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 that creates different modes representing transposase insertion into NFRs and linker regions. This allows ATAC-seq to elucidate multiple layers of information relative to the other assays. Although computational tools exist for DNase-seq, FAIRE-seq and ChIP-seq [8], that can be and are used for ATAC-seq analysis, such as MACS2 [9] and F-Seq [10], these would fall short since they only utilize a subset of information, usually the nucleosome-free signals. To date, there are no dedicated peak-callers specifically to account for ATAC-seq.

We present here HMMRATAC, the Hidden Markov ModeleR for ATAC-seq, a semi-supervised machine learning approach for identifying open chromatin regions from ATAC-seq data. The principle concept of HMMRATAC is built upon "decomposition and integration", whereby a single ATAC-seq dataset is decomposed into different layers of signals corresponding to different chromatin features, and then the relationships between the layers of signals at open chromatin regions are learned and utilized for pattern reorganization (Fig. 1a and 1b). Our method takes advantage of the unique features of

ATAC-seq to more accurately identify the chromatin structure. We found that HMMRATAC was able to identify chromatin architecture and the most likely transcription factor binding sites. Additionally, compared with existing methods used for ATAC-seq analysis, HMMRATAC outperformed them in most tests, including recapitulating active and/or open chromatin regions identified with other assays.

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

Results

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

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

Signal decomposition through the probabilistic approach

In the first ATAC-seq paper by Buenrostro et al., the authors separated the fragments according to their insertion lengths into four populations representing those from the NFRs

114 and those spanning 1, 2 or 3 nucleosomes [5]. Non-overlapping and non-adjacent cutoffs
115 were used. For example, lengths below 100 bp represent nucleosome-free signals and
116 lengths between 180 and 250 bp represents mono-nucleosome signals. We consider such
117 cutoff-based strategy not a general practice for the following two reasons. First, different
118 species or different cell lines may have different nucleosome spacing and these differences
119 may not be properly understood [12, 13]. Secondly, since the fragments whose size falls
120 between the non-overlapping and non-adjacent cutoffs are discarded (e.g. those between
121 100 and 180 bp long), the recall (or sensitivity) to identify accessible regions will be
122 reduced (Additional file 1: Fig. S2 and explained later). As a fact, about 15% of the entire
123 data generated from the original paper are ignored with their strategy. We addressed the
124 above two challenges in a probabilistic approach in HMMRATAC.
125
126 First, we let the data itself identify the optimal parameters for the four different
127 distributions of fragment lengths. Using the distributional assumptions of Buenrostro et al.,
128 we create four distributions to represent four different chromatin features: nucleosome-free
129 regions (NFRs), mono-nucleosomes (1Ns), di-nucleosomes (2Ns) and tri-nucleosomes
130 (3Ns). HMMRATAC uses a mixture model and Expectation-Maximization (EM)
131 algorithm to identify optimum parameters (Fig. 1c; Additional file 1: Fig S3; see method
132 section for more detail). As an example, HMMRATAC identified the optimum fragment
133 lengths representing 1Ns, 2Ns, and 3Ns from ATAC-seq data in the human GM12878 cell
134 line, generated by merging three replicates from [5], as being 195bp, 396bp, and 693bp,
135 respectively. As a demonstration of different nucleosome spacing in different datasets,
136 HMMRATAC determined the optimum fragment lengths for the 1Ns, 2Ns, and 3Ns from

137 ATAC-seq data in human monocytes treated with RPMI for one day [14] as 186bp, 368bp,
138 and 549bp, respectively. Additionally, the values for any of the individual replicates, in
139 the GM12878 data [5], were never more than 5bp off from the average values listed above
140 for the merged datasets (Additional file 1: Fig S3). Not all ATAC-seq datasets show a
141 clear multi-modal pattern, especially datasets from fewer cells. As an example, the
142 fragment distribution generated by Buenrostro and et al. utilizing 500 cells (GM12878
143 cells) per-replicate, did not show as clear of a distribution as 50,000 cells per replicate
144 (Additional file 1: Fig. S4a and S4b) [5]. However, we found our approach can robustly
145 identify the parameters even if the multi-modal pattern is not clear. As we can see from
146 Fig. 1c, the mean and standard deviation values converged after no more than 12 iterations
147 and the parameters converged to very similar values for either the 50,000-cell per-replicate
148 or 500 cell per-replicate ATAC-seq datasets in GM12878.
149
150 Secondly, we propose in HMMRATAC a probabilistic approach to make use of data from
151 the entire spectrum of fragment lengths. We generate four separated signal profiles in 10bp
152 resolution, for NFRs, 1Ns, 2Ns, and 3Ns, allowing each mapped fragment to contribute to
153 all profiles with different weights corresponding to its particular fragment length. For each
154 genomic location that is occupied by a particular fragment, all four signal profiles are
155 incremented by the probabilities that this fragment, assigned with an observed fragment
156 length through read mapping, belongs to each of the four corresponding distributions. We
157 tested the performance of a regular HMMRATAC approach with this probabilistic method
158 utilizing the entire data and an alternative approach utilizing only the subset of data
159 satisfying the cutoffs proposed in Buenrostro et al. The two approaches were compared in

160 terms of precision and recall against the "active regions" true set (Additional file 1: Fig.
161 S2; described in more detail below). We found that our probabilistic method had a
162 considerably better recall and precision compared to the cutoff based method.
163
164 HMMRATAC utilizes the length of an insertion fragment, identified by paired-end
165 sequencing, as an important and critical piece of information in its processes. As such,
166 HMMRATAC can only be used in paired-end sequencing libraries that have not undergone
167 stringent size selection, either physical or computational. Because we observed a decrease
168 in the performance of our method when we used the cutoff based approach, which resulted
169 in the removal of ~15% of all sequencing reads, we wanted to understand the effects of
170 size selection and paired-end sequencing on ATAC-seq peak calling. We performed two
171 *in silico* size selections on the data from Buenrostro et al. The first restricted the fragments
172 to those whose length was under 100bp and the second restricted the fragments to those
173 with lengths under 300bp. We also treated the data as single-end, by unpairing the read
174 mate pairs. We then used MACS2 [9] to call peaks with each data set, including the
175 original unaltered data and compared the performance to predict the "active regions" in
176 chromatin (Additional file 1: Fig S5; see methods and below for more detail). As a result,
177 the original unaltered data performed the best in terms of precision and recall. This
178 indicates that an ATAC-seq assay should be performed without a size selection step and
179 the resulting fragments should undergo paired-end sequencing, regardless of the peak
180 calling algorithm that is used.
181
182 **Integration through Hidden Markov Model**

183

184 After removing reads with low mapping quality, and masking blacklist regions from
185 downstream analysis [15], HMMRATAC identifies a set of training regions throughout
186 the genome to train a 3-states Hidden Markov Model (HMM) with multivariate Gaussian
187 emissions, representing the underlying three chromatin states of 1) the *center* of open
188 chromatin, 2) the *nucleosome regions* and 3) the *background* (Fig. 1a). By default, training
189 regions can be either specified by users or automatically selected where the signal fold-
190 change above background falls within a predetermined range. After signal decomposition,
191 a matrix of 10bp resolution signals for NFRs, 1Ns, 2Ns, and 3Ns in training regions is used
192 to train the parameters in HMM through the Baum-Welch algorithm [16]. The initial state
193 transition probabilities are set equally as 0.33 and the emission parameters for each state
194 are calculated after clustering all locations in the training regions into 3 classes (See
195 methods section for more detail). Fig. 1d shows the final parameters for the HMM on
196 ATAC-seq in the GM12878 cell line, again generated by merging three replicates from [5].
197 Although the parameters will change depending on the choice of training regions and the
198 datasets provided, several key characteristics are evident in the model and remain largely
199 stable with different settings and data. State 3 in the HMM model, which HMMRATAC
200 annotated as the *center* state, showed the highest average signal across the four signal
201 profiles. State 2, annotated as the *nucleosome* state, showed the second highest average
202 values. Additionally, as we hypothesized, the results show the *center* state mainly transits
203 from or to the *nucleosome* state.

204

205 After learning the model, HMMRATAC takes a matrix of genome-wide NFRs, 1Ns, 2Ns,
206 and 3Ns signals and labels each genomic locations with one of the states using the Viterbi
207 algorithm [17]. On the 50K cells ATAC-seq data in GM12878 cell line, there were about
208 4.3 million regions annotated as *background* with an average size of 288 bps (~35% of
209 genome size), about 4.5 million regions annotated as *nucleosomes* with an average size of
210 382 bps (~49%), and about 226 thousand regions annotated as *center regions* with an
211 average size of 504 bps (~3%). HMMRATAC finally connected the *center regions* with
212 their adjacent *nucleosomes* and called a total of 87,440 accessible regions with an average
213 size of 1,831 bps (5% of genome size), excluding those with abnormally high coverage and
214 those below a minimum length (see method section for more details). It should be noted
215 that many of the *nucleosomes* are not adjacent to the *center* state. These may represent
216 non-regulatory nucleosomes that ATAC-seq is able to identify or may be artifacts due to
217 increased transposition frequency.
218
219 **Recapitulating chromatin architecture at functional elements**
220
221 We hypothesized that HMMRATAC could identify chromatin architecture around
222 accessible regions, or ideally, nucleosomes near the regulatory elements. To prove this,
223 we explored the distribution of histone modification ChIP-seq signal around our identified
224 accessible regions. We downloaded and analyzed datasets released by ENCODE in the
225 GM12878 cell line [11], including ChIP-seq for CTCF, active histone marks such as
226 H3K4me1, H3K4me3, H3K9ac, and H3K27ac, repressive histone marks such as
227 H3K9me3 and H3K27me3, as well as sequencing of micrococcal nuclease digestion

228 (MNase-seq). Fig. 2a shows two loci, one on chromosome 8 that was identified as an
229 accessible region by MACS2 and F-Seq but not HMMRATAC and another locus on
230 chromosome 2 that was identified as accessible by all three methods. At the chromosome
231 2 locus, we observed that the *center* state was located in regions characterized by depletion
232 in both active histone marks and MNase digestion. The surrounding *nucleosome* states
233 showed enrichment in active histone marks and MNase digestion. We next plotted the
234 same ChIP-seq, DNase-seq and MNase-seq data around all *center* states and used k-means
235 clustering to group the resulting profiles (Fig. 2b). We found that HMMRATAC states can
236 identify, what appear to be, different types of regulatory elements. For example, the first
237 cluster (C1) characterized by strong enrichment of H3K27ac relative to H3K4me1 seems
238 to indicate active enhancers while the third cluster (C3) marked by high levels of H3K4me1
239 and more moderate levels of H3K27ac would suggest poised enhancers (Fig. 2b). The
240 second cluster (C2), characterized by strong enrichment of CTCF binding and lower levels
241 of histone modifications, seems to suggest that this cluster represents insulators or domain
242 boundaries. Taken together, these results indicate that HMMRATAC can recapitulate the
243 histone architecture around accessible regions.
244
245 The chromosome 8 locus exemplifies some of the advantages of HMMRATAC over other
246 peak callers. Both MACS2 and F-Seq identified this locus as being accessible, despite the
247 lack of other corroborating evidence, such as transcription factor binding and the presence
248 of active chromatin states. It appears that the site was called by these methods because of
249 an enrichment of the mono-nucleosome signal. Because neither MACS2 or F-Seq integrate
250 the various chromatin features present in ATAC-seq data and because they are based on

251 signal enrichment only, as opposed to HMMRATAC which incorporates the structure of
252 the element, they both falsely identified this region as accessible, while HMMRATAC
253 identified the region as being a *nucleosome* state.
254
255 In order to understand the chromatin features that are enriched in each state across the
256 genome, as the unique result HMMRATAC is able to conclude from a single ATAC-seq
257 data, we plotted the overlap of the 3 hidden states with chromatin states and histone mark
258 calls from ChIP-seq in the GM12878 cell line based on ENCODE data (Fig 3a; Additional
259 file 1: Table S1). We found that the chromatin states "Active Promoters", "Strong
260 Enhancer", "Insulator", "Poised Promoter" and "Weak Promoter" states mainly overlaps
261 with our *center* state, and the "Heterochromatin", "Repetitive", "Repressed",
262 "Transcription Elongation", and "Weak Transcription" barely overlaps with the *center*
263 state. Interestingly, the "Strong Enhancer" state had a similar amount of overlap with the
264 *nucleosome* state as the *center* state. Combined with our observation that our *nucleosome*
265 state showed elevated levels of active histone marks near accessible regions, we
266 hypothesized that this state may represent regulatory nucleosomes, those marked by active
267 histone modifications and adjacent to accessible regions. We plotted the H3K4me1 active
268 histone mark around the centers of our *center* states as well as around the centers of the
269 adjacent *nucleosome* states, those directly flanking our *center* states (Fig 3b). As we can
270 see, the *center* states show a bimodal enrichment for the histone mark, consistent with
271 MACS2 and F-Seq peaks. Additionally, our adjacent *nucleosome* states show central
272 enrichment for the histone mark. This indicates that the *nucleosome* state, at least those
273 adjacent to our *center* states, are likely regulatory nucleosomes. As a result, HMMRATAC

274 merges the *center* states with their adjacent *nucleosome* states, both upstream and
275 downstream, to identify the regulatory region as the output.
276
277 Functional regulatory elements that harbor transcription factor binding sites are generally
278 under higher evolutionary constraint [18]. Success in transcription factor ChIP-seq data
279 analysis is often benchmarked by high PhastCons score, a precompiled evolutionary
280 conservation score at every base pair [19], at the peak center or summit, relative to the
281 surrounding regions [20]. We found that the summits of accessible regions detected by
282 HMMRATAC in GM12878 cell line also showed PhastCons enrichment in vertebrates
283 (Fig 3c), indicating that these summits were likely locations of evolutionarily conserved
284 regulatory elements and could represent functional regulatory elements.
285
286 **Method comparison with MACS2 and F-Seq**
287
288 We next sought to compare the performance of HMMRATAC in identifying open
289 chromatin regions to other popular software for ATAC-seq. We chose MACS2 [9] and F-
290 Seq [10], which had previously been shown to be the most accurate peak-callers in
291 identifying open chromatin regions from DNase-seq [21]. MACS2, originally designed
292 for ChIP-seq, utilizes a local Poison model wherein enrichment is calculated relative to a
293 local Poisson model. F-Seq invokes a Gaussian kernel-smoothing algorithm and then
294 calculates enrichment based on the resulting Gaussian distribution. Although none of them
295 utilize the intrinsic and unique features in ATAC-seq data we described before, MACS2
296 has been used to analyze ATAC-seq data in many published works [6, 14, 22-26], and

297 MACS2 and F-Seq is part of the ENCODE analysis pipeline for ATAC-seq data [27]. We
298 tested the performance of the four algorithms using the two ATAC-seq datasets in
299 GM12878 cell line [5] as the previous sections.
300
301 A major limitation in comparing the performance of identifying open chromatin regions is
302 the absence of a valid "gold standard" [2]. To overcome this problem, we chose to compare
303 the performance of HMMRATAC and the other algorithms to three distinct independent
304 datasets as the proxies of true and false set: 1) the active regions and heterochromatins from
305 chromatin states analysis combining multiple histone modification features[28], 2) DNase-
306 seq [3] and 3) FAIRE-seq [4] accessible regions of the same GM12878 cell line.
307
308 ***Enrichment in chromatin states of active regions***
309
310 The first proxy of the true set that we compared the algorithms to is referred to as "active
311 regions." These sites were defined through an integrative chromatin state segmentation
312 using chromHMM [28, 29] algorithm conducted by ENCODE project. 8 histone
313 modifications (H3K27me3, H3K36me3, H4K20me1, H3K4me1, H3K4me2, H3K4me3,
314 H3K27ac and H3K9ac), 1 transcription factor (CTCF) ChIP-seq and 1 whole genome input
315 datasets in the GM12878 cell line were integrated through an unsupervised machine
316 learning approach and each genomic location was annotated as one of 15 mutually-
317 exclusive states. Two resulting states, "active promoters" and "strong enhancers", were
318 merged together to create our list of "active regions". Additionally, the state annotated as
319 "heterochromatin" was used as a false set. We tested different cutoff values for each

320 algorithm (see Methods) and found that HMMRATAC outperformed the other algorithms,
321 in terms of precision, recall, and false positive rate in identifying these "active regions"
322 while avoiding "heterochromatin" (Fig. 4 top row). The area under the curve (AUC) was
323 then calculated for the recall vs. false positive rate curves (AU-ROC) and for the precision-
324 recall curves (AU-PRC) (Table 1), after adding extreme points to the unreachable ends of
325 the curves. HMMRATAC was proved to have better AUC than the two other methods. It
326 worth noting that HMMRATAC calls larger peaks than the other two methods due to its
327 merging of the adjacent nucleosome states to the center state. However, despite the larger
328 average peak size, HMMRATAC identifies a comparable number of accessible regions in
329 base pairs, as shown in Additional file 1:Fig. S6, at comparable levels of sensitivity for
330 active chromatin regions. This would also indicate that many peaks called by the other
331 methods are falling outside of the "gold standard" regions, possibly contributing to their
332 deficits in precision compared to HMMRATAC. Taken together, these results indicate that
333 HMMRATAC is able to identify active chromatin regions with a higher recall and
334 precision and lower false positive rate, than MACS2 or F-Seq.
335
336 ***Consistency with results from DNase-Seq and FAIRE-Seq***
337
338 We then used the accessible chromatin regions detected by two independent assays DNase-
339 seq and FAIRE-seq in GM12878 cell line as cross-validation of ATAC-seq analysis. We
340 downloaded DNaseI hypersensitive sites (DHSs) [3] and FAIRE-seq [4] accessible regions
341 called by ENCODE in the GM12878 cell line [11, 30], then merged DHSs or FAIRE
342 accessible regions that were within 2kb of each other to define our alternative true sets.

343 Because the DNase-seq and FAIRE-seq regions are called from traditional methods as
344 Hotspot [31] and F-Seq, the evaluation based on a single dataset is less reliable compared
345 with the previous test with chromatin state which integrates multiple evidence. We
346 continued to use the heterochromatin states outlined above as our false set, although we
347 eliminated any DHSs or FAIRE regions that overlapped a heterochromatin state region and
348 vice versa. We found that HMMRATAC outperformed the other two methods in terms of
349 AUC-PRC in recapitulating DHSs (Additional file 1: Fig. S7, Table 1). MACS2 performed
350 slightly better in AUC-ROC by 0.001. F-Seq was found to perform the best in terms of
351 AUC-ROC and AUC-PRC with the FAIRE-seq data.
352
353 ***Reproducibility in replicates***
354
355 A common and robust method to gauge a method's performance in analyzing next-
356 generation sequencing data, in particular, ChIP-seq, is the robustness in technical and
357 biological replicates [32]. One such way to measure the reproducibility of peak callers is
358 the Irreproducible Discovery Rate (IDR) [33]. Therefore we sought to test the
359 reproducibility of HMMRATAC and the other algorithms in utilizing ATAC-seq datasets.
360 We calculated the number of reproducible accessible regions, defined as overlapping
361 accessible regions with an IDR score <= 0.05, between each pair of three replicates for
362 each algorithm. We show that when using the individual replicate data (Additional file 1:
363 Table S2), HMMRATAC produced a similar number of reproducible accessible regions as
364 the other two algorithms. Furthermore, the reproducible peaks called by each algorithm
365 showed a high amount of overlap with each other (Additional File 1: Fig. S8).

366

367

368 Furthermore, we hypothesized that the biology (open chromatin) should be consistent even
369 with a smaller number of cells used in ATAC-seq. We tested how well each algorithm
370 performed at recapitulating the accessible regions called from the 50,000 cell per-replicate
371 data and the 500 cell per-replicate data. We found, similar to the results from pair-wise
372 replicate testing, that HMMRATAC was able to identify a similar number of reproducible
373 accessible regions, with IDR <= 0.05, using datasets with a different number of cells
374 (Additional file 1: Table S2). These data indicate that HMMRATAC is comparable to
375 MACS2, and F-Seq in identifying reproducible accessible regions.

376

377 **Application for human monocytes**

378

379 Thus far, we have shown the performance of HMMRATAC and other algorithms with data
380 generated in the human GM12878 cell line. In order to determine our method's
381 performance with different data sets and different cell lines, we analyzed ATAC-seq data
382 generated in human monocytes subjected to several chemical treatments and sequenced at
383 different time points [14]. In addition to ATAC-seq, Novokovic et al. performed ChIP-seq
384 on the same cells, with the same treatments and time points, for the active histone marks
385 histone H3K27ac, H3K4me1 and H3K4me3 and the repressive histone mark histone
386 H3K9me3. We chose five separate data sets generated by that group: RPMI treatment after
387 4 and 24 hours, LPS treatment after 6 days and BG treatment after 4 and 24 hours. Using
388 the active histone modification as the "Gold standard" and the repressive histone

389 modification as the "real negatives" (See Method section for more details), we compared
390 the performance of the algorithms in the same way we compared their performance in
391 human data (Fig 4 2nd to 6th row, Table 1). We found that HMMRATAC has superior
392 precision, sensitivity, specificity, AUC-ROC and AUC-PRC over both MACS2 and F-Seq.
393 Overall, this data indicates that HMMRATAC's superior performance over the other
394 algorithms is consistent across different data sets and different cell lines.
395
396 **Summary of method comparisons**
397
398 The summary is shown in Table 1. We used a total of 16 independent measures in two
399 different cell types and six different conditions to benchmark the performance of the three
400 algorithms, including 8 AUC-ROC and 8 AUC-PRC. We found that HMMRATAC
401 outperformed the other two methods in 7 of 8 AUC-PRC and 6 of 8 AUC-ROC.
402 Furthermore, in all of the 12 evaluations of which the real positives were compiled by
403 integrating multiple histone modification ChIP-seq datasets, HMMRATAC outperformed
404 the other methods constantly. Overall, these results indicate that HMMRATAC is the most
405 accurate and reliable peak-caller for ATAC-seq data analysis.
406
407 **Discussion**
408
409 HMMRATAC is designed to integrate the nucleosome-free and nucleosome-enriched
410 signals derived from a single ATAC-seq dataset in order to identify open chromatin
411 regions. For this reason, HMMRATAC should not be used on ATAC-seq datasets that

412 have undergone either physical or computational size selection. For that same reason,
413 single-end sequence libraries from ATAC-seq cannot be analyzed with HMMRATAC.
414 However, we have shown that size selection results in a decrease in sensitivity and should
415 not be used as a general practice in ATAC-seq protocols. Therefore, a reasonable way to
416 evaluate the quality of ATAC-seq data for HMMRATAC analysis is to check the length
417 distribution of the transposition fragments to ensure that large fragments are present and
418 relatively abundant.
419
420 Identifying transcription factor footprints, wherein the DNA protected by a binding protein
421 shows resistance to nonspecific DNA digestion, is an area of active research. Although
422 footprint identification may be possible with certain datasets or for certain transcription
423 factors, we believe that HMMRATAC identified summits, or the position with the most
424 insertion events within the open region, are more robust to identify potential transcription
425 factor binding sites. In fact, it has been previously shown that DNase-seq signal intensity
426 at a binding motif is a better predictor of transcription factor binding than the strength of a
427 footprint [34], and we believe this principle applies to ATAC-seq as well. However, a
428 major reason for this result from He et al., was that many footprints were artifacts caused
429 by the enzymes' sequence bias. The Tn5 transposase also has a sequence bias, which
430 would need to be corrected before an effective footprint analysis could occur. Several
431 methods exist [35, 36] that could correct the sequence bias from an ATAC-seq library.
432 HMMRATAC could be used downstream of such sequence corrections, to reduce bias' in
433 peak calling, or upstream of the corrections, to identify the reads within peaks that need to
434 be corrected for footprint or other downstream analysis.

435

436 It is possible to extend HMMRATAC to identify differentially accessible regions between
437 two or more conditions. Most of the approaches for identifying such differences from
438 ATAC-seq data would suffer from the same problems mentioned before, namely, that these
439 methods only consider the relative strength of the signal and the differences in signal
440 intensity between conditions. As we've shown here, the incorporation of the structure of
441 regulatory chromatin into a model could increase the accuracy and sensitivity of calling
442 accessible regions. We believe our strategy of "decomposition and integration" can be
443 adopted in a differential accessibility analysis pipeline, by studying the differential NFRs
444 and nucleosomal signals separately and then combining together.

445

446 **Conclusions**

447

448 ATAC-seq shows numerous advantages over other methods for identifying open chromatin
449 regions, such as DNase-seq and FAIRE-seq, owing to its low starting material requirement
450 and simple protocol. For these reasons, ATAC-seq has become one of the standard assays
451 for locating regions of open chromatin, particularly as biomedical research continues to
452 move toward translational research and precision medicine. Although the number of
453 published ATAC-seq datasets increases rapidly, researchers are still using methods
454 originally developed for ChIP-seq to analyze the data. We present HMMRATAC as a
455 dedicated algorithm for the analysis of ATAC-seq data. We take advantage of the nature
456 of the ATAC-seq experiment that due to the lack of size-selection while preparing the DNA
457 library in its protocol, DNA fragments associated with well-positioned nucleosomes

458 around open chromatin will be sequenced as well. Therefore, as a side note, HMMRATAC
459 shouldn't be applied to those ATAC-seq data generated with a stringent size-selection on
460 DNA fragments. Different from current methods where only the read enrichment is
461 considered, HMMRATAC separates and integrates ATAC-seq data to create a statistical
462 model that identifies the chromatin structure at accessible and active genomic regions. We
463 have shown that HMMRATAC outperforms other computational methods in recapitulating
464 open and active chromatin identified with other assays such as chromatin state analysis
465 integrating multiple histone mark ChIP-seq, DNase-seq and FAIRE-seq. As HMMRATAC
466 is a cross-platform and user-friendly algorithm, we envision it becoming the standard for
467 ATAC-seq data analysis pipeline, replacing current methods designed for ChIP-seq
468 analysis.
469

470 ## Methods

471

472 **Preprocessing of ATAC-seq data**

473

474 The human GM12878 cell line ATAC-seq paired-end data used in this study was publicly
475 available and downloaded under six SRA [37] accession numbers SRR891269-
476 SRR891274[5]. There are three biological replicates generated using 50,000 cells per
477 replicate and other three generated using 500 cells per replicate. Each dataset was aligned
478 to the hg19 reference genome using bowtie2 [38]. After alignment, each group of
479 replicates (either 50,000 cells or 500 cells) were merged together, sorted and indexed.
480 Reads that had a mapping quality score below 30 or that were considered duplicates (exact

481 same start and stop position) were removed from the merged files. It should be noted that
482 HMMRATAC will remove duplicate and low mapping quality reads by default, although
483 some other algorithms do not.
484
485 The merged, filtered and sorted BAM files, created as described above, were the input for
486 MACS2. HMMRATAC took the sorted paired-end BAM file and its corresponding BAM
487 index file as the main inputs. In addition, HMMRATAC requires a genome-wide read
488 coverage file in BigWig format. To generate the coverage file, MACS2 was run with the
489 option set for reporting the read coverage in BedGraph format (option –B) and this
490 BedGraph was then converted to a BigWig file using UCSC's bedGraphToBigWig
491 program. F-Seq requires a single-end BED file of read alignment results as its input. To
492 generate this BED file, we converted the paired-end BAM file into a BED file, using an in-
493 house script, and split each read pair into forward and reverse strand reads.
494
495 The human monocyte data from [14] was publicly available and downloaded from the
496 Cistrome Database [7] as aligned BAM files to the hg19 reference genome. The data
497 corresponds to Gene Expression Omnibus accessions GSM2325680, GSM2325681,
498 GSM2325686, GSM2325689, and GSM2325690. One replicate per condition was used
499 and processed in the same way as described above, filtering.
500
501 **The HMMRATAC algorithm**
502
503 ***Segregation of ATAC-Seq signals***

504

505 After the preprocessing step that eliminates duplicate reads and low mapping quality reads,
506 the main HMMRATAC pipeline begins by separating the ATAC-seq signal into four
507 components, each representing a unique feature. These features are nucleosome-free
508 regions (NFRs), which are most likely to occur within the open chromatin region itself,
509 and three nucleosomal features, representing mono-, di- and tri-nucleosomes (1Ns, 2Ns
510 and 3Ns), respectively. We utilize four distributions to represent the four signal tracks: an
511 exponential distribution for the nucleosome-free track and three Gaussian distributions to
512 represent the three nucleosomal signal tracks. These distributions were utilized by
513 Buenrostro et al. to identify their cutoffs for fragment separations [5]. Although sequence
514 read data is discrete, the fact that the expectation value of each Gaussian is large (>100bp)
515 and the number of observations is high, the discrete distribution, such as a Poison
516 distribution, can be approximated as a Gaussian. The mean value of the exponential
517 distribution for NFRs is set as a fixed value at runtime, either as the default value or a user-
518 defined value. The parameters of three Gaussian distributions are initialized at runtime
519 with default or user-defined values and then updated using the expectation maximization
520 (EM) algorithm for Gaussian mixture models. These initial values were the same as those
521 used by Buenrsotro et al. to define their distributions for the chromatin features [5]. Briefly,
522 each nucleosomal fragment, those larger than 100bp, is classified as belonging to one of
523 the nucleosomal distributions based on its weighted probability. Once each fragment has
524 been classified, new means, standard deviations, and weights are calculated for each
525 distribution and become the new values for the mixture model. This process continues
526 iteratively until the change in the mean values between iterations is less than a reasonable

527 epsilon value. At this point, the model is assumed to have reached convergence and the
528 EM process is halted (Fig. 1c). In order to decrease the time required for the EM algorithm,
529 HMMRATAC randomly sub-selects 10% of all the fragments to use as the training data.
530 Once the parameters of the four distributions have been determined, HMMRATAC creates
531 four genome-wide signal tracks. For every genomic position, all of the fragments that
532 occupy that position are determined. Each track is then incremented by the probability that
533 a particular fragment belongs to the tracks' corresponding distribution. The signal tracks
534 are then transformed by square root into continuous space.
535
536 ***Training the Hidden Markov Model of accessible chromosomal regions, and refining***
537 ***the model with the Baum-Welch algorithm***
538
539 Once the four signal tracks have been created, the main HMMRATAC process begins.
540 HMMRATAC first identifies up to 1000 training regions throughout the genome, with
541 which it learns the model. These training regions are determined by scanning the genome
542 for regions whose read coverage fold changes over background falls within a certain range.
543 The default lower and upper limit of fold changes, as used in the Result section, is set
544 between 10 and 20. These regions are then extended by 5KB in either direction. It is also
545 possible for users to provide a BED file containing predefined training regions. The initial
546 HMM has proportional initial probabilities, which are not updated during training, as well
547 as proportional transition probabilities, which are updated during training. The emission
548 probabilities are calculated with k-means clustering [39], by separating the training regions
549 into 3 clusters and calculating the means and covariance matrices for the four signal tracks

for each cluster. We model the emissions as a multivariate Gaussian distribution since this assumption has been successfully adopted in other HMM-based genome segmentation methods such as hiHMM [40, 41] and Segway [42, 43]. The three clusters represent the three states of our model, corresponding to the *center* of the open region, the *nucleosomes,* and a *background* state, respectively. It is also possible to create a model using a different number of states instead of 3, although this option is not generally recommended. Once this initial HMM is created, the transition and emission parameters are updated with the Baum-Welch algorithm [16] until convergence.

Annotating with the Viterbi algorithm

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

open state. We generally consider the nearby nucleosomes to be part of the regulatory region (Fig. 3a and b), and therefore the region is then reported as a peak in gappedPeak format.

It should be noted that at all stages of the algorithm, 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 system 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. Finally, in terms of runtime, HMMRATAC took approx. 4.9 hours to process the GM12878 data set and an average of 5.6 hours to process the monocyte data in our server equipped with AMD Opteron(tm) Processor 6378 and 500G memory.

Post-processing and output

When the report peak option is set for 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.

596 It is also possible to report the average read coverage, the median coverage, and the fold
597 change above the genomic background, the zscore of the average coverage or all values.
598 Additionally, those high coverage regions that are masked from Viterbi decoding (see
599 above), are added back to the final peak file. They are annotated by HMMRATAC as
600 "High Coverage Accessible regions". The user may then keep or discard these regions and
601 explore them as they see fit.
602
603 Additionally, when the report peak option is set, HMMRATAC also supplies an additional
604 file containing the summits of each peak. To calculate a summit, HMMRATAC first
605 smoothes the read coverage signal through the *center* state using a Gaussian smoothing
606 function with a window size of 120bp. It then calculates the position that has the maximum
607 read coverage and reports that position as the summit.
608
609 **Evaluation**
610
611 ***Calculating Precision, Recall, False Positive Rate and F1-scores***
612
613 Let Predicted Positive (PP) represent the number of base pairs identified as accessible
614 regions from any method. Let Real Positives (RP) represent the number of base pairs in
615 the "gold standard" sets (described below). Let Real Negatives (RN) represent the number
616 of base pairs in the "real negative" set described below. Let True Positive (TP) represent
617 the number of overlapping base pairs between a called peak and a real positive. Let False

618 Positive (FP) represent the number of overlapping base pairs between a called peak and a
619 real negative.
620
621 Then *precision* or positive predictive value (PPV) is calculated as: PPV = TP/(TP+FP)
622
623 *Recall*, or true positive rate (TPR), is calculated as: TPR = TP/RP
624
625 *False positive rate* (FPR) is calculated as: FPR = FP/RN
626
627 The number of overlapping base pairs was calculated with the BEDTools intersect tool
628 [45] and custom python scripts.
629
630 ***Compiling the "Gold Standard" lists and "True Negative" set***
631
632 We tested the performance of HMMRATAC and other methods against three "gold
633 standard" datasets. The first one, which we call "Active Regions", is made by combining
634 two chromatin state annotations generated in GM12878 cells using chromHMM [28, 29]
635 (wgEncodeBroadHmmGm12878HMM). These two annotations are "Active Promoters"
636 and "Strong Enhancers".
637
638 The other two "gold standard" test sets that were used in this study were DNaseI
639 hypersensitive sites (wgEncodeAwgDnaseUwdukeGm12878UniPk) and FAIRE
640 accessible regions (wgEncodeOpenChromFaireGm12878Pk) [11, 30]. We found that there

641 were numerous cases throughout the genome where clusters of DNase or FAIRE accessible
642 regions were in close proximity to one another. These clusters tended to have broad
643 enrichment in either DNase-seq or FAIRE-seq signal across the regions. Therefore, we
644 merged all accessible regions that were within 2KB of each other into a single peak and
645 used these merged sets as the "gold standard" sets.
646
647 The "real negative" set was made with the "Heterochromatin" annotation from
648 chromHMM. When using the "active region" set, no modifications were necessary
649 because the three states are mutually exclusive. When using either the DNase or FAIRE
650 sets, any "heterochromatin" states that overlapped (by any amount) a DNase or FAIRE
651 peak were excluded from the final "real negative" set. Additionally, any DNase or FAIRE
652 peak that overlapped any "heterochromatin" state (by any amount) was, likewise, excluded
653 from its respective "gold standard" set.
654
655 The "gold standard" set used for the monocyte data were compiled as follows. For each
656 condition, 2 replicates of the histone marks H3K27ac, H3K4me1, and H3K4me3 from the
657 same study [14] were downloaded from the Cistrome Database [7] as aligned BAM files.
658 Each file was processed by MACS2 with the p-value parameter (-p option) set to 0.15. The
659 resulting peaks were sorted by the –log10(FDR) value and the top 10000 peaks for each
660 sample were pooled together and merged. The "real negative" set for the monocyte data
661 were H3K9me3 peaks from the same study. One replicate of H3K9me3 histone mark ChIP-
662 seq was downloaded from the Cistrome Database, for each condition, as an aligned BAM
663 file. The file was then processed by MACS2 with the p-value parameter set to 0.15. The

664 peaks were sorted by their –log10(FDR) value and the top 10000 peaks were kept. At this
665 point, the "Active Regions" true set was created by eliminating any merged active peak,
666 created as described previously, that overlapped an H3K9me3 peak. This list of "Active
667 Regions" became the "gold standard" set for each monocyte condition. Any H3K9me3
668 peak that did not overlap the "Active Region" set, was kept as the "real negative" set for
669 that condition. The accession numbers for all above histone modification ChIP-seq data
670 are listed in Additional file 2.
671
672 ***Calls from MACS2***
673
674 Version 2.1 of MACS was used to call peaks using the sorted paired-end BAM files created
675 as described above. With each run the input file format (option -f) was set to "BAMPE"
676 and the option for keeping duplicate reads (option --keep-dup) was set to "all", as duplicate
677 reads had already been discarded in pre-processing. MACS2 was run with the local lambda
678 option turned off (option --nolambda) and was run with q-value cutoffs (option -q) set to
679 0.5, 0.1, 0.05 (default), 0.005, 0.0005 and 0.00005. It was also run with a p-value cutoff
680 (option -p) set to 0.6 and 0.3.
681
682 ***Calls from F-Seq***
683
684 Version 1.84 of F-Seq was used to call peaks using the single-end BED files created as
685 described previously. With each run, the output was reported as a BED file (option -of).

686 The standard deviation cutoff (option -t) was set to 0.1, 1, 4 (default), 6, 7, 8, 10, 15 and
687 20.
688
689 ***Calls from HMMRATAC***
690 HMMRATAC was run using the sorted BAM files, BAM index files and genome-wide
691 coverage files created as described previously. The upper limit fold-change for choosing
692 training regions (option -u) was set to 20 with corresponding lower limits (option -l) set to
693 10. These settings are the HMMRATAC defaults. HMMRATAC was also run using
694 blacklisted sites for hg19 [11, 15], which were masked from the program and all subsequent
695 steps. For the precision, recall and false positive rate calculations, the HMMRATAC peaks
696 were filtered by minimum score cutoffs of 0, 5, 10, 15, 20, 25, 30, 35, 40 and 45.
697 Additionally, for two monocyte samples (accession numbers GSM2325689 – RPMI_4h
698 and GSM2325681 – BG_d1) HMMRATAC did not produce a typical model with the
699 default settings (-u 20 –l 10). Instead, the models produced showed the nucleosome state
700 as having the highest NFR signal. We re-ran these samples with more conservative settings
701 (-u 40 –l 20), at which time HMMRATAC produced models that showed the characteristic
702 pattern that we have seen with the other samples. These peak results from HMMRATAC
703 were then filtered by the same cutoffs as the other samples and tested in the same way. All
704 the transition and emission parameters learned by HMMRATAC from GM12878 50k cells,
705 500 cells, five human monocyte samples are summarized in Additional file 1: Table S3.
706
707 ***Calculating reproducibility***
708

709 Reproducibility was determined using the Irreproducible Discovery Rate (IDR) originally
710 described by Li et al. [33]. Peaks were called using the following parameters: the p-value
711 for MACS2 was set to 0.6 for 50,000 cell per-replicate *merged* data and 0.6 for 500 cell
712 per-replicate *merged* data, the standard deviation for F-Seq was set to 0.1 for both *merged*
713 datasets, and the fold-change range for HMMRATAC was set to 10-20 (default) for both
714 *merged* datasets. The parameters for each individual 50,000 cell replicate was the same for
715 F-seq and HMMRATAC and the MACS2 q-value setting was set to 0.5. These parameters
716 were also the most permissive settings used throughout this paper, including in our PRC
717 and ROC curves. The MACS2 setting was changed for the individual replicates because
718 at the permissive settings (-p 0.6 or –p 0.3) it called the entire genome as a peak, so we
719 chose the third most permissive setting instead. The peaks, called by each method, were
720 then sorted by their respective scores (-log(p-value) for F-seq, the signal value for MACS2
721 and HMMRATAC) and the top 100k peaks were retained. These top peaks were the inputs
722 into the IDR software.
723
724 For calculating the 500 to 50,000 cell per-replicate reproducibility, each method was run
725 using the merged file(s) described previously, with the respective parameters as described
726 previously. For each method, each pair of resulting peak files, from the 500 cell per-
727 replicate and 50,000 cell per-replicate datasets, were run through the IDR algorithm using
728 default parameters and each methods' respective score column as the rank. Reproducible
729 peaks between datasets were considered to have an IDR value equal to or below 0.05.
730

731 For replicate testing, each method was run with the same parameters as described
732 previously, using the individual replicate file(s) as their input. Each possible pair of
733 replicates was then run through the IDR algorithm with default parameters. For each
734 replicate pair, the number of peaks whose local IDR score was equal to or below 0.05 was
735 calculated. Peaks with IDR scores equal to or below 0.05 in each of the pairwise
736 comparisons were retained and used to calculate the overlap between the three methods
737 (Additional file 1: Fig. S8).
738

739 **Abbreviations**

740

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

754

755 ## Declarations

756

757 **Funding**

758

759 This work was supported by the startup fund for TL from Jacobs School of Medicine and

760 Biomedical Sciences, University at Buffalo.

761

762 **Availability of data and materials**

763

764 The HMMRATAC source code, executable file, quick start guide and javadocs can be

765 downloaded from https://github.com/LiuLabUB/HMMRATAC. All peaks/regions called

766 by HMMRATAC, F-Seq, and MACS2, as well as the various "gold standard" files, can be

767 downloaded from the following URL: http://biomisc.org/download/HumanData.tar.gz.

768

769 **Authors' contributions**

770

771 EDT and TL conceived and designed the computational method. EDT developed the

772 software package, implemented the method and analyzed the data. EDT and TL wrote the

773 paper. All authors read and approved the final manuscript.

774

775 **Competing interests**

776

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Ethics approval is not applicable for this study.

798 **Tables**

799

800 **Table 1. The Area Under the Curves (AUCs) of method comparisons of**

801 **HMMRATAC, MACS2, and F-Seq.**

802

803 The AUC calculated from the ROC and PRC curves in Fig. 4 and Additional file 1: Fig S7.

804 Because we can't reach the extreme sides of the curves while tuning cutoff values, before

805 we calculated AUC values, we added theoretical extreme points to ROC – (0, 0) and (1,

806 1), and to PRC – (0, 1) and (1, L_{true}/(L_{true} + L_{false})). L_{true} is the total length of real positive

807 set, and the L_{false} is the total length of real negative set. The numbers of the best

808 performance are shown in bold.

809

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

810

811

812

813

814 **Figure Legends**

815

816 **Fig. 1 The HMMRATAC Algorithm.**

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

824

825 **Fig. 2 Chromatin Architecture at HMMRATAC Peaks.**

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

837

838 **Fig. 3 Chromatin features of HMMRATAC states.**

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

848

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

850 Left: Receiver Operator Characteristic (ROC) curves and right: Precision-Recall (PRC)
851 curves for the real positive and real negative pairs of 1. active chromatin states vs.
852 heterochromatin for GM12878 cells, 2. active histone marks (either H3K4me1, H3K4me3
853 or H3K27ac) vs. heterochromatin (H3K9me3) for human monocytes (see method for
854 detail).

855

856

857

858

859

860 **References**

861

862 1. Li G, Reinberg D: **Chromatin higher-order structures and gene regulation.**
863 *Curr Opin Genet Dev* 2011, **21:**175-186.
864 2. Tsompana M, Buck MJ: **Chromatin accessibility: a window into the genome.**
865 *Epigenetics Chromatin* 2014, **7:**33.
866 3. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS,
867 Crawford GE: **High-resolution mapping and characterization of open**
868 **chromatin across the genome.** *Cell* 2008, **132:**311-322.
869 4. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD: **FAIRE (Formaldehyde-**
870 **Assisted Isolation of Regulatory Elements) isolates active regulatory elements**
871 **from human chromatin.** *Genome Res* 2007, **17:**877-885.
872 5. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of**
873 **native chromatin for fast and sensitive epigenomic profiling of open**
874 **chromatin, DNA-binding proteins and nucleosome position.** *Nat Methods*
875 2013, **10:**1213-1218.
876 6. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP,
877 Chang HY, Greenleaf WJ: **Single-cell chromatin accessibility reveals**
878 **principles of regulatory variation.** *Nature* 2015, **523:**486-490.
879 7. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, Zhu M, Wu J, Shi X, Taing L, et
880 al: **Cistrome Data Browser: a data portal for ChIP-Seq and chromatin**
881 **accessibility data in human and mouse.** *Nucleic Acids Res* 2017, **45:**D658-
882 D662.

883 8. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev
884 I, Zhao K: **High-resolution profiling of histone methylations in the human**
885 **genome.** *Cell* 2007, **129:**823-837.
886 9. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum
887 C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq**
888 **(MACS).** *Genome Biol* 2008, **9:**R137.
889 10. Boyle AP, Guinney J, Crawford GE, Furey TS: **F-Seq: a feature density**
890 **estimator for high-throughput sequence tags.** *Bioinformatics* 2008, **24:**2537-
891 2538.
892 11. Consortium EP: **An integrated encyclopedia of DNA elements in the human**
893 **genome.** *Nature* 2012, **489:**57-74.
894 12. Lantermann AB, Straub T, Stralfors A, Yuan GC, Ekwall K, Korber P:
895 **Schizosaccharomyces pombe genome-wide nucleosome mapping reveals**
896 **positioning mechanisms distinct from those of Saccharomyces cerevisiae.** *Nat*
897 *Struct Mol Biol* 2010, **17:**251-257.
898 13. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A: **Determinants**
899 **of nucleosome organization in primary human cells.** *Nature* 2011, **474:**516-
900 520.
901 14. Novakovic B, Habibi E, Wang SY, Arts RJW, Davar R, Megchelenbrink W, Kim
902 B, Kuznetsova T, Kox M, Zwaag J, et al: **beta-Glucan Reverses the Epigenetic**
903 **State of LPS-Induced Immunological Tolerance.** *Cell* 2016, **167:**1354-1368
904 e1314.

905 15. **Human ENCODE Data Analysis Consortium Blacklisted Regions**
906 [http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapabil
907 ity/wgEncodeDacMapabilityConsensusExcludable.bed.gz]
908 16. Baum LE, Petrie T, Soules G, Weiss N: **A Maximization Technique Occurring**
909 **in the Statistical Analysis of Probabilistic Functions of Markov Chains.** *The*
910 *Annals of Mathematical Statistics* 1970, **41:**164-171.
911 17. Viterbi A: **Error bounds for convolutional codes and an asymptotically**
912 **optimum decoding algorithm.** *IEEE Trans Inf Theor* 1967, **13:**260-269.
913 18. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S: **Eukaryotic regulatory element**
914 **conservation analysis and identification using comparative genomics.** *Genome*
915 *Res* 2004, **14:**451-458.
916 19. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K,
917 Clawson H, Spieth J, Hillier LW, Richards S, et al: **Evolutionarily conserved**
918 **elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005,
919 **15:**1034-1050.
920 20. Shin H, Liu T, Duan X, Zhang Y, Liu XS: **Computational methodology for**
921 **ChIP-seq analysis.** *Quant Biol* 2013, **1:**54-70.
922 21. Koohy H, Down TA, Spivakov M, Hubbard T: **A Comparison of Peak Callers**
923 **Used for DNase-Seq Data.** *PLoS One* 2014, **9:**e96303.
924 22. Hansel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di
925 Antonio M, Pike J, Kimura H, Narita M, et al: **G-quadruplex structures mark**
926 **human regulatory chromatin.** *Nat Genet* 2016, **48:**1267-1272.

927 23. Philip M, Fairchild L, Sun L, Horste EL, Camara S, Shakiba M, Scott AC, Viale
928 A, Lauer P, Merghoub T, et al: **Chromatin states define tumour-specific T cell**
929 **dysfunction and reprogramming.** *Nature* 2017, **545:**452-456.
930 24. Denny SK, Yang D, Chuang CH, Brady JJ, Lim JS, Gruner BM, Chiou SH, Schep
931 AN, Baral J, Hamard C, et al: **Nfib Promotes Metastasis through a Widespread**
932 **Increase in Chromatin Accessibility.** *Cell* 2016, **166:**328-342.
933 25. Chen X, Shen Y, Draper W, Buenrostro JD, Litzenburger U, Cho SW, Satpathy
934 AT, Carter AC, Ghosh RP, East-Seletsky A, et al: **ATAC-see reveals the**
935 **accessible genome by transposase-mediated imaging and sequencing.** *Nat*
936 *Methods* 2016, **13:**1013-1020.
937 26. Risca VI, Denny SK, Straight AF, Greenleaf WJ: **Variable chromatin structure**
938 **revealed by in situ spatially correlated DNA cleavage mapping.** *Nature* 2017,
939 **541:**237-241.
940 27. **ENCODE ATAC-Seq analysis pipeline** [https://www.encodeproject.org/atac-
941 seq/]
942 28. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and**
943 **characterization.** *Nat Methods* 2012, **9:**215-216.
944 29. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang
945 X, Wang L, Issner R, Coyne M, et al: **Mapping and analysis of chromatin state**
946 **dynamics in nine human cell types.** *Nature* 2011, **473:**43-49.
947 30. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield
948 NC, Stergachis AB, Wang H, Vernot B, et al: **The accessible chromatin**
949 **landscape of the human genome.** *Nature* 2012, **489:**75-82.

950 31. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL,
951 Stamatoyannopoulos JA: **Chromatin accessibility pre-determines**
952 **glucocorticoid receptor binding patterns.** *Nat Genet* 2011, **43:**264-268.
953 32. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S,
954 Bernstein BE, Bickel P, Brown JB, Cayting P, et al: **ChIP-seq guidelines and**
955 **practices of the ENCODE and modENCODE consortia.** *Genome Res* 2012,
956 **22:**1813-1831.
957 33. Li QH, Brown JB, Huang HY, Bickel PJ: **Measuring Reproducibility of High-**
958 **Throughput Experiments.** *Annals of Applied Statistics* 2011, **5:**1752-1779.
959 34. He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H,
960 Long H, et al: **Refined DNase-seq protocol and data analysis reveals intrinsic**
961 **bias in transcription factor footprint identification.** *Nat Methods* 2014, **11:**73-
962 78.
963 35. Wang JR, Quach B, Furey TS: **Correcting nucleotide-specific biases in high-**
964 **throughput sequencing data.** *BMC Bioinformatics* 2017, **18:**357.
965 36. Martins AL, Walavalkar NM, Anderson WD, Zang C, Guertin MJ: **Universal**
966 **correction of enzymatic sequence bias reveals molecular signatures of**
967 **protein/DNA interactions.** *Nucleic Acids Res* 2018, **46:**e9.
968 37. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence
969 Database C: **The Sequence Read Archive: explosive growth of sequencing**
970 **data.** *Nucleic Acids Res* 2012, **40:**D54-56.
971 38. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat*
972 *Methods* 2012, **9:**357-359.

973 39. MacQueen J: **Some methods for classification and analysis of multivariate**
974 **observations.** In *Proceedings of the Fifth Berkeley Symposium on Mathematical*
975 *Statistics and Probability, Volume 1: Statistics*; *1967; Berkeley, Calif.* University
976 of California Press; 1967: 281-297.
977 40. Sohn KA, Ho JW, Djordjevic D, Jeong HH, Park PJ, Kim JH: **hiHMM: Bayesian**
978 **non-parametric joint inference of chromatin state maps.** *Bioinformatics* 2015,
979 **31:**2066-2074.
980 41. Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A,
981 Tolstorukov MY, Appert A, et al: **Comparative analysis of metazoan**
982 **chromatin organization.** *Nature* 2014, **512:**449-452.
983 42. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS:
984 **Unsupervised pattern discovery in human chromatin structure through**
985 **genomic segmentation.** *Nat Methods* 2012, **9:**473-476.
986 43. Chan RCW, Libbrecht MW, Roberts EG, Bilmes JA, Noble WS, Hoffman MM:
987 **Segway 2.0: Gaussian mixture models and minibatch training.** *Bioinformatics*
988 2018, **34:**669-671.
989 44. **Human ENCODE Data Analysis Consortium Blacklisted Regions**
990 [http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapabil
991 ity/wgEncodeDacMapabilityConsensusExcludable.bed.gz]
992 45. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing**
993 **genomic features.** *Bioinformatics* 2010, **26:**841-842.
994

Fig. 1

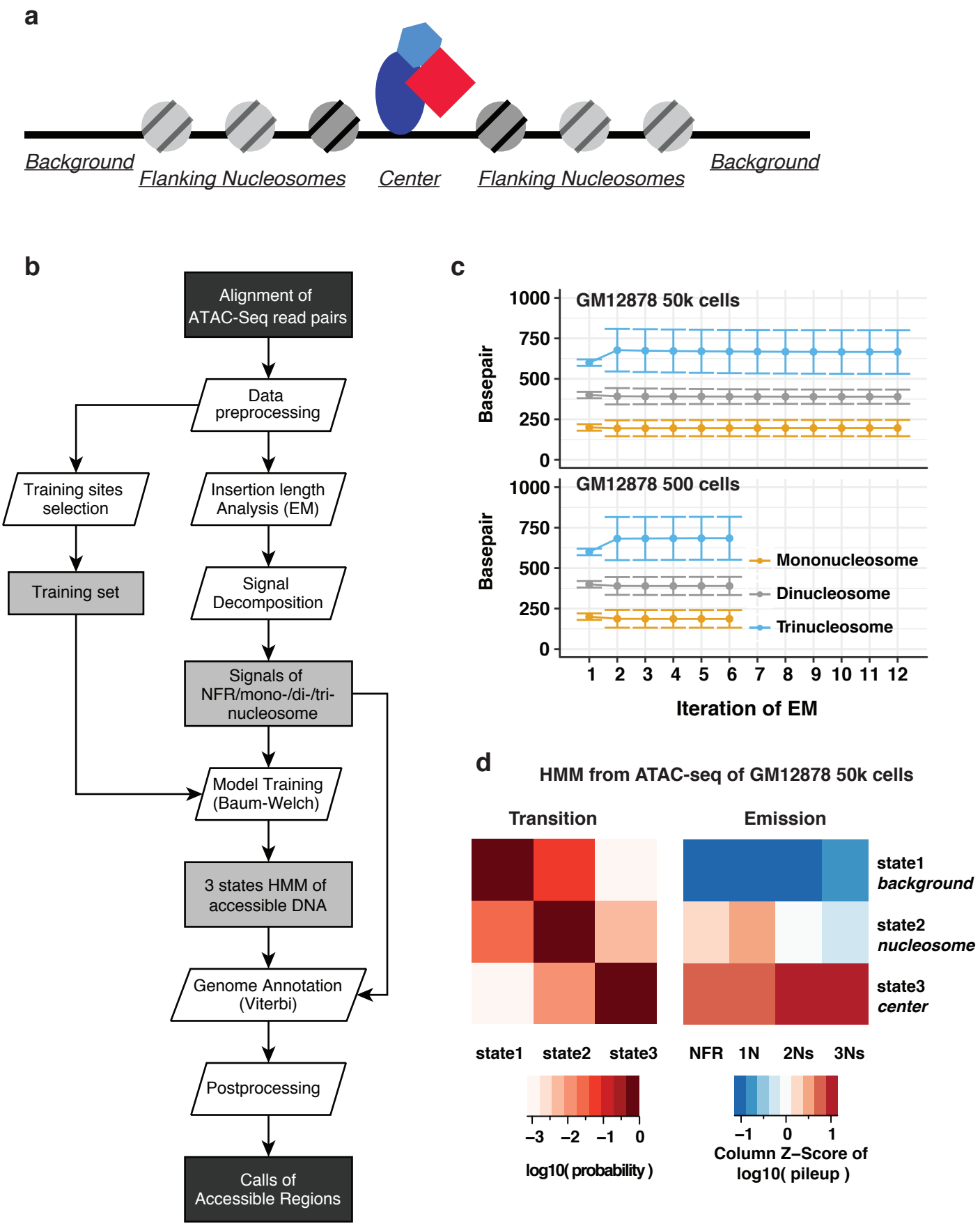

a
Background
Flanking Nucleosomes
Center
Flanking Nucleosomes
Background
b
Alignment of ATAC-Seq read pairs
Data preprocessing
Training sites selection
Insertion length Analysis (EM)
Training set
Signal Decomposition
Signals of NFR/mono-/di-/tri-nucleosome
Model Training (Baum-Welch)
3 states HMM of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

Fig. 2

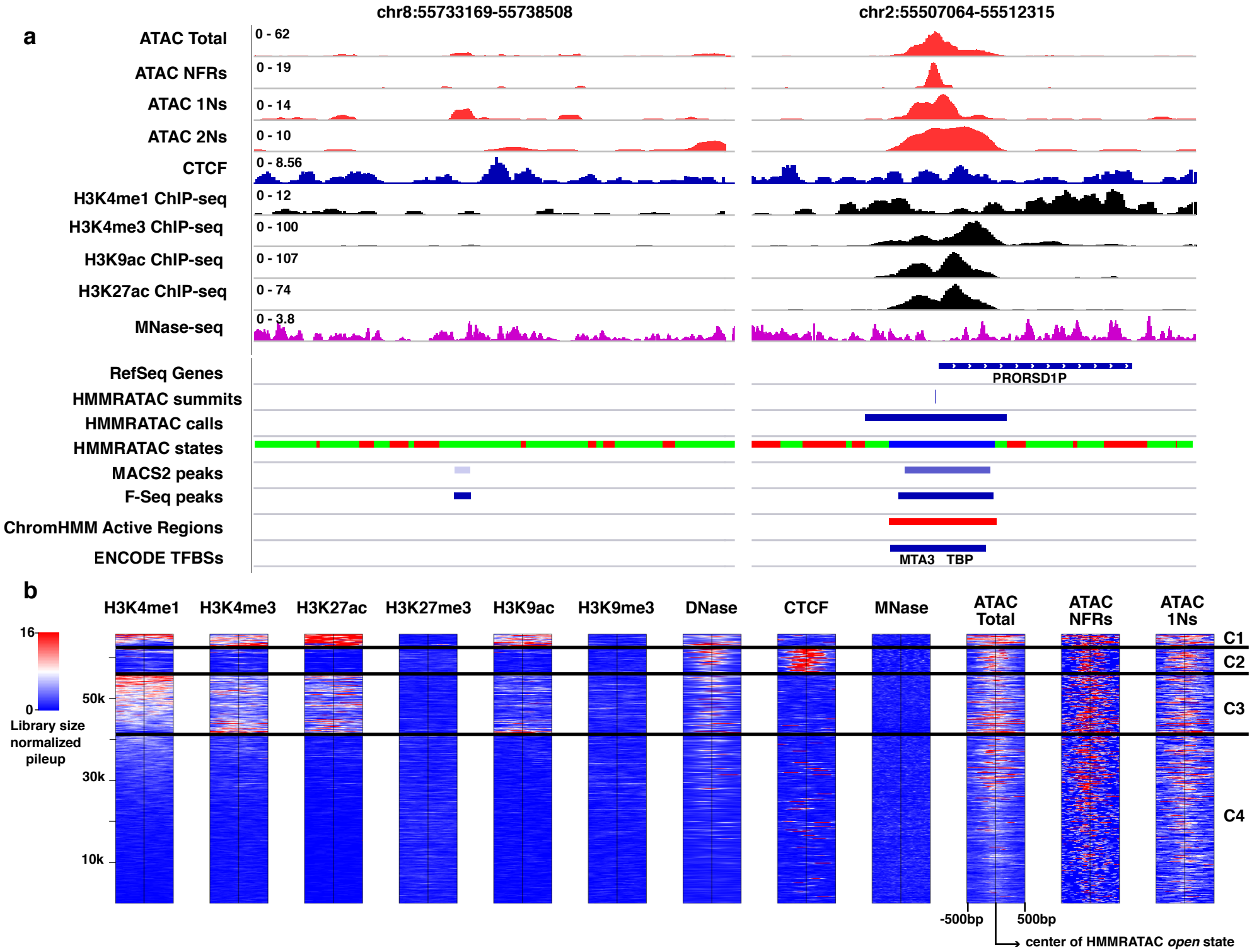

a
chr8:55733169-55738508
chr2:55507064-55512315
ATAC Total
0 - 62
ATAC NFRs
0 - 19
ATAC 1Ns
0 - 14
ATAC 2Ns
0 - 10
CTCF
0 - 8.56
H3K4me1 ChIP-seq
0 - 12
H3K4me3 ChIP-seq
0 - 100
H3K9ac ChIP-seq
0 - 107
H3K27ac ChIP-seq
0 - 74
MNase-seq
0 - 3.8
RefSeq Genes
PRORSD1P
HMMRATAC summits
HMMRATAC calls
HMMRATAC states
MACS2 peaks
F-Seq peaks
ChromHMM Active Regions
ENCODE TFBSs
MTA3
TBP
b
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9ac
H3K9me3
DNase
CTCF
MNase
ATAC Total
ATAC NFRs
ATAC 1Ns
16
0
Library size normalized pileup
50k
30k
10k
C1
C2
C3
C4
-500bp
500bp
center of HMMRATAC open state

Fig. 3

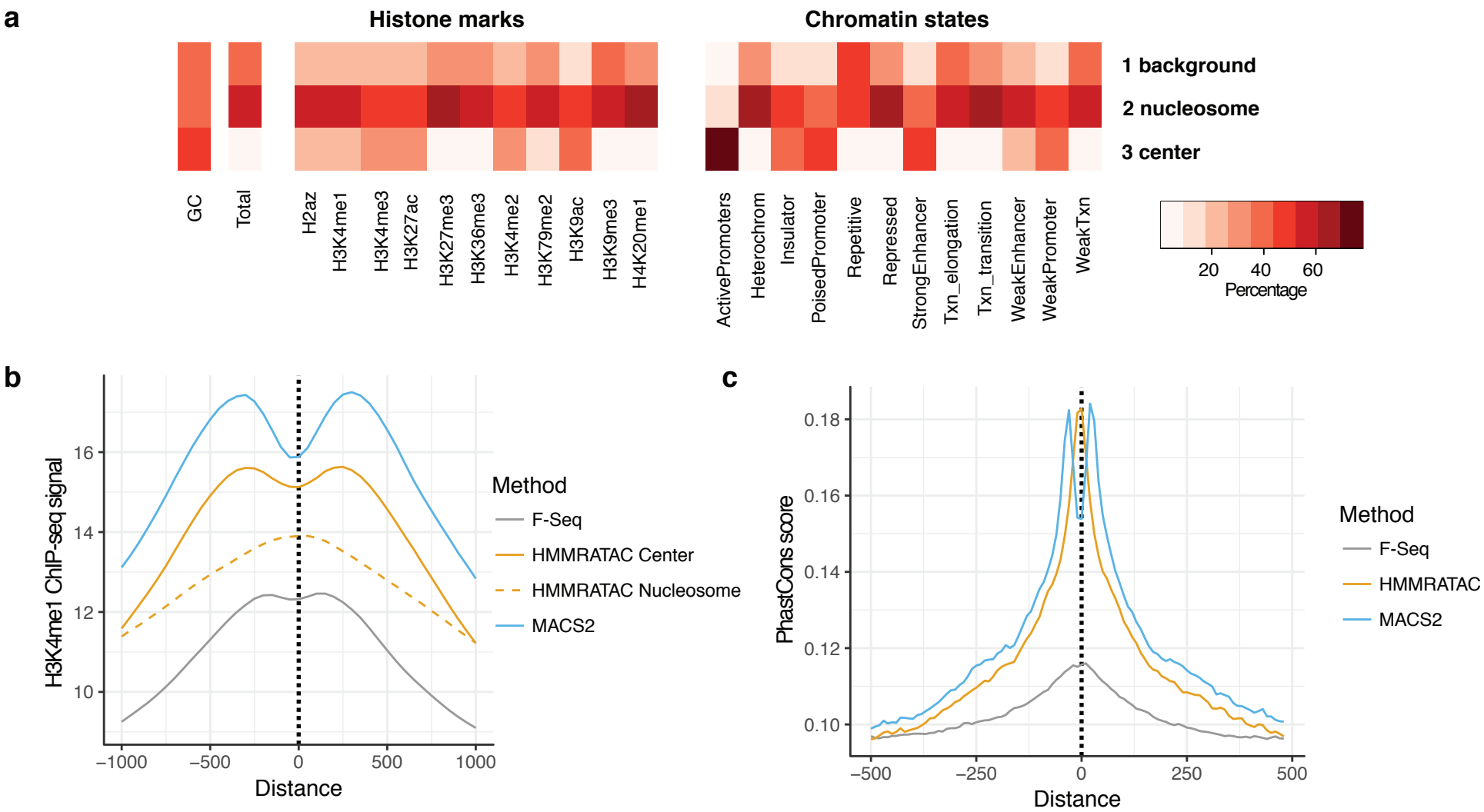
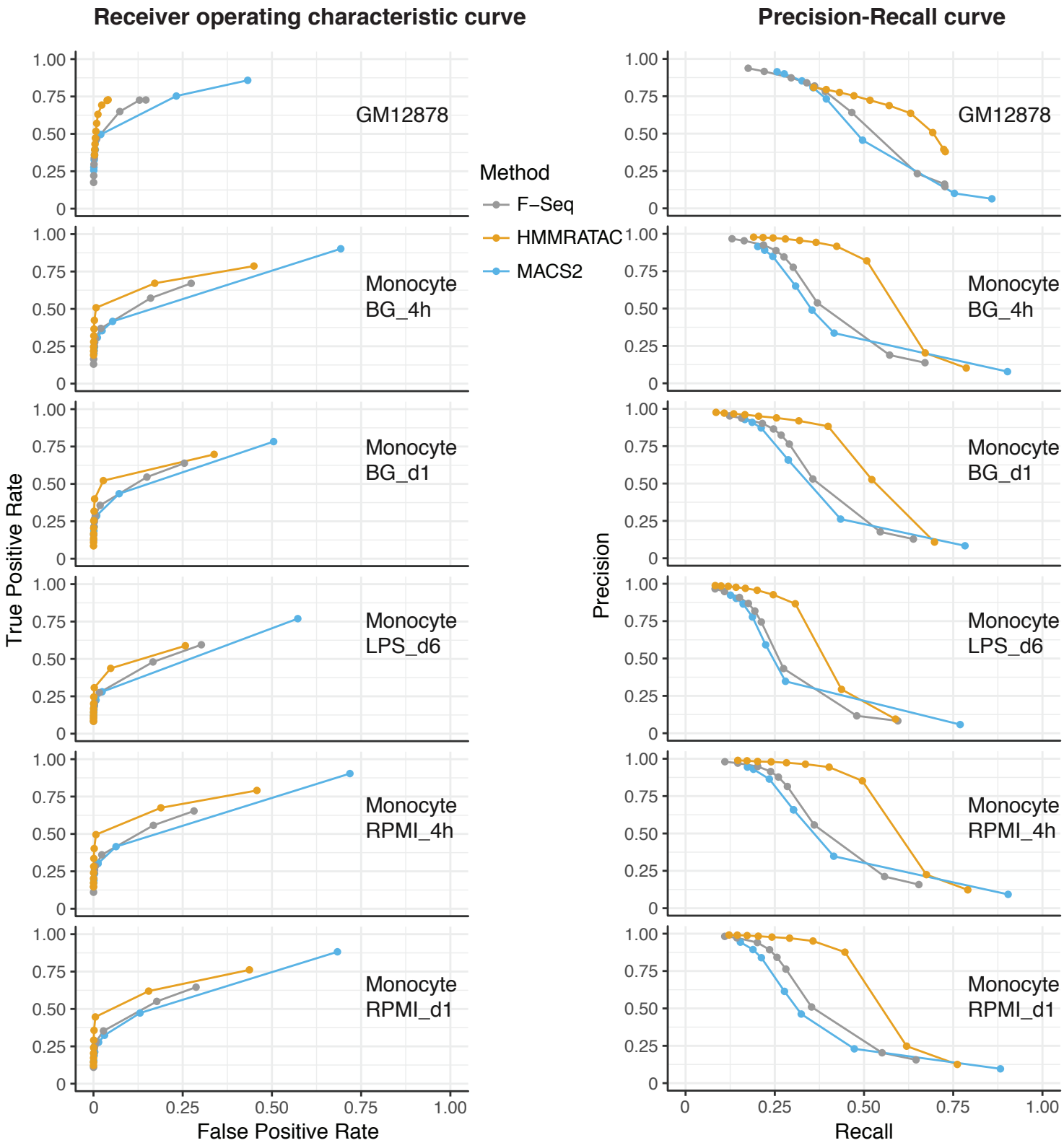

a
Histone marks
Chromatin states
1 background
2 nucleosome
3 center
GC
Total
H2az
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K36me3
H3K4me2
H3K79me2
H3K9ac
H3K9me3
H4K20me1
ActivePromoters
Heterochrom
Insulator
PoisedPromoter
Repetitive
Repressed
StrongEnhancer
Txn_elongation
Txn_transition
WeakEnhancer
WeakPromoter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

Fig. 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