nature portfolio

Peer Review File

Genomic and epigenomic maps of mouse centromeres and pericentromeres

Corresponding Author: Dr Jitendra Thakur

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Centromeres are essential regions of chromosomes required for chromosome segregation. Centromeres are commonly composed of complex arrays of satellite repeats, that assemble CENPA chromatin that builds the kinetochore. Until recently such arrays have been challenging to assemble. In this study the authors apply long read sequencing to assemble the mouse centromere. As the mouse is an important genomic and chromosome biology model system this will have wide impact and facilitate many other studies. However

The core of the mouse centromeres are made of 120 bp minor satellites surrounded by variant satellites, as well as major 234 bp satellites and other repeats on the pericentromere flanks.

The authors use HiFi sequencing and hifiasm to assemble the mouse centromeres. Further details should be provided in the main text. What depth of reads were used, what was their length? How contiguous was the assembly? What QC was performed to ensure that the centromere assembly is correct. Is the C57BL/6J mouse inbred or heterozygous? Where possible, orthogonal validation, eg FISH, should be performed to validate the assembly. Very little information is provided on the genome assembly process and how the assembly was validated – eg the methods provides a single short paragraph on this. As the quality of the assembly is central to all else in the paper, a better job needs to be done on reporting the process and QC steps.

Figure 1 is not terribly useful, mostly showing coloured rectangles. Perhaps show less and add further annotation to make the repeat structure more interpretable, eg using dotplots, Stainedglass, or Moddotplot or similar approaches? Same comment applies to the other figures. As the authors only refer to contigs, does this mean that they have not fully assembled the chromosomes and centromeres? To what extent are higher order repeats detected in the satellite arrays? To what extent are HOR patterns similar to those observed in humans?

The authors profile DNA methylation and show that the centromere region is hypomethylated and transcribed. Human centromeres are reported to have a hypomethylated region coincident with CENPA loading. To some extent this is shown to be similar with lower levels in the minor satellites compared to the pericentromeres. In Fig 4 Im unsure what the y axis of the methylation plots shows. It is labelled as mCpG density and values in the range 0 to ~0.04 shown. Does 0.04 indicate 4% methylation? That is really very low and doesn't seem to be consistent with published work. Heterochromatin in mammals would tend to have much higher levels? Are you dividing by the total width of the interval? In E methylation frequency is shown which looks much more like I would expect. In the human centromeres there is a strong anticorrelation between CENPA occupancy and methylation, so it would also be useful to see this methylation data compared to CENPA occupancy using ChIP-seq data. RNA-seq data re used to profile centromeric transcription and low levels observed that negatively correlate with DNA methylation. The methods contains scant description of how methylation was analysed and presented in the figures, including the difference between density and frequency measures – please rectify.

Finally the authors performed CENPA ChIP analysis aswell as other chromatin modifications. In Fig 6 closeups of the centromere arrays are shown with enrichment in the satellite regions. It would be useful to see these data aligned to the whole chromosome, where the enrichment should be very low, and then very high in comparison with the satellite region. The plots shown in part A also have no x axis scale so its very difficult to appreciate how large these arrays are. From what is shown the enrichment is certainly broader than that reported in human, which only occupies a small section of the alpha

satellites. It would be useful to further see the DNA methylation data added as a track onto these plots. Generally speaking showing these plots for the whole genome or whole chromosome would help interpretability, rather than just showing unassembled ptg contigs, eg the presentation in Fig 7 is much better than in the other figures, although again the figure is shown with no x axis scale, or y axis scales on the tracks, which is not acceptable. The H3K27me3 results are surprising and novel and would benefit from orthologous cytogenetic validation.

Reviewer #2

(Remarks to the Author)
Chaudhry et al. sequenced and partially assembled the Mus musculus C57BL/6J genome using PacBio HiFi sequencing of native DNA extracted from kidneys, then they performed an in-depth analysis of contigs containing centromeric and pericentromeric sequences. They also generated and mapped RNA-seq data from liver, and they generated CUT&RUN data for CENP-A, CENP-B, H3K9me3, and H3K27me3. They describe several general categories of how centromeric and pericentromeric sequences are arranged on different contigs, including a detailed catalog of interspersed genes and transposable elements. Importantly, by analyzing CpG methylation information from PacBio reads, they report that mice centromeres lack the hypomethylated Centromere Dip Regions observed in humans and other vertebrates. Rather, they observe a modest, uniform decrease in methylation and broad CENP-A enrichment at centromeric minor satellite arrays compared to pericentromeric major satellite arrays. They also observe an increase in diverged or structurally variant repeats at the edges of satellite arrays, as seen in other species. Overall, this study provides the most thorough analysis of mouse centromeric and pericentromeric sequences to date. Future studies applying this depth of analysis to more complete assemblies will provide a more comprehensive view of mouse centromere evolution and epigenetics, but the contributions of this study stand on their own. Below are suggestions to clarify and improve the manuscript.

Assembly quality: The main weakness of this study is its highly fragmented assembly, owing to a lack of ultralong nanopore data. Because of this, most centromeric and pericentromeric contigs have not been assigned to particular chromosomes, except for 19 that contain junctions between major satellites and chromosome arms. While this study's partial assembly provides an improvement over this group's previous analysis of unassembled PacBio reads from the same mouse strain (Packiaraj & Thakur Genome Biology 2024), it does not provide a complete picture of these regions and cannot be described as "comprehensive" (as in the last paragraph of the Discussion section). This also makes it difficult to know if the observed sequence organization categories (Figure 7) are truly representative of all mouse centromeres. Two recent studies have generated nearly gapless assemblies of M. musculus genomes (Francis et al. BioRxiv 2024, Liu et al. Science 2024). Of course, Chaudhry et al. would not have had access to these assemblies during execution of their study. Chaudhry et al. also provide a much deeper analysis of centromeric and pericentromeric regions than the other studies, and they add CUT&RUN data, which are not reported in the other studies. The findings of the present study would be strengthened by leveraging the more complete assemblies from Francis et al. and Liu et al., even if just for assigning contigs to chromosomes. However, I imagine this is something the Thakur group or others will report in a future publication.

Sequence analysis: How well can you distinguish MiSat sequences from different contigs/chromosomes? Does this give you a sense for how within-array homogenization compares to between-array homogenization?

Figure 3A: Where are the tRNA genes referred to in the text? They should be labeled clearly.

mCpG analysis: Why were PacBio reads aligned back to the same assembly that they were used to generate, rather than just using the absolute position of each read based on the assembly graph? Why was mCpG density reported, rather than just CG density, alongside CpG methylation frequency? mCpG density is a product of CG density and CpG methylation frequency, and it's not clear to me from the text why the authors chose to report this product. Every time I see mCpG density reported in a figure, it leaves me wondering, is this pattern due primarily to the underlying CG density, or to the methylation frequency?

Array methylation: Note that Gershman et al. Science 2022 reports that active human alpha satellite arrays are typically hypermethylated relative to their neighboring inactive or non-alpha arrays, except for the CDR region, which is hypomethylated. It would be useful to more explicitly compare this human pattern to the MiSat methylation pattern observed in mice (it seems there was a start to this comparison in the second-to-last paragraph of the discussion). The observation that mice lack CDRs is one of the major claims of this paper. To really solidify this result, it would help to see some validation of the alignment and mCpG calling pipeline to know that the reported results in MiSats are not due to something like an alignment artifact.

MaSat motif: In the paragraph introducing the MaSat motif, the sentences need to be rearranged to introduce the Major motif before contrasting it with the CENP-B motif. It would be helpful in the discussion to offer some hypotheses about the possible function of the MaSat motif and ties to prior literature. Perhaps its tracks of poly-A/T provide binding sites for HMGA proteins?
Centromeric inversions: It would be useful to contrast the finding of many MiSat inversions in mice with the near total lack of inversions in active human alpha satellite arrays.

RNA-seq analysis: Figure 5D is confusing. Why not plot this as an x/y plot? Is the association significant compared to a null model? Is the effect size plausible or meaningful?

CUT&RUN analysis: How were alignments processed and filtered? This can affect the apparent localization of CENP-A in centromeric satellite arrays (see e.g. Rhie et al. Nature 2023 Fig 3, where unfiltered reads appear to show CENP-A uniformly distributed across the array, but filtered reads reveal much narrower enrichment in CDRs). It would be interesting to try to discern how CENP-A is localized at finer scales

Figure 7: This figure is overcomplicated and it was not clear to me that this was a schematic until reading the legend. I would suggest simplifying this to highlight the most important results and making it more obvious that it is a schematic and not a direct representation of data, e.g. by plotting tracks as continuous lines rather than as bedgraph-style tracks with coverage variation and noise (which makes them look like data).

Methods:
- It's important to state which antibodies were used for CUT&RUN.
- It's important to state explicitly which tissue type was used for CUT&RUN. Presumably, it's the same pelleted kidney nuclei used for PacBio sequencing, but this is not clear.
- It's important to state how CUT&RUN reads were mapped to contigs. Was multi-mapping allowed? The default settings of alignment algorithms are not usually optimized for mapping to repetitive regions and alignment parameters/filters must be chosen carefully.
- Similarly, it's important to state how Pacbio reads were mapped to contigs for CpG methylation analysis. Was a mapq threshold or any other threshold applied to the reads? Was a threshold applied for mCpG calls to reduce false positives?
- It's also important to state the detailed protocol by which RNA was extracted and sequenced and mapped. Reiterate here that this is total RNA. Was multi-mapping allowed?

Reviewer #3

(Remarks to the Author)
Long-read sequencing allows the assembly of previously unalignable, highly repetitive DNA regions at centromeres and pericentromeres, essential for cell division and genome integrity. Such assemblies and the related epigenetic maps have been generated for humans (PMID: 35357911, PMID: 35357915) and several primate species (PMID: 38570684). Most recently, centromeric and pericentromeric regions have also been assembled in mice (PMID: 39636971; https://doi.org/10.1101/2024.10.24.619615), with some insights into their epigenetic profiles (PMID: 38378611). Here, Chaudhry et al. expanded on their previous work (PMID: 38378611) by assembling both genomic and epigenetic maps of mouse centromeric and pericentromeric regions, including junctions between those arrays, and with telomeres and chromosome arms. They also provided more comprehensive CENP-A (centromere), H3K9me3 (pericentromere), and H3K27me3 (facultative heterochromatin) profiles, and added maps of DNA methylation, CENP-B occupancy (centromeric protein) and transcription. Contrasting different array types at their junctions allows relative comparison of differences in the epigenetic landscapes, which, in my view, is the biggest strength of the paper. Therefore, the data presented here will be of interest to the centromere biology field. My main criticism is that the authors do not fully take advantage of the data, missing an opportunity to add to the ongoing discussions in the field, e.g., about the role of DNA sequence variation in chromatin organization at centromeres or the impact of array homogenization on function. This can be remedied textually by addressing the comments listed below.

Major comments:

1) Given the already published/available studies of mouse centromeric and pericentromeric assemblies (PMID: 39636971 and https://doi.org/10.1101/2024.10.24.619615) the authors should put their findings in perspective of those studies.

2) The authors should help the reader understand if there is a correlation between the number of CENP-B boxes, the level of CENP-B protein, and the CENP-A level at different arrays (homogeneous/junctions). Such a correlation could suggest a link between the DNA sequence and centromere chromatin organization, which has long been debated in the field.

3) The common view in the field is that CENP-B protein binds to CENP-B boxed to support the kinetochore (PMID: 25942623) or pericentromeric chromatin (PMID: 18160038) formation. The authors show the enrichment of CENP-B boxes on MajSat arrays compared to MinSat at their junctions (Fig. 1H) but obtain an opposite enrichment of CENP-B protein at those junctions (Fig. 6A and B). What do the authors think is the reason for that? Perhaps it is due to DNA methylation, known to inhibit CENP-B binding (PMID: 15634350), which the authors show is high at MajSat and low at MinSat (Fig. 4). The authors should discuss this result.

4) (Related to comment 2). The authors show that MinSat 112 and 112-64 dimer arrays contain more CENP-B boxes than 120-mers (Fig. 1H), which appear at least to some extent to be occupied by CENP-B protein (Fig. 6A). At the same time, 112 and 112-64 dimers have less CENP-A despite similar H3K9me3 compared to 120-mers (Fig. 6A). How do the authors explain that given that CENP-B is thought to promote CENP-A binding? Related to that, CENP-B appears enriched at the 112-mer array, but that region is nearly depleted of CENP-A (Fig. 6A; ptg000458l).

5) (Related to comment 2). Homogeneous MajSat arrays appear especially enriched in DNA methylation (Fig. 4B-E). Would this suggest a contribution of DNA sequence? The authors also say that "pockets of lower DNA methylation were also present in homogeneous MaSats (Figure 4B…)". The figure does not appear to show those pockets. Please clarify.

Minor comments:

1) Please add line numbering to facilitate commenting on the manuscript.

2) Pericentromeric maps section - Paragraph 2 – “In contrast, pericentric-chromosomal junction contigs were the largest (up to 76Mb), as expected due to the unique sequences found in the chromosomal arms”. Please provide more explanation as to why this is expected. Presumably, because it’s easier to assemble those regions?

3) Pericentromeric maps section - Paragraph 2 – Please explain what “TLC” stands for.

4) Pericentromeric maps section - Paragraph 3 – “While the CENP-B boxes are essential for recruiting and binding to the centromeric protein CENP-B, the function of the Major motif is not understood”. Please explain what the Major motif is.

5) Pericentromeric maps section - Paragraph 3 – “Recently we have characterized a 12bp motif called the MaSat motif…”. Please explain why this motif is important or relevant.

6) Pericentric region DNA methylation section – Paragraph 1 – “However, the reduction in DNA methylation in centromeric contigs compared to pericentric regions was less pronounced than what has been observed in humans (Altemose et al., 2022)”. Please specify what “less pronounced” means and how was it assessed.

7) Some very small numbers/labels are not readable in each ChIP row (Fig. 5 and 6). Please remove or enlarge.

8) Figure 7 model – there appears to be regions depleted of all marks (CENP-A, CENP-B, H3K9me3 and H3K27me3) in Type 1 and 2 MinSat : MajSat junctions. The authors should comment on why that is.

9) Fig. 7 Type 3 – 112-mer color is mismatched with the legend.

10) Fig. 7 – adding a row showing CENP-B boxes would help conceptualize the contribution of DNA sequence to the epigenetic landscape.

11) Fig. 7 – there are two shades of blue overlayed on one another (CENP-A rows) across the figure – what does that mean?

12) Discussion – paragraph 1 – “Future studies… whether the absence of pericentric MaSats reduces the chromosome segregation efficiency of the mouse Y-chromosomes”. Can the authors provide more context as to why would MaSat presence/absence impact chromosome segregation?

13) Discussion – paragraph 2 – “(…) indicating that pericentric regions are more permissive to non-satellite sequence insertions compared to centromeric regions likely due to their proximity to chromosomal arms”. Please explain or cite the literature on why the proximity of chromosome arms would explain that.

14) Discussion – paragraph 2 – The last two sentences make claims regarding the importance of non-satellite elements at centromeres/pericentromeres, but no citations are provided. Therefore, it is difficult to understand the meaning of the authors’ findings.

15) Please provide more context on why inversions are interesting so the reader can understand the importance of characterizing them.

Reviewer #4

(Remarks to the Author)
This works reports generation of genomic maps of mouse centromeres and their epigenomic annotation. Authors conclude that “This work lays the foundation for research aimed at producing a high quality and fully annotated complete T2T assembly of the mouse genome”. While this work is certainly executed well, it has one major fault. The critical deficiency of this work is the recent publication of T2T assembly of mouse genome (https://www.science.org/doi/10.1126/science.adq8191). There are certainly many distinguishing details of this work – it is not the same strain - and there are other caveats, but within a context of centromeric region assembly, manuscript by Liu et al. vastly supersedes this work.
Functional data - RNA-Seq and CUT&RUN - and analysis reported in this work are certainly interesting, and worthy of publishing, but the manuscript would greatly benefit if all the data would be re-analyzed within the context of T2T mouse assembly. For example, Figures 1 and 2 should be reported as centromeric/pericentromeric regions of specific chromosomes rather than anonymous HiFi contigs. Furthermore, authors can potentially identify discrepancies between their C57 HiFi contigs and published T2T assembly and describe relevance of these finding. And regardless of the extent of differences, an extent of functional annotation in current manuscript exceeds centromere annotation reported by Liu et al. Thus, the reviewer would recommend authors to embrace these new data and reanalyze and re-interpret their finding in the context of T2T mouse assembly. The amount of work should be manageable.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thank you - the authors have addressed my points.

Reviewer #2

(Remarks to the Author)
All of my comments have been sufficiently addressed and the manuscript has been greatly improved. This is an important study that contributes to our evolving understanding of the genomic and epigenetic organization of centromeres across different species.

Sincerely,
Nicolas Altemose

Reviewer #3

(Remarks to the Author)
The authors have addressed all of my comments and provided satisfactory clarifications to the issues raised. I particularly commend their careful evaluation of the assemblies and the resulting claims, including the comparative annotation of their maps against the newly published T2T assemblies. In my view, this study represents a significant advancement in the centromere field by delivering the most detailed analysis of mouse centromeres to date. I therefore support its publication in Nature Communications.

Damian Dudka

Reviewer #4

(Remarks to the Author)
My main concern and suggestion for the authors was to incorporate complete T2T genome assemblies in the analysis. Authors fully addressed my comments and included these novel genomic data without taking shortcuts. As a result, I believe that manuscript became stronger and more relevant. And, as a reviewer, I'm thankful to authors for substantial amount of effort put into data re-analysis. Overall I'm quite satisfied with the outcome.
One minor comment that I have is somewhat odd color scheme selection for Figure 2. I would prefer colors scheme similar to one used on Figure 9c, or two-color scheme. In any case, this is more of a cosmetic suggestion rather comment on content.

Authors: We sincerely thank the reviewers for their time, thoughtful feedback, and incredibly insightful and constructive suggestions. We have carefully addressed each concern and incorporated all relevant suggestions into the revised manuscript. As a result of these revisions, the manuscript length exceeded the journal's word limit, requiring substantial editing to reduce the overall length. All edited sections are highlighted in dark red text. Overall, we highly appreciate the reviewers' comments, which greatly enhanced the clarity, quality, and scope of our study.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Centromeres are essential regions of chromosomes required for chromosome segregation. Centromeres are commonly composed of complex arrays of satellite repeats, that assemble CENPA chromatin that builds the kinetochore. Until recently such arrays have been challenging to assemble. In this study the authors apply long read sequencing to assemble the mouse centromere. As the mouse is an important genomic and chromosome biology model system this will have wide impact and facilitate many other studies. However, the core of the mouse centromeres are made of 120 bp minor satellites surrounded by variant satellites, as well as major 234 bp satellites and other repeats on the pericentromere flanks.

The authors use HiFi sequencing and hifiasm to assemble the mouse centromeres. Further details should be provided in the main text. What depth of reads were used, what was their length? How contiguous was the assembly? What QC was performed to ensure that the centromere assembly is correct. Is the C57BL/6J mouse inbred or heterozygous? Where possible, orthogonal validation, eg FISH, should be performed to validate the assembly. Very little information is provided on the genome assembly process and how the assembly was validated – eg the methods provides a single short paragraph on this. As the quality of the assembly is central to all else in the paper, a better job needs to be done on reporting the process and QC steps.

Authors: We thank the reviewer for highlighting the need to provide these details and an orthogonal validation. In the revised manuscript, we have provided details on the sequencing depth, read length, and assembly metrics. The Hifiasm assemblies for the male C57BL/6J were created using our PacBio HiFi dataset, which contained reads with a mean length of ~9.5 kb and a genome coverage of ~17X. Additionally, the HiFi assemblies produced from the public PacBio HiFi dataset, derived from a female C57BL/6J mouse (Hon et al., 2020), contained reads with a mean length of ~16.6 kb and a genome coverage of 24.6X. In the revised manuscript, we also present a HiFi assembly derived from the latest available PacBio dataset from Liu et al. (2024), offering 65X genome coverage and an average length of ~7.5 Kb (Page 4, lines 190–197).

The C57BL/6J mouse strain we used is an inbred strain, and this information has been mentioned in the revised manuscript (Page 4, line 193 and Page 27, line 798).

Regarding the orthogonal validation, two independent drafts of the mouse T2T genome assembly (Liu at al., 2024 and Francis et al., 2024) constructed using a hybrid approach that integrated PacBio HiFi, Illumina short-read, and Hi-C data became available during the review of this

manuscript. These timely resources served as ideal references for cross-validating and contextualizing our findings. Throughout the revised manuscript, we have incorporated side-by-side comparisons between the T2T and our Hifiasm assemblies to demonstrate the reproducibility and robustness of all of our analyses. All findings, including the structure, distribution, and variation of minor satellites and non-satellite insertions, are supported by both datasets.

Regarding polishing, many widely used polishing tools (e.g., Pilon, Racon, and earlier versions of NextPolish) lead to overcorrections and introduce local variations, which would be counterproductive to the natural variations we are characterizing. While NextPolish2 represents a recent improvement aimed at correcting HiFi-based assemblies, it has not been tested for satellite-rich regions. Since we validated our findings using the recently released two independent mouse T2T assembly drafts (Liu et al., 2024 and Francis et al., 2024), which include polishing steps, we decided instead to leverage the accuracy of HiFi long reads that HiFiasm assembly is known to retain during the assembly process. The T2T assemblies offer high continuity and scaffold-level resolution, whereas the unpolished HiFi assemblies maintain high local sequence accuracy and avoid potential overcorrection artifacts commonly introduced during polishing. Therefore, we combined the complementary strengths of both assemblies to maximize robustness and confidence in our findings, and gained new insights. While our comparative analyses confirmed all the findings presented in the original submission, along with several new insights, we also identified interstitial telomeres that are likely potential artifacts in the Liu et al T2T assembly.

Figure 1 is not terribly useful, mostly showing coloured rectangles. Perhaps show less and add further annotation to make the repeat structure more interpretable, eg using dotplots, Stainedglass, or Moddotplot or similar approaches? Same comment applies to the other figures. As the authors only refer to contigs, does this mean that they have not fully assembled the chromosomes and centromeres? To what extent are higher order repeats detected in the satellite arrays? To what extent are HOR patterns similar to those observed in humans?

Authors: We thank the reviewer for highlighting this. In response, we have revised Figure 1 and related figures to make the repeat organization more interpretable and visually accessible. As recommended by the reviewer, we generated StainedGlass dotplots (revised Figure 3 and Supplementary Figures 6–7) to illustrate the sequence identity patterns within centromeres, pericentromeres, and flanking chromosomal arms (Page 9, lines 330–343). The new figures more clearly convey the spatial organization and transitions between satellite types across genomic regions.

Regarding higher-order repeats (HORs), we thank the reviewer for prompting us to investigate the presence of HORs in mouse centromeres. We have now performed an extensive investigation of HOR organization in mouse centromeres. HORs were detected in 19 of the 21 mouse T2T centromeres, with Cen3 and Cen4 being the only exceptions, both of which are also the most homogeneous centromeres in the assembly. These HORs were validated in both the T2T assembly and our C57BL/6J Hifiasm contigs. Our findings (presented in revised Figure 3 and Supplementary Figures 9, Page 11, lines 362–391, Page 26, lines 714–723) show that, unlike the well-characterized, chromosome-specific longer HORs of human centromeres, the majority of

mouse centromeres contain shorter HORs (e.g., dimers and trimers) that are less abundant and lack strong chromosome specificity. We further explored the spatial distribution and evolutionary relationships of mouse HORs. We observed preferential localization of specific HORs to telomeric or pericentric ends of the centromeres and found phylogenetic separation of previously known Y-centromere HORs and 112-64 dimeric length variants from other HORs. These observations suggest that mouse HORs are in the early stages of evolving from simpler repeat units into more complex structures.

Regarding the reviewer's question on contigs and assembly completeness, the term "contigs" was used specifically in reference to the Hifiasm assemblies, which captured representative centromeric regions but did not span entire chromosomes. In the revised version of the manuscript, the T2T assembly enabled us to provide complete, chromosome-level annotated maps of satellite regions, which were utilized throughout the revised manuscript to contextualize the findings.

The authors profile DNA methylation and show that the centromere region is hypomethylated and transcribed. Human centromeres are reported to have a hypomethylated region coincident with CENPA loading. To some extent this is shown to be similar with lower levels in the minor satellites compared to the pericentromeres. In Fig 4 Im unsure what the y axis of the methylation plots shows. It is labelled as mCpG density and values in the range 0 to ~0.04 shown. Does 0.04 indicate 4% methylation? That is really very low and doesn't seem to be consistent with published work. Heterochromatin in mammals would tend to have much higher levels? Are you dividing by the total width of the interval? In E methylation frequency is shown which looks much more like I would expect. In the human centromeres there is a strong anticorrelation between CENPA occupancy and methylation, so it would also be useful to see this methylation data compared to CENPA occupancy using ChIP-seq data. RNA-seq data re used to profile centromeric transcription and low levels observed that negatively correlate with DNA methylation. The methods contains scant description of how methylation was analysed and presented in the figures, including the difference between density and frequency measures – please rectify.

Authors: mCpG density represents the number of methylated CpG sites normalized by the total length of the genomic interval analyzed. Therefore, a value of 0.04 in our plots denotes that 4% of bases in that region are methylated CpG sites, not 4% methylation at individual CpGs. Methylation frequency, in contrast, measures the average proportion of methylation per CpG site. We have included both mCpG density and DNA methylation frequency in the DNA methylation figures and clearly distinguished between mCpG density and DNA methylation frequency in the figure legends (Figures 6 and Supplementary Figures 12–13). We have expanded the Methods section with a detailed description of the methylation analysis pipeline, including the calculation of both mCpG density and methylation frequency as the reviewer suggested (Page 30, lines 914–924).

To facilitate a direct comparison between chromatin enrichment and DNA methylation, we have added DNA methylation tracks alongside chromatin tracks in Figures 8–9 and Supplementary Figures 15, 16, and 18.

We also added details of the RNA-seq analyses pipeline to the Method section (Page 30, lines 925-937).

Finally the authors performed CENPA ChIP analysis aswell as other chromatin modifications. In Fig 6 closeups of the centromere arrays are shown with enrichment in the satellite regions. It would be useful to see these data aligned to the whole chromosome, where the enrichment should be very low, and then very high in comparison with the satellite region. The plots shown in part A also have no x axis scale so its very difficult to appreciate how large these arrays are. From what is shown the enrichment is certainly broader than that reported in human, which only occupies a small section of the alpha satellites. It would be useful to further see the DNA methylation data added as a track onto these plots. Generally speaking showing these plots for the whole genome or whole chromosome would help interpretability, rather than just showing unassembled ptg contigs, eg the presentation in Fig 7 is much better than in the other figures, although again the figure is shown with no x axis scale, or y axis scales on the tracks, which is not acceptable. The H3K27me3 results are surprising and novel and would benefit from orthologous cytogenetic validation.

Authors: We have now updated the chromatin profiling plots (Figure 8 and Supplementary Figures 15) to display the enrichment patterns of CENP-A, H3K9me3, and H3K27me3 across the entire satellite regions, including the centromere and surrounding chromosomal arms. We have also included zoomed-in views for centromeric regions in Supplementary Figures 17 and 18. We have added x-axis genomic coordinates for the data mapped on the T2T assemblies and the length of the contig for the data mapped to Hifiasm contigs (Figure 8 and Supplementary Figures 15–18). DNA Methylation tracks have been added below chromatin profiles, enabling a direct comparison between chromatin marks and DNA methylation (Figure 8 and Supplementary Figures 15–16 and 18). We have also included details about the y-axis scale for all chromatin tracks in the figure legend. Due to space constraints, the Y-axis values for each mark are clearly specified in the figure legends (Figures 8 and Supplementary Figures 15-18).

We thank the reviewer for recognizing the novelty of the H3K27me3 enrichment observed in the centromeric regions. To further support this finding, we have now confirmed the consistency of this enrichment in mouse T2T and Hifiasm assemblies (Figures 8–9 and Supplementary Figures 15-17). We fully agree that additional orthogonal cytological validation, ideally, technically challenging immuno–fiber FISH, would provide compelling further support. The specific goal of such validation would be to distinguish between high H3K9me3 enrichment with an absence of H3K27me3 on homogeneous MaSats, and reduced H3K9me3 with the presence of H3K27me3 on divergent MaSats. However, designing robust FISH probes to discriminate between these MaSats presented challenges. Divergent MaSat units typically share less than 75% sequence identity, and we found that no single divergent MaSat sequence had sufficient copy number within a confined locus on pericentric regions to ensure a strong and localized FISH signal. As such, we believe that cytogenetic validation will be best pursued in future studies, where possibly a probe cocktail comprising multiple divergent MaSat units can be optimized to visualize regions of elevated H3K27me3 and divergent MaSats simultaneously.

Reviewer #2 (Remarks to the Author):

Chaudhry et al. sequenced and partially assembled the Mus musculus C57BL/6J genome using PacBio HiFi sequencing of native DNA extracted from kidneys, then they performed an in-depth analysis of contigs containing centromeric and pericentromeric sequences. They also generated and mapped RNA-seq data from liver, and they generated CUT&RUN data for CENP-A, CENP-B, H3K9me3, and H3K27me3. They describe several general categories of how centromeric and pericentromeric sequences are arranged on different contigs, including a detailed catalog of interspersed genes and transposable elements. Importantly, by analyzing CpG methylation information from PacBio reads, they report that mice centromeres lack the hypomethylated Centromere Dip Regions observed in humans and other vertebrates. Rather, they observe a modest, uniform decrease in methylation and broad CENP-A enrichment at centromeric minor satellite arrays compared to pericentromeric major satellite arrays. They also observe an increase in diverged or structurally variant repeats at the edges of satellite arrays, as seen in other species. Overall, this study provides the most thorough analysis of mouse centromeric and pericentromeric sequences to date. Future studies applying this depth of analysis to more complete assemblies will provide a more comprehensive view of mouse centromere evolution and epigenetics, but the contributions of this study stand on their own. Below are suggestions to clarify and improve the manuscript.

Assembly quality: The main weakness of this study is its highly fragmented assembly, owing to a lack of ultralong nanopore data. Because of this, most centromeric and pericentromeric contigs have not been assigned to particular chromosomes, except for 19 that contain junctions between major satellites and chromosome arms. While this study's partial assembly provides an improvement over this group's previous analysis of unassembled PacBio reads from the same mouse strain (Packiaraj & Thakur Genome Biology 2024), it does not provide a complete picture of these regions and cannot be described as "comprehensive" (as in the last paragraph of the Discussion section). This also makes it difficult to know if the observed sequence organization categories (Figure 7) are truly representative of all mouse centromeres. Two recent studies have generated nearly gapless assemblies of M. musculus genomes (Francis et al. BioRxiv 2024, Liu et al. Science 2024). Of course, Chaudhry et al. would not have had access to these assemblies during execution of their study. Chaudhry et al. also provide a much deeper analysis of centromeric and pericentromeric regions than the other studies, and they add CUT&RUN data, which are not reported in the other studies. The findings of the present study would be strengthened by leveraging the more complete assemblies from Francis et al. and Liu et al., even if just for assigning contigs to chromosomes. However, I imagine this is something the Thakur group or others will report in a future publication.

Sequence analysis: How well can you distinguish MiSat sequences from different contigs/chromosomes? Does this give you a sense for how within-array homogenization compares to between-array homogenization?

Authors: We thank the reviewer for recognizing the depth of our centromere and pericentromere analysis, as well as the novelty of the accompanying epigenetic and transcriptional profiling. We also agree with the reviewer's important point regarding the limitations of relying solely on a fragmented Hifiasm assembly and pointing to the recent mouse T2T assemblies.

We have now fully incorporated both Liu et al., *Science* 2024 and Francis et al., Biorxiv, 2024 assemblies into our revised analysis. Specifically, all findings from our Hifiasm-based analysis have been confirmed on the T2T assembly, including satellite identity, classification, organization, DNA methylation, chromatin enrichment and transcriptional profiling. Additionally, the continuity and chromosome-level resolution of the T2T assemblies allowed us to assign centromeric and pericentromeric features to specific chromosomes, perform more precise genome-wide quantification, and capture all centromere-pericentromere-chromosome arm transitions. In addition, we conducted new analyses on HOR identification, revealing widespread, albeit relatively short, HORs in mouse centromeres that extend beyond previously reported Y chromosome HORs (Figure 3, Supplementary Figure 9). Based on the reviewer's feedback, we have revised the manuscript to avoid describing our earlier Hifiasm-based analysis as "comprehensive." Instead, we now clearly distinguish between the insights gained from our representative Hifiasm assemblies and the broad chromosome-level findings supported by the T2T assembly.

We thank the reviewer for the question on within-array vs. between-array homogenization. Our sequence identity analyses revealed that within-array identity is indeed higher than between-array identity, particularly for 120-mer MiSat arrays (Figure 2 and Supplementary Figures 6–7). However, multiple centromeres share highly similar sequence identity and HOR distribution patterns, suggesting recent homogenization and duplication of centromeres (Figures 2–3). The presence of variant dimers (112-mers, 112-64 dimers) provided us with distinguishing markers for a subset of centromeres (e.g., Cen2, Cen6, and CenX), which we used to evaluate local sequence homogenization patterns on these centromeres. We observed that 112-64-dimers were the most homogeneous sequences within centromeres (Figures 2B–2C).

Figure 3A: Where are the tRNA genes referred to in the text? They should be labeled clearly.

Authors: We found that tRNA-Lys-AAG, present in the pericentric regions, is located within the non-satellite islands. We have included tRNA tracks in Figure 4B. However, we had to move the text related to tRNA to the supplementary files due to space constraints (Supplementary Page 2, line 64-68).

mCpG analysis: Why were PacBio reads aligned back to the same assembly that they were used to generate, rather than just using the absolute position of each read based on the assembly graph? Why was mCpG density reported, rather than just CG density, alongside CpG methylation frequency? mCpG density is a product of CG density and CpG methylation frequency, and it's not clear to me from the text why the authors chose to report this product. Every time I see mCpG density reported in a figure, it leaves me wondering, is this pattern due primarily to the underlying CG density, or to the methylation frequency?

Authors: We chose to realign the PacBio HiFi reads back to the same assembly used for our satellite region analyses (Hifiasm or T2T) to ensure that all analyses, including methylation and annotation, were consistent with the same satellite regions across all datasets (DNA methylation, CUT&RUN, RNA-seq) for our integrated analyses of specific satellite types (e.g., 120-mer MiSats, MaSats).

Regarding the rationale to report mCpG density, we agree with the reviewer that mCpG density reflects the product of CpG density and methylation frequency. Our reason for emphasizing mCpG density is that it captures the actual methylation burden per unit DNA, which may have more direct implications for chromatin regulation than methylation frequency alone. Two regions may have similar methylation frequencies, yet one could contain considerably more CpG sites, resulting in a higher total number of methylated bases (i.e., a higher mCpG density). These differences are further shown in our new analyses of DNA methylation across different CENP-B box variants, where CG mutations change mCpG density and DNA methylation frequency (Figure 9 and Supplementary Figures 18-19). However, we agree that interpreting mCpG density without considering DNA methylation frequency could be confusing. Therefore, in the revised manuscript, we present DNA methylation frequency alongside mCpG density (Figure 6 and Supplementary Figure 12–13), allowing readers to determine whether observed trends result from differences in CpG content, methylation efficiency, or both.

Array methylation: Note that Gershman et al. Science 2022 reports that active human alpha satellite arrays are typically hypermethylated relative to their neighboring inactive or non-alpha arrays, except for the CDR region, which is hypomethylated. It would be useful to more explicitly compare this human pattern to the MiSat methylation pattern observed in mice (it seems there was a start to this comparison in the second-to-last paragraph of the discussion). The observation that mice lack CDRs is one of the major claims of this paper. To really solidify this result, it would help to see some validation of the alignment and mCpG calling pipeline to know that the reported results in MiSats are not due to something like an alignment artifact.

Authors: We thank the reviewer for the thoughtful suggestion to more explicitly compare mouse satellite methylation patterns to those described in humans, particularly in the context of CDRs reported in Gershman et al. (Science 2022). In humans, CDRs are hypomethylated regions within active alpha satellite HORs, which are embedded within larger hypermethylated inactive HOR arrays, coinciding with CENP-A enrichment. Inactive alpha satellites, which lack CENP-A, are uniformly methylated. In humans, only a limited number of alpha satellites are enriched with CENP-A and function as active centromeres. In contrast, the majority of mouse minor satellites CENP-A enriched. A small percentage of MiSats, which consist of arrays of 112-mer and 112 64-dimeric MiSats, show reduced levels of CENP-A and are probably evolving into inactive centromeric satellites. In terms of DNA methylation, we notice a similar, albeit subtler, trend in mouse centromeres. The 120-mer MiSats, which are significantly enriched with CENP-A, show hypomethylation. In contrast, the flanking 112-64-mer MiSat variants near pericentric regions, which show reduced CENP-A occupancy, exhibit higher DNA methylation, resembling the methylation profile of human inactive alpha satellites described by Gershman et al. Additionally, the mouse pericentric MaSats analogous to human HSat1/2/3, are densely methylated and enriched for H3K9me3, consistent with a repressive chromatin state similar to their human counterparts. Together, our data indicate that while mouse centromeres lack sharply defined CDRs, they exhibit a stratified centromeric methylation landscape where CENP-A–rich 120-mer MiSats are consistently less methylated than their adjacent, lower-CENP-A 112-64-mer variants (Figures 6 and 9 and Supplementary Figures 12, 13, 18 and 19), suggesting an emerging but diffuse pattern of functional methylation stratification in the mouse centromere. We have now included this description in the discussion section (Page 26, lines 725–735).

Additionally, in the revised manuscript, we updated our methylation analysis pipeline to enable multimapping and filtering the mapped reads based on the quality score. The analyses presented in the original submission were performed using the CCSMeth pipeline, which uses the pbmm2 aligner in combination with pb-CpG-tools for methylation calling. In the revised analysis, we adopted the widely used minimap2 aligner with the -N 10 multimapping parameter, applying a mapping quality threshold of 0.8, followed by methylation calling using modkit. The methylation patterns were consistent across both the T2T and Hifiasm assemblies and across both alignment pipelines, reinforcing the robustness of our findings. Due to space constraints, we have reported only the results from the minimap2 + modkit pipeline in the revised manuscript.

Furthermore, we identified large blocks of interstitial telomeric sequences, some extending up to ~80 Kb, within the Liu et al. T2T assembly (Figure 1B and Supplementary Figure 3A–3C), which likely represent assembly artifacts as they are absent in the Francis et al. T2T and all Hifiasm assemblies, including the one generated from the same PacBio data that was used to construct Liu et al T2T, we analyzed. We observed poor alignment of PacBio HiFi reads to centromeric regions (Cen4, Cen9, Cen10, Cen12, and Cen16) harboring these interstitial telomeres (Figure 6 and Supplementary Figure 18). The only additional centromere with poor alignment but lacking interstitial telomeric sequences was Cen3. Notably, Cen3 and Cen4 are among the most homogeneous mouse centromeres and entirely lack higher-order repeats (Figures 2B–2C, 3A, Supplementary Figure 6B). The extreme sequence homogeneity at Cen3 likely complicates the alignment even in the absence of noticeable assembly artifacts. In contrast, such alignment issues were absent in the Francis T2T (Supplementary Figure 12) and Hifiasm assemblies (Supplementary Figure 12). These observations provide additional confidence in the robustness of our alignment pipeline for DNA methylation analyses and underscore its sensitivity in flagging potential reference assembly anomalies.

MaSat motif: In the paragraph introducing the MaSat motif, the sentences need to be rearranged to introduce the Major motif before contrasting it with the CENP-B motif. It would be helpful in the discussion to offer some hypotheses about the possible function of the MaSat motif and ties to prior literature. Perhaps its tracks of poly-A/T provide binding sites for HMGA proteins?
Centromeric inversions: It would be useful to contrast the finding of many MiSat inversions in mice with the near total lack of inversions in active human alpha satellite arrays.

Authors: In the revised manuscript, we have reorganized the paragraph introducing the MaSat motif to clearly define the MaSat consensus motif first (Page 21, lines 613–617). We have also added additional analyses showing that the density of the MaSat Motif is positively correlated with H3K9me3 enrichment and negatively correlated with H3K27me3 enrichment (Figure 9C-9D) (Page 21, lines 617–624)
We thank the reviewer for the insightful comment regarding the functional implications of the MaSat motif. In the Discussion, we now expand on the potential functional significance of the MaSat motif following reviewer's suggestion. We discuss the role of HMGA proteins in chromatin organization and suggest that their interaction with MaSat arrays may contribute to the pericentromeric heterochromatin in the mouse (Page 27, lines 766–779).

Regarding centromeric inversions, we now compare our findings with those from human data in the discussion. We observed frequent inversions within both MiSat arrays in mouse centromeres,

particularly in pericentromeric regions and array boundaries. In contrast, human active alpha satellite arrays—especially those within well-defined HORs—are remarkably strand-polarized and largely lack inversions, as documented in Gershman et al. (Science 2022). This difference may reflect distinct evolutionary pressures or assembly dynamics in the two species, and we now include this comparison to emphasize the unique architectural properties of mouse centromeres (Page 25, lines 684–697).

RNA-seq analysis: Figure 5D is confusing. Why not plot this as an x/y plot? Is the association significant compared to a null model? Is the effect size plausible or meaningful?

Authors: We have now included a correlation plot between RNA levels and DNA methylation levels, which shows a negative correlation between DNA methylation and RNA transcript levels (Figure 7E).

CUT&RUN analysis: How were alignments processed and filtered? This can affect the apparent localization of CENP-A in centromeric satellite arrays (see e.g. Rhie et al. Nature 2023 Fig 3, where unfiltered reads appear to show CENP-A uniformly distributed across the array, but filtered reads reveal much narrower enrichment in CDRs). It would be interesting to try to discern how CENP-A is localized at finer scales.

Authors: We thank the reviewer for the helpful suggestion to incorporate a filtering strategy to strengthen our analysis. In the original submission, all results were based on a non-filtering alignment approach. In the revised manuscript, we now include additional analyses using a filtered alignment strategy to validate the robustness of our findings on chromatin features in highly repetitive satellite regions.

Specifically, we applied k-mer-based filtering using the Meryl algorithm to identify uniquely mappable regions within the satellite-rich T2T assembly. Chromatin profiling was then performed using only these filtered regions. Due to the highly homogeneous sequence composition of mouse centromeric and pericentric satellites, approximately one-third of reads were excluded during this filtering process because they lacked unique k-mers. This result is consistent with previous observations in the human genome, where k-mer filtering also reduced mappability in centromeric regions. Nonetheless, for the subset of reads retained, the chromatin mark profiles remained highly consistent with those obtained using the original, unfiltered alignments (Supplementary Figure 15A–15C). These results demonstrate that the reported patterns are not artifacts of multi-mapping noise. Further details are provided in the revised Methods section (Page 29, lines 864–867). For the remaining figures in the manuscript, we have presented results from the non-filtering approach, as it captures chromatin features on satellites broadly, irrespective of the presence of uniquely mappable regions.

Figure 7: This figure is overcomplicated and it was not clear to me that this was a schematic until reading the legend. I would suggest simplifying this to highlight the most important results and making it more obvious that it is a schematic and not a direct representation of data, e.g. by plotting tracks as continuous lines rather than as bedgraph-style tracks with coverage variation and noise (which makes them look like data).

Authors: We thank the reviewer for pointing out a lack of clarity in this figure, along with providing the specific reason. We have now significantly improved this schematic figure. To emphasize its schematic nature, we have created it using Biorender (Figure 10).

Methods:
- It's important to state which antibodies were used for CUT&RUN.
Authors: The information has been added to the method section (Page 28, lines 846–850).

- It's important to state explicitly which tissue type was used for CUT&RUN. Presumably, it's the same pelleted kidney nuclei used for PacBio sequencing, but this is not clear.

Authors: Thank you for asking for the details. Due to the limited tissue yield from a single pair of kidneys, obtaining high-quality nuclei required for a set of CUT&RUN experiments proved challenging. Therefore, liver tissue, which provides a more abundant source of material, was used for the CUT&RUN experiments. The source tissue for the nuclei has been specified for each experiment in the revised Methods section. (Page 27, lines 801–802, Page 28, line 838).

- It's important to state how CUT&RUN reads were mapped to contigs. Was multi-mapping allowed? The default settings of alignment algorithms are not usually optimized for mapping to repetitive regions and alignment parameters/filters must be chosen carefully.

Authors: In the original submission, results were based on a non-multi-mapping alignment strategy. In the revised manuscript, we now present results generated using Bowtie2 with the -k 10 parameter to include multi-mapping reads, along with a MAPQ ≥10 filter to retain high-confidence alignments. Details are provided in the revised method section (Page 29, lines 861–867). The findings are consistent across both mapping approaches.

- Similarly, it's important to state how Pacbio reads were mapped to contigs for CpG methylation analysis. Was a mapq threshold or any other threshold applied to the reads? Was a threshold applied for mCpG calls to reduce false positives?

Authors: In the original submission, we had not used multimapping or settings for the DNA methylation analysis. In the revised manuscript, we used multimapping (N = 10) and a threshold of 0.8 to retain only high-confidence sites (P ≥ 0.8) for mCpG calls, thereby filtering out methylation calls with confidence below 80%. Details are mentioned in the revised method section (Page 30, lines 914–924).

- It's also important to state the detailed protocol by which RNA was extracted and sequenced and mapped. Reiterate here that this is total RNA. Was multi-mapping allowed?

Authors: We thank the reviewer for this important suggestion. In the revised manuscript, we have now included a detailed description of the RNA extraction, sequencing, and mapping protocols (Page 30, lines 925–937). Specifically, total RNA was extracted using the Trizol method and sequenced on an Illumina platform by Azenta Life Sciences. Raw reads were quality-checked using FastQC (v0.11.9), and adapter sequences and low-quality bases were removed using Trimmomatic. The resulting high-quality reads were aligned to both the T2T and

Hifiasm genome assemblies using STAR (v2.7.11b). To accommodate the repetitive nature of satellites, multi-mapping was explicitly allowed by setting --outFilterMultimapNmax 1000 and --winAnchorMultimapNmax 1000. To retain informative alignments for visualization and quantification, unique mappings were preserved in the final BAM file using --outSAMmultNmax 1. Read coverage was calculated using *bedtools* coverage.

Reviewer #3 (Remarks to the Author):

Long-read sequencing allows the assembly of previously unalignable, highly repetitive DNA regions at centromeres and pericentromeres, essential for cell division and genome integrity. Such assemblies and the related epigenetic maps have been generated for humans (PMID: 35357911, PMID: 35357915) and several primate species (PMID: 38570684). Most recently, centromeric and pericentromeric regions have also been assembled in mice (PMID: 39636971; https://doi.org/10.1101/2024.10.24.619615), with some insights into their epigenetic profiles (PMID: 38378611). Here, Chaudhry et al. expanded on their previous work (PMID: 38378611) by assembling both genomic and epigenetic maps of mouse centromeric and pericentromeric regions, including junctions between those arrays, and with telomeres and chromosome arms. They also provided more comprehensive CENP-A (centromere), H3K9me3 (pericentromere), and H3K27me3 (facultative heterochromatin) profiles, and added maps of DNA methylation, CENP-B occupancy (centromeric protein) and transcription. Contrasting different array types at their junctions allows relative comparison of differences in the epigenetic landscapes, which, in my view, is the biggest strength of the paper. Therefore, the data presented here will be of interest to the centromere biology field. My main criticism is that the authors do not fully take advantage of the data, missing an opportunity to add to the ongoing discussions in the field, e.g., about the role of DNA sequence variation in chromatin organization at centromeres or the impact of array homogenization on function. This can be remedied textually by addressing the comments listed below.

Authors: We thank the reviewer for recognizing the advances our study provides and for suggesting that we put our findings in perspective by utilizing the recent mouse T2T drafts.

Major comments:

1) Given the already published/available studies of mouse centromeric and pericentromeric assemblies (PMID: 39636971 and https://doi.org/10.1101/2024.10.24.619615) the authors should put their findings in perspective of those studies.

Authors: We have now performed all the analyses on T2T assembly by Liu et al, Science, 2024. We have also confirmed major findings on the T2T assembly generated by Francis et al., 2024.

2) The authors should help the reader understand if there is a correlation between the number of CENP-B boxes, the level of CENP-B protein, and the CENP-A level at different arrays (homogeneous/junctions). Such a correlation could suggest a link between the DNA sequence and centromere chromatin organization, which has long been debated in the field.

Authors: We thank the reviewer for this suggestion. We have performed detailed analyses on the number of CENP-B boxes and their contribution to CENP-A and CENP-B binding. Please see our response to the following (Comment#3) comment for details.

3) The common view in the field is that CENP-B protein binds to CENP-B boxed to support the kinetochore (PMID: 25942623) or pericentromeric chromatin (PMID: 18160038) formation. The authors show the enrichment of CENP-B boxes on MajSat arrays compared to MinSat at their junctions (Fig. 1H) but obtain an opposite enrichment of CENP-B protein at those junctions (Fig. 6A and B). What do the authors think is the reason for that? Perhaps it is due to DNA methylation, known to inhibit CENP-B binding (PMID: 15634350), which the authors show is high at MajSat and low at MinSat (Fig. 4). The authors should discuss this result.

Authors: We thank the reviewer for this insightful comment regarding the potential involvement of DNA methylation in CENP-B function. This prompted us to perform detailed analyses, which are now presented in a new Figure (Figure 9).

We found that overall, CENP-B binding strongly correlates positively with CENP-A enrichment, except for the 112-64 dimers, which show reduced CENP-A enrichment compared to the 120-mer MiSat regions, despite having higher CENP-B enrichment. Interestingly, these 112-64-dimers display a higher density of canonical CENP-B boxes and a corresponding increase in CENP-B enrichment compared to the 120-mer MiSats, despite their lower levels of CENP-A. DNA methylation is negatively correlated with CENP-A enrichment and has been shown to inhibit CENP-B binding. Notably, canonical CENP-B boxes contain two CpG sites, whereas many CENP-B box variants we analyzed have mutations in these sites. Accordingly, we observed higher mCpG density in 112-64-dimers compared to flanking 120-mers. These findings indicate that DNA methylation might inhibit CENP-A binding more strongly than CENP-B, or that other elements affect how DNA methylation differently impacts CENP-A and CENP-B binding in regions containing 112-64 dimers at the centromeric-pericentric junctions. Since these variants exhibit higher DNA methylation and lower CENP-A levels compared to the 120-mers, resembling the features of inactive HORs in humans, the 112-64-dimers are probably evolving into elements similar to human inactive HORs. The reason why these variants retain high CENP-B enrichment requires further investigation. While the role of CENP-B in promoting CENP-A chromatin formation is well established, the study by Otake et al. (2020) suggests that CENP-B functions as a molecular switch, depending on the chromatin context, by facilitating the recruitment of either CENP-A assembly factors (e.g., ASH1L) or heterochromatin-promoting proteins (e.g., Suv39h1, HP1). Together, our observations combined with previous studies support a dual regulatory model for CENP-B's influence on CENP-A binding: a sequence-dependent increase in CENP-A binding by CENP-B on 120-mer MiSats, and a methylation-dependent regulation that reduces CENP-A enrichment without a proportional decrease in CENP-B binding on 112-64-dimers in the centromere-pericentric transition zones. (Page 21, lines 598–611 and Page 26–27, lines 741–764).

4) (Related to comment 2). The authors show that MinSat 112 and 112-64 dimer arrays contain more CENP-B boxes than 120-mers (Fig. 1H), which appear at least to some extent to be occupied by CENP-B protein (Fig. 6A). At the same time, 112 and 112-64 dimers have less CENP-A despite similar H3K9me3 compared to 120-mers (Fig. 6A). How do the authors explain

that given that CENP-B is thought to promote CENP-A binding? Related to that, CENP-B appears enriched at the 112-mer array, but that region is nearly depleted of CENP-A (Fig. 6A; ptg000458l).

Authors: We thank the reviewer for bringing this point to our attention. Please see our response to the previous comment (#3)

5) (Related to comment 2). Homogeneous MajSat arrays appear especially enriched in DNA methylation (Fig. 4B-E). Would this suggest a contribution of DNA sequence? The authors also say that "pockets of lower DNA methylation were also present in homogeneous MaSats (Figure 4B…)". The figure does not appear to show those pockets. Please clarify.

Authors: We thank the reviewer for raising another critical point about the possibility of sequence-dependent contribution to chromatin. We have now performed additional analysis to investigate the contribution of the MaSat motif to pericentric contribution (Figure 9, Page 21, lines 613–624).

We found that homogeneous MaSats, predominantly localized near centromeres and in the core regions of the pericentromeres, were more enriched with H3K9me3, a marker of transcriptionally silent constitutive heterochromatin. In contrast, divergent MaSats, concentrated at the pericentric-chromosomal arm transitions, exhibited lower levels of H3K9me3. We found that homogeneous MaSats, which occupy the majority of mouse pericentric regions, contain denser MaSat motifs than divergent MaSats, which are localized near pericentric non-satellite islands and pericentric-chromosomal arm junctions, suggesting that the presence of the MaSat motif may influence H3K9me3 heterochromatin enrichment. Mice HMGA1 and *Drosophila* D1, AT-hook proteins, bind to AT-rich pericentric satellites and suppress position-effect variegation (PEV), indicating their role in H3K9me3-mediated silencing[76]. Given that the MaSat motif also contains poly A/T tracts, it may facilitate the binding of HMGA1 and other AT-hook proteins by binding to the narrow minor groove of poly A/T tracts[77], potentially contributing to the recruitment or stabilization of the H3K9me3 chromatin state at homogeneous MaSats. Interestingly, we observed the opposite pattern for the enrichment of facultative H3K27me3 heterochromatin on pericentric regions. High levels of constitutive H3K9me3 heterochromatin are preferentially associated with highly homogeneous, MaSat motif-rich pericentric MaSats. In contrast, more divergent pericentric satellites exhibit reduced H3K9me3 enrichment, accompanied by a corresponding increase in facultative H3K27me3 heterochromatin. Furthermore, non-satellite islands embedded within MaSat regions are largely devoid of H3K9me3 and instead show preferential enrichment for H3K27me3, indicating a shift toward facultative heterochromatin (Page 27, lines 766–786).

We thank the reviewer for pointing out the lack of evidence for the existence of pockets of low methylation within homogeneous MaSats. Upon examining the overall DNA methylation patterns on T2T and Hifiasm assemblies, we now believe that the so-called "pockets' in the original manuscript are likely minor local methylation variations. Therefore, we have removed the term "pocket" from the revised manuscript.

Minor comments:

1) Please add line numbering to facilitate commenting on the manuscript.

Authors: We have added line numbers to the revised submission.

2) Pericentromeric maps section - Paragraph 2 – “In contrast, pericentric-chromosomal junction contigs were the largest (up to 76Mb), as expected due to the unique sequences found in the chromosomal arms”. Please provide more explanation as to why this is expected. Presumably, because it’s easier to assemble those regions?

Authors: We have now changed the sentence to the following

“In contrast, pericentric-chromosomal junction contigs were the largest (up to 76 Mb), as expected, because the presence of unique sequences in the chromosomal arms allows for the assembly of longer contigs” (Page 6, lines 292–294).

3) Pericentromeric maps section - Paragraph 2 – Please explain what “TLC” stands for.

Authors: The TLC (telocentric) satellites are composed of 146 bp monomeric units and occupy centromeric-telomeric junctions in mice. We have also added a detailed description of TLC-Sats (Page 6, lines 257–261).

4) Pericentromeric maps section - Paragraph 3 – “While the CENP-B boxes are essential for recruiting and binding to the centromeric protein CENP-B, the function of the Major motif is not understood”. Please explain what the Major motif is.

Authors: The 12-bp MaSat motif was identified in our previous study (Thakur and Packiaraj, 2024). The function of this motif is currently unknown.

5) Pericentromeric maps section - Paragraph 3 – “Recently we have characterized a 12bp motif called the MaSat motif…”. Please explain why this motif is important or relevant.

Authors: We have now included a detailed discussion on the importance of the MaSat motif (Page 27, lines 770–786).

6) Pericentric region DNA methylation section – Paragraph 1 – “However, the reduction in DNA methylation in centromeric contigs compared to pericentric regions was less pronounced than what has been observed in humans (Altemose et al., 2022)”. Please specify what “less pronounced” means and how was it assessed.

Authors: Unlike human active HORs, which show almost no DNA methylation signals and form a centromere-depleted region (CDR), mouse centromeric MiSats exhibit a decrease in DNA methylation as compared to MaSats but lack well-defined CDRs. We have now modified the text to reflect this.

"Centromeric MiSats exhibited an apparent reduction in DNA methylation (Figure 6A–6E, Supplementary Figures 12A–12B and 13A–13E). However, the reduction in DNA methylation in centromeric contigs compared to pericentric regions was less pronounced than the clear dip, with almost no DNA methylation observed in human active centromeres" (Page 15, lines 475–479).

7) Some very small numbers/labels are not readable in each ChIP row (Fig. 5 and 6). Please remove or enlarge.

Authors: We have now removed Y-axis labels from the figures displaying the chromatin profiling tracks. We have added the numbers in the Figure legends.

8) Figure 7 model – there appears to be regions depleted of all marks (CENP-A, CENP-B, H3K9me3 and H3K27me3) in Type 1 and 2 MinSat : MajSat junctions. The authors should comment on why that is.

Authors: We thank the reviewer for bringing this mistake to our attention. There should be no such regions. We have now improved the figure to display the boundary of each mark precisely (Figure 10).

9) Fig. 7 Type 3 – 112-mer color is mismatched with the legend.

Authors: We thank the reviewer for bringing this mistake to our attention. We have now modified the legend color to reflect 112-mer and 112-64-dimer colors in the Figure (Figure 10).

10) Fig. 7 – adding a row showing CENP-B boxes would help conceptualize the contribution of DNA sequence to the epigenetic landscape.

Authors: We thank the reviewer for this suggestion. We have now added a row showing CENP-B box density across the regions shown. We have also added a row for the MaSat motif box (Figure 10).

11) Fig. 7 – there are two shades of blue overlayed on one another (CENP-A rows) across the figure – what does that mean?

Authors: We thank the reviewer for bringing this to our attention. We have now modified the Schematic figure and used a single shade of blue for CENP-A, consistent across and within the tracks (Figure 10).

12) Discussion – paragraph 1 – "Future studies… whether the absence of pericentric MaSats reduces the chromosome segregation efficiency of the mouse Y-chromosomes". Can the authors provide more context as to why would MaSat presence/absence impact chromosome segregation?

Authors: We have now added a description to explain why the presence or absence of MaSat can impact chromosome segregation.

"Interestingly, we confirmed that the Y-centromere lacks flanking pericentric MaSats, which agrees with previous findings suggesting that the majority of the mouse Y chromosome is euchromatic (Morgan & Pardo-Manuel de Villena, 2017; Soh et al., 2014). Heterochromatin protein 1 (HP1), which binds to pericentric H3K9me3 heterochromatin, plays a crucial role in recruiting and maintaining the cohesin complex, essential for ensuring proper sister chromatid cohesion and preventing premature separation prior to anaphase (Bernard et al., 2001; Shimura et al., 2011; Yi et al., 2018). Since the mouse Y chromosome lacks detectable pericentric MaSats, and H3K9me3 heterochromatin, future studies investigating the loss rate of the mouse Y chromosome may shed light on whether the absence of pericentric MaSats reduces the chromosome segregation efficiency of the mouse Y chromosome". (Page 25, lines 667–671).

13) Discussion – paragraph 2 – "(…) indicating that pericentric regions are more permissive to non-satellite sequence insertions compared to centromeric regions likely due to their proximity to chromosomal arms". Please explain or cite the literature on why the proximity of chromosome arms would explain that.

Authors: Satellite regions undergo frequent homogenization and may frequently replace transposon insertions. However, as they become more divergent, the homogenization process decreases. As we have shown in the manuscript, diverged MaSats are more frequent towards chromosomal arm junctions. This may make these regions more permissive to accumulating non-satellite insertions at a higher frequency than the satellites present at the core of pericentric regions. We have now modified the discussion to reflect this.

"Our study shows the progressive increase in non-satellite repeat density from centromeres to pericentromeres to chromosomal arm junctions. While non-satellite repeats are present as individual, isolated elements at centromeric and homogeneous pericentric regions, they have expanded into clustered islands within the divergent parts of pericentromeres. This observation suggests that as MaSats divergence increases, non-satellite insertions increase and lead to the formation of islands. Similar non-satellite repeats have been observed in centromeres of other species. For example, *Drosophila* centromeres contain several satellite and non-satellite repeat families[45,68-70]. However, the majority of functional CENP-A enriched regions are located on the islands of retrotransposons[45]. Similarly, retrotransposons interspersed within plant centromeric satellites, which are more divergent than those of mouse centromeres, are part of functional centromeres[41,42,44,71,72]. Given our observation that rare LTRs within homogeneous MiSats are devoid of CENP-A, retrotransposons likely adopt centromeric function when satellites become more divergent and non-satellite repeats cluster into islands during evolution.
" (Page 25–26, lines 699–712).

14) Discussion – paragraph 2 – The last two sentences make claims regarding the importance of non-satellite elements at centromeres/pericentromeres, but no citations are provided. Therefore, it is difficult to understand the meaning of the authors' findings.

Authors: We have now provided citations and modified the discussion section for enhanced clarity (Page 25–26, lines 699–712).

15) Please provide more context on why inversions are interesting so the reader can understand the importance of characterizing them.

Authors: In the revised manuscript, we now provide additional context on the biological relevance of inversions within centromeric MiSats arrays.

"Furthermore, while inversions are present in centromeric and pericentric regions, they are more frequent within centromeric regions. In contrast, human active alpha satellite arrays, especially those within well-defined HOR domains, largely lack inversions (Gershman et al., 2022). This contrast likely reflects differences in satellite sequence homogeneity and evolutionary pressures between species. Human centromeres, being more divergent, may limit the potential for sequence-based inversions. In contrast, the high sequence homogeneity of mouse centromeric MiSats, facilitates inversions via non-allelic homologous recombination. In younger, more homogeneous centromeres such as those in mice, inversion events may have evolutionary significance, particularly by generating structural variations that drive further divergence. As centromeres accumulate more abundant HOR structures and sequence variations, as seen in humans, inversion frequency may decline. Together, we speculate that the inversion frequency within satellites is influenced by sequence homogeneity and centromere maturation, providing valuable insights into the evolutionary trajectory of centromeres" (Page 25, lines 684–697).

Reviewer #4 (Remarks to the Author):

This works reports generation of genomic maps of mouse centromeres and their epigenomic annotation. Authors conclude that "This work lays the foundation for research aimed at producing a high quality and fully annotated complete T2T assembly of the mouse genome". While this work is certainly executed well, it has one major fault. The critical deficiency of this work is the recent publication of T2T assembly of mouse genome (https://www.science.org/doi/10.1126/science.adq8191). There are certainly many distinguishing details of this work – it is not the same strain - and there are other caveats, but within a context of centromeric region assembly, manuscript by Liu et al. vastly supersedes this work.
Functional data - RNA-Seq and CUT&RUN - and analysis reported in this work are certainly interesting, and worthy of publishing, but the manuscript would greatly benefit if all the data would be re-analyzed within the context of T2T mouse assembly. For example, Figures 1 and 2 should be reported as centromeric/pericentromeric regions of specific chromosomes rather than anonymous HiFi contigs. Furthermore, authors can potentially identify discrepancies between their C57 HiFi contigs and published T2T assembly and describe relevance of these finding. And regardless of the extent of differences, an extent of functional annotation in current manuscript exceeds centromere annotation reported by Liu et al. Thus, the reviewer would recommend authors to embrace these new data and reanalyze and re-interpret their finding in the context of T2T mouse assembly. The amount of work should be manageable.

Authors: We thank the reviewer for recognizing the quality of our work and emphasize the timely need to put our findings into context with the recently published mouse T2T assemblies. We also appreciate the suggestions to compare T2T assemblies with Hifiasm assemblies. In

response, we conducted all our analyses using T2T assemblies as references and compared the results with those from Hifiasm assemblies.

Specifically:

- All genome-wide analyses now use chromosome-level coordinates derived from the Liu et al. T2T assembly. We have also cross-validated major findings using the Francis et al. T2T assembly.

- All findings presented in the original submission, including representative centromere/pericentromere organization, presence of non-satellite repeat islands, chromatin profiles, transcript levels, and DNA methylation features, have now been validated on T2T assemblies. We have expanded our analyses to include HOR identification, revealing multiple HORs across mouse centromeres, most of which are short (dimers and trimers), in contrast to the longer HORs observed in human centromeres.

- While most of our findings are supported by new analyses of T2T assemblies, we have identified specific differences between the Hifiasm contigs and the T2T assembly. For instance, interstitial telomeres seen in the Liu et al. mhaESC T2T assembly were not detected in the Francis et al. T2T or any of the Hifiasm assemblies we examined (Figure 1 and Supplementary Figures 1 and 3). Therefore, these interstitial telomeres are likely assembly artifacts, as they were also absent from the Hifiasm assemblies generated from PacBio data used by Liu et al. to produce the T2T assembly. Additionally, we found that the density of length-variant 112-mer MiSats is lower in T2T assemblies compared to Hifiasm assemblies (Supplementary Figure 2).

Taken together, we have fully leveraged recent T2T mouse assemblies to generate comprehensive genomic and epigenomic annotations for mouse satellite regions. While our initial Hifiasm-based assemblies captured representative features, the T2T assemblies allowed us to provide a chromosome-resolved, end-to-end view of mouse centromeric and pericentromeric organization.

Reviewer #1 (Remarks to the Author):

Thank you - the authors have addressed my points.

Authors: Thank you.

Reviewer #2 (Remarks to the Author):

All of my comments have been sufficiently addressed and the manuscript has been greatly improved. This is an important study that contributes to our evolving understanding of the genomic and epigenetic organization of centromeres across different species.

Sincerely,
Nicolas Altemose

Authors: Thank you.

Reviewer #3 (Remarks to the Author):

The authors have addressed all of my comments and provided satisfactory clarifications to the issues raised. I particularly commend their careful evaluation of the assemblies and the resulting claims, including the comparative annotation of their maps against the newly published T2T assemblies. In my view, this study represents a significant advancement in the centromere field by delivering the most detailed analysis of mouse centromeres to date. I therefore support its publication in Nature Communications.

Damian Dudka

Authors: Thank you.

Reviewer #4 (Remarks to the Author):

My main concern and suggestion for the authors was to incorporate complete T2T genome assemblies in the analysis. Authors fully addressed my comments and included these novel genomic data without taking shortcuts. As a result, I believe that manuscript became stronger and more relevant. And, as a reviewer, I'm thankful to authors for substantial amount of effort put into data re-analysis. Overall I'm quite satisfied with the outcome.
One minor comment that I have is somewhat odd color scheme selection for Figure 2. I would prefer colors scheme similar to one used on Figure 9c, or two-color scheme. In any case, this is more of a cosmetic suggestion rather comment on content.

Authors: We thank the reviewer for this thoughtful suggestion regarding the color scheme of Figure 2. The triangular StainedGlass heatmaps indeed provide excellent visualization at the

scale of a few megabases, as shown in Figure 9c. Figure 9 focuses on relatively short regions (a few Mb), where triangular StainedGlass representations preserve resolution and clarity.

In contrast, Figure 2 presents identity maps across much larger genomic regions, often spanning >10 Mb. At this scale, triangular StainedGlass plots lost resolution and failed to convey sequence identity patterns effectively. To address this, we used the HiGlass container to visualize the identity maps generated by StainedGlass. We did revisit the reviewer's suggestion and tested the same default triangular StainedGlass color scheme within HiGlass. However, applying those colors to the HiGlass heatmaps across large genomic regions did not yield optimal visualization. For this reason, we decided to keep the color scheme that we had carefully optimized for the HiGlass-based identity maps. We believe that the colors selected for Figure 2 best highlight the differences in sequence homogeneity across these large regions, enabling clearer interpretation of the data for the intended scale.