HMMRATAC, The Hidden Markov ModeleR 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 where transposase insertion into NFRs and linker regions creates different modes. 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, 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.

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 and mono-nucleosome fragments based on cutoffs identified by Buenrostro et al. We plotted the two types of fragments around the centers of DNaseI hypersensitive sites (DHSs) in the same cell line, identified by ENCODE [9] (Supplementary Figure 1). 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 four distinct regions around an open site: 1) the *center*, characterized by an enrichment of nucleosome-free fragments and flanked by nucleosome fragments; 2) the *flanking regions*, characterized by an enrichment of nucleosome fragments and directly next to the center; 3) the *non-flanking regions*, characterized by lower levels of nucleosome fragments and directly next to the flanking regions; 4) and the *background*, which although not visible on the figure, would be characterized by a depletion in any type of fragments (Fig. 1a; Additional file 1: Fig. S1). This observation coincides with current knowledge that accessible chromatin actively bound by transcription factors is surrounded by highly positioned nucleosomes, then further by unpositioned nucleosomes [10]. 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.

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. 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 and in the reproducibility between replicates.

Results

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 [11, 12]. 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 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. Similar to 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 (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, respectfully. As a demonstration of different nucleosome spacing in different species and datasets, HMMRATAC determined the optimum fragment lengths for the 1Ns, 2Ns, and 3Ns from ATAC-seq data in mouse C2C12 myoblast cells, generated by merging two replicates from [13], as 174bp, 370bp and 623bp, respectfully. Additionally, the values for any of the individual replicates, in either human or mouse (three human replicates [5] and two mouse replicates [13]), were never more than 5bp off from the average values listed above for the merged datasets. 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. S3a and S3b) [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 and 1d, 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 (described in more detail below) (Additional file 1: Fig. S2). We found that our probabilistic method had a considerably better recall and precision compared to the cutoff based method.

Integration through Hidden Markov Model

After removing reads with low mapping quality, and masking blacklist regions from downstream analysis [14], HMMRATAC identifies a set of training regions throughout the genome to train a 4-states Hidden Markov Model (HMM) with multivariate Gaussian

emissions, representing the underlying four chromatin states of 1) the *center* of open chromatin, 2) the *flanking regions*, 3) the *non-flanking regions* and 4) 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 [15]. The initial state transition probabilities are set equally as 0.25 and the emission parameters for each state are calculated after clustering all locations in the training regions into 4 classes (See methods section for more detail). Table 1_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 4 in the table, which HMMRATAC annotated as the *center* state, showed the highest average signal across the four signal profiles. State 3, annotated as the *flanking region* state, showed the second highest average values. Additionally, as we hypothesized, the results show the *center* state mainly transit from or to the *flanking regions* 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 [16]. On the 50K cells ATAC-seq data in GM12878 cell line, there were about 4.5 million regions annotated as *background* with an average size of 254 bps (38% of

genome size), about 2.8 million regions annotated as *non-flanking regions* with an average size of 235 bps (22%), about 3.7 million regions annotated as *flanking regions* with an average size of 303 bps (37%), and about 166 thousand regions annotated as *center regions* with an average size of 494 bps (3%). HMMRATAC finally connected the *center regions* with their adjacent *flanking nucleosomes* and called a total of 93,276 accessible regions with an average size of 1,574 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 *flanking regions* do not actually flank 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 [9], 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 (MNase-seq) and DNaseI digestion (DNase-seq). Fig. 2a shows a locus on chromosome 1 identified as an accessible region by HMMRATAC. We observed that the *center* state was located in regions characterized by depletion in both active histone

marks and MNase digestion, and by enrichment in DNase digestion and CTCF signal. The surrounding *flanking nucleosome* states showed enrichment in active histone marks and MNase digestion, together with depleted DNase digestion and CTCF signal. 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 characterized by strong enrichment of H3K27ac relative to H3K4me1 seems to indicate active enhancers while the third cluster marked by high levels of H3K4me1 and more moderate levels of H3K27ac would suggest poised enhancers (Fig. 2b). The second cluster, 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.

Functional regulatory elements that harbor transcription factor binding sites are generally under higher evolutionary constraint [17]. Success in transcription factor ChIP-seq data analysis is often benchmarked by high PhastCons score, a precompiled evolutionary conservation score at every base pair [18], at the peak center or summit, relative to the surrounding regions [19]. We found that the summits of accessible regions detected by HMMRATAC in GM12878 cell line also showed PhastCons enrichment in vertebrates (Additional file 1: Fig. S4), indicating that these summits were likely locations of evolutionarily conserved regulatory elements and could represent functional regulatory elements.

Method comparison with MACS2, F-seq, and Hotspot2

We next sought to compare the performance of HMMRATAC in identifying open chromatin regions to other popular software for ATAC-seq. We chose MACS2 [20], F-seq [21] and Hotspot2 [22], which had previously been shown to be the most accurate peak-callers in identifying open chromatin regions from DNase-seq [23]. 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. Hotspot2, originally designed for DNase-seq, calculates enrichment in a narrow window relative to a larger local window using the binomial 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, 24-29], and MACS2 and F-seq is part of the ENCODE analysis pipeline for ATAC-seq data [30]. 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 heterochromatins from chromatin states analysis combining multiple histone modification

features[31], 2) DNase-seq [3] and 3) FAIRE-seq [4] accessible regions of the same GM12878 cell line. We also used replicate reproducibility and biological reproducibility to assess each methods' robustness, following the guidelines developed for ChIP-seq by the ENCODE and modENCODE consortiums [32].

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 [31, 33] 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. 3a top panel, Fig. 3b top panel). We also found that HMMRATAC outperformed the other algorithms in terms of F1-scores, the harmonic mean of precision and recall (Fig. 3c top panel). The area under the curve (AUC) was then calculated for the recall vs. false positive rate curves (AU-ROC)

and for the precision vs. recall curves (AU-PRC) (Table 2). HMMRATAC was proved to have better AUC than the three other methods. 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, F-seq or Hotspot.

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 [9, 34], 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 three methods in terms of precision, recall, false positive rate, and F1-score in recapitulating DHSs (Fig. 3a, 3b, and 3c middle panels) as well as in AU-ROC and AU-PRC (Table 2). As for FAIRE regions, we found that HMMRATAC outperformed the other methods in recall vs. false positive rate, and the corresponding AU-ROC for FAIRE accessible regions (Fig. 3a bottom panel, Table 2). F-seq was found to perform the best in terms of F1-score and AU-PRC with the FAIRE-seq data (Fig. 3b, and 3c bottom panel, Table 2).

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) [35]. 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 (Fig. 4a), HMMRATAC produced a higher average number of reproducible accessible regions than the other three algorithms. We found that Hotspot2 performed well in terms of reproducibility while MACS2 and F-seq underperform both HMMRATAC and Hotspot2 (Table 2)

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 more reproducible accessible regions, with IDR <= 0.05, using the different datasets than any of the other algorithms (Fig. 4b). Once again, Hotspot performed well and identified only slightly fewer reproducible accessible regions than HMMRATAC, while MACS2

and F-seq identified fewer (Table 2). These data indicate that HMMRATAC outperforms Hotspot, MACS2, and F-seq in identifying statistically reproducible accessible regions.

The summary of method comparisons is shown in Table 2. We used a total of 8 independent measures to benchmark the performance of the four algorithms, including AUC-ROC, AUC-PRC, and the numbers of reproducible regions as described before. We found that HMMRATAC outperformed the other three methods in 7 of those 8 benchmarks. F-seq outperformed the other methods in 1 of the 8 benchmarks. F-seq also came in second in 5 of the 6 AUC calculations, although it performed the worst in terms of the reproducibility benchmarks. Hotspot2 performed very poorly at all of the AUC calculations but came in second in both reproducibility benchmarks. MACS2 performed very similar to F-seq in all benchmarks but never ranked higher than second in any of them. Overall, these results indicate that HMMRATAC is the most accurate and reliable peak-caller for ATAC-seq data analysis.

Application for mouse cells

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, different cell lines, and different species, we analyzed ATAC-seq data generated in C2C12 myoblast mouse cells [13]. In addition to ATAC-seq, Dell'Orso et al. performed ChIP-seq on the same cells for the active histone mark histone H3K27ac and the repressive histone mark histone H3K27me3. Using the

active histone modification as the “true set” and the repressive histone modification as the “false set”, we compared the performance of the algorithms in the same way we compared their performance in human data (Additional file 1: Fig. S5). We show that HMMRATAC has superior precision, sensitivity, and specificity over both MACS2 and F-seq. Hotspot2, for reasons unknown, was unable to process this data set. Overall, this data indicates that HMMRATAC's superior performance over the other algorithms is consistent across different data sets, different species, and different cell lines.

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. 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.

The Tn5 transposase has an insertion preference for specific DNA sequences, as do nucleases used in similar or complementary assays such as DNase-seq or MNase-seq. We believe that such biases would have little effect on the performance and accuracy of HMMRATAC for the following reasons. First, because HMMRATAC considers the

structure at regulatory chromatin, regions with an increase in transposition frequency caused by a sequence preference and not by a depletion of nucleosomes, would be unlikely to possess the characteristic pattern of nucleosome-free and nucleosome-enriched signals and would not be identified as accessible by HMMRATAC. Second, due to the smoothing strategy during peak calling, either by HMMRATAC or other algorithms and the requirement for minimum peak size, such localized enrichment events are unlikely to influence the peak call.

Identifying transcription factor footprints, wherein the DNA protected by a binding protein shows resistance to nonspecific DNA digestion, is an area of active research. However, we don't believe that the conventional strategies for footprint identification, i.e. looking for a small window of depletion within a broader region of enrichment of cutting events, are reasonable for ATAC-seq. First, the contamination from mitochondrial DNA is a major concern in the common ATAC-seq protocol and it varies with different cell types, so to identify fine features such as footprints requires higher sequencing depth from ATAC-seq than that from other assays. Moreover, although the sequence bias of the Tn5 transposase would have little effect on identifying open regions, it will have a big impact on footprints. 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.

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 a differential accessibility analysis pipeline.

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 with 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. We have also shown that HMMRATAC shows better reproducibility among replicates, as well as in datasets generated from a different number of cells. As HMMRATAC is a cross-platform and user-friendly algorithm, we envision it becoming the standard for ATAC-seq data analysis, replacing current methods based on ChIP-seq analysis.

Methods

Preprocessing of ATAC-seq data

The human GM12878 cell line ATAC-seq paired-end data used in this study was publically available and downloaded under six SRA [36] 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 [37]. 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 and Hotspot. 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 mouse data was publically available and downloaded under the accession numbers SRR2999996 and SRR2999997 [13]. This data was generated from 50,000 C2C12 myoblast cells. Two replicates of ATAC-seq data were aligned to the mm10 reference genome using bowtie2 [31]. All pre-processing steps were identical to the human processing pipeline described above, including merging and 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 fragments (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. 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. 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 and 1d). 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 (Fig. 1b).

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 upper and lower limit of fold changes, as used in 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 [38], by separating the training regions into 4 clusters and calculating the means and covariance matrices for the four signal tracks for each cluster, thereby modeling the emissions as a multivariate Gaussian distribution. These four clusters represent the four states of our model, corresponding to the *center* of the open region, the *flanking nucleosomes*, the outer *non-flanking nucleosomes* and a *background* state, respectively. It is also possible to create a

model using a different number of states instead of 4, 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 [15] until convergence.

Annotating with the Viterbi algorithm

Once the HMM has been created and refined, the Viterbi algorithm [16] is used to classify every genomic position into one of the four 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 from 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. 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 open 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, and therefore the region is then reported as a peak in gappedPeak format.

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.

Post-processing and output

When the report peak option is set for HMMRATAC (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 zscore 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:

PPV = TP/TP+FP

Recall, or true positive rate (TPR), is calculated as:

TPR = TP/RP

False positive rate (FPR) is calculated as:

FPR = FP/RN

The F1-score is the harmonic mean of precision and recall and is calculated according to the equation:

F1 = 2 * ((precision * recall)/(precision + recall)).

The number of overlapping base pairs was calculated with the BEDTools intersect tool [39] 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 [31, 33] (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) [9, 34]. 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 mouse data were H3K27ac ChIP-seq accessible regions from the C2C12 myoblast cell line [13]. The raw sequencing data was downloaded under the SRA accession numbers SRR2999984-5, aligned to the mm10 reference genome using bowtie2 (using default parameters as with all alignments performed in this study). The resulting aligned reads were merged and broad peaks were called with MACS2 using the parameters –g mm, --broad, --nomodel, --extsize 150 and –p 0.12.

The “real negative” set for the mouse data were H3K27me3 peaks from the same cell type [13]. The raw sequencing reads were downloaded under the SRA accession number SRR2999992, aligned to the mm10 reference genome with bowtie2 and processed into broad peaks with MACS2 using the parameters –g mm, --broad, --nomodel, --extsize 150 and –p 0.03.

The p-value parameter for MACS2 broad peak calling, for both histone marks, were determined by setting multiple values and then calculating the number of peaks called versus the parameter value. The point where the number of peaks called began to rapidly increase or decrease relative to the parameter value was considered the best value for broad peak calling (data not shown).

When calling broad peaks from MACS2, for either histone mark, an input file was used. This data was downloaded under the accession number SRR3139303 and aligned to the mm10 genome using bowtie2. The aligned read file was then used as an input control (-c option).

Calls from MACS2, F-Seq, Hotspot, and HMMRATAC

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.

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.

Version 4.1 of Hotspot was used to call peaks from the BAM files created as described previously. Hotspot was run using the shell script provided in the release (runhotspot) with parameters laid out in the token file (runall.token.txt). This script automatically calls several programs in the Hotspot pipeline and outputs a peak BED file. The tag length was set to 50bp (_K_ = 50) and duplicate tags were allowed (_DUPOK_ = T) (considering that they had been removed previously, this setting has little effect). The FDR settings (_FDR_) used were 0.1, 0.05, 0.01 (default) and 0.005. It should be noted that the Hotspot pipeline would not finish or report completed peak BED files at higher or lower cutoffs. Mappable files were downloaded from hotspot website [40] for both mm10 and hg19 and used by the hotspot pipeline (_MAPPABLE_FILE_). Blacklisted regions for hg19 were also used to mask these regions from peak calling [9, 14]. Additionally, Hotspot was unable to process the mouse data with any setting.

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 [9, 14], 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 and 40. Additionally, HMMRATAC was run using non-default settings where the upper limit (option –u) was set to 10 and the lower limit (option –l) was set to 2. The minimum scores of 0 and 5 filtered these peaks.

Calculating reproducibility

Reproducibility was determined using the Irreproducible Discovery Rate (IDR) originally described by Li et al. [35]. Peaks were called using the following parameters: the q-value for MACS2 was set to 0.87 for 50,000 cell per-replicate data and 0.89 for 500 cell per-replicate data, the standard deviation for F-seq was set to 6 for both datasets, Hotspot FDR values were set to 0.05 for both datasets and the fold-change range for HMMRATAC was set to 2-10 for both datasets. The parameters for MACS2 and F-seq were determined by setting multiple values and then calculating the number of peaks called versus the parameter value. The point where the number of peaks called began to rapidly increase or decrease relative to the parameter value was considered the best value for IDR analysis (data not shown).

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. The output files were sorted by their local IDR value and plotted as ranked peaks versus IDR score.

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 and plotted.

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; AUC: area under the curve; 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, FSeq, Hotspot2, and MACS2, as well as the various "gold standard" files, can be downloaded from the following URL: http://biomisc.org/download/HumanData.tar.gz and http://biomisc.org/download/MouseData.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.

Tables

Table 1. Hidden Markov Model trained from GM12878 ATAC-Seq data.

Numbers show transition probabilities from each state to other states, and emission probabilities of NFR, 1Ns, 2Ns, and 3Ns signals for each state.

State index	**1**	**2**	**3**	**4**
Annotation	**Background**	**Non-flanking**	**Flanking**	**Open**
Transition to State 1	0.942	0.027	0.019	0
Transition to State 2	0.025	0.958	0.008	0
Transition to State 3	0.033	0.015	0.967	0.01
Transition to State 4	0	0	0.005	0.99
Mean value of NFR	0.001	0.032	0.037	0.117
Mean Value of 1Ns	0	0.069	0.079	0.21
Mean Value of 2Ns	0.021	0.033	0.07	0.183
Mean Value of 3Ns	0.022	0.006	0.049	0.125

Table 2. Summary of method comparisons of HMMRATAC, MACS2, FSeq, and HotSpot2.

Area under the curves calculated for the three ROC curves (Fig. 3a) and for the three precision vs. recall curves (Fig. 3b), the median number of reproducible peaks from

pairwise replicate testing (Fig. 4a) and the number of reproducible peaks between 500 cells per replicate and 50,000 cells per replicate data with a local IDR value below 0.05 (Fig. 4b). The numbers of the best performance are shown in bold.

	Active regions		**DNase-seq**		**FAIRE-seq**			
Method	**AUC-ROC**	**AUC-PRC**	**AUC-ROC**	**AUC-PRC**	**AUC-ROC**	**AUC-PRC**	**Reproducible Peaks within Replicates**	**Reproducible Peaks 500 to 50K cells**
HMMRATAC	**0.895**	**0.618**	**0.865**	**0.592**	**0.844**	0.550	**43,942**	**20,842**
MACS2	0.815	0.521	0.805	0.554	0.801	0.575	24,317	12,696
F-seq	0.826	0.526	0.817	0.565	0.818	**0.592**	15,008	10,123
Hotspot2	0.581	0.260	0.607	0.362	0.611	0.362	39,787	17,296

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 Expectation Maximization algorithm: **c.** Mean and **d.** standard deviation parameters per EM iteration for three nucleosomal distributions are consistent among datasets using 50,000 and 500 cells per replicate.

Fig. 2 Chromatin Architecture at HMMRATAC Peaks.

a. IGV browser snapshot of various signals around FAM212B locus on chromosome 1. Top three tracks show ATAC-seq signals, after being separated by size, the fourth track shows CTCF ChIP-seq signal, fifth through eighth tracks show various histone modification ChIP-seq signal and ninth track shows MNase-seq signal. The bottom panel shows RefSeq genes, HMMRATAC center and flanking nucleosome annotations (called with default settings), MACS2 peaks (called with default parameters), "Active Promoters" and "Strong Enhancers" state annotations from chromHMM (see methods section). **b.** Various ChIP-seq, DNase-seq and MNase-seq signals plotted across all *open* states. Data were clustered using k-means algorithm into 5 clusters.

Fig. 3 HMMRATAC outperforms other methods in precision, recall and false positive rate.

a. Receiver Operator Characteristic (ROC) plots for the real positive and real negative pairs of 1. Active regions vs. heterochromatin (top panel), 2. DNase peaks vs. heterochromatin (middle panel) and 3. FAIRE peaks vs. heterochromatin (bottom panel). Note: the endpoints, (0,0) and (1,1), are added to the plots for completeness. **b.** Precision vs. Recall plots for the real positive – real negative pairs of 1. Active regions vs. heterochromatin (top panel), 2. DNase peaks vs. heterochromatin (middle panel) and 3. FAIRE peaks vs. heterochromatin (bottom panel). Note: the endpoints, (0,1) and (1,0), are added to the plots for completeness**. c.** Boxplots showing the F1-scores of the four methods for the data shown in B.

Fig. 4 HMMRATAC shows better reproducibility in peak calls, over other methods. **a.** The numbers of reproducible peaks, identified with an IDR value below 0.05, from three pairwise replicate tests for each method. **b.** The numbers of overlapping peaks, identified with 500 cells per replicate and 50,000 cells per replicate, versus local IDR value. The horizontal line represents IDR value of 0.05.

References

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

10. Bell O, Tiwari VK, Thoma NH, Schubeler D: **Determinants and dynamics of genome accessibility.** *Nat Rev Genet* 2011, **12:**554-564.

11. 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.

12. 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.

13. Dell'Orso S, Wang AH, Shih HY, Saso K, Berghella L, Gutierrez-Cruz G, Ladurner AG, O'Shea JJ, Sartorelli V, Zare H: **The Histone Variant MacroH2A1.2 Is Necessary for the Activation of Muscle Enhancers and Recruitment of the Transcription Factor Pbx1.** *Cell Rep* 2016, **14:**1156-1168.

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

15. 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.

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

17. 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.

18. 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.

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

20. 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.

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

22. 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.

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

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

29. 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.

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

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

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. 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.

34. 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.

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

36. 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.

37. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

38. 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.

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

40. **HotSpot Mappability for Human Genome** [www.uwencode.org/proj/hotspot/]

Fig. 1

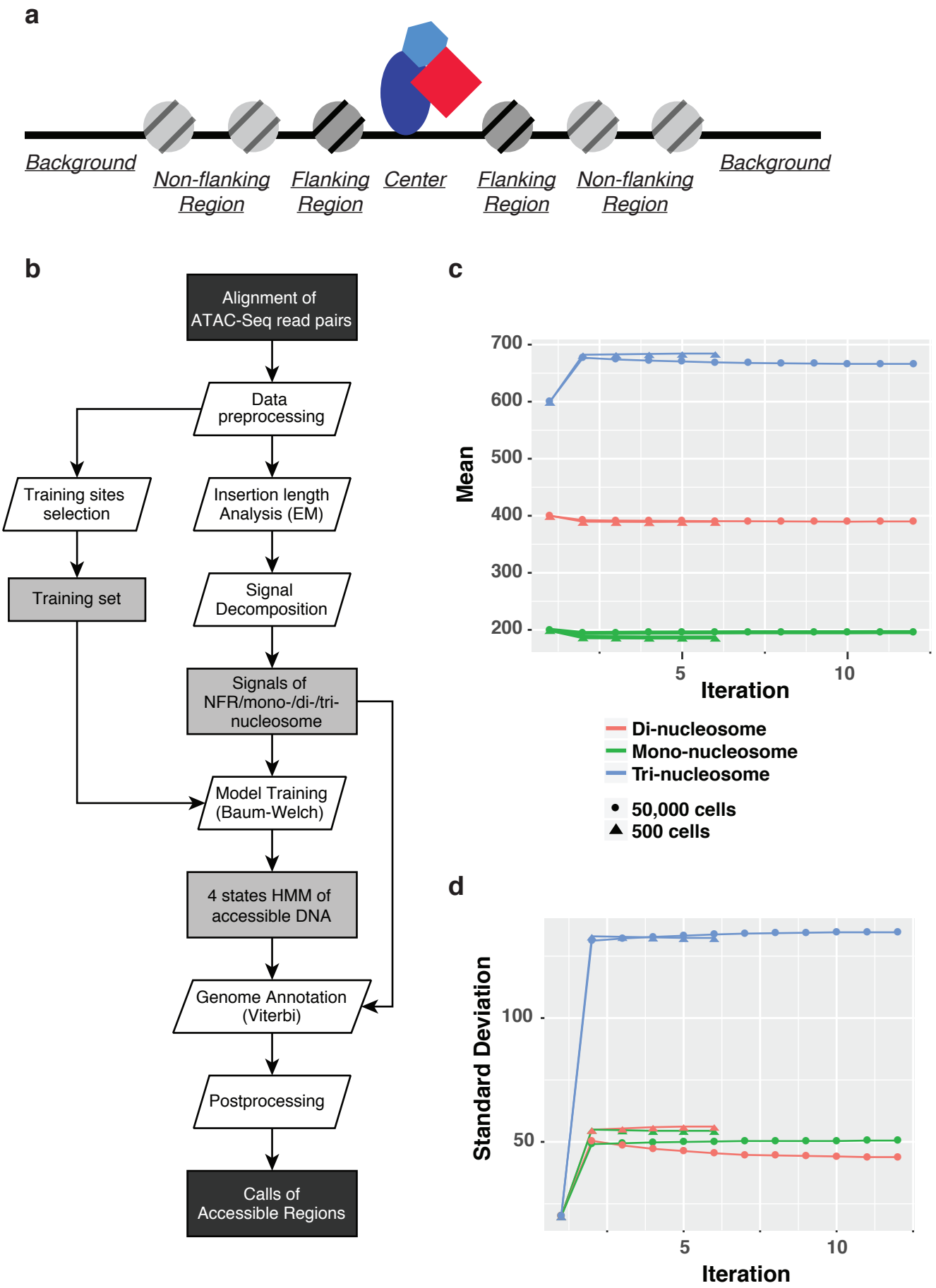

a
Background
Non-flanking Region
Flanking Region
Center
Flanking Region
Non-flanking Region
Background
b
Alignment of ATAC-Seq read pairs
Data preprocessing
Training sites selection
Insertion length Analysis (EM)
Training set
Signal Decomposition
Signals of NFR/mono-/di-/tri-nucleosome
Model Training (Baum-Welch)
4 states HMM of accessible DNA
Genome Annotation (Viterbi)
Postprocessing
Calls of Accessible Regions
c
Mean
700
600
500
400
300
200
5
10
Iteration
Di-nucleosome
Mono-nucleosome
Tri-nucleosome
50,000 cells
500 cells
d
Standard Deviation
100
50
5
10
Iteration

Fig. 2

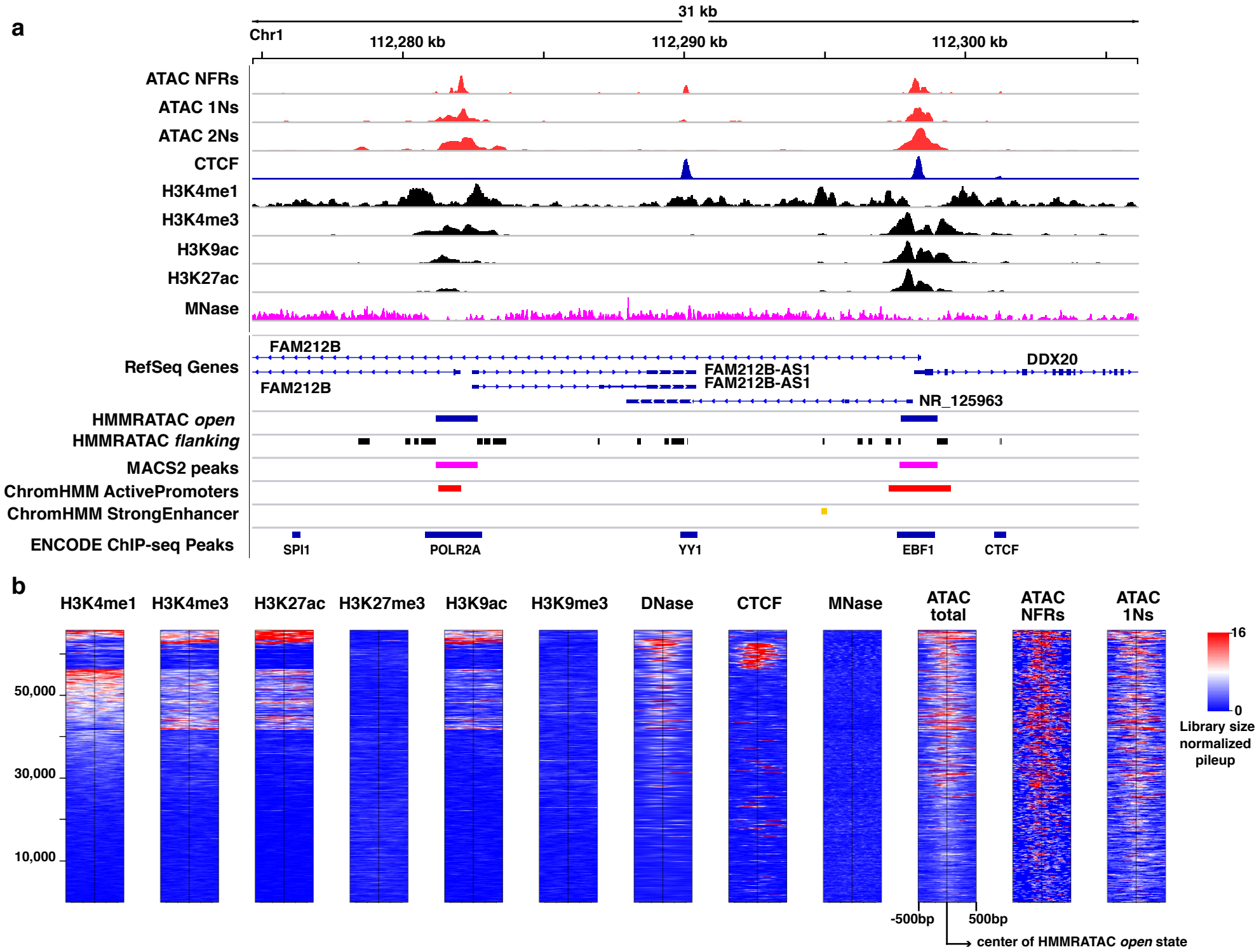

a
31 kb
Chr1
112,280 kb
112,290 kb
112,300 kb
ATAC NFRs
ATAC 1Ns
ATAC 2Ns
CTCF
H3K4me1
H3K4me3
H3K9ac
H3K27ac
MNase
RefSeq Genes
FAM212B
FAM212B-AS1
FAM212B-AS1
FAM212B
DDX20
NR_125963
HMMRATAC open
HMMRATAC flanking
MACS2 peaks
ChromHMM ActivePromoters
ChromHMM StrongEnhancer
ENCODE ChIP-seq Peaks
SPI1
POLR2A
YY1
EBF1
CTCF
b
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9ac
H3K9me3
DNase
CTCF
MNase
ATAC total
ATAC NFRs
ATAC 1Ns
50,000
30,000
10,000
16
0
Library size normalized pileup
-500bp
500bp
center of HMMRATAC open state

Fig. 3

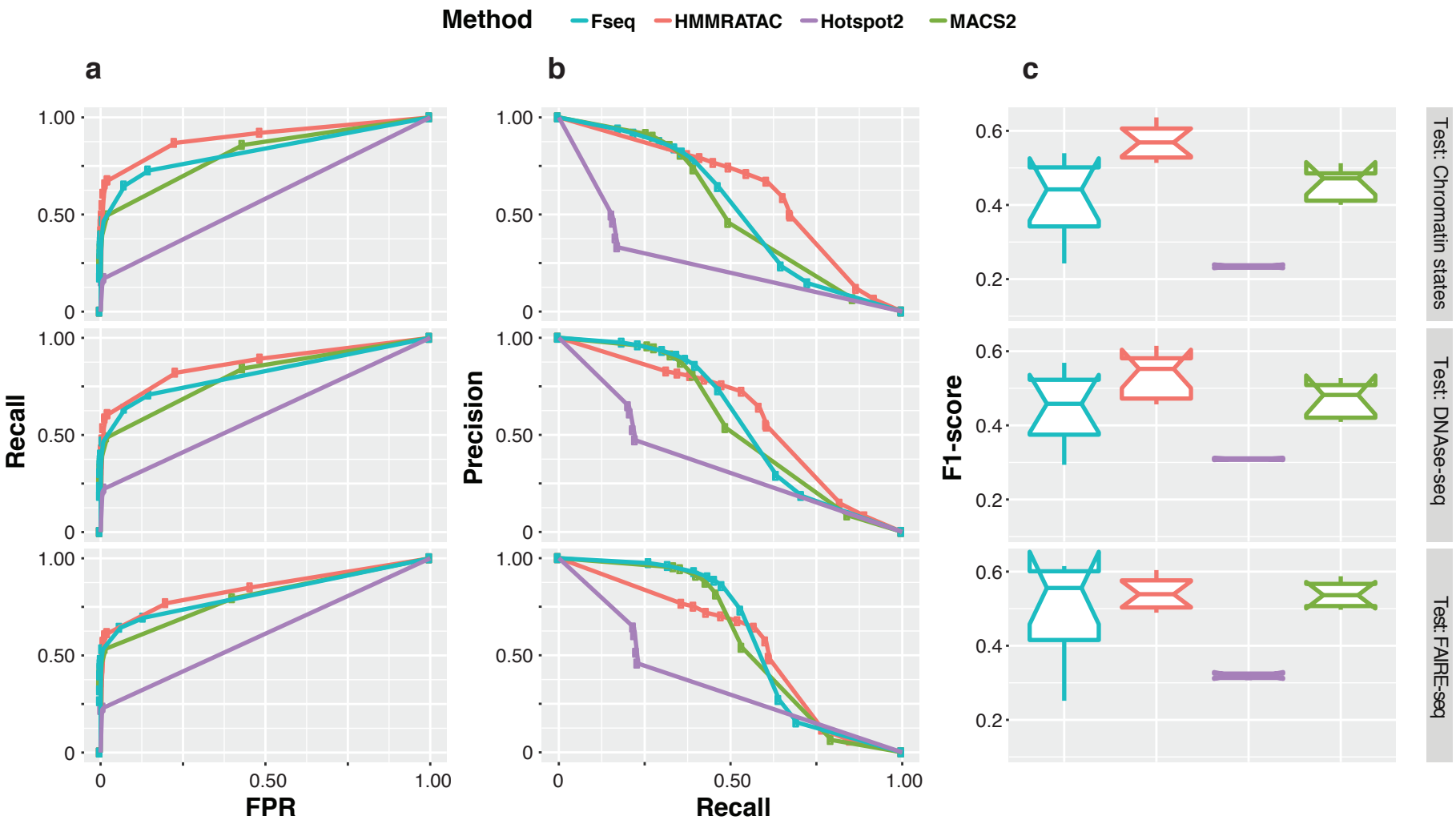
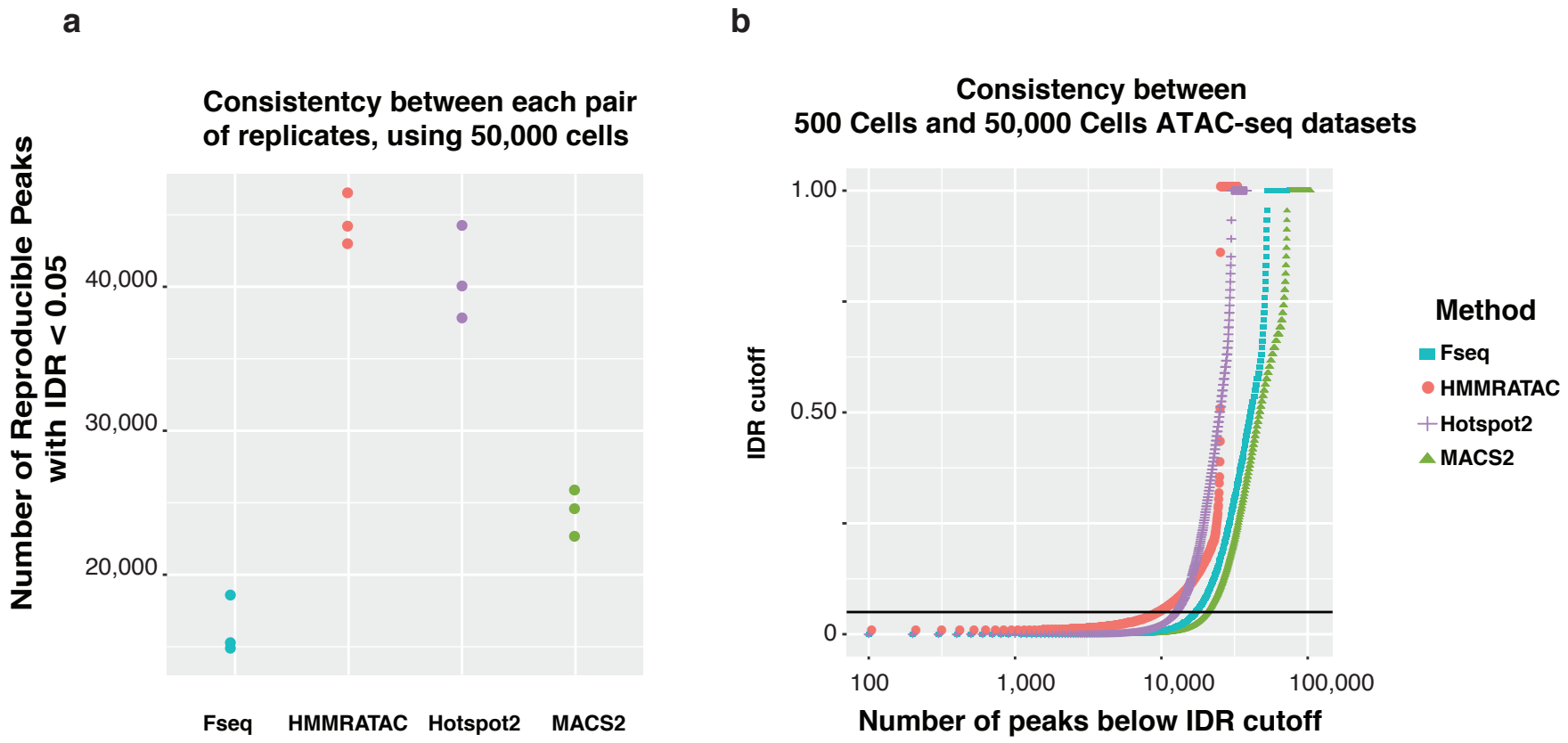

Method
Fseq
HMMRATAC
Hotspot2
MACS2
a
b
c
Recall
FPR
Precision
Recall
F1-score
1.00
0.50
0
0.6
0.4
0.2
Test: Chromatin states
Test: DNAse-seq
Test: FAIRE-seq

Fig. 4


a
Consistentcy between each pair
of replicates, using 50,000 cells
Number of Reproducible Peaks
with IDR < 0.05
40,000
30,000
20,000
Fseq
HMMRATAC
Hotspot2
MACS2
b
Consistency between
500 Cells and 50,000 Cells ATAC-seq datasets
IDR cutoff
1.00
0.50
0
100
1,000
10,000
100,000
Number of peaks below IDR cutoff
Method
Fseq
HMMRATAC
Hotspot2
MACS2