The comments of reviewer 1 are in blue.
Our comments are in purple and italics.

Changes to the software or software documentation are in bold italics, with yellow highlight.
Changes in the manuscript are in bold italics.
Actual text changes in the manuscript are in bold green.

Reviewer: 1

Comments to the Author
I have no further major comments.

I think that the HMM with ZINM emmissions and the classifier idea are nicely presented. This is the take home message for me from this manuscript and all my comments regarding this have been addressed.

However, even though I feel there were attempts to address my comments about benchmarking, this part is still confusing and not helpful. If I have to guess a reason, I'd say it's because the authors did not properly research this niche peak finding field before writing the manuscript and that made it especially difficult to communicate potential pitfalls to them. For example, I had to specifically point to the ENCODE IDR webpage in revision round 2.

Minor comments:
Since I think dwelling further on this benchmarking stuff is not productive for this paper, I have the following suggestion:

restructure the paper slightly by removing some of those comparisons you report in the tables (h3k4me3 and pol2 are especially confusing - In Fig1 in kumar et al. Nat Biotech, all peak callers seem to perform significantly better than what it's reported here, including MACS). This will make for a shorter more concise and easier to read paper.

We have removed the comparison involving H3K4me3 and Pol2. ***Sections 3.2.2 and 3.2.3 have been removed. Tables 3 and 4 have been removed. Where applicable, the text has been updated accordingly.***

If you decide to keep all the comparisons though, you can double check the following:

(1) Do you intersect H3K4me3 and PolII with the single nucleotide annotated TSS? If that's the case, you should intersect with a window around the TSS. h3k4me3 (pol2 chip peaks for the matter) are not expected to be exactly at the TSS.

(2) You count GRO-seq reads in a window around TSS..on the right strand?

The last two points no long need to be addressed.

(3) did you resize all the peaks for the different peak finders before running your comparisons? peak finders with bigger peaks will have more motifs, intersect more TSSs...etc. just because the peaks are biggerThis is more or less a standard practice for benchmarking TF peaks the way you do

We have resized the windows to 500 bp. See below for changes in the text.

(4) I pointed you before to the IDR webpage: how many peaks do you use as input to IDR, did you check the IDR recommendation? I think using more than 150k is not recommended.

There are clear recommendations when using the SPP caller to aim for 150k to 300k peaks, (but such recommendations are not made for MACS for instance). We have limited the initial peaks of JAMM according to these guidelines. ***We have added the following part to the third paragraph of section 2.4.***

For JAMM, we took the top ranking 300,000 peaks as input for IDR. (...) For MACS, peaks scoring lower than 0.05 were kept. For JAMM, the top n peaks in the joint discretization were kept, where n is the number of peaks scoring lower than 0.05 in separate discretizations. For the CTCF benchmark, all peaks were resized to 500 bp from the center of the window.

Table 2 and Figure 7 have been updated accordingly.

(5) As I mentioned before: If you will report JAMM's default results without any filtering, indicate "unthresholded" (otherwise, check the JAMM manuscript, how do they recommend you run IDR+JAMM?)

For H3K36me3 (section 3.2.2) the IDR is not recommended. ***We have added the following sentence to the caption of Figure 6.***

Also note that JAMM is used "unthresholded", as IDR is not recommended for broad signal.

Following all these changes, ***the second paragraph of section 3.2.1 now reads as shown below****.*

Table 2 shows that for most tools, the F_1 score (the harmonic mean of precision and recall) is between 0.34 and 0.41. The exception is JAMM, achieving significantly higher precision than the other tools at the cost of recall. On this dataset, the performance of Zerone is fair, with a good balance between precision and recall.
