HMMRATAC, The Hidden Markov Models for ATAC-seq.

Evan D. Tarbell and Tao Liu

Abstract

ATAC-seq has been widely adopted to identify accessible chromatin regions across the genome. However, current data analysis still utilizes approaches originally designed for ChIP-seq or DNase-seq, without taking into account the transposase digested DNA fragments that contain additional nucleosome positioning information. We present the first dedicated ATAC-seq analysis tool, a semi-supervised machine learning approach named HMMRATAC. HMMRATAC splits a single ATAC-seq dataset into nucleosome-free and nucleosome-enriched signals, learns the unique chromatin structure around accessible regions, and then predicts accessible regions across the entire genome. We show that HMMRATAC outperforms the popular peak-calling algorithms on published human and mouse ATAC-seq datasets.

Keywords

ATAC-seq; peak-calling; Hidden Markov Model; nucleosome-free regions; flanking nucleosomes; decomposition and integration; machine learning; chromatin states

Introduction
The genomes of all known eukaryotes are packaged into a nucleoprotein complex called chromatin. The nucleosome is the fundamental, repeating unit of chromatin, consisting of approximately 147 base pairs of DNA wrapped around an octet of histone proteins [1]. The eviction of nucleosomes into nucleosome-free regions (NFRs) makes DNA more accessible to various DNA binding factors. The binding of these factors to the accessible DNA can exert spatiotemporal control of gene expression, which is critical in the establishment of cellular identity during development, cellular responses to stimuli, DNA replication and other cellular processes [2].

Several assays exist to identify open chromatin regions in a genome-wide manner. These include DNase-seq, which utilizes the DNase I nuclease [3], FAIRE-seq, which utilizes differences in polarity between nucleosome-bound and nucleosome-free DNA [4], and ATAC-seq, which uses a transposase to selectively cut into accessible DNA [5]. Although each of these assays identifies some unique open chromatin regions, they are generally highly correlated in their identifications [5]. Whereas DNase-seq and FAIRE-seq are complicated protocols that require, on average, one million cells, ATAC-seq is a simple three-step protocol, which is optimized for fifty thousand cells, and can be performed on as few as 500 cells or at a single cell level [5, 6]. ATAC-seq has become popular over the years, and the Cistrome Database [7], in an effort to collect all publicly available functional genomics data, has listed nearly 1,500 datasets for human and mouse.
Due to steric hindrance, the Tn5 transposase used in ATAC-seq preferentially inserts into NFRs. However, it is also possible for the transposase to insert into the linker regions between adjacent nucleosomes, resulting in larger (over 150bps) DNA fragments, which correspond to the integer numbers of adjacent nucleosomes [5]. The DNA fragments are constructed in a paired-end library for sequencing, and after mapping both sequenced ends of each fragment to the genome sequence, we can infer their fragment lengths according to the observed mapping locations, or the insertion length. As described in [5], if we plot the observed fragment length versus frequency, we will see a multi-modal distribution that creates different modes representing transposase insertion into NFRs and linker regions. This allows ATAC-seq to elucidate multiple layers of information relative to the other assays. Although computational tools exist for DNase-seq, FAIRE-seq and ChIP-seq [8], that can be and are used for ATAC-seq analysis, such as MACS2 [9] and F-Seq [10], these would fall short since they only utilize a subset of information, usually the nucleosome-free signals. To date, there are no dedicated peak-callers specifically to account for ATAC-seq.

We present here HMMRATAC, the Hidden Markov Modeler for ATAC-seq, a semi-supervised machine learning approach for identifying open chromatin regions from ATAC-seq data. The principle concept of HMMRATAC is built upon “decomposition and integration”, whereby a single ATAC-seq dataset is decomposed into different layers of signals corresponding to different chromatin features, and then the relationships between the layers of signals at open chromatin regions are learned and utilized for pattern reorganization (Fig. 1a and 1b). Our method takes advantage of the unique features of
ATAC-seq to more accurately identify the chromatin structure. We found that HMMRATAC was able to identify chromatin architecture and the most likely transcription factor binding sites. Additionally, compared with existing methods used for ATAC-seq analysis, HMMRATAC outperformed them in most tests, including recapitulating active and/or open chromatin regions identified with other assays.

A typical analysis pipeline for ATAC-seq would begin with aligning the sequencing reads to a reference genome, followed by identification of accessible regions or “peaks” and then downstream analysis such as motif enrichment in the accessible peaks; differential accessibility identification; and association studies with other data sets, such as gene expression data. Quality control measurements would also take place at each step, such as calculating sequence quality during the alignment and performing replicate correlation during peak calling. We envision HMMRATAC becoming the principle peak-calling method in such a pipeline. HMMRATAC could also aid in other procedures in the pipeline, such as identifying likely transcription factor binding sites to be used in motif analysis or aiding in differential accessibility analysis by identifying regions that are accessible in one condition compared to another.

Results

The distinctive pattern of ATAC-seq fragments around known accessible elements
In order to investigate the ATAC-seq profile around accessible genomic regions, we took the ATAC-seq data in the human GM12878 cell line from Buenrostro et al. [5] and separated the fragments into short nucleosome-free (under 100bp) and mono-nucleosome (180-250bp) fragments based on cutoffs identified by Buenrostro et al. These cutoffs were determined by fitting the fragment length distribution to several simulated distributions modeling nucleosome and nucleosome-free transposition frequencies [5]. We plotted the two types of fragments around the centers of DNasel hypersensitive sites (DHSs) in the same cell line, identified by ENCODE [11] (Additional file 1: Fig. S1). We observed a clear, symmetrical pattern around the centers of these sites characterized by an enrichment of the nucleosome-free fragments and flanking enrichment of the mono-nucleosome fragments. We were able to visually discern at least three distinct regions around an open site: 1) the center, characterized by an enrichment of nucleosome-free fragments and flanked by nucleosome fragments; 2) the nucleosome regions, characterized by an enrichment of nucleosome fragments; 3) and the background, characterized by a depletion in any type of fragments (Fig. 1a; Additional file 1: Fig. S1). We hypothesized that this pattern would exist at all open chromatin throughout the genome and that by learning and recognizing this pattern computationally in a single ATAC-seq dataset, we would identify all open chromatin regions with higher confidence than any other existing method.

**Signal decomposition through the probabilistic approach**

In the first ATAC-seq paper by Buenrostro et al., the authors separated the fragments according to their insertion lengths into four populations representing those from the NFRs
and those spanning 1, 2 or 3 nucleosomes [5]. Non-overlapping and non-adjacent cutoffs were used. For example, lengths below 100 bp represent nucleosome-free signals and lengths between 180 and 250 bp represents mono-nucleosome signals. We consider such cutoff-based strategy not a general practice for the following two reasons. First, different species or different cell lines may have different nucleosome spacing and these differences may not be properly understood [12, 13]. Secondly, since the fragments whose size falls between the non-overlapping and non-adjacent cutoffs are discarded (e.g. those between 100 and 180 bp long), the recall (or sensitivity) to identify accessible regions will be reduced (Additional file 1: Fig. S2 and explained later). As a fact, about 15% of the entire data generated from the original paper are ignored with their strategy. We addressed the above two challenges in a probabilistic approach in HMMRATAC.

First, we let the data itself identify the optimal parameters for the four different distributions of fragment lengths. Using the distributional assumptions of Buenrostro et al., we create four distributions to represent four different chromatin features: nucleosome-free regions (NFRs), mono-nucleosomes (1Ns), di-nucleosomes (2Ns) and tri-nucleosomes (3Ns). HMMRATAC uses a mixture model and Expectation-Maximization (EM) algorithm to identify optimum parameters (Fig. 1c; Additional file 1: Fig S3; see method section for more detail). As an example, HMMRATAC identified the optimum fragment lengths representing 1Ns, 2Ns, and 3Ns from ATAC-seq data in the human GM12878 cell line, generated by merging three replicates from [5], as being 195bp, 396bp, and 693bp, respectively. As a demonstration of different nucleosome spacing in different datasets, HMMRATAC determined the optimum fragment lengths for the 1Ns, 2Ns, and 3Ns from
ATAC-seq data in human monocytes treated with RPMI for one day [14] as 186bp, 368bp, and 549bp, respectively. Additionally, the values for any of the individual replicates, in the GM12878 data [5], were never more than 5bp off from the average values listed above for the merged datasets (Additional file 1: Fig S3). Not all ATAC-seq datasets show a clear multi-modal pattern, especially datasets from fewer cells. As an example, the fragment distribution generated by Buenrostro and et al. utilizing 500 cells (GM12878 cells) per-replicate, did not show as clear of a distribution as 50,000 cells per replicate (Additional file 1: Fig. S4a and S4b) [5]. However, we found our approach can robustly identify the parameters even if the multi-modal pattern is not clear. As we can see from Fig. 1c, the mean and standard deviation values converged after no more than 12 iterations and the parameters converged to very similar values for either the 50,000-cell per-replicate or 500 cell per-replicate ATAC-seq datasets in GM12878.

Secondly, we propose in HMMRATAC a probabilistic approach to make use of data from the entire spectrum of fragment lengths. We generate four separated signal profiles in 10bp resolution, for NFRs, 1Ns, 2Ns, and 3Ns, allowing each mapped fragment to contribute to all profiles with different weights corresponding to its particular fragment length. For each genomic location that is occupied by a particular fragment, all four signal profiles are incremented by the probabilities that this fragment, assigned with an observed fragment length through read mapping, belongs to each of the four corresponding distributions. We tested the performance of a regular HMMRATAC approach with this probabilistic method utilizing the entire data and an alternative approach utilizing only the subset of data satisfying the cutoffs proposed in Buenrostro et al. The two approaches were compared in
terms of precision and recall against the “active regions” true set (Additional file 1: Fig. S2; described in more detail below). We found that our probabilistic method had a considerably better recall and precision compared to the cutoff based method.

HMMRATAC utilizes the length of an insertion fragment, identified by paired-end sequencing, as an important and critical piece of information in its processes. As such, HMMRATAC can only be used in paired-end sequencing libraries that have not undergone stringent size selection, either physical or computational. Because we observed a decrease in the performance of our method when we used the cutoff based approach, which resulted in the removal of ~15% of all sequencing reads, we wanted to understand the effects of size selection and paired-end sequencing on ATAC-seq peak calling. We performed two in silico size selections on the data from Buenrostro et al. The first restricted the fragments to those whose length was under 100bp and the second restricted the fragments to those with lengths under 300bp. We also treated the data as single-end, by unpairing the read mate pairs. We then used MACS2 [9] to call peaks with each data set, including the original unaltered data and compared the performance to predict the “active regions” in chromatin (Additional file 1: Fig S5; see methods and below for more detail). As a result, the original unaltered data performed the best in terms of precision and recall. This indicates that an ATAC-seq assay should be performed without a size selection step and the resulting fragments should undergo paired-end sequencing, regardless of the peak calling algorithm that is used.

Integration through Hidden Markov Model
After removing reads with low mapping quality, and masking blacklist regions from downstream analysis [15], HMMRATAC identifies a set of training regions throughout the genome to train a 3-states Hidden Markov Model (HMM) with multivariate Gaussian emissions, representing the underlying three chromatin states of 1) the center of open chromatin, 2) the nucleosome regions and 3) the background (Fig. 1a). By default, training regions can be either specified by users or automatically selected where the signal fold-change above background falls within a predetermined range. After signal decomposition, a matrix of 10bp resolution signals for NFRs, 1Ns, 2Ns, and 3Ns in training regions is used to train the parameters in HMM through the Baum-Welch algorithm [16]. The initial state transition probabilities are set equally as 0.33 and the emission parameters for each state are calculated after clustering all locations in the training regions into 3 classes (See methods section for more detail). Fig. 1d shows the final parameters for the HMM on ATAC-seq in the GM12878 cell line, again generated by merging three replicates from [5]. Although the parameters will change depending on the choice of training regions and the datasets provided, several key characteristics are evident in the model and remain largely stable with different settings and data. State 3 in the HMM model, which HMMRATAC annotated as the center state, showed the highest average signal across the four signal profiles. State 2, annotated as the nucleosome state, showed the second highest average values. Additionally, as we hypothesized, the results show the center state mainly transits from or to the nucleosome state.
After learning the model, HMMRATAC takes a matrix of genome-wide NFRs, 1Ns, 2Ns, and 3Ns signals and labels each genomic locations with one of the states using the Viterbi algorithm [17]. On the 50K cells ATAC-seq data in GM12878 cell line, there were about 4.3 million regions annotated as background with an average size of 288 bps (~35% of genome size), about 4.5 million regions annotated as nucleosomes with an average size of 382 bps (~49%), and about 226 thousand regions annotated as center regions with an average size of 504 bps (~3%). HMMRATAC finally connected the center regions with their adjacent nucleosomes and called a total of 87,440 accessible regions with an average size of 1,831 bps (5% of genome size), excluding those with abnormally high coverage and those below a minimum length (see method section for more details). It should be noted that many of the nucleosomes are not adjacent to the center state. These may represent non-regulatory nucleosomes that ATAC-seq is able to identify or may be artifacts due to increased transposition frequency.

**Recapitulating chromatin architecture at functional elements**

We hypothesized that HMMRATAC could identify chromatin architecture around accessible regions, or ideally, nucleosomes near the regulatory elements. To prove this, we explored the distribution of histone modification ChIP-seq signal around our identified accessible regions. We downloaded and analyzed datasets released by ENCODE in the GM12878 cell line [11], including ChIP-seq for CTCF, active histone marks such as H3K4me1, H3K4me3, H3K9ac, and H3K27ac, repressive histone marks such as H3K9me3 and H3K27me3, as well as sequencing of micrococcal nuclease digestion
Fig. 2a shows two loci, one on chromosome 8 that was identified as an accessible region by MACS2 and F-Seq but not HMMRATAC and another locus on chromosome 2 that was identified as accessible by all three methods. At the chromosome 2 locus, we observed that the center state was located in regions characterized by depletion in both active histone marks and MNase digestion. The surrounding nucleosome states showed enrichment in active histone marks and MNase digestion. We next plotted the same ChIP-seq, DNase-seq and MNase-seq data around all center states and used k-means clustering to group the resulting profiles (Fig. 2b). We found that HMMRATAC states can identify, what appear to be, different types of regulatory elements. For example, the first cluster (C1) characterized by strong enrichment of H3K27ac relative to H3K4me1 seems to indicate active enhancers while the third cluster (C3) marked by high levels of H3K4me1 and more moderate levels of H3K27ac would suggest poised enhancers (Fig. 2b). The second cluster (C2), characterized by strong enrichment of CTCF binding and lower levels of histone modifications, seems to suggest that this cluster represents insulators or domain boundaries. Taken together, these results indicate that HMMRATAC can recapitulate the histone architecture around accessible regions.

The chromosome 8 locus exemplifies some of the advantages of HMMRATAC over other peak callers. Both MACS2 and F-Seq identified this locus as being accessible, despite the lack of other corroborating evidence, such as transcription factor binding and the presence of active chromatin states. It appears that the site was called by these methods because of an enrichment of the mono-nucleosome signal. Because neither MACS2 or F-Seq integrate the various chromatin features present in ATAC-seq data and because they are based on
signal enrichment only, as opposed to HMMRATAC which incorporates the structure of
the element, they both falsely identified this region as accessible, while HMMRATAC
identified the region as being a nucleosome state.

In order to understand the chromatin features that are enriched in each state across the
genome, as the unique result HMMRATAC is able to conclude from a single ATAC-seq
data, we plotted the overlap of the 3 hidden states with chromatin states and histone mark
calls from ChIP-seq in the GM12878 cell line based on ENCODE data (Fig 3a; Additional
file 1: Table S1). We found that the chromatin states “Active Promoters”, “Strong
Enhancer”, “Insulator”, “Poised Promoter” and “Weak Promoter” states mainly overlaps
with our center state, and the “Heterochromatin”, “Repetitive”, “Repressed”,
“Transcription Elongation”, and “Weak Transcription” barely overlaps with the center
state. Interestingly, the “Strong Enhancer” state had a similar amount of overlap with the
nucleosome state as the center state. Combined with our observation that our nucleosome
state showed elevated levels of active histone marks near accessible regions, we
hypothesized that this state may represent regulatory nucleosomes, those marked by active
histone modifications and adjacent to accessible regions. We plotted the H3K4me1 active
histone mark around the centers of our center states as well as around the centers of the
adjacent nucleosome states, those directly flanking our center states (Fig 3b). As we can
see, the center states show a bimodal enrichment for the histone mark, consistent with
MACS2 and F-Seq peaks. Additionally, our adjacent nucleosome states show central
enrichment for the histone mark. This indicates that the nucleosome state, at least those
adjacent to our center states, are likely regulatory nucleosomes. As a result, HMMRATAC
merges the center states with their adjacent nucleosome states, both upstream and downstream, to identify the regulatory region as the output.

Functional regulatory elements that harbor transcription factor binding sites are generally under higher evolutionary constraint [18]. Success in transcription factor ChIP-seq data analysis is often benchmarked by high PhastCons score, a precompiled evolutionary conservation score at every base pair [19], at the peak center or summit, relative to the surrounding regions [20]. We found that the summits of accessible regions detected by HMMRATAC in GM12878 cell line also showed PhastCons enrichment in vertebrates (Fig 3c), indicating that these summits were likely locations of evolutionarily conserved regulatory elements and could represent functional regulatory elements.

Method comparison with MACS2 and F-Seq

We next sought to compare the performance of HMMRATAC in identifying open chromatin regions to other popular software for ATAC-seq. We chose MACS2 [9] and F-Seq [10], which had previously been shown to be the most accurate peak-callers in identifying open chromatin regions from DNase-seq [21]. MACS2, originally designed for ChIP-seq, utilizes a local Poison model wherein enrichment is calculated relative to a local Poisson model. F-Seq invokes a Gaussian kernel-smoothing algorithm and then calculates enrichment based on the resulting Gaussian distribution. Although none of them utilize the intrinsic and unique features in ATAC-seq data we described before, MACS2 has been used to analyze ATAC-seq data in many published works [6, 14, 22-26], and
MACS2 and F-Seq is part of the ENCODE analysis pipeline for ATAC-seq data [27]. We tested the performance of the four algorithms using the two ATAC-seq datasets in GM12878 cell line [5] as the previous sections.

A major limitation in comparing the performance of identifying open chromatin regions is the absence of a valid “gold standard” [2]. To overcome this problem, we chose to compare the performance of HMMRATAC and the other algorithms to three distinct independent datasets as the proxies of true and false set: 1) the active regions and heterochromatin from chromatin states analysis combining multiple histone modification features[28], 2) DNase-seq [3] and 3) FAIRE-seq [4] accessible regions of the same GM12878 cell line.

**Enrichment in chromatin states of active regions**

The first proxy of the true set that we compared the algorithms to is referred to as “active regions.” These sites were defined through an integrative chromatin state segmentation using chromHMM [28, 29] algorithm conducted by ENCODE project. 8 histone modifications (H3K27me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3, H3K27ac and H3K9ac), 1 transcription factor (CTCF) ChIP-seq and 1 whole genome input datasets in the GM12878 cell line were integrated through an unsupervised machine learning approach and each genomic location was annotated as one of 15 mutually-exclusive states. Two resulting states, “active promoters” and “strong enhancers”, were merged together to create our list of “active regions”. Additionally, the state annotated as “heterochromatin” was used as a false set. We tested different cutoff values for each
algorithm (see Methods) and found that HMMRATAc outperformed the other algorithms, in terms of precision, recall, and false positive rate in identifying these “active regions” while avoiding “heterochromatin” (Fig. 4 top row). The area under the curve (AUC) was then calculated for the recall vs. false positive rate curves (AU-ROC) and for the precision-recall curves (AU-PRC) (Table 1), after adding extreme points to the unreachable ends of the curves. HMMRATAc was proved to have better AUC than the two other methods. It worth noting that HMMRATAc calls larger peaks than the other two methods due to its merging of the adjacent nucleosome states to the center state. However, despite the larger average peak size, HMMRATAc identifies a comparable number of accessible regions in base pairs, as shown in Additional file 1:Fig. S6, at comparable levels of sensitivity for active chromatin regions. This would also indicate that many peaks called by the other methods are falling outside of the “gold standard” regions, possibly contributing to their deficits in precision compared to HMMRATAc. Taken together, these results indicate that HMMRATAc is able to identify active chromatin regions with a higher recall and precision and lower false positive rate, than MACS2 or F-Seq.

Consistency with results from DNase-Seq and FAIRE-Seq

We then used the accessible chromatin regions detected by two independent assays DNase-Seq and FAIRE-seq in GM12878 cell line as cross-validation of ATAC-seq analysis. We downloaded DNaseI hypersensitive sites (DHSs) [3] and FAIRE-seq [4] accessible regions called by ENCODE in the GM12878 cell line [11, 30], then merged DHSs or FAIRE accessible regions that were within 2kb of each other to define our alternative true sets.
Because the DNase-seq and FAIRE-seq regions are called from traditional methods as Hotspot [31] and F-Seq, the evaluation based on a single dataset is less reliable compared with the previous test with chromatin state which integrates multiple evidence. We continued to use the heterochromatin states outlined above as our false set, although we eliminated any DHSs or FAIRE regions that overlapped a heterochromatin state region and vice versa. We found that HMMRATAC outperformed the other two methods in terms of AUC-PRC in recapitulating DHSs (Additional file 1: Fig. S7, Table 1). MACS2 performed slightly better in AUC-ROC by 0.001. F-Seq was found to perform the best in terms of AUC-ROC and AUC-PRC with the FAIRE-seq data.

**Reproducibility in replicates**

A common and robust method to gauge a method's performance in analyzing next-generation sequencing data, in particular, ChIP-seq, is the robustness in technical and biological replicates [32]. One such way to measure the reproducibility of peak callers is the Irreproducible Discovery Rate (IDR) [33]. Therefore we sought to test the reproducibility of HMMRATAC and the other algorithms in utilizing ATAC-seq datasets. We calculated the number of reproducible accessible regions, defined as overlapping accessible regions with an IDR score <= 0.05, between each pair of three replicates for each algorithm. We show that when using the individual replicate data (Additional file 1: Table S2), HMMRATAC produced a similar number of reproducible accessible regions as the other two algorithms. Furthermore, the reproducible peaks called by each algorithm showed a high amount of overlap with each other (Additional File 1: Fig. S8).
Furthermore, we hypothesized that the biology (open chromatin) should be consistent even with a smaller number of cells used in ATAC-seq. We tested how well each algorithm performed at recapitulating the accessible regions called from the 50,000 cell per-replicate data and the 500 cell per-replicate data. We found, similar to the results from pair-wise replicate testing, that HMMRATAC was able to identify a similar number of reproducible accessible regions, with IDR <= 0.05, using datasets with a different number of cells (Additional file 1: Table S2). These data indicate that HMMRATAC is comparable to MACS2, and F-Seq in identifying reproducible accessible regions.

Application for human monocytes

Thus far, we have shown the performance of HMMRATAC and other algorithms with data generated in the human GM12878 cell line. In order to determine our method's performance with different data sets and different cell lines, we analyzed ATAC-seq data generated in human monocytes subjected to several chemical treatments and sequenced at different time points [14]. In addition to ATAC-seq, Novokovic et al. performed ChIP-seq on the same cells, with the same treatments and time points, for the active histone marks histone H3K27ac, H3K4me1 and H3K4me3 and the repressive histone mark histone H3K9me3. We chose five separate data sets generated by that group: RPMI treatment after 4 and 24 hours, LPS treatment after 6 days and BG treatment after 4 and 24 hours. Using the active histone modification as the “Gold standard” and the repressive histone
modification as the “real negatives” (See Method section for more details), we compared
the performance of the algorithms in the same way we compared their performance in
human data (Fig 4 2nd to 6th row, Table 1). We found that HMMRATAc has superior
precision, sensitivity, specificity, AUC-ROC and AUC-PRC over both MACS2 and F-Seq.
Overall, this data indicates that HMMRATAc’s superior performance over the other
algorithms is consistent across different data sets and different cell lines.

Summary of method comparisons

The summary is shown in Table 1. We used a total of 16 independent measures in two
different cell types and six different conditions to benchmark the performance of the three
algorithms, including 8 AUC-ROC and 8 AUC-PRC. We found that HMMRATAc
outperformed the other two methods in 7 of 8 AUC-PRC and 6 of 8 AUC-ROC.
Furthermore, in all of the 12 evaluations of which the real positives were compiled by
integrating multiple histone modification ChIP-seq datasets, HMMRATAc outperformed
the other methods constantly. Overall, these results indicate that HMMRATAc is the most
accurate and reliable peak-caller for ATAC-seq data analysis.

Discussion

HMMRATAc is designed to integrate the nucleosome-free and nucleosome-enriched
signals derived from a single ATAC-seq dataset in order to identify open chromatin
regions. For this reason, HMMRATAc should not be used on ATAC-seq datasets that
have undergone either physical or computational size selection. For that same reason, single-end sequence libraries from ATAC-seq cannot be analyzed with HMMRATAC. However, we have shown that size selection results in a decrease in sensitivity and should not be used as a general practice in ATAC-seq protocols. Therefore, a reasonable way to evaluate the quality of ATAC-seq data for HMMRATAC analysis is to check the length distribution of the transposition fragments to ensure that large fragments are present and relatively abundant.

Identifying transcription factor footprints, wherein the DNA protected by a binding protein shows resistance to nonspecific DNA digestion, is an area of active research. Although footprint identification may be possible with certain datasets or for certain transcription factors, we believe that HMMRATAC identified summits, or the position with the most insertion events within the open region, are more robust to identify potential transcription factor binding sites. In fact, it has been previously shown that DNase-seq signal intensity at a binding motif is a better predictor of transcription factor binding than the strength of a footprint [34], and we believe this principle applies to ATAC-seq as well. However, a major reason for this result from He et al., was that many footprints were artifacts caused by the enzymes’ sequence bias. The Tn5 transposase also has a sequence bias, which would need to be corrected before an effective footprint analysis could occur. Several methods exist [35, 36] that could correct the sequence bias from an ATAC-seq library. HMMRATAC could be used downstream of such sequence corrections, to reduce bias’ in peak calling, or upstream of the corrections, to identify the reads within peaks that need to be corrected for footprint or other downstream analysis.
It is possible to extend HMMRATAC to identify differentially accessible regions between two or more conditions. Most of the approaches for identifying such differences from ATAC-seq data would suffer from the same problems mentioned before, namely, that these methods only consider the relative strength of the signal and the differences in signal intensity between conditions. As we’ve shown here, the incorporation of the structure of regulatory chromatin into a model could increase the accuracy and sensitivity of calling accessible regions. We believe our strategy of “decomposition and integration” can be adopted in a differential accessibility analysis pipeline, by studying the differential NFRs and nucleosomal signals separately and then combining together.

**Conclusions**

ATAC-seq shows numerous advantages over other methods for identifying open chromatin regions, such as DNase-seq and FAIRE-seq, owing to its low starting material requirement and simple protocol. For these reasons, ATAC-seq has become one of the standard assays for locating regions of open chromatin, particularly as biomedical research continues to move toward translational research and precision medicine. Although the number of published ATAC-seq datasets increases rapidly, researchers are still using methods originally developed for ChIP-seq to analyze the data. We present HMMRATAC as a dedicated algorithm for the analysis of ATAC-seq data. We take advantage of the nature of the ATAC-seq experiment that due to the lack of size-selection while preparing the DNA library in its protocol, DNA fragments associated with well-positioned nucleosomes
around open chromatin will be sequenced as well. Therefore, as a side note, HMMRATAC shouldn't be applied to those ATAC-seq data generated with a stringent size-selection on DNA fragments. Different from current methods where only the read enrichment is considered, HMMRATAC separates and integrates ATAC-seq data to create a statistical model that identifies the chromatin structure at accessible and active genomic regions. We have shown that HMMRATAC outperforms other computational methods in recapitulating open and active chromatin identified with other assays such as chromatin state analysis integrating multiple histone mark ChIP-seq, DNase-seq and FAIRE-seq. As HMMRATAC is a cross-platform and user-friendly algorithm, we envision it becoming the standard for ATAC-seq data analysis pipeline, replacing current methods designed for ChIP-seq analysis.

Methods

Preprocessing of ATAC-seq data

The human GM12878 cell line ATAC-seq paired-end data used in this study was publicly available and downloaded under six SRA [37] accession numbers SRR891269-SRR891274[5]. There are three biological replicates generated using 50,000 cells per replicate and other three generated using 500 cells per replicate. Each dataset was aligned to the hg19 reference genome using bowtie2 [38]. After alignment, each group of replicates (either 50,000 cells or 500 cells) were merged together, sorted and indexed. Reads that had a mapping quality score below 30 or that were considered duplicates (exact
same start and stop position) were removed from the merged files. It should be noted that HMMRATAC will remove duplicate and low mapping quality reads by default, although some other algorithms do not.

The merged, filtered and sorted BAM files, created as described above, were the input for MACS2. HMMRATAC took the sorted paired-end BAM file and its corresponding BAM index file as the main inputs. In addition, HMMRATAC requires a genome-wide read coverage file in BigWig format. To generate the coverage file, MACS2 was run with the option set for reporting the read coverage in BedGraph format (option –B) and this BedGraph was then converted to a BigWig file using UCSC’s bedGraphToBigWig program. F-Seq requires a single-end BED file of read alignment results as its input. To generate this BED file, we converted the paired-end BAM file into a BED file, using an in-house script, and split each read pair into forward and reverse strand reads.

The human monocyte data from [14] was publicly available and downloaded from the Cistrome Database [7] as aligned BAM files to the hg19 reference genome. The data corresponds to Gene Expression Omnibus accessions GSM2325680, GSM2325681, GSM2325686, GSM2325689, and GSM2325690. One replicate per condition was used and processed in the same way as described above, filtering.

The HMMRATAC algorithm

Segregation of ATAC-Seq signals
After the preprocessing step that eliminates duplicate reads and low mapping quality reads, the main HMMRATAC pipeline begins by separating the ATAC-seq signal into four components, each representing a unique feature. These features are nucleosome-free regions (NFRs), which are most likely to occur within the open chromatin region itself, and three nucleosomal features, representing mono-, di- and tri-nucleosomes (1Ns, 2Ns and 3Ns), respectively. We utilize four distributions to represent the four signal tracks: an exponential distribution for the nucleosome-free track and three Gaussian distributions to represent the three nucleosomal signal tracks. These distributions were utilized by Buenrostro et al. to identify their cutoffs for fragment separations [5]. Although sequence read data is discrete, the fact that the expectation value of each Gaussian is large (>100bp) and the number of observations is high, the discrete distribution, such as a Poison distribution, can be approximated as a Gaussian. The mean value of the exponential distribution for NFRs is set as a fixed value at runtime, either as the default value or a user-defined value. The parameters of three Gaussian distributions are initialized at runtime with default or user-defined values and then updated using the expectation maximization (EM) algorithm for Gaussian mixture models. These initial values were the same as those used by Buenrsotro et al. to define their distributions for the chromatin features [5]. Briefly, each nucleosomal fragment, those larger than 100bp, is classified as belonging to one of the nucleosomal distributions based on its weighted probability. Once each fragment has been classified, new means, standard deviations, and weights are calculated for each distribution and become the new values for the mixture model. This process continues iteratively until the change in the mean values between iterations is less than a reasonable
epsilon value. At this point, the model is assumed to have reached convergence and the EM process is halted (Fig. 1c). In order to decrease the time required for the EM algorithm, HMMRATAC randomly sub-selects 10% of all the fragments to use as the training data. Once the parameters of the four distributions have been determined, HMMRATAC creates four genome-wide signal tracks. For every genomic position, all of the fragments that occupy that position are determined. Each track is then incremented by the probability that a particular fragment belongs to the tracks’ corresponding distribution. The signal tracks are then transformed by square root into continuous space.

**Training the Hidden Markov Model of accessible chromosomal regions, and refining the model with the Baum-Welch algorithm**

Once the four signal tracks have been created, the main HMMRATAC process begins. HMMRATAC first identifies up to 1000 training regions throughout the genome, with which it learns the model. These training regions are determined by scanning the genome for regions whose read coverage fold changes over background falls within a certain range. The default lower and upper limit of fold changes, as used in the Result section, is set between 10 and 20. These regions are then extended by 5KB in either direction. It is also possible for users to provide a BED file containing predefined training regions. The initial HMM has proportional initial probabilities, which are not updated during training, as well as proportional transition probabilities, which are updated during training. The emission probabilities are calculated with k-means clustering [39], by separating the training regions into 3 clusters and calculating the means and covariance matrices for the four signal tracks.
for each cluster. We model the emissions as a multivariate Gaussian distribution since this assumption has been successfully adopted in other HMM-based genome segmentation methods such as hiHMM [40, 41] and Segway [42, 43]. The three clusters represent the three states of our model, corresponding to the center of the open region, the nucleosomes, and a background state, respectively. It is also possible to create a model using a different number of states instead of 3, although this option is not generally recommended. Once this initial HMM is created, the transition and emission parameters are updated with the Baum-Welch algorithm [16] until convergence.

**Annotating with the Viterbi algorithm**

Once the HMM has been created and refined, the Viterbi algorithm [17] is used to classify every genomic position into one of the three states of our model. In our experience, Viterbi has trouble when encountering a very high coverage region. Therefore, regions whose z-scored coverage is above a certain cutoff, either a default or user-defined value, are masked by the Viterbi algorithm. Additionally, it is possible to include a list of genomic regions to mask along with the extremely high coverage regions, such as annotated blacklisted sites. We have found that 91% of the high coverage regions that are masked from the Viterbi algorithm in the GM12878 cells are annotated blacklist regions [44]. A genome-wide BedGraph of all state annotations can be reported by HMMRATAC, although this is not the default behavior. By default, HMMRATAC will take all regions classified as belonging to the center state and whose length is above a default or user-defined threshold and merge them with the nucleosome states located both upstream and downstream of the
open state. We generally consider the nearby nucleosomes to be part of the regulatory region (Fig. 3a and b), and therefore the region is then reported as a peak in gappedPeak format.

It should be noted that at all stages of the algorithm, HMMRATA C does not store the read or coverage information in memory. Any information needed by the program is read from the desired file and used at the specific time. This was incorporated to reduce the memory usage of HMMRATA C. However, because of this feature, it is not recommended to run HMMRATA C simultaneously with the same data files. It is, however, possible to process multiple different data sets in parallel. In terms of system requirements, the only required system component is to have Java 7 or higher installed. We have tested HMMRATA C on OpenJDK 7, standard edition JRE 8 and the standard edition JRE 10. Finally, in terms of runtime, HMMRATA C took approx. 4.9 hours to process the GM12878 data set and an average of 5.6 hours to process the monocyte data in our server equipped with AMD Opteron(tm) Processor 6378 and 500G memory.

Post-processing and output

When the report peak option is set for HMMRATA C (option -p True), the peak region is reported in gappedPeak format. The total peak region reported is a region that was classified as being in the center state of the model (and is longer than a default or user-supplied length cutoff) along with the nucleosome states, both upstream and downstream. Each peak is given a score, which represents the maximum read coverage of the open state.
It is also possible to report the average read coverage, the median coverage, and the fold change above the genomic background, the z-score of the average coverage or all values. Additionally, those high coverage regions that are masked from Viterbi decoding (see above), are added back to the final peak file. They are annotated by HMMRATAC as “High Coverage Accessible regions”. The user may then keep or discard these regions and explore them as they see fit.

Additionally, when the report peak option is set, HMMRATAC also supplies an additional file containing the summits of each peak. To calculate a summit, HMMRATAC first smoothes the read coverage signal through the center state using a Gaussian smoothing function with a window size of 120bp. It then calculates the position that has the maximum read coverage and reports that position as the summit.

**Evaluation**

Calculating Precision, Recall, False Positive Rate and F1-scores

Let Predicted Positive (PP) represent the number of base pairs identified as accessible regions from any method. Let Real Positives (RP) represent the number of base pairs in the “gold standard” sets (described below). Let Real Negatives (RN) represent the number of base pairs in the “real negative” set described below. Let True Positive (TP) represent the number of overlapping base pairs between a called peak and a real positive. Let False
Positive (FP) represent the number of overlapping base pairs between a called peak and a real negative.

Then precision or positive predictive value (PPV) is calculated as: \( \text{PPV} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \)

Recall, or true positive rate (TPR), is calculated as: \( \text{TPR} = \frac{\text{TP}}{\text{RP}} \)

False positive rate (FPR) is calculated as: \( \text{FPR} = \frac{\text{FP}}{\text{RN}} \)

The number of overlapping base pairs was calculated with the BEDTools intersect tool [45] and custom python scripts.

**Compiling the “Gold Standard” lists and “True Negative” set**

We tested the performance of HMMRATAC and other methods against three “gold standard” datasets. The first one, which we call “Active Regions”, is made by combining two chromatin state annotations generated in GM12878 cells using chromHMM [28, 29] (wgEncodeBroadHmmGm12878HMM). These two annotations are “Active Promoters” and “Strong Enhancers”.

The other two “gold standard” test sets that were used in this study were DNaseI hypersensitive sites (wgEncodeAwgDnaseUwdukeGm12878UniPk) and FAIRE accessible regions (wgEncodeOpenChromFaireGm12878Pk) [11, 30]. We found that there
were numerous cases throughout the genome where clusters of DNase or FAIRE accessible regions were in close proximity to one another. These clusters tended to have broad enrichment in either DNase-seq or FAIRE-seq signal across the regions. Therefore, we merged all accessible regions that were within 2KB of each other into a single peak and used these merged sets as the “gold standard” sets.

The “real negative” set was made with the “Heterochromatin” annotation from chromHMM. When using the “active region” set, no modifications were necessary because the three states are mutually exclusive. When using either the DNase or FAIRE sets, any “heterochromatin” states that overlapped (by any amount) a DNase or FAIRE peak were excluded from the final “real negative” set. Additionally, any DNase or FAIRE peak that overlapped any “heterochromatin” state (by any amount) was, likewise, excluded from its respective “gold standard” set.

The “gold standard” set used for the monocyte data were compiled as follows. For each condition, 2 replicates of the histone marks H3K27ac, H3K4me1, and H3K4me3 from the same study [14] were downloaded from the Cistrome Database [7] as aligned BAM files. Each file was processed by MACS2 with the p-value parameter (-p option) set to 0.15. The resulting peaks were sorted by the –log10(FDR) value and the top 10000 peaks for each sample were pooled together and merged. The “real negative” set for the monocyte data were H3K9me3 peaks from the same study. One replicate of H3K9me3 histone mark ChIP-seq was downloaded from the Cistrome Database, for each condition, as an aligned BAM file. The file was then processed by MACS2 with the p-value parameter set to 0.15. The
peaks were sorted by their $-\log_{10}(\text{FDR})$ value and the top 10000 peaks were kept. At this point, the “Active Regions” true set was created by eliminating any merged active peak, created as described previously, that overlapped an H3K9me3 peak. This list of “Active Regions” became the “gold standard” set for each monocyte condition. Any H3K9me3 peak that did not overlap the “Active Region” set, was kept as the “real negative” set for that condition. The accession numbers for all above histone modification ChIP-seq data are listed in Additional file 2.

**Calls from MACS2**

Version 2.1 of MACS was used to call peaks using the sorted paired-end BAM files created as described above. With each run the input file format (option -f) was set to “BAMPE” and the option for keeping duplicate reads (option --keep-dup) was set to “all”, as duplicate reads had already been discarded in pre-processing. MACS2 was run with the local lambda option turned off (option --nolambda) and was run with q-value cutoffs (option -q) set to 0.5, 0.1, 0.05 (default), 0.005, 0.0005 and 0.00005. It was also run with a p-value cutoff (option -p) set to 0.6 and 0.3.

**Calls from F-Seq**

Version 1.84 of F-Seq was used to call peaks using the single-end BED files created as described previously. With each run, the output was reported as a BED file (option -of).
The standard deviation cutoff (option -t) was set to 0.1, 1, 4 (default), 6, 7, 8, 10, 15 and 20.

**Calls from HMMRATAC**

HMMRATAC was run using the sorted BAM files, BAM index files and genome-wide coverage files created as described previously. The upper limit fold-change for choosing training regions (option -u) was set to 20 with corresponding lower limits (option -l) set to 10. These settings are the HMMRATAC defaults. HMMRATAC was also run using blacklisted sites for hg19 [11, 15], which were masked from the program and all subsequent steps. For the precision, recall and false positive rate calculations, the HMMRATAC peaks were filtered by minimum score cutoffs of 0, 5, 10, 15, 20, 25, 30, 35, 40 and 45. Additionally, for two monocyte samples (accession numbers GSM2325689 – RPMI_4h and GSM2325681 – BG_d1) HMMRATAC did not produce a typical model with the default settings (-u 20 –l 10). Instead, the models produced showed the nucleosome state as having the highest NFR signal. We re-ran these samples with more conservative settings (-u 40 –l 20), at which time HMMRATAC produced models that showed the characteristic pattern that we have seen with the other samples. These peak results from HMMRATAC were then filtered by the same cutoffs as the other samples and tested in the same way. All the transition and emission parameters learned by HMMRATAC from GM12878 50k cells, 500 cells, five human monocyte samples are summarized in Additional file 1: Table S3.

**Calculating reproducibility**
Reproducibility was determined using the Irreproducible Discovery Rate (IDR) originally described by Li et al. [33]. Peaks were called using the following parameters: the p-value for MACS2 was set to 0.6 for 50,000 cell per-replicate merged data and 0.6 for 500 cell per-replicate merged data, the standard deviation for F-Seq was set to 0.1 for both merged datasets, and the fold-change range for HMMRATAC was set to 10-20 (default) for both merged datasets. The parameters for each individual 50,000 cell replicate was the same for F-seq and HMMRATAC and the MACS2 q-value setting was set to 0.5. These parameters were also the most permissive settings used throughout this paper, including in our PRC and ROC curves. The MACS2 setting was changed for the individual replicates because at the permissive settings (-p 0.6 or –p 0.3) it called the entire genome as a peak, so we chose the third most permissive setting instead. The peaks, called by each method, were then sorted by their respective scores (-log(p-value) for F-seq, the signal value for MACS2 and HMMRATAC) and the top 100k peaks were retained. These top peaks were the inputs into the IDR software.

For calculating the 500 to 50,000 cell per-replicate reproducibility, each method was run using the merged file(s) described previously, with the respective parameters as described previously. For each method, each pair of resulting peak files, from the 500 cell per-replicate and 50,000 cell per-replicate datasets, were run through the IDR algorithm using default parameters and each methods’ respective score column as the rank. Reproducible peaks between datasets were considered to have an IDR value equal to or below 0.05.
For replicate testing, each method was run with the same parameters as described previously, using the individual replicate file(s) as their input. Each possible pair of replicates was then run through the IDR algorithm with default parameters. For each replicate pair, the number of peaks whose local IDR score was equal to or below 0.05 was calculated. Peaks with IDR scores equal to or below 0.05 in each of the pairwise comparisons were retained and used to calculate the overlap between the three methods (Additional file 1: Fig. S8).

**Abbreviations**

ATAC-seq: assay for transposase-accessible chromatin with sequencing; ChIP-seq: chromatin immunoprecipitation with sequencing; MNase-seq: Micrococcial nuclease digestion with sequencing; FAIRE-seq: formaldehyde assisted isolation of regulatory elements; DNase-seq: DNaseI digestion with sequencing; DNA: deoxyribonucleic acid; HMM: hidden markov model; HMMRATAC: hidden markov modeler for ATAC-seq; NFR: nucleosome free region; DHS: DNaseI hypersensitive sites; ENCODE: encyclopedia of DNA elements; EM: Expectation-Maximization algorithm; 1N: mono-nucleosome; 2N: di-nucleosome; 3N: tri-nucleosome; BP: base-pairs; KB: kilo base-pairs; MB: mega base-pairs; CTCF: CCCTC binding factor; ROC: receiver operator characteristic; AUPRC: area under the precision recall curve; AUROC: area under the receiver operator characteristic curve; IDR: irreproducible discovery rate; PP: predicted positive; RP: real positive; RN: real negative; TP: true positive; FP: false positive; PPV: positive predictive value; TPR: true positive rate; FPR: false positive rate
Declarations

Funding

This work was supported by the startup fund for TL from Jacobs School of Medicine and Biomedical Sciences, University at Buffalo.

Availability of data and materials

The HMMRATAC source code, executable file, quick start guide and javadocs can be downloaded from https://github.com/LiuLabUB/HMMRATAC. All peaks/regions called by HMMRATAC, F-Seq, and MACS2, as well as the various "gold standard" files, can be downloaded from the following URL: http://biomisc.org/download/HumanData.tar.gz.

Authors’ contributions

EDT and TL conceived and designed the computational method. EDT developed the software package, implemented the method and analyzed the data. EDT and TL wrote the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Ethics approval and consent to participate

Ethics approval is not applicable for this study.
Table 1. The Area Under the Curves (AUCs) of method comparisons of HMMRATAC, MACS2, and F-Seq.

The AUC calculated from the ROC and PRC curves in Fig. 4 and Additional file 1: Fig S7. Because we can’t reach the extreme sides of the curves while tuning cutoff values, before we calculated AUC values, we added theoretical extreme points to ROC – (0, 0) and (1, 1), and to PRC – (0, 1) and (1, \(L_{true}/(L_{true} + L_{false})\)). \(L_{true}\) is the total length of real positive set, and the \(L_{false}\) is the total length of real negative set. The numbers of the best performance are shown in bold.

| GM12878 cell line | Human monocytes |
|-------------------|----------------|
|                   | NA | RPMI day 1 | RPMI 4 hours | LPS day 6 | BG 4 hours | BG day 1 |
| Chromatin states  |    |            |              |           |            |          |
| DNase-seq         |    |            |              |           |            |          |
| FAIRE-seq         |    |            |              |           |            |          |
| Histone mark ChIP-seq: H3K4me1, H3K4me3, H3K27ac & H3K9me3 |    |            |              |           |            |          |
| PRC               |    |            |              |           |            |          |
| ROC               |    |            |              |           |            |          |
| PRC               |    |            |              |           |            |          |
| ROC               |    |            |              |           |            |          |
| PRC               |    |            |              |           |            |          |
| ROC               |    |            |              |           |            |          |
| PRC               |    |            |              |           |            |          |
| ROC               |    |            |              |           |            |          |
| PRC               |    |            |              |           |            |          |
| ROC               |    |            |              |           |            |          |
| PRC               |    |            |              |           |            |          |
| ROC               |    |            |              |           |            |          |
| PRC               |    |            |              |           |            |          |
| ROC               |    |            |              |           |            |          |
| PRC               |    |            |              |           |            |          |
| ROC               |    |            |              |           |            |          |
| MACS2             | 0.508 | 0.829 | 0.54 | **0.818** | 0.563 | 0.809 | 0.402 | 0.720 | 0.447 | 0.721 | 0.336 | 0.672 | 0.438 | 0.732 | 0.402 | 0.730 |
| F-Seq             | 0.531 | 0.827 | 0.572 | 0.817 | **0.597** | **0.818** | 0.440 | 0.728 | 0.457 | 0.736 | 0.326 | 0.690 | 0.45 | 0.749 | 0.426 | 0.737 |
| HMMRATAC          | **0.631** | **0.854** | **0.611** | 0.817 | 0.577 | 0.816 | **0.580** | **0.772** | **0.614** | **0.792** | **0.427** | **0.715** | **0.602** | **0.794** | **0.545** | **0.763** |
Figure Legends

Fig. 1 The HMMRATAC Algorithm.

a. A schematic to represent the structure of an accessible region.  b. The algorithmic workflow. Estimation of fragment length parameters through the Expectation Maximization algorithm: c. Mean and standard deviation parameters per EM iteration for three nucleosomal distributions are consistent among datasets using 50,000 and 500 cells. The iteration #1 shows the initial parameters. EM was applied on pooled biological replicates. The parameters per biological replicate were shown in Additional file 1: Fig. S3.

Fig. 2 Chromatin Architecture at HMMRATAC Peaks.

a. IGV browser snapshot of various signals around a negative locus on chromosome 8 and a positive locus on chromosome 2. Top four tracks show ATAC-seq signals, after being separated by size, the fifth track shows CTCF ChIP-seq signal, the sixth through ninth tracks show various histone modification ChIP-seq signal and the tenth track shows MNase-seq signal. The bottom panel shows RefSeq genes, HMMRATAC calls and state annotations (called with default settings, red: background, green: nucleosome, and blue: center state), MACS2 and F-Seq peaks (called with default parameters), the active regions combined from “Active Promoters” and “Strong Enhancers” states from chromHMM (see methods section).  b. Various ChIP-seq, DNase-seq and MNase-seq signals plotted across all distal (> 10kb from gene promoters) open states. Data were clustered using k-means algorithm into 4 clusters.
Fig. 3 Chromatin features of HMMRATAC states.

a. The overlap between either histone modification ChIP-seq peaks or chromatin states in GM12878 and three HMMRATAC states – background, nucleosome, and center. Color shows the percentage of GC, the percentage of total genome length, or the percentage of histone peaks or chromatin states overlapping with each of HMMRATAC states. Overlaps were calculated in basepair level. b. Distribution of GM12878 H3K4me1 ChIP-seq signal in 1000bps window surrounding the summits of F-Seq peaks, MACS2 peaks, or HMMRATAC center and nucleosome states. c. Distribution of Phastcons conservation score within vertebrates in 1000bps window surrounding the summits of F-Seq, MACS2, and HMMRATAC calls.

Fig. 4 Comparison of performance between HMMRATAC, MACS2, and F-Seq

Left: Receiver Operator Characteristic (ROC) curves and right: Precision-Recall (PRC) curves for the real positive and real negative pairs of 1. active chromatin states vs. heterochromatin for GM12878 cells, 2. active histone marks (either H3K4me1, H3K4me3 or H3K27ac) vs. heterochromatin (H3K9me3) for human monocytes (see method for detail).
1. Li G, Reinberg D: Chromatin higher-order structures and gene regulation. *Curr Opin Genet Dev* 2011, 21:175-186.

2. Tsompana M, Buck MJ: Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 2014, 7:33.

3. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: High-resolution mapping and characterization of open chromatin across the genome. *Cell* 2008, 132:311-322.

4. 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, 17:877-885.

5. 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, 10:1213-1218.

6. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ: Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015, 523:486-490.

7. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, Zhu M, Wu J, Shi X, Taing L, et al: Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res* 2017, 45:D658-D662.
8. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**:823-837.

9. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9**:R137.

10. Boyle AP, Guinney J, Crawford GE, Furey TS: **F-Seq: a feature density estimator for high-throughput sequence tags.** *Bioinformatics* 2008, **24**:2537-2538.

11. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57-74.

12. Lantermann AB, Straub T, Stralfors A, Yuan GC, Ekwall K, Korber P: *Schizosaccharomyces pombe* genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of *Saccharomyces cerevisiae.* *Nat Struct Mol Biol* 2010, **17**:251-257.

13. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A: **Determinants of nucleosome organization in primary human cells.** *Nature* 2011, **474**:516-520.

14. Novakovic B, Habibi E, Wang SY, Arts RJW, Davar R, Megchelenbrink W, Kim B, Kuznetsova T, Kox M, Zwaag J, et al: **beta-Glucan Reverses the Epigenetic State of LPS-Induced Immunological Tolerance.** *Cell* 2016, **167**:1354-1368 e1314.
15. Human ENCODE Data Analysis Consortium Blacklisted Regions

[http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz]

16. Baum LE, Petrie T, Soules G, Weiss N: A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. The Annals of Mathematical Statistics 1970, 41:164-171.

17. Viterbi A: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theor 1967, 13:260-269.

18. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S: Eukaryotic regulatory element conservation analysis and identification using comparative genomics. Genome Res 2004, 14:451-458.

19. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005, 15:1034-1050.

20. Shin H, Liu T, Duan X, Zhang Y, Liu XS: Computational methodology for ChIP-seq analysis. Quant Biol 2013, 1:54-70.

21. Koohy H, Down TA, Spivakov M, Hubbard T: A Comparison of Peak Callers Used for DNase-Seq Data. PLoS One 2014, 9:e96303.

22. Hansel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, et al: G-quadruplex structures mark human regulatory chromatin. Nat Genet 2016, 48:1267-1272.
23. Philip M, Fairchild L, Sun L, Horste EL, Camara S, Shakiba M, Scott AC, Viale A, Lauer P, Merghoub T, et al: Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature* 2017, 545:452-456.

24. Denny SK, Yang D, Chuang CH, Brady JJ, Lim JS, Gruner BM, Chiou SH, Schep AN, Baral J, Hamard C, et al: Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. *Cell* 2016, 166:328-342.

25. Chen X, Shen Y, Draper W, Buenrostro JD, Litzenburger U, Cho SW, Satpathy AT, Carter AC, Ghosh RP, East-Seletsky A, et al: ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat Methods* 2016, 13:1013-1020.

26. Risca VI, Denny SK, Straight AF, Greenleaf WJ: Variable chromatin structure revealed by in situ spatially correlated DNA cleavage mapping. *Nature* 2017, 541:237-241.

27. ENCODE ATAC-Seq analysis pipeline [https://www.encodeproject.org/atac-seq/]

28. Ernst J, Kellis M: ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012, 9:215-216.

29. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al: Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011, 473:43-49.

30. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al: The accessible chromatin landscape of the human genome. *Nature* 2012, 489:75-82.
31. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA: Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet 2011, 43:264-268.

32. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012, 22:1813-1831.

33. Li QH, Brown JB, Huang HY, Bickel PJ: Measuring Reproducibility of High-Throughput Experiments. Annals of Applied Statistics 2011, 5:1752-1779.

34. He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H, et al: Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Methods 2014, 11:73-78.

35. Wang JR, Quach B, Furey TS: Correcting nucleotide-specific biases in high-throughput sequencing data. BMC Bioinformatics 2017, 18:357.

36. Martins AL, Walavalkar NM, Anderson WD, Zang C, Guertin MJ: Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. Nucleic Acids Res 2018, 46:e9.

37. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database C: The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 2012, 40:D54-56.

38. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods 2012, 9:357-359.
39. MacQueen J: Some methods for classification and analysis of multivariate
observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical
Statistics and Probability, Volume 1: Statistics; 1967; Berkeley, Calif. University
of California Press; 1967: 281-297.

40. Sohn KA, Ho JW, Djordjevic D, Jeong HH, Park PJ, Kim JH: hiHMM: Bayesian
non-parametric joint inference of chromatin state maps. Bioinformatics 2015,
31:2066-2074.

41. Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A,
Tolstorukov MY, Appert A, et al: Comparative analysis of metazoan
chromatin organization. Nature 2014, 512:449-452.

42. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS:
Unsupervised pattern discovery in human chromatin structure through
 genomic segmentation. Nat Methods 2012, 9:473-476.

43. Chan RCW, Libbrecht MW, Roberts EG, Bilmes JA, Noble WS, Hoffman MM:
Segway 2.0: Gaussian mixture models and minibatch training. Bioinformatics
2018, 34:669-671.

44. Human ENCODE Data Analysis Consortium Blacklisted Regions
[http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz]

45. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing
genomic features. Bioinformatics 2010, 26:841-842.
Fig. 1

(a) Schematic representation of nucleosomes and accessible DNA regions.

(b) Workflow of data preprocessing and analysis:
- Alignment of ATAC-Seq read pairs
- Data preprocessing
- Training sites selection
- Insertion length analysis (EM)
- Signal decomposition
- Signals of NFR/mono-/di-/tri-nucleosome
- Model training (Baum-Welch)
- 3 states HMM of accessible DNA
- Genome annotation (Viterbi)
- Postprocessing
- Calls of accessible regions

(c) Iteration of EM analysis for GM12878 50k and 500 cells.

(d) HMM from ATAC-seq of GM12878 50k cells:
- Transition and emission matrices for states 1, 2, and 3.
- Columns showing log10(probability) and log10(pileup) values.

Legend:
- State 1: Background
- State 2: Nucleosome
- State 3: Center
- NFR: Nucleosome-Free Region
- 1N, 2Ns, 3Ns: Mono-, Di-, Tri-nucleosome

Graphs show basepair distributions and log10(pileup) values for different states and samples.
Fig. 3

(a) Histone marks and Chromatin states

(b) Distribution of H3K4me1 ChIP-seq signal

(c) Distribution of PhastCons score
