

A global view of human centromere variation and evolution

Corresponding Author: Dr Glennis Logsdon

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)

The manuscript Gao S., et. al., 2026 evaluates centromere sequence via assemblies at an unprecedented scale of 2110 complete human centromeres from diverse global population groups categorized into 5 continents. They have developed multiple bioinformatic tools to assemble centromeres carefully and thereafter identify and annotate the alpha-satellite repeat elements that constitute centromere structure. First and foremost, the authors describe their method of collecting a dataset of diverse human centromeres assembled before describing the genetic and epigenetic variation they observe in this dataset. The group uncover 1870 new alpha-satellite HOR variants different from that described in CHM1 and CHM13. The organization of these HORs are also described among the 5 continental groups and among the 23 pairs of chromosomes that make up the human karyotype. They analyze the variation in lengths of alpha-satellite HOR arrays, the sequence variation at both large and small scales that impact the organization of HOR arrays. They also describe the variation in alpha-satellite HOR composition and highlight various cases of pronounced differences in alpha-satellite HOR composition for different chromosomes. The authors bring to light novel centromere recombination events between acrocentric chromosomes 13 and 22.

An exciting aspect of the manuscript is the evaluation of the epigenetic variation of these assembled centromeres among global groups. They describe the centromere dip regions (CDR) in these assembled centromeres. They describe CDR variation for different chromosomes among continent groups. Interestingly, the CDR is often found on sequences that are highly homogenous. The team methylation and CENP-A to describe kinetochore sites in a fraction of the assembled centromeres. The authors examine the evolution of centromere arrays at both ancient and modern scales, highlighting how centromeres are constantly evolving and mutating. The manuscript continues to finally take advantage of the enormous dataset and evaluate mutational dynamics of different features of the centromere and describe evolutionary trajectories of centromere HORs. The group show that the CDRs found on centromeric chromatin that is highly similar also show the highest rates of mutability. The authors underscore the vast variation in sequence and structure of the human centromere. They detect various patterns and propose various directions for the field to investigate. Overall, the work described is robust and exciting for the field to understand and appreciate human centromere variation.

Comments:

1. Please provide clear definitions for the terms centromere, kinetochore, CDR up front.
2. An overarching assumption that the paper starts with is the fact that the CDR is the kinetochore-forming region. They also, however, mention that the kinetochore sites often have multiple CDRs within them in later analyses.
 - a. Therefore, it is unclear in cases where the lengths of kinetochore forming sites are compared, whether this takes into consideration the absolute length of the kinetochore forming chromatin (including CDRs and the space between them) or if this is a summation of the CDR containing chromatin for these chromosomes. This extends to other comparisons between continental groups as well.
 - b. It is also unclear when looking at the similarity and entropy of kinetochore forming chromatin if this takes into consideration the entire stretch of CDR containing chromatin including the HOR array with high CpG methylation or just the CDRs.
 - c. The authors can use the vast assembly data as an opportunity to define what a CDR is, objectively, for the field. This would be a highly significant and beneficial definition for the field.
3. Supplemental information that will make this manuscript a valuable resource for the field is the inclusion of bed files for the boundaries of all the centromeres defined in the assembled centromeres. This will help standardize future analysis of these centromeres.

4. Authors should use a more convincing representative chromosome spread image showing a 13/22 recombined chromosome (figure 2i)
5. When commenting on the number of MEIs, the authors show that chromosome 16 has an elevated incidence of MEIs and correlate this with the known propensity of this chromosome to show nondisjunction during meiosis. It will be useful when speculating on this correlation to show another example of a chromosome that shows a lower incidence of MEIs and does not show a high propensity for nondisjunction in meiosis.
6. Supplemental Figure 77 should be made clearer to show the changes in the CDR vs active HOR array from G1 to G4 for different details analyzed. It is also unclear why the number of dips in CpG methylation is examined here.
7. When analyzing the number of archaic hominin specific Kmers in haplotypes of chromosomes 12 and 21, it looks like the enrichment of these Kmers are specifically outside the CDR. It would be interesting for the authors to speculate on this and how this kmer analysis of ancient and modern centromere haplotypes fit with the model of heightened recombination in highly similar CDR chromatin.
8. The authors should discuss the limitations of using only three neanderthal and one denisovan genome in their kmer analysis. The main crux of the paper involves describing the vast variability in human centromeres. Therefore, using a handful of kmers to determine the evolution of chromosomes 10, 12 and 21 come with the caveat that the kmers used may not be sampling all ARC-kmers completely.
9. The authors do not mention the 28 population groups in any of the analyses and stick to describing similarities and differences in 5 continental groups. The authors may not need to mention the 28 population groups if they are not going to use them in the analysis?
10. There are some occurrences where an acronym is used without being defined. (MEI, SV, SNVs) it will be useful for these to be defined the first time its used.
11. In terms of the mutation rate at the centromere, it seems both that 1) slippage from DNA replication may happen at a higher rate due to the repetitive nature of the sequence and 2) that individual mutations may not be strongly selected against and 3) that recombination may be suppressed due to heterochromatin. It hard to know how the balance of all these processes plays out in the observed sequence-it would be nice to include some discussion of these processes (replication, recombination, repair, selection) and their contribution to the observations.

(Remarks on code availability)

Referee #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Referee #3

(Remarks to the Author)

Gao, Oshima et al. present an exciting follow-up to the HGSC paper from last year (Logsdon et al. Nature 2025), with >2000 fully assembled centromeres from 65 near-complete human genome assemblies drawn from across the globe. A novel pipeline reveals hundreds of major centromere haplotypes as well as >1800 novel alphaSat HOR variants. Using these data, the authors report remarkable variation in kinetochore binding sites, HOR arrays, as well as large structural rearrangements (and a striking chromosome 13-22 recombination present in multiple individuals). Using CENP-A DiMeLo-seq and CENP-A CUT&RUN, the authors validate individual kinetochore sites (occupied by CENP-A), showing that almost 95% of centromeres have only one kinetochore site closely linked to the sequence/layout of the centromere, with kinetochores preferentially positioned in regions with homogenous HOR distributions; analysis of individual kinetochore sites revealed a fairly frequent mobile element insertion within the kinetochore sites of chromosome 16. Finally, the authors cluster haplotypes to create phylogenetic trees for each centromere with chimpanzees as outgroups, revealing a large number of recent evolutionary changes in centromere sequence/structure and also higher than expected mutation rates.

Overall, the paper is well written, the methods are extensively documented, and the figures are well-organised (albeit incredibly dense, and somewhat unsuited to the static 2D PDF format given how data-rich they are, more on that below). The raw data and centromere annotations have been made available for download. It is clear from this work - in line with data from other eukaryotic clades - that centromeres evolve so quickly and so dramatically that the imprint of recent and large-scale evolution is writ large on the genomes of extant human populations, with major implications for models of genome evolution, epidemiology and medicine, but also clear cell biological predictions related to the mode and tempo of kinetochore site evolution.

I have a few comments which if addressed would heighten the utility of this work for others in the field.

The clustering/thresholding approach used to delineate centromere haplotypes from each other is not well described in the methods. This is critical given the complexity and size of satellite regions and the combinatorial diversity of the alphaSat HORs - presumably differently coarse-grained clustering approaches would lead to a varying number of haplotypes, in turn influencing the analyses in Figure 1 and the phylogenetic analyses later in the paper. More importantly, it would be difficult to reconcile these definitions with future analyses coming in from other big global and regional genomics projects.

I have a similar comment with regards to the identification and definition of novel HORs. How can these be validated given the absence of any independent functional or phylogenetic definition? As I understand it each novel genome in the future will then result, almost axiomatically, in the identification of new HORs. Would it not make sense to focus only on those with representation across multiple genomes (or did the authors implement this filter? Perhaps I missed it)

The distribution of alphaSat HORs is absolutely impossible to “read”/infer from most of the figures, particularly Figure 1. I found the stacked representation in Figure 2e at least easier to parse/visualise, but the overlapping colours/lines are simply too small in most of the other panels. Would there be any way to create a software visualisation tool that could accompany the paper? A way to browse through the haplotypes and annotations, focussing either on individuals, or chromosomes across individuals, or specific HORs and their phylogenetic representations, etc. I understand the challenges involved in creating something versatile and also user friendly but I believe it would be well worth the investment - and also future-proof these analyses, allowing others to build on the findings here as new genomes come in from around the world.

(Remarks on code availability)

Referee #4

(Remarks to the Author)

Key results:

In this article, Gao, Oshima et al. have assembled fully and characterized 2110 complete human centromeres from 65 individuals representing 28 different populations from 5 continents. This allows for an unprecedented global view of human centromere diversity, as they combined this dataset with previously described ones, reaching 8340 total centromeres. The development of specific bioinformatic tools led to the discovery of 226 major new centromere haplotypes and 1870 - satellite higher-order repeats variants, refining our knowledge of centromeric diversity across individuals and populations and even on a 4 generation pedigree.

For 7 out of 24 chromosomes it was determined that the new reference human genome T2T-CHM13 does not represent the most common centromere haplotype, highlighting the relevance of this study for the broad scientific community, and the centromere field in particular. On a population-scale, the data presented revealed several mechanisms that contribute to centromeric diversity, such as large and small structural rearrangements, and the insertion of mobile elements in ~30% of centromeres. From an evolutionary point of view, this broad dataset has unveiled ancient (up to ~1 million years ago) and recent (down to 1000 years ago, or ~50 generations) evolutionary events that have shaped the modern human centromere landscape, such as the introgression of archaic hominins' satellites in certain minor haplotypes or the phylogenetic separation of particular haplotype variants in specific populations.

Overall the findings of this article provide multiple supporting evidences for the proposed model of centromere evolution.

Validity:

This manuscript does not present any evident flaws which should prohibit its publication.

Originality and significance:

This is mainly a descriptive article of the assembly and characterization of a wide cohort of human centromeres. These analyses were achieved through the development and/or refinement of a suite of bioinformatics tools specifically developed for these purposes, and that could be re-applied in future research.

Many of the concepts for the main model of centromere evolution have already been put forward in the past. Logsdon et al., 2024 [<https://doi.org/10.1038/s41586-024-07278-3>], in particular, is overall a very similar article, but based on the comparison of only two full genomes and a few chromosomes in other species. The work presented here provides novel conclusions and, to date, the strongest evidence that support the notions that were proposed -but not fully demonstrated- before. The results presented are of interest for a broad audience, as they highlight the high variability of the human centromeres sequences, and more specifically, for the audience in the centromere field, they are of immense relevance.

Data & methodology:

The approach of this work is conceptually simple: to sequence and compare as many centromeres possible from a diversity of individuals. From a technical and methodological point of view the work is extensive and complex. The resulting data is of high quality and it has been presented in a manner that is extremely clear and easy to follow.

Data and code are available and accessible. The 2110 centromere fasta files from <https://figshare.com/s/cb839beea6ed06d641db>, which is supposed to be complete, include 66 centromere fastas with gap. Please double check if they are the final version on this link. Please note that this reviewer is not an expert in computational biology or bioinformatics, so I refrain to judge if the methodology is sufficiently detailed to allow for the results to be fully reproduced.

Conclusions:

The data presented and its interpretation is robust, valid and reliable.

References:

The manuscript references the previous literature appropriately.

Clarity and context:

The abstract is very clear and accessible to a broad audience. Even though the article presents some concepts that are not immediately accessible to everyone, it is written in a way that guides the reader step by step and is very clear and appropriate.

Suggested improvements:

- Centromere genome and epigenome could be influenced by some confounders, e.g. sex, age, and may result in disease status. Do you consider the confounders influence or try to find some associations?
- The description of the CUT&RUN method adapted to long-read is unclear as chromatin fragments enriched with CENP-A seems to be digested as normal CUT&RUN. How does it work for long-reads?
- The FISH analysis presented in figure 2h is not very clear and does not really support fusion of chromosome 13 and 22. In the provided example there is no clear cen13 staining. The FISH signal is also too saturated.
- Line 238: The authors should cite previous works demonstrating that CENP-A is a key epigenetic mark of centromere identity. The cited computational works did not prove that.
- Lines 269 – 270 / Figure 3 d-e: Please define what specific criteria used to define a methylation dip as such. Is it how low the methylation frequency goes compared to the surrounding? What is the minimum length that a dip has to have? For example, Figure 3e states that there are 8 dips. Only zooming in heavily on the image can one see what I assume is the 3rd dip, which is very narrow and not so deep as the others, and was 5th dip, which is wide but very shallow?

Minor comments

- Figure 3 l and m: the description of the plots is missing from the figure legend.
- Figure 4d: there is an error in the labelling of the last node presented.
- Figure 5a: error bars are not defined.
- Line 168, "D1Z7" means the primary alphaSat family in chr1. If the meaning of the numbering scheme is explained clearly at the beginning, there is no need to repeatedly restate these alphaSat family IDs and chromosome names later on.
- Line 278 "MEI", even the full name shown up in the Figure 1 legend, it is better to have the explanation in the full context.

(Remarks on code availability)

Version 1:

Reviewer comments:

Referee #3

(Remarks to the Author)

The authors have satisfactorily addressed my comments, and to the extent that I am able to judge, those of the other reviewers. I particularly commend them on building an interactive tool that enables non-specialist readers to interact directly with complex centromere sequence/structure data.

(Remarks on code availability)

Referee #4

(Remarks to the Author)

In this revised manuscript, Gao et al., have addressed the remarks and answered all questions made by all reviewers, improving several aspects of an already remarkable manuscript. In particular, the creation of a encyclopedia as a way of interactively visualizing the large and complex data here generated is of tremendous value to the field.

We would like to suggest two minor additions that could further improve the quality of the text and figures, and are happy to recommend the publication of this work in Nature.

- Reviewer 1, point 2a) It would be great if the authors could also incorporate an equivalent scheme as the one presented in Reviewer Figure 1, clarifying in a visual way the difference between CDR and CDR dips in the main Figure 5c and 5d for all the readers.

- Reviewer 1, point 2c. I agree with this comment from Reviewer 1, and perhaps the authors' change to the main text could be further improved to erase all possible doubt in the following way:
“(...) identifying the location of the CDR, a region which spans from the first hypomethylated base to the last hypomethylated base and can contain one or multiple methylation dips, in each centromere.”

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Authors' responses to referees' comments

Referee expertise:

Referee #1: chromosome biology

Referee #3: evolutionary genomics

Referee #4: chromosome biology

Referees' comments:

Referee #1 (Remarks to the Author):

The manuscript Gao S., et. al., 2026 evaluates centromere sequence via assemblies at an unprecedented scale of 2110 complete human centromeres from diverse global population groups categorized into 5 continents. They have developed multiple bioinformatic tools to assemble centromeres carefully and thereafter identify and annotate the alpha-satellite repeat elements that constitute centromere structure. First and foremost, the authors describe their method of collecting a dataset of diverse human centromeres assembled before describing the genetic and epigenetic variation they observe in this dataset. The group uncover 1870 new alpha-satellite HOR variants different from that described in CHM1 and CHM13. The organization of these HORs are also described among the 5 continental groups and among the 23 pairs of chromosomes that make up the human karyotype. They analyze the variation in lengths of alpha-satellite HOR arrays, the sequence variation at both large and small scales that impact the organization of HOR arrays. They also describe the variation in alpha-satellite HOR composition and highlight various cases of pronounced differences in alpha-satellite HOR composition for different chromosomes. The authors bring to light novel centromere recombination events between acrocentric chromosomes 13 and 22.

An exciting aspect of the manuscript is the evaluation of the epigenetic variation of these assembled centromeres among global groups. They describe the centromere dip regions (CDR) in these assembled centromeres. They describe CDR variation for different chromosomes among continent groups. Interestingly, the CDR is often found on sequences that are highly homogenous. The team methylation and CENP-A to describe kinetochore sites in a fraction of the assembled centromeres. The authors examine the evolution of centromere arrays at both ancient and modern scales, highlighting how centromeres are constantly evolving and mutating. The manuscript continues to finally take advantage of the enormous dataset and evaluate mutational dynamics of different features of the centromere and describe evolutionary trajectories of centromere HORs. The group show that the CDRs found on centromeric chromatin that is highly similar also show the highest rates of mutability. The authors underscore the vast variation in sequence and structure of the human centromere. They detect various patterns and propose various directions for the field to investigate. Overall, the work described is robust and exciting for the field to understand and appreciate human centromere variation.

Thank you for your helpful and thoughtful comments on our manuscript. We have addressed each comment below, which has undoubtedly improved the quality of the manuscript.

Comments:

1. Please provide clear definitions for the terms centromere, kinetochore, CDR up front.

Thanks for this suggestion. We now define each of these terms in the abstract of the paper as follows (changes in **bold**):

“ABSTRACT

Centromeres are essential **chromosomal regions that ensure** accurate chromosome segregation during cell division, yet their highly repetitive sequence has historically hindered their complete assembly and characterization. Consequently, the full spectrum of centromere diversity across individuals, populations, and evolutionary contexts remains largely unexplored. Here, we address this gap in knowledge by assembling and characterizing 2,110 complete human centromeres from a diverse cohort of individuals representing 5

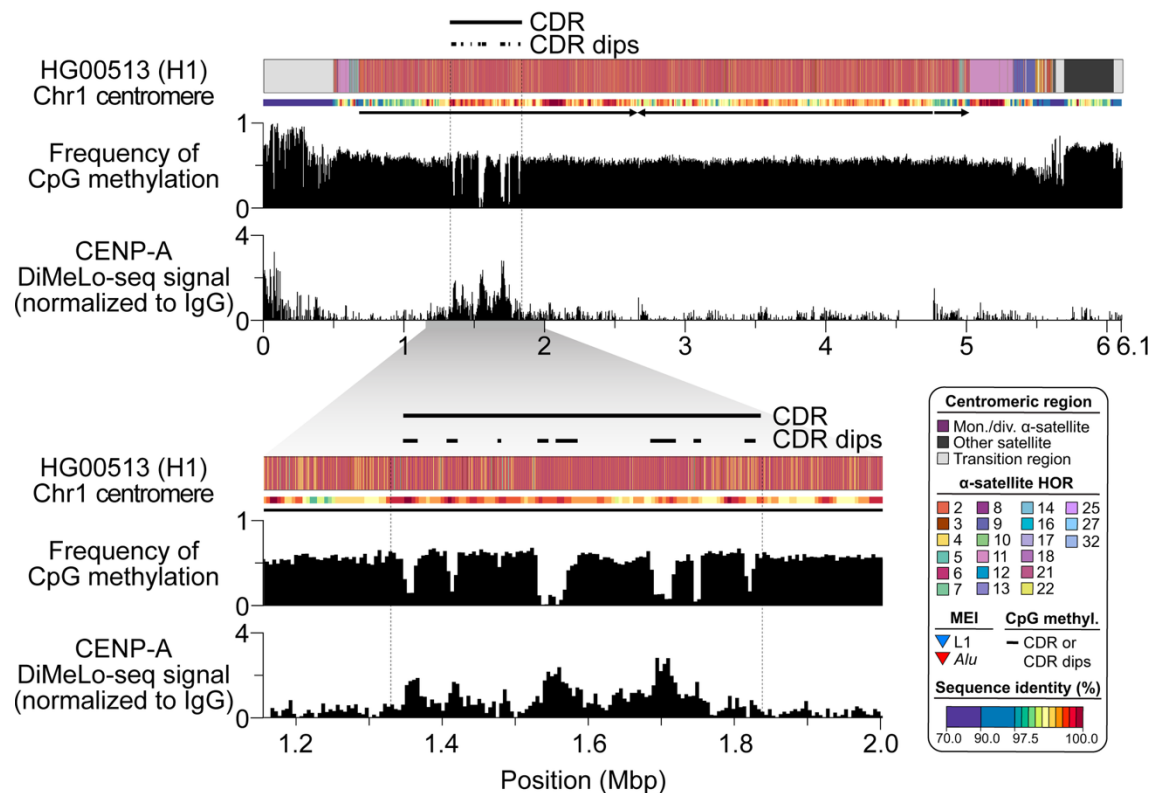
continental groups. By developing a novel suite of bioinformatic tools tailored for centromeric regions, we uncover previously unknown variation within centromeres, including 226 novel centromere haplotypes and 1,870 new α -satellite higher-order repeat (HOR) variants. We find that mobile element insertions are present in 30% of centromeres, with chromosome 16 harboring *Alu* elements within the kinetochore site, **a region marked by nucleosomes containing the histone H3 variant CENP-A and hypomethylated DNA (known as the “centromere dip region”, or CDR) and which serves as the site of spindle microtubule attachment during cell division**, at an 11-fold higher frequency than expected. While most centromeres have a single kinetochore site, 6% of them have di-kinetochores, and $<1\%$ have tri-kinetochores, which we confirm with long-read CENP-A CUT&RUN, DiMeLo-seq, and multi-generational inheritance. We further show that the position of the kinetochore is not random and is, instead, closely associated with the underlying sequence and structure of the centromere. To understand the nature of evolutionary change, we compared 2,110 complete human centromeres to 5,747 complete centromeres recently assembled from the Human Pangenome Reference Consortium. We show that centromeres have a >20 -fold variation in mutation rate, with the most rapidly mutating centromeres on chromosome 1 and the slowest mutating centromeres on chromosome Y. Additionally, a subset of centromeres show evidence of introgression from archaic hominins, shaping their sequence, structure, and evolutionary history. We validate these centromere mutation rates in a four-generation family, spanning 28 family members and 483 accurately assembled centromeres, and show that the kinetochore site is the most rapidly mutating region in the centromere, with 1.8-fold more single-nucleotide variants than the rest of the centromeric α -satellite HOR array on average. We propose a model that reveals an ‘arms race’ between centromeric sequence and proteins, with frequent mutations within the site of the kinetochore that lead to changes in genetic and epigenetic landscapes and, ultimately, rapid evolution of these critically important regions.”

2. An overarching assumption that the paper starts with is the fact that the CDR is the kinetochore-forming region. They also, however, mention that the kinetochore sites often have multiple CDRs within them in later analyses.

a. Therefore, it is unclear in cases where the lengths of kinetochore forming sites are compared, whether this takes into consideration the absolute length of the kinetochore forming chromatin (including CDRs and the space between them) or if this is a summation of the CDR containing chromatin for these chromosomes. This extends to other comparisons between continental groups as well.

Thank you for pointing this out. We define a “CDR” as the absolute length of the kinetochore-forming chromatin domain, from the first hypomethylated base to the last hypomethylated base and inclusive of the sequence in between. This is consistent with the original definition and usage by Gershman et al., *Science*, 2022, who coined the term and described the chromosome X centromere as having one CDR but with multiple hypomethylated regions.

To distinguish between a “CDR” and “CDR dips”, we have included a schematic below to illustrate the difference, and we have modified the main text and Methods to make this clearer.



Reviewer Figure 1. Definition of a “CDR” vs. “CDR dips”. A “CDR” is defined as the region spanning from the first hypomethylated base to the last hypomethylated base and inclusive of the sequence in between. In contrast, “CDR dips” are only the hypomethylated regions themselves within the CDR. CDR dips that are within 150 kbp of each other are considered to be part of the same CDR.

Main text (lines 263-268; changes in **bold**):

“Using the complete centromere assemblies we generated, we sought to uncover how the underlying sequence and structure within centromeres affects the position and distribution of the kinetochore and how this stratifies by continental groups. To do this, we first determined the methylation profile of each centromere by aligning native ONT long-read sequencing data from the same source genome to each assembly and identifying the location of the CDR, **which spans from the first hypomethylated base to the last hypomethylated base**, in each centromere³¹.”

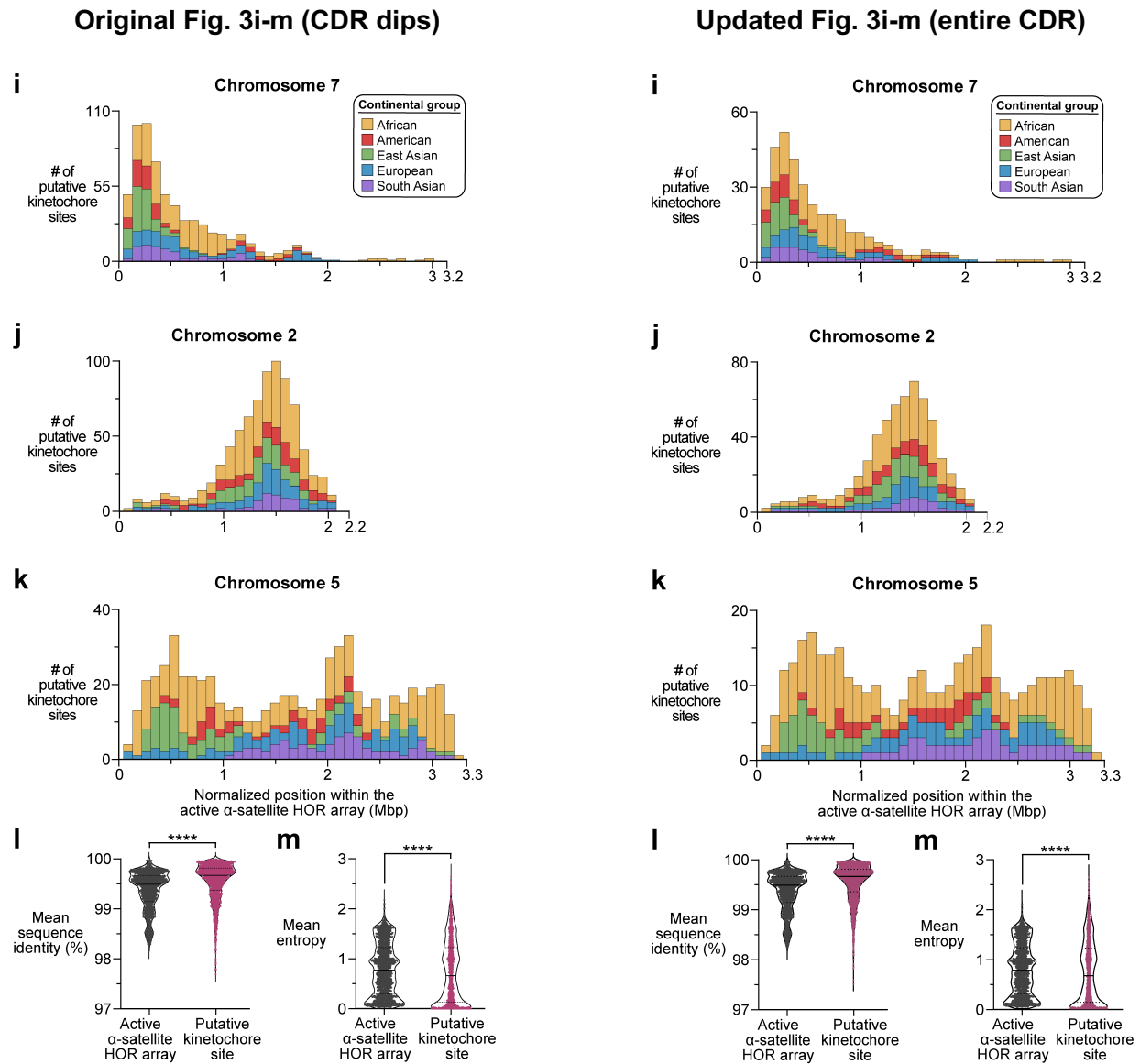
Methods (lines 668-673; changes in **bold**):

“CenMAP uses CDR-Finder³¹ (v1.0.4), a tool to detect hypomethylated regions termed “centromere dip regions” (CDRs²²), **spanning from the first hypomethylated base to the last hypomethylated base and inclusive of the sequence in between**. We manually curated the CDR calls by checking each region outputted in the BED file using IGV⁵¹, removing false positives, recovering missed calls, and **merging calls that were within 150 kbp of each other**.”

b. It is also unclear when looking at the similarity and entropy of kinetochore forming chromatin if this takes into consideration the entire stretch of CDR containing chromatin including the HOR array with high CpG methylation or just the CDRs.

Thank you for pointing this out. While our original analysis only evaluated the sequences within the CDR dips and not the entire CDR itself, we have now redone this analysis on the entire CDR (from the first hypomethylated base to the last hypomethylated base) and updated the figures and supplementary tables accordingly to improve consistency. Importantly, the updated results do not change the conclusions of our original analysis and still show statistically significant differences in sequence identity and entropy between the active α -satellite HOR array and the putative kinetochore sites ($p < 0.0001$, two-tailed Mann-Whitney test).

Below, we show a comparison of the original **Fig. 3i-m** panels on the left (CDR dips) and the updated panels on the right (entire CDR), with subtle differences between the two. We have also updated **Supplementary Figs. 28-30** and **Supplementary Table 11** to reflect these changes as well.



Reviewer Figure 2. Comparison of the distribution, sequence identity, and entropy of CDR dips vs. the entire CDR. Only subtle differences exist between the two sets of results, and the conclusions remain the same.

c. The authors can use the vast assembly data as an opportunity to define what a CDR is, objectively, for the field. This would be a highly significant and beneficial definition for the field.

Thank you for suggesting this. We now define a “CDR” in the main text and Methods as spanning “from the first hypomethylated base to the last hypomethylated base and inclusive of the sequence in between”, as described in our response to Comment 2a (and included below for convenience).

Main text (lines 263-268; changes in **bold**):

“Using the complete centromere assemblies we generated, we sought to uncover how the underlying sequence and structure within centromeres affects the position and distribution of the kinetochore and how this stratifies by continental groups. To do this, we first determined the methylation profile of each centromere by

aligning native ONT long-read sequencing data from the same source genome to each assembly and identifying the location of the CDR, **which spans from the first hypomethylated base to the last hypomethylated base**, in each centromere³¹.”

Methods (lines 668-673; changes in **bold**):

“CenMAP uses CDR-Finder³¹ (v1.0.4), a tool to detect hypomethylated regions termed “centromere dip regions” (**CDRs²²**), **spanning from the first hypomethylated base to the last hypomethylated base and inclusive of the sequence in between**. We manually curated the CDR calls by checking each region outputted in the BED file using IGV⁵¹, removing false positives, recovering missed calls, and **merging calls that were within 150 kbp of each other**.”

3. Supplemental information that will make this manuscript a valuable resource for the field is the inclusion of bed files for the boundaries of all the centromeres defined in the assembled centromeres. This will help standardize future analysis of these centromeres.

Thanks for this suggestion. We now provide the coordinates of all centromeres and their active α -satellite HOR arrays for the HGSC assemblies in **Supplementary Table 7** ($n = 2,110$), HPRC assemblies in **Supplementary Table 22** ($n = 5,747$), and the 4-generation pedigree in **Supplementary Table 23** ($n = 483$). We have also uploaded the BED files with these coordinates, along with their FASTAs and sequence annotations, to figshare, available at the following link: <https://figshare.com/s/cb839beea6ed06d641db>.

Supplementary Table 7. Length of active α -satellite HOR arrays in 2,110 completely and accurately assembled centromeres.

Genome	Chromosome	Coordinate of centromeric region	Coordinate of active α -satellite HOR array
HG00096	1	haplotype1-0000018:120846650-126839919	haplotype1-0000018:121521308-125759934
HG00096	1	haplotype2-0000105:121072185-126614484	haplotype2-0000105:121745614-125567067
HG00096	2	haplotype1-0000026:283801-3479905	haplotype1-0000026:809552-2957405
HG00096	2	haplotype2-0000106:236451-3494750	haplotype2-0000106:758809-2972212
HG00096	3	haplotype2-0000121:76345346-82963448	haplotype2-0000121:78268713-79826353
HG00096	3	haplotype2-0000121:76345346-82963448	haplotype2-0000121:81621059-82379452
HG00096	4	haplotype1-0000031:49455645-53538994	haplotype1-0000031:49955619-51772834
HG00096	4	haplotype1-0000031:49455645-53538994	haplotype1-0000031:52210424-53009688
HG00096	5	haplotype1-0000029:45706213-53389912	haplotype1-0000029:46902778-51648489
HG00096	6	haplotype2-0000117:57813721-62215202	haplotype2-0000117:58357115-61359157
HG00096	7	haplotype2-0000122:58859506-64747156	haplotype2-0000122:60384816-63773280
HG00096	8	haplotype1-0000004:42802921-47440620	haplotype1-0000004:43700949-46713556
HG00096	8	haplotype2-0000100:30941998-34899997	haplotype2-0000100:31835062-34173782
HG00096	9	haplotype2-0000110:44516270-48044169	haplotype2-0000110:45029432-47536154
HG00096	10	haplotype2-0000118:38552782-42152385	haplotype2-0000118:39589884-40153010
HG00096	10	haplotype2-0000118:38552782-42152385	haplotype2-0000118:40189994-41279485
HG00096	10	haplotype2-0000118:38552782-42152385	haplotype2-0000118:41289156-41360205
HG00096	10	haplotype2-0000118:38552782-42152385	haplotype2-0000118:41650481-41652346
HG00096	11	haplotype1-0000034:49760855-55518554	haplotype1-0000034:50764157-54235060

Supplementary Table 22. Coordinates of centromeric regions and active α -satellite HOR arrays in the HPRC genome assemblies.

Genome	Chromosome	Coordinate of centromeric region	Coordinate of active α -satellite HOR array
HG00097	1	HG00097#1#CM094060.1:121483033-128878553	HG00097#1#CM094060.1:122254568-127629795
HG00097	2	HG00097#1#CM094061.1:91548321-95301170	HG00097#1#CM094061.1:92064702-94778658
HG00097	2	HG00097#2#CM094076.1:91860275-94843854	HG00097#2#CM094076.1:92380069-94321314
HG00097	3	HG00097#1#CM094062.1:89852942-97170848	HG00097#1#CM094062.1:91843958-93528091
HG00097	3	HG00097#1#CM094062.1:89852942-97170848	HG00097#1#CM094062.1:95812465-96585197
HG00097	3	HG00097#2#JBIRDE01000040.1:89817073-95522928	HG00097#2#JBIRDE01000040.1:91467216-92497670
HG00097	3	HG00097#2#JBIRDE01000040.1:89817073-95522928	HG00097#2#JBIRDE01000040.1:94451949-94941389
HG00097	5	HG00097#1#CM094063.1:45862451-52024708	HG00097#1#CM094063.1:47409456-47435957
HG00097	5	HG00097#1#CM094063.1:45862451-52024708	HG00097#1#CM094063.1:47468936-50177641
HG00097	5	HG00097#2#CM094077.1:45715550-53913949	HG00097#2#CM094077.1:46914886-51018376
HG00097	5	HG00097#2#CM094077.1:45715550-53913949	HG00097#2#CM094077.1:51548096-51565249
HG00097	5	HG00097#2#CM094077.1:45715550-53913949	HG00097#2#CM094077.1:52143543-52161036
HG00097	8	HG00097#1#CM094064.1:42776987-46022436	HG00097#1#CM094064.1:43685460-45296193
HG00097	8	HG00097#2#CM094079.1:42675598-49305797	HG00097#2#CM094079.1:43569729-48578707
HG00097	10	HG00097#1#CM094066.1:38677297-42332946	HG00097#1#CM094066.1:39732676-41433295
HG00097	10	HG00097#2#CM094081.1:38528602-42826755	HG00097#2#CM094081.1:39584705-42023140
HG00097	11	HG00097#1#CM094067.1:49956836-54652435	HG00097#1#CM094067.1:50951932-53320461
HG00097	11	HG00097#1#CM094067.1:49956836-54652435	HG00097#1#CM094067.1:53371916-53374126
HG00097	11	HG00097#2#CM094082.1:49934463-55568362	HG00097#2#CM094082.1:50944105-54265604
HG00097	12	HG00097#2#CM094083.1:33974542-38227091	HG00097#2#CM094083.1:34834222-37571020
HG00097	13	HG00097#1#JBIRDD01000002.1:4701612-8578377	HG00097#1#JBIRDD01000002.1:6369985-7982572
HG00097	13	HG00097#2#JBIRDE01000006.1:4607158-8097753	HG00097#2#JBIRDE01000006.1:5788905-7508706
HG00097	14	HG00097#1#JBIRDD010000018.1:7462694-10965043	HG00097#1#JBIRDD010000018.1:7984517-10390929

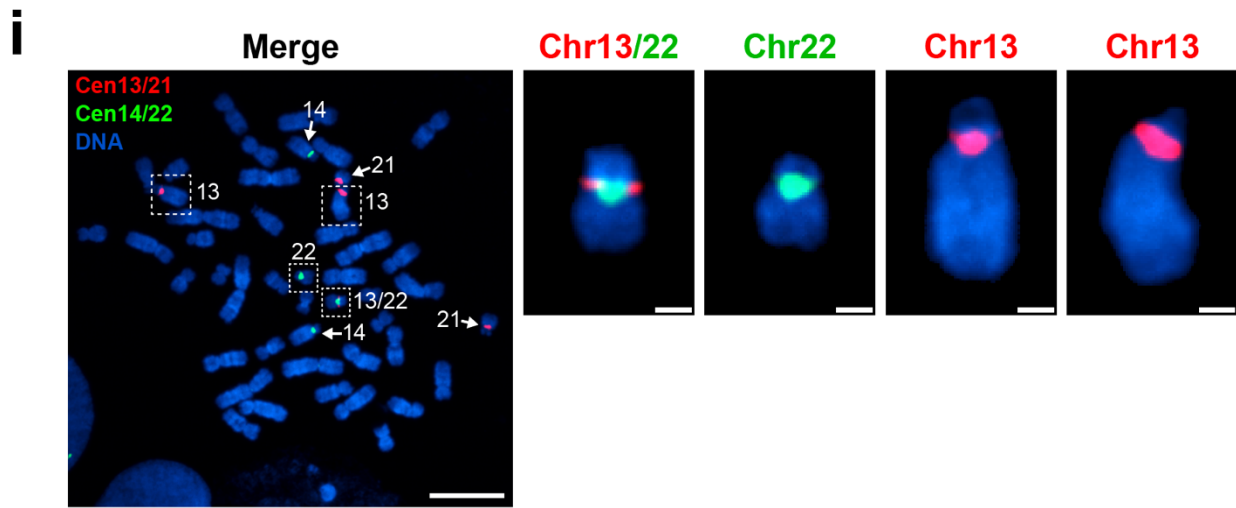
Supplementary Table 23. Coordinates of centromeric regions and active α -satellite HOR arrays in the 4-generation, 28-member pedigree.

Generation	Genome	Chromosome	Coordinate of centromeric region	Coordinate of active α -satellite HOR array
G1	NA12889	1	haplotype1-0000001:120737016-125955414	haplotype1-0000001:121440638-124897181
G1	NA12889	1	haplotype2-0000071:121453763-126214561	haplotype2-0000071:122115737-125169192
G1	NA12889	2	haplotype1-0000007:91953476-94450374	haplotype1-0000007:92472392-93927852
G1	NA12889	2	haplotype2-0000078:91910757-95436445	haplotype2-0000078:92433091-94913906
G1	NA12889	3	haplotype1-0000004:89830229-97490427	haplotype1-0000004:91695158-92992501
G1	NA12889	3	haplotype1-0000004:89830229-97490427	haplotype1-0000004:96695845-96906440
G1	NA12889	3	haplotype2-0000075:89844210-98127508	haplotype2-0000075:91317039-93024029
G1	NA12889	3	haplotype2-0000075:89844210-98127508	haplotype2-0000075:97028580-97543562
G1	NA12889	4	haplotype1-0000014:49473893-54030837	haplotype1-0000014:49973890-51683742
G1	NA12889	4	haplotype1-0000014:49473893-54030837	haplotype1-0000014:52791386-53508136
G1	NA12889	4	haplotype2-0000087:49792657-53987356	haplotype2-0000087:50292623-51325239
G1	NA12889	4	haplotype2-0000087:49792657-53987356	haplotype2-0000087:52656905-53464705
G1	NA12889	5	haplotype1-0000002:46024020-52050068	haplotype1-0000002:47292103-50062198
G1	NA12889	6	haplotype2-0000093:58008996-61441390	haplotype2-0000093:58552367-60591446
G1	NA12889	8	haplotype1-0000006:42638282-46489380	haplotype1-0000006:43532388-45762316
G1	NA12889	8	haplotype2-0000076:43386615-47829790	haplotype2-0000076:44279676-47090206
G1	NA12889	9	haplotype1-0000017:44665330-48351778	haplotype1-0000017:45178497-47843550
G1	NA12889	10	haplotype1-0000027:38614176-42447154	haplotype1-0000027:39669349-41559401
G1	NA12889	10	haplotype1-0000027:38614176-42447154	haplotype1-0000027:41945248-41947113
G1	NA12889	11	haplotype1-0000011:49707541-54132315	haplotype1-0000011:50696139-52850782
G1	NA12889	13	haplotype1-0000026:4226718-6705967	haplotype1-0000026:5677738-6109314
G1	NA12889	13	haplotype2-0000094:4143744-8515992	haplotype2-0000094:6222193-7924410
G1	NA12889	14	haplotype1-0000005:5247508-8093261	haplotype1-0000005:5769321-7520522

4. Authors should use a more convincing representative chromosome spread image showing a 13/22 recombined chromosome (figure 2i).

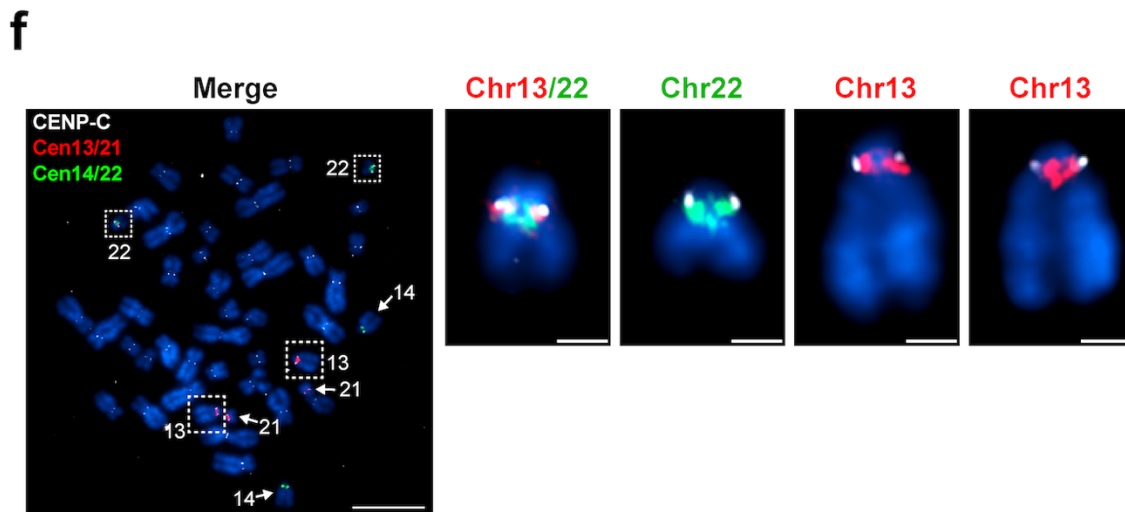
Thanks for this suggestion. To improve the quality of this image, we initiated a collaboration with Dr. Jennifer Gerton (Stowers) and her research scientist, Dr. Tamara Potapova, who are experts in performing FISH on metaphase chromosome spreads. They performed this experiment on the same GM20355 cell line, using fluorescent DNA probes specific to the chromosome 13/21 α -satellite DNA and chromosome 14/22 α -satellite DNA. Their images more clearly show a hybrid chromosome with both chromosome 13/21 α -satellite (red) and

14/22 α -satellite (green), confirming that this chromosome has α -satellite from two chromosomes at the primary constriction (see the updated Fig. 2i).



Updated Fig. 2i. Fluorescent *in situ* hybridization (FISH) on an GM20355 metaphase chromosome spread showing the recombined chromosome 13/22. Fluorescent probes specific to chromosome 13/21 α -satellite (red) and 14/22 α -satellite (green) colocalize on the recombined chromosome 13/22 in 100% of cells ($n=19$). Scale bar (chromosome spread), 10 μm ; scale bar (individual chromosome), 1 μm .

To test if the chromosome 13 α -satellite HOR array is the site of kinetochore assembly (as predicted by our CDR analysis), Dr. Potapova performed immunoFISH on the same cell line, using the same fluorescent DNA probes and an antibody to CENP-C, an inner-kinetochore protein. Her images show that CENP-C colocalizes specifically with the chromosome 13 α -satellite DNA on the hybrid chromosome 13/22, indicating that chromosome 13 α -satellite is the site of kinetochore assembly and confirming our original findings. We have included a representative image from this experiment in **Supplementary Fig. 23f** (shown below).



New Supplementary Fig. 23f. Immunolabeling and fluorescent *in situ* hybridization (immunoFISH) on an GM20355 metaphase chromosome spread, showing colocalization of CENP-C with chromosome 13/21 α -satellite (red), which is adjacent to 14/22 α -satellite (green) on the recombined chromosome 13/22 ($n=30$). Scale bar (chromosome spread), 10 μm ; scale bar (individual chromosome), 1 μm .

We have updated the Methods (lines 903-946) to describe these new experiments and have moved the original FISH image to **Supplementary Fig. 23e**.

5. When commenting on the number of MEIs, the authors show that chromosome 16 has an elevated incidence of MEIs and correlate this with the known propensity of this chromosome to show nondisjunction during meiosis. It will be useful when speculating on this correlation to show another example of a chromosome that shows a lower incidence of MEIs and does not show a high propensity for nondisjunction in meiosis.

We have now revised our manuscript to include an example of a chromosome with a low incidence of MEIs within the putative kinetochore site and concomitant lack of association with meiotic nondisjunction (chromosome 1).

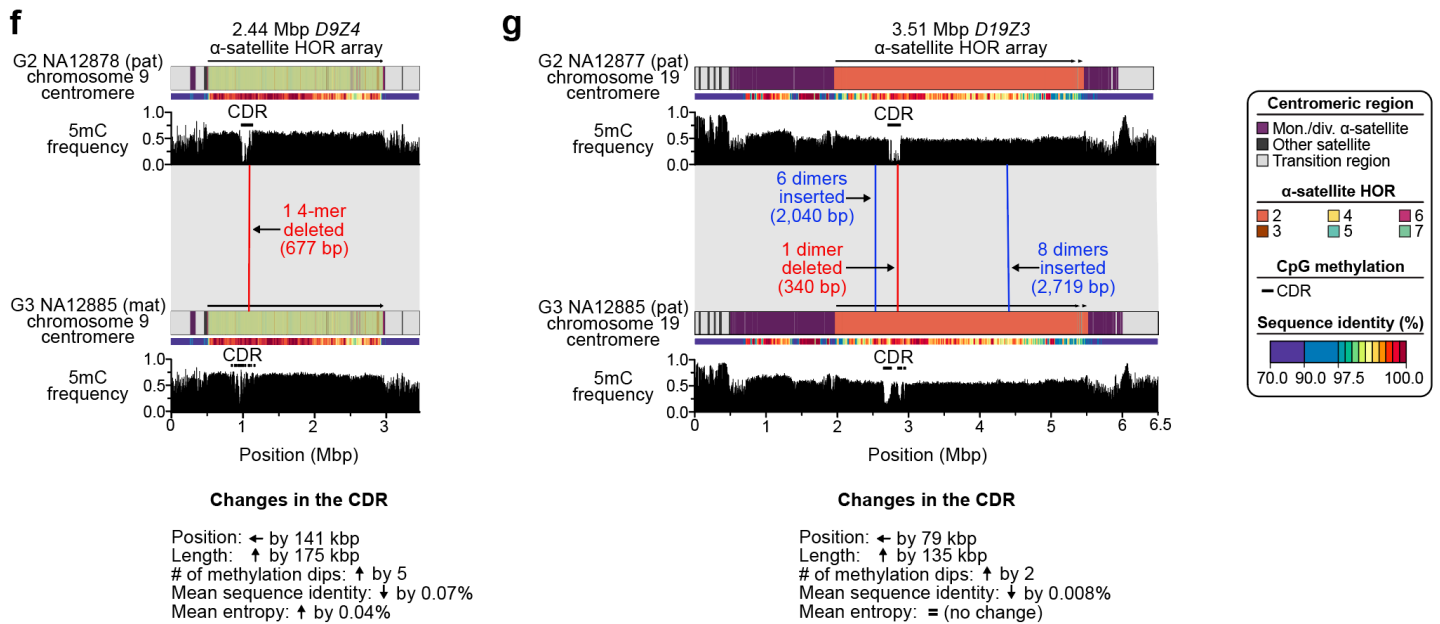
Main text (lines 297-310; changes in **bold**):

“While mobile element insertions (MEIs) within kinetochore sites are rare, they occur frequently within the chromosome 16 centromere, with 14% (16/112) of centromeres harboring an MEI within the putative kinetochore site, which is 17 times more frequent than in the other chromosomes (Supplementary Fig. 28a-d, Supplementary Table 10). This is notable, as chromosome 16 is known to have the highest rate of nondisjunction during meiosis and is the most frequently gained or lost chromosome in aneuploid embryos^{30,31}. **In contrast, chromosomes such as chromosome 1 show a markedly lower incidence of MEIs within the putative kinetochore site (Supplementary Fig. 28a-d, Supplementary Table 10) and are not associated with elevated rates of meiotic nondisjunction, suggesting a potential relationship between MEI burden and segregation fidelity.** Previous studies suggest that errors in kinetochore-microtubule attachments can lead to increased rates of chromosome missegregation³², and MEIs within the kinetochore site may disrupt the 3D chromatin organization required for proper spindle microtubule attachment. Additional studies that test the functional consequences of MEIs within kinetochore sites, such as those found on chromosome 16, will need to be performed to confirm their contribution to increased chromosome missegregation rates.”

6. Supplemental Figure 77 should be made clearer to show the changes in the CDR vs active HOR array from G1 to G4 for different details analyzed. It is also unclear why the number of dips in CpG methylation is examined here.

To more clearly demonstrate the changes in the CDR vs the active α -satellite HOR array across generations, we added two new panels to **Supplementary Fig. 77** showing examples of transmitted centromeres with mutations in or near the CDR, resulting in changes in its position, length, distribution, and sequence composition in the next generation (**Supplementary Fig. 77f,g**). In panel f, we show how a 677-bp deletion in the CDR in a G2 chromosome 9 centromere shifts the central position of the CDR towards the p-arm by 141 kbp, increases the length of the CDR by 175 kbp, increases the # of dips in methylation by 5, reduces the mean sequence identity by 0.07%, and increases the mean entropy by 0.04 in G3. Similarly, in panel g, we show that a 340-bp deletion in the CDR results in a shift in its central position towards the p-arm by ~79 kbp, an increase in length by ~135 kbp, an increase in the # of dips in methylation by 2, and a reduction in the mean sequence identity by 0.008%. All of these statistics are listed in **Supplementary Table 21**.

The # of dips in methylation is examined in our analyses to show how the CDR distribution changes across generations, with some dips in the methylation lost and others gained (as also reported by Xu et al., *bioRxiv*, 2026). These changes are also shown in the two new panels with centromeres (**panels f and g**), where you can see one major dip in methylation changing into 5 or 3, respectively, from G2 to G3.



Supplementary Figure 77f,g. Examples of differences in the putative kinetochore site for the **f)** chromosome 9 and **g)** chromosome 19 centromeres across generations. In panel f, a 677-bp deletion in the CDR in the chromosome 9 centromere shifts the central position of the CDR towards the p-arm by 141 kbp, increases the length of the CDR by 175 kbp, increases the # of dips in methylation by 5, reduces the mean sequence identity by 0.07%, and increases the mean entropy by 0.04. Similarly, in panel g, a 340-bp deletion in the CDR results in a shift in the central position of the CDR towards the p-arm by ~79 kbp, an increase in length by ~135 kbp, an increase in the # of dips in methylation by 2, and a reduction in the mean sequence identity by 0.008%.

7. When analyzing the number of archaic hominin specific Kmers in haplotypes of chromosomes 12 and 21, it looks like the enrichment of these Kmers are specifically outside the CDR. It would be interesting for the authors to speculate on this and how this kmer analysis of ancient and modern centromere haplotypes fit with the model of heightened recombination in highly similar CDR chromatin.

This is a great point. Archaic hominin-specific *k*-mers do appear to be de-enriched in the CDR relative to the rest of the α -satellite HOR array and flanking sequences. This is consistent with our model of centromere evolution, wherein sequences within the CDR are constantly evolving and homogenizing, increasing their sequence identity and lowering their entropy, whereas other sequences (especially those in the transition regions from the active α -satellite HOR array to euchromatin) are accumulating mutations and diverging over time. Your observation that archaic hominin *k*-mers are enriched outside the CDR is in line with this model, as archaic DNA is likely cleared more rapidly due to several rounds of mutations within the CDR. We now speculate on this finding in the Discussion as follows:

Discussion (lines 497-512; changes in **bold**):

“Across chromosomes, we observe strong positional biases of the kinetochore site (p arm-proximal, q arm-proximal, or haplotype-dependent distributions) and show that the sequence underlying the kinetochore is, on average, more homogeneous and of higher sequence identity than the surrounding α -satellite HORs. This association indicates that certain sequence contexts (e.g. highly identical, low-entropy stretches of α -satellite DNA) are preferential substrates for kinetochore establishment and/or maintenance. This finding helps reconcile models in which centromeric proteins and DNA co-evolve (e.g. KARM^{33,34}) and suggests that local sequence homogeneity may facilitate the assembly or maintenance of centromeric chromatin. **Consistent with this model, we observe that archaic hominin-specific *k*-mers are de-enriched within the putative kinetochore site relative to the surrounding α -satellite HOR array and flanking regions. This pattern suggests that sequences within the kinetochore site are subject to continual turnover and homogenization, increasing sequence identity while reducing sequence entropy, whereas adjacent regions, such as the transition between active α -satellite HOR arrays and euchromatin, accumulate**

mutations and diverge over time. The relative enrichment of archaic-derived sequence outside the putative kinetochore site is therefore consistent with a model in which introgressed or ancestral variants are rapidly eroded within kinetochore-associated regions through repeated cycles of mutation and homogenizing recombination.”

8. The authors should discuss the limitations of using only three neanderthal and one denisovan genome in their kmer analysis. The main crux of the paper involves describing the vast variability in human centromeres. Therefore, using a handful of kmers to determine the evolution of chromosomes 10, 12 and 21 come with the caveat that the kmers used may not be sampling all ARC-kmers completely.

Thank you for pointing out the potential caveat of using a limited number of archaic genomes in this analysis. We now discuss this limitation in the Discussion immediately after the paragraph above as follows:

Discussion (lines 512-517; changes in **bold**):

“Although data from three Neanderthal and one Denisovan genome provide only a limited snapshot of archaic hominin-specific *k*-mers, the distinct signatures observed on chromosomes 10 and 21 support the presence of archaic-like centromeres in modern humans. Beyond introgression, alternative scenarios, such as ancestral polymorphism or independent convergent mutation, may also explain this variation. Further work will be warranted to disentangle these possibilities and provide a more comprehensive view of human centromere evolution.”

9. The authors do not mention the 28 population groups in any of the analyses and stick to describing similarities and differences in 5 continental groups. The authors may not need to mention the 28 population groups if they are not going to use them in the analysis?

This is a good point. We have now removed all mention of the 28 population groups from the Abstract and main text.

10. There are some occurrences where an acronym is used without being defined. (MEI, SV, SNVs) it will be useful for these to be defined the first time its used.

This is an excellent suggestion. We now define “MEI”, “SVs”, “SNVs” the first time they are used in the main text, as shown below.

Main text (lines 297-300; changes in **bold**):

“While **mobile element insertions (MEIs)** within kinetochore sites are rare, they occur frequently within the chromosome 16 centromere, with 14% (16/112) of centromeres harboring an MEI within the putative kinetochore site, which is 17 times more frequent than in the other chromosomes (Supplementary Fig. 28a-d, Supplementary Table 10).”

Main text (lines 422-425; changes in **bold**):

“Among these 483 centromeres, 244 of them were transmitted from one generation to the next (Supplementary Fig. 76c, Supplementary Table 15), allowing us to identify *de novo* **single-nucleotide and structural variants (SNVs and SVs, respectively)** that occur within the centromeric regions across generations.”

11. In terms of the mutation rate at the centromere, it seems both that 1) slippage from DNA replication may happen at a higher rate due to the repetitive nature of the sequence and 2) that individual mutations may not be strongly selected against and 3) that recombination may be suppressed due to heterochromatin. It hard to know how the balance of all these processes plays out in the observed sequence-it would be nice to include some discussion of these processes (replication, recombination, repair, selection) and their contribution to the observations.

This is a great suggestion. We have now added the following section to the Discussion:

Discussion (lines 525-541; changes in **bold**):

“The elevated mutation rates we observe at centromeres likely reflect the combined influence of multiple molecular processes acting on highly repetitive DNA. First, the tandemly repeated structure of α -satellite HOR arrays may predispose these regions to increased replication-associated errors, including polymerase slippage and template misalignment, which can generate both small-scale mutations and repeat length variation. Second, the heterochromatic environment of centromeres is known to suppress canonical meiotic recombination, potentially limiting the efficacy of homologous recombination-based repair pathways and altering the balance of DNA repair mechanisms operating in these regions. At the same time, the high sequence identity within α -satellite HOR arrays may facilitate non-allelic homologous recombination or gene conversion-like processes, promoting rapid sequence homogenization while also generating structural variation through unequal exchange. Finally, selective constraints at centromeres are likely complex: while strong purifying selection may act to preserve core features required for kinetochore assembly, many individual nucleotide changes within repetitive arrays may be effectively neutral or only weakly deleterious, allowing mutations to accumulate. The interplay between replication error, recombination suppression and misalignment, DNA repair pathway choice, and relatively relaxed sequence-level constraint likely shapes the distinctive mutational landscape of centromeres, although further work will be required to disentangle the relative contributions of these processes.”

Referee #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

[Thank you for your help in reviewing this manuscript.](#)

Referee #3 (Remarks to the Author):

Gao, Oshima et al. present an exciting follow-up to the HGSVC paper from last year (Logsdon et al. Nature 2025), with >2000 fully assembled centromeres from 65 near-complete human genome assemblies drawn from across the globe. A novel pipeline reveals hundreds of major centromere haplotypes as well as >1800 novel alphaSat HOR variants. Using these data, the authors report remarkable variation in kinetochore binding sites, HOR arrays, as well as large structural rearrangements (and a striking chromosome 13-22 recombination present in multiple individuals). Using CENP-A DiMeLo-seq and CENP-A CUT&RUN, the authors validate individual kinetochore sites (occupied by CENP-A), showing that almost 95% of centromeres have only one kinetochore site closely linked to the sequence/layout of the centromere, with kinetochores preferentially positioned in regions with homogenous HOR distributions; analysis of individual kinetochore sites revealed a fairly frequent mobile element insertion within the kinetochore sites of chromosome 16. Finally, the authors cluster haplotypes to create phylogenetic trees for each centromere with chimpanzees as outgroups, revealing a large number of recent evolutionary changes in centromere sequence/structure and also higher than expected mutation rates.

Overall, the paper is well written, the methods are extensively documented, and the figures are well-organised (albeit incredibly dense, and somewhat unsuited to the static 2D PDF format given how data-rich they are, more on that below). The raw data and centromere annotations have been made available for download. It is clear from this work - in line with data from other eukaryotic clades - that centromeres evolve so quickly and so dramatically that the imprint of recent and large-scale evolution is writ large on the genomes of extant human populations, with major implications for models of genome evolution, epidemiology and medicine, but also clear cell biological predictions related to the mode and tempo of kinetochore site evolution.

I have a few comments which if addressed would heighten the utility of this work for others in the field.

Thank you for your helpful suggestions and comments, which we've addressed below. They have improved the manuscript and increased the utility of the work.

The clustering/thresholding approach used to delineate centromere haplotypes from each other is not well described in the methods. This is critical given the complexity and size of satellite regions and the combinatorial diversity of the alphaSat HORs - presumably differently coarse-grained clustering approaches would lead to a varying number of haplotypes, in turn influencing the analyses in Figure 1 and the phylogenetic analyses later in the paper. More importantly, it would be difficult to reconcile these definitions with future analyses coming in from other big global and regional genomics projects.

This is a good point. We have now updated the Methods to describe how we identify major centromere haplotypes (see below), and we have provided a **new Supplementary Table 12** that lists the coordinates by which we generated each phylogenetic tree (see below). We hope this will enable others to recapitulate this analysis in their own studies.

New Methods section (lines 968-976):

"Identification of major centromere haplotypes

To identify major centromere haplotypes, we first compared the phylogenetic trees from both p- and q-arm and selected the tree that exhibited higher topological congruence with the distribution and architecture of the active α -satellite HOR array as well as higher bootstrap values. Then, we partitioned the tree based on a divergence time of 100 kya. The 100 kya threshold was determined following a comprehensive inspection of all centromere phylogenies, providing an optimal balance between intra-clade sequence homogeneity and population-level stratification. Finally, we manually reviewed each clade to identify haplotypes that were distinct from those in other clades. One haplotype from each clade was selected as a representative of that clade and included in our list of major centromere haplotypes."

Supplementary Table 12. Regions within the CHM13 v2.0 reference genome selected for phylogenetic tree reconstruction.

Chromosome	Coordinates	Arm	Used to identify haplogroup?					
1	chr1:121380600-121400600	p	No					
	chr1:126797634-126817634	q	Yes					
2	chr2:91868340-91888340	p	Yes					
	chr2:94728229-94748229	q	No					
3	chr3:91546303-91566303	p	Yes					
	chr3:96507966-96527966	q	No					
4	chr4:48953201-48973201	p	No					
	chr4:55315653-55335653	q	Yes					
5	chr5:46500000-46520000	p	Yes					
	chr5:51128191-51148191	q	No					
6	chr6:58168698-58188698	p	No					
	chr6:61416786-61436786	q	Yes					
7	chr7:59937150-59957150	p	Yes					
	chr7:63807657-63827657	q	No					
8	chr8:44154876-44174876	p	No					
	chr8:46369500-46389500	q	Yes					
9	chr9:44844000-44864000	p	Yes					
	chr9:47642000-47662000	q	No					
10	chr10:39461256-39481256	p	Yes					
	chr10:41950495-41970495	q	No					
11	chr11:50830000-50850000	p	Yes					
	chr11:54589006-54609006	q	No					
12	chr12:34520000-34540000	p	Yes					
	chr12:37385109-37405109	q	No					
13	--	p	No					
	chr13:17577000-17597000	q	Yes					
14	--	p	No					
	chr14:12798743-12818743	q	Yes					
15	--	p	No					
	chr15:17803100-17823100	q	Yes					
16	chr16:35686921-35706921	p	Yes					
	chr16:37892480-37912480	q	No					
17	chr17:23413013-23433013	p	Yes					
	chr17:27683282-27703282	q	No					
18	chr18:15433210-15453210	p	No					
	chr18:21214412-21234412	q	Yes					
19	chr19:25753231-25773231	p	No					
	chr19:29992317-30012317	q	Yes					
20	chr20:26250625-26270625	p	No					
	chr20:29554713-29574713	q	Yes					
21	--	p	No					
	chr21:11353689-11373689	q	Yes					
22	--	p	No					
	chr22:15902587-15922587	q	Yes					
X	chrX:57322095-57342095	p	Yes					
	chrX:61247427-61267427	q	No					
Y	chrY:10463747-10483747	p	Yes					
	chrY:11376655-11396655	q	No					

I have a similar comment with regards to the identification and definition of novel HORs. How can these be validated given the absence of any independent functional or phylogenetic definition? As I understand it each novel genome in the future will then result, almost axiomatically, in the identification of new HORs. Would it not make sense to focus only on those with representation across multiple genomes (or did the authors implement this filter? Perhaps I missed it)

This is an excellent point. To validate the novel α -satellite HORs that we discovered in each genome, we did indeed implement a filter, requiring that an α -satellite HOR exist at least twice in the genome it was identified in and/or be confirmed in another genome from the broader HGSVC and HPRC dataset. This strict requirement reduced our candidate α -satellite HORs from 2,709 to 1,870 novel α -satellite HORs we report in our study (a 31% reduction). We modified the main text to describe this filtering strategy and make this clearer.

Main text (lines 134-139; changes in **bold**):

“Applying this pipeline to all 2,110 centromeres revealed 226 new major centromere haplotypes (Fig. 1, Supplementary Figs. 3-6) as well as 1,870 new α -satellite HOR variants compared to CHM13 and CHM1 (Supplementary Table 4), **all of which occur at least twice in the genome they were identified in, and 97% of which occur in at least one other genome in our dataset or in the broader set of genome assemblies recently released by the Human Pangenome Reference Consortium (HPRC; Supplementary Table 5; Methods).**”

The distribution of alphaSat HORs is absolutely impossible to “read”/infer from most of the figures, particularly Figure 1. I found the stacked representation in Figure 2e at least easier to parse/visualise, but the overlapping colours/lines are simply too small in most of the other panels. Would there be any way to create a software visualisation tool that could accompany the paper? A way to browse through the haplotypes and annotations, focussing either on individuals, or chromosomes across individuals, or specific HORs and their phylogenetic representations, etc. I understand the challenges involved in creating something versatile and also user friendly but I believe it would be well worth the investment - and also future-proof these analyses, allowing others to build on the findings here as new genomes come in from around the world.

This is such a great suggestion. We took your advice to build an interactive tool to visualize the centromeres and developed a new tool called Cencyclopedia (www.cencyclopedia.com), an interactive and dynamic browser for centromeres (see Reviewer Figs. 3-6 below). This tool, built in collaboration with Yafei Mao's group at Shanghai Jiao Tong University, allows one to click on any centromere described in our paper and visualize its α -satellite HOR annotation, ONT methylation, CDR(s), MEI(s), local sequence identity, orientation, and sequence identity heat map.

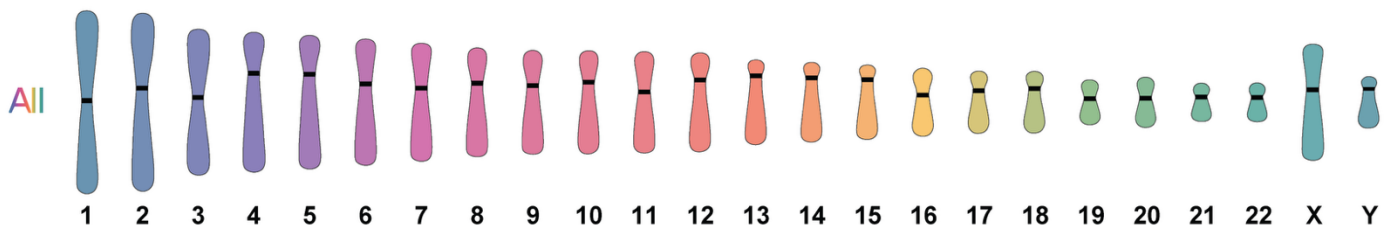


CENCYCLOPEDIA

A CENTROMERE VISUALIZATION TOOL

Cencyclopedia is an interactive visualization tool that allows you to investigate the sequence, structure, epigenetic landscape, and evolutionary relationships among human centromeres

Start by picking a chromosome



Developed by

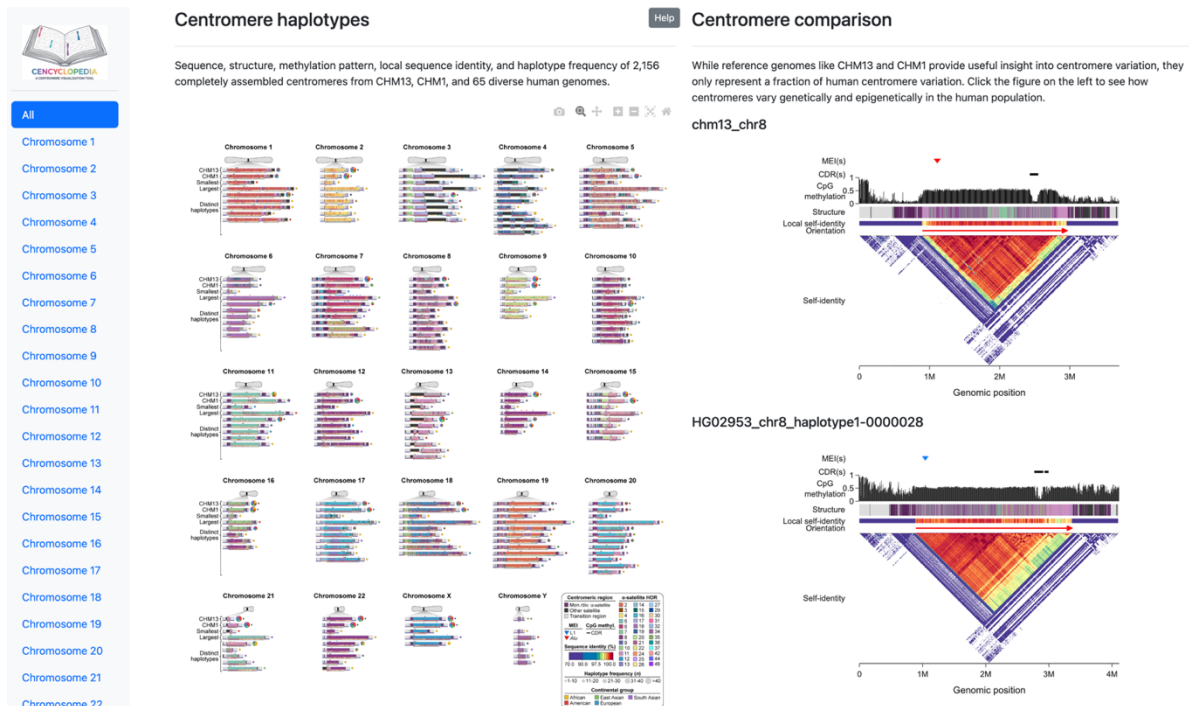
- Keisuke K. Oshima (Logsdon Lab)
- Zikun Yang (Mao Lab)

We thank the anonymous reviewer who suggested we build this interactive centromere visualization tool.

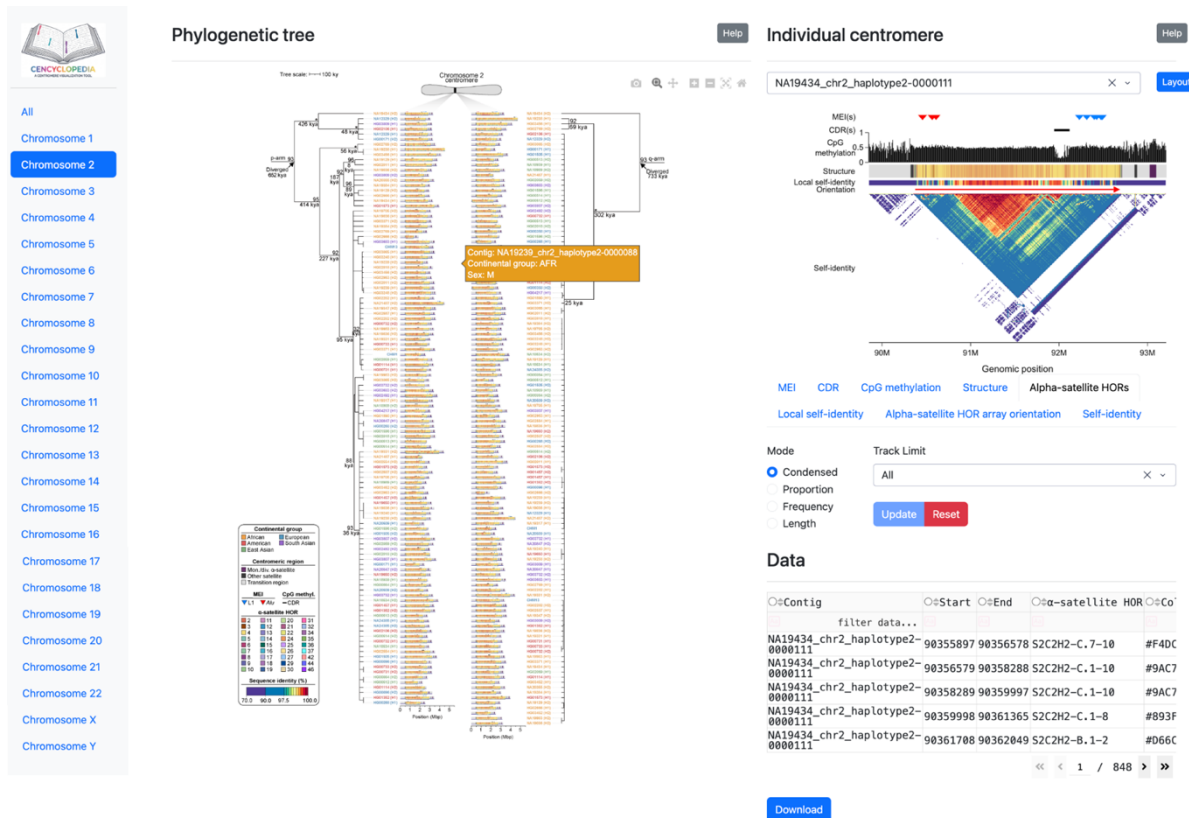
Citation

Gao S, Oshima KK, Chuang SC, Loftus M, Montanari A, Gordon DS, Human Genome Structural Variation Consortium, Human Pangenome Reference Consortium, Hsieh P, Konkel MK, Ventura M, Logsdon GA. A global view of human centromere variation and evolution. *bioRxiv*. 2025. p. 2025.12.09.693231. [doi:10.64898/2025.12.09.693231](https://doi.org/10.64898/2025.12.09.693231)

Reviewer Figure 3. Cencyclopedia, a new interactive and dynamic centromere browser. The home page of www.cencyclopedia.com shows the logo and describes the tool. It also allows for the user to click on “All” or a particular chromosome to visualize centromeres from that chromosome. Finally, we acknowledge the developers of the tool, and we thank you specifically for suggesting we build this tool. We also provide a citation to our preprint.

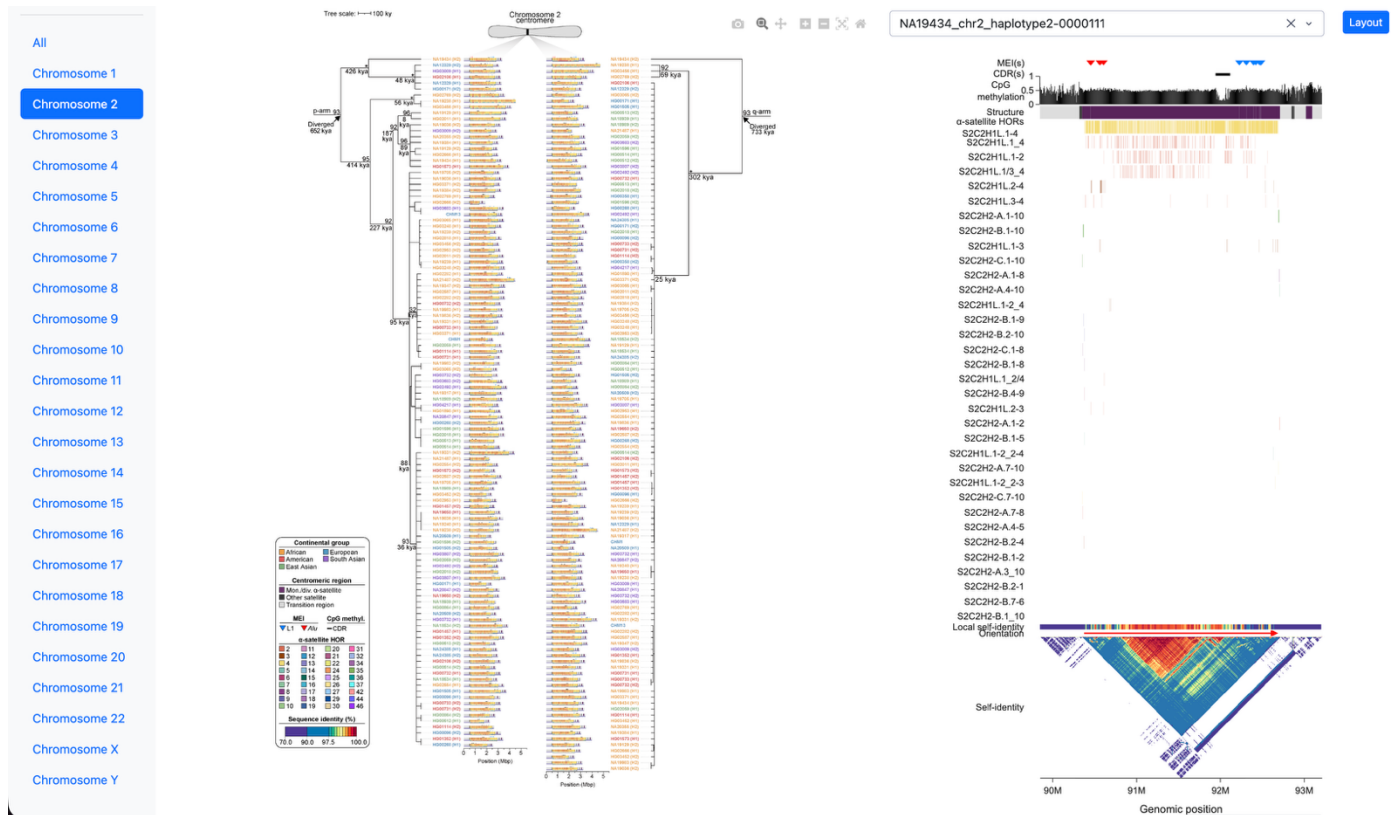


Reviewer Figure 4. Interactive and comparative analyses of centromeres. Clicking on “All” takes the user to a new page that shows the major centromere haplotypes we describe in our manuscript (on the left) and a comparison of the CHM13 centromere and a centromere one selects (on the right). This page is interactive and dynamic; a user can click on any centromere in the figure on the left, and its plot will render in real time on the right to enable the comparison. Additionally, users can hover over the α -satellite HOR annotation track to reveal which HOR it is, and they can zoom in and out of the centromere to visualize different regions.



Reviewer Figure 5. Phylogenetic relationships of centromeres. Clicking on a particular chromosome will bring the user to a new page that shows the phylogenetic tree for that chromosome (on the left) and plots of an

individual centromere (on the right). A user can hover over different chromosomes in the phylogenetic tree to see their ID, continental group, and sex. Clicking on a centromere will render a plot of that centromere in real time and allow for interactive investigation of that centromere as well.



[lab/Snakemake-DiMeLo-seq-Viz](#)); CUT&RUN-Viz (<https://github.com/logsdon-lab/Snakemake-CutNRun>); Pan-centromeric mutation rate estimation (<https://github.com/logsdon-lab/Snakemake-MutationRate>); **and Cencyclopedia** (<https://github.com/logsdon-lab/Cencyclopedia>).”

We note that if there are any issues loading www.cencyclopedia.com (e.g., due to a firewall), the user can access it at this alternative link: <https://cencyclopedia.onrender.com/>. We have also uploaded the source code for users to run on their local computer, available at this link: <https://github.com/logsdon-lab/Cencyclopedia>.

Referee #4 (Remarks to the Author):

Key results:

In this article, Gao, Oshima et al. have assembled fully and characterized 2110 complete human centromeres from 65 individuals representing 28 different populations from 5 continents. This allows for an unprecedented global view of human centromere diversity, as they combined this dataset with previously described ones, reaching 8340 total centromeres. The development of specific bioinformatic tools led to the discovery of 226 major new centromere haplotypes and 1870 -satellite higher-order repeats variants, refining our knowledge of centromeric diversity across individuals and populations and even on a 4 generation pedigree.

For 7 out of 24 chromosomes it was determined that the new reference human genome T2T-CHM13 does not represent the most common centromere haplotype, highlighting the relevance of this study for the broad scientific community, and the centromere field in particular. On a population-scale, the data presented revealed several mechanisms that contribute to centromeric diversity, such as large and small structural rearrangements, and the insertion of mobile elements in ~30% of centromeres. From an evolutionary point of view, this broad dataset has unveiled ancient (up to ~1 million years ago) and recent (down to 1000 years ago, or ~50 generations) evolutionary events that have shaped the modern human centromere landscape, such as the introgression of archaic hominins' satellites in certain minor haplotypes or the phylogenetic separation of particular haplotype variants in specific populations.

Overall the findings of this article provide multiple supporting evidences for the proposed model of centromere evolution.

[Thank you for your thoughtful and helpful comments on our manuscript, which we've addressed below.](#)

Validity:

This manuscript does not present any evident flaws which should prohibit its publication.

Originality and significance:

This is mainly a descriptive article of the assembly and characterization of a wide cohort of human centromeres. These analyses were achieved through the development and/or refinement of a suite of bioinformatics tools specifically developed for these purposes, and that could be re-applied in future research.

Many of the concepts for the main model of centromere evolution have already been put forward in the past. Logsdon et al., 2024 [<https://doi.org/10.1038/s41586-024-07278-3>], in particular, is overall a very similar article, but based on the comparison of only two full genomes and a few chromosomes in other species. The work presented here provides novel conclusions and, to date, the strongest evidence that support the notions that were proposed -but not fully demonstrated- before. The results presented are of interest for a broad audience, as they highlight the high variability of the human centromeres sequences, and more specifically, for the audience in the centromere field, they are of immense relevance.

Data & methodology:

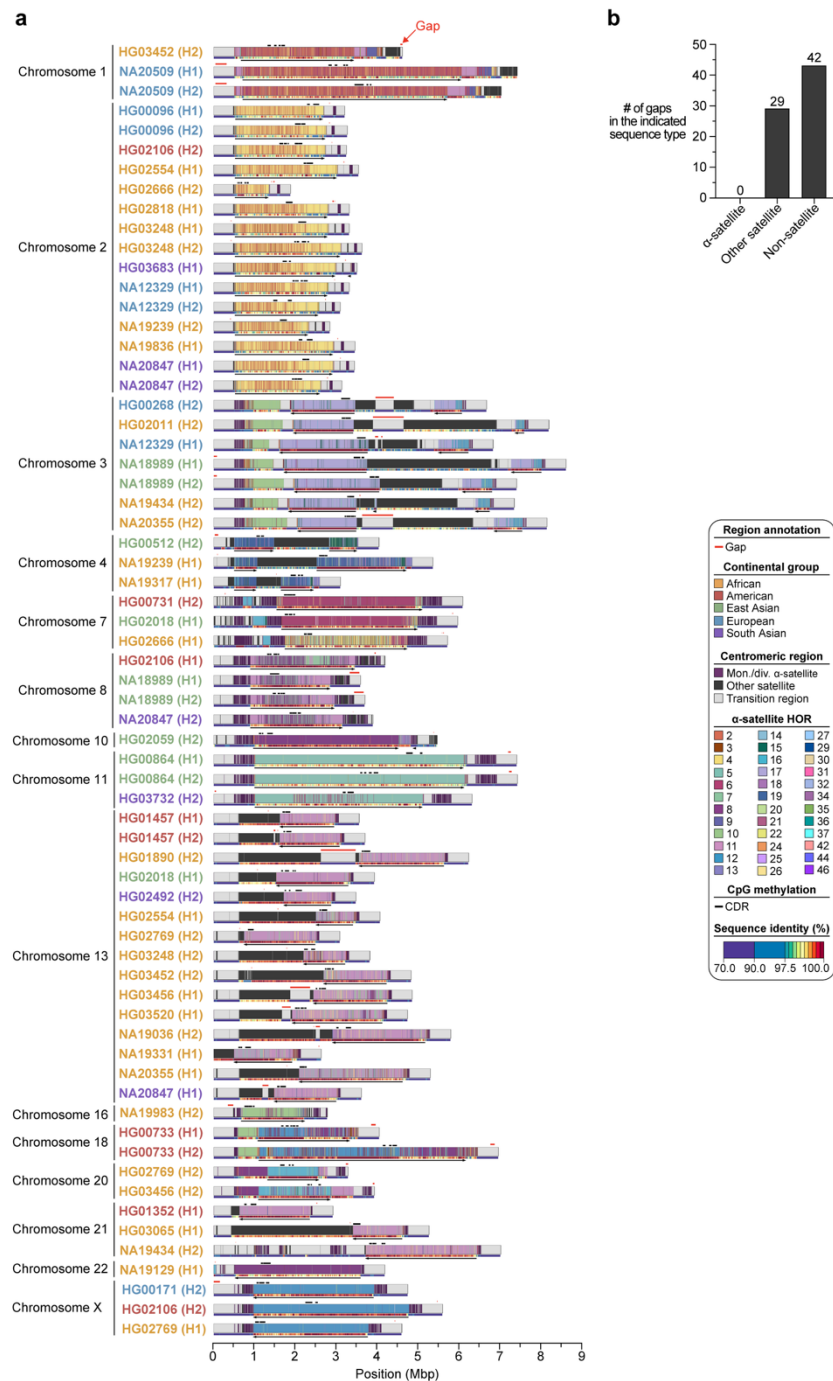
The approach of this work is conceptually simple: to sequence and compare as many centromeres possible from a diversity of individuals. From a technical and methodological point of view the work is extensive and complex. The resulting data is of high quality and it has been presented in a manner that is extremely clear and easy to follow.

[We thank you for noting the novelty and strength of our findings and their potential for broad appeal. We also thank you for your kind words on our data analyses and presentation.](#)

Data and code are available and accessible. The 2110 centromere fasta files from <https://figshare.com/s/cb839beea6ed06d641db>, which is supposed to be complete, include 66 centromere fastas with gap. Please double check if they are the final version on this link. Please note that this reviewer is

not an expert in computational biology or bioinformatics, so I refrain to judge if the methodology is sufficiently detailed to allow for the results to be fully reproduced.

Thank you for pointing out that a subset of the FASTAs that we uploaded to figshare have a gap in sequence. We note that these gaps are not within the centromeric α -satellite HOR arrays themselves but are instead within other satellites or tandem repeats in the pericentromere or between α -satellite HOR arrays. Since these gaps reside outside of the centromeric α -satellite HOR array, they don't affect our analyses nor conclusions. Below, we show the location, sequence composition, and frequency of these gaps (**Reviewer Fig. 7**). Additionally, we updated **Supplementary Note 1** to indicate that gapped regions exist within the pericentromere and between centromeric α -satellite HOR arrays but not within the α -satellite HOR arrays themselves.



Reviewer Figure 7. Location, sequence composition, and quantification of gaps in sequence in 66 centromeric regions. a) Location of gaps in sequence, defined by a string of Ns, in 66 centromeric regions.

The gapped regions are shown as a red bar above the α -satellite HOR array, and all gaps exist outside of the centromeric α -satellite HOR array. **b)** Quantification of gapped regions within the 66 centromeric regions, showing that 29 are within other satellite (e.g., HSats), and 42 are in other non-satellite sequences. We note that some centromeres have more than one gap, but none are within α -satellite.

Supplementary Note 1 (changes in **bold**):

“To confirm the accurate assembly of each human centromere, we assessed each of them with a suite of computational tools and approaches. First, we aligned native long-read Pacific Biosciences high-fidelity (PacBio HiFi) and Oxford Nanopore Technologies (ONT) sequencing data generated from each genome to the corresponding genome assembly and assessed for uniform read depth across each centromere. We found that all **α -satellite HOR arrays** had uniform coverage of both PacBio HiFi and ONT data, suggesting that they were free of large structural errors (Supplementary Table 2). Then, we applied HMM-Flagger¹, a tool that uses a Hidden Markov Model (HMM) to detect anomalies in coverage based on PacBio HiFi alignments, and found that nearly all α -satellite sequences within the centromeres were free of assembly errors (Supplementary Table 2). Next, we ran NucFlag², a tool that assesses for assembly errors based on first and second most common bases in aligned PacBio HiFi or ONT reads, and found that 100% of **α -satellite sequences** were free of such errors (Supplementary Table 2). Then, we applied GAVISUNK³, a tool that compares singly unique nucleotide *k*-mers (SUNKs) within the assembly to those found in native ONT reads, and found that 98% of α -satellite sequences were supported with raw ONT data, with the remaining 2% due to validation gaps³ (Supplementary Table 2). **A subset of centromeres ($n = 66$; 3.1%) have gaps in regions outside of the centromeric α -satellite HOR array (e.g., the pericentromere or in-between α -satellite HOR arrays) but not within the α -satellite HOR array itself.** Finally, we estimated the quality value (QV) of the centromeres with Merqury⁴ using short-read Illumina sequencing data generated from the same source genome and found them to have an average QV of 60.99, indicating that the centromeres were >99.9999% accurate on average (Supplementary Table 3).”

Conclusions:

The data presented and its interpretation is robust, valid and reliable.

References:

The manuscript references the previous literature appropriately.

Clarity and context:

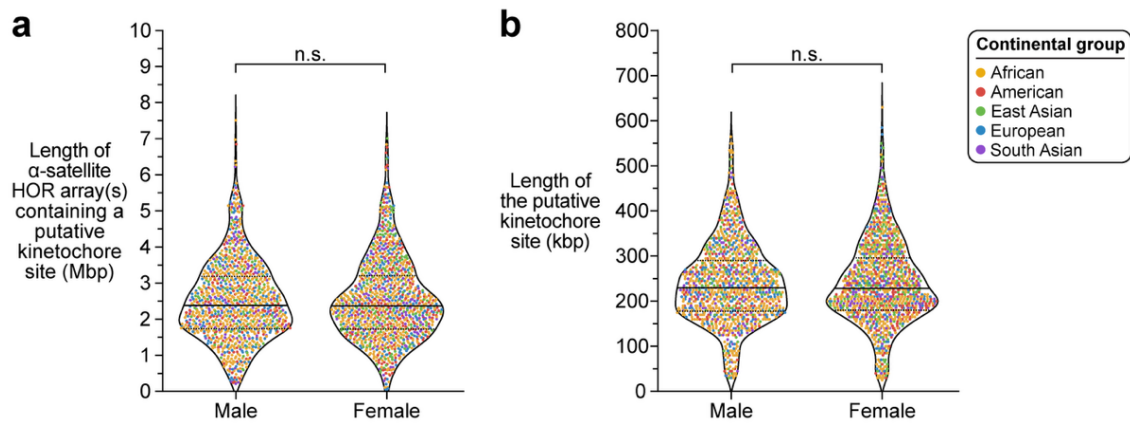
The abstract is very clear and accessible to a broad audience. Even though the article presents some concepts that are not immediately accessible to everyone, it is written in a way that guides the reader step by step and is very clear and appropriate.

Thank you once again for your kind and thoughtful comments on our manuscript.

Suggested improvements:

- Centromere genome and epigenome could be influenced by some confounders, e.g. sex, age, and may result in disease status. Do you consider the confounders influence or try to find some associations?

To address this comment, we analyzed our set of 2,110 centromeres to see if there is any significant association between centromeric α -satellite HOR array length and/or putative kinetochore length and sex. Comparison of these features in both males and females revealed that there is no significant difference, and comparison of their α -satellite HOR array organizations and epigenetic landscapes also reveals no correlation (**Reviewer Fig. 8**).



Reviewer Figure 8. No correlation between centromeric α -satellite HOR array or putative kinetochore length and sex. **a)** Variation in the length of α -satellite HOR arrays containing a putative kinetochore site between males and females. **b)** Variation in the length of putative kinetochore sites between males and females. There is no significant difference between males and females, as determined with a two-tailed Wilcoxon rank-sum test. n.s., not significant.

We, unfortunately, couldn't test if any centromere features are correlated with age in this study, as metadata was not made available for the 1000 Genomes Project (to our knowledge). However, we hypothesize that there may be small differences in centromere length, composition, and organization in rapidly dividing tissues that could accumulate as one ages, leading to increased variability among centromeres. We plan to test this hypothesis in future studies with datasets that are currently being generated, including those with longitudinal data from multi-generational CEPH pedigrees as well as data being generated from individuals of varying ages by the Somatic Mosaicism across Human Tissues (SMAHT) Network.

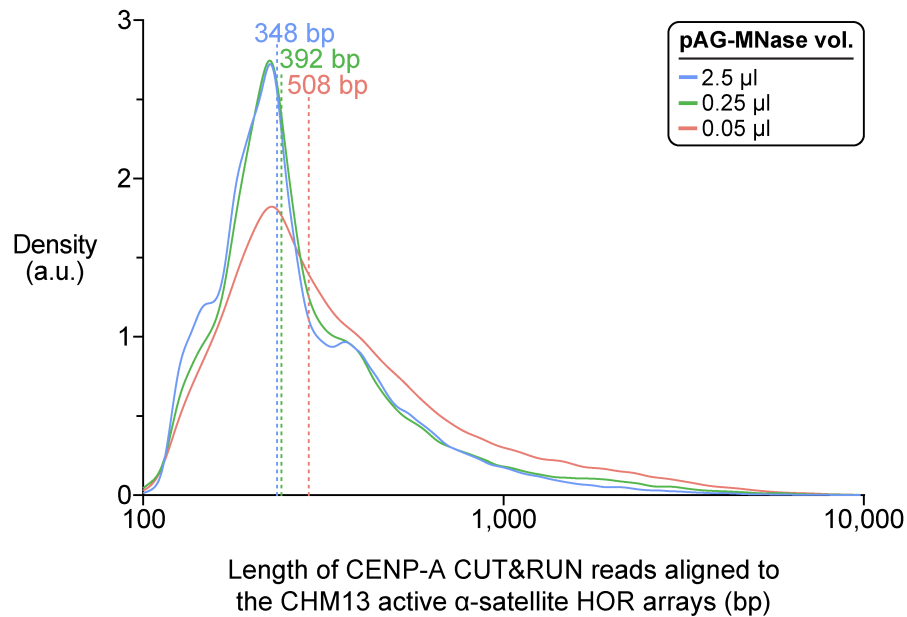
Finally, we hypothesize that there may be significant differences in centromere length, composition, organization, and/or epigenetic landscape in individuals with cancer, infertility, or aneuploidy disorders. However, because this study focuses on only healthy individuals, this is beyond the scope of the current study. With that said, there is a contemporaneous study by our group where we report the finding that a subset of individuals with Down Syndrome (or Trisomy 21) have significantly different chromosome 21 α -satellite HOR array lengths. We hypothesize that this centromere size asymmetry leads to improper pairing and synapse formation during meiosis I, leading to increased rates of non-disjunction. We hope to expand this study to hundreds of more individuals with Down Syndrome to see if there's a significant association between centromere size asymmetry and chromosome 21 non-disjunction risk in the future.

- The description of the CUT&RUN method adapted to long-read is unclear as chromatin fragments enriched with CENP-A seems to be digested as normal CUT&RUN. How does it work for long-reads?

The long-read CUT&RUN assay is very similar to the typical CUT&RUN assay, except there's no size selection step during library prep that would normally select for shorter (100-200 bp) DNA fragments. Instead, all fragments are kept during library prep and sequenced with ONT. While a subset (<30%) of the reads are mononucleosome length (100-200 bp), the majority of reads are actually polynucleosome length (300+ bp), which improves alignment accuracy and mapping quality in repetitive regions.

Although this method doesn't require any modifications to the typical CUT&RUN assay, we tested if we could obtain longer read lengths by reducing the amount of pAG-MNase in the reaction. Therefore, we performed the experiment with three different volumes of pAG-MNase: 2.5 μ L (the original volume), 0.25 μ L (a 10x reduction), and 0.05 μ L (a 50x reduction), keeping the rest of the reaction conditions the same. However, upon ONT sequencing, we observed only a modest increase in read length upon reduction in pAG-MNase volume, with an average read length of 348 bp, 392 bp, and 508 bp for the 2.5 μ L, 0.25 μ L, and 0.05 μ L pAG-MNase volumes tested, respectively (**Reviewer Fig. 9**). While the average read length increased by up to 1.5-fold, the yield was drastically reduced by ~5-fold. Therefore, we decided to perform the CUT&RUN experiment in this

study with the original pAG-MNase volume (2.5 μ L) in order to obtain long CUT&RUN reads while reducing costs.



Reviewer Figure 9. A reduction in pAG-MNase volume increases read lengths but reduces yield in a long-read CENP-A CUT&RUN assay with CHM13 cells. Means are indicated with a dashed line.

We revised the Methods section describing the long-read CUT&RUN assay to note that no size selection step was performed during library prep and that reductions in the amount of pAG-MNase can increase read lengths but reduce yield.

Methods section (lines 819-857; changes in **bold**):

“CENP-A CUT&RUN and ONT long-read sequencing

CUT&RUN was performed following the CUTANA protocol developed by EpiCypher with some modifications. For each reaction, 550,000 cells were harvested and nuclei were isolated following the nuclei extraction protocol for CUTANA assays (EpiCypher). Nuclei were immobilized with activated Concanavalin A (ConA) beads provided in the CUTANA ChIC/CUT&RUN Kit (14-1048, EpiCypher). After binding to ConA beads, nuclei were incubated overnight at 4°C in 50 μ L of Antibody buffer (1X Proteinase Inhibitor, 0.5 mM Spermidine, 0.01% w/v Digitonin, and 2 mM EDTA in Pre-wash buffer) with either mouse anti-CENP-A antibody (ADI-KAM-CC006, Enzo) or a rabbit anti-IgG antibody (ab171870, Abcam) at a 1:50 dilution. An appropriate amount of K-MetSat Panel was added to the reaction before adding the IgG negative control. Following overnight incubation at 4°C, rotating on a nutator at 15 rpm, samples were washed and incubated in Cell Permeabilization Buffer (1X Proteinase Inhibitor, 0.5 mM Spermidine, and 0.01% w/v Digitonin in Pre-wash buffer) with **2.5 μ L pAG-MNase** (15-1016, EpiCypher) for 10 mins at RT. **In our experience, a reduction in pAG-MNase (e.g., to 0.25 μ L or 0.05 μ L) can increase fragment lengths (up to 1.5-fold) but often also reduces sequencing yield (to ~20%). Therefore, we chose to use the original volume of pAG-MNase in our reactions to obtain long reads while reducing sequencing costs.** After pAG-MNase was bound to the primary antibody, 100 mM CaCl_2 (EpiCypher) was added to activate chromatin-targeted digestion at 4°C for 2 hours. Digestion was stopped by adding Stop Buffer (EpiCypher), and samples were transferred to low-bind tubes for DNA purification and extraction following the fragment release and cleanup in the protocol of Hainer & Fazio⁵⁶. Each sample was brought up to 300 μ L with 0.1X TE (EpiCypher), followed by addition of 3 μ L of 10% SDS and 2.5 μ L of 20 mg/ml Proteinase K (P8107S, NEB), then incubated at 70°C for 10 minutes.

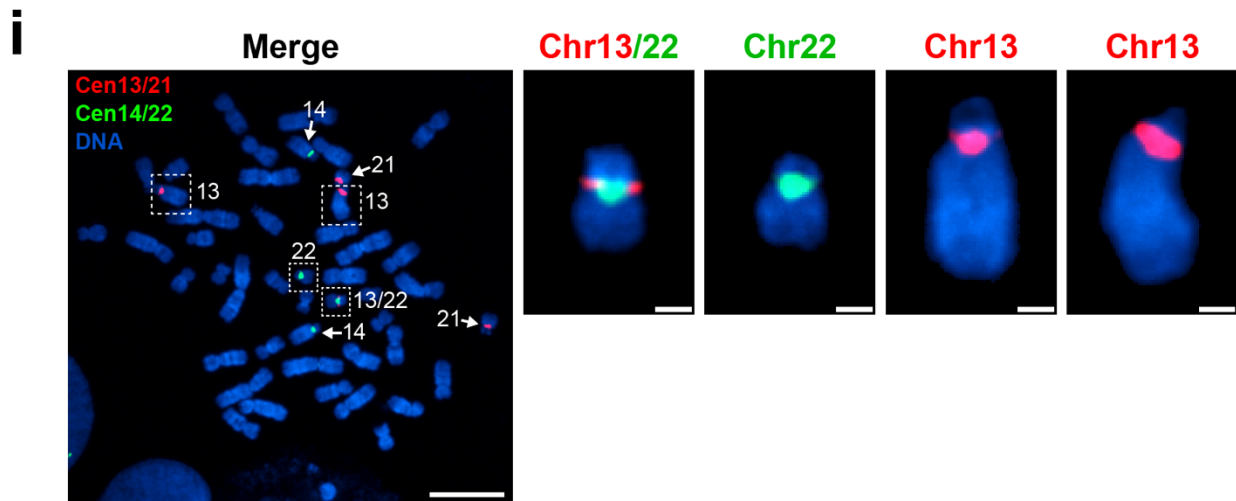
DNA extraction was performed using one phenol:chloroform:isoamyl alcohol (25:24:1 v/v; 15593031, Invitrogen) extraction step, followed by transfer of the aqueous phase to a MaXtract High Density 2 ml tube (129056, QIAGEN), then one chloroform (C2432, SIGMA) extraction step, again transfer the aqueous phase to

a MaXtract tube. DNA was precipitated with 750 μ l of 100% ethanol and 2 μ l of Glycogen (AM9510, AMB) and stored overnight at -20 °C. CUT&RUN-enriched DNA was washed with 100% ethanol and eluted in 0.1X TE.

For ONT library preparation and sequencing, 0.5-5 ng of CUT&RUN-enriched DNA was amplified using CUTANA CUT&RUN Library Prep Kit (14-1001 & 14-1002, EpiCypher). Up to the indexing PCR step, the KOD Xtreme Hot Start DNA polymerase (71975-3, SIGMA) was used for fragment enrichment. Each 50 μ l PCR reaction contains 2X KOD PCR buffer, 0.4 mM dNTPs, 1.5 μ l of i7 primer, 1.5 μ l of i5 primer, and 1 μ l of KOD polymerase. PCR cycling conditions were: 2-min initial denaturation at 94°C; 20 cycles of 10-second denaturation at 98°C, 30-s primer annealing at 55 °C, and 5-min extension at 68°C; followed by a final 5-min extension at 68°C. **We note that we did not perform a size selection step during library preparation in order to preserve long DNA fragments.** The sequencing library was prepared using the Oxford Nanopore Technologies (ONT) SQK-NBD114.24 kit as recommended by Hook et al., unpublished, and sequenced on a PromethION 24 using a R10.4.1 flow cell (FLO-PRO114M, ONT)."

- The FISH analysis presented in figure 2h is not very clear and does not really support fusion of chromosome 13 and 22. In the provided example there is no clear cen13 staining. The FISH signal is also too saturated.

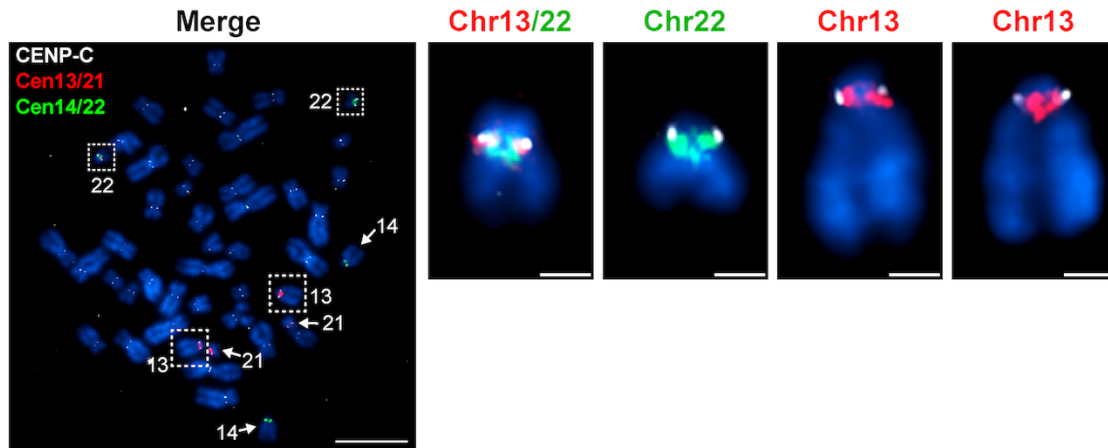
To improve the quality of this image, we initiated a collaboration with Dr. Jennifer Gerton (Stowers) and her research scientist, Dr. Tamara Potapova, who are experts in performing FISH on metaphase chromosome spreads. They performed this experiment on the GM20355 cell line described in our manuscript, using fluorescent DNA probes specific to the chromosome 13/21 α -satellite DNA and chromosome 14/22 α -satellite DNA. Their images, collected with a spinning disk confocal microscope, more clearly show a hybrid chromosome with both chromosome 13/21 α -satellite (red) and 14/22 α -satellite (green), confirming that this chromosome has α -satellite from two chromosomes at the primary constriction (see the updated Fig. 2i).



Updated Fig. 2i. Fluorescent *in situ* hybridization (FISH) on an GM20355 metaphase chromosome spread showing the recombined chromosome 13/22. Fluorescent probes specific to chromosome 13/21 α -satellite (red) and 14/22 α -satellite (green) colocalize on the recombined chromosome 13/22 in 100% of cells ($n=19$). Scale bar (chromosome spread), 10 μ m; scale bar (individual chromosome), 1 μ m.

To test if the chromosome 13 α -satellite HOR array is the site of kinetochore assembly (as predicted by our CDR analysis), Dr. Potapova performed immunoFISH on the same cell line, using the same fluorescent DNA probes and an antibody to CENP-C, an inner-kinetochore protein. Her images show that CENP-C colocalizes specifically with the chromosome 13 α -satellite DNA on the hybrid chromosome 13/22, indicating that chromosome 13 α -satellite is the site of kinetochore assembly and confirming our original findings. We have included a representative image from this experiment in **Supplementary Fig. 23f** (shown below).

f



New Supplementary Fig. 23f. Immunolabeling and fluorescent *in situ* hybridization (immunoFISH) on an GM20355 metaphase chromosome spread, showing colocalization of CENP-C with chromosome 13/21 α -satellite (red), which is adjacent to 14/22 α -satellite (green) on the recombined chromosome 13/22 ($n=30$). Scale bar (chromosome spread), 10 μ m; scale bar (individual chromosome), 1 μ m.

We have updated the Methods (lines 903-946) to describe these new experiments and have moved the original FISH image to **Supplementary Fig. 23e**.

• Line 238: The authors should cite previous works demonstrating that CENP-A is a key epigenetic mark of centromere identity. The cited computational works did not prove that.

Thanks for this really great suggestion. We now cite foundational work demonstrating that CENP-A is the epigenetic mark that confers centromere identity in the following sentence:

Main text (lines 256-258; changes in **bold**):

“Functional centromeres are epigenetically defined by two key features: nucleosomes containing the histone H3 variant CENP-A^{28–30} and underlying domains containing hypomethylated DNA, known as CDRs^{8,22}.

Citations:

28. Palmer, D. K., O'Day, K., Wener, M. H., Andrews, B. S. & Margolis, R. L. A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. *J Cell Biol* **104**, 805–815 (1987).
29. Sullivan, K. F., Hechenberger, M. & Masri, K. Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *J Cell Biol* **127**, 581–592 (1994).
30. Warburton, P. E. *et al.* Immunolocalization of CENP-A suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres. *Current Biology* **7**, 901–904 (1997).

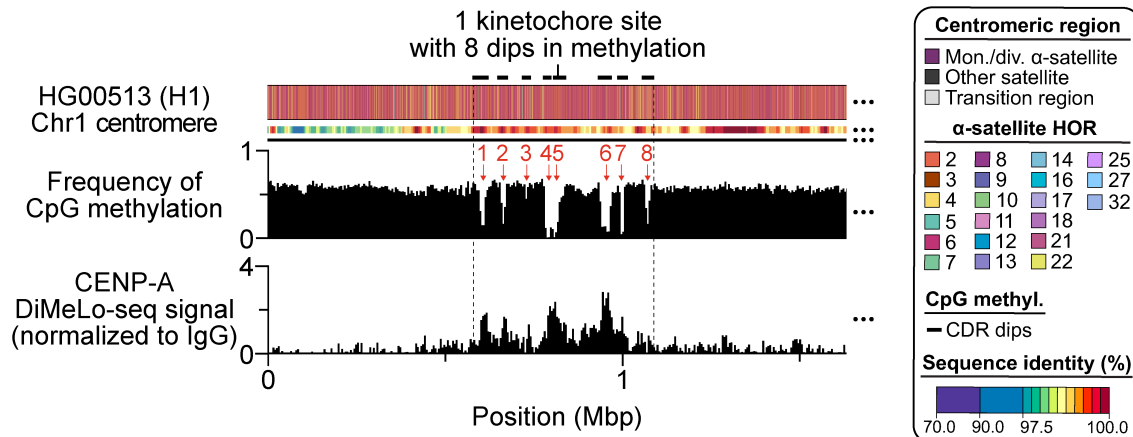
• Lines 269 – 270 / Figure 3 d-e: Please define what specific criteria used to define a methylation dip as such. Is it how low the methylation frequency goes compared to the surrounding? What is the minimum length that a dip has to have? For example, Figure 3e states that there are 8 dipoles. Only zooming in heavily on the image can one see what I assume is the 3rd dip, which is very narrow and not so deep as the others, and a 5th dip, which is wide but very shallow?

In this study, we used CDR-Finder (Mastrorosa *et al.*, *Bioinformatics*, 2024) to automatically detect dips in methylation based on aligned long-read sequencing data. CDR-Finder first bins CpG methylation signals into 5-kbp bins and then calculates the mean methylation frequency in each bin. Then, it subtracts the mean methylation frequency from 1 and applies a peak detection algorithm to detect candidate peaks, requiring that the peak be at least 34% of the mean CpG methylation frequency of the flanking 1-Mbp of sequence (500 kbp upstream and downstream, on average) and have at least 30% topographic prominence (e.g., a 30% drop in

the surrounding methylation frequency). The calls are then manually curated to ensure accuracy. As a result, CDR-Finder requires the following to detect a dip in methylation:

- At least 5 kbp of sequence (this is based on the 5 kbp bins, but it can be adjusted by the user)
- At least a 34% reduction in CpG methylation frequency compared to the flanking 1 Mbp of sequence, on average
- At least a 30% difference in CpG methylation frequency compared to the next highest methylation frequency upstream and downstream

Based on these strategies, we identified 8 dips in methylation in the centromere in Fig. 3, which we've illustrated with the top track and number below for convenience (**Reviewer Fig. 10**):



Reviewer Fig. 10. Annotation of eight dips in methylation in the kinetochore site of the HG00513 (H1) chromosome 1 centromere. Methylation dips are automatically detected with CDR-Finder and manually checked and curated to ensure accuracy.

This centromere in particular has 7 very strong dips in methylation (#s 1-2, 4-8) and 1 weak dip in methylation (#3). However, the weak dip (#3) is still 25 kbp in length and meets the other requirements as well. The shallow dip between #5 and #6 is not considered to be a dip in methylation because it doesn't meet the 34% drop in frequency and 30% prominence criteria. The dip containing both #4 and #5 dips are considered two separate dips due to the increase in methylation frequency that separates the two. We hope this explanation is helpful in understanding how these dips were called.

Minor comments

- Figure 3 l and m: the description of the plots is missing from the figure legend.

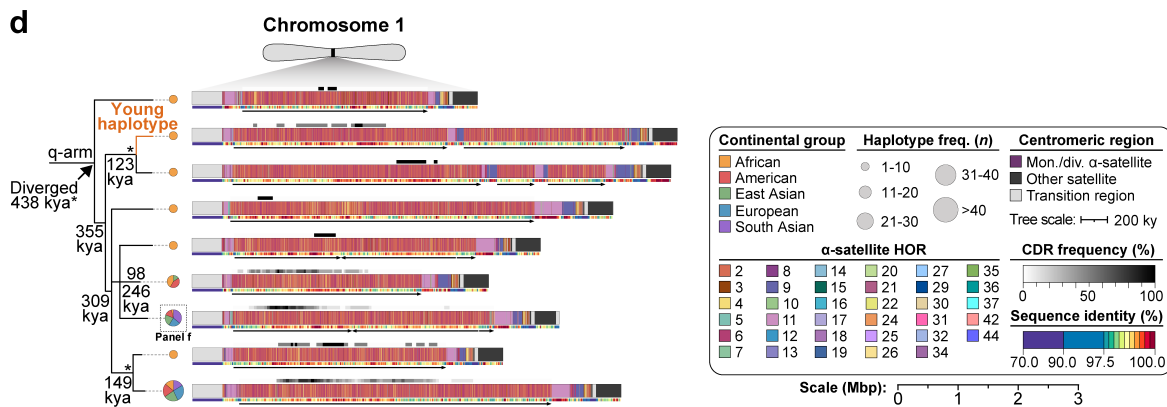
Thanks for pointing this out. We have now updated the legend for Fig. 3l,m to include the description of these plots.

Figure 3 legend (changes in **bold**):

"l,m) Comparison of the mean l) sequence identity and m) entropy between sequences comprising the active α-satellite HOR array vs. those within the CDR (**$n = 2,156$, mean sequence identity = 99.36% vs. 99.55%, mean entropy = 0.80 vs. 0.74, for the active α-satellite HOR array vs. CDR, respectively**). The mean is shown as a solid line, and first and third quartiles are shown as dotted lines. CDR-harboring sequences have higher sequence identity and lower entropy. ****, $p < 0.0001$, **two-tailed Mann-Whitney test**."

- Figure 4d: there is an error in the labelling of the last node presented.

Thanks for drawing our attention to this. We have now updated Fig. 4d to match the phylogenetic tree in Supplementary Fig. 32, as shown below.



Updated Fig. 4d. Maximum-likelihood phylogenetic tree depicting the q-arm topology for chromosome 1 centromere haplogroups along with estimated divergence times and haplotype frequencies. A young haplotype with a large, multi-megabase duplication of the α -satellite HOR array is observed and is estimated to have emerged in the last 10,000 years (or 500 generations).

• Figure 5a: error bars are not defined.

We have now updated the legend for Fig. 5a as follows

Figure 5a legend (changes in **bold**):

“a) Estimated *de novo* mutation rates for different sequences within centromeric regions across four generations of a pedigree (CEPH 1463). The active α -satellite HORs and CDR are mutating 3.4- and 5.9-fold faster than the unique sequences in the pericentromere, respectively. **Error bars, 95% confidence interval.**”

• Line 168, “D1Z7” means the primary alphaSat family in chr1. If the meaning of the numbering scheme is explained clearly at the beginning, there is no need to repeatedly restate these alphaSat family IDs and chromosome names later on.

Thanks for this suggestion. We have now added a **new Supplementary Note 3** to explain this nomenclature, and we reference it on the first usage of the nomenclature in the main text.

New Supplementary Note 3:

“Human α -satellite arrays have an established nomenclature first defined by the Human Genome Project⁵, which follows the convention “DCZA”, where D is the DNA segment, C is the chromosomal assignment, Z is the complexity of DNA (Z for repetitive), and A is the array number. For example, the centromeric α -satellite HOR array on chromosome 1 is called “D1Z7”, indicating that it is a DNA segment on chromosome 1 that is repetitive and the seventh array.”

Main text (lines 182-184; changes in **bold**):

“This is perhaps best exemplified by chromosome 1, where we found that a quarter of the individuals in our dataset have a large-scale inversion in the core of the D1Z7 α -satellite HOR array (**Supplementary Note 3**), spanning 637 kbp to 2.47 Mbp (like CHM13⁸ and CHM1⁹).”

• Line 278 “MEI”, even the full name shown up in the Figure 1 legend, it is better to have the explanation in the full context.

Thanks for suggesting this. We now provide the definition of an “MEI” at the first occurrence in the main text.

Main text (lines 297-300; changes in **bold**):

“While **mobile element insertions (MEIs)** within kinetochore sites are rare, they occur frequently within the chromosome 16 centromere, with 14% (16/112) of centromeres harboring an MEI within the putative kinetochore site, which is 17 times more frequent than in the other chromosomes (Supplementary Fig. 28a-d, Supplementary Table 10).”

Response to Referees' comments

Referees #1/2 were unavailable this time around.

Referee #3 (Remarks to the Author):

The authors have satisfactorily addressed my comments, and to the extent that I am able to judge, those of the other reviewers. I particularly commend them on building an interactive tool that enables non-specialist readers to interact directly with complex centromere sequence/structure data.

We thank you for your kind comments about our manuscript and your appreciation for the development of our new interactive centromere browser, Cencyclopedia, which we hope will enable readers to study centromeres in greater detail.

Referee #4 (Remarks to the Author):

In this revised manuscript, Gao et al., have addressed the remarks and answered all questions made by all reviewers, improving several aspects of an already remarkable manuscript. In particular, the creation of cencyclopedia as a way of interactively visualizing the large and complex data here generated is of tremendous value to the field.

We would like to suggest two minor additions that could further improve the quality of the text and figures, and are happy to recommend the publication of this work in Nature.

Thank you for your thoughtful and helpful comments about our manuscript and Cencyclopedia.

- Reviewer 1, point 2a) It would be great if the authors could also incorporate an equivalent scheme as the one presented in Reviewer Figure 1, clarifying in a visual way the difference between CDR and CDR dips in the main Figure 5c and 5d for all the readers.

We have now incorporated **Reviewer Figure 1** as a **new Supplemental Figure 26** to allow readers to visually distinguish between “CDRs” and “CDR dips”.

- Reviewer 1, point 2c. I agree with this comment from Reviewer 1, and perhaps the authors' change to the main text could be further improved to erase all possible doubt in the following way:
“(…) identifying the location of the CDR, a region which spans from the first hypomethylated base to the last hypomethylated base and can contain one or multiple methylation dips, in each centromere.”

We thank you for this suggestion and have now revised the main text as follows (lines 420-423; **changes in bold**):

“To do this, we first determined the methylation profile of each centromere by aligning native ONT long-read sequencing data from the same source genome to each assembly and **identifying the location of the CDR, which spans from the first hypomethylated base to the last hypomethylated base and can contain one or multiple methylation dips, in each centromere.**”