Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Response to reviewer comments

Editor

> While I understand that a full analysis of mammalian genomes goes beyond the scope of this work I do think that it is important to show some data on mammalian systems to document how well SMAC-seq works in the context of endogenous 5mC methylation and how well you can resolve endogenous from the exogenously introduced GC and 6mA marks. Would it be possible to focus on one particular locus only?

We have generated low-coverage whole-genome m^{6}A-SMAC-seq data for the GM12878 lymphoblastoid cell line. Analysis of average SMAC-seq signal around known chromatin landmarks such as CTCF sites, DNase hypersensitivity regions and TSSs shows that endogenous methylation does not interfere with accessibility measurements in this context.

Reviewer 1:

> This paper marginally extends previous methods, i.e. Nome-Seq and dSMF, that leverage the bias of methyltransferases for open chromatin to map nucleosome position and chromatin accessibility. The main advance over previous methods is the addition of a third methyltransferase specific for m6A. The authors argue the addition of the third methyltransferase has two advantages: higher data resolution due to more frequent methylation sites and less confounding of the accessibility signal with endogenous methylation. For an extension of an existing method, their comparisons to previous methods are underdeveloped, and the paper fails to show the addition of a third methyltransferase meaningfully affects the data produced by the assay. With the exception of a brief mention of the dynamic range of dSMF, they do not compare their data to the previous methods. Also, the examples in the paper primarily reflect coordinated changes activity across multi-kb regions, so the support for the claim that the addition of a third methyltransferase increases the assay resolution is limited. Finally, all experiments are performed in yeast, an organism without endogenous methylation, which does not allow the authors to study how the method interacts with endogenous methylation in practice, a main selling point of their method. Overall, this paper does not provide strong evidence that their method is a significant improvement over previous methods or study a biological phenomenon in enough detail for the biological findings to stand on their own.

We respectfully, but very strongly disagree with the reviewer's assessment of the novelty and usefulness of SMAC-seq (an opinion that is also shared by other reviews of the manuscript, which find our work both "novel and important"). There are multiple major novel aspects of the SMAC-seq approach that provide fundamentally new capabilities that have not been available before, and that also open the door for entirely new approaches for studying chromatin structure.

The ability to profile chromatin accessibility within long single molecules is not a marginal extension of previous methods, it enables qualitatively novel insights into the state of chromatin accessibility in the genome. There are severe limitations to existing NOME-seq/dSMF approaches. First, only fragments $\leq$600bp long can be assayed using short-read sequencing and even if it was possible to sequence longer fragments, bisulfite conversion destroys DNA so longer fragments are just not available to sequence (indeed, the previously published dSMF study had to employ two rounds of PCR for a total of $\geq$40 PCR cycles to capture enough fragments of a few hundred bases in length). This allows one to capture accessibility states within a single regulatory element and nothing more than that. Even then, it is often impossible to do so around regions of interest due to the lack of CG/GC sites. This limitation is even more severe in genomes containing endogenous methylation where only GC can be used. We have observed many promoters where the density of CG/GC dinucleotides is too low to say anything about chromatin accessibility states. Though it was not our intention to explicitly highlight such cases, this issue is clearly visible

in Supplementary Figures 4 and 5 in the previous version of the manuscript where multiple positioned nucleosomes are not covered by even a single CG or GC dinucleotide. Thus the addition of m^6A is indeed a major improvement over previous methods.

Even more important though is that with very long single molecule reads one can profile chromatin states on multikilobase resolution, which allows the accessibility status of distal regulatory elements to be mapped within the same original chromatin fiber. There has never been any way to generate such measurements in the past. In addition, repetitive elements can now be directly studied, which was not possible before (one had to rely on not entirely reliable statistical methods for distributing multimapping reads between different locations in the genome).

More broadly, the transition from short reads to long single-molecule reads is analogous in its importance for functional genomics to the transition from microarrays to short-read sequencing. In the case of mapping chromatin accessibility, it was already possible to do so using tiling microarrays even without short read sequencing. What the transition to short reads described in studies such as Hesselberth et al. 2009 provided was increased accuracy and resolution and eliminating the need to manufacture arrays for any new species assayed, fundamentally new capabilities. Analogously, long reads provide qualitatively new capabilities compared to short reads – the previously mentioned capture of long-range dependencies between chromatin states, and the ability to directly map any DNA modification rather than being limited to 4 bases and enzymatic conversions of only one modified base to another (usually accompanied by DNA degradation/cleavage) as is the case with Illumina sequencing. This ability makes possible the development of a large new class of assays for studying chromatin structure, including single-molecule multiomics assays (for example, simultaneous mapping of accessibility, endogenous methylation and protein occupancy/long-range physical interactions in the same molecules). We also foresee future variations of SMAC-seq employing an expanded repertoire of both DNA modifying enzymes and DNA modifications that they deposit, depending on the particulars of the experimental system being studied. Thus SMAC-seq is properly seen as the first method of this broad entirely new class of technique, furthermore underscoring its importance.

We used yeast for developing SMAC-seq because it is a standard eukaryote model with a compact genome and for which numerous useful external genomic resources (such as chemical nucleosome mapping datasets) are available. There is a long tradition of pioneering functional genomic techniques in yeast, precisely for these reasons. The most recent such example is the D-Nascent method (Müller et al. 2019, *Nature Methods*) that uses BrdU incorporation and nanopore sequencing to profile active replication in the genome (as well as two other analogous studies that have been posted on *bioRxiv*). Additional examples include the first ultra-deep DNAse-seq study (Hesselberth et al. 2009, *Nature Methods*), two of the original RNA-seq method development studies (Nagalakshmi et al. 2008; *Science*, and Wilhelm et al. 2008; *Nature*), high-resolution Hi-C techniques such as Micro-C XL (Hsieh et al. 2016, *Nature Methods*) and Hi-CO (Ohno et al. 2019; *Cell*), and numerous others.

The method is widely applicable to other organisms, including humans. This is precisely because of the sparsity of CG dinucleotides that is a limitation of dSMF/NOME-seq. Potential confounding between CpG methylation and m6A methylation can only happen in the immediate vicinity of CpG dinucleotides given that nanopore sequencing works by reading 5-mers or 6-mers at a time. However, CpG dinucleotides are found on average >30bp from each other in mammalian species.

Also, while we consider this obvious, we nevertheless would like to explicitly point out that the mammalian version of SMAC-seq involves either m6A alone or m6A + GC, without a CG MTase included, and with computational filtering out of the immediate 5-bp vicinity of CG dinucleotides.

More generally, SMAC-seq as an assay is not defined by the combination of the three particular enzymes used in our study but the by use of high-density modifications and long reads. The array of such possible modifications is actually quite large, both given the wide diversity of methyltransferase enzymes that exist in nature and the large chemical space of modifying groups that can potentially be used as alternatives of methylation.

The main challenges for applying the method to mammalian systems are sequencing costs (which is also a main reason, for example, why in the past studies such as the first high-resolution DNase-based digital footprinting mapping by Hesselbreth et al. 2009 were also carried out in yeast) and high-molecular weight DNA isolation. But the introduction of the PromethION platform, with its very high throughput, promises to address the former, while significant progress has already been made on the latter. Additional strategies include Cas9-mediated locus selection, which has also been already successfully applied to selectively study individual loci using nanopore sequencing.

In the revised version of the manuscript, we have generated low-coverage whole-genome m^6A-SMAC-seq data for the GM12878 lymphoblastoid cell line. Analysis of average SMAC-seq signal around known chromatin landmarks such as CTCF sites, DNase hypersensitivity regions and TSSs shows that endogenous methylation does not interfere with accessibility measurements in this context. However, studying mammalian chromatin using SMAC-seq in depth

is the subject of subsequent studies and beyond the scope of the current manuscript.

> 1) What is the expected relative level of methylation from these enzymes? Do they have some innate sequence preference beyond the single-base (EcoGII) or dinucleotide context (M.CviPI, M.SssI)? How does might affect the results (particularly for TF footprinting).

We already examined the sequence bias issue in the original manuscript (Supplementary Figures 10 and 11). M.CviPI and M.SssI exhibit a near-100% methylation efficiency, as also already addressed in the original manuscript (Supplementary Figure 3). Estimating the methylation levels of EcoGII directly is much more difficult because m^6A cannot be read using Illumina sequencing and fully modified templates do not sequence well on the Oxford Nanopore platform. The latter reason is why we have struggled to generate deep sequencing coverage of positive control samplles consisting of genomic DNA treated with EcoGII. We have included such controls in the revised manuscript, which suggest a ~50% methylation efficiency for EcoGII, which is in line with what is reported by NEB (the reason for incomplete methylation efficiency is thought to be steric hindrance by already methylated immediately adjacent nucleotides). We do, however, note that it is quite likely that this is an underestimate due to the underrepresentation of fully methylated fragments in the nanopore sequencing dataset. The original publication describing the M.EcoGII enzyme (PMID: 29228259) reported 50% methylation of naked DNA after 5 minutes of treatment, increasing to 85% after an hour and even 95% if excessive amount of enzyme is used. Our *in situ* treatment is carried out for >20 minutes, thus the efficiency of methylating accessible bases is likely significantly higher than 50%.

We have revised the text to better explain these issues.

> 2) The authors largely rely on theoretical arguments to justify the utility of measuring mA6 to increase resolution, yet Fig 1e shows that they essentially are capturing the same topological features in chromatin (nucleosome positioning, etc). Is there any analysis that yields fundamentally different results due to the inclusion of EcoGII andenosine methyltransferase?

The purpose of Figure 1e was in fact to show that SMAC-seq can faithfully capture the same information as short-read based assays.

More importantly, SMAC-seq actually does provide new insights about general topological features of chromatin, as shown in Figure 4, where we show that the two strands of DNA are differentially accessible around positioned nucleosomes, and where we directly evaluate the dependencies between local nucleosome occupancy states.

We also reiterate that the major fundamentally different new capability provided by SMAC-seq is profiling accessibility within very long single molecules. Measuring the long-range dependencies between the status of distal regulatory elements within the same chromatin fiber has not been previously possible.

The sparseness of CG and GC nucleotides in the genome means that numerous positioned nucleosomes and even entire regulatory elements are either not covered at all using just CG or GC methyltransferases or are only covered by a single informative site, greatly reducing the signal-to-noise ratio. This point was already illustrated in the original manuscript in the Supplementary Figures section, but we have now added additional Supplementary Materials figures illustrating the impact of the addition of m^6A using average accessibility estimates and browser snapshots analogous to those shown in Figure 1.

> 3) Despite only minor modifications to previously published work (mA6 and long reads), the authors do not perform any meaningful comparative evaluation of their method. The authors should systematically compare the sensitivity and resolution of their method to competing methods (ie., dSMF and NOME-seq).

We are afraid that the reviewer misunderstands the capabilities of and the goals behind the development of SMAC-seq. The major fundamentally novel aspect of SMAC-seq is the ability to profile accessibility on a multikilobase scale within single molecules. Given that dSMF and NOME-seq are inherently incapable of accomplishing this task, then it is not clear to us how a meaningful comparison could be carried out.

> 4) The footprinting analysis Supplementary Figs. 35-44 is extremely confusing. Finally, what do the colors (white, yellow, black) signify for the methylation dot plots in Supplementary Figs 36-44? Where is the footprint for each of these factors?

The white color indicates missing data, yellow and black correspond to unmethylated/inaccessible and methylated/accessible as shown elsewhere in the manuscript. The footprint is indicated by location of the ChIP-exo peaks on each strand shown underneath the single molecule data plots. We have revised the figure legends accordingly.

> 5) The authors note that Reb1, Rap1, and ORC1 make strong footprints while Abf1 and Cbf1 make weak footprints. This is striking because Abf1 and Cbf1 are known to have strong occupancy and create extremely well-defined DNase I footprints (Hesselberth 2009). The authors state that TF footprints [with SMAC-seq] is probably dependent on the biophysical properties of individual factors. This is extremely weak justification that undermines a selling point of their paper ("enabling high-resolution footprinting of transcription factors"). Do the authors have any other examples other than Reb1 and Rap1 that create well-defined regions of methylase protection?

We thank the reviewer for pointing out the Hesselberth et al. study.

We based our footprinting analysis on the most reliable transcription factor occupancy datasets available, which is ChIP-exo. We examined footprinting only for factors for which ChIP-exo datasets are available. Thus our original results were not directly comparable to those obtained by Hesselberth et al.

We have now compared SMAC-seq footprinting profiles over the footprints identified by Hesselberth et al. (directly obtained from the Supplementary Tables in that study) with those observed using ATAC-seq and DNAse-seq, and find concordance between all three assays. We have added these results to the Supplementary Materials in the revised text. We also examined SMAC-seq profiles around motifs within the footprints from Hesselberth et al. We still observe robust footprinting by Reb1 and Rap1, and weak footprinting by Abf1 and Cbf1.

The phrasing "biophysical properties of individual factors" is probably not optimal and we have revised it accordingly, as the observed differences are most likely the result of how much access the different enzymes used to profile accessibility have to DNA occupied by proteins (which is a function of the binding of the TF but also of the properties of the enzyme). Our results identifying fine-scale and even strand-specific accessibility patterns around positioned nucleosomes (which are not observed in high-resolution DNAse-seq datsets) suggest that EcoGII can access positions that are occluded to other enzymes used to map chromatin accessibility.

We also note that using small molecules for mapping chromatin accessibility, e.g. as shown in the Umeyama & Ito 2017 DMS-seq study, also results in minimal Abf1 and Cbf1 footprints.

> 6) It is unclear what the resolution of the assay truly is. The average width of the protected DNA in Supplementary Figs is close to 100 base-pairs. The average yeast TF recognizes between 6-10 bp. The purported resolution of your assay is 3 bp. What explains the ¿10-fold difference in theoretical vs observed resolution?

As shown in Figure 4a,b and Figure S35 in the original manuscript, the average width of the protected DNA is much shorter than 100 base pairs, thus it is not entirely clear to us where the statement that it is "close to 100 base-pairs" arises from. Perhaps the reviewer is referring to the Rap1 footprints, which are indeed on the order of 100bp. However, this is simply a reflection of how Rap1 binds to DNA (as also shown by ChIP-exo data).

> 7) The sheer volume of figures (∼70 figures) and analysis is distracting and make this manuscript difficult to read and evaluate.

We understand the reviewer's concern. However, our manuscript introduces an entirely novel approach and data type, that has never been carried out and used before, thus a comprehensive technical evaluation and an exploration of the properties of the datasets were necessary, which is not possible within the limited space allowed for by the main text. We would like to point out that the other reviews of the manuscript actually found the level of detail and thoroughness a major virtue of the manuscript, precisely for these reasons.

Reviewer 2:

> I have concerns regarding the stress response analysis. For instance, ATAC-seq was collected at all time points, but it is never shown that ATAC-seq changes in response to stress correspond with SMAC-seq changes (time point 0 concordance is shown in Supp Fig 15, but what we are interested in is whether the changes across time points are consistent). As well as a systematic analysis, the ATAC-seq time points should also be shown in genome browser plots.

We did generate ATAC-seq datasets for all time points, but we did not include them in the original figures because space there was limited and because ATAC-seq is generally highly concordant with SMAC-seq measurements, a point already established in the text prior to that, thus its inclusion would not have provided much additional important information. We have now added separate supplementary figures showing that changes in accessibility measured by SMAC-seq and by ATAC-seq are similar to each other for the genes shown in the main text figure.

Reviewer 3:

> In the manuscript, "Long-range single-molecule mapping of chromatin accessibility in eukaryotes" Shipony et al. describe the development of SMAC-seq, an approach that combines long-read nanopore sequencing with a chromatin accessibility assay. This work is novel and important. The authors do a nice job explaining the development of their system, and benchmarking their results to previous methods. Specifically, they use a mixture of DNA methyltransferases to mark sites of chromatin accessibility and read out modifications using direct, long-read sequencing. The validate their results by comparing their combined data with the results from similar assays that were previously conducted with short read sequencing, but show how their work can provide insight into the relationships of chromatin accessibility at a longer scale. While I think the TF foot printing aspects of the manuscript are relatively weak and those claims need to get toned down, overall this is an exciting manuscript both for the results that are presented and also for the types of related experiments that the authors discuss at the end of the manuscript. Toning down the TF foot printing claims will not detract from the impact of this work. This is the first genome-wide characterization of chromatin at this scale and therefore will be of great interest to the chromatin community.

We would like to thank the reviewer for their thoughtful comments and have revised the manuscript following their suggestions.

> 1. I appreciate the large amount of relatively raw and partially processed data the authors included in the SI. It was very helpful to get a sense for texture of the data and the different decisions the authors made to interpret these complicated data. They also do a nice job justifying many of their ad hoc decisions (e.g., p = 0.5 for the binary cutoff). It would be helpful if the authors could include a histogram of the read lengths from these data. In some of their analyses, the authors examine regions that are substantially larger than the ∼1.5 kbp median length. From the mean and median in Table S1, it is clear there is a long tailed distribution of read lengths, but how many reads the authors have at different lengths is not straight forward to determine from what is presented. This seems like a key parameter in this experiment and could use additional discussion, especially as some of the regions the authors discuss are >9kb.

We have added a supplementary figure showing the read length distributions. We would also like to point out that the read length distribution can be greatly improved with more sophisticated DNA isolation and size selection, and indeed we have been able to obtain much higher average read lengths in subsequent experiments in different systems. With the natural progression of technology development, we expect yet further improvements in that direction.

> 2. One application that will be of interest is whether SMAC-seq can eventually be used to confidently call the long distance correlation between footprints/sites of TF occupancy. While TF occupancy is a major focus of the introduction, the remaining obstacles and whether or not the authors are close to achieving this goal (at high resolution) are not discussed in this manuscript. It is clear their SMAC-seq data contain footprint information, but it isn't clear if it would be possible to call footprints as in the single molecule analyses, or how good the occupancy estimates are for the "high-resolution foot printing of transcription factor occupancy (p2)" promised in the introduction. The authors demonstrate the footprints are clearly present when one knows where to look (which might be sufficient for many analyses), but it is unclear what sensitivity and specificity of single-molecule footprint calling can be achieved. My general sense is that the authors aren't quite there yet, which is fine but means the claims in the introduction should be toned down.

We share the reviewer's desire for correlating TF occupancy. We did in fact attempt to do such an analysis, the reason none of it was presented in the original manuscript is that binding sites for individual transcription factors are too widely separated in yeast for two of them to be captured by the same reads. It is also true that at present the noise levels in the data are too high for footprinting within a single molecule to be as reliable as we would like it to be; for that goal to be achieved, improved base calling models will be needed (if using plain methylation) and/or modifying DNA with bulkier modifications that induce larger current changes through the pore. We also expect more comprehensive TF binding resources in yeast to become available in the future (analogous to those provided by the ENCODE/modENCODE/mouseENCODE consortia in other model systems), which would enable the global measurement of correlations between the occupancy of different TFs.

> 3. Replicates: The authors have replicates from their untreated diamide experiment but do not present much data to demonstrate how well these replicates agree. While general agreement would be nice to present (e.g. average SMAC-seq signal between replicates and splitting out the replicates in the heat maps in Fig S50-51), I am primarily interested in whether the NMI statistic and the coordinated accessibility values are reproducible (at least in ordering). I appreciate the authors desire to create a statistic to capture coordinated accessibility, and also their logic for how the authors went about it. The problem is that it is very hard to tell how well this metric works. How many of the changes in red and blue in these plots reflect real, reproducible signal between biological replicates that are prepared independently? Reproducibility alone doesnt guarantee these differences represent real events on the chromatin because the artifacts could also be reproducible, but reproducibility would be reassuring. How do the NMI values correlate across the genome between replicates? And more specifically, if the analysis in Fig S47 is repeated on the replicates from the untreated diamide replicates, how well do the ranks of the most significant long range positions compare?

We also share the reviewer's concern about reproducibility. The reason such analysis was not presented in the original manuscript is that, as is common with time courses, the two replicates of the diamide treatment exhibited somewhat different temporal dynamics. We have added the requested comparisons in the revised version of the text. We would also like to point out that the calculation of the NMI statistic relies on repeated samplings from the population of single molecules and then taking the average.

> 4. The authors introduce their technique through a description of classic DNase I accessibility experiments and emphasize regulatory elements. While there is some reason to use this literature to motivate their work, DNase I accessibility is more analogous to the authors previous work with ATAC-seq. The authors discussion of MNase-seq and its predecessors is more relevant to the most successful aspects of this manuscript, yet gets relatively little introduction. There is plenty of literature about larger scale changes to chromatin accessibility that is more directly relevant to this approach. Given the emphasis on genome-wide nucleosome positioning (and the nice results the authors have in this area), I suggest the authors consider referencing works like Yuan et al 2015 PMID:15961632 which had large impact on how the field thinks about the genome-wide distribution of nucleosomes in yeast.

We thank the reviewer for this suggestion, and have revised the manuscript accordingly.

> Minor points: 1. On page 2 the "0% on naked DNA" presumably refers to 0% on *untreated* naked DNA.

We thank the reviewer for pointing out this error. This is indeed what we meant, and have corrected the text to reflect that.

> 2. It would be helpful if the authors emphasize that some of this discussion is about the theoretical resolution rather than actual of the technique (p2 claims about 3bp with m6A) given that most of their analysis uses shared information across windows.

We have changed the text to better clarify that point.

> 3. In Fig 1 the MNase/DNAse label is confusing because it implies division. Also MNAse -¿ MNase.

We have corrected the label.

4. It would be helpful in the Fig 1 legend to clarify that it is *average* signal to distinguish it from the aggregate single molecule signal as was emphasized in the text.

We have altered the legend accordingly.

5. Fig 4 subfigures are misspecified.

We thank the reviewer for pointing out the mislabeling and have corrected the figure.