Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering.

Roham Razaghi[1], Paul W. Hook[1], Shujun Ou[2], Michael C. Schatz[2], Kasper D. Hansen[3], Miten Jain[4], Winston Timp[1,*]

1. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA
2. Department of Biology and Computer Science, Johns Hopkins University, Baltimore, Maryland, USA
3. Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD, USA
4. Department of Bioengineering, Department of Physics, Northeastern University, Boston, MA

*Corresponding author: Winston Timp wtimp@jhu.edu

Abstract

The advent of long-read sequencing methods provides new opportunities for profiling the epigenome - especially as the methylation signature comes for “free” when native DNA is sequenced on either Oxford Nanopore or Pacific Biosciences instruments. However, we lack tools to visualize and analyze data generated from these new sources. Recent efforts from the GA4GH consortium have standardized methods to encode modification location and probabilities in the BAM format. Leveraging this standard format, we developed a technology-agnostic tool, modbamtools to visualize, manipulate and compare base modification/methylation data in a fast and robust way. modbamtools can produce high quality, interactive, and publication-ready visualizations as well as provide modules for downstream analysis of base modifications. Modbamtools comprehensive manual and tutorial can be found at https://rrazaghi.github.io/modbamtools/.

Introduction

Direct single-molecule sequencing methods, e.g. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have recently greatly expanded in throughput and yield. In addition to the canonical base sequencing data that these platforms generate, modifications on the nucleic acids can be measured directly, either via delays in the incorporation of bases (IPD, PacBio (Flusberg et al. 2010)) or perturbations in the electrical current (ONT (Simpson et al. 2017)). These have been accompanied by development of software tools to measure and call modifications within this data, but the output formats of these calls were not standardized precluding easy downstream development. Modification data files have typically been stored as enormous (terabyte scale) tsv/csvs and early efforts to incorporate 5-methylcytosine information from ONT into a “bisulfite-like” BAM file required complex manipulations(Lee et al. 2020).

More recently, the Global Alliance for Genomics and Health (GA4GH) (Rehm et al. 2021) standards group proposed an addition to the BAM file spec, incorporating two new tags (MM and ML) for SAM/BAM alignment files. The MM tag is used to locate the strand and position the modification was observed on, and the ML tag is the probability of each modification being present (http://samtools.github.io/hts-specs). Although these tags were introduced as an adaptation to long-read base modification data, it is anticipated that all technologies will eventually incorporate this file format.

Single molecule base modification callers have rapidly adapted to the new standard format. Currently, for nanopore data, most modification calling tools can output BAM files with tags, including guppy, bonito, Megalodon, and nanopolish (Simpson et al. 2017). Similarly, Primrose, and ccsmeth (Ni et al. 2022) can be used for PacBio reads. An updated list of compatible tools generating these alignment files can be found at https://rrazaghi.github.io/modbamtools/.

Here we introduce modbamtools, a suite of tools to explore modifications in single-molecule data using this new format. With this tool we generate interactive and batch visualization and analysis for methylation frequency and single-molecule methylation. Profiling methylation across individual molecules, we can look at coordination of long-range methylation effects, e.g. enhancer-promoter interactions, and the degree of variation of methylation “noise” within regions. We have also generated modules to phase reads by using genetic variation or through methylation alone via a read clustering approach, to enable exploration of allele-specific methylation and epigenetic heterogeneity.

Results

Usage and Examples

We developed modbamtools, a software package that provides analysis and interactive visualization of single-read base modification data along with other highly used formats for genomic tracks (GTF, bigwig, bedgraph, etc). Modbamtools utilizes core python modules including numpy (van der Walt et al. 2011), pandas (McKinney and Others 2011), scikit-learn (Pedregosa et al. 2011), pysam (Heger et al. 2014), click, plotly (Plotly Technologies Inc., 2015), modbampy, pybigwig (Ryan et al. 2016), pypdf2, pillow, and hdbscan (McInnes et al. 2017). We have made modbamtools easily accessible through PyPI (`pip install modbamtools`).

The tool has three main elements (`calcMeth`, `calcHet`, `cluster`) and a plotting function that allows for interactive plotting of single-read base modification data. This generates a multi-panel plot **(Figure 1)** consisting of an annotation track, methylation frequency track, and single-read plots. The annotation track can display other sets of genomics data including gene models, other epigenetic data (e.g. ENCODE ChIP-seq), and genetic variation. Methylation frequencies along with a smoothed average frequency is plotted on top of the reads similar to a conventional genome browser. The methylation frequency plot shows the per locus frequency of modified to total called bases. Finally, the single-read plots represent each individual single molecule with base modifications indicated as blue for unmodified and red for modified. These figures can be

output as HTML, PDF, PNG, or SVG. The HTML provided is generated with plotly and is interactive, allowing magnification. Multiple plots can be output in batch mode by providing a BED file of regions of interest resulting in a multiple page HTML or PDF report.

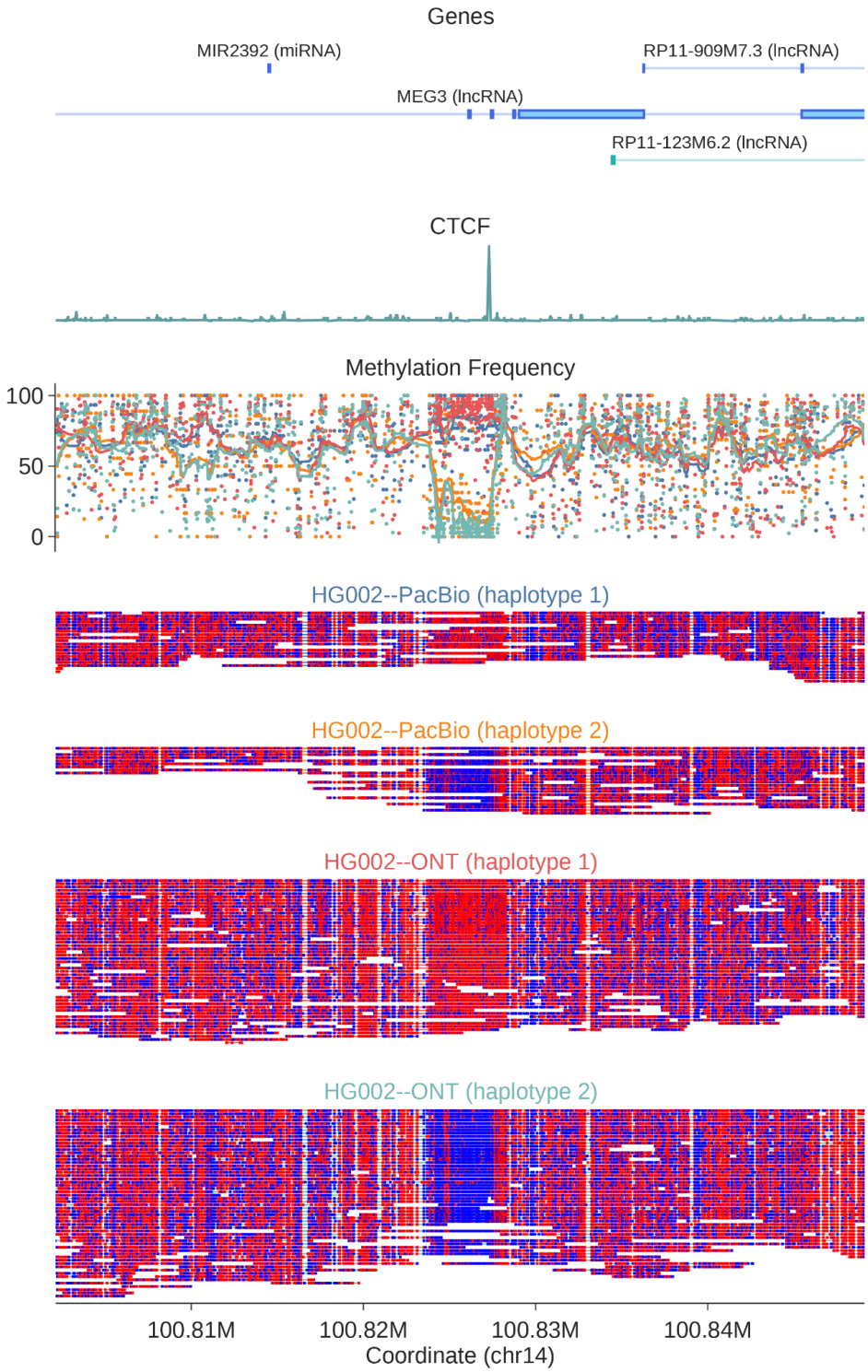

***Figure 1**: Example of modbamtools output on MEG3 (chr14:100,802,132-100,849,111) locus using both PacBio and ONT single-molecule data from the HG002 Genome in a Bottle cell line. "Genes" track shows GENCODE (Release 38, GRCH38) gene models and the "CTCF" track shows CTCF CHiP-seq ENCODE track from GM12878. Methylation frequency track is colored according to platform and haplotype, with colors indicated by the title of the single-molecule plots. In single-molecule plots, each read is a single horizontal bar, with methylated bases shown as red and unmethylated as blue.*

Using appropriate tools, e.g. clair (Zheng et al. 2021) or whatshap (Martin et al. 2016), BAM files can have the haplotypes of reads encoded with the commonly used "HP" tag. Our tool has the ability to group the alignments based on phase tag (HP) in BAM files. Using this HP tag, we can separate reads according to haplotype, plotting each haplotype's methylation frequency as different colored lines and the single reads as separate plot elements. We show an example of this module on methylation calls from the HG002 cell line at the *MEG3* long noncoding RNA (lncRNA), using public single-molecule methylation data from both ONT and PacBio platforms (**Figure 1**). *MEG3* has known monoallelic expression in many tissues and loss of this regulation has been implicated in development of type 2 diabetes mellitus (Rosa et al. 2005; Kameswaran et al. 2014). From this data, we observe clear examples of allele-specific methylation at a CTCF binding site and *MEG3* promoter region.

Beyond clustering according to genomic haplotype, we have implemented a method to cluster single-molecule reads based on methylation status alone using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes et al. 2017). This is a useful feature for regions without many SNPs for phasing reads into haplotypes (Gershman et al. 2022). Clustering can also be used to quantify different cell types or to profile early cancer detection from a heterogeneous sample (Wang et al. 2021; Houseman et al. 2008; Gkountela et al. 2019; Tian et al. 2020). Clustering can be performed either as a part of the plotting command or separately (`--cluster` command) with the input of a batch file for locations used for the clustering. As shown in **Figure 2**, we can cluster the *SNURF* gene promoter based purely on methylation signal at this locus. This paternally imprinted locus can also be phased based on genotyping information (**Supplementary Figure 1**), demonstrating the agreement of our clustering approach with classical methods.

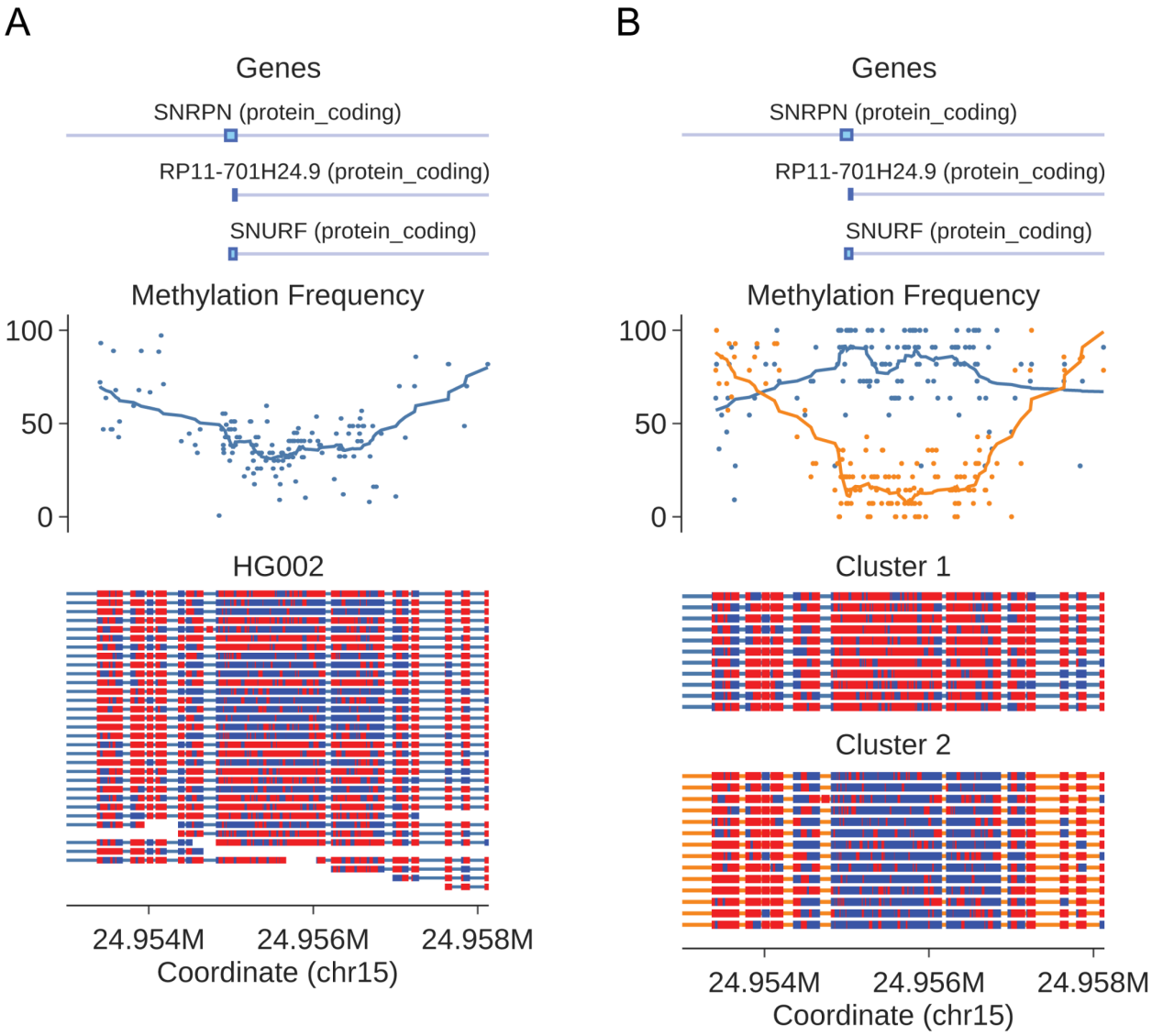

***Figure 2**: A) single molecule methylation profile on gene SNRPN (chr15:24,953,000-24,958,133) from HG002 data as in Figure 1. B) Single molecule methylation profile on gene SNRPN separated into clusters with 'modbamtools plot –cluster'*

Finally, using a BED file of genomic loci, we can profile the average methylation in each location, including methylation on each haplotype. The "calcMeth" module calculates methylation average across each single molecule *first* then aggregates over all molecules which map to that region, rather than averaging CpG methylation per CpG then averaging across the region. This is especially useful with long reads to capture methylation variability more efficiently (**Figure 3**).

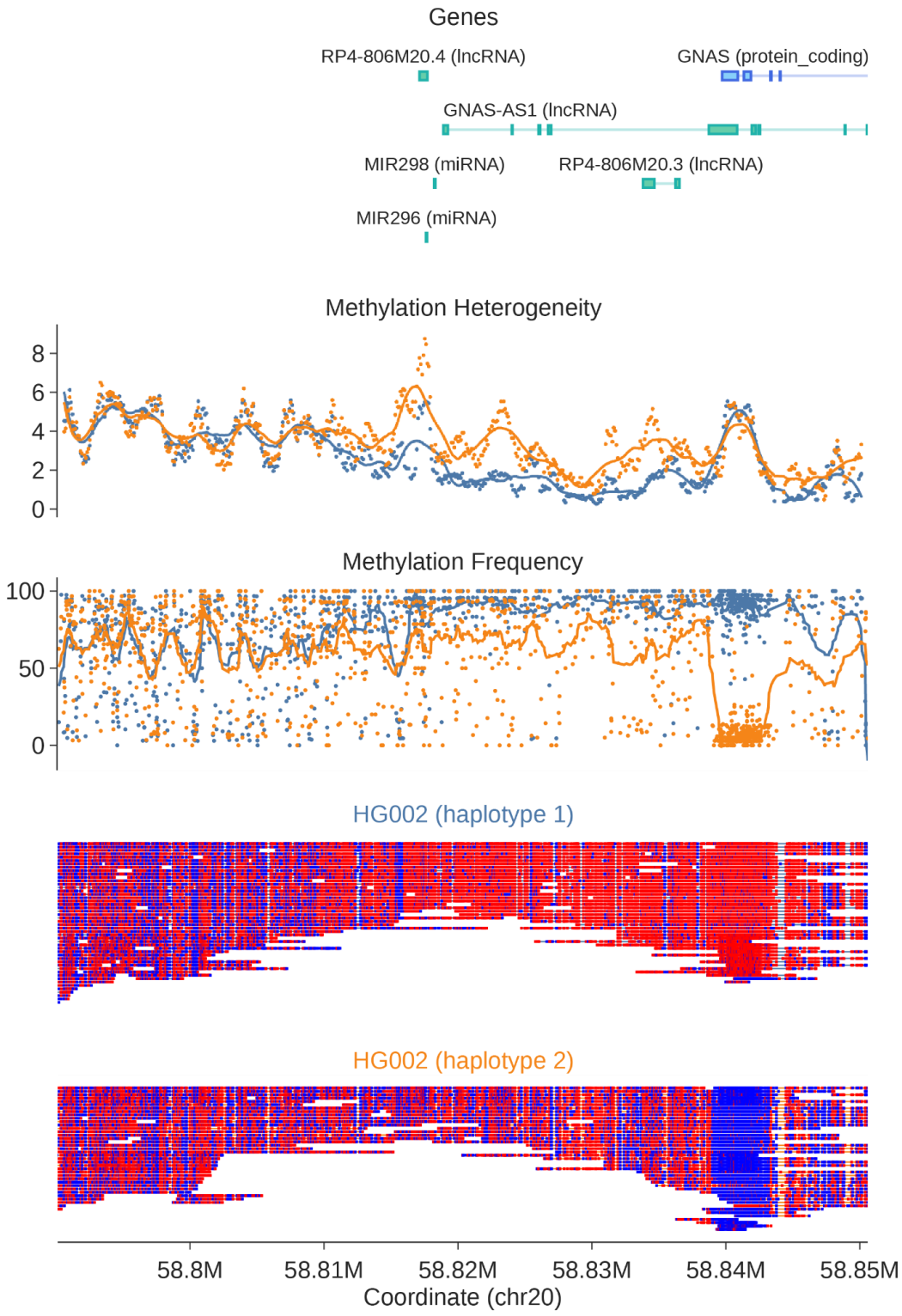

***Figure 3**: Example of modbamtools plot with options for haplotype separation and calculating heterogeneity at the GNAS locus (chr20:58,790,127-58,850,596).*

With single-molecule methylation data, can quantify not only methylation frequency averaged across all reads, but also variability of methylation across individual molecules. A few studies have attempted to address this by proposing different algorithms to quantify this feature (Scherer et al. 2020; Landau et al. 2014; Guo et al. 2017; Landan et al. 2012; Xie et al. 2011). Here, we implemented a module to calculate methylation heterogeneity ("calcHet") that calculates this on genomic regions provided by the user (See **Supplementary Note 1** for detailed methods). Similar to the clustering function, "–heterogeneity" option can be used with plotting command to visualize this; we have plotted it for the *GNAS* locus in **Figure 3**. There we observe areas of clear difference in methylation heterogeneity across the region, suggesting not only a change in methylation but a less ordered epigenetic state on one allele when compared to the other.

Conclusion

Advances in single-molecule sequencing throughput suggest we are at an inflection point where large scale data sets are on the horizon. These data types offer the unique advantage of providing DNA methylation data *as well as* primary sequence - but without tools to take advantage of it, these data will be "left on the table" and not used to their potential. Here we have described a toolset to take advantage of these data, using the newly described modification tags present in the SAM/BAM file specifications. This toolset is compatible with all modern modification callers. Modbamtools provides fast, robust, interactive visualization and analysis for alignment files containing base modification tags.

Acknowledgments

We would like to thank Jared Simpson and Chris Wright for their helpful comments and contributions to the development of modified base alignment files. W.T. has two patents (8,748,091 and 8,394,584) licensed to ONT.

Funding

This study was supported by National Human Genome Research Institute (project no. 5R01HG009190) and National Cancer Institute (project no. 1U01CA253481-01A1)

Data Availability

Publicly available data on cell line HG002 was downloaded from s3://ont-open-data/gm24385_mod_2021.09/extra_analysis/bonito_remora (ONT HG002 WGS) and https://downloads.pacbcloud.com/public/dataset/HG002-CpG-methylation-202202/ (PacBio HG002 WGS). CTCF track was downloaded from https://www.encodeproject.org/experiments/ENCSR000DZN/. GENCODE Release 38 for GRCH38 was used for gene model tracks.

Code Availability and implementation:

modbamtools source code is available at https://github.com/rrazaghi/modbamtools. A manual and tutorial are available at https://rrazaghi.github.io/modbamtools/.

References

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**: 461–465.

Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S, et al. 2022. Epigenetic patterns in a complete human genome. *Science* **376**: eabj5089.

Gkountela S, Castro-Giner F, Szczerba BM, Vetter M, Landin J, Scherrer R, Krol I, Scheidmann MC, Beisel C, Stirnimann CU, et al. 2019. Circulating Tumor Cell Clustering Shapes DNA Methylation to Enable Metastasis Seeding. *Cell* **176**: 98–112.e14.

Guo S, Diep D, Plongthongkum N, Fung H-L, Zhang K, Zhang K. 2017. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet* **49**: 635–642.

Heger A, Belgrad TG, Goodson M, Jacobs K. 2014. pysam: Python interface for the SAM/BAM sequence alignment and mapping format.

Houseman EA, Christensen BC, Yeh R-F, Marsit CJ, Karagas MR, Wrensch M, Nelson HH, Wiemels J, Zheng S, Wiencke JK, et al. 2008. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9**: 365.

Kameswaran V, Bramswig NC, McKenna LB, Penn M, Schug J, Hand NJ, Chen Y, Choi I, Vourekas A, Won K-J, et al. 2014. Epigenetic regulation of the DLK1-MEG3 microRNA cluster in human type 2 diabetic islets. *Cell Metab* **19**: 135–145.

Landan G, Cohen NM, Mukamel Z, Bar A, Molchadsky A, Brosh R, Horn-Saban S, Zalcenstein DA, Goldfinger N, Zundelevich A, et al. 2012. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet* **44**: 1207–1214.

Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, Stevenson K, Sougnez C, Wang L, Li S, et al. 2014. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**: 813–825.

Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, Timp W. 2020. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat Methods* **17**: 1191–1199.

Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schöenhuth A, Marschall T. 2016. WhatsHap: fast and accurate read-based phasing. *bioRxiv* 085050. https://www.biorxiv.org/content/10.1101/085050 (Accessed July 2, 2022).

McInnes L, Healy J, Astels S. 2017. hdbscan: Hierarchical density based clustering. *J Open Source Softw* **2**: 205.

McKinney W, Others. 2011. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing* **14**: 1–9.

Ni P, Xu J, Zhong Z, Zhang J, Huang N, Nie F, Luo F, Wang J. 2022. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *bioRxiv* 2022.02.26.482074. https://www.biorxiv.org/content/10.1101/2022.02.26.482074 (Accessed July 7, 2022).

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**: 2825–2830.

Rehm HL, Page AJH, Smith L, Adams JB, Alterovitz G, Babb LJ, Barkley MP, Baudis M, Beauvais MJS, Beck T, et al. 2021. GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genom* **1**. http://dx.doi.org/10.1016/j.xgen.2021.100029.

Rosa AL, Wu Y-Q, Kwabi-Addo B, Coveler KJ, Reid Sutton V, Shaffer LG. 2005. Allele-specific methylation of a functional CTCF binding site upstream of MEG3 in the human imprinted domain of 14q32. *Chromosome Res* **13**: 809–818.

Ryan D, Gruning B, Ramirez F. 2016. pyBigWig 0.2. 4. *Cited on* 3.

Scherer M, Nebel A, Franke A, Walter J, Lengauer T, Bock C, Müller F, List M. 2020. Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res* **48**: e46.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407–410.

Tian Z, Meng L, Long X, Diao T, Hu M, Wang M, Liu M, Wang J. 2020. DNA methylation-based classification and identification of bladder cancer prognosis-associated subgroups. *Cancer Cell Int* **20**: 255.

van der Walt S, Colbert SC, Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* **13**: 22–30.

Wang J, Li J, Chen R, Yue H, Li W, Wu B, Bai Y, Zhu G, Lu X. 2021. DNA methylation-based profiling reveals distinct clusters with survival heterogeneity in high-grade serous ovarian cancer. *Clin Epigenetics* **13**: 190.

Xie H, Wang M, de Andrade A, Bonaldo M de F, Galat V, Arndt K, Rajaram V, Goldman S, Tomita T, Soares MB. 2011. Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res* **39**: 4099–4108.

Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. 2021. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *bioRxiv* 2021.12.29.474431. https://www.biorxiv.org/content/biorxiv/early/2021/12/30/2021.12.29.474431 (Accessed July 2, 2022).