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

MUSIC

1 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].

Deleted: *MUSIC:*

Formatted: Font: Not Bold, Not Italic

Formatted ... [1]

Formatted: Font: Not Bold, Not Italic

Deleted: *Correction and*

Formatted: Font: Not Bold, Not Italic

Formatted: Heading 1

Deleted: *:*

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Deleted: signal processing

Deleted:

Deleted: makes it challenging to process th ... [2]

Deleted: long enrichments into many small ... [3]

Deleted: many

Deleted: enrichment

Deleted: 100 bps to megabases, that

Deleted: Motivated by these,

Deleted: utilizes a smoothing and enhance ... [4]

Deleted: .

Deleted: smoothing

Deleted: using multiple windows of increa ... [5]

Deleted: smoothed and the candidate enri ... [6]

Deleted: all the scales are merged

Deleted: enriched regions. We compare th ... [7]

Deleted: and show that

Deleted: ERs compared to other methods. ... [8]

Deleted: with MUSIC. We show

Deleted: at the promoters. This observatio ... [9]

Deleted: evaluating the length scale

Deleted: enrichment for genomic signals ... [10]

Deleted: from Gerstein Lab web page

Formatted: Font: +Body (Calibri), 11 pt

Formatted: Font: Not Italic

Formatted: Heading 1

Deleted: *:*

Formatted: Font: Not Italic

Deleted: [cite] of many cell lines and tissues.

Deleted: HMs and TF

Deleted: analyze

Deleted: enrichments

Deleted: [7]. Depending on the target of t ... [11]

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

Deleted: [8]. For most HMs, enrichments are broad. For example, the repressive heterochromatin mark H3k9me3 show enrichments that can extend upto megabases.

Deleted: ,

Deleted: enrichments

Deleted: enrichments

Deleted: [9]

Field Code Changed

Deleted: 10–14

Deleted: There are two main challenges for identification of broad enrichments in the ChIP-Seq read depth signal profiles. First is that unlike transcription factor binding, HM and diffuse binding protein enrichments are observed at much larger length scales and the spectrum of enrichment lengths are different for different types of ChIP-Seq experiments. This makes it necessary to identify the enrichments 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 but in those approaches, the selection of the predefined wavelet functions 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 [15], in the form of loss of signal. This noise causes discontinuities in the identified enrichments. This becomes an important factor especially in the intergenic regions where a large region of enrichment, which may be a single element like a long repressed region, would get broken into many smaller regions. ¶
Many different approaches have been applied for identification of broad enrichments, which include change point identification within the formality of Bayesian inference (BCP, [16]), local island ... [12]

Deleted: (ERs)

Deleted: because of

Deleted: the

Deleted: a

Deleted: reveals

Deleted: as “blobs” that can be

Deleted: [20, 21]. ¶
The

Deleted: are then trimmed and filtered with respect to their significance compared to a normalized control. This procedure

Deleted: Each SSER corresponds to an enriched region at a certain scale. At

Deleted: the enrichments that are

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.

2 RESULTS

2.1 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:

Deleted: enrichments

Deleted: , which

Deleted: .

Deleted: ¶

Deleted: novel

Deleted: broadness level

Deleted: gene promoters.

Deleted: at their promoters

Deleted: independent

Deleted: binding

Deleted: the

Moved down [1]: RESULTS

Formatted: Font: Not Bold, Not Italic

Deleted: :¶

Deleted: motivation for

Moved (insertion) [1]

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Heading 2

Deleted: :

Formatted: Font: Not Italic

Deleted: There are two factors that motivate the novel methodology behind MUSIC: ¶
1. Mapability is an important aspect of read mapping and processing. For example, in the repetitive regions the number of uniquely mapable positions decreases significantly. This, depending on the parameters of the mapping algorithm, causes a systematic decrease of signal at repetitive regions and makes it impossible to evaluate whether a decrease in the signal is due to low mapability or a decrease in the modification levels. This becomes problematic especially in the intergenic and intronic regions which contain many repetitive regions. Consequently, the broadly enriched intergenic and intronic regions will be fragmented into many smaller enriched regions. It is worth noting that this problem is less severe for the punctate enrichments like transcription factor binding.¶
In order to characterize the mapability of different regions, MUSIC generates the genome-wide multi-mapability signal profile. For each position, ... [13]

Deleted: and

Deleted: smoothing

Deleted: analysis.

Deleted: PCR

Deleted: for both samples

Deleted:

Deleted: smoothed

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

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-mapability 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-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 i, MUSIC computes the median of the signal values at highly mapable positions (multi-mapability 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 mapability causes decrease in the signal level, if the signal value at i is higher than its vicinity, then it is highly likely that the mapability did not affect the signal value at i. 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 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.

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.) 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

Deleted: Compare

Deleted: with

Deleted: In summary, MUSIC first generates the mapability corrected signal profile from the ChIP-seq signal (See Methods), where

Deleted: 1.2)

Deleted: of them

Field Code Changed

Deleted: 23

Deleted: .

Deleted: that are used in multiscale filtering follow

Deleted: . This parameter

Deleted: Compared to the kernel based linear filters (which are also used in the wavelet based multiscale decompositions), median filtering has two advantages.

Moved down [2]: 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

Deleted: [24, 25]. ¶

Deleted: each

Deleted: smoothing window length

Deleted: RD

Deleted: enrichment feature

Deleted: , or most compact,

Deleted: enrichment feature

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.

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 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.

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, Polymerase 2, H3k4me1, H3k4me3, H3ke36me3, H3k27me3, and H3k9me3) from 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 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.

Deleted: p-value

Deleted: smoothing

Formatted: Font: Not Italic

Formatted: Heading 3

Deleted: decompositions

Deleted: evaluates the quality of

Deleted: concordance

Deleted: a

Deleted: p-value

Deleted: [26].

Formatted: Font: Not Italic

Formatted: Heading 3

Deleted: *the*

Formatted: Font: Not Italic

Deleted: ***Broadness***

Formatted: Font: Not Italic

Deleted: “

Deleted: ”

Deleted: enrichments

Deleted: many

Deleted: has a broader enrichment

Deleted: small

Deleted: generates the SSERs pileup signal profile; a genome-wide profile that is generated by counting

Deleted: Given the SSER pileup signal, we can translate the pileup signal to the corresponding smoothing window length for easier interpretation of the scale levels.

Deleted: illustrate this

Deleted: and several HMs

Deleted: MUSIC with smoothing

Deleted: lengths starting from

Deleted: to

Deleted: with

Deleted: generated

Deleted: signal

Deleted: signal

Deleted: In this plot, we show corresponding smoothing window length for the SSER pileup signal values on the x-axis.

Deleted: shows

Deleted: least broad enrichments

2.2 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).

2.2.1 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.

Deleted: *ER Identification*

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Deleted: *:*

Formatted: Heading 2

Formatted: Font: Not Italic

Deleted: In order to evaluate the accuracy of ERs, we compared MUSIC with 5 other algorithms that identify ERs from ChIP-Seq data: BCP [16], SPP [27], MACS [28], SICER [17], and PeakRanger [29]. We ran all the algorithms using H3k36me3 and H3k27me3 ChIP-Seq datasets for GM12878 and K562 cell lines from ENCODE project

Deleted: . H3k36me3 is known to mark the bodies of actively transcribed genes [30]. 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 [31] 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:*¶**

2.3 Analysis of Polymerase II and Gene Expression Levels

GOOD FOR MULTI SCALE

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 $l_{start} = 10$ 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 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.

3 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

Formatted: Font: Not Italic

Formatted: Heading 2

Deleted: Next, in order to illustrate a novel 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 [32, 33], the ChIP-Seq enrichment becomes more broad ... [14]

Deleted: broadness

Deleted: enrichments

Deleted: signal for smoothing window lengths

Deleted: to

Deleted: signal

Deleted: promoter

Deleted: dataset

Deleted: Dashboard

Deleted: joint distribution of broadness

Deleted: 4. For the broadness of binding v ... [15]

Deleted: can be seen

Deleted: very small

Deleted: upto

Deleted: pair scale level.

Deleted: have

Deleted: 6kbs

Deleted: divided

Deleted: -

Deleted: plane into 4 regions (log

Deleted: thresholded at

Deleted: 5, and

Deleted: level thresholded at

Deleted: bps) and

Deleted: for

Deleted: fall into each quadrant. The

Deleted: plot in the vicinity of the promot ... [16]

Deleted: binding

Deleted: promoters

Deleted: plots

Deleted: binding

Deleted: Interestingly, we also observed ... [17]

Formatted: Font: Not Italic

Formatted: Heading 1

Deleted: :

Formatted: Font: Not Italic

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.

4 METHODS

We describe signal processing methodology underlying MUSIC in more detail.

4.1 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:

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

Where x_i is the signal value at nucleotide position i, $\text{median}(\{x_i\})$ is the median of the set $\{x_i\}$, m_a is the value of the multi-mapability profile at the position a, and l_c 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 l_c to 2000 bps empirically. This window length depends on the distribution of length of the non-mapable region lengths. Different l_c 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.

Moved (insertion) [2]

Deleted: There are several limitations of MUSIC. Currently MUSIC cannot be run using only the ChIP sample, i.e., without a control sample. As explained in [34], the control experiments has become a standard part of any ChIP analysis. In addition, control experiment enables the ER identification algorithm to correct of read mapping artifacts and the sample specific genomics aberrations like CNVs. We also observed that the sequencing depth is an important factor for accuracy of signal processing framework. Although MUSIC performs well with default parameters on many datasets, we observed that the FDR increase with decreasing sequencing depth. For datasets with low sequencing depth (Smaller than 20 million reads), it may be necessary to re-evaluate the parameters using the parameter selection procedure.¶

Deleted: enrichments

Formatted: Font: Not Italic

Formatted: Heading 1

Deleted: :

Formatted: Font: Not Italic

Deleted: pipeline

Formatted: Heading 2 Char, Font: Not Italic

Deleted: :

Deleted: from observations.

4.2 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:

Deleted: :

Formatted: Heading 2 Char, Font: Not Italic

Deleted: smoothing for generating a

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

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 mapability 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].

Deleted: [22].

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.

4.3 Identification of Scale Specific Enriched Regions

Formatted: Font: Not Italic

Deleted: :

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:

$$x'^s_i = (x_i^s - x_{i-1}^s)$$

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

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

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

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

4.4 Comparison of Smoothed Signal in Candidate Enriched Regions

Deleted: :

Formatted: Font: Not Italic

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, γ. In other words, MUSIC removes the candidate enriched region $[i, j]$ at scale s, if

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

This test is designed as 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.

4.5 Candidate Enriched Region End Trimming using Poisson Distribution Model

Formatted: Font: Not Italic

Deleted: *:*

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:

Deleted: feature

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

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

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 peak list.

4.6 Enriched Region Trimming via p-value Minimization

Formatted: Font: Not Italic

Deleted: *:*

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:

Deleted: feature

Deleted: feature

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

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

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

4.7 Per Strand Concordance Test

Deleted: *:*

Formatted: Font: Not Italic

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.

4.8 P-value Computation and FDR Estimation

Deleted: *:*

Formatted: Font: Not Italic

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:

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

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 peak. Larger values of $l_{p_{val}}$ 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]:

Field Code Changed

Deleted: 26

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

where N_{ERs} is the total number of enriched regions and i is the rank of the 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.

4.9 Multi-Mapability Signal Generation

Deleted: *:*

Formatted: Font: Not Italic

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.

4.10 Parameter Selection for Benchmarking

Formatted: Font: Not Italic

Deleted: :

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 $2 \times 2^{13} = 16000$ 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 peak calls. 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.

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

4.11 Accuracy Measures

Deleted: :

Formatted: Font: Not Italic

For evaluating the accuracy of H3k36me3 peak 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 peaks. We combined these two accuracy measures to compute F-measure, computed as:

Deleted: peaks

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

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

Deleted: H3k4me3 peaks, we used all

Deleted: promoters (TSS

Deleted: transcript ±2500 bps). For these

Deleted: a slightly different approach to compute sensitivity

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

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

Deleted: PPV:

4.12 Datasets and Data Processing

Deleted: $Sensitivity = \frac{\#(S \cap P)}{\#(S)}$¶

$PPV = \frac{\#(P \cap S)}{\#(P)}$¶

Where $\#(S), \#(P), \#(P \cap S)$ represent number of active promoters, number of peaks, and number of peaks that overlap with active promoters, respectively.¶

Formatted: Font: Not Italic

Deleted: :

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:

Formatted: Font: Not Italic

Page 1: [1] Formatted Ozgun 3/9/2014 8:38:00 PM

Title, Border: Bottom: (Single solid line, Accent 1, 1 pt Line width, From text: 9 pt Border spacing:)

Page 1: [2] Deleted Ozgun 3/9/2014 8:38:00 PM

makes it challenging to process the enrichments. Especially for the diffuse enrichments this noise

Page 1: [3] Deleted Ozgun 3/9/2014 8:38:00 PM

long enrichments into many smaller enrichments.

Page 1: [4] Deleted Ozgun 3/9/2014 8:38:00 PM

utilizes a smoothing and enhancement procedure to correct for

Page 1: [5] Deleted Ozgun 3/9/2014 8:38:00 PM

using multiple windows of increasing lengths, i.e. multiple scales. Essentially, at each smoothing scale, the details of the RD signal

Page 1: [6] Deleted Ozgun 3/9/2014 8:38:00 PM

smoothed and the candidate enriched regions are revealed as "blobs". The candidate enriched regions are filtered with respect to significance compared to a normalized control signal. The significantly enriched regions

Page 1: [7] Deleted Ozgun 3/9/2014 8:38:00 PM

enriched regions. We compare the accuracy to

Page 1: [8] Deleted Ozgun 3/9/2014 8:38:00 PM

ERs compared to other methods. Next, we analyze

Page 1: [9] Deleted Ozgun 3/9/2014 8:38:00 PM

at the promoters. This observation indicates that the multiscale analysis can be utilized

Page 1: [10] Deleted Ozgun 3/9/2014 8:38:00 PM

enrichment for genomic signals and can reveal important biological information.

Page 1: [11] Deleted Ozgun 3/9/2014 8:38:00 PM

[7]. Depending on the target of the ChIP-seq assay, the length scale of enrichments can be very different for different experiments, which changes the mode of enrichment analysis. For example, for transcription factor binding, the enrichments are concentrated to

Page 2: [12] Deleted Ozgun 3/9/2014 8:38:00 PM

There are two main challenges for identification of broad enrichments in the ChIP-Seq read depth signal profiles. First is that unlike transcription factor binding, HM and diffuse binding protein enrichments are observed at much larger length scales and the spectrum of enrichment lengths are different for different types of ChIP-Seq experiments. This makes it necessary to identify the enrichments 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 but in those approaches, the selection of the predefined wavelet functions 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 [15], in the form of loss of signal. This noise causes discontinuities in the identified enrichments. This becomes an important factor especially in the intergenic regions where a large region of enrichment, which may be a single element like a long repressed region, would get broken into many smaller regions.

Many different approaches have been applied for identification of broad enrichments, which include change point identification within the formality of Bayesian inference (BCP, [16]), local island identification and clustering (SICER [17]), local thresholding and merging (MACS), using local Poisson statistics to identify broad enrichments (SPP), and wavelet based smoothing and identification of enriched regions (WaveSeq [18]), which is also applied to analysis of ChIP-chip datasets [19].

Page 3: [13] Deleted Ozgun 3/9/2014 8:38:00 PM

There are two factors that motivate the novel methodology behind MUSIC:

1. Mapability is an important aspect of read mapping and processing. For example, in the repetitive regions the number of uniquely mapable positions decreases significantly. This, depending on the parameters of the mapping algorithm, causes a systematic decrease of signal at repetitive regions and makes it impossible to evaluate whether a decrease in the signal is due to low mapability or a decrease in the modification levels. This becomes problematic especially in the intergenic and intronic regions which contain many repetitive regions. Consequently, the broadly enriched intergenic and intronic regions will be fragmented into many smaller enriched regions. It is worth noting that this problem is less severe for the punctate enrichments like transcription factor binding.

In order to characterize the mapability of different regions, MUSIC generates the genome-wide multi-mapability signal profile. For each position, this profile contains the number of reads (of certain length) that can map from any other position in the genome. In order to gain a perspective on the statistics of multi-mapability signal, we aggregated the signal over different elements. This reveals, as expected, that the protein coding exons and promoter regions show the highest mapability (See Figure S1). The multi-mapability signal is utilized by MUSIC in correction of effects of mapability.

2. The length distribution of ERs for broad enrichments usually have a large variance. This makes it necessary to identify the enrichments for a spectrum of scales. For example, for HMs like H3k36me3, H3k27me3, the ERs can extend from several kilobases to hundreds of kilobases. On the other hand, for HMs like H3k4me3 and H3k27ac, which marks the gene promoters and enhancers, the ERs are around kilobases in length. Another interesting example is the RNA Polymerase II, whose enrichments can extend from less than a kilobase to hundreds of kilobases.

Motivated by these facts, we designed MUSIC to account for the effects of mapability and to be scale sensitive. In essence, MUSIC first corrects the RD signal from ChIP experiment for the mapability. MUSIC then computes the multiscale decomposition of the signal by smoothing the signal with multiple smoothing window lengths. In the process of smoothing, fine details in the signal are removed and the broad enrichments are revealed as "blobs" in the smoothed signal, which are detected as the regions between consecutive local minima of the smoothed signal. The blobs in smaller scales are merged with each other as the scale level is increased, where the broader enrichments are revealed. The smoothing window lengths used in decompositions, therefore, specify the length spectrum of the identified enriched regions [22].

These regions are then filtered with respect to significance computed in comparison with the control signal to generate the scale specific enriched regions, SSERs. The SSERs at small scales represent the small enrichments in the signal and the vice versa for SSERs at large scales. With multiple scales, MUSIC can detect SSERs within a spectrum of lengths that can be tuned by adjusting the starting and ending scale levels to be processed by MUSIC.

Page 7: [14] Deleted **Ozgun** **3/9/2014 8:38:00 PM**

Next, in order to illustrate a novel 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 [32, 33], 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.

Page 7: [15] Deleted **Ozgun** **3/9/2014 8:38:00 PM**

4. For the broadness of binding we translated the SSER pileup signal to the corresponding smoothing window length.

Page 7: [16] Deleted **Ozgun** **3/9/2014 8:38:00 PM**

plot in the vicinity of the promoters of all the genes in 4 different parts of the plot.

Page 7: [17] Deleted **Ozgun** **3/9/2014 8:38:00 PM**

Interestingly, we also observed that the low expression and broad polymerase bound promoters have significantly high Pol2s2 signal levels that