Genome Biology

METHOD

Open Access

SeATAC: a tool for exploring the chromatin landscape and the role of pioneer factors

Wuming Gong[1,2*], Nikita Dsouza[1] and Daniel J. Garry[1,2,3,4*]

*Correspondence: gongx030@umn.edu; garry@umn.edu

[1] Cardiovascular Division, Department of Medicine, University of Minnesota, Minneapolis, MN 55455, USA
[2] Lillehei Heart Institute, University of Minnesota, 2231 6Th St SE, Minneapolis, MN 55455, USA
[3] Stem Cell Institute, University of Minnesota, Minneapolis, MN 55455, USA
[4] Paul and Sheila Wellstone Muscular Dystrophy Center, University of Minnesota, Minneapolis, MN 55455, USA

Abstract

Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq) reveals chromatin accessibility across the genome. Currently, no method specifically detects differential chromatin accessibility. Here, SeATAC uses a conditional variational autoencoder model to learn the latent representation of ATAC-seq V-plots and outperforms MACS2 and NucleoATAC on six separate tasks. Applying SeATAC to several pioneer factor-induced differentiation or reprogramming ATAC-seq datasets suggests that induction of these factors not only relaxes the closed chromatin but also decreases chromatin accessibility of 20% to 30% of their target sites. SeATAC is a novel tool to accurately reveal genomic regions with differential chromatin accessibility from ATAC-seq data.

Keywords: Etv2, Oct4, Sox2, Klf4, Ascl1, Nucleosomal DNA, Pioneer factors

Background

Eukaryotic genomes are packed into nucleoprotein called chromatin whose basic unit is the nucleosome, which comprises a histone octamer wrapped around 147 base pairs of DNA [1]. Nucleosomes are arranged into regularly spaced arrays, separated by unwrapped linker DNA whose length varies among species and cell types [2]. The dense nucleosome regions (nucleosome occupied regions, NOR) are tightly packed, whereas the loose nucleosome regions (nucleosome free regions, NFR) are more accessible to transcription factors. It is known that precise location of a nucleosome relative to transcriptional target sites can significantly influence factor binding [3–7]. Thus, the chromatin accessibility plays a critical role in regulating gene expression pattern.

High-throughput sequencing techniques such as MNase-seq [8, 9], chemical mapping [10], DNase-seq [11], FAIRE-seq [12], and ATAC-seq [13] have been developed to assess genome-wide chromatin structure. MNase-seq uses an endo-exonuclease that degrades the accessible linker DNA between nucleosomes and reveals the position of nucleosomes by sequencing the protected DNAs. The chemical cleavage method introduces a cysteine substitution at serine 47 in histone H4 (H4S47C) to localize free radical mediated cleavage of nucleosome DNA, followed by performing a copper ion-mediated Fenton

reaction to cleave nucleosomal DNAs. The cleaved DNA fragments are then sequenced to estimate the position of the center of the nucleosome. DNase-seq digests with DNase I endonuclease and the resulting DNA fragments correspond to open chromatin region. FAIRE-seq uses formaldehyde and phenol–chloroform extraction separation to isolate nucleosome-depleted DNA from chromatin. Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), which utilizes Tn5 transposases to digest accessible genomic DNA, is an efficient and precise method for revealing chromatin accessibility across the genome. Compared with other techniques, ATAC-seq requires less input materials and sample processing time [13, 14] and thus becomes a widely adopted tool for profiling chromatin accessibility of both bulk samples and single cells [15].

The *fragment size profile* of ATAC-seq paired-end reads can be partitioned into reads generated from putative NFR and NOR regions of DNAs, respectively [13]. The reads from the NOR region have clear periodicity of approximately 150 to 200 bp and produced detailed information on nucleosome position and degree of chromatin compaction [13]. This unique feature of ATAC-seq reads have been utilized to infer the nucleosome positions using NucleoATAC [16] and deNOPA [17], which have demonstrated improved performance compared to generic nucleosome calling tools such as DANPOS [18] and NPS [19]. However, to date, there is no published method for the detection of differential chromatin accessibility specifically for ATAC-seq data. Currently, MACS2 [20, 21], which was originally designed for ChIP-seq data, remains the gold standard for analyzing ATAC-seq data [22] and does not consider ATAC-seq-specific properties.

In this study, we engineered a tool, named SeATAC, to estimate the genomic regions with statistically differential chromatin accessibility from multiple ATAC-seq data. Using SeATAC, each genomic region is represented as a V-plot, a dot-plot showing how sequencing reads with different fragment sizes distribute surrounding one or a set of genomic region(s) [23]. The V-plot based analysis has been used to study nucleosome dynamics flanking the transcription factor (TF) binding sites [23, 24], nucleosome phasing near pioneer factors during reprogramming [25], clustering the nucleosome profiles near promoters [26], and examining the distance between nearby nucleosomes [27, 28]. However, the V-plot was derived from and visualized for a set of genomic regions due to the noisy and sparse nature of the sequence reads on genomic regions. The difference of V-plots on individual genomic regions between multiple ATAC-seq datasets have never been evaluated before. For SeATAC, we used a conditional variational autoencoder (CVAE) model to learn the latent representation of the ATAC-seq V-plot [29–31]. With the probabilistic representation of the data, we developed a Bayesian method to evaluate the statistical difference between multiple V-plots. We demonstrated that SeATAC had significantly better performance on six separate tasks compared to MACS2 and/or NucleoATAC on both synthetic and real ATAC-seq datasets. SeATAC is available at https://github.com/gongx030/seatac as an R package.

Results

The SeATAC model

The SeATAC model uses a V-plot with a width of 640-bp genomic region and a height of 640 bp of fragment sizes that covers nucleosome free reads (< 100 bp), mono-nucleosome

reads (between 180 and 247 bp), di-nucleosome reads (between 315 and 473 bp), and tri-nucleosomes (between 558 and 615 bp) [13]. The four groups of ATAC-seq reads represent the majority of total ATAC-seq reads (> 95%) and have been successfully used to segment the genomic structure [13, 32]. To reduce the impact of noise, an array of 5×10 pixels were aggregated together and became a single larger pixel, resulting in an image composed of 128×64 pixels. We named the bins along the genomic region dimension and fragment size dimension as *genomic bins* and *fragment size bins*, respectively. The aggregated reads along the genomic bins were then normalized to a vector that sum to one (Fig. 1a).

We modeled the V-plot $\boldsymbol{x}_{ni}$ of each genomic region i in each sample n as a probabilistic distribution $p(\boldsymbol{x}_{ni}|\boldsymbol{z}_{ni}, s_n)$ conditioned on the sample indicator s_n of each sample, as well as an unobserved latent variable $\boldsymbol{z}_{ni}$ (Fig. 1b). The sample indicator s_n represents the nuisance variation due to the sample-specific fragment size profile. The latent variable $\boldsymbol{z}_{ni}$ is a K dimensional vector of Gaussians representing the remaining variation with respect to the underlying V-plot ($K = 5$). In SeATAC, a neural network serves as a decoder to map the latent variables z_{ni} and sample indicator s_n to an estimated output V-plot. We expected that latent variables provide batch-corrected representations of the V-plot for the differential analysis. We derived an approximation of the posterior distribution of the latent variable $q(\boldsymbol{z}_{ni}|\boldsymbol{x}_{ni}, s_n)$ by training another encoder neural network using variational inference and a scalable stochastic optimization procedure [29, 30]. The variational distribution $q(\boldsymbol{z}_{ni}|\boldsymbol{x}_{ni}, s_n)$ is chosen to be Gaussian with a diagonal covariance matrix, where the mean and covariance are estimated by an encoder neural network applied to $(\boldsymbol{x}_{ni}, s_n)$. The variational evidence lower bound (ELBO) is

$$\log p(x|s) \geq \mathbb{E}_{q(z|x,s)} \log p(x|z,s) - D_{KL}[q(z|x,s) \parallel p(z)]$$

A standard multivariable normal prior $p(\boldsymbol{z}_{ni})$ is used in SeATAC because it can be reparametrized into a way that allows backpropagation to flow through the deterministic nodes [29]. To optimize this lower bound, we used the reparameterization trick to compute low-variance Monte Carlo estimates of the expectations' gradients. Throughout the study, we used Adam optimizer (learning rate = 0.01) with a cosine learning rate schedular with warmup.

SeATAC corrects batch effects in ATAC-seq data

Although the fragment size profile (the fragment size density plot) provided similar fragment length estimation regarding NFR and nucleosomes (mono-nucleosomes, di-nucleosomes, tri-nucleosomes, etc.) [13], the exact pattern differed across ATAC-seq datasets, resulting in different fragment size ranges and density for NFR and nucleosome reads. We assumed that the majority of the batch effects in the ATAC-seq were due to the difference of the fragment size profile [13]. In the SeATAC model, an embedding layer first maps the sample indicator s_n to the fragment size vector g_n and combines with the input V-plot to produce a modified V-plot. Then, convolutional neural networks (CNN) map the modified V-plot to the latent variables. Once the model was optimized, SeATAC used a constant sample indicator s_0 to replace the sample specific indicator s_n to generate a batch-free estimated V-plot.

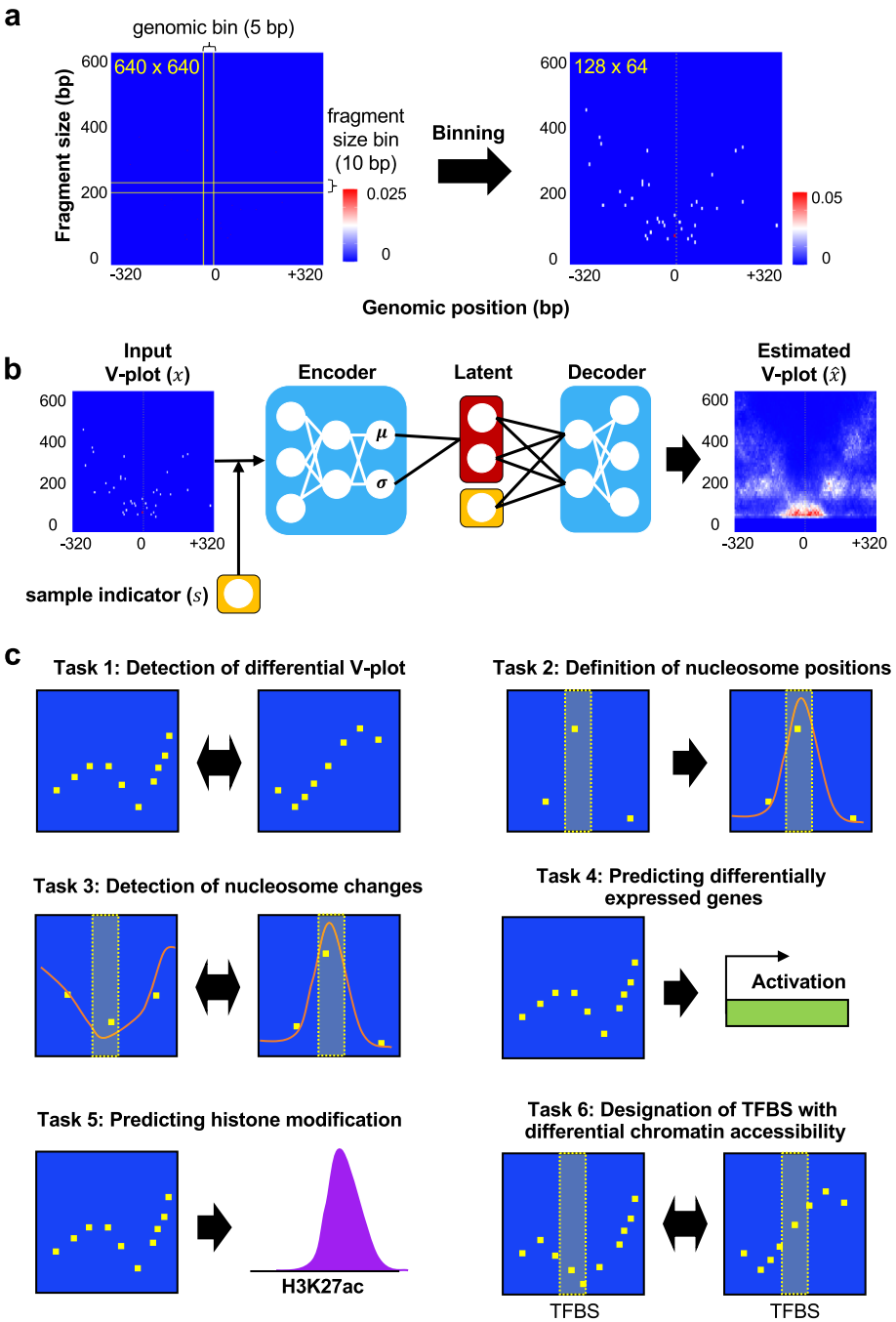

Fig. 1 The SeATAC model and tasks for performance evaluation. **a** A full V-plot has a width of 640-bp genomic region and a height of 640 bp of fragment sizes (left panel). An array of 5 × 10 pixels is aggregated together and become a single larger pixel, resulting in a 128 × 64 pixels image (right panel). The heatmap color indicates the normalized read density. **b** SeATAC models the ATAC-seq V-plot using a conditional variational autoencoder (CVAE) framework. **c** Six separate tasks for evaluating the performance of detecting chromatin accessibility changes. MACS2 was excluded from tasks #2 and #3

We applied SeATAC to a human hematopoietic differentiation dataset with 13 samples [33], and each sample showed a distinct fragment size profile (Additional File 1: Fig. S1a). We randomly sampled 2000 640-bp genomic regions, generated batch-free V-plot,

and computed the aggregated fragment size profile by averaging along each fragment size bin. The corrected fragment size profile became consistent across 13 samples, suggesting that SeATAC was able to successfully correct the batch effects due to difference in fragment size profile, allowing SeATAC, in an unbiased fashion, to compare multiple ATAC-seq samples.

Tasks for performance evaluation

We designed six separate tasks for evaluating the performance of detecting chromatin accessibility changes including: (1) the detection of differential V-plots, (2) the recovery of nucleosome positions from sparse ATAC-seq data, (3) calling differential nucleosomes, (4) predicting differentially expressed genes from ATAC-seq signals near promoters, (5) predicting histone modifications, and (6) the designation of transcriptional factor binding sites (TFBS) following increased chromatin accessibility (Fig. 1c). The task #1 was to determine whether or not the V-plot for a genomic region was different between multiple ATAC-seq samples. Tasks #2 and #3 asked the methods to recover (task #2) and to compare (task #3) nucleosome positions. We excluded MACS2 from these two tasks since MACS2 was not capable of calling the nucleosomes directly. Tasks #1–#3 were evaluated on the datasets down-sampled from a full ATAC-seq dataset. Task #4 was evaluated on paired RNA-seq/ATAC-seq datasets. Task #5 was evaluated on paired ATAC-seq/histone ChIP-seq datasets. Both tasks, #4 and #5, were designed to evaluate how accurate the local ATAC-seq information captured by SeATAC, NucleoATAC, or MACS2 was able to predict the biologically relevant readout such as differentially expressed genes or local histone modification. Task #6 focused on the detection of individual TFBS with differential chromatin accessibility and was evaluated using several ATAC-seq datasets of TF-induced reprogramming.

SeATAC detects differential V-plot

To define a benchmark dataset for testing a differential V-plot, we generated two separate down-sampled datasets (dataset #1 and dataset #2) that included 10% of sequencing reads of a full ATAC-seq dataset (GM12878) by using different random seeds, separately. Then every read in dataset #2 was shifted to 3′ direction by a pre-specified distance (e.g., 100 bp) to generate a new dataset #3. Thus, dataset #1 and dataset #2 should have the identical V-plot for any genomic regions, while dataset #1 and dataset #3 should have different V-plot because the shift size is smaller than the length of nucleosome DNAs and the linker DNAs (Fig. 2a). We used SeATAC, MACS2, and NucleoATAC to compare dataset #1 vs. dataset #2 and dataset #1 vs. dataset #3 and evaluate the performance of calling differential V-plots by computing the receiver operating characteristic (ROC) curves, respectively (Fig. 2b). The SeATAC p-values (p^{SeATAC}), maximum difference of MACS2 pileup (and the maximum difference of NucleoATAC signal were used to rank the differential V-plots (see the " Methods" section). We evaluated the performance on different shift size for dataset #3 (10 to 100 bp with a step size of 10 bp). With a shift size of 50 bp, the average area under the ROC curve (AUC) of SeATAC, MACS2, and NucleoATAC were 0.994, 0.538, and 0.536, respectively (Fig. 2c). The performance of SeATAC was not significantly impacted by the shift size (Additional File 1: Fig. S2a). Moreover, we found that SeATAC had significantly better performance on detecting differential

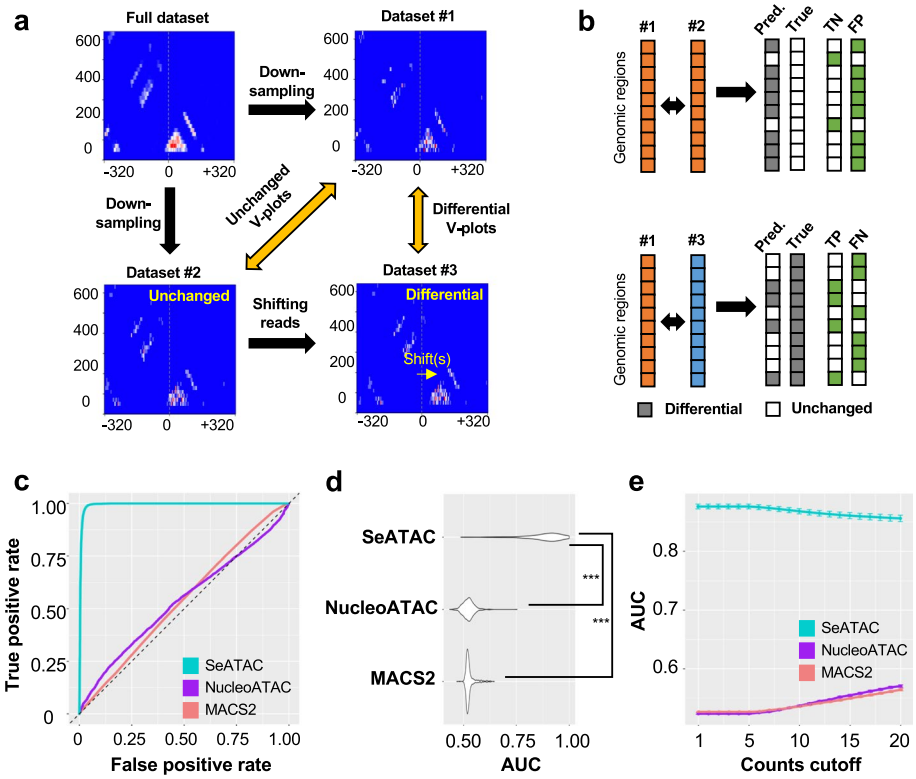

Fig. 2 SeATAC detects differential V-plots. **a** A full ATAC-seq dataset is down-sampled to two separate datasets (dataset #1 and dataset #2) that includes 10% of the sequencing reads. Every read in dataset #2 is shifted to the 3′ direction by a pre-specified distance (e.g., 100 bp) to generate a new dataset #3. The dataset #1 and dataset #2 have the identical V-plot for any genomic regions, while dataset #1 and dataset #3 have different V-plots. **b** Different tools are used to compare dataset #1 vs. dataset #2 and dataset #1 vs. dataset #3 to detect differential V-plots. The true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions are computed. The receiver operating characteristic (ROC) curve is used to evaluate the performance of different tools. **c** The ROC curves for SeATAC, NucleoATAC, and MACS2 with a shift size of 50 bp. **d** The violin plot shows the AUC (area under ROC) of SeATAC, NucleoATAC, and MACS2 on 523 ATAC-seq samples from 20 studies. ***Wilcoxon rank sum test *p*-value < 0.001. **e** The AUC of SeATAC, NucleoATAC, and MACS2 at different read counts cutoff from 1 to 20 (the minimum reads in a V-plot)

V-plots than NucleoATAC and MACS2 on 523 ATAC-seq samples from 20 published studies (Fig. 2d) [25, 34–53]. The read counts of the V-plot had no significant impact on SeATAC performance, suggesting robust performance of SeATAC on detecting differential V-plot (Fig. 2e).

SeATAC recovers nucleosome positions from sparse ATAC-seq data

To evaluate how well SeATAC detected nucleosome positions from sparse ATAC-seq data, we first defined the NFR or NOR positions on a full ATAC-seq dataset (GM12878). A genomic locus was considered as a NOR center if the NucleoATAC signal at this locus was greater than 0.5 and was also greater than any other positions in the flanking 200-bp region. A genomic locus was considered as a NFR center if the NucleoATAC signal at this locus was smaller than 0.01 and was also smaller than any other position in the flanking 200-bp region. There were 9965 and 316,075 NOR and NFR centers in the full ATAC-seq data. We randomly sampled ~ 5000 NOR and NFR centers to evaluate

the performance of nucleosome calling. We down-sampled the full ATAC-seq dataset to 0.1%, 1%, and 10% of the full datasets and used SeATAC and NucleoATAC to estimate the nucleosome signals at each NOR and NFR centers (see the " Methods" section). SeATAC demonstrated overall superior performance on calling nucleosomes from sparse ATAC-seq data with AUR of 0.583, 0.606, and 0.653 for 0.1%, 1%, and 10% down-sampled datasets, respectively, while the AUC for NucleoATAC were 0.503, 0.491, and 0.591, respectively (Fig. 3a). Among ~ 5000 NORs, we identified 2042, 3453, and 22 regions that were called by both SeATAC and NucleoATAC, SeATAC only, and NucleoATAC only as nucleosomes, respectively ($Nuc^{\mathrm{SeATAC}} > 0.5$ or $Nuc^{\mathrm{NucATAC}} > 0.2$). The center of these genomic regions that were called as nucleosomes by SeATAC only showed enriched nucleosome signals supported by both NucleoATAC estimation on the full dataset and an MNase-seq dataset on GM12878 [54] (Fig. 3b). The additional systematic analysis over 523 ATAC-seq samples further supported the notion that

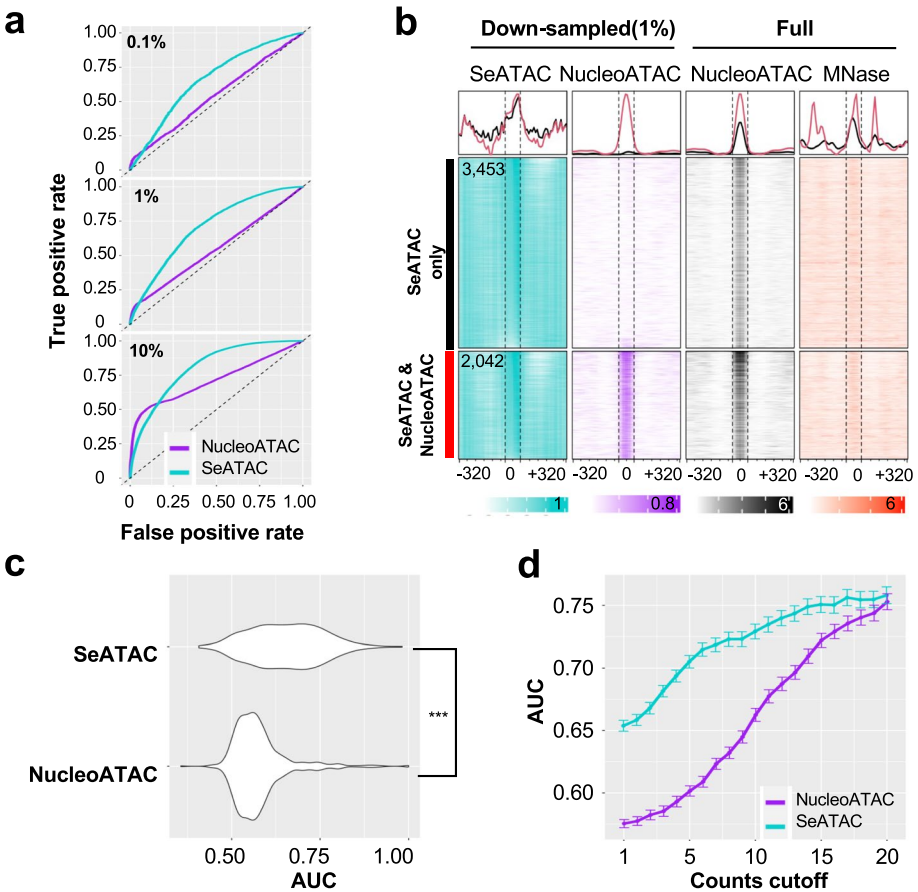

Fig. 3 SeATAC recovers nucleosome positions from sparse ATAC-seq. **a** The ROC curve for recovering nucleosome positions from ATAC-seq with 0.1%, 1%, and 10% of the sequencing reads randomly sampled from the full dataset (GM12878). **b** The heatmaps shows the nucleosome density estimated by SeATAC (blue) and NucleoATAC (purple) on a 1% down-sampled dataset. There are 2042 and 3453 regions (640 bp) identified by both SeATAC/NucleoATAC and by SeATAC only as nucleosomes. The NucleoATAC signal on the full dataset (black) and a MNase-seq dataset on GM12878 (red) for these regions are also shown. **c** The violin plot shows the AUC (area under ROC) of SeATAC and NucleoATAC on 523 ATAC-seq samples from 20 studies. ***Wilcoxon rank sum test *p*-value < 0.001. **d** The AUC of SeATAC and NucleoATAC at different read counts cutoff from 1 to 20 (the minimum reads in a V-plot)

SeATAC had better performance on estimating nucleosomes from sparse ATAC-seq data (Fig. 3c). The systematic analysis also showed that the performance of SeATAC was significantly positively correlated with total number of reads, proper pair rate, and negatively correlated with mitochondria rate, unmapped rate, has unmapped mate rate (*t*-test *p*-value < 0.05) (Fig. 3d and Additional File 1: Fig. S3).

SeATAC detects nucleosome changes

By using the ground truth NOR and NFR centers on full GM12878 ATAC-seq dataset, we could also evaluate how capable SeATAC was regarding the call of the nucleosome change from NFR to NOR. We randomly sampled 5000 NFR/NOR pairs and applied SeATAC and NucleoATAC to evaluate the nucleosome changes at the center of each NFR/NOR pairs on a down-sampled ATAC-seq dataset with 10% of sequencing reads. The nucleosome changes were ranked by SeATAC's *differential central nucleosome score* (δ^{SeATAC}) and NucleoATAC's *differential central signal* (δ^{NucATAC}), respectively (see the " Methods" section). SeATAC demonstrated superior performance on calling nucleosome changes than NucleoATAC with an AUC of 0.904 vs. 0.827 (Fig. 4a). Among ~ 5 k NFR/NOR pairs, SeATAC was able to successfully identify more than 72.9% of genuine NFR/NOR changes compared to NucleoATAC (1278 vs. 739), and these changes were supported by the NucleoATAC signals on the full dataset and an MNase-seq dataset [54] (Fig. 4b, c). Similar to the previous two tasks, we extended the analyses to include 523 ATAC-seq samples and confirmed that SeATAC could significantly more accurately detect the nucleosome changes between ATAC-seq samples (Fig. 4d, e).

SeATAC predicts differentially expressed genes from ATAC-seq signals near promoters

Previous studies have shown that the changes of DNA accessibility over the promoter regions were weakly associated the gene expression changes [55, 56]. However, these studies used the simple ATAC-seq peaks or density as the features for correlation with the RNA-seq levels. We asked whether including more sophisticated ATAC-seq features such as V-plot would improve the accuracy of predicting gene expression changes. We compiled a list of 17 paired RNA-seq/ATAC-seq datasets on temporal reprogramming or cellular differentiation (Table 1). In each dataset, we first compared the RNA-seq of data from any two conditions (e.g., treatment vs. control) and determined a list of significantly up- and down-regulated genes (DESeq2 *q*-value < 0.05 with > two fold change) [57]. Then we used SeATAC, NucleoATAC, and MACS2 to generate *features* from the ATAC-seq signals over the promoter regions to predict whether the underlying genes were up- or down-regulated (Additional File 1: Fig. S4a). We found that the feature dimensions had moderate impact on the prediction performance of MACS2 and increasing the latent dimension from 10 to 20 significantly improved the performance of SeATAC (Wilcoxson rank-sum test *p*-value < 0.05), while extending the promoter regions (e.g., from 640 to 2560 bp) did not improve the performance of SeATAC (Additional File 1: Fig. S4b) [25, 58]. Overall, we found that SeATAC had the best performance of predicting differentially expressed (DE) genes in 16 out of 17 datasets, suggesting that the V-plot representations produced by SeATAC better captured the relationship between DNA accessibility at the promoter and the gene expression changes (Additional File 1: Fig. S4c).

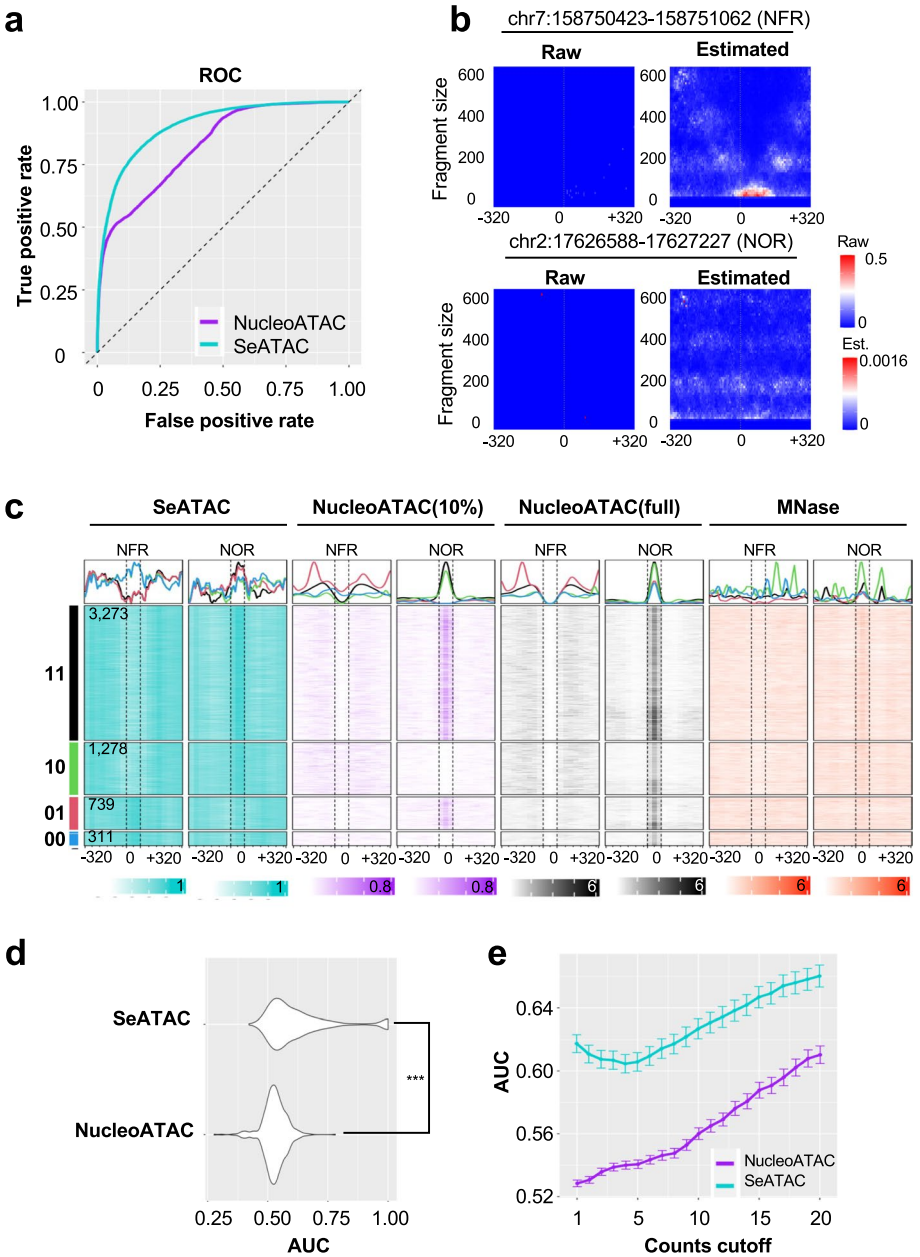

Fig. 4 SeATAC detects nucleosome changes. **a** The ROC curve for detecting nucleosome changes from ATAC-seq with 10% of the sequencing reads from the full dataset (GM12878). **b** The raw and estimated V-plot of a NFR (chr1:113,162,059–113,162,698) and a NOR (chr2:226,653,061–226,653,700) region are shown. The heatmap color indicates the normalized read density. **c** The heatmaps show the nucleosome density of ~ 5000 sampled NOR and NFR regions estimated by SeATAC (blue) and NucleoATAC (purple) on a 10% down-sampled dataset. There are 3276, 1278, 739, and 311 regions that are identified as a change from NFR to NOR (with decreased chromatin accessibility) by both SeATAC and NucleoATAC (11), by SeATAC only (10), by NucleoATAC only (01), and by neither of them (00), respectively. The NucleoATAC signal on the full dataset (black) and a MNase-seq dataset on GM12878 (red) for these regions are also shown. **d** The violin plot shows the AUC (area under ROC) of SeATAC and NucleoATAC on 523 ATAC-seq samples from 20 studies. ***Wilcoxon rank sum test *p*-value < 0.001. **e** The AUC of SeATAC and NucleoATAC at different read counts cutoff from 1 to 20 (the minimum reads in a V-plot)

Table 1 24 ATAC-seq or ATAC-seq/RNA-seq paired datasets used in this study

Name	Tasks	Description	RNA-seq	ATAC-seq	# cond	Refs
Ascl1	4,6	Ascl-1-induced mouse neural reprogramming	GSE43916	GSE101397	4	[25, 59–61]
Duren	1,2,3,4	Retinoic acid (RA)-induced mESC differentiation	GSE136312	GSE136312	6	[35, 62]
Perrin	1,2,3,4	Human adipocyte differentiation	GSE178795	GSE178794	5	[51, 63, 64]
Ramirez (macrophage, monocyte derived, monocyte)	1,2,3,4	Human myeloid differentiation	GSE79044	GSE79019	30	[46, 65, 66]
Liu	1,2,3,4	Human cardiac differentiation	GSE85331	GSE85330	16	[44, 67, 68]
Markov	1,2,3,4	Human early iPSC reprogramming	GSE121052	GSE120992	10	[45, 69, 70]
Schwarz	1,2,3,4	Human iPSC reprogramming	GSE106836	GSE106834	7	[41, 71, 72]
Liu2	1,2,3,4	Human-induced trophoblast stem cell reprogramming	GSE150616	GSE150590	3	[43, 73, 74]
Melendez	1,2,3,4	Human dopaminergic neuron differentiation	GSE153005	GSE153005	4	[75, 76]
Benchetrit	1,2,3,4	Mouse blastocyst cell reprogramming	GSE98124	GSE98124	7	[40, 77]
Wu	1,2,3,4	Human somatic cell reprogramming	GSE147679	GSE147678	11	[39, 78, 79]
Yagi	1,2,3,4	Mouse nature myocytes and myogenic stem cell trans-differentiation	GSE169488	GSE169488	15	[38, 80]
Knaupp	1,2,3,4	Human iPSC reprogramming	GSE101905	GSE101905	9	[37, 81]
Li	1,2,3,4,6	Mouse iPSC reprogramming	GSE93027	GSE93026	14	[36, 82, 83]
Zenere	1,2,3,4	T-help type 1 (Th1) differentiation	E-MTAB-7775 E-MTAB-10423	E-MTAB-10444	13	[34, 84–86]
Maza	1,2,3	Somatic cell trans-differentiation		GSE67298	4	[47, 87]
Qu	1,2,3	Primary human T cells		GSE60682	29	[53, 88]
Denny	1,2,3	Primary tumors and metastases		GSE81255	31	[48, 89]
Zviran	1,2,3	Somatic cell reprogramming		GSE103821	16	[49, 90]
Corces	1,2,3	Hematopoietic and leukemic cells		GSE74912	130	[50, 91]
GM12878	5	Human LCL (GM12878)		GSE47753	1	[13, 92]
Tang	1,2,3	MCF-7 cells with retinoic acid and/or TGF-beta		GSE152749	4	[42, 93]
K562	5	Human K562 cells		GSE170378	1	[58, 94]
Etv2	5,6	Etv2-induced reprogramming and differentiation		GSE168636	17	[95, 96]
Buenrostro		Human hematopoietic differentiation		GSE96771	13	[33, 97]

SeATAC predicts histone modifications

To test whether histone modification changes were associated with chromatin accessibility changes as determined by SeATAC, we compiled the ATAC-seq data from GM12878 and K562 cell lines [13, 58] and used SeATAC to detect the nucleosome changes over NFKB1 binding sites, which were enriched in the GM12878 cell lines [15]. We identified 728 and 1633 NFKB1 binding sites that had decreased chromatin accessibility in K562 in comparison with GM12878 (SeATAC adjusted p-value < 0.05 and $\delta^{\mathrm{NOR}} > 0.2$) in distal and promoter regions, respectively (Additional File 1: Fig. S5a). We observed "dips" of the signals at the NFKB1 binding sites of all three examined euchromatic marks (H3K4me1, H3K4me3, and H3K27ac) (Additional File 1: Fig. S5b).

To further explore whether the local ATAC-seq signal captured by SeATAC, MACS2, or NucleoATAC can be predictive of the histone modification signals, we used the features produced by three tools to train a simple multilayer perceptron (MLP) to predict the H3K27ac, H3K4me1, and H3K4me3 signals (Additional File 1: Fig. S5c). SeATAC had the best performance of predicting 6 out of 7 histone modifications (Additional File 1: Fig. S5d). In summary, tasks #4 and #5 demonstrated that the latent representations produced by SeATAC on ATAC-seq data were significantly more predictive of biological readout such as gene expression and histone modifications than NucleoATAC and MACS2.

SeATAC detects chromatin accessibility changes associated with biological functions

Having established SeATAC's superior performance on three separate tasks using synthetic data, we then applied SeATAC to ATAC-seq datasets of Etv2-induced MEF reprogramming and ES/EB differentiation [98]. Etv2 is an essential transcription factor for the development of cardiac, endothelial, and hematopoietic lineages [99–109]. Moreover, Etv2 has recently been shown to function as a pioneer factor. In these studies, the induction of Etv2 drove embryonic body (EB) and MEFs to an endothelial fate [98]. Therefore, we hypothesized that the relaxed Etv2 binding sites (becoming more accessible) during the Etv2-induced differentiation or reprogramming period and should be closely associated with the endothelial function.

SeATAC identified 5451 and 2142 Etv2 motifs with increased chromatin accessibility from MEF reprogramming (undifferentiated MEFs vs. Flk1$^+$ cells at 7 days post induction) and EB differentiation (D2.5 EB vs. Flk1$^+$ cells at 12 h post induction) ATAC-seq data, respectively (adjusted p-value < 0.05 and $\delta^{\mathrm{NOR}} < -0.2$). Interestingly, SeATAC identified 2776 and 1626 relaxed Etv2 motifs that were detected by neither MACS2 nor NucleoATAC. The aggregated V-plot of 1626 SeATAC-only Etv2 binding sites showed increased NFR reads, while the aggregated V-plot of 222 MACS2-only and 2305 NucleoATAC-only Etv2 binding sites did not show significant changes from undifferentiated EBs to Flk1$^+$ cells from 12 h post induction (Fig. 5c). The aggregated V-plot of SeATAC-only, MACS2-only, and NucleoATAC-only Etv2 binding sites from MEF reprogramming also showed a similar pattern (Additional File 1: Fig. S6a). Moreover, the pathway analysis showed that the relaxed Etv2 binding sites identified by SeATAC were more significantly associated with Gene Ontology terms related to endothelial development and cell migration (Fig. 5d).

We examined two additional ATAC-seq datasets of Ascl1-induced neural reprogramming [25] (undifferentiated MEFs vs. 22 days post induction of Ascl1) and OSK (Oct4, Sox2, and Klf4)-induced reprogramming [36]. We found that the relaxed Ascl1 binding sites identified by SeATAC were more significantly associated with Gene Ontology terms related to neurogenesis and neuron migration (Additional File 1: Fig. S7), while the relaxed OSK binding sites were more significantly associated with Gene Ontology terms related to stem cell development, fibroblast growth factor receptor signaling pathways, and canonical Wnt signaling pathway (Additional File 1: Fig. S8) [36, 110]. These results suggested that SeATAC was able to identify TFBS with differential chromatin accessibility and closely related biological functions. Importantly, these differential TFBS were missed by conventional tools such as MACS2 and NucleoATAC.

Induction of pioneer factors cause both chromatin relaxation and closure

Previous studies showed that pioneer factors such as Etv2, Ascl1, and OSK could recognize their target DNA sequences in compacted chromatin, recruit chromatin remodelers, and trigger the relaxation of the adjacent chromatin landscape to accommodate non-pioneer transcription factors [111, 112]. In Etv2-induced EB differentiation and MEF reprogramming, although the overall Etv2 motif associated chromatin accessibility significantly increased, as suggested by chromVAR analysis [113], SeATAC showed that among the Etv2 motifs with differential chromatin accessibility, ~ 30% (24.6% in MEFs and 35.1% in EBs) of the Etv2 motifs showed decreased chromatin accessibility during Etv2-induced differentiation (Fig. 6a, c and Additional File 1: Fig. S11). We found that a majority of the Etv2 motifs with decreased chromatin accessibility were located near the promoter regions (Fig. 6b) and marked by euchromatic marks such as H3K4me1, H3K4me2, H3K27ac, and P300 (Fig. 6d). The decrease of chromatin accessibility was also coupled with the decrease of Brg1 (SMARCA4) density, a key SWI/SNF-related chromatin-remodeling complex that facilitates chromatin relaxation (Fig. 6d) [114]. Additionally, we found that the genes, which harbor Etv2 binding sites with decreased chromatin accessibility in the promoter regions (− 5000 to + 1000 bp surrounding the TSS), including Brachyury (T) and Mycn, were more likely to be down-regulated during the differentiation process (Fig. 6e–g, Additional File 1: Fig. S6c), suggesting the Etv2 may regulate gene expression by reducing the chromatin accessibility of their binding sites.

(See figure on next page.)

Fig. 5 SeATAC detects Etv2 binding sites with increased chromatin accessibility during Etv2-induced EB differentiation and MEF reprogramming. **a, b** The Venn diagrams show the number of Etv2 motifs with increased chromatin accessibility identified by SeATAC, MACS2, and NucleoATAC, in **a** Etv2-induced MEF reprogramming (undifferentiated MEFs vs. Flk1$^+$ cells at 7 days post-induction) and **b** Etv2-induced EB differentiation (D2.5 EB vs. Flk1$^+$ cells at 12 h post-induction). **c** The aggregated V-plot includes 1626, 222, and 2305 Etv2 motifs with increased chromatin accessibility identified by SeATAC only, MACS2 only and NucleoATAC only in ATAC-seq data of Etv2-induced EB differentiation (day 2.5 EB vs. Flk1$^+$ cells at 12 h post-induction). Both raw V-plots and estimated V-plots are shown. The heatmap color indicates the normalized read density for raw counts (top) and the estimated read density for estimated read counts (bottom). **d** The barplots show the Gene Ontology (GO) terms that are significantly associated with the genes where the promoters (− 5000 to + 1000 bp region flanking the TSS) have Etv2 motifs with increased chromatin accessibility, identified by SeATAC, MACS2, and NucleoATAC. The *y*-axis showed the adjusted *p*-value of the pathway analysis

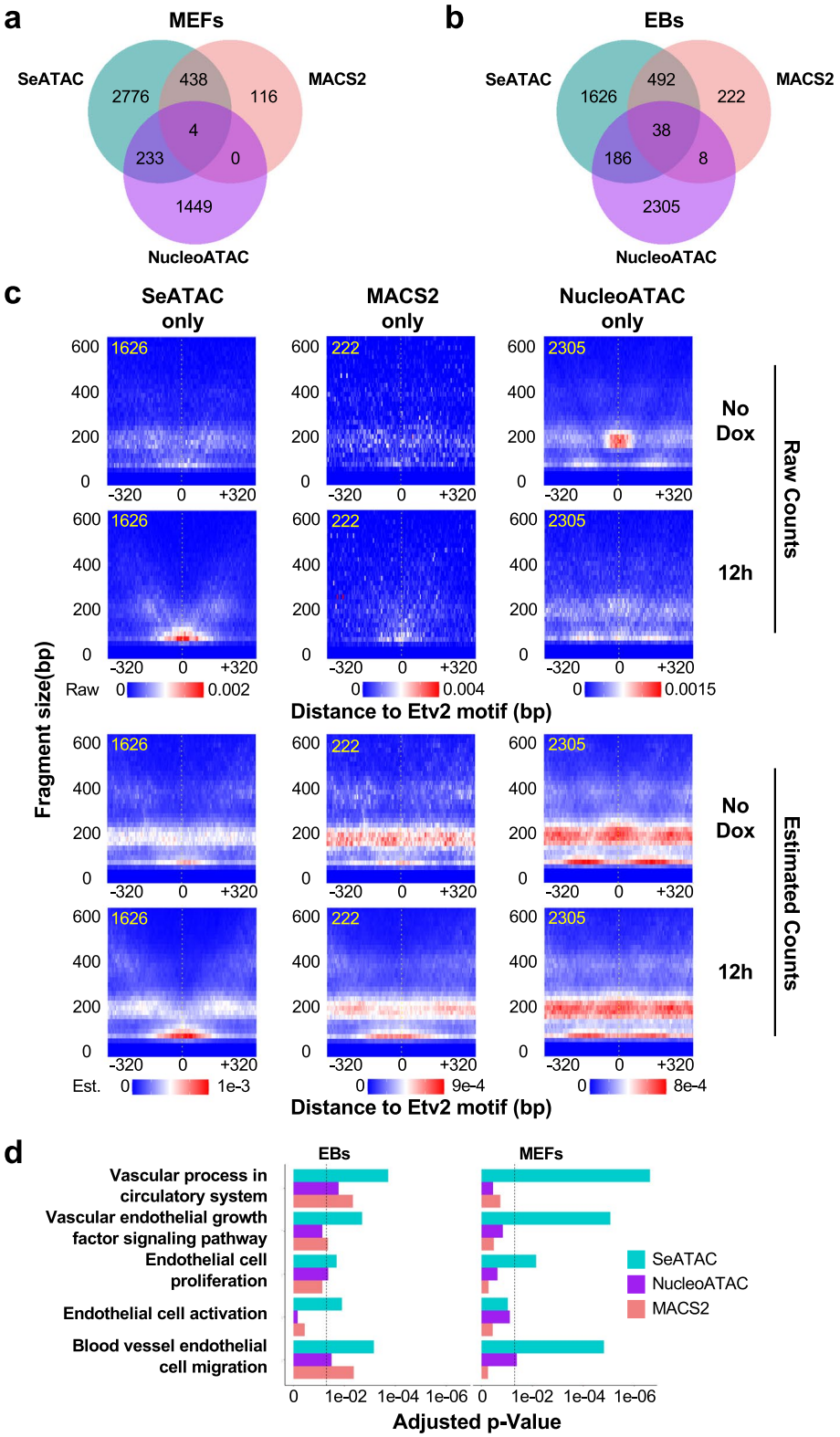

Fig. 5 (See legend on previous page.)

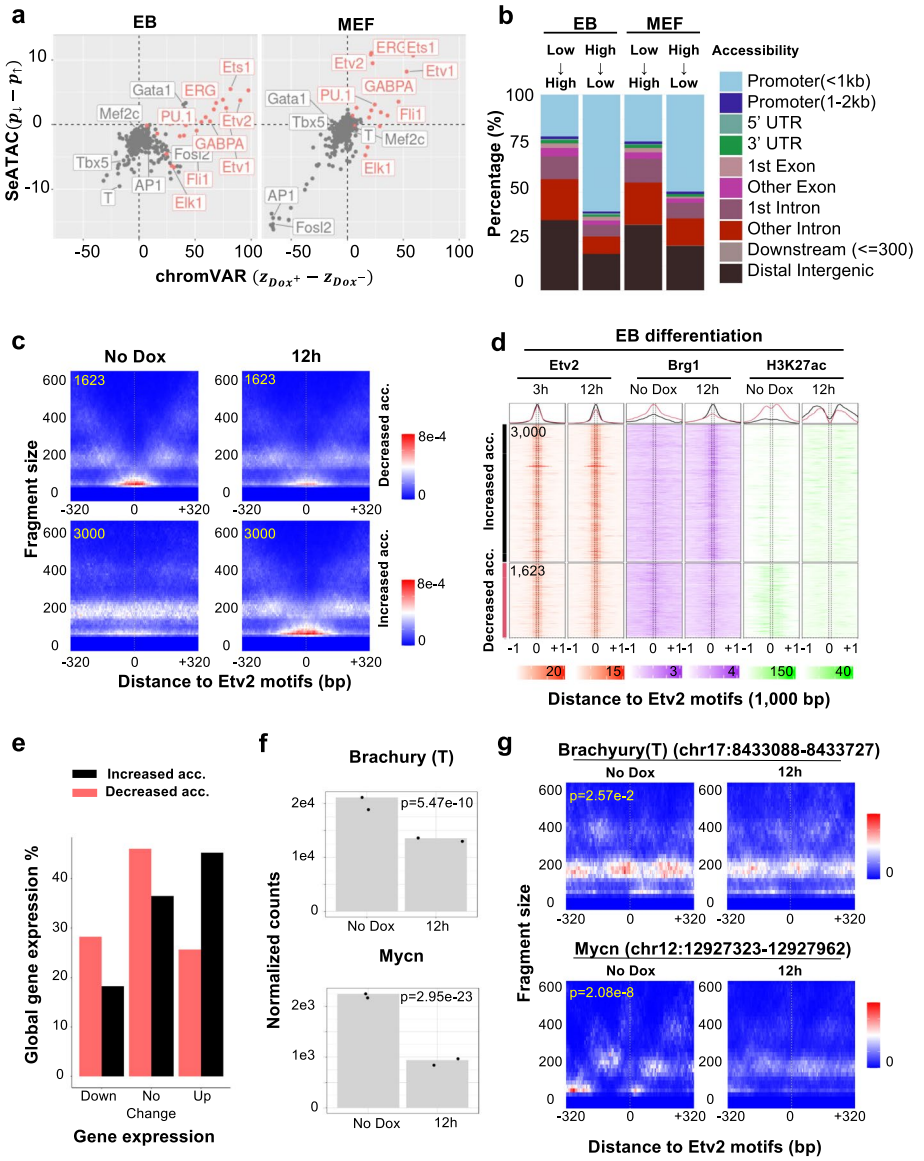

Fig. 6 Inducing Etv2 causes both chromatin relaxation and closure at Etv2 binding sites. **a** The dot plots compare the changes of motif associated chromatin accessibility estimated by chromVAR (*x*-axis) and the difference of the percent of TFBS with decreased or increased chromatin accessibility estimated by SeATAC (*y*-axis). $\boldsymbol{z_{Dox^+}}$ and $\boldsymbol{z_{Dox^-}}$ are the normalized deviation score of Dox$^+$ condition (Flk1$^+$ cells at 7 days post-induction for MEF reprogramming or Flk1$^+$ cells at 12 h post-induction for EB differentiation) and Dox$^-$ condition (undifferentiated MEFs or D2.5 EBs). $\boldsymbol{p_\downarrow}$ and $\boldsymbol{p_\uparrow}$ are the percent of TFBS that shows decreased or increased chromatin accessibility in Dox$^+$ condition compared with the Dox$^-$ condition. **b** The barplots show the genomic distribution of Etv2 binding sites with decreased (NFR- > NOR) or increased (NOR- > NFR) chromatin accessibility in EB differentiation or MEF reprogramming. The change of chromatin accessibility is estimated by SeATAC. **c** The aggregated V-plot include 3000 and 1623 Etv2 binding sites that have increased (NOR- > NFR) or decreased (NFR- > NOR) chromatin accessibility during MEF reprograming. The heatmap color indicates the estimated read density. **d** The heatmaps show the Etv2, Brg1, H3K27ac ChIP-seq of 3000 and 1623 Etv2 binding sites that have increased (NOR- > NFR) or decreased (NFR- > NOR) chromatin accessibility at day 2.5 EB (Brg1 and H3K27ac), 3 h post Etv2 induction (Etv2), and 12 h post Etv2 induction (Etv2, Brg1, and H3K27ac). The change of chromatin accessibility is estimated by SeATAC. **e** The barplots show the percent of genes that were down-regulated, up-regulated, or not changed between day 2.5 EB and 12 h post Etv2 induction. **f–g** Brachyury (T) and Mycn **(f)** are significantly down-regulated during the Etv2-induced differentiation and **(g)** have Etv2 motifs that become significantly less accessible during differentiation at their promoter region (− 5000 to + 1000 bp region flanking the TSS). The heatmap color indicates estimated read density

The analysis of the ATAC-seq dataset of Ascl1-induced neural reprogramming [25] revealed among Ascl1 motifs with differential chromatin accessibility, 19.8% showed decreased chromatin accessibility (Fig. S9a and S9d). Similar to Etv2 motifs, the Ascl1 motifs with decreased chromatin accessibility were marked by euchromatic histone marks (Fig. S9b and S9c) and were present in the promoters of genes that were down-regulated during the reprogramming, including Hmga2 [115], Egfr [116], and Elf4 [117], as well as Notch signaling member Hes1 [118] (Additional File 1: Fig. S9e). The analysis of the ATAC-seq dataset of OSK-induced reprogramming [47] also revealed that the OSK motifs that became less accessible during the reprograming were marked by euchromatic marks in MEFs, more likely located at the promoter regions, and present at the promoters of down-regulated genes during reprogramming, including Maf [119] and Smad3 [120] (Additional File 1: Fig. S10).

These results clearly showed that pioneer factors could recognize DNA sequences in both closed and open chromatin structure and alter the chromatin landscape in a context dependent manner.

Discussion

SeATAC employed a conditional variational autoencoder framework to model the ATAC-seq-specific V-plot while addressing the batch effect in the ATAC-seq data, allowing an unbiased comparison across multiple samples. The convolutional neural network (CNN) blocks used in the encoder network allowed SeATAC to robustly estimate the posterior distribution of the latent variables by considering ATAC-seq specific fragment size profile, resulting in superior performance on several tasks such as detecting differential V-plot, recovering nucleosome positions, detecting nucleosome changes, predicting differentially expressed genes, predicting histone modifications, and calling TFBS with differential chromatin accessibility compared to conventional methods such as MACS2 and NucleoATAC.

When applying ATAC-seq datasets on TF-induced differentiation and reprogramming methods, SeATAC more accurately identified TFBS with differential chromatin accessibility, resulting in a more significant association with the underlying biological function. Surprisingly, we found that the induction of pioneer factors such as Etv2, Ascl1, Oct4, Sox2, and Klf4 not only relaxed the compacted chromatin surrounding the respective binding sites but also resulted in the reduction of chromatin accessibility near 20% ~ 30% of the binding sites. The mechanism of pioneer factor induced chromatin closure and their roles in lineage specification has never been explored before and it warrants further investigation.

SeATAC was designed as a tool to model the local ATAC-seq data as a V-plot and to provide more accurate information regarding the local chromatin accessibility changes, such as nucleosomal positions and nucleosome phasing [95, 121]. However, SeATAC could not directly predict the global outcome (e.g., gene expression changes) based on the changes of local chromatin accessibility. Although SeATAC was able to determine a significant amount of pioneer factor-induced decreasing of chromatin accessibility, the functional role of these events need to be confirmed by further experiments such as mutagenesis followed by ATAC-seq or ATAC-PCR, especially for

the distal enhancers such as the ZPA regulatory sequence (ZRS) in limb development [121].

The SeATAC framework can be extended to model single cell ATAC-seq (scATAC-seq) data and to investigate the V-plot dynamics in the scATAC-seq data [15]. More sophisticated neural architecture such as attentions or transformer encoders [122] can be used to replace CNN layers to better model the dependences of ATAC-seq reads on V-plot [123]. Although throughout this study, a default width of 640 bp was used for the V-plot, a wider V-plot (e.g., 2048 bp) can be potentially used to model more nucleosomes at a specific locus and distant dependencies.

Conclusion

In the present study, we presented a novel algorithm SeATAC for the detection of genomic regions with differential chromatin accessibility and nucleosome positions. We believe that SeATAC provides an accurate and powerful way of revealing chromatin dynamics from the ATAC-seq data and be a valuable tool to examine the chromatin landscape and the functional role of epigenetic regulators.

Methods

Neural architecture

For each genomic region i in S ATAC-seq samples, the V-plot with the dimension of $W \times H \times 1$ from each sample was stacked together at the channel dimension to form an array $\boldsymbol{x}_i \in \mathrm{R}^{W \times H \times S}$, where W is the number of genomic bins, H is the number of fragment size bins, and S is the sample size. SeATAC used W=128 and H=64 by default. An embedding layer first maps the sample indicator $\boldsymbol{s} \in \mathrm{Z}^S$ to a fragment size array $\boldsymbol{g} \in \mathrm{R}^{1 \times H \times S}$. An encoder neural network then maps the modified V-plot $(\boldsymbol{x}_i + \boldsymbol{g})$ to latent variables with the mean of $\boldsymbol{z}_i \in \mathrm{R}^{K \times S}$ and the standard deviation of $\boldsymbol{\sigma} \in \mathrm{R}^{K \times 1}$, where K is the dimension of the latent variable ($K = 5$ by default). The encoder network consists of four convolutional neural networks (CNN) blocks, where each block consists of a CNN layer (filter of 16, stride of 2 and kernel size of 3), a batch normalization layer and a Rectified linear Unit (ReLU) activation layer. The output of the CNN blocks is flattened and mapped to latent variables with the mean of $\boldsymbol{z}_i$ and standard deviation of $\boldsymbol{\sigma}$ by a dense layer. The decoder neural network first maps concatenated latent variable $\boldsymbol{z}_i$ and sample indicator $\boldsymbol{s}$ to a vector of 128 by a dense layer, followed by four transposed CNN blocks, where each block consists of a transposed CNN layer (filter of 16, stride of 2 and kernel size of 3), a batch normalization layer, and a Rectified linear Unit (ReLU) activation layer. The output of the CNN blocks feed into a final softmax activation layer to normalize the values in each genomic bin to a vector that sum to one. In this study, we employed the binary cross entropy loss to minimize the difference between input and the estimated V-plot.

Task #1: Detection of differential V-plots

SeATAC

With its probabilistic representation of the data, SeATAC provides a natural way of testing differential V-plot, while intrinsically controlling for nuisance factors. We used the SeATAC model to approximate the posterior probability of the batch-free latent variable $\boldsymbol{z}$. For each genomic region and a pair of ATAC-seq samples with latent variables of mean of (z_{ak}, z_{bk}) and variance of $\left(\sigma_{ak}^2, \sigma_{bk}^2\right)$, where $k = 1, ..., K$ and K is the dimension of the latent variables, we constructed a χ^2 variable Q by standardizing the difference between $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$:

$$Q = \sum_{i=1}^{K} \frac{(z_{ak} - z_{bk})^2}{\sigma_{ak}^2 + \sigma_{bk}^2}$$

This χ^2 variable Q measures the standardized distance between a pair of V-plot on the latent space and a χ^2 test with K degree of freedom was used to compute a p-value of the difference between two V-plot [124] (p^{SeATAC}).

MACS2

We used MACS2 (v2.1.1) [20, 21] to compare two BAM files (file1.bam and file2.bam) twice by swapping the control and treatment samples, using the following parameters: "macs2 callpeak -q 0.05 –call-summits -f BAMPE –nomodel -t file1.bam -c file2.bam –keep-dup all" and "macs2 callpeak -q 0.05 –call-summits -f BAMPE –nomodel -t file2.bam -c file1.bam –keep-dup all". The maximum absolute values of the difference of pileup signals that overlapped with a 640-bp genomic region was used as the difference of nucleosome signals for this genomic region.

NucleoATAC

We used NucleoATAC (v0.3.4 with default parameters) [16] to estimate the nucleosome signal of two BAM files separately and calculated the difference of estimated nucleosome signal for genomic regions. The maximum absolute values of the difference of nucleosome signals that overlapped with a 640-bp genomic region was used as the difference of nucleosome signals for this genomic region.

Task #2: Estimating the nucleosome signals

SeATAC

For any genomic region, SeATAC generates estimated V-plot $\widehat{\boldsymbol{x}} \in \mathrm{R}^{W \times H}$ based on the latent variables $\boldsymbol{z}$ and a constant sample indicator s_0, from which we computed the *central NFR score*:

$$w^{NFR} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \widehat{x}_{ij}$$

where N is the number of central genomic bins and M is the number of fragment size bins for NFR. The central genomic bins were defined as the genomic bins which distance to the V-plot center (d_i) is less than 50 bp ($-50 \leq d_i \leq 50$), and fragment size bins for

NFR (f_j) were defined as the fragment size less than 150 bp ($f_j \leq 150$). The *center nucleosome score* was defined as:

$$Nuc^{SeATAC} = 1 - w^{NFR}$$

The central nucleosome score (Nuc^{SeATAC}) was used as the nucleosome score estimated by SeATAC to rank the nucleosomes.

NucleoATAC

We used NucleoATAC to estimate the nucleosome signal from the input BAM files. We defined the *central NucleoATAC signal* as the average NucleoATAC signal over the 100-bp region flanking the V-plot center.

$$Nuc^{\mathrm{NucATAC}} = \frac{1}{100}\sum_{i=1}^{100} h_i$$

where h_i is the NucleoATAC signal at position i. We used central NucleoATAC signal (Nuc^{NucATAC}) to rank the nucleosomes for this task.

Task #3: Detection of nucleosome changes

SeATAC

For any genomic region between a pair of ATAC-seq samples (i,j), SeATAC computed the *differential central nucleosome score* by:

$$\delta^{\mathrm{SeATAC}} = \log Nuc_j^{\mathrm{SeATAC}} - \log Nuc_i^{\mathrm{SeATAC}}$$

δ^{SeATAC} quantitatively measures how estimated nucleosome signal changes from sample i to i over the 100-bp regions flanking the center.

NucleoATAC

For any genomic region between a pair of ATAC-seq samples (i,j), the *differential central NucleoATAC signal* ($\delta^{\mathrm{NucATAC}}$) was defined as the difference of the average NucleoATAC signal over the 100-bp region flanking the V-plot center between a pair of ATAC-seq samples:

$$\delta^{\mathrm{SeATAC}} = Nuc_j^{\mathrm{NucATAC}} - Nuc_i^{\mathrm{NucATAC}}$$

Task #4: Predicting differentially expressed genes from ATAC-seq signals near promoters

SeATAC

The latent representations of the V-plots centering at transcription start sites (TSS) were used as the input for training a logistic regression model.

MACS2

The principal components (PC) of the differential pileup signals (MACS2's pileup output) centering at TSS were used as the input for training a logistic regression model.

NucleoATAC

The principal components (PC) of the differential nucleosome signals (NucleoATAC's smooth signal output) centering at TSS were used as the input for training a logistic regression model.

We kept the feature dimensions the same for three tools. Simple logistical regression models were built to predict the up- or down-regulated genes (a classification task) from the corresponding features, followed by a fivefold cross-validation (CV) to evaluate the prediction.

Task #5: Predicting histone modifications

SeATAC

The latent representations of the V-plots were used as the input for training a MLP model.

MACS2

The principal components (PC) of the pileup signals (MACS2's pileup output) were used as the input for training a MLP model.

NucleoATAC

The principal components (PC) of the nucleosome signals (NucleoATAC's smooth signal output) were used as the input for training a MLP model. We kept the feature dimensions the same for three tools. MLP models were built to predict the observed histone modifications (H3K27ac, H3K4me1, and H3K4me3) (Additional File 1: Fig. 5c). We trained the MLP model on 50,000 randomly sampled genomic regions from GM12878 and tested it on 20,000 randomly sampled genomic regions from K562 (H3K27ac, H3K4me1, and H3K4me3), Etv2-induced reprogramming (H3K27ac only), and Etv2-induced EB differentiation datasets (H3K27ac only) [95]. The mean squared error between known and predicted histone modification signals was used to quantitatively evaluate the prediction performance.

Task #6: Designation of TFBS with increased chromatin accessibility

SeATAC

Between a pair of ATAC-seq samples (i, j), SeATAC determined that a TFBS became more accessible in sample j compared with sample i if the adjusted p-value < 0.05 and $\delta^{\mathrm{SeATAC}} < -0.2$.

MACS2

First, we used MACS2 to compare the sample i (file1.bam) and sample j (file2.bam) using the following parameters: "macs2 callpeak -q 0.05 –call-summits -f BAMPE –nomodel -c file1.bam -t file2.bam –keep-dup all". Then we computed MACS2 p-value for a specific TFBS as the minimum p-values of all summits that overlapped with the 100-bp region flanking this TFBS (p^{MACS2}). MACS2 determined that a TFBS became more accessible if adjusted $p^{\mathrm{MACS2}} < 0.05$.

NucleoATAC

NucleoATAC determined that a TFBS became more accessible in sample j compared with sample i if the $\delta^{\mathrm{NucATAC}} < -0.4$.

Input data processing

Ascl1-induced MEF reprogramming (Ascl1)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE101397) [25, 59]. The sequence reads for undifferentiated MEFs was obtained from GSM2701947. For MEFs at day 22 post Ascl1 induction the sequence reads were pooled from three replicates (GSM2701979, GSM2701980, and GSM2701981). MACS2 identified 123,271 peaks for undifferentiated MEFs and MEFs at day 22 post Ascl1 induction and motifmatchr identified 71,616 canonical Ascl1 motif binding sites. The RNA-seq dataset was downloaded from NCBI GEO database (GSE43916) [60, 61].

Retinoic acid (RA)-induced mESC differentiation (Duren)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE136312) [35, 62]. The samples of day 0, day 2, day 4, day 10, and day 20 post differentiation were used in downstream analysis.

Human adipocyte differentiation (Perrin)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE178795 and GSE178794) [51, 63, 64]. The samples of day 0, day 2, day 4, and day 14 post differentiation were used in the downstream analysis.

Human myeloid differentiation (Ramirez)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE79019 and GSE79044) [46, 65, 66]. The 30 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Human cardiac differentiation (Liu)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE120992 and GSE121052) [44, 67, 68]. The 16 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Human early iPSC reprogramming (Markov)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE85330 and GSE85331) [45, 69, 70]. The 10 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Human iPSC reprogramming (Schwarz)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE106834 and GSE106836) [41, 71, 72]. The 7 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Human-induced trophoblast stem cell reprogramming (Liu2)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE150590 and GSE150616) [43, 73, 74]. The three samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Human dopaminergic neuron differentiation (Melendez)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE153005) [75, 76]. The four samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Mouse blastocyst cell reprogramming (Benchetrit)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE98124) [40, 77]. The 7 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Human somatic cell reprogramming (Wu)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE147678 and GSE147679) [39, 78, 79]. The 11 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Mouse nature myocytes and myogenic stem cell trans-differentiation (Yagi)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE169488) [38, 80]. The 15 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Human iPSC reprogramming (Knaupp)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (GSE101905) [37, 81]. The 9 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

OSK-induced MEF reprogramming (Li)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE93026) [36]. The sequence reads for the undifferentiated MEF samples were pooled from two replicates (GSM2442671 and GSM2442671) and for the MEFs at day 7 post-OSK induction were pooled from two replicates (GSM2442705 and GSM2442706). We used motifmatchr to identify 282,789 putative binding sites for Oct4, Sox2, or Klf4 for the downstream analysis. The RNA-seq datasets were downloaded from NCBI GEO database (GSE93027) [36, 82, 83].

T-help type 1 (Th1) differentiation (Zenere)

The ATAC-seq and RNA-seq datasets were downloaded from NCBI GEO database (E-MTAB-7775, E-MTAB-10423, and E-MTAB-10444) [34, 84–86]. The 13 samples with both ATAC-seq and RNA-seq profiles were used in the downstream analysis.

Somatic cell trans-differentiation (Maza)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE67298) [47, 87]. The four samples from the ATAC-seq dataset were used in the downstream analysis.

Primary human T cells (Qu)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE60682) [53, 88]. The 29 samples from the ATAC-seq dataset were used in the downstream analysis.

Primary tumors and metastases (Denny)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE81255) [48, 89]. The 31 samples from the ATAC-seq dataset were used in the downstream analysis.

Somatic cell reprogramming (Zviran)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE103821) [49, 90]. The 16 samples from the ATAC-seq dataset were used in the downstream analysis.

Hematopoietic and leukemic cells (Corces)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE74912) [50, 91]. The 130 samples from the ATAC-seq dataset were used in the downstream analysis.

GM12878

EBV-transformed lymphoblastoid cell line (LCL) ATAC-seq data were downloaded from NCBI GEO database (GSE47753) [33, 97]. The sequence reads from three replicates of 50 k cell sample (GSM1155957, GSM1155958, and GSM1155959) were pooled and used for the downstream analysis. The 86,004 peaks called by MACS2 (v2.1.1) [20, 21] were used for the downstream analysis.

MCF-7 cells with retinoic acid and/or TGF-beta (Tang)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE152749) [42, 93]. The four samples from the ATAC-seq dataset were used in the downstream analysis.

Human K562 cells (K562)

The K562 ATAC-seq dataset was downloaded from NCBI GEO database (GSE170378) [58, 94].

Etv2-induced MEF reprogramming and ES/EB differentiation (Etv2)

The ATAC-seq dataset was downloaded from NCBI GEO database (GSE168636) [96, 98]. Sequence reads for undifferentiated MEFs were pooled from two replicates (GSM5151877 and GSM5151879) and for $Flk1^{+}$ MEFs at day 7 post-Etv2 induction were pooled from two replicates (GSM5151861 and GSM5151863). Sequence

reads for undifferentiated EBs were pooled from two replicates (GSM5151873 and GSM5151875), and for Flk1$^+$ EBs at day 2.5 post Etv2 induction were pooled from two replicates (GSM5151869 and GSM5151871). MACS2 (v2.1.1) identified 57,732 peaks for undifferentiated MEFs and Flk1 + MEFs at day 7 post Etv2 induction and 36,114 peaks for undifferentiated EBs and Flk1 + EBs at day 2.5 post Etv2 induction. We used motifmatchr (v1.16.0) to obtain 20,822 and 24,935 putative Etv2 motif binding regions for MEFs and EBs, respectively [25, 59, 36].

Human hematopoietic differentiation (Buenrostro)

The dataset was downloaded from NCBI GEO database (GSE96771). A union set with 491,437 peaks defined by the original authors were used for the downstream analysis [33, 97].

For ATAC-seq data, the sequencing reads where mapped to the mouse and human genome (mm10 or hg19) using Bowtie2 (v2.2.4) [125]. The ATAC-seq reads lied on chromosome Y and mitochondria were excluded [126]. ChromVAR (v1.10) [113] were used for transcription factor based chromatin accessibility analysis; 322 transcription factors compiled in the Homer database were used for the chromVAR analysis. The pathway analysis was performed using R packages clusterProfiler and ChIPseeker [127, 128]. For RNA-seq, the sequencing reads were mapped to the mouse and human genome (mm10 or hg19) using kallisto (v0.46.0) [129].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02954-5.

Additional file 1: Fig. S1. SeAAQTAC corrects batch effects in fragment size distributions in ATAC-seq data. **Fig. S2.** SeATAC detects differential V-plot with different shift size in synthetic data. **Fig. S3.** The performance of SeATAC on different ATAC-seq quality indices. **Fig. S4.** SeATAC accurately predicted differentially expressed genes from ATAC-seq signals surrounding the transcription start sites. **Fig. S5.** SeATAC predicts histone modifications. **Fig. S6.** SeATAC detects Etv2 binding sites with increased chromatin accessibility during Etv2 induced EB differentiation and MEF reprogramming. **Fig. S7.** SeATAC detects Ascl1 binding sites with increased chromatin accessibility during Ascl1 induced MEF reprograming. **Fig. S8.** SeATAC detects OSK binding sites with increased chromatin accessibility during OSK induced reprograming. **Fig. S9.** Induction of Ascl1 causes both chromatin relaxation and closure at Ascl1 binding sites. **Fig. S10.** Induction of OSKcauses both chromatin relaxation and closure at OSK binding sites. **Fig. S11.** SeATAC detects Etv2 binding sites with significantly altered chromatin accessibility during the Etv2 induced reprogramming.

Additional file 2. Review history.

Acknowledgements
We acknowledge the Minnesota Supercomputing Institute for providing computational resources.

Review history
The review history is available as Additional File 2.

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

Authors' contributions
WG conceptualized the study, designed the algorithm, led the data analysis, and wrote the manuscript with the input from all authors. ND performed the data analysis. DJG supervised the study. All authors read and approved the final manuscript.

Funding
These studies were supported by funding from NHLBI (P01HL160476), Department of Defense (W81XWH2110606) and Minnesota Regenerative Medicine.

Availability of data and materials
The latest version of SeATAC is freely available as a R package on GitHub (https://github.com/gongx030/seatac) under the GNU 3 license [130], and the source code used to obtain the results presented in this article are available on GitHub (https://github.com/gongx030/seatac_manuscript) and as a Zenodo archive with DOI https://doi.org/10.5281/zenodo.7819334 [131]. The datasets used in this project are listed in Table 1.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
DJG is the founder of NorthStar Genomics. The remaining authors declare no competing interests.

Received: 26 April 2022 Accepted: 27 April 2023
Published online: 22 May 2023

References

1. Richmond TJ, Davey CA. The structure of DNA in the nucleosome core. Nature. 2003;423:145–50.
2. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. 2013;20:267–73.
3. Martinez-Campa C, Politis P, Moreau J-L, Kent N, Goodall J, Mellor J, et al. Precise nucleosome positioning and the TATA box dictate requirements for the histone H4 tail and the bromodomain factor Bdf1. Mol Cell. 2004;15:69–81.
4. Lomvardas S, Thanos D. Nucleosome sliding via TBP DNA binding in vivo. Cell. 2001;106:685–96.
5. Churchman LS, Weissman JS. Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature. 2011;469:368–73.
6. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013;339:950–3.
7. Weber CM, Ramachandran S, Henikoff S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol Cell. 2014;53:819–30.
8. Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008;132:887–98 Available from:http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&;id=18329373&retmode=ref&cmd=prlinks.
9. Cui K, Zhao K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. Methods Mol Biol. 2012;833:413–9.
10. Voong LN, Xi L, Sebeson AC, Xiong B, Wang J-P, Wang X. Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. Cell. 2016;167:1555-1570.e15.
11. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132:311–22.
12. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007;17:877–85.
13. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213–8.
14. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods. 2017;14:959–62.
15. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486–90.
16. Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res. 2015;25(11):1757–70. https://doi.org/10.1101/gr.192294.115.
17. Xu B, Li X, Gao X, Jia Y, Liu J, Li F, et al. DeNOPA: decoding nucleosome positions sensitively with sparse ATAC-seq data. Brief Bioinform. 2021;23:bbab469.
18. Chen K, Xi Y, Pan X, Li Z, Kaestner K, Tyler J, Dent S, He X, Li W. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res. 2013;23(2):341–51. https://doi.org/10.1101/gr.142067.112.
19. Zhang Y, Shin H, Song JS, Lei Y, Liu XS. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. BMC Genomics. 2008;9:537.
20. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012;7:1728–40.
21. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
22. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol. 2020;21:22.
23. Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. Epigenome characterization at single base-pair resolution. Proc Natl Acad Sci U S A. 2011;108:18318–23.
24. Qu J, Yi G, Zhou H. p63 cooperates with CTCF to modulate chromatin architecture in skin keratinocytes. bioRxiv. 2019;140:525667.

25. Wapinski OL, Lee QY, Chen AC, Li R, Corces MR, Ang CE, et al. Rapid chromatin switch in the direct reprogramming of fibroblasts to neurons. Cell Rep. 2017;20:3236–47.
26. Gutin J, Sadeh R, Bodenheimer N, Joseph-Strauss D, Klein-Brill A, Alajem A, et al. Fine-resolution mapping of TF binding and chromatin interactions. Cell Rep. 2018;22:2797–807.
27. Brahma S, Henikoff S. RSC-associated subnucleosomes define MNase-sensitive promoters in yeast. Mol Cell. 2019;73:238-249.e3.
28. Bao X, Rubin AJ, Qu K, Zhang J, Giresi PG, Chang HY, et al. A novel ATAC-seq approach reveals lineage-specific reinforcement of the open chromatin landscape via cooperation between BAF and p63. Genome Biol. 2015;16:284.
29. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. 2013.
30. Kingma DP, Mohamed S, Rezende DJ, Welling M. Semi-supervised learning with deep generative models. arXiv. org. 2014;56:cs.LG-9 (Available from: arXiv.org).
31. Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems. 2015;28.
32. Tarbell ED, Liu T. HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. Nucleic Acids Res. 2019;21:175.
33. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. Cell. 2018;173(6):1535 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96771.
34. Zenere A, Rundquist O, Gustafsson M, Altafini C. Using high-throughput multi-omics data to investigate structural balance in elementary gene regulatory network motifs. Bioinformatics. 2021;38:173–8.
35. Duren Z, Chen X, Xin J, Wang Y, Wong W. Time course regulatory analysis based on paired expression and chromatin accessibility data. Genome Res. 2020;30:gr.257063.119.
36. Li D, Liu J, Yang X, Zhou C, Guo J, Wu C, et al. Chromatin accessibility dynamics during iPSC reprogramming. Cell Stem Cell. 2017;21:819-833.e6.
37. Knaupp AS, Buckberry S, Pflueger J, Lim SM, Ford E, Larcombe MR, et al. Transient and permanent reconfiguration of chromatin and transcription factor occupancy drive reprogramming. Cell Stem Cell. 2017;21:834-845.e6.
38. Yagi M, Ji F, Charlton J, Cristea S, Messemer K, Horwitz N, et al. Dissecting dual roles of MyoD during lineage conversion to mature myocytes and myogenic stem cells. Gene Dev. 2021;35:1209–28.
39. Wu L, Zhao G, Xu S, Kuang J, Ming J, Wu G, et al. The nuclear factor CECR2 promotes somatic cell reprogramming by reorganizing the chromatin structure. J Biol Chem. 2021;296:100022.
40. Benchetrit H, Jaber M, Zayat V, Sebban S, Pushett A, Makedonski K, et al. Direct induction of the three pre-implantation blastocyst cell types from fibroblasts. Cell Stem Cell. 2019;24:983-994.e7.
41. Schwarz BA, Cetinbas M, Clement K, Walsh RM, Cheloufi S, Gu H, et al. Prospective isolation of poised iPSC intermediates reveals principles of cellular reprogramming. Cell Stem Cell. 2018;23:289-305.e5.
42. Tang Y, Xiong S, Yu P, Liu F, Cheng L. Direct conversion of mouse fibroblasts into neural stem cells by chemical cocktail requires stepwise activation of growth factors and Nup210. Cell Rep. 2018;24:1355-1362.e3.
43. Liu X, Ouyang JF, Rossello FJ, Tan JP, Davidson KC, Valdes DS, et al. Reprogramming roadmap reveals route to human induced trophoblast stem cells. Nature. 2020;586:101–7.
44. Liu Q, Jiang C, Xu J, Zhao M-T, Bortle KV, Cheng X, et al. Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hiPSCs and hESCs. Circ Res. 2017;121:376–91.
45. Markov GJ, Mai T, Nair S, Shcherbina A, Wang YX, Burns DM, et al. AP-1 is a temporally regulated dual gatekeeper of reprogramming to pluripotency. Proc National Acad Sci. 2021;118:e2104841118.
46. Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A, Mortazavi A. Dynamic gene regulatory networks of human myeloid differentiation. Cell Syst. 2017;4:416-429.e3.
47. Maza I, Caspi I, Zviran A, Chomsky E, Rais Y, Viukov S, et al. Transient acquisition of pluripotency during somatic cell transdifferentiation with iPSC reprogramming factors. Nat Biotechnol. 2015;33(7):769–74.
48. Denny SK, Yang D, Chuang C-H, Brady JJ, Lim JS, Grüner BM, et al. Nfib Promotes metastasis through a widespread increase in chromatin accessibility. Cell. 2016;166:328–42.
49. Zviran A, Mor N, Rais Y, Gingold H, Peles S, Chomsky E, et al. Deterministic somatic cell reprogramming involves continuous transcriptional changes governed by Myc and epigenetic-driven modules. Cell Stem Cell. 2019;24:328-341.e9.
50. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;48:1193–203.
51. Perrin HJ, Currin KW, Vadlamudi S, Pandey GK, Ng KK, Wabitsch M, et al. Chromatin accessibility and gene expression during adipocyte differentiation identify context-dependent effects at cardiometabolic GWAS loci. Plos Genet. 2021;17:e1009865.
52. Sanford EM, Emert BL, Coté A, Raj A. Gene regulation gravitates toward either addition or multiplication when combining the effects of two signals. Elife. 2020;9:e59388.
53. Qu K, Zaba LC, Giresi PG, Li R, Longmire M, Kim YH, et al. Individuality and variation of personal regulomes in primary human T cells. Cell Syst. 2015;1:51–61.
54. Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Res. 2012;22:1735–47.
55. Kiani K, Sanford EM, Goyal Y, Raj A. Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations. Mol Syst Biol. 2022;18:e10979.
56. González AJ, Setty M, Leslie CS. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. Nat Genet. 2015;47:1249–59.
57. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.
58. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
59. Wapinski OL, Lee QY, Chen AC, Li R, Corces MR, Ang CE, et al. Rapid chromatin switch in the direct reprogramming of fibroblasts to neurons. Datasets Gene Expression Omnibus . 2017;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE101397.

60. Wapinski OL, Vierbuchen T, Qu K, Lee QY, Chanda S, Fuentes DR, et al. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. Datasets Gene Expression Omnibus . 2013;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43916.
61. Wapinski OL, Vierbuchen T, Qu K, Lee QY, Chanda S, Fuentes DR, et al. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. Cell. 2013;155:621–35.
62. Duren Z, Chen X, Xin J, Wang Y, Wong W. Time course regulatory analysis based on paired expression and chromatin accessibility data. Datasets Gene Expression Omnibus . 2020;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136312.
63. Perrin HJ, Currin KW, Vadlamudi S, Pandey GK, Ng KK, Wabitsch M, et al. Chromatin accessibility and gene expression during adipocyte differentiation identify context-dependent effects at cardiometabolic GWAS loci. Datasets Gene Expression Omnibus . 2021;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178794.
64. Perrin HJ, Currin KW, Vadlamudi S, Pandey GK, Ng KK, Wabitsch M, et al. Chromatin accessibility and gene expression during adipocyte differentiation identify context-dependent effects at cardiometabolic GWAS loci. Datasets Gene Expression Omnibus . 2021;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178795.
65. Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A, Mortazavi A. Dynamic gene regulatory networks of human myeloid differentiation. Datasets Gene Expression Omnibus . 2017;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79019.
66. Ramirez RN, El-Ali NC, Mager MA, Wyman D, Conesa A, Mortazavi A. Dynamic gene regulatory networks of human myeloid differentiation. Datasets Gene Expression Omnibus . 2017;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79044.
67. Liu Q, Jiang C, Xu J, Zhao M-T, Bortle KV, Cheng X, et al. Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hiPSCs and hESCs. Datasets Gene Expression Omnibus . 2017;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85331.
68. Liu Q, Jiang C, Xu J, Zhao M-T, Bortle KV, Cheng X, et al. Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hiPSCs and hESCs. Datasets Gene Expression Omnibus . 2017;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85330.
69. Markov GJ, Mai T, Nair S, Shcherbina A, Wang YX, Burns DM, et al. AP-1 is a temporally regulated dual gatekeeper of reprogramming to pluripotency. Datasets Gene Expression Omnibus . 2021;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121052.
70. Markov GJ, Mai T, Nair S, Shcherbina A, Wang YX, Burns DM, et al. AP-1 is a temporally regulated dual gatekeeper of reprogramming to pluripotency. Datasets Gene Expression Omnibus . 2021;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120992.
71. Schwarz BA, Cetinbas M, Clement K, Walsh RM, Cheloufi S, Gu H, et al. Prospective isolation of poised iPSC intermediates reveals principles of cellular reprogramming. Datasets Gene Expression Omnibus . 2018;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106834.
72. Schwarz BA, Cetinbas M, Clement K, Walsh RM, Cheloufi S, Gu H, et al. Prospective isolation of poised iPSC intermediates reveals principles of cellular reprogramming. Datasets Gene Expression Omnibus . 2018;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106836.
73. Liu X, Ouyang JF, Rossello FJ, Tan JP, Davidson KC, Valdes DS, et al. Reprogramming roadmap reveals route to human induced trophoblast stem cells. Datasets Gene Expression Omnibus . 2020;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150590.
74. Liu X, Ouyang JF, Rossello FJ, Tan JP, Davidson KC, Valdes DS, et al. Reprogramming roadmap reveals route to human induced trophoblast stem cells. Datasets Gene Expression Omnibus . 2020;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150616.
75. Meléndez-Ramírez C, Duran RC-D, Barrios-García T, Giacoman-Lozano M, López-Ornelas A, Herrera-Gamboa J, et al. Dynamic landscape of chromatin accessibility and transcriptomic changes during differentiation of human embryonic stem cells into dopaminergic neurons. Sci Rep-uk. 2021;11:16977.
76. Meléndez-Ramírez C, Duran RC-D, Barrios-García T, Giacoman-Lozano M, López-Ornelas A, Herrera-Gamboa J, et al. Dynamic landscape of chromatin accessibility and transcriptomic changes during differentiation of human embryonic stem cells into dopaminergic neurons. Datasets Gene Expression Omnibus . 2021;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE153005.
77. Benchetrit H, Jaber M, Zayat V, Sebban S, Pushett A, Makedonski K, et al. Direct induction of the three pre-implantation blastocyst cell types from fibroblasts. Datasets Gene Expression Omnibus . 2019;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98124.
78. Wu L, Zhao G, Xu S, Kuang J, Ming J, Wu G, et al. The nuclear factor CECR2 promotes somatic cell reprogramming by reorganizing the chromatin structure. Datasets Gene Expression Omnibus . 2021;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147679.
79. Wu L, Zhao G, Xu S, Kuang J, Ming J, Wu G, et al. The nuclear factor CECR2 promotes somatic cell reprogramming by reorganizing the chromatin structure. Datasets Gene Expression Omnibus . 2021;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147678.
80. Yagi M, Ji F, Charlton J, Cristea S, Messemer K, Horwitz N, et al. Dissecting dual roles of MyoD during lineage conversion to mature myocytes and myogenic stem cells. Datasets Gene Expression Omnibus . 2021;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169488.
81. Knaupp AS, Buckberry S, Pflueger J, Lim SM, Ford E, Larcombe MR, et al. Transient and permanent reconfiguration of chromatin and transcription factor occupancy drive reprogramming. Datasets Gene Expression Omnibus . 2017;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE101905.
82. Li D, Liu J, Yang X, Zhou C, Guo J, Wu C, et al. Chromatin accessibility dynamics during iPSC reprogramming. Datasets Gene Expression Omnibus . 2017;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE93026.
83. Li D, Liu J, Yang X, Zhou C, Guo J, Wu C, et al. Chromatin accessibility dynamics during iPSC reprogramming. Datasets Gene Expression Omnibus . 2017;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE93027.

84. Zenere A, Rundquist O, Gustafsson M, Altafini C. Using high-throughput multi-omics data to investigate structural balance in elementary gene regulatory network motifs. Datasets ArrayExpress. 2021;https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-10444.
85. Zenere A, Rundquist O, Gustafsson M, Altafini C. Using high-throughput multi-omics data to investigate structural balance in elementary gene regulatory network motifs. Datasets ArrayExpress. 2021;https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-10423.
86. Zenere A, Rundquist O, Gustafsson M, Altafini C. Using high-throughput multi-omics data to investigate structural balance in elementary gene regulatory network motifs. Datasets ArrayExpress. 2021;https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-7775.
87. Maza I, Caspi I, Zviran A, Chomsky E, Rais Y, Viukov S, et al. Transient acquisition of pluripotency during somatic cell transdifferentiation with iPSC reprogramming factors. Datasets Gene Expression Omnibus . 2015;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67298.
88. Qu K, Zaba LC, Giresi PG, Li R, Longmire M, Kim YH, et al. Individuality and variation of personal regulomes in primary human T cells. Datasets Gene Expression Omnibus . 2015;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60682.
89. Denny SK, Yang D, Chuang C-H, Brady JJ, Lim JS, Grüner BM, et al. Nfib promotes metastasis through a widespread increase in chromatin accessibility. Datasets Gene Expression Omnibus . 2016;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81255.
90. Zviran A, Mor N, Rais Y, Gingold H, Peles S, Chomsky E, et al. Deterministic somatic cell reprogramming involves continuous transcriptional changes governed by Myc and epigenetic-driven modules. Datasets Gene Expression Omnibus . 2019;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103821.
91. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Datasets Gene Expression Omnibus . 2016;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74912.
92. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Datasets Gene Expression Omnibus. 2013;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47753.
93. Tang Y, Xiong S, Yu P, Liu F, Cheng L. Direct conversion of mouse fibroblasts into neural stem cells by chemical cocktail requires stepwise activation of growth factors and Nup210. Datasets Gene Expression Omnibus . 2018;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152749.
94. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. Datasets Gene Expression Omnibus . 2012;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE170378.
95. Gong W, Das S, Sierra-Pagan JE, Skie E, Dsouza N, Larson TA, et al. ETV2 functions as a pioneer factor to regulate and reprogram the endothelial lineage. Nat Cell Biol. 2022;24:672–84.
96. Gong W, Das S, Sierra-Pagan JE, Skie E, Dsouza N, Larson TA, et al. ETV2 functions as a pioneer factor to regulate and reprogram the endothelial lineage. Datasets Gene Expression Omnibus . 2022;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE168636.
97. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. datasets gene expression omnibus. 2018;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96771.
98. Gong W, Das S, Sierra-Pagan JE, Skie E, Dsouza N, Larson TA, et al. ETV2 functions as a pioneer factor to regulate and reprogram the endothelial lineage. Nature Cell Biology. 2022;24(5):672–84.
99. Garry DJ, Olson EN. A common progenitor at the heart of development. Cell. 2006;127:1101–4.
100. Shi X, Richard J, Zirbes KM, Gong W, Lin G, Kyba M, et al. Cooperative interaction of Etv2 and Gata2 regulates the development of endothelial and hematopoietic lineages. Dev Biol. 2014;389:208–18.
101. Gong W, Rasmussen TL, Singh N, Koyano-Nakagawa N, Pan W, Garry DJ. Dpath software reveals hierarchical haemato-endothelial lineages of Etv2 progenitors based on single-cell transcriptome analysis. Nat Commun. 2017;8:14362.
102. Koyano-Nakagawa N, Kweon J, Iacovino M, Shi X, Rasmussen TL, Borges L, et al. Etv2 is expressed in the yolk sac hematopoietic and endothelial progenitors and regulates lmo2 gene expression. Stem cells (Dayton, Ohio). 2012;30:1611–23; http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&;id=22628281&retmode=ref&cmd=prlinks.
103. Rasmussen TL, Kweon J, Diekmann MA, Belema-Bedada F, Song Q, Bowlin K, et al. ER71 directs mesodermal fate decisions during embryogenesis. Development (Cambridge, England). 2011;138:4801–12.
104. Rasmussen TL, Martin CM, Walter CA, Shi X, Perlingeiro R, Koyano-Nakagawa N, et al. Etv2 rescues Flk1 mutant embryoid bodies. Genesis (New York, NY : 2000). 2013;51:471–80.
105. Koyano-Nakagawa N, Shi X, Rasmussen TL, Das S, Walter CA, Garry DJ. Feedback mechanisms regulate Ets variant 2 (Etv2) gene expression and hematoendothelial lineages. J Biol Chem. 2015;290:28107–19.
106. Shi X, Wallis AM, Gerard RD, Voelker KA, Grange RW, Depinho RA, et al. Foxk1 promotes cell proliferation and represses myogenic differentiation by regulating Foxo4 and Mef2 factors. J Cell Sci. 2012;125(Pt 22):5329–37.
107. Liu F, Li D, Yu YYL, Kang I, Cha M-J, Kim JY, et al. Induction of hematopoietic and endothelial cell program orchestrated by ETS transcription factor ER71/ETV2. EMBO Rep. 2015;16:654–69.
108. Singh BN, Koyano-Nakagawa N, Gong W, Moskowitz IP, Weaver CV, Braunlin E, et al. A conserved HH-Gli1-Mycn network regulates heart regeneration from newt to human. Nat Commun. 2018;9:4237.
109. Singh BN, Tahara N, Kawakami Y, Das S, Koyano-Nakagawa N, Gong W, et al. Etv2-miR-130a-Jarid2 cascade regulates vascular patterning during embryogenesis. PLoS One. 2017;12:e0189010.
110. Chronis C, Fiziev P, Papp B, Butz S, Bonora G, Sabri S, et al. Cooperative binding of transcription factors orchestrates reprogramming. Cell. 2017;168(3):442-459.e20.
111. Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell. 2015;161:555–68.

112. Zaret KS. Pioneer transcription factors initiating gene network changes. Annu Rev Genet. 2020;54:1–19.
113. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017;14:975–8.
114. King HW, Klose RJ. The pioneer factor OCT4 requires BRG1 to functionally mature gene regulatory elements in mouse embryonic stem cells. eLife. 2017;6:e22631.
115. Treutlein B, Lee QY, Camp JG, Mall M, Koh W, Shariati SAM, et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature. 2016;534:391–5.
116. Narayanan A, Gagliardi F, Gallotti AL, Mazzoleni S, Cominelli M, Fagnocchi L, et al. The proneural gene ASCL1 governs the transcriptional subgroup affiliation in glioblastoma stem cells by directly repressing the mesenchymal gene NDRG1. Cell Death Differ. 2019;26:1813–31.
117. Park NI, Guilhamon P, Desai K, McAdam RF, Langille E, O'Connor M, et al. ASCL1 reorganizes chromatin to direct neuronal fate and suppress tumorigenicity of glioblastoma stem cells. Cell Stem Cell. 2017;21:209-224.e7.
118. Somasundaram K, Reddy SP, Vinnakota K, Britto R, Subbarayan M, Nambiar S, et al. Upregulation of ASCL1 and inhibition of Notch signaling pathway characterize progressive astrocytoma. Oncogene. 2005;24:7073–83.
119. Cevallos RR, Edwards YJK, Parant JM, Yoder BK, Hu K. Human transcription factors responsive to initial reprogramming predominantly undergo legitimate reprogramming during fibroblast conversion to iPSCs. Sci Rep-uk. 2020;10:19710.
120. Toh C-XD, Chan J-W, Chong Z-S, Wang HF, Guo HC, Satapathy S, et al. RNAi reveals phase-specific global regulators of human somatic cell reprogramming. Cell Rep. 2016;15:2597–607.
121. Koyano-Nakagawa N, Gong W, Das S, Theisen JWM, Swanholm TB, Ly DV, et al. Etv2 regulates enhancer chromatin status to initiate Shh expression in the limb bud. Nat Commun. 2022;13:4221.
122. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv.org. 2017;cs.CL. Available from: arXiv.org
123. Parmar N, Vaswani A, Uszkoreit J, Kaiser Ł, Shazeer N, Ku A, et al. Image transformer. Arxiv. 2018;
124. Held L, Ott M. On p-values and Bayes factors. Annu Rev Stat Appl. 2018;5:393–419.
125. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
126. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: identification of problematic regions of the genome. Sci Rep-uk. 2019;9:9354.
127. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31:2382–3.
128. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for comparing biological themes among gene clusters. Omics J Integr Biology. 2012;16:284–7.
129. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
130. Gong W, Dsouza N, Garry DJ. SeATAC: a tool for exploring the chromatin landscape and the role of pioneer factors. GitHub. 2023;https://github.com/gongx030/seatac.
131. Gong W, Dsouza N, Garry DJ. SeATAC: a tool for exploring the chromatin landscape and the role of pioneer factors. Zenodo. 2023;https://doi.org/10.5281/zenodo.7819334.

Publisher's Note