nature portfolio

Peer Review Information

Journal: Nature Methods
Manuscript Title: DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide
Corresponding author names: Aaron F. Straight, Aaron Streets

Reviewer Comments & Decisions:

Decision Letter, initial version:

Subject: Decision on Nature Methods submission NMETH-A46738
Message:

14th Sep 2021

Dear Aaron,

Your Article, "DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide", has now been seen by 4 reviewers. As you will see from their comments below, although the reviewers find your work of considerable potential interest, they have raised a number of critical concerns. We are interested in the possibility of publishing your paper in Nature Methods, but would like to consider your response to these concerns before we reach a final decision on publication.

We therefore invite you to revise your manuscript to fully address these concerns, particularly the concern about low methylation efficiency and sequence coverage.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised paper within 8 weeks. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please update your reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

IMAGE INTEGRITY
When submitting the revised version of your manuscript, please pay close attention to our href="https://www.nature.com/nature-research/editorial-policies/image-integrity">Digital Image Integrity Guidelines. and to the following points below:

-- that unprocessed scans are clearly labelled and match the gels and western blots presented in figures.
-- that control panels for gels and western blots are appropriately described as loading on sample processing controls
-- all images in the paper are checked for duplication of panels and for splicing of gel lanes.

Finally, please ensure that you retain unprocessed data and metadata files after publication, ideally archiving data in perpetuity, as these may be requested during the peer review and production process or after publication if any issues arise.

DATA AVAILABILITY
We strongly encourage you to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific and community-recognized repositories; a list of repositories is provided here: http://www.nature.com/sdata/policies/repositories

All novel DNA and RNA sequencing data, protein sequences, genetic polymorphisms, linked genotype and phenotype data, gene expression data, macromolecular structures, and proteomics data must be deposited in a publicly accessible database, and accession codes and associated hyperlinks must be provided in the "Data Availability" section.

Refer to our data policies here: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data

To further increase transparency, we encourage you to provide, in tabular form, the data underlying the graphical representations used in your figures. This is in addition to our data-deposition policy for specific types of experiments and large datasets. For readers, the source data will be made accessible directly from the figure legend. Spreadsheets can be submitted in .xls, .xlsx or .csv formats. Only one (1) file per figure is permitted: thus if there is a multi-paneled figure the source data for each panel should be clearly labeled in the csv/Excel file; alternately the data for a figure can be included in multiple, clearly labeled sheets in an Excel file. File sizes of up to 30 MB are permitted. When submitting source data files with your manuscript please select the Source Data file type and use the Title field in the File Description tab to indicate which figure the source data pertains to.

Please include a “Data availability” subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: “The data that support the findings of this study are available from the corresponding author upon request”, describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY

Please include a “Code Availability” subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement “available upon request” allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

SUPPLEMENTARY PROTOCOL

To help facilitate reproducibility and uptake of your method, we ask you to prepare a step-by-step Supplementary Protocol for the method described in this paper. We encourage authors to share their step-by-step experimental protocols on a protocol sharing platform of their choice and report the protocol DOI in the reference list. Nature Research's Protocol Exchange is a free-to-use and open resource for protocols; protocols deposited in Protocol Exchange are citable and can be linked from the published article. More details can found at www.nature.com/protocolexchange/about.

ORCID

Nature Methods is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Best regards,
Lei

Lei Tang, Ph.D.
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Despite the long history of ChIP and related methods to map factors on chromatin, there is still space for further improvements. For example, the recent development of CUT&Run and CUT&Tag substantially improved the scalability of antibody-based binding site detection. Here the authors describe what is essentially a reboot of DamID, using antibody-directed methylation to study the binding of the factors in chromatin. The key advance here is (1) the integration of directed methylation with updated in situ approaches (like CUT&Run) with (2) native long-read sequencing (nanopore) to determine sites of methylation to infer sites of binding.

Overall, the authors do a nice job arguing for the promise of their approach and I am convinced that native long read sequencing is the future for analyses of complex genomic regions of this sort. More specifically, the authors nicely outline a number of applications one could imagine for this approach (see Figure 1) in the introduction and discussion. I'm generally excited about this work and am impressed by the effort the authors put into this system.

I have the following major concerns:

1. The authors' claim of proportionality. The authors argue that their direct sequencing provides true proportionality (unlike ChIP etc.) because it does not rely on PCR amplification (ex lines 38-39, 230-245, 511 and Figure 1b). But techniques like ChIP/DamID do not provide true proportionality for reasons besides amplification, such as epitope masking and general chromatin accessibility. Indeed, the authors see this type of bias in their own data in the form of footprints of nucleosomes/TFs as well as the observation that even the untagged pA-Hia5 control shows 6-fold preference for regions of CTCF binding (line 280), demonstrating that there are many factors influencing reactivity. The authors acknowledge this as an issue on line 564, but do not weaken their strong statements about proportionality. Finally, while the new approach clearly correlates with previous data and shows a linear relationship that the authors use to form their conclusion, the degree of agreement is not very impressive (Figure 3e).
2. The authors claim that they can measure heterogeneity at the single cell level. While long read sequencing holds the promise of this type of analysis, the system must work well to interpret individual reads (or small clusters of reads) in practice. In fact, the authors acknowledge that they are not quite able to confidently measure instances of multiple CTCF binding on the same read (line 354-355), which is the type of event one would like to be able to measure. Another place where this type of single read analysis is important to the conclusions of the paper is the CENP-A data. The authors interpret their

CENP-A findings to suggest that there is not a stereotyped arrangement of CENP-A nucleosomes across cells, but it is unclear how much of the heterogeneity the authors observe is technical versus biological. Would homogenous positions appear homogenous in their assay? In principle the authors could have supported their interpretation using the system depicted in Figure 2 had this been a better-defined and relatively homogeneous system. The chromatin appears quite heterogenous (a range of different numbers of nucleosomes per DNA), so once again the heterogeneity observed in Figure 2 could be due to biochemical heterogeneity in the nucleosomes, or it could be due technical artifacts in the assay. (The authors should show the data referenced in line 769 characterizing the saturation of these arrays with AvaI). In Figure 2d and e and Figure S2h and I, how do the authors interpret the reads in the center of the heat map with nearly all blue? For Figure 2d, if they are CA chromatinized, we would expect protection. If they are not CA chromatinized, they why are they getting methylated? These are the most extreme examples, but the bigger point is that even with clustering, it is hard to assign and biochemically interpret the reads, yet for one to make statements about biological heterogeneity this interpretation must be robust. The aggregated reads look good, but it is not clear if the authors have the system working well enough to interpret individual reads or small clusters of reads. So, am not convinced by the authors' conclusion that these data confirm "that DiMeLo-seq is capable of profiling heterogeneity in protein-DNA interactions at the single cell level."

3. How well is this approach actually working? While the authors touch on many of the key metrics, they are inconsistent and leave some room for concern. For example, the authors at times use only highly confident calls of 0.9 for the mA (line 257), and then other times 0.6 (line 348). The authors claim a resolution of about 200 bp for their approach, but then reference a range that is approximately 400 bp (line 338). Furthermore, they claim the data drops off from target sites over about 75 bp (line 137), but with on-target methylation rates of only 0.2-0.4%, it isn't clear this number of As in 75 bp would be sufficient to give the stated sensitivity values (up to ~50%, line 521). The metrics should generally be explained more clearly (which was easy to follow only for the cLAD analysis, and even there it could be better in lines 258-261 and in the methods).

Other points:

4. The results presented in Figure 2b are concerning: how does the third lane, which lacks Hia5, have such high mA signal? Taken as the authors present it, this is a major red flag. I'm tempted to assume this is a mistake in labeling, but if so, can the other labels be trusted?

5. Is the level of density of data in Figure 2c for the H3 chr + CA-directed pA-Hia5 accurately depicted? It seems like that bar would need to be much higher given the description of how this plot was made. If the authors are cutting data from the figure without clearly indicating it in the figure/legend, that is of course a major issue.

6. The sentence on line 310 "The in-phase…" is long, complex, and not straight forward to parse.

7. The authors claim they can use this approach to observe single molecules up to hundreds of kb in length (line 507), which is cool. Where is the data supporting this? More generally, are the reads for this approach shorter on average than other ONP approaches? It would be nice to see some histograms rather than just a summary table.
8. Line 574 “The method is also compatible with in vivo expression…” this either needs to be demonstrated or the wording changed to clarify that this is just in theory.
9. I’m confused why the free pA-Hia5 has such dramatically higher activity with these substrates. In general, one of my biggest concerns is whether the efficiency of directed methylation is sufficient. The footprints observed when chromatin is treated with free Hia5 are beautiful (Figure S2d and e). By contrast the directed activity is much weaker and seems to define the sites of CA chromatin relatively poorly. Is it the lower local concentration of enzyme, or that the fusion has lower activity in general?

Reviewer #2:
Remarks to the Author:
The manuscript “DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide” describes a chromatin profiling method that combines antibody tethering of a DNA methyltransferase with long-read sequencing. This method builds on previous work using untargeted methyltransferases to footprint DNA-bound proteins and histones in chromatin on extended chromatin reads. The innovation of tethering the methyltransferase gives information on footprinting when a particular chromatin protein is bound. The authors report extensive optimizing of the method for cells, and apply this to in vitro nucleosome arrays, to lamin-associated domains, to the insulator protein CTCF, and to the centromeric histone CENPA. While this is an interesting method with high potential, it is not clear that it currently has sufficient coverage.

This issue first arises on line 354: “While the lack of signal at a site may be the result of the sensitivity of our assay rather than the vacancy of a CTCF site, this analysis demonstrates the potential to analyze coordinated binding patterns on single molecules.” This is not demonstrated; at the current methylation density it does not appear that coordinated binding can be inferred. The sensitivity of the assay may improve with higher methylation density, but can more methylation be achieved with the tethered strategy?

Similarly, in characterizing distribution of CENPA nucleosomes in centromeres, it seems that the sparse overall methylation may be mis-interpreted as the absence of a centromeric nucleosome. On line 487, they state ”It is important to note, however, that an absolute quantification of CENP-A nucleosome

density in centromeres will require further characterization of the single-molecule sensitivity of DiMeLo-seq in these regions." This seems like a major limitation that should either be explained or remedied. They also describe heterogeneity of centromeric nucleosomes within an HOR, but could this heterogeneity just be an artifact of sparse methylation?

Reviewer #3:
Remarks to the Author:
In this manuscript by Altemose, et al, a new method, DiMeLo-seq, for long-read based mapping of protein-DNA interactions at genome-scale is introduced. The authors present this method as an advance over current methods that rely on short read data and PCR-based seq library amplifications thereby limiting their utility in studying repetitive sequences and DNA modifications in a truly quantitative manner. In this method, antibody directed DNA adenine methylation is achieved by fusing Protein A to a nonspecific deoxyadenosine methyltransferase (Hia5) to catalyze mA in DNA "proximal to the targeted chromatin-associated proteins". The authors present several case studies to demonstrate the overall utility of this approach, supporting the study with extensive variable tests. This technique is an important and interesting advance in studying DNA-protein interactions, taking advantage of long-read sequencing (ONT) and the newly released human genome assemblies. I have a few minor comments that would increase the utility of this approach to a broader audience of scientists.

In the introduction the authors mention "proximal". Can the authors provide an estimated range to guide scientists in choosing this technique? More specifically, how much direct binding domain information is lost?

The authors state that a large footprint for binding for LMNB1 means lower coverage for sequencing – how does this impact the CENP-A assay on 601? A demonstration of exactly how to estimate targeted coverage would be powerful for the reader interesting in applying this technique to determine how much sequencing may be required for a particular DNA-Protein interaction assessment. The authors should provide a method in the main part of the paper to estimate adequate coverage before sequencing. How does error correction/error rates of the different basecallers for modified (and unmodified) nucleotides impact this coverage estimate? How to take such information into account when estimating targeted coverage would be useful for the reader as the tools and flow cell chemistry continue to improve.

Steric hinderance is brought up a few times, particularly with respect to CTCF with 3' cohesin – would this apply to other targeted proteins and thus present a limitation, or perhaps provide an added benefit to demonstrate immediate protein-protein interactions at specific loci? Determining whether this observed bias to 5' of the binding motif of CTCF is due to 3' cohesin OR the antibody binding location (3' or 5') is important to determine to appropriately interpret the data. Is there an antibody for CTCF that binds 3' to test the influence of antibody binding site vs local protein environment? Given that such an antibody may not be available, this is merely a minor comment rather than a major requirement for publication.

Can the authors provide an estimate of how much the input requirement is reduced with the use of concanavalin-A coated beads? Along these lines, can the authors estimate whether efficiencies are impacted by lightly fixed vs frozen cells? It is hard looking at all of the variables in Supp Tables 1 and 2 to assess the impact of each variable on projected success rates, particularly since there is a range of sequencing depth across samples.

Other minor comments:

Figure 2d inset – should be larger (or include a zoom of 0-400, for example) to illustrate the location of the dyad position

Figure 2e – is this single reads or an aggregate? Not sure why some of the red circles are faded – if it is a single read, it should be there or not – correct? Or is this noise in the data due to the basecalls?

Lines 351-358 – Figure 4d – for clarity for the reader, can the authors indicate on the figure (by clustering) which molecules represent CTCF at both sites, one site or neither site?

Lines 34-36 – the authors should add an example (or prospective use) of why assessing DNA-protein interactions and DNA methylation simultaneously provides an advantage (for a broader scientific audience.)

Line 134: spell out "1"

Lines 226-7 – "also"…"as well" is redundant in the same sentence

Lines 843-845… "as described in figure." The figure number is missing.

Reviewer #4:
Remarks to the Author:
In their manuscript “DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide”, Altemose et al. outline the development, optimization, and validation of the DiMeLo-seq method. DiMeLo-seq relies on antibody directed methylation of Adenine which can be read out by Oxford Nanopore Technologies sequencers. This is an exciting advance over previous methods that used similar but untargeted strategies to map open chromatin. The experiments are well described and the detailed description of the methods optimization provides an interesting insight into how sensitivity and specificity of DiMeLo-seq were dialed in. Further, the development of the AlphaHOR-RES enrichment strategy could provide a useful tool for the sequencing of centromeric regions in general.

In experiments targeting Lamin B1, CTCF, and H3K9me3 in a range of different human cell lines, the authors clearly outline the advantages and disadvantages of using long reads for this type of analysis over established short read methods like CUT&RUN or ChIP-seq.

Overall this is a strong and very interesting manuscript and approach. The few main concerns I have are all linked to the preferential modification of open chromatin by the Hia5 methylase and the relatively low sequencing coverage achieved in some of the experiments.

Major points:

CTCF binding analysis: The authors show that CTCF peaks show a 22-fold enrichment when a methylase targeted by an antibody is used but also a 6-fold enrichment when a free methylase is used to methylate DNA. The authors reasonably argue that this is due to CTCF peaks generally falling within open chromatin.
Because of this however, it might be challenging to define CTCF binding peaks using only DiMeLo-seq data. I would therefore ask for two or three additions to this analysis:

1) In addition to comparing mA/A within CTCF peaks vs outside of CTCF peaks, I would like to see the addition of mA/A in ATAC seq peaks (i.e. open chromatin) that don’t overlap CTCF peaks. Since this experiment was performed in GM12878, I am 99% confident ATAC-seq data for this cell line already exists.

2) If the read coverage is anywhere close to high enough to do this, I would like the authors to attempt to call CTCF peaks based on DiMeLo-seq data alone and compare how these peaks relate to CTCF ChIP-seq peaks. Whatever method the authors chose to do this and how successful this ends up being will give the reader an idea whether DiMeLo-seq will be valuable in enriching short read data sets for now or whether it can already be used as a stand-alone assay. The same is true for the H3K9me3 data, i.e. can you determine enrichment areas, aka peaks from DiMeLo-seq data alone.

3) If the coverage isn't high enough to do any type of peak calling, would it be feasible to generate this data by using a PromethION run on the CTCF targeted DNA? You could get ~30x coverage out of a single run which could really help gaining a deeper understanding of the data. I really think this is highly optional though since asking for additional experiments is bad form.

Minor points:

1) Figure 4:
- Color choice. While blue and orange are a good combination in general (and are maintained throughout the manuscript which is great), the orange doesn't come through on a computer screen so the right side of Fig. 4c is basically just an off-white rectangle. Maybe increase the intensity of the individual points or increase the contrast in some other way to make the underlying data visible.
- Legend inserts. The legend inserts are given without units. "A" content for example is given as a color gradient from 1000-10000 but it is not clear from the legend panel or the figure legend what that actually means. More information would be appreciated.

2) This is a subjective thing so I would leave this up to the authors but I think adding a short sentence for each paragraph stating how many reads across how many replicates were generated for each experiment, what average lengths they are and what genome coverage that corresponds to would add context to the experiments. All this data is currently in the supplement so you have to go looking for it.

Author Rebuttal to Initial comments

We thank the reviewers for their constructive feedback and suggestions to improve the manuscript. We have addressed all of the concerns of the referees with new data, analysis and changes to the text. Major text revisions are highlighted in the revised manuscript. Please find below our point-by-point responses (in black) to the reviewer's comments (copied below in

blue, *italics*).

Reviewer #1:

Remarks to the Author:

Despite the long history of ChIP and related methods to map factors on chromatin, there is still space for further improvements. For example, the recent development of CUT&Run and CUT&Tag substantially improved the scalability of antibody-based binding site detection. Here the authors describe what is essentially a reboot of DamID, using antibody-directed methylation to study the binding of the factors in chromatin. The key advance here is (1) the integration of directed methylation with updated in situ approaches (like CUT&Run) with (2) native long-read sequencing (nanopore) to determine sites of methylation to infer sites of binding.

Overall, the authors do a nice job arguing for the promise of their approach and I am convinced that native long read sequencing is the future for analyses of complex genomic regions of this sort. More specifically, the authors nicely outline a number of applications one could imagine for this approach (see Figure 1) in the introduction and discussion. I'm generally excited about this work and am impressed by the effort the authors put into this system.

We thank the reviewer for their enthusiasm for our work and for their very careful reading of our manuscript and helpful feedback.

I have the following major concerns:

1. The authors' claim of proportionality. The authors argue that their direct sequencing provides true proportionality (unlike ChIP etc.) because it does not rely on PCR amplification (ex lines 38-39, 230-245, 511 and Figure 1b). But techniques like ChIP/DamID do not provide true proportionality for reasons besides amplification, such as epitope masking and general chromatin accessibility. Indeed, the authors see this type of bias in their own data in the form of footprints of nucleosomes/TFs as well as the observation that even the untagged pA-Hia5
control shows 6-fold preference for regions of CTCF binding (line 280), demonstrating that there are many factors influencing reactivity. The authors acknowledge this as an issue on line 564, but do not weaken their strong statements about proportionality. Finally, while the new approach clearly correlates with previous data and shows a linear relationship that the authors

use to form their conclusion, the degree of agreement is not very impressive (Figure 3e).

We thank the reviewer for raising this point about proportionality. We agree with the reviewer that like any method, ours is also subject to variability resulting from accessibility, epitope masking and other experimental biases. Our intention with the discussion of proportionality is to point out that because the method is a single-molecule measurement, the proportion of reads that contain a binding event can be used to estimate the proportion of cells that contain the binding event more accurately than methods that rely on PCR. We agree that there are other factors that prevent direct proportionality. For example, because of the imperfect sensitivity of this single-molecule method, the slope of the line relating % of molecules methylated to % of molecules bound is less than 1, as illustrated in our Fig. 3e. However, because DiMeLo-seq involves no amplification, one does not have to account for systematic error due to amplification bias or for the high levels of random error due to phenomena like runaway amplification common in nonlinear amplification workflows. To support this claim, we compared the correlation in Figure 3e to a similar plot using coverage data from bulk DamID, which relies on PCR (new Supplementary Fig. 5f). This analysis demonstrates how DiMeLo-seq's
amplification-free, single-molecule approach correlates more strongly with single-cell measurements than does bulk DamID's amplification-based approach. We have also revised the text referred to above and Figure 3 to soften the claims of proportionality and focus more on the improved ability of DiMeLo-seq for estimation of single-cell interaction frequency. The revised text is copied below (starting at line 308).

"Because DiMeLo-seq directly probes unamplified genomic DNA, each sequencing read represents a single, native DNA molecule from a single cell, sampled independently and with uniform probability from the population of cells. This allows for estimation of absolute protein-DNA interaction frequencies, i.e. the proportion of cells in which a site is bound by the target protein, without needing to account for the amplification bias inherent to other protein-DNA mapping methods. … This revealed a nearly linear relationship between the two interaction frequency estimates, with a simple linear model achieving an R^2 of 0.71, compared to an R^2 of 0.31 when scDamID-based interaction frequencies are compared to bulk conventional DamID coverage (Fig. 3e, Supplementary Fig. 5f)."

2. The authors claim that they can measure heterogeneity at the single cell level. While long

read sequencing holds the promise of this type of analysis, the system must work well to interpret individual reads (or small clusters of reads) in practice. In fact, the authors acknowledge that they are not quite able to confidently measure instances of multiple CTCF binding on the same read (line 354-355), which is the type of event one would like to be able to measure. Another place where this type of single read analysis is important to the conclusions of the paper is the CENP-A data. The authors interpret their CENP-A findings to suggest that there is not a stereotyped arrangement of CENP-A nucleosomes across cells, but it is unclear how much of the heterogeneity the authors observe is technical versus biological. Would homogenous positions appear homogenous in their assay? In principle the authors could have supported their interpretation using the system depicted in Figure 2 had this been a better-defined and relatively homogeneous system. The chromatin appears quite heterogenous (a range of different numbers of nucleosomes per DNA), so once again the heterogeneity observed in Figure 2 could be due to biochemical heterogeneity in the nucleosomes, or it could be due technical artifacts in the assay. (The authors should show the data referenced in line 769 characterizing the saturation of these arrays with AvaI). In Figure 2d and e and Figure S2h and I, how do the authors interpret the reads in the center of the heat map with nearly all blue? For Figure 2d, if they are CA chromatinized, we would expect protection. If they are not CA chromatinized, they why are they getting methylated? These are the most extreme examples, but the bigger point is that even with clustering, it is hard to assign and biochemically interpret the reads, yet for one to make statements about biological heterogeneity this interpretation must be robust. The aggregated reads look good, but it is not clear if the authors have the system working well enough to interpret individual reads or small clusters of reads. So, am not convinced by the authors' conclusion that these data confirm "that DiMeLo-seq is capable of profiling heterogeneity in protein-DNA interactions at the single cell level."

We thank the reviewer for this important comment about single-molecule sensitivity. We agree that distinguishing between technical and biological variability is critical for interpretation of single-molecule and single-cell heterogeneity. To address this concern we have returned to our reconstituted chromatin system with the goal of producing a more homogeneous system, where instead of creating arrays of 18x601 sequence we have created a 730 bp template with 1x601 at the center. We have purified this mononucleosomal construct using glycerol gradients after digestion with BsiWI, which will cut any naked DNA without a nucleosome assembled. Using this purification strategy we have confidently isolated a homogeneous population of nucleosomes that we can use to measure our assay's sensitivity. Using this substrate with

free-floating pA-Hia5 we measure 97% of sequenced molecules to have nucleosomes. When we perform DiMeLo-seq on this substrate targeting for CENP-A nucleosomes we identify nucleosomes on ~65% of molecules. Importantly when we use IgG targeting we identify ~5% of molecules to contain nucleosomes. Therefore we estimate our *in vitro* sensitivity to be 65% with a FDR of 5% (Fig 2a-d and Supplementary Fig 2c-e 2). With these measurements we have adjusted our claims about CENP-A nucleosome density in the text (lines 563 - 575). We acknowledge that the sensitivity and specificity of DiMeLo-seq and signal-to-background ratios depend on the choice of thresholds of methylation probability scores for calling modified bases as well as abundance of detected methylation. We have included ROC curves to describe this dependence (Supplementary Fig. 2g,h). Based on the on-target methylation rate of 65% for CENP-A-directed methylation on mononucleosomes *in vitro*, we have now removed the discussion of the apparent lack of stereotyped patterns of nucleosome positions *in vivo* (please also refer to our response to Reviewer 2's third comment).

In addition we have now added data from native gel electrophoresis after digestion with AvaI that cuts 18x601 arrays into mononucleosomes as shown in Supplementary Figure 3c and referenced previously in line 769, now line 988. This digestion-based assay demonstrates that the 18x601 reconstituted chromatin arrays are not 100% saturated; however, we cannot identify the number or positions of nucleosome-bound 601 sites using this assay as we can using our free-floating pA-Hia5 and DiMeLo-seq assay.

Finally, in response to the comment regarding highly methylated reads in Fig. 2d, we think that these hypermethylated reads likely result from off-target or background methylation of non-chromatinized DNA. When we cluster reads in our plots by the position of methylated A's, these hypermethylated reads cluster together and appear to be prevalent, but they make up < 2% of reads in DiMeLo-seq libraries and only appear in experiments involving the 18x601 array DNA, and never in our *in situ* experiments. We have now updated the figures (Fig. 2f,g, Supplementary Fig. 3g,h,k,l) showing individual reads clustered based on inferred nucleosome positions reflecting the updated binning and thresholding based on our single-nucleosome experiments (1x601 experiments). Any read showing methylation on > 60% of a 400 bp region centered at the 601 dyad is inferred to not have a nucleosome at that dyad position to account for hypermethylated regions/reads while clustering reads hierarchically (Methods, lines 1100 - 1105). This is consistent with the maximum % methylation observed on free pA-Hia5 reads from 1x601 chromatin (Supplementary Fig. 2k).

3. How well is this approach actually working? While the authors touch on many of the key

metrics, they are inconsistent and leave some room for concern. For example, the authors at times use only highly confident calls of 0.9 for the mA (line 257), and then other times 0.6 (line 348). The authors claim a resolution of about 200 bp for their approach, but then reference a range that is approximately 400 bp (line 338). Furthermore, they claim the data drops off from target sites over about 75 bp (line 137), but with on-target methylation rates of only 0.2-0.4%, it isn't clear this number of As in 75 bp would be sufficient to give the stated sensitivity values (up to ~50%, line 521). The metrics should generally be explained more clearly (which was easy to follow only for the cLAD analysis, and even there it could be better in lines 258-261 and in the methods).

This is a good question. On one hand, the resolution and sensitivity of our method are critical performance parameters that are important for experimental design and interpretation, but on the other hand, these metrics are not fixed, and can vary depending on the target, the target's footprint, the antibody, and the analysis parameters that may, for example, be chosen to maximize sensitivity or specificity depending on the question. Nonetheless, we agree that it is important to clarify our choice of parameters and to be consistent when reporting the performance specifications of the method. To this end, we have made the following changes in the revised manuscript:

To address the question of modification probability thresholds, we have added a new methods section called "Modification calling thresholds" to discuss our considerations for choosing specific thresholds for certain tasks and explaining the consequences (lines 1056 - 1072). We have provided new plots estimating the false positive rates, false negative rates, and false discovery rates for different thresholds on (i) methylation probability scores (Supplementary Fig. 2f, 4b), and (ii) mA abundance for distinguishing on-target vs off-target methylation for LMNB1 (Supplementary Fig. 5b-e) ,CTCF (Supplementary Fig. 6h), and CENP-A detection on chromatin *in vitro* (Supplementary Fig. 2f-h). We now use a threshold of 0.6 for 50bp binned Megalodon methylation probability scores for detecting mA's on reads from *in vitro* treated chromatin with an FDR of 0.05 (false detection rate of detecting mA's on untreated DNA reads compared to free pA-Hia5 treated DNA reads) (lines 1083 - 1099). Using CENP-A mononucleosomes, we estimate an on-target methylation rate of 65% (Fig. 2d) (percentage of CENP-A DiMeLo-seq reads identified as methylated on at least 20% of bins at the threshold of 0.6 for 50 bp bins (Supplementary Fig 2g,h,k)). We no longer set a mA probability threshold of 0.6 for CTCF sensitivity calculations (previously line 348), and we instead maintain a consistent mA probability threshold of 0.75 for all CTCF Megalodon analysis. We note that the false positive rate for mA calls on unmethylated genomic DNA is extremely low across all mA

probability thresholds (Supplementary Fig. 4a-b), and that the major source of background noise in DiMeLo-seq likely comes from off-target methylation, rather than false-positive modification calls. Setting higher thresholds on mA probability scores is effectively selecting for regions with higher mA density. The choice of threshold for a particular protein depends on the desired sensitivity vs specificity, as higher thresholds effectively discard many real methyladenine calls. For LMNB1, which has an enormous binding footprint, a higher threshold was used because the resulting loss of sensitivity did not affect the ability to distinguish between reads from on-target vs off-target regions (Supplementary Fig. 5e), but it did provide a greater aggregate
on-target:off-target mA/A ratio across all reads (Supplementary Fig. 5c).

We added this text to the LMNB1 results section (starting on line 261):

"These performance metrics depend on the choice of mA score threshold (Supplementary Fig. 5c), which was chosen to balance sensitivity and specificity in distinguishing regions with on-target and off-target methylation. We note that this threshold does not primarily serve to reduce false-positive mA calls, which occur at an extremely low rate (Supplementary Fig. 4a,b; see full discussion of threshold evaluation in Methods). To confirm that this optimization would apply to other types of proteins, we also examined the results of different protocol variations targeting the protein CTCF and found them to be concordant (Methods, Supplementary Fig. 6a)."

Regarding the question of resolution, in the revised manuscript, instead of claiming a fixed "single-molecule" resolution of 200 bp, we focus the discussion on measurements that can help us to characterize resolution. For *in vitro* DiMeLo-seq, we report that ~70% of CENP-A-directed methylation on 1x601 CENP-A chromatin falls within 250 bp on either side of the 601 dyad at line 139. Similarly, we state only that the predicted CTCF peak center for approximately 70% of single molecules falls within +/- 200 bp of the CTCF binding motif center at lines 397- 400. While this "range" is 400 bp, because we know the location of the motif, we can conclude that 70% of the molecules contain peaks which are 200 bp from the motif center. Regarding sensitivity, the 0.2-0.4% mA detection rate that the reviewer is referring to comes from the LMNB1 optimization data generated from low coverage sequencing and a stringent mA probability threshold of 0.9.

In practice, we have measured this detection rate between 12% (as measured in the CTCF data, see figure 4a for example) and over 40% in our in vitro validation studies (see figure 2c for example).

Other points:

4. The results presented in Figure 2b are concerning: how does the third lane, which lacks Hia5, have such high mA signal? Taken as the authors present it, this is a major red flag. I'm tempted to assume this is a mistake in labeling, but if so, can the other labels be trusted?

We thank the reviewer for bringing up this point for clarification. This is not a mistake in labeling. The result presented in Supplementary Fig. 3e (previously Figure 2b) compares the intensity of anti-methyladenine immunofluorescence signal on chromatin-coated beads after incubation with primary antibodies and/or pA-Hia5 (and SAM), with the aim of testing the specificity of our antibody-guided methylation approach. In this experiment, CENP-A, H3, and IgG antibodies were of rabbit origin, and so was the anti-methyladenine antibody. Because of this, the fluorophore-conjugated anti-rabbit-IgG secondary antibody (which is intended to bind only to the anti-methyladenine antibody) could also bind to the anti-histone antibody (which was used to localize pA-Hia5). To minimize undesired cross-reactivity, we incubated the beads with 2M NaCl at 55 C for 1hr after methyltransferase activation (incubation with SAM) was complete, in order to denature chromatin and remove as much of the anti-histone primary antibody as possible prior to incubation of the DNA-coated beads with the anti-methyladenine antibody. To demonstrate that the high fluorescence signal we observe in lane 1 (CENP-A chromatin + CENP-A antibody + pA-Hia5) could not be explained by cross-reactivity, we include lane 3, which lacks pA-Hia5 and therefore should have no actual methyladenines. The low level of signal that we do see in lane 3 is explainable as secondary antibody cross-reactivity due to imperfect removal of the primary anti-CENP-A antibody. However, comparing lanes 1 and 3, it is clear that the high signal in lane 1 is specific to the presence of pA-Hia5. To avoid confusion, we have included a note in the methods (line 1035) explaining the interpretation of these signals in light of the low level of antibody cross-reactivity.

5. Is the level of density of data in Figure 2c for the H3 chr + CA-directed pA-Hia5 accurately depicted? It seems like that bar would need to be much higher given the description of how this plot was made. If the authors are cutting data from the figure without clearly indicating it in the figure/legend, that is of course a major issue.

We thank the reviewer for bringing up this point. The data is cut off in the histograms

displayed in 2C. We had adjusted the y axis to be able to better visualize differences between our experimental conditions. In our updated manuscript, we have kept the axis on this plot, but also included cumulative distribution functions to additionally illustrate the level of methylation in each sample and the differences between samples. In addition, we have updated the figure legend to reflect that the histogram is cut off at 20.

6. The sentence on line 310 "The in-phase…" is long, complex, and not straight forward to parse.

We agree this sentence needs to be simpler and have replaced this sentence at line 390 with "both A and CpG are preferentially methylated in linker DNA."

7. The authors claim they can use this approach to observe single molecules up to hundreds of kb in length (line 507), which is cool. Where is the data supporting this? More generally, are the reads for this approach shorter on average than other ONP approaches? It would be nice to see some histograms rather than just a summary table.

We have added Supplementary Figure 13 with read length histograms. Depending on the choice of cleanup method for the LSK110 library preparation described in the Methods section "Nanopore library preparation and sequencing," we achieved an N50 of ~20 kb with standard bead-based cleanup and ~50 kb with pelleting the DNA for cleanup. The read lengths for DiMeLo-seq are consistent with read lengths achieved with other ONT approaches. The library preparation, rather than the DiMeLo-seq protocol, is the main factor in determining read length. For AlphaHOR-RES, smaller fragment sizes are recovered compared to the standard
DiMeLo-seq protocol because the genome is digested with restriction enzymes and purified with column cleanups that are required to clean gel-extracted DNA.

8. Line 574 "The method is also compatible with in vivo expression…" this either needs to be demonstrated or the wording changed to clarify that this is just in theory.

We have now added an experiment demonstrating this capability. We created a stably transduced HEK293T cell line with an inducible EcoGII-LMNB1 fusion protein, and we have added the following text to the results and discussion sections. This experiment is now also included in Fig. 3b, Supplementary Figure 5a, Supplementary Table 1, and Methods.

Results (starting at line 282):

We also generated a stably transduced line expressing a direct fusion between EcoGII and LMNB1 *in vivo*, as in MadID (Sobecki et al. 2018), then we detected mAs with nanopore sequencing (Supplementary Fig 5). This *in vivo* approach produced threefold more on-target methylation compared to *in situ* DiMeLo-seq with pAG-EcoGII (Fig. 3b). This likely owes to the effectively longer incubation time during which methyl groups can be deposited on adenines *in vivo* (15 h) compared to *in situ* (2 h), as well as to chromatin dynamics in vivo that may make a greater fraction of the genome accessible to the methyltransferase (van Schaik et al. 2020). However, compared to *in situ* DiMeLo-seq with pA-Hia5, the *in vivo* EcoGII-LMNB1 approach produced 36% less on-target methylation and 25% more off-target methylation, although this *in vivo* performance is expected to vary with different fusion proteins and their expression levels.

Discussion (starting at line 686):
We also demonstrated that DiMeLo-seq can read out methyladenines deposited by *in vivo* expression of protein-MTase fusions, as in conventional DamID (van Steensel and Henikoff 2000) or MadID (Sobecki et al. 2018), instead of antibody targeting *in situ*. This may prove useful for investigating more transient protein-DNA interactions, or proteins that lack suitable antibodies, in cases where the biological system being studied can be readily genetically modified.

Methods (starting at line 870):
Creation and induction of stable cell lines for in vivo *DiMeLo-seq*
Stable HEK293T cell lines were created by retroviral transduction followed by drug selection. Retroviral plasmids containing DDdegron-EcoGII-V5linker-LMNB1 were obtained from Addgene (#122083; Sobecki et al. 2018). Retroviruses were produced in the Phoenix Ampho packaging cell line (obtained from the UC Berkeley cell culture facility). Phoenix cells were seeded in standard growth medium (DMEM with 10% FBS and 1X P/S) in a T75 flask 24 hours before transfection, aiming for 70% confluence at the time of transfection. 25 µg of plasmid DNA was combined with 75 µl FUGENE-HD transfection reagent in 1200 µl optiMEM and incubated for 10 minutes, then added to the media. After 12 hours, the media was replaced with fresh media, and the cells were incubated at 32 °C with 5% CO_2 and 100% humidity to help preserve viral

particles. 36 hours later, the virus-containing media was harvested and centrifuged at 1800 rpm for 5 minutes to remove any Phoenix cells. The media was supplemented with 10 µl/ml of 1 M HEPES and 4 µg/ml of polybrene. For HEK293T cells, 2.5 ml of this media was added to each well of a 6-well plate containing adhered cells at 40-50% confluence. Plates were spinoculated in a centrifuge with a swinging-bucket plate rotor at 1300x*g* for 1 hour at room temperature, then incubated at 37 °C overnight. The media was replaced the next morning. After 24 hours, puromycin was added to the media at a concentration of 1 µg/ml and the media was replenished every 48 hours for 10 days. Surviving cells were expanded and frozen for later use. 15 hours prior to harvesting, 1 µM Aqua-Shield-1 reagent (AOBIOUS AOB6677, made to 0.5 mM stock) was added to the media to stabilize protein expression. DNA was harvested using an NEB Monarch Genomic DNA Purification Kit (T3010S), sheared to a target of 8 kb using a Covaris g-tube (Covaris 520079), and purified with a Circulomics SRE XS kit (SS-100-121-01), then barcoded and library prepped with method 1 described below.

9. I'm confused why the free pA-Hia5 has such dramatically higher activity with these substrates. In general, one of my biggest concerns is whether the efficiency of directed methylation is sufficient. The footprints observed when chromatin is treated with free Hia5 are beautiful (Figure S2d and e). By contrast the directed activity is much weaker and seems to define the sites of CA chromatin relatively poorly. Is it the lower local concentration of enzyme, or that the fusion has lower activity in general?

It is entirely expected that incubation with a high concentration of untethered pA-Hia5 produces higher methylation levels in accessible DNA regions than does tethered pA-Hia5. In theory, the same stretch of DNA can come in contact with many distinct untethered pA-Hia5 molecules as they diffuse freely around and collide with the DNA. In contrast, when pA-Hia5 is tethered to an antibody, and all free-floating Hia5 is washed away, each binding region may only have one or a few pA-Hia5 molecules able to methylate the surrounding linker DNA. It is also quite possible that the steric constraints imposed by the tether can reduce the methylation rate per pA-Hia5 molecule. However, these lower methylation levels are still sufficient to detect most binding events. As described above and in the updated text (line 132), we show with *in vitro* experiments on CENP-A mononucleosomes that we can detect methylation surrounding 65% of CENP-A mononucleosomes with tethered pA-Hia5 (vs. 97% for untethered pA-Hia5) (Figure 2d). Nucleosome calling requires at least 20% of 50 bp bins to have a probability of methylation greater than 0.6 (Supplementary Fig. 2f,g,h,k). In the revised version of the paper,

we also describe additional changes to the *in situ* protocol that yield 50-60% more on-target methylation, increasing sensitivity without sacrificing specificity (described further in responses below).

Reviewer #2:
Remarks to the Author:
The manuscript "DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide" describes a chromatin profiling method that combines antibody tethering of a DNA methyltransferase with long-read sequencing. This method builds on previous work using untargeted methyltransferases to footprint DNA-bound proteins and histones in chromatin on extended chromatin reads. The innovation of tethering the methyltransferase gives information on footprinting when a particular chromatin protein is bound. The authors report extensive optimizing of the method for cells, and apply this to in vitro nucleosome arrays, to lamin-associated domains, to the insulator protein CTCF, and to the centromeric histone CENPA. While this is an interesting method with high potential, it is not clear that it currently has sufficient coverage.

We thank the reviewer for their assessment of our work. We have performed additional experiments and analyses to increase our coverage and improve the sensitivity of DiMeLo-seq, as described below.

This issue first arises on line 354: "While the lack of signal at a site may be the result of the sensitivity of our assay rather than the vacancy of a CTCF site, this analysis demonstrates the potential to analyze coordinated binding patterns on single molecules." This is not demonstrated; at the current methylation density it does not appear that coordinated binding can be inferred. The sensitivity of the assay may improve with higher methylation density, but can more methylation be achieved with the tethered strategy?

This reviewer raises an important concern about the efficiency of methylation and mA detection and how this relates to the sensitivity of the method. This performance metric is particularly important when determining the needed coverage (sequencing depth) of a target locus. To address this concern we performed additional rounds of optimization in order to achieve increased methylation efficiency, as now described in the Methods (starting on line 817):

To increase methylation efficiency, the following protocol changes were made and used when

targeting LMNB1 and CTCF for experiments indicated in Supplementary Table 2 and Supplementary Table 3: (1) changed pA-Hia5 binding to 2 hours at 4℃, (2) increased activation time to 2 hours, (3) replenished SAM halfway through activation by adding an additional 800 µM final concentration, (4) reduced Spermidine in the activation buffer from 0.5 mM to 0.05 mM. We refer to the protocol with these changes as protocol v2.

This v2 protocol achieves 50-60% more on-target methylation for both LMNB1 and CTCF without reducing the on-target:off-target ratio (Fig. 3b and Supplementary Fig. 6a). Furthermore, the sensitivity in classifying single molecules as bound by CTCF is estimated to be at least 54%, when using a threshold that achieves 94% specificity (Supplementary Figure 6h, Methods).
Similarly, the single-molecule sensitivity for LMNB1 is at least 59% with 94% specificity (Supplementary Fig. 5e).

After improving the targeted methylation protocol, we then investigated whether sensitivity could alternatively be improved with the PacBio Sequel IIe system, which has advertised improved base calling accuracy with the circular consensus sequencing technique. We sequenced a DiMeLo-seq sample targeting CTCF in GM12878 and an unmethylated GM12878 DNA control with PacBio. Contrary to our expectation, PacBio detected more methylation in the unmethylated control than ONT detected. Both methods still produce similar enrichment profiles around ChIP-seq peaks, but PacBio showed lower accuracy in detecting methylated adenines. We have added these results to Supplementary Figure 9, to the CTCF results section at lines 454 - 461, and to the Methods (lines 933-954, 1247-1262).

Regarding the potential to infer coordinated binding at adjacent CTCF sites on single molecules, we further investigated this potential by sequencing the CTCF-targeted ONT libraries to significantly higher coverage (25X) in order to gain more statistical power for distinguishing signal from technical noise. Fig. 4 has now been updated with these high-coverage CTCF data obtained using the v2 protocol. While we do not have 100% sensitivity, this increased coverage now demonstrates the potential to statistically infer joint binding events. We also softened the language and included proper cavates about sensitivity. Additionally, we added a concrete demonstration of the ability to observe differential binding of two adjacent CTCF sites on single molecules (from a region within the HLA locus on chr6; Supplementary Fig. 8a). Each site contains a heterozygous SNP disrupting the CTCF binding motif on one haplotype or the other. On reads spanning both sites, one can clearly observe methyladenines surrounding one site or the other, consistent with the haplotype phasing of that

read (Supplementary Fig. 8a).

Similarly, in characterizing distribution of CENPA nucleosomes in centromeres, it seems that the sparse overall methylation may be mis-interpreted as the absence of a centromeric nucleosome. On line 487, they state "It is important to note, however, that an absolute quantification of CENP-A nucleosome density in centromeres will require further characterization of the single-molecule sensitivity of DiMeLo-seq in these regions." This seems like a major limitation that should either be explained or remedied. They also describe heterogeneity of centromeric nucleosomes within an HOR, but could this heterogeneity just be an artifact of sparse methylation?

We agree with the point raised by this reviewer and also by reviewer 1 regarding the claim of positional heterogeneity of CENP-A nucleosomes between different chromatin fibers. However, one clear new insight from our study is that we can accurately measure the density of CENP-A nucleosomes across the centromeric region of the chromosome. Using DiMeLo-seq we observe that the CENP-A containing region of the centromere is much smaller than used for previous estimates of CENP-A density (~100kb vs ~1 Mb). Thus, although the total number of nucleosomes may be the same between our study and the work of Bodor et. al., the density of nucleosomes in the active centromere region is approximately 1:4 CENP-A to H3 rather than the earlier estimate of 1:25 (lines 569 - 575). We have modified the text to temper the claims of heterogeneity and emphasize the ability to accurately assess CENP-A nucleosome density, and we have removed the text quoted in the reviewer's comment. Additionally, based on our *in vitro* estimate of CENP-A on-target methylation sensitivity, we have included that the estimate of CENP-A density at the CDR is likely a lower limit (lines 636 -639 in the discussion).

Reviewer #3:
Remarks to the Author:
In this manuscript by Altemose, et al, a new method, DiMeLo-seq, for long-read based mapping of protein-DNA interactions at genome-scale is introduced. The authors present this method as an advance over current methods that rely on short read data and PCR-based seq library amplifications thereby limiting their utility in studying repetitive sequences and DNA modifications in a truly quantitative manner. In this method, antibody directed DNA adenine methylation is achieved by fusing Protein A to a nonspecific deoxyadenosine methyltransferase (Hia5) to catalyze mA in DNA "proximal to the targeted chromatin-associated proteins". The authors present several case studies to demonstrate the overall

utility of this approach, supporting the study with extensive variable tests. This technique is an important and interesting advance in studying DNA-protein interactions, taking advantage of long-read sequencing (ONT) and the newly released human genome assemblies. I have a few minor
comments that would increase the utility of this approach to a broader audience of scientists.

We thank the reviewer for their constructive comments on our work.

In the introduction the authors mention "proximal". Can the authors provide an estimated range to guide scientists in choosing this technique? More specifically, how much direct binding domain information is lost?

We agree with the reviewer that analysis of the degree to which direct binding information is lost with DiMeLo-seq will help guide scientists who use this technique. In vitro, on chromatin containing a single nucleosome, we observe that the methylation probability is highest closest to the nucleosome footprint. In this controlled *in vitro* system without adjacent nucleosomes, we observe peak methylation probability at ~100 bp on either side of the nucleosome dyad with 70% of methylation signal coming from within ~250 bp on either side of the nucleosome dyad (Fig. 2c) (line 139). We have also compared the CTCF binding footprint we observe to ChIP-exo and DNase I footprinting in the CTCF results section at lines 360 - 366.

The authors state that a large footprint for binding for LMNB1 means lower coverage for sequencing – how does this impact the CENP-A assay on 601? A demonstration of exactly how to estimate targeted coverage would be powerful for the reader interesting in applying this technique to determine how much sequencing may be required for a particular DNA-Protein interaction assessment. The authors should provide a method in the main part of the paper to estimate adequate coverage before sequencing. How does error correction/error rates of the different basecallers for modified (and unmodified) nucleotides impact this coverage estimate? How to take such information into account when estimating targeted coverage would be useful for the reader as the tools and flow cell chemistry continue to improve.

The large footprint of LMNB1 was particularly useful for rapid protocol optimization, because we did not need high resolution protein mapping for this purpose. Instead ~0.2X coverage was sufficient because we were calculating the on-target and off-target methylation in large 100-kb bins of the genome. We agree that a discussion about sequencing depth would be very useful to guide scientists using this method, although target coverage will certainly depend on the protein

target and the biological question. To provide the reader with an approach to estimate target sequence depth, we focused on the CTCF analysis. We increased CTCF sequencing depth to ~25X coverage. This increased coverage allowed us to evaluate the effect of sequencing depth on calling CTCF peaks de novo. We have added Supplementary Figure 6f, showing the ROC curves from peak calling at various depths from 5X-25X, using CTCF ChIP-seq peaks as the reference for true CTCF binding sites. Increasing depth from 5X to 25X does increase the area under the curve from 0.82 to 0.92, but the increase in the AUC with coverage begins to plateau at ~15X coverage (Supplementary Fig. 6f inset). For calling CTCF peaks with DiMeLo-seq data,
~15X coverage, with an AUC of 0.90, would likely suffice. However, for single-site analyses, higher coverage may be desired and enrichment strategies like AlphaHOR-RES can make achieving higher coverage more cost effective. Regarding the point about basecaller error rate, we tried two basecallers, Megalodon and Guppy, and found that both had very low FPR (Supplementary Fig. 4). We therefore don't expect the necessary sequencing depth for an experiment to be drastically affected by the basecalling algorithm.

Steric hinderance is brought up a few times, particularly with respect to CTCF with 3' cohesin – would this apply to other targeted proteins and thus present a limitation, or perhaps provide an added benefit to demonstrate immediate protein-protein interactions at specific loci? Determining whether this observed bias to 5' of the binding motif of CTCF is due to 3' cohesin OR the antibody binding location (3' or 5') is important to determine to appropriately interpret the data. Is there an antibody for CTCF that binds 3' to test the influence of antibody binding site vs local protein environment? Given that such an antibody may not be available, this is merely a minor comment rather than a major requirement for publication.

We thank the reviewer for this suggestion to target the N-terminus of CTCF. We ran a DiMeLo-seq experiment with two samples with: (1) antibody targeting CTCF C-terminus, and (2) antibody targeting CTCF N-terminus. We concluded that the antibody binding site does in fact contribute to the peak asymmetry. We have added Supplementary Figure 6e with this result, as well as discussion at lines 368 - 377. This asymmetry is not as prominent with the protocol updates we implemented for increased sensitivity for the revision, likely owing to the fact that we reduced the spermidine concentration to decondense the chromatin at activation and increased the activation time, allowing more time for methylation saturation in the vicinity of pA-Hia5. Nevertheless, we still see an asymmetry that the antibody binding site is causing.

Can the authors provide an estimate of how much the input requirement is reduced with the use of concanavalin-A coated beads? Along these lines, can the authors estimate whether

efficiencies are impacted by lightly fixed vs frozen cells? It is hard looking at all of the variables in Supp Tables 1 and 2 to assess the impact of each variable on projected success rates, particularly since there is a range of sequencing depth across samples.

Input requirements for DiMeLo-seq depend on a multitude of factors: the desired coverage, the desired fragment length distribution, the genome size, the ploidy of the cell type, and the efficiency of the DNA extraction and library prep protocols being used. For the experiments numbered 64-66 in our Supplementary Table 1, we used conA beads with 500k, 430k, and 500k cells each for HEK293T (~triploid), GM12878 (diploid), and Hap1 (haploid), respectively. The final DNA yield was 20%, 39%, and 75%, respectively, of what one would theoretically expect from the input number of cells, after accounting for ploidy (estimated as 3*ploidy pg per cell). This variance in DNA recovery may have to do with the propensity for each cell type to bind conA beads and resist nuclear envelope rupture, or possibly to do with relative cell sizes and the binding capacity of the conA beads. Further replicates would be needed to establish this empirical efficiency more precisely for any given cell type. For the conA bead experiment with 430k GM12878 cells, we yielded 500 ng of DNA after extraction, and ~200 ng after library prep. If prepared with an lsk-110 kit, this would be enough to load a minION flowcell twice while maintaining high pore occupancy (100 ng per loading). Each loading of a flowcell yields ~9 Gb on average, so this amount of DNA would provide 6x coverage of the human genome. Thus, based on our empirical results from this replicate, we can estimate that around 200k diploid cells are needed for 3x human genome coverage. For a line/protocol with higher recovery efficiency, this could fall closer to 100k cells per 3x human genome coverage. We hope this can provide
some guidance to answer the referee's question. We are actively working to determine the lower limits of input for future studies, and we note that ONT's flowcell loading requirements are continuing to decrease as more efficient library prep chemistries are being released.

We have added this information to the Discussion (lines 651 - 656) and we have added further details about the concanavalin A bead experiments to the Methods section (lines 824 - 854).

Regarding the efficiency of DiMeLo-seq on a frozen sample, we can compare experiments #40 and #54 from Supplementary Table S1, which were processed identically and as part of the same batch, with comparable coverage, with the only difference being the use of fresh cells in experiment #40 and frozen cells in experiment #54. The frozen sample achieves comparable SNR and on-target methylation compared to the fresh sample (SNR 23 fresh, 28 frozen; mA/A p>0.9 = 0.003 for both). Similarly, we can compare experiment #62 (fixed) to experiments #56

and #57 (unfixed), which were processed identically in the same batch, with comparable coverage. We note that this entire batch appeared to have anomalously higher background methylation compared to all other batches with a similar protocol, but we can still observe that the fixed sample 62 shows somewhat higher SNR and on-target methylation (SNR 9.7; mA/A p>0.9 = 0.0036) compared to the unfixed samples #56 and #57 (SNR 7.7, 6.9, respectively; mA/A p>0.9 = 0.0024 for both). We conclude from these observations that freezing cells in DMSO/FBS-containing media or lightly fixing them (0.1% PFA for 2 min) does not hinder the ability to perform DiMeLo-seq. However, we note that the frozen/fixed samples had mean read lengths that were ~25% shorter compared to their fresh counterparts. More replicates would be needed to determine if this shorter read length is systematically true, which may owe to increased DNA shearing due to freezing/fixation.

Other minor comments:

Figure 2d inset – should be larger (or include a zoom of 0-400, for example) to illustrate the location of the dyad position

We thank the reviewer for this suggestion to improve clarity. This figure has now been updated (Fig. 2f,g).

Figure 2e – is this single reads or an aggregate? Not sure why some of the red circles are faded – if it is a single read, it should be there or not – correct? Or is this noise in the data due to the basecalls?

We thank the reviewer for this clarifying comment. The cartoon insets in Figure 2G, Supplementary Figure 3H, and Supplementary Figure 3L represent our interpretation of nucleosome positions for the corresponding cluster of reads. We had previously included the faded red circles to represent that the methylation we observe on the array could come from a nucleosome on either side. When we see methylation signal on both sides, we can be confident that the signal comes from a nucleosome positioned between signal peaks. Given the ambiguity of the faded circles, we have removed them and now only show the nucleosome positions where we see signal on both sides, and thus can be sure of the position.

Lines 351-358 – Figure 4d – for clarity for the reader, can the authors indicate on the figure (by clustering) which molecules represent CTCF at both sites, one site or neither site?

We think the reviewer's suggestion does help clarify the original Figure 4d. This figure is now Figure 4c. We have added a cartoon of protein binding to indicate the binding pattern the reader should take away from each cluster.

Lines 34-36 – the authors should add an example (or prospective use) of why assessing DNA-protein interactions and DNA methylation simultaneously provides an advantage (for a broader scientific audience.)

We thank the reviewer for this suggestion to add an example of why simultaneous measurement of DNA-protein interactions and DNA methylation may be useful. Instead of modifying the introduction, we added Figure 4d and Supplementary Figure 8 to show the example case of using DiMeLo-seq to measure the effects of haplotype-specific epigenetic variation on protein binding. For example, CpG methylation on the paternal allele in the IGF2/H19 Imprinting Control Region (ICR) prevents CTCF binding; only on the unmethylated maternal allele can CTCF bind within the ICR, enabling monoallelic expression of H19. On the paternal allele, because methylation prevents CTCF binding within the ICR, IGF2 is instead expressed. Additional examples are shown in Supplementary Fig. 8 for loci on the X chromosome where CpG methylation on the active or inactive X chromosome is associated with reduced CTCF binding on that homolog.

Line 134: spell out "1"

Thank you for noting this. We have deleted this sentence while addressing other reviewer concerns.

Lines 226-7 – "also"…"as well" is redundant in the same sentence

Thank you. We have removed "as well".

Lines 843-845… "as described in figure." The figure number is missing.

Thank you. We have fixed this.

Reviewer #4:
Remarks to the Author:

In their manuscript "DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide", Altemose et al. outline the development, optimization, and validation of the DiMeLo-seq method. DiMeLo-seq relies on antibody directed methylation of Adenine which can be read out by Oxford Nanopore Technologies sequencers. This is an exciting advance over previous methods that used similar but untargeted strategies to map open chromatin. The experiments are well described and the detailed description of the methods optimization provides an interesting insight into how sensitivity and specificity of DiMeLo-seq were dialed in. Further, the development of the AlphaHOR-RES enrichment strategy could provide a useful tool for the sequencing of centromeric regions in general.

In experiments targeting Lamin B1, CTCF, and H3K9me3 in a range of different human cell lines, the authors clearly outline the advantages and disadvantages of using long reads for this type of analysis over established short read methods like CUT&RUN or ChIP-seq.

Overall this is a strong and very interesting manuscript and approach. The few main concerns I have are all linked to the preferential modification of open chromatin by the Hia5 methylase and the relatively low sequencing coverage achieved in some of the experiments.

Major points:

CTCF binding analysis: The authors show that CTCF peaks show a 22-fold enrichment when a methylase targeted by an antibody is used but also a 6-fold enrichment when a free methylase is used to methylate DNA. The authors reasonably argue that this is due to CTCF peaks generally falling within open chromatin.
Because of this however, it might be challenging to define CTCF binding peaks using only DiMeLo-seq data. I would therefore ask for two or three additions to this analysis:

1) In addition to comparing mA/A within CTCF peaks vs outside of CTCF peaks, I would like to see the addition of mA/A in ATAC seq peaks (i.e. open chromatin) that don't overlap CTCF peaks. Since this experiment was performed in GM12878, I am 99% confident ATAC-seq data for this cell line already exists.

The reviewer brings up the important point that we do have background methylation preferentially in open chromatin, and as CTCF sites are in open chromatin, it could make resolving true binding sites difficult. We have added Supplementary Figure 6c, detailing the fraction of adenines methylated in ATAC-seq peaks that do not overlap CTCF ChIP-seq peaks as compared to ATAC-seq peaks that do overlap CTCF ChIP-seq peaks. We do see

significantly higher methylation in the CTCF-targeted sample when a CTCF ChIP-seq peak is also present in open chromatin. The free pA-Hia5 fraction of adenines methylated also increases, as CTCF sites are in the particularly open chromatin; however, the increase is not to the same degree as our CTCF-targeted sample. We have added discussion of this concern and new analysis at lines 351 - 353.

2) If the read coverage is anywhere close to high enough to do this, I would like the authors to attempt to call CTCF peaks based on DiMeLo-seq data alone and compare how these peaks relate to CTCF ChIP-seq peaks. Whatever method the authors chose to do this and how successful this ends up being will give the reader an idea whether DiMeLo-seq will be valuable in enriching short read data sets for now or whether it can already be used as a stand-alone assay. The same is true for the H3K9me3 data, i.e. can you determine enrichment areas, aka peaks from DiMeLo-seq data alone.

We thank the reviewer for this suggestion, and we are currently developing new single-molecule peak calling algorithms for DiMeLo-seq data. In the revised manuscript, we increased our CTCF coverage to ~25X and implemented a simple method of calling peaks to address this question. For each adenine in the reference, we calculated the average probability of methylation across all reads that covered that base. We then computed the average probability of methylation in 200 bp bins genome wide and classified each bin for various thresholds of mean mA probability as TP, FP, TN, FN based on whether a CTCF ChIP-seq peak overlapped that bin. We created an ROC curve (Supplementary Figure 6f), and with 25X coverage were able to achieve an AUC of 0.92. This de novo peak calling is now discussed in the Results (lines 378 - 384) and in the Methods (lines 1188-1198). We need to sequence H3K9me3 more deeply to have sufficient coverage for de novo peak calling so we only included CTCF peak calling in this manuscript.

3) If the coverage isn't high enough to do any type of peak calling, would it be feasible to generate this data by using a PromethION run on the CTCF targeted DNA? You could get ~30x coverage out of a single run which could really help gaining a deeper understanding of the data. I really think this is highly optional though since asking for additional experiments is bad form.

We ran a new CTCF-targeted experiment with our v2 protocol for increased sensitivity and ran two MinION flowcells to increase our overall CTCF coverage to 25X to call peaks. All CTCF analysis in this paper is now with this increased coverage, and increased coverage allowed us to evaluate the effects of sequencing depth on peak calling and to map

haplotype-specific protein-DNA interactions at specific loci in Figure 4d and Supplementary Figure 8.

Minor points:

1) Figure 4:
- Color choice. While blue and orange are a good combination in general (and are maintained throughout the manuscript which is great), the orange doesn't come through on a computer screen so the right side of Fig. 4c is basically just an off-white rectangle. Maybe increase the intensity of the individual points or increase the contrast in some other way to make the underlying data visible.

We appreciate the reviewer's suggestion and have increased the orange intensity.

- Legend inserts. The legend inserts are given without units. "A" content for example is given as a color gradient from 1000-10000 but it is not clear from the legend panel or the figure legend what that actually means. More information would be appreciated.

We thank the reviewer for pointing out that the base abundance scales are unclear. We have added a description to the legend for Figure 4 (line 421 - 423).

2) This is a subjective thing so I would leave this up to the authors but I think adding a short sentence for each paragraph stating how many reads across how many replicates were generated for each experiment, what average lengths they are and what genome coverage that corresponds to would add context to the experiments. All this data is currently in the supplement so you have to go looking for it.

This is useful information for the reader and it can be a bit tedious to look in the supplement for sample details, but we would like to keep this information in the supplement to save space in the main text. We have added more informative read length histograms to Supplementary Figure
13. We have added a section entitled "Sample summary metrics" to the Methods at lines 730 - 732, so the reader can refer to one place in the text to find sample information. The section states: "Sequencing summary metrics for samples included in this study can be found in Supplementary Table 1, Supplementary Table 2, Supplementary Table 3, and Supplementary Figure 13."

Decision Letter, first revision:

Subject: AIP Decision on Manuscript NMETH-A46738A
Message:

Our ref: NMETH-A46738A

6th Mar 2022

Dear Aaron,

Thank you for submitting your revised manuscript "DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide" (NMETH-A46738A). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Methods, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

[REDACTED]

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements as soon as possible. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't. Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know

specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Methods Please do not hesitate to contact me if you have any questions.

Best regards,
Lei

Lei Tang, Ph.D.
Senior Editor
Nature Methods

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #1 (Remarks to the Author):

I am impressed by the serious and thorough responses that the authors made to every point that I raised. I apologize to the authors for my delay getting this review completed. This is a very nice manuscript that is a valuable contribution to the field.

Reviewer #2 (Remarks to the Author):

The revised manuscript “DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide” has addressed my concerns, mainly by improving the method so that targeted methylation is more efficient. My main concerns with the first version was questioning whether methylation was complete enough to infer co-binding events on a molecule, and the revisions to the text are appropriate for these potential limitations.

Reviewer #3 (Remarks to the Author):

In this revision additional information, details and clarifications have been provided in both the main text and supplemental data. With respect to previous comments, the authors have effectively addressed each with additional data, controls, and/or figures. Notably, the authors have added information to make this exciting technique of broad interest (e.g. the H19/Igf2 locus in Fig 4, PacBio vs ONT comparisons, depth of coverage analyses, calling thresholds, in vivo capabilities, etc). In doing so, the authors have presented a strong study that represents an important methodological advance.

Final Decision Letter:

Subject: Decision on Nature Methods submission NMETH-A46738B
Message:

24th Mar 2022

Dear Aaron,

I am pleased to inform you that your Article, "DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide", has now been accepted for publication in Nature Methods. Your paper is tentatively scheduled for publication in our June or July print issue, and will be published online prior to that. The received and accepted dates will be 3rd Aug 2021 and 24th Mar 2022 . This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

[REDACTED]

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Please note that *Nature Methods* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Your paper will now be copyedited to ensure that it conforms to Nature Methods style. Once proofs are generated, they will be sent to you electronically and you will be asked to send a corrected version within 24 hours. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email, phone and fax) of someone who will be able to check the proofs and deal with any last-minute problems.

If, when you receive your proof, you cannot meet the deadline, please inform us at rjsproduction@springernature.com immediately.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Once your paper has been scheduled for online publication, the Nature press office will be in touch to

confirm the details.

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number NMETH-A46738B and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Methods. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.
As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

Nature Research journals encourage authors to share their step-by-step experimental protocols on a protocol sharing platform of their choice. Nature Research's Protocol Exchange is a free-to-use and open resource for protocols; protocols deposited in Protocol Exchange are citable and can be linked from the

published article. More details can found at www.nature.com/protocolexchange/about.

Please note that you and any of your coauthors will be able to order reprints and single copies of the issue containing your article through Nature Research Group's reprint website, which is located at http://www.nature.com/reprints/author-reprints.html. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points.

Best regards,
Lei

Lei Tang, Ph.D.
Senior Editor
Nature Methods