Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation

Sebastian Preissl [1,2,11], Rongxin Fang[1,3,11], Hui Huang[1,4], Yuan Zhao[1,3], Ramya Raviram[1], David U. Gorkin [1,2], Yanxiao Zhang [1], Brandon C. Sos[4,5], Veena Afzal[6], Diane E. Dickel [6], Samantha Kuan[1], Axel Visel [6,7,8], Len A. Pennacchio [6,7,9], Kun Zhang [5] and Bing Ren [1,2,10]*

[1]Ludwig Institute for Cancer Research, La Jolla, CA, USA. [2]Center for Epigenomics, Department of Cellular and Molecular Medicine, University of California, San Diego, School of Medicine, La Jolla, CA, USA. [3]Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA, USA. [4]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA, USA. [5]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. [6]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [7]US Department of Energy Joint Genome Institute, Walnut Creek, CA, USA. [8]School of Natural Sciences, University of California Merced, Merced, CA, USA. [9]Comparative Biochemistry Program, University of California Berkeley, Berkeley, CA, USA. [10]Institute of Genomic Medicine, Moores Cancer Center, University of California San Diego, School of Medicine, La Jolla, CA, USA. [11]These authors contributed equally: Sebastian Preissl and Rongxin Fang. *e-mail: biren@ucsd.edu

a. Nuclei permeabilization (b) → Tagmentation at 37°C (Buffer (c); Tn5 amount (d,h); time (e); quenching (d)) → Release DNA from Tn5 after FANS (d) → PCR, Pooling, Sequencing

d.

Tagmentation		Fragment release			qPCR
Tn5 amount	EDTA[1]	EDTA/ MgCl2[2]	SDS[3]	SDS/ Triton-X100[4]	Relative fragments
1x	+		+	+	20.2
1x	+	+			14.3
1x	+		+		14.0
2x	+	+			12.9
2x	+		+		11.3
1x		+			10.0
2x	+		+	+	9.2
1x			+	+	6.9
1x			+		4.4
2x	+				1.3
1x	+				0.8
1x					1.0

[1] Tn5 reaction was quenched with 20 mM EDTA prior to FANS

[2] Fragments were released by adding EDTA (50 mM) and incubation for 30 min at 50ºC. Prior to PCR EDTA was quenched with 50 mM MgCl$_2$.

[3] Fragments were released by adding SDS (0.02 %) and incubation for 7 min at 55ºC.

[4] Fragments were released by adding SDS (0.02 %) and incubation for 7 min at 55ºC. Prior to PCR SDS was quenched with 1 % Triton-X100.

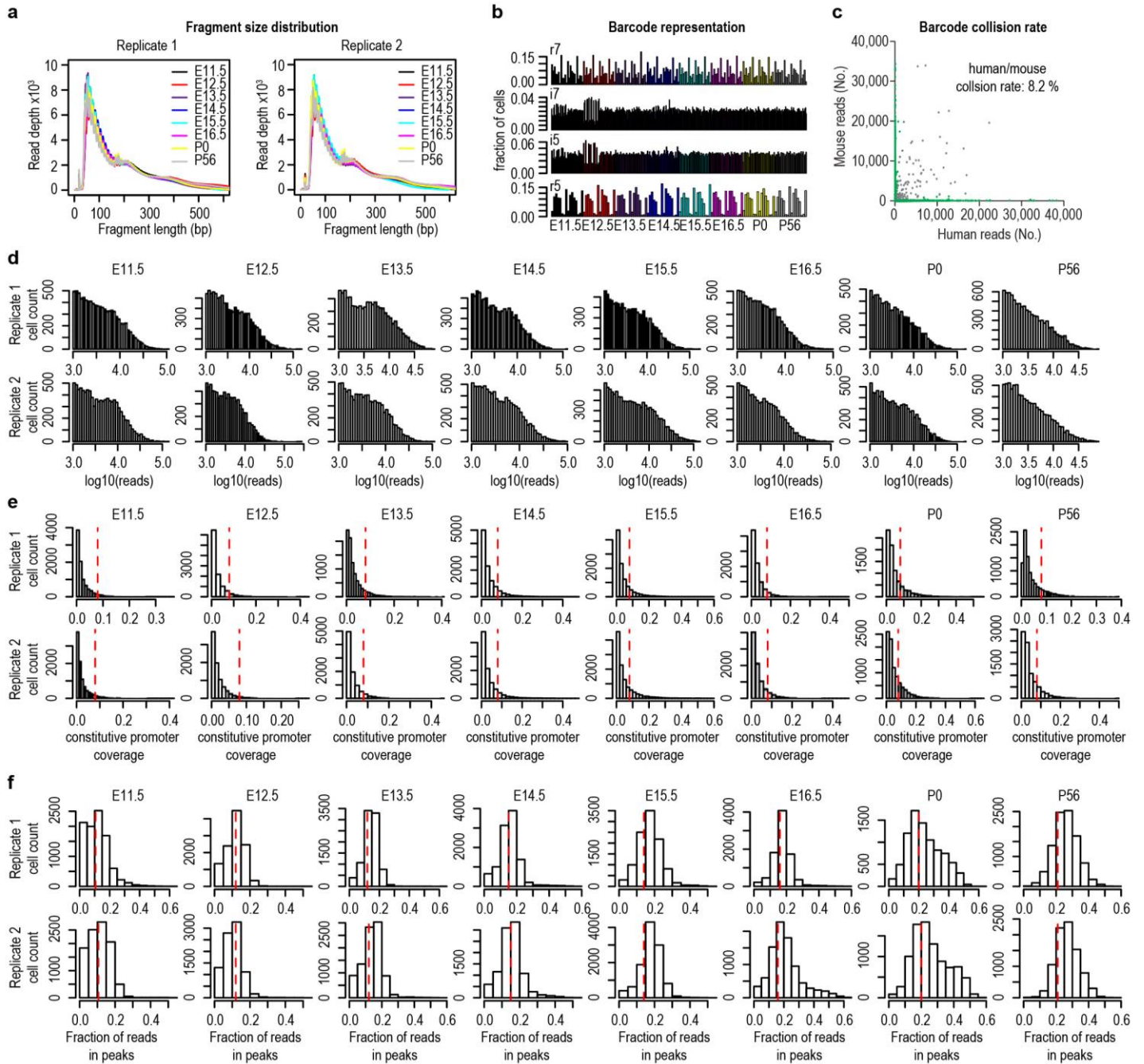**Supplementary Figure 1**

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 (n = 1 experiment). **c.** Tagmentation was facilitated by high salt concentrations in reaction buffer (n = 1 experiment; Wang, Q. *et al. Nature protocols, 2013,* doi:10.1038/nprot.2013.118: Sos, B. C. *et al. Genome biology*, 2016, doi:10.1186/s13059-016-0882-7). **d.** Maximum amount of fragments per nucleus could be recovered when quenching Tn5 by EDTA prior to FANS and denaturation of Tn5 after FANS 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 (n = 1 experiment). **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 (n = 1 experiment). **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 increased the number of fragments per nucleus by 3 fold. Further increase resulted only in minor improvements (n = 1 experiment). **i.** Dot blot illustrating the amount of library from 25 nuclei per well. 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). **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 (n = 1 experiment; snATAC libraries for forebrain samples showed comparable nucleosomal patterns: n = 16 experiments).
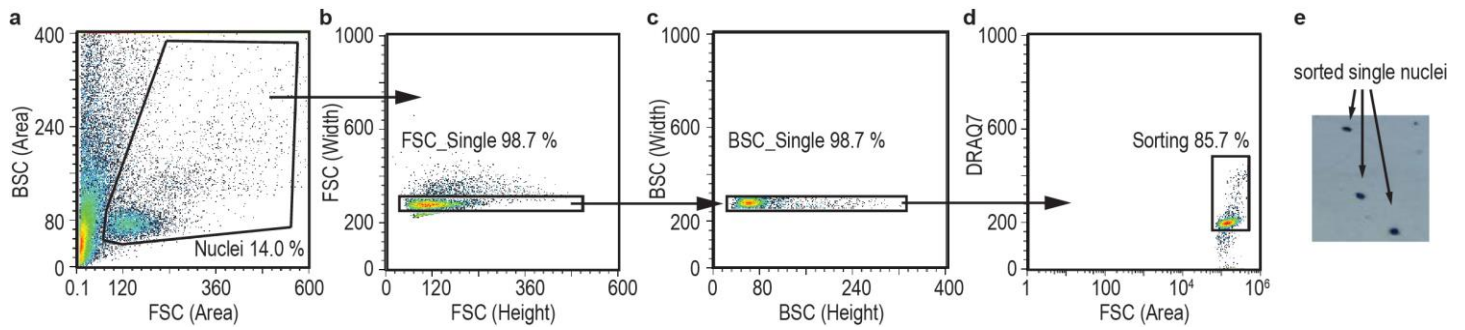
Supplementary Figure 2

Isolation 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 FANS was done with Trypan Blue staining under a microscope.

Supplementary Figure 3

Overview of snATAC-seq sequencing data and quality filtering for single nuclei.

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 nuclei sharing the same nuclei barcode, single nuclei 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 nucleus. The red line indicates fraction of reads in peak regions in corresponding bulk ATAC-seq data sets from the same biological sample. Nuclei with lower reads in peak
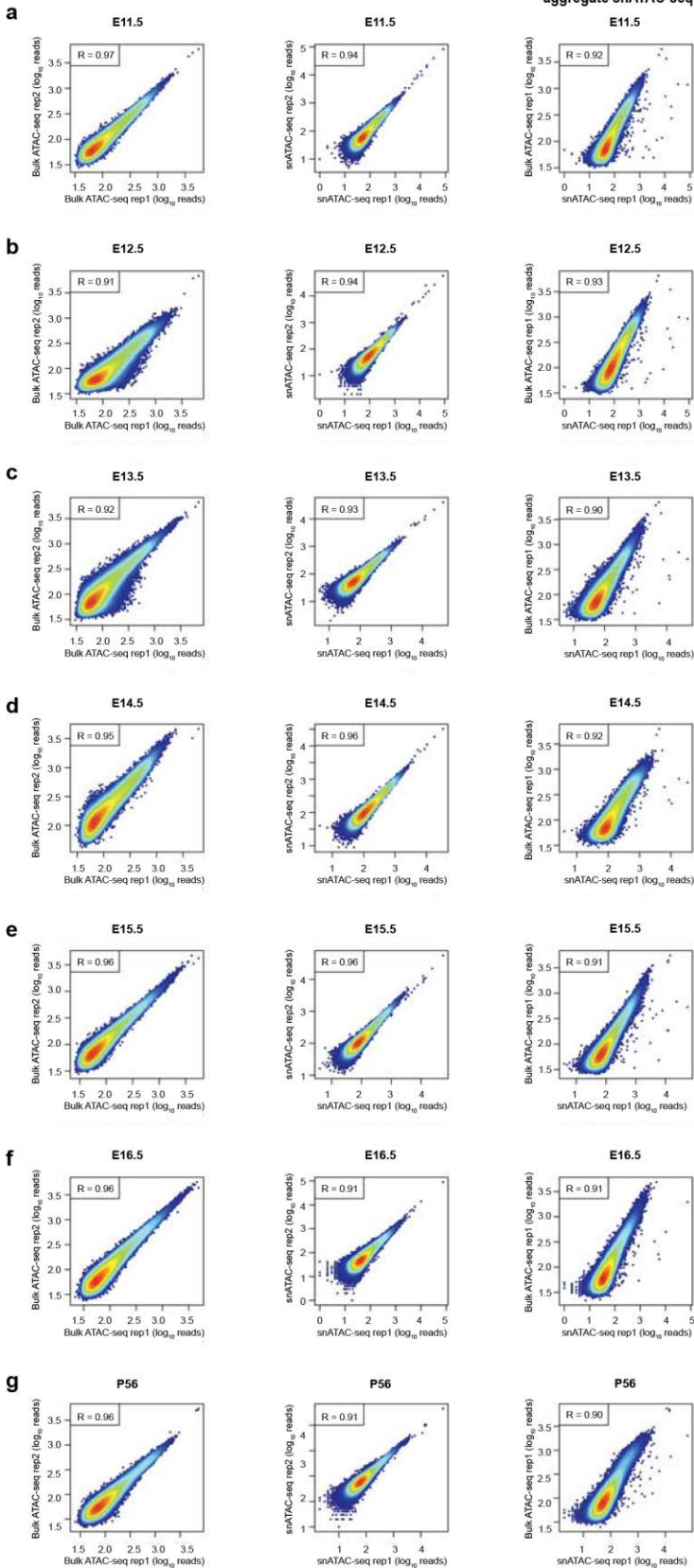
ratios coverage than the bulk ATAC-seq data set were discarded from downstream analysis. For bulk ATAC-seq data generated by the ENCODE consortium were processed

(https://www.encodeproject.org/search/?type=Experiment&lab.title=Bing+Ren%2C+UCSD&assay_title=ATAC-seq&organ_slims=brain).

Supplementary Figure 4

**snATAC-seq data sets are robust and reproducible.**

Pearson correlation of chromatin accessibility profiles from two independent experiments derived from bulk ATAC-seq (left column) and from aggregate snATAC-seq after aggregating single nuclei profiles (middle column) is shown in each plot. In the right column the correlation between bulk ATAC-seq and aggregate snATAC-seq are displayed for the experiment on the first set of forebrain tissues. 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, and **h.** P56. For bulk ATAC-seq data generated by the ENCODE consortium were processed (https://www.encodeproject.org/search/?type=Experiment&lab.title=Bing+Ren%2C+UCSD&assay_title=ATAC-seq&organ_slims=brain).
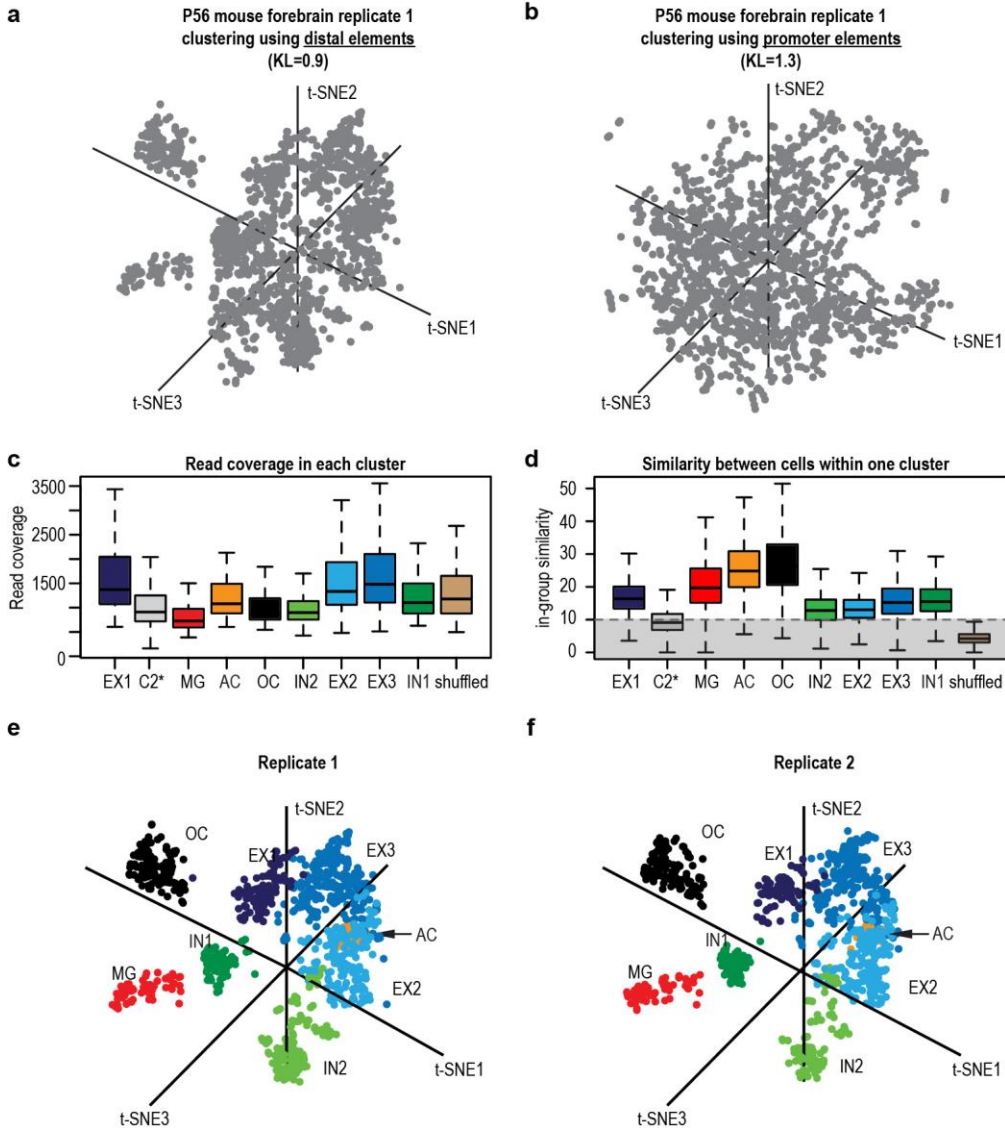
Supplementary Figure 5

Clustering strategies, quality control of clusters and clustering result for individual experiments 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 (sample size for cluster is EX1: 190, C2: 946, MG: 126, AC: 120, OC: 252, IN2: 320, EX2: 366, EX3: 519, IN1: 195, shuffled: 199; 25% quantile is EX1: 1076, C2: 665, MG: 595, AC: 884.25, OC: 755, IN2: 754, EX2: 106, EX3: 1104, IN1: 881, shuffled: 880; median value is EX1: 1372, C2: 855, MG: 726, AC: 1079, OC: 871, IN2: 899, EX2: 1334, EX3: 1482, IN1: 1102, shuffled: 1178; 75% quantile is EX1: 2045, C2: 1196, MG: 972, AC: 1489, OC: 1188, IN2: 1134, EX2: 1929, EX3: 2102, IN1: 1496, shuffled: 1652) **d** Box plot of similarity analysis between 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 (sample size is EX1: 190, C2:946, MG:126, AC:120, OC: 252, IN2: 320, EX2: 366, EX3: 519, IN1: 195, shuffled: 199; 25% quantile is EX1: 13.34, C2: 6.84, MG: 15.15, AC: 19.89, OC: 20.60, IN2: 9.88, EX2: 10.53, EX3: 11.81, IN1: 12.58, shuffled: 3.02; median is EX1: 16.34, C2: 9.12, MG: 19.68, AC: 24.835, OC: 26.23, IN2: 12.77, EX2: 13.00, EX3: 15.23, IN1: 15.50, shuffled: 4.20; 75% quantile is EX1: 20.07, C2: 11.74, MG: 25.58, AC: 30.860, OC: 32.95, IN2: 16.11, EX2: 16.02, EX3: 19.46, IN1: 19.25, shuffled: 5.56) **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.
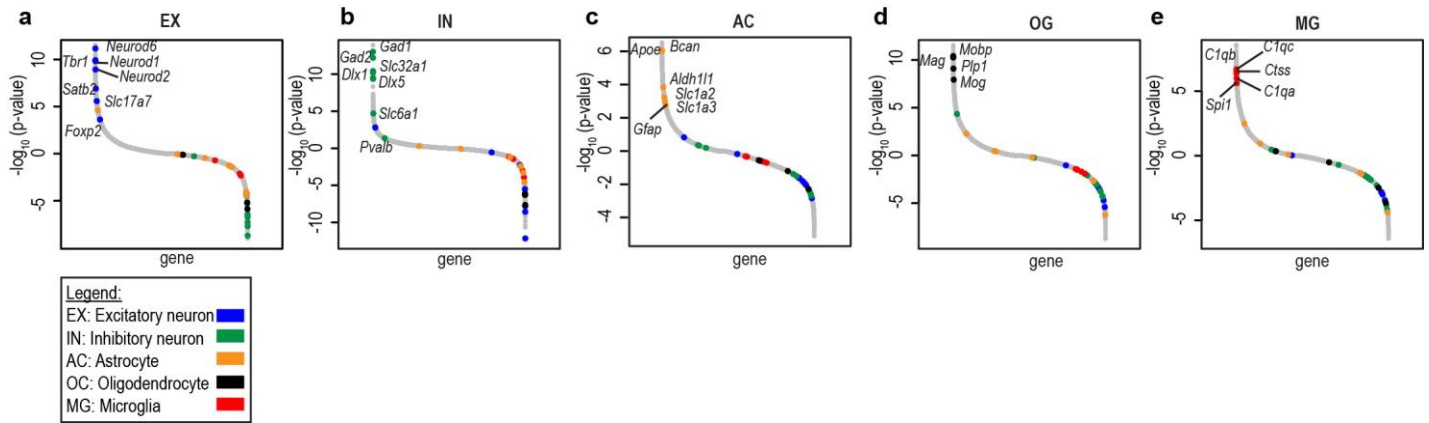
Supplementary Figure 6

Ranking of gene loci (TSS ± 10 kb) 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. For each cell type, data from two experiments (n = 2) were used to carry out the negative binomial test.
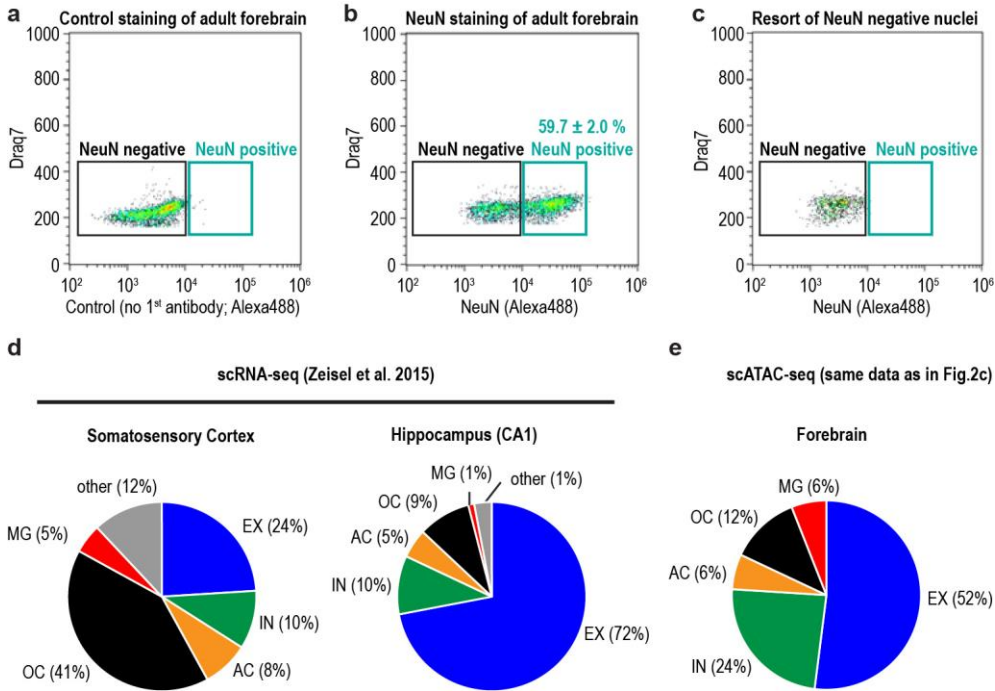
Supplementary Figure 7

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[22] (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.
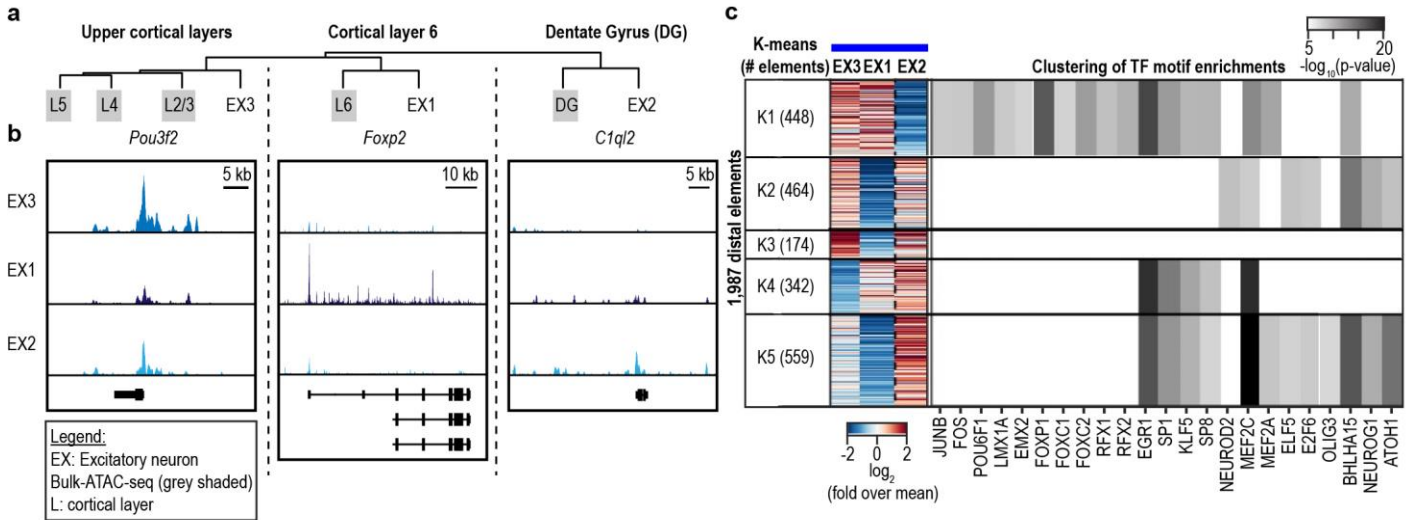
Supplementary Figure 8

Subclassification 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. Statistical test for motif enrichment: One-tailed Fisher's Exact test; displayed p-values are Bonferroni corrected for multiple testing[59].
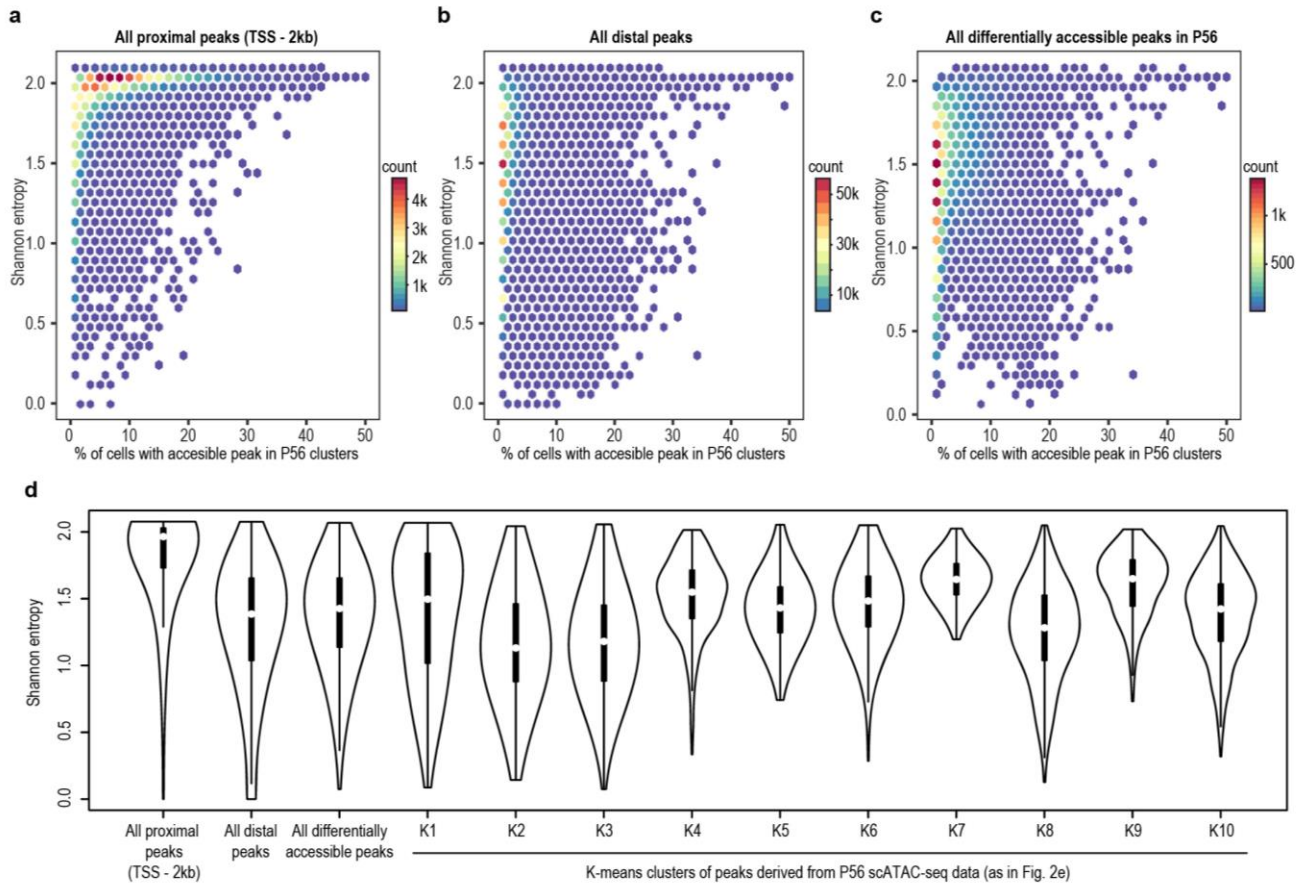
Supplementary Figure 9

Cell-type-specificity and coverage of the *cis* 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 nuclei 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, all genomic elements that separate two clusters as well as subsets identified from k-means clustering of genomic elements depending on chromatin accessibility in adult forebrain are displayed (related to Fig. 2e). (all proximal peaks n = 14,262 (minimum/median/maximum; 0/1.96/2.08), all distal peaks n = 140,102 (0/1.38/2.08), all differentially accessible peaks n = 4,980 (0.07/1.4/2.06), K1 n = 529 (0.08/1.49/2.06), K2 n = 586 (0.14/1.13/2.04), K3 n = 737 (0.07/1.18/2.05), K4 n = 270 (0.33/1.55/2.01), K5 n = 601 (0.74/1.43/2.05), K6 n = 513 (0.28/1.48/2.05), K7 n = 538 (1.19/1.64/2.02), K8 n = 490 (0.13/1.28/2.05), K9 n = 282 (0.73/1.65/2.02), K10 n = 434 (0.32/1.42/2.04). TSS: transcriptional start site.
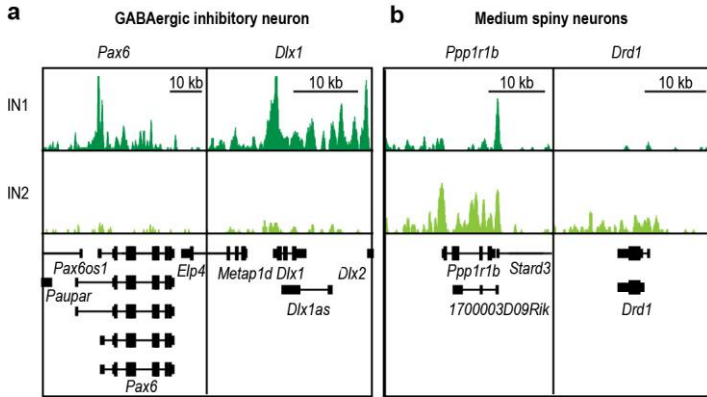
Supplementary Figure 10

Distinct chromatin accessibility profiles of two GABAergic neuron clusters.

IN2 is depleted for chromatin accessibility at the genes *Pax6* and *Dlx1* (**a**), but enriched for marker genes of medium spiny neurons as compared to IN1 cluster (**b**).
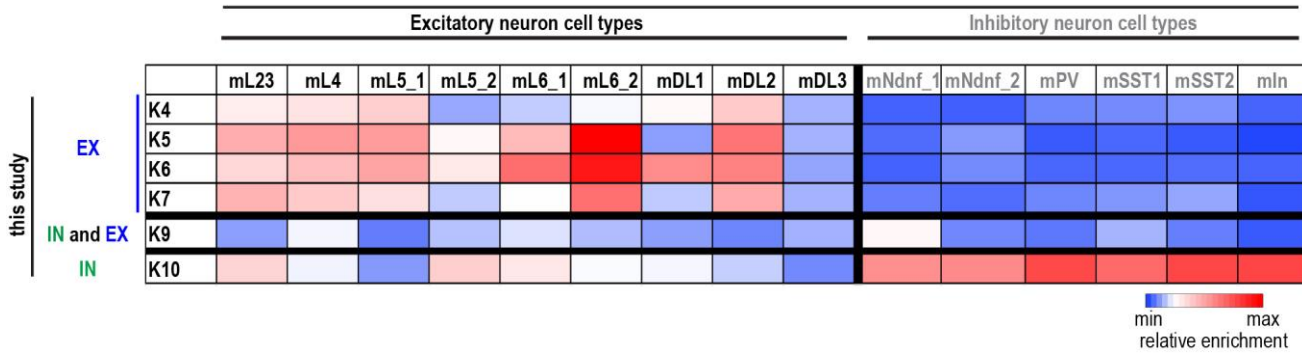
Supplementary Figure 11

Comparison of chromatin accessibility and differentially methylated regions in neuronal subtypes.

Displayed is the fraction of cell-type specific differentially methylated[29] that overlapped with genomic elements accessible in excitatory (EX) and inhibitory neurons (IN). This analysis illustrates that cis regulatory elements specific for inhibitory neurons and excitatory neurons, respectively, could be identified by both methods. Clusters (K) from this study are the same as in Fig.2e (m: mouse; L: layer; DL: deep layer).
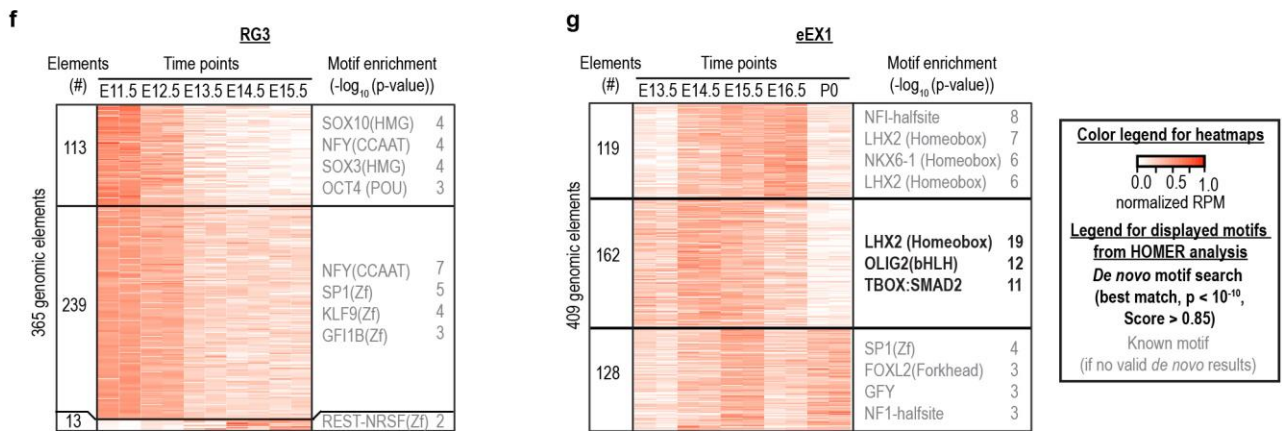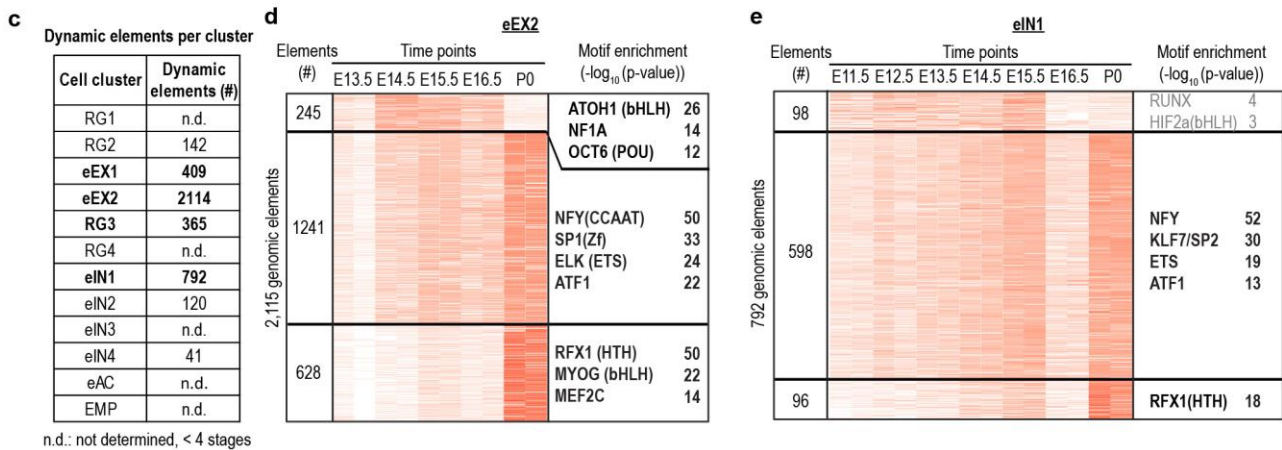
a Reads in peaks per cell cluster at each time point

Time point / Cell cluster	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

Time point / Cell cluster	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

d eEX2 — 2,115 genomic elements — Time points: E13.5 E14.5 E15.5 E16.5 P0 — Motif enrichment (-log₁₀ (p-value))

Elements (#)	Motif enrichment (-log$_{10}$(p-value))	
245	ATOH1 (bHLH)	26
	NF1A	14
	OCT6 (POU)	12
1241	NFY(CCAAT)	50
	SP1(Zf)	33
	ELK (ETS)	24
	ATF1	22
628	RFX1 (HTH)	50
	MYOG (bHLH)	22
	MEF2C	14

e eIN1 — 792 genomic elements — Time points: E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 P0 — Motif enrichment (-log$_{10}$(p-value))

Elements (#)	Motif enrichment	
98	RUNX	4
	HIF2a(bHLH)	3
598	NFY	52
	KLF7/SP2	30
	ETS	19
	ATF1	13
96	RFX1(HTH)	18

f RG3 — 365 genomic elements — Time points: E11.5 E12.5 E13.5 E14.5 E15.5 — Motif enrichment (-log$_{10}$(p-value))

Elements (#)	Motif enrichment	
113	SOX10(HMG)	4
	NFY(CCAAT)	4
	SOX3(HMG)	4
	OCT4 (POU)	3
239	NFY(CCAAT)	7
	SP1(Zf)	5
	KLF9(Zf)	4
	GFI1B(Zf)	3
13	REST-NRSF(Zf)	2

g eEX1 — 409 genomic elements — Time points: E13.5 E14.5 E15.5 E16.5 P0 — Motif enrichment (-log$_{10}$(p-value))

Elements (#)	Motif enrichment	
119	NFI-halfsite	8
	LHX2 (Homeobox)	7
	NKX6-1 (Homeobox)	6
	LHX2 (Homeobox)	6
162	LHX2 (Homeobox)	19
	OLIG2(bHLH)	12
	TBOX:SMAD2	11
128	SP1(Zf)	4
	FOXL2(Forkhead)	3
	GFY	3
	NF1-halfsite	3

Color legend for heatmaps: 0.0 0.5 1.0 normalized RPM

Legend for displayed motifs from HOMER analysis
De novo motif search (best match, p < 10⁻¹⁰, Score > 0.85)
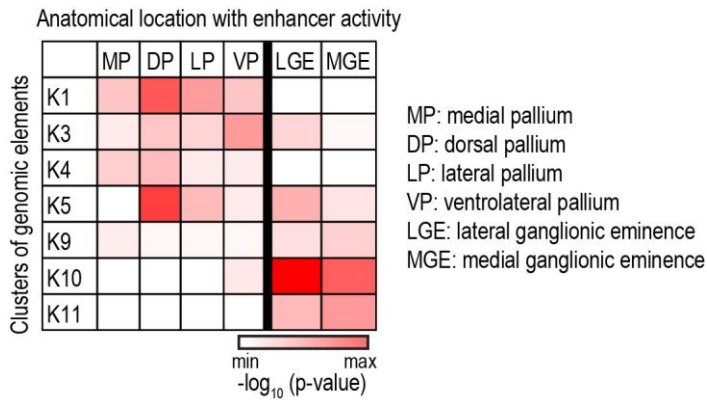Known motif (if no valid *de novo* results)

Supplementary Figure 12

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. Statistical test for motif enrichment: hypergeometric test. P-values

were not corrected for multiple testing[50]. (e: embryonic; RG: Radial glia; EX: Excitatory neuron; IN: Inhibitory neuron; EMP: Erythromyeloid progenitor cell; AC: Astrocyte).

Anatomical location with enhancer activity

MP: medial pallium
DP: dorsal pallium
LP: lateral pallium
VP: ventrolateral pallium
LGE: lateral ganglionic eminence
MGE: medial ganglionic eminence

Supplementary Figure 13

Distal genomic element clusters are associated with distinct anatomical locations in the developing forebrain.

Displayed is the enrichment of clusters of open chromatin for enhancers that are active in distinct regions of the developing forebrain (n = 95)[47]. As expected elements mainly associated with radial glia and excitatory neuron cell-types (Fig.2e, K1,3,4) were enriched for pallial subregions, whereas inhibitory neuron associated elements (Fig.2e, K9-11) were enriched in LGE and MGE regions. Clusters with less than 5 overlapping elements were excluded from the analysis. Binomial testing was used for statistical analysis. The p-values were not corrected. Anatomically annotated enhancers: n = 146[47]; open chromatin regions: K1: n = 880; K3: n = 1838; K4: n = 1015; K5: n = 1276; K9: n = 1042; K10: n = 1238; K11: n = 623.