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

# ABSTRACT

We present MUSIC, a method for identification of enriched regions (ERs) in the genome-wide read depth (RD) signal profiles from ChIP-seq experiments. The basic motivation behind MUSIC is twofolds: First, the systematic noise introduced by non-uniform read mapability 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 1000 base pairs to mega bases, 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 mapability 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 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) polymerase. These results suggest MUSIC as a promising tool for multiscale analysis of ERs in ChIP-Seq datasets. MUSIC is available for download at: http://archive.gersteinlab.org/proj/MUSIC/music.html

# INTRODUCTION

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 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 [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) in the 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 few megabases. Another interesting example is 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 large for many HMs. This makes it necessary to identify the ERs at different scales. A widely used method for identifying the 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 aspect. The selection of the predefined wavelet functions, however, are not justified for their choice. Second, the signal profiles contain systematic noise introduced to the read depth signal by the repeat regions with low mapability [9, 16], in the form of loss of signal. This noise causes discontinuities in the identified ERs. This becomes an important factor especially in the intergenic regions where a large ERs, which may mark a long regulatory region, gets broken 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 mapability correction at the nucleotide resolution so as to correct for the spurious loss of signal at the regions with low mapability. 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.

First, to evaluate the accuracy of ERs, we concentrate on H3k36me3, a well characterized HM that gets enriched on expressed gene bodies. Thus, we use the expressed gene bodies as gold standard. We compare the accuracy of ERs with several methods with respect to accuracy, in terms of consistency with expressed regions, and reproducibility between biological replicates. We show that ERs identified by MUSIC have higher F-measure and higher reproducibility compared to other methods. Next, in order to present a utility of the SSERs identified by MUSIC, we concentrate on Polymerase II ChIP-Seq dataset. Motivated by the basic observation that the stalled polymerase tend to show punctate enrichments (SSERs at small scales) and elongating polymerase to show broad enrichments (SSERs at higher scales), we computed the SSERs for Polymerase II ChIP-Seq dataset using MUSIC. Then using the SSERs, we estimate the length scale for polymerase binding at all protein coding genes. We demonstrate that the genes with less broad polymerase binding have significantly lower expression (close to 0) than the genes that are bound with more broad polymerase at the promoters. 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.

Paper is organized as follows. We first present MUSIC algorithm and lay out the steps of the algorithm. Then we present comparison of MUSIC with other ER identification algorithms. We finally present the joint processing of the polymerase data with gene expression levels.

# RESULTS

## MUSIC ALGORITHM

Figure 1 shows the flowchart of 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-mapability profile. The multi-mapability profile quantifies at each position, the number of reads that gets mapped non-uniquely. Therefore, for a position that is uniquely mapable, the multi-mapability value is 1. For the repeat regions, multi-mapability value increases. Fig. S1 shows aggregation of multi-mapability profile around different genomic elements. It should be noted that multi-mapability 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 mapability using the multi-mapability profile. The correction can be formulated as following:

where and are the uncorrected and corrected signal values, respectively, at position , is the value of multi-mapability profile at position , is the length of median filter utilized in correction which is by default set to 2000 base pairs, and is the average multi-mapability signal value over the exonic regions, which we identified as the most mapable regions in the genome (See Fig S1). In summary, for each position , MUSIC computes the median of the signal values at highly mapable positions (multi-mapability signal smaller than ) within vicinity of . Then MUSIC compares this value with the signal value at and assigns the maximum to the corrected value. The basic idea behind this correction is that since we know that low mapability causes decrease in the signal level, if the signal value at is higher than its vicinity, then it is highly likely that the mapability did not affect the signal value at . Otherwise, it is replaced by the median signal value at mapable 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 mapability corrected ChIP profile to compute multiscale decomposition of ChIP signal at multiple length scales (Step 4.) For this, MUSIC uses window lengths beginning with and ending at and performs sliding window based median filtering. The window length is increased multiplicatively between consecutive scales, thus, the window lengths form a geometric series:

where is the multiplicative factor between consecutive window lengths, which is set to 1.5 by default.

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 ), MUSIC discards this candidate enriched region (Refer to Methods.) This way, MUSIC avoids utilizing the regions identified from oversmoothed signal profiles.

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 ER. 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. MUSIC utilizes the SSERs for processing the enrichments in the signal.

### 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 from 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. 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 from the correction are thresholded with respect to 0.05 for identification of the significant ERs.

### 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, Polymerase 2, H3k4me1, H3k4me3, H3ke36me3, H3k27me3, and H3k9me3) from ENCODE project for K562 cell line with window length parameters bps, Mbp, and (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 for different datasets. As expected, CTCF, a punctate binding transcription factor, 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 H3k4me3, H3k4me1, and H3k36me3 have maxima at certain scales, which indicates that these HMs get enriched at specific length scales that are observed frequently. Finally 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.

## Comparison with Other Methods

In order to evaluate the accuracy of ERs, we compared MUSIC with 5 other algorithms that identify ERs from ChIP-Seq data: BCP [17], SPP [26], MACS [27], SICER [18], and PeakRanger [28]. We ran all the algorithms 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 ENCODE RNA-seq dashboard [30] and thresholded the expression levels of the transcripts and filtered the transcripts with low expression. The expressed transcripts are then merged to generate the gold standard set of expressed regions. Rather than selecting one expression threshold, we selected thresholds between 0 and 1 RPKM increasing with steps of 0.01 so as to evaluate the accuracy of peak 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).

### Accuracy Measures

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 b 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 mapability correction, we ran ER identification without mapability correction and computed the F-measure of the ERs. Fig 3c shows the F-measure versus RPKM threshold. Using mapability map significantly increases the accuracy of peak calls and shows the importance of utilizing the mapability correction in ER identification.

We also evaluated the reproducibility of the peaks generated by the peak callers. We used the replicates generated by ENCODE with the same HM datasets to assess reproducibility of peak calling. Figure 3d shows the average of fraction of the overlapping regions to the total coverage of each replicate. MUSIC has higher reproducibility for H3k27me3 and H3k36me3 than all other methods except for K562 H3k36me3 dataset, where BCP has slightly higher reproducibility than MUSIC. For K562, MUSIC has highest reproducibility for H3k27me3. For H3k36me3, BCP has slightly higher reproducibility than MUSIC. Overall, MUSIC has higher or comparable reproducibility with respect to other peak callers.

## Analysis of Polymerase II and Gene Expression Levels

Next, in order to illustrate a utility for the SSERs identified by MUSIC, 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., elongating or stalled [31, 32], the ChIP-Seq enrichment becomes more broad and more punctate for elongating and stalled polymerase, respectively. In addition, the stalled and elongating polymerase can be distinguished by comparing the detected amount of transcription at the polymerase binding.

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 bps, 2.5 Mbps, and . 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 elongating form of polymerase, Pol2s2, from 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.

# DISCUSSION

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 mapability correction for mediating the effects of the data. Mapability is an important aspect of peak calling from next generation sequencing data especially for identifying the broad domains of enrichment since the read depth profiles are highly correlated with the mapability map. We showed that MUSIC outperforms other methods in terms of accuracy of H3k36me3 peaks 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-mapability can be viewed as an impulse noise [33, 34].

Our results We believe that MUSIC is an important tool for multiscale identification and analysis of ERs in ChIP-Seq datasets.

# METHODS

We describe signal processing methodology underlying MUSIC in more detail.

## Mapability Correction

Given the read depth signal at each nucleotide position, MUSIC generates the per nucleotide multi-mapability signal and corrects for the mapability based loss of signal using following filtering:

Where is the signal value at nucleotide position , is the median of the set , is the value of the multi-mapability profile at the position , and is the window length used in mapability aware filtering. Using this filtering, MUSIC infers the signal values for positions with low mapability using the median of the values at nearby positions with multi-mapability signal lower than 1.2. We selected this value since it is the smallest multi-mapability signal profile value, i.e. most mapable, over exons and promoters as shown in Fig S1. We set the window length to 2000 bps, empirically. This window length depends on the distribution of length of the non-mapable region lengths. Different values did not seem to affect the results too much on 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.

## 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:

Where is the value of the decomposition at scale level for which the smoothing window length is , and is the mapability corrected signal profile. The window length is chosen from a geometric series with the factor 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.

## Identification of 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:

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

Where and are the sets of positions of minima and maxima of , respectively. The scale specific candidate enriched regions of are identified as the regions between the consecutive minima.

## 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 ratio threshold, In other words, MUSIC removes the candidate enriched region at scale , if

This test is designed as a simple and efficient check to evaluate whether the signal within the candidate region identified at the scale level 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 is set to 4.

## 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 of the Poisson distribution, MUSIC selects a threshold that satisfies 5% false positive rate:

Where represents the cumulative distribution function of , which is distributed as Poisson with mean For a region with start and end at positions and , respectively, the trimmed end coordinates are given as:

Where and 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 peak list.

## Enriched Region 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:

where represents the p-value for the peak starting at and ending at with the length of p-value window given by (Refer to p-value computation.)

## 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:

where and is the total signal on the positive and negative strand within the start and end coordinates of the ER, respectively.

## 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 () and control sample () that overlap with the region, then compute one tailed p-value as:

Where and are the normalized read counts for the region:

where is the length of the p-value computation window and refers to the p-value value for the peak. Larger values of increase the significance of regions (See Parameter Selection for Benchmarking). We perform multiple hypothesis correction by false discovery rate estimation (q-values) using the Benjamini-Hochberg procedure [25]:

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

## Multi-Mapability Signal Generation

MUSIC can generate per nucleotide multi-mapability signal profiles. For this it is required to have a read mapping program installed on the system. Currently MUSIC uses bowtie2 [35], 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-mapability RD signal profile. The regions with high signal corresponds to regions with low mapability. We generated multi-mapability profiles for hg19 genome assembly for read lengths of 36, 50, 76, 100, and 200 bps that are available for download with MUSIC.

## Parameter Selection for Benchmarking

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

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

Thus, should be small enough to ensure detection of the smallest enrichments that we expect to observe and 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 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 to 1000 bps. For choosing , 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 to bps.

The other parameters to set is . This parameter tunes the p-values of the SSERs and the final set of ERs. Generally, increasing increases the power of identification (See p-value computation) but also increases FDR. In addition, depending on the sequencing depth, can be used to avoid saturation of the peak calls. To select , assessed the p-values computed using different 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 values, we divided chromosome 1 into bins of 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 . It can be observed that as increases, the p-values corresponding to same fold change decreases. Our basic idea is to choose 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 to 1750 base pairs.

The remaining parameters, namely , and are set to their default values. These parameters define quality of the identified ERs and are set by trial and error.

## Accuracy Measures

For evaluating the accuracy of H3k36me3 peak calls, we computed sensitivity, positive predictive values:

Where is the coverage of ERs, is the coverage of expressed gene bodies and is the coverage of the overlap between expressed gene bodies and peaks. We combined these two accuracy measures to compute F-measure, computed as:

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

where and represent the coverage of the ERs identified from replicate 1, , and replicate 2, .

##  Datasets and Data Processing

We downloaded ENCODE ChIP-Seq from UCSC genome browser. The RNA-seq expression quantifications are downloaded from ENCODE RNA Dashboard. 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.

***References:***