We would like to thank the reviewers for taking the time to review our software and manuscript, and we value their feedback. For the sake of clarity, and in order to highlight the changes implemented in this new version of the manuscript, we used the following color code.

The comments of reviewer 1, reviewer 2 and reviewer 3 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.

We have made a major modification that was not directly requested by the reviewers, but that will prove useful to answer their concerns and to improve the quality, the documentation and the reproducibility of our work. ***We have created a Docker container where the users can run Zerone, reproduce the benchmark and view exactly how the training of the SVM was performed.*** *This allows us to 1. move the technical detail away from the main text, 2. ensure that our benchmark is reproducible, 3. provide a working computer environment where Zerone runs natively, and 4. stand for higher reproducibility standards in bioinformatics. We use the Docker container (https://hub.docker.com/r/nanakiksc/zerone/) to answer several of the points below, and we invite the reviewers to downoad it if they want to verify that their concerns are properly addressed. The following sentence has been appended to the abstract.*

The information to reproduce the benchmark and the figures is stored in a public Docker container that can be downloaded from https://hub.docker.com/r/nanakiksc/zerone/.

We will add the exact commit ID when the manuscript is suitable for publication.

Reviewer: 1

**General Comments:

The authors present Zerone, a chip-seq peak finder. There are two main advances offered in my opinion. First, the negative multinomial distribution is used to model read counts thereby accounting for replicate dependencies. In that way Zerone can handle ChIP-seq replicates. Second, the authors offer an innovative solution to ChIP-seq quality control namely to learn what model parameters represent high quality datasets using a supervised framework.

The program is very fast, easy to use and requires no user adjusted parameters. However, the manuscript and figures need major restructuring and the validation efforts can be improved. Generally, the manuscript can be restructured to highligh better the main advances offered by Zerone.

Major Comments:

1-The classification step in Zerone is used not to give a quality score for each peak but to give a decision on the whole peak finding process. This should be mentioned more clearly in the text, it's an important difference compared to IDR.

We have added the following sentence at the end of Section 2.3.

Unlike quality control methods based on individual peaks (such as the IDR for instance) the quality control implemented in Zerone is 'all-or-none', i.e. the profile is rejected or accepted as a whole.

It should say also what datasets precisely were used in the training and what criteria that authors used to assign positive and negative cases as those things might be somewhat subjective. The SVM effectively learned what the

authors think is a good discretization and not some ground truth. This point can be considered a disadvantage and should be discussed.

We fully agree with this point, the lack of ground truth is indeed one of the main limitations of our machine learning approach. In the case of experimental ChIP-seq data there is no obvious way to make an objective classification and the best workaround is to allow the users to retrain the classifier, as suggested below. For lack of time, we have not yet implemented this feature, but we acknowledge that the labels of the training sets are subjective. The 'Quality control' section now starts as follows.

We used a machine learning strategy to identify discretization failures. The true status (success or failure) of experimental ChIP-seq data is not known because success is partly subjective and because there is no gold standard for protein binding in live cells. We prepared an experimental data set where we labelled the output of Zerone as positive (success) or negative (failure) based on empirical criteria (see associated Docker container for detail). The definition of success is thus subjective, but the training is performed on representative data.

The next paragraph now starts as shown below.

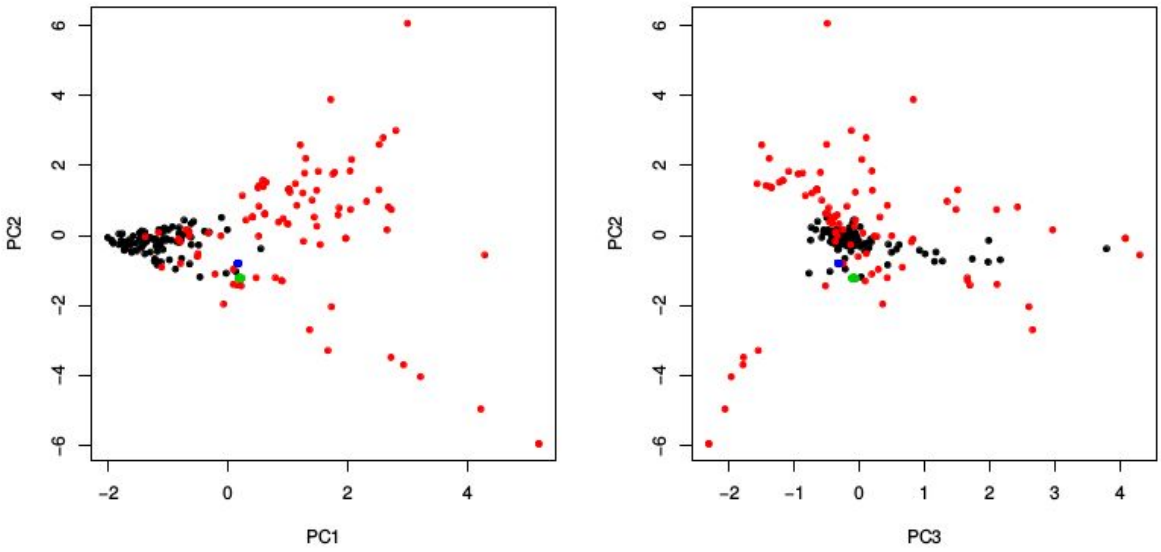
We discretized 144 replicated ChIP-seq experiments together with their respective input control (see associated Docker container). We labelled the output of the discretization as positive (91 cases) or negative (53 cases), based on visual inspection and on the available literature about the chromatin features.

The list of ChIP-seq profiles used for training the SVM is available in the Docker container.

Would it have been better to train the SVM on properly simulated datasets?

Good point. Even though simulated data can be used to establish a ground truth, it does not make the definition of success and failure less arbitrary. An additional limitation of simulated data is that experimental biases such as protocol and experimenter effects cannot be simulated properly (we are not aware of reliable methods to do so). Their characteristics are unknown beforehand and have an impact on the summary statistics used to train the SVM. Those are the reasons why we chose to use real data sets representing different types of ChIP-seq signal.

That said, we agree that simulated data is useful for calibration purposes. We have tried to generate ChIPs-seq data sets with ART (https://www.niehs.nih.gov/research/resources/software/biostatistics/art/), the R script chip-seq-simu.r (http://www.gersteinlab.org/proj/chip-seq-simu/), and the Bioconductor R package ChIPsim (https://www.bioconductor.org/packages/release/bioc/html/ChIPsim.html). In our case, it is critical to generate paired replicate files, together with a paired control file. Since none of the software mentioned above offers this option, we wrote a custom R script for this purpose. We generated a collection of simulated data sets by means of a Markov chain that used parameters that were learned from actual data. For instance, we ran Zerone on an actual CTCF data set consisting of one control and two replicate ChIP-seq profiles, and used the model parameters learned by Zerone's HMM to simulate new, random CTCF data sets of similar characteristics. Unfortunately, this approach is far from providing the diversity observed in real data sets, as shown in the illustration below.



This Principal Component Analysis reproduces the data of Figure 2 from the manuscript. Each dot represents a profile in the space of the features used to train the SVM after projection on the first three principal components. The black dots are positive examples, the red dots are negative examples, the green dots are resampled profiles of CTCF (narrow peaks), the blue dots are resampled profiles of H3K36me3 (wide domains). 100 blue dots and 100 green dots are plotted on the figure, but they cluster very close to each other.

In conclusion, it is not straightforward to sample the space of ChIP-seq profiles with simulated data, and because the definition of success would be equally subjective (which is the point under discussion), we decided to not include simulated data sets.

2-Validation and comparisons to other peak finders:

- The recommended parameters should be used for each dataset/peak finder (MACS: --broad-cutoff 0.1 –broad for H3K36me3 and JAMM –r region for the same dataset).

*The benchmark has been repeated with these options (note that MACS' option --broad-cutoff 0.1 is already default when --broad is set) and the **manuscript and figures have been updated accordingly**. This information has also been added to the Docker container. Note that the performance of JAMM has worsened slightly with these parameters.*

- Polymerase is also expected to be found in inter-and intra-genic regions. Therefore, precision is not really convincing. A somewhat more plausible alternative would be to define active genes and inactive genes from gene expression (maybe using GROseq) and use them as positive and negative TSS sets. The same goes for H3K36me3.

*Thanks for this suggestion. We have used a GROseq data set to separate active from inactive TSS and have **added this information to Table 3 and Table 4** (see the request to include H3K4me3 below). **We have updated the text accordingly**.*

- More than one dataset should be used for each example to make sure Zerone's behavior is consistent.

We haven't addressed this point. A look at Figure 7 of the manuscript shows that it takes about 3 days to discretize a ChIP-seq profile with BayesPeak. Adding just two profiles per chromatin feature represents over 24 days of computation for BayesPeak alone (we run benchmark on the same machine and sequentially in order to

avoid interference or competition for computer resources), which we have not done for the sake of time. If, however, the editor feels that the current benchmark is unacceptable as is, we would be happy to run more tests, understanding that it would take some time.

- How does Zerone behave with narrow histone modification datasets? For example H3K4me3. Because those are narrower than H3K36me3 but broader than CTCF, it might be the case they pose problems in terms of replicate reproducibility depending on the resolution zerone operates on

We have added H3K4me3 to the benchmark. ***Figure 7 now contains the performance results of BayesPeak, JAMM, MACS and Zerone on this data set. Section 3.2.3 now describes this benchmark.***

- JAMM gives a high number of peaks because it expects the user to use IDR for quality control (which should be indicated when discussing benchmarking results) and MACS can be relaxed to give a high number of peaks too. So why not compare MACS+IDR and/or JAMM+IDR versus the quality control step by Zerone. This quality control step is an important highlight of the paper, so it should be compared to what's available.

Thanks for the suggestion: the improvement in precision of MACS is significant at almost no cost for for the recall. This is a more honest benchmark for MACS. IDR is a peak-based QC score, whereas Zerone implements a profile-based (all-or-none) correction. We not see how to compare the two (but we strongly agree that IDR should be used for the benchmark). JAMM was included as the only discretizer able to merge replicates. Running replicates separately and then using another tool for the merge would defeat the purpose of including JAMM in the benchmark.

Table 2, Table 3 and Table 4 now show the result of running MACS with relaxed peak definition following by IDR trimming. The text has been updated accordingly.

- Definitions of accuracy and f-score the authors used should be stated. How were negative sets defined for CTCF? Peaks without a motif?

The legend of Table 2 and Table 3 has been changed in order to give more formal definitions of precision, recall and F1 score.

3- The zerone-related methods should be expanded to include a better explanation of the preprocessing steps that zerone does (how does it handle fragment length issues in chip-seq, what bin size does it use..etc. This is completely absent) as well as a more detailed description of the HMM in the main text. The intuition behind using a negative multinomial distribution in relation to handling replicates could be better explained.

We have fused the first two sections of the methods in a single section called 'Model and parameter estimation'. We have added a methods section called 'ChIP-seq preprocessing'. We have added formal definitions of the distribution and the HMM*. Paragraph 4 of section 2.1 now gives the formula of the emission model. Paragraph 7 now gives the formula of log-likelihood to be maximized by the Baum-Welch algorithm. Paragraph 9 of the same section reads as follows.*

The fitting process resolves conflicts between replicates. Say that a ChIP-seq peak is present in only one of them; the signal will be locally high only for this replicate and low for the others. Because of the conflict, the local log-likelihood will be low for all the possible states but there will still be an optimum that corresponds to the 'least unlikely' state. The final call depends on whether the weight of evidence is higher for the presence of the peak or for its absence. If the conflict is strong, the confidence in the final call will be weak, which can lead to a rejection of the profile as a whole if such cases are too frequent (see below).

The section 'ChIP-seq preprocessing' reads as follows.

Mapped reads are binned in fixed-step windows (default 300 bp) by their mid-point and PCR duplicates (i.e. reads mapping to the same location in the same orientation) are removed. The window size should not be smaller than the sonication fragment length and it should be set so that there are on average more than 3-4 mapped reads per window. Zerone decompresses gzipped input files if required. There is no upper limit to the number of input files to discretize simultaneously, but there must be at least one negative control and one ChIP-seq experiment.

4- The "other" methods should be expanded and consolidated in one section…how did the authors scan for motifs? Which genome annotations were used? what was the definition of promoter regions…etc.?

The methods section now contains a single subsection called 'Benchmark data sets and conditions'. We have added more specific information about the files and genome annotations. *The full detail is available on the Docker.*

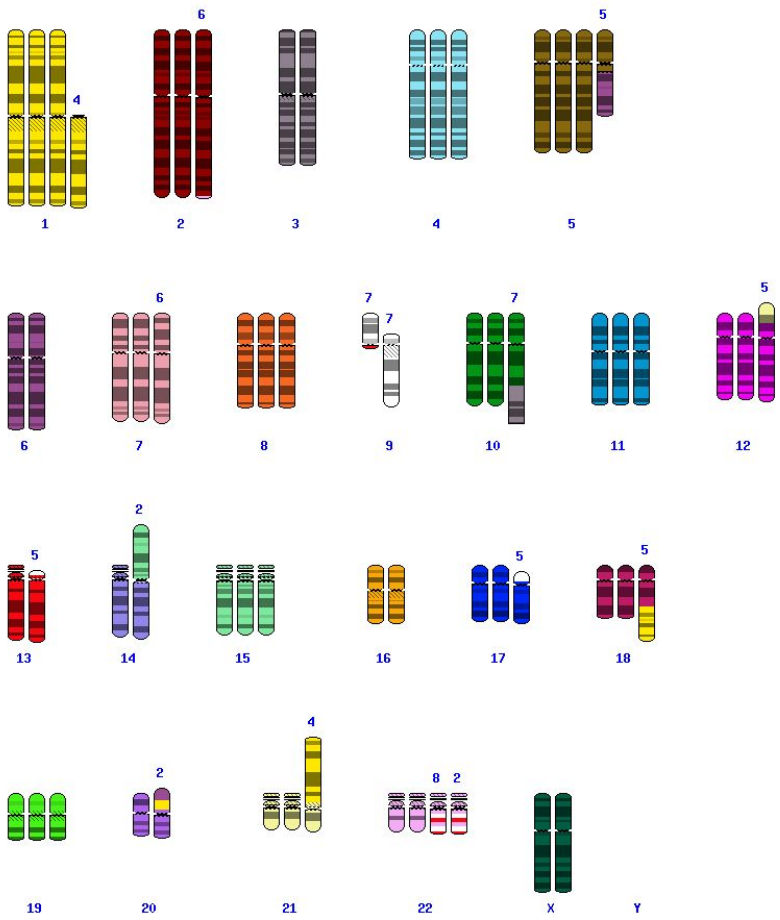
5- Figures 1,2,3 and 5 are unclear. Missing axis labels and so on. Please fix them, I didn't find them useful because it was impossible to see/understand the results represented there.

The figures have been fixed.

Minor Comments:

1- It is not clear to me why the authors chose to fix the shape parameter of the emission distribution from the control profiles and not IP profiles. For example, ChIP.seq input represents the amount of material available to the antibody at a given location, so it will typically be enriched in very different locations than the IP experiment and its counts will be inherently differently affected by biases. Can the authors at least discuss this point in more details? also pi (the zero inflation parameter) is fixed from the control profiles but I would expect zero inflation to be more probable in the IP experiment because there will be locations that are mappable but just weren't chipped while input is more uniformally distributed. Finally, how does the program behave when no control files are supplied? When I ran it without control, it didn't stop but told me the discretization has low quality.

Thanks for bringing this up; this is a general and important issue. For the sake of this discussion, it would be useful to know the sources supporting the claim that the 'ChIP-seq input represents the amount of material available to the antibody'. What about copy number variations? The vast majority of human cell lines are cancer cells and the perturbation of the karyotype has a large impact on every experiment (for illustration purposes the image below is the karyotype of the cell line K562 where all the ChIP-seq experiments discussed in the manuscript except one were performed; source: http://www.ncbi.nlm.nih.gov/sky/). Amplification of a short fragment typically gives rise to a spurious peak in the profile. What about the sex of the cell line? This has a major effect on the number of reads expected on the X chromosome. What about G+C content, which is known to affect PCR efficiency? What about mappability? About 20% of mammalian genomes are unmappable and 10-15% have dubious mappability, with huge consequence on the expected number of reads. What about "input controls" in the strict sense? In this case the sample never comes in contact with an antibody before sequencing... If the biases are thought to be different between the control and the experiment, then why run a control experiment at all?



The question is rather which control was run. There are several options available (input chromatin, mock IP with nonimmune serum, mock IP with irrelevant antibody, specific IP in a knock-out system etc.). We trust the experimenters to choose the most appropriate to correct the biases in their experiments.

That said, we cannot expect that literally everything except the target sites behaves exactly the same in the control as in the experiment. Differences are bound to exist and in this sense it would be justified to estimate the baseline on both the experiments and the controls. Actually, we tried it in an early prototype of Zerone and we observed that fitting the parameters alpha and pi together with the other parameters prevented the convergence of the Baum-Welch cycles or led to aberrant solutions. The reason is that the presence of peaks in the ChIP-seq experiments could also be explained by an overdispersed baseline distribution (a very small value of alpha). So fitting alpha and pi separately dramatically increases the robustness and the speed of Zerone. As discussed above, this solution implicitly assumes that controls are good proxies for experimental biases, which is up to the experimenters.

We have added the following text in the discussion (end of the second paragraph).

Note that control profiles play a key role in the process. In order to evaluate the quality of the discretization, Zerone implicitly assumes that the user has provided controls that properly capture systematic biases such as batch effects, mappability and copy number variations.

Thanks for pointing out that Zerone was running without control profile. ***The new version does not run without at least one control file.***

2- (section 2.5) the parameters used for each peak finder should be indicated

We have put this information in the Docker container.

3- (section 2.1) The authors should acknowledge previous implementations of Zero inflated negative binomial distributions in general and specifically in peak finding (for example: doi:10.1186/gb-2011-12-7-r67), especially since it is now widely demonstrated in many papers that the Poisson distribution is inadequate to model read counts from sequencing experiments

The link above points to a manuscript describing ZINBA, a ChIP-seq discretizer using the Zero Inflated Negative Binomial (ZINB) distribution. We initially planned to benchmark Zerone against ZINBA, but unfortunately, ZINBA is no longer maintained and it is broken under the current version of R. We had to run it on legacy versions of R, but we kept bumping into technical difficulties because of this. We abandoned the benchmark and removed the references to ZINBA in the text. We should indeed have kept one for the reference to the ZINB distribution.

The citation has been added at the end of the first paragraph of section 2.1.

4- (section 2.1) The negative binomial distribution models sequencing biases only implicitly. This can be mentioned in the last paragraph in contrast to previous approaches where such biases are modeled explicitly.

This is a good point. The end of the corresponding paragraph now reads as follows.

Note that the biases are not modelled explicitly from local features of the genome (e.g. G+C content), but implicitly by adjusting the variance of the distribution. Also, the ZINM distribution models the statistical dependence between replicates and thus yields more accurate probabilities than assuming independence.

5- The pareto front idea was used before to inspect signal-to-noise ratio by a program called deeptools: https://github.com/fidelram/deepTools/wiki/QC - I agree that in general this is a nice agnostic way to compare peak finders but I'm not sure I agree that a good peak finder should always be close to the line on the pareto front. This should perhaps be only the case if the dataset is almost free of biases and has very straightforward single peaks (like a very clean ChIP-seq or chip-exo dataset)

Thanks for the reference to deeptools, which were were not aware of. We will do our best to rephrase the argument as we understand it: for profiles with clearly defined peaks, the Pareto front is expected to have a kink (a rapid increase followed by a slow growth in our representation), whereas for more diffuse signal the curve may be smooth or even close to flat. In the extreme, the curve would only capture mappability artifacts and discretizations close or far from the front would be equally meaningless.

This would indeed be the case, but this has little to do with the curve itself. Measuring precision, recall, accuracy etc*. on this data set would be equally meaningless, wouldn't it? For intermediate cases, the authors of the link above give the following example (at the very bottom of the page) "H3K27me3 is a mark that yields broad domains instead of narrow peaks, it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed." But it does mean that the signal is more diffuse and less straightforward to discretize.*

Provided the experiment is valid, we do not see in which case a discretization capturing less signal (ChIP-seq reads) for the same amount of called targets would be better. That said, we would be happy to discuss this issue in the manuscript and bring the necessary corrections if we are given such an example.

6- Perhaps browser snapshots of chip-seq signal and zerone discretization in select locations might be helpful in better judging its behavior.

This is a good idea, unfortunately the revisions have inflated the manuscript to 8 pages and we are running out of space.

7- I can easily imagine an extention to zerone in which the user defines what they think are high and low quality datasets to retrain the SVM. Is this already possible by Zerone?

This is an excellent suggestion and we have started to work on implementing this feature (this is a separate branch on the Github repository). We will need to design a user-friendly interface and to internalize the SVM training (so far this was done separately), which will require more time than allotted for this revision. We are also considering the idea of creating a database of ChIP-seq profiles to train the classifier, where users could submit their profiles together with their opinion about it. Any additional feedback on how to implement any of those ideas is more than welcome.

Reviewer: 2

Comments to the Author
The authors design, implement and analyze an approach to solving a long-pending problem of combining different quality sequencing (ChIP-seq) data. Their solution is elegant and combines a number of important techniques from machine learning and statistics to arrive at the best to date result which is a ChIP-seq discretizer with a built-in quality controller. Additionally, the implemented software is also the fastest one.
The presentation is very clear and well-written. I would only suggest that a fair comparison of Zerone to the existing solutions should include a meta-server solution. Meta-servers have become rather popular in bioinformatics due to their apparent power coming from multiple sources. And Zerone is in a sense a meta approach. Here, a relevant example is to benchmark it against Peak Finder Metaserver (Kruczyk M et al, BMC Bioinformatics, 2013).

PFMS uses outdated versions of most of the peak finders. Comparing Zerone against these outdated versions would be unfair since newer versions are supposed to be improved (for the benchmarks, we used only the most recent versions of each software). Also, after following the installation instructions of PFMS, the compilation of two of its dependencies (cisGenome-2.0 and HPeak-1.1) fails. In the past we thought about including precisely these two programs in the benchmarks, but we faced the same problem. Even when downloading the programs from the developers' websites, they only compiled after editing the sources ourselves (some libraries were not included or linked). We think that the average user may not have the necessary C programming skills to fix the compilation errors that appear during PFMS installation and therefore it is not relevant to compare Zerone against it (since the user would never face the dilemma of choosing one of these two discretizers over the other).

On the other hand, ***we provide a fully running Docker container*** *where the users can replay our benchmark. Installing Docker is straightforward, and Docker is establishing itself as a standard service for virtual machines in bioinformatics. The information is centralized in a single Makefile so that the user simply has to type 'make' and see the whole benchmark being replayed. If they want to change a parameter to see how it impacts the benchmark, they can just edit the Makefile and run it again. If they want to add input data sets, they just have to append their path or the link to download them.*

Overall, we believe that a publicly available virtual machine is a more user-friendly solution than PFMS for similar benefits.

The Supplement is more than adequate. Noticed one typo only:
(ZINB)is should be (ZINB) is

Thanks. The typo has been corrected. We also realized that section 6 described an older version of the algorithm, which is no longer used in Zerone, so this section has been removed. Section 5 is also not used in the present version of the software, but because it is generally useful it has been turned into an appendix.

Minor comments.
Fig. 1 is somewhat difficult to read since the choice of graphical representation does not make the vertical bars easily discernible.

Figure 1 is now in color.

Section 2.5
add comma before "respectively".

Section 3.1
(c.f. section 2.3) should be (c.f. Section 2.3)

for instance by discussing
for instance, by discussing

Section 3.2
the second higest recall
the second highest recall

Fig. 6
The overlap for BayesPeak is only visible at 400% blow of the figure. In print it is not. Refine the picture.

The figure aims to show that the results of BayesPeak on the two replicates are mostly indistinguishable. We have updated the figure legend to make this explicit.

Note that the discretizations made by BayesPeak overlap completely and are indistinguishable in this representation.

Section 4
disretized
discretized

First, the statistical
Firstly, the statistical

and

Second,
Secondly,

The first form is acceptable, but the explicit adverbial preferred in formal texts.

Thanks for pointing this out. ***The corrections have been made****.*

Reviewer: 3

Comments to the Author
In this manuscript entitled "Zerone: a ChIP-seq discretizer for multiple replicates with built-in quality control" the authors proposed a new ChIP-seq discretizer with built-in quality control using hidden Markov model and

zero-inflated negative multinomial (ZINM) emissions to determine the locations of enriched areas. The authors described that this method provides a better fit to the ChIP-seq data than Possion and NB distribution that have been widely used by other methods by comparing three ChIP-seq callers. The method has key processes or ideas for improvement of ChIP-seq analysis and the authors also proposed graphical representation of discretizations. However, the paper need be more improved in manuscript organization and logical flows in method.

Major comments:
1. ZINM is proposed to model the counts of sequenced reads of ChIP-seq and it seems reasonable for ChIP-seq data. The authors showed that ZINM distribution is better fitted over the all ranges in 300bp window than other two distributions using one human leukemia cell line 562 only in Fig. 1. The authors should discuss the reason of the choice (process) of 300bp window size and the concordance in other cell lines. Do the numbers on the X axis of the figure indicate the number of reads? It seems like read count is too small. Why not check the distribution on samples treated by IP with antibody, not a mock control data (that is treated by the IP without any antibody)?

The choice for 300 bp windows is guided by the size of the sonication fragments (between 200 and 300 bp generally, even though this step is notoriously hard to standardize exactly) and by the sequencing depth so that the average number of reads per bin is not too low. We are not aware that the cell line in itself (i.e. irrespective of the experimental conditions) may have a major impact on the optimal window size. It would be useful for this discussion to know the sources supporting this claim. We would be happy to discuss this issue in more detail in the text, provided we can cite such references.

The section 'ChIP-seq preprocessing' contains a brief discussion of how to choose an appropriate window size *(see minor point 3 raised by reviewer 1).*

The numbers of the X axis in Figure 1 are indeed read counts. The average is around 3-4 reads, and since there are 10 million windows of 300 bp in the human genome, this corresponds to a total of 30-40 million reads. Nowadays a typical Illumina run would be about 5 times as much, but... time flies and ENCODE data was obtained on older machines (this file was released in May 2012).

The description of the file says "datatype: ChipSeq; datatype description: Chromatin IP Sequencing; treatment description: No special treatment or protocol applies" (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM935601), from which it is difficult to gather exactly what was done (we assumed that it was crosslinked, sonicated and uncrosslinked before sequencing). We are not aware of experiments where the IP is performed without antibody, but we have seen experiments where the IP is performed with nonimmune antibodies. The issue is that peaks are clearly visible on some of these profiles depending on whether the antibodies were from mouse or rabbit. Because we could not decide which of them was the 'right' negative control with an IP, we opted for one where no IP was performed.

In our experience, the ZINB gives a very good fit for all the negative controls we looked at, including for large window sizes like 3 kb. For instance, we did not choose the file wgEncodeSydhTfbsK562InputRawData.fastq because it gave a better fit than other mock profiles used in this study. Our hands-on experience is that Figure 1 is representative, but if you strongly feel that this negative control is inappropriate, please point us to another publicly available file and we will redo Figure 1 with this data set, together with the adjustments in the text.

2. The authors should check the ENCODE accession numbers of three data sets that represent signals with different patterns used to compare Zerone with other discretizers in 2.5. The accession numbers of ENSCR000DWE, ENCSR000DWB and ENCSR000EHP represent CTCF, H3K36me3, Pol2, respectively.

Thanks. We had not noticed that one of the datasets was incorrectly referenced in the paper, ***this has been fixed****.*

3. Although Zerone included automatic quality control, the most novel feature of Zerone, to solve some issues originating from replicates by combining ChIP-seq data of replicates in a single discretized profile and then

resolving by maximizing the likelihood, it is not clear to explain what the user can identify and what the user can use for further analysis via this process. The authors need to be shown the 8 summary statistics of output from quality control process the authors referred in the manuscript at least in the supplementary information. The authors should explain accurately the advantages of this step.

After mentioning the features, we have added the following description (second paragraph of section 2.3)

The features consist of entries from the transition matrix ($Q_{3,1}$, $Q_{3,2}$ and $Q_{1,3}$, indicating the size of the targets), the maximum value of the ratios between (p_2 , . . . , p_r) for each pair of states (indicating the signal to noise ratio), the average posterior probability of targets (indicating the confidence of the discretizer) and the amount of targets.

4. For automatic quality control, the authors suggest the good example in Fig. 3 that considering one of two replicates is the proper way to handle low quality data of replicates. Is the reliable peak (true positive occupancy) in analyzing an experiment with replicates decided by the user manually or automatically? Can the user distinguish between the discordant peaks and concordant peaks in replicates from files after automatic quality control?

The quality control is a global ‘accept’ or ‘reject’ recommendation, it does not focus on any peak in particular. The confidence for each peak is computed during the Baum-Welch cycles as a posterior probability. Each window is associated with a probability of being in the state ‘high’; the higher this probability, the higher the confidence that the window is a target. There is currently no direct way to identify discordant peaks as such. However, the user can now identify high confidence peaks, which are by definition concordant.

We have added an optional --confidence flag allowing the user to request the confidence of each window. However, this option is not available when windows are merged. The documentation and the tutorial have been updated to describe this feature.

5. Other tools such as MACS have a sift process for signal profiling because ChIP-seq data form a bimodal distribution pattern near binding site or peak. This step does not appear in this manuscript. Explain that it could be fine for Zerone without this critical step.

Thanks for giving us the opportunity to discuss this important point. The short answer is that Zerone is not a peak finder. When developing Zerone, we took the decision early on to make it agnostic to the type of signal because we felt that there was a need for ‘exploratory’ tools. The general window-based HMM approach is well suited for this purpose, but it involves that targets can be mapped only to a window, and not to a binding site.

In the case of transcription factors, the distribution of ChIP-seq reads is bimodal, but it is still centered around the binding site, which belongs to one and only one window. In effect, the bimodality corresponds to a spread of the signal, which will ‘leak’ to the neighboring windows in the worst case. This in turn may lead to false positives and justifies addressing the issue. One option was to shift the reads in order to reduce the spread, but this immediately conflicted with the decision of making Zerone agnostic to the type of signal. Indeed, shifting is not meaningful for domain-like features for instance. In addition, misestimating the amount of shifting required may actually increase the spread, and requiring the user to specify it would again defeat the purpose of writing a general agnostic tool. Instead, we tackled this issue with the third ‘intermediate’ state of the HMM. There are less reads in windows contiguous to those containing a binding site and they typically end up in the ‘intermediate’ state instead of the ‘high’ state. This also explain why we consider that only the ‘high’ state corresponds to real targets.

In summary, shifting is critical to call binding sites, but to call enriched windows we favored a more universal approach.

6. The authors should describe the data set information and the statistics of analysis process such as library size, read size, sequencing depth, the total number of mapped reads, coverage of features (covering motifs) because

library size and the number of mapped reads (or sequencin= depth) are important factors for choice of window size and calling for broad or sharp peak, respectively.

Table 1 now describes the data sets used in the benchmark.

7. The authors should describe specifically advantages or/and comparison of Zerone's method over Irreproducible Discovery Rate (IDR) developed by the ENCODE consortium for concordance of peak in replicates or reproducibility. For example, ENCODE developers do not recommend the IDR method for broad peaks.

Thanks for highlighting this. We have added the following paragraph to the discussion (third paragraph).

The quality control (QC) implemented in Zerone goes beyond the Irreproducible Discovery Rate (IDR, Li et al., 2011) in several ways. Zerone measures the quality of the discretization and not only the consistency between replicates. Also, the QC of Zerone is neither limited to a specific type of profile (e.g. sharp peaks), nor to a preset number of replicates. Finally, issuing an 'all-or-none' call about the discretization is better practice than silently ignoring the regions that differ between experiments (see e.g. Fig. 3).

Minor comments:
1. It is better to add legends or/and use color with Fig 2, 3, 5.
2. The manuscript is generally well written, but also contain many typos and grammatical errors (e.g. catpuer in 20 line in 3.2, disretized in 23 line in 4. first phrase In 3.3).

*Thanks. **We fixed the typos and hope that none remains**.*

4. Describe output files finally generated by Zorone. For example, for MACS, there are a BED format file containing the peak locations, r script and so on.

*The two possible output formats of Zerone **are now described in the tutorial**, available at the GitHub repository. One of these formats is BED-like. The output value of the zerone() function in the Zerone R package is now described in the R documentation.*

5. Describe Zorone parameters. For example, for MACS, shift size that is used to improve the spatial resolution of inferred TF binding sites can changed by the user.

***Zerone parameters are now described in the tutorial**, available at the GitHub repository. There are actually very few parameters on purpose. We prefer to not describe them in the manuscript since features may be added or changed in the future, making the text outdated. Instead, we prefer to centralize the information on the Github repository to make sure that the description always fits with the software at hand.*
