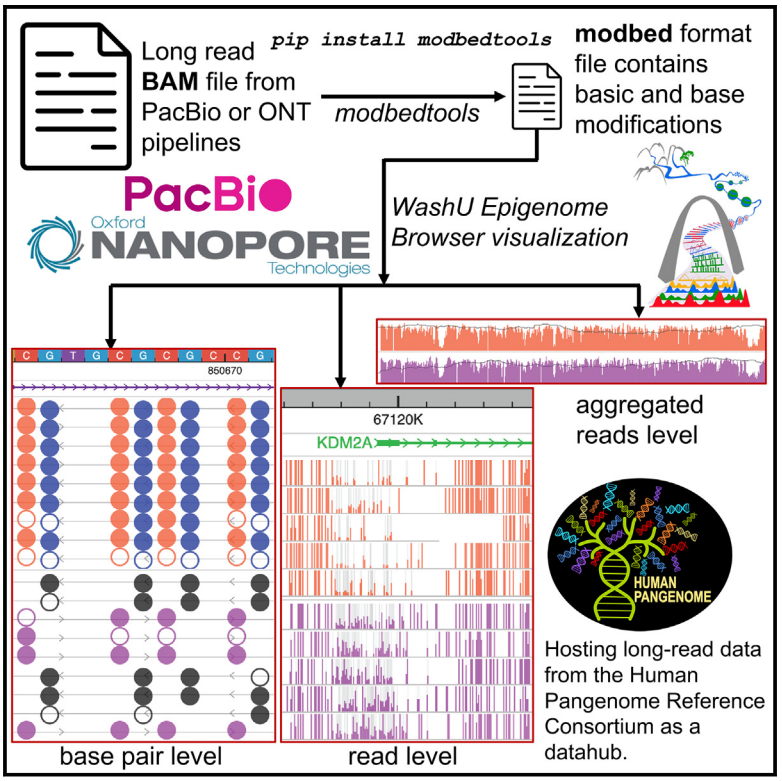
Cell Genomics

Modbed track: Visualization of modified bases in single-molecule sequencing

Graphical abstract



Highlights

- Reduced file size for long-read DNA modification data using modbed vs. BAM
- Provided command line tool for converting long-read BAM to modbed
- Automatic methylation view style switching based on the size of the viewing region
- Long-read methylation data hub for the Human Pangenome Reference Consortium

Authors

Daofeng Li, Xiaoyu Zhuo, Jessica K. Harrison, Shane Liu, Ting Wang

Correspondence

dli23@wustl.edu (D.L.), twang@wustl.edu (T.W.)

In brief

The new modbed track type introduced by the WashU Epigenome Browser provides visualization of base modification (DNA methylation) details from single-molecule (long-read) sequencing technologies such as PacBio and Oxford Nanopore, both at single-read level and at aggregated level from multiple reads across a dynamic range of resolutions.

Li et al., 2023, Cell Genomics 3, 100455
December 13, 2023
https://doi.org/10.1016/j.xgen.2023.100455

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

Modbed track: Visualization of modified bases in single-molecule sequencing

Daofeng Li,[1,*] Xiaoyu Zhuo,[1] Jessica K. Harrison,[1] Shane Liu,[1,2] and Ting Wang[1,3,4,*]
[1]Department of Genetics, The Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St. Louis, MO, USA
[2]Computer Science and Engineering Division, University of Michigan, Ann Arbor, MI, USA
[3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
[4]Lead contact
*Correspondence: dli23@wustl.edu (D.L.), twang@wustl.edu (T.W.)
https://doi.org/10.1016/j.xgen.2023.100455

SUMMARY

Recent advances in long-read sequencing technologies have not only dramatically increased sequencing read length but also have improved the accuracy of detecting chemical modifications to the canonical nucleotide bases, thus opening exciting venues to investigate the epigenome. Currently, the ability to visualize modified bases from long-read sequencing data in genome browsers is still limited, preventing users from easily and fully exploring these type of data. To address this limitation, the WashU Epigenome Browser introduces the modbed track type, which provides visualization of modification details in each single read as well as aggregated modifications of individual or multiple molecules across a dynamic range of resolutions. The modbed file can be uploaded for visualization as a local track or viewed with an accessible URL freely on the WashU Epigenome Browser at https://epigenomegateway.wustl.edu/.

INTRODUCTION

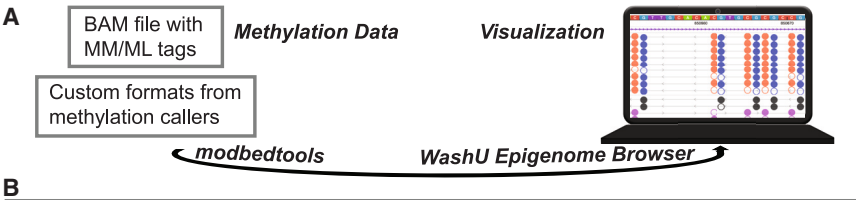
Third-generation long-read sequencing or single-molecule real-time sequencing technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have made rapid advancements in terms of read length, throughput, yield, and analysis methods.[1,2] These technologies directly sequence a single molecule of polynucleotides in real time and produce considerably longer reads than those from second-generation sequencing platforms. Long-read sequencing has impacted all major areas of genomics. For example, longer reads enable better genome assembly. The Human Pangenome Reference Consortium now routinely uses long-read sequencing to build reference-quality genomes of individuals to better represent human diversity.[3–5]

In addition to generating longer reads, single-molecule-based sequencing can also detect chemical modifications of the canonical bases. For example, PacBio can detect 5-methylcytosine (5mC) modifications without bisulfite treatment by analyzing the kinetic signatures composed of pulse width and inter-pulse duration.[6] ONT sequencing quantifies the readout of electrolytic signals that are sensitive to base modifications and thus can be used to detect chemical modifications such as 5mC.[7] Investigators took advantage of these new capabilities to invent technologies that convert chromatin biological signals into base modifications.[8–10] For example, Fiber-seq labels open chromatin with N^6-adenine (6mA), thus enabling the study of regulatory DNA and nucleosome positioning at single chromatin fiber resolution. These technologies generate profiles of modifications to canonical DNA bases in multi-kilobase chromatin molecules. A few tools allow visualization of such data, for example, desktop software IGV,[11] command line tool modbamtools,[12] and the web-based genome browser Jbrowse2.[13] They usually display single sequencing reads in a heatmap with each row representing chemical modifications from a read and use red/blue color in columns to indicate methylation status. Information such as methylation percentage and statistics of modified/canonical bases over each read are missing in such representations.

WashU Epigenome Browser is a web-based genomic data exploration tool.[14–16] A major development goal of the browser is to support efficient, versatile, and more expressive visualization of newly evolved genomic approaches and data types, e.g., Hi-C,[17] whole-genome bisulfite sequencing,[18] and 3D chromatin modeling.[19] Here, we present the "modbed" track for visualizing base modifications in single-molecule long-read sequencing data. This new track type provides visualization of modification details for each molecule, as well as aggregated modifications of individual or multiple molecules across a dynamic range of resolutions.

DESIGN

We extended the common browser extended data (BED) format to create the modbed format to describe long-read modification data (Figure 1). Long-read methylation data generated from PacBio and ONT analysis pipelines (BAM files) can be directly converted to the modbed format using the associated command line tool "modbedtools" (supplemental information, methods S1) (Figure 1A) and



B

```
chr11 5173273 5195306 id score + -110,-266,-1459,... -1096,-1543,...
chr11 5174507 5194585 id score - 223,605,607,613,... 123,307,544,...
chr11 5174543 5196481 id score + 187,271,508,570,... 568,656,664,...
```

columns: ***1***,*chromosome;* ***2-3***, *start and end position of an aligned long read;* ***4***, *read id ;* ***5***, *a scorecan be used to sort read vertically;* ***6***, *strand;* ***7-8***, *modified and canonical base positions, relative to start.* ***-*** *indicates base on reverse strand.*

Figure 1. Overview and data format of the modbed track

(A) Methylation data in BAM format with MM/ML tags or custom formats from different methylation callers can be converted to modbed format using the modbedtools software and be visualized in the WashU Epigenome Browser.

(B) DNA modification profile data are encoded in a BED-like format. Each row represents a molecule or chromatin fiber mapped to the reference genome. The first 3 columns record chromosomal location. The 4th column contains the read ID or any identifier for that read. The 5th column is a score that can be used to sort reads vertically in the browser in ascending or descending order. The 6th column is the strand the read mapped to the genome. The last 2 columns indicate methylated/modified and unmethylated/unmodified/canonical base relative to the start position (2nd column) of the read, separated by commas, respectively.

visualized at the WashU Epigenome Browser (Figure 1B). This format significantly reduces track file size by 57-fold on average, compared with the original long-read SAM or BAM files (Table S1). This approach preserves essential information of base modification for visualization while reducing data transfer, thus improving the performance of web-based genome browsers to fetch and process data. The modbed file is compressed and indexed using Tabix[20] to allow fast random retrieval of data from any genomic region. The modbed file contains information about both modified and canonical bases for calculating modification frequency (such as methylation level) when data need to be aggregated.

RESULTS

Multiple modes of visualizations are provided to the users as a function of the size of the displayed genomic region or based on user settings. We used the modbed track to display both long-read methylation data from ONT and PacBio platforms (Figures 2, 3, and 4), data from methylation caller software such as NanoMethPhase[21] (Figure 5A), and Fiber-seq data (Figure 5B). The methylation status of each long read can be visualized as a heatmap in a similar style as what IGV and Jbrowse2 provide, where default or user-defined colors indicate methylated or unmethylated status (Figures 3B and 5). When users zoom out to view a larger region where a single pixel on the computer screen spans multiple base pairs in the genome, the browser will calculate and display an aggregated signal for each pixel using the number of modified bases over the sum of modified and canonical bases in the read represented by that pixel. This aggregated signal can be displayed as a bar graph where the height of each bar represents the modification frequency. The shade of gray for the bar graph background indicates the base (for example, C for 5mC or A for 6mA) density of each pixel window (Figure 3). At the highest base-pair resolution, the modification status of each DNA base on a single molecule is displayed as filled (for methylated) or open (for unmethylated) circles (Figure 4), whereas the strand information is indicated by different colors. In addition to displaying stacks of single-read or chromatin fiber, information across many reads mapped to the same genomic region can also be aggregated and summarized into a view that is reminiscent of a typical wiggle track (Figure 2). Here, methylation from ONT and PacBio is shown as a bar plot over the CpG context, where each pixel displays the frequency of modification across all reads, whereas the gray line in the background indicates read coverage. The light-green highlighting box indicates the promoter region of *KDM2A*, which shows low methylation of aggregated signal from all reads as bar plot (Figure 2), as heatmap (Figure 3B), and as bar plot over each long read (Figure 3A).

DISCUSSION

Overview

The modbed track is the latest addition to the WashU Epigenome Browser for visualizing long-read/single-molecule sequencing data. It provides novel and customizable visualizations to facilitate the investigation of epigenetic information encoded in the long

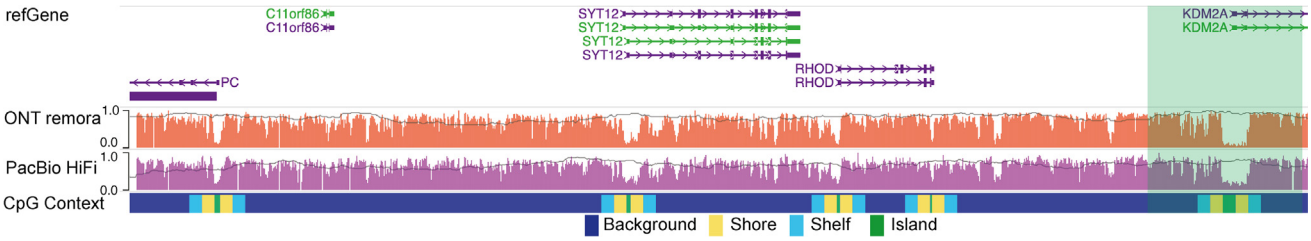


Figure 2. Oxford Nanopore and PacBio methylation data are displayed as modbed tracks

Aggregated signal of multiple long reads when zoomed out to a larger region. From top to bottom are the refGene track and two DNA methylation data tracks from Nanopore and PacBio data generated by the Human Pangenome Reference consortium. The height of each bar indicates methylation percentage where 1 means fully methylated and 0 means fully unmethylated. The gray line represents the reads coverage. The bottom track is a CpG context track showing the locations of CpG islands, shelves, or shores. The light-green box indicates the highlighted region shown in Figure 3.

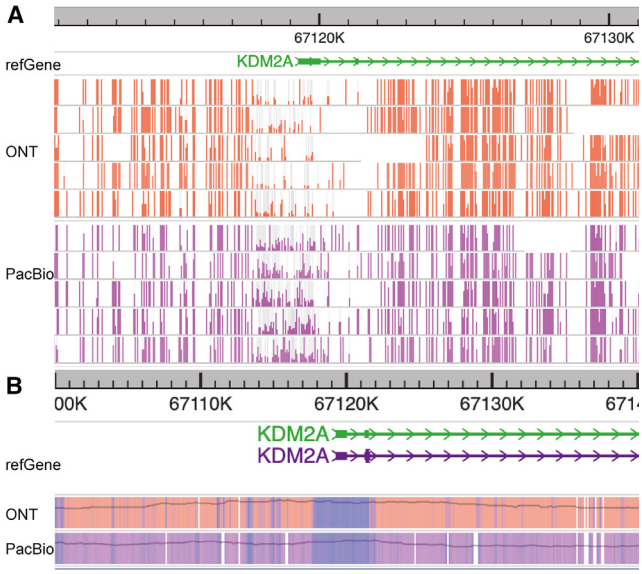


Figure 3. Zoomed-in view of the same long-read methylation data in the region highlighted in the light-green box in Figure 2

(A) Zooming into the promoter region of *KDM2A* gene, displaying methylation status over each long read in the bar plot. Each segment is a long read showing aggregated signal in each (computer) screen pixel, where the height of each bar correlates with modification frequency/methylation level, and gray shade of the background of each bar indicates the base (for example, C for 5mC or A for 6mA) density.

(B) Same summary view while switching to heatmap style, focusing on the promoter region of *KDM2A* gene. Blue indicates low methylation, orange/purple for high methylation, and gray line indicates read density.

reads. The recent introduction of "MM" and "ML" tags to the format specification of SAM format[22] standardized the representation of modified bases. Converting a standardized BAM file to modbed format file preserves information about mapping and base modification while significantly reducing file size, making it faster for transferring and parsing and improving performance in a web-based genome browser platform and user experience. The modbed track has a flexible and expressive design, and we anticipate its wide adoption in the community to enhance investigations using long-read-based DNA methylation data as well as data generated by a wide range of new technologies that rely on the detection of modified bases by long-read sequencing.

Limitations

Although file sizes of modbed track files are reduced significantly to reduce web transfer of data, the alignment details of each long read, including sequence variations, are no longer preserved. Users need to consult the original BAM files for alignment information. Another limitation is that for genome-wide background methylation signal data, for example, based on the Fiber-seq data given only methylated adenine bases as input, iterating over the whole genome's bases to calculate background adenine bases can be slow; thus multi-thread support is required for the "addbg" module.

Future directions

Although converting long-read BAM files to modbed files with modbedtools is a heavy I/O (input/output) process, it might still benefit from multi-threading programming. We are considering adding multi-threading support for the future versions of modbedtools to increase the speed of generating modbed files for both "bam2mod" and "addbg" modules. Additional improvements include supporting multiple types of modification simultaneously (e.g., DNA methylation and hydroxymethylation together) and preserving haplotype information of long reads.

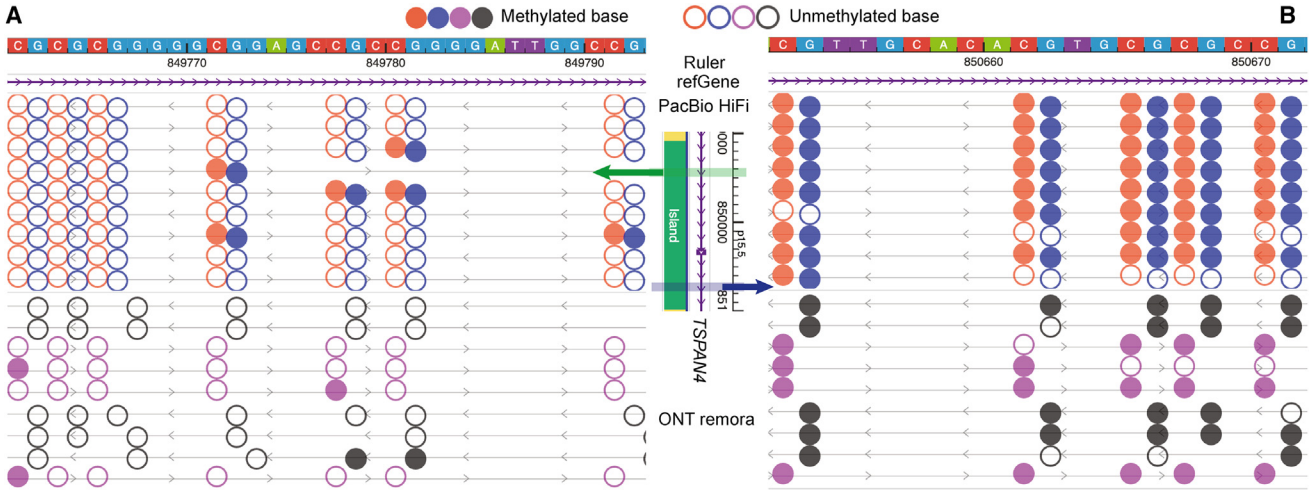


Figure 4. Base-pair-level visualization of DNA modifications from PacBio HiFi and ONT remora methylation data

Shown are data in a lowly methylated region (A) and in a highly methylated region (B). The top two tracks are ruler and refGene tracks. A filled or open circle indicates a methylated or unmethylated base; the circle color represents the strand (forward strand = orange/purple; reverse strand = blue/black for PacBio/ONT). Note that PacBio CpG methylation calls of circular consensus sequencing (CCS) reads represent the predicted methylation status of the CpG site as a unit. To faithfully present it, we plotted the methylation prediction of CCS reads on the C base of both strands at each CpG site by enabling the specific option in the associated modbedtools command line tool.

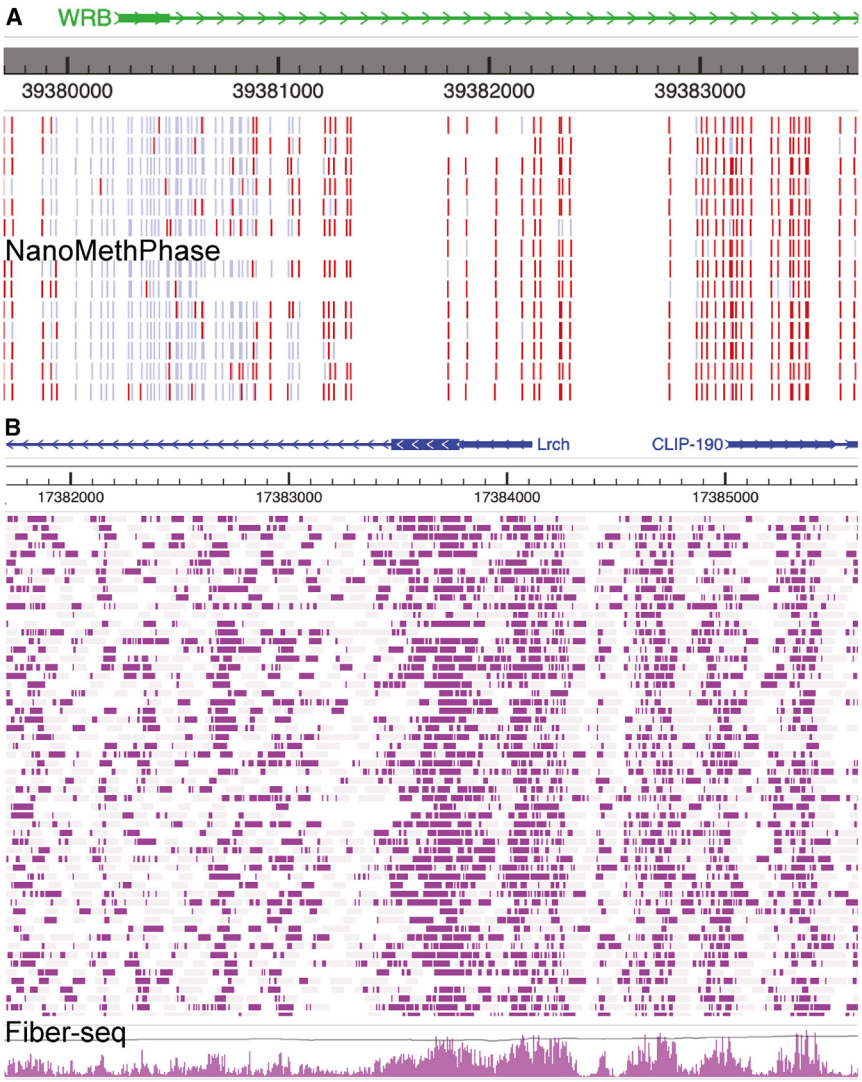


Figure 5. Visualizing other long-read assay data using modbed track

(A) Visualizing NanoMethPhase[21] data as a modbed track. Methylation signals from each read are shown in heatmap style, where red means methylated, and blue means unmethylated.

(B) Fiber-seq data[23] are displayed as modbed tracks. From top to bottom are the refGene track, ruler track, and a heatmap style of Fiber-seq data from the GEO database: GSM4411218, where each row represents a fiber read, and purple color means methylated adenine bases. Below the heatmap, the same data are displayed as a wiggle-style track, where the height of each bar represents the adenine methylation level. URLs for the visualizations of each panel or part of each panel can be found in Methods S1 of the supplemental information file.

We will also keep updating our long-read methylation datahub collection as the Human Pangenome Reference Consortium releases more single-molecule long-read sequencing data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100455.

ACKNOWLEDGMENTS

We thank the Human Pangenome Reference Consortium for providing the free download of PacBio and Nanopore methylation data. We thank Eric Haugen and John A. Stamatoyannopoulos for helpful discussions on Fiber-seq data visualization. This work was supported by NIH (R01HG007175, U01CA200060, U01HG009391, UM1HG011585, U41HG010972, U24HG012070, UM1DA058219, and U24NS132103).

AUTHOR CONTRIBUTIONS

T.W. conceived and oversaw all aspects of the study, supervised research, and provided constructive feedback. D.L. developed the visualization module in the browser. D.L. and X.Z. developed the modbedtools package. J.K.H. and S.L. provided key contributions and useful feedback. The manuscript was written by D.L. and T.W. with additional contributions and edits from X.Z. and J.K.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: August 7, 2023
Revised: September 28, 2023
Accepted: November 4, 2023
Published: December 6, 2023

REFERENCES

1. Lucas, M.C., and Novoa, E.M. (2023). Long-read sequencing in the era of epigenomics and epitranscriptomics. Nat. Methods *20*, 25–29.
2. Kovaka, S., Ou, S., Jenike, K.M., and Schatz, M.C. (2023). Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. Nat. Methods *20*, 12–16.
3. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446.
4. Jarvis, E.D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F., Vollger, M.R., Porubsky, D., et al. (2022). Semi-automated assembly of high-quality diploid human reference genomes. Nature *611*, 519–531.
5. Liao, W.W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312–324.
6. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods *7*, 461–465.
7. Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods *14*, 407–410.
8. Abdulhay, N.J., McNally, C.P., Hsieh, L.J., Kasinathan, S., Keith, A., Estes, L.S., Karimzadeh, M., Underwood, J.G., Goodarzi, H., Narlikar, G.J., and Ramani, V. (2020). Massively multiplex single-molecule oligonucleosome footprinting. Elife *9*, e59404.
9. Lee, I., Razaghi, R., Gilpatrick, T., Molnar, M., Gershman, A., Sadowski, N., Sedlazeck, F.J., Hansen, K.D., Simpson, J.T., and Timp, W. (2020). Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat. Methods *17*, 1191–1199.
10. Wang, Y., Wang, A., Liu, Z., Thurman, A.L., Powers, L.S., Zou, M., Zhao, Y., Hefel, A., Li, Y., Zabner, J., and Au, K.F. (2019). Single-molecule long-read sequencing reveals the chromatin basis of gene expression. Genome Res. *29*, 1329–1342.
11. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.
12. Razaghi, R., Hook, P.W., Ou, S., Schatz, M.C., Hansen, K.D., Jain, M., and Timp, W. (2022). Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering. Preprint at bioRxiv.
13. Diesh, C., Stevens, G.J., Xie, P., De Jesus Martinez, T., Hershberg, E.A., Leung, A., Guo, E., Dider, S., Zhang, J., Bridge, C., et al. (2023). JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome Biol. *24*, 74.
14. Li, D., Hsu, S., Purushotham, D., Sears, R.L., and Wang, T. (2019). WashU Epigenome Browser update 2019. Nucleic Acids Res. *47*, W158–W165.
15. Li, D., Purushotham, D., Harrison, J.K., Hsu, S., Zhuo, X., Fan, C., Liu, S., Xu, V., Chen, S., Xu, J., et al. (2022). WashU Epigenome Browser update 2022. Nucleic Acids Res. *50*, W774–W781.
16. Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E.A., Koebbe, B.C., Nielsen, C., Hirst, M., Farnham, P., et al. (2011). The Human Epigenome Browser at Washington University. Nat. Methods *8*, 989–990.
17. Zhou, X., Lowdon, R.F., Li, D., Lawson, H.A., Madden, P.A.F., Costello, J.F., and Wang, T. (2013). Exploring long-range genome interactions using the WashU Epigenome Browser. Nat. Methods *10*, 375–376.
18. Zhou, X., Li, D., Lowdon, R.F., Costello, J.F., and Wang, T. (2014). methylC Track: visual integration of single-base resolution DNA methylation data on the WashU EpiGenome Browser. Bioinformatics *30*, 2206–2207.
19. Li, D., Harrison, J.K., Purushotham, D., and Wang, T. (2022). Exploring genomic data coupled with 3D chromatin structures using the WashU Epigenome Browser. Nat. Methods *19*, 909–910.
20. Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics *27*, 718–719.
21. Akbari, V., Garant, J.M., O'Neill, K., Pandoh, P., Moore, R., Marra, M.A., Hirst, M., and Jones, S.J.M. (2021). Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. Genome Biol. *22*, 68.
22. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
23. Stergachis, A.B., Debo, B.M., Haugen, E., Churchman, L.S., and Stamatoyannopoulos, J.A. (2020). Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science *368*, 1449–1454.
24. Li, D., and Wang, T. (2023). Example modbed files can be used for visualization on the WashU Epigenome Browser. Zenodo. https://doi.org/10.5281/zenodo.10035814.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Example modbed files	This paper	Zenodo: 10.5281/zenodo.10035814
One-click URLs for visualization	This paper	Zenodo: 10.5281/zenodo.10035814
Long read methylation data	Human Pangenome Reference Consortium[3]	https://humanpangenome.org/
Software and algorithms		
Modbedtools	This paper	https://github.com/lidaof/modbedtools
WashU Epigenome Browser	Li et al.[14,15]	https://epigenomegateway.wustl.edu/browser/
Tabix	Li et al.[20]	https://www.htslib.org/doc/tabix.html
Other		
Fruit fly Fiber-seq data	NCBI GEO Database[23]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4411218
NanoMethPhase example data	Akbari et al.[21]	https://github.com/vahidAK/NanoMethPhase/tree/master/Example_Data

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and code should be directed to and will be fulfilled by the Lead Contact, Ting Wang (twang@wustl.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The modbed track is one of the supported visualization data types for the WashU Epigenome Browser, which can be freely accessed from https://epigenomegateway.wustl.edu/. The instructions and documentations for the modbed track can be found at https://github.com/lidaof/modbedtools. Example modbed files that can be used for visualization on the WashU Epigenome Browser have been deposited in Zenodo.[24]