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

Reviewer 1:

> I still have major concerns about the data and analysis performed within this manuscript.
>
> Chiefly, I am still unconvinced by the claims of dramatically increased resolution using EcoGII. In the revised manuscript no additional quantitative analysis regarding the putative increase in sensitivity and resolution gained by including EcoGII were performed other than providing a couple of additional browser shots.

We are disappointed that the reviewer feels we have not successfully communicated these points in the previously submitted version of the manuscript. In the previous version, we did attempt to demonstrate the added value of EcoGII (i.e. m^6A modification) in several other figures (Supplementary Figures 1, 7 and 8, 25 and 26, 44 in the previous version of the manuscript, illustrating the order-magnitude difference between the sparse coverage provided by GC dinucleotides and the dense coverage provided by m^6A, and providing examples of biological phenomena that only become discernible with the addition of m^6A, such as the strand asymmetric protection patterns around positioned nucleosomes). We have now further expanded upon these analyses demonstrating the utility of m^6A.

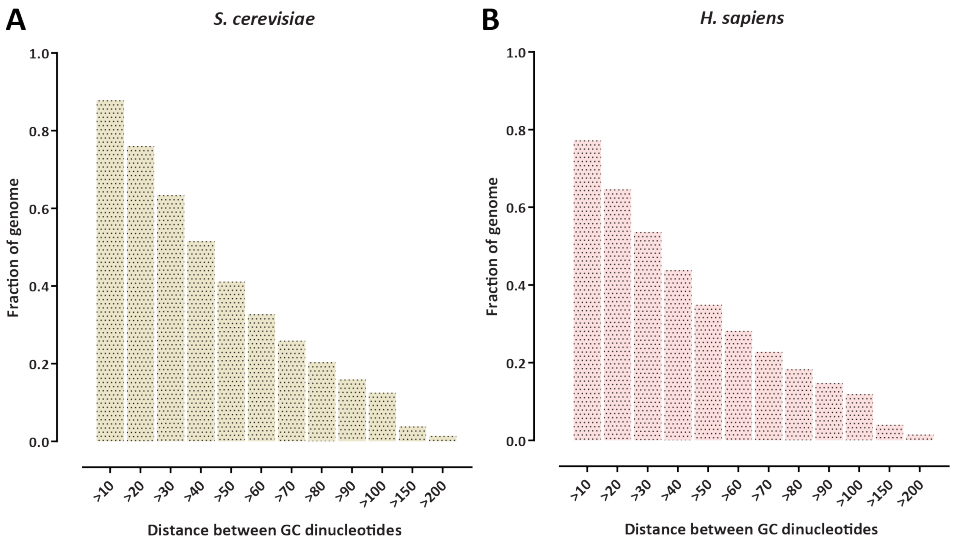
We would like to first note that the browser shots that we presented show only a small number of the many regions of the genome that cannot be probed with methods other than m^6A. The regions highlighted were not "cherry-picked" from around the genome, but were instead obtained by examining a random single locus on chrXIV. We have now provided many more snapshots showing some of the numerous regulatory regions where CG/GC methylation fails to capture accessibility while m^6A does (Supplementary Figures 18–47). Given the large number of regions that are devoid of GC dinucleotides, this selection of regions is necessarily still only a small subset of regions of the many regions in the genome that cannot be assayed using that sequence context. In general, these tracks also show how the sparse nature of GC/CG dinucleotides degrades the correlation of these accessibility profiles with DNAse seq (described further below).

To further highlight this point, we have included additional global statistics for the regions lacking GC dinucleotides in the yeast and human genomes (Supplementary Figure 2, also reproduced here as Response To Reviewers Figure 1). This analysis shows that more than 50% of each genome consists of regions with GC dinucleotides spaced at least 30 bp apart (and even 40 bp in the case of *S. cerevisiae*), which is longer than the length of a typical nucleosome linker, and ∼10% of each genome consists of regions where neighboring GC dinucleotides are 100 or more base pairs apart.

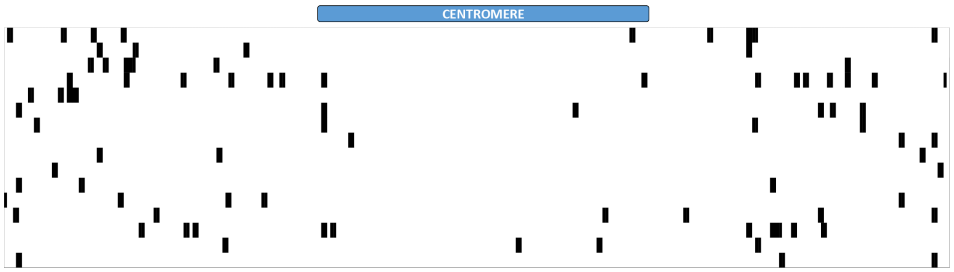
We have also highlighted the nucleotide content of centromeres in yeast, which are highly AT rich and as a result can only be assayed using m^6A (Supplementary Figure 3, also reproduced here as Response To Reviewers Figure 2).

Next, we demonstrate the utility of m^6A and the limitations of approaches using CG/GC by comparing CG/GC to m^6A signal in terms of their correlation with DNAse-seq signal at the base pair level. While DNAse-seq measures only regions of accessible chromatin (and generally not linkers etc., see discussion below), higher correlation with DNAse-seq signal still suggests more highly resolved accessibility data in regions of DNAse-seq accessibility. Indeed, this is what we observe for m^6A (Supplementary Figures 49 and 50, also reproduced here as Response To Reviewers Figures 2 and 4). At all sequencing depths, m^6A is much more highly correlated with DNAse-seq at the base pair level, suggesting a large and robust improvement in the ability to capture accessibility using m^6A compared with CG/GC-based measurements. GC-based measurements only converge with those using m^6A at very high densities of GC nucleotides, but only a small fraction of genomes has such properties.

These additional examples and analyses further demonstrate that m^6A provides a major boost in terms of allowing the accurate capture of accessibility profiles along the genome as well as enabling the assaying of the whole genome in the first place (rather than leaving complete gaps in coverage where there are no GC sequence contexts), including well known highly biologically relevant examples, such as centromeres (that are highly AT-rich). In addition, our long read analysis provides, for the first time, coverage of repetitive regions, unmappable with short reads (which take up up to a quarter of mammalian genomes and even higher fractions in other organisms).



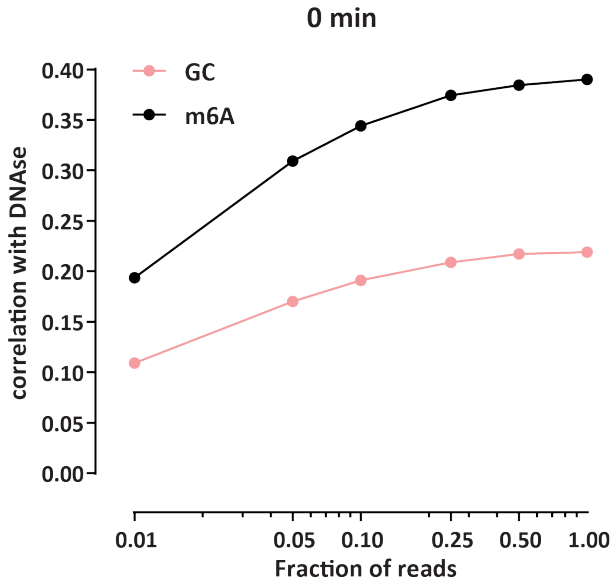
Response To Reviewers Figure 1: GC positions alone are insufficient to provide proper coverage of the genome in the context of a methylation-based assay for profiling chromatin accessibility. Shown is the distribution of the fraction of the genome that contains no GC dinucleotides closer to each other than the indicated distance. The mitochondrial genome was excluded from the calculation. (a) *Saccharomyces cerevisiae*; (b) *Homo sapiens*. More than 50% of each genome consists of regions with GC dinucleotides spaced at least 30 bp apart (and 40 bp in the case of *S. cerevisiae*), i.e longer than a typical nucleosome linker.



Response To Reviewers Figure 2: Important functional elements in the yeast genome, such as, in this example, centromeres, are almost completely devoid of GC/GC dinucleotides. Shown is the distribution of GC dinucleotides (black rectangles) around *S. cerevisiae* centromeres.

Finally, as articulated in the previous version of the text, SMAC-seq is not defined by the use of m^6A. EcoGII with a SAM donor is used here because this is the best enzyme for the purpose of mapping chromatin accessibility that is currently available, i.e. it meets the criteria of a high density of modification without context specificity, ability to modify dsDNA rather than requiring ssDNA templates, fast action, etc. We anticipate that the enzymatic toolkit will vastly expand in the future. The defining feature and major advance of SMAC-seq is the coupling of dense DNA modifications with long-read sequencing, which in the words of Reviewer #2 provides "an important advance over related methods, both because of the longer reads and the better coverage and resolution provided by m6a...".

> I get the impression that the potential extra signal of measuring m6A appears to be completely dampened by the inefficient kinetics of the m6A methyltransferase EcoGII and the extremely high error rate in calling modified bases. A back of the envelope calculation would suggest that only that only 40% of the A position that are accessible would produce signal (50% methylation rate times ∼20% base call error rate; both parameters which are not fully known or characterized).



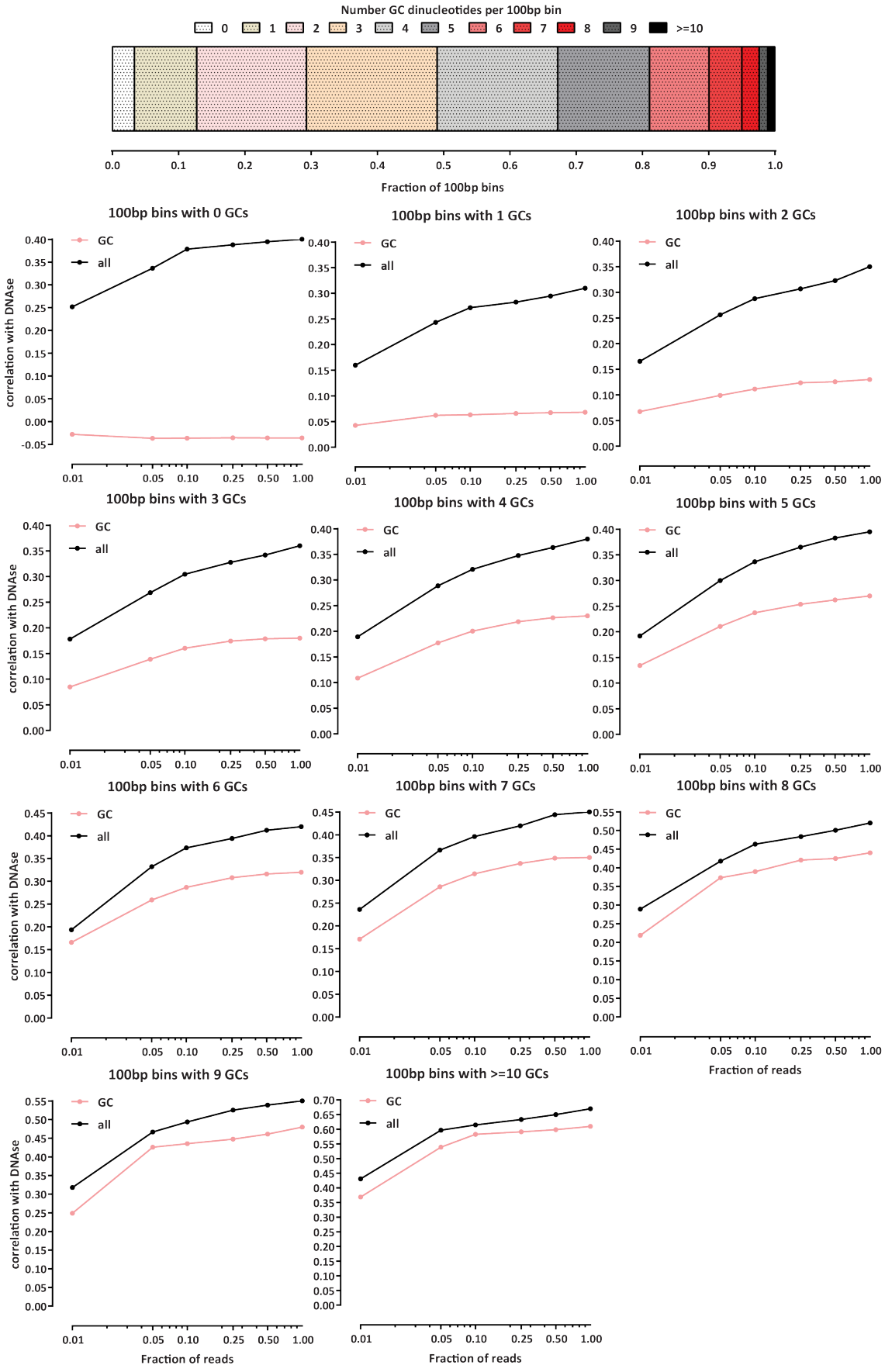
Response To Reviewers Figure 3: Impact of the addition of m^{6}A on assay resolution. Shown are Spearman correlation values between average methylation calls and smoothed (over 10bp) DNAse-seq tracks for each position in the yeast genome (without filtering out positions that are not uniquely mappable) for different subsamplings of SMAC-seq reads. Due to the sparseness of GC dinucleotides in the genome results, using GC methylation alone captures the accessibility signal much more poorly than what is enabled by the dense coverage provided by m^{6}A, as also shown above in Supplementary Figures 18–47.

As was discussed in the previous version of the manuscript, we believe that the methylation rate for EcoGII is greater than 50%. The reported methylation efficiency by NEB is 50% after 5 minutes of treatment, but is reported to increase to 85% with more prolonged incubation. In the SMAC-seq protocol applied here, we use ∼22 minutes of incubation, therefore we anticipate the methylation efficiency of accessible chromatin to be greater than the 50% that the reviewer is suggesting. In addition, the overall error rate we report of 20% (that we estimate from naked DNA and comparisons to dSMF) already includes both the enzymatic inefficiency and the base-calling errors.

Perhaps the easiest way to quantify the signal to background level of m^{6}A is to look at the modulation depth we observe around well positioned nucleosome sites (Figure 1e). This plot depicts the raw signal obtained per base probed around well positioned nucleosomes and thus the depth of modulation of our signal is a global measure of our per-base signal compared to background. This modulation depth, and thus signal to background, is similar for CpG/GpC and m6A methylation states. Therefore the average signal generated from an individual m^{6}A site across all well positioned nucleosomes is similar to that generated by a CpG or GpC site. However, because m^{6}A sites occur much more frequently (on average, in *S. cerevisiae*, 1 in 34 bases are CpGs, 1 in 26 are GpC and 1 in 3 are As), we are able to obtain signal at far more locations across the genome. Therefore with similar signal to background characteristics to other methods for probing chromatin accessibility at nucleosome linkers, we are able to obtain accessibility signal from approximately an order of magnitude more probe sites when compared to GpC methylation. In our minds (and in the minds of the other reviewers), this improvement is clearly a significant advance.

> Here, the authors could try to calibrate the error rates using PacBio sequencing (that can also detect m6A), where each individual template is read multiple times (vs once on a nanopore sequencer). Although I am not sure how that is going to profoundly improve the analysis.

We did consider using PacBio sequencing when initiating the development of SMAC-seq. However, because of its more limited read lengths and much more limited ability to read arbitrary modifications, PacBio is not a suitable platform for developing long-read assays for mapping chromatin structure. In the end we agree with the reviewer that this addition would not provide substantial improvements to our analysis.


Number GC dinucleotides per 100bp bin
0
1
2
3
4
5
6
7
8
9
>=10
Fraction of 100bp bins
100bp bins with 0 GCs
100bp bins with 1 GCs
100bp bins with 2 GCs
100bp bins with 3 GCs
100bp bins with 4 GCs
100bp bins with 5 GCs
100bp bins with 6 GCs
100bp bins with 7 GCs
100bp bins with 8 GCs
100bp bins with 9 GCs
100bp bins with >=10 GCs
GC
all
correlation with DNAse
Fraction of reads

> The low sensitivity of detection is particularly problematic for downstream interpretation because it lends itself to a very low signal to noise ratio. While the authors provide a correlation analysis between the signal specifically within DNaseI or ATAC-seq peaks, they do not address how much signal is localizing to these regions vs. non-accessible regions. Figure 1 From the browser shots provided in the manuscript indicate a seems to be very low enrichment within regulatory DNA regions (interpretation of Figure 1, Supplementary Fig 80). Given the authors claim about mapping accessible chromatin (e.g. title: "Long-range single molecule mapping for chromatin accessibility...") such analysis should be performed.

We may not fully understand the reviewer's concern here, and would suggest there may be some misunderstanding of the nature of methylation-based assays for mapping chromatin accessibility leading to a belief that SMAC-seq signal ought to be identical to ATAC-seq or DNAse-seq data. However, unlike ATAC and DNase, which enrich strongly for accessible, nucleosome-depleted chromatin – i.e. focal regions of perhaps 250–500 bases separated by large regions of inaccessibility – SMAC-seq provides a readout of both histone linker accessibility (i.e. regions between nucleosomes that occur across most nucleosomal DNA) as well as regions of nucleosome-depleted chromatin, and it does so along the length of chromatin fibers. The information obtained is a combination of what both the ATAC/DNAse-seq and MNase-seq assays provide.

Because the nature of the SMAC-seq signal includes methylation of linkers, and because what is being read out is the accessibility state of the population of chromatin fibers (rather than strongly enriching for accessible regions through size-selection or other means), we do not expect a strong concentration of signal only in regions that are DNAse-hypersensitive, as we believe the reviewer is suggesting. We expect to obtain enrichment in DNAse-hypersenstive regions, but also enrichment in nucleosome linker regions, with a signal distribution that ranges from 0 to 1 but with non-zero background levels, reflecting background accessibility levels (e.g. see *bioRxiv* 639971 for an orthogonal way of measuring absolute occupancy/accessibility levels).

In other words, the browser snapshots the reviewer refers to contain both small peaks corresponding to nucleosome linker segments and large peaks corresponding to open chromatin regions; the large peaks do not stand out as much as they would if the small peaks were not there. This, however, is a feature of SMAC-seq, not a weakness, because both positioned nucleosomes and accessible regions can be assayed in the same dataset. In effect, we would respectfully submit that the reviewer is measuring this method with an inappropriate yardstick.

Indeed all of the above mentioned characteristics are general features of methylation-based assays. NOMe-seq, which relies on a bisulfite sequencing readout, and has a very high per-base signal-to-noise ratio, exhibits very similar background accessibility levels to what we observe.

> Because the sensitivity is so low, it is really unclear what novel chromatin biology this would enable. The potential for long reads to illuminate the chromatin structure of individual templates over kilobases of genome is tantalizing. However, the authors do not provide any convincing analysis that demonstrates the interpretability of individual reads and mostly rely on aggregate plots over hundreds of templates.

We respectfully submit that the examples already provided in the text, such as rDNA accessibility and anticorrelation with an accessible element (See Figures 3 and 5) are indeed a demonstration of exactly what the reviewer requests. The identification of distinct states fundamentally requires the clustering of individual single molecules such that patterns (i.e. distinct, highly occupied/accessible states) can be discerned (a single read in isolation would not be sufficient to identify a significantly occupied/accessible state even with perfect, absolutely accurate base calling). But the fact that clustering is used does not imply that the individual traces are no longer reporting chromatin structure of individual templates over kilobases of the genome, or that these states could have been found

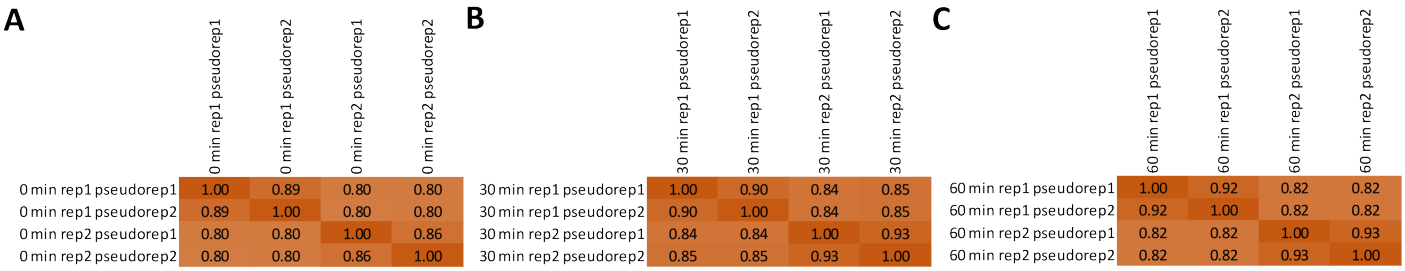
---

Response To Reviewers Figure 4 *(preceding page)*: Impact of the addition of m^6A on assay resolution. Shown are Spearman correlation values between average methylation calls and smoothed (over 10bp) DNAse-seq tracks for each position in the yeast genome (without filtering out positions that are not uniquely mappable) for different subsamplings of SMAC-seq reads and for different 100 bp bins of the genome, separated by their density of GC dinucleotides. Due to the sparseness of GC dinucleotides in the genome, using GC methylation alone captures the accessibility signal much more poorly than what is enabled by the dense coverage provided by m^6A, as also shown above in Supplementary Figures 18–47. At high density of GC dinucleotides, the GC alone and the SMAC-seq readouts begin to converge. However, only a small fraction of the genome has such properties, thus m^6A is overall essential for proper capture of accessibility patterns genome-wide.

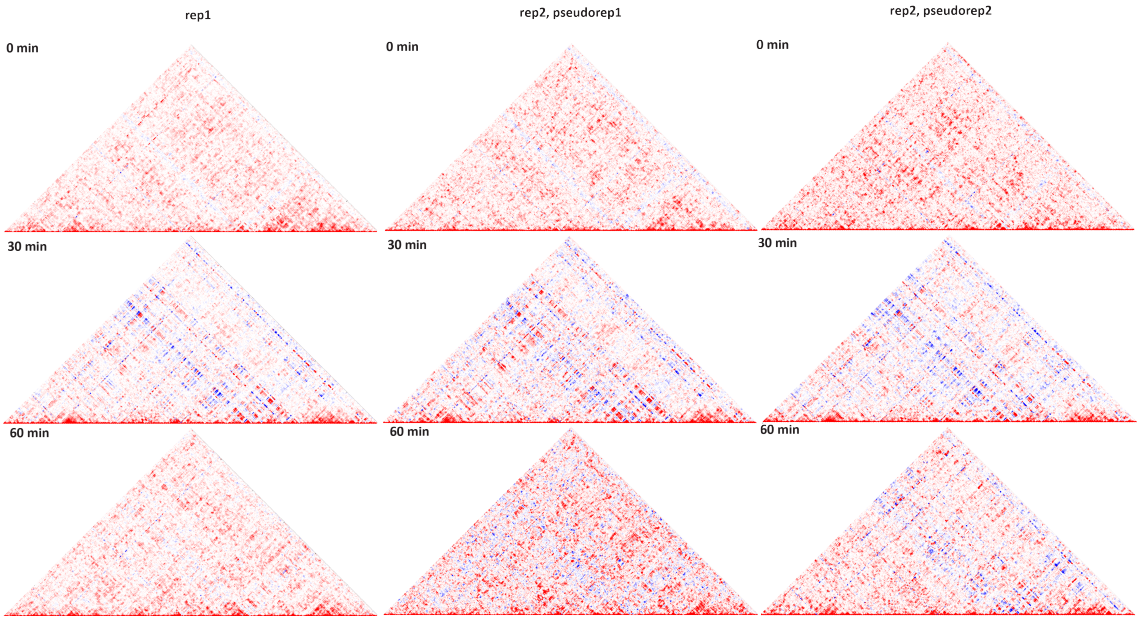
using bulk methods (indeed the single molecule nature of the data is absolutely required to parse these states). It unfortunately remains unclear to us why the reviewer does not accept this application as the other reviewers readily do.

> Finally, I am still concerned about the reproducibility and overall robustness of SMAC-seq. While the authors performed some replicates, they did not perform any meaningful analysis looking at concordance of signal between them (despite reviewers). This is surprising given the aforementioned challenges associated with methylation activity and base calling.

To further quantify the reproducibility of our data, we have added additional supplementary figures assessing the reproducibility of SMAC-seq signal between biological replicates and between pseudoreplicates (Supplementary Figures 50 and 106, also reproduced here as Response To Reviewers Figures 5 and 6). These analyses show high degree of correlation at the individual base pair level ($r^2 \geq 0.8$) between biological replicates, only slightly lower than what we see in pseudoreplicates. We also observe similar coaccessibility patterns across replicates, especially at the peak of our diamide response time course (30 min).



Response To Reviewers Figure 5: Correlation at the base pair level between replicates and pseudoreplicates. Shown are the Pearson r^2 values for the average methylation calls for each position in the yeast genome between pseudoreplicates (generated by randomly splitting reads in two halves) of the same and different biological replicates (tracks generated as shown in Figure 1).



Response To Reviewers Figure 6: Coordinated changes in chromatin accessibility and nucleosomal occupancy during the yeast stress response around the *HSP82* gene. Shown are NMI profiles in the vicinity of the *HSP82* gene for the first replicate of the diamide time course as well as for pseudoreplicates (generated by randomly splitting reads in two halves) from the second replicate of the diamide time course.