[46,47].
231 We also found strong enrichment for the known olfactory neuron regulator LIM homeobox
232 factor LHX2 in module M5 which was associated with GABAergic neurons in the olfactory
233 bulb (OBGA1) (Supplementary Table 9)[48].
234

235 **Integrative analysis of chromatin accessibility and gene expression across mouse**
236 **brain cell types**

237

238 To dissect the transcriptional regulatory programs responsible for cell-type specific gene
239 expression patterns in the mouse cerebrum, we carried out integrative analysis combining
240 the single nucleus ATAC-seq collected in the current study with single cell RNA-seq data
241 from matched brain regions[1]. Enhancers can be linked to putative target genes by
242 measuring co-accessibility between enhancer and promoter regions of putative target
243 genes and co-accessible sites tend to be in physical proximity in the nucleus[49]. Thus, we
244 first identified pairs of co-accessible cCREs in each cell cluster using Cicero[49] and inferred
245 candidate target promoters for distal cCRE located more than 1 kbp away from annotated
246 transcription start sites in the mouse genome (Fig. 4a, see **Methods**)[40]. We determined
247 a total of 813,638 pairs of cCREs within 500 kbp of each other, and connected 261,204
248 cCREs to promoters of 12,722 genes (Supplementary Table 10).
249

250 Next, we sought to identify the subset of cCREs that might increase expression of putative
251 target genes and therefore function as putative enhancers in neuronal or non-neuronal
252 types. To this end, we first identified distal cCREs for which chromatin accessibility was
253 correlated with transcriptional variation of the linked genes in the joint cell clusters as
254 defined above (Fig. 2a). We computed Pearson correlation coefficients (PCC) between
255 the normalized chromatin accessibility signals at each cCRE and the RNA expression of
256 the predicted target genes across these cell clusters (Fig. 4a, b). As a control, we
257 randomly shuffled the cCREs and the putative target genes and computed the PCC of

258 the shuffled cCRE-gene pairs (Fig. 4b, see **Methods**). This analysis revealed a total of
259 129,404 pairs of positively correlated cCRE (putative enhancers) and genes at an
260 empirically defined significance threshold of FDR < 0.01 (Supplementary Table 10).
261 These included 86,850 putative enhancers and 10,604 genes (Fig. 4b). The median
262 distances between the putative enhancers and the target promoters was 178,911 bp (Fig.
263 4c). Each promoter region was assigned to a median of 7 putative enhancers (Fig. 4d),
264 and each putative enhancer was assigned to one gene on average. To investigate how
265 the putative enhancers may direct cell-type specific gene expression, we further classified
266 them into 38 modules, by applying non-negative matrix factorization to the matrix of
267 normalized chromatin accessibility across the above joint cell clusters. The putative
268 enhancers in each module displayed a similar pattern of chromatin accessibility across
269 cell clusters to expression of putative target genes (Fig 4e, Supplementary Table 11 and
270 13). This analysis revealed a large group of 12,740 putative enhancers linked to 6,373
271 genes expressed at a higher level in all neuronal cell clusters than in all non-neuronal cell
272 types (module M1, top, Fig. 4e). It also uncovered modules of enhancer-gene pairs that
273 were active in a more restricted manner (modules M2 to M38, Fig 4e). For example,
274 module M33 was associated with perivascular microglia (PVM). Genes linked to putative
275 enhancers in this module were related to immune gene and the putative enhancers were
276 enriched for the binding motif for ETS-factor PU.1, a known master transcriptional
277 regulator of this cell lineage (Fig. 4e, f, Supplementary Table 13 and 14)[50]. Similarly,
278 module M35 was strongly associated with oligodendrocytes (OGC) and the putative
279 enhancers in this module were enriched for motifs recognized by the SOX family of
280 transcription factors (Fig. 4e, f, Supplementary Table 14)[47]. We also identified module
281 M15 associated with several cortical glutamatergic neurons (IT.L2/3,IT.L4,IT.L5/6,IT.L6),
282 in which the putative enhancers were enriched for sequence motifs recognized by the
283 bHLH factors NEUROD1 (Fig. 4e, f, Supplementary Table 14)[51]. Another example was
284 module M10 associated with medium spiny neurons (MSN1 and 2), in which putative
285 enhancers were enriched for motif for the MEIS factors, which play an important role in
286 establishing the striatal inhibitory neurons (Fig. 4e, f, Supplementary Table 14)[52]. Notably
287 and in stark contrast to the striking differences at putative enhancers, the chromatin
288 accessibility at promoter regions showed little variation across cell types (Fig. 4g). This is

289 consistent with the paradigm that cell-type-specific gene expression patterns are largely
290 established by distal enhancer elements[42,53].
291
292 **Distinct groups of transcription factors act at the enhancers and promoters in the**
293 **pan-neuronal gene module**
294
295 As shown above, genes associated with module M1 are preferentially expressed in both
296 glutamatergic and GABAergic neurons, but not in glial cell types (Fig. 4e). *De novo* motif
297 enrichment analysis of the 12,740 cCREs or putative enhancers in this module showed
298 dramatic enrichment of sequence motifs recognized by the transcription factors CTCF,
299 RFX, MEF2 (Supplementary Table 15), as well as many known motifs for other
300 transcription factors (Fig. 4f, Fig. 5a, Supplementary Table 14). CTCF is a ubiquitously
301 expressed DNA binding protein with a well-established role in transcriptional insulation
302 and chromatin organization[54]. Recently, it was recognized that CTCF also promotes
303 neurogenesis by binding to promoters and enhancers of proto-cadherin alpha gene
304 cluster and facilitating enhancer-promoter contacts[55,56]. In the current study we found
305 putative enhancers with CTCF motif for 2,601 genes that were broadly expressed in both
306 inhibitory and excitatory neurons (Fig. 4e, 5b), and involved in multiple neural processes
307 including axon guidance, regulation of axonogenesis, and synaptic transmission (Fig. 5c,
308 Supplementary Table 16). For example, we found one CTFC peak overlapping a distal
309 cCRE positively correlated with expression of *Lgi1* which encodes a protein involved in
310 regulation of presynaptic transmission[57] (Fig 5d). The RFX family of transcription factors
311 are best known to regulate the genes involved in cilium assembly pathways[58].
312 Unexpectedly, we found the RFX binding sequence motif to be strongly enriched at the
313 putative enhancers for genes encoding proteins that participate in postsynaptic
314 transmission, postsynaptic transmembrane potential, mitochondrion distribution, and
315 receptor localization to synapse (Fig. 5c, Supplementary Table 16). For example, we
316 found RFX motif in a distal cCRE positively correlated with expression of *Kif5a* which
317 encodes a protein essential for $GABA_A$ receptor transport (Fig. 5e)[59]. This observation
318 thus suggests a role for RFX family of transcription factors in regulation of synaptic
319 transmission pathways in mammals. Similar to CTCF and RFX, the MEF2 family

320 transcription factors have also been shown to play roles in neurodevelopment and mental
321 disorders[60]. Consistent with this, the genes associated with putative enhancers containing
322 MEF2 binding motifs were selectively enriched for those participating in positive
323 regulation of synaptic transmission, long-term synaptic potentiation, and axonogenesis
324 (Fig. 5c, Supplementary Table 16). For example, we found a distal cCRE harboring a
325 MEF2 motif positively correlated with expression of *Cacng2* which encodes a calcium
326 channel subunit that is involved in regulating gating and trafficking of glutamate receptors
327 (Fig 5f)[61]. Notably, in types with high accessibility levels, cCREs and promoters of putative
328 target genes also showed low levels of DNA methylation (Fig. 5d-f, see companion
329 manuscript by Liu, Zhou et al. 2020[37]).
330
331 Interestingly, motif analysis of promoters of genes linked to cCREs in the module M1
332 revealed the potential role of very different classes of transcription factors in neuronal
333 gene expression. Among the top ranked transcription factor motifs are those recognized
334 by CREB (cAMP-response elements binding protein), NF-κB, STAT3 and CLOCK
335 transcription factors (Supplementary Table 17). Enrichment of CREB binding motif in
336 module M1 gene promoters is consistent with its well-documented role in synaptic activity-
337 dependent gene regulation and neural plasticity[62,63]. Enrichment of NF-κB[64], STAT3[65]
338 and CLOCK[66] binding motifs in the module M1 gene promoters is interesting, too, as it
339 suggests potential roles for additional extrinsic signaling pathways, i.e. stress, interferon,
340 circadian rhythm, respectively, in the regulation of gene expression in neurons.
341
342 **Non-coding variants associated with neurological traits and diseases are enriched**
343 **in the human orthologs of the mouse brain cCREs in a cell type-specific manner**
344
345 Genome-wide association studies (GWASs) have identified genetic variants associated
346 with many neurological disease and traits, but interpreting the results have been
347 challenging because most variants are located in non-coding parts of the genome that
348 often lack functional annotations and even when a non-coding regulatory sequence is
349 annotated, its cell-type specificity is often not well known[67,68]. To test if our maps of
350 cCREs in different mouse brain cell type could assist the interpretation of non-coding risk

351 variants of neurological diseases, we identified orthologs of the mouse cCREs in the
352 human genome by performing reciprocal homology search[69]. For this analysis, we found
353 that for 69.2% of the cCREs, human genome sequences with high similarity could be
354 identified (> 50 % of bases lifted over to the human genome, Fig. 6a). Supporting the
355 function of the human orthologs of the mouse brain cCREs, 83.0% of them overlapped
356 with representative DNase hypersensitivity sites (rDHSs) in the human genome[41,42]. Next,
357 we performed linkage disequilibrium (LD) score regression (LDSC)[70] to determine
358 associations between different brain regions and distinct GWAS traits (Fig. 6b, Extended
359 Data Figure 5). We found a significant enrichment of cCREs from 36 out of 45 brain
360 regions for risk variants of Schizophrenia (Fig. 6b). In fact, most neurological traits
361 showed widespread enrichment across brain regions, but a few like ADHD (Attention
362 deficit hyperactivity disorder) showed some regional enrichment patterns (Fig. 6b).
363
364 We also performed LDSC analysis and found significant associations between 20
365 neuronal and non-neuronal traits and cCREs found in one or more major cell types (Fig.
366 6c). We observed widespread and strong enrichment of genetic variants linked to
367 psychiatric and cognitive traits such as major depressive disorder, intelligence,
368 neuroticism, educational attainment, bipolar disorder and schizophrenia in cCREs across
369 various neuronal cell types (Fig. 6c). Other neurological traits, such as attention deficit
370 hyperactivity disorder, chronotype, autism spectrum disorder and insomnia were
371 associated with specific neuronal cell-types in cerebral nuclei and hippocampus (Fig. 6c).
372 Schizophrenia risk variants were not only enriched in cCREs in all excitatory neurons, but
373 also in certain inhibitory neuron sub-types (Fig. 6c)[71]. The strongest enrichment of
374 heritability for bipolar disorder was in elements mapped in excitatory neurons from
375 isocortex (Fig. 6c). Risk variants of tobacco use disorder showed significant enrichment
376 in the cell types from striatum, a cerebral nucleus previously implicated in addiction (Fig.
377 6c)[72]. Interestingly, cCREs of non-neuronal mesenchymal cells were not enriched for
378 neurological traits but showed enrichment for cardiovascular traits such as coronary
379 artery disease (Fig. 6c). Similarly, variants associated with height were also significant in
380 these cell types (Fig. 6c). cCREs in microglia were significantly enriched for variants
381 related to immunological traits like inflammatory bowel disease, Crohn's disease and

382 multiple sclerosis (Fig. 6c). Notably, most of these patterns were not apparent in the peaks
383 called on aggregated bulk profiles from brain regions (Fig. 6b, Extended Data Fig. 5),
384 demonstrating the value of cell type resolved open chromatin maps which was also
385 highlighted by a recent study using single cell ATAC-seq profiling of human brain which
386 focusing on Alzheimers' and Parkinson's disease[73].
387
388 **DISCUSSION**
389
390 Understanding the cellular and molecular genetic basis of brain circuit operations is one
391 of the grand challenges in the 21st century[12,74]. In-depth knowledge of the transcriptional
392 regulatory program in brain cells would not only improve our understanding of the
393 molecular inner workings of neurons and non-neuronal cells, but could also shed new
394 light into the pathogenesis of a spectrum of neuropsychiatric diseases[75]. In the current
395 study, we report comprehensive profiling of chromatin accessibility at single cell resolution
396 in the mouse cerebrum. The chromatin accessibility maps of 491,818 cCREs, probed in
397 813,799 nuclei and 160 sub-types representing multiple cerebral cortical areas and
398 subcortical structures, are the largest of its kind so far. The cell type annotation based on
399 open chromatin landscape showed strong alignment with those defined based on single
400 cell transcriptomics[1], which allowed us to jointly analyse the two molecular modalities
401 across major cell types in the brain and identify putative enhancers for over 10,604 genes
402 expressed in the mouse cerebrum. We further characterized the cell-type-specific
403 activities of putative enhancers, inferred their potential target genes, and predicted
404 transcription factors that act through these candidate enhancers to regulate specific gene
405 modules and molecular pathways.
406
407 We identified one large group of putative enhancers for genes that are broadly expressed
408 in GABAergic and glutamatergic neurons, but at low levels or are silenced in all glial cell
409 types. A significant fraction of these cCREs are bound by CTCF in the mouse brain
410 (Figure 5)[43]. Recently, it was shown that CTCF is involved in promoter selection in the
411 proto-cadherin gene cluster by promoting enhancer-promoter looping[55,56]. Our data now
412 suggest that CTCF could regulate a broader set of neuronal genes than previously

413 demonstrated[55,76], which need to be verified in future experiments. In addition, the RFX
414 family of transcription factors was described to regulate cilia in sensory neurons[58]. Our
415 data suggest a more widespread role for RFX family of transcription factors in the brain
416 in regulation of synaptic transmission. Consistent with this proposal, deletion of *Rfx4*
417 gene in mouse was shown to severely disrupt neural development[77]. We have previously
418 shown that RFX motif was enriched in elements that were more accessible after birth
419 compared to prenatal time points in both GABAergic and glutamatergic neuronal types[31].
420 RFX was also found to be strongly enriched in mouse forebrain enhancers with increased
421 activity after birth[78]. Similar to CTCF and RFX, the MEF2 family transcription factors have
422 been demonstrated to play roles in neurodevelopment and mental disorders[60]. The MEF2
423 motif was enriched at enhancers with higher chromatin accessibility in late forebrain
424 development in mice coinciding with synapse formation[78].
425
426 Thus, our results are consistent with the notion that cell identity is encoded in distal
427 enhancer sequences, executed by sequence-specific transcription factors during different
428 stages of brain development. The reference maps of cCREs for the mouse cerebrum
429 would not only help to understand the mechanisms of gene regulation in different brain
430 cell types, but also enable targeting and purification of specific neuronal or non-neuronal
431 cell types or targeted gene therapy[28,79]. In addition, the maps of cCREs in the mouse
432 brain would also assist the interpretation of non-coding risk variants associated with
433 neurological diseases[73]. The datasets described here represent a rich resource for the
434 neuroscience community to understand the molecular patterns underlying diversification
435 of brain cell types in complementation to other molecular and anatomical data.

ACKNOWLEDGEMENTS

We thank Josh Huang (CSHL) for critical reading of the manuscript. We thank Drs. Ramya Raviram, Yanxiao Zhang, Guoqiang Li and James Hocker for discussions and all other members of the Ren laboratory for valuable inputs. We would like to extend our gratitude to the QB3 Macrolab at UC Berkeley for purification of the Tn5 transposase. This study was supported by NIH Grant U19MH11483. Work at the Center for Epigenomics was also supported by the UC San Diego School of Medicine.

AUTHOR CONTRIBUTIONS

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

COMPETING INTERESTS

B.R. is a co-founder and consultant of Arima Genomics, Inc.. J.R.E is on the scientific advisory board of Zymo Research, Inc

Figure 1

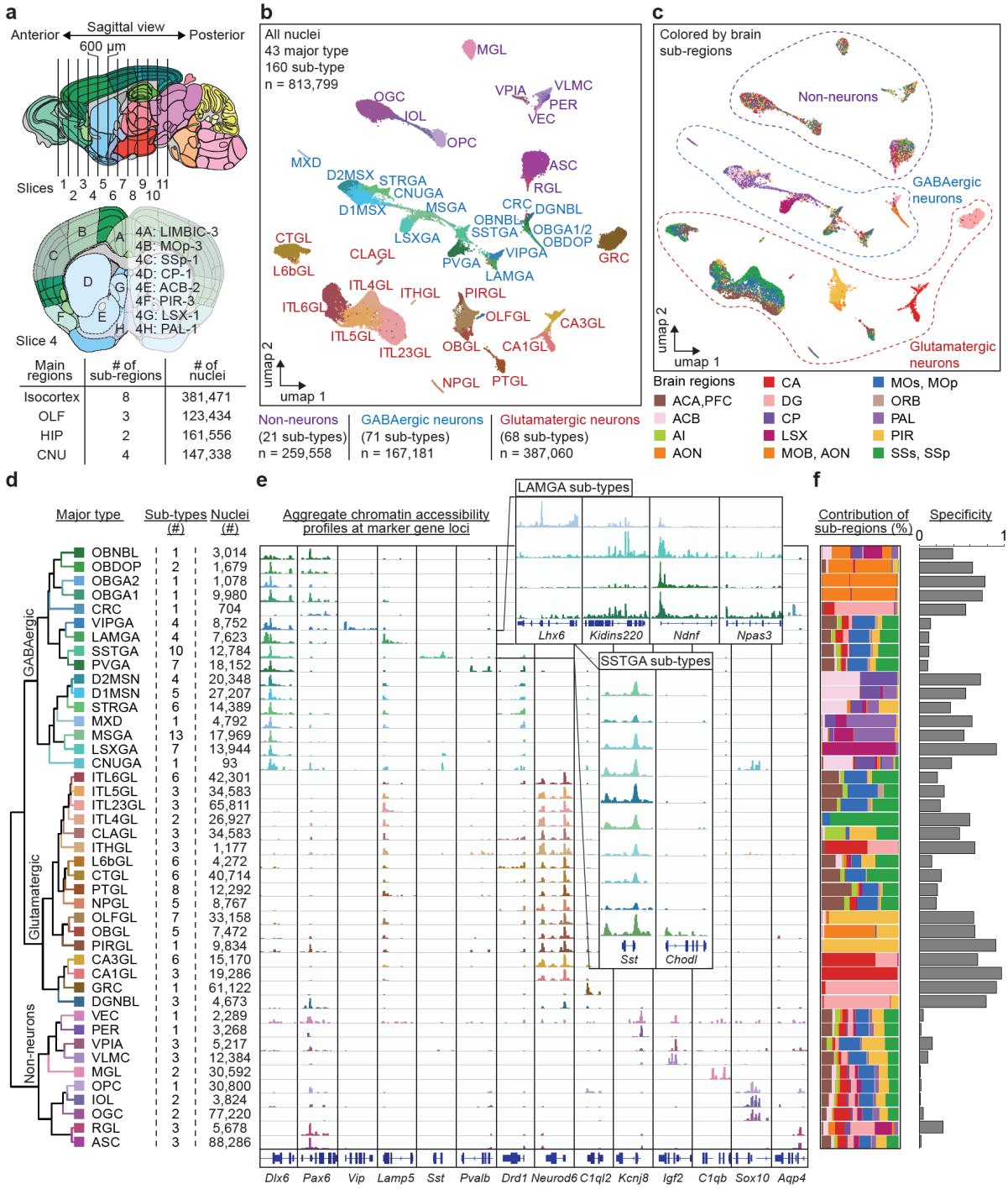


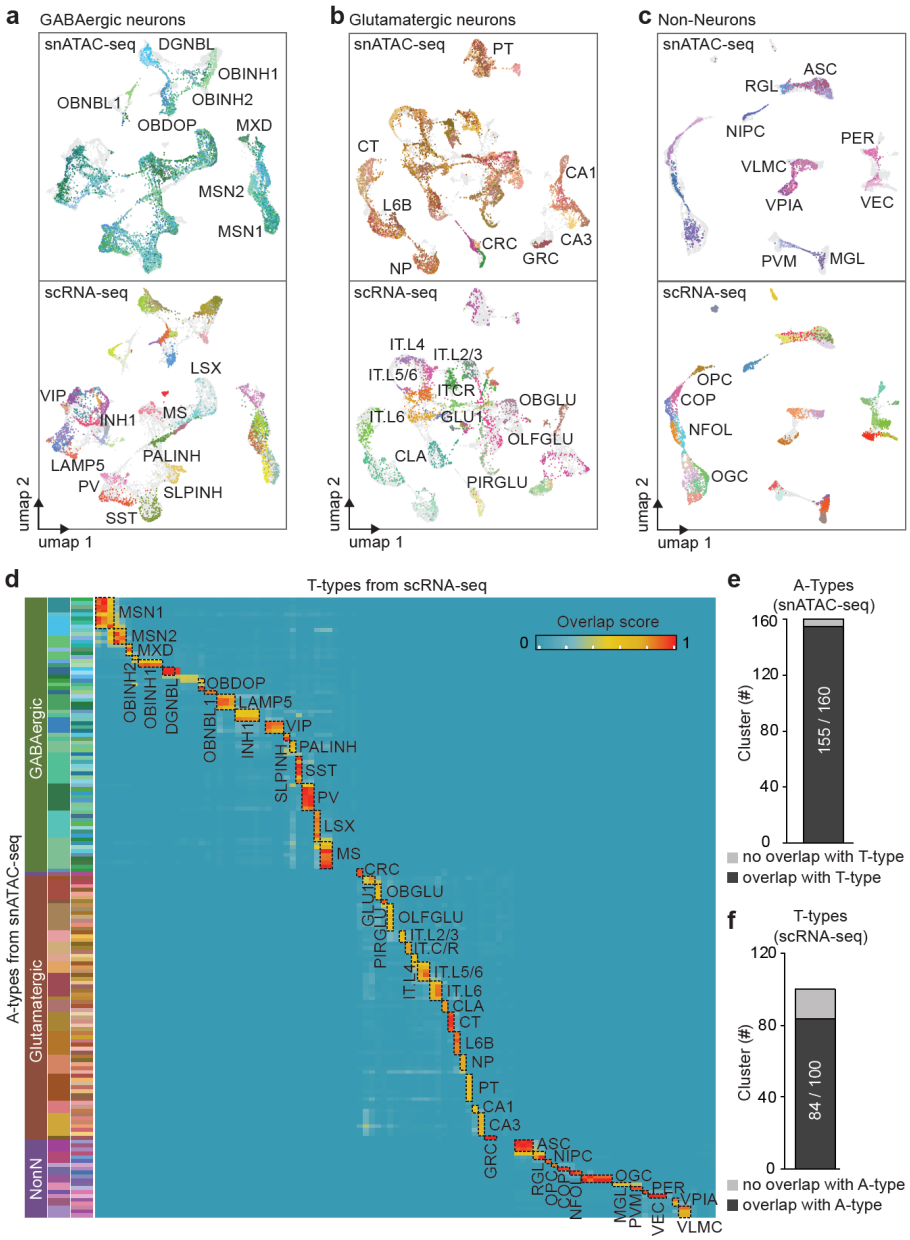
Figure 1: Chromatin accessibility profiling, clustering and annotation of over 800,000 nuclei in adult mouse cerebrum.

a Schematic of sample dissection strategy. The brain regions studied were dissected from 600 µm-thick coronal slices generated from 8-week-old mouse brains (top panel). A total

460 of 45 regions were dissected according to the Allen Brain Reference Atlas[32]. Shown is
461 the frontal view of slice 4 and the dissected brain regions (middle panel, alphabetically
462 labeled). For example, dissection region 4B: MOp-3 denotes part 3 of the primary motor
463 cortex (MOp) region which corresponds to region B from slice 4. The dissected regions
464 represent 17 sub regions from four main brain areas: isocortex, olfactory bulb (OLF),
465 hippocampus (HIP) and cerebral nuclei (CNU). A detailed list of regions can be found in
466 Supplementary Table 1. **b** Uniform manifold approximation and projection (UMAP)[80]
467 embedding and clustering analysis of snATAC-seq data from 813,799 nuclei, revealing
468 43 major types and 160 sub-types assigned to non-neuronal (21, purple), GABAergic (71,
469 blue/green) and Glutamatergic neuron clusters (68, red/brown). Clusters were annotated
470 based on chromatin accessibility at promoter regions and gene bodies of canonical
471 marker genes. Each dot in the UMPA represents a nucleus and the nuclei are colored
472 and labeled by major cluster ID. For example, ITL23GL denotes excitatory neurons from
473 cortex layer 2/3. For a full list and description of cluster labels see Supplementary Table
474 3. **c** Same embedding as in **b** but colored by sub-regions, e.g. SSp (primary
475 somatosensory cortex). For a full list of brain regions see Supplementary Table 1. Dotted
476 lines demark major cell classes. **d** Hierarchical organization of cell clusters based on
477 chromatin accessibility depicting level 1 and 2 clusters (left panel). Each major type
478 represents 1-10 sub-types (middle). Total number of nuclei per major type ranged from
479 93 to 88,286 nuclei (right). For a full list and description of cluster labels see
480 Supplementary Table 2. **e** Genome browser tracks of aggregate chromatin accessibility
481 profiles for each major cell cluster at selected marker gene loci that were used for cell
482 cluster annotation. The inlets highlight the 10 subtypes of *Sst+* (SSTGA) inhibitory
483 neurons including Chodl-Nos1 neurons (bottom track in SSTGA inlet)[35] and 4 subtypes
484 of *Lamp5+* (LAMGA) inhibitory neurons including *Lhx6* positive putative chandelier like
485 cells (top track in LAMGA inlet)[3]. For a full list and description of cluster labels see
486 Supplementary Table 3. **f** Bar chart representing the relative contributions of sub-regions
487 to major clusters. Color code is the same as in **b**. Based on these relative contributions,
488 an entropy-based specificity score was calculated to indicate if a cluster was restricted to
489 one or a few of the profiled regions (high score) or broadly distributed (low score). Several
490 neuronal types showed high regional specificity whereas non-neuronal types were mostly

491 unspecific. Glutamatergic neurons showed higher regional specificity than GABAergic
492 neurons consistent with transcriptomic analysis[3].

Figure 2



493

494 **Figure 2: Alignment of chromatin-based cell clustering to scRNA-seq-based cell**
495 **type taxonomy.**

496 **a-c** SnATAC-seq data were integrated with scRNA-seq profiles from matched brain
497 regions[1] using the Seurat package[38]. Uniform manifold approximation and projections
498 (UMAPs)[80] illustrate co-embedding of snATAC-seq and scRNA-seq datasets from three
499 main cell classes, namely **c** GABAergic neurons, **d** glutamatergic neurons, and **e** non-
500 neurons (top: colored by snATAC-seq clusters (A-type), bottom: colored by scRNA-seq

clusters (T-type); labelling denotes integrated A/T-types). **d** Heatmap illustrating the overlap between A-type and T-type cell cluster annotations. Each row represents a snATAC-seq sub-type (total of 160 A-types) and each column represents scRNA-seq cluster (total of 100 T-types). The overlap between original clusters and the joint cluster was calculated (overlap score) and plotted on the heatmap. Joint clusters with an overlap score of >0.5 are highlighted using black dashed line and labeled with joint cluster ID. For a full list of cell type labels and description see Supplementary Table 4. **e, f** Bar plots indicating the number of clusters that overlapped (dark grey) and that did not overlap (light grey) with clusters from the other modality. **e** 155 out of 160 A-types had a matching T-type. **f** 84 out of 100 T-types had a matching A-type.

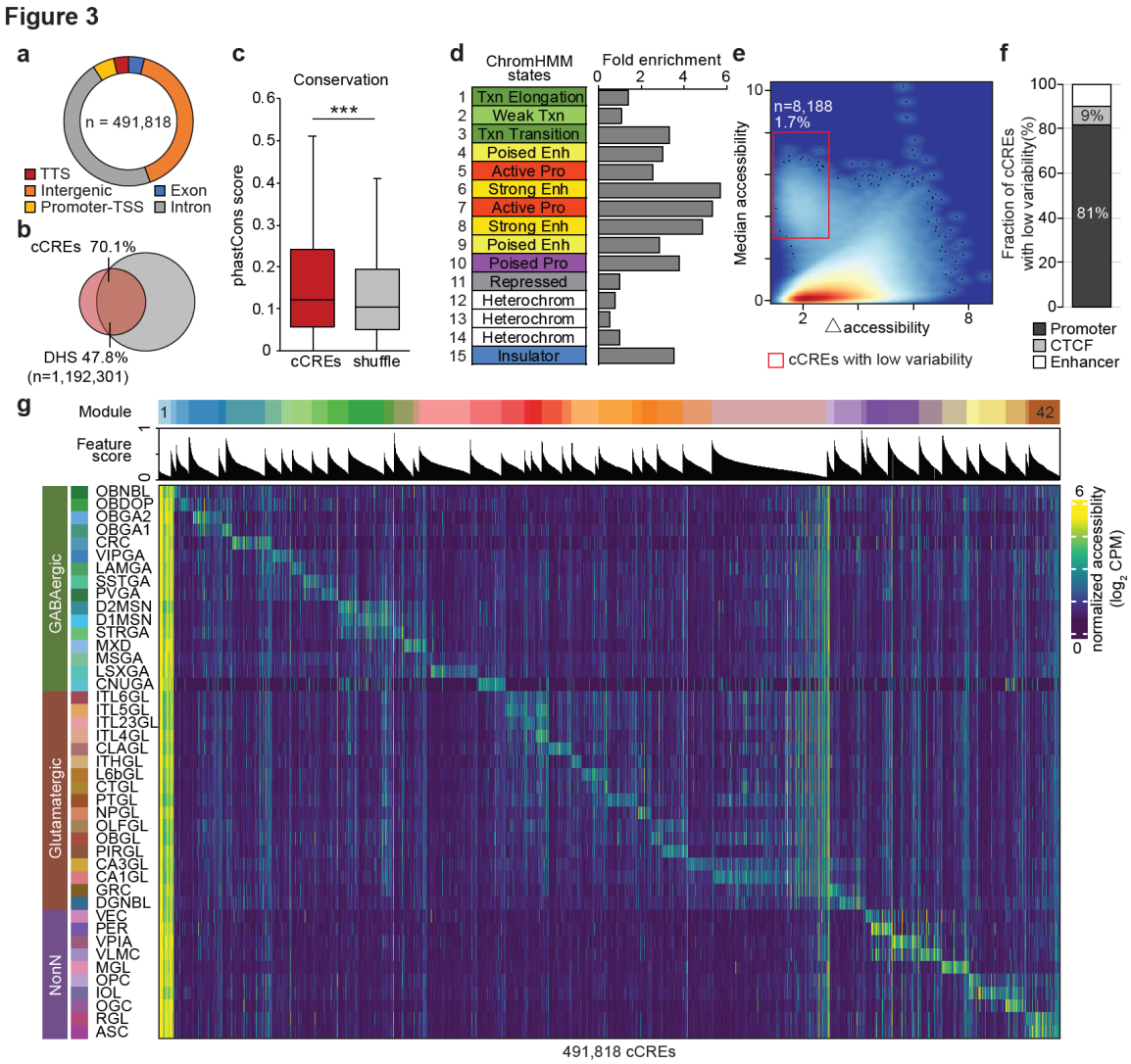


Figure 3: Characterization of candidate *cis* regulatory elements identified in mouse cerebral cell types.

a Fraction of the identified cCREs that overlap with annotated transcriptional start sites (TSS), introns, exons, transcriptional termination sites (TTS) and intergenic regions in the mouse genome. **b** Venn diagram showing the overlap between cCREs and DNase hypersensitive sites (DHS) from developmental and adult mouse tissue from the SCREEN database[42]. **c** Box-Whisker plot showing that sequence conservation measured by PhastCons score[81] is higher for cCREs than the controls consisting of GC-matched random genomic sequences (*** p <0.001, Wilcoxon rank sum test, the box is drawn from lower quartile (Q1) to upper quartile (Q3) with a horizontal line drawn in the middle to denote the median, whiskers with maximum 1.5 IQR). **d** Enrichment analysis of cCREs

524 with a 15-state ChromHMM model[45] in the mouse brain chromatin[43]. **e** Density map
525 showing two main groups of elements based on the median accessibility and the range
526 of chromatin accessibility variation (maximum – minimum) across cell clusters for each
527 cCRE. Each dot represents a cCRE. Red box highlights elements with low chromatin
528 accessibility variability across clusters. **f** 81 % of sites with low variability (red box in **e**)
529 overlapped promoters, 10 % enhancers and 9 % CTCF regions. **g** Heatmap showing
530 association of 43 major cell types (rows) with 42 *cis* regulatory modules (top). Each
531 column represents one of 491,818 cCREs. These cCREs were combined into *cis*
532 regulatory modules based on accessibility patterns across major cell types. For each
533 cCRE a feature score was calculated to represent the specificity for a given module.
534 Module 1 comprised invariable elements and was enriched for promoters. For a full list
535 and description of cell cluster labels see Supplementary Table 3, for a full list of cluster-
536 module association see Supplementary Table 7 and for association of cCREs to modules
537 see Supplementary Table 8. CPM: counts per million.
538

Figure 4

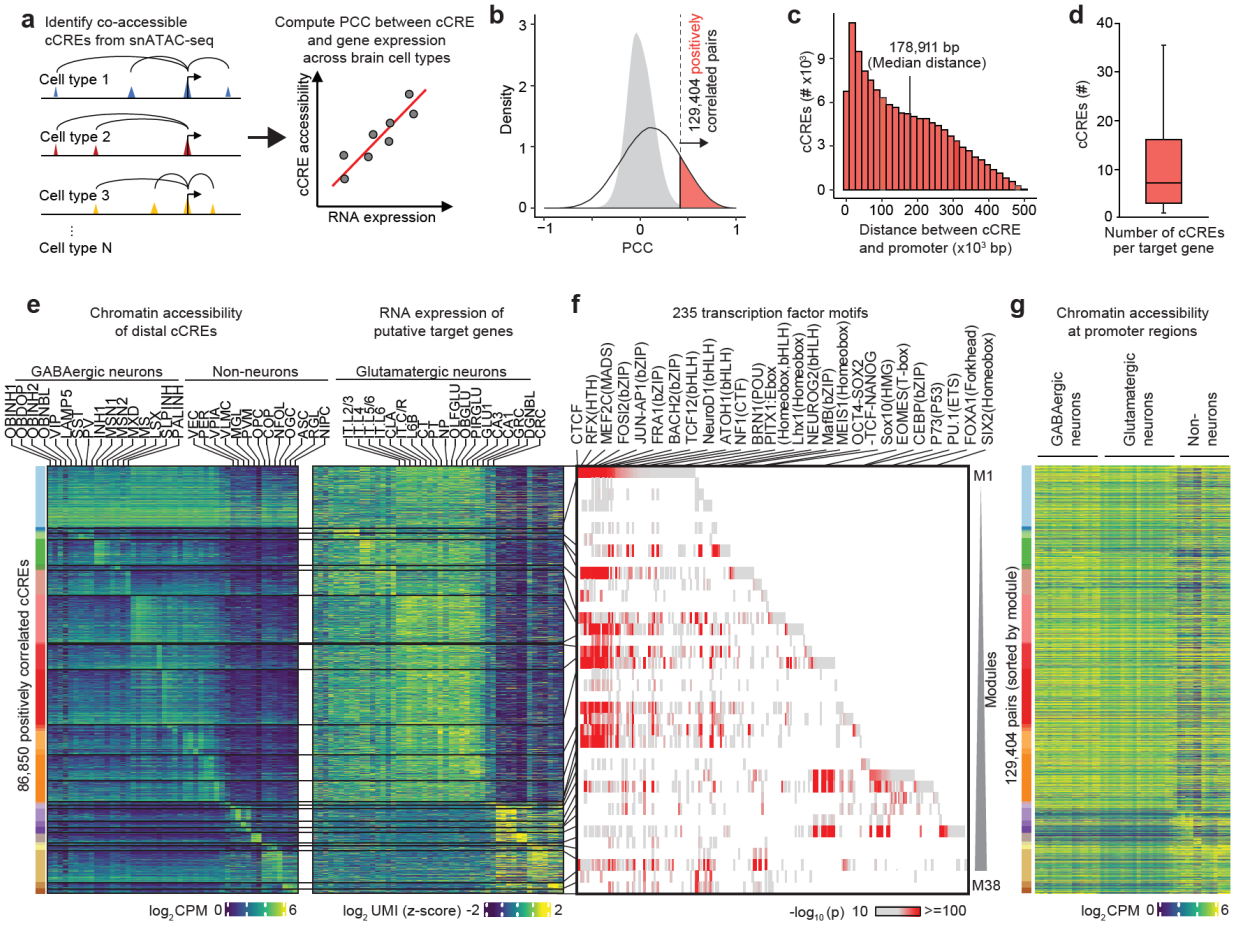


Figure 4: Identification and characterization of putative enhancer-gene pairs. a Schematic overview of the computational strategy to identify cCREs that are positively correlated with transcription of target genes. The cCREs were first assigned to putative target gene promoters in specific cell clusters using co-accessibility analysis with Cicero[49]. Next, chromatin accessibility at cCREs was correlated with RNA-seq signals of the putative target gene across different cell clusters (PCC: Pearson correlation coefficient). **b** Detection of putative enhancer-gene pairs. 129,404 pairs of positively correlated cCRE and genes (highlighted in orange) were identified using an empirically defined significance threshold of FDR<0.01 (see **Methods**). Grey filled curve shows distribution of PCC for randomly shuffled cCRE-gene pairs. **c** Histogram illustrating distance between positively correlated distal cCRE and putative target gene promoters. Median distance was 178,911 bp. **d** Box-Whisker plot showing that genes were linked with a median of 7 putative enhancers (box is drawn from Q1 to Q3 with a horizontal line

553 drawn in the middle to denote the median, whiskers with maximum 1.5 IQR). **e** Heatmap
554 of chromatin accessibility of 86,850 putative enhancers across cell clusters (left) and
555 expression of 10,604 linked genes (right). Note genes are displayed for each putative
556 enhancer separately. For association of modules with cell types see Supplementary Table
557 11 and association of individual putative enhancer with modules see Supplementary
558 Table 13. CPM: counts per million, UMI: unique molecular identifier. **f** Enrichment of
559 known transcription factor motifs in distinct enhancer-gene modules. Displayed are known
560 motifs from HOMER[46] with enrichment p-value $<10^{-10}$. Motifs were sorted based on
561 module. For full list see Supplementary Table 14. **g** Accessibility at promoter regions
562 across joint A/T-types, same order as **e**.
563

Figure 5

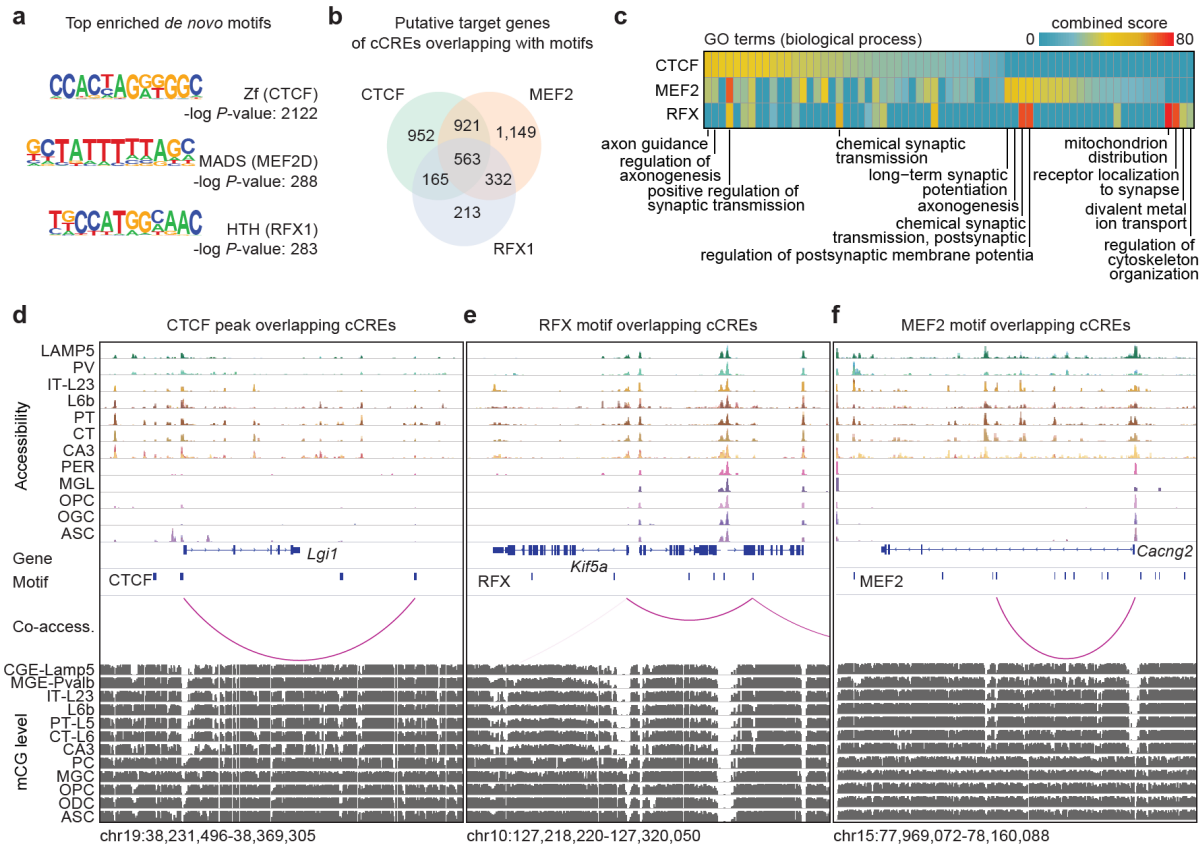


Figure 5: Transcription factors involved in a pan neuronal gene regulatory program.

a Enrichment of sequence motifs for CTCF, MEF2 and RFX from de novo motif search in the putative enhancers of module M1 using HOMER[46]**.** For a full list see Supplementary Table 16. **b** Venn diagram illustrating the overlap of putative target genes of cCREs containing binding sites for MEF2, RFX and CTCF, respectively. **c** Gene ontology (GO) analysis of the putative target genes of each factor in module M1 was performed using Enrichr[82]. The combined score is the product of the computed p value using the Fisher exact test and the z-score of the deviation from the expected rank[82]. **d-f** Examples distal cCRE overlapping peaks/motifs and positively correlated putative target genes. For CTCF, cCREs were intersected with peak calls from ChIP-seq experiments in the adult mouse brain[43] (**d**) and cCREs overlapping RFX (**e**) and MEF2 (**f**) were identified using de novo motif search in HOMER[46]. Genome browser tracks displaying chromatin accessibility, mCG methylation levels (see companion manuscript by Liu, Zhou et al. 2020[37]) and positively correlated cCRE and genes pairs.

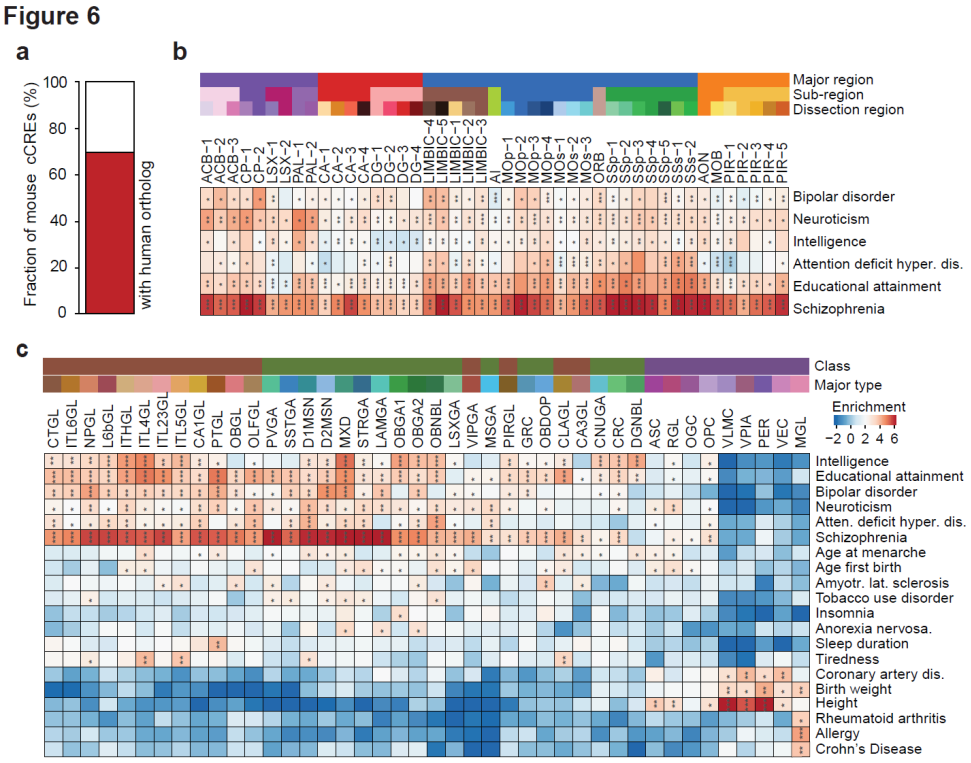
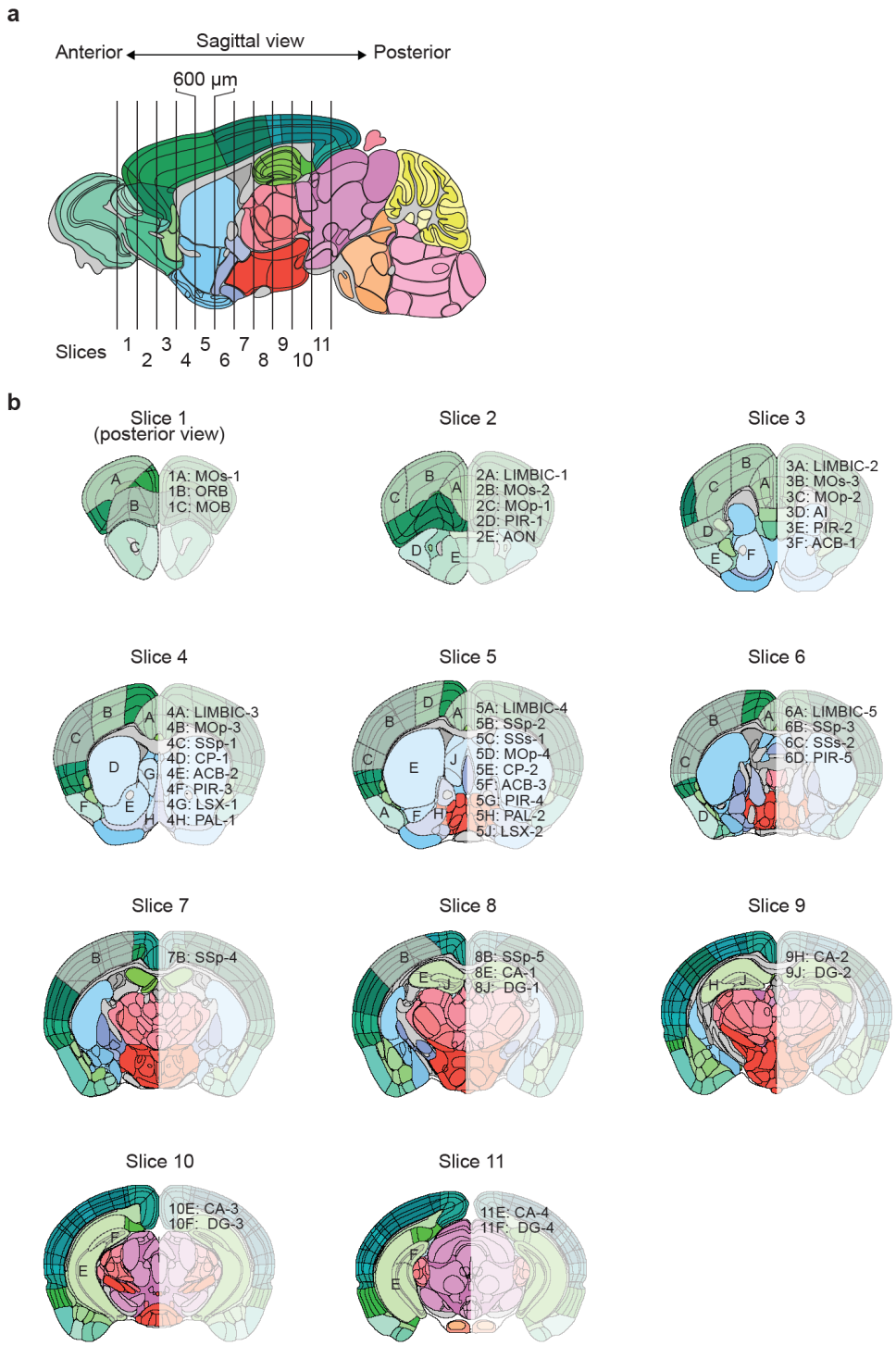


Figure 6: Association of different brain regions and cell types with risk variants for neurological diseases and traits.

a For 69.2 % of cCREs identified in the current study, we found a human ortholog (> 50 % of bases lifted over to the human genome). **b** Brain-region-specific enrichment of sequence variants associated with indicated neurological traits and diseases (* FDR < 0.05, ** FDR < 0.01, ***FDR < 0.001). Displayed are all regions and all tested phenotypes with at least one significant association. **c** Enrichment of sequence variants associated with the indicated traits/disease in the human orthologs of cCREs in major mouse cerebral cell types (* FDR < 0.05, ** FDR < 0.01, ***FDR < 0.001). Displayed are all major cell clusters and tested traits/diseases with at least one significant association (FDR < 0.05). A detailed list of regions can be found in Supplementary Table 1 and a full list of cell cluster labels can be found in Supplementary Table 3.

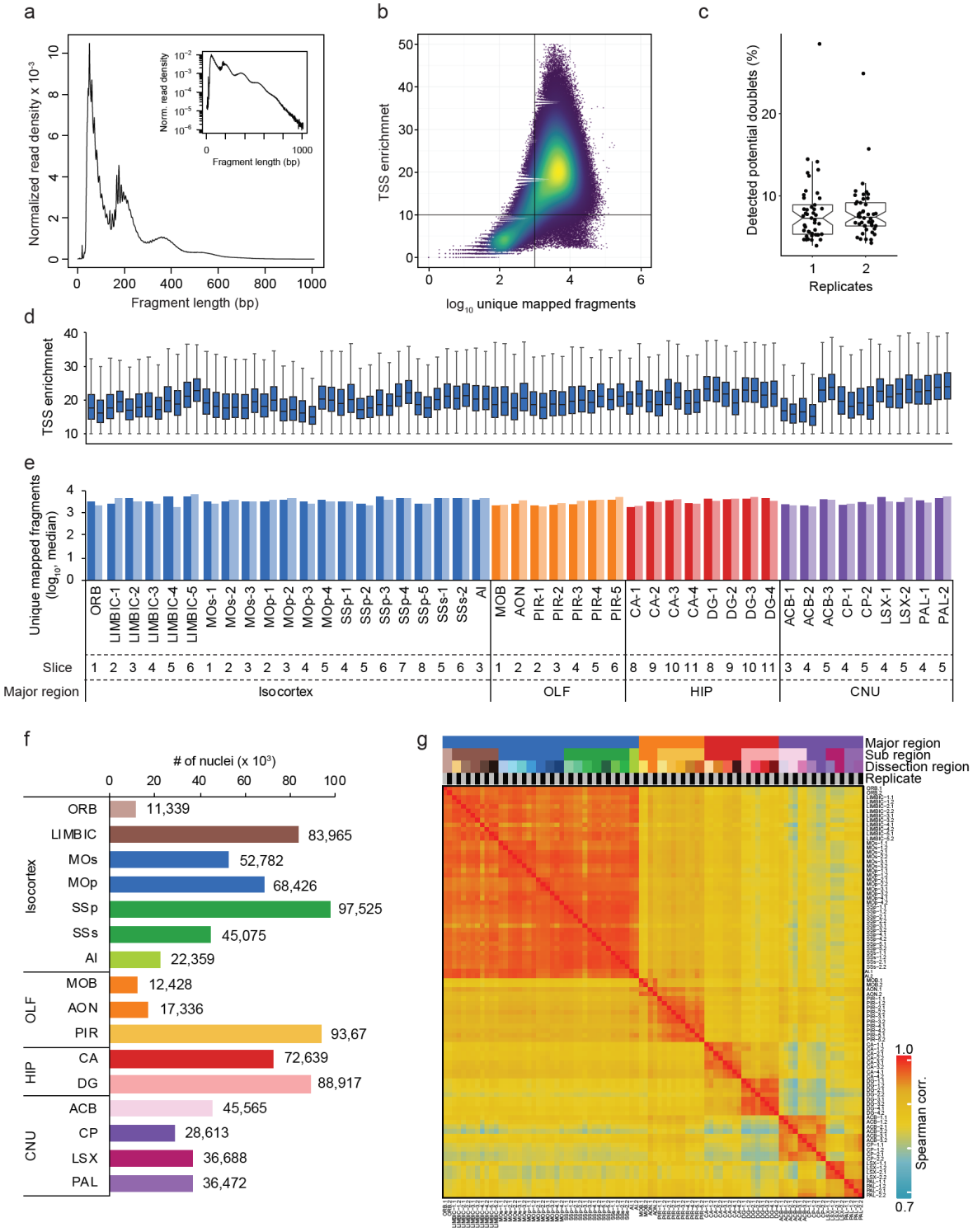
Extended Data Figure 1



Extended Data Figure 1: Maps of mouse brain regions that were dissected in the current study. a Schematic of brain sample dissection strategy. Mouse brains were cut into 600 µm thick coronal slices; **b** 45 regions were dissected from eleven coronal slices according to the Allen Brain Reference Atlas[32]. Shown is the frontal view of slice 1-11 and

598 isolated regions. For example, dissection region 1A: MOs-1 denotes part 1 of the
599 secondary motor cortex (MOs) region which corresponds to region A from slice 1. A
600 detailed list of regions can be found in Supplementary Table 1.
601

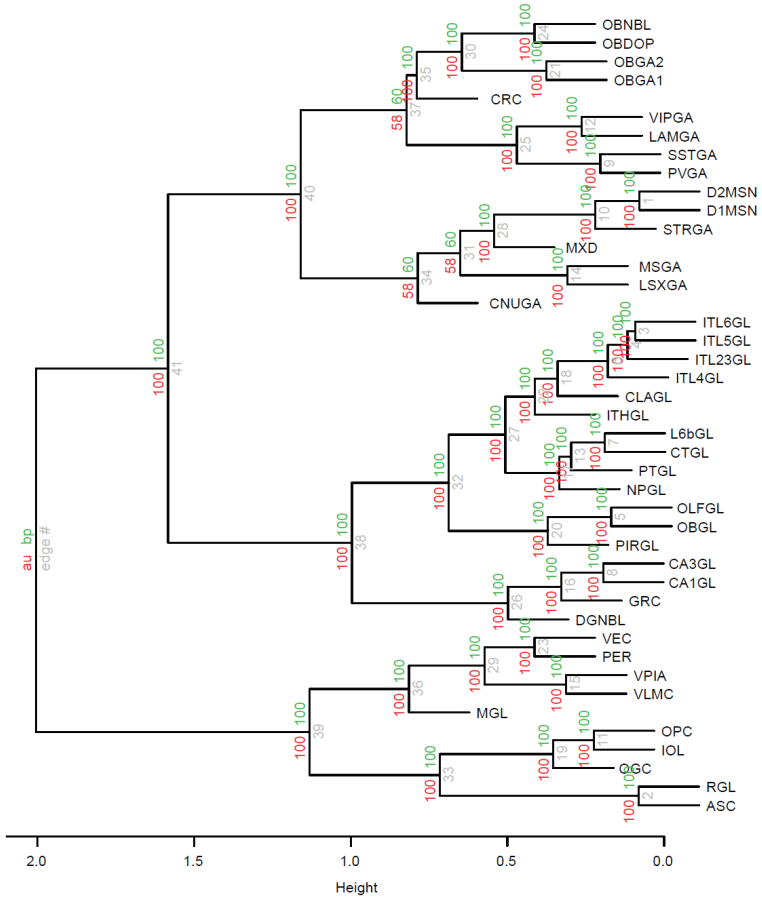
Extended Data Figure 2



Extended Data Figure 2: Quality metrics of snATAC-seq datasets. a Fragment size distribution of a typical snATAC-seq library. **b** Dot-blot illustrating fragments per nucleus and individual TSS (transcriptional start site) enrichment. Nuclei in the upper right

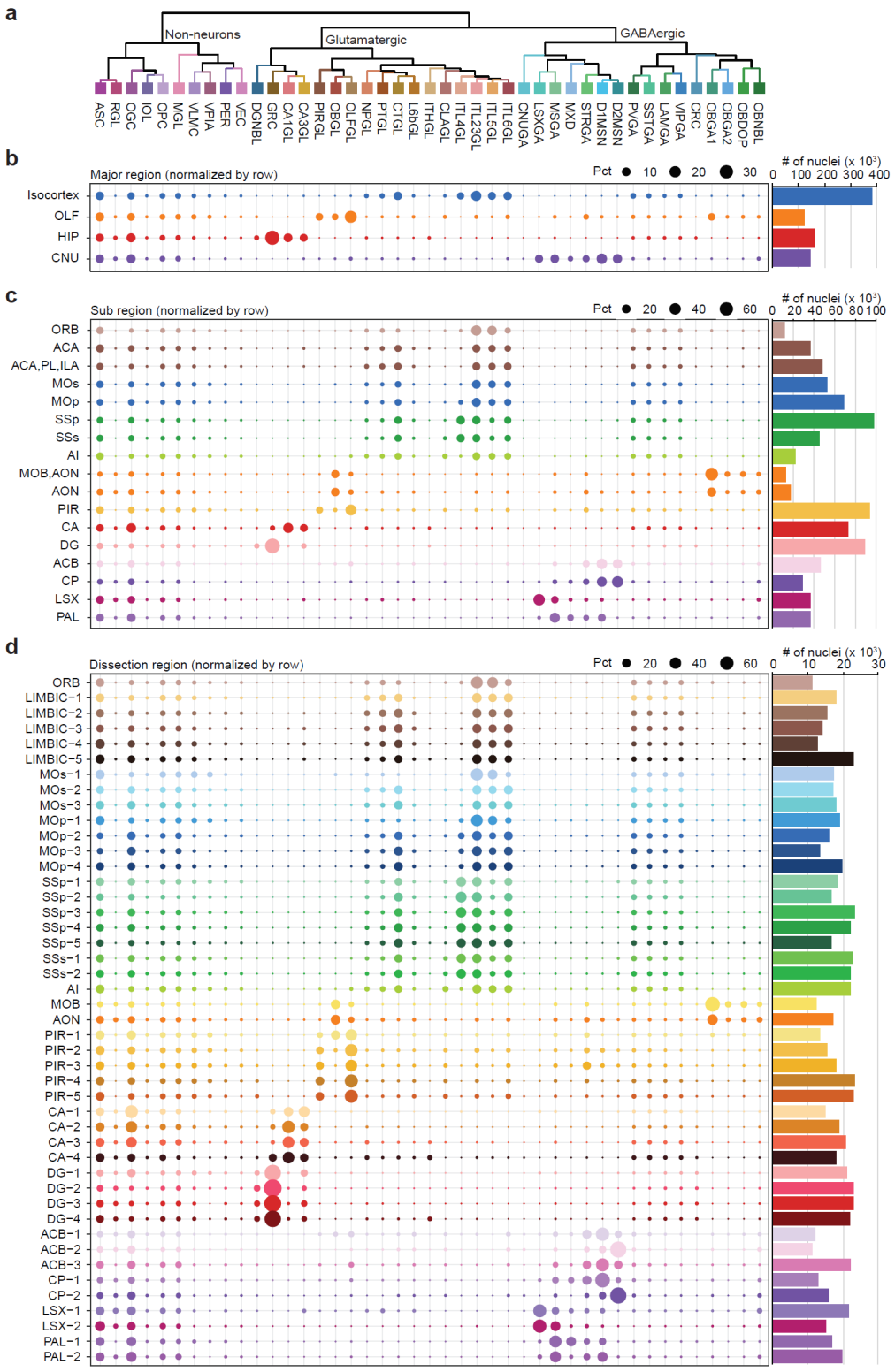
606 quadrant were selected for analysis. **c** Fraction of potential barcode collisions detected in
607 snATAC-seq libraries using a modified version of Scrublet[33] (the box is drawn from lower
608 quartile (Q1) to upper quartile (Q3) with a horizontal line drawn in the middle to denote
609 the median, whiskers with maximum 1.5 IQR). Potential barcode collisions were removed
610 for downstream processing. **d** Distribution of TSS enrichment (the box is drawn from lower
611 quartile (Q1) to upper quartile (Q3) with a horizontal line drawn in the middle to denote
612 the median, whiskers with maximum 1.5 IQR) and **e** number of uniquely mapped
613 fragments/nucleus for individual libraries. **f** Number of nuclei passing quality control for
614 sub-regions. **g** Spearman correlation matrix of snATAC-seq libraries.

Extended Data Figure 3



Extended Data Figure 3: Hierarchical dendrogram of the major cell types. Dendrogram for major cell types was constructed using 1000 rounds of bootstrapping for major cell types using R package pvclust[83]. Nodes are labeled in grey, approximately unbiased (AU) p-values (in red) and bootstrap probability (BP) values (in green) are labeled at the shoulder of the nodes, respectively. For a full list and description of cell cluster labels see Supplementary Table 3.

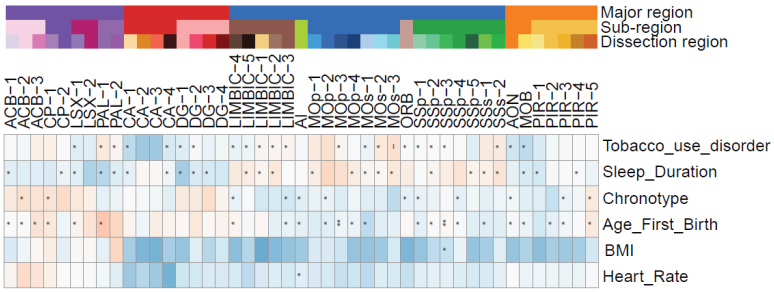
Extended Data Figure 4



Extended Data Figure 4: Relative cell cluster proportion at region resolution. a Cluster dendrogram based on chromatin accessibility. **b-d** Major cell-type composition in

b the four major regions, **c** the sub-regions and **d** the dissected regions. Indicated are row normalized percentages (pct) of clusters per major region and the total number of nuclei for each major region. Bar plots to the right show total number of nuclei sampled for each region.

Extended Data Figure 5



Extended Data Figure 5: GWAS enrichment for additional traits in open chromatin of distinct cell types.

Brain region specific enrichment of indicated GWAS traits (* FDR < 0.05, ** FDR <0.01, ***FDR < 0.001). Displayed are all brain regions and all tested phenotypes with at least one significant association.

635 **TABLES**

636 **Supplementary Table1:** Sample and dissection summary

637 **Supplementary Table 2:** Metadata table for nuclei

638 **Supplementary Table 3:** Cell cluster annotation

639 **Supplementary Table 4:** Overlap score for integration of snATAC-seq and scRNA-seq
640 clusters

641 **Supplementary Table 5:** List of the genomic locations of cCREs

642 **Supplementary Table 6:** Cluster assignment of cCREs

643 **Supplementary Table 7:** Association of *cis* regulatory modules with major cell types

644 **Supplementary Table 8:** Module assignment of cCREs

645 **Supplementary Table 9:** Known motif enrichment in *cis* regulatory modules

646 **Supplementary Table 10:** Summary of gene-cCRE correlations

647 **Supplementary Table 11:** Association of modules with joint cell clusters

648 **Supplementary Table 12:** Association of modules with individual putative enhancers

649 **Supplementary Table 13:** Gene Ontology analysis of candidate target genes of putative
650 enhancers

651 **Supplementary Table 14:** Known motif enrichment in putative enhancers

652 **Supplementary Table 15:** *De novo* motif enrichment in module M1 putative enhancers

653 **Supplementary Table 16:** Gene Ontology analysis of candidate target gene of putative
654 enhancers with motif sites in module M1

655 **Supplementary Table 17:** Known motif enrichment in candidate target promoters of
656 putative enhancers

657 **Supplementary Table 18:** Primer sequences and nuclei barcodes for version 1 and 2
658 indexing schemes.

659 **METHODS**

660 **Tissue preparation and nuclei isolation**

661 All experimental procedures using live animals were approved by the SALK Institute
662 Animal Care and Use Committee under protocol number 18-00006. Adult C57BL/6J male
663 mice were purchased from Jackson Laboratories. Brains were extracted from 56-63 day
664 old mice and sectioned into 600 µm coronal sections along the anterior-posterior axis in
665 ice-cold dissection media.[14,15] Specific brain regions were dissected according to the
666 Allen Brain Reference Atlas[32] (Extended Data Figure 1) and nuclei isolated as
667 described.[15]

668

669 **Single nucleus ATAC-seq**

670 Single nucleus ATAC-seq was performed as described with steps optimized for
671 automation[15,31,34]. A step-by-step-protocol for library preparation is available here:
672 https://www.protocols.io/view/sequencing-open-chromatin-of-single-cell-nuclei-sn-
673 pjudknw/abstract.

674 Brain nuclei were pelleted with a swinging bucket centrifuge (500 x g, 5 min, 4°C; 5920R,
675 Eppendorf). Nuclei pellets were resuspended in 1 ml nuclei permeabilization buffer (5 %
676 BSA, 0.2 % IGEPAL-CA630, 1mM DTT and cOmpleteTM, EDTA-free protease inhibitor
677 cocktail (Roche) in PBS) and pelleted again (500 x g, 5 min, 4°C; 5920R, Eppendorf).
678 Nuclei were resuspended in 500 µL high salt tagmentation buffer (36.3 mM Tris-acetate
679 (pH = 7.8), 72.6 mM potassium-acetate, 11 mM Mg-acetate, 17.6% DMF) and counted
680 using a hemocytometer. Concentration was adjusted to 1,000-4,500 nuclei/9 µl, and
681 1,000-4,500 nuclei were dispensed into each well of a 96-well plate. For tagmentation, 1
682 µL barcoded Tn5 transposomes[34] were added using a BenchSmart™ 96 (Mettler Toledo,
683 RRID:SCR_018093, Supplementary Table 18), mixed five times and incubated for 60 min
684 at 37 °C with shaking (500 rpm). To inhibit the Tn5 reaction, 10 µL of 40 mM EDTA were
685 added to each well with a BenchSmart™ 96 (Mettler Toledo, RRID:SCR_018093) and
686 the plate was incubated at 37 °C for 15 min with shaking (500 rpm). Next, 20 µL 2 x sort
687 buffer (2 % BSA, 2 mM EDTA in PBS) were added using a BenchSmart™ 96 (Mettler
688 Toledo, RRID:SCR_018093). All wells were combined into a FACS tube and stained with
689 3 µM Draq7 (Cell Signaling). Using a SH800 (Sony), 20 nuclei were sorted per well into

690 eight 96-well plates (total of 768 wells) containing 10.5 µL EB (25 pmol primer i7, 25 pmol
691 primer i5, 200 ng BSA (Sigma). Preparation of sort plates and all downstream pipetting
692 steps were performed on a Biomek i7 Automated Workstation (Beckman Coulter,
693 RRID:SCR_018094). After addition of 1 µL 0.2% SDS, samples were incubated at 55 °C
694 for 7 min with shaking (500 rpm). 1 µL 12.5% Triton-X was added to each well to quench
695 the SDS. Next, 12.5 µL NEBNext High-Fidelity 2× PCR Master Mix (NEB) were added
696 and samples were PCR-amplified (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72°C
697 60 s) × 12 cycles, held at 12 °C). After PCR, all wells were combined. Libraries were
698 purified according to the MinElute PCR Purification Kit manual (Qiagen) using a vacuum
699 manifold (QIAvac 24 plus, Qiagen) and size selection was performed with SPRI Beads
700 (Beckmann Coulter, 0.55x and 1.5x). Libraries were purified one more time with SPRI
701 Beads (Beckmann Coulter, 1.5x). Libraries were quantified using a Qubit fluorimeter (Life
702 technologies, RRID:SCR_018095) and the nucleosomal pattern was verified using a
703 Tapestation (High Sensitivity D1000, Agilent). Libraries generated with indexing version
704 1[34] (Supplementary Table 1) were sequenced on a HiSeq2500 sequencer
705 (RRID:SCR_016383, Illumina) using custom sequencing primers, 25% spike-in library
706 and following read lengths: 50 + 43 + 37 + 50 (Read1 + Index1 + Index2 + Read2).
707 Libraries generated with indexing version 2 (Supplementary Table 1) were sequenced on
708 a HiSeq4000 (RRID:SCR_016386, Illumina) using custom sequencing primers with
709 following read lengths: 50 + 10 + 12 + 50 (Read1 + Index1 + Index2 + Read2). Indexing
710 primers and sequencing primers are in Supplementary Table 18. The nuclei indexing
711 version (v1 or v2) used for each library is indicated in Supplementary Table 1.
712

713 **Nuclei indexing schemes**

714 To generate snATAC-seq libraries we used initially an indexing scheme as described
715 before (Version 1).[29,31] Here, 16 p5 and 24 p7 indexes were combined to generate an
716 array of 384 indexes for tagmentation and 16 i5 as well as 48 i7 indexes were combined
717 for an array of 768 PCR indexes. Due to this library design, it is required to sequence all
718 four indexes to assign a read to a specific nucleus with long reads and a constant base
719 sequence for both indices reads between i and p barcodes. Therefore, the resulting

720 libraries were sequenced with 25% spike-in library on a HiSeq2500 (RRID:SCR_016383)
721 and these read lengths: 50+43+37+50.[31]
722 To generate libraries compatible with other sequencers and not requiring spike-in libraries
723 or custom sequencing recipes, we modified the library scheme (Version 2). For this, we
724 used 384 individual indices for T7 and combined with one T5 with a universal index
725 sequence for tagmentation (for a total of 384 tagmentation indexes). For PCR, we used
726 768 different i5 indexes and combined with a universal i7 primer index sequence.
727 Tagmentation indexes were 10 bp and PCR indexes 12 bp long. We made sure, that the
728 hamming distance between every two barcodes was >=4, the GC content between 37.5-
729 62.5 % and the number of repeats <= 3. The resulting libraries were sequenced on a
730 HiSeq4000 with custom primers and these read lengths: 50+10+12+50 (Supplementary
731 Table 18).
732
733 **Processing and alignment of sequencing reads**
734 Paired-end sequencing reads were demultiplexed and the cell index transferred to the
735 read name. Sequencing reads were aligned to mm10 reference genome using bwa[84].
736 After alignment, we used the R package ATACseqQC (1.10.2)[85] to check for fragment
737 length contribution which is characteristic for ATAC-seq libraries. Next, we combined the
738 sequencing reads to fragments and for each fragment we performed following quality
739 control: 1) Keep only fragments quality score MAPQ > 30; 2) Keep only the properly
740 paired fragments with length <1000bp. 3) PCR duplicates were further removed with
741 SnapTools (https://github.com/r3fang/SnapTools, RRID:SCR_018097)[34]. Reads were
742 sorted based on the cell barcode in the read name.
743
744 **TSS enrichment calculation**
745 Enrichment of ATAC-seq accessibility at TSSs was used to quantify data quality without
746 the need for a defined peak set. The method for calculating enrichment at TSS was
747 adapted from previously described. TSS positions were obtained from the GENCODE
748 database v16 (RRID:SCR_014966)[40]. Briefly, Tn5 corrected insertions (reads aligned to
749 the positive strand were shifted +4 bp and reads aligned to the negative strand were
750 shifted –5 bp) were aggregated ±2,000 bp relative (TSS strand-corrected) to each unique

751 TSS genome wide. Then this profile was normalized to the mean accessibility ±1,900-
752 2,000 bp from the TSS and smoothed every 11bp. The max of the smoothed profile was
753 taken as the TSS enrichment.
754
755 **Doublet removal**
756 We used a modified version of Scrublet (RRID:SCR_018098)[33] to remove potential
757 doublets for every dataset independently. Peaks were called using MACS2 for aggregate
758 accessibility profiles on each sample. Next, cell-by-peak count matrices were calculated
759 and used as input, with default parameters. Doublet scores were calculated for both
760 observed nuclei $\{x_i\}$ and simulated doublets $\{y_i\}$ using Scrublet (RRID:SCR_018098)[33].
761 Next, a threshold θ is selected based on the distribution of $\{y_i\}$, and observed nuclei with
762 doublet score larger than θ are predicted as doublets. To determine θ, we fit a two-
763 component mixture distribution by using function normalmixEM from R package mixtools.
764 The lower component contained majority of embedded doublet types, and the other
765 component contained majority of neo-typic doublets (collision between nuclei from
766 different clusters. We selected the threshold θ where the $p_1 \cdot pdf(x, \mu_1, \sigma_1) = p_2 \cdot$
767 $pdf(x, \mu_2, \sigma_2)$. This value suggested that the nuclei have same chance of belonging to
768 both classes.
769
770 **Clustering and cluster annotation**
771 We used an iterative clustering strategy using the snapATAC package
772 (RRID:SCR_018097) with slight modifications as detailed below.[34] For round 1 clustering,
773 we clustered and finally merged single nuclei to three main cell classes: non-neurons,
774 GABAergic neurons and glutamatergic neurons. For each main cell class, we preformed
775 another round of clustering to identify major cell types. Last, for each major cell types, we
776 performed a third round of clustering to find sub-types.
777 Detailed description for every step is listed below:
778 1) Nuclei filtering
779 Nuclei with >=1,000 uniquely mapped fragments and TSS (transcription start site)
780 enrichment >10 were filtered for individual dataset. Second, potential barcode collisions
781 were also removed for individual dataset.

782 2) Feature bin selection

783 First, we calculated a cell-by-bin matrix at 500 kb resolution for every dataset
784 independently and subsequently merged the matrices. Second, we converted the cell-by-
785 bin count matrix to a binary matrix. Third, we filtered out any bins overlapping with the
786 ENCODE blacklist (mm10,
787 http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-
788 mouse/mm10.blacklist.bed.gz). Fourth, we focused on bins on chromosomes 1-19, X and
789 Y. Last, we removed the top 5% bins with the highest read coverage from the count matrix.

790 3) Dimensionality reduction

791 SnapATAC applies a nonlinear dimensionality reduction method called diffusion maps,
792 which is highly robust to noise and perturbation.[34] However, the computational time of the
793 diffusion maps algorithm scales exponentially with the increase of number of cells. To
794 overcome this limitation, we combined the Nyström method (a sampling technique)[86] and
795 diffusion maps to present Nyström Landmark diffusion map to generate the low-
796 dimensional embedding for large-scale dataset.

797 A Nyström landmark diffusion maps algorithm includes three major steps:

798 1. sampling: sample a subset of K ($K \ll N$) cells from N total cells as "landmarks".
799 2. embedding: compute a diffusion map embedding for K landmarks;
800 3. extension: project the remaining N-K cells onto the low-dimensional embedding as
801 learned from the landmarks to create a joint embedding space for all cells.

802 Having more than 800,000 single nuclei at the beginning, we decided to apply this
803 strategy on the level 1 and 2 clustering. 10,000 cells were sampled as landmarks and the
804 remaining query cells were projected onto the diffusion maps embedding of landmarks.
805 Later for the level III clustering, diffusion map embeddings were directly calculated from
806 all nuclei.

807 4) Principal Component (PC) selection

808 To determine the number of principal components to include for downstream analysis, we
809 generated "Elbow plot", to rank all principal components based on the percentage of
810 variance explained by each one. For each round of clustering, we selected the top 10-20
811 principal components that captured the majority of the variance.

812 5) Graph-based clustering

Using the selected significant components, we next construct a K Nearest Neighbor (KNN) Graph. Each cell is a node and the k-nearest neighbours of each cell were identified according to the Euclidian distance and edges were drawn between neighbours in the graph. Next, we applied the Leiden algorithm on the KNN graph using python package leidenalg (https://github.com/vtraag/leidenalg)[87]. 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 300 rounds of Leiden clustering with randomized starting seed s. let M^{s} denote the $\mathrm{N} \times \mathrm{N}$ connectivity matrix resulting from applying Leiden algorithm to the dataset D^{s} with different seeds. The entries of M^{s} are defined as follows:

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

Let I^{s} be the $\mathrm{N} \times \mathrm{N}$ identicator 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 normalised sum of all connectivity matrices of all the perturbed D^{s}.

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

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 minimal 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.

840 Finally, for every cluster, we tested whether we could identify differential features
841 compared to all other nuclei (background) and to the nearest nuclei (local background)
842 using the function 'findDAR'.
843 6) Visualization
844 For visualization we applied Uniform Manifold Approximation and Projection (UMAP)[80].
845
846 **Regional specificity**
847 For each cell type, fraction of nuclei is first calculated from each brain regions. Then, we
848 use function 'entropyDiversity' from R package BioQC (cite) to calculate regional diversity
849 for each cell types and minus the value by 1 as specificity.
850
851 **Identification of reproducible peak sets in each cell cluster**
852 We performed peak calling according to the ENCODE ATAC-seq pipeline
853 (https://www.encodeproject.org/atac-seq/). For every cell cluster, we combined all
854 properly paired reads to generate a pseudobulk ATAC-seq dataset for individual
855 biological replicates. In addition, we generated two pseudo-replicates which comprise half
856 of the reads from each biological replicate. We called peak for each of the four dataset
857 and a pool of both replicates independently. Peak calling was performed on the Tn5-
858 corrected single-base insertions using the MACS2[39] with these parameters: --shift -75 --
859 extsize 150 --nomodel --call-summits --SPMR --keep-dup all -q 0.01. Finally, we extended
860 peak summits by 250 bp on either side to a final width of 501 bp for merging and
861 downstream analysis. To generate a list of reproducible peaks, we kept peaks that 1)
862 were detected in the pooled dataset and overlapped >=50% of peak length with a peak
863 in both individual replicates or 2) were detected in the pooled dataset and
864 overlapped >=50% of peak length with a peak in both pseudo-replicates.
865 To account for differences in performance of MACS2[39] based on read depth and/or
866 number of nuclei in individual clusters, we converted MACS2 peak scores (-log10(q-
867 value)) to "score per million"[88]. We filtered reproducible peaks by choosing a "score per
868 million" cut-off of 2 was used to filter reproducible peaks.
869 We only kept reproducible peaks on chromosome 1-19 and both sex chromosomes, and
870 filtered ENCODE mm10 blacklist regions (mm10,

871 http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-
872 mouse/mm10.blacklist.bed.gz). A union peak list for the whole dataset obtained by
873 merging peak sets from all cell clusters using BEDtools (RRID:SCR_006646)[89].
874 Lastly, since snATAC-seq data are very sparse, we selected only elements that were
875 identified as open chromatin in a significant fraction of the cells in each cluster. To this
876 end, we first randomly selected same number of non-DHS regions (~ 670k elements)
877 from the genome as background and calculated the fraction of nuclei for each cell type
878 that that showed a signal at these sites. Next, we fitted a zero-inflated beta model and
879 empirically identified a significance threshold of FDR < 0.01 to filter potential false positive
880 peaks. Peak regions with FDR < 0.01 in at least one of the clusters were included into
881 downstream analysis.
882
883 **Computing chromatin accessibility scores**
884 Accessibility of cCREs in individual clusters was quantified by counting the fragments in
885 individual clusters normalized by read depth (counts per million: CPM).
886 For each gene, we summed counts within the gene body + 2kb upstream to calculate
887 "gene activity score (GAS)" using Seurat
888 (https://satijalab.org/seurat/v3.1/atacseq_integration_vignette.html,
889 RRID:SCR_016341)[38], GAS were used for visualization and integrative analysis with
890 single cell RNA-seq.
891
892 **Integrative analysis of single nucleus ATAC-seq and single cell RNA-seq for mouse**
893 **brain**
894 For integrative analysis, we downloaded level 5 clustering data from the Mouse Brain
895 Atlas website (http://mousebrain.org)[1]. First, we filtered brain regions that matched
896 samples profiled in this study using these attributes for "Region": "CNS", "Cortex",
897 "Hippocampus", "Hippocampus,Cortex", "Olfactory bulb", "Striatum dorsal", "Striatum
898 ventral", "Dentate gyrus", "Striatum dorsal,Striatum ventral", "Striatum dorsal, Striatum
899 ventral, Dentate gyrus", "Pallidum", "Striatum dorsal, Striatum ventral, Amygdala",
900 "Striatum dorsal, Striatum ventral", "Telencephalon", "Brain", "Sub ventricular zone,
901 Dentate gyrus"

902 Second, we manually subset cell types into three groups by checking the attribute in
903 “Taxonomy_group”: Non-neurons: "Vascular and leptomeningeal cells", "Astrocytes",
904 "Oligodendrocytes", "Ependymal cells", "Microglia", "Oligodendrocyte precursor cells",
905 "Olfactory ensheathing cells", "Pericytes", "Vascular smooth muscle cells", "Perivascular
906 macrophages", "Dentate gyrus radial glia-like cells", "Subventricular zone radial glia-like
907 cells", "Vascular smooth muscle cells", "Vascular endothelial cells", "Vascular and
908 leptomeningeal cells"; GABAergic neurons: "Non-glutamatergic neuroblasts",
909 "Telencephalon projecting inhibitory neurons", "Olfactory inhibitory neurons",
910 "Glutamatergic neuroblasts", "Cholinergic and monoaminergic neurons", "Di- and
911 mesencephalon inhibitory neurons", "Telencephalon inhibitory interneurons",
912 "Peptidergic neurons"; Glutamatergic neurons: "Dentate gyrus granule neurons", "Di- and
913 mesencephalon excitatory neurons", "Telencephalon projecting excitatory neurons"
914 We performed integrative analysis with single cell RNA-seq using Seurat 3.0
915 (RRID:SCR_016341) to compare cell annotation between different modalities[38]. We
916 randomly selected 200 nuclei (and used all nuclei for cell cluster with <200 nuclei) from
917 each cell cluster for integrative analysis. We first generated a Seurat object in R by using
918 previously calculated gene activity scores, diffusion map embeddings and cell cluster
919 labels from snATAC-seq. Then, variable genes were identified from scRNA-seq and used
920 for identifying anchors between these two modalities. Finally, to visualize all the cells
921 together, we co-embedded the scRNA-seq and snATAC-seq profiles in the same low
922 dimensional space.
923 To quantify the similarity between cell clusters from two modalities, we calculated an
924 overlapping score as the sum of the minimum proportion of cells/nuclei in each cluster
925 that overlapped within each co-embedding cluster[10]. Cluster overlaps varied from 0 to 1
926 and were visualized as a heat map with snATAC-seq clusters in rows and scRNA-seq
927 clusters in columns.
928

929 **Identification of *cis* regulatory modules**

930 We used Nonnegative Matrix Factorization (NMF)[90] to group cCREs into *cis* regulatory
931 modules based on their relative accessibility across major clusters. We adapted NMF
932 (Python package: sklearn[91]) to decompose the cell-by-cCRE matrix V (N×M, N rows:

933 cCRE, M columns: cell clusters) into a coefficient matrix H (R×M, R rows: number of
934 modules) and a basis matrix W (N×R), with a given rank R:

935 $$V \approx WH,$$

936 The basis matrix defines module related accessible cCREs, and the coefficient matrix
937 defines the cell cluster components and their weights in each module. The key issue to
938 decompose the occupancy profile matrix was to find a reasonable value for the rank R
939 (i.e., the number of modules). Several criteria have been proposed to decide whether a
940 given rank R decomposes the occupancy profile matrix into meaningful clusters. Here we
941 applied two measurements "Sparseness"[92] and "Entropy"[93] to evaluate the clustering
942 result. Average values were calculated from 100 times for NMF runs at each given rank
943 with random seed, which will ensure the measurements are stable.
944 Next, we used the coefficient matrix to associate modules with distinct cell clusters. In the
945 coefficient matrix, each row represents a module and each column represents a cell
946 cluster. The values in the matrix indicate the weights of clusters in their corresponding
947 module. The coefficient matrix was then scaled by column (cluster) from 0 to 1.
948 Subsequently, we used a coefficient > 0.1 (~95th percentile of the whole matrix) as
949 threshold to associate a cluster with a module.
950 In addition, we associated each module with accessible elements using the basis matrix.
951 For each element and each module, we derived a basis coefficient score, which
952 represents the accessible signal contributed by all cluster in the defined module. In
953 addition, we also implemented and calculated a basis-specificity score called "feature
954 score" for each accessible element using the "kim" method[93]. The feature score ranges
955 from 0 to 1. A high feature score means that a distinct element is specifically associated
956 with a specific module. Only features that fulfil both following criteria were retained as
957 module specific elements:

958 1. feature score greater than median + 3 standard deviation;
959 2. the maximum contribution to a basis component is greater than the median of all
960 contributions (i.e. of all elements of W).

961

962 **Dendrogram construction for mouse brain cell types**

963 First, we calculated for cCRE the median accessibility per cluster and used this value as
964 cluster centroid. Next, we calculated the coefficient of variant (CV) for the cluster centroid
965 of each element across major cell types. Finally, we only kept variable elements with CV
966 larger than 1.5 for dendrogram construction.
967 We used the set of variable features defined above to calculate a correlation-based
968 distance matrix. Next, we performed linkage hierarchical clustering using the R package
969 pvclust (v.2.0)[83] with parameters method.dist="cor" and method.hclust="ward.D2". The
970 confidence for each branch of the tree was estimated by the bootstrap resampling
971 approach.
972
973 **Motif enrichment**
974 We performed both *de novo* and known motif enrichment analysis using Homer (v4.11,
975 RRID:SCR_010881)[46]. For cCREs in the consensus list, we scanned a region of ± 250
976 bp around the center of the element. And for proximal/promoter regions, we scanned a
977 region of ± 1000 bp around the transcriptional start site.
978
979 **GREAT analysis**
980 Gene ontology annotation of cCREs was performed using GREAT (version 4.0.4,
981 RRID:SCR_005807)[94] with default parameters. GO Biological Process was used for
982 annotations.
983
984 **Gene ontology enrichment**
985 We perform gene ontology enrichment analysis using R package Enrichr
986 (RRID:SCR_001575)[82]. Gene set library "GO_Biological_Process_2018" was used with
987 default parameters. The combined score is defined as the p-value computed using the
988 Fisher exact test multiplied with the z-score of the deviation from the expected rank.
989
990 **Predicting enhancer-promoter interactions**
991 First, co-accessible regions are identified for all open regions in each cell cluster
992 (randomly selected 200 nuclei, and used all nuclei for cell cluster with <200 nuclei)
993 separately, using Cicero[49] with following parameters: aggregation k = 10, window size =

994 500 kb, distance constraint = 250 kb. In order to find an optimal co-accessibility threshold
995 for each cluster, we generated a random shuffled cCRE-by-cell matrix as background and
996 identified co-accessible regions from this shuffled matrix. We fitted the distribution of co-
997 accessibility scores from random shuffled background into a normal distribution model by
998 using R package fitdistrplus[95]. Next, we tested every co-accessibility pairs and set the
999 cut-off at co-accessibility score with empirically defined significance threshold of
1000 FDR<0.01.
1001 CCRE outside of ± 1 kb of transcriptional start sites (TSS) in GENCODE mm10 (v16,
1002 RRID:SCR_014966).[40] were considered distal. Next, we assigned co-accessibility pairs
1003 to three groups: proximal-to-proximal, distal-to-distal, and distal-to-proximal. In this study,
1004 we focus only on distal-to-proximal pairs. We further used RNA expression from matched
1005 T-types to filter pairs that were linked to non-expressed genes (normalized UMI > 5).
1006 We calculated Pearson's correlation coefficient (PCC) between gene expression and
1007 cCRE accessibility across joint RNA-ATAC clusters to examine the relationship between
1008 co-accessibility pairs. To do so, we first aggregated all nuclei/cells from scRNA-seq and
1009 snATAC-seq for every joint cluster to calculate accessibility scores ($\log_2$ CPM) and
1010 relative expression levels ($\log_2$ normalized UMI). Then, PCC was calculated for every
1011 gene-cCRE pair within a 1 Mbp window centered on the TSS for every gene. We also
1012 generated a set of background pairs by randomly selecting regions from different
1013 chromosomes and shuffling of cluster labels. Finally, we fit a normal distribution model
1014 and defined a cut-off at PCC score with empirically defined significance threshold of
1015 FDR<0.01, in order to select significant positively correlated cCRE-gene pairs.
1016

1017 **GWAS enrichment**

1018 To enable comparison to GWAS of human phenotypes, we used liftOver with settings "-
1019 minMatch=0.5" to convert accessible elements from mm10 to hg19 genomic
1020 coordinates.[69] Next, we reciprocal lifted the elements back to mm10 and only kept the
1021 regions that mapped to original loci. We further removed converted regions with length >
1022 1kb.
1023 We obtained GWAS summary statistics for quantitative traits related to neurological
1024 disease and control traits: Heart Failure[96], Type 1 Diabetes[97], Age First Birth and Number

1025 Children Born[98], Lupus[99], Primary Biliary Cirrhosis[100], Tiredness[101], Crohns_Disease[102],
1026 Inflammatory Bowel Disease[102], Ulcerative_Colitis[102], Asthma[103], Attention Deficit
1027 Hyperactivity Disorder[104], Heart Rate[105], Celiacs Disease[106], HOMA-B[107], HOMA-IR[107],
1028 Childhood Aggression[108], Atopic Dermatitis[109], Allergy[110], HDL_Cholesterol[111],
1029 LDL_Cholesterol[111], Total Cholesterol[111], Triglycerides[111], Autism Spectrum Disorder[112],
1030 Birth Weight[113], Bipolar Disorder[114], Multiple Sclerosis[115], Insomnia[116], Vitamin D[117],
1031 Primary Sclerosing Cholangitis[118], Vitiligo[119], Chronotype[120], Sleep Duration[120],
1032 Alzheimer's Disease[121], BMI[122], Neuroticism[123], Type 2 Diabetes[124], Stroke[125], Fasting
1033 Glucose[126], Fasting Insulin[126], Child Sleep Duration[127], Coronary Artery Disease[128], Atrial
1034 Fibrillation[129], Rheumatoid Arthritis[130], Educational Attainment[131], Chronic Kidney
1035 Disease[132], Obsessive Compulsive Disorder[133], Post Traumatic Stress Disorder[134],
1036 Schizophrenia[135], Age At Menopause[136], Age At Menarche[137], Tobacco use disorder
1037 (ftp://share.sph.umich.edu/UKBB_SAIGE_HRC/, Phenotype code: 318)[138],
1038 Intelligence[139], Alcohol Usage[140], Fasting Proinsulin[141], Head Circumference[142],
1039 Microalbuminuria[143], Extraversion[144], Birth Length[145], Amyotrophic Lateral Sclerosis[146],
1040 Anorexia Nervosa[147], HbA1c[148], Major Depressive Disorder[149], Height[150].
1041 We prepared summary statistics to the standard format for Linkage disequilibrium (LD)
1042 score regression. We used homologous sequences for each major cell types as a binary
1043 annotation, and the superset of all candidate regulatory peaks as the background control.
1044 For each trait, we used cell type specific (CTS) LD score regression
1045 (https://github.com/bulik/ldsc) to estimate the enrichment coefficient of each annotation
1046 jointly with the background control[70].
1047

1048 **External datasets**

1049 We listed all the datasets we used in this study for intersection analysis:
1050 rDHS regions for both hg19 and mm10 are obtained from SCREEN database
1051 (https://screen.encodeproject.org)[41,42].
1052 ChromHMM[43,45] states for mouse brain are download from GitHub
1053 (https://github.com/gireeshkbogu/chromatin_states_chromHMM_mm9), and coordinates
1054 are LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to mm10 with default
1055 parameters[69].

1056 PhastCons[81] conserved elements were download from the UCSC Genome Browser
1057 (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phastCons60way/).
1058 CTCF binding sites are download from Mouse Encode Project[43]
1059 http://chromosome.sdsc.edu/mouse/). CTCF binding sites from cortex and olfactory bulb
1060 were used in this study. Peaks are extended ± 500 bp from the loci of peak summits and
1061 used LiftOver to mm10[69].
1062

1063 **Statistics**

1064 No statistical methods were used to predetermine sample sizes. There was no
1065 randomization of the samples, and investigators were not blinded to the specimens being
1066 investigated. However, clustering of single nuclei based on chromatin accessibility was
1067 performed in an unbiased manner, and cell types were assigned after clustering. Low-
1068 quality nuclei and potential barcode collisions were excluded from downstream analysis
1069 as outlined above. For significance of ontology enrichments using GREAT, Bonferroni-
1070 corrected binomial p values were used[94]. For ontology enrichment using Enrichr the
1071 combined score which represents the product of the p-value computed using the Fisher
1072 exact test multiplied with the z-score of the deviation from the expected rank was used[82].
1073 For significance testing of enrichment of *de novo* motifs, a hypergeometric test was used
1074 without correction for multiple testing[46].
1075

1076 **Data availability**

1077 Demultiplexed data can be accessed via the NEMO archive (NEMO, RRID:SCR_016152)
1078 here: http://data.nemoarchive.org/biccn/grant/cemba/ecker/chromatin/scell/raw/
1079 Processed data are available on our web portal and can be explored here:
1080 http://catlas.org/mousebrain
1081

1082 **Code availability**

1083 Custom code and scripts used for analysis can be accessed here:
1084 https://github.com/YoungLeeBBS/snATACutils and https://github.com/r3fang/SnapATAC.

REFERENCES

1 Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014 e1022, doi:10.1016/j.cell.2018.06.021 (2018).

2 Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015-1030 e1016, doi:10.1016/j.cell.2018.07.028 (2018).

3 Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78, doi:10.1038/s41586-018-0654-5 (2018).

4 Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, doi:10.1126/science.aau5324 (2018).

5 Eng, C. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235-239, doi:10.1038/s41586-019-1049-y (2019).

6 *The Rat Nervous System,* . 4th edition edn, (2015).

7 Harris, K. D. & Shepherd, G. M. The neocortical circuit: themes and variations. *Nat Neurosci* **18**, 170-181, doi:10.1038/nn.3917 (2015).

8 Huang, Z. J. Toward a genetic dissection of cortical circuits in the mouse. *Neuron* **83**, 1284-1302, doi:10.1016/j.neuron.2014.08.041 (2014).

9 Douglas, R. J. & Martin, K. A. Neuronal circuits of the neocortex. *Annu Rev Neurosci* **27**, 419-451, doi:10.1146/annurev.neuro.27.070203.144152 (2004).

10 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68, doi:10.1038/s41586-019-1506-7 (2019).

11 Luo, C. *et al.* Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv* (2019).

12 Ecker, J. R. *et al.* The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. *Neuron* **96**, 542-557, doi:10.1016/j.neuron.2017.10.007 (2017).

13 Lee, D. S. *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* **16**, 999-1006, doi:10.1038/s41592-019-0547-z (2019).

14 Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600-604, doi:10.1126/science.aan3351 (2017).

15 Yao, Z. *et al.* An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *bioRxiv* (2020).

16 Kepecs, A. & Fishell, G. Interneuron cell types are fit to function. *Nature* **505**, 318-326, doi:10.1038/nature12983 (2014).

17 Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat Neurosci* **21**, 120-129, doi:10.1038/s41593-017-0029-5 (2018).

18 Yap, E. L. & Greenberg, M. E. Activity-Regulated Transcription: Bridging the Gap between Neural Activity and Behavior. *Neuron* **100**, 330-348, doi:10.1016/j.neuron.2018.10.013 (2018).

19 Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-286, doi:10.1038/nrg3682 (2014).

20 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

21 Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39-55, doi:10.1016/j.cell.2013.09.011 (2013).

22 Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914, doi:10.1126/science.aab1601 (2015).

23 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490, doi:10.1038/nature14590 (2015).

24 Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* **37**, 916-924, doi:10.1038/s41587-019-0147-6 (2019).

25 Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution. *Nat Commun* **9**, 3647, doi:10.1038/s41467-018-05887-x (2018).

26 Chen, X., Miragaia, R. J., Natarajan, K. N. & Teichmann, S. A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat Commun* **9**, 5345, doi:10.1038/s41467-018-07771-0 (2018).

27 Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80, doi:10.1038/nbt.4038 (2018).

28 Graybuck, L. T. *et al.* Prospective, brain-wide labeling of neuronal subclasses with enhancer-driven AAVs. *bioRxiv* (2019).

29 Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324 e1318, doi:10.1016/j.cell.2018.06.052 (2018).

30 Sinnamon, J. R. *et al.* The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Res* **29**, 857-869, doi:10.1101/gr.243725.118 (2019).

31 Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* **21**, 432-439, doi:10.1038/s41593-018-0079-3 (2018).

32 Wang, Q. *et al.* The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*, doi:10.1016/j.cell.2020.04.007 (2020).

33 Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291 e289, doi:10.1016/j.cels.2018.11.005 (2019).

34 Fang, R. *et al.* Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *bioRxiv* (2019).

35 Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* **19**, 335-346, doi:10.1038/nn.4216 (2016).

36 Hochgerner, H., Zeisel, A., Lonnerberg, P. & Linnarsson, S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat Neurosci* **21**, 290-299, doi:10.1038/s41593-017-0056-2 (2018).

1175 37 Liu, H. *et al.* DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution.
1176 *bioRxiv* (2020).
1177 38 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-
1178 1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
1179 39 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137,
1180 doi:10.1186/gb-2008-9-9-r137 (2008).
1181 40 Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE.
1182 *Genome Biol* **7 Suppl 1**, S4 1-9, doi:10.1186/gb-2006-7-s1-s4 (2006).
1183 41 Consortium, E. P. A user's guide to the encyclopedia of DNA elements (ENCODE).
1184 *PLoS Biol* **9**, e1001046, doi:10.1371/journal.pbio.1001046 (2011).
1185 42 Consortium, E. P. An integrated encyclopedia of DNA elements in the human
1186 genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
1187 43 Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome.
1188 *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).
1189 44 Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome.
1190 *Nature* **515**, 355-364, doi:10.1038/nature13992 (2014).
1191 45 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and
1192 characterization. *Nat Methods* **9**, 215-216, doi:10.1038/nmeth.1906 (2012).
1193 46 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors
1194 prime cis-regulatory elements required for macrophage and B cell identities. *Mol*
1195 *Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
1196 47 Glasgow, S. M. *et al.* Mutual antagonism between Sox10 and NFIA regulates
1197 diversification of glial lineages and glioma subtypes. *Nat Neurosci* **17**, 1322-1329,
1198 doi:10.1038/nn.3790 (2014).
1199 48 Kolterud, A., Alenius, M., Carlsson, L. & Bohm, S. The Lim homeobox gene Lhx2
1200 is required for olfactory sensory neuron identity. *Development* **131**, 5319-5326,
1201 doi:10.1242/dev.01416 (2004).
1202 49 Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-
1203 Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-871 e858,
1204 doi:10.1016/j.molcel.2018.06.044 (2018).
1205 50 Kierdorf, K. *et al.* Microglia emerge from erythromyeloid precursors via Pu.1- and
1206 Irf8-dependent pathways. *Nat Neurosci* **16**, 273-280, doi:10.1038/nn.3318 (2013).
1207 51 Nord, A. S., Pattabiraman, K., Visel, A. & Rubenstein, J. L. R. Genomic
1208 perspectives of transcriptional regulation in forebrain development. *Neuron* **85**, 27-
1209 47, doi:10.1016/j.neuron.2014.11.011 (2015).
1210 52 Yuan, F. *et al.* Efficient generation of region-specific forebrain neurons from human
1211 pluripotent stem cells under highly defined condition. *Sci Rep* **5**, 18550,
1212 doi:10.1038/srep18550 (2015).
1213 53 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human
1214 epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).
1215 54 Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome
1216 topology and function. *Nat Rev Genet* **15**, 234-246, doi:10.1038/nrg3663 (2014).
1217 55 Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N. & Yagi, T. CTCF is required
1218 for neural development and stochastic expression of clustered Pcdh genes in
1219 neurons. *Cell Rep* **2**, 345-357, doi:10.1016/j.celrep.2012.06.014 (2012).

1220 56 Guo, Y. *et al.* CTCF/cohesin-mediated DNA looping is required for protocadherin
1221 alpha promoter choice. *Proc Natl Acad Sci U S A* **109**, 21081-21086,
1222 doi:10.1073/pnas.1219280110 (2012).
1223 57 Boillot, M. *et al.* LGI1 acts presynaptically to regulate excitatory synaptic
1224 transmission during early postnatal development. *Sci Rep* **6**, 21769,
1225 doi:10.1038/srep21769 (2016).
1226 58 Choksi, S. P., Lauter, G., Swoboda, P. & Roy, S. Switching on cilia: transcriptional
1227 networks regulating ciliogenesis. *Development* **141**, 1427-1441,
1228 doi:10.1242/dev.074666 (2014).
1229 59 Nakajima, K. *et al.* Molecular motor KIF5A is essential for GABA(A) receptor
1230 transport, and KIF5A deletion causes epilepsy. *Neuron* **76**, 945-961,
1231 doi:10.1016/j.neuron.2012.10.012 (2012).
1232 60 Assali, A., Harrington, A. J. & Cowan, C. W. Emerging roles for MEF2 in brain
1233 development and mental disorders. *Curr Opin Neurobiol* **59**, 49-58,
1234 doi:10.1016/j.conb.2019.04.008 (2019).
1235 61 Shi, Y. *et al.* Functional comparison of the effects of TARPs and cornichons on
1236 AMPA receptor trafficking and gating. *Proc Natl Acad Sci U S A* **107**, 16315-16319,
1237 doi:10.1073/pnas.1011706107 (2010).
1238 62 Lopez de Armentia, M. *et al.* cAMP response element-binding protein-mediated
1239 gene expression increases the intrinsic excitability of CA1 pyramidal neurons. *J*
1240 *Neurosci* **27**, 13909-13918, doi:10.1523/JNEUROSCI.3850-07.2007 (2007).
1241 63 Zhou, Y. *et al.* CREB regulates excitability and the allocation of memory to subsets
1242 of neurons in the amygdala. *Nat Neurosci* **12**, 1438-1443, doi:10.1038/nn.2405
1243 (2009).
1244 64 Mattson, M. P. & Camandola, S. NF-kappaB in neuronal plasticity and
1245 neurodegenerative disorders. *J Clin Invest* **107**, 247-254, doi:10.1172/JCI11916
1246 (2001).
1247 65 Dziennis, S. & Alkayed, N. J. Role of signal transducer and activator of
1248 transcription 3 in neuronal survival and regeneration. *Rev Neurosci* **19**, 341-361,
1249 doi:10.1515/revneuro.2008.19.4-5.341 (2008).
1250 66 Fontenot, M. R. *et al.* Novel transcriptional networks regulated by CLOCK in
1251 human neurons. *Genes Dev* **31**, 2121-2135, doi:10.1101/gad.305813.117 (2017).
1252 67 Pickrell, J. K. Joint analysis of functional genomic data and genome-wide
1253 association studies of 18 human traits. *Am J Hum Genet* **94**, 559-573,
1254 doi:10.1016/j.ajhg.2014.03.004 (2014).
1255 68 Maurano, M. T. *et al.* Systematic localization of common disease-associated
1256 variation in regulatory DNA. *Science* **337**, 1190-1195,
1257 doi:10.1126/science.1222794 (2012).
1258 69 Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids*
1259 *Res* **45**, D626-D634, doi:10.1093/nar/gkw1134 (2017).
1260 70 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from
1261 polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295,
1262 doi:10.1038/ng.3211 (2015).
1263 71 Skene, N. G. *et al.* Genetic identification of brain cell types underlying
1264 schizophrenia. *Nat Genet* **50**, 825-833, doi:10.1038/s41588-018-0129-5 (2018).

72 Volkow, N. D. & Morales, M. The Brain on Drugs: From Reward to Addiction. *Cell* **162**, 712-725, doi:10.1016/j.cell.2015.07.046 (2015).
73 Corces, M. R. *et al.* Single-cell epigenomic identification of inherited risk loci in Alzheimer's and Parkinson's disease. *bioRxiv* (2020).
74 initiative, B. BRAIN 2025 Report. (2014).
75 Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177**, 162-183, doi:10.1016/j.cell.2019.01.015 (2019).
76 Canzio, D. & Maniatis, T. The generation of a protocadherin cell-surface recognition code for neural circuit assembly. *Curr Opin Neurobiol* **59**, 213-220, doi:10.1016/j.conb.2019.10.001 (2019).
77 Zhang, D. *et al.* Identification of potential target genes for RFX4_v3, a transcription factor critical for brain development. *J Neurochem* **98**, 860-875, doi:10.1111/j.1471-4159.2006.03930.x (2006).
78 Gorkin, D. U. *et al.* Systematic mapping of chromatin state landscapes during mouse development. *bioRxiv* (2017).
79 Mich, J. K. *et al.* Functional enhancer elements drive subclass-selective expression from mouse to primate neocortex. *bioRxiv* (2020).
80 Leland McInnes, J. H., Nathaniel Saul, Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3(29)**, 861, doi:https://doi.org/10.21105/joss.00861 (2018).
81 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).
82 Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128, doi:10.1186/1471-2105-14-128 (2013).
83 Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540-1542, doi:10.1093/bioinformatics/btl117 (2006).
84 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
85 Ou, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169, doi:10.1186/s12864-018-4559-3 (2018).
86 Bouneffouf, D. B., I. Theoretical analysis of the Minimum Sum of Squared Similarities sampling for Nyström-based spectral clustering. . *2016 International Joint Conference on Neural Networks (IJCNN)*, 3856–3862, doi:10.1109/ijcnn.2016.7727698 (2016).
87 Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233, doi:10.1038/s41598-019-41695-z (2019).
88 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, doi:10.1126/science.aav1898 (2018).
89 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

90 Li, Y. E. *et al.* Identification of high-confidence RNA regulatory elements by combinatorial classification of RNA-protein binding sites. *Genome Biol* **18**, 169, doi:10.1186/s13059-017-1298-8 (2017).

91 Fabian Pedregosa, G. V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12(85)**, 2825−2830 (2011).

92 Hoyer, P. O. Non-negative Matrix Factorization with Sparseness Constraints. *J. Mach. Learn. Res.* **5**, 1457–1469 (2004).

93 Kim, H. & Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **23**, 1495-1502, doi:10.1093/bioinformatics/btm134 (2007).

94 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

95 Delignette-Muller, M. L. & Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *2015* **64**, 34, doi:10.18637/jss.v064.i04 (2015).

96 Arvanitis, M. *et al.* Genome-wide association and multi-omic analyses reveal ACTN2 as a gene linked to heart failure. *Nat Commun* **11**, 1122, doi:10.1038/s41467-020-14843-7 (2020).

97 Aylward, A., Chiou, J., Okino, M.-L., Kadakia, N. & Gaulton, K. J. Shared genetic contribution to type 1 and type 2 diabetes risk. *bioRxiv* (2018).

98 Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet* **48**, 1462-1472, doi:10.1038/ng.3698 (2016).

99 Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat Genet* **47**, 1457-1464, doi:10.1038/ng.3434 (2015).

100 Cordell, H. J. *et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat Commun* **6**, 8019, doi:10.1038/ncomms9019 (2015).

101 Deary, V. *et al.* Genetic contributions to self-reported tiredness. *Mol Psychiatry* **23**, 609-620, doi:10.1038/mp.2017.5 (2018).

102 de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256-261, doi:10.1038/ng.3760 (2017).

103 Demenais, F. *et al.* Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet* **50**, 42-53, doi:10.1038/s41588-017-0014-7 (2018).

104 Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet* **51**, 63-75, doi:10.1038/s41588-018-0269-7 (2019).

105 den Hoed, M. *et al.* Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet* **45**, 621-631, doi:10.1038/ng.2610 (2013).

106 Dubois, P. C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42**, 295-302, doi:10.1038/ng.543 (2010).

107 Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42**, 105-116, doi:10.1038/ng.520 (2010).
108 Pappa, I. *et al.* A genome-wide approach to children's aggressive behavior: The EAGLE consortium. *Am J Med Genet B Neuropsychiatr Genet* **171**, 562-572, doi:10.1002/ajmg.b.32333 (2016).
109 Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat Genet* **47**, 1449-1456, doi:10.1038/ng.3424 (2015).
110 Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat Genet* **49**, 1752-1757, doi:10.1038/ng.3985 (2017).
111 Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283, doi:10.1038/ng.2797 (2013).
112 Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* **51**, 431-444, doi:10.1038/s41588-019-0344-8 (2019).
113 Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248-252, doi:10.1038/nature19806 (2016).
114 Hou, L. *et al.* Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Hum Mol Genet* **25**, 3383-3394, doi:10.1093/hmg/ddw181 (2016).
115 International Multiple Sclerosis Genetics, C. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214-219, doi:10.1038/nature10251 (2011).
116 Jansen, P. R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat Genet* **51**, 394-403, doi:10.1038/s41588-018-0333-3 (2019).
117 Jiang, X. *et al.* Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat Commun* **9**, 260, doi:10.1038/s41467-017-02662-2 (2018).
118 Ji, S. G. *et al.* Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat Genet* **49**, 269-273, doi:10.1038/ng.3745 (2017).
119 Jin, Y. *et al.* Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. *Nat Genet* **48**, 1418-1424, doi:10.1038/ng.3680 (2016).
120 Jones, S. E. *et al.* Genome-Wide Association Analyses in 128,266 Individuals Identifies New Morningness and Sleep Duration Loci. *PLoS Genet* **12**, e1006125, doi:10.1371/journal.pgen.1006125 (2016).
121 Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-1458, doi:10.1038/ng.2802 (2013).
122 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).

123 Luciano, M. *et al.* Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat Genet* **50**, 6-11, doi:10.1038/s41588-017-0013-8 (2018).
124 Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* **50**, 1505-1513, doi:10.1038/s41588-018-0241-6 (2018).
125 Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* **50**, 524-537, doi:10.1038/s41588-018-0058-3 (2018).
126 Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**, 659-669, doi:10.1038/ng.2274 (2012).
127 Marinelli, M. *et al.* Heritability and Genome-Wide Association Analyses of Sleep Duration in Children: The EAGLE Consortium. *Sleep* **39**, 1859-1869, doi:10.5665/sleep.6170 (2016).
128 Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet* **49**, 1385-1391, doi:10.1038/ng.3913 (2017).
129 Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet* **50**, 1234-1239, doi:10.1038/s41588-018-0171-3 (2018).
130 Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-381, doi:10.1038/nature12873 (2014).
131 Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539-542, doi:10.1038/nature17671 (2016).
132 Pattaro, C. *et al.* Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat Commun* **7**, 10023, doi:10.1038/ncomms10023 (2016).
133 International Obsessive Compulsive Disorder Foundation Genetics, C. & Studies, O. C. D. C. G. A. Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol Psychiatry* **23**, 1181-1188, doi:10.1038/mp.2017.154 (2018).
134 Duncan, L. E. *et al.* Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. *Mol Psychiatry* **23**, 666-673, doi:10.1038/mp.2017.77 (2018).
135 Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).
136 Day, F. R. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat Genet* **47**, 1294-1303, doi:10.1038/ng.3412 (2015).
137 Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat Genet* **49**, 834-841, doi:10.1038/ng.3841 (2017).

138 Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335-1341, doi:10.1038/s41588-018-0184-y (2018).
139 Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* **50**, 912-919, doi:10.1038/s41588-018-0152-6 (2018).
140 Schumann, G. *et al.* KLB is associated with alcohol drinking, and its gene product beta-Klotho is necessary for FGF21 regulation of alcohol preference. *Proc Natl Acad Sci U S A* **113**, 14372-14377, doi:10.1073/pnas.1611243113 (2016).
141 Strawbridge, R. J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**, 2624-2634, doi:10.2337/db11-0415 (2011).
142 Taal, H. R. *et al.* Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat Genet* **44**, 532-538, doi:10.1038/ng.2238 (2012).
143 Teumer, A. *et al.* Genome-wide Association Studies Identify Genetic Loci Associated With Albuminuria in Diabetes. *Diabetes* **65**, 803-817, doi:10.2337/db15-1313 (2016).
144 van den Berg, S. M. *et al.* Meta-analysis of Genome-Wide Association Studies for Extraversion: Findings from the Genetics of Personality Consortium. *Behav Genet* **46**, 170-182, doi:10.1007/s10519-015-9735-5 (2016).
145 van der Valk, R. J. *et al.* A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Hum Mol Genet* **24**, 1155-1168, doi:10.1093/hmg/ddu510 (2015).
146 van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet* **48**, 1043-1048, doi:10.1038/ng.3622 (2016).
147 Watson, H. J. *et al.* Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat Genet* **51**, 1207-1214, doi:10.1038/s41588-019-0439-2 (2019).
148 Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* **14**, e1002383, doi:10.1371/journal.pmed.1002383 (2017).
149 Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668-681, doi:10.1038/s41588-018-0090-3 (2018).
150 Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649, doi:10.1093/hmg/ddy271 (2018).