AtacWorks: A deep convolutional neural network toolkit for epigenomics

Avantika Lal[2,*], Zachary D. Chiang[1,*], Nikolai Yakovenko[2], Fabiana M. Duarte[1], Johnny Israeli[2†], Jason D. Buenrostro[1†]

[1]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
[2]NVIDIA Corporation, 2788 San Tomas Expy, Santa Clara, California 95051, USA.
*Equal Contributions
†Co-corresponding

Correspondence should be addressed to: jisraeli@nvidia.com or jason_buenrostro@harvard.edu.

Abstract

We introduce AtacWorks (https://github.com/clara-genomics/AtacWorks), a method to denoise and identify accessible chromatin regions from low-coverage or low-quality ATAC-seq data. AtacWorks uses a deep neural network to learn a mapping between noisy ATAC-seq data and corresponding higher-coverage or higher-quality data. To demonstrate the utility of AtacWorks, we train a model on data from four blood cell types and show that this model accurately denoises and identifies peaks from low-coverage bulk sequencing of different individuals, cell types, and experimental conditions. Further, we show that the deep learning model can be generalized to denoise low-quality data, aggregate single-cell ATAC-seq profiles, and Tn5 insertion sites for transcription factor footprinting. Finally, we apply our deep learning approach to denoise single-cell ATAC-seq data from hematopoietic stem cells to identify differentially-accessible regulatory elements between rare lineage-primed cell subpopulations.

Introduction

Within a single cell, the eukaryotic genome is hierarchically organized to form a gradient of chromatin accessibility ranging from compact, inactive heterochromatin to nucleosome-free regions that regulate the expression of genes. Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq) directly measures chromatin accessibility to quantify the relative activity of DNA regulatory elements across the genome[1]. ATAC-seq has been applied to identify the effects of transcription factors on chromatin, construct cellular regulatory networks, and localize epigenetic changes underlying diverse development and disease-associated transitions[2–4]. Recently, the development of single-cell ATAC-seq methods have made it possible to measure accessible chromatin in individual cells, enabling epigenomic analysis of rare cell types within heterogeneous tissues[5].

The resolution of localizing epigenetic changes using ATAC-seq depends on both the signal-to-noise ratio of accessible chromatin and the depth of sequencing coverage. Technical parameters such as the overall quality of cells and tissues, or the nuclei extraction method[6] can lead to over-digestion of chromatin, resulting in attenuated measurements of accessibility. Importantly, these issues are exacerbated in single-cell experiments, where primary tissues may vary in quality and key cell types may be exceedingly rare.

To address these experimental limitations, we introduce AtacWorks, a deep learning-based toolkit that takes as input a low-coverage or low-quality ATAC-seq signal, and denoises it to produce a higher-resolution or higher-quality signal. To accurately predict chromatin accessibility at each position on the genome, an AtacWorks-trained model incorporates information from a surrounding region spanning several kilobases (6 kb

for the models presented here). Based on the denoised signal, the model also predicts the genomic locations of accessible regulatory elements. We apply AtacWorks to subsampled low-coverage bulk ATAC-seq and aggregated single-cell ATAC-seq from a small number of cells. On both types of experiments, we show that AtacWorks improves the resolution of the chromatin accessibility signal and the identification of regulatory elements. Further, AtacWorks is able to improve data from cell types not present in the training set, demonstrating that our deep learning model learns generalizable features of chromatin accessibility. We also apply AtacWorks to a bulk ATAC-seq dataset with low signal-to-noise ratio, and show that our method increases enrichment of accessible chromatin. Finally, we apply AtacWorks to aggregated single-cell ATAC-seq from rare subpopulations of lineage-priming hematopoietic stem cells (HSCs) to identify epigenetic changes at key developmental regulatory elements.

Results

A deep learning framework for denoising low-coverage data

AtacWorks trains a deep neural network to learn a mapping between noisy, low-coverage or low-quality ATAC-seq data and matching high-coverage or high-quality ATAC-seq data from the same cell type. Once this mapping is learned, the trained model is applied to denoise similar low-coverage or low-quality datasets. To do this, AtacWorks uses the Resnet (residual neural network) architecture, which has been used extensively for natural image classification and localization[7]. Our model consists of multiple stacked residual blocks, each composed of three convolutional layers and a skip connection that bypasses intermediate layers (Fig. 1a). These skip connections allow propagation of the input through the layers of the network to avoid vanishing gradients[7], enabling deeper and more accurate networks to be trained. Given a noisy ATAC-seq signal track as input, the model simultaneously performs two tasks: denoising (predicting an improved signal track) and peak calling (predicting the genomic location of accessible regulatory elements).

Using AtacWorks, we trained deep learning models with bulk ATAC-seq data from FACS-isolated human blood cells[2]. We took ATAC-seq datasets from 4 cell types (B cells, NK cells, $CD4^+$ and $CD8^+$ T cells) and sampled each to a depth of 50 million reads (25 million read pairs) to produce ‘clean’, high-coverage data. Peaks for each clean dataset were identified using MACS2 (see Methods). We then subsampled each clean dataset to multiple lower sequencing depths ranging from 0.2 million to 20 million reads (Supplementary Fig. 1). For each depth, we trained a model to take as input the low-coverage ATAC-seq signal and reconstruct both the clean ATAC-seq signal and peaks.

We tested the performance of these models on ATAC-seq data from erythroid cells (erythroblast)[2], which were not included in the training set. We subsampled reads from erythroid cells to the same depths as the training data. For each sequencing depth, we then applied the trained deep learning model to the corresponding subsampled erythroid cell dataset to obtain a predicted high-coverage signal track and peak calls. We evaluated the predicted high-coverage signal tracks produced by AtacWorks by comparing them to a clean (50 million read) erythroid cell signal. We found that at all sequencing depths, both the Pearson correlation and MSE (Mean Squared Error) between the predicted and clean signal tracks were substantially greater than that between the noisy and clean signal (Fig. 1b, Supplementary Table 1, Supplementary Fig. 2). We also found that our method presents significantly higher performance than smoothing using linear regression (Supplementary Table 2).

Next, we evaluated the peak calls produced by AtacWorks from each sequencing depth, and found that both the AUPRC and AUROC of calls were superior to MACS2[8] calls from the same subsampled data (Fig. 1c, Supplementary Table 1, Supplementary Fig. 2). Overall in this analysis, AtacWorks produced output data of

quality equivalent to (on average) 2.6x the number of reads in the input data based on Pearson correlation, and 4.2x the number of reads in the input data based on AUPRC (Supplementary Table 1).

To show that the model is not simply recapitulating the training set, we calculated performance metrics on a held-out chromosome not used for training (chromosome 10), and obtained highly similar results to those computed on the whole genome (Fig. 1b and 1c, Supplementary Table 1). Further, we also found that the results were highly robust to the training data used (Supplementary Table 3). We also evaluated the model only on differential peaks present in either the training or test set. We find AtacWorks improves both the signal track fidelity and peak calling in these regions (Supplementary Table 1); we also confirmed that AtacWorks identifies cell-type-specific peaks not present in the training data (Fig. 1b).

Finally, we obtained similar results by applying this trained model to low-coverage ATAC-seq data from additional cell types (Monocytes; Supplementary Table 4 and Human Peyer's Patch data from ENCODE; Supplementary Table 5, Supplementary Fig. 3).

AtacWorks generalizes to diverse applications

In addition to low-coverage data, AtacWorks can improve signal quality in ATAC-seq datasets with low signal-to-noise ratio. To demonstrate this, we trained AtacWorks on paired high and low quality ATAC-seq datasets from FACS-isolated human monocytes[2] (Supplementary Table 7, Methods). Both datasets had similar sequencing depth (approximately 20 million reads); however, one had higher signal-to-noise ratio measured by the enrichment of insertions at transcription start sites (TSSs). We trained AtacWorks to learn a mapping from the low-quality signal to the high-quality signal. We then applied this trained model to denoise low-quality bulk ATAC-seq data of similar depth from Erythroid cells. AtacWorks reduced background noise (Fig. 1f) and improved the enrichment at TSSs (Fig. 1e), producing a signal track and peak calls more similar to those obtained from the higher-quality data (Supplementary Table 8, Fig. 1f).

AtacWorks can also be applied to enhance the aggregated ATAC-seq signal from small populations of cells in a single-cell ATAC-seq experiment. To demonstrate this, we obtained single-cell ATAC-seq data from human blood cells, sequenced using the dsci-ATAC-seq protocol[9]. We randomly selected 90 cells (~200,000 reads) or 450 cells (~1 million reads) of the same type to obtain noisy data, and applied AtacWorks for denoising and peak calling. We showed that both a model trained on bulk ATAC-seq data with a similar number of reads, and a model trained on different cell types in the same single-cell dataset, were effective at denoising and peak calling from low numbers of cells (Supplementary Tables 9 and 10, Supplementary Fig. 4, Methods).

AtacWorks performs denoising at single-base pair resolution, and is therefore adaptable for transcription factor footprinting. "Footprinting" leverages the fact that transcription factor-bound DNA is inaccessible in order to identify characteristic insertion signatures and predict binding across the genome. However, these approaches traditionally require over 100 million reads[10], prohibiting their use on data from less abundant cell types. We trained models using high-coverage (100 million reads) ATAC-seq data from FACS-sorted NK cells[2], downsampled to a range of lower sequencing depths (0.2 - 70 million reads), and tested them on ATAC-seq data from HSCs downsampled to the same sequencing depths. We used signal tracks of a higher resolution (see Methods), in order to identify transcription factor-specific patterns of Tn5 insertion frequency across the genome. At each sequencing depth, AtacWorks improved the signal track for HSCs (Supplementary Table 6), producing a high-resolution denoised track in which the characteristic footprint spanning the CTCF motif is clearer than in the low-coverage input (Fig. 1g, Supplementary Fig. 5).

Deep learning enhances the resolution of single-cell studies

In order to demonstrate the biological applications of our method, we sought to denoise low-coverage single-cell ATAC-seq signal from cell types that are present at very low frequencies and are not possible to experimentally isolate. Previous single-cell studies of FACS-isolated bone marrow mononuclear cells (BMMCs) have observed epigenetic variability within immunophenotypically-defined cellular populations, suggesting that many hematopoietic stem and progenitor cells lie along a continuum of differentiation states (Fig. 2a)[11,12]. Though it is possible to quantify differences in transcription factor motif accessibility across the entire population[13], these cell states are so granular that there is often not enough sequencing coverage to pinpoint when specific regulatory elements are activated or repressed throughout differentiation. Of particular importance, HSCs are thought to include rare subpopulations of cells primed towards lymphoid, myeloid and erythroid lineages[11,14,15]; however, analysis of the underlying changes of individual regulatory elements at single-cell resolution has not yet been possible.

We reasoned we could identify epigenetic signatures of lineage primed HSCs using AtacWorks. To do this, we generated dscATAC-seq data from FACS-isolated HSCs (see Methods) and collected published bead-enriched $CD34^+$ data[9]. To jointly analyze these single-cell data, we used a bulk reference-guided approach (see Methods) to project these cells into a shared latent space, and visualized them using UMAP for dimensionality reduction (Fig. 2b). This analysis localized FACS-isolated HSCs to the top of the hierarchy. We also confirmed that HSCs localized in this region exhibited variability of sequence motifs corresponding to GATA2 (Fig. 2c) and MESP1 (Fig. 2d), previously implicated as TF motifs correlated with epigenetic lineage priming[12].

To generate low-coverage lineage-primed epigenetic profiles for our model, we selected three distant FACS-isolated HSC profiles and selected the 50 most similar bead-enriched $CD34^+$ cells for each. For each subsample of 50 aggregated cells, we performed signal denoising using AtacWorks and visualized the denoised aggregate chromatin accessibility profiles surrounding genes known to be involved in hematopoietic differentiation[12] (Fig. 2f and Supplementary Fig. 6). Using this approach, we observed considerable chromatin accessibility differences among HSC subpopulations. Though the function of these differential peaks have yet to be experimentally validated, these results demonstrate the unique capacity of the deep learning to reveal novel biological insights from sparse single-cell ATAC-seq data.

Discussion

ATAC-seq has become an increasingly valuable tool enabling high-resolution characterization of the epigenome, providing insights into mechanisms underlying gene expression changes associated with development, evolution, and disease. However, technical limitations in tissue quality, assay performance, and sequencing coverage constrain our ability to measure the diverse chromatin states represented across the genome. These limitations also pertain to emerging single-cell ATAC-seq methods, as cell types of interest are often difficult to experimentally isolate and present at low frequencies in heterogeneous contexts.

Deep learning methods present as a powerful tool to address these limitations; deep learning models have been widely successful for denoising speech[16], and for inpainting (filling missing data) in images[17,18]. An earlier study[19] demonstrated that simple convolutional neural networks are successful at denoising or peak calling from ChIP-seq data. However, this method (Coda) is optimized for broad peak calling from ChIP-seq of histone modifications averages over 25 base pair windows of the genome, precluding its application to single base-pair maps of chromatin accessibility. Here we present AtacWorks, a toolkit to train and apply deep learning models that improve both bulk and single-cell ATAC-seq data. Our model is based on resnet architectures that have shown great success in image classification. AtacWorks produces results at single-base pair resolution, enabling its use in TF footprinting, and uses the same model to perform both signal denoising and peak calling.

We demonstrate that AtacWorks improves the quality of noisy ATAC-seq data, and also improves the accuracy of open chromatin identification. We do not use the DNA sequence as a model input, and therefore the model does not depend on cell-type specific sequence motifs. Moreover, the model that we present here is applicable across cell and tissue types, individuals, and even across experimental datasets and protocols. As such, we anticipate that this deep learning framework will also be applicable to other high-resolution functional genomics experiments such as CUT&RUN[20]. Further, we show that this model can be used to reduce experimental noise, to identify transcription factor binding at low coverage, to identify open chromatin in small populations of cells, and identify separate lineages in complex mixtures of cells.

Lastly, we apply AtacWorks to denoise aggregate single-cell chromatin accessibility profiles from rare subpopulations of hematopoietic stem cells. We demonstrate that AtacWorks can denoise aggregate chromatin accessibility tracks from as few as 50 cells to identify regulatory elements specific to these subpopulations that might be otherwise missed due to the sparsity of the data. Overall, we anticipate that AtacWorks will be a synergistic approach to enhance the quantity and quality of biological information generated by single-cell ATAC-seq workflows.

Figures:

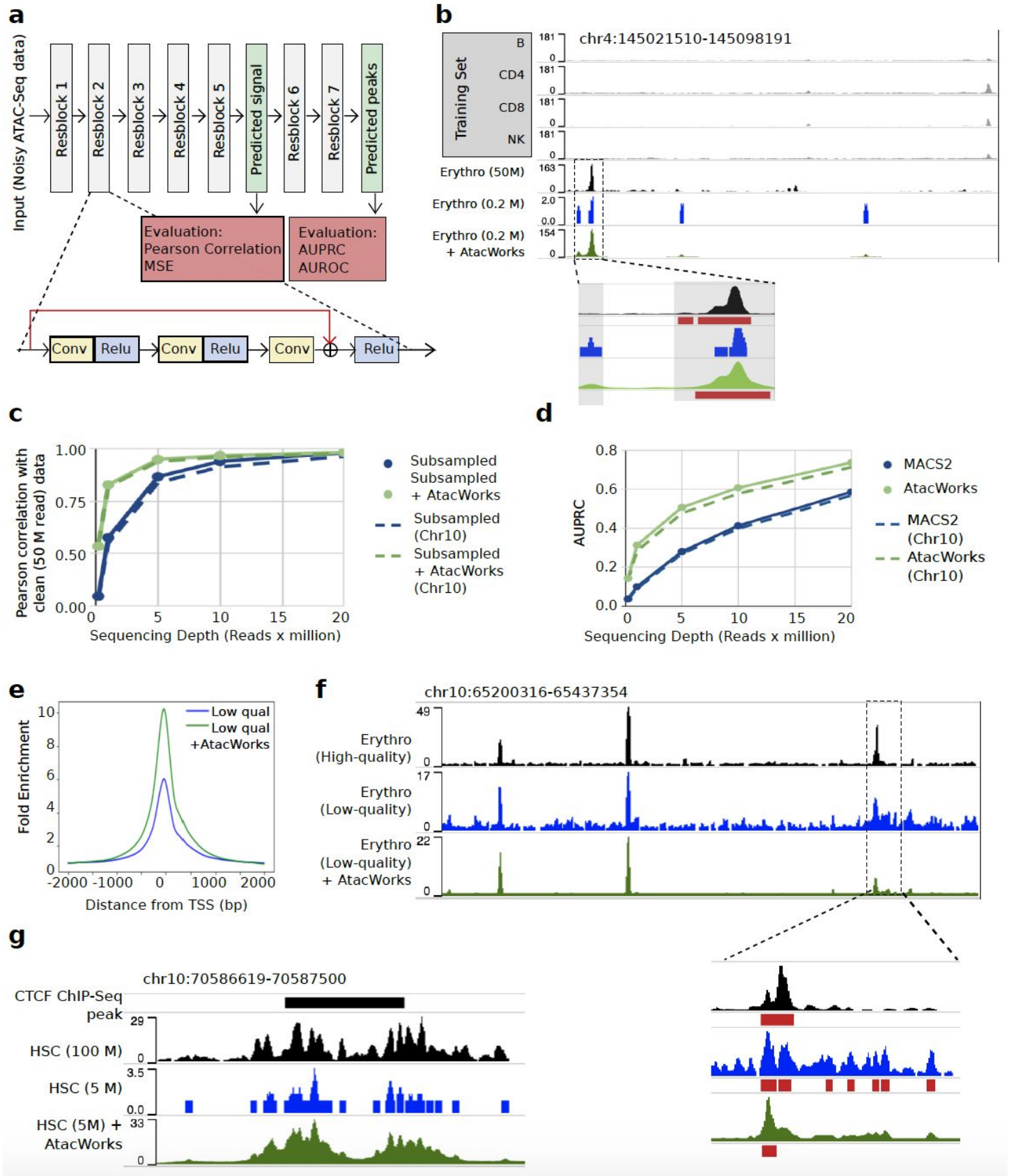

Figure 1. A deep learning approach to denoise ATAC-seq data.

A. Schematic of the resnet architecture used in AtacWorks, as well as a residual block composed of 1-dimensional convolutional layers ('Conv') and ReLU activation functions ('ReLU').

B. Pearson correlation between a clean ATAC-seq signal track (50 million reads) and subsampled data for Erythroid cells, before (blue) and after (green) denoising with an AtacWorks model trained on 4 other cell types. The X-axis shows the number of reads in the subsampled data. Whole lines show metrics over the whole genome while dotted lines show metrics over chromosome 10.

C. AUPRC for MACS2 (blue) and AtacWorks (green) showing their performance in peak calling on the subsampled data, using peaks called by MACS2 subcommands on the clean (50 million reads) signal track as truth. Whole lines show metrics over the whole genome while dotted lines show metrics over chromosome 10.

D. Signal tracks for the 4 cell types used for training the AtacWorks model, Erythroid cells (at a depth of 50 million reads), Erythroid cells (subsampled to 0.2 million reads), and Erythroid cells (subsampled to 0.2 million reads) after denoising by AtacWorks. ATAC-seq signal is shown for a region around the *GYPA* gene on chromosome 4. Red bars below the tracks show peak calls by MACS2 (for the 50 M and 0.2 M read tracks) and AtacWorks (for the denoised track)

E. Average ATAC-seq signal over 4000-base windows centered on transcription start sites, in low-quality ATAC-seq data from erythroid cells, before (blue) and after (green) denoising with AtacWorks.

F. Signal tracks for high-quality ATAC-seq data from Erythroid cells (high enrichment at TSS), low-quality ATAC-seq data from Erythroid cells (low enrichment at TSS), and low-quality ATAC-seq data from Erythroid cells after denoising by AtacWorks. ATAC-seq signal is shown for a region on chromosome 10. Red bars below the tracks show peak calls by MACS2 (for the upper two tracks) and AtacWorks (for the denoised track)

G. Signal tracks around a single CTCF binding site, from HSCs (at a depth of 100 million reads), HSCs (subsampled to 5 million reads), and HSCs (subsampled to 5 million reads) after denoising by AtacWorks. The upper black bar shows the CTCF binding site.

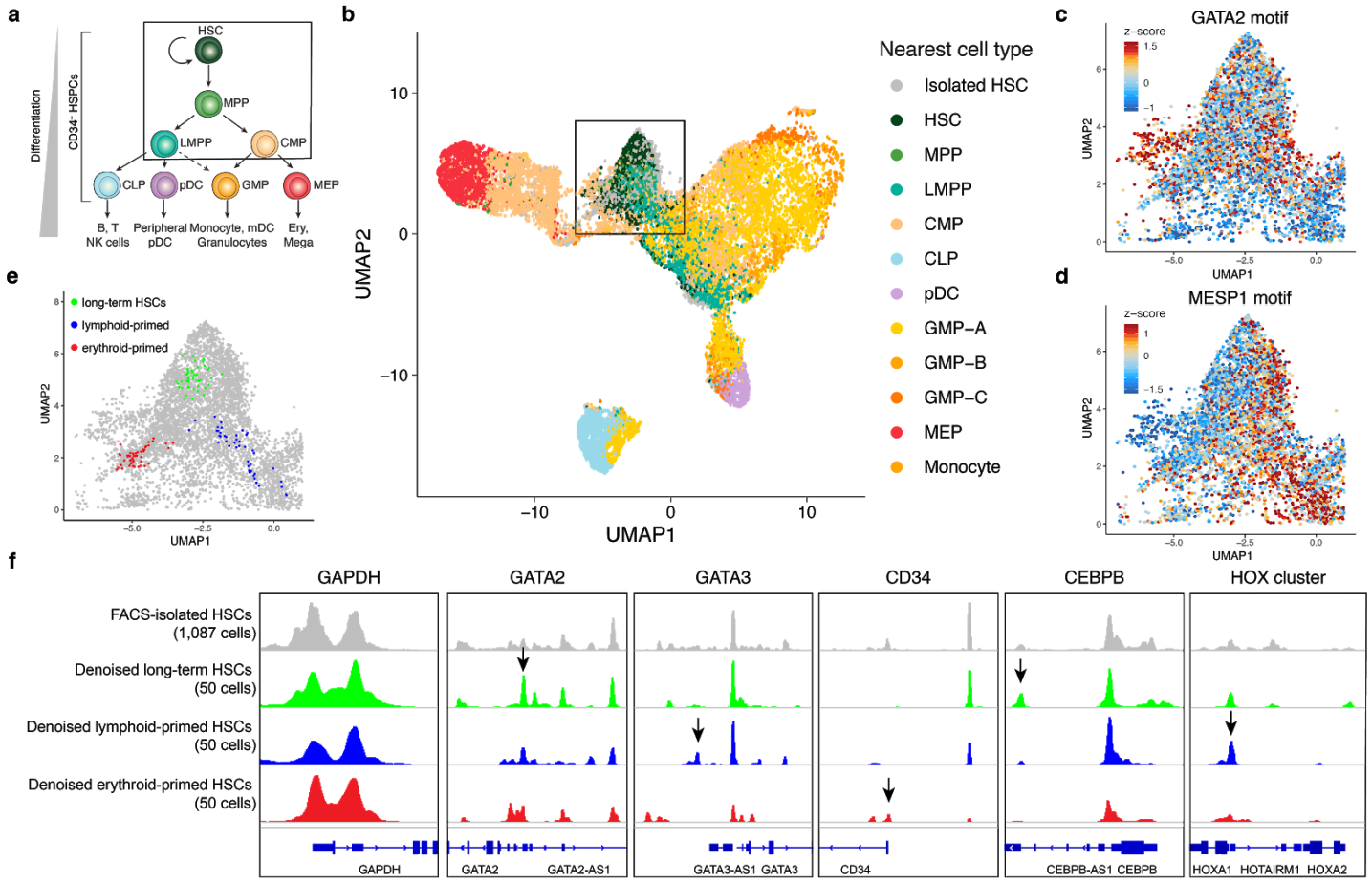

Figure 2. AtacWorks identifies regulatory elements associated with lineage primed HSCs.

A. A schematic of the classical hierarchy of human hematopoietic differentiation.

B. A UMAP dimensionality reduction of single-cell ATAC-seq profiles from bead-enriched CD34$^+$ bone marrow progenitor cells (ref) (*n* = 28,505) and hematopoietic stem cells (HSCs) isolated via fluorescence-activated cell sorting (*n* = 1,087). The bead-enriched CD34$^+$ cells are colored by the most correlated cell type from a FACS-isolated single-cell ATAC-seq reference[12]. The box indicates the region containing the majority of the isolated HSCs.

C,D. Single cells are colored by chromVAR transcription factor motif accessibility z-scores (ref), for (C) GATA2 and (D) MESP1.

E. Three 50-cell subsamples, each generated by selecting a single HSC and identifying the 50 most similar cells from the CD34$^+$ experiment.

F. Aggregate chromatin accessibility profiles from isolated HSCs (*n* = ~1,087) and three denoised subsamples of CD34$^+$ cells (*n* = 50 each) surrounding genes known to be involved in human hematopoietic differentiation[12].

Methods

Input data for AtacWorks

AtacWorks models can be trained using one or more pairs of matching ATAC-seq datasets. Each pair consists of two ATAC-seq datasets from the same biological sample or cell type, a 'clean' dataset of high sequencing coverage or quality, and a second, 'noisy' dataset of lower coverage or lower quality. A low-coverage dataset can be generated computationally, by randomly subsampling a fraction of reads or cells from the high-coverage dataset, or experimentally, by carrying out an identical ATAC-seq experiment using fewer cells or lower sequencing depth.

The model requires three specific inputs for each such pair of datasets:

1. a signal track representing the number of sequencing reads mapped to each position on the genome in the noisy dataset.
2. a signal track representing the number of sequencing reads mapped to each position on the genome in the clean dataset.
3. The genomic positions of peaks called on the clean dataset. These can be obtained by using MACS2 or any other peak caller.

The model learns a mapping from (1) to both (2) and (3); in other words, from the noisy signal track, it learns to predict both the clean signal track, and the positions of peaks in the clean dataset.

AtacWorks accepts input data in standard genomic formats: bigWig files for signal tracks, and BED or narrowPeak files for peak positions.

Processing data into genomic intervals

The input dataset is first divided into training, validation and holdout sets each comprising selected chromosomes. These are then divided into non-overlapping intervals. Each interval represents a single training example.

Before feeding these intervals into the neural network, we also add data for additional bases at either end of each interval. This is done because our model uses convolutional filters which predict a value at each base using information from the neighboring bases on either side. Without adding additional bases at the ends of each interval, the model would lack information to make accurate predictions for points at the edges of the interval. However, the model does not output predictions for these additional bases.

The intervals are passed to the deep learning model in batches of a fixed size. The order of intervals is shuffled after every epoch of training.

Deep Learning Model in AtacWorks

AtacWorks uses a Resnet (residual neural network) model consisting of multiple stacked residual blocks. Each residual block includes three convolutional layers; the input to the first layer is added to the output of the third layer through a skip connection (Fig. 1a).

For each position in the given interval, the model performs two tasks; a regression or denoising task (predicting the ATAC-seq signal at each position in the clean dataset) and a classification or peak calling task (predicting the likelihood that each position is part of a peak).

In order to perform both tasks, the input is passed through several residual blocks, followed by a regression output layer that returns the predicted ATAC-seq signal at each position in the input. The regression output is then passed through another series of residual blocks followed by a classification output layer that returns a prediction for whether each base in the input is part of a peak.

We train our model in the PyTorch neural network framework [21], utilizing the Adam optimizer [22]. We optimize the denoising task with a regression loss, and the peak classification task with a binary cross entropy (BCE) loss (details below).

We use the rectified linear unit (ReLU) activation function throughout the network, except for the classification output layer, which uses a sigmoid activation function. The sigmoid activation is necessary for our network to return a value between 0 and 1 for each input base, which is interpreted as the probability of that base being part of a peak.

Our model does not utilize batch normalization [23] for the convolutional layers, as these did not improve accuracy on either the regression or classification tasks in our experiments. We include the option to use dilated convolutional layers throughout the model to increase the receptive field of the model without increasing the parameter count. This approach has been effective in image classification tasks where a larger receptive field is desirable [24].

We experimented with other convolutional neural network architectures, including the U-net [25], and chose this architecture based on its robust performance in both denoising and peak calling tasks on several datasets.

Model Training in AtacWorks

The deep learning model is trained using a multi-part loss function, comprising a weighted sum of three individual loss functions:

1. Mean squared error (for the regression output)
2. 1 - Pearson correlation coefficient (for the regression output)
3. Binary cross-entropy (for the classification output)

The relative importance of these loss functions can be tuned by assigning different weights to each.

To train the model, one or more chromosomes of each input dataset is held out as a validation set. At the end of each epoch of training, the performance on the validation set is measured. The model with the best validation set performance is saved and used.

Model Evaluation in AtacWorks

The performance of the model is measured using the following metrics:

For regression (denoising):

1. Pearson correlation
2. Mean Squared Error

For classification (peak calling):

1. Area under the Precision-Recall Curve (AUPRC)
2. Area under the Receiver Operating Characteristic (AUROC)

Our model outputs the probability of belonging to a peak, for each position in the genome. In order to obtain predicted peaks, we must set a probability threshold above which a base is said to be a peak. Similarly, MACS2 produces a p-value for each position and the final peak calls depend on a user-defined probability threshold. Therefore we have chosen to use the AUPRC and AUROC metrics which evaluate the performance of the peak calling methods over the entire range of possible thresholds.

Data output

AtacWorks produces results in the form of bigWig and BED files which can be visualized using a genome browser.

Code availability

AtacWorks is available at https://github.com/clara-genomics/AtacWorks.

Data sources

Bulk ATAC-seq data for various blood cell types was obtained from [2]. We used B cells, NK cells, CD4+ and CD8+ T cells for training, and Erythroid cells and Monocytes for testing. For the experiment on transcription factor footprinting, we used NK cells for training and HSCs for testing.

We downloaded publicly available ATAC-seq data for the human Peyers Patch from ENCODE (https://www.encodeproject.org/experiments/ENCSR017RQC/). We also downloaded ChIP-Seq results for CTCF binding in HSCs (https://www.encodeproject.org/files/ENCFF344RYC/)

dsci-ATAC data for blood cells was obtained from [9]. We used data from CD4+ T cells, CD8+ T cells, pre-B cells and Monocytes for experiments as described below.

dscATAC-seq data for bead-enriched $CD34^+$ cells was obtained from [9].
scATAC-seq data for FACS-isolated PBMCs was obtained from [12].
ATAC-seq data for bulk-guided clustering was also obtained from [2].

A list of transcription start sites for hg19 were obtained from the UCSC Genome Browser.

Generation of single-cell ATAC-seq data for FACS-isolated HSCs

Cryopreserved human bone marrow mononuclear cells were purchased from Allcells (catalog number BM, CR, MNC, 10M). Cells were quickly thawed in a 37°C water bath, rinsed with culture medium (RPMI 1640 medium supplemented with 15% FBS) and then treated with 0.2 U/μL DNase I (Thermo Fisher Scientific) in 2 mL of culture medium at room temperature for 15 min. After DNase I treatment, cells were filtered with a 40 μm cell strainer, washed with MACS buffer (1x PBS, 2 mM EDTA and 0.5% BSA), and cell viability and concentration were measured with trypan blue on the TC20 Automated Cell Counter (Bio-Rad). Cell viability was greater than 90% for all samples. CD34 positive cells were then bead enriched using the CD34 MicroBead Kit UltraPure (Miltenyi Biotec, catalog number 130-100-453) following manufacturer's instructions. The enriched population was then simultaneously stained with CD45, Lineage cocktail, CD34, CD38, CD45RA and CD90 antibodies in MACS buffer for 20 min at 4°C. Stained cells were then washed with MACS buffer and the CD45+ Lin- CD38-

CD34+ CD45RA- CD90+ fraction (HSCs) was FACS sorted using a MoFlo Astrios Cell Sorter (Beckman Coulter). Single-cell ATAC-seq data was then generated for the sorted HSCs using the dscATAC-seq Whole Cell protocol as previously described[9].

Data preparation for experiments

BAM files for bulk ATAC-seq were downsampled to a fixed number of reads using SAMtools v.1.9 [26]. For paired-end sequencing data, read pairs are kept intact, so, for example, 100,000 read pairs are selected to obtain a total of 200,000 sequencing reads.

For single-cell ATAC-seq experiments, a number of cells of the chosen cell type were randomly selected and all reads from those cells were extracted from the BAM file.

To identify the exact location of Tn5 insertions with base pair resolution, each ATAC-seq read was converted to a single genomic position corresponding to the first base pair of the read. Previous work has demonstrated that the Tn5 transposase inserts adapters separated by 9 bp, so reads aligning to the + strand were offset by +4 bp, while reads aligning to the - strand were offset by -5 bp. Each cut site location was extended by 100 bp in either direction, except for transcription factor footprinting experiments where each cut site was extended by 5 bp in either direction. The bedtools genomecov function[27] was then used to convert the list of locations into a genome coverage track containing the ATAC-seq signal at each genomic position.

To call peaks from clean and noisy signal tracks, MACS2 subcommands bdgcmp and bdgpeakcall were run with the ppois parameter and a -log10(p-value) cutoff of 3. BED files with equal coverage over all chromosomes were provided as a control input track.

In all our experiments, we used chromosome 20 of each ATAC-seq dataset for validation and held out chromosome 10 of each ATAC-seq dataset for testing. The rest of the autosomes were used for training. Each chromosome was split into non-overlapping genomic intervals of length 50,000 bases.

Generation of paired high and low quality tracks

Paired high and low quality chromatin accessibility tracks were computationally generated from the same experiment in order to minimize the impact of potential batch effects. Previously published bulk ATAC-seq tracks from monocytes and erythroid cells[2] were split by donor and biological replicate, and then quantified using a TSS enrichment score. Tracks were then visually classified as high or low enrichment, and then aggregated based on classification and cell type to form the paired high and low quality tracks (Supplementary Table 7). We confirmed previous findings that biological variability across samples from the same cell type, but different donors was minimal.

Parameters used for training AtacWorks models

The hyperparameters to use for the final model were chosen based on validation set performance. We note that deeper and larger models can produce slightly better results; however, larger models are also expensive and time-consuming to train.

Number of residual blocks: 7 total (5 followed by regression output, then 2 more followed by classification output)
Kernel size: 51 (for all convolutional layers)
Dilation: 8 (for all convolutional layers)

Number of channels: 15 (for all convolutional layers)
Batch size: 64
Metric to select best model: AUROC

Weights for loss functions:
Mean squared error: 0.0005 (except for transcription factor footprinting experiments, where this was set to 0.05)
1 - Pearson correlation coefficient: 1
Binary cross-entropy: 1

The models were trained using Pytorch version 1.2.0 in Python version 3.5.2. All models were trained with a constant learning rate of 2×10^{-4} for 25 epochs.

Unless otherwise specified, a probability cutoff of 0.5 was used to call peaks from the probability values produced by AtacWorks.

AtacWorks took 2.7 minutes per epoch to train on one ATAC-seq dataset, and 7 minutes to test on a different whole genome, using 8 Tesla V100 16GB GPUs in an Nvidia DGX-1 server.

dsci-ATAC experiments

We performed two experiments to validate AtacWorks using dsci-ATAC data[9].
In the first experiment (Supplementary Table 9), we took a cluster of 20378 CD4+ T cells (~ 50 million reads). We randomly selected 90 CD4+ T cells (~200,000 reads) or 450 CD4+ T cells (~1 million reads), and applied a model trained on bulk ATAC-seq data from 4 cell types with a similar number of reads. Since CD4+ T cells were also among the cell types used to train the model, we evaluated the results only on chromosome 10, which was not used in training.

In the second experiment (Supplementary Table 10), we selected 4 cell types from the dsci-ATAC dataset - CD4+ T cells, CD8+ T cells, pre-B cells, and Monocytes. We randomly sampled 6000 cells of each type to obtain 'clean' data, and further subsampled 90 and 450 cells of each type to obtain noisy data. 3 cell types - CD4+ T cells, CD8+ T cells, and pre-B cells - were used for training a deep learning model. This model was tested on noisy data of the fourth cell type (Monocytes).

Data visualization

We used the WashU epigenome browser (http://epigenomegateway.wustl.edu/browser/) for visualization. The HSC denoised subsamples (Fig. 2f) were visualized using the Integrative Genomics Viewer [28].

Bulk-guided projections

The bulk-guided UMAP projection of single cells (Fig. 2B) was performed as previously described. In brief, a common set of peaks (k = 156,311) was used to create a vector of read counts for each single-cell profile. Principle-component analysis (PCA) was run on previously-published bulk ATAC-seq data ([2]) to generate eigenvectors capturing variations in regulatory element accessibility across cell types. Each single cell was then projected in the same space as these eigenvectors by multiplying their counts vector by the common PCA loading coefficients. The resulting projection scores were scaled and centered prior to being visualized using

UMAP. Predicted labels for the CD34$^+$ cells were derived by correlating their projected single-cell scores with those of a reference set of FACS-isolated PBMCs [12] and assigning the label of the closest match.

Transcription factor motif accessibility z-scores

Motif accessibility z-scores (Fig. 2b-d) were computed using chromVAR[13]. The method calculates gain or loss in accessibility within peaks that share a common transcription factor motif while adjusting for GC content and overall region accessibility. The single cells were scored using a provided file called "human_pwms_v1" that is a list of human transcription factor motifs curated from the CIS-BP database.

Denoising CD34$^+$ subsamples

Three CD34$^+$ subsamples were generated by selecting a single HSC and identifying the 50 most similar single-cell profiles from the CD34$^+$ bead-enriched sample. These subsamples were converted from BAM to bigWig format as described earlier, and then denoised using a model trained on bulk ATAC-seq from 4 human blood cell types, not including HSCs.

Acknowledgments

We thank Eric Xu and Joyjit Daw for contributing to the code for AtacWorks. We thank members of the Buenrostro lab and NVIDIA team for insightful comments throughout the development of this work. We thank Ronald Lebofsky and Giulia Schiroli for assistance in generating dscATAC-seq data. J.D.B., Z.D.C., and F.M.D. acknowledge support by the Allen Distinguished Investigator Program through the Paul G. Allen Frontiers Group. This work was further supported by the Chan Zuckerberg Initiative and the NIH Director's New Innovator award. Z.D.C. is supported by the Quantitative Biology Initiative at Harvard University.

References

1. 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).
2. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* **48**, 1193–1203 (2016).
3. Yoshida, H. *et al.* The cis-Regulatory Atlas of the Mouse Immune System. *Cell* **176**, 897–912.e20 (2019).
4. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).
5. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
6. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* **14**, 959–962 (2017).

7. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).
8. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
9. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
10. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
11. Yu, V. W. C. *et al.* Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell* **168**, 944–945 (2017).
12. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535–1548.e16 (2018).
13. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
14. Rodriguez-Fraticelli, A. E. *et al.* Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).
15. Pei, W. *et al.* Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).
16. Pascual, S., Bonafonte, A. & Serrà, J. SEGAN: Speech Enhancement Generative Adversarial Network. *arXiv [cs.LG]* (2017).
17. Yang, C. *et al.* High-Resolution Image Inpainting using Multi-Scale Neural Patch Synthesis. *arXiv [cs.CV]* (2016).
18. Liu, G. *et al.* Image Inpainting for Irregular Holes Using Partial Convolutions. *arXiv [cs.CV]* (2018).
19. Koh, P. W., Pierson, E. & Kundaje, A. Denoising genome-wide histone ChIP-seq with convolutional neural networks. *Bioinformatics* **33**, i225–i233 (2017).
20. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, (2017).

21. Adam, P. *et al.* Automatic differentiation in pytorch. in *Proceedings of Neural Information Processing Systems* (2017).

22. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).

23. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv [cs.LG]* (2015).

24. Kudo, Y. & Aoki, Y. Dilated convolutions for image classification and object localization. in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)* 452–455 (2017).

25. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science* 234–241 (2015). doi:10.1007/978-3-319-24574-4_28

26. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

27. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

28. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

Supplementary Materials

Supplementary Table 1: Performance of AtacWorks on bulk ATAC-seq data from erythroid cells. The model used here is a resnet model trained on bulk ATAC-seq data from 4 blood cell types (CD4+ T cells, CD8+ T cells, B cells, and NK cells). All metrics were calculated separately on the whole genome, on chromosome 10 (not used for training the model), and on differential peaks (peaks present only in either the training data or the test data).

Supplementary Table 2: Performance of AtacWorks on bulk ATAC-seq data from erythroid cells. The resnet model used to generate Supplementary Table 1 is compared against linear regression for denoising. Both models were trained on bulk ATAC-seq data from the same 4 cell types. All metrics were calculated separately on the whole genome and on chromosome 10 (not used for training the model).

Supplementary Table 3: Performance of AtacWorks on bulk ATAC-seq data from erythroid cells. Two models are compared: a resnet model trained on bulk ATAC-seq data from 4 cell types (same as Supplementary Table 1), and a resnet model trained on bulk ATAC-seq data from only 1 cell type (CD4+ T cells). All metrics were calculated separately on the whole genome and on chromosome 10 (not used for training the model).

Supplementary Table 4: Performance of AtacWorks on bulk ATAC-seq data from Monocytes. The model used here is the same as in Supplementary Table 1. All metrics were calculated separately on the whole genome, on chromosome 10 (not used for training the model).

Supplementary Table 5: Performance of AtacWorks on ENCODE bulk ATAC-seq data from the human Peyer's Patch. The resnet model used here is the same as in Supplementary Table 1. All metrics were calculated separately on the whole genome, on chromosome 10 (not used for training the model).

Supplementary Table 6: Performance of AtacWorks on high-resolution bulk ATAC-seq data from HSCs (Hematopoietic Stem Cells). A resnet model was trained on high-resolution bulk ATAC-seq data from NK (Natural Killer) cells. All metrics were calculated separately on the whole genome and on chromosome 10 (not used for training the model).

Supplementary Table 7: Breakdown of how monocyte and erythroid ATAC-seq tracks (^{2}) were aggregated across donors and replicates to create paired high and low quality training data.

Supplementary Table 8: Performance of AtacWorks on low-quality bulk ATAC-seq signal from Erythroid cells. A resnet model was trained on paired low- and high-quality ATAC-seq data from Monocytes. All metrics were calculated separately on the whole genome and on chromosome 10 (not used for training the model).

Supplementary Table 9: Performance of AtacWorks on single-cell ATAC-seq (sci-ATAC) data from CD4+ T cells. The model used here is the same as in Supplementary Table 1. All metrics were calculated on chromosome 10 (not used for training the model).

Supplementary Table 10: Performance of AtacWorks on single-cell ATAC-seq (sci-ATAC) data from Monocytes. A resnet model was trained on sci-ATAC data from 3 other cell types (CD4+ T cells, CD8+ T cells, and pre-B cells). All metrics were calculated separately on the whole genome, on chromosome 10 (not used for training the model), and on differential peaks (peaks present only in either the training data or the test data).

Supplementary Figure 1: Clean (black) and noisy (blue) ATAC-seq signal tracks for the 4 ATAC-seq datasets (CD4+ T cells, CD8+ T cells, B cells and NK cells) used for training the deep learning model.

Supplementary Figure 2: Clean (black), noisy (blue) and denoised (green) ATAC-seq signal tracks for bulk ATAC-seq data from Erythroid cells. Detailed views of two peaks are shown. Below the noisy (blue) signal tracks, the heatmaps show the negative log of the p-value returned by MACS2 for each position and the red bars show peak calls by MACS2 using a p-value cutoff of 10^{-3}. Below the denoised (green) signal tracks, the heatmaps show the probability returned by AtacWorks (representing the probability that each position is part of a peak) and the red bars show peak calls by AtacWorks using a probability cutoff of 0.5.

Supplementary Figure 3: Clean (black), noisy (blue) and denoised (green) ATAC-seq signal tracks for ENCODE bulk ATAC-seq data from the Peyer's Patch. Detailed views of two regions are shown. Below the noisy (blue) signal tracks, the heatmaps show the negative log of the p-value returned by MACS2 for each position and the red bars show peak calls by MACS2 using a p-value cutoff of 10^{-3}. Below the denoised (green) signal tracks, the heatmaps show the probability returned by AtacWorks (representing the probability that each position is part of a peak) and the red bars show peak calls by AtacWorks using a probability cutoff of 0.5.

Supplementary Figure 4: Clean (black), noisy (blue) and denoised (green) ATAC-seq signal tracks for single-cell ATAC-seq (sci-ATAC) data from CD4+ T cells. A detailed view of one peak is shown. Below the noisy (blue) signal track, the heatmap shows the negative log of the p-value returned by MACS2 for each position and the red bar shows peak calls by MACS2 using a p-value cutoff of 10^{-3}. Below the denoised (green) signal tracks, the heatmap shows the probability returned by AtacWorks (representing the probability that each position is part of a peak) and the red bar shows peak calls by AtacWorks using a probability cutoff of 0.5.

Supplementary Figure 5: Heatmaps showing the signal at 10,000 genomic regions surrounding CTCF motifs (rows) in clean (100 million read), noisy (downsampled to 5 million reads) and denoised signals.

Supplementary Figure 6: Noisy and denoised ATAC-seq signal tracks for the three aggregated subsamples from Fig. 2. Each subsample was generated by selecting a single HSC and identifying the 50 most similar cells from the $CD34^{+}$ experiment. A track for the aggregated FACS-isolated HSCs is also provided for reference. Selected regions contain genes that are known to be involved in human hematopoietic differentiation.