

ARTICLE

<https://doi.org/10.1038/s41467-018-07771-0>

OPEN

A rapid and robust method for single cell chromatin accessibility profiling

Xi Chen ¹, Ricardo J. Miragaia ^{1,4}, Kedar Nath Natarajan ^{1,5} & Sarah A. Teichmann ^{1,2,3}

The assay for transposase-accessible chromatin using sequencing (ATAC-seq) is widely used to identify regulatory regions throughout the genome. However, very few studies have been performed at the single cell level (scATAC-seq) due to technical challenges. Here we developed a simple and robust plate-based scATAC-seq method, combining upfront bulk Tn5 tagging with single-nuclei sorting. We demonstrate that our method works robustly across various systems, including fresh and cryopreserved cells from primary tissues. By profiling over 3000 splenocytes, we identify distinct immune cell types and reveal cell type-specific regulatory regions and related transcription factors.

¹ Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ² EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ³ Theory of Condensed Matter, Cavendish Laboratory, 19 JJ Thomson Ave, Cambridge CB3 0HE, UK. ⁴ Present address: MedImmune, Sir Aaron Klug Building, Granta Park, Cambridge CB21 6GH, UK. ⁵ Present address: Functional Biology and Metabolism Unit, Biochemistry and Molecular Biology, SDU, 5230 Odense, Denmark. Correspondence and requests for materials should be addressed to S.A.T. (email: st9@sanger.ac.uk)

Due to its simplicity and sensitivity, ATAC-seq¹ has been widely used to map open chromatin regions across different cell types in bulk. Recent technical developments have allowed chromatin accessibility profiling at the single cell level (scATAC-seq) and revealed distinct regulatory modules across different cell types within heterogeneous samples^{2–9}. In these approaches, single cells are first captured by either a microfluidic device³ or a liquid deposition system⁷, followed by independent tagmentation of each cell. Alternatively, a combinatorial indexing strategy has been reported to perform the assay without single cell isolation^{2,4,9}. However, these approaches require either a specially engineered and expensive device, such as a Fluidigm C1³ or Takara ICELL8⁷, or a large quantity of customly modified Tn5 transposase^{2,4,5,9}.

Here, we overcome these limitations by performing upfront Tn5 tagging in the bulk cell population, prior to single-nuclei isolation. It has been previously demonstrated that Tn5 transposase-mediated tagmentation contains two stages: (1) a tagging stage where the Tn5 transposome binds to DNA, and (2) a fragmentation stage where the Tn5 transposase is released from DNA using heat or denaturing agents, such as sodium dodecyl sulfate (SDS)^{10–12}. As the Tn5 tagging does not fragment DNA, we reasoned that the nuclei would remain intact after incubation with the Tn5 transposome in an ATAC-seq experiment. Based on this idea, we developed a simple, robust and flexible plate-based scATAC-seq protocol, performing a Tn5 tagging reaction^{6,13} on a pool of cells (5000–50,000) followed by sorting individual nuclei into plates containing lysis buffer. Tween-20 is subsequently added to quench the SDS in the lysis buffer¹⁴, which otherwise will interfere the downstream reactions. Library indexing and amplification are done by PCR, followed by sample pooling, purification and sequencing. The whole procedure takes place in one single plate, without any intermediate purification or plate transfer steps (Fig. 1a). With this easy and quick workflow, it only takes a few hours to prepare sequencing-ready libraries, and the method can be implemented by any laboratory using standard equipment.

Results

Benchmark and comparison to Fluidigm C1 scATAC-seq. We first tested the accuracy of our sorting by performing a species mixing experiment, where equal amounts of HEK293T and NIH3T3 cells were mixed, and scATAC-seq was performed with our method. Using a stringent cutoff (Online Methods), we recovered 307 wells, among which 303 wells contain predominantly either mouse fragments ($n = 136$) or human fragments ($n = 167$). Only 4 wells are categorised as doublets (Fig. 1b).

To compare our plate-based method to the existing Fluidigm C1 scATAC-seq approach, we performed side-by-side experiments, where cultured K562 and mouse embryonic stem cells (mESC) were tested by both approaches. We used three metrics to evaluate the quality of the data generated by both methods (Fig. 1c and Supplementary Figure 1a). Our plate-based method has higher library complexity (library size estimated by the Picard tool), comparable or lower amount of mitochondrial DNA, and higher signal-to-noise ratio measured by fraction of reads in peaks (FRiP) (Fig. 1c). In addition, visual inspection of the read pileup from the aggregated single cells suggested both methods were successful, but data generated from our plate-based method exhibited higher signal (Fig. 1d, e).

The main difference between our method and Fluidigm C1 is the Tn5 tagging strategies. The plate-based method performed Tn5 tagging using a population of cells, whereas it was done in individual microfluidic chambers in the Fluidigm C1. It is possible that the upfront Tn5 tagging is more efficient than tagging in microfluidic chambers.

Validation using different cryopreserved cells. To evaluate the generality of our method, we tested the plate-based method on cryopreserved cells from four tissues: human and mouse skin fibroblasts (hSF and mSF)¹⁵ and mouse cardiac progenitor cells (mCPC) at embryonic day E8.5 and E9.5¹⁶. Cells were revived from liquid nitrogen, and our plate-based method was carried out immediately after revival. The library complexities varied among cell types (Fig. 2a). We obtained median library sizes ranging from 52,747 (mSF) to 104,608.5 (mCPC_E8.5) unique fragments (Fig. 2a). The amount of mitochondrial DNA also varied across cell types but was low in all samples (<13%). All four tested samples had very high signal-to-noise ratio, with a median FRiP ranging from 0.50 (mSF) to 0.60 (hSF) (Fig. 2a). The insert size distributions of the aggregated single cells from all four samples exhibited clear nucleosomal banding patterns (Fig. 2b), which is a feature of high quality ATAC-seq libraries¹. Finally, visual inspection of aggregate of single cell profiles showed clear open chromatin peaks around expected genes (Fig. 2c, d). Details of all tested cells/tissues are summarised in Supplementary Data 1.

Profiling chromatin accessibility of mouse splenocytes. After this validation of the technical robustness of our plate-based method, we further tested it by generating the chromatin accessibility profiles of 3648 splenocytes (after red blood cell removal) from two C57BL/6Jax mice. In total, we performed two 96-well plates and nine 384-well plates. By setting a stringent quality control threshold (>10,000 reads and >90% mapping rate), 3385 cells passed the technical cutoff (>90% successful rate) (Supplementary Figure 3b). The aggregated scATAC-seq profiles exhibited good coverage and signal and resembled the bulk data generated from 10,000 cells by the Immunological Genome Project (ImmGen)¹⁷ (Fig. 3a). The library fragment size distribution before and after sequencing both displayed clear nucleosome banding patterns (Fig. 3b and Supplementary Figure 2a). In addition, sequencing reads showed strong enrichment around transcriptional start sites (TSS) (Fig. 3c), further demonstrating the quality of the data was high.

Importantly, for the majority of the cells, less than 10% (median 2.1%) of the reads were mapped to the mitochondrial genome (Supplementary Figure 3a). Overall, we obtained a median of 643,734 reads per cell, whereas negative controls (empty wells) generated only ~100–1000 reads (Supplementary Figure 3b). In most cells, more than 98% of the reads were mapped to the mouse genome (Supplementary Figure 3b), indicating low level of contamination. The median of estimated library sizes is 31,808.5 (Supplementary Figure 3c). At the sequencing depth of this experiment, the duplication rate of each single cell library is ~95% (Supplementary Figure 3d), indicating that the libraries were sequenced to near saturation. Down-sampling the raw reads (from the fastq files) and repeating the analysis suggest that at 20–30% of our current sequencing depth, the majority of the fragments would have already been captured (Supplementary Figure 4a and b). Therefore, in a typical scATAC-seq experiment, ~120,000 reads per cell are sufficient to capture most of the unique fragments, with higher sequencing depth still increasing the number of detected unique fragments (Supplementary Figure 3e).

Next, we examined the data to analyse signatures of different cell types in the mouse spleen. Reads from all cells were merged, and a total of 78,048 open chromatin regions were identified by peak calling with q -values less than 0.01¹⁸ (Methods). We binarised peaks as “open” or “closed” (Methods) and applied a Latent Semantic Indexing (LSI) analysis to the cell-peak matrix for dimensionality reduction² (Methods). Consistent with previous findings², the first dimension is primarily influenced by sequencing depth (Supplementary Figure 3f). Therefore, we only focused on the second dimension and upwards and

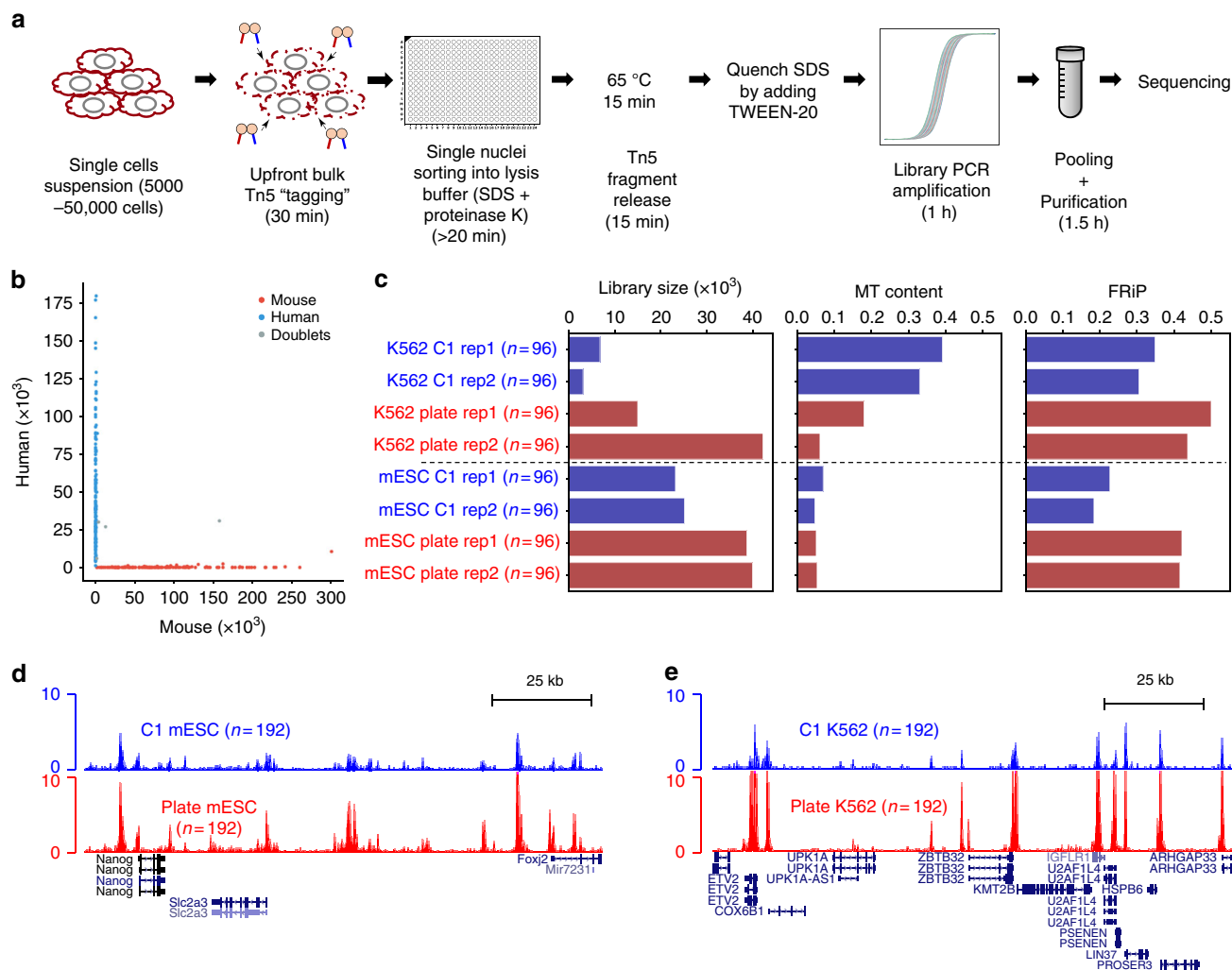


Fig. 1 Simple and robust analysis of chromatin status at the single cell level. **a** Schematic view of the workflow of the scATAC-seq method. Tagmentation is performed upfront on bulk cell populations, followed by sorting single-nuclei into 96/384-well plates containing lysis buffer. The lysis buffer contains a low concentration of proteinase K and SDS to denature the Tn5 transposase and fragment the genome. Tween-20 is added to quench SDS¹⁴. Subsequently, library preparation by indexing PCR is performed, and the number of PCR cycles needed to amplify the library is determined by quantitative PCR (qPCR) (Supplementary Figure 2b). **b** Species mixing experiments to show the accuracy of FACS. Equal amounts of HEK293T (Human) and NIH3T3 (Mouse) cells were mixed, and scATAC-seq was performed as described in **a**. Successful wells with more than 90% of reads uniquely mapped to either human or mouse were categorised as singlets ($n = 303$). Otherwise, they will be categorised as doublets ($n = 4$) (see Methods). **c** Comparison of the median library size (estimated by the Picard tool), fraction of mitochondrial DNA (MT content) and fraction of reads in peaks (FRiP) in single cells from either C1 (blue) or plate-based (red) scATAC-seq approach. **d** UCSC genome browser tracks displaying the signal around the *Nanog* gene locus from the aggregate of mESCs obtained from Fluidigm C1 (top) and plate (bottom). **e** The same type of tracks as **d** around the *ZBTB32* gene locus in K562 cells

visualised the data by t-distributed stochastic neighbour embedding (t-SNE)¹⁹. We did not observe batch effects from the two profiled spleens, and several distinct populations of cells were clearly identified in the t-SNE plot (Fig. 3d). Read counts in peaks near key marker genes (e.g. *Bcl11a* and *Bcl11b*) suggested that the major populations are B and T lymphocytes, as expected in this tissue (Fig. 4a). In addition, we found a small number of antigen-presenting cell populations (Supplementary Figure 5), consistent with previous analyses of mouse spleen cell composition²⁰.

To systematically interrogate various cell populations captured in our experiments, we applied a spectral clustering technique²¹ which revealed 12 different cell clusters (Fig. 4b). Reads from cells within the same cluster were merged together to form ‘pseudo-bulk’ samples and compared to the bulk ATAC-seq data sets generated by ImmGen (Supplementary Figures 6 and 7). Cell clusters were assigned to the most similar ImmGen cell type

(Fig. 4b and Supplementary Figure 7). In this way, we identified most clusters as different subtypes of B, T and Natural Killer (NK) cells, as well as a small population of granulocytes (GN), dendritic cells (DC) and macrophages (MF) (Fig. 4b and Supplementary Data 2). An aggregate of all single cells within the same predicted cell type agrees well with the ImmGen bulk ATAC-seq profiles (Supplementary Figure 8). Remarkably, the aggregate of as few as 55 cells (e.g. the predicted MF cell cluster) already exhibited typical bulk ATAC-seq profiles (Supplementary Figure 8). This finding opens the door for a different ATAC-seq experimental design, where Tn5 tagging can be performed upfront on large populations of cells (e.g. 5000–50,000 cells). Subsequently, cells of interest (for example, marked by surface protein antibodies or fluorescent RNA/DNA probes) can be isolated by FACS, and libraries generated for subsets of cells only. This will be a simple and fast way of obtaining scATAC-seq profiles for rare cell populations.

To test the feasibility of this idea, we stained mouse splenocytes with an anti-CD4 antibody conjugated with PE and performed tagmentation afterwards. The PE signal remained after tagmentation (Supplementary Figure 9), allowing us to specifically sort out CD4-positive T cells from the rest of the splenocytes for analysis (we named these “TagSort” libraries). As a control, we first purified CD4 T cells using an antibody-based depletion method (Methods), and subsequently performed scATAC-seq on the purified CD4 T cells (we named these “SortTag” libraries). The data of CD4 T cells generated from these two strategies agree very well (Fig. 4c). The library complexity is comparable with median library sizes of 30,953 and 25,830, respectively (Fig. 4c, top left panel). The binding signals around open chromatin peaks are highly correlated (Pearson $r=0.96$) (Fig. 4c, top right panel). Visual inspection of read pileup profiles around the *Cd4* gene locus from single cell aggregates suggested the data are of good quality (Fig. 4c, bottom panel).

This experiment serves as a proof-of-principle test where staining of a surface marker can be done before Tn5 tagging, and a specific population can be sorted by FACS afterwards for scATAC-seq analysis. It should be noted that we have only tested CD4—an abundant marker in a subpopulation of splenocytes. Other surface markers in different tissues would need to be investigated individually. In addition, the ability to investigate rare cell populations using this approach is limited by the frequency of the rare cell types and the amount of cells that can be tagged upfront.

The spectral clustering was able to distinguish different cell subtypes, such as naive and memory CD8 T cells, naive and regulatory CD4 T cells and CD27+ and CD27– NK cells (Fig. 4b). Previous studies have identified many enhancers that are only accessible in certain cell subtypes, and these are robustly identified in our data. Examples are the *Ilr2b* and *Cd44* loci in memory CD8 T cells²² and *Ikzf2* and *Foxp3* in regulatory T cells²³ (Supplementary Figure 10a and b). Interestingly, our clustering approach successfully identified two subtle subtypes of NK cells (CD27– and CD27+ NK cells), as determined by their open chromatin profiles (Fig. 4b, d). It has been shown that, upon activation, NK cells can express CD83²⁴, a well-known marker for mature dendritic cells²⁵. In mouse spleen, *Cd83* expression was barely detectable in the two NK subpopulations profiled by the ImmGen consortium (Supplementary Figure 10c). However, in our data, the *Cd83* locus exhibited different open chromatin states in the two NK clusters (Fig. 4d). Multiple ATAC-seq peaks were observed around the *Cd83* locus in the CD27+ NK cell cluster but not in the CD27– NK cluster (Fig. 4d). This suggests that *Cd83* is in a transcriptionally permissive state in the CD27+ NK cells, and the CD27+ NK cells have a greater potential for rapidly producing CD83 upon activation. This may partly explain the functional differences between CD27+ and CD27– NK cell states²⁶.

Finally, we investigated whether we could identify the regulatory regions that define each cell cluster. To this end, we trained a logistic regression classifier using the spectral

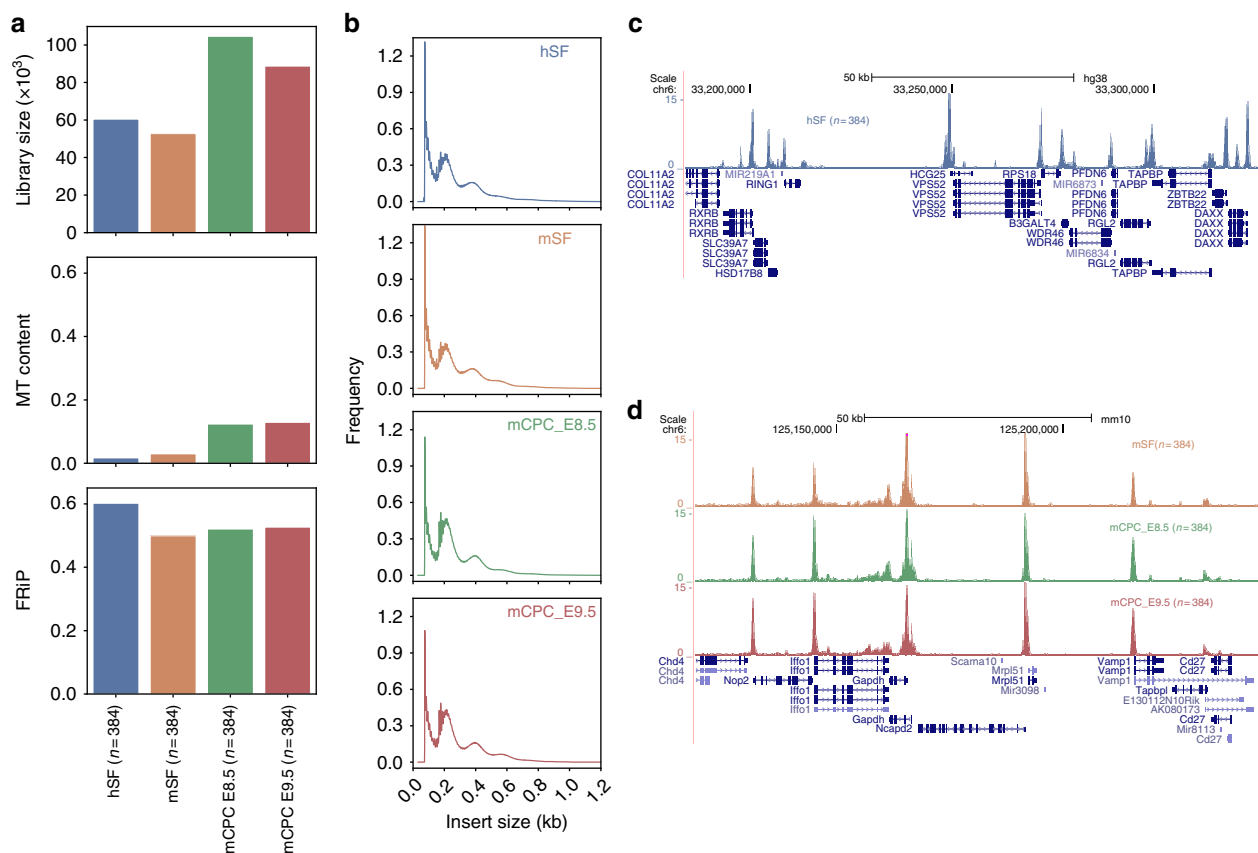


Fig. 2 Plate-based scATAC-seq worked robustly on cryopreserved cells from primary tissues. **a** Comparison of the median library size (estimated by the Picard tool), fraction of mitochondrial DNA (MT content) and fraction of reads in peaks (FRiP) in cryopreserved single cells from four different tissues: human skin fibroblasts (hSF), mouse skin fibroblasts (mSF), mouse cardiac progenitor cells (mCPC) at embryonic day E8.5 and E9.5. **b** Insert size frequencies from the aggregated data of the cells from the four tissues. **c, d** UCSC genome browser tracks displaying the signal around the *RPS18* gene locus from the aggregate of hSFs **c** and around the *Gapdh* gene locus from the aggregate of mSFs, mCPC_E8.5 and mCPC_E9.5 **d**

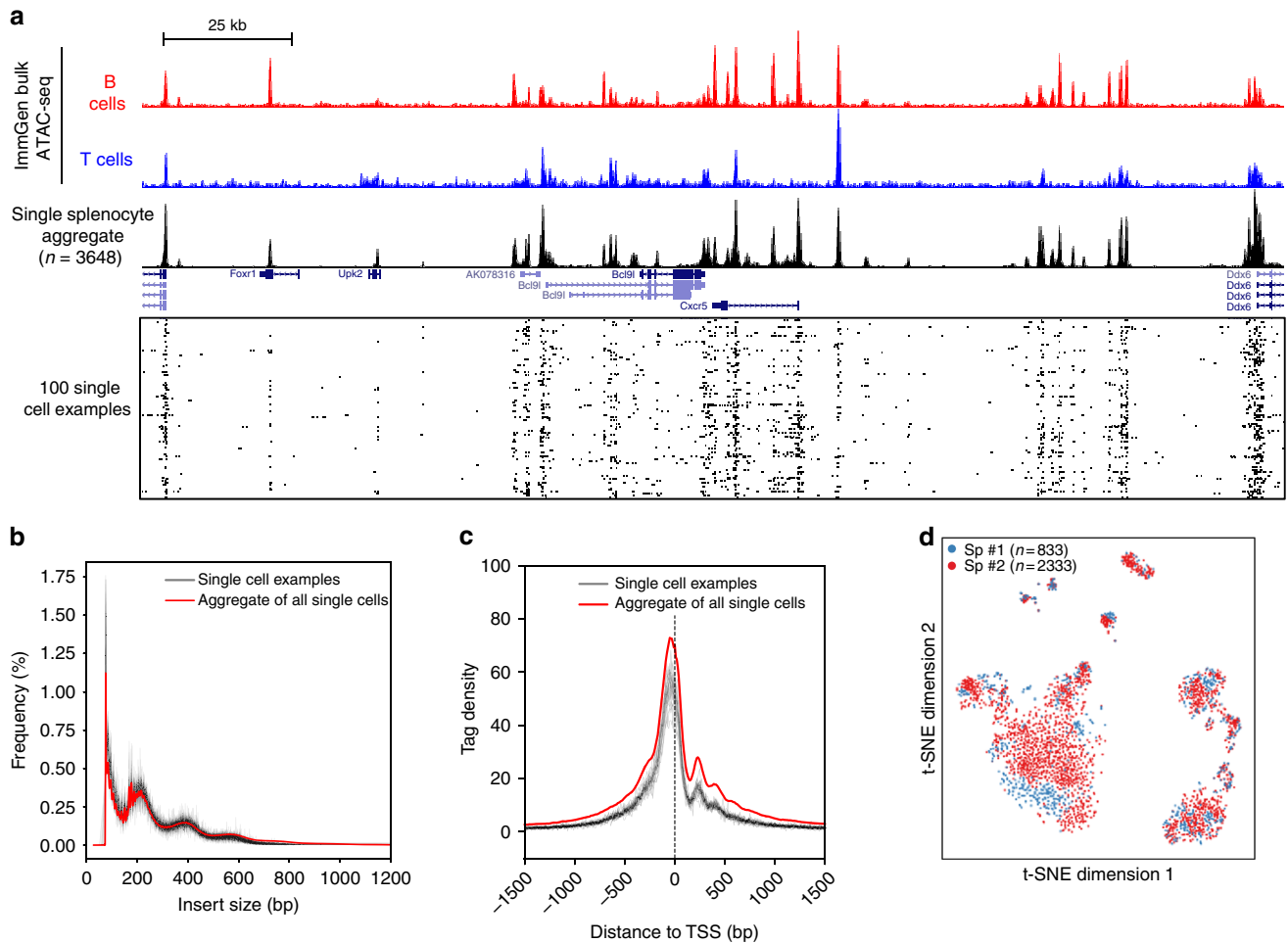


Fig. 3 Plate-based scATAC-seq applied to over 3000 mouse splenocytes. **a** UCSC genome browser tracks displaying the signal around the *Cxcr5* gene locus from the aggregate of all single cells in this study. Bulk ATAC-seq profiles from the ImmGen consortium are also shown. Randomly selected 100 single cell profiles are shown below the aggregated profile. **b**, **c** Insert size frequencies **b** and sequencing read distributions across transcriptional start sites **c** of libraries from the aggregated data (the red line) and individual single cells (grey lines, 24 examples are shown). **d** A two-dimensional projection of the scATAC-seq data using t-SNE. Colours represent two different batches, showing excellent agreement between batches. Sp, spleen

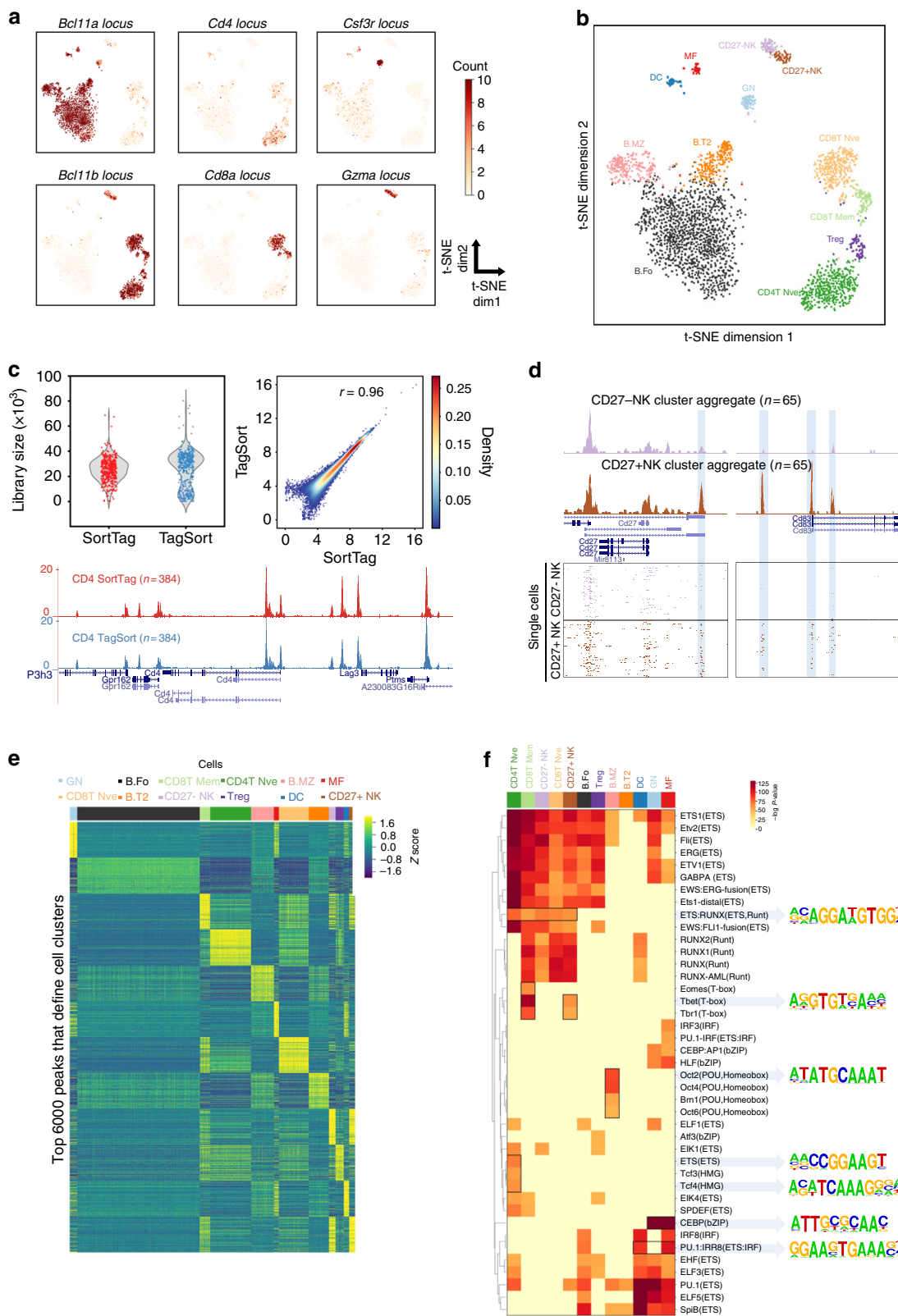
clustering labels and the binarised scATAC-seq count data (Methods). From the classifier, we extracted the top 500 open chromatin peaks (marker peaks) that can distinguish each cell cluster from the others (Fig. 4e and Methods). By looking at genes in the vicinity of the top 50 marker peaks, we recapitulated known markers, such as *Cd4* for the helper T cell cluster (cluster 3), *Cd8a* and *Cd8b1* for the cytotoxic T-cell cluster (cluster 6) and *Cd9* for marginal zone B cell cluster (cluster 4) (Supplementary Figure 11 and Supplementary Data 3). These results are consistent with our correlation-based cell cluster annotation (Fig. 4b).

Whereas the peaks at TSS are useful for cell-type annotation, the majority of the cluster-specific marker peaks are in intronic and distal intergenic regions, in line with the global peak distribution (Supplementary Figure 12). To identify transcription factors that are important for the establishment of these marker peaks, we investigated them in more detail by motif enrichment analysis using HOMER²⁷. The full results of these motif enrichment analyses are included in Supplementary Data 4. As expected, different ETS motifs and ETS-IRF composite motifs were significantly enriched in marker peaks of many clusters (Fig. 4f), consistent with the notion that ETS and IRF transcription factors are important for regulating

immune activities²⁸. Furthermore, we found motifs that were specifically enriched in certain cell clusters (Fig. 4f). Our motif discovery is consistent with previous findings, such as the importance of T-box (e.g. *Tbx21*) motifs in NK²⁹ and CD8T memory cells³⁰ and POU domain (e.g. *Pou2f2*) motifs in marginal zone B cell³¹. This suggests that our scATAC-seq data are able to identify known gene regulation principles in different cell types within a tissue.

Discussion

In recent years, other methods, such as DNase-seq³², MNase-seq³³ and NOMe-seq^{34,35}, have investigated chromatin status at the single cell level. However, due to its simplicity and reliability, ATAC-seq currently remains the most popular technique for chromatin profiling. Several recent studies have demonstrated the power of using scATAC-seq for investigating regulatory principles, e.g. brain development^{4,9}, Mouse sci-ATAC-seq Atlas³⁶ and pseudotime inference³⁷. The combined multi-omics approaches also began to emerge, such as sci-CAR-seq³⁸, scCAT-seq³⁹ and piATAC-seq⁸. Our study added on top of those methods to provide a simple and easy-to-implement scATAC-seq approach that can successfully detect different cell populations, including subtle and rare cell subtypes, from a complex tissue. More



importantly, it is able to reveal key gene regulatory features, such as cell-type-specific open chromatin regions and transcription factor motifs, in an unbiased manner. Future studies can utilise this method to unveil the regulatory characteristics of novel and rare cell populations and the mechanisms behind their transcriptional regulation.

Methods

Ethics statement. The mice were maintained under specific pathogen-free conditions at the Wellcome Trust Genome Campus Research Support Facility (Cambridge, UK). These animal facilities are approved by and registered with the UK Home Office. All procedures were in accordance with the Animals (Scientific Procedures) Act 1986. The protocols were approved by the Animal Welfare and Ethical Review Body of the Wellcome Trust Genome Campus.

Fig. 4 Identification of different cell types and cell-type-specific open chromatin regions and transcription factor motifs. **a** The same t-SNE plot as in Fig. 3d, coloured by the number of counts in the peaks near indicated gene locus. **b** The same t-SNE plot as in Fig. 3d coloured by spectral clustering and cell-type annotation. **c** Comparisons of spleen CD4 T cells scATAC-seq obtained by two strategies. TagSort: cells were stained with anti-CD4-PE, tagged with Tn5 and CD4-PE-positive cells were sorted for scATAC-seq; SortTag: CD4 T cells were purified first and scATAC-seq was performed on the purified cells. Top: comparison of library size and binding signal correlation (pearson $r = 0.96$) around called peaks; bottom: UCSC genome browser tracks of the indicated single cell aggregates around the *Cd4* gene locus. **d** UCSC genome browser tracks around *Cd27* and *Cd83* gene loci, displaying the aggregate (top panel) and single cell (bottom panel) signals of the two NK clusters. ATAC-seq peaks specific to the CD27+ NK cells are highlighted. For visual comparison reason, we randomly choose 65 out of 75 CD27- NK cells. **e** Z-score of normalised read counts in the top 500 peaks that distinguish each cell cluster based on the logistic regression classifier, across each peak (row) in each cell (column). Top 500 marker peaks were picked per cell cluster, so there are $500 \times 12 = 6000$ peaks in the heatmap. Cells are ordered by cluster labels. **f** Heatmap representation of transcription factor motif (rows) enrichments (binomial test p -values) in the top 500 marker peaks in different cell clusters (columns). Some key motifs are enclosed by black rectangles and motif logos are shown to the right. Motif names are taken from the HOMER software suite

Cell isolation. For splenocytes, the spleen from a C57BL/6Jax mouse was mashed by a 2-ml syringe plunger through a 70 μ m cell strainer (Fisher Scientific 10788201) into 30 ml 1X DPBS (Thermo Fisher 14190169) supplied with 2 mM EDTA and 0.5% (w/v) BSA (Sigma A9418). Cells were centrifuged down, supernatant was removed, and the cell pellet was briefly vortexed. 5 ml 1X RBC lysis buffer (Thermo Fisher 00-4300-54) was used to resuspend the cell pellet, and the cell suspension was vortexed again, and left on bench for 5 min to lyse red blood cells. Then 45 ml 1X DPBS was added, and cells were centrifuged down. Volume of 30 ml 1X DPBS was used to resuspend the cell pellet. The cell suspension was passed through a Miltenyi 30 μ m Pre-Separation Filter (Miltenyi 130-041-407), and the cell number was determined using C-chip counting chamber (VWR DHC-N01). All centrifugations were done at 500 \times g, 4 $^{\circ}$ C, 5 min. For human and mouse skin fibroblasts, cells were extracted as previously described¹⁵. For mouse cardiac progenitor cells, cells were extracted as previously described¹⁶. Cells were cryopreserved in 90% FBS and 10% DMSO and stored in liquid nitrogen until experiments.

Plate-based single-cell ATAC-seq (scATAC-seq). A detailed step-by-step protocol can be found in Supplementary Methods. Briefly, 50,000 cells were centrifuged down at 500 \times g, 4 $^{\circ}$ C, 5 min. Cell pellets were resuspended in 50 μ l tagmentation mix (33 mM Tris-acetate, pH 7.8, 66 mM potassium acetate, 10 mM magnesium acetate, 16% dimethylformamide (DMF), 0.01% digitonin and 5 μ l of Tn5 from the Nextera kit from Illumina, Cat. No. FC-121-1030). The tagmentation reaction was done on a thermomixer (Eppendorf 538400039) at 800 rpm, 37 $^{\circ}$ C, 30 min. The reaction was then stopped by adding equal volume (50 μ l) of tagmentation stop buffer (10 mM Tris-HCl, pH 8.0, 20 mM EDTA, pH 8.0) and left on ice for 10 min. A volume of 200 μ l 1X DPBS with 0.5% BSA was added and the nuclei suspension was transferred to a FACs tube. DAPI (Thermo Fisher 62248) was added at a final concentration of 1 μ g/ μ l to stain the nuclei.

Species mixing experiments. A total of 25,000 HEK293T (Human, ATCC[®] CRL-3216[™]) and 25,000 NIH3T3 (Mouse, ATCC[®] CRL-1658[™]) cells were mixed together, and scATAC-seq was performed as described in Supplementary Methods. The obtained sequencing reads were mapped to a concatenated genome of mouse and human by hisat2⁴⁰. One 384-well plate was performed. We first set a technical cutoff where a successful well must contain more than 10,000 total reads and more than 90% of reads are mapped to the concatenated genome. In all, 307 wells were marked as successful. Among the successful wells, we calculated the ratio of reads that mapped to the human genome and the mouse genome. If the ratio is larger than 10, the well is categorised as containing human single cells; if the ratio is less than 0.1, the well is categorised as containing mouse single cells; otherwise, the well is categorised as containing human-mouse doublets.

Plate scATAC-seq on CD4+ T cells (TagSort vs. SortTag). For the “TagSort” strategy, 50,000 splenocytes were stained with anti-Mouse CD4-PE (eBioscience cat no. 12-0043-82) at room temperature for 30 min according to the manufacturer’s instructions. The stained cells were washed with ice-cold 1X PBS twice and pelleted down at 500 \times g, 4 $^{\circ}$ C, 5 min. Experiments were carried out following the procedures described in Supplementary Methods. DAPI and PE double-positive cells were sorted into a 384-well plate for library construction. For the “SortTag” strategy, CD4+ T cells were purified first from mouse splenocytes using the Naive CD4 T-Cell Isolation Kit, Mouse (Miltenyi, cat. no. 130-104-453) following the manufacturer’s instruction without the anti-CD44 depletion step. The purified CD4 T cells were processed according to the procedures described in Supplementary Methods.

scATAC-seq using Fluidigm C1. Experiments were performed as previously described³ using the medium-sized (1862x) Open App chip. We followed the manufacturer’s instructions described in the “ATAC Seq No Stain (Rev C)” from the Fluidigm ScriptHub (<https://www.fluidigm.com/c1openapp/scripthub>), except that we replace the detergent NP-40 in the original protocol with digitonin so that the final concentration of digitonin in the reaction chamber is 0.005%. After

collecting the pre-amplified material from the Fluidigm chip, the libraries were indexed by library PCR for 14 cycles as previously described³.

Costs involved in plate-based and Fluidigm C1 scATAC-seq. For our plate-based scATAC-seq method, most reagents and buffers are available in a standard molecular biology lab. Exceptions are the Tn5 transposase, which can be purchased from Illumina (Cat No. FC-121-1030), and the PCR master mix, which can be purchased from various vendors (we used the 2X NEBNext[®] High-Fidelity 2X PCR Master Mix from NEB). As the Tn5 tagging reaction was performed upfront at the bulk level, the Tn5 cost per cell depends on how many cells are sorted during the sorting. Based on our experience, when 50,000 cells are used at the beginning, two to eight 384-well plates can be sorted. Therefore, the cost of Tn5 is negligible. The major cost per unit for the plate-based scATAC-seq is the PCR master mix used during library amplification. Currently, 10 μ l of PCR master mix are needed per cell in a 20 μ l library amplification reaction, but we have been successfully and consistently generated libraries from half of the volume described in the protocol. For scATAC-seq using the Fluidigm C1, all the aforementioned reagents are needed, and a microfluidic chip is required per 96 cells.

Hands-on time for plate-based vs. C1 scATAC-seq approaches. For our plate-based scATAC-seq method, the most time-consuming part is the lysis plate preparation (mixing lysis buffer and indexing primers). For maximum efficiency, this can be done upfront in bulk, and the lysis plate is stable in -80 $^{\circ}$ C for a long time. Another time/labour-consuming step is the pooling of single cell libraries after PCR using a multi-channel pipette. We provide online advice to perform the whole procedure in minutes. This information is included in the accompanying GitHub page: https://github.com/dbrg77/plate_scATAC-seq. For scATAC-seq using the Fluidigm C1, an extra ~ 4 h of C1 runtime are needed.

qPCR for library amplification. After assembly of the 20 μ l PCR reaction (see Supplementary Methods), a pre-amplification step was performed on a PCR machine (Alpha Cyclor 4, PCRmax) with 72 $^{\circ}$ C 5 min, 98 $^{\circ}$ C 5 min, 8 cycles of [98 $^{\circ}$ C 10 s, 63 $^{\circ}$ C 30 s, 72 $^{\circ}$ C 20 s]. Of the product, 19 μ l of pre-amplified library was transferred to a 96-well-qPCR plate, 1 μ l 20X EvaGreen (Biotium #31000) was added, and qPCR was performed on an ABI StepOnePlus system with the following cycle conditions: 98 $^{\circ}$ C 1 min, 20 cycles of [98 $^{\circ}$ C 10 s, 63 $^{\circ}$ C 30 s, 72 $^{\circ}$ C 20 s]. Data were acquired at 72 $^{\circ}$ C. We qualitatively chose the cycle number to where the fluorescence signals just about to start going up (Supplementary Figure 1b). In this study, a total of 18 cycles were used to amplify the libraries.

Sequencing data processing. All sequencing data were processed using a pipeline written in snakemake⁴¹. The software/packages and the exact flags used in this study can be found in the ‘Snakefile’ provided in the GitHub repository https://github.com/dbrg77/plate_scATAC-seq. Briefly, reads were trimmed with cutadapt⁴² to remove the Nextera sequence at the 3’-end of short inserts. The trimmed reads were mapped to the reference mouse genome (UCSC mm10) using hisat2⁴⁰. Reads with mapping quality less than 30 were removed by samtools⁴³ (-q 30 flag) and deduplicated using the MarkDuplicates function of the Picard tool (<http://broadinstitute.github.io/picard>). All reads from single cells were merged together using samtools, and the merged BAM file was deduplicated again. Peak calling was performed on the merged and deduplicated BAM file by MACS2¹⁸. For bulk ATAC-seq and single cell aggregate coverage visualisation, bedGraph files generated from MACS2 callpeak were converted to bigWig files and visualised via UCSC genome browser. For individual single cell ATAC-seq visualisation, aligned reads from individual cells were converted to bigBed files. A count matrix over the union of peaks was generated by counting the number of reads from individual cells that overlap the union peaks using coverageBed from the bedTools suite⁴⁴.

Public ATAC-seq data processing. FASTQ files were all downloaded from the European Nucleotide Archive (ENA). The ImmGen bulk ATAC-seq data (study accession PRJNA392905) and the scATAC-seq data using Fluidigm C1 (study accessions PRJNA274006 and PRJNA299657) were processed in the same way as described in this study. The ‘Snakefile’ used to process the data can be found at the same GitHub repository.

Bioinformatics analysis. Codes used to carry out all the analyses were provided as Jupyter Notebook files, which can be found in the same GitHub repository. Briefly, downsampling was performed by randomly selecting a fraction of reads from the original FASTQ files using seqtk (<https://github.com/lh3/seqtk>), and the same pipeline was run on the sub-sampled FASTQ files. For binarising the scATAC-seq data, peak calling was performed on reads merged from all cells, and we labelled the peak ‘1’ (open) if there was at least one read overlapping the peak, and ‘0’ (closed) otherwise. Latent semantic indexing analysis was performed by first normalising the binarized count matrix by term frequency inverse document frequency (TF-IDF) and then performing a Singular-Value Decomposition (SVD) on the normalised count matrix. Only the 2nd–50th-dimensions after the SVD were passed to t-SNE for visualisation. To compare with ImmGen bulk ATAC-seq data, a reference peak set was created by taking the union of peaks from the peak calling results of aggregated scATAC-seq (this study) and different samples of ImmGen bulk ATAC-seq using mergeBed from the bedTools suite⁴⁴. All comparisons were done based on this reference peak set. The annotatePeaks.pl from HOMER²⁷ was used to assign genes to peaks. Latent semantic indexing, spectral clustering and logistic regression were carried out using Scikit-learn⁴⁵.

Code availability. The code used for the analysis is available on the Github repository https://github.com/dbrg77/plate_scATAC-seq.

Data and availability

The sequencing data have been deposited at ArrayExpress, accession number E-MTAB-6714. The UCSC genome browser tracks containing both the ImmGen bulk ATAC-seq and scATAC-seq from this study can be viewed via this link: http://genome-euro.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=dbrg77&hgS_otherUserSessionName=mSpleen_scATAC_cluster.

Received: 10 October 2018 Accepted: 13 November 2018

Published online: 17 December 2018

References

- 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).
- Cusanovich, D. A. et al. Epigenetics. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
- Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
- Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
- Cusanovich, D. A. et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
- Ryan Corces, M. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
- Mezger, A. et al. High-throughput chromatin accessibility profiling at single-cell resolution. *Nat. Commun.* **9**, 3647 (2018).
- Chen, X. et al. Joint single-cell DNA accessibility and protein epitope profiling reveals environmental regulation of epigenomic heterogeneity. Preprint at <https://doi.org/10.1101/310359> (2018).
- 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).
- Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
- Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
- Goryshin, I. Y., Jendrisak, J., Hoffman, L. M., Meis, R. & Reznikoff, W. S. Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes. *Nat. Biotechnol.* **18**, 97–100 (2000).
- Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
- Goldenberger, D., Perschil, I., Ritzler, M. & Altwegg, M. A simple ‘universal’ DNA extraction procedure using SDS and proteinase K is compatible with direct PCR amplification. *PCR Methods Appl.* **4**, 368–370 (1995).
- Hagai, T. et al. Gene expression variability across cells and species shapes innate immunity. Preprint at <https://doi.org/10.1101/137992> (2017).
- Jia, G. et al. Single cell RNA-seq and ATAC-seq indicate critical roles of Isl1 and Nkx2-5 for cardiac progenitor cell transition states and lineage settlement. Preprint at <https://doi.org/10.1101/210930> (2017).
- Heng, T. S. P. & Painter, M. W., Immunological Genome Project Consortium. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
- Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
- van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
- Cesta, M. F. Normal structure, function, and histology of the spleen. *Toxicol. Pathol.* **34**, 455–465 (2006).
- von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
- Bevington, S. L. et al. Inducible chromatin priming is associated with the establishment of immunological memory in T cells. *EMBO J.* **35**, 515–535 (2016).
- Samstein, R. M. et al. Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151**, 153–166 (2012).
- Mailliard, R. B. et al. IL-18-induced CD83+CCR7+ NK helper cells. *J. Exp. Med.* **202**, 941–953 (2005).
- Zhou, L. J. & Tedder, T. F. Human blood dendritic cells selectively express CD83, a member of the immunoglobulin superfamily. *J. Immunol.* **154**, 3821–3835 (1995).
- Hayakawa, Y. & Smyth, M. J. CD27 dissects mature NK cells into two subsets with distinct responsiveness and migratory capacity. *J. Immunol.* **176**, 1517–1524 (2006).
- Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- Smale, S. T. Transcriptional regulation in the immune system: a status report. *Trends Immunol.* **35**, 190–194 (2014).
- Simonetta, F., Pradier, A. & Roosnek, E. T-bet and eomesodermin in NK cell development, maturation, and function. *Front. Immunol.* **7**, 241 (2016).
- Intlekofer, A. M. et al. Effector and memory CD8+ T cell fate coupled by T-bet and eomesodermin. *Nat. Immunol.* **6**, 1236–1244 (2005).
- Martin, F. & Kearney, J. F. Marginal-zone B cells. *Nat. Rev. Immunol.* **2**, 323–335 (2002).
- Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142–146 (2015).
- Lai, B. et al. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* **562**, 281–285 (2018).
- Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *elife* **6**, e23203 (2017).
- Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
- Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324 (2018).
- Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
- Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
- Liu, L. et al. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Preprint at <https://doi.org/10.1101/316208> (2018).
- Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
- Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
- Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

We thank Jong-Eun Park, Johan Henriksson, Tzachi Hagai, Tomas Gomez, Kerstin Meyer, Roser Vento, Lira Mamanova and all others from the Teichmann group for the

inspiring discussion of the method, the critical reading of the manuscript and the computational help. We also thank Natalia Kunowska, Qianxin Wu and Andrew Knights for the helpful discussion related to the Tn5 transposase. We thank Bee Ling Ng, Chris Hall, Jennie Graham and Sam Thompson for the excellent support of FACS. We thank the DNA pipeline from the Wellcome Sanger Institute for the Illumina sequencing support. We thank Guangshuai Jia and Jens Preussner from Thomas Braun's lab for sharing the mouse cardiac progenitor cells. We thank Aik Ooi for the initial help with the experimental setup. X.C. is funded by the FET-OPEN grant MRG-GRAMMAR 664918, K.N.N. by the Wellcome Trust Strategic Award "Single cell genomics of mouse gastrulation" and S.A.T. by the European Research Council grant ThDEFINE. Wellcome trust core facilities are supported by grant WT206194.

Author contributions

X.C., K.N.N. and S.A.T. conceived the project. X.C. designed the protocol. X.C., R.J.M. and K.N.N. performed the experiments. X.C. carried out the computational analysis. S.A.T. supervised the entire project. All authors contributed to the writing.

Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-018-07771-0>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

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



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

© The Author(s) 2018