


METHOD

Open Access



txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility

Hao Zhang^{1,2†}, Ryan M. Mulqueen^{3†}, Natalie Iannuzo¹, Dominique O. Farrera⁴, Francesca Polverino^{2,5,6}, James J. Galligan⁴, Julie G. Ledford^{1,2}, Andrew C. Adey^{3,7,8,9*} and Darren A. Cusanovich^{1,2*} 

[†]Hao Zhang and Ryan M. Mulqueen contributed equally to this work.

*Correspondence: adey@ohsu.edu; darrenc@arizona.edu

¹ Department of Cellular and Molecular Medicine, University of Arizona, Tucson, AZ, USA

² Asthma & Airway Disease Research Center, University of Arizona, Tucson, AZ, USA

³ Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA

⁴ Department of Pharmacology and Toxicology, University of Arizona, Tucson, AZ, USA

⁵ Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, University of Arizona, Tucson, AZ, USA

⁶ Banner - University Medicine North, Pulmonary - Clinic F, Tucson, AZ, USA

⁷ Cancer Early Detection Advanced Research Center, Oregon Health & Science University, Portland, OR, USA

⁸ Oregon Health & Science University, Knight Cancer Institute, Portland, OR, USA

⁹ Oregon Health & Science University, Knight Cardiovascular Institute, Portland, OR, USA

Abstract

We develop a large-scale single-cell ATAC-seq method by combining Tn5-based pre-indexing with 10x Genomics barcoding, enabling the indexing of up to 200,000 nuclei across multiple samples in a single reaction. We profile 449,953 nuclei across diverse tissues, including the human cortex, mouse brain, human lung, mouse lung, mouse liver, and lung tissue from a club cell secretory protein knockout (CC16^{-/-}) model. Our study of CC16^{-/-} nuclei uncovers previously underappreciated technical artifacts derived from remnant 129 mouse strain genetic material, which cause profound cell-type-specific changes in regulatory elements near many genes, thereby confounding the interpretation of this commonly referenced mouse model.

Keywords: Single-cell, ATAC-seq, Chromatin accessibility, Combinatorial indexing, Molecular hashing, CC16

Background

Chromatin accessibility measurement has become a widely used method to understand gene regulation and identify cell types and states. A common technique is the “assay for transposase-accessible chromatin using sequencing” (ATAC-seq) [1], in which a hyperactive mutant of the Tn5 transposase inserts sequencing adapters into sterically open (“accessible”) regions of chromatin. After mapping the locations of these insertions, the resulting pile-up of genome-aligned reads identifies loci that are putatively active in gene regulation [1]. Performed at single-cell resolution (scATAC-seq), this assay has generated catalogs of genome-wide DNA regulatory sites, dynamic chromatin reorganization through development [2], and whole organism cell atlases [2, 3].

Most modern single-cell methods generate data on hundreds to thousands of cells in parallel to enable proper characterization of heterogeneous or dynamic cellular systems. Two general strategies have been developed to generate data at this scale. First, cells can be isolated into individual reaction vessels—plate wells, micro-wells, or droplets. This has most commonly been implemented with microfluidics



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

platforms, such as the commercialized products of 10× Genomics [4]. Second, iterative split-pool barcoding, as is seen in “single-cell combinatorial indexing” (sci) strategies, can index single cells while never isolating individual cells during the molecular reactions [5–7]. However, choosing one of these two approaches requires researchers to accept tradeoffs in terms of throughput and data quality. Microfluidic approaches generally have superior data quality, while combinatorial indexing benefits from flexibility, increased scalability, and cost efficiencies.

One strategy to boost the scalability of microfluidic approaches has been to “pre-index” cells or nuclei before loading them on a microfluidic device. In this way, aliquots of cells/nuclei are provided with a specific cellular/nuclear barcode via one of a variety of strategies and then aliquots are pooled before loading on a microfluidic device. The pre-index can be used along with the droplet barcode to deconvolute individual cells at the data analysis stage. This allows multiple samples to be processed in parallel and can enable some “overloading” of the droplets. For example, a single-nucleus barcoding approach (SnuBar) [8] was previously demonstrated to allow for pre-indexing of nuclei in a scATAC-seq approach. However, individual molecules are not labeled in this strategy and thus droplets with multiple nuclei could not be discriminated, somewhat limiting the overall throughput. In another approach, a chimeric single-cell method combining a droplet-microfluidic system with molecular-level pre-indexing (called “dsciATAC-seq”) was previously developed, which improved the throughput of the microfluidic platform without sacrificing the data quality [9]. In this case, because the pre-indexing occurs at the molecular level (rather than the nuclear level), droplets containing multiple nuclei can still be computationally deconvoluted. However, this technique was formulated using the BioRad system, which requires an overloading of beads to reduce the frequency of empty droplets due to the design of the beads [10, 11]. This bead overloading strategy necessitates the consolidation of bead multiplets using computational inference and therefore may introduce technical artifacts. In contrast, 10× Genomics, another commercialized platform widely embraced for single-cell data generation, takes advantage of deformable hydrogel beads that can be closely packed in the microfluidic channels, which allows ~80% loading efficiency of a single bead per droplet without overloading beads [12]. Here we demonstrate a method that combines 96-well plate-indexed tagmentation with 10× Gel Bead-In EMulsions (GEM) barcoding to substantially improve the throughput of the 10× platform by overloading nuclei and enabling the multiplexing of up to 96 samples in a single reaction (Fig. 1a). We call this method 10×-compatible (or TenX-compatible) Combinatorial Indexing ATAC-seq (txci-ATAC-seq). We use this strategy to generate up to 200,000 cells in a single 10× reaction (~22-fold increase in cell throughput as compared to the standard 10× Chromium scATAC-seq at a constant collision rate) and apply it to study the heterogeneity of chromatin accessibility in five primary samples, including human and mouse brain, human and mouse lung, and mouse liver, demonstrating the robustness of this approach. The scalability and flexibility of txci-ATAC-seq make it suitable for single-cell atlas efforts, population-scale studies, and experiments implementing replicates and proper study design.

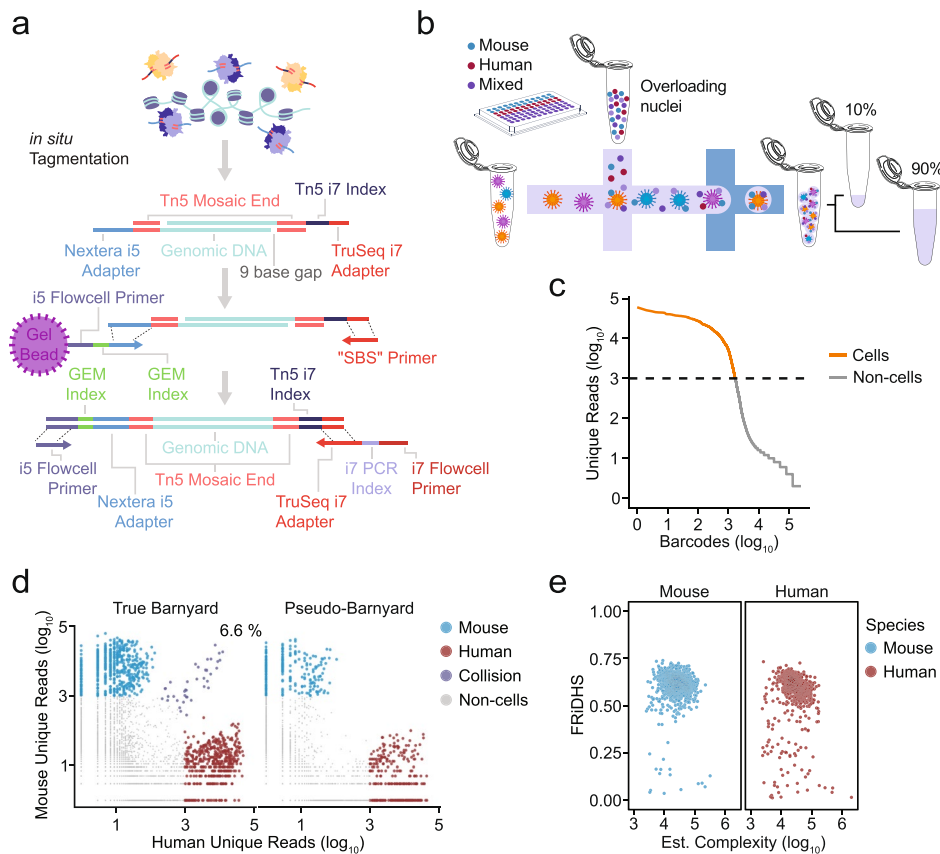


Fig. 1 txcI-ATAC-seq generates high-quality single-cell ATAC libraries at high throughput. **a** Schematic of molecular details of txcI-ATAC-seq library generation. **b** Experimental workflow for txcI-ATAC-seq barnyard library generation. After 96-plex tagmentation, nuclei are overloaded on a 10X Chromium microfluidics device. Following nucleus encapsulation in the formed droplets, 10% of the GEMs can be used for quality control and the remaining 90% for data analysis. **c-e** txcI-ATAC-seq QC metrics for human (GM12878) and mouse (CH12) cell lines supplemented with SBS primer during in-droplet PCR. **c** “Knee” plot showing the unique reads (\log_{10} scale) against the rank of each barcode (\log_{10} scale) ordered from most unique reads (left) to least (right). The dashed line indicates the threshold (1000 reads) used to identify cell barcodes (orange points). **d** Scatter plots showing the number of unique reads mapped to either the human or mouse genome for both true and pseudo-barnyard experiments. Values were \log_{10} -transformed after adding a pseudo-count of 1 to all values. The percentage shown in the true barnyard panel (6.6%) represents the estimated collision rate. **e** Scatter plots showing the FRIDHS against the estimated complexity for each cell barcode detected as either mouse (blue) or human (red) cell. The estimated complexity is shown on a \log_{10} scale

Results

Coupling droplet-based microfluidics with indexed transposition enables the overloading of nuclei

In order to implement a strategy analogous to dsciATAC-seq [9] on the 10× platform, we first conducted a pilot experiment, tagging nuclei using 96 barcoded Tn5 reactions (similar to our previous sci-ATAC-seq workflows [5]) followed by pooling all nuclei and processing samples through a largely unmodified standard 10× workflow (except that we overloaded the sample with 75,000 nuclei in a lane instead of the recommended 15,300 maximum capacity). The single-cell resolution and the degree of barcode collisions (i.e., instances where one barcode represents the contents of two or more cells) were evaluated using a “barnyard” experiment in which we mixed human and mouse nuclei—using

either cell lines (human GM12878 nuclei mixed with mouse CH12.LX nuclei) or tissues (human lung nuclei mixed with mouse lung nuclei). Two mixing strategies were designed on the same 96-well plate: the nuclei from the two species were either pooled during tagmentation (“true barnyard”, which was used to reflect the rate of detected collisions caused by both pre- and post-pooling events) or after the tagmentation reaction (“pseudo-barnyard”, which was used to reflect the rate of detected collisions caused by post-pooling events only) (Additional file 1: Fig. S1a). A mixed species experiment such as this (Fig. 1b) allows for an accurate estimation of collision rate since each index is expected to align uniquely to either the human or mouse reference genome. Indexes with cross-alignment indicate collisions and allow us to empirically scale cells loaded during droplet formation. The tagmentation reactions were performed with either a modified version of the “Omni” ATAC-seq protocol [13] or the 10× protocol (see “Methods”). After performing indexed tagmentation on a 96-well plate and pooling all nuclei, 75,000 nuclei from the pool were loaded onto a single 10× lane. The sample and cell-specific information of the resulting libraries was deconvoluted using the combination of three barcodes introduced during the workflow: a PCR barcode (i7) used to distinguish different lanes of the 10×, a GEM barcode introduced in the droplet, and a Tn5 barcode introduced during tagmentation (Fig. 1a). To account for the collisions originating from the same species, all the collision rates reported in this study were estimated using the equation described in [14]. Unexpectedly, regardless of barnyard type (true vs pseudo), the initial experiment exhibited an extremely high collision rate, i.e., 46.0% in a true-barnyard experiment mixing two cell lines, 44.4% in a pseudo-barnyard of cell lines, and 40.1% in a true barnyard mixing lung tissues (Additional file 1: Fig. S1b). We also tested a second tagmentation buffer (provided in the 10× kit), but obtained similar results (47.4% estimated collision rate with a true barnyard of cell lines). However, limiting our measurement to GEMs with a single-Tn5 barcode demonstrated a remarkably reduced collision rate across all tested samples and buffers (4.7%, 3.3%, 4.4%, and 8.6% for the true barnyard of cell lines, pseudo-barnyard of cell lines, true barnyard with 10× buffer, and true barnyard with lung tissue, respectively). These results suggested that most multi-plets were not arising from pre-pooling events but instead were a consequence of cross-contamination due to Tn5 barcode swapping within droplets (Additional file 1: Fig. S1c).

We tested three different strategies to eliminate this apparent in-droplet barcode-swapping (Additional file 1: Fig. S1d; see [Methods](#) for details): (1) adding a second round of tagmentation with an additional (unamplifiable) duplex DNA prior to pooling to exhaust excess Tn5 (“Decoy DNA”); (2) supplementing the GEM reaction with a blocking oligo containing a reverse complement sequence of the Tn5 adapter and an inverted dideoxythymidine (“dT”) at the 3’ end to inhibit the use of free Tn5 adapters as amplification primers (“Blocking oligo”); or (3) supplementing the GEM reaction with a reverse primer to enable exponential amplification instead of linear amplification in the droplet PCR (“SBS primer”) with the goal of outcompeting barcode-swapping. To facilitate better optimization of experiments in overloaded droplets without imposing a significant burden of sequencing for each condition tested, we also developed a method to sample a subset of droplets after in-droplet amplification. To do so, we took 10% of the volume of droplets immediately after amplification (but before breaking the droplets) and processed both the 10% sample and 90% in parallel (Fig. 1b). In this way, we could

first sequence 10% of the loaded cells to evaluate data quality and subsequently sequence the remaining 90% if warranted. We then tested all three strategies head-to-head (Additional file 1: Fig. S2a) and used a conservative cutoff of 1000 reads to identify cells for all conditions (Fig. 1c and Additional file 1: Fig. S2b). While all three tested strategies mitigated some of the barcode swapping, we found that the SBS primer was most efficient—reducing the estimated collision rate of cell lines from 46.0% to 6.6% in the true barnyard and resulting in no collision cells observed in the pseudo-barnyard wells (Fig. 1d and Additional file 1: Fig. S2c). Similar results were also seen in the lung barnyard (Additional file 1: Fig. S2c), with a collision rate of 11.1% for the true barnyard and only a single collision observed in the pseudo-barnyard when spiking in the SBS primer. We also used the fraction of reads mapping to the ENCODE-defined DNase I hypersensitive sites (FRiDHS) and the estimated library complexity (see “Methods” for calculations) to evaluate the performance across all three blocking conditions. Considering the data generated for cell lines, we found that the SBS primer provided the highest FRiDHS scores (a median of 61.5% for mouse cells and 60.3% for human cells, Fig. 1e and Additional file 1: Fig. S2d) and a comparable complexity (a median of 25,504.1 for mouse and 27,298.6 for human, Fig. 1e) with Decoy DNA but a higher complexity than Blocking oligo (Additional file 1: Fig. S2e). Coherent trends were also observed in the lung tissues (Additional file 1: Fig. S2d,e). Interestingly, the SBS primer strategy also caused a shift in the fragment size distribution relative to the other conditions, indicating the exponential amplification of GEM reactions is biased toward small fragments given the same number of amplification cycles (Additional file 1: Fig. S2f). A reduced number of cycles in droplet PCR, however, can partially recover the large fragment sizes (Additional file 1: Fig. S2g). Nonetheless, by optimizing this hybrid protocol of barcoded transposition followed by GEM amplification, we successfully developed a novel protocol that enables multiplexing of multiple samples and unbiased profiling of chromatin accessibility at extremely high throughput on the 10× Genomics platform. Having established a working protocol, we next sought to apply it to complex tissues to evaluate the assay’s performance. Below, we described the results from five primary samples.

Profiling chromatin accessibility of human and mouse brain tissue

To evaluate the performance of txci-ATAC-seq in complex tissues, we initially generated chromatin accessibility profiles for human cortex and mouse whole brain samples using a true-barnyard scheme with two separate experiments to test nuclei inputs of 25,000 (~1.5X the maximum recommended input) and 75,000 (~4.5X the maximum recommended input) on the microfluidic device (see “Methods”). Libraries were sequenced to an average depth of 45,622 unique reads per cell, with an estimated saturation rate of 60.3% unique reads (Additional file 1: Fig. S3a). We observed an estimated collision rate of 0.6% and 1.3% in the 25,000 and 75,000 inputs, respectively (Additional file 1: Fig. S3b), which resulted in a ~24-fold increase in the throughput of a standard 10× workflow at a comparable collision rate. A majority of droplet barcodes were assigned to a single 10× nucleus barcode, with 78.94% and 60.38% of droplets containing a single nucleus for 25,000 and 75,000 nuclei loadings, respectively (Additional file 1: Fig. S3c). Overall, we captured 17,257 and 61,171 cells for the 25,000 and 75,000 nuclei loadings, respectively (Additional file 1: Fig. S3d). To understand sample complexity, dimensionality

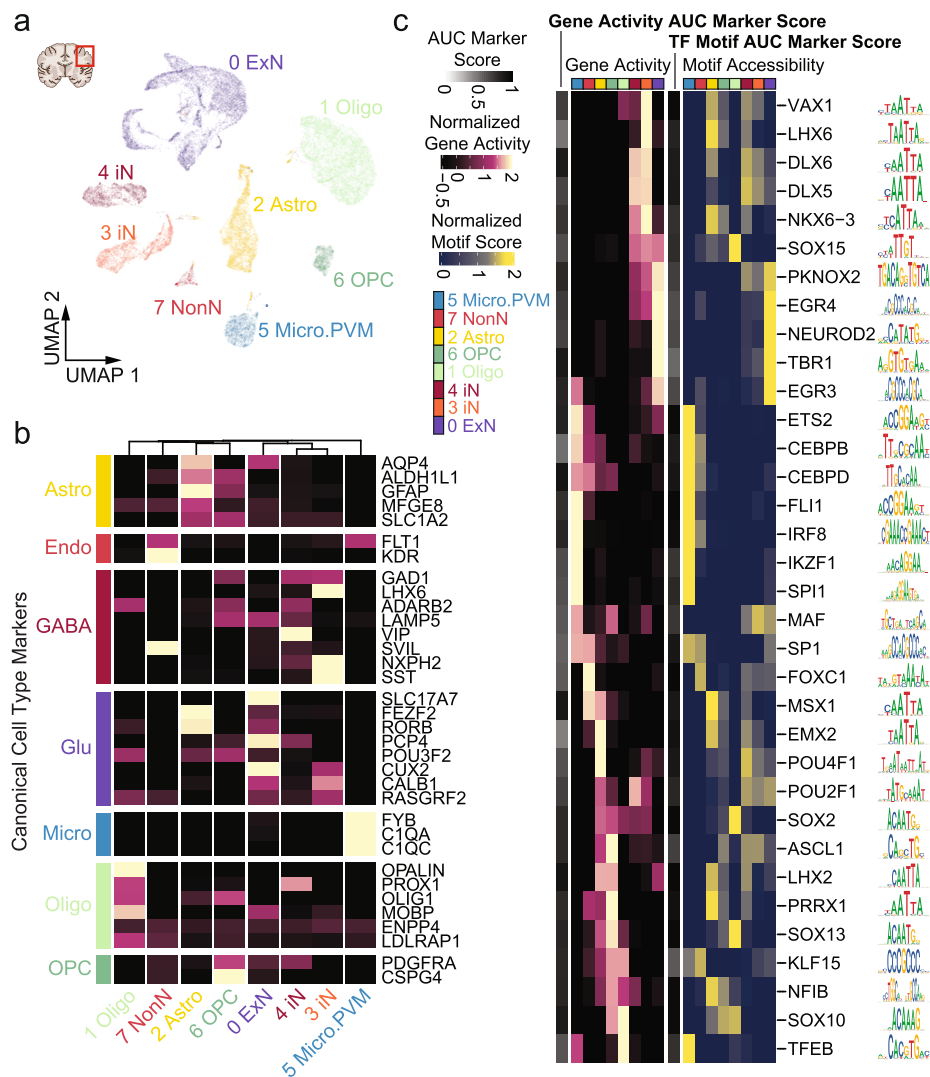


Fig. 2 Cell type identification and marker assessment in human cortex sample. **a** UMAP projection of human cortex nuclei ($n = 28,663$). Nuclei are colored by their predicted cell type. **b** Heatmap of z-scored average gene activity score per cluster for canonical markers from Brain Map datasets. Astro: astrocytes; Endo: endothelial cells; ExN: excitatory neurons; GABA: GABAergic; Glu: Glutamatergic; iN: inhibitory neurons; Micro: microglia; Micro.PVM: microglia and perivascular macrophages; NonN: Non-neuronal; Oligo: oligodendrocytes; OPC: oligodendrocyte progenitor cells. **c** De novo determination of TF marker genes through chromatin accessibility-derived gene activity (left) and TF motif usage (right). Z-scored average gene activity score and TF motif usage per cluster are plotted for the top 5 markers within each cluster. TF markers are ranked by AUC reported from one vs. rest Wilcoxon rank sum test. TF motifs are shown on the right as SeqLogos alongside heatmap rows

reduction [15] and clustering [16] were performed on the human (Fig. 2a) and mouse (Fig. 3a) cells separately. We also identified and removed cryptic doublets within species to filter out the barcode collisions passing our initial species alignment filter (“Methods”) [17]. We generated gene activity scores (akin to a surrogate for gene expression) using *cis*-co-accessibility networks (CCANs) anchored on promoter regions [18]. A label-transfer algorithm then assigned cell types in comparison to published RNA datasets [19–21]. The high percentage of cells assigned to the same RNA-defined cell type

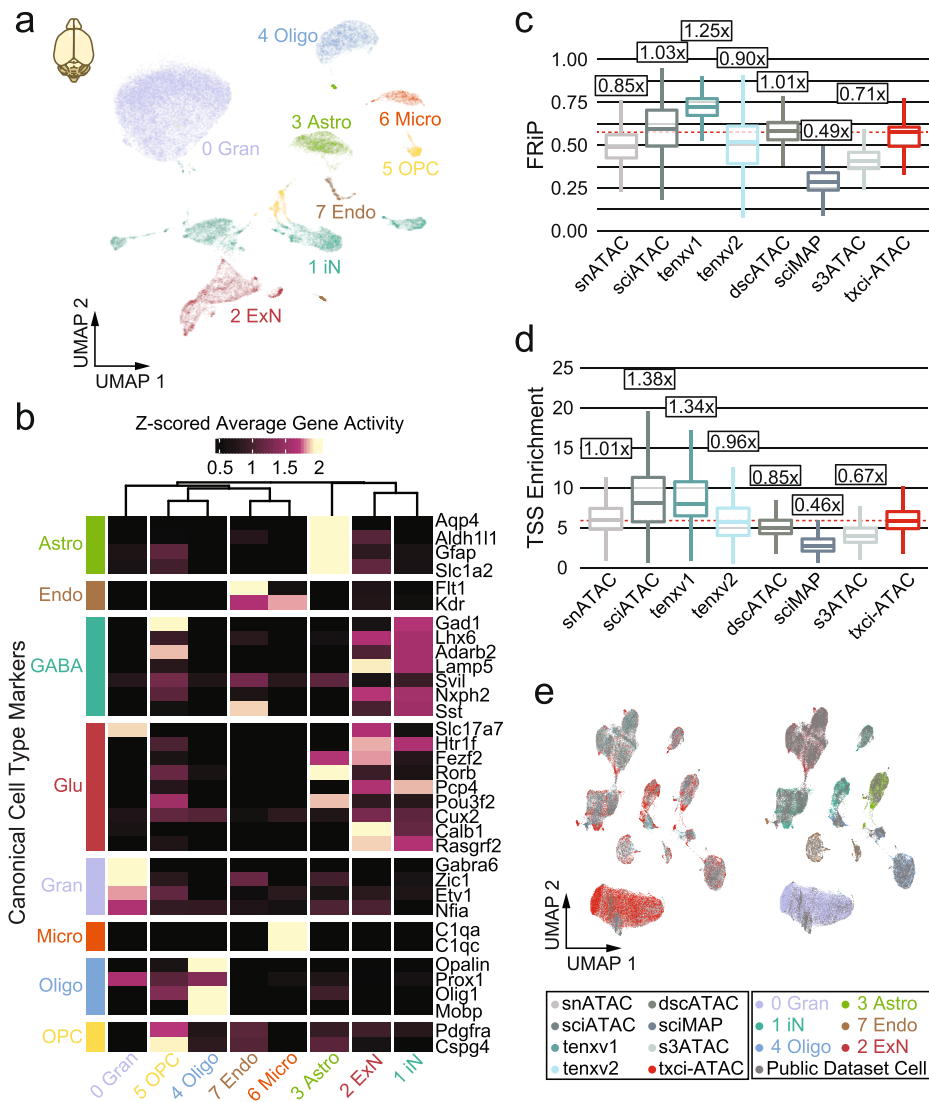


Fig. 3 Cell type identification and marker assessment in mouse whole brain sample. **a** UMAP projection of mouse brain nuclei ($n = 49,765$). Nuclei are colored by their predicted cell type. **b** Z-scored average gene activity score per cluster plotted as a heatmap. Clusters are arranged by hierarchical clustering. Marker sets are from Brain Map marker genes. **c** Boxplots of FRIP per technology using a unified peak set. Numbers over the boxplot reflect the fold-change of medians in comparison to txci-ATAC-seq. Stippled line is the median value for txci-ATAC-seq. **d** Boxplots of transcription start site (TSS) enrichment across technologies. Numbers over the boxplot reflect the fold-change of medians in comparison to txci-ATAC-seq. Stippled line is the median value for txci-ATAC-seq. **e** LIGER integrated UMAP projection of technologies ($n = 75,845$ cells) colored by technology (left) and cell type (right)

per cluster supported the specificity of the label-transfer approach (Additional file 1: Fig. S3e,f). We corroborated the assigned labels by examining the cluster-wise mean gene activity scores for canonical RNA markers of cell types (Figs. 2b and 3b) [20, 21]. We next sought to define marker transcription factors (TFs) per cluster de novo by implementing an average “area under the curve” (AUC) value [22] across both gene activity and motif accessibility [23] scores in the human cortex (Fig. 2c). This approach allows for either gene activity or motif accessibility to be informative. For example, we found that

the two human inhibitory neuron clusters could be distinguished by gene activity of LIM Homeobox 6 (LHX6), while motif usage differences between them were not significant and the motif is most accessible in astrocytes. In this case, the lack of distinction in motif usage is likely driven by other TFs of the LIM family that share a very similar motif, such as LIM Homeobox 2 (LHX2).

Since the mouse brain is a commonly profiled benchmark tissue of scATAC-seq methods, we compared our data to publicly available datasets for combinatorial indexing (snATAC-seq [24], sci-ATAC-seq [25], sci-MAP-seq [25], and s3-ATAC-seq [26]) and droplet-based (dscATAC-seq [9], 10× scATAC-seq v1 [27], and v2 [28]) chemistries. With all datasets merged, we uncovered a unified peak set of 344,258 features of open chromatin in the mouse brain. txci-ATAC-seq performed comparably to the other technologies in terms of the fraction of reads in peaks (FRiP) (Fig. 3c) and transcription start site (TSS) enrichment at the level of individual cells (Fig. 3d), demonstrating the fourth-best FRiP (out of 8) and the fourth-best TSS enrichment. Notably, we observed that txci-ATAC recovered a full spectrum of insert sizes in brain samples compared to the other techniques (Additional file 1: Fig. S3g). Estimating unique reads given a constant sequencing depth per cell (Additional file 1: Fig. S3h), we noted that txci-ATAC-seq fell between the high-content ATAC-seq preparations (such as 10× scATAC-seq v2 chemistry or s3-ATAC-seq) and combinatorial methods (like snATAC-seq and sci-ATAC-seq). Also, txci-ATAC-seq integrated readily with other technologies on the unified peak set, with the exception of a notable increase in granule cells (Fig. 3e), potentially reflecting a higher concentration of cerebellum tissue during initial brain dissociation. We found that the genomic coverage of ATAC-seq fragments across a pan-excitatory neuronal marker *Slc17a7* showed a similar distribution across compared technologies and our own (Additional file 1: Fig. S4a). We further validated our technology's ability to capture reproducible chromatin accessibility patterns within the mouse brain by calculating the average gene activity score per gene and correlating our dataset to existing technologies via Spearman correlation (Additional file 1: Fig. S4b). Our results revealed that all compared techniques generated consistent signals across genes (Spearman's $\rho \geq 0.86$ and p -value ≤ 0.001 for all comparisons). Overall, txci-ATAC-seq enabled detailed epigenomic characterization of cell types in brain tissues, including the de novo definition of marker TFs by leveraging a combination of gene activity and TF motif usage. In tissue-matched comparisons across technologies, we found that txci-ATAC-seq performed equivalently in quality control metrics of library complexity and ATAC signals.

Profiling chromatin accessibility of lung and liver tissue

To test the robustness of this strategy in different biological contexts, we multiplexed mouse lung and liver samples on a single 96-well plate with two replicates for each tissue (Fig. 4a). The last two rows of the plate were set up as a true-barnyard design by mixing mouse nuclei with human lung nuclei to estimate the internal collision rate for each sample. Two loading inputs (100,000 and 200,000 nuclei per lane) were tested and sequenced separately. Using a conservative cutoff of 1000 reads to define a bona fide cell barcode (Additional file 1: Fig. S5a,b), we recovered 67,251 (67.3%) and 104,987 (52.5%) nuclei from the 100,000 and 200,000 inputs, respectively (Additional file 1: Fig. S5c). Since these libraries were sequenced to an average depth of 6418.9 and 4014.8 unique

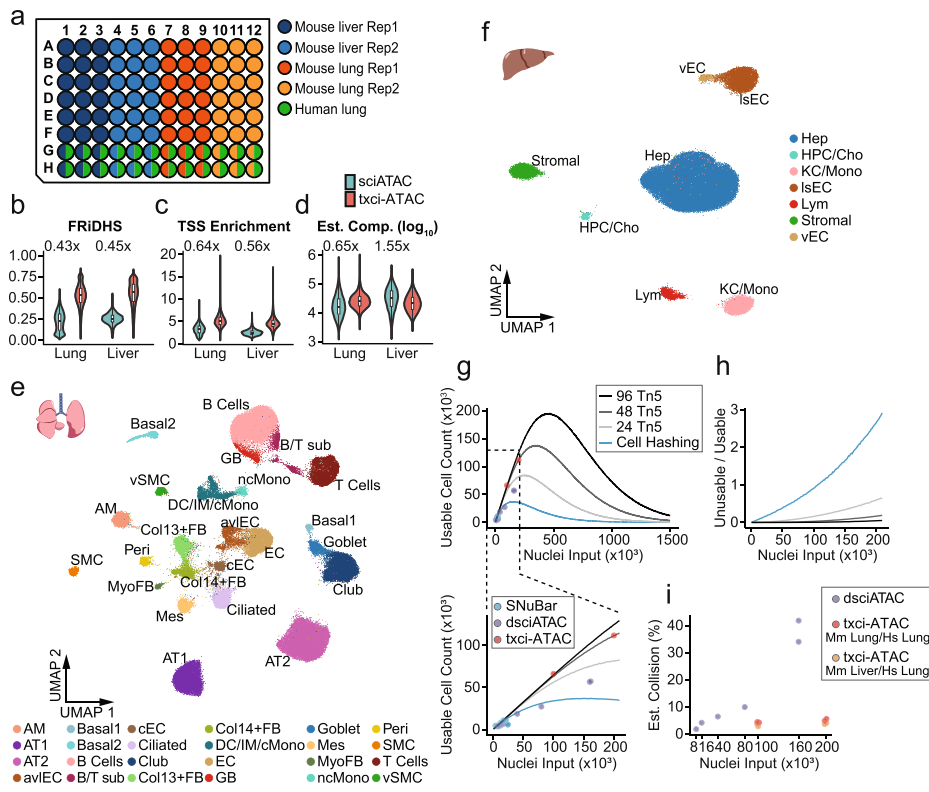


Fig. 4 txcI-ATAC-seq generates high-quality single-cell ATAC-seq data on multiple tissues in parallel at scale. **a** Well assignment showing the multiplexing of primary samples. Rows 7 and 8 provide an estimate of the empirical collision rate for each sample by mixing human lung nuclei with mouse nuclei isolated from each corresponding tissue. **b-d** The comparison of quality metrics between sciATAC-seq and txcI-ATAC-seq for each cell in mouse lung and liver tissue. The **(b)** FRiDHS, **(c)** TSS enrichment score, and **(d)** estimated complexity (on a \log_{10} scale) indicate the performance of single-cell ATAC-seq methods. The numbers over the violin plots reflect the fold-change in median compared to txcI-ATAC-seq. **e** UMAP visualization of mouse lung nuclei ($n = 73,280$) integrating two replicates across two loading inputs. Nuclei are colored by their predicted cell type. **f** UMAP visualization of mouse liver nuclei ($n = 63,429$) integrating two replicates across two loading inputs. Abbreviations: AM, alveolar macrophages; AT1, alveolar type 1 epithelial cells; AT2, alveolar type 2 epithelial cells; avIEC, arterial/venous/lymphatic endothelial cells; B/T sub, B and T cell subpopulation; cEC, capillary endothelial cells; Col13 + FB, collagen type XIII α 1 chain positive fibroblasts; Col14 + FB, collagen type XIV α 1 chain positive fibroblasts; DC/IM/cMono, dendritic cells/interstitial macrophages/classical monocytes; EC, endothelial cells; GB, germinal B cells; Hep, hepatocytes; HPC/Cho, hepatic progenitor cells/cholangiocytes; KC/Mono, Kupffer cells/monocytes; IsEC, liver sinusoidal endothelial cells; Lym, lymphocytes; Mes, mesothelial cells; MyoFB, myofibroblasts; ncMono, nonclassical monocytes; Peri, pericytes; SMC, smooth muscle cells, vEC, venous endothelial cells, vSMC, vascular smooth muscle cells. **g** Theoretical and empirical number of deconvolutable cells recovered across various nuclei loading inputs for molecular and cellular hashing strategies. The simulated cell recovery for either molecular hashing at different numbers of Tn5 barcodes (black and gray lines) or cellular hashing (blue line) strategies were compared with the observed cell recovery obtained from the txcI-ATAC (indexed with 96 Tn5 barcodes, red dots), dsciATAC (indexed with 48 Tn5 barcodes, slate blue dots), and SNUBar (blue dots) datasets. For txcI-ATAC and dsciATAC, the data were processed using the same pipeline, and the cell threshold was determined using *K*-means clustering. For SNUBar, the recovery values were taken directly from the original paper. The lower panel zooms in on the range highlighted by the dashed box in the upper panel. **h** Unusable-to-usable cell ratio derived from the simulated data presented in panel **(g)**. Unusable refers to cells that had to be discarded because they were in unresolvable multiplets. Usable cells include singlets and cells that can be unambiguously demultiplexed from multiplets for molecular hashing strategies. The colors are consistent with panel **(g)**. **i** Estimated collision rate calculated from the cells determined by *K*-means clustering for both txcI-ATAC and dsciATAC datasets across various loading inputs. A mixture of human K562 and murine 3T3 cells was used to identify multiplets in the dsciATAC datasets (slate blue). For txcI-ATAC, the collision rate was evaluated using either a mixture of mouse and human lung cells (red) or a mixture of mouse liver and human lung cells (orange). The data points from the txcI-ATAC datasets were jittered to enhance the visibility of individual values

reads per cell for the 100,000 and 200,000 input libraries, respectively (21.4% and 14.7% saturated, Additional file 1: Fig. S5d,e), the slightly lower recovery rate observed for the 200,000 nuclei input may be due to the lower per-cell sequencing depth resulting in some likely cells failing to pass the read depth threshold (Additional file 1: Fig. S5b,e). Collision rate estimates showed that pushing the loading throughput from 100,000 to 200,000 nuclei only raised the average rate from 3.6% to 4.4% (Additional file 1: Fig. S5f). Overall, these libraries increased the yield of usable nuclei by nearly 22-fold in comparison to the standard 10× Chromium scATAC-seq at the same collision rate. While the collision rate appeared to be tissue-dependent within this experiment (with an average of 4.6% for lung and 3.4% for liver), the fold increase in the number of cells that could be processed at a 10×-equivalent collision rate aligned well with what we observed in brain tissues. In addition, we again compared a series of quality metrics between our txci-ATAC-seq data and previously obtained sci-ATAC-seq data on the same tissues [3] and demonstrated that the data generated with txci-ATAC-seq had a substantially higher quality than the original combinatorial indexing assay: the median FRiDHS increased from 22.8% to 53.0% for lung and from 25.5% to 56.5% for liver (Fig. 4b); the median TSS enrichment score increased from 3.2 to 5.1 for lung and from 2.5 to 4.5 for liver (Fig. 4c); the aggregated TSS enrichment increased from 10.7 to 22.2 for lung and from 7 to 18.2 for liver (Additional file 1: Fig. S5g); the median complexity increased from 16,472.2 to 25,338.4 for lung while it decreased from 33,123.4 to 21,362.2 for liver (Fig. 4d). After filtering out low-quality nuclei and putative doublets (see [Methods](#)), we generated chromatin accessibility profiles for 152,508 primary cells, including 73,280 mouse lung nuclei, 63,429 mouse liver nuclei, and 15,799 human lung nuclei (59,348 of the nuclei recovered from the 100,000 input library and 93,160 of the nuclei recovered from the 200,000 input library).

To dissect the diverse chromatin landscapes present in these heterogeneous tissues, we performed an iterative peak calling and clustering method to parse out the distinct cell populations. In brief, we called peaks on aggregated reads for all cells, scored individual cells for insertion events in these reference peaks, and then carried out dimensionality reduction and cluster identification using Seurat [29]. A second round of peak calling was performed on cells from each cluster separately, and the peaks identified for all clusters were then merged and used as a reference set to perform dimensionality reduction again and re-cluster the cells. The associated cell type for each cluster was predicted by label transfer using previously published single-cell/single-nucleus RNA-seq (scRNA-seq/snRNA-seq) and sci-ATAC-seq datasets from mouse lung tissue (Additional file 1: Fig. S6a-d; [3, 30]), mouse liver tissue (Additional file 1: Fig. S7a-d; [3, 31]), and human lung tissue (Additional file 1: Fig. S8; [32–34]). The predicted labels were further manually curated according to the gene activity scores (by summing the read counts in gene bodies and promoters [19]) of marker genes that exhibited cell-type-specific expression patterns using scRNA-seq data browsers of mouse lung (Additional file 1: Fig. S6e; [30, 35]) and liver (Additional file 1: Fig. S7e; [31]). The top motifs in each cell type were also evaluated to further confirm the annotated cell types in the murine lung (Additional file 1: Fig. S6f) and liver tissue (Additional file 1: Fig. S7f). For example, the TEA domain (TEAD) motifs were highly accessible in AT1 cells [36]; tumor protein p63 (TP63) motif displayed increased accessibility in basal cells [37]; and hepatocyte nuclear factor (HNF)

motifs were specifically accessible in hepatic lineage cells [38, 39]. As a result, we identified 24 clusters representing distinct cell types in mouse lung tissue (Fig. 4e) and 7 clusters in mouse liver tissue (Fig. 4f). Even relatively rare cell types such as goblet cells (1335 cells, 1.8% of total), pericytes (833 cells, 1.1% of total), and myofibroblasts (366 cells, 0.5% of total) in mouse lung tissue were identified, in contrast to the previous sci-ATAC-seq atlas. To evaluate the performance of txci-ATAC-seq in cell type prediction, we randomly subsampled (without replacement) our mouse lung data to have the same number of cells as that in sci-ATAC-seq 1000 times and ran cell type prediction using label transfer. As compared to the combinatorial indexing assay, txci-ATAC-seq exhibited improved prediction accuracy (Additional file 1: Fig. S9). Further validating our approach in human lung tissue, we identified 9 distinct clusters (Additional file 1: Fig. S10a) and found that the human lung nuclei exhibited consistent clustering (Additional file 1: Fig. S10b) and data quality (Additional file 1: Fig. S10c-e), regardless of which mouse sample they were mixed with in the barnyard experiment. In addition, while we did observe some stratification of mouse hepatocytes according to the individual mouse replicate, no other cell type showed evidence of batch effects, between either mouse replicates or nuclei loading inputs (Additional file 1: Fig. S11).

To gain insight into the improvement of Tn5-based molecular hashing approaches (such as our txci-ATAC-seq) in cell recovery as compared to the cellular hashing methods, we simulated the cell recovery outcome as a function of loading input spanning from 1000 to 1.5 million nuclei for both hashing strategies, modeling cells in droplets under a Poisson distribution (see “Methods”). The theoretical estimates of cellular hashing were evaluated using the cell recovery reported in the SNUBar paper (Fig. 4g). Similarly, the evaluations of molecular hashing estimates using different numbers of Tn5 barcodes were conducted by comparing them with the cell recovery measured in both txci-ATAC (96 Tn5 barcodes) and dsciATAC (48 Tn5 barcodes) datasets across a range of input nuclei [9]. Because a bead overloading strategy was employed for the dsciATAC method, we identified bead multiplets (which refers to droplets containing more than one bead barcode) in both droplet-based assays using the bead-based ATAC processing (bap) package [40] and then merged the bead barcodes inferred to have been present in the same droplet prior to downstream analysis (see “Methods”). The cell thresholds were automatically determined by performing K -means clustering on \log_{10} -transformed read counts for each dataset (Additional file 1: Fig. S12a). While it had been previously reported that ~13–21% of bead barcodes could be derived from multi-bead droplets for the 10 \times platform [40], we only inferred ~3–4.1% of bead barcodes affected by bead multiplets in our txci-ATAC system (Additional file 1: Fig. S12b), suggesting that this phenomenon has a minor impact on our txci-ATAC data. Comparing molecular and cellular hashing strategies, the theoretical maximum of cells recovered that can be deconvoluted by molecular hashing was ~5 times higher than that achieved by cell hashing (Fig. 4g). When using 96 barcodes for molecular hashing (as demonstrated for our txci-ATAC-seq method), the number of deconvoluted cells from a single lane of the 10 \times instrument was maximized at an input of 470,680 total nuclei, yielding data from 194,970 usable cells. Cellular hashing, however, only reached a maximum of 37,134 usable cells upon loading 161,292 nuclei. Notably, txci-ATAC-seq empirical data not only achieved a substantially higher cell recovery rate than dsciATAC (averaging 61.0% in txci-ATAC vs. 39.1%

in dsciATAC), but also exhibited a closer alignment with the theoretical expectations (Fig. 4g). A further evaluation of the ratio of collision cells (i.e., cells that could not be deconvoluted) to usable cells through simulation revealed that the increased throughput in the cell hashing also requires the added cost of sequencing a substantially higher number of unusable cells than molecular hashing strategies (Fig. 4h). Comparing the two molecular hashing strategies head-to-head, txci-ATAC-seq consistently demonstrated lower collision rates than dsciATAC (Fig. 4i). The improved performance of txci-ATAC in both cell recovery and collision rate reduction remained evident even using a simple universal read depth threshold for determining cells (Additional file 1: Fig. S12c,d). In addition, a significantly higher library complexity was observed for txci-ATAC compared to that obtained by dsciATAC (Additional file 1: Fig. S12e). While this may be partially explained by differences in sample types, the lower complexity measured in dsciATAC could be attributed to the fundamental difference in single-cell platforms employed by the txci-ATAC and dsciATAC as a lower recovery of unique fragments was reported in the Bio-Rad system compared to the 10× [10].

Development of Phased-txci-ATAC-seq to improve multiplexing capability

While the conventional txci-ATAC-seq method enables sample multiplexing, its efficiency is hindered by the labor-intensive nature of nuclei washing and counting procedures, constraining the number of samples (typically no more than 12) that can be processed in a single day. To address this limitation, we developed a “phased” protocol variant (Phased-txci-ATAC-seq) that effectively decouples sample processing from library preparation. Specifically, on the sample processing day, nuclei are isolated, quantified, and then cryopreserved in either PCR tube strips or a 96-well plate. On the designated library preparation day, the frozen nuclei are thawed and directly subjected to tagmentation via a modified transposition reaction. This alteration eliminates the need for intermediate steps such as nuclei washing and quantification on the library preparation day, empowering us to gradually accumulate samples (potentially up to 96) over time and subsequently process all samples in parallel during the library preparation phase of the protocol. To evaluate the performance of the phased version of our protocol, we applied it to mouse lung nuclei that were isolated from either wild-type (WT) or club cell secretory protein-deficient (*CC16*^{-/-}) mice with three replicate lungs for each genotype. The standard txci-ATAC-seq was also performed on the same samples separately. *CC16* is a secreted protein encoded by the *Scgb1a1* gene that is produced predominantly by club cells, an epithelial cell type of the airways. This “pneumoprotein” plays an important role locally in protecting the lung against oxidant injury [41] and inflammatory diseases, such as asthma [42] and chronic obstructive pulmonary disease (COPD) [43]. It has also been linked with more systemic effects on human health as evidenced by its association with overall cancer risk [44]. After processing and pooling all samples for each protocol (Additional file 1: Fig. S13a), we loaded 50,000 and 100,000 nuclei on the 10× Genomics platform for Phased-txci-ATAC-seq and used 100,000 and 200,000 nuclei as inputs for the standard assay. The removal of low-quality nuclei and predicted doublets resulted in similar recovery rates between the two protocols with 44.3% nuclei (10,937 WT nuclei and 11,213 *CC16*^{-/-} nuclei) at the 50,000 input and 43.6% nuclei (21,688 WT nuclei and 21,961 *CC16*^{-/-} nuclei) at the 100,000 input for the phased

protocol, compared to 50.0% nuclei (24,962 WT nuclei and 25,011 CC16^{-/-} nuclei) at the 100,000 input and 52.1% nuclei (51,536 WT nuclei and 52,627 CC16^{-/-} nuclei) at the 200,000 input for the standard txci-ATAC-seq (Additional file 1: Fig. S13b). An examination of QC metrics demonstrated that both assays can provide high-quality single-cell data despite a slightly lower complexity observed in the phased version (Additional file 1: Fig. S13c-f). Using the iterative clustering strategy and label transfer with a scRNA-seq reference followed by manual curation with scRNA-seq marker genes [30, 35] (Additional file 1: Fig. S13g) and top motifs (Additional file 1: Fig. S13h), we identified 23 distinct cell clusters in mouse lungs profiled by the standard txci-ATAC-seq (Fig. 5a) and then used them to further annotate the Phased-txci-ATAC-seq lungs. While the cellular heterogeneity in mouse lungs has been characterized by other atlas efforts [45], the enhanced scale and quality of our data enabled the identification of certain rare cell types lacking well-defined regulatory DNA signatures, such as pulmonary neuroendocrine cells (PNECs) and basal cells. A joint embedding of both assays revealed that the phased protocol recapitulated the mouse lung heterogeneity in chromatin accessibility characterized by the standard protocol (Additional file 1: Fig. S13i) with minimal batch effects (Additional file 1: Fig. S13j).

(See figure on next page.)

Fig. 5 Chromatin accessibility dynamics induced by CC16 deficiency and genetic variants. **a** UMAP visualization of WT and CC16^{-/-} mouse lung nuclei ($n = 154,136$) across two loading inputs by integrating six animals with three replicates from each group. Nuclei are colored by their predicted cell type. The abbreviation of cell labels was described in Fig. 4 except for aEC (arterial endothelial cells), Endo-like (endothelial-like cells), and PNEC (pulmonary neuroendocrine cells). **b** The number of differential peaks identified between CC16^{-/-} and WT samples for each cell type. The blue bars indicate the peaks less accessible in the knockout samples and the red bars represent the more accessible peaks. **c** Aggregated chromatin accessibility surrounding the *Scgb1a1* (CC16 gene) locus in club and goblet cells per sample. The aggregated accessibility signal for each sample was normalized by the scaling factor that was computed as the number of cells in the sample multiplied by the mean sequencing depth for the cells in that sample. The WT tracks are labeled in blue and the knockout ones are in red. The genomic regions for the significantly less accessible peaks identified in CC16^{-/-} samples per cell type are highlighted by green shade (five peaks in club cells and two peaks in goblet cells). The associated adjusted p-value is shown above the tracks at their corresponding peak region. Adjusted p-values less than 0.0001 are given four asterisks. The peak annotation for the promoter region of *Scgb1a1* is colored red. **d** Chromosomal distribution of the midpoint of differential peaks identified on chromosomes 8 and 19 with the genomic location and density estimate plotted on the x- and y-axis, respectively. **e** Chromosomal distribution of SNVs identified on chromosomes 8 and 19 for both WT (blue) and CC16^{-/-} (red) samples. The regions between the dashed lines indicate the SNV hotspots where the knockout samples exhibited a substantially higher number of SNVs than WT samples. The y-axis shows the Phred-scaled quality score generated by BCFtools. **f** Heatmap showing the Jaccard similarity between the hotspot SNVs identified in CC16^{-/-} lungs and the SNPs derived from 36 different strains on chromosome 8 (lower triangle) and 19 (upper triangle). **g** "Functional" motifs for which gains or losses of the motif instances are associated with significant changes in chromatin accessibility. The motifs associated with increased chromatin accessibility ("opening") are shown in red and those associated with decreased chromatin accessibility ("closing") are colored in green. The y-axis represents the Student's *t*-test statistic value. Two motif families (one for transcriptional activators and one for transcriptional repressors) are highlighted on the x-axis. **h** Cell-type-specific enrichment for the motifs that explain chromatin accessibility changes in SNV hotspots. The bar plot next to the enrichment heatmap shows the total number of differential peaks located in the SNV hotspots for each cell type, which is stratified by the peaks that can be explained by the SNV-driven difference in motif presence (red) and unexplained peaks (blue). The values next to the bars denote the percentage of peaks explained. Only the cell types with more than 10 differential peaks identified in the SNV hotspots are shown

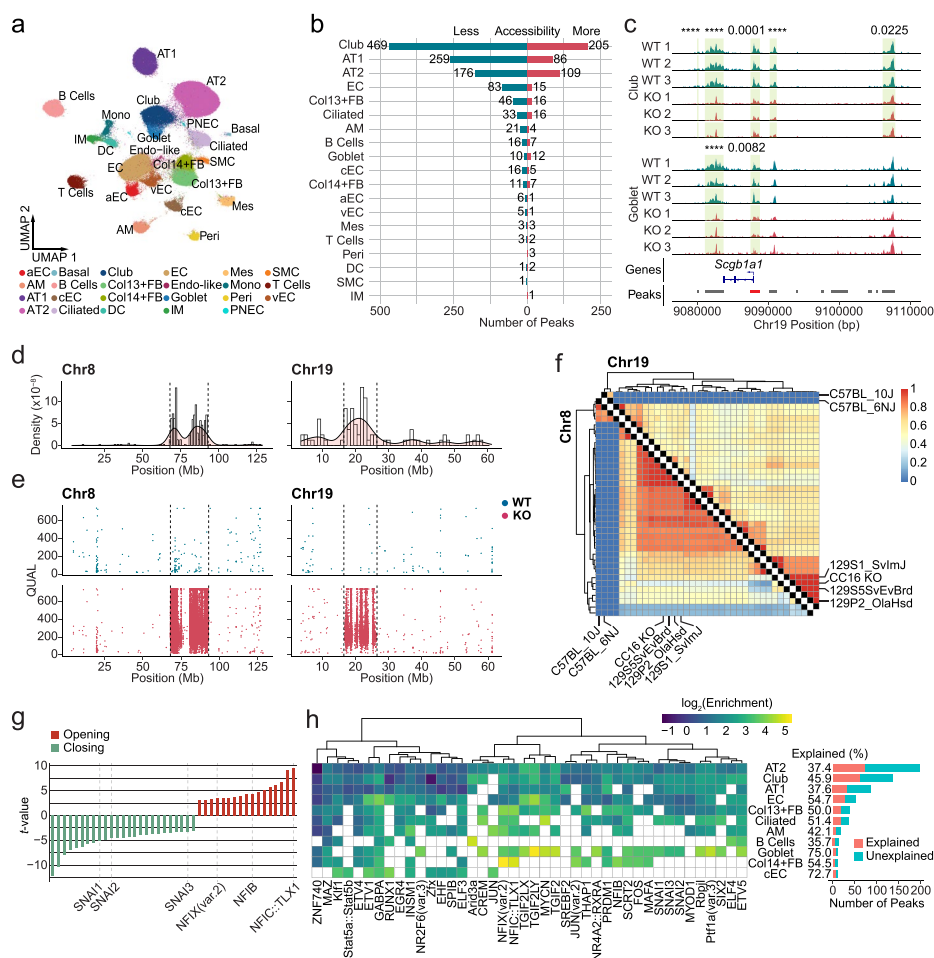


Fig. 5 (See legend on previous page.)

Cell-type-specific regulation of chromatin accessibility in CC16^{-/-} mouse lungs

We next used the cells profiled by the standard assay to explore meaningful differences in chromatin accessibility between WT and CC16^{-/-} mice. We were initially interested to identify differences of biological import, including (1) chromatin accessibility of the *Scgb1a1* locus being restricted to club and goblet cells (Additional file 1: Fig. S14a), (2) significantly differentially accessible peaks across a variety of cell types (Fig. 5b and Additional file 2: Table S1)—many of which were only identifiable at such high throughput (Additional file 1: Fig. S14b,c)—(3) some evidence for potential autoregulation of CC16 (Fig. 5c and Additional file 1: Fig. S14d), and (4) differential peaks enriched for TF motifs (Additional file 1: Fig. S15a,b and Additional file 3: Table S2) and molecular pathways (Additional file 1: Fig. S15c and Additional file 4: Table S3). However, we also noted an unexpected number of differential peaks in two genomic loci on chromosomes 8 and 19 (Fig. 5d and Additional file 1: Fig. S15d). While the *Scgb1a1* gene is located on chromosome 19 (mm10 chr19:9,083,636–9,087,958), this chromosome-specific enrichment of differential peaks was unexpected. To better understand the concentration of signal in these two loci, we carried out variant calling on the ATAC-seq data and identified 5909

single-nucleotide variants (SNVs) differing from the reference genome in WT samples and 50,054 SNVs in $CC16^{-/-}$ samples (Additional file 1: Fig. S15e). The vast majority of SNVs in the $CC16^{-/-}$ samples were located in two hotspots (Fig. 5e) on chromosome 8 ($n = 37,822$; 75.6%) and chromosome 19 ($n = 5,610$; 11.2%), essentially perfectly matching the locations where the differential peaks were identified. To trace the origin of the $CC16^{-/-}$ SNVs, we further mapped the hotspot SNVs to the single-nucleotide polymorphism (SNP) profiles that were previously defined in 36 different mouse strains relative to the C57BL/6 J mouse reference genome [46]. Almost all of our identified SNVs matched to the SNPs identified in the three 129 strain references (Fig. 5f) on chromosome 8 (an average of 96.3% of SNVs matched the SNPs defined in each of the 129 strains) and chromosome 19 (an average of 96.6% SNVs matched the SNPs defined in each of the 129 strains). Given that the $CC16^{-/-}$ mice were generated using 129-derived embryonic stem (ES) cells, we conclude that the hotspot SNVs are remnants of the 129 genome, a common problem with knockout models [47]. Notably, we found ~90% of SNVs residing in intronic and intergenic regions for both WT and $CC16^{-/-}$ samples (Additional file 1: Fig. S15f), suggesting ATAC-seq may have been a particularly powerful choice of assay for capturing such genetic variation and thus may serve as a cost-effective alternative to whole genome sequencing in genotyping knockout models.

Although the SNV-driven phenotype confounded the analysis of *Scgbla1* effects, it provided an opportunity to explore the extent and mechanism by which genetic variants can modulate chromatin accessibility, even in a cell-type-specific manner. To this end, we took all the peaks that were differentially accessible in the hotspot regions (413 peaks) and looked for TF motifs that were gained or lost due to SNVs. The functional motifs were defined as those whose chromatin accessibility exhibited a significant positive or negative correlation with the gain or loss of the motif in the knockout mice relative to the WT mice. We identified 42 functional motifs and found that gaining the motifs for the transcriptional activators, e.g., certain members of the nuclear factor I (NFI) family and the ETS-domain family, tended to increase chromatin accessibility (Fig. 5g and Additional file 1: Fig. S16a). On the other hand, gaining a repressor motif, such as motifs for the Snail and Scratch families, was likely to reduce chromatin accessibility (Fig. 5g and Additional file 1: Fig. S16b). Finally, we investigated whether there were cell-type-specific enrichments for specific functional motifs being gained or lost (Fig. 5h). We observed that gains and losses of NFI TFs, including NFIB, NFIC::TLX1, and NFIX (var.2), were highly enriched in both $Col13^+$ and $Col14^+$ fibroblasts and similarly, gains and losses of the ARID3A motif, which is required for B cell lineage development [48], was highly enriched in differential peaks in B cells. In sum, functional motifs being gained or lost were able to account for a substantial number of differentially accessible peaks observed in the SNV hotspot regions in different cell types—ranging from 35.7% of differentially accessible peaks from the regions for B cells to 75% of differentially accessible peaks from the regions for goblet cells (Fig. 5h).

Discussion

Limited throughput, prohibitive cost, and variance between batches have put some limitations on the implementation of single-cell techniques, which nonetheless are proving invaluable resources for studying health and disease. To reduce those limitations, we paired combinatorial indexing with a droplet-based microfluidic system to substantially increase the scalability of the commercial single-cell device by loading up to 200,000 nuclei in a single emulsion reaction. In addition, a “phased” version protocol was developed, which greatly improved the multiplexing capability. The scalability and flexibility allow txci-ATAC-seq to establish unbiased regulatory definitions across various disease, genetic, and/or environmental states. Other strategies do exist for multiplexing samples on microfluidic single-cell platforms, such as membrane barcoding-based approaches that tag cellular or nuclear membrane components [49–51] and genetic deconvolution of samples [52]. However, those methods index at the cellular/nuclear level and so the scalability is restricted by the maximum number of singlets that can be generated because multiplets cannot be deconvoluted. Conversely, the molecular indexing strategy used in our design along with that previously implemented on a different commercial instrument (dsciATAC-seq) [9] and on RNA profiling (scifi-RNA-seq) [53] allows for multiplets to be deconvoluted, resulting in the ability to load substantially more nuclei per lane and therefore provide larger-scale sample multiplexing and increased cost savings. By quantifying the number of hashing barcodes in each droplet, we can empirically assess cell recovery in both cellular and molecular indexing strategies using our txci-ATAC-seq datasets alone. As expected, deconvoluting cells based on molecular indexing exhibited markedly higher cell recovery compared to the cellular indexing deconvolution strategy in which multiplets must be discarded, especially with increasing loading inputs (Additional file 1: Fig. S17). In addition, compared to the theoretical cell recovery, the total cells deconvoluted by txci-ATAC aligned closely with the simulated cell recovery achieved through the molecular hashing strategy. However, for cellular hashing, a noticeable reduction in the number of single-Tn5 droplets was observed at the larger loading inputs compared to the theoretical recovery, suggesting that the cellular hashing recovery rate would be more adversely affected than molecular hashing by multiplet rates that exceed our expectations. Given our results and the reported metrics for scifi-RNA-seq [53], we are confident that an even higher throughput is achievable with txci-ATAC-seq (even more so if one were to leverage 384 barcoded transposition reactions). Based on current costs, we estimate that a standard 10× ATAC-seq run costs approximately \$0.175 per cell. For txci-ATAC-seq, assuming that commercial Tn5 was purchased and loaded, the approximate cost is \$0.015 per cell, representing a 12-fold reduction in cost relative to the standard workflow. If one were to produce their own Tn5 following previously published protocols [54–56], the cost per cell would drop by 40% as we estimate that the commercial Tn5 accounts for \$0.006 per cell of the total cost. We estimate that the cost per cell for cellular hashing would be approximately \$0.047. In addition, cellular-level indexing results in a large number of cell multiplets, which cannot be used but still require extra costs for sequencing. Also, the antibody-based cell hashing approaches necessitate additional expenditures for acquiring hashing antibodies [49, 50].

We believe that implementing the SBS probe, designed to counteract Tn5 barcode-swapping, has the promise of re-engineering other single-cell techniques that leverage the 10× scATAC platform into multiplexable and cost-effective super-loading assays, including scCUT&Tag (a method that utilizes Tn5 to map the genomic locations of TF binding and histone modifications) [57] and ISSAAC-seq (a multi-omics approach enabling simultaneous profiling of chromatin accessibility and gene expression) [58]. We also note that the design of txci-ATAC-seq is potentially applicable to existing single-cell methods employing a combinatorial indexing framework, such as sci-MET [59], CRISPR-sciATAC [60], and sci-CAR [61].

The improved study design and statistical rigor made possible by more cost-effective inclusion of replicates and larger sample sizes with techniques such as txci-ATAC-seq will be essential for realizing the full potential of single-cell approaches. In addition, the scalability of scATAC-seq techniques also plays an important role in identification of peaks for rare cell populations. In the absence of a comprehensive catalog of regulatory elements, peak calling is an essential step to define features in both bulk and single-cell ATAC-seq data analysis. The power to call peaks, however, heavily depends on the number of reads used [56]. For single-cell data, that means profiling the accessible regions from a rare cell population is not only limited by the sequencing depth per cell but also by the number of cells captured from that population. Therefore, an ultra-high throughput scATAC-seq method, like txci-ATAC-seq, will enable finer definitions of peaks and should better characterize particularly dynamic or heterogeneous systems.

The CC16^{-/-} mice characterized here have been used by several groups to investigate the role of CC16 in COPD and infectious diseases [43, 62, 63]. We found that the remnant 129 genetic material elicited profound changes in chromatin accessibility (in a cell-type-specific manner in many instances), requiring caution when evaluating the existing congenic knockout models. In addition, we identified 42 different motifs gained or lost in at least one differentially accessible peak from the 129 strain regions, which were capable of explaining the observed accessibility changes for 37.5% of those peaks. The remainders may have been caused by more subtle changes in motif affinity, trans effects, or may not have been tested in our analysis (as we required both gained and lost events for a given motif to be considered).

There are several caveats worth keeping in mind when interpreting our results. First, to marry microfluidics and combinatorial indexing on the 10× system, we converted the in-droplet linear amplification into an exponential amplification. This could result in major differences in amplification behavior. However, we have not systematically tested the optimal number of cycles in this regime. In addition, our analysis approach is based on the assumption that each droplet contains at most one bead barcode. It is worth noting, however, that we identified ~4% of bead barcodes derived from bead multiplets in our data. The resulting artifact “cells” may slightly confound some of our interpretations.

Conclusions

Taken together, txci-ATAC-seq provides unprecedented opportunities to generate unbiased single-cell atlases of chromatin accessibility for large cohorts with various genetic backgrounds or case–control studies, thus establishing reliable references of single-cell chromatin landscapes in a variety of experimental settings. We hope that this method

will encourage more widespread adoption of scATAC-seq, a powerful technique for understanding organismal development and disease processes.

Methods

Cell lines

The GM12878 (Coriell Cell Repository) and CH12.LX (kind gift from the Sherman Weissman lab) cells were cultured at 37 °C with 5% CO₂ in RPMI 1640 medium (GIBCO, cat. no. 11875–093) containing 15% FBS (GIBCO, cat. no. 10437–028), 100 U/ml Penicillin Streptomycin (GIBCO, cat. no. 15140–122). Cells were counted and split into either 300,000 (GM12878) or 100,000 (CH12.LX) cells/ml three times a week. The cell lines used in this study were not authenticated or checked for mycoplasma contamination.

Human and mouse brain tissue samples

Human cortex samples from the middle frontal gyrus were sourced from the Oregon Brain Bank from a 50-year-old female of normal health status. Samples were collected by an OHSU neuropathologist, placed into a labeled cassette, and cryopreserved in an airtight container in a –80 °C freezer. The duration of time between the time of death and brain biopsy sample freezing, or post-mortem interim (PMI), was <24 h.

Mouse brain tissue was collected as discarded tissue from mice used for unrelated studies approved by the OHSU IACUC. Whole mouse brains were dissected from sacrificed C57BL/6 J mice and flash-frozen in an isopentane-LN₂ double-bath and stored at –80 °C.

Mouse lung and liver tissue samples

All animal activity was approved by the University of Arizona IACUC. Mice were euthanized via exsanguination followed by cervical dislocation to ensure death. For the samples used to evaluate the performance of txci-ATAC-seq in Fig. 4, whole mouse lungs and liver were dissected from 2 male C57BL/6 J mice that were 24 weeks old.

For the samples used to study the CC16-mediated chromatin dynamics in Fig. 5, age-matched (~8 weeks) WT and CC16^{-/-} male mice on a C57BL/6 J background (as described in [64, 65]) were used to dissect whole lungs. Three replicates from each genotype were profiled. All six animals were born and raised in the same room and were tested to be specific-pathogen-free according to standard protocols using sentinel mice from the same room.

The dissected samples were flash-frozen in liquid nitrogen and then transferred to –80 °C for long-term storage.

Human lung tissue samples

De-identified lung pieces were provided by the Arizona Donor Network from two deceased male donors (a 36-year-old American Indian and a 62-year-old Hispanic Latino) as soon as possible after the time of death. All human lung samples were quickly frozen in the –80 °C freezer and stored there prior to nuclear extraction.

Nuclei isolation

Nuclei isolation of cell lines

The nuclei isolation followed the procedures described in [13]. The cells were collected and washed with 1 × PBS (pH 7.4, Gibco, cat. no. 10–010-023) supplemented with 0.1% BSA (New England Biolabs, cat. no. B9000S) and then resuspended in 200 µl of ATAC-seq lysis buffer, which was made by supplementing ATAC resuspension buffer (RSB) with detergents (see below). RSB buffer is 10 mM Tris–HCl (pH 7.5, Invitrogen, cat. no. 15567027), 10 mM NaCl (Invitrogen, cat. no. AM9759), and 3 mM MgCl₂ (Invitrogen, cat. no. AM9530G) in nuclease-free water. RSB was made in bulk and stored at 4 °C long-term. On the day of the experiment, the ATAC lysis buffer was made by adding 0.1% IGEPAL (Sigma, cat. no. I3021), 0.01% digitonin (Invitrogen, cat. no. BN2006), and 0.1% Tween-20 (Bio-Rad, cat. no. 1610781) to RSB. The detergent percentages reported are final concentrations. After resuspending cell pellets in the lysis buffer, they were incubated on ice for 3 min, and then the lysis was stopped by adding 1 ml RSB containing 0.1% Tween-20. The nuclei were counted with a hemocytometer by diluting 10 µl nuclei in 40 µl of 2 × Omni TD Buffer (20 mM Tris HCl pH 7.5, 10 mM MgCl₂ and 20% Dimethyl Formamide) followed by adding 50 µl Trypan blue solution. In our previous report [56], we found that adding nuclei straight to Trypan blue solution will cause inflation of nuclei and diluting nuclei in TD buffer before exposure to Trypan blue improves the nuclei integrity. Following counting, we centrifuged nuclei at 500 r.c.f for 10 min at 4 °C and removed the supernatant. Then, the nuclei were either used to perform downstream experiments directly or resuspended in a nuclei-freezing buffer (NFB) containing 50 mM Tris-HCl (pH 8.0, Invitrogen, cat. no. 15568025), 5 mM Magnesium Acetate (Sigma, cat. no. 63052), 25% glycerol (VWR, cat. no. RC3290-32), 0.1 mM EDTA (Fisher, cat. no. AM9260G), 5 mM DTT (Fisher, cat. no. P2325), and 2% (v/v) protease inhibitor (Sigma, cat. no. P8340) for storage. The NFB was adopted from [66] and we previously used this buffer for preservation of nuclei for sci-ATAC-seq [2, 3, 67]. After diluting in NFB, 1 ml aliquots of the nuclei were flash-frozen in liquid nitrogen and then transferred to a liquid nitrogen dewar for long-term storage.

Nuclei isolation from brain tissue

At the time of nuclei dissociation, 50 ml of nuclei isolation buffer (NIB-HEPES) was freshly prepared with final concentrations of 10 mM HEPES–KOH (Fisher Scientific, BP310-500 and Sigma Aldrich 1,050,121,000, respectively), pH 7.2, 10 mM NaCl (Fisher Scientific S271-3), 3 mM MgCl₂ (Fisher Scientific AC223210010), 0.1% (v/v) IGEPAL CA-630 (Sigma Aldrich I3021), 0.1% (v/v) Tween-20 (Sigma-Aldrich P-7949), and diluted in PCR-grade Ultrapure distilled water (Thermo Fisher Scientific 10,977,015). After dilution, two tablets of Pierce™ Protease Inhibitor Mini Tablets, EDTA-free (Thermo Fisher A32955) were dissolved and suspended to prevent protease degradation during nuclei isolation.

An at-bench dissection stage was set up prior to nuclei extraction. A petri dish was placed over dry ice, with fresh sterile razors pre-chilled by dry-ice embedding; 7 ml capacity Dounce homogenizers were filled with 2 ml of NIB-HEPES buffer and held on wet ice. Dounce homogenizer pestles were held in ice-cold 70% (v/v) ethanol (Decon Laboratories Inc 2701) in 15 ml tubes on ice to chill. Immediately prior to use, pestles

were rinsed with chilled distilled water. For tissue dissociation, mouse and human brain samples were treated similarly. The still-frozen block of tissue was placed on the clean pre-chilled petri dish and roughly minced with the razors. Razors were then used to transport roughly 1 mg of the minced tissue into the chilled NIB-HEPES buffer within a Dounce homogenizer. Suspended samples were given 5 min to equilibrate to the change in salt concentration prior to douncing. Tissues were then homogenized with 5 strokes of a loose (A) pestle, another 5-min incubation, and 5–10 strokes of a tight (B) pestle. Nuclei were transferred to a 15-ml conical tube and pelleted with a 400 r.c.f centrifugation at 4 °C in a centrifuge for 10 min. The supernatant was removed and pellets were resuspended in 5 ml of ATAC-PBS buffer (APB) consisting of 1X PBS (Thermo Fisher 10,010) and 0.04 mg/ml (f.c.) of bovine serum albumin (BSA, Sigma Aldric A2058). Samples were then filtered through a 35- μ m cell strainer (Corning 352,235). A 10 μ l aliquot of suspended nuclei was diluted in 90 μ l APB (1:10 dilution) and manually counted on a hemocytometer with Trypan Blue staining (Thermo Scientific T8154). The stock nuclei suspension was then diluted to a concentration of 2857 nuclei/ μ l in APB. Dependent on experimental schema, pools of tagged nuclei were combined to allow for the assessment of pure samples and to test index collision rates.

Nuclei isolation of human lung, mouse lung, and mouse liver tissue

The human and mouse samples were dissected and stored at -80 °C. The nuclei isolation procedure of lung and liver tissues was performed following the single-nucleus isolation protocol described in [68]. To do so, we cut a ~ 0.1 – 0.2 g piece from either human or mouse samples removed from -80 °C and kept it on dry ice until use. The tissue block was thawed almost completely on ice for 1 min and then injected with 1 ml of cell lysis buffer, which was made of 1 \times cComplete protease inhibitor cocktail (1 tablet per 10 ml solution, Sigma-Aldrich, Cat. 11,836,153,001) in Nuclei EZ prep buffer (Sigma-Aldrich, Cat. NUC101), into the center of the tissue with a 30-G needle and syringe. Following lysis buffer injection, the tissue was chopped into small pieces with scissors and then transferred along with the lysing buffer into a gentleMACS C tube (Miltenyi Biotec, Cat. 130–096-334). An additional 1 ml of lysing buffer was added into the C tube to make a final volume of 2 ml. The minced tissue was then homogenized using a gentleMACS tissue dissociator by running the “m_lung_01” program followed by the first 20 s of the “m_lung_02” program. After homogenization, tissue lysate was briefly centrifuged to reduce foam and then passed through a 40- μ m cell strainer in a 50-ml tube. After passing the sample through, the strainer was rinsed with 4 ml of washing buffer (PBS with 1% BSA). The nuclei were counted with a hemocytometer (see “[nuclei isolation of cell lines](#)” for details) and centrifuged at 500 r.c.f for 5 min at 4 °C. Then, we removed the supernatant and resuspended the nuclei in the NFB to make a concentration of 4–5 million nuclei/ml; 1 ml aliquots of the nuclei were flash-frozen in liquid nitrogen and then transferred to a liquid nitrogen dewar for long-term storage.

Sample multiplexing

A 96-well plate pre-loaded with 5 μ l of 500 nM pre-indexed Tn5 transposase per well (iTSM plate, kind gift of Illumina Inc.) was used to multiplex samples and perform bar-coded transposition. Before using, the iTSM plate was thawed on ice and briefly mixed

at 1400 rpm for 30 s on a pre-chilled thermomixer and then quickly spun to collect the enzyme at the bottom of the wells. To avoid sequencing with a custom recipe, the Tn5 enzyme was loaded with a common Tn5ME-A and a custom Tn5ME-B containing a partial sequence of i7 TruSeq primer (see Additional file 5: Table S4 for oligo sequence) and an 8-bp unique barcode (Additional file 5: Table S5). Both Tn5ME-A and Tn5ME-B were annealed to the Tn5MErev (Additional file 5: Table S4) before loading to Tn5.

Barnyard experiments

Two different barnyard settings were designed to estimate the total collisions arising from pre- and/or post-pooling events. To test the total collision rate, the human and mouse cells were mixed in the same well at a 1:1 ratio to perform barcoded transposition (“true barnyard”). The collision rate driven by events downstream of pooling was evaluated by performing barcoded transposition on wells containing pure species (“pseudo-barnyard”) and pooling the human and mouse nuclei afterward. Detailed information about the cell sources used in each barnyard assay and each figure is shown in Additional file 5: Table S6.

Optimization of txci-ATAC-seq protocol

Coupling barcoded transposition with standard 10× protocol

The nuclei isolated from human and mouse lungs were removed from the liquid nitrogen dewar (see “[Nuclei isolation of human lung, mouse lung, and mouse liver tissue](#)” for details) and then thawed in the water bath at 37 °C for 1 to 2 min until a tiny ice crystal remained. After thawing, the nuclei stored in 1 ml freezing buffer were diluted with 3 ml RSB supplemented with 0.1% Tween-20 and 0.1% BSA (RSB washing buffer) and then centrifuged at 500 r.c.f for 10 min in a pre-chilled (4 °C) swinging bucket centrifuge. The nuclei pellet was resuspended with another 1 ml of RSB washing buffer and then transferred to a 1.5-ml LoBind tube through a 40-µm Flowmi Cell strainer (Bel-Art SP Scienceware, Cat. 14–100-150). The filtered nuclei were pelleted at 500 r.c.f for 5 min in a pre-chilled fixed-angle centrifuge and then resuspended in 25 µl of 1.25 × Tagment DNA Buffer (Nextera XT Kit, Illumina Inc. FC-131–1024). For cell cultures, the human and mouse nuclei were freshly isolated as described in “[Nuclei isolation of cell lines](#)” and resuspended in 50 µl of 1 × Nuclei Buffer (10× Genomics, PN-2000207). Then, we counted nuclei for each sample and added 5000 nuclei diluted in 20 µl of 1.25 × Tagment DNA buffer to each well of the iTSM plate (see “[Sample multiplexing](#)” for details), except for the wells used to test the 10× reagents in which 5000 nuclei diluted in 5 µl of 1 × Nuclei Buffer were added to a mixture of 7 µl of ATAC Buffer B (10× Genomics, PN-2000193) and 3 µl of barcoded Tn5. The plate layout and well IDs for each barnyard condition are shown in Additional file 1: Fig. S1 and Additional file 5: Table S6. The tagmentation was performed at 55 °C for 30 min on a thermocycler with a heated lid. To quench the Tn5 activity, we added a 2 × Tagmentation Stop Buffer containing 40 mM EDTA (Invitrogen™, Cat. AM9260G) and 1 mM Spermidine (Sigma-Aldrich, Cat. S0266-1G) to the transposition reactions at a 1:1 ratio and incubated the plate on ice for 15 min. We found that stopping the transposition reaction was unnecessary and thereby removed this step from our final txci-ATAC-seq protocol. All nuclei were pooled and

centrifuged at 500 r.c.f for 10 min. After aspirating the supernatant, nuclei were resuspended in 400 μ l 1 \times Nuclei Buffer and pelleted again. Then, we carefully removed the supernatant and resuspended nuclei in 30 μ l 1 \times Nuclei Buffer. After quantification of nuclei with a hemocytometer, 75,000 nuclei were taken and diluted in 1 \times Nuclei Buffer to make a total volume of 15 μ l, which underwent the standard 10 \times Chromium Next GEM protocol (v1.1, Document No. CG000209 Rev D from Steps 2 to 4) except following steps. For Sample Index PCR (step 4.1), we substituted the Single Index N Set A with a 25 μ M i7 TruSeq primer and added 2.5 μ l of customized i7 primer (Additional file 5: Table S7) to each 10 \times library followed by performing 8 cycles of PCR amplification. The resulting library was sequenced on a NextSeq 550 Platform (Illumina Inc.) using a Mid Output Kit with the following cycles: Read 1, 50 cycles; i7 index, 8 cycles; i5 index, 16 cycles; and Read 2, 77 cycles.

Blocking barcode-swapping

Flash-frozen (human and mouse lung samples and human cell line, see “[Nuclei isolation](#)” for details) and fresh nuclei (mouse cell line, see “[Nuclei isolation](#)” for details) were used to test the efficiency of strategies to block barcode-swapping. The flash-frozen nuclei were thawed, washed, and filtered following the procedures described in the “Coupling barcoded transposition with standard 10 \times protocol” section. Both flash-frozen and freshly isolated nuclei were resuspended in 100 μ l of PBS containing 0.04% BSA (PBSB) and quantified using a hemocytometer (See “[Nuclei isolation of cell lines](#)” for details). After counting, the nuclei were diluted in PBSB to a concentration of 2857 per μ l (20,000 nuclei per well in 7 μ l) and then mixed with a Tagmentation buffer solution (TBS, which was modified from the Omni protocol [13]) followed by transferring to the iTSM plate (see “[Sample multiplexing](#)” for details). Each 13 μ l of TBS contains 12.5 μ l of Illumina Tagment DNA Buffer, 0.25 μ l of 1% Digitonin in DMSO (Promega (2%), Cat. PRG9441), and 0.25 μ l of 10% Tween-20 (Bio-Rad, Cat. 1,610,781) in nuclease-free water. The barcoded transposition reaction was performed at 37 $^{\circ}$ C for 30 min on a thermocycler with a heated lid at 47 $^{\circ}$ C. Each blocking condition was assigned to 8 columns leading to a total of two 96-well plates for all 3 conditions. The plate layout and well IDs for each barnyard design in each blocking condition are shown in Additional file 1: Fig. S2a and Additional file 5: Table S6. After tagmentation, the nuclei used to test the Decoy DNA were transferred to a new 96-well plate with a multi-channel pipette, and 2.5 μ l of 50 μ M duplex DNA (see Additional file 5: Table S8 for the oligo sequence) was added to each well followed by incubating at 55 $^{\circ}$ C for 10 min. Then, we added the 2 \times Tagmentation Stop Buffer (see “[Coupling barcoded transposition with standard 10 \$\times\$ protocol](#)” for details) to the transposition reactions at a 1:1 ratio for all three blocking conditions and incubated the plates on ice for 15 min. Subsequently, the nuclei from the same blocking condition were pooled together and pelleted at 500 r.c.f for 10 min at 4 $^{\circ}$ C. After removal of supernatant from each tube, the nuclei were washed with 500 μ l of 1 \times Nuclei Buffer (10 \times Genomics, PN-2000207) with centrifugation of 500 r.c.f for 5 min at 4 $^{\circ}$ C and resuspended in 25 μ l of 1 \times Nuclei Buffer. Then, we counted nuclei with Trypan blue on a hemocytometer and diluted 100,000 nuclei in 1 \times Nuclei Buffer to make a total of 15 μ l for each blocking condition. The resulting three aliquots of nuclei were run on separate lanes of the 10 \times as per the manufacturer’s instructions

(10× Chromium Next GEM Single Cell ATAC protocol v1.1, Document No. CG000209 Rev D) with the following modifications. During GEM Generation and Barcoding (Step 2.1a), the nuclei dedicated to evaluating the Blocking oligo were mixed with the Master Mix supplemented with 2.5 µl of 100 µM DNA oligo incorporating an inverted dT at the 3'-end (see Additional file 5: Table S8 for the oligo sequence); and the nuclei dedicated to testing the SBS primer were mixed with the Master Mix supplemented with 2.5 µl of 25 µM full SBS primer (Additional file 5: Table S8) for in-droplet exponential amplification. After GEM PCR (Step 2.5a), a 10 µl PCR product (10% GEM) was slowly aspirated and transferred to a new PCR tube and subjected to Post GEM Incubation Cleanup in parallel with the 90% sample. Following cleanup, we performed the Sample Index PCR on the 10% sample (step 4.1) by supplementing the PCR mixes of SBS primer, Decoy DNA, and Blocking oligo with 2.5 µl of 25 µM barcoded i7 TruSeq primer (Additional file 5: Table S7), which was used to replace the Single Index N Set A. The PCR mixes were amplified and monitored on a Bio-Rad CFX Connect Real-time cycler. The amplification was stopped when it appeared to be leveling off (i.e., the SBS primer was stopped at 4 cycles; the Decoy DNA and Blocking oligo were stopped at 15 cycles). To monitor the relative efficiencies of amplification in our initial test, we ended up introducing two different barcoded SBS primers in the SBS condition: one barcode was used for in-droplet amplification and another barcode was used for final library sample indexing. Both barcodes were assigned to thousands of reads per cell, indicating that both reactions were working. However, the theoretical expectation for the ratio between the two barcodes was 1/16 (because the second primer was used for 4 cycles of PCR). When we examined the ratio in our actual data, it was consistently ~1/3, indicating that the sample index amplification is not perfectly efficient (Additional file 1: Fig. S18). Therefore, in subsequent experiments using lung and liver tissues, we reduced the in-droplet PCR to 8 cycles and added an additional cycle of PCR for sample indexing. The resulting libraries with 10% GEM were pooled together with a library from an unrelated experiment to balance nucleotide diversity through the fixed sequence at the Tn5Merev region in Read 2, and then sequenced on a NextSeq 550 Platform (Illumina Inc.) using a Mid Output Kit with the following cycles: Read 1, 50 cycles; i7 index, 10 cycles; i5 index, 16 cycles; and Read 2, 92 cycles. While 8 cycles in i7 index and 77 cycles in Read 2 were sufficient for the libraries generated in this study, we ran 10 and 92 cycles for those two steps, respectively, to accommodate the other library.

txci-ATAC-seq using brain tissue samples

Tagmentation plates were prepared by the combination of 1430 µl of TBS with 770 µl nuclei solution. The TBS recipe was described in “[Blocking barcode-swapping](#)”, but a different version of Digitonin (Bivision 2082–1) was used here. This solution was mixed briefly on ice; 20 µl of the mixture was placed into the 96-well iTSM plate (see “[Sample multiplexing](#)” for details). Tagmentation was performed at 37 °C for 60 min on a 300 r.c.f Eppendorf ThermoMixer with a lid heated to 65 °C. Following this incubation, plate temperature was brought down with a 5-min incubation on ice to stop the reaction. Tagmented nuclei were then pooled into a single 15-ml conical tube; 5 ml of tagmentation wash buffer (TMG) was prepared consisting of a final concentration of 10 mM Tris

acetate pH 7.5 (Sigma 93,352 and Sigma A6283, respectively), 5 mM magnesium acetate (Sigma M5661), and 10% (v/v) glycerol (Sigma G5516), diluted in PCR grade water; 1 ml of TMG was added on top of the chilled tagged nuclei. Nuclei were pelleted at 500 r.c.f for 10 min at 4 °C. Most of the supernatant was removed with care not to disturb the pellet. Then 500 µl of TMG was added to the pellet and the tube was once again spun at 500 r.c.f for 5 min at 4 °C; 490 µl was removed leading to a low volume of concentrated nuclei. Loading buffer was prepared by combining two 5 × stock buffers and diluting them to 1 × in water (buffer 1 consisted of 50 mM Tris acetate pH 7.6, 25 mM magnesium acetate, and 50% (v/v) dimethyl formamide; buffer 2 consisted of 50% (v/v) glycerol, 100 mM NaCl, 50 mM Tris–HCl pH 7.5, 0.1 mM EDTA (Fisher Scientific AM9260G), and 1 mM DTT (VWR 97061–340)). The nuclear pellet was resuspended with an additional 30 µl of loading buffer. An aliquot of 2 µl of sample was diluted 20–50 × and quantified with Trypan Blue on a hemocytometer. Depending on the experiment, a 14 µl nuclei solution containing the desired amount of nuclei in the loading buffer was then combined with 1 µl of 75 µM short SBS oligo (Additional file 5: Table S8).

The 10× Chromium was then run with the custom nuclei solution as per the manufacturer's instructions (10× Document CG000209 Rev D) with the following adaptations. In step 2.4e, during GEM aspiration and transfer, 100 µl GEM volume was split into two tubes, with one receiving 10 µl and the other 90 µl (henceforth referred to as 10% and 90% samples). In step 2.5.a, GEM incubation cycles were limited to 6. For Pre-PCR wash elution (Step 3.2.j), the 10% sample was eluted in 8.5 µl whereas the 90% sample was eluted in 32.5 µl. For step 3.2.n, the 10% sample had 8 µl transferred to a new strip, while the 90% sample had 32 µl transferred to a new strip. At step 4.1.b, the sample Index PCR mix was split with 11.5 µl and 46 µl being combined with the 10% and 90% samples, respectively. For step 4.1.c, 1 µl and 2 µl of a 10 µM i7 TruSeq primer was used, respectively. For step 4.1.d, 8 and 7 PCR cycles were used, respectively. Libraries were then checked for quality and quantified by Qubit DNA HS assay (Agilent Q32851) and TapeStation D5000 (Agilent 5067–5589) following the manufacturer's instructions. Libraries were then diluted and sequenced on a NextSeq 500 Mid flow cell or a NovaSeq 6000 S4 flow cell (Illumina Inc.).

txci-ATAC-seq using human lung, mouse lung, and mouse liver tissue samples

Flash-frozen nuclei isolated from human lung, mouse lung, and mouse liver tissues were thawed, washed, and filtered following the procedures described in “Coupling barcoded transposition with standard 10× protocol”, and then resuspended in 150 µl PBSB (PBS containing 0.04% BSA). To count nuclei, we added 1.5 µl of 300 µM DAPI to 150 µl of PBSB containing nuclei for a final concentration of 3 µM DAPI, and incubated the nuclei on ice for 5 min. Then, we loaded 10 µl on a Countess Cell Counting Chamber Slide to count the nuclei with Countess II Automated Cell Counter.

After counting nuclei, we diluted the samples with PBSB to a concentration of 2857 per µl and mixed 7 µl of nuclei solution (20,000 nuclei) with 13 µl of TBS (see “[Blocking barcode-swapping](#)” for details) for each well. This 20 µl nuclei/transposition mixture was then added to each well of the iTSM plate pre-loaded with 5 µl of barcoded Tn5 per well (see “[Sample multiplexing](#)” for details) to make a total volume of 25 µl reaction per well. For native samples shown in Fig. 4, 20,000 mouse nuclei were added to each well from

rows A to F. But for rows G and H, 10,000 mouse nuclei were mixed with 10,000 human nuclei and then transferred to each well to estimate the empirical collision rate for each sample. For WT and *CC16^{-/-}* lungs shown in Fig. 5, each well was loaded with 20,000 nuclei. The well IDs for each sample in each experiment are specified in Additional file 5: Table S6. After loading nuclei, the iTSM plate was sealed and briefly shaken at 1000 rpm for 1 min on a pre-chilled thermomixer. The barcoded transposition was performed at 37 °C for 1 h on a thermocycler with a heated lid at 47 °C. At the end of incubation, the plate was briefly centrifuged at 500 r.c.f for 10 s and then chilled on ice for 5 min to stop the transposition reaction. After quenching enzyme activity, the nuclei were pooled into a 12-tube strip and then transferred to a 15-ml conical tube preloaded with 400 µl tagmentation washing buffer (TMG, which contains 10 mM Tris acetate pH 7.8 (Boston BioProducts, Cat. BB-2412), 5 mM magnesium acetate (Sigma, Cat. 63,052-100ML), and 10% (v/v) glycerol (VWR, Cat. RC3290-32) diluted in nuclease-free water. Subsequently, we added 50 µl/well of TMG to the first row of the plate and pipetted them throughout the whole plate to wash out the residual nuclei remaining in the plate. After washing the last row of the plate, the TMG was transferred to the same conical tube that was used to collect the barcoded nuclei. The pooled nuclei were then centrifuged at 500 r.c.f for 10 min in a pre-chilled swinging-bucket centrifuge at 4 °C. After aspirating the supernatant, the nuclei were resuspended in 500 µl TMG and then transferred to a 1.5-ml LoBind tube through a 40 µm Flowmi Cell strainer. The nuclei suspension was then centrifuged at 500 r.c.f for 5 min in a pre-chilled fixed-angle centrifuge at 4 °C. After centrifugation, 400 µl of supernatant was removed. The 100 µl of supernatant left from the first aspiration was then carefully removed by pipetting with a P200 pipette tip to avoid disturbing the nuclei pellet. The nuclei were resuspended with 30 µl of loading buffer supplemented with 5 µM short SBS oligo (see Additional file 5: Table S8 for the oligo sequence). The loading buffer was prepared as described above in the “txci-ATAC-seq using brain tissue samples” section (final concentrations: 10% (v/v) glycerol, 20 mM NaCl, 10 mM Tris-HCl pH 7.5, 0.02 mM EDTA, 0.2 mM DTT, 10 mM Tris acetate pH 7.6, 5 mM magnesium acetate, and 10% (v/v) dimethyl formamide). After counting nuclei using a hemocytometer (see “Nuclei isolation of cell lines” for details), the volume of solution, containing the appropriate number of nuclei, was taken and diluted with the loading buffer supplemented with 5 µM short SBS oligo to make a total volume of 15 µl, which was subsequently used as an input into the 10× Chromium Controller. The GEM generation, Barcoding, and Post GEM Incubation Cleanup were performed following steps 2 and 3 described in the 10× Chromium Next GEM Single Cell ATAC protocol (v1.1, Document No. CG000209 Rev D) except for step 2.5, in which 8 cycles were used for GEM incubation. For Sample Index PCR (step 4.1), we substituted the Single Index N Set A (10× Genomics) with 25 µM i7 TruSeq primer containing an 8 bp custom barcode (Additional file 5: Table S7) and added 2.5 µl of customized i7 primer to each 10× library. The PCR was performed following the 10× protocol shown in Step 4.1 but with 5 total cycles. The Double Sided Size Selection was then conducted as described in Step 4.2 shown in the 10× protocol. Following the size selection, the txci-ATAC-seq libraries were quantified by Qubit 1X dsDNA HS Assay Kit (Invitrogen, Cat. Q33231) and run on a 6% PAGE gel to check the library quality. To balance nucleotide diversity of the fixed sequence at the Tn5Merev region in Read 2, we pooled these libraries with 5%

of bulk ATAC libraries (from an unrelated experiment) and sequenced them on a Next-Seq 550 Sequencer (Illumina Inc.) using a High Output Kit with following cycles: Read 1, 51 cycles; i7 index, 10 cycles; i5 index, 16 cycles; and Read 2, 78 cycles. The txci-ATAC-seq library only has 8 bp of i7 barcode, but we ran 10 cycles in i7 index to accommodate the barcode length of the bulk ATAC libraries. In cases where txci-ATAC-seq libraries are sequenced alone, we recommend either spiking in an appropriate amount of PhiX as per the manufacturer's instruction or performing dark cycles for the cycles from 9 to 27 in Read 2.

Phased-txci-ATAC-seq

To decouple sample processing from library preparation, the nuclei freshly isolated from WT and CC16^{-/-} mouse lungs (see “[Nuclei isolation of human lung, mouse lung, and mouse liver tissue](#)” for details) were diluted in NFB (see [Nuclei isolation of cell lines](#)) at 3175 nuclei/μl. For each sample, 6.3 μl of diluted nuclei (20,000 nuclei) were added to each well of an 8-tube strip for a total of 8 wells. Then, the nuclei were flash-frozen in liquid nitrogen and transferred to -80 °C for storage. The paired WT and CC16^{-/-} samples were processed together but each pair was processed on a separate day. On the designated library preparation day, the nuclei flash-frozen in the tube strips were thawed on ice and 13.7 μl of transposition buffer (which contains 12.5 μl of 2X Illumina Tagment DNA Buffer, 0.7 μl of 10 × PBS, 0.25 μl of 1% Digitonin, 0.25 μl of 10% Tween-20) was added to each well containing nuclei followed by adding 5 μl of 500 nM pre-indexed Tn5 transposase per well. Then, the barcoded transposition reaction was performed on all six samples simultaneously by incubating at 37 °C for 60 min. Since each sample was distributed into 8 wells, a total of 48 Tn5 barcodes were used. As described above in the txci-ATAC-seq protocol, the barcoded nuclei were then cooled down on ice, pooled, washed, and loaded on the 10× Chromium Controller with either 50,000 or 100,000 nuclei in a lane. The well IDs of Tn5 barcodes assigned to each sample are shown in Additional file 5: Table S6, and the TruSeq i7 index used for each loading input is provided in Additional file 5: Table S7.

Data processing and analysis

Raw code for the brain analysis is available at <https://github.com/adeylab/txci-atac>. Raw code for the cell line and lung/liver datasets is available at <https://github.com/cusanovichlab/txciatac>. The specific programs (and their version) used in data analyses were as follows: bcl2fastq (v2.19.0 for brain analysis and v2.20.0.422 for the other samples, Illumina Inc.), Trimmomatic (v0.36) [69], SAMtools and tabix (v1.7 for brain analysis and v1.10 for the other samples) [70, 71], BWA-MEM (v0.7.15-r1140) [72], Bowtie2 (v2.4.1) [73], Perl (v5.16.3) [74], MACS2 (v.2.2.7.1 for brain analysis and v2.1.2 for the other samples) [75], bedtools (v2.28.0) [76], Python (2.7.13 [77] and 3.6.7 [78]), PyPy (5.10.0), pybedtools (0.7.10) [79], R (v4.1.1) [80], cisTopic (v0.3.0) [15], Cicero (v1.3.4.10) [18], Signac (v1.0.0 for brain analysis and v1.5.0 for the other samples) [19], Presto (v1.0.0) [22], chromVAR (v1.16.0) [23], Seurat (v4.1.0) [29], corrplot (v0.92) [81], LIGER (v1.0.0) [82], uwot (v0.1.8) [83], Harmony (v1.0) [84], irlba (v2.3.5) [85], mclust (v5.4.9) [86], bap2 [40], edgeR (v3.40.0) [87], rGREAT (v2.0.2) [88], KEGGREST (v1.38.0) [89], BCFtools

(v1.15.1) [90], GATK (4.3.0.0) [91], MOODS (1.9.4) [92], ggplot2 (v3.3.5) [93], and ComplexHeatmap (v2.5.5) [94].

Computational analysis of brain samples

Preprocessing for brain tissues

After sequencing, data was converted from bcl format to FastQ format using bcl2fastq with the following options “–with-failed-reads”, “–no-lane-splitting”, “–fastq-compression-level=9”, and “–create-fastq-for-index-reads”. Data were then demultiplexed, aligned, and de-duplicated using the in-house scitools pipeline [95]. Briefly, FastQ reads were assigned to their expected primer index sequence allowing for sequencing error (Hamming distance ≤ 2) and indexes were concatenated to form a “cellID”. Reads that could be assigned unambiguously to a cellID were then aligned to reference genomes. Paired reads were first aligned to a concatenated hybrid genome of hg38 and GRCm38 (“mm10”, Genome Reference Consortium Mouse Build 38 (GCA_000001635.2)) with BWA-MEM. Reads were then de-duplicated to remove PCR and optical duplicates by a Perl script aware of cellID, chromosome number, read start coordinate, read end coordinate, and strand. From there, the putative single-cells were distinguished from debris and error-generated cellIDs by both unique reads and percentage of unique reads.

Barnyard analysis for brain tissues

With single-cell libraries distinguished, we next quantified contamination between nuclei during library generation. We calculated the read count of unique reads per cellID aligning to either human reference or mouse reference chromosomes (Additional file 1: Fig. S3b). CellIDs with $\geq 90\%$ of reads aligning to a single reference genome were considered bona fide single cells. Those not passing this filter were considered collisions. The collision rate was estimated using the equation in [14] to account for cryptic collisions (two cells from the same species). Bona fide single-cell cell IDs were then split from the original FastQ files to be aligned to the proper hg38 or mm10 genomes with BWA-MEM as described above. Human and mouse assigned cellIDs were then processed in parallel for the rest of the analysis. After alignment, reads were again de-duplicated to obtain proper estimates of library complexity.

Dimensionality reduction for brain tissues

Pseudo-bulked data (agnostic of cellID) was then used to call read pile-ups or “peaks” via MACS2 with the option “–keep-dup all”. Narrowpeak bed files were then merged by overlap and extended to a minimum of 500 bp for a total of 350,261 peaks for human and 292,304 peaks for mouse. A scitools Perl script was then used to generate a sparse matrix of peaks \times cellID to count the occurrence of reads within peak regions per cell. FRiP was calculated as the number of unique, usable reads per cell that are present within the peaks out of the total number of unique, usable reads for that cell for each peak bed file. Tabix-formatted files were generated using samtools and tabix. The count matrix and tabix files were then input into a SeuratObject for Signac processing. We performed LDA-based dimensionality reduction via cisTopic with 28 and 30 topics for human and mouse cells, respectively. The number of topics was selected after generating 25 separate models per species with topic counts of 5,10,20–30,40,50,55,60–70 and selecting the

topic count using `selectModel` based on the second derivative of model perplexity. Cell clustering was performed with Signac “`FindNeighbors`” and “`FindClusters`” functions on the topic weight \times cellID data frame. For the “`FindClusters`” function call, resolution was set to 0.01 and 0.02 for human and mouse samples, respectively. The respective topic weight \times cellID was then projected into two-dimensional space via UMAP by the function “`umap`” in the `uwot` package. To check for putative doublets within species, we then ran `scrublet` analysis and removed the `scrublet`-identified doubles from further analysis [17]. A second iteration of sub-clustering was performed on each cluster to better ascertain cell type diversity. This was done as described above with the data subset to just the cells within the respective cluster for both `cisTopic` model building and UMAP projection. Resolution per subcluster was set post hoc based on cell separation in UMAP projection. CCANs and the resulting gene activities were generated through the Signac wrapper of Cicero. Genome-wide accessibility of known TF motifs was calculated per cell using the JASPAR database (release 8) [96] via `chromVAR`.

Cell type identification for brain tissues

For cell type identification, we used previously existing single-cell RNA datasets of the human M1 cortex [97] and mouse whole cortex and hippocampus [98, 99]. We applied the Signac label transfer strategy between the annotated single-cell RNA with our gene activity scores at the level of our sub-clustered cell groups. For cell type refinement, we plotted the average gene activity score per subcluster for a set of RNA-defined marker genes, as well as markers defined within our datasets on the gene activity scores using the Signac “`FindMarkers`” function as described above. Subcluster dendrograms were generated by using base R functions `dist` and `hclust` through running Z-scored average gene activity on internally-defined markers and based on “`ward.D2`” clustering of Euclidean distance. The resultant dendrogram was used for both pre-defined and internally defined marker sets. Results were plotted via `ComplexHeatmap`.

TF marker ranking

TFs were ranked for specificity across sub-clusters, based on combined motif accessibility (generated through `chromVAR`) and gene activity (generated through `cicero`). AUC values were determined per cluster via the Wilcoxon test as reported by the “`wilcoxauc`” function in `Presto`. An average AUC of motif accessibility and gene activity was used for ranking TFs. A set of top 5 markers per sub-cluster was filtered for duplicates and then plotted via `ComplexHeatmap`.

Comparison across scATAC-seq mouse brain datasets

FastQ files for sciATAC-seq, sciMAP, snATAC-seq, dscATAC-seq, and s3-ATAC-seq were downloaded via the SRA toolkit (SRX9850743, SRX9850744, GSM2668124, GSE123581, GSM5289637, respectively); 10 \times scATAC-seq v1 and v2 chemistries FastQ files were obtained through the 10 \times Genomics website. Files were then demultiplexed following the original author’s instructions to generate a `scitools` analogous cellID and were processed through the `scitools` pipeline as described above. Briefly, after alignment to a consistent mouse reference genome (GRCm38), files were treated to de-duplication in parallel before merging. For each dataset, cellIDs were filtered to those with at least

1000 unique reads and then merged into a single bam file. Peaks were called as previously described, resulting in 344,258 regions of accessibility. Per cell, FRiP was calculated using this peak set. TSS enrichment values were calculated for all cells using the method established by the ENCODE project (<https://www.encodeproject.org/data-standards/terms/enrichment>), whereby the aggregate distribution of reads ± 1000 bp centered on the set of TSSs generates 100 bp windows at the flanks of the distribution as the background and then the maximum window centered on the TSS is used to calculate the fold-enrichment over the outer flanking windows. Signac was then used to generate a SeuratObject as described above and a genomic coverage plot on aggregated technology signals was generated through the Signac function “CoveragePlot” as described previously for a pan-excitatory neuronal marker *Slc17a7* [21]. The number of ATAC fragments mapping in annotated genes or 2 kb upstream (promoter regions) was calculated via the Signac function “GeneActivity”, normalized by read depth per cell, and aggregated across methods. Transcripts greater than 500 kb were filtered out. Average gene and promoter coverage was then correlated across methods in a pairwise manner via Spearman correlation with R base function “cor”. Data were plotted via corrplot R package. The integration across technologies was performed using LIGER, which was chosen for its performance as a top-performing method from a recent integration comparative study [100]. Briefly, the counts matrix was binarized and filtered to peaks with at least 50 cells showing accessibility, and cells were filtered to those with 500 or more peaks accessible. 3695, 4781, 3031, 6492, 906, 4492, 8295, and 38,605 scATAC profiles passed filters for sciATAC-seq, sciMAP, snATAC-seq, dscATAC-seq, tenxv1, tenxv2, s3-ATAC-seq, and txci-ATAC-seq, respectively. The top 50,000 peaks were chosen based on the “vst” method with the Signac function “FindVariableFeatures”. The remaining peaks and cells were then integrated with LIGER functions “RunOptimizeALS” with $k=20$, $\lambda=5$, and split by method. LIGER function “RunQuantileNorm” was run and then the Signac function “RunUMAP” was used. The resulting integration was then plotted with Seurat “DimPlot” function for both cell type and method (Fig. 3e).

Computational analysis of human lung, mouse lung, and mouse liver tissue samples

There were some deviations from the analysis of the brain samples, which are detailed below.

Preprocessing

Fastq files were generated using bcl2fastq with the following options: “-ignore-missing-bcls”, “-no-lane-splitting”, and “-create-fastq-for-index-reads”. Then, we modified the fastq files by attaching the first 8 bp (Tn5 barcodes) of Read 2 to the header and removing the first 27 bp (8 bp of Tn5 barcodes + 19 bp of Tn5 mosaic end) from Read 2 with a custom python script. Barcodes that did not perfectly match any of the expected barcodes were converted to the closest matching barcode if the edit distance was no greater than 2. Barcodes matching more than 1 expected barcode after correction were removed. After barcode correction, we demultiplexed samples based on a combination of Tn5 barcodes and i7 sample indices and generated a combined barcode for each read by concatenating the i7 sample index, 10 \times bead barcode, and Tn5 barcode. Next, we removed the sequence adapters and low-quality reads using trimmomatic with following

parameters: “LEADING:3; TRAILING:3; SLIDINGWINDOW:4:10; MINLEN:20” and then mapped the trimmed reads to a hybrid hg38/mm10 reference genome using Bowtie2 with a maximum fragment length of 2000 pb (-X 2000) and 1 base trimmed from the 3' end of each read (-3 1). Following mapping, only the reads confidently ($\text{MAPQ} \geq 10$) aligned to the assembled nuclear chromosomes and in proper pairs (determined by “-f3” and “-F12” options in SAMtools) were preserved for downstream analysis. To eliminate PCR duplicates, we removed all fragments that possessed the same combined barcode and identical start and end coordinates, keeping a random representative read for each end of the molecule using a custom script.

Peak calling

The deduplicated bed files were used to call peaks with MACS2, considering a 200 bp window centered on the read start using the parameters “-nomodel -keep-dup all -ext-size 200 -shift -100”. Because each peak may have multiple summits and will therefore be listed multiple times in the resulting peak bed file, the peaks output from MACS2 were then merged into a single peak set for each sample using bedtools “merge”. The consolidated peaks were then intersected with the ENCODE blacklist (mm10 [101] or hg38 ENCFF356LFX) to remove signal-artifact regions using bedtools “intersect” with “-v” option.

Calculation of ATAC-seq QC metrics

FRiDHS

The FRiDHS score was determined using orthogonal peak references identified in DNase-seq data. The GM12878 DHS peaks combined the two replicates of narrowPeak-formatted files obtained from the ENCODE consortium (ENCSR000EMT). The CH12.LX DHS peaks combined the two replicates of narrowPeak-formatted files obtained from the ENCODE consortium (ENCSR000CMQ). The mouse lung DHS peaks combined the three replicates of narrowPeak-formatted files obtained from the ENCODE consortium (ENCSR000CNM). The mouse liver DHS peaks combined the 14 replicates of narrowPeak-formatted files obtained from the ENCODE consortium (ENCSR000CNI). The human lung DHS peaks combined the narrowPeak-formatted files obtained from two separate DNase-seq data but from the same individual (ENCODE Donor Accession: ENCDO845WKR; ENCODE Experiment Accession: ENCSR164WOF and ENCSR058VBM). The overlapping peaks between replicate bed files were consolidated using bedtools “merge”, and the peaks overlapped with ENCODE blacklist (mm10 [101] or hg38 ENCFF356LFX) were removed using bedtools “intersect” with “-v” option. We removed the reads mapping to the non-nuclear genome and performed deduplication before calculating FRiDHS. The reads overlapping the DHS peak reference were counted using the “BedTool.intersect” function from pybedtools with “u = True”.

TSS enrichment

The human and mouse TSS coordinates were obtained from the Gencode human reference v39 [102] and Gencode mouse reference vM23 [103], respectively. To build TSS references, we first collected the most upstream base (accounting for strand) of each transcript using a custom R script, and then only the TSSs of gene types and transcript

types listing the following terms were included: “protein_coding”, “lncRNA”, “IG_C_gene”, “IG_D_gene”, “IG_J_gene”, “IG_LV_gene”, “IG_V_gene”, “IG_V_pseudogene”, “IG_J_pseudogene”, “IG_C_pseudogene”, “TR_C_gene”, “TR_D_gene”, “TR_J_gene”, “TR_V_gene”, “TR_V_pseudogene”, and “TR_J_pseudogene”. We also excluded transcripts with a tag of “readthrough_transcript” or “PAR”. These filters were similar to the filtering strategy used by the 10× single-cell ATAC-seq pipeline [104]. The TSS enrichment score for each cell was calculated using the TSSEnrichment function in the Signac package.

Estimated complexity

The nuclear genome mapped reads and deduplicated reads were used to estimate the complexity for each cell using the same calculation as Picard [105] implemented in R.

Collision rate estimation

For each combined barcode, we quantified the number of deduplicated reads mapping to the human and mouse genome and filtered out the combined barcodes with fewer than 1000 total reads. The collision barcodes were determined as the cell barcodes that had more than 10% of reads aligned to the minor genome. Since the cell doublets can be generated by either two cells from the same species or cells from distinct species, the observed collisions only reflect approximately half of the collision events that in fact occur in the experiment. To this end, we estimated the actual collision rate using the equation in [14].

Dimensionality reduction and clustering

An iterative peak-calling strategy was used to perform dimensionality reduction and cluster cells. The first round of clustering was performed with a pseudo-bulk peak reference, which was identified by calling peaks on deduplicate reads from identified cells (≥ 1000 reads). Then, a binarized peak (column) by cell (row) matrix was generated by scoring the peaks defined in the previous step for overlap with reads from each cell. The low complexity cells and features were removed using Signac “CreateChromatinAssay” function by setting “min.cells = 50 and min.features = 200” for mouse samples and setting “min.cells = 15 and min.features = 200” for human samples followed by filtering out the cells considered as outliers for QC metrics (DHS region reads $> 20,000$, FRiDHS < 0.2 and TSS enrichment score < 2). Potential cell doublets were identified by performing a modified version of the Scrublet workflow [17] on each txci-ATAC-seq library separately. In brief, we transformed the filtered cell/peak matrix with the term-frequency inverse-document-frequency (TF-IDF) algorithm by computing $\log(\text{TF} \times \text{IDF})$ as described in [106] and then calculated the first 30 components for PCA using the irlba R package. Simulated cell doublets were created by randomly sampling 50% of observed cells from the original matrix and summing them with another 50% of randomly sampled cells. The matrix of simulated doublets was then binarized and transformed with the same TF-IDF implementation. Subsequently, we projected the transformed doublets into the PCA space generated by the observed data and performed L2-normalization on the resulting matrix including both observed and simulated cells with Seurat “L2Dim” function. The L2-normalized reduction was then used to compute the fraction of simulated doublet neighbors for each cell using Seurat “FindNeighbors” function with dimensions 2 to

30 and setting “k.param” as 129 (mouse lung nuclei from the 100,000 lane), 166 (mouse lung nuclei from the 200,000 lane), 120 (mouse liver nuclei from the 100,000 lane), 147 (mouse liver nuclei from the 200,000 lane), 62 (human lung nuclei from the 100,000 lane), and 74 (human lung nuclei from the 200,000 lane). We derived the “k.param” values using the k_{adj} equation in Scrublet. Finally, a doublet score was calculated for each cell using the appropriate equations described in Scrublet.

Assuming a bimodal distribution, a threshold for doublet scores was calculated with the simulated cells by identifying the boundary between the doublets incorporating highly similar cells (“embedded”) and the doublets of dissimilar cells (“neotypic”) using the mclust R package [86] with less than 5% uncertainty that the doublets were classified into the “neotypic” category. After removing the doublets detected, we computed the latent semantic indexing (LSI) matrix by running singular value decomposition (SVD) on the TF-IDF normalized matrix using Signac and then clustered cells using Seurat “FindNeighbors” function with the dimensions of reduction from 2 to 30 followed by Seurat “FindClusters” implementing the SLM algorithm with default resolution. For tissues with replicates, the LSI matrix was integrated by individual with Harmony [84] prior to cell clustering.

Each cell cluster from this first round of clustering was then used to identify peaks independently, and all cluster peaks were merged into a single reference set. Subsequently, a second round of clustering was performed using this updated peak set. With the same workflow, we used in the first round of clustering, a binarized count matrix generated with cluster-identified peaks was created and used to perform normalization, dimension reduction, integration (for tissues with replicates), and clustering, except that the resolution parameter used to determine the community size was set differently for each tissue (i.e., 0.8, 0.2, and 0.3 were used for mouse lung tissue, mouse liver tissue, and human lung tissue, respectively). Regarding liver samples, we decided to consolidate clusters 0, 7, 8, and 9 because of no visible separation between them in 2D UMAP space. For visualization purposes, the data was projected into a two-dimensional space via Seurat “RunUMAP” function with 30 dimensions (excluding the first component, which represented the sequencing depth).

Cell type annotation

The cell types associated with each cluster were predicted by label transfer using publicly available sc/snRNA-seq and sci-ATAC-seq data. Only the cell types including at least 50 cells in the reference dataset were used to infer the cell types in the query dataset. To annotate cell types with transcriptome data, we used previously published data from steady state mouse liver, mouse lung, and healthy human lung samples to construct an “integrated” reference for each tissue in each species using the Seurat scRNA-seq integration pipeline. In all cases, the 5000 most variable genes across reference samples were selected to find integration “anchors”. The mouse lung reference was built by integrating three samples (a scRNA-seq sample and two replicate samples of snRNA-seq) from a single study [30]. The mouse liver reference was created by integrating samples generated with three different protocols (snRNA-seq and scRNA-seq using cells isolated via either ex vivo or in vivo enzymatic digestion method) from a single study as well (<https://www.livercellatlas.org/download.php>; [31]). The human lung reference was established

by integrating two scRNA-seq datasets obtained from two independent studies [32, 33]. After creating RNA-seq references, we estimated transcriptional activity across the genes selected for integration by quantifying the txci-ATAC-seq counts in both the 2-kb region upstream and the gene body of each gene using the Signac “GeneActivity” function. The prediction of cell type was then achieved by performing canonical correlation analysis on the gene activity scores calculated from ATAC-seq data along with the integrated scRNA-seq reference using Seurat “FindTransferAnchors” function followed by transferring annotations from reference to query cells using “TransferData” function in which the 2nd to 30th components of the LSI matrix calculated on ATAC-seq data was used to compute the weights of the local neighborhood of anchors.

For annotating cells with a chromatin reference, we downloaded the fastq files of mouse sci-ATAC-seq data from [3] (GEO accession number: Lung RepA, GSM3034631; Lung Rep B, GSM3034632; Liver, GSM3034630) and mapped them to the mm10 reference genome using Bowtie2. In terms of the human reference, the cell by bin (5 kb) matrices were downloaded from four lung samples [34] (GEO accession number: GSE165659) and binarized prior to cell type prediction. To ensure that the same features were measured in the reference and query datasets, we summarized the reads from txci-ATAC-seq to either the peaks identified from the pseudo-bulk sci-ATAC-seq data (Mouse) or 5-kb genomic windows (human) for each query cell and only retained the features that were detected in at least 50 (mouse) or 15 (human) cells in both datasets. The label transfer was performed using Seurat “FindTransferAnchors” function with `reference.reduction = "lsi"` and `reduction = "lsiproject"` followed by “MapQuery” function with `reference.reduction = "lsi"`.

Cell type labels transferred to each cell were aggregated by applying a majority vote strategy to each cluster. The top cluster-specific marker genes identified (based on gene activity scores) using the Seurat “FindAllMarkers” function were interactively evaluated for cell-type-specific expression using online scRNA-seq data browsers of mouse lung [30, 35] and liver [31] to arrive at our final cell type annotations. The color palette used for cell types in the UMAPs was selected from colors available in the ArchR package [107].

Simulation of cell recovery

The cell recovery was modeled using the Poisson distribution with the assumption that (1) a total of 100,000 bead-containing droplets were generated by the 10× microfluidic system based on the observed mean nuclei per droplet and (2) a 35% cell loss occurs, attributable to the factors such as the dead volume, cell bursting, and encapsulation into empty droplets (consistent with guidelines provided by the 10× Chromium Next GEM Single Cell ATAC protocol (v1.1, Document No. CG000209 Rev D). We simulated the nuclei loading inputs ranging from 1000 to 1.5 million. The mean number of nuclei per droplet (λ) was defined as the following equation:

$$\lambda = \frac{N}{100,000} \times 0.65, \text{ in which } N \text{ is the number of nuclei loading input.}$$

The collision rate in a given droplet was defined as the probability of at least one duplicate barcode present in that droplet (P) and was calculated with the following equation:

$$P = 1 - \frac{m \times (m - 1) \times (m - 2) \cdots \times (m - n + 1)}{m^n} \text{ if } n \leq m$$

$$P = 1 \text{ if } n > m$$

The m represents the number of Tn5 barcodes used for pre-indexing, and n represents the number of cells within a given droplet. To evaluate the effectiveness of usable cell recovery in each hashing strategy, we discarded the droplets containing more than one cell for the cellular hashing simulation. However, for molecular hashing, the droplets with a Tn5 barcode collision rate exceeding 10% were considered as undemultiplexable. As a result, the droplets containing more than 4 cells in a 96-pre-indexing system, those containing more than 3 cells in a 48-pre-indexing system, and those containing more than 2 cells in a 24-pre-indexing system were excluded from the count of usable cells.

Comparison to dsciATAC

The fastq files of dsciATAC datasets at 8000, 16,000, 40,000, 80,000, and 160,000 nuclei inputs tagged with 48 Tn5 barcodes were downloaded from the Gene Expression Omnibus (GEO) database under accession number GSE123581. The sequenced barcodes were parsed and assigned to the closest known sequence using a custom Python script allowing for up to one mismatch per 6-mer (Tn5) or 7-mer (Bio-Rad SureCell bead) barcode. The reads with corrected barcodes were then trimmed and mapped to a hybrid hg38/mm10 reference genome as described in the “[Preprocessing](#)” section. To consolidate the bead barcodes within the same droplet for both dsciATAC and txci-ATAC data, we added the bead barcodes to a BAM tag, identified bead multiplets using the `bap2` package, and then merged the inferred bead barcodes using a custom Python script that accounted for Tn5 barcodes. After this, the reads were filtered and deduplicated as detailed in the “[Preprocessing](#)” section and subsequently used to quantify the species-specific mapping for each barcode ID (a combination of bead barcode, after merging bead barcodes from bead multiplets, and Tn5 barcode). To identify cell-associated barcodes within both dsciATAC and txci-ATAC datasets, we tested two distinct methods: (1) implementing a universal cutoff of 1000 unique reads across all datasets and (2) automatically determining a cutoff with K -means clustering for each dataset independently. For the latter approach, K -means clustering was applied to the \log_{10} -transformed unique reads from the barcode ID with a minimum of 100 unique reads using a value of k equal to two and 50 random sets. The collision rate was calculated as described in the “[Collision rate estimation](#)” section considering multiplets arising from either a single species or both species. For wells containing only cells from a single species in txci-ATAC-seq data, the multiplets were detected in each tissue separately using a customized Scrublet workflow, explained in the “[Dimensionality reduction and clustering](#)” section, and then eliminated prior to counting recovered cells. To conservatively identify singlets from those wells, if the multiplet rate inferred by Scrublet was lower than the empirical collision rate established from the wells designated as a true barnyard, the latter rate was used to remove the potential multiplets from the wells with pure species.

Identification of differential peaks

An edgeR-based pseudo-bulk method was used to identify differential peaks between WT and CC16^{-/-} mouse lungs for each cell type. To do so, we aggregated the reads for all cells from the same replicate in a cluster-wise manner, which resulted in three biological replicates for each genotype per cell type. The lowly accessible peaks in each differential test were filtered out using the “filterByExpr” function with default parameters followed by calculating normalization factors with “calcNormFactors()”. Then, we estimated dispersions using “estimateDisp()” with a design matrix and “robust = TRUE” and performed hypothesis testing using the quasi-likelihood F-test. The Benjamini and Hochberg (BH) method [108] was used to control the false discovery rate (FDR).

Variant calling and filtering

The variant calling was performed using the BCFtools “mpileup” command followed by “call” command using multiallelic calling model (-m) by grouping all replicates from the same genotype (-G). Only the SNVs that met the following criteria were used for downstream analyses: a Phred-scaled quality score (QUAL) of at least 20, a sum of read depth (DP) across all three replicates of at least 10, and the same genotype in at least two of three replicates.

Motif analysis

The motif position frequency matrices obtained from the JASPAR database (version 2020) [109] were used for all motif analyses. To identify differentially active motifs between cell types, a per-cell motif activity score was computed first for each motif using chromVAR implemented in Signac. The differential testing was then performed on the chromVAR z-score for each cell type using Seurat “FindAllMarkers” function. For motif enrichment analysis, we applied the Signac “FindMotifs” function to all differentially accessible peaks per cell type to identify the enriched motifs using a GC-content-matched set of peaks created from the accessible peaks as a background. The multiple testing correction was performed with the BH procedure [108].

To identify the SNV-driven gains and losses in motif matching, we first generated alternative DNA sequences over the differentially accessible peaks from SNV hotspots (mm10 chr8:68,000,000–93,000,000 and mm10 chr19:16,500,000–26,500,000) for WT and CC16^{-/-} samples by replacing the reference bases at the variation sites with the hotspot SNVs identified in each genotype using GATK “FastaAlternateReferenceMaker” tool. Then, we matched the motifs against the alternative DNA sequences using the MOODS package with a *p*-value cutoff of 0.0001. To identify functional motifs capable of accounting for the chromatin accessibility changes, we tested for associations between the log₂(fold-change) of differentially accessible peaks and the gain or loss of motifs in CC16^{-/-} background using the student’s *t*-test and controlling for multiple testing with the BH method [108]. When counting differences in accessibility that might be explained by specific motif-disrupting SNVs, we only considered the instances that exhibit a coherent change in chromatin accessibility with the overall motif effect to be explanatory (i.e., depending on whether gained/lost motifs are positively or negatively associated with

peak accessibility, each of the potential explanatory instances for that motif also needs to display concordant increases or decreases in accessibility to be counted as explanatory).

Functional analysis

The KEGG pathways enrichment analysis was performed by applying rGREAT on all differential peaks identified for each cell type using both binomial and hypergeometric tests. To control both tests, we used a previously implemented two-threshold approach [110] to define the significant pathways by requiring a stringent 10% FDR threshold for at least one test, but allowing for a more relaxed threshold (unadjusted p -value of 0.05) for the other test. The gene sets of KEGG pathways were retrieved using the KEGGREST package.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-03150-1>.

Additional file 1: Fig S1. Collision rates of standard 10X protocol coupled with combinatorial indexing. **Fig S2.** Exponential amplification during GEM PCR enables deconvolution of cells in the same droplet. **Fig S3.** Evaluating the performance of txc1-ATAC-seq on brain samples. **Fig S4.** Consistency of scATAC-seq mouse brain datasets. **Fig S5.** Quality metrics of txc1-ATAC-seq in the native lung and liver samples with loading 100,000 or 200,000 nuclei. **Fig S6.** Cell type annotation of mouse lung samples. **Fig S7.** Cell type annotation of mouse liver samples. **Fig S8.** Cell type annotation of human lung sample with label transfer. **Fig S9.** Comparison of prediction accuracy between txc1-ATAC-seq and sci-ATAC-seq in mouse lung cells. **Fig S10.** Characterization of cellular heterogeneity in human lung tissue. **Fig S11.** txc1-ATAC-seq are robust to batch effects. **Fig S12.** Comparison of txc1-ATAC with dsciATAC using a bead-merging strategy. **Fig S13.** Phased-txc1-ATAC-seq improves multiplexing capability without sacrificing data quality. **Fig S14.** Chromatin accessibility changes induced by CC16^{-/-} deficiency in mouse lung. **Fig S15.** Functional analysis and regulatory variant identification in CC16 deficient mouse. **Fig S16.** SNP-driven differences in motif usage alter chromatin accessibility. **Fig S17.** Comparison of empirical and theoretical cell recovery between molecular and cellular hashing strategies. **Fig S18.** Examination of efficiency for in-droplet and sample index PCR.

Additional file 2: Table S1. Differentially accessible peaks identified between CC16^{-/-} and WT samples for each cell type. This table is provided as a separate file. Column 1: Cell type in which the test was performed. Column 2: Peak region. Columns 3-5: Log₂(fold-change), raw p -value, and FDR-adjusted p -value, respectively. Columns 6-11: Log₂-transformed counts per million (CPM) for each sample, computed using the normalized library sizes. The CPM values for WT samples are shown in columns 6-8, and the CPM values for CC16^{-/-} samples are shown in columns 9-11.

Additional file 3: Table S2. Enriched motifs in more accessible and less accessible peaks in response to CC16 deficiency for each cell type. This table is provided as a separate file. Column 1: Cell type in which the test was performed. Column 2: Changing direction of differentially accessible peaks that were used to perform the test. Column 3: Motif ID. Column 4: Motif name. Column 5: The number of differential peaks that contain the motif identified. Column 6: The number of background peaks that contain the motif identified. Column 7: The percentage of differential peaks that contain the motif identified. Column 8: The percentage of background peaks that contain the motif identified. Column 9: The ratio of the observed frequency of the motif in differential peaks to the expected frequency calculated by the background peaks. Column 10: Raw p -value. Column 11: FDR-adjusted p -value.

Additional file 4: Table S3. KEGG pathways enriched in differential peaks between CC16^{-/-} and WT samples for each cell type. This table is provided as a separate file. Column 1: Cell type in which the test was performed. Column 2: KEGG pathway ID. Column 3: Description of KEGG pathway. Column 4: Fraction of non-gap base pairs in the genome that lie in the regulatory domain of a gene with the annotation. Column 5: Actual number of differential peaks with the annotation. Column 6: Fold enrichment of number of differential peaks with the annotation. Column 7: Uncorrected p -value from the binomial test over genomic regions. Column 8: FDR-adjusted p -value for the binomial test. Column 9: Mean absolute distance of input regions to TSS of genes in a gene set. Column 10: Actual number of genes linking to a differential peak with the annotation. Column 11: Number of genes in the genome with the annotation. Column 12: Fold enrichment of number of genes linking to a differential peak with the annotation. Column 13: Uncorrected p -value from the hypergeometric test over genes. Column 14: FDR-adjusted p -value for the hypergeometric test.

Additional file 5: Table S4. Sequences of Tn5 linker oligos. The 'N' bases shown in the Tn5ME-B sequence represent the Tn5 barcodes. **Table S5.** Tn5 barcode sequences. Column 1 shows the well ID for each well on the iTSM plate. Column 2 shows the sequences of Tn5 barcodes assigned to each well. Column 3 is the 12 numerical labels for the plate columns. Column 4 is the 8 alphabetical labels for the plate rows. This table is provided as a separate file. **Table S6.** Well IDs of Tn5 barcodes assigned to each sample or species-mixing condition in each experiment. Column 1 shows the figure number for each experiment. Column 2 shows the experimental conditions employed in each figure. Column 3 indicates the species-mixing condition. Column 4 shows the cell source of human samples. Column 5 shows the number of human nuclei loaded to each well. Column 6 shows the cell source of mouse

samples. Column 7 shows the number of mouse nuclei loaded to each well. Column 8 indicates the nuclei preparation method (Fresh vs. Frozen). Column 9 is the well ID on the iTSM plate (see Tn5 barcode sequences in Additional file 5: Table S5) assigned to each sample or barnyard experiment. This table is provided as a separate file. **Table S7.** TruSeq i7 index sequences used for each library in Sample Index PCR. Column 1 shows the index ID. Column 2 shows the oligo sequence. Column 3 indicates the barcode sequence assigned to each library shown in Column 4. **Table S8.** DNA oligonucleotides used to block barcode swapping. Each row provides the sequence of an oligo used in the barcode swapping blocking tests. The lowercase letters shown in the full SBS primer represent the barcode sequence. For Decoy DNA, the strands A and B were annealed to form a duplex DNA.

Additional file 6: Review history.

Acknowledgements

We thank J. Galina-Mehlman, R. Sprissler, and B. Fransway from the University of Arizona Genetics Core for their invaluable assistance. We thank C. Romanoski from the University of Arizona for helpful advice and insightful comments on the manuscript. We also would like to thank C. Thornton from Oregon Health & Science University for help in cell type identification of mouse and human brain tissues.

Review history

The review history is available as Additional file 6.

Peer review information

Hamish King and Veronique van den Berghe were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

H.Z., R.M.M., A.C.A., and D.A.C. conceived the study. H.Z. performed the experiments and data analysis for the txc1-ATAC-seq of the cell line barnyard, human lung, naive mouse lung and liver, CC16 knockout lungs, as well as Phased-txc1-ATAC-seq under the supervision of D.A.C. R.M.M. performed the experiments and data analysis for the txc1-ATAC-seq of human and mouse brain under the supervision of A.C.A. N.I. and J.G.L. provided the WT and CC16 knockout lungs. D.O.F. and J.J.G. provided the naive mouse lung and liver samples. F.P. provided the human lung sample. N.I., J.G.L., D.O.F., J.J.G., and F.P. aided in the development of nuclei isolation protocol for human lung, mouse lung, and mouse liver samples. The paper was written by H.Z., R.M.M., A.C.A., and D.A.C. All authors read, edited, and approved the final manuscript.

Funding

Financial support was provided by R35GM137896 (NIH/NIGMS) for D.A.C., R35GM124704 (NIH/NIGMS) and RF1MH128842 (NIH/NIMH, BICCN) for A.C.A., R35GM137910 (NIH/NIGMS) for J.J.G., T32ES007091 (NIH/NIEHS) for D.O.F., and R01HL142769 (NIH/NHLBI) for J.G.L.

Availability of data and materials

A step-by-step protocol of txc1-ATAC-seq is available at <https://doi.org/10.17504/protocols.io.dm6gp3o68vzp/v1> [111]. The datasets generated during the current study are available in the Gene Expression Omnibus repository (Brain samples: GSE245957 [112]; other samples: GSE231708 [113]). All raw sequence data for human cortex samples have been deposited in the database of Genotypes and Phenotypes (dbGaP), under accession phs003497.v1: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003497.v1.p1 or search for "phs003497.v1.p1" at <http://www.ncbi.nlm.nih.gov/gap> [114] under restricted user access (human genetic data for research use only). All source code is freely available under the MIT license (the analysis pipeline for the brain samples is available at Github: <https://github.com/adeylab/txci-atac> [115] and OSF: <https://doi.org/https://doi.org/10.17605/OSF.IO/VNPWB> [116]; the analysis pipeline for the cell line and lung/liver samples is available at Github: <https://github.com/cusanovichlab/txciatac> [117] and OSF: <https://doi.org/https://doi.org/10.17605/OSF.IO/YA7SE> [118]). The sci-ATAC-seq [119], sci-MAP [119], snATAC-seq [120], dscATAC-seq [121], and s3-ATAC-seq [122] of mouse brain datasets were downloaded from GEO accession GSM5021069, GSM5021070, GSM2668124, GSE123581, and GSM5289637, respectively. The 10x scATAC-seq v1 [27] and v2 [28] datasets of mouse brain were obtained from <https://www.10xgenomics.com/resources/datasets/fresh-cortex-from-adult-mouse-brain-p-50-1-standard-1-2-0> and <https://www.10xgenomics.com/resources/datasets/8k-adult-mouse-cortex-cells-atac-v2-chromium-x-2-standard>, respectively. The mouse lung scRNA-seq [123] datasets were downloaded from GEO accession GSE145998 and the mouse lung sci-ATAC-seq [124] datasets were downloaded from GEO accession GSM3034631 and GSM3034632. The scRNA-seq [125] and sci-ATAC-seq [124] datasets of mouse liver were downloaded from <https://www.livercellatlas.org/download.php> and GEO accession GSM3034630, respectively. The scRNA-seq datasets of human lung were obtained from <https://www.synapse.org/#Synapse:syn21041850> [126] and GEO accession GSE135893 [127]. The sci-ATAC-seq datasets of human lung were downloaded from GEO accession GSE165659 [128]. The dscATAC datasets of human K562 and mouse NIH/3T3 cell line mixture were downloaded from GEO accession GSE123581 [121].

Declarations

Ethics approval and consent to participate

All animal use was approved by the Institutional Animal Care and Use Committee of the University of Arizona or OHSU and adhered to the NIH Guide for the Care and Use of Laboratory Animals. The details of the animal study reported here are in compliance with ARRIVE guidelines.

De-identified human brain and lung tissue specimens were obtained from the Oregon Brain Bank (brain) or the Arizona Donor Network (lung). The Oregon Brain Bank and the Arizona Donor Network collected the samples under their own IRB-approved protocols and obtained appropriate consents. The studies presented here were reviewed by the OHSU IRB (for brain) and the University of Arizona IRB (for lung) and determined not to be human subjects research requiring separate IRB-approved protocols because we only had access to de-identified tissue samples.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 12 May 2023 Accepted: 20 December 2023

Published online: 22 March 2024

References

- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013;10:1213–8.
- Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, et al. A human cell atlas of fetal chromatin accessibility. *Science*. 2020;370. <https://doi.org/10.1126/science.aba7612>
- Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*. 2018;174:1309–1324.e18.
- Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol*. 2019;37:925–36.
- Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. 2015;348:910–4.
- Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. 2017;357:661–7.
- Vitak SA, Torkency KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods*. 2017;14:302–8.
- Wang K, Xiao Z, Yan Y, Ye R, Hu M, Bai S, et al. Simple oligonucleotide-based multiplexing of single-cell chromatin accessibility. *Mol Cell*. 2021;81:4319–4332.e10.
- Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol*. 2019;37:916–24.
- De Rop FV, Hulselmans G, Flerin C, Soler-Vila P, Rafels A, Christiaens V, et al. Systematic benchmarking of single-cell ATAC-sequencing protocols. *Nat Biotechnol*. 2023. <https://doi.org/10.1038/s41587-023-01881-x>.
- Kartha VK, Duarte FM, Hu Y, Ma S, Chew JG, Lareau CA, et al. Functional inference of gene regulation using single-cell multi-omics. *Cell Genom*. 2022;2. <https://doi.org/10.1016/j.xgen.2022.100166>
- Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049.
- Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods*. 2017;14:959–62.
- Bloom JD. Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ*. 2018;6: e5578.
- Bravo González-Blas C, Minnoye L, Papisokrati D, Aibar S, Hulselmans G, Christiaens V, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods*. 2019;16:397–400.
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36:411–20.
- Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst*. 2019;8:281–291.e9.
- Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell*. 2018;71:858–871.e8.
- Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. *Nat Methods*. 2021. <https://doi.org/10.1038/s41592-021-01282-5>.
- Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Grayback LT, et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature*. 2019;573:61–8.
- Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. 2007;445:168–76.
- Korsunsky I, Nathan A, Millard N, Raychaudhuri S. Presto scales Wilcoxon and auROC analyses to millions of observations. <https://doi.org/10.1101/653253>
- Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods*. 2017;14:975–8.
- Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci*. 2018;21:432–9.
- Thornton CA, Mulqueen RM, Torkency KA, Nishida A, Lowenstein EG, Fields AJ, et al. Spatially mapped single-cell chromatin accessibility. *Nat Commun*. 2021;12:1274.
- Mulqueen RM, Pokholok D, O'Connell BL, Thornton CA, Zhang F, O'Roak BJ, et al. High-content single-cell combinatorial indexing. *Nat Biotechnol*. 2021;39:1574–80.
- Fresh cortex from adult mouse brain. Datasets. 10X Genomics. Available: <https://www.10xgenomics.com/resources/datasets/fresh-cortex-from-adult-mouse-brain-p-50-1-standard-1-2-0>
- 8k adult mouse cortex cells, ATAC v2, Chromium X. Datasets. 10X Genomics. Available: <https://www.10xgenomics.com/resources/datasets/8k-adult-mouse-cortex-cells-atac-v2-chromium-x-2-standard>
- Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184:3573–3587.e29.

30. Koenitzer JR, Wu H, Atkinson JJ, Brody SL, Humphreys BD. Single-nucleus RNA-sequencing profiling of mouse lung. Reduced dissociation bias and improved rare cell-type detection compared with single-cell RNA sequencing. *Am J Respir Cell Mol Biol*. 2020;63:739–747.
31. Guilliams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell*. 2022;185:379–396.e38.
32. Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature*. 2020;587:619–25.
33. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv*. 2020;6:eaba1972.
34. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell*. 2021;184:5985–6001.e19.
35. Guo M, Morley MP, Jiang C, Wu Y, Li G, Du Y, et al. Guided construction of single cell reference for human and mouse lung. *Nat Commun*. 2023;14:4566.
36. Penkala IJ, Liberti DC, Pankin J, Sivakumar A, Kremp MM, Jayachandran S, et al. Age-dependent alveolar epithelial plasticity orchestrates lung homeostasis and regeneration. *Cell Stem Cell*. 2021;28:1775–1789.e5.
37. Bilodeau C, Shojaie S, Goltsis O, Wang J, Luo D, Ackerley C, et al. TP63 basal cells are indispensable during endoderm differentiation into proximal airway cells on acellular lung scaffolds. *NPJ Regen Med*. 2021;6:12.
38. Li J, Ning G, Duncan SA. Mammalian hepatocyte differentiation requires the transcription factor HNF-4alpha. *Genes Dev*. 2000;14:464–74.
39. Yu D-D, Jing Y-Y, Guo S-W, Ye F, Lu W, Li Q, et al. Overexpression of hepatocyte nuclear factor-1beta predicting poor prognosis is associated with biliary phenotype in patients with hepatocellular carcinoma. *Sci Rep*. 2015;5:13319.
40. Lareau CA, Ma S, Duarte FM, Buenrostro JD. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat Commun*. 2020;11:866.
41. Mango GW, Johnston CJ, Reynolds SD, Finkelstein JN, Plopper CG, Stripp BR. Clara cell secretory protein deficiency increases oxidant stress response in conducting airways. *Am J Physiol*. 1998;275:L348–56.
42. Li X, Guerra S, Ledford JG, Kraft M, Li H, Hastie AT, et al. Low CC16 mRNA expression levels in bronchial epithelial cells are associated with asthma severity. *Am J Respir Crit Care Med*. 2023;207:438–51.
43. Laucho-Contreras ME, Polverino F, Gupta K, Taylor KL, Kelly E, Pinto-Plata V, et al. Protective role of club cell secretory protein-16 (CC16) in the development of COPD. *Eur Respir J*. 2015;45:1544–56.
44. Guerra S, Vasquez MM, Spangenberg A, Halonen M, Martinez FD. Serum concentrations of club cell secretory protein (Clara) and cancer mortality in adults: a population-based, prospective cohort study. *Lancet Respir Med*. 2013;1:779–85.
45. Duong TE, Wu Y, Sos BC, Dong W, Limaye S, Rivier LH, et al. A single-cell regulatory map of postnatal lung alveologenesis in humans and mice. *Cell Genom*. 2022;2. <https://doi.org/10.1016/j.xgen.2022.100108>
46. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*. 2011;477:289–94.
47. Eisener-Dorman AF, Lawrence DA, Bolivar VJ. Cautionary insights on knockout mouse studies: the gene or not the gene? *Brain Behav Immun*. 2009;23:318–24.
48. Webb CF, Bryant J, Popowski M, Allred L, Kim D, Harriss J, et al. The ARID family transcription factor bright is required for both hematopoietic stem cell and B lineage development. *Mol Cell Biol*. 2011;31:1041–53.
49. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol*. 2018;19:224.
50. Gaublotte JT, Li B, McCabe C, Knecht A, Yang Y, Drokhyansky E, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat Commun*. 2019;10:2907.
51. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods*. 2019;16:619–26.
52. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*. 2018;36:89–94.
53. Datlinger P, Rendeiro AF, Boenke T, Senekowitsch M, Krausgruber T, Barreca D, et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat Methods*. 2021;18:635–42.
54. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res*. 2014;24:2033–40.
55. Hennig BP, Velten L, Racke I, Tu CS, Thoms M, Rybin V, et al. Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3*. 2018;8:79–89.
56. Zhang H, Rice ME, Alvin JW, Farrera-Gaffney D, Galligan JJ, Johnson MDL, et al. Extensive evaluation of ATAC-seq protocols for native or formaldehyde-fixed nuclei. *BMC Genomics*. 2022;23:214.
57. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat Biotechnol*. 2021;39:825–35.
58. Xu W, Yang W, Zhang Y, Chen Y, Hong N, Zhang Q, et al. ISSAAC-seq enables sensitive and flexible multimodal profiling of chromatin accessibility and gene expression in single cells. *Nat Methods*. 2022;19:1243–9.
59. Mulqueen RM, Pokholok D, Norberg SJ, Torkenczy KA, Fields AJ, Sun D, et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat Biotechnol*. 2018;36:428–31.
60. Liscovitch-Brauer N, Montalbano A, Deng J, Méndez-Mancilla A, Wessels H-H, Moss NG, et al. Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nat Biotechnol*. 2021;39:1270–7.
61. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. 2018;361:1380–5.
62. Iannuzzo N, Insel M, Marshall C, Pederson WP, Addison KJ, Polverino F, et al. CC16 deficiency in the context of early-life infection results in augmented airway responses in adult mice. *Infect Immun*. 2022;90: e0054821.
63. Rojas-Quintero J, Laucho-Contreras ME, Wang X, Fucci Q-A, Burkett PR, Kim S-J, et al. CC16 augmentation reduces exaggerated COPD-like disease in Cc16-deficient mice. *JCI Insight*. 2023;8. <https://doi.org/10.1172/jci.insight.130771>

64. Stripp BR, Lund J, Mango GW, Doyen KC, Johnston C, Hultenby K, et al. Clara cell secretory protein: a determinant of PCB bioaccumulation in mammals. *Am J Physiol*. 1996;271:L656–64.
65. Zhai J, Insel M, Addison KJ, Stern DA, Pederson W, Dy A, et al. Club cell secretory protein deficiency leads to altered lung function. *Am J Respir Crit Care Med*. 2019;199:302–12.
66. Saunders A, Core LJ, Sutcliffe C, Lis JT, Ashe HL. Extensive polymerase pausing during *Drosophila* axis patterning enables high-level and pliable transcription. *Genes Dev*. 2013;27:1146–58.
67. Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferrer R, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*. 2018;555:538–42.
68. Joshi N, Misharin A. Single-nucleus isolation from frozen human lung tissue for single-nucleus RNA-seq. protocols.io. Available: <https://www.protocols.io/view/single-nucleus-isolation-from-frozen-human-lung-ti-zu8f6zw> (2019).
69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–20.
70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence alignment/map format and SAM-tools. *Bioinformatics*. 2009;25:2078–9.
71. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. 2011;27:718–9.
72. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 [cited 23 Jun 2022]. doi:<https://doi.org/10.48550/arXiv.1303.3997>
73. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.
74. Wall L, Christiansen T, Orwant J. Programming Perl. "O'Reilly Media, Inc."; 2000.
75. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9:R137.
76. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
77. van Rossum G. Python Reference Manual. 1995.
78. Van Rossum G, Drake FL. Python 3 Reference Manual: (Python Documentation Manual Part 2). CreateSpace; 2009.
79. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*. 2011;27:3423–4.
80. R Core Team. R: A Language and Environment for Statistical Computing. 2021. Available: <https://www.R-project.org/>
81. Wei T, Simko V. R package "corrplot": visualization of a correlation matrix. In: (Version 0.92) [Internet]. 2021. Available: <https://github.com/taiyun/corrplot>
82. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*. 2019;177:1873–1887.e17.
83. Melville J. The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction [R package uwot version 0.1.11]. 2021 [cited 23 Jun 2022]. Available: <https://CRAN.R-project.org/package=uwot>
84. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16:1289–96.
85. Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices [R package irlba version 2.3.5]. 2021 [cited 24 Jun 2022]. Available: <https://CRAN.R-project.org/package=irlba>
86. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J*. 2016;8:289–317.
87. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40.
88. Gu Z, Hübschmann D. rGREAT: an R/Bioconductor package for functional enrichment on genomic regions. *Bioinformatics*. 2022. <https://doi.org/10.1093/bioinformatics/btac745>.
89. Tenenbaum D, Maintainer B. KEGGREST: client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). R package version 1.38.0. 2022.
90. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011. Available: <https://academic.oup.com/bioinformatics/article-abstract/27/21/2987/217423>
91. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303.
92. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics*. 2009;25:3181–2.
93. Wickham H. ggplot2: elegant graphics for data analysis. Springer Science & Business Media; 2009.
94. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32:2847–9.
95. Sinnamon JR, Torkenczy KA, Linhoff MW, Vitak SA, Mulqueen RM, Pliner HA, et al. The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Res*. 2019;29:857–69.
96. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res*. 2018;46:D1284.
97. Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature*. 2021;598:111–9.
98. Yao Z, van Velthoven CTJ, Nguyen TN, Goldy J, Sedeno-Cortes AE, Baftizadeh F, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*. 2021;184:3222–3241.e26.
99. Kozareva V, Martin C, Osorno T, Rudolph S, Guo C, Vanderburg C, et al. A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. *Nature*. 2021;598:214–9.
100. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods*. 2022;19:41–50.
101. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: identification of problematic regions of the genome. *Sci Rep*. 2019;9:9354.

102. Gencode human reference v39. Available: https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_39/gencode.v39.annotation.gtf.gz
103. Gencode mouse reference vM23. Available: http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M23/gencode.vM23.annotation.gtf.gz
104. 10X Genomics Build Notes for Reference Packages. Available: https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/build#hg19_1.2.0
105. Picard toolkit. Broad Institute, GitHub repository. 2019. Available: <http://broadinstitute.github.io/picard/>
106. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;1888–1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>
107. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet*. 2021;53:403–11.
108. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. 1995;57:289–300.
109. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2020;48:D87–92.
110. Zhou X, Cain CE, Myrthil M, Lewellen N, Michelini K, Davenport ER, et al. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol*. 2014;15:547.
111. Zhang H, Mulqueen RM, Adey AC, Cusanovich DA. Ten(10)X-compatible Combinatorial Indexing ATAC sequencing (txci-ATAC-seq). *protocols.io*. Available: <https://doi.org/10.17504/protocols.io.dm6gp3o68vzp/v1> (2023).
112. Zhang H, Mulqueen RM, Iannuzo N, Farrera DO, Polverino F, Galligan JJ, et al. txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE245957> (2023).
113. Zhang H, Mulqueen RM, Iannuzo N, Farrera DO, Polverino F, Galligan JJ, et al. txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE231708> (2023).
114. Zhang H, Mulqueen RM, Iannuzo N, Farrera DO, Polverino F, Galligan JJ, et al. txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility. *Datasets*. Database of Genotypes and Phenotypes. Available: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003497.v1.p1 (2023).
115. Zhang H, Mulqueen RM, Iannuzo N, Farrera DO, Polverino F, Galligan JJ, et al. txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility. *GitHub*. Available: <https://github.com/adeylab/txci-atac> (2023).
116. Zhang H, Mulqueen RM, Iannuzo N, Farrera DO, Polverino F, Galligan JJ, et al. txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility. *OSF*. 2023. <https://doi.org/10.17605/OSF.IO/VNPWB>.
117. Zhang H, Mulqueen RM, Iannuzo N, Farrera DO, Polverino F, Galligan JJ, et al. txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility. *GitHub*. Available: <https://github.com/cusanovichlab/txciatac> (2023).
118. Zhang H, Mulqueen RM, Iannuzo N, Farrera DO, Polverino F, Galligan JJ, et al. txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility. *OSF*. 2023. <https://doi.org/10.17605/OSF.IO/YA75E>.
119. Thornton CA, Mulqueen RM, Torkenczy KA, Nishida A, Lowenstein EG, Fields AJ, et al. Spatially mapped single-cell chromatin accessibility. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164849> (2021).
120. Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2668124> (2018).
121. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE123581> (2019).
122. Mulqueen RM, Pokholok D, O'Connell BL, Thornton CA, Zhang F, O'Roak BJ, et al. High-content single-cell combinatorial indexing. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5289637> (2021).
123. Koenitzer JR, Wu H, Atkinson JJ, Brody SL, Humphreys BD. Comparison of single cell and single nucleus RNASeq approaches in mouse lung. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145998> (2020).
124. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111586> (2018).
125. Williams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Datasets*. Liver Cell Atlas: Mouse StSt. Available: <https://www.livercellatlas.org/download.php> (2022).
126. Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Datasets*. Synapse. Available: <https://www.synapse.org/#Synapse:syn21041850> (2020).
127. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, et al. Single-cell RNA-sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135893> (2019).
128. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A cell atlas of chromatin accessibility across 25 adult human tissues. *Datasets*. Gene Expression Omnibus. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE165659> (2021).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.