HMMRATAC: a Hidden Markov ModeleR for ATAC-seq.

Evan D. Tarbell[1] and Tao Liu[1]*

[1] Department of Biochemistry, University at Buffalo, Buffalo, NY, 14203, USA
* To whom correspondence should be addressed. Tel: 716-829-2749; Fax: 716-849-6890; Email: tliu4@buffalo.edu

ABSTRACT

ATAC-seq has been widely adopted to identify accessible chromatin regions across the genome. However, current data analysis still utilizes approaches initially 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.

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(2). 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(3).

Several assays exist to identify open chromatin regions in a genome-wide manner. These include DNase-seq, which utilizes the DNase I nuclease(4), FAIRE-seq, which utilizes differences in polarity between nucleosome-bound and nucleosome-free DNA(5), and ATAC-seq, which uses a transposase to cut into accessible DNA selectively(6). Although each of these assays identifies some unique open chromatin regions, they are generally highly correlated in their identifications(6). Whereas DNase-seq and FAIRE-seq are complex 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(6,7). ATAC-seq has become popular over the years, and the Cistrome Database(8), 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. 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(6), 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 fact 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(9), that can be and are used for ATAC-seq analysis, such as MACS2(10) and F-Seq(11), 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 recognition. Our method takes advantage of the unique features of ATAC-seq to identify the chromatin structure more accurately. 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 using aligner such as BOWTIE2(12) or BWA(13), followed by identification of accessible regions or “peaks” by HMMRATAC, and then downstream analysis such as motif enrichment using MEME(14) or footprint identification in the accessible peaks with

CENTIPEDE(15); differential accessibility analysis with Diffbind(16); and association studies with other data sets, such as with gene expression data with BETA(17). 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(18). We envision HMMRATAC becoming the principle peak-calling method in such a pipeline.

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(19) accession numbers SRR891269-SRR891274. 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(12). 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. F-Seq requires a single-end BED file of 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(20) was publicly available and downloaded from the Cistrome Database(8) 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, including merging and filtering on the BAM files.

The HMMRATAC algorithm

The HMMRATAC algorithm is built upon the idea of “decomposition and integration”, and based on the observation of distinct nucleosome organization at accessible chromatin (see **Results**, **Figure 1A,** and **Supplemental Figure S1**). A single ATAC-seq dataset is decomposed into

different layers of signals according to the sizes of DNA fragments digested by Tn5 transposase, and the patterns of the multiple layers of information at the open chromatin regions are learned and utilized for predicting accessible regions across the whole genome(**Figure 1B**). The detail steps of HMMRATAC algorithm is described as follows.

Segregation of ATAC-seq signals

After the preprocessing step that eliminates duplicate reads and low mapping quality reads, the main HMMRATAC pipeline (**Figure 1B**) 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(6). 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. 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 (**Figure 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, 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(21,22) and Segway(23,24). The three clusters represent the three states of our model, corresponding to the *center* of the open region with overall high enrichment of all of the four types of signals, the *nucleosomes* with low NFR signal but moderate nucleosomal signals, and a *background* state with low enrichment of any signals, 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(25) until convergence (**Figure 1D**).

Annotating with the Viterbi algorithm

Once the HMM has been created and refined, the Viterbi algorithm(26) 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 provided by the ENCODE project. 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 (downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz). 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 *center* state. We generally consider the nearby nucleosomes to be part of the regulatory region (See **Results**), and therefore the region is then reported as a peak in gappedPeak format (https://genome.ucsc.edu/FAQ/FAQformat.html#format14). Each reported region is given a score, which represents the maximum read coverage of the *center* state. Additionally, HMMRATAC also supplies a file containing the summits of each region. To calculate a summit, HMMRATAC first filters the read coverage signal throughout 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.

Software implementation

HMMRATAC software is implemented using JAVA language. It should be noted that at all stages of the algorithm, HMMRATAC 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 HMMRATAC. However, because of this feature, it is not recommended to run HMMRATAC 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 component is to have Java 7 or higher installed. We have tested HMMRATAC on OpenJDK 7, standard edition JRE 8 and the standard edition JRE 10. In terms of hardware requirement, although all results involved in this study were generated on our server equipped with AMD Opteron(tm) Processor 6378 and 500GB memory, HMMRATAC can successfully process human ATAC-seq datasets on modest personal computers such as the author's laptop equipped with a 2.8GHz Intel Core i7 processor and 16 GB of DDR3 memory. The actual maximum memory usage and the runtime of HMMRATAC mainly depend on the parameter settings. As an example, when the window size for Viterbi decoding (option --window) was set to 2,500 kbps, the maximum memory usage decreased from 24GB, when using the default window size of 25,000 kbps, down to 13 GB, on the same merged GM12878 dataset.

Evaluation

Calls from MACS2

MACS version 2.1 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

F-Seq version 1.84 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 version 1.1 was used to call peaks using the sorted BAM file. 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. The blacklisted sites for hg19 were provided, 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 **Supplemental Table S2**.

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

We tested the performance of HMMRATAC and other methods on ATAC-seq datasets of GM12878 cells against three proxies of “gold standard” datasets. The first one, which we call “active regions”, is made by combining two chromatin states “Active Promoters” and “Strong Enhancers” generated in GM12878 cells using chromHMM(27,28) (downloaded from UCSC genome browser track wgEncodeBroadHmmGm12878HMM). The other two sets were DNaseI hypersensitive sites (UCSC track wgEncodeAwgDnaseUwdukeGm12878UniPk) and FAIRE

accessible regions (wgEncodeOpenChromFaireGm12878Pk)(29) called from DNase-seq and FAIRE-seq on the GM12878 cell line. Because the DNase-seq and FAIRE-seq regions were called by the ENCODE project(30) from traditional methods including Hotspot(31) and F-Seq, the evaluation based on either of them is less unbiased compared with the test based on chromatin states which integrate multiple independent experiments. We found that there were numerous cases throughout the genome where clusters of DNaseI or FAIRE regions were in close proximity to one another. These clusters tended to have broad enrichment in either DNase-seq or FAIRE-seq signals. Therefore, we merged all DNaseI or FAIRE regions that were within 2KB of each other into a single region and used these merged sets as the proxies to the gold standard.

The "real negative" set was made with the "Heterochromatin" state annotation from chromHMM. When using the "active region" set as "gold standard", no modifications were necessary because the three states are mutually exclusive. When using either the DNaseI or FAIRE sets, any "heterochromatin" states that overlapped (by any amount) a DNaseI or FAIRE region were excluded from the final "real negative" set. Additionally, any DNaseI 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 human monocyte ATAC-seq data were compiled as follows. For each condition, 2 replicates of H3K27ac, H3K4me1 and H3K4me3 ChIP-seq datasets, one replicate of H3K9me3 from the same study(20) were downloaded from the Cistrome Database(8) 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 kept. The "active regions" true set was created as the genomic regions overlapped with either H3K27ac, H3K4me1 or H3K4me3 but not overlapped with H3K9me3 peaks. This list of "active regions" became the proxy of "gold standard" set for each 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 **Supplemental Table S3**.

Calculating Precision, Recall, False Positive Rate and Area Under the Curves

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. Let Real Negatives (RN) represent the number of base pairs in the "real negative" set. 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 the *precision* or positive predictive value (PPV) is calculated as PPV = TP/(TP+FP). The *Recall*, or true positive rate (TPR), is calculated as TPR = TP/RP. The *false positive rate* (FPR) is calculated as FPR = FP/RN. The number of overlapping base pairs was calculated with the BEDTools intersect tool(32). After the precisions, recalls and false positive rates were calculated for accessible regions predicted by a given method under various cutoff values, the approximate area under the curve (AUC) was calculated for the *recall* vs. *false positive rate* (ROC) curves and for the *precision-recall* (PR) curves, after adding extreme points to the unreachable ends of the curves. The extreme points to the ROC curves are TPR=0 and FPR=0, and TPR=1 and FPR=1; the extreme points to the PR curves are TPR=0 and PPV=1, and TPR=1 and PPV=L_{true}/(L_{true} + L_{false}) where the L_{true} is the total length of real positive set and the L_{false} is the total length of the real negative set.

RESULTS

The distinctive pattern of ATAC-seq fragments around known accessible elements

In order to investigate the ATAC-seq signal profile around accessible genomic regions, we took the ATAC-seq data in the human GM12878 cell line from Buenrostro et al. (6)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(6). We plotted the two types of fragments around the centers of DNaseI hypersensitive sites (DHSs) in the same cell line, identified by ENCODE(30) (**Supplemental Figure 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 fragments (**Figure 1A; Supplemental Figure S1**). Based on this observation and existing knowledge on the nucleosome positioning around the accessible chromatin (2), we hypothesized that such pattern would exist at the open chromatin throughout the genome and that by learning and recognizing this pattern computationally in a single ATAC-

seq dataset, we would identify the 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(6). 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 adequately understood(33,34). 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 (**Supplemental Figure S2**). 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. As an example, HMMRATAC identified the optimum fragment lengths representing 1Ns, 2Ns, and 3Ns from ATAC-seq data in the human GM12878 cell line(6), generated by merging three replicates, as being 195bp, 396bp, and 693bp, respectively (**Figure 1C**). 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(20) as 186bp, 368bp, and 549bp, respectively. Additionally, the values for any of the individual replicates, in the GM12878 data, were never more than 5bp off from the average values listed above for the merged datasets (**Supplemental Figure 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 (**Supplemental Figure S4A and S4B**). However, we found our approach can robustly identify the parameters even if the multi-modal pattern is not clear. As we can see from **Figure 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 datasets.

Secondly, we propose in HMMRATAC a probabilistic approach to make use of data from the entire spectrum of fragment lengths. We generate four distinct 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 (**Supplemental Figure S2**). We found that our probabilistic method had considerably better recall and precision compared to the cutoff based method. This result proved that the traditional way to discard reads between cutoffs of fragment sizes would hurt the prediction power.

HMMRATAC utilizes the length of an insertion fragment, identified by paired-end sequencing, as an essential 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(10) to call peaks with each data set, including the original unaltered data and compared the performance to predict the accessible chromatin (**Supplemental Figure S5**). 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, 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* (**Figure 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(25). 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** for detail). **Figure 1D** shows the final parameters for the HMM on ATAC-seq in the GM12878 cell line, by merging three replicates from 50k cells(6). 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 mostly 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 location with one of the states using the Viterbi algorithm(26). 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 human genome), about 4.5 million regions annotated as *nucleosomes* with an average size of 382 bps (~49% of human genome), and about 226 thousand regions annotated as *center regions* with an average size of 504 bps (~3% of human genome). 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 human genome), excluding those with abnormally high coverage and those below a minimum length. 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(30), including

ChIP-seq for CTCF, active histone marks such as H3K4me1, H3K4me3, H3K9ac, and H3K27ac, and repressive histone marks such as H3K9me3 and H3K27me3, as well as sequencing of micrococcal nuclease digestion (MNase-seq). **Figure 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 of active histone marks and MNase digestion. We next plotted the same ChIP-seq, DNase-seq and MNase-seq data around all distal (> 10kb from any gene promoter) *center* states and used k-means clustering to group the resulting profiles (**Figure 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 weak or poised enhancers (**Figure 2B**). The second cluster (C2), characterized by strong enrichment of CTCF binding, lower levels of H3K4me1, H3K4me3 and acetylation histone marks and a higher level of H3K27me3, seems to suggest that this cluster represents insulators or domain boundaries. The fourth cluster (C4), similar to C2, suggested accessible regions also exist in the DNA elements with repressive functions in euchromatin, potentially bound by factors rather than CTCF. To summarize, these results indicate that HMMRATAC recapitulates 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 these methods called the site 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 concluded from a single ATAC-seq data, we plotted the overlap of the 3 hidden states with chromHMM states and histone mark calls from ChIP-seq in the GM12878 cell line based on ENCODE data (**Figure 3A; Supplemental Table S1**). We found that the chromatin states “Active Promoters”, “Strong Enhancer”, “Insulator”, “Poised Promoter” and

"Weak Promoter" states mainly overlap 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 (**Figure 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(35). Success in transcription factor ChIP-seq data analysis is often benchmarked by high PhastCons score, a precompiled evolutionary conservation score at every base pair(36), at the peak center or summit, relative to the surrounding regions(37). We found that the summits of accessible regions detected by HMMRATAC in GM12878 cell line also showed PhastCons enrichment in vertebrates (**Figure 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 methods for ATAC-seq. We chose MACS2(10) and F-Seq(11), which had previously been shown to be the most accurate peak-callers in identifying open chromatin regions from DNase-seq(38). 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(7,20,39-43), and MACS2 and F-Seq is part of the ENCODE analysis pipeline for ATAC-seq data (https://github.com/kundajelab/atac_dnase_pipelines). We tested the

performance of the four algorithms using the two ATAC-seq datasets in GM12878 cell line(6) as the previous sections.

A major limitation in comparing the performance of identifying open chromatin regions is the absence of a valid "gold standard" (3). 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 independent histone features (27), 2) DNase-seq (4) and 3) FAIRE-seq (5) accessible regions of the same GM12878 cell line, originally called from traditional methods such as Hotspot (31) and F-Seq.

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(27,28) algorithm conducted by the 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" (**Figure 4** top row). The area under the curve (AUC) was then calculated for the recall vs. false positive rate curves (AUC-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 **Supplemental Figure S6**, at comparable levels of sensitivity for "active 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) (4) and FAIRE-seq (5) accessible regions called by ENCODE in the GM12878 cell line(30), then merged DHSs or FAIRE accessible regions that were within 2kb of each other to define our alternative true sets. 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 AU-PRC in recapitulating DHSs (**Supplemental Figure 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 AU-PRC with the FAIRE-seq data. Because the DNase-seq and FAIRE-seq regions were initially been called from traditional methods such as Hotspot(31) and F-Seq, the evaluations based on them are expected to be more biased to traditional approaches.

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(20). 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 (**Figure 4** the second to the sixth row, **Table 1**). We found that HMMRATAC has superior precision, sensitivity, specificity, AUC-ROC and AU-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 AU-PRC. We found that HMMRATAC outperformed the other two methods in 6 of 8 AUC-ROC and 7 of 8 AU-PRC. 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 consistently. 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 specific 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(44), 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(45,46) 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.

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 initially 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.

DATA AVAILABILITY

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.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

FUNDING

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

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.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interests.

REFERENCES

1. Li, G. and Reinberg, D. (2011) Chromatin higher-order structures and gene regulation. *Curr Opin Genet Dev*, **21**, 175-186.
2. Bell, O., Tiwari, V.K., Thoma, N.H. and Schubeler, D. (2011) Determinants and dynamics of genome accessibility. *Nat Rev Genet*, **12**, 554-564.
3. Tsompana, M. and Buck, M.J. (2014) Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*, **7**, 33.
4. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311-322.
5. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res*, **17**, 877-885.

6. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, **10**, 1213-1218.

7. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y. and Greenleaf, W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486-490.

8. Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res*, **45**, D658-D662.

9. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823-837.

10. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.

11. Boyle, A.P., Guinney, J., Crawford, G.E. and Furey, T.S. (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537-2538.

12. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357-359.

13. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

14. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, **37**, W202-208.

15. Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y. and Pritchard, J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res*, **21**, 447-455.

16. Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389-393.

17. Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C.A., Zhang, Y. and Liu, X.S. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat Protoc*, **8**, 2502-2515.

18. Qin, Q., Mei, S., Wu, Q., Sun, H., Li, L., Taing, L., Chen, S., Li, F., Liu, T., Zang, C. *et al.* (2016) ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics*, **17**, 404.

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

20. Novakovic, B., Habibi, E., Wang, S.Y., Arts, R.J.W., Davar, R., Megchelenbrink, W., Kim, B., Kuznetsova, T., Kox, M., Zwaag, J. *et al.* (2016) beta-Glucan Reverses the Epigenetic State of LPS-Induced Immunological Tolerance. *Cell*, **167**, 1354-1368 e1314.

21. Sohn, K.A., Ho, J.W., Djordjevic, D., Jeong, H.H., Park, P.J. and Kim, J.H. (2015) hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*, **31**, 2066-2074.

22. Ho, J.W., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A., Tolstorukov, M.Y., Appert, A. *et al.* (2014) Comparative analysis of metazoan chromatin organization. *Nature*, **512**, 449-452.

23. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A. and Noble, W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods*, **9**, 473-476.

24. Chan, R.C.W., Libbrecht, M.W., Roberts, E.G., Bilmes, J.A., Noble, W.S. and Hoffman, M.M. (2018) Segway 2.0: Gaussian mixture models and minibatch training. *Bioinformatics*, **34**, 669-671.

25. Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970) A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. Math. Statist.*, **41**, 164-171.

26. Viterbi, A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**, 260-269.

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

28. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43-49.

29. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75-82.

30. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

31. John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager, G.L. and Stamatoyannopoulos, J.A. (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet*, **43**, 264-268.

32. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
33. Lantermann, A.B., Straub, T., Stralfors, A., Yuan, G.C., Ekwall, K. and Korber, P. (2010) Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. *Nat Struct Mol Biol*, **17**, 251-257.
34. Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z. and Sidow, A. (2011) Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516-520.
35. Liu, Y., Liu, X.S., Wei, L., Altman, R.B. and Batzoglou, S. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res*, **14**, 451-458.
36. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034-1050.
37. Shin, H., Liu, T., Duan, X., Zhang, Y. and Liu, X.S. (2013) Computational methodology for ChIP-seq analysis. *Quant Biol*, **1**, 54-70.
38. Koohy, H., Down, T.A., Spivakov, M. and Hubbard, T. (2014) A comparison of peak callers used for DNase-Seq data. *PLoS One*, **9**, e96303.
39. Hansel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat Genet*, **48**, 1267-1272.
40. Denny, S.K., Yang, D., Chuang, C.H., Brady, J.J., Lim, J.S., Gruner, B.M., Chiou, S.H., Schep, A.N., Baral, J., Hamard, C. *et al.* (2016) Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. *Cell*, **166**, 328-342.
41. Philip, M., Fairchild, L., Sun, L., Horste, E.L., Camara, S., Shakiba, M., Scott, A.C., Viale, A., Lauer, P., Merghoub, T. *et al.* (2017) Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature*, **545**, 452-456.
42. Chen, X., Shen, Y., Draper, W., Buenrostro, J.D., Litzenburger, U., Cho, S.W., Satpathy, A.T., Carter, A.C., Ghosh, R.P., East-Seletsky, A. *et al.* (2016) ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat Methods*, **13**, 1013-1020.
43. Risca, V.I., Denny, S.K., Straight, A.F. and Greenleaf, W.J. (2017) Variable chromatin structure revealed by in situ spatially correlated DNA cleavage mapping. *Nature*, **541**, 237-241.
44. He, H.H., Meyer, C.A., Hu, S.S., Chen, M.W., Zang, C., Liu, Y., Rao, P.K., Fei, T., Xu, H., Long, H. *et al.* (2014) Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods*, **11**, 73-78.

45. Wang, J.R., Quach, B. and Furey, T.S. (2017) Correcting nucleotide-specific biases in high-throughput sequencing data. *BMC Bioinformatics*, **18**, 357.
46. Martins, A.L., Walavalkar, N.M., Anderson, W.D., Zang, C. and Guertin, M.J. (2018) Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic Acids Res*, **46**, e9.

TABLE AND FIGURES LEGENDS

Table 1. The Area Under the Curves (AUCs) of method comparisons of HMMRATAC, MACS2, and F-Seq.

The AUC calculated from the Receiver Operator Characteristic (ROC) curves and precision-recall curves (PRCs) in **Figure 4** and **Supplemental Figure S7**. Because we can't reach the extreme sides of the curves while tuning cutoff values, before we calculated AUC values, we added the following points to ROC curves– (0, 0) and (1, 1), and to PRCs – (0, 1) and (1, L_{true}/(L_{true} + L_{false})). The L_{true} is the total length of the real positive set, and the L_{false} is the total length of the real negative set. The numbers of the best performance are shown in bold.

Figure 1. The HMMRATAC Algorithm.
(A) A schematic to represent the structure of an accessible region. **(B)** The algorithmic workflow. **(C)** Estimation of fragment length parameters through the Expectation-Maximization algorithm. Mean and standard deviation parameters per EM iteration for three nucleosomal distributions are consistent among datasets using 50,000 and 500 cells. The parameters for NFR signals are fixed during EM training so they were not plotted. Iteration #1 shows the initial parameters. EM was applied to pooled biological replicates. The parameters per biological replicate were shown in **Supplemental Figure S3**. **(D)** The transition and emission parameters of the Hidden Markov Model trained through Baum-Welch algorithm on the 50,000 GM12878 cells ATAC-seq data.

Figure 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* state, green: *nucleosome* state, 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**). **(B)** ChIP-seq of H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K9ac, and H3K9me3, DNase-seq and MNase-seq signals (log2 transformed library size normalized pileup) plotted across all distal (> 10kb from gene promoters) *center* states. Data were clustered using k-means algorithm into 4 clusters.

Figure 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.

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

Left: Receiver Operator Characteristic (ROC) curves and right: precision-recall curves (PRCs) 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 **Methods** for detail).

Table 1

	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									
	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	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 1

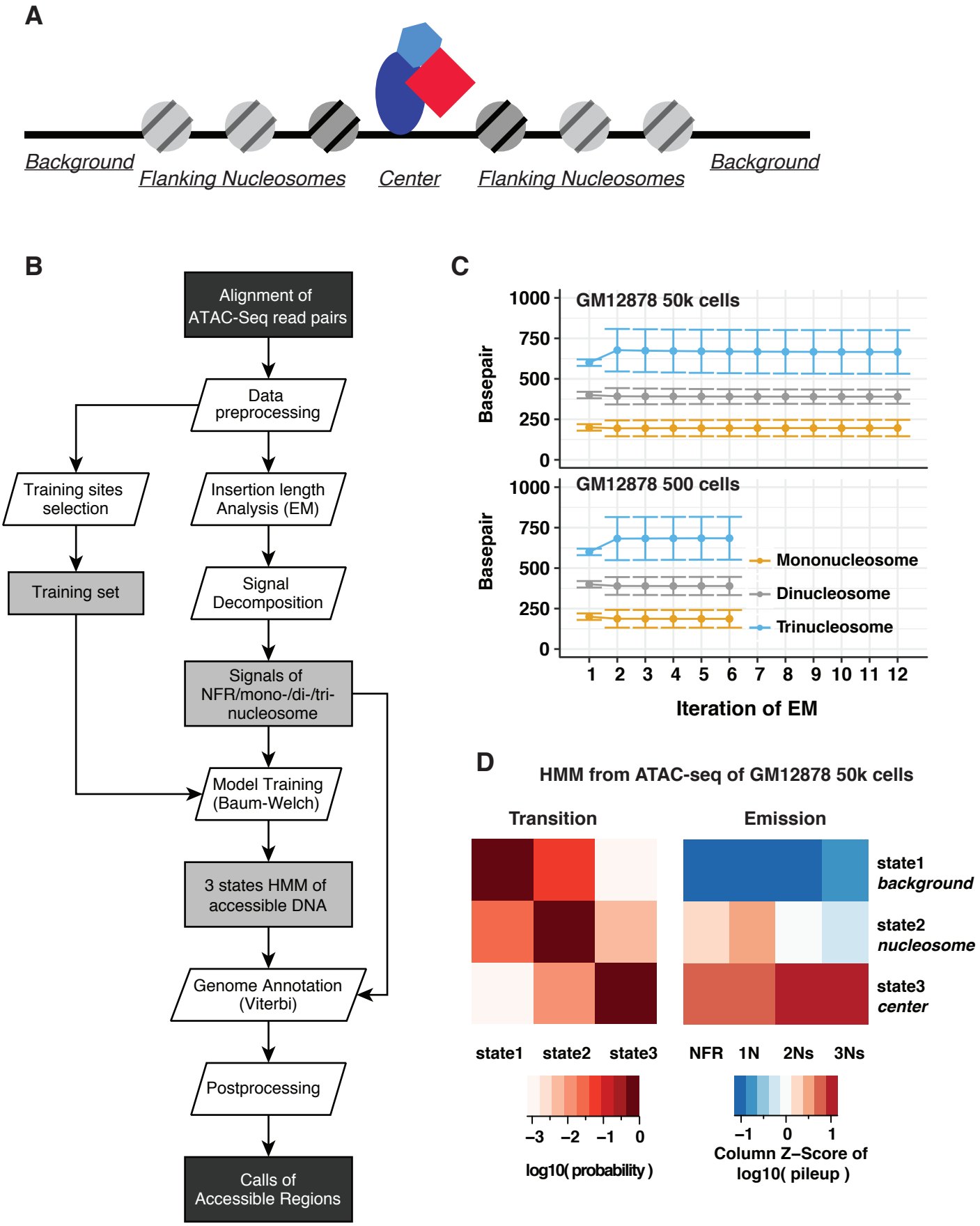

A
Background
Flanking Nucleosomes
Center
Flanking Nucleosomes
Background
B
Alignment of ATAC-Seq read pairs
Data preprocessing
Training sites selection
Insertion length Analysis (EM)
Training set
Signal Decomposition
Signals of NFR/mono-/di-/tri-nucleosome
Model Training (Baum-Welch)
3 states HMM of accessible DNA
Genome Annotation (Viterbi)
Postprocessing
Calls of Accessible Regions
C
GM12878 50k cells
GM12878 500 cells
Basepair
1000
750
500
250
0
Mononucleosome
Dinucleosome
Trinucleosome
1 2 3 4 5 6 7 8 9 10 11 12
Iteration of EM
D
HMM from ATAC-seq of GM12878 50k cells
Transition
Emission
state1 background
state2 nucleosome
state3 center
state1 state2 state3
NFR 1N 2Ns 3Ns
−3 −2 −1 0
log10(probability)
−1 0 1
Column Z−Score of log10(pileup)

Figure 2

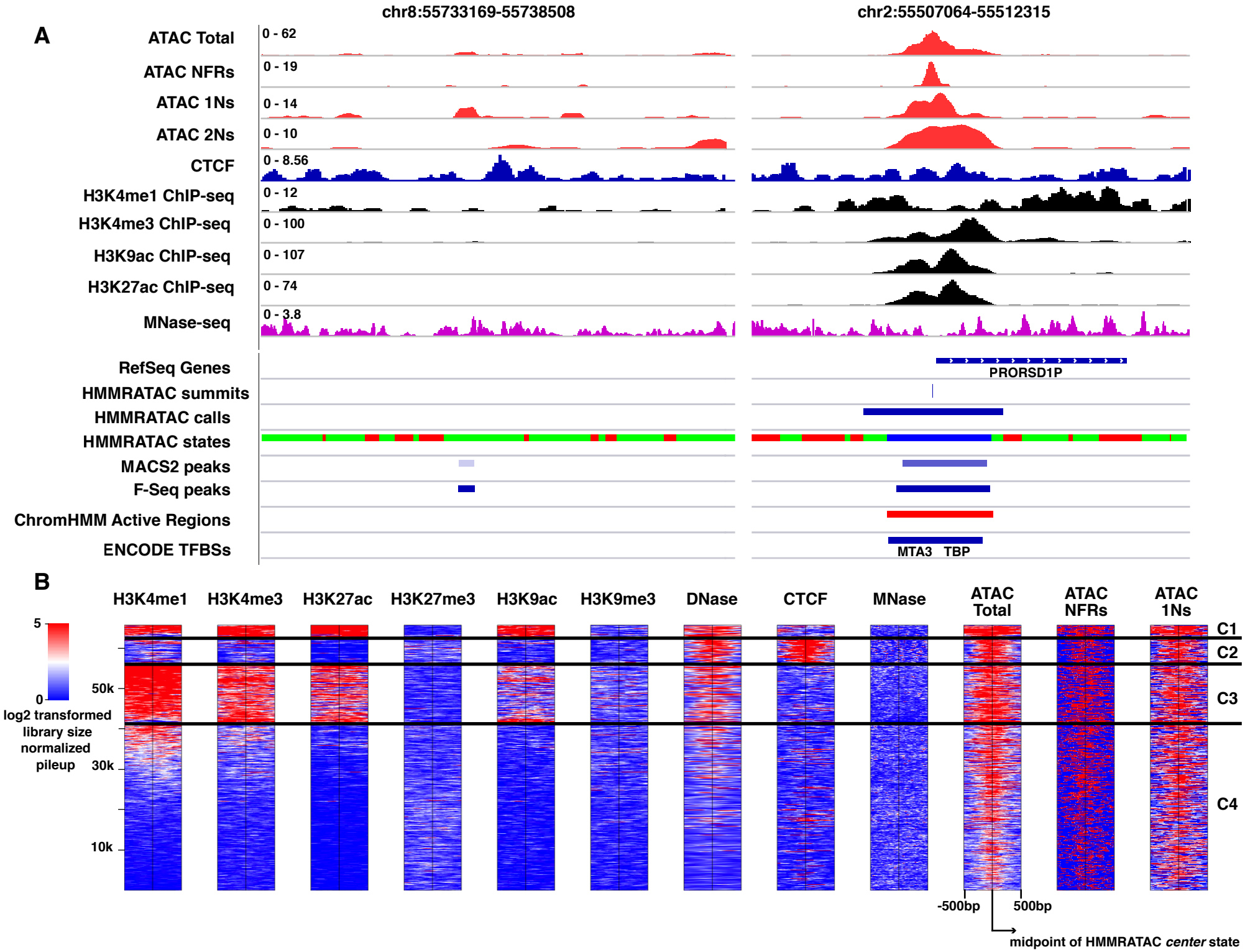

A
chr8:55733169-55738508
chr2:55507064-55512315
ATAC Total
ATAC NFRs
ATAC 1Ns
ATAC 2Ns
CTCF
H3K4me1 ChIP-seq
H3K4me3 ChIP-seq
H3K9ac ChIP-seq
H3K27ac ChIP-seq
MNase-seq
0 - 62
0 - 19
0 - 14
0 - 10
0 - 8.56
0 - 12
0 - 100
0 - 107
0 - 74
0 - 3.8
RefSeq Genes
HMMRATAC summits
HMMRATAC calls
HMMRATAC states
MACS2 peaks
F-Seq peaks
ChromHMM Active Regions
ENCODE TFBSs
PRORSD1P
MTA3
TBP
B
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9ac
H3K9me3
DNase
CTCF
MNase
ATAC Total
ATAC NFRs
ATAC 1Ns
5
0
log2 transformed library size normalized pileup
50k
30k
10k
C1
C2
C3
C4
-500bp
500bp
midpoint of HMMRATAC center state

Figure 3

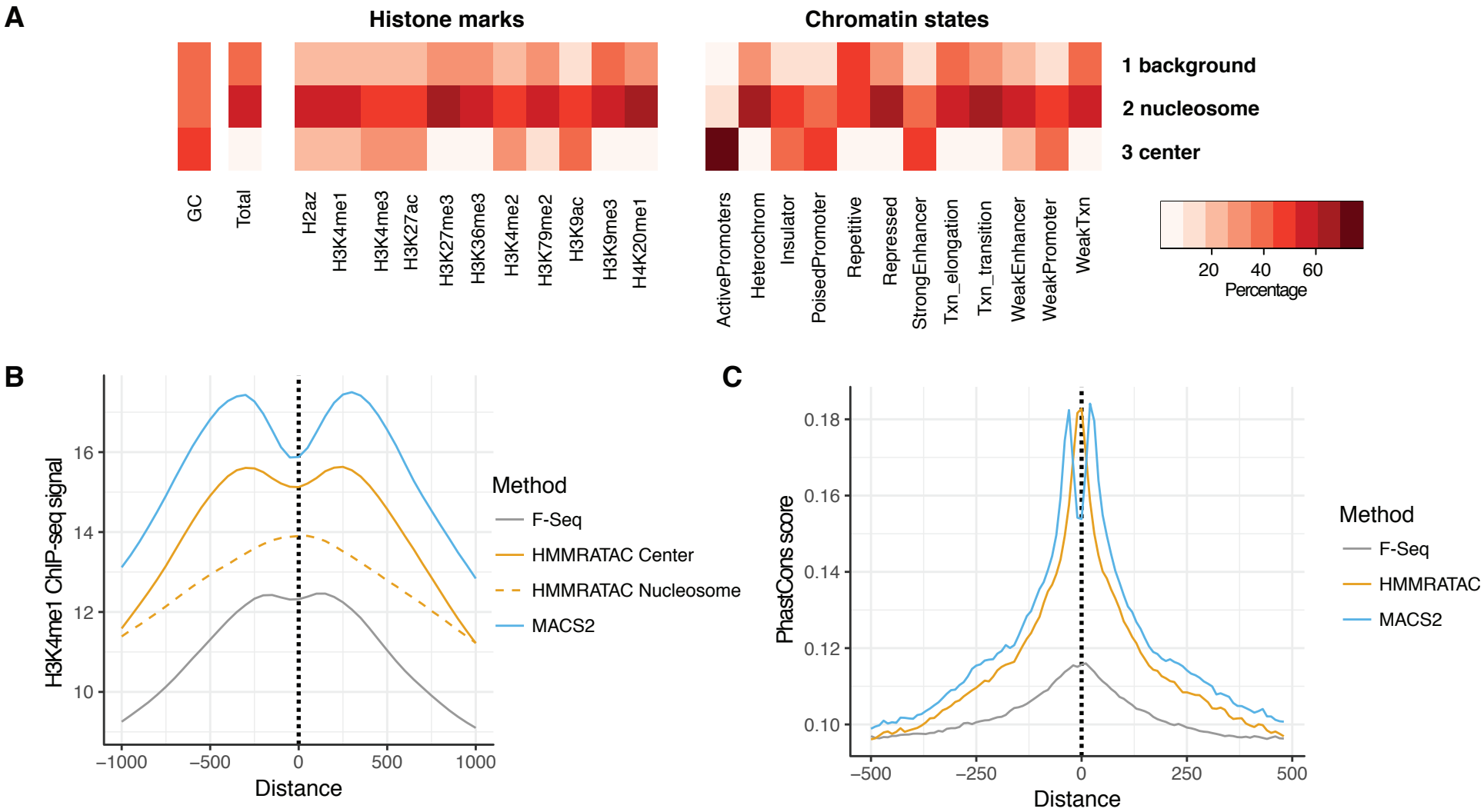
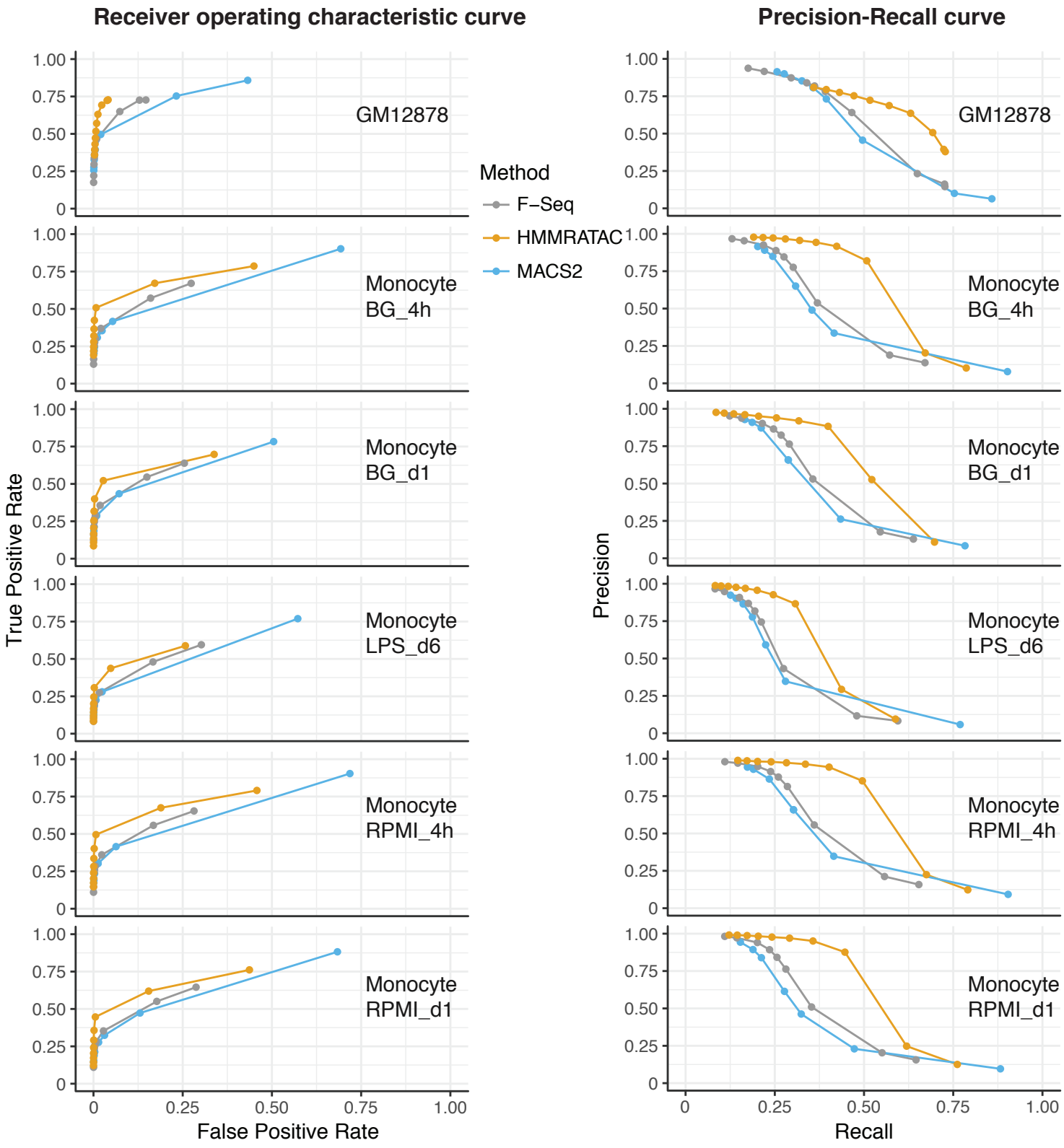

A
Histone marks
Chromatin states
1 background
2 nucleosome
3 center
GC
Total
H2az
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K36me3
H3K4me2
H3K79me2
H3K9ac
H3K9me3
H4K20me1
ActivePromoters
Heterochrom
Insulator
PoisedPromoter
Repetitive
Repressed
StrongEnhancer
Txn_elongation
Txn_transition
WeakEnhancer
WeakPromoter
WeakTxn
20
40
60
Percentage
B
H3K4me1 ChIP-seq signal
16
14
12
10
−1000
−500
0
500
1000
Distance
Method
F-Seq
HMMRATAC Center
HMMRATAC Nucleosome
MACS2
C
PhastCons score
0.18
0.16
0.14
0.12
0.10
−500
−250
0
250
500
Distance
Method
F-Seq
HMMRATAC
MACS2

Figure 4


Receiver operating characteristic curve
Precision-Recall curve
GM12878
Monocyte BG_4h
Monocyte BG_d1
Monocyte LPS_d6
Monocyte RPMI_4h
Monocyte RPMI_d1
Method
F−Seq
HMMRATAC
MACS2
True Positive Rate
False Positive Rate
Precision
Recall
0
0.25
0.50
0.75
1.00