HMMRATAC: a Hidden Markov ModeleR for ATAC-seq

Evan D. Tarbell[3] and Tao Liu[1,2]*

[1] Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, 14263, USA

[2] Department of Biochemistry, University at Buffalo, Buffalo, NY, 14203, USA

[3] Enhanced Pharmacodynamics LLC., Buffalo, NY, 14203, USA

* To whom correspondence should be addressed. Tel: 716-845-1300(5034); Email: tao.liu@roswellpark.org

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 considering 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 ATAC-seq datasets. We find that single-end sequenced or size-selected ATAC-seq datasets result in a loss of sensitivity compared to paired-end datasets without size-selection.

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 transcription factors (2). Transcription factor bindings 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). The genome-wide investigations on nucleosome positioning found the interplay between transcription factor binding to nucleosome organization at accessible chromatin where the binding sites are nucleosome-depleted and surrounded by precisely phased nucleosomes (4).

38
39 Several assays exist to identify open chromatin regions in a genome-wide manner. These
40 include DNase-seq, which utilizes the DNase I nuclease (5), FAIRE-seq, which utilizes
41 differences in polarity between nucleosome-bound and nucleosome-free DNA (6), and ATAC-
42 seq, which uses a transposase to cut into accessible DNA selectively (7). Although each of these
43 assays identifies some unique open chromatin regions, they are generally highly correlated in
44 their identifications (7). Whereas DNase-seq and FAIRE-seq are complex protocols that require,
45 on average, one million cells, ATAC-seq is a simple three-step protocol, which is optimized for
46 fifty thousand cells and can be performed on as few as 500 cells or at a single cell level (7,8).
47 ATAC-seq has become popular over the years, and the Cistrome Database (9), in an effort to
48 collect all publicly available functional genomics data, has listed nearly 1,500 datasets for human
49 and mouse.
50
51 Due to steric hindrance, the Tn5 transposase used in ATAC-seq preferentially inserts into NFRs.
52 However, it is also possible for the transposase to insert into the linker regions between adjacent
53 nucleosomes, resulting in larger (over 150 bps) DNA fragments, which correspond to the integer
54 numbers of adjacent nucleosomes (7). The DNA fragments are constructed in a paired-end
55 library for sequencing, and after mapping both sequenced ends of each fragment to the genome
56 sequence, we can infer their fragment lengths according to the observed mapping locations, or
57 the insertion length. As demonstrated in (7), if we plot the observed fragment length versus
58 frequency, we will see a multi-modal distribution that creates different modes representing
59 transposase insertion into NFRs and linker regions between nearby phased nucleosomes up to
60 +4 nucleosome. This fact allows ATAC-seq to elucidate multiple layers of information relative to
61 the other assays. Although computational tools exist for DNase-seq, FAIRE-seq and ChIP-seq
62 (10), that can be and are used for ATAC-seq analysis, such as MACS2 (11) and F-Seq (12),
63 these would fall short since they only utilize a subset of information, usually the nucleosome-free
64 signals. NucleoATAC (13) was designed to separate an ATAC-seq dataset into its nucleosome-
65 free and nucleosome components then to identify nucleosome positions. However, NucleoATAC
66 only refines a predefined peak set, rather than utilize the nucleosomal information in the peak
67 calling process. To date, there are no dedicated peak-callers specifically to account for both NFR
68 and nucleosomal signals at the same time from ATAC-seq.
69
70 We present here HMMRATAC, the Hidden Markov ModeleR for ATAC-seq, a semi-supervised
71 machine learning approach for identifying open chromatin regions from ATAC-seq data. The
72 principle concept of HMMRATAC is built upon “decomposition and integration”, whereby a single
73 ATAC-seq dataset is firstly decomposed into different layers of coverage signals corresponding to
74 the sequenced DNA fragments originated from NFRs or nucleosomal regions; and then the

relationships between the layers of signals at open chromatin regions are learned in a Hidden Markov Model and utilized for predicting open chromatins. 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 (14) or BWA (15), then the identification of accessible regions or “peaks” by HMMRATAC, and then downstream analysis such as motif enrichment using MEME (16) or footprint identification in the accessible peaks with CENTIPEDE (17); differential accessibility analysis with Diffbind (18); and association studies with other data sets, such as with gene expression data with BETA (19). Quality control measurements would also take place at each step, such as calculating sequence quality before alignment using FASTQC (20) or after the alignment and performing replicate correlation during peak calling using ATACSeqQC (21). 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 (22) 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 (14). 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 (23) was publicly available and downloaded from the Cistrome Database (9) 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 (2-4,7) (see **Results**, **Figure 1A,** and **Supplemental Figure S1**). A single ATAC-seq dataset is firstly decomposed into different layers of coverage signals corresponding to the sequenced DNA fragments originated from NFRs or nucleosomal regions; and then the relationships between the layers of signals at open chromatin regions are learned in a Hidden Markov Model and utilized for predicting accessible regions across the whole genome (**Figure 1B**). The detailed steps of HMMRATAC algorithm are 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 ($P_s(\beta|x)$) for the nucleosome-free track and three Gaussian distributions ($P_m(\mu_m, \sigma_m|x), P_d(\mu_d, \sigma_d|x), P_t(\mu_t, \sigma_t|x)$) to represent the mono-, di- and tri-nucleosomal signal tracks, where x represents the observed insertion length of a reads pair, β, μ_m, μ_d, and $\mu_t$represent the means, $\sigma_m, \sigma_d$and $\sigma_t$represent the standard deviations of the corresponding distribution.

$$P_s(\beta|x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$$

$$P_m(\mu_m, \sigma_m|x) = \frac{1}{\sigma_m\sqrt{2\pi}} e^{-\frac{(x-\mu_m)^2}{2\sigma_m^2}}$$

146 $$P_d(\mu_d, \sigma_d|x) = \frac{1}{\sigma_d\sqrt{2\pi}} e^{-\frac{(x-\mu_d)^2}{2\sigma_d^2}}$$

147 $$P_t(\mu_t, \sigma_t|x) = \frac{1}{\sigma_t\sqrt{2\pi}} e^{-\frac{(x-\mu_t)^2}{2\sigma_t^2}}$$

148 These distributions were utilized by Buenrostro et al. to identify their cutoffs for fragment
149 separations (7). Although sequence read data is discrete, the fact that the expectation value of
150 each Gaussian is large (>100 bps) and the number of observations is high, the discrete
151 distribution, such as a Poison distribution, can be approximated as a Gaussian. The mean value
152 of the exponential distribution for NFRs is set as a fixed value at runtime, either as the default
153 value or a user-defined value. The parameters of three Gaussian distributions are initialized at
154 runtime with default or user-defined values and then updated using the expectation maximization
155 (EM) algorithm for Gaussian mixture models. These initial values were the same as those used
156 by Buenrostro et al. (7) to define their distributions for the chromatin features. Briefly, each
157 nucleosomal fragment, those larger than 100 bps, is classified as belonging to one of the
158 nucleosomal distributions based on its weighted probability. Once each fragment has been
159 classified, new means, standard deviations, and weights are calculated for each distribution and
160 become the new values for the mixture model. This process continues iteratively until the change
161 in the mean values between iterations is less than a reasonable epsilon value. At this point, the
162 model is assumed to have reached convergence, and the EM process is halted (**Figure 1C**). In
163 order to decrease the time required for the EM algorithm, HMMRATAC randomly sub-selects
164 10% of all the fragments to use as the training data. Once the parameters of the four distributions
165 have been determined, HMMRATAC creates four genome-wide signal tracks. For every genomic
166 position, all of the fragments that occupy that position are determined. Each track is then
167 incremented by the probability calculated in the above equation that a particular fragment belongs
168 to the tracks' corresponding distribution. We then use a square root transformation to get the
169 data into continuous space. Although the transformations are commonly done with logarithmic
170 transformations, we chose square root because our data is all positive values but also contains
171 zeroes.
172
173 ***Training the Hidden Markov Model of accessible chromosomal regions, and refining the***
174 ***model with the Baum-Welch algorithm***
175
176 Once the four signal tracks have been created, the main HMMRATAC process begins.
177 HMMRATAC first identifies up to 1000 training regions throughout the genome, with which it
178 learns the model. These training regions are determined by scanning the genome for regions
179 whose read coverage fold changes over background falls within a certain range. The default
180 lower and upper limit of fold changes, as used in the Result section, is set between 10 and 20.

181 These regions are then extended by 5 kbps in either direction, in order to ensure that background
182 regions are included in the training set. It is also possible for users to provide a BED file
183 containing predefined training regions. The initial HMM has proportional initial probabilities, which
184 are not updated during training, as well as proportional transition probabilities, which are updated
185 during training. The emission probabilities are calculated with k-means clustering, by separating
186 the training regions into 3 clusters with random initialization and calculating the means and
187 covariance matrices for the four signal tracks for each cluster. The four signal tracks, across the
188 training regions, are the only input to the k-means algorithm. We model the emissions as a
189 multivariate Gaussian distribution since this assumption has been successfully adopted in other
190 HMM-based genome segmentation methods such as hiHMM (24,25) and Segway (26,27). The
191 three clusters represent the three states of our model, corresponding to the *center* of the open
192 region with overall high enrichment of all of the four types of signals, the *nucleosomes* with low
193 NFR signal but moderate nucleosomal signals, and a *background* state with low enrichment of
194 any signals, respectively. It is also possible to create a model using a different number of states
195 instead of 3, although this option is not generally recommended. Once this initial HMM is created,
196 the transition and emission parameters are updated with the Baum-Welch algorithm (28) until
197 convergence (**Figure 1D**). Furthermore, to demonstrate the robustness of the training approach,
198 we ran HMMRATAC three times on the same GM12878 ATAC-seq data with different random
199 initializations for k-means and found that the k-means centers and HMM parameters were
200 virtually identical, as shown in **Supplemental Table S1**.
201
202 ***Annotating with the Viterbi algorithm***
203
204 Once the HMM has been created and refined, the Viterbi algorithm (29) is used to classify every
205 genomic position into one of the three states of our model. In our experience, when Viterbi
206 encountered a high coverage region, it was interrupted and mistakenly called the entire regions
207 following the high coverage region as either a peak or a background region. Therefore, regions
208 whose z-scored coverage is above a certain cutoff, either a default or user-defined value, are
209 masked prior to the Viterbi algorithm. Additionally, it is possible to include a list of genomic
210 regions to mask along with the extremely high coverage regions, such as annotated blacklisted
211 sites provided by the ENCODE project. We have found that 91% of the high coverage regions
212 that are masked from the Viterbi algorithm in the GM12878 cells are annotated blacklist regions
213 (downloaded from (30)). A genome-wide BedGraph of all state annotations can be reported by
214 HMMRATAC, although this is not the default behavior. By default, HMMRATAC will take all
215 regions classified as belonging to the *center* state and whose length is above a default or user-
216 defined threshold and merge them with the *nucleosome* states located both upstream and
217 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 (31). 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 120 bps. 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™ 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 ATAC-seq dataset. We recognize that HMMRATAC takes longer than most algorithms and are currently optimizing it in order to decrease runtime. We will also introduce a parallel computing ability, which we believe will dramatically reduce the runtime. These efforts are active and ongoing and will be incorporated into future releases.

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

255 option for keeping duplicate reads (option --keep-dup) was set to “all”, as duplicate reads had
256 already been discarded in pre-processing. MACS2 was run with the local lambda option turned
257 off (option --nolambda) and was run with q-value cutoffs (option -q) set to 0.5, 0.1, 0.05 (default),
258 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.
259
260 ***Calls from F-Seq***
261
262 F-Seq version 1.84 was used to call peaks using the single-end BED files created as described
263 previously. With each run, the output was reported as a BED file (option -of). The standard
264 deviation cutoff (option -t) was set to 0.1, 1, 4 (default), 6, 7, 8, 10, 15 and 20.
265
266 ***Calls from HMMRATAC***
267
268 HMMRATAC version 1.2 was used to call peaks using the sorted BAM file. The upper limit fold-
269 change for choosing training regions (option -u) was set to 20 with corresponding lower limits
270 (option -l) set to 10. These settings are the HMMRATAC defaults. The blacklisted sites for hg19
271 were provided, which were masked from the program and all subsequent steps. For the
272 precision, recall and false positive rate calculations, the HMMRATAC peaks were filtered by
273 minimum score cutoffs of 0, 5, 10, 15, 20, 25, 30, 35, 40 and 45. Additionally, for two monocyte
274 samples (accession numbers GSM2325689 – RPMI_4h and GSM2325681 – BG_d1)
275 HMMRATAC did not produce a typical model with the default settings (-u 20 –l 10). The ideal
276 model should show the highest NFR, 1Ns, 2Ns and 3Ns signals in the third (center) state and the
277 second highest values of all four signals in the second (nucleosome) state. In this case, we found
278 that the second state (nucleosome) had the highest NFR signal while the third (center) state had
279 the highest 1Ns, 2Ns, and 3Ns signals. Re-running it with more stringent parameters (-u 40 –l 20)
280 solved that problem and produced a model that had the highest signal values in the third state
281 and the second highest in the second state. These peak results from HMMRATAC were then
282 filtered by the same cutoffs as the other samples and tested in the same way. All the transition
283 and emission parameters learned by HMMRATAC from GM12878 50k cells, 500 cells, five
284 human monocyte samples are summarized in **Supplemental Table S2**.
285
286 ***Compiling the “Gold Standard” lists and “True Negative” set***
287
288 We tested the performance of HMMRATAC and other methods on ATAC-seq datasets of
289 GM12878 cells against three proxies of “gold standard” datasets. The first one, which we call
290 “active regions”, is made by combining two chromatin states “Active Promoters” and “Strong
291 Enhancers” generated in GM12878 cells using chromHMM (32,33) (downloaded from UCSC

genome browser track wgEncodeBroadHmmGm12878HMM). The other two sets were DNaseI hypersensitive sites (UCSC track wgEncodeAwgDnaseUwdukeGm12878UniPk) and FAIRE accessible regions (wgEncodeOpenChromFaireGm12878Pk) (34) 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 (35) from traditional methods including Hotspot (36) and F-Seq (12), 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 2 kbps 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 (23) were downloaded from the Cistrome Database (9) as aligned BAM files. Each file was processed by MACS2 with the p-value parameter (-p option) set to 0.15. We set this loose cutoff, so that, when combining active and heterochromatin marks, the true and false sets can be more exclusive. In order to identify a loose cutoff, we performed MACS2 cutoff-analysis (callpeak --cutoff-analysis) approach to identify the critical cutoff below which the genomics background noises start to be picked up by peak caller. 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**.

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

330

331 Let Predicted Positive (PP) represent the number of base pairs identified as accessible regions
332 from any method. Let Real Positives (RP) represent the number of base pairs in the "gold
333 standard" sets. Let Real Negatives (RN) represent the number of base pairs in the "real
334 negative" set. Let True Positive (TP) represent the number of overlapping base pairs between a
335 called peak and a real positive. Let False Positive (FP) represent the number of overlapping
336 base pairs between a called peak and a real negative.

337

338 Then the *precision* or positive predictive value (PPV) is calculated as PPV = TP/(TP+FP). The
339 *Recall*, or true positive rate (TPR), is calculated as TPR = TP/RP. The *false positive rate* (FPR) is
340 calculated as FPR = FP/RN. The number of overlapping base pairs was calculated with the
341 BEDTools intersect tool (37). After the precisions, recalls and false positive rates were calculated
342 for accessible regions predicted by a given method under various cutoff values, the approximate
343 area under the curve (AUC) was calculated for the *recall* vs. *false positive rate* (ROC) curves and
344 for the *precision-recall* (PR) curves, after adding extreme points to the unreachable ends of the
345 curves. The extreme points to the ROC curves are TPR=0 and FPR=0, and TPR=1 and FPR=1;
346 the extreme points to the PR curves are TPR=0 and PPV=1, and TPR=1 and PPV=L_{true}/(L_{true} +
347 L_{false}) where the L_{true} is the total length of real positive set and the L_{false} is the total length of the
348 real negative set.

349

350 **RESULTS**

351

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

353

354 In order to investigate the ATAC-seq signal profile around accessible genomic regions, we took
355 the ATAC-seq data in the human GM12878 cell line from Buenrostro et al. (7) and separated the
356 fragments into short nucleosome-free (under 100 bps) and mono-nucleosome (180-250 bps)
357 fragments based on cutoffs identified by Buenrostro et al. These cutoffs were determined by
358 fitting the fragment length distribution to several simulated distributions modeling nucleosome and
359 nucleosome-free transposition frequencies (7). We plotted the two types of fragments around the
360 centers of DNaseI hypersensitive sites (DHSs) in the same cell line, identified by ENCODE (35)
361 (**Supplemental Figure S1**). We observed a clear, symmetrical pattern around the centers of
362 these sites characterized by an enrichment of the nucleosome-free fragments and flanking
363 enrichment of the mono-nucleosome fragments. We were able to visually discern at least three
364 distinct regions around an open site: 1) the *center*, characterized by an enrichment of
365 nucleosome-free fragments and flanked by nucleosome fragments; 2) the *nucleosome regions*,

366 characterized by an enrichment of nucleosome fragments; 3) and the *background*, characterized
367 by a depletion in any fragments (**Figure 1A; Supplemental Figure S1**). Based on this
368 observation and existing knowledge on the nucleosome positioning around the accessible
369 chromatin (2-4), we hypothesized that such pattern would exist at the open chromatin throughout
370 the genome and that by learning and recognizing this pattern computationally in a single ATAC-
371 seq dataset, we would identify the open chromatin regions with higher confidence than any other
372 existing method.
373
374 **Signal decomposition through the probabilistic approach**
375
376 In the first ATAC-seq paper by Buenrostro et al., the authors separated the fragments according
377 to their insertion lengths into four populations representing those from the NFRs and those
378 spanning 1, 2 or 3 nucleosomes (7). Non-overlapping and non-adjacent cutoffs were used. For
379 example, lengths below 100 bps represent nucleosome-free signals and lengths between 180
380 and 250 bps represents mono-nucleosome signals. We consider such cutoff-based strategy not a
381 general practice for the following two reasons. First, different species or different cell lines may
382 have different nucleosome spacing and these differences may not be adequately understood
383 (38,39). Secondly, since the fragments whose size falls between the non-overlapping and non-
384 adjacent cutoffs are discarded (e.g. those between 100 and 180 bps long), the recall (or
385 sensitivity) to identify accessible regions will be reduced (**Supplemental Figure S2**). As a fact,
386 about 15% of the entire data generated from the original paper are ignored with their strategy. We
387 addressed the above two challenges in a probabilistic approach in HMMRATAC.
388
389 First, we utilize the Expectation-Maximization (EM) algorithm to identify the optimal parameters
390 for the four different distributions of fragment lengths. Using the distributional assumptions of
391 Buenrostro et al., we create four distributions to represent four different chromatin features:
392 nucleosome-free regions (NFRs), mono-nucleosomes (1Ns), di-nucleosomes (2Ns) and tri-
393 nucleosomes (3Ns). HMMRATAC uses a mixture model and Expectation-Maximization (EM)
394 algorithm to identify optimum parameters. As an example, HMMRATAC identified the optimum
395 fragment lengths representing 1Ns, 2Ns, and 3Ns from ATAC-seq data in the human GM12878
396 cell line (7), generated by merging three replicates, as being 195 bps, 396 bps, and 693 bps,
397 respectively (**Figure 1C**). As a demonstration of different nucleosome spacing in different
398 datasets, HMMRATAC determined the optimum fragment lengths for the 1Ns, 2Ns, and 3Ns from
399 ATAC-seq data in human monocytes treated with RPMI for one day (23) as 186 bps, 368 bps,
400 and 549 bps, respectively. Additionally, the values for any of the individual replicates, in the
401 GM12878 data, were never more than 5 bps off from the average values listed above for the
402 merged datasets (**Supplemental Figure S3**). Not all ATAC-seq datasets show a clear multi-

403 modal pattern, especially datasets from fewer cells. As an example, the fragment distribution
404 generated by Buenrostro and et al. utilizing 500 cells (GM12878 cells) per-replicate, did not show
405 as clear of a distribution as 50,000 cells per replicate (**Supplemental Figure S4A and S4B**).
406 However, we found our approach can robustly identify the parameters even if the multi-modal
407 pattern is not clear. As we can see from **Figure 1C**, the mean and standard deviation values
408 converged after no more than 12 iterations and the parameters converged to very similar values
409 for either the 50,000-cell per-replicate or 500 cell per-replicate datasets.
410
411 Secondly, we propose in HMMRATAC a probabilistic approach to make use of data from the
412 entire spectrum of fragment lengths. We generate four distinct signal profiles in 10 bps resolution,
413 for NFRs, 1Ns, 2Ns, and 3Ns, allowing each mapped fragment to contribute to all profiles with
414 different weights corresponding to its particular fragment length. For each genomic location that is
415 occupied by a particular fragment, all four signal profiles are incremented by the probabilities that
416 this fragment, assigned with an observed fragment length through read mapping, belongs to each
417 of the four corresponding distributions. We tested the performance of a regular HMMRATAC
418 approach with this probabilistic method utilizing the entire data and an alternative approach
419 utilizing only the subset of data satisfying the cutoffs proposed in Buenrostro et al. The two
420 approaches were compared in terms of precision and recall against the "active regions" true set
421 (**Supplemental Figure S2**). We found that our probabilistic method had considerably better
422 recall and precision compared to the cutoff based method. This result proved that the traditional
423 way to discard reads between cutoffs of fragment sizes would hurt the prediction power.
424
425 HMMRATAC utilizes the length of an insertion fragment, identified by paired-end sequencing, as
426 an essential and critical piece of information in its processes. As such, HMMRATAC can only be
427 used in paired-end sequencing libraries. Because we observed a decrease in the performance of
428 our method when we used the cutoff based approach, which resulted in the removal of ~15% of
429 all sequencing reads, we wanted to understand the effects of size-selection and paired-end
430 sequencing on ATAC-seq peak calling. We performed two *in silico* size-selections on the data
431 from Buenrostro et al. The first restricted the fragments to those whose length was under 100
432 bps and the second restricted the fragments to those with lengths under 300 bps. We also
433 treated the data as single-end, by unpairing the read mate pairs. We then used MACS2 (11) to
434 call peaks with each data set, including the original unaltered data and compared the
435 performance to predict the accessible chromatin (**Supplemental Figure S5**). As a result, the
436 original unaltered data performed the best in terms of precision and recall. This indicates that an
437 ATAC-seq assay should better be performed without a size selection step and the resulting
438 fragments should undergo paired-end sequencing, regardless of the peak calling algorithm that is
439 used.

440

441 **Integration through Hidden Markov Model**

442

443 After removing reads with low mapping quality, and masking blacklist regions from downstream
444 analysis, HMMRATAC identifies a set of training regions throughout the genome to train a 3-
445 states Hidden Markov Model (HMM) with multivariate Gaussian emissions, representing the
446 underlying three chromatin states of 1) the *center* of open chromatin, 2) the *nucleosome regions*
447 and 3) the *background* (**Figure 1A**). By default, training regions can be either specified by users
448 or automatically selected where the signal fold-change above background falls within a
449 predetermined range. After signal decomposition, a matrix of 10 bps resolution signals for NFRs,
450 1Ns, 2Ns, and 3Ns in training regions is used to train the parameters in HMM through the Baum-
451 Welch algorithm (28). The initial state transition probabilities are set equally as 0.33 and the
452 emission parameters for each state are calculated after clustering all locations in the training
453 regions into 3 classes (see **Methods** for detail). **Figure 1D** shows the final parameters for the
454 HMM on ATAC-seq in the GM12878 cell line, by merging three replicates from 50k cells (7).
455 Although the parameters will change depending on the choice of training regions and the
456 datasets provided, several key characteristics are evident in the model and remain mostly stable
457 with different settings and data. State 3 in the HMM model, which HMMRATAC annotated as the
458 *center* state, showed the highest average signal across the four signal profiles. State 2,
459 annotated as the *nucleosome* state, showed the second highest average values. Additionally, as
460 we hypothesized, the results show the *center* state mainly transits from or to the *nucleosome*
461 state.

462

463 After learning the model, HMMRATAC takes a matrix of genome-wide NFRs, 1Ns, 2Ns, and 3Ns
464 signals and labels each genomic location with one of the states using the Viterbi algorithm (29).
465 On the 50K cells ATAC-seq data in GM12878 cell line, there were about 4.3 million regions
466 annotated as *background* with an average size of 288 bps (~35% of the human genome), about
467 4.5 million regions annotated as *nucleosomes* with an average size of 382 bps (~49% of the
468 human genome), and about 226 thousand regions annotated as *center regions* with an average
469 size of 504 bps (~3% of human genome). HMMRATAC finally connected the *center regions* with
470 their adjacent *nucleosomes* and called a total of 87,440 accessible regions with an average size
471 of 1,831 bps (5% of the human genome), excluding those with abnormally high coverage and
472 those below a minimum length. It should be noted that many of the *nucleosomes* are not adjacent
473 to the *center* state. These may represent nucleosomes away from regulatory elements that
474 ATAC-seq is able to identify or may be artifacts due to increased transposition frequency, an
475 issue which can be caught by quality assessment tools such as ATACseqQC (21).

476

477 **Recapitulating chromatin architecture at functional elements**
478
479 We hypothesized that HMMRATAC could identify chromatin architecture around accessible
480 regions, or ideally, nucleosomes near the regulatory elements. To prove this, we explored the
481 distribution of histone modification ChIP-seq signal around our identified accessible regions. We
482 downloaded and analyzed datasets released by ENCODE in the GM12878 cell line (35),
483 including ChIP-seq for CTCF, active histone marks such as H3K4me1, H3K4me3, H3K9ac, and
484 H3K27ac, and repressive histone marks such as H3K9me3 and H3K27me3, as well as
485 sequencing of micrococcal nuclease digestion (MNase-seq). **Figure 2A** shows two loci, one on
486 chromosome 8 that was identified as an accessible region by MACS2 and F-Seq but not
487 HMMRATAC and another locus on chromosome 2 that was identified as accessible by all three
488 methods. At the chromosome 2 locus, we observed that the *center* state was located in regions
489 characterized by depletion in both active histone marks and MNase digestion. The surrounding
490 *nucleosome* states showed enrichment of active histone marks and MNase digestion. We next
491 plotted the same ChIP-seq, DNase-seq and MNase-seq data around all distal (> 10 kbps from
492 any gene promoter) *center* states and used k-means clustering to group the resulting profiles
493 (**Figure 2B**). We found that HMMRATAC states can identify, what appears to be, different
494 chromatin features at the accessible genomic regions. For example, the first cluster (C1)
495 characterized by strong enrichment of H3K27ac relative to H3K4me1 seems to indicate active
496 enhancers while the third cluster (C3) marked by high levels of H3K4me1 and more moderate
497 levels of H3K27ac would suggest weak or poised enhancers (**Figure 2B**). The second cluster
498 (C2), characterized by strong enrichment of CTCF binding, lower levels of H3K4me1, H3K4me3
499 and acetylation histone marks and a higher level of H3K27me3, seems to suggest that this cluster
500 represents insulators or domain boundaries. The fourth cluster (C4), similar to C2, suggested
501 accessible regions also exist in the DNA elements with repressive functions in euchromatin,
502 potentially bound by factors rather than CTCF. To summarize, these results indicate that
503 HMMRATAC recapitulates the histone architecture around accessible regions.
504
505 The chromosome 8 locus exemplifies some of the advantages of HMMRATAC over other peak
506 callers. Both MACS2 and F-Seq identified this locus as being accessible, despite the lack of
507 other corroborating evidence, such as transcription factor binding and the presence of active
508 chromatin states. It appears that these methods called the site because of an enrichment of the
509 mono-nucleosome signal. Because neither MACS2 or F-Seq integrate the various chromatin
510 features present in ATAC-seq data and because they are based on signal enrichment only, as
511 opposed to HMMRATAC which incorporates the structure of the element, they both falsely
512 identified this region as accessible, while HMMRATAC identified the region as being a
513 *nucleosome* state.

514
515 In order to understand the chromatin features that are enriched in each state across the genome,
516 as the unique result HMMRATAC concluded from a single ATAC-seq data, we plotted the overlap
517 of the 3 hidden states with chromHMM states and histone mark calls from ChIP-seq in the
518 GM12878 cell line based on ENCODE data (**Figure 3A; Supplemental Table S4**). We found that
519 the chromatin states "Active Promoters", "Strong Enhancer", "Insulator", "Poised Promoter" and
520 "Weak Promoter" states mainly overlap with our *center* state, and the "Heterochromatin",
521 "Repetitive", "Repressed", "Transcription Elongation", and "Weak Transcription" barely overlaps
522 with the *center* state. Interestingly, the "Strong Enhancer" state had a similar amount of overlap
523 with the *nucleosome* state as the *center* state. Combined with our observation that our
524 *nucleosome* state showed elevated levels of active histone marks near accessible regions, we
525 hypothesized that this state may represent positioned nucleosome nearby regulatory elements,
526 those marked by active histone modifications and adjacent to accessible regions. We plotted the
527 H3K4me1 active histone mark around the centers of our *center* states as well as around the
528 centers of the adjacent *nucleosome* states, those directly flanking our *center* states (**Figure 3B**).
529 As we can see, the *center* states show a bimodal enrichment for the histone mark, consistent with
530 MACS2 and F-Seq peaks. Additionally, our adjacent *nucleosome* states show central enrichment
531 for the histone mark. This indicates that the *nucleosome* state, at least those adjacent to our
532 *center* states, are likely nucleosomes close to regulatory elements. As a result, HMMRATAC
533 merges the *center* states with their adjacent *nucleosome* states, both upstream and downstream,
534 to identify the regulatory region as the output.
535
536 Functional regulatory elements that harbor transcription factor binding sites are generally under
537 higher evolutionary constraint (40). Success in transcription factor ChIP-seq data analysis is often
538 benchmarked by high PhastCons score, a precompiled evolutionary conservation score at every
539 base pair (41), at the peak center or summit, relative to the surrounding regions (42). We found
540 that the summits of accessible regions detected by HMMRATAC in the GM12878 cell line also
541 showed PhastCons enrichment in vertebrates (**Figure 3C**), indicating that these summits were
542 likely locations of evolutionarily conserved regulatory elements and could represent functional
543 regulatory elements.
544
545 **Method comparison with MACS2 and F-Seq**
546
547 We next sought to compare the performance of HMMRATAC in identifying open chromatin
548 regions to other popular methods for ATAC-seq. We chose MACS2 (11) and F-Seq (12) which
549 had previously been shown to be the most accurate peak-callers in identifying open chromatin
550 regions from DNase-seq (43). MACS2, originally designed for ChIP-seq, utilizes a local Poison

551 model wherein enrichment is calculated relative to a local Poisson model. F-Seq invokes a
552 Gaussian kernel-smoothing algorithm and then calculates enrichment based on the resulting
553 Gaussian distribution. Although none of them utilize the intrinsic and unique features in ATAC-
554 seq data we described before, MACS2 has been used to analyze ATAC-seq data in many
555 published works (23,44-48), and MACS2 and F-Seq is part of the ENCODE analysis pipeline for
556 ATAC-seq data (https://github.com/kundajelab/atac_dnase_pipelines). We didn't include ZINBA,
557 an algorithm based on the zero-inflated negative binomial model, in this manuscript, although it
558 was used to analyze the GM12878 ATAC-seq dataset in Buenrostro et al.'s original work (7).
559 ZINBA was reported as less sensitive than other methods while being applied to DNase-seq data
560 (43) and we found the same conclusion (**Supplemental Figure S6**). Since this manuscript is not
561 for reviewing all available methods, we decided to only show our improvement over the current
562 best practices in the field. We tested the performance of the above algorithms using the two
563 ATAC-seq datasets in the GM12878 cell line as the previous sections.
564
565 A major limitation in comparing the performance of identifying open chromatin regions is the
566 absence of a valid "gold standard" (3). To overcome this problem, we chose to compare the
567 performance of HMMRATAC and the other algorithms to three distinct independent datasets as
568 the proxies of true and false set: 1) the active regions and heterochromatin from chromatin states
569 analysis combining multiple independent histone features (32), 2) DNase-seq (5) and 3) FAIRE-
570 seq (6) accessible regions of the same GM12878 cell line, originally called from traditional
571 methods such as Hotspot (36) and F-Seq (12).
572
573 ***Enrichment in chromatin states of active regions***
574
575 The first proxy of the true set that we compared the algorithms to is referred to as "active
576 regions." These sites were defined through an integrative chromatin state segmentation using
577 chromHMM (32,33) algorithm conducted by the ENCODE project. 8 histone modifications
578 (H3K27me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3, H3K27ac, and H3K9ac),
579 1 transcription factor (CTCF) ChIP-seq and 1 whole genome input datasets in the GM12878 cell
580 line were integrated through an unsupervised machine learning approach and each genomic
581 location was annotated as one of 15 mutually-exclusive states. Two resulting states, "Active
582 Promoters" and "Strong Enhancers", were merged together to create our list of "active regions".
583 Additionally, the state annotated as "Heterochromatin" was used as a false set. In **Figure 2A** and
584 **Supplemental Figure S7**, we showed some examples of consistent true positive calls from all
585 algorithms and those false positive peaks called by MACS2 or F-Seq but not HMMRATAC, with
586 default settings from each algorithm. A general observation is that conventional peak calling
587 approaches can be misled by high ATAC-seq nucleosomal signals and call regions with no NFR

588 signal as peaks. Full peak calls from default setting can be accessed through a UCSC genome
589 browser track hub (http://biomisc.org/data/HMMRATAC_hub/hub.txt). However, the total number
590 of such cases in the genome is strongly dependent on the cutoffs used by the individual
591 algorithms. By tweaking cutoffs, false positives may disappear while sacrificing sensitivity.
592 Therefore, we decided to test different cutoff values for each algorithm (see **Methods**). We found
593 that HMMRATAC outperformed the other algorithms, in terms of precision, recall, and false
594 positive rate in identifying these "active regions" while avoiding "heterochromatin" (**Figure 4** top
595 row). The area under the curve (AUC) was then calculated for the recall vs. false positive rate
596 curves (Area Under Curve of Receiver Operator Characteristics; AUCROC) and for the precision-
597 recall curves (Area Under Precision-Recall Curve; AUPRC) (**Table 1**), after adding extreme
598 points to the unreachable ends of the curves. HMMRATAC was proved to have better AUC than
599 the two other methods. It worth noting that HMMRATAC calls larger peaks than the other two
600 methods due to its merging of the adjacent nucleosome states to the center state. However,
601 despite the larger average peak size, HMMRATAC identifies a comparable number of accessible
602 regions in base pairs, as shown in **Supplemental Figure S8**, at comparable levels of sensitivity
603 for "active regions". This would also indicate that many peaks called by the other methods are
604 falling outside of the "gold standard" regions, possibly contributing to their deficits in precision
605 compared to HMMRATAC. Additionally, we retrieved 257 GM12878 super-enhancers from the
606 dbSUPER (49) as a true set, assuming these regions are bound by an array of transcription
607 factors and always been broadly accessible, and used the "heterochromatin" regions as false for
608 an extra test. We found that HMMRATAC again performs better than other algorithms
609 (**Supplemental Figure S9**). To study the effect of size-selection, we performed an *in silico* size-
610 selection to restrict the fragments to those whose length was under 250 bps, then we tested the
611 performance of HMMRATAC, MACS2, and F-Seq as shown before. We found that although the
612 overall performance on the partial data decreases compared with full data set without size-
613 selection, HMMRATAC can still outperform others (**Supplemental Figure S10**). Taken together,
614 these results indicate that HMMRATAC is able to identify active chromatin regions with higher
615 recall and precision and lower false positive rate, than MACS2 and F-Seq.
616
617 ***Consistency with results from DNase-Seq and FAIRE-Seq***
618
619 We then used the accessible chromatin regions detected by two independent assays DNase-seq
620 and FAIRE-seq in GM12878 cell line as cross-validation of ATAC-seq analysis. We downloaded
621 DNaseI hypersensitive sites (DHSs) (5) and FAIRE-seq (6) accessible regions called by
622 ENCODE in the GM12878 cell line (35), then merged DHSs or FAIRE accessible regions that
623 were within 2 kbps of each other to define our alternative true sets. We continued to use the
624 heterochromatin states outlined above as our false set, although we eliminated any DHSs or

625 FAIRE regions that overlapped a heterochromatin state region and *vice versa*. We found that
626 HMMRATAC outperformed the other two methods in terms of AUPRC in recapitulating DHSs
627 (**Supplemental Figure S11, Table 1**). MACS2 performed slightly better in AUCROC by 0.001. F-
628 Seq was found to perform the best in terms of AUCROC and AUPRC with the FAIRE-seq data.
629 Because the DNase-seq and FAIRE-seq regions were initially been called from traditional
630 methods such as Hotspot (36) and F-Seq (12), the evaluations based on them are expected to be
631 more biased to traditional approaches.
632
633 **Application for human monocytes**
634
635 Thus far, we have shown the performance of HMMRATAC and other algorithms with data
636 generated in the human GM12878 cell line. In order to determine our method's performance with
637 different data sets and different cell lines, we analyzed ATAC-seq data generated in human
638 monocytes subjected to several chemical treatments and sequenced at different time points (23).
639 In addition to ATAC-seq, Novakovic et al. performed ChIP-seq on the same cells, with the same
640 treatments and time points, for the active histone marks histone H3K27ac, H3K4me1 and
641 H3K4me3, and the repressive histone mark histone H3K9me3. We chose five separate data sets
642 generated by that group: RPMI treatment after 4 and 24 hours, LPS treatment after 6 days and
643 BG treatment after 4 and 24 hours. Using the active histone modification as the “Gold standard”
644 and the repressive histone modification as the “real negatives” (See Method section for more
645 details), we compared the performance of the algorithms in the same way we compared their
646 performance in human data (**Figure 4** the second to the sixth row, **Table 1**). We found that
647 HMMRATAC has superior precision, sensitivity, specificity, AUCROC and AUPRC over both
648 MACS2 and F-Seq. Overall, this data indicates that HMMRATAC's superior performance over the
649 other algorithms is consistent across different data sets and different cell lines.
650
651 **Summary of method comparisons**
652
653 The summary is shown in **Table 1**. We used a total of 16 independent measures in two different
654 cell types and six different conditions to benchmark the performance of the three algorithms,
655 including 8 AUCROC and 8 AUPRC. We found that HMMRATAC outperformed the other two
656 methods in 6 of 8 AUCROC and 7 of 8 AUPRC. Furthermore, in all of the 12 evaluations of which
657 the real positives were compiled by integrating multiple histone modification ChIP-seq datasets,
658 HMMRATAC outperformed the other methods consistently. Overall, these results indicate that
659 HMMRATAC is the most accurate and reliable peak-caller for ATAC-seq data analysis.
660
661 **DISCUSSION**

662
663 HMMRATAC is designed to integrate the nucleosome-free and nucleosome-enriched signals
664 derived from a single ATAC-seq dataset in order to identify open chromatin regions. For this
665 reason, HMMRATAC should not be used on ATAC-seq datasets that have undergone either
666 physical or computational size selection. For that same reason, single-end sequence libraries
667 from ATAC-seq cannot be analyzed with HMMRATAC. However, we have shown that size
668 selection results in a decrease in sensitivity and should not be used as a general practice in
669 ATAC-seq protocols. Therefore, a reasonable way to evaluate the quality of ATAC-seq data for
670 HMMRATAC analysis is to check the length distribution of the transposition fragments to ensure
671 that large fragments are present and relatively abundant.
672
673 Identifying transcription factor footprints, wherein the DNA protected by a binding protein shows
674 resistance to nonspecific DNA digestion, is an area of active research. Although footprint
675 identification may be possible with certain datasets or for specific transcription factors, we believe
676 that HMMRATAC identified summits, or the position with the most insertion events within the
677 open region, are more robust to identify potential transcription factor binding sites. In fact, it has
678 been previously shown that DNase-seq signal intensity at a binding motif is a better predictor of
679 transcription factor binding than the strength of a footprint (50), and we believe this principle
680 applies to ATAC-seq as well. However, a major reason for this result from He et al., was that
681 many footprints were artifacts caused by the enzymes' sequence bias. The Tn5 transposase also
682 has a sequence bias, which would need to be corrected before an effective footprint analysis
683 could occur. Several methods exist (51,52) that could correct the sequence bias from an ATAC-
684 seq library. HMMRATAC could be used downstream of such sequence corrections, to reduce
685 bias' in peak calling, or upstream of the corrections, to identify the reads within peaks that need to
686 be corrected for footprint or other downstream analysis.
687
688 It is possible to extend HMMRATAC to identify differentially accessible regions between two or
689 more conditions. Most of the approaches for identifying such differences from ATAC-seq data
690 would suffer from the same problems mentioned before, namely, that these methods only
691 consider the relative strength of the signal and the differences in signal intensity between
692 conditions. As we've shown here, the incorporation of the structure of regulatory chromatin into a
693 model could increase the accuracy and sensitivity of calling accessible regions. We believe our
694 strategy of "decomposition and integration" can be adopted in a differential accessibility analysis
695 pipeline, by studying the differential NFRs and nucleosomal signals separately and then
696 combining together.
697

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. 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.

ACKNOWLEDGEMENTS

We thank Dr. Michael Buck, Dr. Marc Halfon, and Dr. Yijun Sun for their expertise and assistance throughout all aspects of our study.

FUNDING

This work was supported by the startup fund from the University at Buffalo, the Buffalo Blue Sky Award from the University at Buffalo, the NIH/NCI Roswell Park CCSG grant (P30CA016056), and the NIH/NCI IOTN DMRC grant (U24CA232979).

735 **AUTHORS' CONTRIBUTIONS**

736

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

740

741 **CONFLICT OF INTEREST**

742

743 The authors declare that they have no conflict of interests.

744

745 **REFERENCES**

746

747 1. Li, G. and Reinberg, D. (2011) Chromatin higher-order structures and gene regulation.
748 *Current opinion in genetics & development*, **21**, 175-186.
749 2. Bell, O., Tiwari, V.K., Thoma, N.H. and Schubeler, D. (2011) Determinants and dynamics
750 of genome accessibility. *Nature reviews. Genetics*, **12**, 554-564.
751 3. Tsompana, M. and Buck, M.J. (2014) Chromatin accessibility: a window into the genome.
752 *Epigenetics & chromatin*, **7**, 33.
753 4. Taberlay, P.C., Statham, A.L., Kelly, T.K., Clark, S.J. and Jones, P.A. (2014)
754 Reconfiguration of nucleosome-depleted regions at distal regulatory elements
755 accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res*, **24**,
756 1421-1432.
757 5. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S.
758 and Crawford, G.E. (2008) High-resolution mapping and characterization of open
759 chromatin across the genome. *Cell*, **132**, 311-322.
760 6. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. (2007) FAIRE
761 (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory
762 elements from human chromatin. *Genome Res*, **17**, 877-885.
763 7. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013)
764 Transposition of native chromatin for fast and sensitive epigenomic profiling of open
765 chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, **10**, 1213-
766 1218.
767 8. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P.,
768 Chang, H.Y. and Greenleaf, W.J. (2015) Single-cell chromatin accessibility reveals
769 principles of regulatory variation. *Nature*, **523**, 486-490.

770 9. Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L.
771 *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin
772 accessibility data in human and mouse. *Nucleic Acids Res*, **45**, D658-D662.
773 10. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G.,
774 Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the
775 human genome. *Cell*, **129**, 823-837.
776 11. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum,
777 C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq
778 (MACS). *Genome biology*, **9**, R137.
779 12. Boyle, A.P., Guinney, J., Crawford, G.E. and Furey, T.S. (2008) F-Seq: a feature density
780 estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537-2538.
781 13. Schep, A.N., Buenrostro, J.D., Denny, S.K., Schwartz, K., Sherlock, G. and Greenleaf,
782 W.J. (2015) Structured nucleosome fingerprints enable high-resolution mapping of
783 chromatin architecture within regulatory regions. *Genome Res*, **25**, 1757-1770.
784 14. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat*
785 *Methods*, **9**, 357-359.
786 15. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-
787 Wheeler transform. *Bioinformatics*, **25**, 1754-1760.
788 16. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W.
789 and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic*
790 *Acids Res*, **37**, W202-208.
791 17. Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y. and Pritchard, J.K. (2011)
792 Accurate inference of transcription factor binding from DNA sequence and chromatin
793 accessibility data. *Genome Res*, **21**, 447-455.
794 18. Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J.,
795 Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R. *et al.* (2012) Differential oestrogen
796 receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389-
797 393.
798 19. Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C.A., Zhang,
799 Y. and Liu, X.S. (2013) Target analysis by integration of transcriptome and ChIP-seq data
800 with BETA. *Nat Protoc*, **8**, 2502-2515.
801 20. Andrews, S. (2019).
802 21. Ou, J., Liu, H., Yu, J., Kelliher, M.A., Castilla, L.H., Lawson, N.D. and Zhu, L.J. (2018)
803 ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-
804 seq data. *BMC Genomics*, **19**, 169.

805 22. Kodama, Y., Shumway, M., Leinonen, R. and International Nucleotide Sequence
806 Database, C. (2012) The Sequence Read Archive: explosive growth of sequencing data.
807 *Nucleic Acids Res*, **40**, D54-56.
808 23. Novakovic, B., Habibi, E., Wang, S.Y., Arts, R.J.W., Davar, R., Megchelenbrink, W., Kim,
809 B., Kuznetsova, T., Kox, M., Zwaag, J. *et al.* (2016) beta-Glucan Reverses the Epigenetic
810 State of LPS-Induced Immunological Tolerance. *Cell*, **167**, 1354-1368 e1314.
811 24. Sohn, K.A., Ho, J.W., Djordjevic, D., Jeong, H.H., Park, P.J. and Kim, J.H. (2015) hiHMM:
812 Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*, **31**,
813 2066-2074.
814 25. Ho, J.W., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A.,
815 Tolstorukov, M.Y., Appert, A. *et al.* (2014) Comparative analysis of metazoan chromatin
816 organization. *Nature*, **512**, 449-452.
817 26. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A. and Noble, W.S. (2012)
818 Unsupervised pattern discovery in human chromatin structure through genomic
819 segmentation. *Nat Methods*, **9**, 473-476.
820 27. Chan, R.C.W., Libbrecht, M.W., Roberts, E.G., Bilmes, J.A., Noble, W.S. and Hoffman,
821 M.M. (2018) Segway 2.0: Gaussian mixture models and minibatch training.
822 *Bioinformatics*, **34**, 669-671.
823 28. Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970) A Maximization Technique
824 Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The*
825 *Annals of Mathematical Statistics*, **41**, 164-171.
826 29. Viterbi, A. (1967) Error bounds for convolutional codes and an asymptotically optimum
827 decoding algorithm. *IEEE Trans. Inf. Theor.*, **13**, 260-269.
828 30.
829 https://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wg
830 EncodeDacMapabilityConsensusExcludable.bed.gz
831 31. https://genome.ucsc.edu/FAQ/FAQformat.html#format14.
832 32. Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and
833 characterization. *Nat Methods*, **9**, 215-216.
834 33. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang,
835 X., Wang, L., Issner, R., Coyne, M. *et al.* (2011) Mapping and analysis of chromatin state
836 dynamics in nine human cell types. *Nature*, **473**, 43-49.
837 34. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E.,
838 Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B. *et al.* (2012) The accessible
839 chromatin landscape of the human genome. *Nature*, **489**, 75-82.
840 35. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human
841 genome. *Nature*, **489**, 57-74.

842 36. John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager,
843 G.L. and Stamatoyannopoulos, J.A. (2011) Chromatin accessibility pre-determines
844 glucocorticoid receptor binding patterns. *Nat Genet*, **43**, 264-268.
845 37. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing
846 genomic features. *Bioinformatics*, **26**, 841-842.
847 38. Lantermann, A.B., Straub, T., Stralfors, A., Yuan, G.C., Ekwall, K. and Korber, P. (2010)
848 Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning
849 mechanisms distinct from those of Saccharomyces cerevisiae. *Nature structural &*
850 *molecular biology*, **17**, 251-257.
851 39. Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z. and Sidow, A. (2011)
852 Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516-520.
853 40. Liu, Y., Liu, X.S., Wei, L., Altman, R.B. and Batzoglou, S. (2004) Eukaryotic regulatory
854 element conservation analysis and identification using comparative genomics. *Genome*
855 *Res*, **14**, 451-458.
856 41. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K.,
857 Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved
858 elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034-1050.
859 42. Shin, H., Liu, T., Duan, X., Zhang, Y. and Liu, X.S. (2013) Computational methodology for
860 ChIP-seq analysis. *Quantitative biology (Beijing, China)*, **1**, 54-70.
861 43. Koohy, H., Down, T.A., Spivakov, M. and Hubbard, T. (2014) A Comparison of Peak
862 Callers Used for DNase-Seq Data. *PLoS One*, **9**, e96303.
863 44. Hansel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di
864 Antonio, M., Pike, J., Kimura, H., Narita, M. *et al.* (2016) G-quadruplex structures mark
865 human regulatory chromatin. *Nat Genet*, **48**, 1267-1272.
866 45. Denny, S.K., Yang, D., Chuang, C.H., Brady, J.J., Lim, J.S., Gruner, B.M., Chiou, S.H.,
867 Schep, A.N., Baral, J., Hamard, C. *et al.* (2016) Nfib Promotes Metastasis through a
868 Widespread Increase in Chromatin Accessibility. *Cell*, **166**, 328-342.
869 46. Philip, M., Fairchild, L., Sun, L., Horste, E.L., Camara, S., Shakiba, M., Scott, A.C., Viale,
870 A., Lauer, P., Merghoub, T. *et al.* (2017) Chromatin states define tumour-specific T cell
871 dysfunction and reprogramming. *Nature*, **545**, 452-456.
872 47. Chen, X., Shen, Y., Draper, W., Buenrostro, J.D., Litzenburger, U., Cho, S.W., Satpathy,
873 A.T., Carter, A.C., Ghosh, R.P., East-Seletsky, A. *et al.* (2016) ATAC-see reveals the
874 accessible genome by transposase-mediated imaging and sequencing. *Nat Methods*, **13**,
875 1013-1020.
876 48. Risca, V.I., Denny, S.K., Straight, A.F. and Greenleaf, W.J. (2017) Variable chromatin
877 structure revealed by in situ spatially correlated DNA cleavage mapping. *Nature*, **541**,
878 237-241.

879 49. Khan, A. and Zhang, X. (2016) dbSUPER: a database of super-enhancers in mouse and
880 human genome. *Nucleic Acids Res*, **44**, D164-171.
881 50. He, H.H., Meyer, C.A., Hu, S.S., Chen, M.W., Zang, C., Liu, Y., Rao, P.K., Fei, T., Xu, H.,
882 Long, H. *et al.* (2014) Refined DNase-seq protocol and data analysis reveals intrinsic bias
883 in transcription factor footprint identification. *Nat Methods*, **11**, 73-78.
884 51. Wang, J.R., Quach, B. and Furey, T.S. (2017) Correcting nucleotide-specific biases in
885 high-throughput sequencing data. *BMC Bioinformatics*, **18**, 357.
886 52. Martins, A.L., Walavalkar, N.M., Anderson, W.D., Zang, C. and Guertin, M.J. (2018)
887 Universal correction of enzymatic sequence bias reveals molecular signatures of
888 protein/DNA interactions. *Nucleic Acids Res*, **46**, e9.
889
890
891 **TABLE AND FIGURES LEGENDS**
892
893 **Table 1. The Area Under the Curves (AUCs) of method comparisons of HMMRATAC,**
894 **MACS2, and F-Seq.**
895
896 The AUC calculated from the Receiver Operator Characteristic (ROC) curves and Precision-
897 Recall (PR) curves in **Figure 4** and **Supplemental Figure S11**. Because we can't reach the
898 extreme sides of the curves while tuning cutoff values, before we calculated AUC values, we
899 added the following points to ROC curves– (0, 0) and (1, 1), and to PR curves – (0, 1) and (1,
900 $L_{true}/(L_{true} + L_{false})$). The L_{true} is the total length of the real positive set, and the L_{false} is the total
901 length of the real negative set. The numbers of the best performance are shown in bold.
902
903 **Figure 1. The HMMRATAC Algorithm.**
904 **(A)** A schematic to represent the structure of an accessible region. **(B)** The algorithmic
905 workflow. **(C)** Estimation of fragment length parameters through the Expectation-Maximization
906 algorithm. Mean and standard deviation parameters per EM iteration for three nucleosomal
907 distributions are consistent among datasets using 50,000 and 500 cells. The parameters for NFR
908 signals are fixed during EM training so they were not plotted. Iteration #1 shows the initial
909 parameters. EM was applied to pooled biological replicates. The parameters per biological
910 replicate were shown in **Supplemental Figure S3**. **(D)** The transition and emission parameters of
911 the Hidden Markov Model trained through the Baum-Welch algorithm on the 50,000 GM12878
912 cells ATAC-seq data.
913
914 **Figure 2. Chromatin Architecture at HMMRATAC Peaks.**

915 **(A)** IGV browser snapshot of various signals around a negative locus on chromosome 8 and a
916 positive locus on chromosome 2. Top four tracks show ATAC-seq signals, after being separated
917 by size, the fifth track shows CTCF ChIP-seq signal, the sixth through ninth tracks show various
918 histone modification ChIP-seq signal and the tenth track shows MNase-seq signal. The bottom
919 panel shows RefSeq genes, HMMRATAC calls and state annotations (called with default settings,
920 red: *background* state, green: *nucleosome* state, and blue: *center* state), MACS2 and F-Seq
921 peaks (called with default parameters), the active regions combined from "Active Promoters" and
922 "Strong Enhancers" states from chromHMM (see **Methods**). **(B)** ChIP-seq of H3K4me1,
923 H3K4me3, H3K27ac, H3K27me3, H3K9ac, and H3K9me3, DNase-seq and MNase-seq signals
924 (log2 transformed library size normalized pileup) plotted across all distal (> 10 kbps from gene
925 promoters) *center* states. Data were clustered using k-means algorithm into 4 clusters.
926
927 **Figure 3. Chromatin features of HMMRATAC states.**
928 **(A)** The overlap between either histone modification ChIP-seq peaks or chromatin states in
929 GM12878 and three HMMRATAC states – *background*, *nucleosome,* and *center*. The color
930 shows the percentage of GC, the percentage of total genome length, or the percentage of histone
931 peaks or chromatin states overlapping with each of HMMRATAC states. Overlaps were
932 calculated in basepair level. **(B)** Distribution of GM12878 H3K4me1 ChIP-seq signal in 1000 bps
933 window surrounding the summits of F-Seq peaks, MACS2 peaks, or HMMRATAC *center* and
934 *nucleosome* states. **(C)** Distribution of Phastcons conservation score within vertebrates in 1000
935 bps window surrounding the summits of F-Seq, MACS2, and HMMRATAC calls.
936
937 **Figure 4. Comparison of performance between HMMRATAC, MACS2, and F-Seq**
938 Left: Receiver Operator Characteristic (ROC) curves and right: precision-recall (PR) curves for
939 the real positive and real negative pairs of 1. active chromatin states vs. heterochromatin for
940 GM12878 cells, 2. active histone marks (either H3K4me1, H3K4me3 or H3K27ac) vs.
941 heterochromatin (H3K9me3) for human monocytes (see **Methods** for detail).
942

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

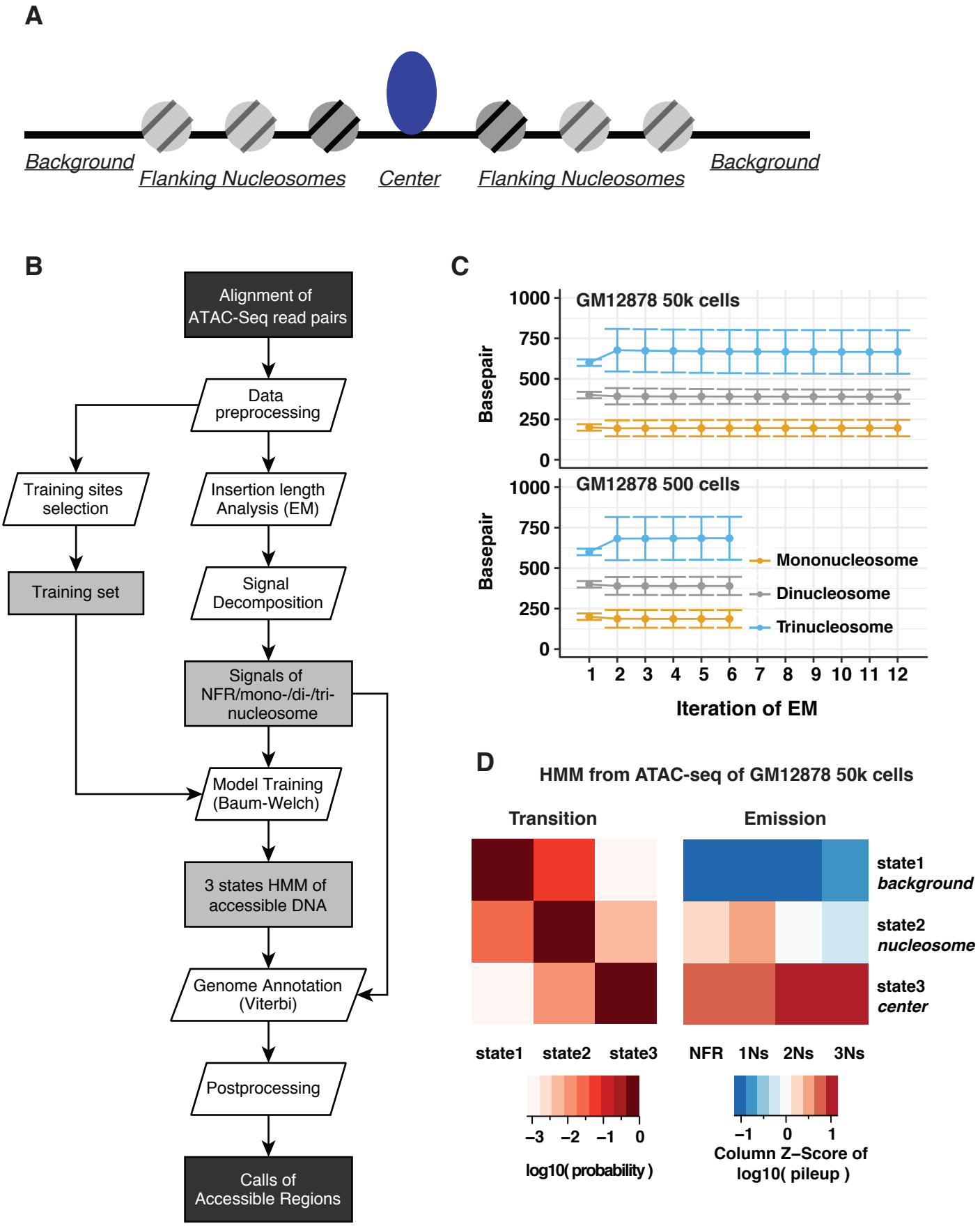

A
Background
Flanking Nucleosomes
Center
Flanking Nucleosomes
Background
B
Alignment of ATAC-Seq read pairs
Data preprocessing
Training sites selection
Insertion length Analysis (EM)
Training set
Signal Decomposition
Signals of NFR/mono-/di-/tri-nucleosome
Model Training (Baum-Welch)
3 states HMM of 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 1Ns 2Ns 3Ns
−3 −2 −1 0
log10(probability)
−1 0 1
Column Z−Score of log10(pileup)

Figure 2

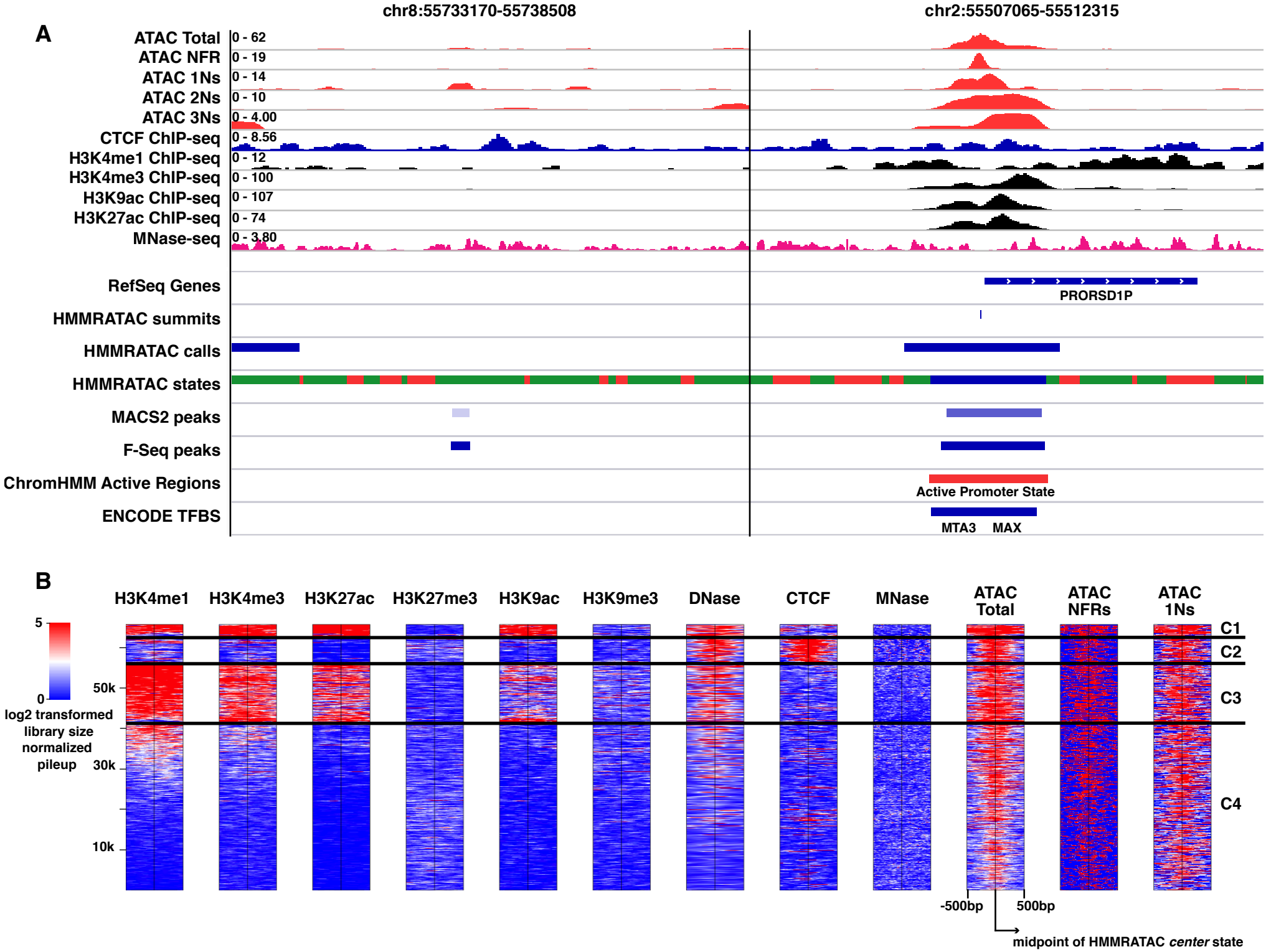

A
chr8:55733170-55738508
chr2:55507065-55512315
ATAC Total
ATAC NFR
ATAC 1Ns
ATAC 2Ns
ATAC 3Ns
CTCF ChIP-seq
H3K4me1 ChIP-seq
H3K4me3 ChIP-seq
H3K9ac ChIP-seq
H3K27ac ChIP-seq
MNase-seq
0 - 62
0 - 19
0 - 14
0 - 10
0 - 4.00
0 - 8.56
0 - 12
0 - 100
0 - 107
0 - 74
0 - 3.80
RefSeq Genes
PRORSD1P
HMMRATAC summits
HMMRATAC calls
HMMRATAC states
MACS2 peaks
F-Seq peaks
ChromHMM Active Regions
Active Promoter State
ENCODE TFBS
MTA3
MAX
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

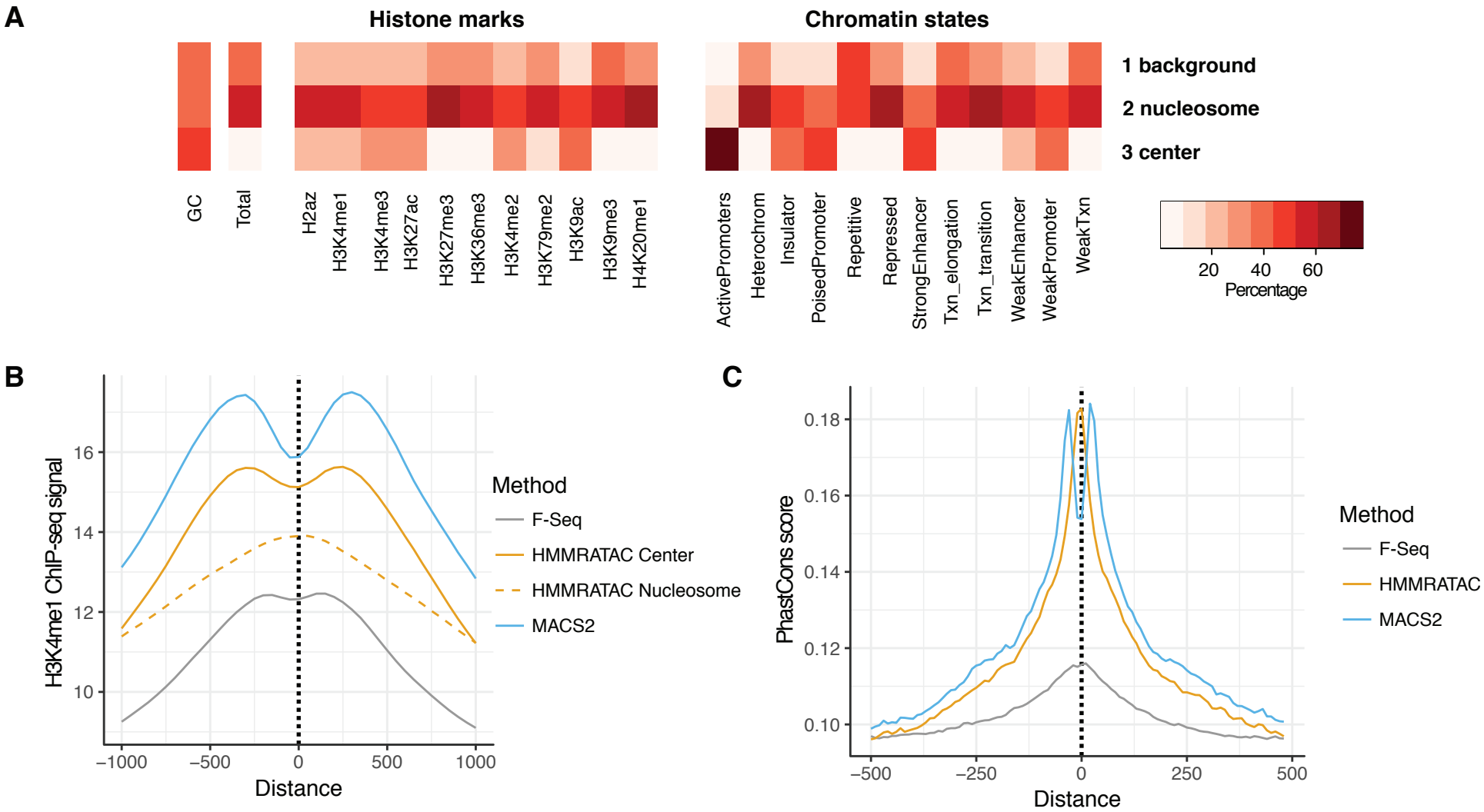
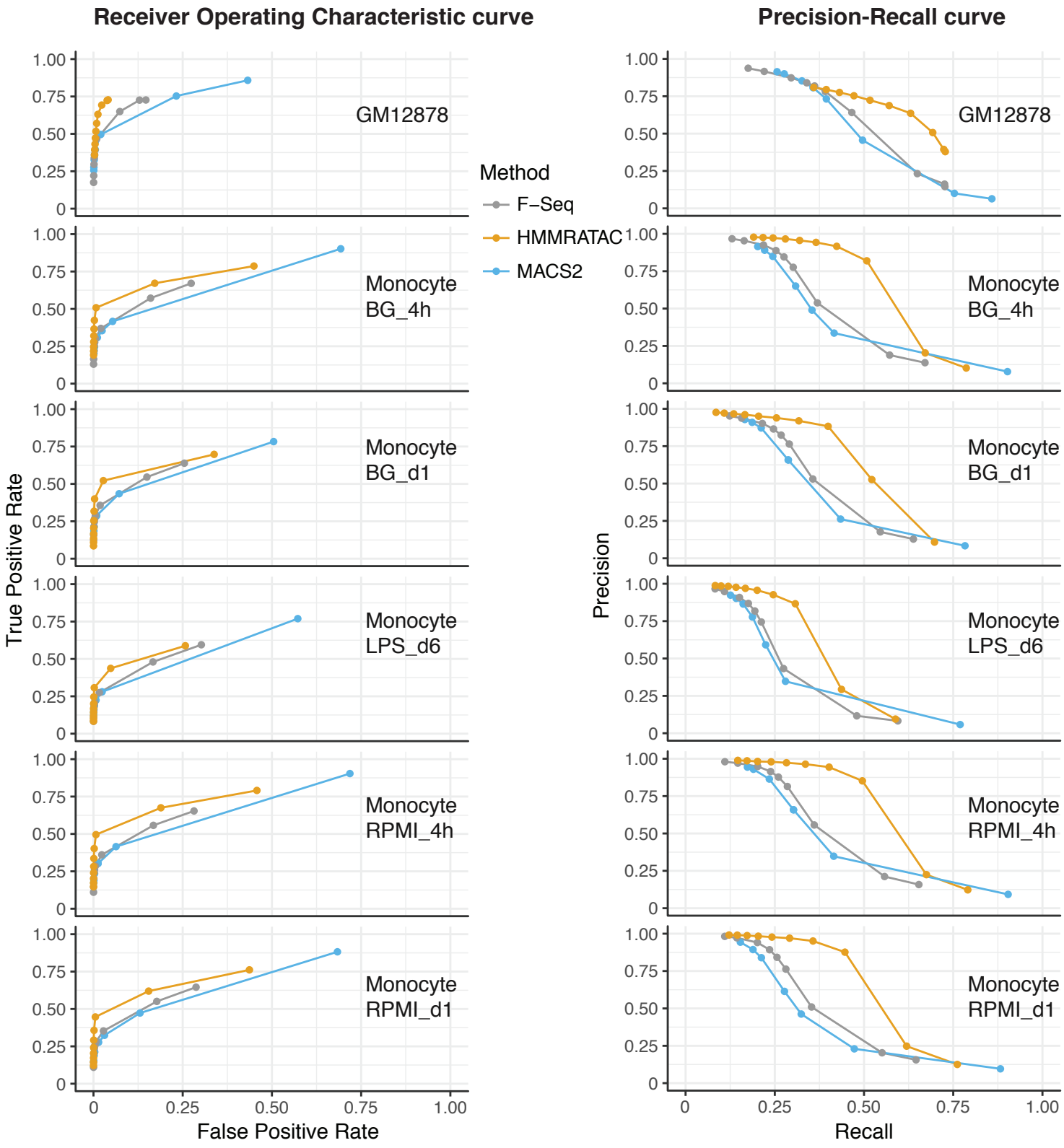

A
Histone marks
Chromatin states
1 background
2 nucleosome
3 center
GC
Total
H2az
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K36me3
H3K4me2
H3K79me2
H3K9ac
H3K9me3
H4K20me1
ActivePromoters
Heterochrom
Insulator
PoisedPromoter
Repetitive
Repressed
StrongEnhancer
Txn_elongation
Txn_transition
WeakEnhancer
WeakPromoter
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