Response Letter for Manuscript ID BIOM2020344P - 'Efficient Detection and Classification of Epigenomic Changes Under Multiple Conditions'

Pedro L. Baldoni, Naim U. Rashid, Joseph G. Ibrahim

20 February, 2021

Reviewer 1

The authors faithfully referenced the relevant literature. However, it may be better to cite the recent review from a comprehensive comparison of tools for differential ChIP-seq analysis, and discuss the reason for selecting the candidate DPCs in the benchmark part.

> We thank the reviewer for this constructive comment. We cite in our paper the benchmarking article 'A comprehensive comparison of tools for differential ChIP-seq analysis' (Briefins in Bioinformatics, 2016). However, we authors are not aware of a more recent comprehensive review of DPCs algorithms in the literature. As suggested we add the following sentence to our real data analysis section 'We compared the genome-wide performance of the presented model with the widely used DPCs ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR, a list that covers a variety of algorithmic approaches for both broad and short data sets.' We believe that such a list covers the most used algorithms for DPC in thie field of bioinformatics.

In the third challenge at "Third, the analysis of ChIP-seq data...", the author may also need to discuss the potential sequence GC content bias except for local total reads abundance bias, since GC bias may lead to false-positive peak call or differential peak call as mentioned in Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data.

> We acknowledge the issue of GC-content bias in the referenced paragraph. We cite the work of Teng and Irizarry (2017) that proposed a correction method for GC-content bias in ChIP-seq analysis. Their method is applicable to any peak calling algorithm that measures the enrichment of reads in bins, such as our method. Later in the article, we point out to the reader that our method is able to account for GC-content bias in the same manner as suggested by Teng and Irizarry (2017) by using their count-normalizing denominators in our model as offsets. We describe this process in detail in Section B1.3 of the Web Appendix.

The author may also need to consider citing the MACS2 GitHub page along with the Zhang et al.2008 since the first MACS version does not support broad peak calling.

> This has been addressed. In the second paragraph of Section 5, we reference the current GitHub repository of MACS2 https://github.com/macs3-project/MACS.

1. As mentioned before, GC bias is a nonlinear covariate that may lead to false-positive peak, mixNBHMM includes a GLM-based framework for adjusting local read abundance bias, could the mixNBHMM handle extra GC content in the bins, would correction of GC bias increase the specificity and sensitivity of mixNBHMM?

> As mentioned above, our method is able to account go GC-bias via model offsets. These model offsets can be obtained by methods such as the one proposed by Teng and Irizarry (2017). We present a simulation study in Section B1.3 of the Web Appendix showing that epigraHMM leads

to higher sensitivity/specificity when accounting for GC-bias even after read count normalization using the count-normalizing denominators from Teng and Irizarry (2017). However, for the broad marks that we analized in our manuscript, we did not observe a substantial effect of GC-content on read count distribution and did not observe any substantial improvement in accounting for GC-content in our model via offsets when utilizing the method from Teng and Irizarry (2017), a method that was developed only for narrow marks characterized by sharp peaks. In fact, the authors do not recommend the use of their method when analyze epigenomic marks with broad enrichment. We did not explore alternative methods for GC-bias correction for broad marks as we believe this task is beyond the scope of our work.

2. What's the exact form of weighted NB regression used in the parameter estimation part? What is X and Y in the regression?

In the main text, we refer the reader to Section B5 of the Web Appendix. To estimate the model coefficients during the t^{th} iteration of the EM-algorithm, one needs to find the optimal values $\beta_1^{(t)}$, $\beta_3^{(t)}$, $\lambda_1^{(t)}$, and $\lambda_3^{(t)}$ that maximize the quantity $Q_1\left(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)}\right) + Q_2\left(\boldsymbol{\delta}, \boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)}\right) + Q_3\left(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)}\right)$. This problem is analogous to estimating parameters in a weighted NB regression model, since Q_1, Q_2, and Q_3 can be seen as weighted log-likelihood functions from a NB generalized linear model with means and dispersions parametrized as described in the Section 3.1 of the main article. Exact forms of the Q-functions are presented in Equation B.4 of Web Appendix B5.

3. What are the BIC measurement for Table 1 and corresponding web tables? Are these results showing the optimal component?

Results presented in Table 1 pretain to the full model with all possible combinatorial patterns. This is the 'optimal' model in the sense that the simulated data contains all possible combinatorial patterns of enrichment and background among the condition. The point of Figure 1 is to show that our efficient estimation method via fast RCEM leads to unbiased parameter estimates. Together with Figure 2, Table 1 raises the point that increasing the number of technical or biological replicates leads to better results in terms of sensitivity/specificity of peak calls as well as in the sense of asymptotic efficiency of parameter estimates. We leave the discussion of model selection via BIC in our model to Section B2 of the Web Appendix where we present a simulation study.

4. In section 4.1.1, there is no clear evidence in Table 1 and corresponding web tables showing the trend that higher SNR has a lower bias as there are only two levels of SNR in one replicate cases, more levels of SNR may be needed for this, also what is the SNR level for the real data from ENCODE?

We thank the reviewer for point this out. We believe that comparing the relative bias of parameter estimates across different SNR levels is difficult because different SNRs are associated with different parameter values (some larger than others) in our simulation study. Although we did observe an improvement in performance in terms of model sensitivity/specificity of peak calls in higher SNR data sets, the main driving quantity for the reduction of parameter bias is the number of replicates included in our model. For this reason, we rephrase the mentioned sentence in our paper as "Depending on the number of conditions, the observed relative bias and the range of the reported percentiles tended to decrease as more replicates were included in the analyses. This effect was particularly significant in scenarios with four conditions with respect to parameters β_3 and λ_3. In scenarios with higher number of conditions, these results highlight the importance of experimental replicates to achieve precise parameter estimates.". The estimated SNR level from ENCODE was 3.21 (ratio of mean counts between peak and background regions) for H3K37me3. This is now reported in our manuscript. We present the estimate SNR value for other scenarios in our Web Appendix (Web Tables 2-3)

5. In real data benchmarking, some histone marks have a clear classification of broad or narrow peaks marks, others are not, for example, H3K27ac and H3K4me3 are the predictors for super-enhancer or super promoter, the genome-wide peaks are constituted with a mixture of

broad and narrow peaks, mixNBHMM could handle the independent narrow or broad peaks, could mixNBHMM handle the heterogenous peaks situation?

We believe our method is able to handle such situations. Note, for instance, in Figure 4D of our manuscript that we report differential peak calls from H3K4me3 in which both narrow and broad differential peaks are deemed to be differential by epigraHMM at FDR level of 0.05 (see rightmost peaks). ChIPComp, for instance, missed the broad peak on the right while called as differential all narrow peaks on the left. In fact, note that the left most region of Figure 4D is comprised by a super enhancer region, in which several peaks are present in the cell line Hepg2. Our method was able to call this super enhancer region as differential while also covering the broad peak on the right of the panel.

6. What is the exact difference between nominal FDR and observed FDR? Is nominal FDR the FDR estimated in the alpha or total FDR mentioned in the method part?

Nominal FDR is the total FDR as described in the last paragraph of the Methods section. The observed FDR is the actual proportion of false discoveries (i.e. the proportion of genomic windows incorrectly classified as differentially enriched out of the total number of genomic windows classified as differential). Once a nominal FDR level is specified, an α cutoff level can be determined and used to classify genomic windows into differential and non-differential. Then, the observed FDR can be straightforwardly calculated as the proportion of false discoveries. We make this distinction clear in the last paragraph of the 'Methods' section as well as in the first paragraph of the 'Simulation Results' section.

7. For the benchmark measurement of real data, how about spearman correlation and LFC metrics for the differential broad peaks calling? Using differentially transcribed genes from LFC (>2) of H3K36me3 sounds like a cyclic logic for differential H3K36me3 peak call, authors may need to consider using the matched differentially expressed genes as the gold standard to evaluate the differential transcribed genes called from differential H3K36me3 activity. In addition, gene ontology at least should be provided for interpreting the biological significance of those differential peaks covered genes, are there any liver (HepG2) or Cervical (HelaS3) related terms predicted?

We believe that Spearman correlation and LFC are not good metrics to compare peak callers with broad marks for two reasons. First, methods that tend to call narrow peaks within brod regions of enrichment will exhibit an unrealistic advantage in performance in terms of LFC. This is so because these methods tend to call as narrow peaks any genomic window exhibiting excessively high read count enrichment (or, LFC in differential enrichment). For broad marks, interest lies not on these single punctate windows but instead in broad regions of enrichment. Second, and due to a similar reason, these methods will also exhibit unrealistic advantage in performance in terms of Spearman correlation because broad regions of enrichment are much noisier than punctate regions and, thefore, read count correlations calculated from broad peaks will be attenuated in comparison to correlations calculated from narrow and punctate windows.

Regarding the use of 'matched differentially expressed genes as the gold standard to evaluate the differential transcribed genes called from differential H3K36me3 activity', we did not observe a sufficient correspondance between gene expression levels and H3K36me3 enrichment for the cancer cell lines analyzed in our paper. For instance, we have found several instances were genes were expressed but no enrichment of H3K36me3 was observed. Below, we present some of these problematic cases obtained from the UCSC Genome Browser. Note in Figure 1, for example, that both cell lines Helas3 and Hepg2 were not enrichment for H3K36me3 in the rightmost region of the plot while the associated gene body was clearly expressed. Note in Figure 2, for example, the lack of agreement between the transcription level of the reported gene and the enrichment for H3K36me3 in cell line Hepg2. Due to these reasons, we decided not to use gene expression levels when benchmarking H3K36me3 as suggested by the reviewer.

Regarding the biological significance of our method, Figure 5E of our paper now presents an analysis

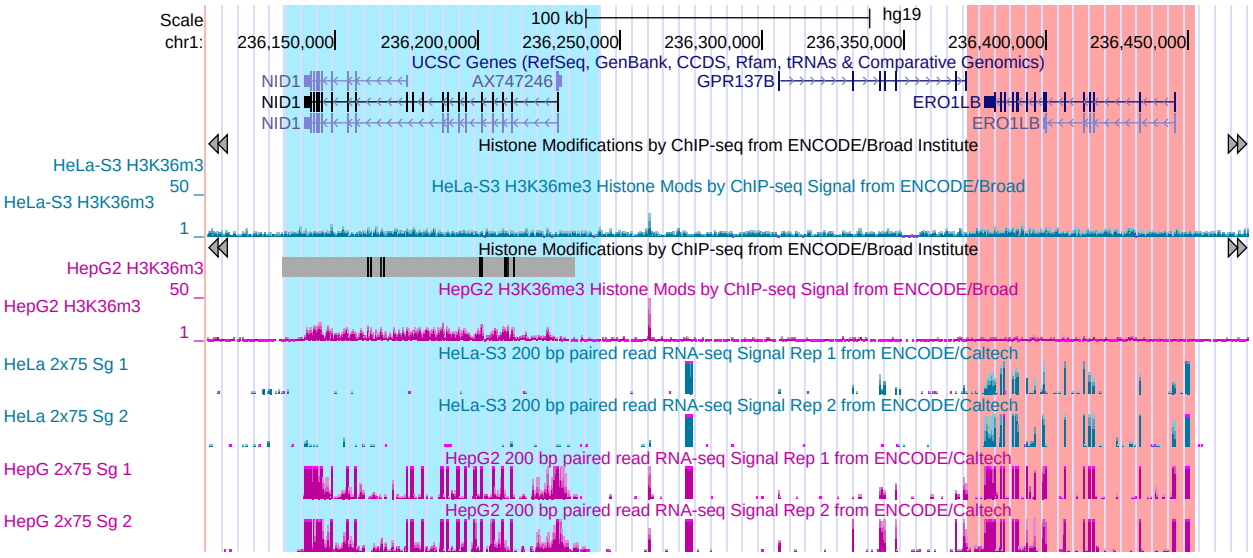

Figure 1: First example of nonconcordant gene expression level and H3K36me3 ChIP-seq enrichment in Helas3

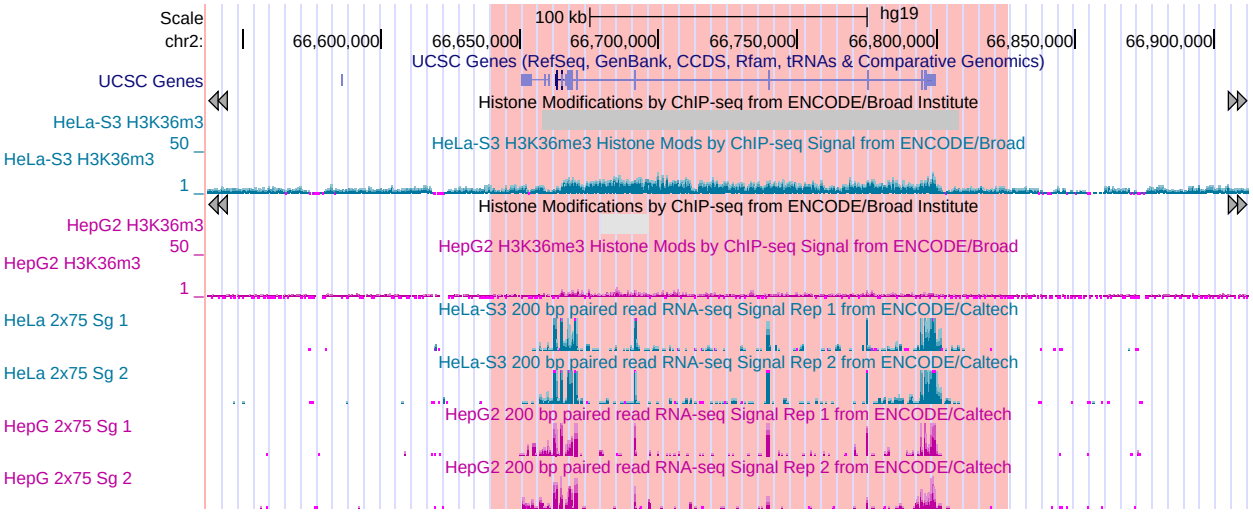

Figure 2: Second example of nonconcordant gene expression level and H3K36me3 ChIP-seq enrichment in Helas3

of known marker genes for cancer cell line Helas3. Specifically, we show that all known marker genes of Helas3 had associated differential enrichment for the H3K36me3 mixture model component. The presented differential associated enrichment level shown in Figure 5E is the average mixture model posterior proportion associated with the mixture component that model the enrichment of H3K36me3 alone. These results highlight the benefit of the chosen modeling approach that allows not only the detection of differential regions but also the correctly classification of differential enrichment pattern, in this case the gene transcription level associated with H3K36me3.

Authors may also need to consider improving the following regions of the paper to make it be more accessible to the readers (minor).

1. In the introduction part, "To quantify local genome activity" may be changed to "To quantify genome-wide epigenome activity".

I have changed "To quantify local genome activity" to "To quantify local epigenomic activity on a genome-wide scale".

2. enhancer EZH2 is not a proper term for the description of the chromatin regulator since enhancer is a kind of cis-element, this should be corrected as Enhancer of zeste homolog 2 (EZH2).

This has been addressed

3. For the layout of figures, web figure 1 could be Figure 1, and Figure 1 could be web figure 1, the workflow along with the method in the main text could make things clearer.

We thank the reviewer for this comment. We added Web Figure 1 to our main manuscript and it is now shown in Figure 1 of the paper.

4. In section 3.1, the "N x 1 vector" and "NM x 1 vector" may be replaced with N vector and N x M matrix.

This has been addressed

5. In section 4.2.1, how would we deal with the failure case from RSEG? Were they removed in the benchmark?

This has been addressed and made clear in text. For RSEG, the metrics were calculated based on the instances where it ran without issues (71 out of 100 simulations).

6. In the supplementary documents (Web appendices), the figure or table labels can be prefixed with Web or Supplementary.

This has been addressed. The tables and figures of the supp. are already indexed with 'Web' in front of it (e.g. Web Figure 1).

7. B22 section occurs twice in the web appendices.

This has been addressed

8. On page 11 of main text, "t th step" of the EM algorithm can be "t th iteration" of the EM algorithm.

This has been addressed

9. In section 4.1, genome length can be changed to window number.

This has been addressed

10. How many threads are the software using for Figure 2C, 3E, 4E? Are they using similar memories?

All analyses in the paper were conducted using a single CPU per task without multithreading. Figures 4E and 3E now exhibit the peak memory comsumption by each method along with the

total computing time. We observed that in broad marks (500bp) our method was among the most efficient ones, requiring only 5GB of memory and 30 min for a genome-wide analysis of 4 samples (2 conditions, 2 replicates per condition). In narrow marks (250bp), our method required nearly 10GB of memory for a computation time of about 1 hour. We believe that these are reasonable quantities and quite comparable to other efficient algorithms. For example, the widely used method DiffBind required nearly as much memory even though it is a method that relies on peaks called a priori (MACS2 in our analysis) and had a much lower performance in broad histone modification marks.

11. In 1st line of the 16th page, two technical replicates seem not right, usually, we computed the differential peak regions from biological replicates.

This has been addressed. The analyzed ENCODE data pertain to isogenic replicates (such as https://www.encodeproject.org/experiments/ENCSR000AOL/) because they are derived from cell lines of the same biosample/donor. These are neither technical nor biological replicates (https://www.encodeproject.org/data-standards/terms/), and we make this distinction in the referenced text.

12. The main text appendix for Q function and forward and backward probability can be merged into the B5 in web appendices.

This has been addressed

13. In Figure 5D. RNA-seq tracks may be added to see if it is a highly expressed gene set near the differential peak regions.

We have added the gene expression level of the covered gene body in Figure 5F of our manuscript.

14. In B3, certain combinatorial "patters" should be "patterns".

This has been addressed

15. In short differential peak calling, what might be the reason that mixNBHMM is worse than other methods?

We believe our method is quite comparable to other DPC in short differential peak calling. These results are shown in Figure 4 of our paper as well as in the Web Appendix (Web Figures 25-27). In terms of LFC and Spearman correlation, our method performed quite comparably to ChIPComp, a widely used method for narrow marks. Examples in Figure 4B, 4D, and 4F clearly shows that our model is correctly calling differential peaks similar to other methods.. We did see, however, a slight increase in memory consumption of 5GB when analyzing such marks due to the higher resolutions of 250bp. However, such an increase is still well within the available resourses of most high performance computing clusters and even in some personal laptops nowadays.

16. BIC features as the model selection is a novel way for inference of chromatin segmentation, which is worth discussing in the discussion.

As mentioned in our paper, we added a discussion in Web Appendix regarding the model selection via BIC. Web Figures 8-10 include results from our simulation study exploring the presented strategy to choose the optimal number of combinatorial patterns via BIC. Due to the lack of space, we have added the following sentence to our Discussion "In situations where certain combinatorial patterns are rare, prunning of specific combinatorial patterns is possible via model selection via BIC (Web Appendix B2)".

17. Biological replicate insight is really important, which is the leading factor in differential peak calling, this provides a guideline for future ChIP-seq experiment design about the necessity of generating biological replicate for differential epigenome study, which is worth mentioning in the discussion.

We thank the reviwer for the constructive comment. Due to the lack of space in our paper, we added the following brief sentence to our discussion "Our results highlight the importance of the

inclusion of technical/biological replicates in the analysis of epigenomic data".

Reviewer 2

1. There are also other methods that can also perform combinatorial binding pattern detection such as jMOSAiCS (https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r38). dPCA (https://www.pnas.org/content/110/17/6789), SignalSpider (https://academic.oup.com/bioinformatics/article/31/1/17/2366199). The proposed method should be compared to these methods. In addition. HMM-based peak callers have been widely developed including histoneHMM (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4347972/); HPeak (https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-369); https://academic.oup.com/bioinformatics/article/24/20/2344/258202. chromHMM (http://compbio.mit.edu/ChromHMM/). These similar methods diminishes the novelty of mixNBHMM, which is also based on HMM and for the same peak-calling/genome segmentation purpose. Moreover, the difference in terms of methodology among all these HMM-based peak-calling methods and mixNBHMM are not clear.

TO DO.

2. mixNBHMM does not consider background control sample in peak-calling, which may result highly false positive. While other methods such as ChIPComp consider both IP and control. I am curious how the simulations are set for these methods considering controls, and how the control samples are analyzed in the real ENCODE data analysis.

Input control samples were utilized by the methods that are designed and capable to do so. These methods include ChIPComp, THOR, diffReps, and DiffBind. We added the following sentence to the second paragraph of Section 5 of our paper: "Results presented in this section pertain to the analysis of two cell lines, namely Helas3 and Hepg2. Analyses were adjusted by input control samples in methods that are designed to do so (ChIPComp, DiffBind, diffReps, and THOR)". Similarly, input controls were also generated in our simulation study and accounted for in these algorithms. Throughout our analysis, we did not observe any significant improvement in performance by accounting for input control effect in differential peak calling. This effect was also reported by others (Lun & Smyth, 2015).

3. According to Figure 2A/Figure 3A, mixNBHMM has significantly underestimated FDR. While other methods such as RSEG has better FDR control. The results are contrary to the claim mixNBHMM outperforms other competing methods.

In Figure 2A, the observed FDR is shown for all methods and for all assessed nominal FDR levels, for all methods. The observed FDR shown in Figure 2A is quite close to the nominal FDR chosen (note vertical lines on the panel and corresponding nominal FDR points). In Figure 3A, no FDR was or is currently reported. Instead, we show 1-specificity on the x-axis. In Figure 3B, we show the FDR for all methods. As we can see, the observed and nominal FDR levels reported by our method were quite close to each other (0.05, 0.07, 0.08, 0.09, 0.10). In comparison, RSEG exhibited observed FDR levels of 0.23, 0.63, 0.69, 0.74,0.80 for nominal FDR levels of 0.01, 0.05, 0.10, 0.15, and 0.20. We thank the reviewer for this comment but we do believe our results are superior than the suggested FDR control of RSEG.

4. In the real data analysis for the ENCODE data, there is no gold standard for DPC. However, nearest differential expressed gene among multiple conditions can be used to evaluate DPC because DPC and nearest DE genes could have the same change direction and magnitude.

We thank the reviewer for the constructive comment. As we replied to the first anonymous reviewe, we did not observe a sufficient agreement between gene expression levels for the analyzed cell lines in our paper (see Figures 1 and 2 above). In fact, not all the analyzed marks in this paper are associated with the expression level of the neared gene. For instance, H3K27me3 and EZH2 are two epigenomic marks known to exhibit read enrichment in intergenic regions without

any nearby gene body. We believe that this analysis would be suboptimal to the one presented in this paper.

5. The importance of mixNBHMM lies on its ability to detect different combinatorial pattern of enrichment across cell lines. However, the biological meaning of combinatorial pattern is not clear, which limits the general application of mixNBHMM. A biological interpretation of the finding results are needed to demonstrate the potential biological application of the method, e.g. finding novel TF cofactors.

As we discuss above, Figure 5E of our paper now presents an analysis of known marker genes for cancer cell line Helas3. Specifically, we show that all known marker genes of Helas3 had associated differential enrichment for the H3K36me3 mixture model component. The presented differential associated enrichment level shown in Figure 5E is the average mixture model posterior proportion associated with the mixture component that model the enrichment of H3K36me3 alone. These results highlight the benefit of the chosen modeling approach that allows not only the detection of differential regions but also the correctly classification of differential enrichment pattern, in this case the gene transcription level associated with H3K36me3.

Reviewer 3

The real data analysis section is another simulation study. The authors assumed that they know the truth about the data and compared the result they get from mixNBHMM to the results they get from other competing pipelines. For instance they provided criteria ROC (Figure 3A), false positive rates, etc. If it is a real "real dataset" then nobody knows the truth and the authors must remove all comparisons to other methods and provide a real data analysis. If the authors know the truth, then they must move this section to simulation study and provide another real dataset and analyze it without comparing to other methods.

We appreciate the comment for our reviewer 3. Although Figure 3 of our paper does pertain to a simulation study, our paper also includes an extensive section on the analysis of real samples from the ENCODE project. We compare our method with competing algorithms in both simulated and real data scenarios. We utilized data driven metric to benchmark our algorithm with other and observed superior performance throughout our analysis. Section 5 of our paper presents results from the analysis of the ENCODE project.