Bioinformatics, 2024, **40(12)**, btae733
https://doi.org/10.1093/bioinformatics/btae733
Advance Access Publication Date: 9 December 2024
Original Paper

Genome analysis

Identification and annotation of centromeric hypomethylated regions with CDR-Finder

Francesco Kumara Mastrorosa [1], Keisuke K. Oshima [2], Allison N. Rozanski [1], William T. Harvey [1], Evan E. Eichler [1,3], Glennis A. Logsdon [1,4,*]

[1] Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, United States
[2] Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States
[3] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, United States
[4] Present address: Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States

*Corresponding author. Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Clinical Research Building 500, 415 Curie Blvd, Philadelphia, PA 19104, United States. E-mail: glogsdon@pennmedicine.upenn.edu.

Associate Editor: Peter Robinson

Abstract

Motivation: Centromeres are chromosomal regions historically understudied with sequencing technologies due to their repetitive nature and short-read mapping limitations. However, recent improvements in long-read sequencing allow for the investigation of complex regions of the genome at the sequence and epigenetic levels.

Results: Here, we present Centromere Dip Region (CDR)-Finder: a tool to identify regions of hypomethylation within the centromeres of high-quality, contiguous genome assemblies. These regions are typically associated with a unique type of chromatin containing the histone H3 variant CENP-A, which marks the location of the kinetochore. CDR-Finder identifies the CDRs in large and short centromeres and generates a BED file indicating the location of the CDRs within the centromere. It also outputs a plot for visualization, validation, and downstream analysis.

Availability and implementation: CDR-Finder is available at https://github.com/EichlerLab/CDR-Finder.

1 Introduction

Centromeres are chromosomal regions essential for the segregation of sister chromatids during cell division. Centromeres are typically composed of near-identical tandem repeats known as α-satellite, with other types of satellites (e.g. β-satellite, ɤ-satellite, HSat1A, HSat1B, HSat2, and HSat3) found within pericentromeric regions (Miga and Alexandrov 2021, Logsdon and Eichler 2022). Within the centromere, α-satellite repeats are organized into higher-order repeat (HOR) arrays, which vary in composition and length and constitute the functional unit of the centromeres.

Recently, long-read sequencing technologies such as Pacific Biosciences (PacBio) high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) long-read sequencing, as well as newly developed genome assembly algorithms (Cheng *et al.* 2021, Rautiainen *et al.* 2023), have led to the reconstruction of the most complex regions of the human genome, including centromeres (Miga *et al.* 2020, Logsdon *et al.* 2021, Altemose *et al.* 2022, Logsdon *et al.* 2024b), telomeres (Nurk *et al.* 2022), segmental duplications (Vollger *et al.* 2022), and other repetitive regions (Nurk *et al.* 2022, Rhie *et al.* 2023). These (Miga *et al.* 2020, Logsdon *et al.* 2021, Altemose *et al.* 2022, Nurk *et al.* 2022, Logsdon *et al.* 2024b) and other (Gershman *et al.* 2022) studies have also revealed the genetic and epigenetic features of human centromeres. For instance, centromeres are typically hypermethylated throughout the α-satellite HOR array except for a small hypomethylated region called the centromere dip region (CDR) (Gershman *et al.* 2022). The CDR was shown to coincide with a unique type of chromatin containing the centromeric histone H3 variant CENP-A (Logsdon *et al.* 2021, Altemose *et al.* 2022), which marks the site of the kinetochore. This finding was confirmed with several experimental assays (Miga *et al.* 2020, Logsdon *et al.* 2021, Altemose *et al.* 2022).

Here, we present CDR-Finder, a tool to identify and annotate the CDRs based on the CpG methylation data obtained from either PacBio HiFi or ONT data. CDR-Finder detects regions of hypomethylation within sequence-resolved centromeric α-satellite HOR arrays, determines their coordinates and size, and outputs a plot showing the sequence composition, mean CpG methylation frequency, and fold coverage of the total and methylated sequencing data per region analyzed. CDR-Finder can be used to assess both large and small centromeres, as long as the centromere is fully assembled and sequence-resolved.

Received: 7 June 2024; **Revised:** 26 November 2024; **Editorial Decision:** 30 November 2024; **Accepted:** 6 December 2024

2 Materials and methods

CDR-Finder requires three input files: a FASTA file of the genome assembly, a BED file containing the coordinates of the region of interest (e.g. the centromere) within the assembly, and a BAM file containing alignments of PacBio HiFi or ONT data with methylation tags to the same genome assembly. CDR-Finder first converts the BAM file to a bedMethyl file using modkit (https://github.com/nanoporetech/modkit), which lists the position and frequency of modified bases for each CpG. Then, it divides the region of interest within the methylBED file into sequential 5-kbp bins and calculates the average CpG methylation frequency for each bin. Bins without an assigned value are excluded. Finally, CDR-Finder runs RepeatMasker (Smit *et al.* 2013–2015) on the region of interest to identify the location of the α-satellite sequences (annotated as “ALR/Alpha”), and it calculates the mean CpG methylation frequency across each bin containing α-satellite.

To identify the CDR(s), our tool first selects bins with a CpG methylation frequency less than the median frequency of all α-satellite sequences in the region of interest. While this frequency can be specified by the user, in our experience, a frequency of 0.34 identifies most CDRs with high precision and recall (described in the example below). Frequencies >0.34 often fail to detect CDRs with shallow hypomethylation, and frequencies <0.34 often miscall CDRs due to variation in sequencing coverage. Then, CDR-Finder further refines the bins with a low CpG methylation frequency to those that also have a minimal dip prominence [defined topographically (Kirmse and de Ferranti 2017)] from the median. In our experience, a minimal dip prominence of 0.30 often removes low-confidence calls when the methylation levels are uniformly low across a subregion. Finally, CDR-Finder evaluates the boundaries of each candidate CDR by calculating the mean CpG methylation frequency and then extending each call boundary to the mean CpG methylation frequency ± a specific value. In our experience, the mean minus one standard deviation is usually sufficient to capture the entire CDR in each call.

3 Results

CDR-Finder is organized as a configurable Snakemake pipeline. To run it, the user should first clone the repository from https://github.com/EichlerLab/CDR-Finder. Then, the user should modify the configuration file, config.yaml, to specify the sample name, the path to sample’s genome assembly, the path to the BED file containing a region(s) of interest, and the path to the BAM file containing the alignment of PacBio HiFi or ONT data with methylation tag to the sample’s genome assembly.

In most cases, default parameters can be maintained, but the sequencing coverage and methylation-calling algorithm will affect the ability of CDR-Finder to detect CDRs accurately. As such, the parameters may need to be adjusted accordingly. While CDR-Finder can run on any sequence containing α-satellite DNA, we recommend that the user run it on an α-satellite HOR array with additional flanking sequence on both sides to ensure that the centromere is completely traversed and to observe the transition in methylation patterns between the centromeric and pericentromeric regions. Since CDR detection is based on α-satellite sequences only, the additional flanking sequence will not affect the results of the analysis. We also recommend that the user verify the accuracy of each call based on the coverage data in the plot generated by CDR-Finder and in the original methyl-reads alignment. The user should exclude calls in regions with low coverage and other potential false positives.

The pipeline can be run using conda or singularity with the following commands: `snakemake -np—sdm conda -c 4` or `snakemake -np—sdm apptainer conda -c 4`, respectively. It will output a BED file with the coordinates of each CDR call as well as a plot showing the CpG methylation frequency of the region of interest, the overall sequencing coverage and methylated sequencing coverage, and annotation showing the location of the CDR call and the sequence composition of the region.

Figure 1 reports an example of a CDR call for the T2T-CHM13 chromosome 8 centromere (Logsdon *et al.* 2021) characterized by CDR-Finder (Fig. 1). We used the original chromosome 8 centromere α-satellite HOR array coordinates with ∼500 kbp additional sequence on both sides as a target region [chr8:43 746 447–47 020 471 in the T2T-CHM13 v2.0 genome (Nurk *et al.* 2022)] as well as the original ONT data generated from the same genome. CDR-Finder called only one CDR, as expected (chr8:45 789 626–45 899 626, size = 110 kbp). The same T2T-CHM13 chromosome 8 CDR was originally described with a similar size (73 kbp) (Logsdon *et al.* 2021). However, on that occasion, CpG methylation was detected with a different method, and the

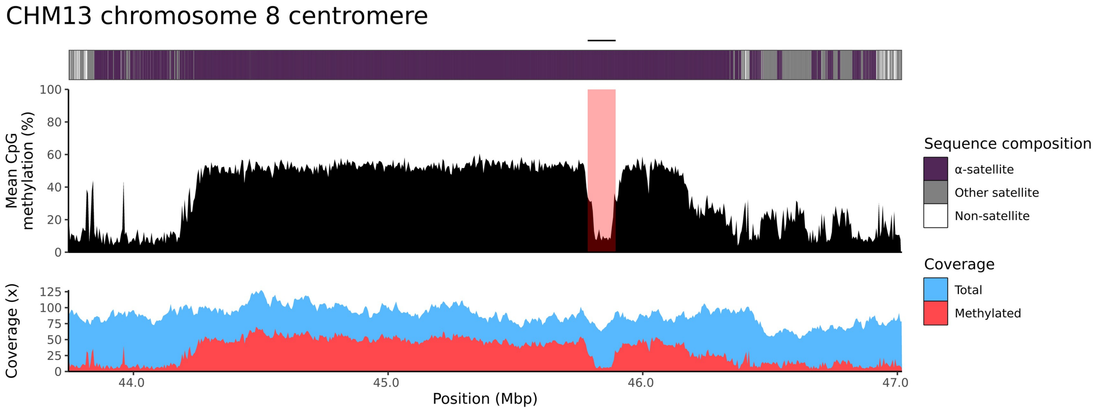

Figure 1. Detection of a CDR in the T2T-CHM13 chromosome 8 centromere with CDR-Finder. CDR-Finder generates a plot with the annotation of the region (top), mean CpG methylation frequency (middle), and the corresponding read coverage (both total and methylated reads) across the region (bottom). The CDR is highlighted in the hypomethylated region and indicated with a bar on top.

CDR was considered as the region with the lowest CpG methylation levels.

To evaluate CDR-Finder's ability to detect CDRs in diverse human centromeres, we ran the tool on 200 completely and accurately assembled centromeres from 15 randomly selected samples from the Human Genome Structural Variation Consortium (Logsdon *et al.* 2024a) with default parameters. We manually inspected all calls, counting the number of CDRs that were correctly called, partially called (where the CDR could be extended on one or both sides), incorrectly called, or not called. Our test showed that 98.7% (443/449) of the CDRs were correctly called, indicating a precision of 0.99, and only 43 CDRs were not called (8.7%), indicating a recall of 0.91. In 28 cases (6.3%), the CDR(s) were only partially called, which could be corrected by tweaking the parameters for the specific case. The six erroneous calls could be easily excluded with a visual inspection of the CDR-Finder plots (Supplementary Table S1).

4 Conclusion

We developed CDR-Finder, a user-friendly method to identify CDRs in complete centromeres. The tool analyzes CpG methylation data from both ONT and PacBio HiFi data and outputs the coordinates of the CDR calls and a plot with related read coverage and highlighted CDR windows. In our experience, both PacBio HiFi and ONT data perform similarly with this tool (Mastrorosa *et al.* 2024). However, the user can modify the parameters based on specific cases and quality of the data available. We provide an extensive explanation of the parameters with test cases and common issues on the CDR-Finder GitHub page (https://github.com/EichlerLab/CDR-Finder).

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc. The other authors declare no competing interests.

Funding

This research was supported, in part, by funding from the National Institutes of Health (NIH) [R01HG010169 to E.E. E. and R00GM147352 to G.A.L.]. E.E.E. is an investigator of the Howard Hughes Medical Institute. This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

Data availability

The T2T-CHM13 v2.0 genome assembly, PacBio HiFi data, and ONT data are available at: https://github.com/marbl/CHM13. The HGSVC genome assemblies, PacBio HiFi data, and ONT data are available at: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release. The ONT data used in our tests were basecalled with Guppy v6.3.7, which detects methylated cytosines during basecalling, and were aligned to the T2T-CHM13 v2.0 reference genome using winnowmap2 (Jain *et al.* 2022). The following command was used for ONT read alignment and processing: `winnowmap -W CHM13_repetitive_k15.txt -y—eqx -ax map-ont -s 4000 -t {threads} -I 10g {ref.fasta} {reads.fastq} | samtools view -u -F 2308 - | samtools sort -o {output.bam} -`.

References

Altemose N, Logsdon GA, Bzikadze AV *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* 2022;**376**:eabl4178. https://doi.org/10.1126/science.abl4178

Cheng H, Concepcion GT, Feng X *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021;**18**:170–5. https://doi.org/10.1038/s41592-020-01056-5

Gershman A, Sauria MEG, Guitart X *et al.* Epigenetic patterns in a complete human genome. *Science* 2022;**376**:eabj5089. https://doi.org/10.1126/science.abj5089

Jain C, Rhie A, Hansen NF *et al.* Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods* 2022;**19**:705–10. https://doi.org/10.1038/s41592-022-01457-8

Kirmse A, de Ferranti J. Calculating the prominence and isolation of every mountain in the world. *Progress Phys Geography Earth Environ* 2017;**41**:788–802. https://doi.org/10.1177/0309133317738163

Logsdon GA, Ebert P, Audano PA *et al.* Complex genetic variation in nearly complete human genomes. bioRxiv, https://doi.org/10.1101/2024.09.24.614721, 2024a, preprint: not peer reviewed.

Logsdon GA, Eichler EE. The dynamic structure and rapid evolution of human centromeric satellite DNA. *Genes (Basel)* 2022;**14**:92. https://doi.org/10.3390/genes14010092

Logsdon GA, Rozanski AN, Ryabov F *et al.* The variation and evolution of complete human centromeres. *Nature* 2024b;**629**:136–45. https://doi.org/10.1038/s41586-024-07278-3

Logsdon GA, Vollger MR, Hsieh P *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* 2021;**593**:101–7. https://doi.org/10.1038/s41586-021-03420-7

Mastrorosa F, Kumara AN, Rozanski WT *et al.* Complete chromosome 21 centromere sequences from a down syndrome family reveal size asymmetry and differences in kinetochore attachment. bioRxiv, https://doi.org/10.1101/2024.02.25.581464, 2024, preprint: not peer reviewed.

Miga KH, Alexandrov IA. Variation and evolution of human centromeres: a field guide and perspective. *Annu Rev Genet* 2021;**55**:583–602. https://doi.org/10.1146/annurev-genet-071719-020519

Miga KH, Koren S, Rhie A *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 2020;**585**:79–84. https://doi.org/10.1038/s41586-020-2547-7

Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53. https://doi.org/10.1126/science.abj6987

Rautiainen M, Nurk S, Walenz BP *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnol* 2023;**41**:1474–82. https://doi.org/10.1038/s41587-023-01662-6

Rhie A, Nurk S, Cechova M *et al.* The complete sequence of a human Y chromosome. *Nature* 2023;**621**:344–54. https://doi.org/10.1038/s41586-023-06457-y

Smit AFA, Hubley R, Green P. *RepeatMasker Open-4.0*. 2013–2015. http://www.repeatmasker.org

Vollger MR, Guitart X, Dishuck PC *et al.* Segmental duplications and their variation in a complete human genome. *Science* 2022;**376**:eabj6965. https://doi.org/10.1126/science.abj6965

Bioinformatics, 2024, 40, 1–3
https://doi.org/10.1093/bioinformatics/btae733
Original Paper