ARTICLES
https://doi.org/10.1038/s41587-019-0147-6
nature biotechnology

Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility

Caleb A. Lareau[1,2,3,7], Fabiana M. Duarte[1,2,7], Jennifer G. Chew[4,7], Vinay K. Kartha[1,2], Zach D. Burkett[4], Andrew S. Kohlway[4], Dmitry Pokholok[5], Martin J. Aryee[1,3,6], Frank J. Steemers[5], Ronald Lebofsky[4*] and Jason D. Buenrostro[1,2*]

Recent technical advancements have facilitated the mapping of epigenomes at single-cell resolution; however, the throughput and quality of these methods have limited their widespread adoption. Here we describe a high-quality (10^5 nuclear fragments per cell) droplet-microfluidics-based method for single-cell profiling of chromatin accessibility. We use this approach, named 'droplet single-cell assay for transposase-accessible chromatin using sequencing' (dscATAC-seq), to assay 46,653 cells for the unbiased discovery of cell types and regulatory elements in adult mouse brain. We further increase the throughput of this platform by combining it with combinatorial indexing (dsciATAC-seq), enabling single-cell studies at a massive scale. We demonstrate the utility of this approach by measuring chromatin accessibility across 136,463 resting and stimulated human bone marrow-derived cells to reveal changes in the *cis*- and *trans*-regulatory landscape across cell types and under stimulatory conditions at single-cell resolution. Altogether, we describe a total of 510,123 single-cell profiles, demonstrating the scalability and flexibility of this droplet-based platform.

Although the primary sequence of the eukaryotic genome is largely invariant across cells in an organism, the quantitative expression of genes is tightly regulated to define the functional identity of cells. Eukaryotic cells use diverse mechanisms to regulate gene expression, including an immense repertoire (>10^6) of DNA regulatory elements[1,2]. These DNA regulatory elements are established and maintained by the combinatorial binding of transcription factors (TFs) and chromatin remodelers, which function together to recruit transcriptional machinery and drive cell-type-specific gene expression[3,4]. DNA regulatory elements, which are characterized by their functional roles (for example, promoters, enhancers and insulators), are marked by a diverse array of histone and DNA modifications[4]. Both classical observations[5] and recent genome-wide efforts[2] have shown that active regulatory elements are canonically nucleosome free and accessible to transcriptional machinery. Thus, methods that measure chromatin accessibility by combining sensitivity to enzymatic digestion with sequencing[6–8] provide an integrated map of chromatin states that encompasses a diverse repertoire of functional regulatory elements[2,5].

Methods to assay chromatin accessibility genome-wide have been used for a variety of applications including the discovery of (1) cell-type-specific *cis*-regulatory elements, (2) master TFs that shape the regulatory landscape or (3) mechanisms for disease-relevant non-coding genetic variation[2,9,10]. However, these 'epigenomic' approaches are generally applied to bulk samples, limiting their resolution when considering the regulatory diversity underlying heterogeneous cell populations. In parallel, methods to measure the transcriptomes of single cells have been used to discover new cell types[11] and new functional cell states[12,13], and provide additional motivation for the development of tools to measure chromatin regulation at single-cell resolution[14].

Technological innovations have enabled the development of single-cell epigenomic methods[14–16]; however, these approaches remain relatively low throughput and have high costs. Assay for transposase-accessible chromatin using sequencing (ATAC-seq)[8,17] is particularly promising for single-cell studies owing to the relative simplicity of the experimental protocol and its widespread use. Previous efforts have adapted ATAC-seq to profile chromatin accessibility in single cells, either by individually isolating cells[18] or by the combinatorial addition of DNA barcodes[19], to enable de novo deconvolution of cell types and the discovery of cell-type-specific regulatory factors[20,21]. However, these current methods for single-cell ATAC-seq (scATAC-seq)[18,19] either remain relatively low throughput (hundreds to thousands of cells per experiment) or provide low-complexity data (thousands of fragments per cell). Therefore, new methods for sensitive, scalable and high-throughput profiling are needed to measure the full repertoire of regulatory diversity across normal and diseased tissues.

To meet the challenges of assaying chromatin states across the breadth and depth of complex cell populations within tissues, we report the development of dscATAC-seq. In brief, our approach utilizes a droplet microfluidic device to individually isolate and barcode single transposed cells. We demonstrate that this approach results in substantially higher data quality than existing methods, and describe an approach to improve cell throughput and the efficiency of cell capture by super-loading barcoded beads into droplets. Furthermore, we extend this droplet barcoding approach by combining it with barcoded transposition[19] and super-loading of cells into droplets, to develop droplet-based single-cell combinatorial indexing for ATAC-seq (dsciATAC-seq), providing chromatin accessibility profiles at substantially improved throughput. We apply these approaches to generate accessibility profiles of 510,123

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [3]Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [4]Bio-Rad, Digital Biology Group, Pleasanton, CA, USA. [5]Illumina, San Diego, CA, USA. [6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [7]These authors contributed equally: Caleb A. Lareau, Fabiana M. Duarte, Jennifer G. Chew. *e-mail: ronald_lebofsky@bio-rad.com; jason_buenrostro@harvard.edu

cells, which include (1) a reference map of chromatin accessibility in the mouse brain (46,653 cells) and (2) an unbiased map of human hematopoietic states in the bone marrow (60,495 cells), isolated cell populations from bone marrow and blood (52,873 cells), and bone marrow cells in response to stimulation (75,958 cells). These unbiased chromatin accessibility profiles provide new insights into the regulators defining cells within these tissues. Finally, we find that pooled stimulus of human bone marrow-derived cells uncovers mechanisms underlying genetic variants that lead to human disease. Overall, this new approach for high-throughput single-cell epigenomics charts a clear course toward obtaining an epigenomic atlas across normal tissues and provides new opportunities for single-cell epigenomic profiling at a massive scale.

Results

scATAC-seq implemented on a droplet microfluidic device. In this work, we describe a method for single-cell profiling of chromatin accessibility using droplet microfluidics and ATAC-seq. Consistent with previously described methods for bulk ATAC-seq, nuclei are first transposed using Tn5 transposase to integrate sequencing adaptors into regions of open chromatin[8,17,22]. Importantly, previous studies have described that transposed nuclei and DNA remain intact after transposition[19,23]. We therefore leverage this finding and use intact transposed nuclei as input material for a droplet microfluidics device, which simultaneously encapsulates transposed chromatin with PCR reagents and barcoded beads into a single droplet (Fig. 1a). Each bead contains clonal copies of oligonucleotides that encode a common PCR primer sequence and a bead-specific DNA barcode. After encapsulation, we perform droplet PCR to add cell-identifying DNA barcodes to transposed chromatin, and the resulting pool of PCR products are then collected and prepared for sequencing. We refer to this droplet-based scATAC-seq platform as dscATAC-seq.

To develop a robust and high-sensitivity platform, we optimized the concentration of Tn5 transposase (Fig. 1b and Supplementary Fig. 1a–c). We found that increasing the total abundance and concentration of Tn5, which is the same enzyme contained within a widely available commercial product (Methods), substantially improved the total number of nuclear fragments, including improvements to the fraction of reads at transcription start sites (TSSs) and distal elements (Fig. 1b and Supplementary Fig. 1a–c). Furthermore, we also adapted previously described transposition methods to reduce the proportion of mitochondrial reads[17,22] (Methods). Altogether, these optimizations, combined with droplet encapsulation and PCR, provide a platform for high-yield and high-efficiency single-cell epigenomic profiling.

To optimize cell capture and throughput, we developed a joint experimental and computational strategy to super-load beads into droplets. Our computational strategy, which we call bead-based scATAC processing (BAP), determines bead barcodes with a high overlap of Tn5 insertion positions along the genome to identify and merge barcodes within a common droplet (Supplementary Figs. 1d and 2a,b). This analytical approach enables loading of beads at higher density (which increases the number of droplets with one or more beads) by identifying single cells with more than one bead barcode (Supplementary Fig. 2c,d and Supplementary Note). To validate our approach, we included a library of random oligonucleotides in a dscATAC-seq experiment, enabling us to define true-positive bead pairs on the basis of overlap of these exogenous sequences (Supplementary Fig. 2b,e–j). Using these orthogonal readouts, the unique Tn5 insertions across single cells and the random oligonucleotides introduced in this experiment, we computed precision-recall and receiver-operating-characteristic curves to verify the accuracy and precision of the BAP approach (mean area under the receiver-operating-characteristic curve (AUROC) = 1.000 and mean area under the precision-recall curve (AUPRC) = 0.997) (Supplementary Fig. 2k; Methods). We also found consistent experimental results across a range of bead concentrations without loss of data quality (Supplementary Fig. 2l–o). To compare the efficacy of our approach with that of other similar methods, we uniformly processed cell line data (from GM12878 and K562) generated using dscATAC-seq and four other recently published approaches[18,19,24,25]. We found that chromatin accessibility from bulk ATAC-seq[8] and DNase–seq[2] and the aggregate chromatin accessibility across the different single-cell technologies[18,19,24,25] were highly correlated (Fig. 1c,d). We also observed a collision rate of <2% (defined by >10% alternate species) when using 800 beads per microliter and 5,000 beads per microliter (Fig. 1e and Supplementary Fig. 2n,o). Notably, this estimated collision rate (<2%) is considerably lower than that of other previously described high-throughput single-cell combinatorial indexing for ATAC-seq (sciATAC-seq) methods[19,24] (>5%) (Supplementary Fig. 3a). Our dscATAC-seq method achieved improved library complexity per cell and numbers of cells per experiment without compromising the proportion of reads mapping to the nuclear genome (Fig. 1f and Supplementary Fig. 3b,c), both of which are common quality metrics for scATAC-seq experiments. Notably, dscATAC-seq recapitulated known variation in the activity of TF binding motifs across single GM12878 cells, as previously reported[18,26] (Supplementary Fig. 3d). Taking these results together, our new methodology provides an approach for high-resolution profiling of chromatin accessibility across thousands of single cells.

Epigenomic diversity of the adult mouse brain. We sought to determine whether our approach could be applied to large-scale efforts to identify cell types within complex tissues de novo. Thus, we applied the dscATAC-seq platform to whole-brain tissues derived from two mice using our super-loaded bead concentration (5,000 beads per microliter). Over 12 experimental libraries, we observed a median cell capture of 5,324/5,600 (95%), consistent with our theoretical expectation (Supplementary Fig. 2d). Cells that passed additional stringent quality filters had a median of 34,046 unique nuclear reads, 58.8% of reads in peaks and an average of 2.5 bead barcodes per cell for 46,653 total cells (Supplementary Fig. 4a).

To characterize differences in chromatin accessibility across cell types, we first reduced the dimensionality of our mouse brain profiles by computing *k*-mer deviation scores (7-mers) using the chromVAR algorithm[26]. Cell clusters were identified using the Louvain modularity method built from a cell nearest-neighbor graph using the 7-mer scores, which uncovered 27 cell clusters. We then used these 7-mer features to map each cell to a two-dimensional representation with *t*-distributed stochastic neighbor embedding (*t*-SNE) (Fig. 2a). Importantly, these clusters were largely uncorrelated with known technical confounders (Supplementary Fig. 4b–d), and we observed a largely consistent pattern when compared to dimensionality reduction and clustering using the latent semantic index (LSI) of our dscATAC-seq data as has been previously performed[27,28] (Supplementary Fig. 4e). For comparison with previous techniques, we also analyzed published sciATAC-seq data from two mouse brains[27], where we identified 13 clusters using the same computational approach (Supplementary Fig. 4f). We attribute the lower number of clusters to the smaller number of cells assayed (5,744 cells), the lower library complexity (median 14,681) and the smaller fraction of reads in peaks per cell (median 30.0%) (Supplementary Fig. 4g,h).

To annotate these clusters, we calculated per-cluster promoter-region chromatin accessibility scores (weighted-sum of chromatin accessibility in the 200-kb region centered on a TSS) (Supplementary Fig. 5a and Supplementary Table 1). Of note, dimension reduction using promoter-region chromatin accessibility scores for all genes resulted in reduced resolution of neuronal subclusters (Supplementary Fig. 5b). We therefore used previously annotated marker genes for mouse brain to correlate our promoter-region

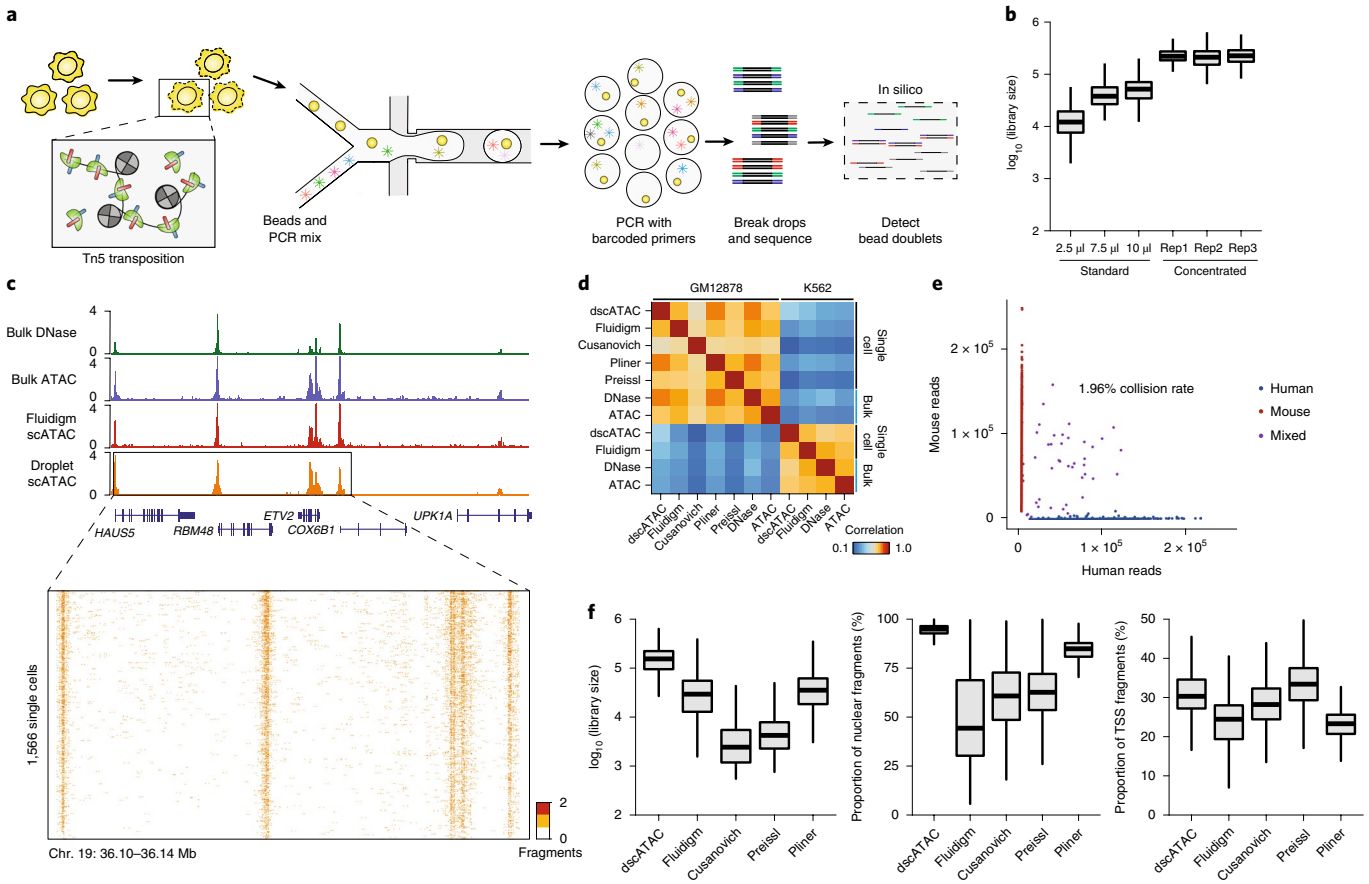

Fig. 1 | dscATAC-seq enables high-resolution characterization of open chromatin regions in single cells. a, Schematic of dscATAC-seq. Cells are transposed with Tn5 transposase, and transposed chromatin is then barcoded and amplified in a microfluidic device. **b**, Comparison of per-cell library sizes using different Tn5 conditions in K562 cells. Three replicates (Rep1, Rep2, Rep3) are shown for the concentrated enzyme mixture ($n = 500$ cells per replicate). **c**, Comparison of the aggregate chromatin accessibility profiles from GM12878 cells using different technologies, and visualization of single-cell chromatin accessibility profiles from dscATAC-seq. The aggregate chromatin accessibility profile from dscATAC-seq is representative of at least ten replicates. **d**, Spearman correlation of reads in chromatin accessibility peaks across bulk and single-cell technologies for GM12878 and K562 cells ($n = 1$ replicate for each). **e**, The number of unique fragments aligning to the human or mouse genome using human (GM12878) and mouse (3T3) cells at 800 beads per microliter. **f**, Quality metrics of scATAC-seq methods for GM12878 cells. The median library size for dscATAC-seq was 165,204 reads (left; all reads reported passed quality filters) as compared to profiles generated from the Fluidigm C1 (ref. [18]) (50,443 reads) and sciATAC-seq methods (4,641 reads, Cusanovich[19]; 6,225 reads, Preissl[24]; 46,730 reads, Pliner[25]). The median fraction of mapped nuclear fragments for dscATAC-seq is 95% (middle). In box plots center lines indicate the median, box limits indicate the first and third quartiles and whiskers indicate 1.5× interquartile range (IQR). The sample size for each method is shown in Supplementary Fig. 3c.

chromatin accessibility scores to a recently described single-cell transcriptomic atlas of cell types across nine regions of the adult mouse brain[29]. We then used the highest correlation to the clusters from single-cell RNA sequencing (scRNA-seq) to partition the dscATAC-seq clusters into the major cell types from the mouse brain. These clusters included microglia (MG1), oligodendrocytes (OG1), oligodendrocyte progenitor cells (OPCs; P1), astrocytes (A1), endothelial cells (E1), inhibitory neurons (IN01–IN05) and excitatory neurons (EN01–EN17) (Fig. 2a). Pooled ATAC-seq signal (Fig. 2b) and promoter-region chromatin accessibility scores (Supplementary Fig. 5c,d) at known cell-type-specific gene markers further validated the cluster assignments. Interestingly, we also observed consistently higher library complexity and a higher ratio of distal to promoter reads per cell for annotated neurons as compared to other cell types, suggesting that neurons may have overall increased chromatin accessibility at distal regulatory elements (Supplementary Fig. 5e–g).

To refine cluster annotations, we employed an optimal matching algorithm to link our promoter accessibility scores to two published scRNA-seq datasets[29,30] (Fig. 2c). Here we identified multiple scRNA-seq clusters to be highly correlated with each of our scATAC-seq clusters, likely reflecting the nature of the annotations, which classify cell types both by expression signatures and from regions of the brain. To define the most likely pairs, we employed the Gale–Shipley algorithm to maximize the global correlation (Spearman) of our cluster assignments to scRNA-seq clusters (Supplementary Table 2; Methods). Differentially enriched genes in each scATAC-seq cluster provided further insights into the putative cell identities (Fig. 2d). For instance, enrichment of chromatin accessibility in *Sst* from the IN04 cluster suggests that this cluster corresponds to Sst$^+$ (somatostatin-expressing) neurons, a defined subset of GABAergic inhibitory neurons with high levels of spontaneous activity[31]. Furthermore, chromatin accessibility for *Syt6*, a marker of layer 6 pyramidal neurons[32], was enriched in EN12. *Htr1a* and *Htr2c*, which encode serotonin receptors and are known markers of serotonin neurons[33], had enriched accessibility in EN10 and EN07, respectively. *Lhx1*, encoding a TF that is enriched in the suprachiasmatic nucleus, which maintains synchrony among circadian oscillator neurons[34], had enriched accessibility in EN04.

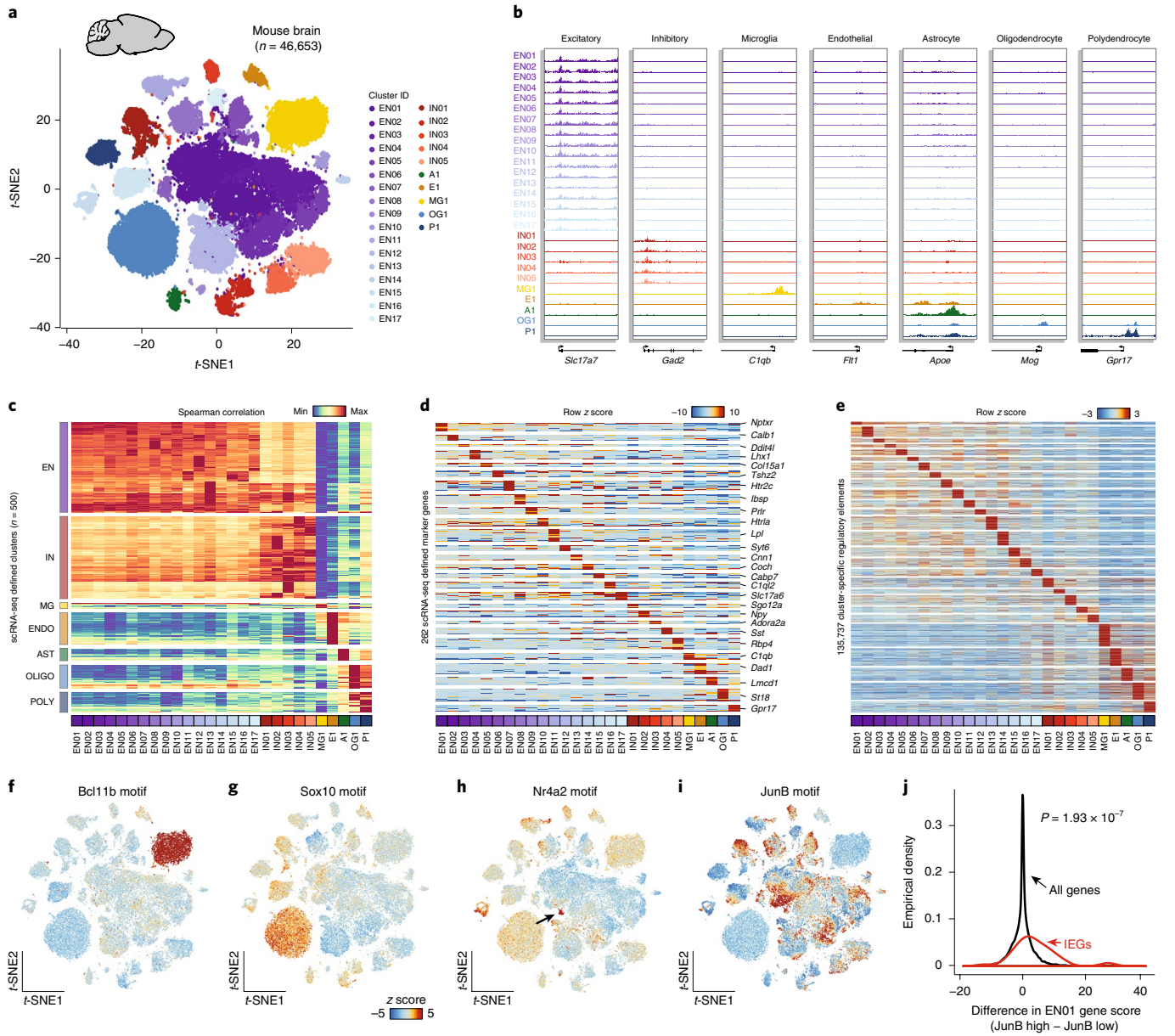

Fig. 2 | De novo classification of cell types in the mouse brain. a, A *t*-SNE visualization of cells ($n = 46{,}653$) derived from two whole mouse brains across 12 experimental batches. Cells are colored by their identity across 27 clusters. **b**, Aggregate chromatin accessibility profiles per cluster surrounding the promoter region of known marker genes. Aggregate profiles were combined over 12 experimental libraries. **c**, A correlation matrix of clusters from mouse brains defined by scRNA-seq[29] and dscATAC-seq. Margin labels indicate cell class. Each row is normalized to the minimum and maximum from Spearman correlation of the marker genes derived using scRNA-seq. **d**, Chromatin accessibility scores for promoter regions from previously defined brain marker genes in mice. The marker genes for each cluster are indicated. **e**, Chromatin accessibility signal across 135,737 cell-type-specific peaks within clusters defined in the mouse brain. **f**–**h**, Cluster-specific activity of known TF regulators in the mouse brain. Plots depict the chromVAR deviation score for each of the transcription factor motifs, including for Bcl11b (**f**; microglia), Sox10 (**g**; oligodendrocytes) and Nr4a2 (**h**, dopaminergic neurons). **i**, Within-cluster variation shown by the JunB motif. **j**, Comparison of promoter-region chromatin-accessibility scores between cells from the JunB-high and JunB-low groups of the EN01 cluster. Empirical densities of 47 annotated IEGs as compared to all annotated genes are shown. The *P* value is from a two-tailed, two-sample Kolmogorov–Smirnov ($n = 24{,}360$ genes).

In addition to the inference of cell types, our approach also enabled the unbiased identification of 135,737 cell-type-specific chromatin regulatory elements (Fig. 2e and Supplementary Table 3), which further validates the unique identity of each cell cluster and provides a general resource for defining regulatory elements as cell-type-specific reporters as part of the effort to better understand the mouse brain[35].

To further utilize the underlying chromatin data of our resource, we sought to examine cell-type-specific TF regulators within each cluster using deviations of TF motifs. We observed strong enrichment of the motifs for Bcl11b (Fig. 2f) and Sox10 (Fig. 2g) in the microglia and oligodendrocyte clusters, respectively. These known master regulators further validate the clusters assigned using our approach. Next, we identified highly specific activity for the Nr4a2 motif (Fig. 2h), suggesting that the EN13 cluster comprises dopaminergic neurons, given the critical role of Nr4a2 in the development and maintenance of the dopaminergic system[36]. In addition to observing cluster-specific TFs, we found

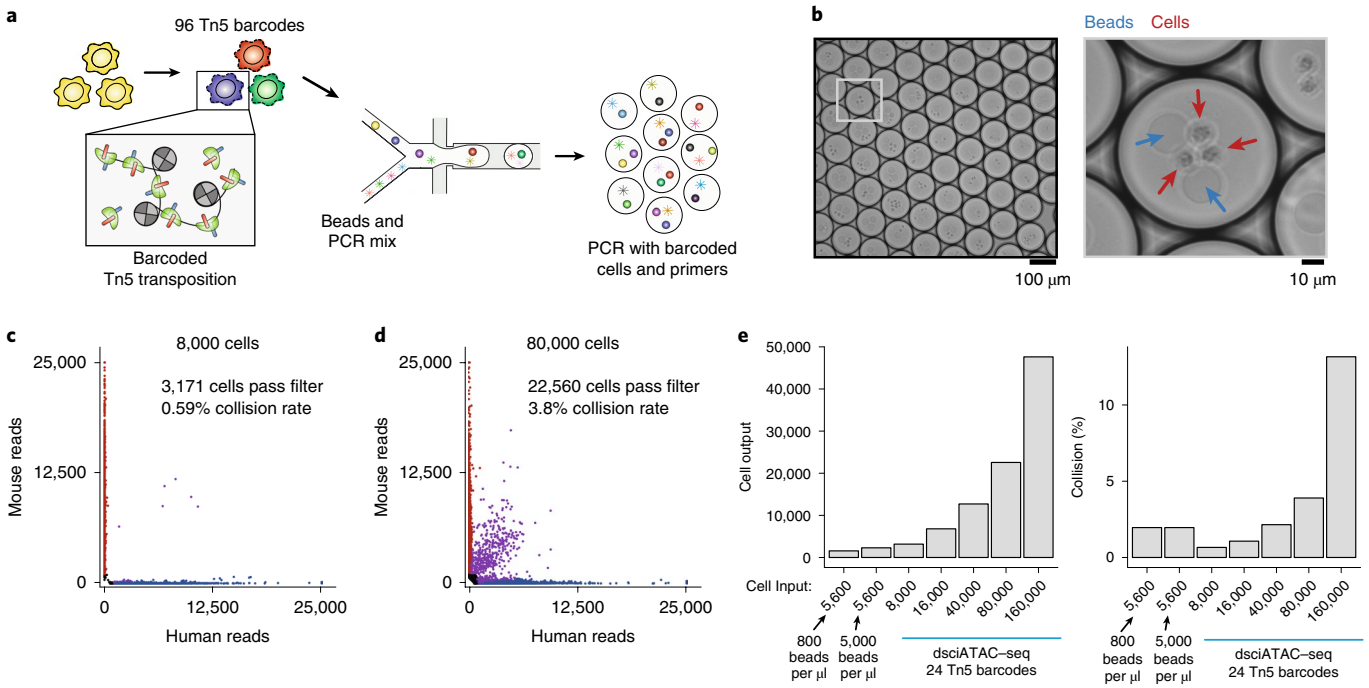

Fig. 3 | dsciATAC-seq enables massive-scale single-cell experiments. a, Schematic of dsciATAC-seq. Cells are transposed with barcoded Tn5, pooled and then further processed through a droplet PCR microfluidic device. **b** Representative images of droplets containing multiple beads and cells. Blue arrows indicate beads and red arrows indicate transposed nuclei. **c**,**d**, The number of unique reads aligning to the mouse or human genome from dsciATAC-seq profiles of human (K562) and mouse (3T3) cells with inputs of 8,000 (**c**) and 80,000 (**d**) cells. **e**, Summary of species mixing and cell yields at various cell inputs.

considerable within-cluster variability for the JunB motif specific to neuron clusters (Fig. 2i). We hypothesized that this variability might reflect neural activity driving expression of immediate early genes (IEGs)[29,37]. Indeed, this hypothesis was supported by a statistically significant enrichment of accessibility for 47 previously annotated IEGs in cells from the JunB-high group (z score > 0) as compared to the JunB-low group (z score < 0) within the EN01 cluster (two-sample Kolmogorov–Smirnov test, $P = 1.93 \times 10^{-7}$; Fig. 2j). Altogether, we observe that the dscATAC-seq platform provides a powerful means for defining and annotating cell types and states and identifying cell-type-specific chromatin features.

Droplet-based sciATAC-seq for massive-scale single-cell studies. Although the dscATAC-seq approach can be scaled to generate data for large numbers of cells by simply performing the experiment across many replicates, as shown above (Fig. 2), we reasoned that we could further increase cell throughput by surpassing Poisson loading of cells in droplets (one cell per droplet). We therefore sought to combine this approach with combinatorial indexing[19,23] to improve throughput and enable multiplexing of multiple samples in a given experiment. To achieve this, we developed dsciATAC-seq, wherein Tn5 transposase is loaded with barcoded DNA adaptors to add well-specific DNA barcodes to open chromatin. Following barcoded transposition, transposed cells were pooled and loaded in our droplet microfluidics device at high density to simultaneously encapsulate multiple Tn5-barcoded cells with multiple beads in each droplet (Fig. 3a,b). Thus, each individual cell may be identified by both the droplet-specific bead barcode and the well-specific Tn5 barcode, enabling an increase in cell throughput proportional to the initial number of Tn5 barcodes used in the experiment. Using our droplet-based platform with barcoded Tn5 reactions increases the number of theoretical barcode combinations to enable greater throughput of cells or samples (if cells originate from different samples).

We first implemented this technology with 24 transposase barcodes and generated high-quality chromatin accessibility profiles for up to 50,000 cells in a single well of the device (representing one experimental sample). Analysis of species mixing (using Tn5-barcode-aware parsing; Methods) confirmed that we could increase cell throughput approximately tenfold while maintaining a collision rate lower than 5% when using 24 transposase barcodes (Fig. 3c–e and Supplementary Fig. 6a) and showed a further reduction in the overall detected collision rates at large cell inputs with 48 barcodes (Supplementary Fig. 6b). Altogether, in this cell titration experiment, we generated 274,144 single-cell profiles demonstrating the massive scalability of this approach. Notably, to perform this experiment, we purified and in vitro assembled Tn5 transposase with different barcodes separately[38]. As this proof-of-concept experiment did not utilize the optimized Tn5 concentration described in Fig. 1b, we observed fewer reads per cell but maintained a high fraction of reads in peaks (72.2%). Together, these experiments demonstrate that barcoded Tn5 can enable super-Poisson loading of cells into droplets to achieve substantially greater throughput for generating epigenomic profiles from between 10^4 and 10^5 single cells per experiment.

Chromatin accessibility profiling of human bone marrow. Barcoded Tn5 transposition enables a substantially increased cell throughput and provides an opportunity to multiplex scATAC-seq for multiple conditions or samples. Notably, tissue-scale perturbations[39] have been used to uncover diverse cell response dynamics[40]. We therefore reasoned that pooled stimulation across heterogeneous cell types within bone marrow mononuclear cells (BMMCs) would provide unique avenues to understand the functional roles of epigenomic diversity within human bone marrow. To achieve this, we used dsciATAC-seq with 96 transposase barcodes to profile BMMCs from two human donors before (untreated controls)

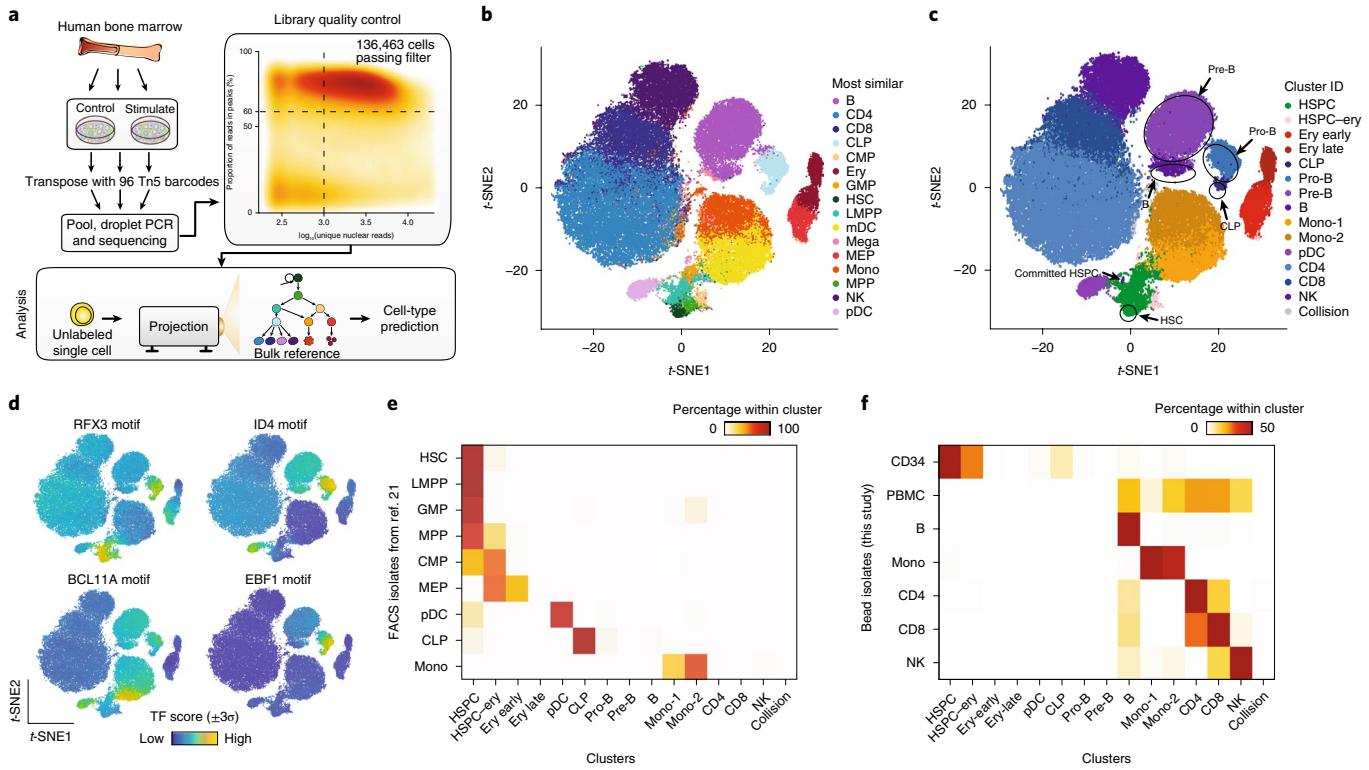

Fig. 4 | dsciATAC-seq of human bone marrow cells reveals the major lineages of hematopoietic differentiation. a, Schematic of the experimental and computational workflows used to assess dsciATAC-seq data from BMMCs. Ninety-six Tn5 transposase barcodes were used to define different donors and stimulation conditions. The plot for library quality control displays a summary of data that passed quality filters across all assayed cells. The total number of cells that passed filters of 60% reads in peaks and 1,000 unique nuclear reads was 136,463. **b**,**c**, Two-dimensional *t*-SNE embedding of single BMMCs without stimulation ($n = 60{,}495$ cells). Cells are colored by the most correlated cell type from a bulk ATAC-seq reference (**b**) or 15 de novo-defined cluster assignments covering known hematopoietic cell types (**c**). Cell types covering the B cell differentiation trajectory are highlighted. **d**, Single cells are colored by TF motif accessibility scores, computed using chromVAR[26], for the motifs of RFX3, ID4, BCL11A and EBF1. **e**,**f**, Confusion matrices showing the percentage overlap of published scATAC-seq from FACS-isolated subsets[21] (**e**) and dscATAC-seq data from bead-isolated subsets generated in this study (**f**) with clusters derived from dsciATAC-seq. Cell types: common myeloid progenitor, CMP; erythroid, Ery; granulocyte-monocyte progenitor, GMP; lymphoid-primed multipotent progenitor, LMPP; megakaryocyte, Mega; multipotent progenitor, MPP; plasmacytoid dendritic cells, pDC.

and after stimulation, producing chromatin accessibility profiles for a total of 136,463 cells that passed quality filters (Fig. 4a and Supplementary Fig. 7a–f).

The reference map of 60,495 resting cells (untreated controls) revealed the major lineages of hematopoietic differentiation de novo. To analyze these reference datasets, we projected the untreated BMMCs onto hematopoietic development trajectories using a reference-guided approach, whereby single cells were scored by principal components trained on bulk sorted hematopoietic ATAC-seq profiles[21] (Fig. 4a; Methods). With this approach, we were able to visualize and predict cell labels given the bulk reference map of epigenomic states (Fig. 4b). Furthermore, using the Louvain modularity method, we identified 15 distinct clusters from the 60,495 resting cells, which recapitulated the major constitutive cell types in the human hematopoietic system (Fig. 4c). These de novo-derived single-cell clusters reflected changes in chromatin accessibility mediated by key lineage-specific TFs, corresponding to altered motif accessibility for TFs associated with the stepwise progression of B cell development from hematopoietic stem cells to mature B cells (Fig. 4c,d). While embedding and cell clustering in our approach were performed on the basis of bulk projections in keeping with our previous work, we noted a considerable concordance of these data using our de novo *k*-mer strategy (Supplementary Fig. 7g–i). Furthermore, we observed unexpected epigenomic heterogeneity across TF motifs, including those of CEBPD and BCL11A, within monocyte clusters (Mono-1 and Mono-2), which likely reflects the heterogeneous developmental transitions from myeloid progenitors to mature monocytes, myeloid dendritic cells (mDCs) and granulocytes (Supplementary Fig. 8a).

To validate the clusters and cell-type annotations from our approach, we assigned previously obtained scATAC-seq profiles from progenitors in human bone marrow and peripheral blood monocytes isolated using fluorescence-activated cell sorting (2,034 cells)[21] to the clusters defined using our method. We classified these published single-cell data to clusters on the basis of the minimum Euclidean distance of a single-cell profile to a cluster medoid. We observed considerable overlap between each isolated subset and its corresponding cluster in the dsciATAC-seq dataset, validating the annotations for progenitor cell types (Fig. 4e). Furthermore, we performed dscATAC-seq on $CD34^+$ bone marrow progenitor cells, peripheral blood mononuclear cells (PBMCs) and bead-enriched subpopulations of PBMCs to derive a total of 52,873 cells, which validated our cluster label assignments for mature cell types (Fig. 4f and Supplementary Fig. 8b). We also used an orthogonal approach to visually validate these findings by dimensionality reduction using the uniform manifold projection (UMAP) algorithm[41], which allowed for data to be projected onto the dsciATAC-seq base dimensionality (Supplementary Fig. 8c–f). Collectively, we have used this approach to define a reference epigenomic atlas of cell states within hematopoietic cells in the human bone marrow, highlighting the

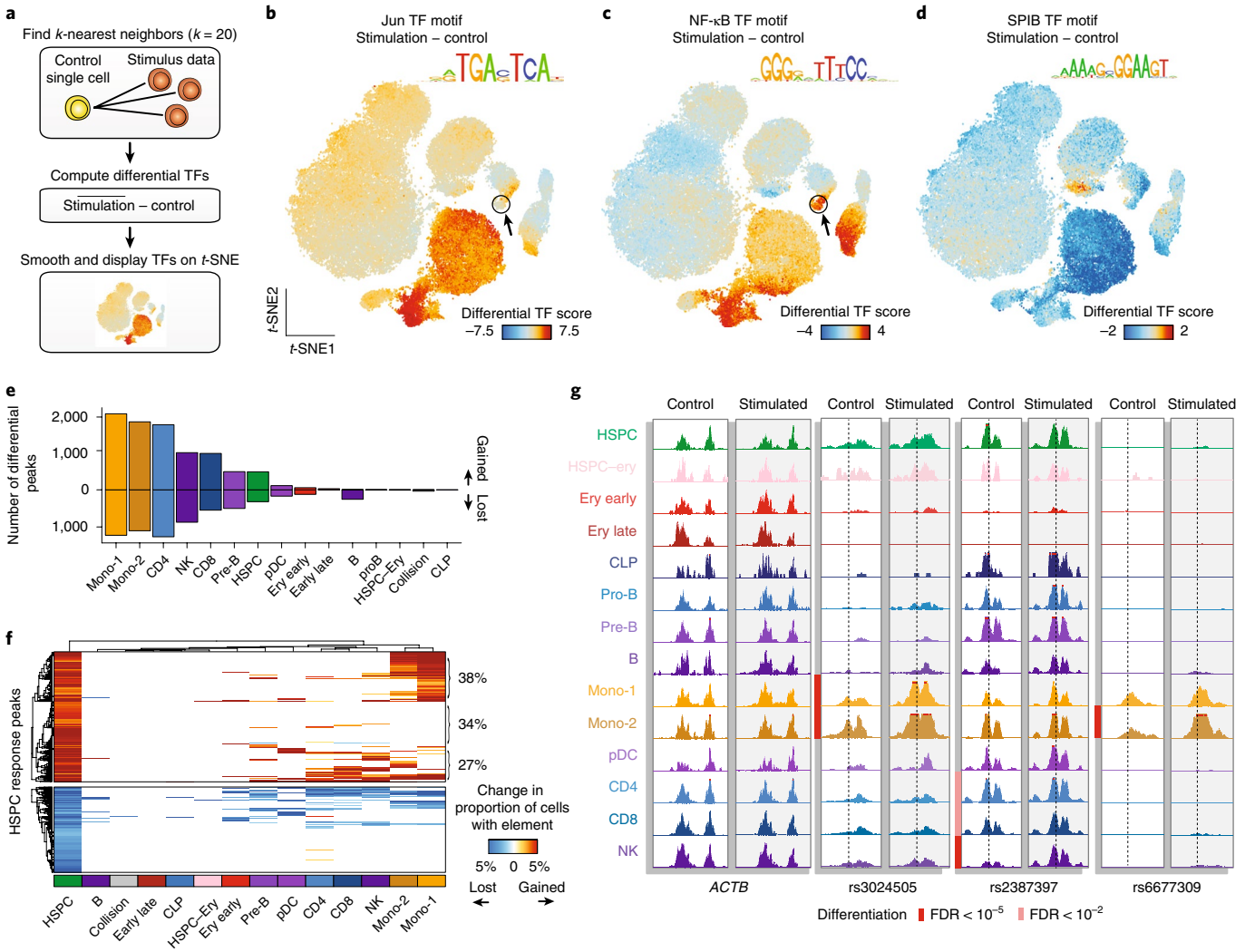

Fig. 5 | Identification of stimulus-response regulators in human bone marrow. a, Schematic depicting the computational workflow for comparing stimulus versus control single-cell data. **b**–**d**, Differential TF deviation scores for Jun (**b**), NF-κB (**c**) and SPIB (**d**) motifs in response to stimulation for $n = 60{,}495$ resting cells. **e**, Summary of the number of differential chromatin accessibility peaks across each cluster at an FDR of 1% after a two-sided permutation test. Bars above the zero line represent chromatin accessibility peaks that were gained, while bars below the zero line represent chromatin accessibility peaks that were lost. **f**, Hierarchical clustering of peaks gained (top) or lost (bottom) across clusters, restricted to the differential peaks identified in HSPCs. **g**, Locus-specific views of the *ACTB* promoter and three fine-mapped variants identified through genome-wide association studies. The dashed line represents the location of the single-nucleotide polymorphism (SNP) in each window.

applicability of our combinatorial approach to the generation of accurate large-scale epigenomic maps to define cell types within primary human tissues.

Regulatory consequences of multi-lineage stimulation. Our multiplexed, dsciATAC-seq method further provides a unique opportunity to decipher regulatory consequences of perturbation without concerns for batch effects, which can confound experimental results. To characterize the response of each immune cell cluster to stimulation conditions, we explored the differences between our untreated control cells and ex vivo cultured and lipopolysaccharide (LPS)-stimulated BMMCs (Fig. 4a). To determine *trans*-acting regulators altered in response to perturbation, we developed an analytical strategy wherein we computed differential TF scores by (1) defining a *k*-nearest neighbor map connecting stimulated cells to control cells and (2) calculating, for each cell, the difference between the TF score for the cell and the average score for the 20 nearest stimulated cells (Fig. 5a). Interestingly, we found significant and highly correlated epigenomic responses to both ex vivo culture and LPS stimulation (Supplementary Fig. 9a–e), suggesting that the effects of ex vivo culture dominate those induced by LPS. For clarity we simply refer to these conditions as 'stimulation' for downstream analysis. With this stimulation data representing the full spectrum of bone marrow hematopoietic cell states, we found cell-type-specific induction of a diverse repertoire of TF motifs (Fig. 5b–d and Supplementary Fig. 9f–j). The differential activity observed included increased accessibility at the Jun and NF-κB motifs, which was largely localized to human hematopoietic stem and progenitor cells (HSPCs) (Fig. 5b,c), depletion of accessibility at the SPIB motif in myeloid cell types (Fig. 5d) and relatively weak induction of accessibility at the IRF8 (myeloid) and MAFF (megakaryocyte–erythrocyte progenitor (MEP) and common lymphoid progenitor (CLP) to pre-B) motifs (Supplementary Fig. 9i,j). Interestingly, accessibility at the Jun and NF-κB motif was largely correlated in HSPCs, with the exception of CLPs and cells from early erythroid differentiation, where cells appeared to respond exclusively by induction of accessibility at the NF-κB motif (Fig. 5b,c).

Next, we examined the *cis*-regulatory consequences of stimulation across our multi-lineage defined cell states. To compute differential chromatin accessibility peaks within each cluster, we devised a permutation test for each peak, permuting control and perturbation cell labels, which allowed us to improve the robustness of our statistical methods by considering each cell as an independent observation (Supplementary Fig. 9k–l; Methods). This analysis revealed a total of 9,638 distinct stimulus-responsive chromatin accessibility peaks (false-discovery rate (FDR) of 1%; Supplementary Table 4). Interestingly, we broadly observed an increase in the total number of accessible peaks, represented by the Mono-1 cluster with 2,114 peaks gained as compared to 1,264 peaks lost (binomial $P < 2.2 \times 10^{-16}$) (Fig. 5e). The global increase in chromatin accessibility upon stimulation was also corroborated by an approximately 20% increase in the average library complexity per cell. The most prominent cell types that responded to stimulation included the two monocyte clusters and the CD4$^+$ T cell cluster. Unexpectedly, we also observed 501 chromatin accessibility peaks gained in the HSPC cluster, and approximately 34% of these gained peaks were unique to HSPCs (Fig. 5f), thus uncovering an HSPC-specific signature of stimulus response. Altogether, considering the TF motif and peak-specific analyses, we find that HSPCs respond to stimulus using the NF-κB and Jun motifs to drive an HSPC-specific response. This finding provides support for reports suggesting that HSPCs are responsive to interferon-mediated immune signaling[42,43], and may be used to further characterize the regulatory basis of interferon signaling in HSPCs to facilitate discovery of chemical inhibitors that will enable ex vivo expansion and gene editing of hematopoietic stem cells[44] for hematopoietic stem cell transplantation.

We further hypothesized that using this approach to uncover cell-type-specific changes resulting from stimulation could elucidate mechanisms in the relevant cell types and regulatory regions encompassing variants implicated in genome-wide association studies[45,46]. With this in mind, we observed stimulus-responsive chromatin accessibility peaks near the *IL10* locus in monocytes overlapping the pleiotropic locus for the rs302405 variant, which is associated with type 1 diabetes (posterior probability (PP) = 0.38), Crohn's disease (PP = 0.40) and ulcerative colitis (PP = 0.41), and increased chromatin accessibility at variant rs2387397, which is associated with celiac disease (PP = 0.32), within the natural killer (NK) and T cell clusters (Fig. 5g). Additionally, we observed a Mono-2 stimulation-specific peak overlapping rs6677309, which is a fine-mapped variant that is associated with multiple sclerosis (PP = 0.49), near the *CD58* locus (Fig. 5g). Interestingly, CD58 presentation by activated monocytes has been shown to cause expansion of CD56$^+$ NK cells[47], which may promote an autoimmune response in multiple sclerosis[48]. Overall, this single experiment comprising 60,495 resting and 75,968 stimulated cells enabled unbiased discovery of regulatory changes across various stages of hematopoietic differentiation and unbiased identification of the regulatory consequences of ex vivo perturbation across multiple lineages, providing new opportunities to better define cell types within complex tissues as well as their role in stem cell therapy and autoimmune disease.

Discussion

In the genomics era of cell atlases, a major goal of single-cell methods is to provide unbiased classification of cell types and the epigenomic, transcriptomic and proteomic features that define them[49]. We find that scATAC-seq maps can provide information-rich measurements of cells (10^5 fragments per cell), which enable the identification of cell types and their underlying regulatory elements. Furthermore, previous work has suggested that activity of regulatory elements may be a more accurate reflection of cell potential and perhaps provide higher cell-type specificity than measurements of gene expression[17]. The scATAC-seq approach described here produces single-cell profiles at higher throughput, improved yield and higher sequencing efficiency than previous scATAC-seq methods, providing a robust platform for identifying new cell types within heterogeneous tissues. We expect that the combination of this scATAC-seq approach with scRNA-seq profiling will provide a more accurate definition of cell types and that further integration of these data[21,27,50] will enable opportunities to define mechanistic models of gene regulation to better understand their function.

We present a series of technological innovations leading to a high-throughput epigenomic profiling approach that enables super-Poisson loading of cells and beads into microfluidic droplets. To achieve this, we have developed a computational approach to identify droplets with multiple barcoded beads and paired this approach with combinatorial indexing by barcoded transposition to add multiple cells to each droplet. Combining these approaches dramatically improves cell throughput to approximately 25,000 cells per well (100,000 cells per droplet device), which we expect may be further improved with optimizations of the approach and additional Tn5 barcodes. More generally, we expect this conceptual framework of combinatorial indexing coupled with a microfluidics device to be compatible with other methods for high-throughput PCR (for example, microwells[51]) and other single-cell genomics assays leveraging combinatorial indexing for cell barcoding[52–54].

This approach allows for multiplexing of many samples in a single experiment. In this work, we multiplex control and perturbation conditions across an entire tissue, enabling us to define shared and cell-type-specific regulatory changes induced upon stimulation across diverse cell types. These advances for multiplexing experiments, along with advances in high-throughput sequencing, provide new opportunities to define not only cell-type-specific chromatin accessibility, but also changes across diverse genetic and environmental conditions. As such, we expect this approach to be used to profile epigenomic variation across healthy individuals or cohorts of patients with disease to determine the functional roles of the regulatory elements and cell types underlying common traits or diseases[55]. Altogether, these advances enable a new era of single-cell epigenomic studies at a massive scale, providing a powerful new tool to connect the vast repertoire of DNA regulatory elements to function.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0147-6.

Received: 17 December 2018; Accepted: 6 May 2019;
Published online: 24 June 2019

References

1. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
2. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
3. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
4. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837 (2013).
5. Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. *Science* **193**, 848–856 (1976).
6. Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
7. Hesselberth, J. R. et al. Global mapping of protein–DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
8. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
9. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).

10. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
11. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
12. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
13. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
14. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: recording the past and predicting the future. *Science* **358**, 69–75 (2017).
15. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142–146 (2015).
16. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
17. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
18. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
19. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
20. Cusanovich, D. A. et al. The *cis*-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
21. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548 (2018).
22. Corces, M. R. et al. An improved ATAC–seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
23. Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
24. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
25. Pliner, H. A. et al. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
26. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. Chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
27. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324 (2018).
28. Chen, X., Miragaia, R. J., Natarajan, K. N. & Teichmann, S. A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **9**, 5345 (2018).
29. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
30. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).
31. Urban-Ciecko, J. & Barth, A. L. Somatostatin-expressing neurons in cortical networks. *Nat. Rev. Neurosci.* **17**, 401–409 (2016).
32. Ullrich, B. & Südhof, T. C. Differential distributions of novel synaptotagmins: comparison to synapsins. *Neuropharmacology* **34**, 1371–1377 (1995).
33. Meneses, A. Serotonin, neural markers, and memory. *Front. Pharmacol.* **6**, 143 (2015).
34. Hatori, M. et al. Lhx1 maintains synchrony among circadian oscillator neurons of the SCN. *eLife* **3**, e03357 (2014).
35. Visel, A. et al. A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895–908 (2013).
36. Kadkhodaei, B. et al. Nurr1 is required for maintenance of maturing and adult midbrain dopamine neurons. *J. Neurosci.* **29**, 15923–15932 (2009).
37. Yap, E.-L. & Greenberg, M. E. Activity-regulated transcription: bridging the gap between neural activity and behavior. *Neuron* **100**, 330–348 (2018).
38. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
39. Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
40. Bodenmiller, B. et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat. Biotechnol.* **30**, 858–867 (2012).
41. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. (2018). https://doi.org/10.1038/nbt.4314.
42. Essers, M. A. G. et al. IFNα activates dormant haematopoietic stem cells in vivo. *Nature* **458**, 904–908 (2009).
43. Espín-Palazón, R. et al. Proinflammatory signaling regulates hematopoietic stem cell emergence. *Cell* **159**, 1070–1085 (2014).
44. Petrillo, C. et al. Cyclosporine H overcomes innate immune restrictions to improve lentiviral transduction and gene editing in human hematopoietic stem cells. *Cell Stem Cell* **23**, 820–832 (2018).
45. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
46. Ulirsch, J.C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
47. Lopez, R. D., Waller, E. K., Lu, P. H. & Negrin, R. S. CD58/LFA-3 and IL-12 provided by activated monocytes are critical in the in vitro expansion of $CD56^{+}$ T cells. *Cancer Immunol. Immunother.* **49**, 629–640 (2001).
48. Laroni, A. et al. Dysregulation of regulatory $CD56^{bright}$ NK cells/T cells interactions in multiple sclerosis. *J. Autoimmun.* **72**, 8–18 (2016).
49. HCA Consortium. *The Human Cell Atlas White Paper* (2017).
50. Stuart, T. et al. Comprehensive integration of single cell data. Preprint at https://doi.org/10.1101/460147 (2018).
51. Han, X. et al. Mapping the mouse cell atlas by microwell-seq. *Cell* **173**, 13071091– (2018).
52. Vitak, S. A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).
53. Mulqueen, R. M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
54. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
55. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12**, 529–541 (2011).

Acknowledgements

We thank members of the Buenrostro lab for useful discussions and critical assessment of this work. We thank D. Norton at Bio-Rad for enabling the collaboration with the Broad Institute and Harvard University. We would also like to acknowledge other Bio-Rad colleagues for establishing and providing droplet-related consumables, including D. Do, B. Zhang, P. Pattamatta, S. Cater, L. Frenz, D. Greiner and J. Agresti. We also recognize L. Christiansen, A. Yunghans and L. Watson from the Illumina Assay Development team, in addition to F. Zhang and F. Schlesinger at Illumina, for their bioinformatics contributions. We are grateful to the Zhang lab (Broad Institute) for providing the Tn5 for combinatorial experiments. J.D.B., C.A.L., F.M.D. and V.K.K. acknowledge support by the Allen Distinguished Investigator Program through the Paul G. Allen Frontiers Group. This work was further supported by the Chan Zuckerberg Initiative. C.A.L. is supported by an NIH F31 grant (F31CA232670).

Author contributions

F.M.D., J.G.C. and A.S.K. generated the data. C.A.L., V.K.K. and Z.D.B. analyzed the data. F.J.S. proposed the droplet scATAC-seq approach and oversaw the proof-of-concept studies performed by D.P. M.J.A. assisted in the development of computational resources. C.A.L., F.M.D. and J.D.B. wrote the manuscript with input from all authors. R.L. and J.D.B. jointly supervised this work.

Competing interests

Work by D.P. and F.J.S. was performed at Illumina. Work by J.G.C., Z.D.B., A.S.K. and R.L. was performed at Bio-Rad. J.D.B. holds patents related to ATAC-seq.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0147-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.L. or J.D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Statistics. All statistical tests and corresponding summary values (for example, *P* values and sample sizes) are described in the corresponding sections below and in the figure legends.

Cell lines. GM12878 (Coriell Institute for Medical Research) human lymphoblastoid cells were maintained in RPMI 1640 medium modified to include 2 mM L-glutamine (ATCC), 15% FBS (ATCC) and 1% penicillin–streptomycin (Pen–Strep) (ATCC). K562 (ATCC) human chronic myelogenous leukemia cells were maintained in IMDM (ATCC) supplemented with 10% FBS and 1% Pen–Strep. NIH/3T3 (ATCC) mouse embryonic fibroblast cells were maintained in DMEM (ATCC) supplemented with 10% calf bovine serum and 1% Pen–Strep. All cell lines were maintained at 37 °C and 5% CO_2 at recommended densities and were collected at mid-log phase for all experiments. All cells in suspension were collected using standard cell culture procedure and adherent cells were detached using TrypLE Express Enzyme (Gibco). After collection, cells were washed twice with ice cold 1× PBS (Gibco) supplemented with 0.1% BSA (Millipore Sigma). Cells were then filtered with a 35-µm cell strainer (Corning) and cell viability and concentration were measured with Trypan blue on the TC20 Automated Cell Counter (Bio-Rad). Cell viability was greater than 90% for all samples. See the Nature Research Reporting Summary for more information.

Mouse tissues. Flash-frozen adult mouse whole-brain tissue was purchased from BrainBits (SKU: C57AWB). Nuclei isolation was performed using the Omni-ATAC protocol for isolation of nuclei from frozen tissues[22]. Nuclei permeability and concentration were measured with Trypan blue on the TC20 Automated Cell Counter. For all samples, over 95% of the nuclei were permeable to Trypan blue, meaning that the nuclei isolation was successful.

Human peripheral blood and bone marrow cells. Cryopreserved human BMMCs, isolated bone marrow $CD34^+$ stem/progenitor cells, PBMCs, and isolated peripheral blood $CD4^+$, $CD8^+$, $CD14^+$, $CD19^+$ and $CD56^+$ cells were purchased from Allcells (see Supplementary Table 5 for catalog numbers and donor information). Cells were quickly thawed in a 37 °C water bath, rinsed with culture medium (IMDM medium supplemented with 10% FBS and 1% Pen–Strep) and then treated with 0.2 U µl^{-1} DNase I (Thermo Fisher Scientific) in 10 ml of culture medium at 37 °C for 30 min. After DNase I treatment, cells were washed once with medium and then twice with ice-cold 1× PBS with 0.1% BSA. Cells were then filtered with a 35-µm cell strainer (Corning) and cell viability and concentration were measured with Trypan blue on the TC20 Automated Cell Counter (Bio-Rad). Cell viability was greater than 80% for all samples.

Human BMMC stimulation. BMMCs were quickly thawed in a 37 °C water bath, rinsed with culture medium (RPMI 1640 medium supplemented with 15% FBS and 1% Pen–Strep) and then treated with 0.2 U µl^{-1} DNase I in 10 ml of culture medium at 37 °C for 30 min. After DNase I treatment, cells were washed with medium once and filtered with a 35-µm cell strainer. Cell viability and concentration were measured with Trypan blue on the TC20 Automated Cell Counter. Cell viability was greater than 90% for all samples. Cells were plated at a concentration of 1×10^6 cell per milliliter, rested at 37 °C and 5% CO_2 for 1 h and then either incubated in serum-containing medium (RPMI 1640 medium supplemented with 15% FBS and 1% Pen–Strep) at 37 °C and 5% CO_2 for 6 h (ex vivo culture) or treated with 20 ng ml^{-1} LPS (tlrl-3pelps, Invivogen) for 6 h (LPS stimulation). After stimulation, cells were washed twice with ice-cold 1× PBS with 0.1% BSA and cell viability and concentration were measured with Trypan blue on the TC20 Automated Cell Counter. As a control, we processed cells immediately after counting, without any incubation.

Cell lysis and tagmentation. For a detailed description of tagmentation protocols and buffer formulations, refer to the SureCell ATAC-Seq Library Prep Kit (17004620, Bio-Rad) User Guide (10000106678, Bio-Rad). Collected cells and tagmentation buffers were chilled on ice. For cell lines, a protocol based on Omni-ATAC was followed[22]. In brief, washed and pelleted cells were lysed with the Omni-ATAC lysis buffer containing 0.1% NP-40, 0.1% Tween-20, 0.01% digitonin, 10 mM NaCl, 3 mM $MgCl_2$ and 10 mM Tris-HCl pH 7.4 for 3 min on ice. The lysis buffer was diluted with ATAC-Tween buffer that contains 0.1% Tween-20 as a detergent. Cells were collected and resuspended in Omni Tagmentation Mix. This mix is formulated with ATAC Tagmentation Buffer and ATAC Tagmentation Enzyme, both of which are included in the SureCell ATAC-Seq Library Prep Kit. The Omni Tagmentation Mix was buffered with 1× PBS supplemented with 0.1% BSA. Cells were mixed and agitated on a ThermoMixer (5382000023, Eppendorf) for 30 min at 37 °C. Tagmented cells were kept on ice before encapsulation.

For PBMCs and BMMCs, lysis was performed simultaneously with tagmentation. Washed and pelleted cells were resuspended in Whole Cell Tagmentation Mix containing 0.1% Tween-20, 0.01% digitonin, 1× PBS supplemented with 0.1% BSA, ATAC Tagmentation Buffer and ATAC Tagmentation Enzyme. Cells were tagmented using a thermal protocol and maintained as described in the Omni-ATAC protocol above.

For mouse tissues, nuclei were washed with ATAC-Tween buffer containing 0.1% Tween-20, 10 mM NaCl, 3 mM $MgCl_2$ and 10 mM Tris-HCl pH7.4 and then processed according to the whole-cell protocol above.

Optimized Tn5 concentration. To test whether the concentrated Tn5 (part of the SureCell ATAC-Seq Library Prep Kit, 17004620, Bio-Rad) performed better than the standard commercial Tn5 enzyme (TDE1, 15027865, Illumina), we prepared dscATAC-seq libraries for K562 cells using different amounts of TDE1 and our new concentrated Tn5. K562 cells were prepared and lysed as specified in the Omni-ATAC protocol described above. Cells were then resuspended in Omni Tagmentation Mix containing ATAC Tagmentation Buffer and either (1) different amounts of TDE1 (2.5, 7.5 or 10 µl in a 50 µl reaction; Fig. 1b) or (2) the concentrated Tn5 (2.5 µl in a 50 µl reaction, three replicates; Fig. 1b). Cells were mixed and agitated on a ThermoMixer for 30 min at 37 °C. Tagmented cells were kept on ice before encapsulation and libraries were prepared using our standard method as described below. The top 500 cells sorted by library complexity are shown for all comparisons.

Droplet library preparation and sequencing. For a detailed protocol and complete formulations, refer to the SureCell ATAC-Seq Library Prep Kit User Guide (10000106678, Bio-Rad). Tagmented cells or nuclei were loaded onto a ddSEQ Single-Cell Isolator (12004336, Bio-Rad). scATAC-seq libraries were prepared using the SureCell ATAC-Seq Library Prep Kit (17004620, Bio-Rad) and SureCell ATAC-Seq Index Kit (12009360, Bio-Rad). Bead barcoding and sample indexing were performed in a C1000 Touch thermal cycler with a 96-Deep Well Reaction Module (1851197, Bio-Rad); PCR conditions were as follows: 37 °C for 30 min, 85 °C for 10 min, 72 °C for 5 min, 98 °C for 30 s, eight cycles of 98 °C for 10 s, 55 °C for 30 s and 72 °C for 60 s, and a single 72 °C extension for 5 min to finish. Emulsions were broken and products were cleaned up using Ampure XP beads (A63880, Beckman Coulter). Barcoded amplicons were further amplified using a C1000 Touch thermal cycler with a 96-Deep Well Reaction Module; PCR conditions were as follows: 98 °C for 30 s, seven to nine cycles (cycle number depending on the cell input, Table 21 of the User Guide) of 98 °C for 10 s, 55 °C for 30 s and 72 °C for 60 s, and a single 72 °C extension for 5 min to finish. PCR products were purified using Ampure XP beads and quantified on an Agilent Bioanalyzer (G2939BA, Agilent) using the High-Sensitivity DNA kit (5067-4626, Agilent). Libraries were loaded at 1.5 pM on a NextSeq 550 (SY-415-1002, Illumina) using the NextSeq High Output Kit (150 cycles; 20024907, Illumina) and sequencing was performed using the following read protocol: read 1, 118 cycles; i7 index read, 8 cycles; read 2, 40 cycles. A custom sequencing primer (part of the SureCell ATAC-Seq Library Prep Kit) is required for read 1.

dsciATAC-seq methods. *Assembly of indexed Tn5 transposome complexes.* To generate indexed Tn5 transposome complexes, we modified the Illumina Nextera Read 1 Adapter to contain a 6-nucleotide barcode (96 distinct barcodes; see Supplementary Table 6 for barcode sequences). Each indexed oligonucleotide was mixed with the Illumina Nextera Read 2 Adapter and annealed to a 15-nucleotide mosaic-end complementary oligonucleotide (5′ phosphorylated and 3′ dideoxy-C) (Supplementary Table 6). All oligonucleotides were purified by high-performance liquid chromatography (IDT). For the annealing reaction, oligonucleotides were mixed at a 1:1:2 molar ratio (read 1:read 2:complementary mosaic end) at 100 µM final concentration in 50mM NaCl. The mixture was incubated at 85 °C, ramped down to 20 °C at a rate of −1 °C min^{-1} and then held at 20 °C for 2 min. After being diluted 1:1 in glycerol, the annealed oligonucleotide mixture was mixed 1:1 with 14.8 µM purified Tn5 (Tn5 was purified as previously described[38]). The Tn5–oligonucleotide mixture was incubated for 30 min at room temperature and then kept at −20 °C before the tagmentation reactions.

Species mixing controls. Human and mouse cell lines were processed and lysed using the Omni-ATAC-seq protocol as described above. For the 24-plex control experiment in Fig. 3 and Supplementary Fig. 6, K562 and NIH/3T3 cells were mixed at a 1:1 ratio and tagmented with Tn5 loaded with indexed oligonucleotides 1–3, 13–15, 25–27, 37–39, 49–51, 61–63, 73–75 and 85–87 (Supplementary Table 6) in 50-µl reactions (10 µl of indexed Tn5 per reaction) with 25,000 cells each. Cell line tagmentation buffer components and reaction conditions were the same as described above. After the tagmentation reaction, all cells were pooled, washed with tagmentation buffer without Tn5 and processed using our standard protocol for droplet library preparation and sequencing. Different cell numbers were used as input, as indicated in Fig. 3 and Supplementary Fig. 6.

For the 48-plex control experiment in Supplementary Fig. 6, K562 and NIH/3T3 cells were mixed at a 1:1 ratio and tagmented with Tn5 loaded with indexed oligonucleotides 1–6, 13–18, 25–30, 37–42, 49–54, 61–66, 73–78 and 85–90 (Supplementary Table 6) in 50-µl reactions (10 µl of indexed Tn5 per reaction) with 25,000 cells each. Cell line tagmentation buffer components and reaction conditions were the same as described above. After the tagmentation reaction, all cells were pooled, washed with tagmentation buffer without Tn5 and processed using our standard protocol for droplet library preparation and sequencing. Different cell numbers were used as input, as indicated in Supplementary Fig. 6.

Human BMMC stimulations. BMMCs from two donors were stimulated and washed as described above. For the experiment in Figs. 4 and 5, BMMCs were tagmented with Tn5 loaded with indexed oligonucleotides 1–96 in 20-µl reactions (4 µl of indexed Tn5 per reaction) with 8,000 cells each (control, ex vivo culture and LPS stimulation as described above). BMMC tagmentation buffer components and reaction conditions were the same as described above. After the tagmentation reaction, all cells were pooled, washed with tagmentation buffer without Tn5 and processed using our standard protocol for droplet library preparation and sequencing. Pooled cells were split into 16 different samples for droplet library preparation, with varying cell inputs (20,000, 40,000 or 80,000 cells). After sequencing, data from all 16 samples were merged for the analyses.

Sequencing data for the dsciATAC-seq experiments were processed with BAP as described below using the '--tn5-aware' flag that inhibits cell merging across different Tn5 barcodes.

Bioinformatics preprocessing. *Raw read processing.* Per-read bead barcodes were parsed and trimmed using UMI-TOOLs (https://github.com/CGATOxford/UMI-tools)[56], and the remaining read fragments were aligned using BWA (http://bio-bwa.sourceforge.net/) on the Illumina BaseSpace online application. Constitutive elements of the bead barcodes were assigned to the closest known sequence, allowing for up to one mismatch per 6-mer or 7-mer (mean > 99% parsing efficiency across experiments). For the dsciATAC-seq experiments, bead barcodes were parsed using a custom Python script aware of the 96 possible Tn5 barcodes. All experiments were aligned to the hg19 or mm10 reference genome (or a combined reference genome in the case of species mixing experiments).

To identify systematic biases (i.e. reads aligning to an inordinately large number of barcodes) and deduplicate reads with barcode awareness, and to perform bead merging (see below), we developed the BAP tool. This software uses as input a .bam file for a given experiment with a bead barcode identifier indicated by a SAM tag. We generalized this preprocessing pipeline to handle other datasets (Fluidigm C1 and sciATAC-seq) to enable consistent comparisons across various technologies (Fig. 1). For additional details regarding this pipeline, including dynamic inference of per-library thresholds, see the Supplementary Note.

Identification of multiple beads per droplet. An integral part of our technique relies on the robust identification of pairs of bead barcodes that share exact insertions at a rate that exceeds what may be expected by chance. We note that our procedure readily enables multiple beads per droplet (Supplementary Fig. 2). First, highly abundant barcodes were detected in the experiment by quantifying each unique barcode sequence among nuclear-mapping reads. Our knee-calling algorithm then established a per-experiment bead threshold. Next, sequencing reads that were assigned to a bead barcode and passed filtering were deduplicated using the insert positions of the paired-end reads (as previously implemented in Picard tools).

After initial deduplication, we further removed paired-end reads that mapped to more than six bead barcodes, reasoning that these represented technical confounders. Next, for each pair of bead barcodes that passed the bead filtering step, we computed the Jaccard index over the insertion positions of reads, providing a measure of how similar the Tn5 insertions were for any pair of bead barcodes. From these pairwise Jaccard index statistics, we performed a second knee call to determine pairs that were likely to have originated from the same droplet (Supplementary Fig. 2f). Finally, to assign droplet-level barcodes, we looped over the original bead barcodes in order of their original nuclear read abundance. For a given bead barcode, if it paired with any other bead barcodes that passed the pairwise knee, those bead barcodes were 'merged' into one droplet barcode. This iteration repeats until all bead barcodes have been assigned to precisely one droplet barcode. To facilitate comparisons without droplet merging (Supplementary Fig. 2n,o), our pipeline includes a '--one-to-one' flag, which maps one bead barcode onto one droplet barcode; this option was employed primarily to process other scATAC-seq datasets that would not have beads requiring merging. Additional details regarding this procedure and comparisons in Supplementary Fig. 2k are discussed in the Supplementary Note.

Species mixing analysis. We carried out the same quantification procedure for all species mixing datasets analyzed in this work. Namely, reads were mapped to a hybrid hg19–mm10 reference genome using BWA. Cells were identified using BAP knee calling (described above). The output of this pipeline yields the number of unique nuclear reads mapping to the mouse and human genomes, which were compared between each cell. We further excluded cells with fewer than 1,000 reads mapping to either the human or mouse genome and identified collisions as those that had less than 10× enrichment over the minor genome. The overall collision rate is reported as the number of annotated collision cells over the total number of cells compared (mouse + human + collisions).

Peak calling. For each scATAC-seq experimental sample, chromatin accessible summits were called using MACS2 callpeak with custom parameters that have been described previously[17]. To generate a non-overlapping set of peaks per analysis, we first extended summits of each experiment to 500-bp windows (±250 bp). We combined these 500-bp peaks, ranked them by their summit significance value and retained specific non-overlapping peaks on the basis of this ordering. We further removed peaks that overlapped the ENCODE blacklist and a custom mitochondrial blacklist generated by aligning a synthetic mitochondrial DNA genome to the nuclear genome (https://github.com/buenrostrolab/mitoblacklist).

Library complexity estimation. Per-cell library complexities were estimated using the Lander–Waterman equation[57] with a custom R function translated from a previously established Java function implemented in Picard tools. Per-cell counts of the total number of mapped nuclear reads that passed quality filters and the number of unique nuclear reads served as inputs. Thus, library complexity represents a metric that estimates the total number of unique nuclear reads from the cell independently of sequencing depth.

Comparison to public datasets. To benchmark the dscATAC-seq platform against existing datasets, we downloaded raw sequencing data (.fastq format) for GM12878 cells from three different combinatorial indexing scATAC-seq methods[19,24,25] and 384 cells processed with the Fluidigm C1 (ref. [18]) from GEO. All datasets were processed using the same pipeline, which included BWA alignment and downstream processing with BAP using the '--one-to-one' flag that skips bead merging. We note that, in all three combinatorial indexing scATAC-seq experiments, GM12878 cells were mixed with mouse cells. As such, we compared only annotated human cells (>9:1 ratio of human:mouse cells) from these experiments for downstream analysis.

To determine the correlation between scATAC-seq experiments, we used a merged peak set comprising 175,581 combined DNase–seq hypersensitivity peaks from GM12878 and K562 made available through the ENCODE Project. The sum of single cells (agnostic to cell ID) was compared against bulk DNase–seq profiles generated from ENCODE and Omni-ATAC[22]. To score the fraction of reads in peaks across single-cell experiments, we used only the GM12878 DNase–seq peak set (124,321 peaks) to ensure that peak selection did not bias our quantification and comparison of technologies.

Validation of multiple beads per droplet inference. To validate our ability to merge cells marked by multiple droplet beads, we introduced a diverse library of random oligonucleotides (14 nucleotides long; see Supplementary Table 6 for the full sequence) to our microfluidic reaction (Supplementary Fig. 2). Human PBMCs were processed with this library of random oligonucleotides at bead concentrations of 200, 800 and 5,000 beads per microliter, spanning the ranges used for the data presented in this work. The random oligonucleotides were spiked into the cells at a final concentration of 5 nM after the tagmentation reaction, and samples were processed and sequenced using our standard protocol (described above). Among pairs of beads that were merged, the average number of oligonucleotides observed per bead ranged from 792 to 1,979 per experiment.

We reasoned that bead barcodes sharing a noticeable overlap of these oligonucleotides (Supplementary Fig. 2a,b) would be barcodes from two beads contained in the same droplet. We identified reads containing our random oligonucleotide by first identifying the 15-bp constant sequence and subsequently parsing the 14 bases downstream of the constant sequence (Supplementary Table 6). For each experiment, we called a knee on the bead-barcode pairwise Jaccard indices (for each observed 14-base oligonucleotide) and computed the overlap of random sequences observed (Supplementary Fig. 2e) for barcodes passing the nuclear read knee. Pairs of bead barcodes that passed the oligonucleotide-overlap knee were annotated as true positives.

Next, we computed our BAP metric in a pairwise manner for each bead barcode using the overlap of pairs of inserts over each fragment (or paired-end read). This produces a metric for all pairs of bead barcodes with at least 500 unique nuclear reads observed per barcode (Supplementary Fig. 2f). Using the true positives defined from the random oligonucleotide data and a continuous overlap metric from BAP, we computed precision-recall and receiver-operating curves (AUROC = 1.000 and AUPRC = 0.997; Supplementary Fig. 2k). We further compared other possible metrics for bead merging, including Pearson and Spearman correlation and a Jaccard index over reads in peaks, but found that our approach was the most robust and specific (Supplementary Fig. 2k). We note that the library of random oligonucleotides provides a completely orthogonal measure of bead overlap as compared to the nuclear DNA fragments used in the BAP algorithm.

Theory of beads and droplet concentrations. In this setting, we are interested in estimating the number of beads per droplet at variable bead concentrations using observed data. Given that our observed data do not yield any droplets with 0 beads (cells not captured) and that any measurement with greater than six beads cannot be relied on (six is the physical limit for beads; thus, observed values likely reflect merged droplets), the observed number of beads per droplet is modeled by a double-truncated Poisson distribution. The probability density function of a double-truncated Poisson distribution for a single observation can be written as follows.

$$\Pr(Y_i = y_i \mid c_1 \le y_i \le c_2) = \frac{\lambda^{y_i}}{y_i! \sum_{k=c_1}^{c_2} \lambda^{y_i}/k!}$$

Here c_1 is our lower bound (1 in our case) of the empirical data and c_2 is the upper bound (in our case 6) for observed numbers of beads per droplet y. Let

$i \in \{1, 2, \ldots, n\}$. Then, we observe n cells and y_i denotes the number of beads per drop for cell i. The log likelihood (l) of observing a value can thus be computed as follows.

$$l(\lambda \mid y) = \sum_{i=1}^{n} y_i \log(\lambda) - \sum_{i=1}^{n} \log(y_i!) - \log\left(\sum_{k=c_1}^{c_2} \frac{\lambda^k}{k!}\right)$$

Here, a closed-form solution of λ (parameter of the Poisson distribution indicating the mean number of beads per cell) is impossible. Thus, we estimate the value using the optim() function in R, providing the maximum-likelihood estimate (MLE).

Given the MLE estimate for λ, we can calculate the proportion of droplets with 0 beads p using the Poisson probability density function

$$p = \exp\{-\lambda\}$$

We can then approximate the number of droplets with a barcode as $1 - p$. Empirical values of λ were determined using GM12878 and mouse brain data at different bead concentrations (800 and 5,000 beads per microliter) and were found to be robust across the various datasets analyzed.

De novo *k*-mer clustering. Here we computed bias-corrected deviation z scores for K k-mers and a set of S samples (dscATAC-seq cells) with P peaks computed via the chromVAR methodology. Our implementation utilizes a binarized matrix $\boldsymbol{M}$ (dimensions P by K) in which $m_{i,k}$ is 1 if k-mer k is present in peak i and 0 otherwise on the basis of the reference genome annotation. For all applications, we used $k = 7$, resulting in $K = 8{,}192$ ($4^7/2$) 7-mers. We note that the division by two is to account for reverse-complement k-mers that would be identical as both strands of the reference genome are considered when building $\boldsymbol{M}$. Using the matrix of fragment counts in peaks $\boldsymbol{X}$ (dimensions P by S), where $x_{i,j}$ represents the number of fragments from peak i in sample j, we produce a deviation score matrix $\boldsymbol{Z}$ of dimensions S samples (rows) and K 7-mers (columns).

The matrix $\boldsymbol{Z}$ is computed using an expectation of peak accessibility based on technical confounders present in assays (differential PCR amplification or variable Tn5 tagmentation conditions). This is achieved by generating 50 background peaks intrinsic to the set of epigenetic data being examined. The full details describing the computation of $\boldsymbol{Z}$ have been previously described in the chromVAR manuscript[26]. Finally, as many of the 8,192 7-mers are highly correlated, we then use the top principal components of the matrix $\boldsymbol{Z}$ as input for downstream processes, including the Louvain clustering and t-SNE embedding.

Cell-type-specific promoter-region chromatin accessibility scores and regulatory region analysis in mouse brain. To define cluster-specific regulatory elements and promoter region chromatin accessibility scores, we defined pseudo-bulk cell types by aggregating the counts per cell over each of the annotated cluster definitions. First, the peak × cell-type counts matrix ($\boldsymbol{X}$) was count-per-million normalized, and peaks with overall mean counts per million > 1 were retained. This filtered peak × cell-type counts matrix was then z score transformed. Explicitly, for cell type j and peak i, our transformed statistic was

$$z_{i,j} = \frac{x_{i,j} - \text{mean}(x_{i,\star})}{\text{sd}(x_{i,\star})}$$

We identified 135,737 cell-type-specific chromatin accessibility peaks with $z_{i,j} > 3$ in at least one cell type (some value j), which were assigned to clusters on the basis of maximum z score value ($\text{argmax}_j z_{i,j}$). Peaks were separated and clustered on the basis of the population with the maximum value in Fig. 2e. An identical procedure was used for the promoter region chromatin accessibility scores × cell-type counts matrix starting with the annotated set of 310 marker genes from a previous scRNA-seq analysis of mouse brain[29], resulting in 262 genes for which the $z_{i,j} > 3$ criterion was met for the promoter gene scores (Fig. 2d).

Promoter region chromatin accessibility scores. To annotate our de novo clusters from the whole mouse brain, we computed per-cluster promoter region chromatin accessibility scores representing a weighted sum of chromatin accessibility around the TSS of each gene in our reference data. Specifically, for gene g and cluster i, we define a chromatin accessibility score g_i from the following.

$$g_i = \sum_{j \in J} x_{i,j} * e^{-d_j/k}$$

Here $x_{i,j}$ represents the counts-per-million normalized chromatin accessibility count for cluster i and chromatin accessibility peak j. Accessibility peaks used per gene J were restricted to those within 100,000 bp of a corresponding TSS, and d_j represents the distance (in base pairs) between the TSS and the center of peak j. The scaling constant, k, was fixed to 5,000 for all chromatin accessibility score computations.

Mouse brain cluster annotation. To annotate the dscATAC-seq mouse brain clusters in a data-driven manner based on the molecular signature of the distinct cell types in the brain, we used a resource containing scRNA-seq data for 690,000 individual cells sampled from nine regions of the adult mouse brain[29], which identified 565 subclusters within the broad classes of cell types in the brain. The list of cell types includes neurons, astrocytes, microglia, oligodendrocytes, polydendrocytes and components of the vasculature. We note that many of these subclusters are from analysis of specific brain regions and further reclustering within broadly defined clusters, leading to a large number of clusters. We use this data resource to (1) assign each one of our clusters to one of the broad cell classes identified in their study and (2) further refine the annotation by identifying which gene expression signature (within the 565 subclusters) provides an optimal match to each one of our dscATAC-seq clusters. To do this, we first obtained the union of the class_marker and type_marker genes identified in the scRNA-seq study (total of 310 unique genes)[29]. We then calculated the Spearman correlation coefficient between the per-cluster promoter-region chromatin accessibility scores (27 clusters) and the aggregated scRNA-seq signal per cluster (565 clusters) at those 310 marker genes. We then employed the Gale–Shipley algorithm to assign an optimal matching of scRNA-seq clusters to our scATAC-seq clusters (Supplementary Table 2). Here, the Gale–Shipley algorithm assigns pairs that maximize the global utility of the matches, noting our utility function was Spearman correlation. To classify the 27 dscATAC-seq clusters, we used the broad class assignment of the most correlated scRNA-seq cluster, except for the 'Neuron' class, which was further divided into excitatory and inhibitory neurons on the basis of the annotation of *Slc17a7* and *Gad1*, respectively. We then performed the same computational approach using another scRNA-seq dataset with 262 clusters[30] to validate the robustness of our approach (Supplementary Table 2). When displaying the overall correlation structure (Fig. 2c), we restricted the scRNA-seq clusters to those that had one or more class matches to the scATAC-seq data (500 of 565 clusters).

Bulk-guided clustering. Bulk-guided clustering of single cells (Fig. 4) was performed as previously described[21]. In brief, a matched peak set ($k = 156{,}311$ peaks) was used for both BMMC dsciATAC-seq ($n = 136{,}463$ single cells) and bulk ATAC-seq profiles previously generated for sorted hematopoietic cell populations (16 cell types)[17,21,46]. Principal-component analysis was first run on quantile-normalized bulk ATAC-seq data generating principal components capturing variation across cell types. Single cells were then projected in the space of these bulk-trained principal components by multiplying the scATAC-seq reads in the peaks matrix with the matrix of peaks with principal-component loading coefficients to yield a matrix of single-cell projection scores (cells × principal components). The derived single-cell scores were then scaled and centered, and the corresponding single-cell data were visualized using t-SNE. Predicted labels for single cells were obtained by correlating projected single-cell scores with bulk principal-component scores and choosing the most correlated bulk cell type on the basis of the Pearson correlation coefficient. To define clusters for the control (unstimulated) BMMC dataset (Fig. 4c), Louvain clustering was performed using the igraph package where the 20 nearest neighbors per cell were used for embedding.

Single-cell classification. To assign the most similar clusters generated from the 15 clusters of the control (unstimulated) BMMC dataset (Fig. 4c) to the additional datasets (Fig. 4e,f), the medoids of each per-cluster principal component were determined over all cells assigned from the Louvain clustering at baseline. Next, for every cell in each of the new datasets (that is, the FACS-sorted populations and the bead-isolated populations), we assigned a reference cluster on the basis of the minimum Euclidean distance between each cell's principal components and the medoids of the clusters.

Analysis of differential TF motifs. To compute differential TF scores under normal and stimulation conditions, we determine the 20 nearest stimulus-condition neighbors for each single cell in the resting condition using the bulk-guided principal-component scores and a Pearson correlation distance metric. To calculate differential TF motifs, we subtract the mean of the 20 stimulus cells by the TF score for each cell in the normal condition. Finally, to suppress noise in the comparison, we smooth the differential TFs by taking the mean of the 20 nearest neighbors in the control condition. Again, the nearest neighbors were calculated using the bulk-guided principal-component scores, with Pearson correlation as a distance metric.

Differential peak identification in bone marrow stimulation. We devised a permutation test that assessed whether the proportion of cells with an accessibility element was different between the stimulated and resting conditions, controlling for overall differences in accessibility (using measures at promoters). First, we filtered our consensus peak set such that a given peak was accessible in at least 1% of cells irrespective of stimulation or resting condition. Then, for an individual regulatory element i, we determined the proportion of cells in the resting (P_r) and stimulated (P_s) conditions that observed one or more fragments overlapping the accessibility peak. Next, we computed the proportion of all promoters annotated in our dataset for both resting (P'_r) and stimulated (P'_s). Our observed differential statistic thus is given by

$$\frac{P_s}{P'_s} - \frac{P_r}{P'_r}$$

To determine statistical significance, we permuted the stimulation and resting labels 1,000 times to generate a permuted distribution. We observed the corresponding z statistic (Supplementary Fig. 9l) to be centered with a largely Gaussian distribution. After converting these z statistics to P values using a standard normal distribution, we computed a per-cluster FDR and established a significance threshold of 1% uniformly across clusters. We further computed an effect size of the difference between stimulated and resting, given simply by $P_s - P_r$. We summarized the differential association in Fig. 5g where the red bars (FDR $< 10^{-5}$) and the pink bar (FDR $< 10^{-2}$) represent the statistical significance of the change in chromatin accessibility for each cell-type cluster.

Overlap with fine-mapped GWAS SNPs. To identify regulatory regions affected by our stimulation conditions that may be relevant for human disease, we overlapped differential peaks identified per cell type with SNPs identified through genetic fine-mapping studies of 21 immune traits as previously described[45]. Specifically, we downloaded the per-SNP metadata available online (http://pubs.broadinstitute.org/pubs/finemapping/dataportal.php) and intersected differentially accessible peaks with annotated positions of fine-mapped variants with PP > 0.3 computed by PICS[45] across all reported traits.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw sequencing files and processed files for all data generated in this study were deposited at Gene Expression Omnibus (GEO) under accession number GSE123581. UCSC genome browser tracks for the datasets generated in this study are available from the following websites: mouse brain, https://s3.us-east-2.amazonaws.com/jasonbuenrostro/2018_mouse_brain/hub.txt; BMMC dsciATAC-seq, https://s3.us-east-2.amazonaws.com/jasonbuenrostro/2018_BM_htsci/hub.txt; stimulated BMMC dsciATAC-seq, https://s3.us-east-2.amazonaws.com/jasonbuenrostro/2018_BM_htsci_stim/hub.txt.

Code availability

Complete code and documentation for the BAP software suite developed in this study is available at https://github.com/buenrostrolab/bap. Scripts corresponding to the analyses contained in this paper are provided at https://github.com/buenrostrolab/dscATAC_analysis_code.

References

56. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
57. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).

nature research

Corresponding author(s): Jason Buenrostro

Last updated by author(s): Apr 16, 2019

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Complete code and documentation for the software suite developed in this study (bap - bead-based ATAC-seq processing tool) is available on GitHub under the following weblink: https://github.com/buenrostrolab/bap. Scripts corresponding to the analyses contained in this paper are further provided at: https://github.com/buenrostrolab/dscATAC_analysis_code.
Data analysis	Complete code and documentation for the software suite developed in this study (bap - bead-based ATAC-seq processing tool) is available on GitHub under the following weblink: https://github.com/buenrostrolab/bap. Scripts corresponding to the analyses contained in this paper are further provided at: https://github.com/buenrostrolab/dscATAC_analysis_code.

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

Data

Policy information about availability of data

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

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

Raw sequencing files and processed files for all data generated in this study were deposited at Gene Expression Omnibus (GEO) under accession number GSE123581. UCSC genome browser tracks for the datasets generated in this study are available through the following weblinks:
Mouse brain: https://s3.us-east-2.amazonaws.com/jasonbuenrostro/2018_mouse_brain/hub.txt
BMMC dsciATAC-seq: https://s3.us-east-2.amazonaws.com/jasonbuenrostro/2018_BM_htsci/hub.txt
BMMC-stim dsciATAC-seq: https://s3.us-east-2.amazonaws.com/jasonbuenrostro/2018_BM_htsci_stim/hub.txt

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

Sample size	We used 2 biological replicates (2 independent donors) for both the mouse brain and the bone marrow mononuclear cells experiments. Results were highly correlated across biological replicates; additional samples were not considered due to the high cost of these experiments.
Data exclusions	All datasets generated in this study were filtered using standard quality thresholds commonly used for ATAC-seq data. All filters used are specified in the Methods.
Replication	In Figures S4c and S7a, we show that cells from 2 biological replicates are distributed evenly among clusters. In Figure S4d we show that cells from different technical replicates are distributed evenly among clusters.
Randomization	Randomization is not relevant to our study, as our statistical tools are not dependent on randomization.
Blinding	Blinding is not relevant to our study, as our statistical tools are not dependent on blinding.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	GM12878: Coriell Institute for Medical Research. K562 and NIH/3T3: ATCC.
Authentication	Cell lines were authenticated at the respective repositories (ATCC or Coriell Institute for Medical Research) prior to shipping.
Mycoplasma contamination	Cell lines were tested for Mycoplasma contamination at the respective repositories (ATCC or Coriell Institute for Medical Research) prior to shipping. Furthermore, Mycoplasma contamination is easily identifiable in ATAC-seq data, and we did not observe any evidence of Mycoplasma DNA in any of our datasets.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell line was used in this study.