Single nucleus analysis of the chromatin landscape in mouse forebrain development

Sebastian Preissl[1*], Rongxin Fang[1*], Yuan Zhao[1], Ramya Raviram[1], Yanxiao Zhang[1], Brandon C. Sos[2], Hui Huang[1,3], David U. Gorkin[1,4], Veena Afzal[5], Diane E. Dickel[5], Samantha Kuan[1], Axel Visel[5,6,7], Len A. Pennacchio[5,6,7], Kun Zhang[2] and Bing Ren[1,8†]

[1]Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, CA 92093, USA

[2]Department of Bioengineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

[3]UCSD Biomedical Sciences Graduate Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

[4]Department of Cellular and Molecular Medicine, Center for Epigenomics, University of California, San Diego, School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA

[5]Functional Genomics Department, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA

[6]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

[7]School of Natural Sciences, University of California, Merced, Merced, California, USA

[8]University of California, San Diego School of Medicine, Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, and Moores Cancer Center, 9500 Gilman Drive, La Jolla, CA 92093-0653, USA

*these authors contributed equally

† corresponding author

ABSTRACT

Genome-wide analysis of chromatin accessibility in primary tissues has uncovered millions of candidate regulatory sequences in the human and mouse genomes[1-4]. However, the heterogeneity of biological samples used in previous studies has prevented a precise understanding of the dynamic chromatin landscape in specific cell types. Here, we show that analysis of the transposase-accessible-chromatin in single nuclei isolated from frozen tissue samples can resolve cellular heterogeneity and delineate transcriptional regulatory sequences in the constituent cell types. Our strategy is based on a combinatorial barcoding assisted single cell assay for transposase-accessible chromatin[5] and is optimized for nuclei from flash-frozen primary tissue samples (snATAC-seq). We used this method to examine the mouse forebrain at seven development stages and in adults. From snATAC-seq profiles of more than 15,000 high quality nuclei, we identify 20 distinct cell populations corresponding to major neuronal and non-neuronal cell-types in foetal and adult forebrains. We further define cell-type specific *cis* regulatory sequences and infer potential master transcriptional regulators of each cell population. Our results demonstrate the feasibility of a general approach for identifying cell-type-specific *cis* regulatory sequences in heterogeneous tissue samples, and provide a rich resource for understanding forebrain development in mammals.

MAIN

A significant fraction of the non-coding DNA in the mammalian genome encodes transcriptional regulatory elements that play fundamental roles in mammalian development and human disease[3,6]. Identification of these sequences and characterizing their dynamic activities in specific cell types is a major goal in biology. Analysis of chromatin accessibility in primary tissues using assays such as DNase-seq[2,4] and ATAC-seq[7,8] has led to annotation of millions of candidate *cis* regulatory elements in the human and mouse genomes[1,3]. Yet, the catalogue of current *cis* regulatory elements, based primarily on analysis of bulk, heterogeneous biological samples, lacks precise information regarding cell-type- and developmental-stage-specific activities of each element. *In-vivo* lineage tracing using INTACT mouse models[8,9] and isolation of particular cell types based

on specific protein markers can address this limitation to some degree and in limited cell types[10,11]. But a more general strategy is necessary to study primary tissues from all stages of development in human as well as in other species.

In theory, single cell based chromatin accessibility studies can be used for unbiased identification of subpopulations in a heterogeneous population, and proof of principle has been reported using cultured mammalian cells and cryopreserved blood cell-types[5,12,13] . However, there has been no report that such approaches have been successfully used to dissect transcriptional regulatory landscapes in primary tissues. The main difficulty is that primary tissues are typically preserved either by formalin fixed paraffin embedding or flash freezing, conditions that are not amenable for isolating intact single cells. Here, we show that it is possible to isolate single nuclei from frozen tissues and assay chromatin accessibility in each nucleus in a massively parallel manner.

We adopted a combinatorial single cell ATAC-seq (scATAC-seq) strategy[5] and optimized it for frozen tissue sections (Methods). Compared to the previous protocol[5], key modifications were made to maximally preserve nuclei integrity during sample processing and optimize transposase-mediated fragmentation of chromatin in individual nuclei (Extended Data Fig. 1-3). We applied this improved protocol, hereafter referred to as snATAC-seq (single nuclei ATAC-seq), to mouse forebrain in the 8-week-old adult mouse (P56) and in seven developmental stages at 1-day intervals starting from embryonic day 11.5 (E11.5) to birth (P0) (Fig.1a, b). Sequencing libraries were sequenced to almost saturation as indicated by a read duplication rate of 36-73% per sample (Extended Data Table 1). We filtered out low quality datasets using three stringent quality control criteria including read depth (Extended Data Fig. 3d), recovery rate of constitutively accessible promoters in each nucleus (Extended Data Fig. 3e), and signal-over-noise ratio estimated by fraction of reads in peak regions (Method; Extended Data Fig. 3f). In total, 15,767 high-quality snATAC-seq datasets were obtained. The median read depth per nucleus ranged from 9,275 – 18,397, median promoter coverage was 11.6%, and the median fraction of reads in peak regions was 22% (Extended Data Table 2, 3). Our protocol maintains the

extraordinary scalability of combinatorial indexing, while featuring a ~ 6 fold increase in read depth per nucleus (Extended Data Table 3). The high quality of the single nuclei chromatin accessibility maps was supported by a high concordance between the aggregate snATAC-seq data and bulk ATAC-seq data ($R > 0.9$), and high reproducibility between biological replicates ($R > 0.91$, Fig. 1c, Extended Data Fig. 4).

The snATAC-seq profiles from each forebrain tissue arise from a mixture of distinct cell types. We reasoned that cells of the same type should share higher similarity in the open chromatin profiles than cells from different cell types. Based on this assumption, we developed a computational framework to uncover distinct cell types from the snATAC-seq datasets without prior knowledge of cell types in the tissue. Specifically, we first determined the open chromatin regions from the bulk ATAC-seq profiles of mouse forebrain tissue in seven fetal development time points and in adults, resulting in a total of 154,364 combined open chromatin regions that were detectable in one or more stages (Fig. 1d, Methods, Supplementary Table 1; Zhao et al. manuscript in preparation). Next, we constructed a binary matrix of open chromatin regions, using 0 and 1 to indicate absence and presence of a read at each open chromatin region, respectively, in each nucleus (Fig. 1d). Third, we calculated the pairwise similarity between cells using a Jaccard index. After applying a non-linear dimensionality reduction method, t-SNE[14], the Jaccard index matrix was projected to a low-dimension space to reveal cell clusters (Fig. 1d)[15]. Finally, we filtered out any cluster with abnormal sequencing depth or in-group similarity compared to other clusters[5]. We performed the clustering using the 140,103 distal elements (outside 2 kb upstream of refSeq transcription start site), since previous studies have shown highly cell type-specific chromatin accessibility profiles at enhancer regions[16], and that such sequences were more effective at classifying cell types than promoter or transcriptomics data[12] (Extended Data Fig. 5a, b).

To demonstrate the effectiveness of the above approach in uncovering cell type-specific chromatin landscapes from heterogeneous tissue samples, we first analysed the 3,033 snATAC-seq profiles obtained from the adult forebrain (Extended Data Table 2). As

negative controls, we included 200 “shuffled” nuclear profiles (Extended data Fig.5c, d, Methods). Initially, nine discrete cell populations, in addition to the group representing shuffled cells, were uncovered (Extended Data Fig. 5c, d). The cluster C2 (including 946 nuclei), like the shuffled cells, exhibited significantly lower intra-group similarity than other clusters, and thus were not included in further downstream analysis (Extended Data Fig. 5c, d). None of the t-SNE dimensions was correlated with read depth (R < 0.3 for all dimensions) and the clustering results were reproducible between two biological replicates (Extended data Fig. 5e, f). To categorize the final eight cell populations, we analysed transposase accessible chromatin at known cell type-specific gene loci and compared it to published data from sorted excitatory neurons[8], GABAergic neurons[9], microglia[17] and NeuN negative nuclei which mostly comprise non-neuronal cells including astrocytes and oligodendrocytes[18] (Fig. 2b, Extended Data Fig. 6a-c). Three cell populations and the sorted excitatory neurons showed high accessibility at the gene locus of the terminal neuronal differentiation factor *Neurod6* and other excitatory neuron-specific genes[19] (Fig. 2b, Extended Data Fig. 7a). Likewise, two cell clusters and the sorted GABAergic neurons showed similar accessibility at the GABA synthesis enzyme *Gad1* locus (Fig. 2b, Extended Data Fig. 7b)[20]. Using this strategy we were able to identify an astrocyte subpopulation according to the accessibility at the *Apoe* locus and other known astroglia markers[21], an oligodendrocyte subpopulation based on the myelin-associated glycoprotein *Mog* and other oligodendrocyte marker genes[22], and a microglia subpopulation based on complement factors including *C1qb* (Fig. 2b, Extended Data Fig. 7c-e)[17]. The categorization of cell groups was further confirmed by hierarchical clustering, with one remarkable exception that the inhibitory neuron cluster 2 (IN2) clustered with excitatory neurons (Fig. 2c). According to snATAC-seq data, the adult mouse forebrain consisted of 52% excitatory neurons, 24% inhibitory neurons, 12% oligodendrocytes and 6% astrocytes and microglia, respectively (Fig. 2d). Since the cell type proportion varies between different forebrain regions, for example cortex and hippocampus[19], the percentages derived from snATAC-seq represent an average of all forebrain region with numbers in between region-specific values (Extended Data Figure 6d, e; Fig.2e). The predominance of neuronal nuclei derived from adult forebrain tissue was confirmed by

flow cytometry analysis using staining against the post-mitotic neuron marker NeuN[18] (Extended Data Fig. 6b). Of note, the proportion of NeuN positive nuclei was lower than the total neuronal proportion derived from snATAC-seq (Extended Data Fig. 6b, e; Fig. 2e).

To further delineate the *cis*-regulatory landscape in each cell population of the adult forebrain, we plotted the frequency a cis regulatory element was accessible in a nucleus against the cell type specificity index of the element measured by the Shannon entropy of normalized read counts (Extended Data Fig. 8). Overall, proximal promoter elements were accessible in more cell types (Median value of 4.2 % for proximal elements vs. 0.4 % for distal elements) while the distal enhancer elements showed significantly higher cell type-specificity (Extended Data Fig. 8a, b, d). Next, to identify accessible chromatin regions that distinguish different cell populations, we developed a feature selection method (Methods), and used it to identify a total of 4,980 genomic elements that could separate the 8 nuclei populations in adult mouse forebrain (Fig. 2e, Extended Data Fig. 8c, d). We next performed k-means clustering against the 4,980 genomic regions and conducted motif enrichment analysis of each cluster of elements (Fig. 2e, f, Extended Data Fig. 8d, Supplementary Table 1). As expected, we observed enrichment of known transcription factor motifs in open chromatin in each cell population, including ETS-factor PU.1 for microglia[23], SOX family of proteins for oligodendrocytes[24], bHLH factors for excitatory neurons and DLX homeodomain factors for inhibitory neurons (Fig. 2f)[25]. Our analysis also revealed that MEIS binding motif was enriched in a subset of elements specific to IN2. Previous reports showed that MEIS2 plays a major role in generation of medium spiny neurons, the main GABAergic neurons in the striatum[26]. Accordingly, we identified gene loci of *Ppp1r1b* and *Drd1,* markers of medium spiny neurons, to be highly accessible in IN2 but not IN1 (Extended Data Fig. 9)[26]. Next, we asked if we could further separate the excitatory neurons to classes that reflect different anatomical areas. Hierarchical clustering with published bulk ATAC-seq data from different cortical layers and from dentate gyrus[9,27] showed that clusters EX1-3 might resemble different anatomical regions (Extended Data Fig. 10a). EX1 and EX3 represented upper and lower

cortical layers, respectively, whereas EX2 showed properties of dentate gyrus neurons (Extended Data Fig. 10a, b). Regions specific for EX1 and 3 were enriched for motifs from the Forkhead family and EX3 was enriched for motifs recognized by MEF2C (Extended Data Fig. 10c, Supplementary Table 1), which has been shown to play an important role in hippocampus mediated memory[28].

We next examined the snATAC-seq profiles derived from foetal mouse forebrains from seven developmental stages (Fig. 1b) that include key events from the onset of neurogenesis to gliogenesis[29]. From 12,733 high-quality single nuclei ATAC-seq profiles we identified 12 distinct sub-populations (Fig. 3a) that exhibit dynamic abundance through development (Fig. 3a-c). Based on accessibility profiles at gene loci of known marker genes, we assigned these cell populations to radial glia, excitatory neurons, inhibitory neurons, astrocytes and erythromyeloid progenitors (EMP) (Fig. 3b)[23,30]. Interestingly, the EMP cluster was restricted to E11.5, whereas the astrocyte cluster was present after E16.5 and expanded dramatically around birth (Fig. 3b, c)[23,29] , highlighting two developmental processes: invasion of myeloid cells into the brain prior to neurogenesis, and gliogenesis succeeding neurogenesis after E16.5[29]. Mature excitatory neurons (eEX2) were indicated by increased accessibility at the post-mitotic neuron marker gene *Neurod6* and absence of signal at the Notch effector *Hes5,* a marker gene for neuronal progenitors (Fig. 3b, c)[29,30]. This cell type expanded in abundance between E12.5 and E13.5 and followed the emergence of early differentiating neurons (eEX1, Fig. 3b, c). Remarkably, inhibitory-neuron-like cells were already present at E11.5 (Fig. 3b).

To identify the transcriptional regulatory sequences in each sub-population, we identified 16,364 genomic elements that show cell-population-specific chromatin accessibility and can best separate the sub-cell populations (Fig. 4a, Supplementary Table 2). We clustered these elements using k-means and performed gene ontology analysis of each cluster using the GREAT[31]. We also conducted *de novo* motif search for each group of elements to uncover transcription factor motifs enriched in cell type-specific open chromatin regions (Fig. 4b, c). Our analysis showed that genomic elements that were

mostly associated with radial glia like cell groups (Fig. 4a, RG1-4) fell into regulatory regions of genes involved in early forebrain developmental processes including “Forebrain regionalization” (Fig. 4b, K1), “Central nervous system development” (Fig. 4b, K3) or “Forebrain development” (Fig. 4b, K5). These elements were enriched for homeobox motifs corresponding to LHX-transcription factors including LHX2 (Fig. 4c, K1, 3, 5), which is critical for generating the correct neuron number by regulating proliferation of neural progenitors[32] and for temporal promoting of neurogenesis over astrogliagenesis[33]. Remarkably, one of these cluster was also enriched for both the proneural bHLH transcription factor ASCL1 (*Mash1*) and its co-regulator POU3F3 (*Brn1*) (Fig. 4c, K5)[34]. ASCL1 has been described to be required for normal proliferation of neural progenitor cells[35] and implicated in a DLX1/2 associated network that promotes GABAergic neurogenesis[36]. In line with this, associated genomic elements were also accessible in one inhibitory neuron cluster (eIN2, Fig. 4c, K5).

We identified transcriptional regulators that were specifically associated either with neurogenesis or gliogenesis during forebrain development. For example, the early astrocyte (eAC)-specific elements were located in open chromatin regions near genes involved in “glia cell fate commitment” and the top enriched transcription factor motif was NF1-halfsite (Fig.4a-c, K2). Previous studies showed that NF1 transcription factor NF1A alone is capable for specifying glia cells to the astrocyte lineage[24]. NFIX is another NF1 family member with proneural function[37]. This motif is enriched together with the bHLH transcription factor NEUROD1 binding sites mainly in open chromatin regions found in the excitatory neuron cell population (Fig4c, K4, 12, 13)[30]. Based on chromatin accessibility profiles at marker gene loci, we have previously assigned two cell clusters to excitatory neuron lineage (eEX1, eEX2, Fig.3b). Compared to cluster eEX2, eEX1 showed increased accessibility at both radial glia associated open chromatin (Fig.4a, K4; Fig.3b) and chromatin regions associated with “CNS neuron differentiation” (Fig.4a, K12). In addition, eEX1 nuclei preceded the emergence of eEX2 nuclei during development (Fig.3c). These findings indicate that eEX1 might represent a transitional state during excitatory neuron differentiation.

The bHLH transcription factor family consists of several subfamilies that recognize different DNA motifs[38]. NEUROD1 is part of a subfamily that binds to a central CAT motif whereas other factors such as TCF12 preferentially bind to a CAG motif[38]. Our snATAC-seq data revealed an enrichment of the TCF12-binding motif in regions associated with "Cortex GABAergic interneuron differentiation" in contrast to the excitatory neuron associated enrichment for NEUROD1 (Fig.4a-c, K4, 11, 12, 13)[25,30,39].

Analysis of inhibitory neuron cluster eIN3 specific genomic elements showed a remarkable enrichment for genes associated with "Skeletal muscle organ development" (Fig4a, b, K8). More detailed analysis revealed that the underlying genes were transcriptional regulators *Mef2c/d* and *Foxp1/2* as well as the dopamine receptors Drd2/3 indicating differentiating striatal medium spiny neurons[40,41]. This finding was consistent with the enrichment for MEIS-homeodomain factors in these regions (Fig.4c, K8) comparable to the medium spiny neuron cluster in adult forebrain (Fig.2e, f, K8; Extended Data Fig.9).

Lastly, genomic elements specific to the EMP cluster were associated with genes involved in "Myeloid cell development" (Fig.4a-c, K14) and enriched for motifs of the ubiquitous AP-1 transcription factor complexes that have been described to play a role in shaping the enhancer landscape of macrophages[42].

Next, we attempted to identify dynamic elements within a given cell clusters (Extended Data Fig.11). Our analysis revealed between 41 and 2,114 dynamic genomic elements for each cell type (Extended Data Fig.11c-g). Regions that are more accessible after birth (P0) compared to early time points were enriched for the RFX1 motif in the GABAergic neuron including the cluster eIN1 as well as in the excitatory neuron cluster eEX2 (Extended Data Fig.11d, e) indicating a general role of the evolutionary conserved RFX factors in perinatal adaptation of brain cells. Several family members including RFX1 are expressed in the brain and have been implicated to regulate cilia e.g. in sensory

neurons[43].

While assessment of open chromatin plays an important role in predicting regulatory elements in the genome[1,3], it does not provide direct information of functional activity. To address this point, we asked if cell-type specific transposase accessible chromatin in the embryonic forebrain overlaps with known enhancers validated by transgenic mouse assays[44]. We focused our analysis on all genomic elements with validated functional activity in the forebrain and a subset shown to be active only in the subpallium[45,46]. The subpallium is a brain region that gives rise to GABAergic and cholinergic neurons[45]. In total, 63.1 % (275/436) of all forebrain enhancer and 64.8% (59/91) of subpallial enhancer were represented in our subset of genomic elements, respectively, indicating a high degree of sensitivity. Next, we calculated the relative enrichment of subpallial enhancers over total forebrain enhancers for each cluster. Remarkably, subpallial enhancers were only enriched in clusters K9-11, which were assigned to the GABAergic neuron lineage (Fig.4d, e). This analysis confirms a high specificity and sensitivity of snATAC-seq experiments in identifying sub-cell populations and their underlying regulatory elements.

Taken together, we demonstrate here that snATAC-seq can be used to dissect heterogeneity and delineate gene regulatory sequences in complex tissues such as forebrain. Using this strategy, we were able to resolve the heterogeneity of primary tissue samples, and uncover both the cell types and the regulatory elements in each cell type without prior knowledge. The snATAC-seq approach will be a valuable tool for analysis of tissue biopsies and will help to pave the way to a better understanding of gene regulation in mammals.

AUTHOR CONTRIBUTIONS

Study was conceived and designed by B.R., S.P., R.F.; Study was overseen by B.R.; Experiments performed by S.P., B.C.S, H.H.; Tissue collection by V.A., D.E.D., S.P.; Sequencing performed by S.K.; Computational strategy developed by R.F.
Data analysis performed by S.P., R.F., R.R., Y.Z., D.U.G; Manuscript written by S.P., R.F. and B.R. with input from all authors.

ACKNOWLEDGMENTS

This study was funded in part by the National Human Genome Research Institute as part of the Encyclopedia of DNA Elements (ENCODE) project (U54HG006997) and supported by funding from the Ludwig Institute for Cancer Research and NIH (2P50 GM085764). S.P. was supported by a postdoctoral fellowship from the Deutsche Forschungsgemeinschaft (DFG, PR 1668/1-1). RR was supported by a Ruth L. Kirschstein National Research Service Award NIH/NCI T32 CA009523. We thank Bin Li for bioinformatic support. We thank Molly He and Trina Osothprarop for providing the Tn5 enzyme. We thank Derek Gao for sequencing on the MiSeq. This study was supported by the NIH grant U01MH098977 to K.Z. Research conducted at the E.O. Lawrence Berkeley National Laboratory was performed under U.S. Department of Energy Contract DE-AC02-05CH11231, University of California.

FIGURES

Fig.1

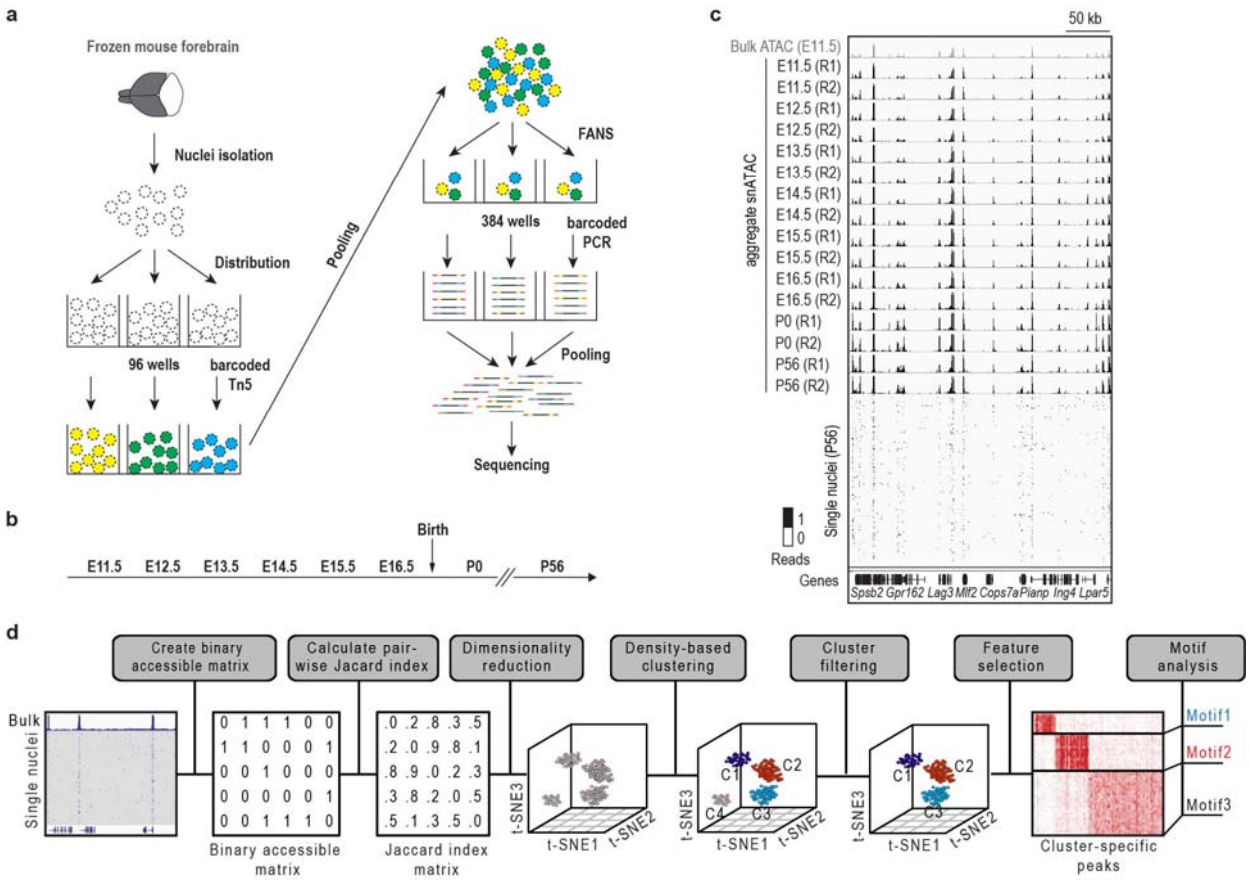


Figure 1: Experimental overview and computational analysis strategy for snATAC-seq. a Following nuclei isolation from frozen forebrain tissue biopsies, tagmentation of 4,500 permeabilized nuclei was carried out using barcoded Tn5 in 96 wells. After pooling, 25 nuclei were sorted in 384 wells and PCR-amplified to introduce second barcodes. FANS: Fluorescence assisted nuclei sorting. **b** Overview of investigated time points during mouse development. E: embryonic day; P: postnatal day; **c** Chromatin accessibility profiles of aggregate snATAC-seq (black tracks) agree with bulk ATAC-seq (grey, top track) and are consistent between biological replicates. **d** Framework of computational analysis of snATAC-seq data.

Fig.2

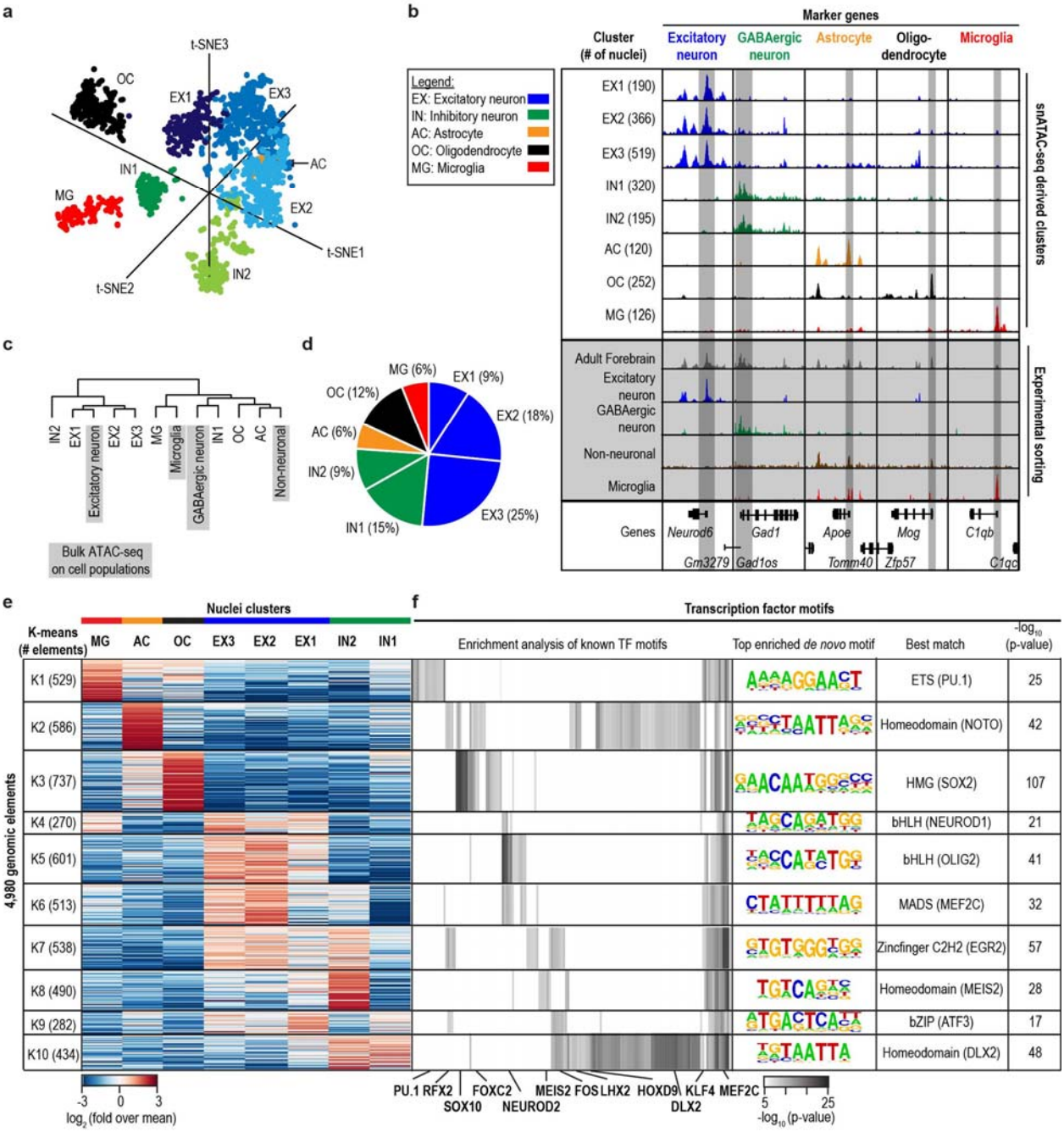


Figure 2: Deconvolution of p56 forebrain and identification of potential master regulators. a Clustering of single nuclei of both replicates revealed 8 different cell groups in adult forebrain. **b** Aggregate chromatin accessibility profiles for each cell cluster and bulk ATAC-seq for sorted cell populations or whole forebrain at marker gene loci (Bulk

data are grey shaded). **c** Hierarchical clustering of aggregate single cell data and sorted bulk data sets. **d** Cellular composition of adult forebrain derived from snATAC-seq data. **e** K-means clustering of 4,980 genomic elements based on chromatin accessibility and **f** enrichment analysis for transcription factor motifs.

Fig.3

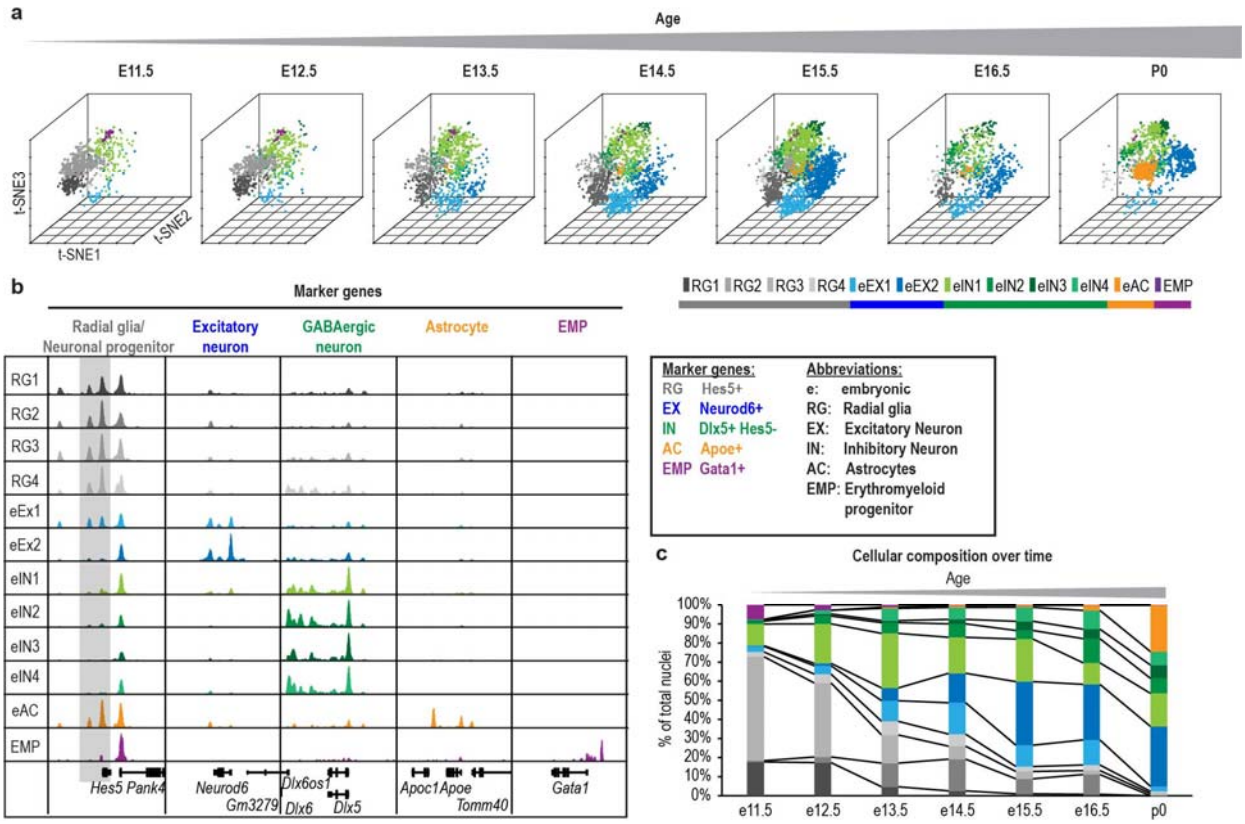


Figure 3: SnATAC-seq analysis reveals cellular heterogeneity during embryonic forebrain development. a Clustering of single nuclei from both replicates revealed 12 different cell groups with changing relative abundance. **b** Aggregate chromatin accessibility profiles for cell clusters and at marker gene loci used to assign cell types. For better visualization, *Hes5* gene locus is grey shaded. **c** Quantification of cellular composition during forebrain development.

Fig.4

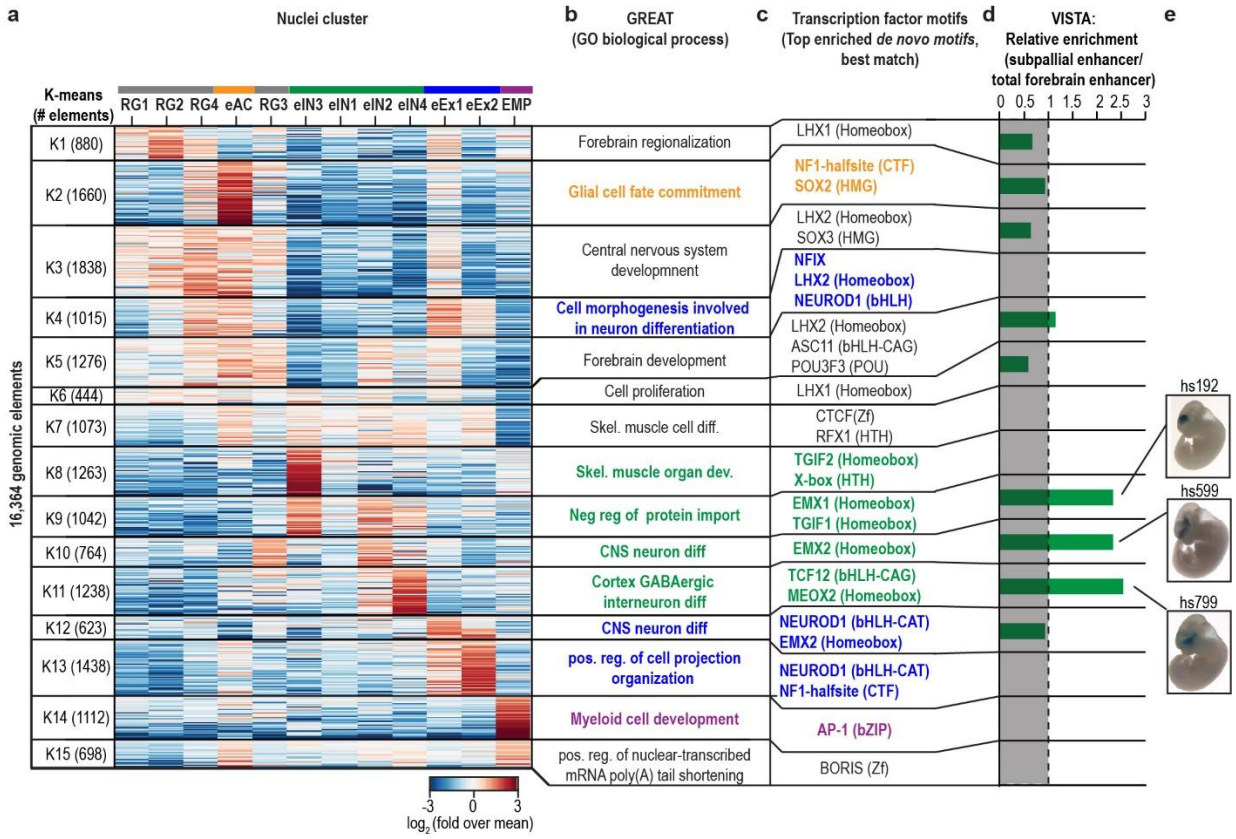
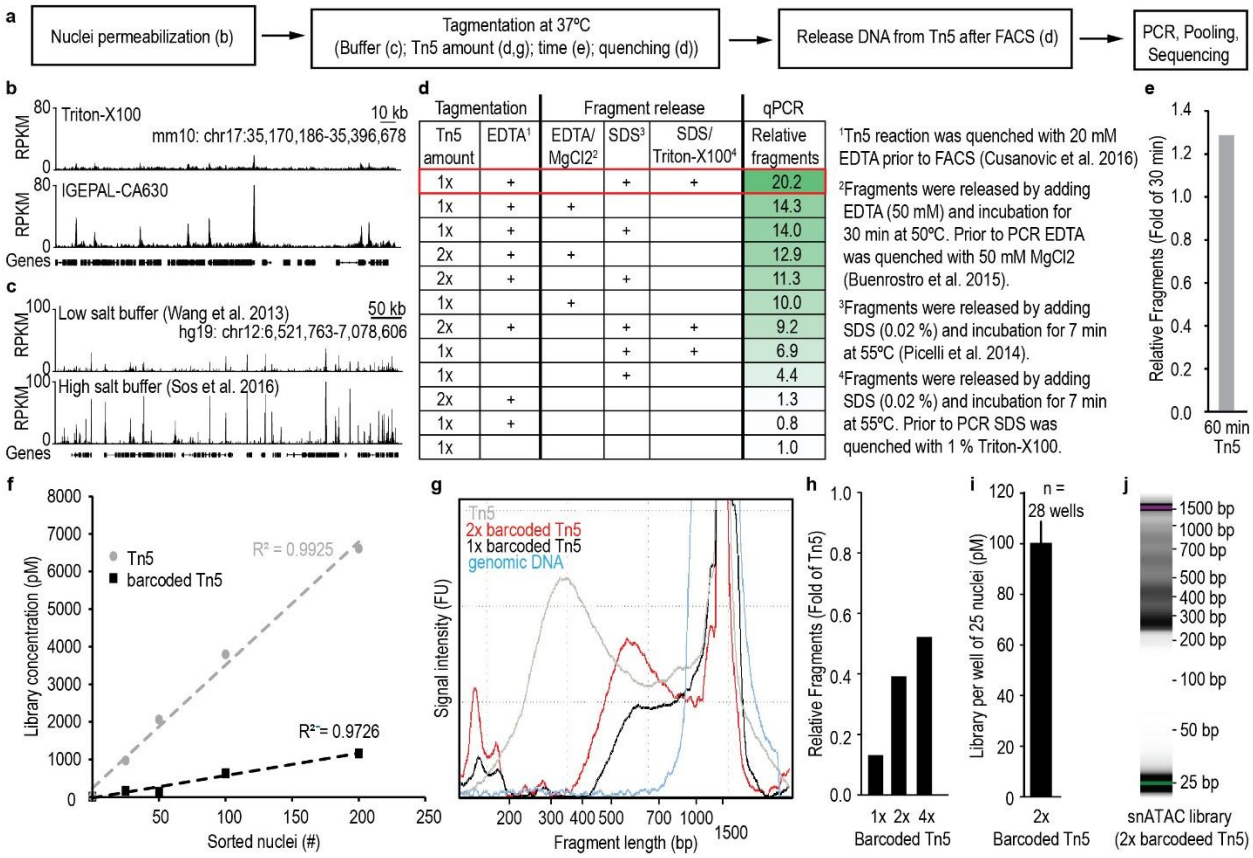


Figure 4: SnATAC-seq revealed genomic elements and transcriptional regulators of lineage specification in the developing forebrain. a K-means clustering of 16,364 genomic elements based on chromatin accessibility. **b** Gene ontology analysis using GREAT and **c** transcription factor enrichment. **d** Enrichment of enhancers that were functionally validated as part of the VISTA database. **e** Representative pictures of transgenic mouse embryos showing LacZ reporter gene expression under control of the indicated subpallial enhancers. Pictures were downloaded from the VISTA database[44].

EXTENDED DATA FIGURES

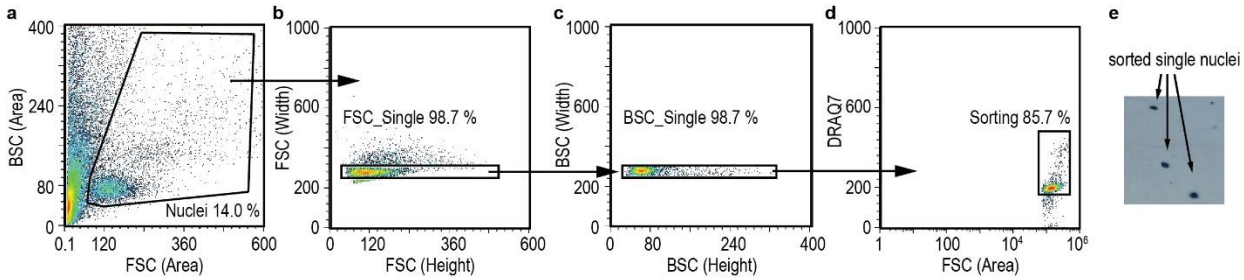
Extended Data Fig.1



Extended Data Figure 1: SnATAC-seq protocol optimization. a Overview of critical steps for the snATAC-seq procedure for nuclei from frozen tissues. **b** IGEPAL-CA630 but not Triton-X100 was sufficient for tagmentation of frozen tissues. **c** Tagmentation was facilitated by high salt concentrations in reaction buffer[47,48]. **d** Maximum amount of fragments per nucleus could be recovered when quenching Tn5 by EDTA prior to FACS and denaturation of Tn5 after FACS by SDS. Finally, SDS was quenched by Triton-X100 to allow efficient PCR amplification. **e** Increasing tagmentation time from 30 min to 60 min can result in more DNA fragments per nucleus. **f** Number of sorted nuclei was highly correlated with the final library concentration. Tn5 loaded with barcoded adapters showed less efficient tagmentation as compared to Tn5 without barcodes. Wells were amplified for 13 cycles, purified and libraries quantified by qPCR using standards with known molarity. **g** Tagmentation with barcoded Tn5 was less efficient and resulted in larger

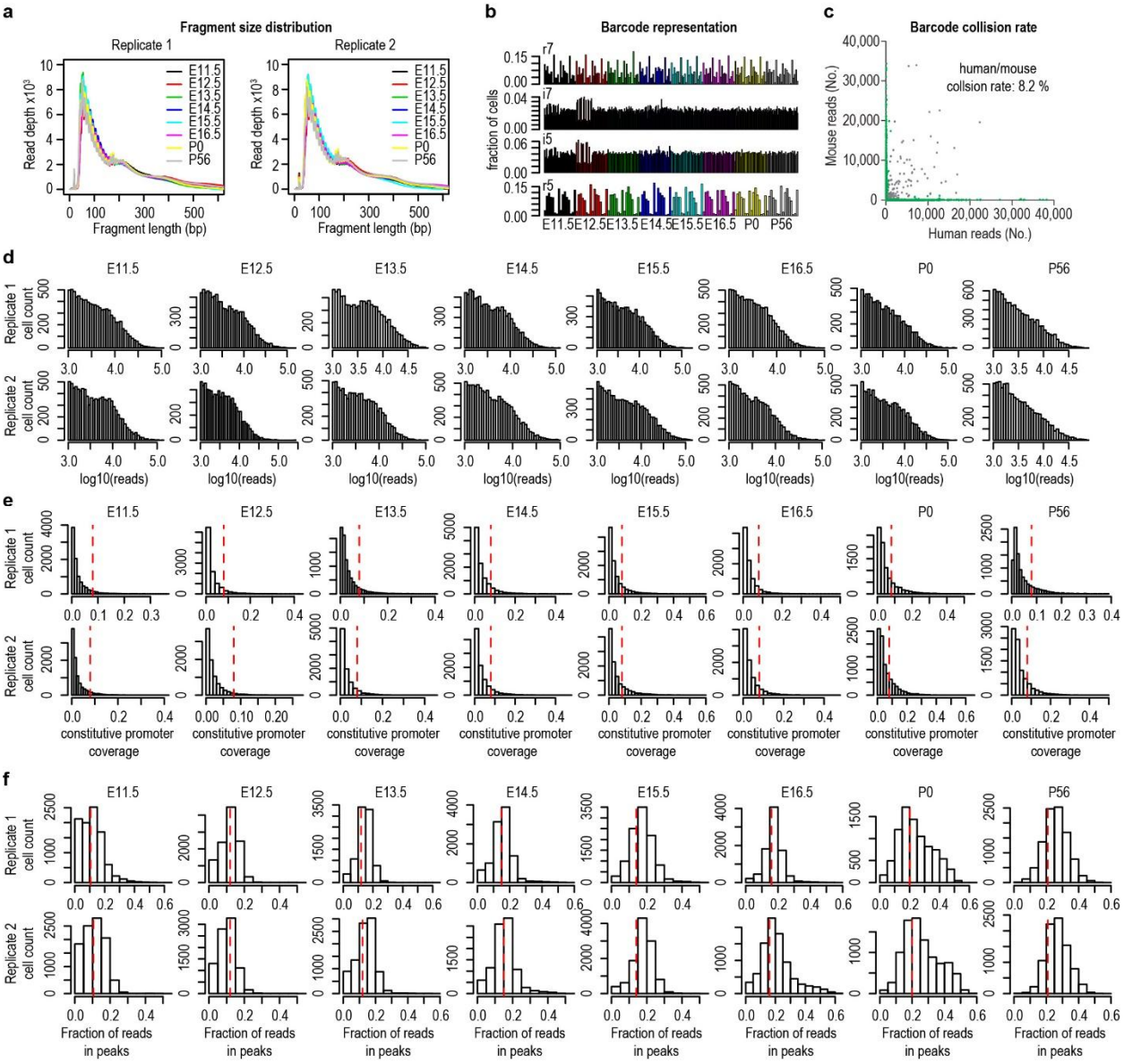
fragments than Tn5 (550 bp vs. 300 bp). Ratio for barcoded Tn5 was based on concentration of regular Tn5. **h** Doubling the concentration of barcoded Tn5 significantly increased the number of fragments per nucleus by 3 fold. Further increase resulted only in minor improvements. **i** Generated amount of library from 25 nuclei per well was reproducible between single wells. Each well was amplified for 11 cycles and quantified by qPCR. This output was used to calculate the number of required PCR cycles for snATAC-seq libraries to prevent overamplification (n = 28 wells, average ± SEM). **j** Size distribution of a successful snATAC-seq library from a mixture of E15.5 forebrain and GM12878 cells shows a nucleosomal pattern. SnATAC-seq was performed including all the optimization steps described above with barcoded Tn5 in 96 well format.

Extended Data Fig.2



Extended Data Figure 2: Sorting of single nuclei after tagmentation. a-d Density plots illustrating the gating strategy for single nuclei. First, big particles were identified (**a**), then duplicates were removed (**b**, **c**) and finally, nuclei were sorted based on high DRAQ7 signal (**d**), which stains DNA in nuclei. **e** Verification of single cell suspension after FACS was done with Trypan Blue staining under a microscope.

Extended Data Fig.3

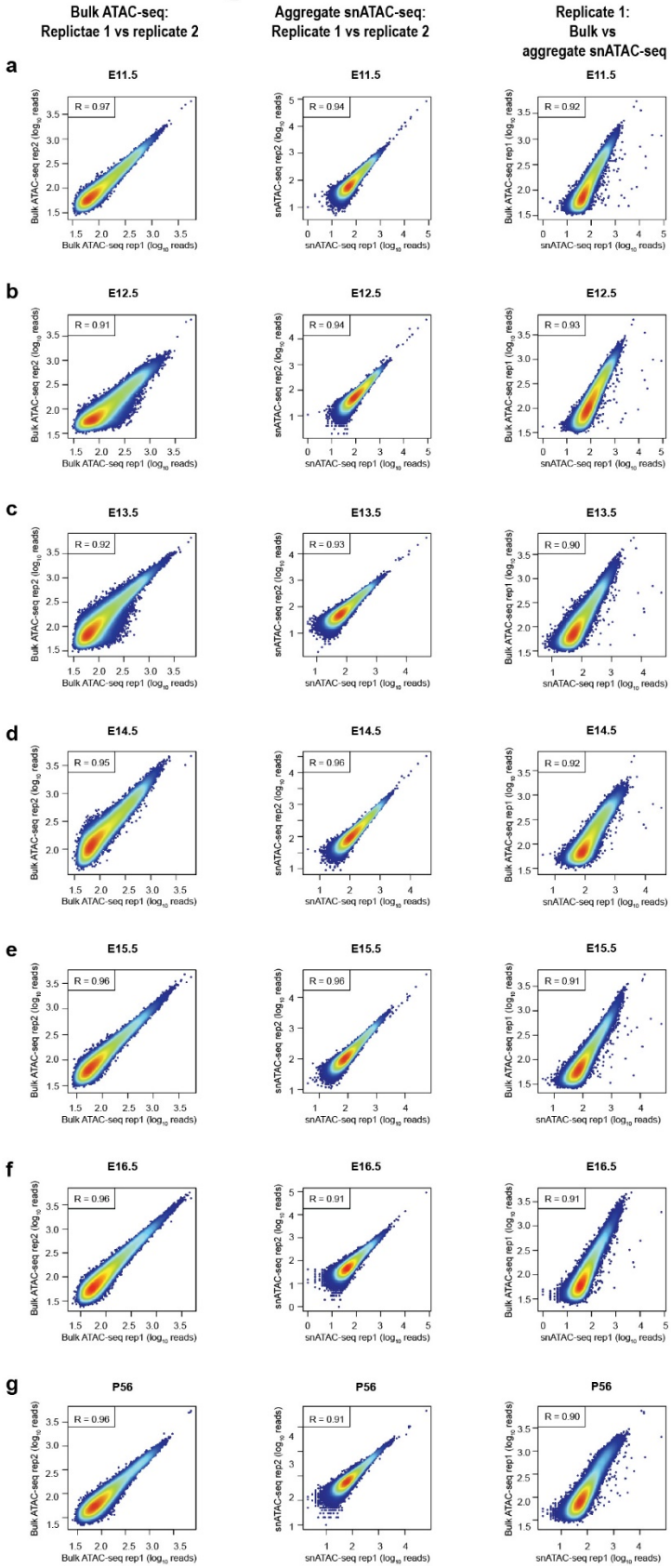


Extended Data Figure 3: Overview of sequencing data and quality filtering for single cells.

a Distribution of insert sizes between reads pairs derived from sequencing of snATAC-seq libraries indicates nucleosomal patterning. **b** Individual barcode representation in the final library shows variability between barcodes. **c** To assess the probability of two cells sharing the same cell barcode, single cell ATAC-seq was performed on a 1:1 mixture of human GM12878 cells and mouse E15.5 forebrain nuclei. A collision was indicated by <

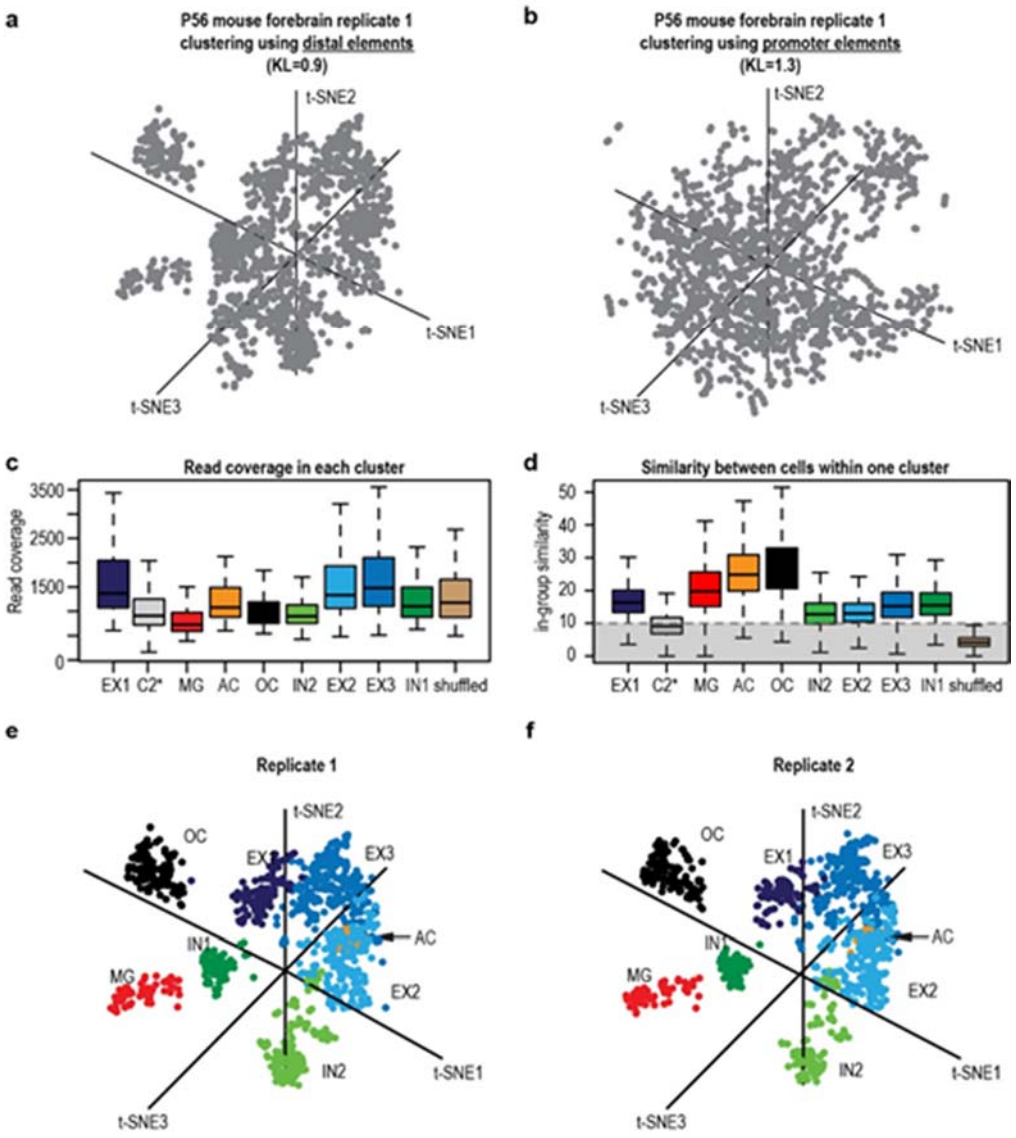
90% of all reads mapping to either the mouse genome (mm9) or the human genome (hg19). We identified 8.2% of these barcode collision events. **d** Read coverage per barcode combination after removal of potential barcodes with less than 1,000 reads. **e** Constitutive promoter coverage for each single cell. The red line indicates the constitutive promoter coverage in corresponding bulk ATAC-seq data sets from the same biological sample. Cells with less coverage than the bulk ATAC-seq data set were discarded. **f** Fraction of reads falling into peaks for each single cell. The red line indicates fraction of reads in peak regions in corresponding bulk ATAC-seq data sets from the same biological sample. Cells with lower reads in peak ratios coverage than the bulk ATAC-seq data set were discarded from downstream analysis. Bulk ATAC-seq data for E11.5-P0 were reanalysed (Zhao et al. manuscript in preparation)

Extended Data Fig.4



Extended Data Figure 4: Pearson correlation plots of bulk and aggregate single-nuclei ATAC-seq data sets. Pearson correlation of chromatin accessibility profiles from two biological replicates derived from bulk ATAC-seq (left column) and from aggregate snATAC-seq after aggregating single cell profiles (middle column). In the right column the correlation between bulk ATAC-seq and aggregate snATAC-seq are displayed for biological replicate 1. Data are displayed from forebrain tissues from following time points: **a** E11.5, **b** E12.5 **c** E13.5 **d** E14.5 **e** E15.5 **f** E16.5 **g** P0 **h** P56. Bulk ATAC-seq data for E11.5-P0 were reanalysed (Zhao et al. manuscript in preparation).

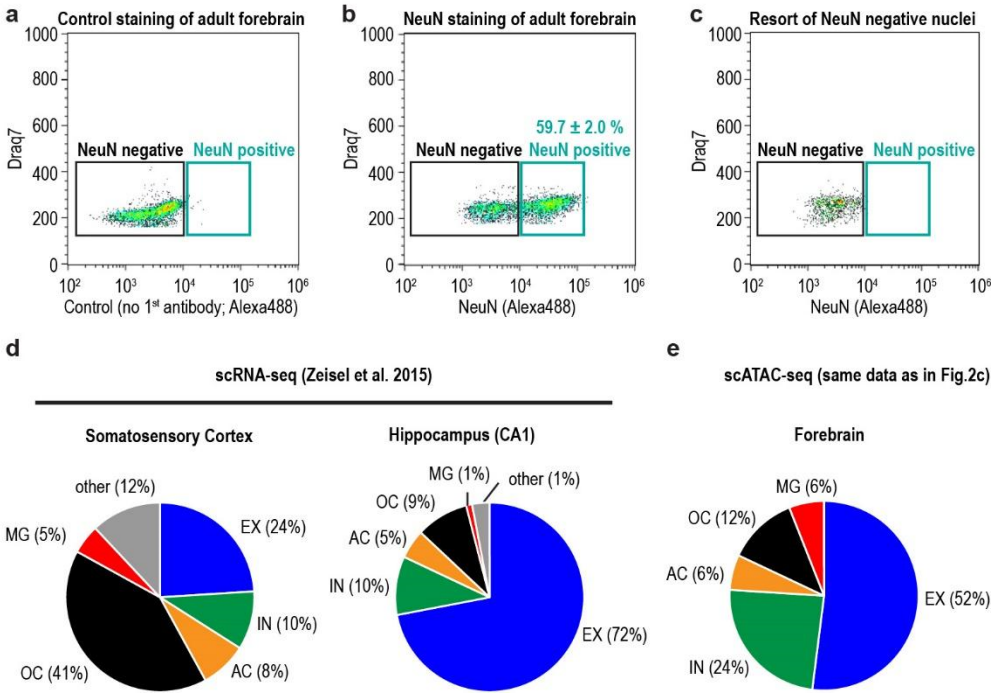
Extended Data Fig.5



Extended Data Figure 5: Clustering strategies, quality control of clusters and clustering result for individual replicates in adult forebrain. a, b T-SNE visualization of clustering using **a** distal elements (regions outside 2 kb of refSeq transcriptional start sites) or **b** promoter regions (KL: Kullback-Leibler divergence reported by t-SNE).**c** Box plot of read coverage for each cluster. **d** Box plot of similarity analysis between two any two given cells in a cluster. Cluster C2 was discarded before downstream analysis due to low its intra-group similarity (median < 10). As a negative control, randomly shuffled cells were included in the analysis displaying exceptionally low in-group similarity. **e, f** T-SNE

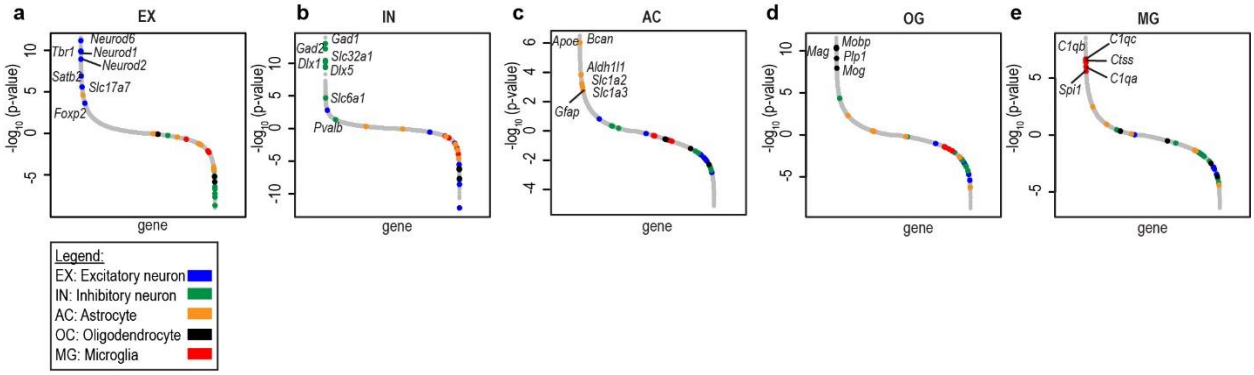
visualization of single cells from **e** replicate 1 and **f** replicate 2. The projection and color coding is the same as in Fig. 2d.

Extended Data Fig.6



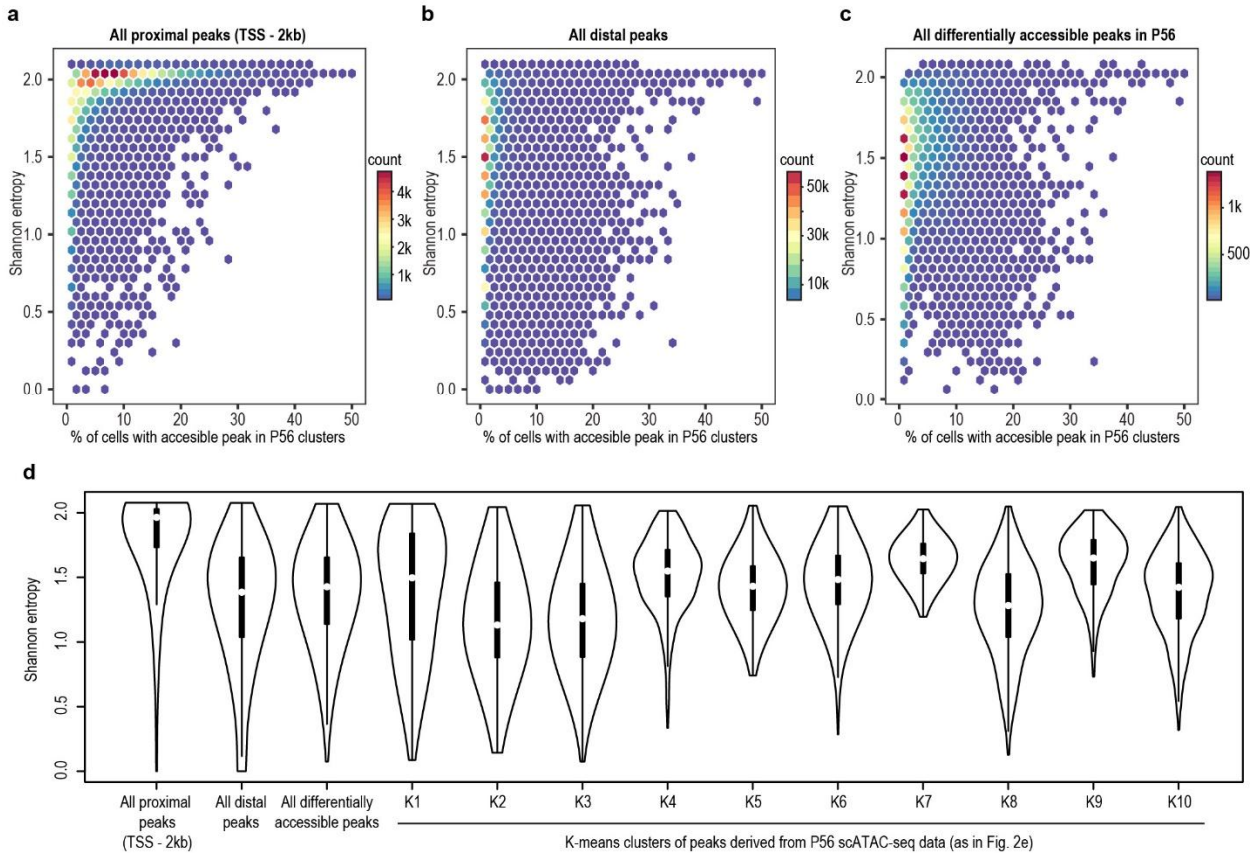
Extended Data Figure 6: Flow cytometric analysis of adult mouse forebrain and comparison to single cell RNA-seq data from different brain regions a-c Dot blots illustrating nuclei from adult forebrain stained for flow cytometry with Alexa488 conjugated secondary antibodies. **a** Displayed are representative blots for experiments without antigen specific primary antibody and **b** with antibodies recognizing the post-mitotic neuron marker NeuN[18] (n=3, average ± SEM)**. c** NeuN negative nuclei were sorted for ATAC-seq experiments and purity (> 98%) was confirmed by flow cytometry of the sorted population. **d** Relative composition of different forebrain regions derived from single cell RNA-seq shows region specific differences[19]**. e** Relative composition derived from snATAC-seq (compare to Fig.2c) of adult forebrain shows values in between.

Extended Data Fig.7



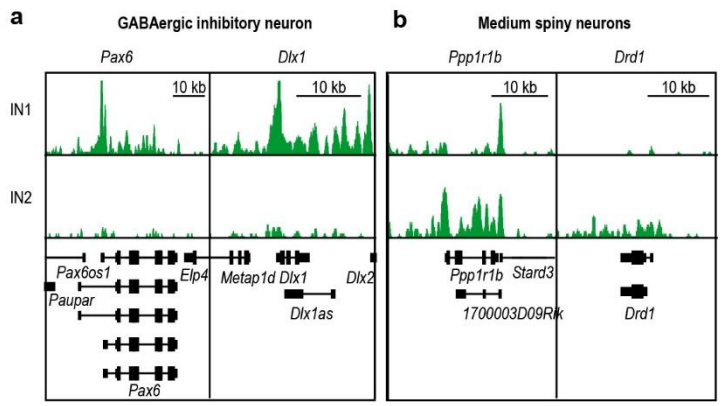
Extended Data Figure 7: Ranking of gene loci (TSS ± 10kb) compared to other clusters in adult forebrain. Negative binomial test shows enrichment for **a** excitatory neuron markers **b** inhibitory neuron markers **c** astrocyte markers **d** oligodendrocyte markers and **e** microglia markers extending the examples shown in Fig. 2b. Please note for general assignment accessibility profiles for Ex1-3 and IN1/2 were merged, respectively.

Extended Data Fig.8



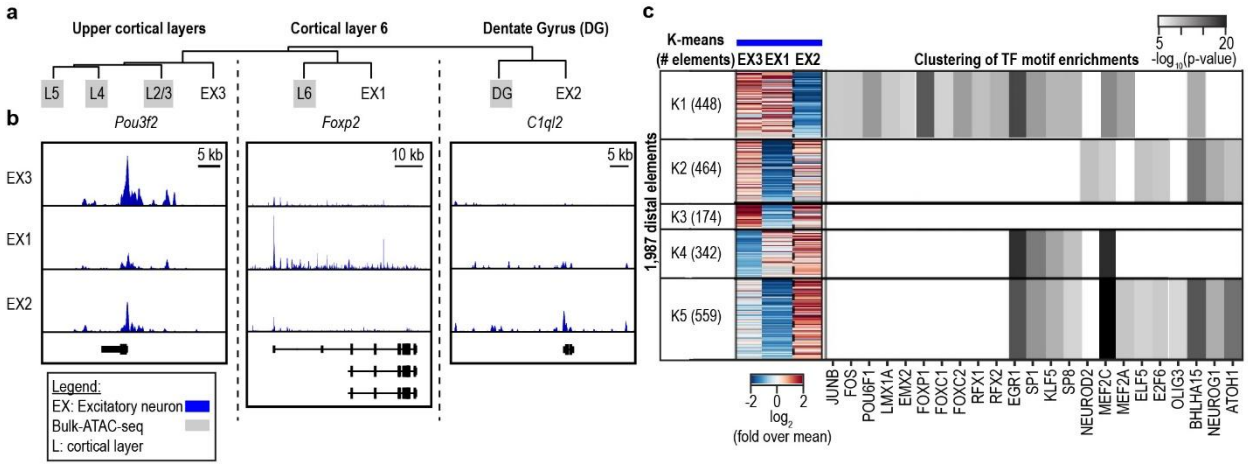
Extended Data Figure 8: Cell-type specificity of genomic elements and per cell coverage of elements. a-c Graphs illustrate cell-type specificity of genomic elements as measured by Shannon entropy based on normalized read counts for each cell-type and percentage of cells in which a genomic element was called accessible as indicated by presence of at least 1 read overlapping with the element a peak. Analysis was performed for the adult forebrain (P56) against **a** TSS-proximal genomic elements (TSS - 2kb), **b** distal elements and **c** the subset of genomic elements that separated two cell clusters. **d** Violin plots illustrate higher cell-type specificity for distal elements compared to proximal elements indicated by significantly lower Shannon entropy value ($p < 2.2e-16$). In addition, distribution of all genomic elements that separate two clusters as well as subsets representing subsets identified from k-means clustering of genomic elements depending on chromatin accessibility in adult forebrain (related to Fig. 2e). TSS: transcriptional start site.

Extended Data Fig.9



Extended Data Figure 9: Distinct chromatin accessibility profiles of two GABAergic neuron clusters. IN2 is depleted for *Pax6* and *Dlx1* (**a**) but enriched for markers of medium spiny neurons as compared to IN1 cluster (**b**).

Extended Data Fig.10



Extended Data Figure 10: Sub-classification of excitatory neurons into hippocampal and cortical neuron types. a Hierarchical clustering of aggregate single cell data for excitatory neuron cluster and sorted bulk data sets corresponding to different anatomical regions (grey shaded). **b** Chromatin accessibility at marker gene loci. **c** K-means clustering of promoter distal genomic elements and enrichment analysis for transcription factor motifs.

Extended Data Fig. 11

a Reads in peaks per cell cluster at each time point

Cell cluster \ Time point	E11.5	e12.5	e13.5	e14.5	e15.5	e16.5	p0
RG1	1,077,714	925,916	253,274	166,469	125,898	20,118	1,401
RG2	34,368	212,006	767,951	1,262,027	1,186,931	356,975	7,909
eEX1	161,911	208,487	668,184	1,320,188	2,114,083	506,033	265,608
eEX2	13,135	67,121	382,525	953,300	4,811,953	858,302	1,550,411
RG3	2,512,140	1,759,240	697,854	446,290	448,451	56,178	39,685
RG4	122,932	239,670	378,461	463,678	336,554	84,909	51,263
eIN1	475,169	918,028	1,211,824	1,164,612	2,916,576	300,165	762,899
eIN2	81,280	203,870	323,656	474,479	587,100	387,242	346,968
eIN3	15,504	50,410	51,024	163,359	751,039	155,228	291,443
eIN4	23,537	83,326	275,184	400,720	935,362	285,780	320,223
eAC	0	0	35,917	79,090	148,759	79,918	1,168,309
EMP	174,341	49,393	20,082	11,265	14,659	5,549	6,050

b Nuclei per cell cluster at each time point

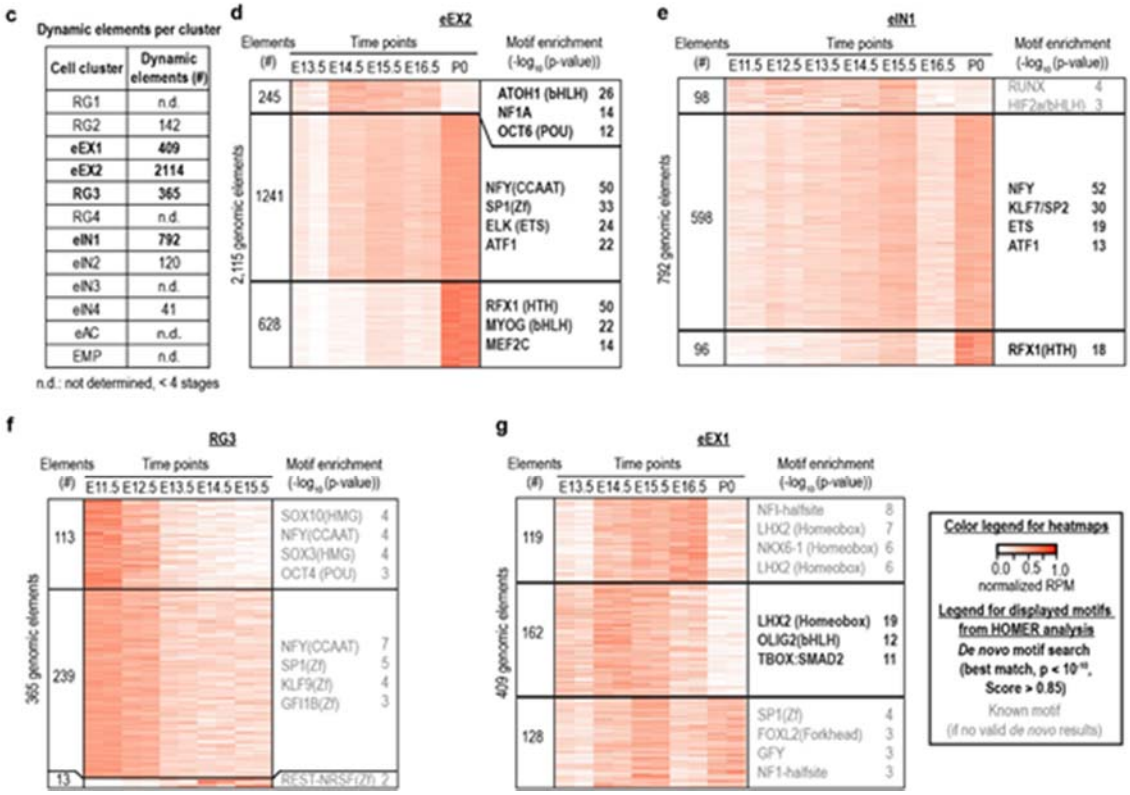
Cell cluster \ Time point	e11.5	e12.5	e13.5	e14.5	e15.5	e16.5	p0
RG1	216	189	62	49	41	9	1
RG2	7	32	161	312	303	109	3
eEX1	36	47	141	309	427	138	63
eEX2	4	14	87	290	1274	308	744
RG3	661	420	191	129	145	22	26
RG4	30	51	92	119	100	31	25
eIN1	137	224	370	354	854	118	412
eIN2	19	45	73	133	168	134	186
eIN3	4	11	14	44	184	53	165
eIN4	7	22	83	117	281	103	168
eAC	0	0	8	19	39	29	575
EMP	92	30	16	7	11	2	5

Color code for A, B: > 50 nuclei; > 50 nuclei and > 250,000 reads in peaks

c Dynamic elements per cluster

Cell cluster	Dynamic elements (#)
RG1	n.d.
RG2	142
eEX1	409
eEX2	2114
RG3	365
RG4	n.d.
eIN1	792
eIN2	120
eIN3	n.d.
eIN4	41
eAC	n.d.
EMP	n.d.

n.d.: not determined, < 4 stages



Extended Data Figure 11: Dynamics of chromatin accessibility within distinct cell groups. a Number of reads in peaks per developmental time point for a specific nuclei cluster. **b** Number of nuclei per time point for a specific nuclei cluster. For analysis of dynamics only cell clusters with > 3 stages with > 50 nuclei and > 250,000 reads in peaks were considered. **c** Overview of dynamic elements identified per cell cluster (see methods) **d-g** K-means clustering and motif enrichment analysis for nuclei clusters with > 200 dynamic genomic elements.

EXTENDED DATA TABLES

Extended Data Table 1

Description	DCC Biosample	Barcodes	Raw reads for both replicates (M)	Uniquely mapped reads	Reads with valid barcodes	Estimated PCR duplication rate	Mitochondrial reads	Final read count (M)	Reads for each replicate (M)
E115_Rep1	ENCBS745VOM	Set_1; p5 only 7 bp called	683	82%	46%	49%	4%	116	54
E115_Rep2	ENCBS720JOH	Set_2; p5 only 7 bp called							62
E125_Rep1	ENCBS974GPH	Set_1	491	85%	71%	56%	5%	113	68
E125_Rep2	ENCBS617DPG	Set_2							46
E135_Rep1	ENCBS273RVL	Set_1	486	84%	46%	36%	2%	109	57
E135_Rep2	ENCBS793ZRY	Set_2 for p5, p7 and i7; i5: S502-S511;X528-X535							52
E145_Rep1	ENCBS002AAA	Set_1	436	87%	58%	39%	4%	118	61
E145_Rep2	ENCBS014AAA	Set_2							57
E155_Rep1	ENCBS811RMD	Set_1	476	84%	93%	53%	2%	160	72
E155_Rep2	ENCBS841GER	Set_2							88
E165_Rep1	ENCBS156DHU	Set_1	545	88%	34%	39%	4%	86	43
E165_Rep2	ENCBS512LKA	Set_2							43
P0_Rep1	ENCBS038AAA	Set_1	361	85%	83%	52%	2%	108	51
P0_Rep2	ENCBS554AAA	Set_2							57
P56_Rep1		Set_1	680	89%	65%	73%	4%	90	47
P56_Rep2		Set_2							44
hu_GM12878_mu_E155_mix		p5:1-8, p7:1-12 i5: S502-S511; i7: S701-S715	27	87%	89%	38%	2%	13	

Barcode sets (for sequences see Supplementary Table 1)

Set_2	
p5	1-8
p7	1-12
i5	X520-X535
i7	N731-X754

Set_1	
p5	1-8
p7	1-12
i5	S502-S511,X512-X519
i7	N701-N729,N730

Extended Data Table 1: Sequencing statistics for single nuclei ATAC-seq libraries. General overview of sequencing for single cell ATAC-seq libraries including PCR duplication rates and fraction of mitochondrial reads. Please note that paired end reads were treated as separate reads. Replicate 1 and 2 were sequenced together and single cell datasets were assigned based on replicate specific barcode combinations (Set_1 or Set_2). One exception was E13.5 where replicate 1 and 2 were sequenced on separate lanes. Please note that for E11.5 7 out of 8 bp were detected for the p5 barcode. M: million

Extended Data Table 2

Sample	Cells pass QC (#)	Reads per cell (median)	Promoter coverage (median)	Fraction of reads in peaks (median)
E11.5_Rep1	528	16,023	10.6%	17.5%
E11.5_Rep2	685	16,499	11.1%	17.5%
E12.5_Rep1	781	18,397	11.2%	15.7%
E12.5_Rep2	303	17,945	10.3%	14.9%
E13.5_Rep1	651	16,891	11.8%	18.4%
E13.5_Rep2	646	14,669	11.1%	18.7%
E14.5_Rep1	976	14,440	11.3%	19.8%
E14.5_Rep2	905	14,489	11.5%	19.2%
E15.5_Rep1	1,591	16,789	12.2%	20.4%
E15.5_Rep2	2,235	17,840	12.9%	20.5%
E16.5_Rep1	317	12,758	11.0%	23.2%
E16.5_Rep2	738	10,460	11.4%	29.3%
P0_Rep1	1,044	9,275	13.3%	36.0%
P0_Rep2	1,333	8,703	12.6%	37.4%
P56_Rep1	1,569	12,509	11.6%	27.1%
P56_Rep2	1,465	12,689	11.5%	26.4%
Average	985	14,399	11.6%	22.6%
Total	15,767			

Extended Data Table 2: Overview of single nuclei ATAC-seq data after filtering out low quality cells. Overview of cells that pass quality control and general properties of data sets including promoter coverage and fraction of reads in peaks.

Extended Data Table 3:

Parameter	this study	Cusanovich et al. 2015[5]	Buenrostro et al. 2015[12]
Single cell strategy	Combinatorial indexing	Combinatorial indexing	Microfluidics
Sample type	Frozen tissues	Cell lines	Cell lines
Total cells/nuclei (per experiment)	15,676 (303-2,235)	15,814 (533-1,459)	1886 (n.r.)
Reads per cell/nucleus (median)	9,375-18,397	1,390-3,094	73,000 (average)
Fraction of reads in peaks	15.7%-37.4%	41-59 %	>15%
Promoter coverage per cell/nucleus	11.60%	n.r.	9.40%

Extended Data Table 3: Comparison of single nuclei ATAC-seq with previously published initial single cell/nuclei ATAC-seq studies. Table illustrating several characteristics of single nuclei/cell ATAC_seq library. n.r. not reported

METHODS

Mouse tissues

All animal experiments were approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee or the University of California, San Diego, Institutional Animal Care and Use Committee. Forebrains from embryonic mice (E11.5-E16.5) and early postnatal mice (P0) were dissected from one pregnant female or one litter at a time and combined. For breeding, animals were purchased from Charles River Laboratories (C57BL/6NCrl strain) or Taconic Biosciences (C57BL/6NTac strain) for E14.5 and P0. Breeding animals for other time points were received from Charles River Laboratories (C57BL/6NCrl). Dissected tissues were flash frozen in a dry ice ethanol bath. For the adult time point (P56), the forebrain from 8-week old male C57BL/6NCrl mice (Charles River Laboratories) were dissected and flash frozen in liquid nitrogen separately. Tissues were pulverized in liquid nitrogen using pestle and mortar. For each time point two biological replicates were processed (n = 2 per time point).

Transposome generation

To generate A/B transposomes, A and B oligos were annealed to common pMENTs oligos (95°C 2 min, 14°C ∞ (cooling rate: 0.1°C/s)) separately (Supplementary Table 2). Next, barcoded transposons were mixed in a 1:1 molar ratio with unloaded transposase Tn5 which was generated at Illumina. Mixture was incubated for 30 min at room temperature. Finally, A and B transposomes were mixed. For combinatorial barcoding we used 8 different A transposons and 12 distinct B transposons which eventually resulted in 96 barcode combinations (Supplementary Table 2)[49].

Combinatorial barcoding assisted single nuclei ATAC-seq

Combinatorial ATAC-seq was performed as described previously with modifications[5]. 5-10 mg frozen tissue was transferred to a 1.5 ml Lobind tube (Eppendorf) in 1 ml NPB (5 % BSA (Sigma), 0.2 % IGEPAL-CA630 (Sigma), cOmplete (Roche), 1 mM DTT in PBS) and incubated for 15 min at 4 °C. Nuclei suspension was filtered over a 30 µm Cell-Tric

(Sysmex) and centrifuged for 5 min with 500 x g. Nuclei pellet was resuspended in 500 µl of 1.1x DMF buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF) and nuclei were counted using a hemocytometer. Concentration was adjusted to 500 /µl and 4500 nuclei were dispensed into each well of a 96 well plate. For tagmentation, 1 µl barcoded Tn5 transposome (0.25 µM, Supplementary Table 2)[49] was added to each well, mixed 5 times and incubated for 60 min at 37°C with shaking (500 rpm). To quench the reaction 10 µl 40 mM EDTA were added to each well and plate was incubated at 37°C for 15 min with shaking (500 rpm). 20 µl sort buffer (2 % BSA, 2 mM EDTA in PBS) were added to each well and all wells combined afterwards. Nuclei suspension was filtered using a 30 µm CellTric (Sysmex) into a FACS tube and 3 µM Draq7 (Cell Signalling) was added. Using a SH800 sorter (Sony) 25 nuclei were sorted per well into 4 96-well plates (total of 384 wells) containing 18.5 µl EB (50 pM Primer i7 (Supplementary Table 2), 200 ng BSA (Sigma)). Sort plates were shortly spun down. After addition of 2 µl 0.2 % BSA samples were incubated at 55°C for 7 min with shaking (500 rpm). 2.5 µl 10% Triton-X was added to each well to quench SDS. Finally, 2 µl 25 µM Primer i5 (Supplementary Table 2) and 25 µl NEBNext® High-Fidelity 2X PCR Master Mix (NEB) and samples were PCR amplified for 11 cycles (72°C 5 min, 98°C 30 s,[98°C 10 s, 63°C 30 s, 72°C 60 s] x 11, 72°C ∞). Following PCR, all wells were combined (around 15.5 mL) and mixed with 80 ml PB including pH-indicator (1:2500, Qiagen) and 4 ml Na-Acetate (3 M, pH = 5.2). Purification was carried out on 4 columns following the MinElute® PCR Purification Kit manual (Qiagen). DNA was eluted with 15 µl EB and eluate from all four columns was combined in a LoBind Tube (Eppendorf). For Ampure XP Bead (Beckmann Coulter) cleanup 170 µl EB buffer and 110 µl Ampure XP Beads (0.55x) were added to 30 µl eluate. After incubation at room temperature for 5 min and magnetic separation supernatant was transferred to a new tube and another 190 µl Ampure XP Beads (1.5x) were added. After incubation beads were washed twice on the magnet using 500 µl 80 % EtOH. After drying the beads for 7 min at room temperature library was eluted with 20 µl EB (Qiagen). Libraries were quantified using Qubit fluoromoeter (Life technologies) and nucleosomal pattern was verified using Tapestation (High Sensitivity D1000, Agilent). 25 pM library was loaded per lane of a HiSeq2500

sequencer (Illumina) using custom sequencing primers (Supplementary Table 2)[49] and following read lengths: 50 + 43 + 37 + 50 (Read1 + Index1 + Index2 + Read2). The first 8 bp of Index1 correspond to the p7 barcode and the last 8 bp to the i7 barcode. The first 8 bp of Index2 correspond to the i5 barcode and the last 8 bp to the p5 barcode. Since Index1 and 2 each contain 2 barcodes separated by a common linker sequence, we generated a spike-in library using different transposon and PCR primer sequences to balance the bases within each detection cycle (Supplementary Table 2). For the human-mouse mixture experiment, E15.5 forebrain and GM12878 nuclei were mixed in a 1:1 ratio prior to tagmentation. Samples were processed as above with the exceptions that just 96 wells were used after nuclei sorting and PCR amplification was performed for 13 cycles. The final library was loaded at 15 pM and sequenced using a MiSeq (Illumina) with following read lengths: PE 44 + 43 + 37 +44 (Read1 + Index1 + Index2 + Read2).

Cell culture

GM12878 (Coriell Institute for Medical Research) cells were cultured in RPMI1640 medium (Thermo Fisher Scientific) containing 2 mM L-glutamine (Thermo Fisher Scientific), 15% foetal bovine serum (Gemini Bioproducts) and 1 % Penicillin-Streptomicin (Thermo Fisher Scientific) in T25 Flasks (Corning) at 37°C under 5% carbon dioxide. For the snATAC-seq mixture experiment, cells were harvested by centrifugation, washed with PBS (Thermo Fisher Scientific) and resuspended in NPB (5 % BSA (Sigma), 0.2 % IGEPAL-CA630 (Sigma), cOmplete (Roche), 1 mM DTT in PBS). Samples were incubated 5 min at 4 °C and finally nuclei were pelleted by centrifugation (500g, 5min, 4 °C). Nuclei pellet was resuspended in 500 µl of 1.1x DMF buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF) and nuclei were counted using a hemocytometer.

NeuN negative sorting

10 mg adult forebrain tissue (P56) were resuspend in 500 µl lysis buffer (0.5% BSA, 0.1% Triton-X, cOmplete (Roche), 1 mM DTT in PBS) and incubated for 10 min at 4°C. After spinning down (5 min, 500 x g) sample was resuspended in 500 µl staining buffer (0.5%

BSA in PBS). Nuclei suspension was incubated with anti-NeuN antibody (1:5000, MAB377, Lot 2806074, EMD Millipore) for 30 min at 4°C. After centrifugation nuclei were resuspend in 500 µl staining buffer (0.5% BSA in PBS) containing anti-mouse Alexa488-antibody (1:1000, A11001, Lot 1696425, Thermo Fisher Scientific). After incubate for 30 min at 4°C, nuclei were pelleted (5 min 500 x g) and resupended in 700 ul sort buffer (1% BSA, 1mM EDTA in PBS). After filtration into a FACS tube 5 ul DRAQ7 (Cell Signalling Technologies) was added and NeuN-negative nuclei were sorted using a SH800 sorter (Sony) into 5% BSA (Sigma) in PBS.

ATAC-seq

ATAC-seq was performed on 20,000 sorted nuclei as described previously with minor modifications[50]. After adding IGEPAL-CA630 (Sigma) in a final concentration of 0.1 % nuclei were pelleted for 15 min at 1000 x g. Pellet was resupended in 19 µl 1.1x DMF buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF). After addition of 1 µl Tn5 transposomes (0.5 µM) tagmentation was performed at 37°C for 60 min with shaking (500 rpm). Next, samples were purified using MinElute columns (Qiagen), PCR-amplified for 8-10 cycles with NEBNext® High-Fidelity 2X PCR Master Mix (NEB, 72°C 5 min, 98°C 30 s,[98°C 10 s, 63°C 30 s, 72°C 60 s] x cycles, 72°C ∞). Amplified libraries were purified using MinElute columns (Qiagen) and Ampure XP Bead (Beckmann Coulter). Sequencing was carried out on a HiSeq2500 or 4000 (50 bp PE, Illumina).

DATA ANALYSIS

Single nuclei ATAC-seq data processing pipeline:

1. Alignment:

   Paired-end sequencing reads were aligned to mm10 reference genome using Bowtie2 in paired-end mode with following parameters "bowtie2 -p 5 -t -X2000 --no-mixed --no-discordant"[51].

2. Alignment filtration:

   Non-uniquely mapped (MAPQ < 30) and improperly paired (flag = 1804) alignments were filtered.

3. Barcode error correction:

   Each full barcode consists of four separate 8 bp long indexes (i5, i7, p5, p7). Reads with barcode combinations containing more than 1 mismatch for any index were removed. Barcodes with ≤ 1 mismatch were changed to its closest barcode.

4. Reads separation:

   Reads were separated into individual cells based on the barcode combination (Extended Data Table 1, Supplementary Table 2).

5. Mark and remove PCR duplicates:

   For individual cells, we sorted reads based on the genomic coordinates using "samtools sort"[52], then marked and removed PCR duplicates using Picard tools (MarkDuplicates, https://broadinstitute.github.io/picard/).

6. Mitochondrial reads removal:

   Reads mapped to the mitochondrial genome were filtered.

7. Adjusting position of Tn5 insertion:

   All reads aligning to the + strand were offset by +4 bp, and all reads aligning to the - strand were offset -5 bp[53].

8. Quality assessment of each single cell:

   Calculate coverage of constitutively accessible promoters (promoters that are accessible across all tissues/cell line from ENCODE DHS), number of reads and signal-over-noise ratio estimated by "reads in peaks" ratio for each cell.

9. Cell selection:

   We only kept cells that pass our threshold (1) coverage of constitutively accessible promoter > 10%; 2) number of reads > 1,000; 3) reads in peak ratio greater than estimation from corresponding bulk ATAC-seq level (Zhao et al. manuscript in preparation).

10. Replicates separation:

    Selected cells were separated into two replicates based on the predefined barcode combination.

Single nuclei ATAC-seq cluster analysis:

Cluster analysis partitions cells into groups such that cells from the same group have higher similarity than cells from different groups. Here, we developed a pipeline to obtain cell clusters. We first generated a catalogue of accessible chromatin regions using bulk ATAC-seq data (Zhao et al. manuscript in preparation) and created a binary accessible matrix. Chromatin sites were 1 for a given cell if there was a read detected within the peak region. Next, we calculated pair-wise Jaccard index between every two cells on the basis of overlapping open chromatin regions. Next, we applied a non-linear dimensionality reduction method (t-SNE) to map the high-dimensional structure to a 3-D space[14]. This transforms high-dimensional structures to dense data clouds in a low-dimensional space, allowing partitioning of cells using a density-based clustering method[15]. We then identified the optimal number of cell clusters using the Dunn index[54]. Finally, we compared our cluster results to those of “shuffled” to further verify our cluster result is not driven by library complexity or other confounding factors.

1. Determining accessible chromatin sites in single cells

   To catalogue accessible chromatin sites in individual cells, we first created a reference map of open chromatin sites determined by bulk ATAC-seq (Zhao et al. manuscript in preparation). The chromatin accessibility maps from different time points (from E11.5 to P56) were merged into a single reference file using BEDtools[55]. For clustering of

single cells, we have tested clustering performance using accessible promoters (2kb upstream of TSS) and distal elements, respectively, and found that clusters by distal elements outperformed promoters with lower Kullback-Leibler divergence (Extended Fig. 5). Therefore, we decided to only focus on distal genomic elements as features to perform clustering. Reads in individual cells overlapping with accessible sites were identified. We generated an accessible matrix of the reads counts overlapping each individual accessible sites (columns) in each cell (row).

2. Binary Accessible Matrix

   We next converted the chromatin accessibility matrix to a binary matrix $M_{N\times D}$ in which M_{ij} is 1 if any read in cell i mapped to region j.

3. Jaccard Index Matrix

   Jaccard index matrix $J_{N\times N}$ were calculated between every two cells in which J_{ij} measures the commonly shared open chromatin regions between cell C_i and C_j as following:

$$J_{ij} = \frac{|M_i \cap M_j|}{|M_i \cup M_j|}$$

   Diagonal elements of $J_{N\times N}$ are set to be 0 as required by t-SNE analysis.

4. Dimensionality reduction using t-SNE

   Using Jaccard index matrix $J_{N\times N}$ as input, we next applied t-SNE to map the N-dimensional data to a 3-D space[14]. Since t-SNE has a non-convex objective function, it is possible that different runs yield different solutions[14]. Thus, we ran t-SNE several times with different initiations and used the result with the lowest Kullback-Leibler divergence and best visualization. In a previous study sequencing depth was a confounding factor and highly correlated with the first principle component of PCA analysis (Pearson correlation >0.95)[5]. However, we did not observe correlation between sequencing depth and any of the t-SNE dimension. We expected that the

coherent structure of the open chromatin landscape of cells with high similarity would rely on a continuous and smooth 3-D structure and cells for different groups would locate to distinct parts of the plot. We used t-SNE to transform the high-dimensional structures to dense data clouds in the 3-D space[14]. Finally, we applied a density-based clustering method to identify different cell populations within the embedded 3-D space[15].

5. Density-based clustering

We applied a density-based clustering method to partition cells into groups in the embedded 3-D space[15]. The method identifies cluster centres that are characterized by two properties: 1) high local density ρ_i and 2) large distance δ_i from points of higher density, which are centers of the clusters[15]. Any cells that showed values above defined thresholds (ρ_0, δ_0) were considered as centers of cluster. Next, the rest of cells were assigned to the center as described here[15]. Clearly, different thresholds (ρ_0, δ_0) will generate different number of clusters. To find the optimal number of clusters, we adopted the method developed by Habib et al to evaluate the quality of different cluster results[54].

6. Number of clusters

In detail, Habib's method applied the Dunn index (Algorithm 2) to quantify the quality of cluster result as following[48]:

$$DB = \frac{min_{1\le i<j\le n}\Delta(C_i, C_j)}{max_{1\le k\le n}\Delta(C_k)}$$

in which $\Delta(C_i, C_j)$ represents the inter-cluster distance between cluster C_i and C_j, $\Delta(C_k)$ represents the intra-cluster distance of cluster C_k. We used the "MaxStep" distance (Algorithm 1) also developed by Habib et al to calculate the distance for Dunn index[54]. Finally, we iterated all possible (ρ_0, δ_0) combinations that yield different clusters and

calculated its Dunn index. The clustering result with the highest Dunn index was chosen as final cluster (Algorithm 3).

Algorithm 1: calculate max step distance (MaxStep) (adopted from Habib et al.)

Input: Pairwise Euclidean distance D in embedded 3-D space
Output: the pairwise MaxStep distance D'
$D' = D$
Let n be the number of data points
for k from 1 to n do
    for i from 1 to $n-1$ do
        for j from $i+1$ to n do
            $D'(i,j) = \min\left(D'(i,j), \max\left(D'(i,k), D'(k,j)\right)\right)$
        end
    end
end
return D'

Algorithm 2: Calculation of the Dunn index (Dunn) (adopted from Habib et al.)

Input: Pairwise Euclidean distance in embedded 3-D space (D), cluster assignment (C)
Output: Dunn index (θ)
C_uniq = unique(C)
$n = |\, C\_uniq \,|$
Let Δ_{in} be an empty array of length n
Let $\Delta_{between}$ be an empty matrix of size $n \times n$
for i from 1 to n do
    Let ii be the index of data whose cluster id is $C\_uniq(i)$
    $\Delta_{in}(i) = max(MaxStep(D(ii,ii)))$
end
for i from 1 to $n-1$ do
    for j from $i+1$ to n do
        Let ii be the index of data whose cluster id is $C\_uniq(i)$ or $C\_uniq(j)$
        $\Delta_{between}(i,j) = \max\left(MaxStep(D(ii,ii))\right)$
    end
end
$\theta = \min(\Delta_{between}) / \max(\Delta_{in})$
return θ

Algorithm 3: Cluster assignment

Input: local density (ρ) and local distance (δ) for every cell; Pairwise Euclidean distance in embedded 3-D space (D).
Output: cluster assignment (C)
Let n be the total number of cells
Let C_{best} be an empty array of length n
Let $Dunn_{best}$=$-INF$
for ρ_0 from 0 to max (ρ) do
    for δ_0 from 0 to max (δ) do
        choose cells whose $\rho(i)$ and $\delta(i)$ is greater than ρ_0, δ_0 as $Centers$
        C = cluster_assignment(D, $Centers$)*
        if $Dunn(D, C) > Dunn_{best}$ do
            $Dunn_{best}$=$Dunn(D, C)$
            $C_{best} = C$
        end
    end
end
return C_{best}

*cluster_assignment(D, $Centers$) is as described here _[16]

7. "Shuffled" cells

Due to the limited genome coverage of each single cell, cells may cluster based on their sequencing depth rather than 'true' co-variation[5]. To verify that our cluster results are not driven by such artefacts, we compared our results to a simulated data set. For this data set in which binary accessible sites within each cell were randomly shuffled across all accessible sites. In other words, we shuffled the data and removed the biological clustering, but maintained the distribution of sequencing depth across cells. "Shuffled" cells were uniformly distributed as a "ball" in the embedded 3-D space without clear partition of cells. However, we did observe that one of the directions becomes correlated with sequencing depth (Pearson correlation 0.55 for t-SNE3) and there is a small portion of cells that tend to form a cluster but did not pass the cut-off (ρ_0, δ_0) used for the P56 forebrain data set.

Identification of cluster-specific features

We next developed a computational method which combines stability selection with

LASSO[56] to identify genomic elements (features) that potentially distinguish cells belonging to different clusters. LASSO regression enables sparse feature selections through the use of L1 penalty. However, LASSO regression often does not result in a robust set of selected features and is sensitive to data perturbation. This is especially true when features are correlated as the case here. To overcome these limitations, we adopted stable lasso to robustly identify features that distinguish every two cell clusters (Algorithm 4)[56]. Finally, we combined all identified features that distinguish different cell types.

Algorithm 4: Cluster specific features selection

Input: $X \epsilon R^{(n,p)}(binary\ matrix)$, $Y \epsilon \{0,1\}^n(cluster\ label)$, $\alpha(subsampling\ rate)$, $\beta(perturbation\ rate)$, $T(iteration)$
Output: importance score for each feature
for $t = 1$ to T do
    Randomly perturb the data:
        Draw a subset (X_t, Y_t) of α of (X, Y)
        Draw a vector $w \sim U([\beta, 1]^p)$
        Re-weight the features: $X'_t = X_t \cdot w$
    Compute the LASSO path of length $\alpha \cdot n$
    Keep the selection matrix $S_t \in {}^{3p,\alpha \cdot n}$ where ix

$S_t(i,j) = \{1, \quad if\ the\ ith\ feature\ selected\ at\ jth\ step\ 0, \quad otherwise$
end for
Compute the feature importance ix

$$f_i = \frac{1}{n\alpha T} \sum_{j=1}^{n/2} \sum_{t=1}^{T} S_t(i,j)$$

Bulk ATAC-seq

Paired-end sequencing reads were aligned to the mm10 reference genome using Bowtie2 in paired-end mode with following parameters “bowtie2 -p 5 -t -X2000 --no-mixed --no-discordant[51] and PCR duplicates were removed using SAMtools[52] . Next, mitochondrial reads were removed and the position of alignments adjusted[53]. For visualization the *bamCoverage* utility from deepTools2 was used[57].

Hierarchical clustering of ATAC-seq profiles in adult forebrain

DeepTools2 was used for correlation analysis and hierarchical clustering of ATAC-seq profiles from cell clusters and sorted cell-types in the adult forebrain[57]. First, we computed read coverage for each data set against the merged list of genomic elements that separate two cell clusters in the adult forebrain using the *multiBamSummary* utility. Next we used *plotCorrelation* to generate hierarchical clustering using Spearman correlation coefficient between two clusters[57].

Accessibility analysis and clustering of genomic elements

To cluster genomic elements based on their accessibility profile we used these promoter distal elements that were capable to distinguish two cell clusters. For each feature we extended the summits identified by MACS2[58] in both directions by 250 bp and generated a union set of elements using *mergeBED* functionality of BEDTools v2.17.0[55]. Next, we intersected cluster specific bam files with the peak list using the *coverageBED* functionality of BEDTools v2.17.0[55]. We discarded elements that had less than five reads on average. After adding a pseudocount of one we calculated cluster-specific RPM (reads per million sequenced reads) values for each genomic element. We divided the RPM value for a given cluster by the average value of all clusters (fold over mean) and finally log2 transformed the data. The generated matrix was used for k-means clustering of the elements using Ward's method. We performed this analysis for all adult clusters, the excitatory neuron clusters and the 12 developmental cell clusters, respectively. A list of elements for each analysis can be found in Supplementary Table 1.

Motif enrichment analysis

To identify potential regulators of chromatin accessibility we performed motif analysis using the AME utility of the MEME suite[59]. A P-Value cut-off of $< 10^{-5}$ was chosen for known motifs from the JASPAR database (JASPAR_CORE_2016_vertebrates.meme)[60]. For identification of *de novo* motifs HOMER tools was used with default settings[61].

Annotation of genomic elements

The GREAT algorithm was used to annotate distal genomic elements using following settings to define the regulatory region of a gene: Basal+extension (constitutive 1 kb upstream and 0.1 kb downstream, up to 500 kb max extension)[31]. Gene ontology categories “Molecular Function” and “Biological Processes” were used.

Analysis of dynamic chromatin accessibility within a cell cluster

First, the ATAC-seq reads were counted in all peaks for each stage, cell type and replicate. For each cell cluster, only stages with more than 250,000 reads overlapping ATAC-seq peaks and more than 50 nuclei were used for dynamic analysis. Peaks with greater than 1 read per million reads (RPM) in at least 2 samples were kept. We used edgeR[62] to assess the significance of difference between adjacent stages for cell clusters with at least 4 out of 7 stages passing filtering criteria. P-values were corrected using the Bonferroni method. Peaks with a Bonferroni p-value less than 0.05 were called dynamic peaks. The total number of dynamic peaks in each cell type are listed in (Extended Data Fig. 11c). For each cell type, the read counts in each peak were normalized into a unit vector (i.e values were divided by the square root of the sum of the squares of the values). K-means was used for clustering of cell clusters with more than 200 dynamic elements (K=3). Motif enrichment analysis was performed for each peak cluster using HOMER[61].

VISTA analysis

Genomic locations of 484 VISTA validated elements[44] were downloaded from https://enhancer.lbl.gov using the search term “forebrain”. Genomic locations were converted from mm9 to mm10 using the *liftOver* tool (minimum rematch ratio of 0.95)[63]. 91 of these were showed specific activity in the subpallium[45]. To identify developmental clusters that are enriched for subpallial enhancers we first calculated the ratio of elements per k-means cluster overlapping with the total forebrain enhancer list and the subpallial subset separately. Finally, we calculated the relative enrichment using the ratio of subpallial over the complete forebrain regions.

External data sets

Published ATAC-seq data of sorted excitatory neurons (GSM1541964, GSM1541965)[8], GABAergic neurons (GSM2333635, GSM2333636)[9], microglia (GSM2104286)[17] , neurons of the dentate gyrus (GSM2179990, GSM2179991)[27] and distinct cortical layers (Layer2/3: GSM2333632, GSM2333633; Layer 4: GSM2333644, GSM2333645; Layer 5: GSM2333641, GSM2333642, Layer 6, GSM2333638, GSM2333639)[9] were reprocessed. In addition, bulk ATAC-seq data for embryonic time points were analysed for comparison (https://www.encodeproject.org/search/?searchTerm=atac+forebrain, Zhao et al. manuscript in preparation)

Data availability

Raw and processed data have been deposited to NCBI Gene Expression Omnibus with the accession number GSE1000333.

REFERENCES

1 Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364, doi:10.1038/nature13992 (2014).
2 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).
3 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
4 Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nature genetics* **47**, 1393-1401, doi:10.1038/ng.3432 (2015).
5 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).
6 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).
7 Wu, J. *et al.* The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652-657, doi:10.1038/nature18606 (2016).
8 Mo, A. *et al.* Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron* **86**, 1369-1384, doi:10.1016/j.neuron.2015.05.018 (2015).
9 Gray, L. T. *et al.* Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *eLife* **6**, doi:10.7554/eLife.21883 (2017).

10 Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905, doi:10.1126/science.1237905 (2013).
11 Gilsbach, R. *et al.* Dynamic DNA methylation orchestrates cardiomyocyte development, maturation and disease. *Nature communications* **5**, 5288, doi:10.1038/ncomms6288 (2014).
12 Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics* **48**, 1193-1203, doi:10.1038/ng.3646 (2016).
13 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490, doi:10.1038/nature14590 (2015).
14 van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J Mach Learn Res* **9**, 2579-2605 (2008).
15 Rodriguez, A. & Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science* **344**, 1492-1496, doi:10.1126/science.1242072 (2014).
16 Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007-1012, doi:10.1126/science.1246426 (2014).
17 Matcovitch-Natan, O. *et al.* Microglia development follows a stepwise program to regulate brain homeostasis. *Science* **353**, aad8670, doi:10.1126/science.aad8670 (2016).
18 Huttner, H. B. *et al.* The age and genomic integrity of neurons after cortical stroke in humans. *Nature neuroscience* **17**, 801-803, doi:10.1038/nn.3706 (2014).
19 Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142, doi:10.1126/science.aaa1934 (2015).
20 La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580 e519, doi:10.1016/j.cell.2016.09.027 (2016).
21 Rousseau, A. *et al.* Expression of oligodendroglial and astrocytic lineage markers in diffuse gliomas: use of YKL-40, ApoE, ASCL1, and NKX2-2. *Journal of neuropathology and experimental neurology* **65**, 1149-1156, doi:10.1097/01.jnen.0000248543.90304.2b (2006).
22 Pernet, V., Joly, S., Christ, F., Dimou, L. & Schwab, M. E. Nogo-A and myelin-associated glycoprotein differently regulate oligodendrocyte maturation and myelin formation. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **28**, 7435-7444, doi:10.1523/JNEUROSCI.0727-08.2008 (2008).
23 Kierdorf, K. *et al.* Microglia emerge from erythromyeloid precursors via Pu.1- and Irf8-dependent pathways. *Nature neuroscience* **16**, 273-280, doi:10.1038/nn.3318 (2013).
24 Glasgow, S. M. *et al.* Mutual antagonism between Sox10 and NFIA regulates diversification of glial lineages and glioma subtypes. *Nature neuroscience* **17**, 1322-1329, doi:10.1038/nn.3790 (2014).
25 Nord, A. S., Pattabiraman, K., Visel, A. & Rubenstein, J. L. Genomic perspectives of transcriptional regulation in forebrain development. *Neuron* **85**, 27-47, doi:10.1016/j.neuron.2014.11.011 (2015).

26 Yuan, F. *et al.* Efficient generation of region-specific forebrain neurons from human pluripotent stem cells under highly defined condition. *Scientific reports* **5**, 18550, doi:10.1038/srep18550 (2015).
27 Su, Y. *et al.* Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nature neuroscience* **20**, 476-483, doi:10.1038/nn.4494 (2017).
28 Barbosa, A. C. *et al.* MEF2C, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 9391-9396, doi:10.1073/pnas.0802679105 (2008).
29 Martynoga, B., Drechsel, D. & Guillemot, F. Molecular control of neurogenesis: a view from the mammalian cerebral cortex. *Cold Spring Harbor perspectives in biology* **4**, doi:10.1101/cshperspect.a008359 (2012).
30 Pollen, A. A. *et al.* Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55-67, doi:10.1016/j.cell.2015.09.004 (2015).
31 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).
32 Subramanian, L. *et al.* Transcription factor Lhx2 is necessary and sufficient to suppress astrogliogenesis and promote neurogenesis in the developing hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E265-274, doi:10.1073/pnas.1101109108 (2011).
33 Hsu, L. C. *et al.* Lhx2 regulates the timing of beta-catenin-dependent cortical neurogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 12199-12204, doi:10.1073/pnas.1507145112 (2015).
34 Castro, D. S. *et al.* Proneural bHLH and Brn proteins coregulate a neurogenic program through cooperative binding to a conserved DNA motif. *Developmental cell* **11**, 831-844, doi:10.1016/j.devcel.2006.10.006 (2006).
35 Castro, D. S. *et al.* A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. *Genes & development* **25**, 930-945, doi:10.1101/gad.627811 (2011).
36 Long, J. E., Cobos, I., Potter, G. B. & Rubenstein, J. L. Dlx1&2 and Mash1 transcription factors control MGE and CGE patterning and differentiation through parallel and overlapping pathways. *Cerebral cortex* **19 Suppl 1**, i96-106, doi:10.1093/cercor/bhp045 (2009).
37 Heng, Y. H. *et al.* NFIX regulates neural progenitor cell differentiation during hippocampal morphogenesis. *Cerebral cortex* **24**, 261-279, doi:10.1093/cercor/bhs307 (2014).
38 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
39 Hori, K. *et al.* A nonclassical bHLH Rbpj transcription factor complex is required for specification of GABAergic neurons independent of Notch signaling. *Genes & development* **22**, 166-178, doi:10.1101/gad.1628008 (2008).
40 Tian, X., Kai, L., Hockberger, P. E., Wokosin, D. L. & Surmeier, D. J. MEF-2 regulates activity-dependent spine loss in striatopallidal medium spiny neurons. *Molecular and cellular neurosciences* **44**, 94-108, doi:10.1016/j.mcn.2010.01.012 (2010).

41 Onorati, M. *et al.* Molecular and functional definition of the developing human striatum. *Nature neuroscience* **17**, 1804-1815, doi:10.1038/nn.3860 (2014).
42 Ghisletti, S. *et al.* Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* **32**, 317-328, doi:10.1016/j.immuni.2010.02.008 (2010).
43 Choksi, S. P., Lauter, G., Swoboda, P. & Roy, S. Switching on cilia: transcriptional networks regulating ciliogenesis. *Development* **141**, 1427-1441, doi:10.1242/dev.074666 (2014).
44 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic acids research* **35**, D88-92, doi:10.1093/nar/gkl822 (2007).
45 Silberberg, S. N. *et al.* Subpallial Enhancer Transgenic Lines: a Data and Tool Resource to Study Transcriptional Regulation of GABAergic Cell Fate. *Neuron* **92**, 59-74, doi:10.1016/j.neuron.2016.09.027 (2016).
46 Visel, A. *et al.* A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895-908, doi:10.1016/j.cell.2012.12.041 (2013).
47 Sos, B. C. *et al.* Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome biology* **17**, 20, doi:10.1186/s13059-016-0882-7 (2016).
48 Wang, Q. *et al.* Tagmentation-based whole-genome bisulfite sequencing. *Nature protocols* **8**, 2022-2032, doi:10.1038/nprot.2013.118 (2013).
49 Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nature genetics* **46**, 1343-1349, doi:10.1038/ng.3119 (2014).
50 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. *Nature methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).
51 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).
52 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
53 Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11**, R119, doi:10.1186/gb-2010-11-12-r119 (2010).
54 Habib, N. *et al.* Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*, doi:10.1126/science.aad7038 (2016).
55 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).
56 Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**, 267-288 (1996).
57 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research* **44**, W160-165, doi:10.1093/nar/gkw257

(2016).

58 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

59 Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research* **34**, W369-373, doi:10.1093/nar/gkl198 (2006).

60 Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research* **44**, D110-115, doi:10.1093/nar/gkv1176 (2016).

61 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

62 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

63 Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic acids research* **45**, D626-D634, doi:10.1093/nar/gkw1134 (2017).