Joint profiling of histone modifications and transcriptome in single cells from mouse brain

Chenxu Zhu, Yanxiao Zhang, Yang Eric Li, Jacinta Lucero, M. Margarita Behrens and Bing Ren

Genome-wide profiling of histone modifications can reveal not only the location and activity state of regulatory elements, but also the regulatory mechanisms involved in cell-type-specific gene expression during development and disease pathology. Conventional assays to profile histone modifications in bulk tissues lack single-cell resolution. Here we describe an ultra-high-throughput method, Paired-Tag, for joint profiling of histone modifications and transcriptome in single cells to produce cell-type-resolved maps of chromatin state and transcriptome in complex tissues. We used this method to profile five histone modifications jointly with transcriptome in the adult mouse frontal cortex and hippocampus. Integrative analysis of the resulting maps identified distinct groups of genes subject to divergent epigenetic regulatory mechanisms. Our single-cell multiomics approach enables comprehensive analysis of chromatin state and gene regulation in complex tissues and characterization of gene regulatory programs in the constituent cell types.

Methods have been described to profile histone modifications in single cells one mark at a time. However, different histone modifications vary greatly in their cellular specificity and relationships to cell-type-specific gene expression, leading to varying degrees of success in resolving cellular heterogeneity from complex tissues. Thus, it is very difficult or nearly impossible to integrate datasets of multiple histone marks in different cell types from complex tissues generated using such approaches. Moreover, to better understand the gene regulatory mechanisms, it is necessary to assess the transcriptional profiles along with chromatin states from the same cells. Thus, a single-cell approach that can jointly assay both histone modifications and gene expression would be highly desirable.

Here we present a highly scalable strategy, Paired-Tag (parallel analysis of individual cells for RNA expression and DNA from targeted tagmentation by sequencing), for joint analysis of gene expression and histone modifications in single cells. Paired-Tag extends a technique that we previously reported, parallel analysis of individual cells for RNA expression and DNA accessibility by sequencing (Paired-seq), for co-assay of open chromatin and gene expression in single cells, by adapting the cleavage under targets and tagmentation (CUT&Tag) strategy. To demonstrate the utility of Paired-Tag, we used it to map transcriptome and histone modifications jointly from the adult mouse frontal cortex and hippocampus, and generated, for the first time to our knowledge, cell-type-resolved maps of chromatin state and transcriptome for 22 mammalian brain cell types, providing insight into the gene regulatory programs of different groups of genes in these cell types.

Results

Overview of Paired-Tag. Paired-Tag includes the following steps (Fig. 1a and Extended Data Fig. 1a). First, permeabilized nuclei are incubated with antibodies against specific histone modifications to target the binding of protein A-fused Tn5 transposase to chromatin. Tagmentation reaction and reverse transcription are then sequentially performed. Specifically, reactions are carried out in 12 different wells, each with a well-specific DNA barcode included in the transposase adaptors and reverse transcription primers, to label different samples or replicates (first round of barcodes). Next, a ligation-based combinatorial barcoding strategy is used to introduce the second and third rounds of DNA barcodes to the nuclei, by sequentially attaching well-specific DNA barcodes to the 5′-end of both chromatin DNA fragments and complementary DNA (cDNA) from reverse transcription in 96-well plates. Finally, the barcoded nuclei are divided into sublibraries and lysed, and the chromatin DNA and cDNA are purified, amplified and split into two sequencing libraries, one for each modality (Supplementary Table 1 and Methods).
Fig. 1: Overview of Paired-Tag. a. Schematic of Paired-Tag workflow. See main text for description. b. Single-cell, joint analysis of histone modifications and transcriptome in adult mouse frontal cortex and hippocampus. Paired-Tag assays were performed with antibodies against H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K9me3, and compared the aggregate profiles of each histone modification with published ChIP–seq datasets of this cell line 35:

We first performed Paired-Tag with ~10,000 HeLa cells each with antibodies against H3K4me1, H3K27ac, H3K27me3 and H3K9me3, and compared the aggregate profiles of each histone modification with published ChIP–seq datasets of this cell line 35:

-60% of called peaks from Paired-Tag overlapped with those identified from ChIP–seq datasets generated from the same cell line (65.9% for H3K4me1, 65.7% for H3K27ac, 59.6% for H3K27me3 and 64.0% for H3K9me3). The genomic distributions of the histone
We next sequenced a total of ~70,000 (SPLiT-seq)34 and 10x scRNA-seq (Extended Data Fig. 2g). The capture efficiency of nuclear RNA in Paired-Tag is also on Snap25 (simultaneous indexing and tagmentation-based ChIP-seq)18 and nuclei with low sequence coverage and removing potential doublets data for ~20,000 nuclei from the same tissues. After filtering out ASCs, microglia, and endothelial and ependymal cells (Fig. 1d–f, two groups of oligodendrocytes (OGCs), two groups of astrocytes Snap25 hippocampal neuron types (Extended Data Fig. 1b and Extended Data Fig. 2a). The aggregate single-cell Paired-Tag DNA profiles and bulk profiles generated in parallel showed excellent agreement (PCC: 0.79–0.96; Extended Data Fig. 2b). Paired-Tag generated datasets with high mapping rates: >95% of H3K4me1 and H3K27ac; ~72% of H3K27me3; and >85% of H3K4me3, H3K9me3 and RNA reads can be mapped to the reference genome, with an estimated random barcode collision rate of ~8.2% (Extended Data Fig. 2c,d). To estimate the library complexities of Paired-Tag datasets, we sequenced a fraction of representative nuclei to near saturation (PCR duplication rates: 81.4% for H3K4me1, 81.2% for H3K4me3, 80.2% for H3K27ac, 57.8% for H3K27me3, 58.2% for H3K9me3, 63.6% for RNA). We recovered up to ~20,000 unique loci mapped per nucleus for DNA profiles (median number of unique loci per nucleus, H3K4me1: 19,332 and 17,357; H3K4me3: 2,571 and 2,046; H3K27ac: 4,460 and 4,543; H3K27me3: 2,565 and 2,499; H3K9me3: 16,404 and 18,497, for frontal cortex and hippocampus, respectively) and up to ~15,000 unique molecular identifiers (UMIs) per nucleus for RNA profiles with 85% of them assigned to genic regions (median numbers, 14,295 and 8,185 UMIs, corresponding to 2,400 and 1,855 genes, for frontal cortex and hippocampus, respectively) (Fig. 1c and Extended Data Fig. 2e–g). The histone modification profiles have a similar number of unique loci captured per nucleus as the recently developed combinatorial barcoding and targeted chromatin release (CoBATCH)95, and are higher than for single cell itChiP (simultaneous indexing and tagmentation-based ChiP-seq)18 and high-throughput single-cell chromatin immunoprecipitation followed by sequencing (HT-scChiP-seq)18 (Extended Data Fig. 2e). The capture efficiency of nuclear RNA in Paired-Tag is also on par with other commonly used single cell RNA-seq (scRNA-seq) assays, such as split-pool ligation-based transcription sequencing (SPLIT-seq)14 and 10x scRNA-seq (Extended Data Fig. 2g).

Histone modification maps of cortical and hippocampal cell types in adult mouse brain. We next sequenced a total of ~70,000 nuclei to moderate depth. In parallel we also generated Paired-seq11 data for ~20,000 nuclei from the same tissues. After filtering out nuclei with low sequence coverage and removing potential doublets (Methods), we recovered 64,849 nuclei with matched DNA and RNA profiles (Extended Data Fig. 2h–k and Supplementary Table 2). Compared with 10x scRNA-seq, Paired-Tag RNA datasets have more reads mapped to genic regions and captured more intrusive reads (Extended Data Fig. 2k,l). We clustered these nuclei into 22 cell groups based on their transcriptionome profiles and assigned them to seven cortical neuron types (Snap25+, Sata2+, Gad1–), four hippocampal neuron types (Snap25+, Slc17a7+ or Prox1+), three inhibitory neuron types (Gad1/Gad2+) and eight non-neuron cell types (Snap25+) including oligodendrocyte precursor cells (OPCs), two groups of oligodendrocytes (OGCs), two groups of astrocytes (ASCs), microglia, and endothelial and ependymal cells (Fig. 1d–f, Extended Data Fig. 3a–c and Supplementary Table 3). We also compared our transcriptomic profiles with previously published scRNA-seq datasets from the same brain regions (reference dataset)36 and found excellent agreement (Extended Data Fig. 3d–g). Specifically, 16 of the 22 clusters can be uniquely assigned to a corresponding cluster (or several closely related subclusters) of the reference dataset. Some of the subclusters here matched multiple subclusters of the reference dataset, which includes: the frontal cortex (FC) L5 and L6 matched with TEGLU10 and 20; the FC L2/3 and L4 matched TEGLU7 and 8; and the hippocampus (HC) cornu ammonis (CA) 1 and subiculum clusters in our datasets fell into two CA1 neuron groups (TEGLU21/24 and TEGLU23) of the reference dataset (Extended Data Fig. 3e).

Comparison of histone modification- and transcriptome-based cell clustering. We also clustered the Paired-Tag profiles based on datasets of different histone marks (Fig. 2 and Methods). For fair comparisons, we first performed independent cell clustering based on transcriptomic profiles for each Paired-Tag dataset (Extended Data Fig. 3b). As expected, reduced cell number lowered the resolution in identifying cell types: for H3K4me1, H3K27ac and H3K9me3 Paired-Tag datasets (>11,000 nuclei), 18 clusters were revealed (four closely related cell subtypes identified from the 65,000 merged dataset were grouped together: FC L6 was merged with FC L5; HC subiculum cells were merged with HC CA1; InNeu-Sst and InNeu-Pvalb were classified as InNeu-MGE; and two OGC subtypes were identified as one group); for H3K27me3 Paired-Tag dataset with 6,534 nuclei, 14 clusters were revealed (FN, HC CA2/3, InNeu-Pvalb and endothelial were further mixed with closely related cell groups). Next, with histone modification H3K4me1-, H3K4me3- and H3K27ac-based clustering, we revealed 18, 3 and 15 clusters, respectively (Extended Data Fig. 4a–c). Sixteen groups of H3K4me1-based and 12 of H3K27ac-based clustering matched well with those from RNA-based clustering (Fig. 2c). Two cortical neuron clusters (L4 and L5) in H3K4me1- and H3K27ac-based clustering matched with L2/3, L4 and L5 groups of RNA-based clustering. Additionally, RNA-based clustering can further separate ASC subgroups; in the H3K27ac dataset, RNA profiles can identify FC near projecting excitatory neurons (NP), HC CA1 and CA2/3 from FC corticothalamic excitatory neurons (CT) and HC CA groups of DNA-based clustering (Fig. 2c). For promoter histone mark H3K4me3, only neuronal cells from two different brain regions and the non-neurons can be distinguished (Extended Data Fig. 4b). For H3K27me3-based clustering, all cortical excitatory neurons formed a single cluster (Extended Data Fig. 4d). For H3K9me3 only, the major non-neuron cell types can be separated, while all neuronal cell types were grouped together as a single cluster (Extended Data Fig. 4e). These results suggest that cell clustering based on Paired-Tag histone modification profiles varies considerably depending on the histone marks used, and that the repressive histone marks do not resolve the cell types as well as the active histone marks H3K4me1 and H3K27ac but better than the promoter mark H3K4me3 (Extended Data Fig. 4f,g).

We further performed joint clustering based on histone modification and gene expression profiles from the same cells (Fig. 2b and Methods). Generally, joint clustering with the two modalities identified more cell groups for most histone marks (Fig. 2c). For example, we identified FC NP and ASC subgroups, which are presented in RNA-based clustering of H3K4me1 and H3K27ac Paired-Tag datasets but absent from the corresponding histone mark-based clustering results (two ASC subtypes were grouped together in H3K4me1 and H3K27ac datasets, FC NP was merged with FC CT in H3K27ac dataset); on the other hand, we identified the subiculum cell group by H3K4me1 DNA-based clustering, but not from RNA-based clustering. Interestingly, for H3K27ac datasets, the subiculum group was missing from both DNA and RNA groups and joint clustering of both modalities helped to identify them. For H3K27me3 the histone modification profiles can only separate major cell groups
and, thus, the joint clustering mainly recapitulated the same cell clustering as RNA-based analysis. For H3K9me3 and H3K4me3, the resolution of joint clustering was even poorer than RNA-based clustering alone, as the histone signal can only separate neurons from non-neuron cells or neurons from different regions.

The inconsistency of cell clustering based on different histone marks indicates that it is important to use the transcriptome profiles to construct the cell-type-specific epigenome maps. We therefore generated genome-wide maps of each histone modification along with gene expression profiles in each of the 22 mouse brain cell types identified based on clustering of the RNA components of Paired-Tag datasets (Fig. 1f). To facilitate the dissemination of these maps, we set up a web portal (http://catlas.org/pairedTag) to enable interactive exploration of the chromatin state in each brain cell type.

Integrative analysis of chromatin states and gene expression at gene promoters across different brain cell types. To investigate the relationship between chromatin states and cell-type-specific gene expression, we aggregated the Paired-Tag signals of each histone modification at annotated gene promoters (~1,500 base pairs (bp) to +500 bp) in each of the brain cell types. For this analysis, we mainly examined the 18 cell groups with at least 50 cells and at least 50,000 combined unique reads for all molecular modalities in this cell group. A total of 17,398 genes (GENCODE GRCm38.p6) with sufficient levels of transcription (reads per kilobase of transcript per million mapped reads (RPKM) > 1 in at least one cell type) or promoter occupancy (counts per million (CPM) > 1 for any histone mark in at least one cell group; Methods) were retained for subsequent analysis. Using k-means clustering, we categorized these gene promoters into seven groups with distinct combinations of histone modification: class I promoters were associated with the heterochromatin mark H3K9me3 (13.1% of all tested genes), class II-a and II-b groups were associated with the polycomb repressive histone modification: class I promoters were associated with the heterochromatin mark H3K9me3 (13.1% of all tested genes), class II-a and II-b groups were associated with the polycomb repressive histone mark H3K27me3 (9.2% of all tested genes) and the remaining four groups of promoters were associated with variable levels of active histone marks H3K4me1 and H3K27ac (77.6% of all tested genes) (Fig. 3a–c, Extended Data Fig. 5a–c and Supplementary Table 4).

We carried out Gene Ontology (GO) analysis and found distinct functional categories of genes within each group (Fig. 3d and Supplementary Table 5). For example, genes in class I were strongly enriched for sensory-related pathways, including olfactory receptor genes (Olf, 647 out of 730 detected) and vomeronasal receptor genes (Vom, 189 out of 201 detected). Olfactory receptor genes were previously shown to be associated with heterochromatin marks during the process of olfactory receptor choice in olfactory sensory neurons. Our data suggest that olfactory receptor genes are silenced by similar mechanisms in the frontal cortex and hippocampus (Extended Data Fig. 5d). H3K27me3-repressed genes were further divided
Class II-a genes were repressed in all cell clusters and class II-b genes were repressed in a more restricted manner. GO analysis revealed that II-a group genes were enriched for terms of general developmental processes such as pattern specification process and embryonic organ development, while II-b group genes were enriched for terms including morphogenesis of an epithelium (Fig. 3d, Extended Data Fig. 5e,f and Supplementary Table 5). Genes in III-a group were characterized by active chromatin state at promoters in all cell types (10.4% of class III genes), while genes in III-b group were expressed in all neuronal cell types (5.9% of class III genes) and genes in III-c group were glial-expressed (31.0% of class III genes). Group III-d genes (52.6% of class III genes) were marked (Fig. 3d).
by active chromatin state in a cell-type-specific manner, with corresponding cell-type-specific expression patterns (Fig. 3a–c). These genes were enriched for GO terms with more specific cellular processes: for example, hippocampal neuron-expressed genes were enriched for learning or memory and microglia-expressed genes were enriched for inflammatory response (Fig. 3d and Supplementary Table 5). We also carried out pseudotime analysis on OGC lineages based on the transcriptome profile and assigned the promoter chromatin states to the differentially expressed genes (Fig. 3e, Extended Data Fig. 5,g,h and Methods). The chromatin accessibilities and densities of active chromatin marks (H3K4me1, H3K4me3 and H3K27ac) were concordant with expression levels of these genes (Fig. 3f). For example, Sox5 and Pdgfra were highly expressed in progenitor populations, agreeing with the role of Sox5 in activating Pdgfra expression to maintain the immature state of OGC progenitor cells. These results demonstrated the utility of Paired-Tag in studying epigenetic regulatory programs during development and cell differentiation.

Integrative analysis of chromatin state at distal elements across brain cell types. Cis-regulatory elements (CREs) are marked with highly cell-type-specific chromatin states that strongly correlate with cell-type-specific gene expression. Recently, a comprehensive analysis of chromatin accessibility from the adult mouse cerebrum identified 491,818 candidate CREs (cCREs)\(^{(39)}\). We found that 286,168 (58.2%) distal CREs from this list showed sufficient levels of Paired-Tag signals in at least one cell group and one or more histone marks (CPM > 1, and more than 1,500 bp upstream and 500 bp downstream from the transcription start sites (TSSs)). To characterize the chromatin state of these cCREs across different brain cell types, we again performed k-means clustering with the aggregate Paired-Tag signals of different histone marks in each of the 18 cell clusters defined above (Methods). We categorized these cCREs into eight groups: two were marked by H3K3me3 either in all cell clusters (class eI-a, 16.3% of all CREs) or selectively in neuronal cells (class eI-b, 4.9% of all CREs), and two were marked with H3K27me3 (eII-a, 5.5% and eII-b, 3.1% of all CREs) primarily in all neuronal cell clusters or in a more restricted manner (eII-b elements). The remaining four groups (class eIII-a to eIII-d) were marked by variable levels of H3K4me1 and H3K27ac modifications in different cell clusters (Fig. 4a–c and Supplementary Table 6). As expected, negligible H3K4me3 signals were detected for these promoter-distal cCREs (Fig. 4a). Similar to the promoter groups, the subclass of cCREs with H3K27ac mark in one or a few cell groups comprised the largest fraction (class eIIII-d, 37.1% of all CREs). H3K9me3-marked cCREs reside preferentially in intergenic regions (eI-a and eI-b), while cCREs marked by relatively invariant H3K4me1 and H3K27ac levels tend to reside in genic regions (eIII-a) (Extended Data Fig. 6a). Class eII-b cCREs were significantly enriched for CpG island (CGI) regions (5.4%, \(P < 2.2 \times 10^{-16}\)) and eII-a cCREs were less enriched (2.0%, \(P = 0.002\)). The two H3K9me3-marked groups were depleted from CGI regions (0.16% and 0.12%, \(P < 2.2 \times 10^{-16}\)). For the active cCRE groups, class eII-a cCREs displayed the highest enrichment for CGI regions (14.1%, \(P < 2.2 \times 10^{-16}\)) while the other subclasses of eIIII cCREs were not enriched for CGI (Extended Data Fig. 6b).

To identify potential TFs that act on the above classes of cCREs, we performed de novo motif enrichment analysis (Fig. 4d). The heterochromatin-associated eI-a and eI-b group cCREs were enriched for recognition motifs of well-known transcriptional repressors EVX1 and MAFG, which play critical roles in embryogenesis\(^{(40)}\) and in the central nervous system\(^{(41)}\), respectively (Fig. 4d). The two polycomb-repressed cCRE groups eI-a and eI-b were enriched for LHX motifs and the group eIIII-d cCREs with dynamic H3K27ac across all clusters were enriched for motif matched with CCCTC-binding factor (CTCF) recognition (Fig. 4d). We also performed enrichment analysis of known TF motifs followed by k-means clustering and revealed distinct modules (Fig. 4e and Supplementary Table 7). For example, the heterochromatin (eI-a) and inhibitory neuron (eIIII-d) groups were enriched for the Ascl1 motif. Ascl1 can function as a pioneer factor targeting closed chromatin to activate the neurogenic gene expression programs\(^{(42)}\) as well as to induce the generation of GABAergic neurons\(^{(43)}\).

To infer potential regulators for each cell lineage by leveraging the joint profiles, we calculated the TF motif enrichment in each cell type, and compared it with expression levels of the corresponding TF genes (Extended Data Fig. 6c). More than half of the TFs (65%) showed a positive correlation between gene expression levels and motif enrichment in the cCREs across different cell types (Supplementary Table 8). For example, one of the top-ranked TFs, Fli1, was restricted in microglia and endothelial cells. Fli1 is known to activate chemokines to mediate the inflammatory response in endothelial cells\(^{(44)}\) and was recently found to be in a coordinated gene expression module associated with Alzheimer’s disease\(^{(45)}\). Other highly ranked TFs including Sox9/10, Mef2c and Neurod2 are known to play a critical role in the development of neuronal systems. We further characterized the chromatin states based on the five histone marks using chromHMM. Eight chromatin states were recovered, including active and weak promoters, active and primed enhancers, H3K27me3-associated and H3K9me3-associated heterochromatin, and two categories without detected signals from these five histone profiles (Fig. 4f,g and Extended Data Fig. 6d,e). The promoter regions are less variable compared with active enhancers across brain cell types. The H3K9me3-associated heterochromatin regions are generally invariable among all excitatory neuron cell types; in non-neuronal cells these chromatin regions adopt other chromHMM states (Fig. 4h).

Integrative analysis of chromatin state and gene expression connects distal cCREs to putative target genes. Distal regulatory elements, including enhancers and silencers, control cell-type-specific transcriptional programs during development or in response to stimuli. Imaging-based tools and chromosome conformation capture techniques have been extensively used to elucidate the interplay between promoters and distal CREs\(^{(46)}\). Recently, efforts were made to predict putative targets for cell-type-specific enhancers from complex tissues using single-cell genomics tools\(^{(47)}\), but less attention was paid to the repressive CREs\(^{(48,49)}\). The epigenetic and transcriptional states from the same cells provide an excellent opportunity to connect both the candidate active and repressive cCREs to their putative target genes. We first identified putative promoter–CRE pairs based on co-occupancy of H3K4me1 reads between cCRE and TSS-proximal regions (–1,500 bp to +500 bp) across all cells. We then calculated the pairwise Spearman’s correlation coefficients (SCCs) between the gene expression levels of the putative target genes and the histone mark levels of the cCREs across cell clusters (Fig. 5a and Methods). We identified 32,252 cCRE–gene pairs where H3K27ac levels at the distal cCREs positively correlated with gene expression, and 15,199 cCRE–gene pairs where H3K27me3 levels at the cCREs negatively correlated with expression of linked genes (false discovery rate (FDR) < 0.05; Fig. 5b,c, Extended Data Fig. 7a–f, Supplementary Table 9 and Methods). Interestingly, a significant fraction of H3K27ac-associated cCRE–gene pairs were in common with the H3K27me3-associated cCRE–gene pairs (\(P < 2.2 \times 10^{-18}, 2.621\) observed compared with 185 randomly expected) (Fig. 5d and Supplementary Table 9). The cCREs in these shared pairs tended to be in the eII-b group (Fig. 5e), and to target genes that were enriched for development processes such as gliogenesis and forebrain development (Extended Data Fig. 7g). These results are consistent with the recent finding that transition between PR2-associated silencers and active enhancers occurs during differentiation\(^{(44)}\). Despite the potentially shared fraction, CREs of the inactive pairs are more

---

**ARTICLES**

**NATURE METHODS**

---

288

NATURE METHODS | VOL 18 | MARCH 2021 | 283–292 | www.nature.com/naturemethods
enriched in intergenic regions as well as being more distal to their targets (Extended Data Fig. 7h–j).

We next compared the categories of cCREs with those of their putative target genes. Interestingly, promoters of target genes tended to be in similar groups with cCREs: for example, target genes of class eI-a and eI-b cCREs were strongly enriched in promoters of class II-a and II-b genes (Fig. 5f). These genes are enriched in those with functions in development processes (Fig. 3c and Extended Data Fig. 5f). We then compared the chromatin state of cCREs with the promoters of the putative target genes: cCREs and promoters from the active pairs displayed higher concordance for their H3K27ac levels, but not for the inactive pairs; on the other hand, higher concordance for H3K27me3 levels was only observed from the inactive pairs (Fig. 5g and Extended Data Fig. 7k). We then grouped the cCREs with linked genes according to their H3K27 acetylation and methylation states (Fig. 5h,i, Extended Data Fig. 7l). Target genes of neuron-specific cCRE groups are enriched in GO terms including modulation of synaptic transmission; genes linked to cCRE groups of glial cells are enriched for terms including gliogenesis, morphogenesis...
of epithelium, neuron projection morphogenesis and so on (Fig. 5j). For the inactive pairs, only a small fraction showed strong regulatory processes acting at different categories of genes and cCREs. Through integrative analysis, we identified distinct epigenetic regulatory processes acting at different categories of genes and cCREs. Paired-Tag enables the joint profiling of transcriptome and epigenome in single cells and can be used to identify enhancers linked to expression of gene targets.

**Discussion**

Here we report a high-throughput method for simultaneous profiling of histone modifications and gene expression in single cells.

**Fig. 5 | Correlative analysis of chromatin state and gene expression links distal candidate cis-regulatory elements to putative target genes.**

- **a.** Schematics for identifying potential targets for cCREs. Putative CRE–gene pairs were first determined by calculating the co-occupancy of H3K4me1 reads in gene promoter and distal CRE regions. The Spearman correlation coefficients (SCCs) between expression levels of genes and H3K27ac or H3K27me3 levels of cCREs were then used to identify CRE–gene pairs. **b.** Density plot showing the distribution of correlations between histone modification levels of cCREs and expression of potential target genes. The cutoffs (FDR = 0.05) for identifying potential H3K27ac-associated active and H3K27me3-associated inactive CRE–gene pairs are also indicated. **c.** Representative genome browser view of Olig1 and Olig2 gene locus; both H3K27ac- and H3K27me3-associated cCREs are shown. TSS-proximal regions are marked with green boxes and cCREs are indicated with blue (H3K27ac-specific), brown (H3K27me3-specific) or purple (shared) boxes. **d, e.** Venn diagram showing the overlap between predicted H3K27ac- and H3K27me3-associated CRE–gene pairs. **f.** Heatmap showing the enrichments of predicted target genes of each promoter group defined in **Fig. 3a** linked by cCREs of each CRE group defined in **Fig. 4.** **g.** SCCs between read densities of cCREs and promoters of putative target genes across cell types for H3K27ac and H3K27me3. **h.** Heatmap showing the histone modification levels at cCREs with potential active roles in expression of putative target genes. cCREs were grouped using K-means clustering based on histone modification levels. **i.** Heatmap showing the expression levels of corresponding putative target genes of cCREs in **h.** **j.** Top enriched GO terms for genes in **i** and the top enriched de novo motifs for cCREs from each group in **h.** NS, not significant.
construction compared with other commercial single-cell omics platforms. Compared with single cell ATAC-seq (scATAC-seq) or joint ATAC-seq–gene expression assays, our method can reveal the functional states of cCREs to provide mechanistic insights of regulatory programs for each cell type from heterogeneous cellular environments. Computational methods have been developed for the integrated analysis of single-cell datasets from multiple modalities; however, it is still challenging to integrate different histone modifications with distinct biological features. The ability of Paired-Tag to achieve unbiased cell-type classification from the transcriptional states provided a unique chance to integrate previous single-cell atlases.

The protein A-Tn5 strategy can also be adapted to droplet-based platforms, such as SNARE-seq\(^8\), to allow efficient generation of single-cell dual-omics profiles (Supplementary Note). While the current Paired-Tag protocol surveys only single histone marks with nuclear transcriptome, with additional modifications, it should be possible to monitor multiple histone marks from the same cells, to help reveal crosstalks and combinatorial effects of different histone marks during gene expression. In addition, the barcoding strategy here can be further extended to the analysis of other molecular modalities such as DNA base modifications and TF binding.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-021-01060-3.

Received: 9 September 2020; Accepted: 4 January 2021; Published online: 15 February 2021

**References**

1. Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* **569**, 345–354 (2019).
2. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316**, 1697–1702 (2007).
3. Crawford, G. E. et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).
4. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epidemic profiling of open chromatin. DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
5. Consortium, E. P. et al. Expanded encyclopedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
6. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
7. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
8. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142–146 (2015).
9. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
10. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
11. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
12. Hainer, S. J., Boskovic, A., McCannell, K. N., Rando, O. J. & Fazzio, T. G. Profiling of pluripotency factors in single cells and early embryos. *Cell* **177**, 1319–1329 e1311 (2019).
13. Harada, A. et al. A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat. Cell Biol.* **21**, 287–296 (2019).
14. Kaya-Oktar, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
15. Carter, B. et al. Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). *Nat. Commun.* **10**, 3747 (2019).
46. Yu, M. & Ren, B. The three-dimensional organization of mammalian genomes. *Annu. Rev. Cell Dev. Biol.* **33**, 265–289 (2017).
47. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 e858 (2018).
48. Ngan, C. Y. et al. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nat. Genet.* **52**, 264–272 (2020).
49. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat. Genet.* **52**, 254–263 (2020).
50. Wang, Y., Lin, L., Lai, H., Parada, L. F. & Lei, L. Transcription factor Sox11 is essential for both embryonic and adult neurogenesis. *Dev. Dyn.* **242**, 638–653 (2013).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Methods

Cell culture and processing. HeLa S3 (human, ATCC CCL-2.2) cells were cultured according to standard procedures in DMEM (ThermoFisher Scientific, 10569010) supplemented with 10% fetal bovine serum (ThermoFisher Scientific, 16000044) and 1% penicillin-streptomycin (ThermoFisher Scientific, 10378016) at 37 °C with 5% CO₂. Cells were not authenticated or tested for mycoplasma. To prepare nuclei, HeLa S3 cells were collected by centrifugation, washed with PBS (ThermoFisher Scientific, 10010-23) and counted using a BioRad TC20 cell counter. The cells were then resuspended in cold nuclei permeabilization buffer (10 mM Tris-HCl pH 7.4 (Sigma, T1119), 1 mM EDTA (Sigma, E6008), 0.1% Triton X-100 (Sigma, T9284)). The cell suspension was then filtered by 30-μm Cell-Tiri (Systex) and spun down for 10 min at 1,000 g, 4 °C before proceeding to Paired-Tag experiments.

Processing of biospecimens. Male C57BL/6J mice were purchased from Jackson Laboratories at 8 weeks of age and maintained in the Salk animal barrier facility on 12-h dark/light cycles with controlled temperature (20–22 °C) and humidity (30–70%), and with food ad libitum, for 4 weeks before dissection. The frontal cortex and hippocampus were dissected and snap-frozen in dry ice. All protocols were approved by the Salk Institute’s Institutional Animal Care and Use Committee.

Single-cell suspensions were prepared from Douncing of the frozen tissues, in Douncing buffer with protease/RNase inhibitor cocktail (DBI: 0.25 M sucrose (Sigma, S7903), 25 mM KCl (Sigma, P9333), 5 mM MgCl₂ (Sigma, 63069), 1 mM EDTA, RNase OUT, 0.01% IGEPAL CA630, 0.01% Digitonin and 2 mM EDTA). Barcoded DNA adaptor oligos were mixed with a mixture of 10 mM dNTPs, 0.5 U/l SUPERase IN, 0.5 U/l RNase OUT, 0.01% IGEPA El AP630 and 0.01% Digitonin, and this was repeated twice. The tagmentation reaction was initiated by adding 2 μl of 250 mM MgCl₂, (Sigma, M7522) and was carried out at 350 rpm, 37 °C for 60 min in a ThermoMixer (Eppendorf). The reaction was quenched by adding 16.5 μl of 40.5 mM EDTA. Nuclei were then spun down at 1,000 g, 4 °C for 10 min and proceeded to reverse transcription immediately.

Reverse transcription. Reverse transcription was carried out similarly to that previously described with minor modifications. Briefly: nuclei pellets were resuspended in 20μl of reverse transcription buffer (1× Buffer RT, 0.5 mM dNTPs (NEB, N0447S), 0.5 U/l SUPERase IN, 0.5 U/l RNase OUT, 2.5 μM barcoded Ti5 primer and 1 μl Maximax Reverse H Minus Reverse Transcriptase (Invitrogen, EP0751)). The reverse transcription was performed in a thermocycler with the following program: Step 1: 50 °C for 5 min; Step 2: 28 °C × 12 s, 15 °C × 45 s, 20 °C × 30 s, 32 °C × 2 min, 42 °C × 5 min, go to Step 2 for additional two times; Step 3: 35 °C × 10 min and hold at 12 °C. After the reaction, the nuclei were transferred and pooled into 1.5-ml tubes prewashed with 5% BSA in PBS and cooled on ice for 2 min, with 4.8 μl of 5% Triton X-100 (Sigma, T9284). Nuclei were then spun down at 1,000 g, 4 °C for 10 min and proceeded to ligation-based combinatorial barcoding immediately.

Ligation-based combinatorial barcoding. Nuclei were resuspended and mixed in 1 ml of 1× NEBuffer 3.1 and then transferred to ligation mix (2,262 μl of H₂O, 500 μl of T4 DNA Ligase Buffer, 50 μl of T4 DNA Ligase (NEB, M0202L)). Each 40 μl of the ligation mix was distributed to Barcode-plate-R02 using a multichannel pipette and incubate at 300 rpm, 37 °C for 30 min in a ThermoMixer (Eppendorf). Then, 10 μl of R02-Blocking-Solution (264 μl of 100 μM Blocker-R02 oligo, 250 μl of 10× T4 Ligase Buffer, 468μl of ultrapure H₂O) was added to each well using a multichannel pipette and the reactions were continued for another 30 min. The nuclei were then pooled and spun down at 1,000 g, 4 °C for 10 min. The second round of ligation was then carried out similarly to the first round, except that after 30 min of the ligation reaction, Termination-Solution (264 μl of 100 μM R04 Terminator oligo, 250 μl of 0.5 M EDTA and 236 μl of ultrapure H₂O) was added to quench the reaction.

Typically, 10,000 to 30,000 nuclei could be recovered after ligation-based barcoding. Nuclei were then resuspended in PBS, counted and aliquoted to sublibraries containing 2,000 to 5,000 nuclei. The size of sublibraries positively correlates with potential barcode collision rates35; hence, we used 3,500 nuclei for each sublibrary, which gave us a 0.2% collision rate (expected estimate 3.2%). To further reduce the collision rate, nuclei sorting can be used to replace dilution-based aliquoting36. Sublibraries were diluted to 35 μl with PBS. Then, 5 μl of 4 M NaCl (Sigma, S7653), 5 μl of 10% SDS (Invitrogen, 15535-035) and 10 μl of 10 mM dNTPs (NEB, P81075) were added and mixed and then lysed at 850 rpm, 55 °C for 2 h in a Thermomixer. The lysed solution was cooled to room temperature and then purified with 1×SPRI beads (Beckman Coulter, B23119) and eluted in 12.5 μl of H₂O. The purified DNA can be stored at −20 °C or −80 °C for up to 4 weeks.

Preamplification of barcoded DNA/cDNA. First, 1.5 μl of 10× TD buffer and 0.5 μl of 1 mM dCTP (NEB, N0447S) were added into 12.5 μl of purified DNA/cDNA mix and then incubated at 37 °C for 60 min. Second, 5 μl of 10 mM dCTP was added with a ramp of 95 °C for 30 s followed by 30 cycles of 95 °C for 10 s, 68 °C for 10 s and 72 °C for 10 s, with 1 μl of 10× KAPA buffer, 0.6 μl of 10 mM dNTPs, 0.6 μl of 10 μg/ml Anchor-FokI-GSH-Oligo and 0.6 μg of KAPA HiFi HS (KAPA, KK2502) was added and the linear amplification was performed in a thermocycler with the following program: Step 1: 98 °C × 3 min; Step 2: 98 °C × 15 s, 72 °C × 60 s; Step 3: 72 °C × 2 min and hold at 12 °C. Amplified products were purified with SPRI double-size selection (10 μl + 32.5 μl, 0.2× + 0.6×) and were eluted in 35 μl of H₂O.

Endonuclease digestion and second adaptor tagging. First, we transferred 17 μl of each of the purified amplified products into two tubes for DNA and RNA library construction, respectively. Then, we added 2.5 μl of 10× Cutsmart buffer (NEB, M72045), 1 μl of SbfI-HF (NEB, R3642), 1 μl of FokI (NEB, R01095) and 3.5 μl of H₂O to a DNA-tube. Next we added 2 μl of 10× Cutsmart buffer and 1 μl of NotI-HF (NEB, R3189) to an RNA-tube. The digestion reaction was incubated at 37 °C for 12 h. We used 1.25 μl of 5× DNA and 25 μl for RNA SPRI beads to purify the digestion product and eluted in 10 μl H₂O. For the DNA part, 2 μl of 10× T4 DNA Ligase Buffer, 2 μl of P5 Adaptor Mix, 4 μl of H₂O and 2 μl of T4 DNA Ligase were added and ligation reactions were carried out with the program: 4 °C for 10 min, 10 °C for 15 min, 16 °C for 15 min, 25 °C for 45 min.
The ligation product was then purified with 1.25X (25 μl) SPRI beads and eluted in 30 μl of H2O. For the RNA part, we added 10.5 μl of 2X Tagmentation Buffer (66 mm Tris-Ac, pH 7.8 (ThermoFisher Scientific, BP-152), 22 mm MgAc (Sigma, M2545), 133 mm KAc (Sigma, P5708) and 32% DMAP (EMD Millipore, DX1739)), and 0.5 μl of 0.05 mg/mL Tn5-N5 was added and tagmentation reactions were carried out at 550 r.p.m., 37°C for 10 min in a ThermoMixer followed by clean-up using QIAquick PCR purification kit and elution in 30 μl of 0.1X Elution Buffer (QIAGEN).

Indexed PCR and sequencing. We prepared the PCR mix (30 μl of purified P5-tagged product, 10 μl of 5X Q5 buffer, 1 μl of 10 mM dNTPs, 0.5 μl of 50 μM P5 Universal primer for DNA or N5 primer for RNA, 2.5 μl of 10 P7 primer, 5 μl of H2O and 1 μl of NEB Q5 DNA Polymerase (NEB, M0491)) and ran the following program: Step 1: 72°C × 5 min; Step 2: 98°C × 30 s; Step 3: 98°C × 10 s, 63°C × 30 s, 72°C × 1 s; Step 2 and 3 repeated Step 2 an additional 8–13 times to reach 10 mM concentration; Step 3: 72°C × 1 min and hold at 12°C. We cleaned up the library using 0.9X (45 μl) SPRI beads. The final libraries were sequenced using a NextSeq 550 or NovaSeq 6000 platform (Illumina) with the following read lengths: paired-end 50 ± 7 + 100 or PE 100 ± 7 + 100 (Read 1 + Index 1 + Read 2).

Paired-seq. An optimized Paired-seq protocol based on previously the described method was used to generate pair chromatin accessibility and transcriptome. Briefly, nuclei were isolated similarly to Paired-Tag procedures and then counted, and 1.2 million permeabilized nuclei were aliquoted into 12 tubes and spun down at 1,000 × g for 10 min and then resuspended in 4°C of 1.1X Tagmentation Buffer (36.7 mm Tris-Ac, pH 7.8, 12.1 mm MgAc, 73.3 mm KAc, and 17.8% DMAP) with 75 μM PitStop2 (ref. 1). Barcoded Tn5 was then added and reactions were carried out at 37°C, 550 r.p.m. for 30 min in a ThermoMixer. Reactions were quenched by adding 5 μl of 40 mM EDTA and nuclei were spun down at 1,000g, 4°C for 10 min. Starting from reverse transcription, the steps are the same as in the Paired-Tag procedures described above.

Quality control of Paired-Tag data. DNA reads were mapped to mm10 or hg19 using bowtie2 (ref. 54) with default parameters; only Read 1 was used for Quality control of Paired-Tag data. DNA reads were mapped to mm10 or Paired-Tag procedures described above. 4 °C for 10 min. Starting from reverse transcription, the steps are the same as in the Paired-Tag procedures described above.

Preprocessing of Paired-Tag data. Cellular barcodes and the linker sequences are read by Read 2. The first bases of barcode (BC) no. 1, BC no. 2 and BC no. 3 should locate within the 84th–87th, 47th–50th and 10th–13th bases of Read 2. We identify the positions of barcodes by matching the linker sequences adjacent to the cellular barcodes. A bowtie reference index was generated with all possible cellular barcode combinations (96 × 96 × 12) and barcode sequences were mapped to the cellular barcode reference using bowtie with the parameters: -v 1 -t 1 –m 1 (reads with more than 1 barcode mismatch and can be assigned to more than 1 cell were discarded). Nextera adaptor sequences were trimmed from 3′ of DNA and RNA libraries. Poly-DT sequences were further trimmed from 3′ of RNA libraries and low-quality reads (minimal length: L = 30, minimal base calling quality: Q = 30) were excluded from further analysis.

Analysis of Paired-tag data. Evaluation of collision rate. Reads from species-mixing test were extracted based on cellular barcodes (BC no. 1 = 06 or 12; Extended Data Fig. 2a) and mapped to a reference genome using STAR56 with the combined reference genome (GRCh37 for human and GRCm38 for mouse). Duplicates were removed based on the mapped position, cellular barcode, PCR index and UMI. For evaluation of the collision rate, nuclei with less than 90% UMI mapped to one species were classified as mixed cells.

Reads mapping. Cleaned reads were first mapped to a mouse GRCm38 reference genome with STAR56 (v2.7.6a) for RNA or bowtie2 (ref. 1) for DNA. Mapped DNA reads of H3K4me1, H3K4me3, H3K27ac and H3K27me3 were further filtered by MAPQ > 10. Duplicates were removed based on the mapped position, cellular barcode, PCR index and UMI. We used BC no. 1 for the identification of the origin of samples. Low-cov-erage nuclei were removed from further analysis (<1,000 transcripts and <300 unique DNA reads). Before generating the cell-counts matrices, DNA bam files were further filtered by removing high-pileup positions (cutoff = 10) regardless of cellular barcode, PCR index and UMI.

Clustering of Paired-Tag profiles. RNA alignment files were converted to a matrix with cells as columns and genes as rows. DNA alignment files were converted to a matrix with cells as columns and 3-kb bins (instead of peaks) as rows. Cells with less than 200 features in both DNA and RNA matrices were removed. The DNA matrix was further filtered by removing the 5% highest covered bins. The clustering of single cells based on RNA profiles was performed with the Seurat package. Briefly, cell-to-gene counts were normalized and variable genes were selected for dimension reduction by principal component analysis (PCA), and batch effects were corrected with harmonization, visualized with UMAP and clustered with the Louvain algorithm. Cell groups with high expression levels of marker genes from multiple major cell types were considered as doublets and excluded from further analyses. Co-embedding of Paired-Tag RNA profile and published scRNA-seq dataset was performed using the Seurat package. To compare the clustering results from different studies, overlap coefficients (O) were calculated according to the number of cells with the labels from the Paired-Tag dataset (A), from Zeisel et al. (B) and from co-embedding (C) (i.e. row index, j = column index):

\[ O_{ij} = \min \left( \frac{A_i \cap C_j}{A_i}, \frac{B_i \cap C_j}{B_i} \right) \]

To visualize the single-cell DNA profiles, cell-to-bin (5-kb bin size) matrices were binarized (B) and converted to cell-to-cell similarity laccard matrices (S) by snapATAC:

\[ S_{ij} = \frac{|B_i \cap B_j|}{|B_i| \cdot |B_j|} \]

The coverage biases were then normalized with the runNormlaccard function of snapATAC, followed by dimension reduction by PCA, batch effect correction with harmonization, visualization with UMAP and cell clustering by the Louvain algorithm.

To cluster single cells with joint modalities, transcript counts of each cell were first filtered by removing the median of the corresponding gene across all cells and log-transformed (L). Cell-to-cell Euclidean distance matrices (E) of RNA profiles were calculated from the normalized cell-to-gene counts matrices:

\[ E_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x}_k)^2} \]

Jaccard distance matrices (D) were then converted from S (D = 1 – S). The two distance matrices (D, E) were rescaled to [0,1] (D* = (1 – D)/n, E* = (1 – E)/n). Dimension reduction (PCA) was then performed on I, followed by batch effect correction with harmonization, visualization with UMAP and cell clustering by the Louvain algorithm. Cell types were annotated based on average expression levels of marker genes from transcriptomic profiles.

Classification of promoter and CRE modules. To classify genes according to epigenetic states of promoters, gene expression (RPKM) and reads densities of promoters (CPM) were summarized from aggregated profiles based on transcriptome-based clustering. Genes with RPKM > 1 for expression and CPM > 1 for promoters in at least one cluster were retained for analysis. Genes were first grouped by k-means clustering based on reads densities of three histone marks (H3K27ac, H3K27me3 and H3K9me3) (k = 4). Each group was then subjected to secondary k-means clustering based on gene expression, resulting in the seven promoter groups in Fig. 3.

To classify CRE into different groups, first, the CRE list was from CEMBA and was extended for 1,000 bp (500 bp in both directions). cCRE overlaps with promoter regions (~1,500bp to +500bp of TSS) were excluded from further analysis. CRE reads densities of four histone marks were then summarized from aggregated profiles based on transcriptome-based clustering. cCREs with CPM > 1 in at least one cluster or one histone profile were retained for analysis. CREs were first grouped by k-means clustering based on reads densities of three histone marks (H3K27ac, H3K27me3 and H3K9me3) (k = 4). Each group was then subjected to secondary k-means clustering based on k27ac reads densities, resulting in the eight CRE groups in Fig. 4.

Motif enrichment and GO analysis. Motif enrichment for each cell type. Motif enrichment for each cell type and histone modifications were carried out using ChromVAR. Briefly, mapped reads were converted to cell-to-bin matrices with a bin size of 1,000 bp for four histone profiles. Reads for each bin were summarized from all cells of the same groups from transcriptome-based clustering. GC bias and background peaks were calculated and motif enrichment score for each cell type was then computed using the `computeDeviations` function of ChromVAR.

Motif enrichment for each CRE module: Motif enrichment for each CRE module was analyzed using Homer (v.4.11). We scanned a region of ±200 bp around the center of the element for both de novo and known motif enrichment analysis (from the JASPAR database). We used the total peak list as the background for motif enrichment analysis of cCREs.

GO enrichment. GO annotation was performed with Homer (v.4.11) with default parameters. Gene set library ‘Biological process’ was used. GO terms with more than 500 total genes in the list were excluded from the ‘top enriched GO terms’.

Pseudotime analysis. Pseudotime analysis was performed with MonoCl3 (ref. 1). Nuclei annotated as OPC, Oligo-MFOL and Oligo-MOL were used for analysis (933, 580, 790, 450 and 1,131 nuclei from H3K4me1, H3K4me3, H3K27ac, etc.).
Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502 (2019).

53. Mulqueen, R. M. et al. Improved single-cell ATAC-seq reveals chromatin dynamics of in vitro corticogenesis. Preprint at bioRxiv https://doi.org/10.1101/637256 (2019).

54. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

55. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).

56. Dobin, A. & Gingeras, T. R. Mapping RNA-seq reads with STAR. Curr. Protoc. Bioinformatics 51, 11 11–11 (2015).

57. Stuart, T. et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902 e1821 (2019).

58. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 16, 1289–1296 (2019).

59. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform manifold approximation and projection. J. Open Source Softw. 3, 861 (2018).

60. Fang, R. et al. SnapATAC: a comprehensive analysis package for single cell ATAC-seq. Preprint at bioRxiv https://doi.org/10.1101/615179 (2020).

61. Scheper, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods 14, 975–978 (2017).

62. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010).

63. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 46, D260–D266 (2018).

64. 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 (2010).

65. Ernst, J. & Kells, M. Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. 12, 2478–2492 (2017).
Extended Data Fig. 1 | Overview of Paired-Tag. a, Schematics for 2nd adaptor tagging of DNA and RNA libraries. For DNA libraries, amplified products were digested with a type IIIS restriction enzyme – FokI, and the cohesive end was then used to ligate the P5 adaptor. For RNA libraries, N5 adaptor was added by tagmentation. b, Table showing the numbers of overlapped peaks across different histone marks between Paired-Tag and ENCODE ChIP-seq or DNase-seq datasets. The total numbers of peaks identified for each dataset are also indicated. c, d, Heatmaps showing the Pearson’s Correlation Coefficients of genome-wide reads distribution (in 10-kb bins) (c) between Paired-Tag datasets and ENCODE ChIP-seq or DNase-seq datasets, and (d) between replicates of Paired-Tag datasets and ENCODE ChIP-seq datasets from HeLa cells. e, Scatter plot showing the Pearson’s correlation coefficients of Paired-Tag RNA dataset and in-house generated nuclei RNA-seq from HeLa cells.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Performances of Paired-Tag in single-nucleus analysis from mouse brain. a, Schematics showing the sample multiplexing strategy in this study. Different samples or replicates were labeled by the 1st round of Paired-Tag cellular barcode (Sample Barcode) located in reverse transcription primers and transposome oligos. b, Heatmap showing the pair-wise Pearson’s correlation coefficients of genome-wide reads distribution for different histone marks from single-cell Paired-Tag datasets (indicated with ‘sc’, aggregated from all cells shown in Fig. 2a) and bulk datasets. c, Boxplots showing the mapping rates (upper panels) and the fraction of reads uniquely mapped to the reference genome (bottom panels) of DNA profiles of different histone marks and RNA profiles in frontal cortex and hippocampus. The boxes were drawn from lower quartile (Q1) to upper quartile (Q3) with the middle line denote the median, whiskers with maximum 1.5 IQR, outliers were indicated with dots. For Frontal Cortex, n = 7,781 (h3K4me1), 3,509 (H3K4me3), 7,584 (H3K27ac), 3,891 (H3K27me3), 6,560 (H3K9me3), 6,551 (ChromAcc) from dissections of 2 different mice s and n = 35,876 (RNA) from dissections of 4 different mice; for Hippocampus, n = 5,181 (H3K4me1), 3,956 (H3K4me3), 4,165 (H3K27ac), 2,643 (H3K27me3), 5,484 (H3K9me3), 7,544 (ChromAcc) from dissections of 2 different mice and n = 28,973 (RNA) from dissections of 4 different mice. d, Scatter plots showing the proportion of human and mouse RNA reads in each cell (left panel) and the fraction of human reads in DNA and RNA libraries for each cell (right panel) in the species-mixing experiment. Barcodes with less than 80% reads from the same species were identified as mixed cells, the 230 mixed cells of RNA profiles (left panel) were excluded from plotting of the right panel. e, Numbers of unique loci per nucleus for deeply sequenced H3K4me1, H3K4me3, H3K27ac, H3K27me3 and H3K9me3 DNA profiles down-sampled to different levels. f, Numbers of unique loci per nucleus for deeply sequenced Paired-seq DNA profiles down-sampled to different levels. g, Numbers of UMI per nucleus for the deeply sequenced RNA sub-library down-sampled to different levels. For comparison, the numbers of unique loci per cell from the stand-alone high-throughput scChIP-seq assays and the numbers of UMI per cell from scRNA-seq assays were also shown, indicated by dots with labels. h, Violin plots showing the numbers of unique loci mapped per nucleus for all sequenced DNA libraries (average 35k sequenced reads/nuclei with ~40–60% PCR duplication rates). Median numbers, H3K4me1: 5,770 and 5,443, H3K4me3: 1,392 and 1,081, H3K27ac: 1,842 and 1,803, H3K27me3: 904 and 925, H3K9me3: 6,563 and 7,182, chromatin accessibility: 3,170 and 4,381, for frontal cortex and hippocampus, respectively. i, Violin plots showing the fraction of reads inside peaks for different histone marks and brain regions. For Frontal Cortex, n = 7,781 (H3K4me1), 3,509 (H3K4me3), 7,584 (H3K27ac), 3,891 (H3K27me3), 6,560 (H3K9me3), 6,551 (ChromAcc) from dissections of 2 different mice; for Hippocampus, n = 5,181 (H3K4me1), 3,956 (H3K4me3), 4,165 (H3K27ac), 2,643 (H3K27me3), 5,484 (H3K9me3), 7,544 (ChromAcc) from dissections of 2 different mice. j, Violin plots showing the numbers of UMI and genes detected per nucleus for all sequenced RNA libraries (average 30k sequenced reads/nuclei with ~40–60% PCR duplication rates). Median numbers, 4,215 and 3,568 RNA UMI per nucleus for frontal and hippocampus, respectively. k, Violin plots showing the fraction of reads mapped to annotated gene regions (GENCODE GRCm38.p6) and (l) fraction of intronic reads for Paired-Tag RNA datasets and 10X scRNA-seq datasets (10k Brain Cells from an E18 Mouse, V3). n = 35,876 (Frontal Cortex) and 28,973 (Hippocampus) from dissections of 4 different mice. For (h-l), the violin plots were drawn from lower quartile (Q1) to upper quartile (Q3) with the middle line denote the median, whiskers with maximum 1.5 IQR, outliers were indicated with dots.
Extended Data Fig. 3 | Annotation of cell types by Paired-Tag transcriptomic profiles. **a,** UMAP embedding of Paired-Tag transcriptomic profiles and stacked bar plots showing the fraction of cells from different regions or replicates (dissections from different mice) in each cell type. **b,** UMAP embedding of transcriptomic profiles from individual Paired-Tag and Paired-seq datasets. The color of cell types was the same as in Fig.1d. **c,** Dot plots showing the expression of marker genes for each mouse brain cell type measured from Paired-Tag RNA profiles. The size of the dots represents the fraction of cells positively detect the transcripts and the color of the dots represents the average. **d,** UMAPs showing the co-embedding of single-nucleus gene expression from Paired-Tag RNA profiles and the previously published scRNA-seq datasets of the same tissues. **e,** Heatmaps showing the confusion matrices of the overlap between cell type annotations based on Paired-Tag RNA profiles and from the previously published scRNA-seq datasets. The circles left side indicating RNA clusters and were colored according to Fig.1d. **f,** Boxplots showing the Pearson’s correlation coefficients for all genes, variable genes and invariable genes for matched and non-matched cell types between Paired-Tag RNA profiles and the previously published scRNA-seq datasets. The boxes were drawn from lower quartile (Q1) to upper quartile (Q3) with the middle line denote the median, whiskers with maximum 1.5 IQR, outliers were indicated with dots. n = 22 cell types. **g,** Scatter plot showing the expression levels of variable genes in Astrocytes measured by Paired-Tag RNA profiles and the published scRNA-seq datasets.
Extended Data Fig. 4 | Histone marks-based single-cell clustering. **a–e**, UMAP embeddings based on (a) H3K4me1, (b) H3K4me3, (c) H3K27ac, (d) H3K27me3 and (e) H3K9me3 DNA profiles and stacked bar plots showing the fraction of cells from each region or replicate. **f**, UMAP embeddings based on Paired-Tag H3K27ac DNA profiles down-sampled to different sequencing depth (11,749 nuclei, 100–1,500 loci/nuclei). **g**, UMAP embeddings based on Paired-Tag H3K27ac DNA profiles of different numbers of sub-sampled nuclei (median 1,826 loci/nuclei, 200–10,000 nuclei). For visualization, cells were colored according to clustering results from Fig. 1d.
Extended Data Fig. 5 | Gene expression and promoter epigenetic states. a, Violin plots showing reads densities of the five histone marks in Group II-a and III-b promoters. Colors represent cell types the same as in (e). Group II-a promoters were repressed by H3K27me3 in all cell types; genes in III-b were activated by H3K27ac in neuron cells, with comparable H3K27me3 levels in all cell groups. b, Boxplots showing the expression levels of genes grouped by their promoter DNA reads densities for different histone marks. The boxes were drawn from lower quartile (Q1) to upper quartile (Q3) with the middle line denote the median, whiskers with maximum 1.5 IQR, outliers were indicated with dots. n = 2,900 genes for the first 5 groups and n = 2,898 for the 6th group of each histone modification. c, Heatmap showing the Spearman’s correlation coefficients of gene expression and promoter histone modification levels within each cell type. d, Genome browser view of aggregated Paired-Tag profiles showing the three Olfr gene clusters in chr7 was silenced by H3K9me3. e, 3D-scatter plot showing the PCA embedding of aggregated RNA profiles. PC1 differs neuron cells from glial cells and PC2 mainly separates different non-neuron cell types. f, Scatter plot showing the loadings of the first 2 PCs for each gene. Genes from group II-b and III-d were colored in brown and blue, respectively. g, UMAP embedding of 4,659 OPC and Oligodendrocytes nuclei used for pseudotime analysis. h, Expression of marker genes alone the pseudotime.
Extended Data Fig. 6 | Histone modification states in mouse neuron cell types. a, Stacked bar plots showing the fraction of genomic regions for CREs of each group in Fig. 4a. b, Line plots showing the densities of CREs from different groups around CpG islands. c, Scatter plot showing the Spearman’s correlation coefficients of TF motif enrichment and TF gene expression across cell types. TFs with significant positive correlations (FDR < 0.05) between expression and motif enrichment for both H3K4me1 and H3K27ac were colored in red. d, Heatmap showing the emission probability of each histone mark across the 8 chromatin states identified by chromHMM. e, Heatmap showing the fold enrichment of the 8 chromatin states around transcription start sites of FC L2/3 cell cluster.
Extended Data Fig. 7 | Identification of putative CRE-gene pairs. a, Bar charts showing the fraction of predicted H3K27ac- and H3K27me3- associated cCRE-gene pairs supported by the CEMBA datasets. P-value, two-tailed Fisher’s exact test. b–e, Bar charts showing the numbers of cCREs per targeted genes for (b) H3K27ac- and (c) H3K27me3- associated cCRE-gene pairs, and the numbers of predicted targeted genes per cCRE for (d) H3K27ac- and (e) H3K27me3- associated cCRE-gene pairs. f, Representative genome browser view of Gad2 gene locus, both H3K27ac- and H3K27me3-associated cCREs were shown. TSS-proximal region is indicated with green box and cCREs are marked with blue (H3K27ac-specific), brown (H3K27me3-specific) or purple (shared) boxes. g, Top enriched de novo TF motifs and GO terms of cCREs in H3K27ac-specific, shared and H3K27me3-specific pairs. h, Stacked bar plots showing the distribution of genomic regions for cCREs with potential active and repressive functions. P-value, two-tailed Fisher’s exact test. i, j, Bar charts showing the distribution of the distance between cCRE and TSS of predicted target genes from (i) H3K27ac- and (j) H3K27me3- associated cCRE-gene pairs. k, Spearman’s correlations coefficients between reads densities of cCREs and promoters of putative target genes across cell types for H3K4me1 and H3K9me3. P-value, two-tailed Wilcoxon signed-rank test. The boxes were drawn from lower quartile (Q1) to upper quartile (Q3) with the middle line denote the median, whiskers with maximum 1.5 IQR. n = 22 cell types. l, Heatmap showing the histone modification levels at cCREs with potential repressive roles in expression of putative target gene. cCRE were grouped using K-means clustering based on histone modification levels. m, Heatmap showing the expression levels of corresponding putative target genes of cCREs in (l). n, Top enriched Gene Ontology terms for genes in (m) and the top enriched de novo motifs for cCREs from each group in (l).
Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

**Data collection**

Illumina bc12/fast2.

**Data analysis**

Bowtie2(2.2.9), STAR(2.5.3a), HOMER(4.10), SnapATAC(1.0), Seurat(3.0), ChromVAR(3.12), Louvain (iGraph1.2.6), Monocle3(0.2.3)

The custom scripts for the analysis can be found at https://github.com/cexhu/Preserved-Tag

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

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

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

The sequencing data obtained in this study have been deposited to the NCBI Gene Expression Omnibus (GEO) [http://www.ncbi.nlm.nih.gov/geo/] under accession number GSE152050. The processed data can be accessed from the web portal [http://carles.org/PreservedTag]. All other data are available upon request.

CEMBA datasets were available from NEMO [https://nemoanalytics.org] with the accession number of KRIX SCR_016252.

ENCODE [https://www.encodeproject.org/] datasets were downloaded with the accession numbers: H3K4me1 (ENCSCR000APW), H3K27ac (ENCSCR000AOC), H3K27me3 (ENCSCR000DTY), H3K9me3 (ENCSCR000AOQ), DNase-seq (ENCSCR000ZQO).

The other external datasets were downloaded from NCBI Gene Expression Omnibus (GEO) [http://www.ncbi.nlm.nih.gov/geo/], with the accession numbers: SPLIT-seq (GSE110823), CoBATCH (GSE129335), iChIP (GSE109762) and HT-scChiP-seq (GSE117309).
10X scRNA-seq datasets were downloaded from 10x genomics website [https://www.10xgenomics.com/].

Field-specific reporting

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

- Life sciences  □ Behavioural & social sciences  □ Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

| Sample size | Sample size was determined based on prior published data from similar experiments (Rosenberg et al., Science, 2018; Cao et al., Science, 2018, Preiss et al., Nat. Neuroscience, 2018): to evaluate the robustness of method, tissues dissected from at least two individuals are required for each experiment. |

| Data exclusions | Low quality single nuclei (low number of reads, low number of genes captured) were excluded from downstream analysis as outlined in the Methods section. Genes (Gencode GRCm38.65) with sufficient levels of transcription (RPKM > 1) or promoter occupancy (CPM > 1 for histone marks in at least one cell group) were retained for subsequent analysis. The cutoff is set by reads number distribution of cells (minimal and maximal 5% of the populations) and the minimal required number for analyses (200 features for single-cell clustering and CPM > 1 for gene expression analyses). |

| Replication | Histone modification datasets were obtained from dissections of two different mice and RNA datasets were obtained from dissections of four different mice. All datasets from independent experiments have similar results. |

| Randomization | Allocation was random. |

| Blinding | The experiments were not blinded as the identities of histone modification and tissue dissection are needed for evaluating the sensitivities and specificities of the method. The clustering of single-nuclei were unsupervised and annotated afterwards. |

Reporting for specific materials, systems and methods

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

Materials & experimental systems

- n/a
- Involved in the study

- Antibodies
- Eukaryotic cell lines
- Palaeontology and archaeology
- Animals and other organisms
- Human research participants
- Clinical data
- Dual use research of concern

Methods

- n/a
- Involved in the study

- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Antibodies

- Antibodies used: H3K4me1 (Abcam, ab8895, 1:25), H3K4me3 (Abcam, ab8580, 1:25), H3K27ac (Abcam, ab4729, 1:25), H3K27me3 (Active Motifs, 39035, 1:25), H3K9me3 (Abcam, ab898, 1:25).

- Validation: The specificities of the four antibodies were compared to published ENCODE ChIP-seq data from the same cell lines (HeLa S1, H3K4me1 (ENC5000010AF), H3K4me3 (ENC5000010OF), H3K27ac (ENC5000010AC), H3K27me3 (ENC5000010TY), H3K9me3 (ENC5000010QO)). Validation of the antibodies can be accessed from manufacturer’s website: H3K4me1, Specific for mono methylated Lysine 4 of histone H3. Does not recognize di- or tri methyl Lysine 4 nor methylation at Lysine 9. Reacts with: Cow, Human and predicted to work with: Mouse, Pig, Saccharomyces cerevisiae, Tetrahymena, Xenopus laevis, Drosophila melanogaster, Plants, Mammals, Plasmodium falciparum, Xenopus tropicalis, Candida albicans (https://www.abcam.com/histone-h3-mono-methyl-k4-antibody-chip-grade-ab8895.html); H3K4me3, Immunogen is synthetic peptide within Human Histone H3 aa 1-100 (tri methyl K4) conjugated to keyhole limpet haemocyanin. Reacts with: Cow, Human and predicted to work with: Mouse, Rat, Rabbit, Pig, Saccharomyces cerevisiae, Tetrahymena, Xenopus laevis, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Indian muntjac, Oikopleura.
Plants, Zebrasfish, Mammals, Trypanosoma cruzi, Common marmoset, Rice, Xenopus tropicalis (https://www.abcam.com/histone-h3-tri-methyl-k4-antibody-chip-grade-ab8580.html);

H3K27ac, immunogen is synthetic peptide corresponding to Human Histone H3 aa 1-100 (acetyl K27) conjugated to keyhole limpet haemocyanin. Reacts with: Mouse, Cow, Human and predicted to work with: Rat, Chicken, Xenopus laevis, Arabidopsis thaliana, Drosophila melanogaster, Monkey, Zebrasfish, Plasmodium falciparum, Rice, Cyanidioschyzon merolae (https://www.abcam.com/histone-h3-acetyl-k27-antibody-chip-grade-ab4729.html);

H3K27me3, This Histone H3 trimethyl lys27 antibody was raised against a peptide including trimethyl-lysine 27 of histone H3. Reactivity: Human, Mouse, Wide Range Predicted (https://www.activemotif.com/catalog/details/39155/histone-h3-trimethyl-lys27-antibody-pub);

H3K9me3, Histone H3 (tri methyl K9) antibody (ab8898) is specific for Histone H3 tri methyl Lysine 9. Shows slight cross-reactivity with tri methyl K27, which shares a similar epitope (please see Western blot image). Does not react with mono or di methylated K9. Reacts with: Mouse, Rat, Human, Saccharomyces cerevisiae, Indian muntjac and predicted to work with: Chicken, Xenopus laevis, Drosophila melanogaster, Mammals, Xenopus tropicalis, Cyanidioschyzon merolae (https://www.abcam.com/histone-h3-tri-methyl-k9-antibody-chip-grade-ab8898.html).

### Eukaryotic cell lines

Policy information about cell lines

| Cell line source(s) | HeLa S3 cells from ATCC (CCL-2.2). |
|---------------------|-----------------------------------|
| Authentication      | Cells were not authenticated.     |
| Mycoplasma contamination | Cells were not tested for mycoplasma. |
| Commonly misidentified lines (See iCLAC register) | None of the cell lines used are listed in the iCLAC database. |

### Animals and other organisms

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

| Laboratory animals | Male C57BL/6J mice were purchased from Jackson Laboratories at 8 weeks of age and maintained in the Salk animal barrier facility on 12-hr dark-light cycles, under the temperature of 20-22 °C and humidity of 30-70 % with food ad libitum for four weeks before dissection. The frontal cortex and hippocampus were dissected and snap-frozen in dry ice. All protocols were approved by the Salk Institute's Institutional Animal Care and Use Committee (IACUC). |
|--------------------|--------------------------------------------------------------------------------|
| Wild animals       | The study did not involve wild animals.                                         |
| Field-collected samples | No commonly misidentified cell lines were used.                              |
| Ethics oversight   | All protocols were approved by the Salk Institute's Institutional Animal Care and Use Committee (IACUC). |

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