Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability Corrected Multiscale Signal Processing Framework

Arif Harmanci[1,2], Joel Rozowsky[1,2], Mark Gerstein[1,2,3,*]

1 Program in Computational Biology and Bioinformatics, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
2 Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
3 Department of Computer Science, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
*Corresponding authors: Mark Gerstein pi@gersteinlab.org

ABSTRACT

We present MUSIC, a method for the identification of enriched regions (ERs) in the genome-wide read depth (RD) signal profiles from ChIP-seq experiments. The basic motivation behind MUSIC is twofold: First, the systematic noise introduced by non-uniform read mappability causes fragmentation of ERs, especially for the diffuse binding proteins and histone modifications. Second, ChIP-Seq assays have a large spectrum of ER lengths, e.g. H3K36me3 marks the active gene bodies whose lengths range from kilo to mega base pairs, which makes it necessary to analyze the signal at multiple length scales. MUSIC first applies a correction filter to mediate the effects of non-uniform read mappability while preserving the signal in enriched regions. MUSIC then performs median filtering based multiscale decomposition to the corrected signal. Our multiscale approach is adapted from scale-space analysis in signal processing literature for feature identification in discrete signals. At each scale, MUSIC identifies the scale specific significant ERs then merges these to generate the final set of ERs. When compared with other ER identification methods, MUSIC performs favorably in terms of accuracy and reproducibility of the ERs. Analysis of RNA Polymerase II binding ChIP-Seq data using the scale specific ERs reveals that there is a clear distinction between the expression levels of genes with punctate bound (stalled) and broadly bound (elongating) forms of polymerase. MUSIC is a useful tool that enables multiscale identification and analysis of ERs in ChIP-Seq datasets and is available for download at: http://music.gersteinlab.org

SHORT ABSTRACT

We present MUSIC, a signal processing methodology for identification of enriched regions in ChIP-Seq data. MUSIC first applies a correction filter to the read depth signal for removing the systematic noise introduced by non-uniform mappability then performs a multiscale decomposition on the corrected signal, which is then used to identify ERs at multiple length scales. When compared with other ER identification methods, MUSIC performs favorably in terms of accuracy and reproducibility of the ERs.

Analysis of the RNA Polymerase II ChIP-Seq data reveals the clear distinction in the stalled and elongating forms of polymerase. MUSIC is available online at http://music.gersteinlab.org

Keywords:

ChIP-Seq, Enriched Regions, Non-linear Scale Space Analysis, Mappability correction, RNA Polymerase,

1 BACKGROUND

With the recent advancements in sequencing technologies, chromatin immuniprecipitation based enrichment of the DNA sequences followed by sequencing (ChIP-seq) [1, 2] has become the mainstream experimental method for genome-wide measurement of the locations DNA binding proteins (e.g. transcription factors) and posttranslational modifications of the histone proteins, or histone modifications [3, 4] (HMs). Consortium projects such as ENCODE [5] and Roadmap Epigenomics Project [6] generated ChIP-Seq datasets to map the chromatin states of many cell lines and tissues [7]. These substantially increased the number of publicly available ChIP-Seq datasets for diverse set of histone modification and transcription factor binding profiles. Following the sequencing, it is necessary to computationally process the read depth (RD) signal profile to identify the enriched regions (ERs) across genome [8].

Depending on the target of the ChIP-seq assay, the length scale of ERs can vary extensively for different experiments, which changes the ER identification workflow. For example, for transcription factor binding, the ERs are observed at punctate regions of protein binding of length hundreds of nucleotides [9]. For most HMs, ERs are broad. For example, the ERs for repressive heterochromatin mark H3K9me3 can extend upto a few megabases. Another interesting example is the RNA Polymerase II, which binds to the promoters and gene bodies for the purpose of mRNA transcription whose ERs can extend over the whole gene bodies or can be punctate and concentrated close to gene promoters. Development of efficient computational methods for identification and characterization of the broad ERs is necessary for understanding the regulatory effects of the HMs and diffuse DNA binding proteins on gene expression as more evidence is brought to light that these epigenetic factors are major driving factors in pluripotency [10], and for disease manifestation like cancerogenesis [11–15].

There are two main challenges for identification of broad ERs. First, unlike transcription factor binding, broad ERs are observed at longer length scales and the length spectrum of ERs are broad for many HMs. This makes it necessary to identify the ERs at different scales. A widely used method for identifying the ERs in HM signal profiles is smoothing the signal profile with a kernel of constant size and shape and using a null model (e.g., Poisson or negative binomial) to identify the significantly enriched regions. It is, however, not clear how the kernel size and shape should be selected**.** The multiscale approaches proposed by the wavelet based methods address this issue. Yet, the reasoning and motivation for which wavelet functions are used in these methods are generally not well established.

Second, the signal profiles contain systematic noise introduced to the read depth signal by the repeat regions with low mappability [9, 16], in the form of loss of signal. This noise causes discontinuities in the identified ERs. This is an important factor especially in the the intergenic regions where a large ERs, which may mark a long regulatory region, gets fragmented into smaller ERs.

Many different approaches have been applied for identification of broad ERs, which include change point identification within the formality of Bayesian inference (BCP, [17]), local island identification and clustering (SICER [18]), local thresholding and merging (MACS), using local Poisson statistics to identify broad ERs (SPP), and wavelet based smoothing and identification of enriched regions (WaveSeq [19]), which is also applied to analysis of ChIP-chip datasets [20].

In this paper, we present MUSIC, a method to identify enriched regions in ChIP-Seq experiments. MUSIC first uses mappability correction at the nucleotide resolution so as to correct for the spurious loss of signal in the regions with low mappability. Next, MUSIC performs a multiscale decomposition of the corrected RD signal. This decomposition is adopted from the scale-space filtering theory in signal processing [21], which is used widely for signal segmentation, smoothing, and enhancement. Unlike the wavelet based multiscale approaches that use linear filtering, we take an approach to multiscale decomposition using the non-linear median filtering. Basically, MUSIC exploits the fact that at each decomposition, smoothing with the certain window length removes the small details in the signal (like small peaks and small valleys) and the candidate enriched regions in the signal are detected as the regions between consecutive local minima of the smoothed signal [22, 23]. MUSIC then identifies the significantly enriched regions at each scale, which yields the scale specific enriched regions (SSERs). In general, at smaller scales, the SSERs correspond to more punctate binding/modification levels compared to SSERs at higher scales, which represent the broader ones. To identify the final set of ERs, MUSIC merges the SSERs from all the scales.

In order to evaluate the accuracy of the identified ERs by MUSIC, we performed benchmarking experiments to compare the accuracy and reproducibility of the ERs identified by MUSIC with numerous other ER identification methods. We concentrated on factors whose ERs manifest at different (i.e., broad, puncate, and point binding) length scales so as to make a thorough comparison with a variety of accuracy metrics. We show that MUSIC performs favorably in the comparisons.

Next, we concentrate on the RNA Polymerase II ChIP-Seq datasets. Motivated by the basic observation that the stalled polymerase tends to show punctate enrichments (SSERs at small scales) and that elongating Polymerase tends to show broad enrichments (SSERs at higher scales), we computed the SSERs for the RNA Polymerase II ChIP-Seq dataset using MUSIC. Then using the identified SSERs, we estimate the length scale for polymerase binding for all protein coding genes. We demonstrate that the genes with punctate polymerase binding have significantly lower expression (close to 0) than the genes that show more boundly bound polymerase. We corroborate this observation with the ChIP-Seq data for elongating (phosphorylated) form of Pol2. We conclude that the length scale of binding of polymerase at the gene promoters as identified by MUSIC is indicative of its state, i.e., stalled or elongating.

The paper is organized as follows. We first present the MUSIC algorithm and laying out the steps of the algorithm. Then we present a comparison of MUSIC with other ER identification algorithms. We finally present the analysis of the RNA Polymerase II data with gene expression levels.

2 RESULTS AND DISCUSSION

2.1 MUSIC Algorithm

Figure 1 shows the flowchart for MUSIC (See Methods for more details.) Here we summarize each step briefly. The input to MUSIC are the sets of reads from the ChIP and control samples (Steps 1 and 2), the set of window lengths to be used in multiscale decomposition, and the multi-mappability profile. The multi-mappability profile quantifies at each position, the average number of reads that gets mapped non-uniquely (See Methods). Therefore, for a position that is uniquely mappable, the multi-mappability value is 1. For the repeat regions, multi-mappability value increases. Fig. S1 shows aggregation of multi-mappability profile around different genomic elements for different read lengths. It should be noted that multi-mappability signal is computed once for each read length (See Methods.) MUSIC first preprocesses the reads and filters the duplicates. Then MUSIC computes a scaling factor using linear regression between the ChIP and control signal profiles. The slope of the regression is used as a normalization factor for control.

Then, in Step 3, the ChIP and normalized control signal profiles are generated, and the ChIP profile is filtered and corrected with respect to mappability using the multi-mappability profile. The correction can be formulated as following:

$$\tilde{x}_i = \overbrace{\max[x_i, \underbrace{\text{median}\left(\{x_a\}_{a\in[i-l_c/2, i+l_c/2]} \mid m_a < \overline{m}_{\text{exonic}}\right)}_{\substack{\text{Median of the signal values at highly mappable} \\ \text{positions around } i}}]}^{\substack{\text{Maximum of the signal value at } i \text{ and} \\ \text{the median signal at highly mappable positions}}}$$

where x_i and $\tilde{x}_i$ are the uncorrected and corrected signal values, respectively, at position i, m_a is the value of multi-mappability profile at position a, l_c is the length of median filter utilized in correction which is by default set to 2000 base pairs, and $\overline{m}_{\text{exonic}}$ is the average multi-mappability signal value over the exonic regions, which we identified as the most mappable regions in the genome (See Fig S1). In summary, for each position i, MUSIC computes the median of the signal values at highly mappable positions (multi-mappability signal smaller than $\overline{m}_{\text{exonic}}$) within l_c vicinity of i. Then MUSIC compares this value with the signal value at i and assigns the maximum to the corrected value. The basic idea behind this correction is that since we know that low mappability causes decrease in the signal level, if the signal value at i is higher than its vicinity, then it is highly likely that the mappability did not affect the signal value at i. Otherwise, it is replaced by the median signal value at mappable positions. It should be noted that maximum filtering, also known as dilation in image processing, is used for feature enhancement in images [24].

MUSIC then performs median filtering to the mappability corrected ChIP profile to compute multiscale decomposition of ChIP signal at multiple length scales (Step 4, Fig S2.) For this, MUSIC uses window lengths beginning with l_{start} and ending at l_{end} and performs sliding window based median filtering. The window length is increased multiplicatively between consecutive scales, thus, the window lengths form a geometric series:

$$\{l_{start}, \lfloor l_{start} \times \sigma \rfloor, \lfloor l_{start} \times \sigma^2 \rfloor, \cdots, l_{end}\}$$

where σ is the multiplicative factor between consecutive window lengths, which is set to 1.5 by default. $\lfloor l_{start} \times \sigma \rfloor$ denotes the largest integer value that is smaller than $l_{start} \times \sigma$, which is necessary since the window lengths are integer values. The multiplicative factor tunes how finely MUSIC samples the scale spectrum. For small σ, MUSIC analyzes large number of scale lengths, however, this also increases the run time.

For smoothed signal at each scale, MUSIC identifies all the local extrema, i.e., local minima and local maxima (Step 4 in Fig. 1). The regions between the consecutive local minima are marked as the candidate enriched regions. Due to the nature of smoothing process, the signal may become oversmoothed at large scales (long windows) which causes overmerging of the enriched regions. To avoid this, it is necessary to remove the regions with oversmoothed signal. For each enriched region, MUSIC computes the fraction of the maximum of smoothed RD signal (at the corresponding scale) to the maximum of the unsmoothed ChIP signal within the boundaries of the enriched region. If this fraction is smaller than the smoothed versus unsmoothed signal ratio threshold (denoted by γ), MUSIC discards this candidate enriched region (refer to Methods.)

The regions identified from the consecutive minima are rough and it is necessary to identify the location of densest signal enrichment within each region. To achieve this, MUSIC performs a Poisson background based thresholding and p-value minimization to trim the ends and identifies the densest regions of signal enrichment in the ERs. Step 5 in Fig 1 illustrates the trimmed ends of the candidate enriched regions. Finally, MUSIC computes the p-value from a binomial test for each trimmed region and filters out those whose p-values are larger than 0.05. We refer to the remaining regions as the scale specific enriched regions (SSERs). SSERs contain all the information about the enrichments in the signal over a spectrum of length scales.

2.1.1 Identification of ERs

MUSIC utilizes SSERs to identify enriched regions in the genome. For this, the candidate ERs are computed by merging the SSERs identified in all the scales (Step 6 in Fig. 1). MUSIC then filters out the ERs with respect to discordance of the signal levels on positive and negative strands. For this, MUSIC computes the amount of signal mapping to positive and negative strand in each ER and filters out the ERs for which the counts of reads that map to positive and negative strand within a factor of 2 of each other (See Methods.)

For each of the remaining ERs, MUSIC computes the p-value from binomial test using the number of reads in the ChIP and normalized control samples. The multiple hypothesis correction is performed by the Benjamini-Hochberg procedure [25]. The q-values computed after the correction are thresholded with respect to 0.05 for identification of the significant ERs.

2.1.2 SSER Pileup Scale and Evaluation of Broadness of Enrichment

The scale dependence of SSERs is a useful property for evaluating the broadness of enrichment. Each SSER represents a locally enriched region at a certain length scale. Therefore, the signal around a position that is covered by large number of SSERs (at different scales) is more broadly enriched than the signal around a position that is covered by less number of SSERs. Following this basic observation, MUSIC pools the SSERs from all the scales and counts the number of SSERs covering each position, which quantifies the broadness of enrichment at each position in the genome. We refer to this value as the SSER Pileup Scale of the position.

To evaluate the spectrum of enrichment length scales specific to different datasets, we processed multiple ChIP-Seq datasets (CTCF, RNA Polymerase 2, H3K4me1, H3K4me3, H3ke36me3, H3K27me3, and H3K9me3) from the ENCODE project for K562 cell line with window length parameters $l_{start} = 100$ bps, $l_{end} = 2.5$ Mbp, and $\sigma = 1.5$ (Total of 25 scales) and computed the SSER pileup scales for the positions on chromosome 1. Figure 2 shows the distribution of SSER pileup scales, i.e., the pileup scale spectrum of all the positions on chromosome 1 for different datasets. We use this plot to assess the scale length characteristics of different datasets. CTCF, a punctate binding transcription factor, has a maximum frequency at the smallest pileup scales compared to other datasets. This suggests, as expected, that CTCF has the most punctate ERs compared to other datasets. H3K4me3 and H3K4me1, active promoter and enhancer HM marks, show broader enrichments than CTCF. H3K36me3 and H3K27me3, which mark active and repressed gene bodies, show broader enrichments and finally H3K9me3, an HM associated with large heterochromatin domains, shows the broadest enrichments. Another interesting observation is that the plots for H3K4me3, H3K4me1, and H3K36me3 datasets have maxima at certain scales, which indicates that these HMs get enriched at specific length scales that are observed frequently. Finally, the RNA Polymerase II signal profiles show a high frequency of enrichments at small scales that shows more gradual decrease in frequency as the scale increases.

2.2 Comparison with Other Methods

In order to evaluate the accuracy of ERs, we compared MUSIC with 8 other algorithms that identify ERs from ChIP-Seq data: DFilter\cite, ZINBA\cite, F-Seq\cite, BCP [17], SPP [26], MACS [27], SICER [18], and PeakRanger [28]. For the detailed list of parameters that were used to run the compared methods, see Section 4. For comparing the accuracy of the methods on identification of more punctate ERs, we used H3K4me3 ChIP-Seq dataset from ENCODE. Finally, we compared the accuracy of the methods with respect to identification of point binding events, we used the CTCF ChIP-Seq dataset.

2.2.1 Comparison of Broad ER Identification

For comparing the performance of methods on broad marks, we ran all the algorithms (in broad ER identification mode) using H3K36me3 and H3K27me3 ChIP-Seq datasets for GM12878 and K562 cell lines from ENCODE project [5]. H3K36me3 is known to mark the bodies of actively transcribed genes [29]. We use this observation to build a gold standard set for H3K36me3 as the bodies of expressed transcripts. We downloaded the transcript quantifications (in RPKMs) from Djebali et al [30] and removed the transcripts with low expression. The bodies of the expressed transcripts are then merged to generate the gold standard set for H3K36me3 ERs. Rather than selecting one expression threshold for identifying the expressed transcripts, we selected thresholds between 0 and 1 RPKM increasing with steps of 0.01 so as to evaluate the accuracy of ER calls against multiple gold standard sets identified at different levels of expression. For these comparisons, we ran MUSIC and other methods with default parameter settings (See Methods).

We observed that MUSIC tends to identify longer ERs compared to other methods and that different methods have very different total ER coverage. To measure the accuracy of identified ERs, it is necessary to account for the difference in the coverage of the identified ERs. We used sensitivity (the fraction of the coverage of correctly predicted ERs to the coverage of the gold standard set) and positive predictive value (the fraction of the coverage of correctly predicted ERs to the coverage of identified ERs). To summarize these accuracy values in one measure, we chose F-measure that is computed as the harmonic mean of sensitivity and positive predictive value (See Methods). Having one measure of accuracy enables us to easily compare the accuracy of methods with changing RPKM thresholds.

Figures 3a and 3b show the F-measure for the H3K36me3 ERs from different methods with respect to the changing RPKM cutoffs. MUSIC has higher F-measure than all the other methods for GM12878 at all expression cutoffs, followed by BCP. For K562, MUSIC has higher F-measure than all other methods for expression cutoffs smaller than 0.8 then falls slightly below BCP. For assessing the importance of mappability correction, we ran ER identification without mappability correction and computed the F-measure of the ERs. Fig 3e shows the F-measure versus RPKM threshold. Using mappability map significantly increases the accuracy of identified ERs and shows the importance of utilizing the mappability correction in ER identification.

2.2.2 Comparison of Punctate ER Identification

For comparing the methods for identification of punctate ERs with smaller length scales, we first chose to compare the methods on H3K4me3 HM, which marks the promoters of active genes. As a gold standard, we utilized the promoters of the active genes (RPKM > 0.5) as the gold standard. We identified the ERs that have at least 5% overlap with the promoter region (2 kb region around the annotated transcription start site). For this comparison, we sorted the top 20,000 ERs with respect to the score reported by each method then computed the overlap of the ERs with active promoters. Starting from the top ERs, we plotted fraction of active promoters that are identified correctly versus fraction of ERs that overlap with active promoters. These are shown in Fig 3c and 3d, respectively for K562 and GM12878 cell lines. MUSIC performs favorably compared to other methods, followed by DFilter and SICER.

For comparing the accuracy of methods with respect to identification of ERs for point binding factors like transcription factors, we used CTCF transcription factor, which has a well-studied motif associated with it. We identified the ERs using each method (using TF peak calling mode when available). As a gold standard we utilized the motif datasets from ENCODE project {\cite Pouya's NAR motif paper}. As an accuracy measure, we ranked the top 2000 peaks then computed the fraction of ERs with a motif around 150 base pairs of the summit of ERs. The results are summarized in Table S2. We observed that MUSIC, SPP, MACS, and DFilter perform very similarly followed by other methods.

2.2.3 Comparison of Reproducibility of ERs

We also evaluated the reproducibility of the ERs. For this comparison, we used the replicates generated by the ENCODE Project. For H3K36me3, H3K27me3, and H3K4me3, we computed the reproducibility as the average of fraction of the overlapping regions to the total coverage of each replicate (See Methods) shown in Figure 3f. Overall reproducibility of MUSIC is higher than the other method. In addition, for broader marks (H3K36me3 and H3K27me3), MUSIC has the the top or the second top reproducibility compared to other methods. For H3K4me3, all methods have comparable reproducibility.

2.3 Analysis of the RNA Polymerase II and Gene Expression Levels

Next, we concentrated on the Polymerase II binding data from ENCODE project. Polymerase shows distinct patterns of binding such that the depending on the state of polymerase, i.e., the genes with broadly bound polymerase are being actively transcribed and show higher levels of expression compared to the genes that are bound punctate by the Polymerase [31, 32]. This makes the Polymerase data suitable for the multiscale analysis using MUSIC.

For evaluating the relation between the expression and the length scale of binding, we processed Polymerase ChIP-Seq data for K562 cell line from ENCODE project using MUSIC and computed the SSERs pileup scale using parameters $l_{start} = 10$ bps, $l_{end} = 2.5$ Mbps, and $\sigma = 1.5$. Then, for each protein coding gene, we assigned the broadness of polymerase binding as the maximum of the SSER pileup scale within the gene body. We then quantified the gene expression levels in RPKMs using the RNA-seq datasets from ENCODE Project. Finally, we plotted the 2 dimensional histogram of binding scale and gene expression level for each gene, which is shown in Fig. 4a. In the plot, two components are revealed: One component is at the low log expression levels (Smaller than 0.1) and has a maximum frequency at scale length of 950 base pairs. This component corresponds to the stalled polymerase, which has punctate enrichment profile and produce very little or no transcripts. The second component is observed at log RPKMs greater than 0.5 with a peak of scale level at around 6 kilobases. With the elongating polymerase and high expression levels, this component is associated with actively transcribed genes.

To study these components further, we focused on the two components of polymerase binding and gene expression levels: For the genes with stalled polymerase, we selected genes with scale between 150 bps and 2.3 kbps and low expression (log(RPKM)<0.1). For the genes with elongating polymerase, we selected the genes with pileup scale greater than 950 base pairs with high expression (log(RPKM) > 0.1). We performed aggregation of the ChIP-Seq RD signal for the elongating form of polymerase, Pol2s2, from the ENCODE project, around the promoters of genes in both sets. The motivation is that signal for Pol2s2 marks the location of elongating polymerase, which should associate with the promoters that we marked as elongating and not with the promoters that are bound by the stalled polymerase. Fig 4b shows the aggregation plots. As expected, for the punctate bound and low expression genes, the aggregation plot shows very little Pol2s2 binding. In contrast, the high expression and broad bound promoters show a substantially higher Pol2s2 binding that extends into the gene body.

3 CONCLUSIONS

We present a novel method, MUSIC, for the identification of enriched regions in ChIP-Seq experiments. MUSIC utilizes a multiscale decomposition of the ChIP-seq signal profile in conjunction with a novel mappability correction for mediating the effects of the data. Mappability is an important aspect of ER identification from next generation sequencing data especially for identifying the broad domains of enrichment since the read depth profiles are highly correlated with the mappability map. We showed that MUSIC outperforms other methods in terms of accuracy of H3K36me3 ERs in comparison with the expressed transcripts identified from the expression data from ENCODE project. An important advantage of MUSIC is that the users can specify the scales that they would like to concentrate on, which is done using the begin and end scale parameters for the multiscale filtering. With the diverse enrichment characteristics of the targets for ChIP-Seq experiments, we believe this customizability will prove very useful for processing the datasets generated using ChIP-Seq experiments for which broad binding profiles are observed.

Compared to the kernel based linear filters (which are also used in the wavelet based multiscale decompositions), multiscale decomposition using median filtering has two advantages. First, at low noise levels, median smoothing preserves the edges, i.e. sharpness of increase and decrease of the RD signal at the ends of enriched regions, in the signal better than the linear filters. Secondly, the median smoothing is more tolerant to the burst or impulse noise compared to the linear filters. This is important for the enriched region identification since the systematic noise added by multi-mappability can be viewed as an impulse noise [33, 34]. For example, in another context, a recent study [35] uses linear Gaussian filtering based multiscale decomposition to generate multiscale representations of the genomic signals. The non-uniform mappability of the genome should be expected to substantially affect the linear multiscale representation because of the impulsive noise introduced by the lowly mappable regions. The median coupled with the mappability correction utilized by MUSIC is much more robust to impulsive noise than Gaussian filtering. In addition, the multiscale framework utilized in MUSIC can be extended to other non-linear multiscale decompositions. The multiscale representation presented in [35], however, relies heavily on Gaussian filtering [36].

We also processed the RNA Polymerase II using MUSIC. The RNA Polymerase II is suitable for multiscale analysis because the Polymerase, unlike other DNA binding proteins, shows a large spectrum of ER lengths. Furthermore the broadness of binding of the Polymerase indicative of its state, i.e., stalled or elongating. We showed that there is a significant distinction between the expression levels of genes that are bound broadly by the Polymerase compared to the genes that are bound punctate.

4 METHODS

We describe signal processing methodology underlying MUSIC in more detail.

4.1 Input Normalization

It is necessary to normalize the control signal profile with respect to ChIP-Seq profile because the read depths can be different. For each chromosome, MUSIC first divides the chromosome into 10,000 base pair bins then computes the total ChIP-seq and control signal in each window. Finally, it estimates the normalization factors as the slope of the minimum squared error estimate of the slope

$$\rho = \underset{\rho'}{\operatorname{argmin}} \left\{ \sum_i (w_i - \rho' \cdot c_i)^2 \right\}$$

Where w_i and c_i represent the total signal in i^th bin for ChIP and control samples, respectively.

4.2 Mappability Correction Filter

Given the read depth signal at each nucleotide position, MUSIC corrects for the loss of signal caused by low mappability using following filtering:

$$\tilde{x}_i = \max\left[x_i, \mathrm{median}\left(\{x_a\}_{a\in[i-l_c/2, i+l_c/2]} \mid m_a < \overline{m}_{\mathrm{exonic}}\right)\right]$$

Where x_i is the signal value at nucleotide position i, $\mathrm{median}(\{x_i\})$ is the median of the set $\{x_i\}$, m_a is the value of the multi-mappability profile at the position a, and l_c is the window length used in mappability aware filtering. Using this filtering, MUSIC infers the signal values for positions with low mappability using the median of the values at nearby positions with multi-mappability signal lower than $\overline{m}_{\mathrm{exonic}}$, which is 1.2. We selected this value since it is the smallest multi-mappability signal profile value, i.e. most mappable, over exons and promoters as shown in Fig S1. We set the window length l_c to 2000 bps, empirically. This window length depends on the distribution of length of the non-mappable region lengths. Different l_c values did not seem to have a significant effect on the results for the human genome.

This filtering is inspired from the dilation operation in image processing, which is a morphological filter and has been used, in combination with other filters, for image enhancement. In our experiments, we also observed that the operation defined above tends to enhance the significant enriched regions.

4.3 Multiscale Decomposition by Median Filtering

MUSIC utilizes a median filtering based multiscale decomposition. We selected to use median filtering since it has many applications in signal processing for performing signal smoothing with edge preserving. Given a window length, i.e. the scale, median filtering can be formulated as:

$$x_i^s = \mathrm{median}\left(\{\tilde{x}_a\}_{a\in\left[i-\frac{l_s}{2}, i+\frac{l_s}{2}\right]}\right), l_s \in (l_{start}, \lfloor l_{start} \times \sigma \rfloor, \cdots, l_{end})$$

Where x_i^s is the i^{th} value of the decomposition at scale level s for which the smoothing window length is l_s, and $\tilde{x}$ is the mappability corrected signal profile. The window length l_s is chosen from a geometric series with the factor σ to ensure that the larger scales do not dominate the identified SSERs [21].

The multiscale decomposition enables automatic identification of blobs in the signal profiles at different scales with very small computational requirement. MUSIC uses a fast and efficient method to implement the median filtering by storing the histogram of the signal values in the current window and processes only the new and obsolete signal values that enter and leave the current window to update the histogram when moved to the next window.

4.4 Identification of Candidate Scale Specific Enriched Regions

After the multiscale decomposition, MUSIC identifies all the local minima in the decomposition. MUSIC utilizes regions between minima points as the regions of enrichment. For this, MUSIC computes the derivative of the signal at each point as the difference between consecutive values:

$$x'^s_i = (x^s_i - x^s_{i-1})$$

where x'^s_i is the derivative of the smoothed signal x^s_i. MUSIC assigns the local extrema at the points where the derivative changes sign:

$$I_{min} = \{i \mid x'^s_i < 0, x'^s_{i-1} > 0\}$$

$$I_{max} = \{i \mid x'^s_i > 0, x'^s_{i-1} < 0\}$$

Where I_{min} and I_{max} are the sets of positions of minima and maxima of x^s_i, respectively. The scale specific candidate enriched regions of x^s_i are identified as the regions between the consecutive minima.

4.5 Comparison of Smoothed Signal in Candidate Enriched Regions

For the candidate enriched regions in each smoothing scale, MUSIC uses the value of smoothed signal levels and unsmoothed signal levels for assessing the quality of enriched region. A scale specific candidate enriched region is filtered if the ratio of the maximum of smoothed signal to the maximum of the unsmoothed signal within the candidate region is higher than the smoothing statistic threshold, γ. In other words, MUSIC removes the candidate enriched region $[i, j]$ at scale s, if

$$\frac{\max(\{x^s_a\}_{a\in[i,j]})}{\max(\{x_a\}_{a\in[i,j]})} < \gamma.$$

The comparison between the ratio on the right and γ offers a simple and efficient check to evaluate whether the signal within the candidate region identified at the scale level s is severely smoothed. This way, MUSIC efficiently detects and avoids overmerging of consecutive regions that have high signal enrichment and are close to each other. In addition, MUSIC removes the enriched regions whose signal levels are severely smoothed. By default γ is set to 4, refer to Methods for selection of γ.

4.6 Candidate Enriched Region End Trimming using Poisson Distribution Model

MUSIC trims the ends of the candidate enriched regions using a Poisson null model for the signal distribution. For this, MUSIC divides genome into 1 megabase windows and for each 1 megabase window estimates the mean of all the values. Using this as the mean parameter μ of the Poisson distribution, MUSIC selects a threshold that satisfies 5% false positive rate:

$$\tau = \underset{t}{\operatorname{argmin}}\{F_{X_\mu}(t) > 0.95\}, X_\mu \sim Poisson(\mu))$$

Where F_{X_μ} represents the cumulative distribution function of X_μ, which is distributed as Poisson with mean μ. For a region with start and end at positions i and j, respectively, the trimmed end coordinates are given as:

$$i' = \underset{a}{\operatorname{argmin}}(x_a > \tau), a \in [i, j]$$

$$j' = \underset{a}{\operatorname{argmax}}(x_a > \tau), a \in [i, j]$$

Where i' and j' are the trimmed start and end coordinates, respectively. The regions for which the signal level does not pass the threshold are removed from the candidate ER list.

4.7 Candidate Enriched Region End Trimming via p-value Minimization

MUSIC fine-tunes the ends of the merged ERs using a p-value minimization procedure. This maximizes the compactness of the merged regions. The end-refined merged regions are the candidate regions of enrichment before p-value computation. The end trimming can be formulated as:

$$i' = \underset{a}{\operatorname{argmin}}\big(p(a, j \mid l_{p_{val}} = (\mathrm{j - a + 1}))\big), a \in [i, j]$$

$$j' = \underset{a}{\operatorname{argmin}}\big(p(i', a \mid l_{p_{val}} = (\mathrm{a - i' + 1}))\big), a \in [i', j]$$

where $p(a, b \mid l_{p_{val}})$ represents the p-value for the region starting at a and ending at b with the length of p-value window given by $l_{p_{val}}$ (Refer to p-value computation.)

4.8 Per Strand Concordance Test

For each ER, MUSIC computes the total signal on positive and negative strands and filters out the enriched regions for which there is high discordance between the signals:

$$\min\left(\frac{\sum_i x_i^+}{\sum_i x_i^-}, \frac{\sum_i x_i^-}{\sum_i x_i^+}\right) < 0.5$$

where $\sum_i x_i^+$ and $\sum_i x_i^-$ is the total signal on the positive and negative strand within the start and end coordinates of the ER, respectively.

4.9 P-value Computation and FDR Estimation

We use one-tailed binomial test to compute the p-values for each candidate enriched region. We first count the number of reads in the chip sample (n_{chip}) and control sample ($n_{control}$) that overlap with the region, then compute one tailed p-value as:

$$p = \sum_{r = n\prime_{chip} + 1}^{n\prime_{chip} + n\prime_{conrol}} \binom{n'_{chip} + n'_{control}}{r} 0.5^{(n'_{chip} + n'_{control})}$$

Where n'_{chip} and $n'_{control}$ are the normalized read counts for the region:

$$n'_{chip} = \frac{n_{chip}}{l_{chip}} \times l_{p_{val}}$$

$$n'_{control} = \frac{n_{control}}{l_{control}} \times l_{p_{val}}$$

where $l_{p_{val}}$ is the length of the p-value computation window and p refers to the p-value value for the ER. It should be noted that the larger values of $l_{p_{val}}$ increase the significance of all the regions and the false positive rate (See Parameter Selection for Benchmarking.) We perform multiple hypothesis correction by false discovery rate estimation (q-values) using the Benjamini-Hochberg procedure [25]:

$$q_i = p_i \times \frac{N_{ERs}}{i}$$

where N_{ERs} is the total number of enriched regions and i is the rank of the ER in the ER list sorted with respect to increasing p-value. By default, MUSIC uses default q-value cutoff of 0.05. The filtered ERs are reported in BED format with their q-values in the score field.

4.10 Summit and Trough Identification

For DNA-binding protein ChIP-Seq data, e.g. transcription factors, MUSIC reports the location of the highest signal level within the ER as the summit of the signal, which can be used as the binding position. An important consideration in ER identification is the identification of valleys (or troughs) in the signal. For example, the troughs in H3K4me3 and H3K27ac ERs may correspond to the nucleosome free regions in promoters and enhancers, respectively, where the transcription factors can interact with DNA and regulate transcription. Therefore, identification of the troughs (in addition to the summits) is an important side information about each ER. Our analysis, however, shows that much of the troughs in ChIP-Seq signal is caused by the decrease in the mappability of the genome (See Fig S6). MUSIC reports one trough position in each peak (for punctate peaks) by determining the smallest position within the top two tallest peaks such that the average multi-mappability around the trough is smaller than exonic multi-mappability (m_e).

4.11 Multi-Mappability Signal Generation

MUSIC can generate the multi-mappability signal profiles. For this, MUSIC utilizes an existing read mapping tool. Currently MUSIC uses bowtie2 [37], a very popular short read mapping algorithm, by default. MUSIC first fragments all the chromosomes to the read length of interest, maps all the fragments to the genome using bowtie2 with 2 mismatches and reporting of maximum of top 5 multimapping positions per fragment. Then MUSIC uses the mapped reads to build the multi-mappability RD signal profile. The regions with high signal corresponds to regions with low mappability. We generated multi-mappability profiles for hg19 genome assembly for read lengths of 36, 50, 76, and 100 bps that are available for download with MUSIC.

4.12 Parameter Selection for Benchmarking

There are 3 parameters associated with MUSIC, starting scale window length (l_{begin}), ending scale window length (l_{end}), and the p-value computation window length ($l_{p_{val}}$). For selecting l_{begin} and l_{end}, we utilize a basic property of the median filtering (See Fig S3). In order to detect an enrichment of length l it is necessary to ensure following:

$$l_{begin} < 2 \times l$$

Similarly, in order to distinguish between two enriched regions that are l base pairs away from each other, it is necessary to ensure following:

$$l_{end} < 2 \times l$$

Thus, l_{begin} should be small enough to ensure detection of the smallest enrichments that we expect to observe and l_{end} should be set to a value to detect each individual enrichment separately without overmerging (See Fig S3b and S3c.) We are assuming that the basic enriched units are the gene bodies, therefore, we choose l_{begin} using the length distribution of gene bodies, shown in Fig S3e. As most of the genes have length longer than 512 bps (log value of 9), we set l_{begin} to 1000 bps. For choosing l_{end}, we computed the cumulative distribution of gene-gene distances, shown in Fig S3d. Evaluating this plot, we observe that 10% cutoff at around log distance of 12.5. As a suitable compromise with the gene length distribution (The median is at log value of 15), we set l_{end} to $2 \times 2^{13} \approx 16000$ bps.

The other parameters to set is $l_{p_{val}}$. This parameter tunes the p-values of the SSERs and the final set of ERs. Generally, increasing $l_{p_{val}}$ increases the power of identification (See p-value computation) but also increases FDR. In addition, depending on the sequencing depth, $l_{p_{val}}$ can be used to avoid saturation of the identified ERs [26]. To select $l_{p_{val}}$, assessed the p-values computed using different $l_{p_{val}}$ values and Fold change (the number of chip sample reads divided by number of normalized control reads). Fold change is generally independent of the sequencing depth and represents an unbiased estimate of enrichment. For different $l_{p_{val}}$ values, we divided chromosome 1 into bins of $l_{p_{val}}$ base pairs and computed the p-value and the fold change in each bins. Fig S4 shows the scatter plot of p-value versus fold change for different values of $l_{p_{val}}$. It can be observed that as $l_{p_{val}}$ increases, the p-values corresponding to same fold change decreases. Our basic idea is to choose $l_{p_{val}}$ such that the windows that show significant enrichment with respect to fold change (above 2) are also significant with respect to p-value (log p-value smaller than -3) and that the windows that do not show significant fold change (below 1.5) do not have significant p-values. Using these criteria, we set $l_{p_{val}}$ to 1750 base pairs.

For punctate marks (like H3K4me3 and H3K27ac), MUSIC is set to run at the lower scale spectrum with $l_{begin} = 100, l_{end} = 2000$. This way MUSIC aims at identifying small ERs and at identifying the enrichments at a reasonable expected length range of several kilobases. We set $l_{p_{val}}$ to 1750 base pairs.

For transcription factors, for which point binding events occur at almost single base pair resolution, MUSIC is set to run at a single scale with $l_{begin} = 100, l_{end} = 200$, which amounts to utilizing MUSIC at one scale. It should be noted that the utility of mappability correction and multiscale decomposition is most effective for identification of more broad ERs.

The remaining parameter, namely γ, is the threshold on the ratio of the maximum of the smoothed signal and the unsmoothed signal on an SSER. This parameter enables MUSIC to avoid overmerging segments by comparing the signal level in the smoothed signal and the original signal. To visualize the effect of changing γ on the identified SSERs, we computed the SSERs for H3K36me3 and H3K4me3 marks for K562 cell line. We then identified the computed the smoothing statistic (as given in Section 4.5) for each SSER. Then we plotted the cumulative distribution of all the SSERs with respect to the smoothing statistic, which is shown in Fig. S5. It can be seen that for H3K4me3 that the distribution is much more skewed toward smaller γ than H3K36me3, which is expected since H3K4me3 has much narrower ERs than H3K36me3. To be as inclusive as possible, we choose γ=4 (98% and 90% of the SSERs respectively for H3K4me3 and H3K36me3 pass the smoothing statistic test) as a suitable parameter to balance the tradeoff between being inclusive in the identified SSERs and overmerging the ERs.

The final parameter to set is, σ, which is the multiplicative factor between the consecutive scales. Higher values of σ decreases the runtime of MUSIC but important information can be lost since sampling of the scale space is sparsified. For example, the SSERs that can be identified at a mid-scale scale in between can get lost. We evaluated several different values for σ and realized that for $\sigma > 2$, MUSIC uses a very sparse set of scales that miss many ERs. As a suitable compromise, we chose to use $\sigma = 1.5$. It should be noted that it may be useful to utilize smaller values for σ when more punctate ERs are being analyzed. For example, we used $\sigma = 1.1$ for performing the scale spectrum analysis in Figure 2.

4.13 Parameters Used for Peak Calling Methods in Benchmarking

The most recent versions of the tools are downloaded from the website and documentations are followed for running the tool in the correct mode.

1. BCP [\cite]: For histone marks (H3K36me3, H3K27me3, and H3K4me3), we used BCP_HM tool with command line options: -f 200 -w 200 -p 0.05. For CTCF dataset, we used BCP_TF tool with command line options: -e 10 -p 0.00000001
2. PeakRanger[\cite]: For H3K36me3, H3K27me3, and H3K4me3 datasets, we used ‘ccat’ option fro broad peak calling. For CTCF peaks, we used ‘ranger’ option.
3. ZINBA[\cite]: For broad marks, we used the unrefined ERs from ZINBA with ‘broad’ flag on as explained in the documentation. For H3K4me3 and CTCF peaks, we used the refined peaks with ‘broad’ flag turned off.
4. F-Seq[\cite]: For broad marks and CTCF, F-Seq is run in default mode.
5. SICER[\cite]: For broad marks, SICER is run with the with command options: hg19, w=200, fragment_size=150, 0.74, g=600, FDR=0.01. For CTCF, SICER is run with smaller gap size of g=200.
6. SPP[\cite]: For broad marks, SPP is run in broad mode using get.broad.enrichment.clusters(…). For CTCF, the peak calling mode is run using find.binding.positions(…).
7. DFilter[\cite]: For H3K36me3 and H3K27me3, DFilter is run with command line options ‘-nonzero -bs=200 -ks=40 -std=2’ then removed the peaks that has score smaller than 2. For H3K4me3, DFilter is run using ‘-bs=100 -ks=100 -dir -std=2’ then peaks with score smaller than 6 are removed. For CTCF, we ran DFilter with ‘-nonzero -bs=25 -ks=50 -pm=300 -std=2’.
8. MACS: For broad marks, MACS is run with options ‘--broad -g hs’. For CTCF, MACS is run with ‘-g hs -q 0.01’.

4.14 Accuracy Measures

For evaluating the accuracy of H3K36me3 ER calls, we computed sensitivity, positive predictive values:

$$Sensitivity = \frac{covg(P \cap G)}{covg(G)}$$

$$PPV = \frac{covg(P \cap G)}{covg(P)}$$

Where $covg(P)$ is the coverage of ERs, $covg(G)$ is the coverage of expressed gene bodies and $covg(P \cap G)$ is the coverage of the overlap between expressed gene bodies and ERs. We combined these two accuracy measures to compute F-measure, computed as:

$$F - measure = \frac{2 \times Sensitivity \times PPV}{(Sensitivity + PPV)}$$

For assessing the reproducibility of the identified ERs from two biological replicates, we use the average overlap fraction between the ERs:

$$Overlap\ Fraction = \left(\frac{covg(P_1 \cap P_2)}{2 \times covg(P_1)} + \frac{covg(P_1 \cap P_2)}{2 \times covg(P_2)}\right)$$

where $covg(P_1)$ and $covg(P_2)$ represent the coverage of the ERs identified from replicate 1, and replicate 2, respectively.

For H3K4me3 ER accuracy assessment, we sorted the top 20,000 ERs identified by each method. Then we overlapped the identified ERs with the promoters of active genes (RPKM > 0.5), which are defined as the 2000 base pair vicinity of the annotated transcription start site. We enforced that the overlap between the promoter region and the peaks at least 5% of the length of the peak. Then, starting from top 1000 ERs, we computed the fraction of active promoters recovered and fraction of ERs that overlap with active promoters for the top peaks. At each step, we increased peak number by 1000.

For CTCF peaks, we sorted the top 2000 peaks from each method, then computed the fraction of peaks whose summit overlaps with a known CTCT motif within 150 base pair vicinity.

4.15 Datasets and Data Processing

The ChIP-Seq datasets for H3k36me3, H3K27me3, H3K4me3 modifications, and CTCF are obtained from ENCODE project [\cite] through UCSC genome browser. The transcript quantifications and RNA-seq datasets are downloaded from Djebali et al [30]. For the transcript quantifications, we used the average RPKM values for the transcripts from two replicates that satisfied the reproducibility criteria that iIDR smaller than 0.1. The transcript and gene annotations are obtained from Harrow et al [38]. The CTCF motifs are downloaded from [\cite].

Competing Interests

Authors declare no competing interests.

Authors' Contributions

AH and JR designed the methodology and experimental setup with input from MG. AH implemented the code, performed the analysis, and wrote the manuscript. All authors edited and approved the final manuscript.

Acknowledgements

Authors thank Akdes Serin Harmanci for constructive comments on the study and experiment design. This study was supported by grants from NIH.

Figure Captions

Figure 1: Flowchart of MUSIC. H3K36me3 ChIP and control data in region chr1:55,170,679-55,240,996 for K562 cell line is used for illustrating the signal processing steps. ChIP and control reads (represented by short horizontal lines) are filtered for duplicates (red colored) and control signal is normalized with respect to ChIP signal (1). RD profiles are generated (2). ChIP-seq profile is corrected for mappability (labeled "Mappability Corrected Signal" profile) using the multi-mappability profile. Note the region indicated between the dashed lines that has low signal because of low mappability is filled with correction (3). 7 scale decomposition of the ChIP-seq signal is computed. Under each decomposition, the ERs with the corresponding local minima are shown (4). Connected window shows the processing performed for generating the SSERs at each scale. The mappability corrected signal is smoothed, the local minima are identified and the candidate ERs are formed (shown in grey), then candidate ERs are trimmed and filtered (shown in red) with respect to significance to identify the SSERS (shown in green.) SSERs for each scale is shown under the corresponding decomposition (5). The final set of ERs are formed by merging the SSERs (6).

Figure 2: Distribution of SSER pileup scale for CTCF, RNA Polymerase II (Pol2b), and several different HMs. The length scale is between 100 bps and 2.5 megabases as shown on X-axis. Y-axis shows the log frequency.

Figure 3: F-measure vs RPKM threshold for H3K36me3 ERs for GM12878 (a) and K562 (b). The F-measure versus RPKM cutoff with (red) and without (blue) mappability correction (c). Average of the overlap fractions of replicates for H3K36me3 and H3K27me3 ERs identified by each method (d).

Figure 4: The 2 dimensional normalized histogram of pileup scale versus log gene expression levels for the protein coding genes. The first component, stalled polymerase binding is indicated on the graph with "Stalled". The second component is indicated on the graph with "Elongating" (a). Aggregation of Pol2s2 signal around promoters of the genes that are bound by stalled polymerase ("Punctate-Low") and around the promoters of the genes bound with elongating polymerase ("Broad-High") (b).

Figure S1: The aggregation of the multi-mapability signal profiles over promoters, exon, introns, and random regions with at least 1 read that is mapped in any control dataset. 4 different read lengths, as indicated in the textbox in bottom right corner of each plot, are used: 36 bp, 50 bp, 76 bp, and 100 bps.

Aggregation over promoters is performed in a strand specific manner such that left side corresponds to the upstream of the promoter and the right side corresponds to the downstream.

Figure S2: Illustration of smoothing process. 4 scale decomposition of HM signal using 1 kb, 4kb, 16 kb and 64 kb long smoothing windows. The local extrema in the decompositions are highlighted with red markers on each smoothed signal. The three broad peaks can be identified with the 3 maxima at the decomposition scale with smoothing window length 64kb.

Figure S3: Illustration of gene-gene distance computation (a). Illustration of failure of detection when the smoothing window length (filter length) is longer than the twice the size of gene body (b). Illustration of overmerging when the smoothing window length is longer than twice the size of gene-gene distance (c). The cumulative distribution plot for intergene distance for genes that are at least 2.5kb apart (d). The cumulative distribution plot for gene lengths from protein coding gene annotations from GENCODE (e).

Figure S4: The plot for p-value versus fold change using the different p-value normalization window lengths. Fold changes 1.5 and 2 are indicated with vertical red dashed lines and p-value of 0.05 (log p-value of -3) is marked with horizontal red dashed line.

Figure S5: The cumulative distribution of the significant SSERs with respect to increasing γ. Y-axis shows the cumulative fraction of SSERs whose smoothed versus unsmoothed signal ratio is smaller than γ on x-axis.

Figure S6: The aggregation of multi-mappability signal over the MUSIC H3K36me3 ERs that are not identified by MACS (Marked with ‘MUSIC specific’). X-axis shows the distance from the mid-point of the region. The random regions are generated by translating the MAC specific regions to 3’ end by 10,000 bps.

References