Response to reviewers for “Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability Corrected Multiscale Signal Processing Framework”

Response Letter

-- Ref1 – General Remarks --

Reviewer Comment	This manuscript presents an improved method for identifying significantly enriched regions from ChIP-seq experiments, called MUSIC. The current methods for analysis of ChIP-seq data suffer from 2 major limitations. First, none of them account for repetitive DNA sequence or other regions that can affect ChIP-seq peak calling due to mappability issues. Second most of the available programs struggle with accurately calling peaks for broadly distributed histones, such as H3K36me3 and H3K4me1. The authors present compelling evidence to show that MUSIC offers a solution to both of these issues, and is a significant improvement over most previous methods. The paper is well written and the data are presented clearly and the conclusions are well supported by the results. I only have minor comments, which can be addressed in a revised version. Overall, I’m very pleased with this manuscript and I look forward to using MUSIC as my preferred ChIP-seq analysis package.
Author Response	We thank the reviewer for the constructive comments. We address the comments of the reviewer in a point by point manner below.
Excerpt From Revised Manuscript	

-- Ref1.1 – Smoothed Signal Track Outputs --

Reviewer Comment	MUSIC outputs a list of peaks and corresponding enrichment scores. This is fine, but it would be very helpful if the program also outputted a track of the processed (smoothed) data that could be loaded onto a Browser, so that the degree of smoothing could be visualized. A “before” and “after” version of these tracks would be particularly useful.
Author Response	We agree with the reviewer that the visualization of the data is an important aspect of assessing the results and would make MUSIC much more useful. For this, we added an option to MUSIC to write the smoothed signal tracks (for each track) in bedGraph format with the output ER’s. These can easily be

	loaded to a genome viewer to visualize the files locally or uploaded to other genome browsers like UCSC Genome Browser. We updated the manuscript to present that the smoothed tracks can be outputted as bedGraph files.
Excerpt From Revised Manuscript	Section 2.1.1: For each ER, MUSIC also computes a summit (See Methods Section 4.10), and a trough in the ER. The summits represent the point of strongest binding/modification in the ER and troughs represent the point where there is a depletion of signal, which may represent the nucleosome-free regions (See Methods Section 4.10.) Finally, in order to visualize the processed tracks, MUSIC has an option to save the smoothed signal profiles at each decomposition scale in bedGraph format that can be loaded to a genome browser.

-- Ref1.2 – Troughs in the Signal --

Reviewer Comment	Second, I’m struggling a bit to assess how much of the smoothing removes details in a given signal that can sometimes be informative. For example, when ChIP-seq data is plotted in aggregate, it’s clear that transcription factors often bind in the “trough” of a bimodal histone-peak, corresponding to the nucleosome free region. Does MUSIC smooth this out? How often are such regions called as a single histone-peak with MUSIC? Do the other programs (MACS, etc) tend to split these regions into 2 called peaks? A comparative analysis would be helpful.
Author Response	The referee raised an important point. We agree that the troughs can be significant since they may mark the nucleosome free regions where regulatory factors (like TFs) can interact with DNA. The identification of these troughs are especially hard since the decrease in signal can either be related to either real decrease in histone modification levels, or simply a decrease in the mappability. Our inspection of the ChIP-Seq signal profiles, however, shows that the dips in the signal are most frequently caused by a decrease in mappability. In fact, it is very hard to distinguish between the non-mappable troughs and mappable troughs. Therefore, MUSIC currently merges these regions together because it aims at identifying the ER as a complete unit, for example, for H3K36me3 marks, the whole gene body. The other peak callers do not generally merge these regions and tend to oversegment the signals. In order to quantify the amount of decrease in the signal with respect to mappability versus the real nucleosome-free regions, we concentrated on H3K36me3 and identified the regions that were identified in ERs identified by MUSIC that were not identified in MACS ERs. Then we performed aggregation of the multi-mappability signal and observed that there is a very significant decrease in the mappability compared to a set of control regions (See Supplementary Figure S6). This result suggests that the regions that MUSIC merges (but MACS does not) are significantly enriched in regions that have low

	mappability. However, since we believe that this is a very important point for analysis of punctate histone marks like H3K27ac and H3K4me3, we also added a functionality to punctate ER identification mode of MUSIC to discover and report the largest trough in the signal in each ER that has good mappability (at the level of exonic mappability) requirement in each identified ER. We believe this is a valuable addition to the functionalities of MUSIC. We added a paragraph in the manuscript summarizing the above result.
Excerpt From Revised Manuscript	Methods Section 4.10: For DNA-binding protein ChIP-Seq data, e.g. transcription factors, MUSIC reports the location of the highest signal level within the ER as the summit of the signal, which can be used as the binding position. An important consideration in ER identification is the identification of valleys (or troughs) in the signal. For example, the troughs in H3K4me3 and H3K27ac ERs may correspond to the nucleosome free regions in promoters and enhancers, respectively, where the transcription factors can interact with DNA and regulate transcription. Therefore, identification of the troughs (in addition to the summits) is an important piece of additional information for each ER. Our analysis, however, shows that much of the troughs in ChIP-Seq signal is caused by the decrease in the mappability of the genome (See Fig S6). MUSIC reports one trough position in each peak by determining the smallest position within the top two tallest peaks such that the average multi-mappability around the trough is smaller than exonic multi-mappability (m_e). No troughs are reported if there is only one summit in the ER.

-- Ref 2 – General Remarks --

Reviewer Comment	Harmanci et al. present a new algorithm called MUSIC to identify enriched regions in the ChIP-seq experiments. MUSIC aims to correct the systematic noise introduced by non-uniform read mappability and devices a smoothing strategy to merge fragmented enriched regions in ChIP-seq experiments. Furthermore, they applied MUSIC at multiple length scales to automatically consider both the narrow and broad peaks. They compared the performance of MUSIC with several peak-finding algorithms on H3K36me3. Using RNA-seq signals as a gold standard, they showed that MUSIC achieved better F-measures than the existing methods. In particular, they investigated the RNA ploymerase II binding ChIP-seq data and showed distinct expressions of genes with different length scale of binding peaks. For a computational method, it provides some new features such as smoothing peaks using read mappability and considering multiple length scales. The major concern is that the performance assessment is not as thorough as it can be.

	Some details of the parameters set should be provided as well.
Author Response	We thank the referee for constructive comments. The referee’s main concern is that MUSIC is not compared to other methods in terms of more punctate events such as H3K4me3 and TF’s. We updated the benchmark section with the requested comparisons in terms of addition of methods and more datasets using comparison with different metrics. We also reorganized the parameter selection section in order to clarify the presentation and added a section to present the parameters of other methods used in comparisons. We address the comments of the reviewer in a point by point fashion below.
Excerpt From Revised Manuscript	

-- Ref 2.1 – ChIP/Input Normalization Factor Computation --

Reviewer Comment	In page 4, "The MUSIC computes a scaling factor using linear regression between the ChIP and control signal profiles. The slope of the regression is used as normalization factor for control." It is unclear how this regression was done. A brief explanation would be helpful for readers to understand how this was done.
Author Response	We agree with the reviewer that this is an important point in the paper that needs to be clarified. For this, we added Methods Section 4.1 to explain the computation of the input normalization factor in full detail.
Excerpt From Revised Manuscript	Methods Section 4.1: It is necessary to normalize the control signal profile with respect to ChIP-Seq profile because the read depths can be different. For each chromosome, MUSIC first divides the chromosome into 10,000 base pair bins then computes the total ChIP-seq and control signal in each window. Finally, it estimates the normalization factors as the slope of the minimum squared error estimate of the slope: $\rho = \underset{\rho'}{\operatorname{argmin}} \left\{ \sum_{i} (w_i - \rho' \cdot c_i)^2 \right\}$ where w_i and c_i represent the total signal in ith bin for ChIP and control samples, respectively. The normalization procedure aims to match the background signal level in the ChIP sample to the control sample.

-- Ref 2.2 – Parameter Selection --

Reviewer Comment	In page 4, how are the parameters of l(start) and l(end) determined? Also, how are the default values of gamma and tau determined? In Methods, it is noted that these parameters are set by trial and error. What is this "trial

	and error" procedure? How to judge what parameter values perform better? Is there any general guidance of choosing the values? Does the choice impact the results?
Author Response	We agree with the reviewer that the selection of parameters is an essential part of MUSIC's workflow and should be clarified. We summarize the parameter selection procedure here briefly: When selecting l(begin) and l(end) for broad marks (H3K36me3, H3K27me3), we utilize the fact a median filter of length l removes all the features of length smaller than l/2 within it (See Methods Section 4.11 and Supplementary Figure S3). Given the distribution of gene-gene distances, we selected l(end) to avoid overmerging of consecutive ERs. Similarly, using the gene length distribution, we selected l(begin) to avoid missing small ERs. Following the discussion in Methods Section 4.11, we set l(begin) to 1000 bps and l(end) to 16000 bps. For punctate marks like H3K4me3, the enrichments are expected at scales of at most several kbs (around the promoters) thus we set l(end) to 2000bp's. We set l(start) to 100bp's so as not to miss any small ERs. For most transcription factors, there is almost no concept of multiscale processing since the binding is assumed to happen at a specific motif and the ERs extend most several hundred base pairs (See Figure 2). For TFs, we use l(start)=100bps and l(end)=200bps. Tau is estimated (for all the modes) as the threshold that satisfies 5% false positive rate under the null model that the reads are distributed as a Poisson distribution with a mean estimated from the one megabase sliding windows across the genome, as defined in equation in Methods Section 4.6. Thus it is not a free parameter. Gamma is the threshold of the smoothing statistic that is introduced to avoid overmerging of the ERs by oversmoothing of the signal in the decomposition. In principle, this oversmoothing test is a proxy for a statistical test that would compare the distribution of signal in the regions at smaller scales that get merged into regions at the higher scales and would determine if there is a significant shift in the signal levels: We expect that as the signal is smoothed, it will diffuse out and become smaller. We realized, however, that this would be computationally too costly and implemented the test with thresholding the simple test statistic presented in Methods Section 4.5. For illustrating how different gamma values change the smoothing levels, we plotted the distribution of p-values of regions with respect to the smoothing statistics for each SSER in a large scale

	decomposition (See Supp Fig. 5). Following these, we decided gamma=4 (where we capture around 90% of the SSERs) is a reasonable value for thresholding the smoothing ratio statistic. For selecting sigma, the interscale multiplicative factor, we evaluated different values and observed that above 2, MUSIC starts missing too many SSERs. We chose 1.5 as a reasonable value for sigma. We reorganized the Methods Section 4.12 itemizing the above points to more clearly explain the selection procedure for these parameters.
Excerpt From Revised Manuscript	Methods Sections 4.12: **Selection of l_{begin} and l_{end}** For punctate marks (like H3K4me3 and H3K27ac), MUSIC is set to run at the smaller scale spectrum with $l_{begin} = 100$, $l_{end} = 2000$. This way MUSIC aims at identifying small ERs and at identifying the enrichments at a reasonable expected length range of several kilobases. For transcription factors, for which point binding events occur at almost single base pair resolution, MUSIC is set to run at very small scales with $l_{begin} = 100$, $l_{end} = 200$. It should be noted that the utility of mappability correction and multiscale decomposition is most effective for identification of more broad ERs. **Selection of l_{pval}** The punctate histone marks (like H3K4me3) and transcription factors (like CTCF) have much more punctate ERs than broad histone marks. In addition, the ERs are observed at a much smaller spectrum of length scales (See Fig 2). Therefore, the procedure that we used for broad marks with large scale spectrum is not very suitable for these marks. For example most of the transcription factor peaks are smaller than 500 base pairs. Therefore, for CTCF, we set l_{pval} to 500 bps. Similarly, H3K4me3, which marks the promoters, extend over the promoters of genes and has ERs of several kilobases. For H3K4me3, we set l_{pval} to 2000 bps. **Selection of γ** γ is the threshold on the ratio of the maximum of the smoothed signal and the unsmoothed signal on an SSER. This parameter enables MUSIC to avoid overmerging segments by comparing the signal level in the smoothed signal and the original signal. To visualize the effect of changing γ on the identified SSERs, we computed the SSERs for H3K36me3 and H3K4me3 marks for K562 cell line. We then identified the computed the smoothing ratio (as given in Section 4.5) for each SSER. Then we plotted the cumulative distribution of all the SSERs with respect to the smoothing statistic, which is shown in Fig. S5. It can be seen that for H3K4me3 that the distribution is much more skewed toward smaller γ than H3K36me3, which is expected since H3K4me3 has much narrower ERs than H3K36me3. To be as inclusive as possible, we choose γ=4 (98% and 90% of the SSERs respectively for H3K4me3 and H3K36me3 pass the smoothing statistic test) as a suitable parameter to balance the tradeoff between being inclusive in the identified SSERs and overmerging the ERs. **Selection of σ** The final parameter to set is, σ, which is the multiplicative factor between the consecutive

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

-- Ref 2.3 –Comparison to TFs and DHSs--

Reviewer Comment	When evaluating the performance of MUSIC, the authors selected H3K36me3 and used RNA-seq signals as the gold standard. Clearly MUSIC outperformed the other methods. This is not completely unexpected because MUSIC tends to identify long enriched regions. What about a comparison on signals with narrow peaks of TFs and DHS? There are many TF ChIP-seq available and their motifs are also known. It would be interesting to see whether MUSIC recovers peaks of these TFs containing the motifs.
Author Response	The referee brings up an important point. We have added a new benchmarking section to the manuscript (Sections 2.2.2 and 2.2.3) for comparing the methods with respect to their accuracies for the transcription factor CTCF. MUSIC performs as one of the best methods calibrated as a function of the fraction of top peaks containing known CTCF sequence motifs.
Excerpt From Revised Manuscript	Section 2.2: For comparing the accuracy of methods with respect to identification of ERs for point binding factors like transcription factors, we used the transcription factor, CTCF, which has a well-studied motif associated with it. We identified the ERs using each method (using the TF peak calling mode when available). As a gold standard we utilized the motif datasets from the ENCODE project [34]. In order to measure accuracy, we ranked the top 2000 peaks and then computed the fraction of ERs with a motif within 150 base pairs of the summit of ERs. The results are summarized in Table S2. We observed that MUSIC, SPP, MACS, and DFilter perform very similarly followed by other methods.

-- Ref 2.4 –Zinba, F-seq, DFilter--

Reviewer Comment	There are several recently developed methods that should be included for comparison, such as Zinba, F-seq and DFilter. These methods also provide flexibility of detecting peaks at different length.
Author Response	We thank the reviewer for pointing out these methods. We have added the mentioned methods (ZINBA, F-Seq, and DFilter) in our ER identification comparisons, updated the results, and highlighted the manuscript. In the benchmark comparisons, the parameters were selected from the documentation for each method that was best suited for the comparison. We added one section to the manuscript (Methods Section 4.13) on the details of the options (including parameters) used to run the each of the other programs in the benchmarking. The results show that

	MUSIC performs favorably compared to all other methods for the analysis of the broad marks.
Excerpt From Revised Manuscript	Section 2.2:In order to evaluate the accuracy of ERs, we compared MUSIC with 8 other algorithms that identify ERs from ChIP-Seq data: DFilter [26], ZINBA [27], F-Seq [28], BCP [17], SPP [29], MACS [30], SICER [18], and PeakRanger [31]. A detailed list of the parameters used to run each method are presented in the Methods Section 4.12 and 4.13. Methods Section 4.13: The most recent versions of the tools are downloaded from the website and documentations are followed for running the tool in the correct mode. 1. BCP [17]: For histone marks (H3K36me3, H3K27me3, and H3K4me3), we used BCP_HM tool with command line options: -f 200 -w 200 -p 0.05. For CTCF dataset, we used BCP_TF tool with command line options: -e 10 -p 0.00000001 2. PeakRanger [31]: For histone marks, we used ‘ccat’ option for broad peak calling. For CTCF peaks, we used ‘ranger’ option. 3. ZINBA [27]: For broad histone marks (H3K36me3, H3K27me3), we used the unrefined ERs from ZINBA with ‘broad’ flag on as explained in the documentation. For H3K4me3 and CTCF peaks, we used the refined peaks with ‘broad’ flag turned off. 4. F-Seq [28]: For histone marks and CTCF, F-Seq is run in default mode. 5. SICER [18]: For broad marks, SICER is run with the with command options: hg19, w=200, fragment_size=150, 0.74, g=600, FDR=0.01. For CTCF, SICER is run with smaller gap size of g=200. 6. SPP [29]: For broad marks, SPP is run in broad mode using get.broad.enrichment.clusters(…). For CTCF, the peak calling mode is run using find.binding.positions(…). 7. DFilter [26]: For H3K36me3 and H3K27me3, DFilter is run with command line options ‘-nonzero -bs=200 -ks=40 -std=2’ then removed the peaks that has score smaller than 2. For H3K4me3, DFilter is run using ‘-bs=100 -ks=100 -dir -std=2’ then peaks with score smaller than 6 are removed. For CTCF, we ran DFilter with ‘-nonzero -bs=25 -ks=50 -pm=300 -std=2’. 8. MACS [30]: For broad marks, MACS is run with options ‘--broad -g hs’. For CTCF, MACS is run with ‘-g hs -q 0.01’.

-- Ref 2.5 –H3K4me3 and TF Comparisons--

Reviewer Comment	F-measure is only one simple statistics and a better performance on H3K36me3 alone is not sufficient to demonstrate that MUSIC is superior to other methods. There are additional criteria for performance comparison such as comparing active promoters overlapping with H3K4me3 peaks called by the methods, percentage of peaks located within 50bp of motifs for TFs et al.. A thorough comparison can be found in the DFilter and Zinbe papers.
Author Response	We thank the reviewer for the suggested comparisons. We have updated the benchmark section with comparisons of H3K4me3

	(i.e. enrichment of active TSS'es around identified H3K4me3 peaks) and TF experiments (i.e. enrichment of motif around 150 bp of the identified peak summit as was used in ZINBA paper as well as several previous papers). The results demonstrate that MUSIC performs favorably for H3K4me3 peaks and performs comparably to the best performing methods in terms of motif enrichment.
Excerpt From Revised Manuscript	Section 2.2: For comparing the methods for identification of punctate ERs with smaller length scales, we first chose to compare the methods on the H3K4me3 histone modification, which marks the promoters of active genes. As a gold standard, we utilized the promoters of the active genes (RPKM > 0.5) as the gold standard positives. We identified the ERs that have at least 5% overlap with the promoter region (2 kb region around the annotated transcription start site). For this comparison, we sorted the top 20,000 ERs with respect to the score reported by each method then computed the overlap of the ERs with active promoters. Starting from the top ERs, we plotted fraction of active promoters that are identified correctly versus fraction of ERs that overlap with active promoters. These are shown in Fig 3c and 3d, respectively for the K562 and GM12878 cell lines. MUSIC performs favorably compared to other methods, followed by DFilter and SICER. For comparing the accuracy of methods with respect to identification of ERs for point binding factors like transcription factors, we used the transcription factor, CTCF, which has a well-studied motif associated with it. We identified the ERs using each method (using the TF peak calling mode when available). As a gold standard we utilized the motif datasets from the ENCODE project [34]. In order to measure accuracy, we ranked the top 2000 peaks and then computed the fraction of ERs with a motif within 150 base pairs of the summit of ERs. The results are summarized in Table S2. We observed that MUSIC, SPP, MACS, and DFilter perform very similarly followed by other methods.

-- Ref 2.6 –Study by Knijnenburg *et al.*--

Reviewer Comment	There was a recent paper published by Knijnenburg *et al.* Nature Methods, 11, 689-694, 2014 that provides a multiscale representation of genomic signals. Can the authors comment on that study and compare the multiscale features of MUSIC with Knijnenburg study?
Author Response	We thank the reviewer for pointing this relevant paper, which was published very close to our submission. The *Knijnenburg study* utilizes a linear Gaussian filtering based multiscale decomposition to summarize and visualize the genomic signals. An important difference in the methodologies is that MUSIC performs the multi-mappability correction before performing the multiscale decomposition. We observed that this increases accuracy of the identified ERs significantly as shown in Figure 3e. *Knijnenburg et al* makes the assumption that the genomics signal is never smaller than the mappability signal, which is not hold true for most of the public datasets we have analyzed. When Gaussian decomposition is utilized, the lowly mappable regions will see a high decrease in signal levels and this may substantially distort the tree based segmentation used in their decomposition. MUSIC aims to correct for these in the mappability correction stage. In addition, the non-linear median filtering based decomposition

	used in MUSIC is better tuned to identify the edges in the signal. In order for the tree based decomposition in *Knijnenburg et al* to successfully work, there are several constraints that has to be met (See Babaud et al). For example, the scale space has to be sampled densely and completely (starting from very bottom scale going to very top scale) so that the full tree can be generated. In MUSIC, however, there are no constraints on the scale levels (l(begin), l(end), and sigma) and the user can change these parameter freely. Similar arguments can be made for l(begin) and l(end). We have added a reference to the study and updated the Conclusion Section to briefly reflect our discussion above.
Excerpt From Revised Manuscript	Conclusion Section: For example, a recent study [39] uses linear Gaussian filtering based multiscale decomposition to compute multiscale representations of genomic signals. The non-uniform mappability of the genome should be expected to substantially affect the representation since mappability is utilized in a post-processing step after representation is computed unlike MUSIC, where the mappability correction is performed before decomposition is computed.